# DATA FUSION STRATEGY TO SUPPORT AGRICULTURAL VULNERABILITY ASSESSMENTS


**IVÁN DARÍO LÓPEZ GÓMEZ**

Doctoral Thesis in Telematics Engineering

Thesis Supervisor:
Juan Carlos Corrales
PhD in Computer Sciences

University of Cauca
School of Electronic and Telecommunications Engineering
Department of Telematics
Research Line in e-@mbiente
Popayán, Colombia, April 2021

IVÁN DARÍO LÓPEZ GÓMEZ

# DATA FUSION STRATEGY TO SUPPORT AGRICULTURAL VULNERABILITY ASSESSMENTS

Thesis submitted to the school of electronic and telecommunications engineering of the University of Cauca for the degree of

PhD. in:
Telematics Engineering

Thesis Supervisor:
Juan Carlos Corrales
PhD. in Computer Sciences

Popayán, Colombia
2021

# Acknowledgments

I would first like to acknowledge *Ministry of Science, Technology and Innovation (Minciencias Colombia)* for PhD scholarship granted to me through *"Convocatoria Doctorados Nacionales 727 de 2015"*, University of Cauca for supporting this research through project ID 4618, and the project *AVA (Agriculture, Vulnerability, and Adaptation)* from the *Climate and Development Knowledge Network (CDKN)* as a conceptual framework for this research. Additionally, this work has also been supported by project *"Alternativas Innovadoras de Agricultura Inteligente para sistemas productivos agrícolas del departamento del Cauca soportado en entornos de IoT - ID 4633"* financed by Convocatoria 04C-2018 *"Banco de Proyectos Conjuntos UEES-Sostenibilidad"* of project *"Red de formación de talento humano para la innovación social y productiva en el Departamento del Cauca InnovAcción Cauca"*.

I would also like to thank my thesis supervisor, Dr. Juan Carlos Corrales, for his guidance through each stage of the process and for inspiring my interest in the development of this work; Dr. Apolinar Figueroa Casas for his advice and valuable contributions around agricultural vulnerability and food security; Dr. Jacques Avelino for his supervision and support during the research internship at the Tropical Agricultural Research and Higher Education Center (CATIE), Turrialba, Costa Rica; the members of the Telematics Engineering Group of the University of Cauca for their continuous feedback, especially Dr. Alvaro Rendón and Dr. Gustavo Ramirez for their support in my doctoral training; and each person who in one way or another participated in the development of this doctoral thesis.

Finally, I would like to thank God for allowing me to face this great challenge with dedication and responsibility in order to culminate one more stage of academic and personal formation. Also, to my parents, family, colleagues, and friends for their unconditional support.

# Structured Abstract

**Background:** Identifying crop species and varieties adaptable to climate change impacts is one of the main aspects of climate vulnerability assessments. This estimation involves processing, integrating, and analyzing many information sources to provide accurate and timely responses. However, designing this evaluation, examine the information gathered, and reaching agreements among all stakeholders and experts, often requires considerable effort in time, money, and people.

**Aims:** Propose a data fusion strategy to support climate vulnerability assessments by identifying the adaptability of crops in a territory in the short term.

**Methods:** This strategy follows the Joint Directors of Laboratories (JDL) data fusion model guidelines. It was evaluated and validated through a case study in Colombia's upper Cauca river basin. For this purpose, we identified Climate, Soil, Water Quality, Productive Alliances, and Production as the most relevant data sources to be integrated. Using metrics such as Mean IR (Average Imbalance Ratio), SCUMBLE (Score of ConcUrrence among iMBalanced LabEls), TCS (Theoretical Complexity Score), among others, we evaluated the combined datasets according to their theoretical complexity. The adaptability of crops in a territory was addressed as a multi-label learning problem, assessing the performance of different multi-label classification models with both test and actual data.

**Results:** Comparing the predicted crops with the actual ones, we obtained a 98% similarity without considering crop ranking using the Binary Relevance approach and the Random Forest and XGBoost algorithms. While with a more exhaustive test involving order, we obtained a maximum similarity of 67% using Binary Relevance and Random Forest.

**Conclusions:** Evaluating the meta-features of a data source reduced time and effort in the implementation and training of a large number of predictive models. Identifying a central dataset (Agronet) to label the Combined Data Sources (CDS) was a key finding for this study. The multi-label exploratory analysis provided key metrics such as TCS to identify those combined datasets (climate, soil, water quality, productive alliances, and production) that might be most appropriate to train subsequent predictive models. We selected the BR-RF (Binary Relevance - Random Forest) model to perform the crop prediction by evaluating and validating the predictive models.

**Keywords:** Climate vulnerability assessment, Climate change, Crop production, Data processing, Data fusion, Machine learning, Multi-label classification, Multi-label dataset, Sustainable agriculture.

# Resumen Estructurado

**Antecedentes:** Identificar especies y variedades de cultivos adaptables a los impactos del cambio climático es uno de los principales aspectos de las evaluaciones de vulnerabilidad climática. Esta estimación implica el procesamiento, la integración y el análisis de muchas fuentes de información para ofrecer respuestas precisas y oportunas. Sin embargo, el diseño de estas evaluaciones, la revisión de la información recopilada y la consecución de acuerdos entre todos los stakeholders, suele requerir un esfuerzo considerable en términos de tiempo, dinero y personas.

**Objetivos:** Proponer una estrategia de fusión de datos para soportar evaluaciones de vulnerabilidad climática, identificando la adaptabilidad de cultivos en un territorio a corto plazo.

**Métodos:** Esta estrategia sigue las directrices del modelo de fusión de datos JDL (Joint Directors of Laboratories). Esta fue evaluada y validada a través de un estudio de caso en la cuenca alta del río Cauca en Colombia. Para este propósito, se identificaron Clima, Suelo, Calidad del Agua, Alianzas Productivas y la Producción de cultivos como las fuentes de datos más relevantes a ser integradas. Utilizando métricas como Mean IR (Average Imbalance Ratio), SCUMBLE (Score of ConcUrrence among iMBalanced LabEls), TCS (Theoretical Complexity Score), entre otras, evaluamos los conjuntos de datos combinados según su complejidad teórica. La adaptabilidad de los cultivos en un territorio se abordó como un problema de aprendizaje multi-etiqueta evaluando el rendimiento de diferentes modelos de clasificación con datos reales y datos de prueba.

**Resultados:** Comparando los cultivos predichos con los reales, obtuvimos un 98% de similitud sin tener en cuenta el ranking de los cultivos utilizando el enfoque de Relevancia Binaria y los algoritmos Random Forest y XGBoost. Mientras que con

una prueba más exhaustiva que incluyó el orden, obtuvimos una similitud máxima del 67% utilizando Binary Relevance y Random Forest.

**Conclusiones:** La evaluación de las meta-características de una fuente de datos redujo el tiempo y el esfuerzo en la implementación y el entrenamiento de un gran número de modelos predictivos. Un hallazgo clave fue identificar un conjunto de datos central (Agronet) para etiquetar todas las fuentes de datos combinadas (CDS). Adicionalmente, el análisis exploratorio multi-etiqueta proporcionó métricas clave como el Puntaje de Complejidad Teórica (TCS) para identificar aquellos CDS (clima, suelo, calidad del agua, alianzas productivas y producción) que podrían ser los más apropiados para entrenar futuros modelos predictivos. Finalmente, seleccionamos el modelo BR-RF (Binary Relevance - Random Forest) para llevar a cabo la predicción de cultivos a partir de la evaluación y validación de modelos predictivos.

**Palabras Clave:** Evaluación de vulnerabilidad climática, Cambio climático, Producción de cultivos, Procesamiento de datos, Fusión de datos, Aprendizaje automático, Clasificación multi-etiqueta, Conjunto de datos multi-etiqueta, Agricultura sostenible.

# Contents

# List of Figures

# List of Tables

# List of Equations

# Glossary

**A**

ANDA
Archivo Nacional de Datos ..... 52, 110
ANOVA
Analysis of Variance ..... VIII, 44, 46, 80, 81
ARL
Average Risk Level of water contamination ..... 65
AVA
Agriculture, Vulnerability, and Adaptation ..... 5, 2, 33

**B**

BR
Binary Relevance ..... 13
BRPLUS
Binary Relevance Plus ..... 13
BR-RF
Binary Relevance - Random Forest ..... 7, 9, V, 83, 84, 85, 100

**C**

CA
Climate Adaptability ..... 25, 65, 67, 105, 108
CART
Classification and Regression Trees ..... 45, 79, 81, 83, 84
CDS
Combined Data Sources ..... 7, 9, VIII, 44, 45, 79, 80, 81, 84, 85, 99, 100
CGD
Combined Global Dataset ..... V, VIII, 72, 77, 79, 80, 84, 85

**Ch**

CH
Cardinality Histogram ..... 73

**C**

CNN
Convolutional Neural Networks ..... 29
CORPOICA
Corporación Colombiana de Investigación Agropecuaria ..... 53, 89, 95, 96, 111
CSA
Climate-Smart Agriculture ..... 8
CVA
Climate Vulnerability Assessment ..... 2, 3

**D**

DANE
Departamento Administrativo Nacional de Estadística ..... 54, 55, 56, 89, 110, 111

**K**

KNN
K-Nearest Neighbors ..... 29

**L**

LB
Label Bar ..... 74
LC
Label Concurrence ..... 74
LP
Label Powerset ..... 13
LR
Logistic Regression ..... VII, 37, 58, 62, 63, 66, 67

**M**

m.a.s.l.
meters above sea level ..... 49
MAJORITY
Majority Class Prediction ..... 45, 81
MeanIR
Average IRLbl ..... 43
MIO
Managed Information Object ..... 47, 90, 91
MLC
Multi-Label Classification ..... 11, 12
MLD
Multi-Label Dataset ..... V, VII, VIII, 12, 13, 41, 42, 43, 44, 74, 75, 76, 77, 100
ML-KNN
Multi-Label KNN ..... 29
MLL
Multi-Label Learning ..... 12, 100
MODIS-EVI
Vegetation Index Products (EVI) ..... 27
MODIS-NDVI
Vegetation Index Products (NDVI) ..... 27

**N**

NBAA
Number of Beneficiaries Approved by Alliance ..... 91
NHA
Number of Hectares per Alliance ..... 91
NN
Neural Networks ..... 29

**P**

PICOC
Population, Intervention, Comparison, Outcomes, Context ..... 19

**R**

RAKEL
Random k-labelsets ..... 13
RANDOM

**S**



**T**



**U**



**X**

# Chapter 1

# Introduction

## 1.1. Context

The total habitable area on Earth corresponds to 10.4 billion hectares, and agriculture develops on approximately 50% of this surface (4 billion hectares for livestock and 1.1 billion hectares for crops) [1]. Agriculture not only provides food and raw materials but also employment opportunities for a large percentage of the world population [2]. Although agriculture contributes approximately 5% to 7% of Gross Domestic Product (GDP) in modern economies, as this percentage increases, the economic system becomes more vulnerable. On the other hand, prospective analyses forecast four world trends around 2050. The first one is the projection of the world population, which is expected to be around 9 billion people [3]. The second trend is linked to changes in food consumption habits, caused especially by the effects of marketing and improving the family economy in emerging countries. A third trend is associated with decreases in arable land (0.19 ha/person in 2016 compared to 0.37 ha/person in 1960). Finally, the latest trend is related to decreases in agricultural productivity due to global warming, where about 2 billion hectares of arable land will be in degradation process.

Population growth and food availability constitute a close and dynamic relationship that has become more complex over time. An imbalance in food supply and demand could negatively impact this relationship and directly affect a country's food security [4], [5]. World Bank statistics indicate a rural population in the world ascending to about 3.4 billion inhabitants (approximately 45% of the world population) [6]. These inhabitants are especially important for food production, as well as raw materials for the industry, but they also play a role of great importance in the conservation of the

planet's natural resources. Besides, estimates indicate that in 2050 about 120 million families dedicated to growing coffee will be negatively affected by factors such as increased temperature, climate variations, and soil degradation. These and other effects will directly impact the biological development of this crop. Like coffee, many other crops will no longer adapt to the conditions of a given territory [7]. Production of ten main crops such as barley, cassava, corn, palm oil, rapeseed, rice, sorghum, soybeans, sugarcane, and wheat has been affected, which represent key food sources for human beings [8]. These food sources constitute 83% of all calories produced on arable land and understanding how much can be affected has become an urgent task for researchers around the world.

## 1.2. Motivation

Food security promotes the interaction between four essential elements: food availability, food access, food utilization, and vulnerability [9]. In the agricultural context, the latter aspect refers to the degree of a system's susceptibility to climate change's adverse effects [10]. Agricultural vulnerability is a function of the system's exposure, sensitivity, and adaptive capacity [11]–[13]. These aspects make it necessary to explore alternatives to mitigate or adapt to the phenomena of variability and climate change, especially in the future, where it will be necessary to feed more people on a lesser portion of land [3], [14]. Measuring agricultural vulnerability is essential for executing sustainable actions and making decisions to develop food security scenarios [15], [16]. Therefore, determining the degree of agricultural vulnerability represents a guide for sustainability and adaptability focused on changing future conditions.

Vulnerability is a fundamental concept in analyzing the risks of climate variability in agriculture [17]. In recent years, different Climate Vulnerability Assessments (CVAs) have been proposed [13], from the use of surveys, databases, and structured indicators [18], to the development of a network of economic, water, and crop models to simulate agricultural markets [19]; adaptation strategies in socioeconomic and environmental scenarios combined with possible climate impacts of the agricultural sector [20]; and finally, methodologies to measure agricultural vulnerability such as AVA (Agriculture, Vulnerability, and Adaptation) [21]. Some results of this last methodology support this thesis considering the study area and the data availability for analysis.

On the other hand, climatic factors are not the only element affecting agricultural production and crop adaptability. Aspects such as environmental, economic, social, cultural, political, among others, can significantly influence agricultural processes in a particular region. Identifying crop species and varieties that adapt to the impacts of climate change is one of the most economical and environmentally friendly strategies for food security [14]. In this sense, the use of appropriate tools, methodologies, or instruments strengthens the agricultural sector, especially the process of adaptation to changing conditions [22]. As stated in USAID's Climate Change and Development Strategy [23], *"Climate adaptation requires that we utilize aspects such as science, technology, innovation, and the best available information to understand and respond to unavoidable impacts".* Therefore, this thesis addresses these aspects in order to determine, in the short term, the crop adaptability in a territory.

## 1.3. Problem Statement

Climate vulnerability assessments have shown significant results in different aspects. Some of these focus on identifying entry points to address climate stress factors, highlighting opportunities to take advantage of in a changing climate, and identifying adaptation measures [13]. However, designing these assessments often requires enormous efforts in time, money, and people. For instance, the interaction between experts in different areas makes it difficult to reach a consensus. Likewise, more specific problems have emerged, such as the fusion of data sources. Under these circumstances, a successful CVA requires accurate information to cover each vulnerability dimension [22].

The accelerated development of Information Technologies (IT) and the massive data generation have increased the volume and the variety of data sources used in CVAs. One of the fundamental aspects of a data fusion process is precisely the variety of data sources. In agriculture, as in many other domains, data inputs correspond to numerous sources such as sensors, structured and unstructured databases, plain text files, multimedia files, and reports. Additionally, many of these data sources have restricted access and are not freely available. Only authorized personnel, usual members of specific organizations, can use such data for a CVA. In this sense, open data allows many more groups of people and organizations to benefit from the availability of these sources [24]. Given these points, data fusion represents a non-

trivial activity in agricultural vulnerability assessments considering aspects like quantity, diversity, and access restrictions to data sources.

## 1.4. Research Questions

From the previous aspects, this doctoral thesis raises the following research question:

- *How to create a data fusion strategy to support agricultural vulnerability assessments, considering data analysis characteristics such as variety, value, and viability?*

## 1.5. Research Purpose and Objectives

The aim of this research is to propose a data fusion strategy to support agricultural vulnerability assessments. This purpose was achieved through the following specific objectives.

1. Determine the viability of the attributes or variables in each agricultural vulnerability dimension in terms of relevance through statistical methods or machine learning techniques.

2. Structure one or more datasets from different dimensions and scales using one or more data integration methods.

3. Build a crop estimation model based on data analysis characteristics such as variety, value, and viability.

4. Implement a prototype from the proposed strategy and evaluate it using different performance metrics.

## 1.6. Contributions

The main contributions of this doctoral thesis are mentioned below, which were obtained from the previously mentioned objectives 1, 2, and 3. The objective 4 gathers the results of the first three objectives.

- A set of data sources corresponding to:

- A set of pre-processed data sources from different agricultural vulnerability dimensions.
- A set of combined datasets from the pre-processed data sources.

– A data fusion strategy composed of:

- A multi-dimensional data preparation process to support Climate Vulnerability Assessments (CVA).
- A method to identify relationships and integrate data sources from different dimensions.
- A multi-label training and evaluation scheme for crop prediction in a given municipality.

Figure 1.1 shows the relations among contributions explained above. From the multi-dimensional data preparation process, we obtain the pre-processed data sources. These sources are combined through the integration process and finally, through the multi-label classifiers, we predict the most suitable crops depending on different conditions in a territory.

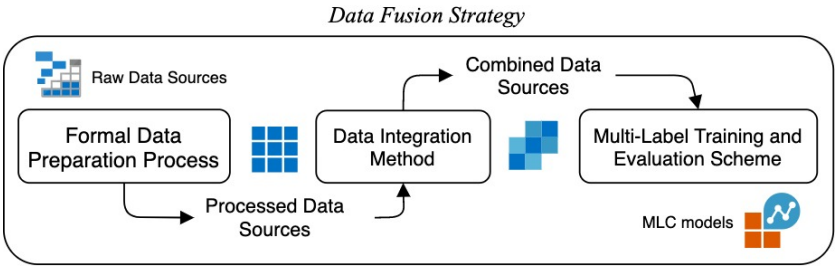



**Figure 1.1.** Doctoral thesis contributions.

## 1.7. Outline

This research is composed of eight chapters which are described below.

– **Chapter 2. State of the Art.** Definition of main concepts, overview of related works, and identification of gaps around the research problem.

– **Chapter 3. Data Fusion Strategy.** Overview of the data fusion strategy and its components.

- **Chapter 4. Data Assessment.** Evaluation of data sources and definition of a preparation process to improve data quality.

- **Chapter 5. Data Integration and Analysis.** Identification of spatial-temporal relationships between data sources, production of new datasets (combined data sources) with a more synthesized and reliable added value, and application of machine learning techniques to train a set of models and estimate one or more target variables (agricultural crops). Additionally, performance evaluation of the trained models using different metrics.

- **Chapter 6. Process Refinement and User Interface.** Details about the refinement process, the storage system, and the software prototype implemented from the data fusion strategy.

- **Chapter 7. Conclusions and Future Work.** Analysis of results highlighting the main research contributions. Furthermore, relevant recommendations for the development of future work are presented.

## 1.8. Publications

### 1.8.1 Accepted Papers

- **López, I.D.**, Figueroa, A. and Corrales, J.C. (2020), "Multi-Dimensional Data Preparation: A Process to Support Vulnerability Analysis and Climate Change Adaptation," in IEEE Access (JCR Q1), vol. 8, pp. 87228-87242, doi: 10.1109/ACCESS.2020.2992255.
- **López, I.D.**, Grass, J.F., Figueroa, A. and Corrales, J.C. (2021), "A proposal for a multi-domain data fusion strategy in a climate-smart agriculture context," in International Transactions in Operational Research (ITOR) (JCR Q2), doi: 10.1111/itor.12899.
- **López, I.D.**, Figueroa, A. and Corrales, J.C. (2021), "Multi-Label Data Fusion to Support Agricultural Vulnerability Assessments," IEEE Access (JCR Q1), doi: 10.1109/ACCESS.2021.3089665.

### 1.8.2 Other Published Papers

- Lasso E., Valencia O., Corrales D.C., **López I.D.**, Figueroa A., Corrales J.C. (2018), "A Cloud-Based Platform for Decision Making Support in Colombian Agriculture: A Study Case in Coffee Rust," in Angelov P., Iglesias J., Corrales

J. (eds) Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change. AACC'17 2017. Advances in Intelligent Systems and Computing, vol 687. Springer, Cham. doi: 10.1007/978-3-319-70187-5_14.

- **López I.D.**, Valencia C.H., Corrales J.C. (2018), "A Data Mining Tool for Water Uses Classification Based on Multiple Classifier Systems," in Nicosia G., Pardalos P., Giuffrida G., Umeton R. (eds) Machine Learning, Optimization, and Big Data. MOD 2017. Lecture Notes in Computer Science, vol 10710. Springer, Cham. doi: 10.1007/978-3-319-72926-8_30.
- **López I.D.**, Corrales J.C. (2018), "A Smart Farming Approach in Automatic Detection of Favorable Conditions for Planting and Crop Production in the Upper Basin of Cauca River," in Angelov P., Iglesias J., Corrales J. (eds) Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change. AACC'17 2017. Advances in Intelligent Systems and Computing, vol 687. Springer, Cham. doi: 10.1007/978-3-319-70187-5_17.
- Plazas J.E., **López I.D.**, Corrales J.C. (2017), "A Tool for Classification of Cacao Production in Colombia Based on Multiple Classifier Systems," in Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science, vol 10405. Springer, Cham. doi: 10.1007/978-3-319-62395-5_5.
- **López I.D.**, Figueroa A., Corrales J.C. (2017), "Adaptive Prediction of Water Quality Using Computational Intelligence Techniques," in Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science, vol 10405. Springer, Cham. doi: 10.1007/978-3-319-62395-5_4.
- **López I.D.**, Valencia C.H., Corrales J.C. (2016), "Predicting Water Quality based on Multiple Classifier Systems," in 21st Century Watershed Technology Conference and Workshop Improving Water Quality and the Environment Conference Proceedings, 3-9 December 2016, IKIAM Universidad Regional Amazónica Quito, Ecuador. doi: 10.13031/wtcw.2016008.

# Chapter 2

# State of the Art

This chapter presents the background around the main topics addressed in this doctoral thesis. In the first part, we define the concepts of Climate-Smart Agriculture, Data Fusion vs. Data Integration, and Multi-Label Classification. Subsequently, we present the related works and the current state of knowledge using two methodologies for literature reviews. The first corresponds to *Petersen's* proposal for systematic mapping [25] and the second is the *Kitchenham's* guidelines for systematic reviews [26]. Finally, for each topic of the literature review, we identify the shortcomings of the related works and mention our contributions.

## 2.1. Background

This section presents the definitions of the three main concepts in this doctoral thesis such as Climate-Smart Agriculture, Data Fusion vs. Data Integration, and Multi-Label Classification.

### 2.1.1 Climate-Smart Agriculture (CSA)

In 2010, the Food and Agriculture Organization (FAO) introduced the concept of Climate-Smart Agriculture (CSA) [14]. This is a synergistic approach for transforming and reorienting agricultural systems to support food security under the new realities of climate change [27]. In summary, CSA promotes the adoption of coordinated actions by farmers, researchers, the private sector, civil society, and policy makers to address the new realities of climate change. As shown in Figure 2.1, CSA is based on three fundamental pillars. The first is sustainably increasing agricultural productivity

and incomes. The second is to adapt and build resilience to climate change. And the third is to reduce and/or eliminate Greenhouse Gas (GHG) emissions where possible.

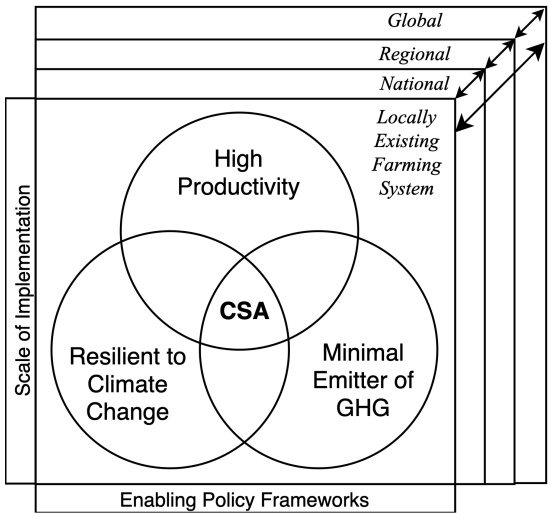



**Figure 2.1.** The objectives of Climate-Smart Agriculture. Adapted from [28].

Figure 2.1 indicates that CSA can be implemented at different agricultural scales, from farm-lots to more complex structures such as global agro-ecosystems [28]. This approach provides tools to help stakeholders identify, at local, national, and international levels, agricultural strategies in line with site conditions. In other words, the CFS does not correspond to a set of practices that can be applied universally, but an approach that involves the integration of different elements in local contexts [27]. Also, considering that agriculture is traditionally site and context-specific, CSA can be applied through a customized intervention for each stakeholder. Furthermore, maintaining an integrative functioning of the different CSA components must be supported by policy frameworks at each level of the agricultural scale [28].

CSA includes both on-farm and off-farm actions as catalytic factors, covering different aspects such as policies, institutions, investments, technologies, methodologies, tools, and development practices. These aspects integrate multi-stakeholder processes and networks that allow the exchange of knowledge from different actors such as local institutions, farmers, climatologists, indigenous communities, among others. An agricultural vulnerability assessment, formally known as Climate Vulnerability Assessment (CVA), is a key approach to integrate the above aspects in a synergistic way, therefore, this concept is defined below.

***Climate Vulnerability Assessment (CVA)***

Different areas and disciplines have involved experts such as scientists, decision-makers, farmers, among others stakeholders, to propose a large number of Climate Vulnerability Assessments (CVAs). These stakeholders are responsible for designing a CVA to understand three important questions: who or what is vulnerable to climate change and variability, why and how they are vulnerable, and what opportunities exist to reduce these vulnerabilities. CVAs are methodologies designed to meet the specific needs of a strategy (globally or at a country level), project (at regional or sectoral level), or activity (specific organizations or sites) [13]. A CVA covers a wide range of methodologies in different fields, from descriptions of climate implications in communities to technical analysis of infrastructure under climate change scenarios [29].

There are several methods to conduct a CVA with different levels of complexity. Among the most widely used are desk reviews, stakeholder and expert workshops, community-based approaches, and additional specialized analysis (vulnerability indexes, simulations, modeling, impact analysis, map generation, among others) [10]. These methods comprise several stages such as conduct literature reviews; identify stakeholders; evaluate the information needs of stakeholders; evaluate the roles and capacities of stakeholders; select data, methods, and tools adjusted to the spatial and temporal scales; and design evaluations on CSA objectives [22]. In this regard, Information and Communication Technologies (ICTs) play a significant role for CVAs by developing simple and robust scientific tools to guide farmers' decision-making. Therefore, improving the forecasts reliability allows farmers to make better use of climate information, take preventive measures, and minimize the effects of extreme events [30].

### 2.1.2 Data Fusion (DF) vs. Data Integration (DI)

Different approaches suggest that Data Fusion (DF) differs from the concept of Data Integration (DI). Below, we present two points of view: traditional databases and data fusion modeling. First, for conventional database approaches, data integration refers to unified access to data that resides in multiple and autonomous sources [31]. These approaches establish three main stages for DI: schema alignment, record linkage, and data fusion. From this point of view, DF represents a step of DI, which is related to data quality management. By detecting conflicts in the data sources, DF identifies the correct values used in the DI process [32]. In contrast, data fusion modeling

approaches propose similar aspects between DF and DI [33], [34]. The intersection point is to integrate and organize data from multiple sources to obtain a unified view around processed information, which is ready to be consumed by different applications [35]. However, the level of abstraction when combining data is the starting point for differentiating such concepts. While DI prioritizes the generation of consumable data, DF focuses on data abstraction at different levels [36].

The definition of DF that guides this study is the one presented in data fusion modeling approaches. Data fusion refers to the process of combining data or information to estimate or predict the state of some aspect of the universe over some past, current, or future time period [37]. In the agricultural context related to this study, DF corresponds to the global process around the data. This process includes preparing, integrating, and analyzing data from multiple dimensions of agricultural vulnerability. Preparation refers to cleaning up raw data sources. Integration corresponds to finding spatio-temporal relationships between data sources and generating a more synthesized and complete data set. Finally, the analysis process establishes the precision of the integration results and verifies the added value generation.

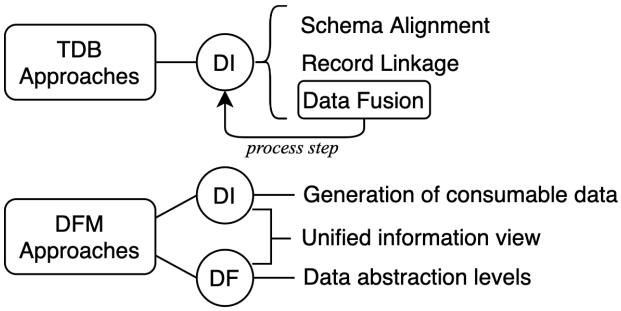



**Figure 2.2.** Classification scheme of Data Fusion (DF) and Data Integration (DI) concepts according to Traditional Database (TDB) and Data Fusion Modeling (DFM) approaches. Adapted from [38].

### 2.1.3 Multi-Label Classification (MLC)

In Machine Learning (ML), traditional classification algorithms induce a model from training examples to predict the value of a single label (single-label classification). This label can contain two classes (binary) or multiple classes(multi-class) [39], [40]. An example of binary classification corresponds to determining the presence of a disease in a crop, and the possible classes are 1 (presence) and 0 (absence). On the other hand, an example of multi-class classification corresponds to determining the production level of a crop, where the possible classes are 0 (low), 1 (intermediate), 2

(high), and 3 (very high). However, this type of classification is not sufficient for many real-world situations where the problems lie in predicting values of several labels at the same time, also called Multi-Label Classification (MLC) [41], [42]. Therefore, correlations between all labels play a critical role in obtaining more realistic predictions. An example of MLC corresponds to predicting the set of crops produced in a territory in the future. In this case, there exists a label in each crop, and the possible classes correspond to 1 (produced) and 0 (no produced).

Multi-label classification is part of the paradigm called Multi-Label Learning (MLL) [43]. Formally, MLC is the task of classifying an instance into several categories or classes simultaneously [44]. The first input of MLC data is a Multi-Label Dataset (MLD) [45]. Table 2.1 shows instances or examples (1,2,3,4), input attributes ($x_1$, $x_2$, $x_3$, $x_4$), labels or classes ($\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$), and labelsets ($\{\lambda_1, \lambda_4\}$, $\{\lambda_3, \lambda_4\}$, $\{\lambda_1\}$, $\{\lambda_2, \lambda_3, \lambda_4\}$) are the main elements of an MLD. The number of labels in an MLD can be hundreds or even thousands, and sometimes it can be higher than the number of instances and input attributes. MLC problems are usually more difficult than single-label problems considering their generality [46]. Therefore, MLC algorithms are developed considering one of two approaches. The former is a *transformation*, where MLDs are modified to apply binary or multi-class classification algorithms. The latter is an *adaptation*, where a known algorithm or technique is altered to fit the MLD [47], [48].

| Example | Attributes | Label set |
|---|---|---|
| 1 | $x_1$ | $\{\lambda_1, \lambda_4\}$ |
| 2 | $x_2$ | $\{\lambda_3, \lambda_4\}$ |
| 3 | $x_3$ | $\{\lambda_1\}$ |
| 4 | $x_4$ | $\{\lambda_2, \lambda_3, \lambda_4\}$ |

**Table 2.1.** Example of a Multi-Label Data set (MLD). Quoted from [49].

By handling multi-label data, we can perform exploratory analysis and also employ different learning tasks such as ranking, clustering, and classification. The latter corresponds to the most studied problem in multi-label learning. There are several approaches or strategies to address classification tasks. These strategies build multi-label models using a Machine Learning base algorithm. In Table 2.2, we briefly mention the strategies used in our research.

| MLC Strategy | Description | Reference |
|---|---|---|
| Binary Relevance (BR) | This strategy produces several binary datasets from an MLD, usually one for each label or one for each pair of labels. | [49] |
| Binary Relevance Plus (BRPLUS) | BRPLUS or BR+ increments the binary classifiers feature space to let them discover existing label dependency by themselves. | [50] |
| Ensemble of Classifier Chains (ECC) | ECC aims to exploit the correlations between labels. These predictions are summed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which form the final predicted multi-label set. | [51] |
| Label Powerset (LP) | It transforms the MLD into a multi-class dataset, taking each label set as a class identifier. | [48] |
| Hierarchy Of Multi-label classifiER (HOMER) | This strategy divides the initial learning task into several simpler subtasks. It builds a label hierarchy from a given labels set, and employs a base multi-label classifier to address the resulting subproblems. | [52] |
| Random k-labelsets (RAKEL) | It builds an ensemble of LP classifiers. Each classifier is trained using a different small random subset of the labels set. | [53] |

**Table 2.2.** MLC strategies used in this research.

## 2.2. Related Works

This section establishes a starting point from the knowledge generated around Data Fusion and Multi-Label Classification in Climate-Smart Agriculture contexts. In this sense, it is important to adopt certain guidelines in this research field in order to build a general classification scheme, and therefore, identify the types and amount of research available. Considering the above, this doctoral thesis is guided by a Systematic Mapping [25] to establish general topics according to trend analysis, and a Systematic Literature Review [26] to identify specific research gaps. For systematic mapping, we used *Bibliometrix* [54], an R-tool for comprehensive science mapping analysis. On the other hand, to develop a more exhaustive process such as the systematic literature review, we use *Parsifal* [55], an online tool for planning, conducting, and reporting the review. The results at each phase of the systematic mapping and systematic literature review are presented below. A more detailed description of the process is provided in Appendix A.

### 2.2.1 Systematic Mapping

To establish an overview and identify trends in research around data fusion and multi-label classification in agriculture, we applied a systematic mapping of the literature. Each process steps and their results are described below.

***Definition of Research Questions***

In this step, we defined the research scope by identifying the guiding questions of the systematic mapping. For each topic, we generated several research questions to answer at the end of the mapping. These questions should be sufficiently general to obtain an overview of the research topics. The research questions are mentioned below.

- *RQ1.* Which research topics are related to data fusion and multi-label classification in agriculture?
- *RQ2.* How is the research trend around data fusion and multi-label classification in agriculture?

***Conduct Search***

We identified the primary studies using search strings by querying the Scopus, ScienceDirect, Google Scholar, IEEE Digital Library, and Springer Link databases. Initially, we obtained a total of 4,657 papers, 1,266 for data fusion and 3,391 for multi-label classification. The base search strings for DF and MLC are shown below.

- *Search String for DF.* ("data fusion" OR "data integration") AND agricultur*
- *Search String for MLC.* "multi-label classification" AND agricultur*

***Screening of Papers for Inclusion and Exclusion***

To select studies relevant to the search strings, we established inclusion and exclusion criteria in both research topics. After applying this selection process, we obtained a total of 2,228 papers, 275 for data fusion and 1,953 for multi-label classification. Inclusion and exclusion criteria are presented below.

- *Inclusion Criteria.* Original articles, literature reviews, books, books chapters, and conference proceedings describing empirical studies regarding *data fusion* or *multi-label classification* applied in agriculture. Documents written in English and related to computer science. The abstract explicitly mentions *data fusion* or *multi-label classification*. Where several papers reported the same study, only the most recent was included. Where several studies were reported in the same paper, each relevant study was treated separately. If several articles reported the same study, we only included the most recent one.

– *Exclusion Criteria.* Documents outside the field of computer science and agriculture. Documents other than original articles, literature reviews, books, books chapters, and conference proceedings. Documents written in a language other than English.

### *Keywording of Abstracts*

Based on the main keywords contained in the title and abstract, we generated the study classification schemes presented in Figure 2.3 and Figure 2.4. These diagrams show a hierarchical grouping of the most closely related topics. In our case, we identified three major groups of topics related to DF (satellite imagery, sensors, and decision-making systems) and two main groups related to MLC (multi-label learning and deep learning).

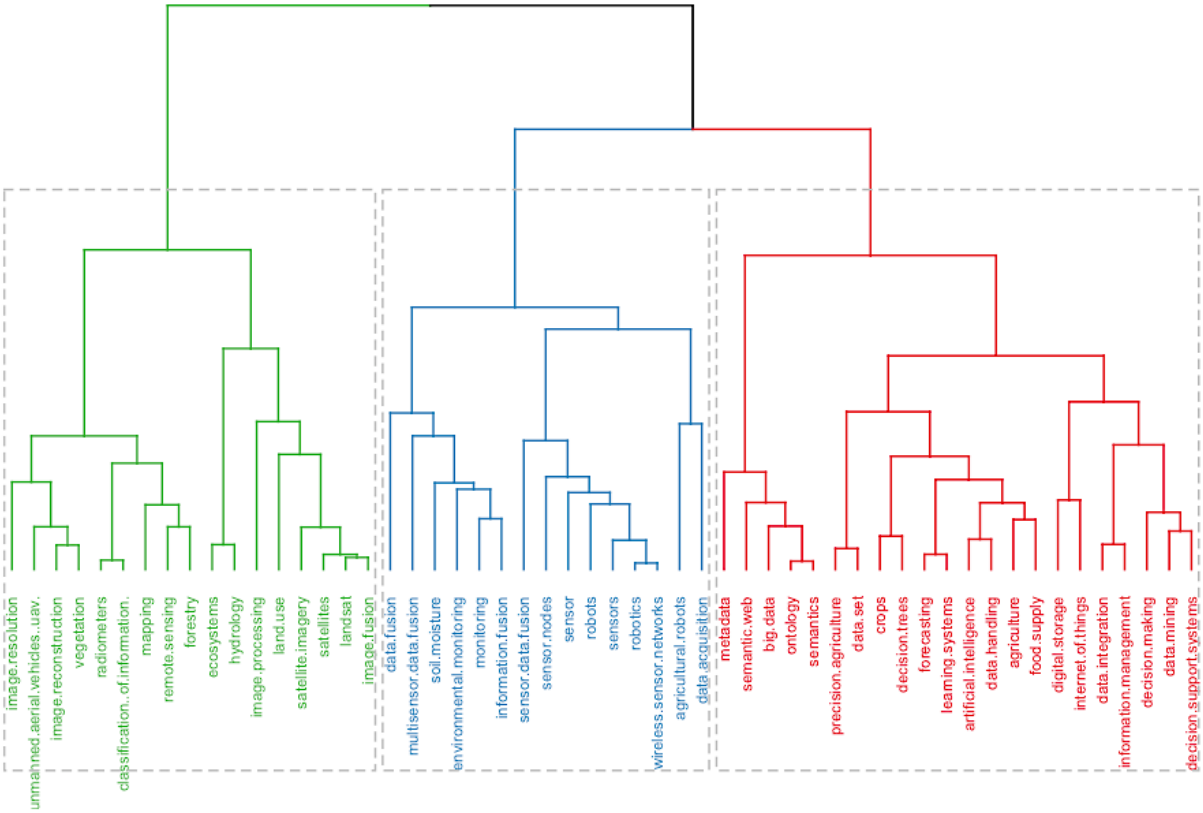



**Figure 2.3.** Classification scheme for data fusion in agriculture.

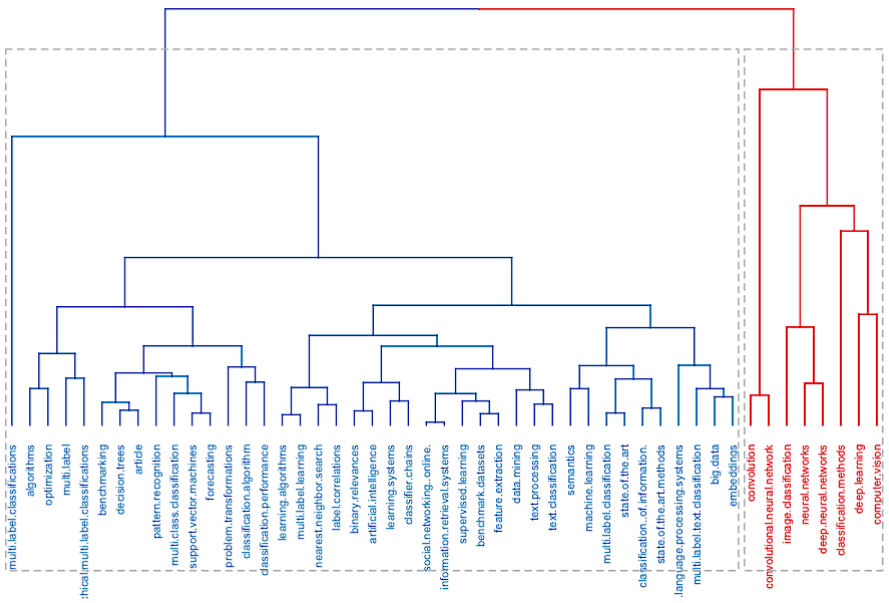



**Figure 2.4.** Classification scheme for multi-label classification in agriculture.

### *Data Extraction and Mapping Process*

We established the number of research papers in a 10-year time window (2011-2020). Figure 2.5 shows a significant increase in annual scientific production around MLC in the last 10 years (Annual Growth Rate of 7.17%). These results represent a trend towards the generation of new multi-label learning approaches in different knowledge domains. In contrast, there was no evidence of a significant increase in annual scientific production around DF (Annual Growth Rate of -5.37%). It could be affirmed considering that the first research works in DF were developed several years before the works in MLC.

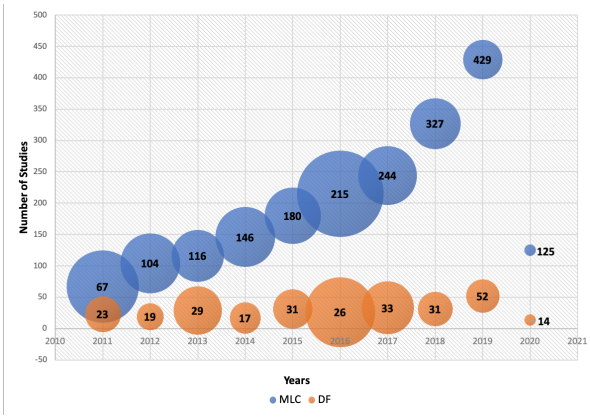



**Figure 2.5.** Annual scientific production around Data Fusion (DF) and Multi-Label Classification (MLC) in different knowledge domains (systematic literature mapping). The size of the circumference indicates the number of citations per year.

On the other hand, we build a conceptual structured map for both topics (Figure 2.6). This map is a categorical data reduction technique and it based on Multiple Correspondence Analysis method (MCA) [56], where dimension 1 explained 49.57% of total variance and dimension 2 explained 14.9%. In Figure 2.6 (a), we observed three topic clusters related to DF in agriculture. Cluster relates topics such as data acquisition and sensor networks to multi-sensor data fusion. Likewise, data fusion acts as a bridge with cluster 2 (satellite image processing) and 3 (decision-making systems). In this last cluster, we visualized some more relevant topics such as artificial intelligence, data mining, learning systems, and prediction, closely related to agriculture, crops, and food supply. This is evidence of research interest in these topics and, in turn, a potential link to the MLC-related clusters shown in Figure 2.6 (b). In this case, the interrelated clusters were DF cluster 3 and MLC cluster 1. The aforementioned indicates that multi-label learning techniques can be applied in combination with a data fusion approach in agricultural contexts.

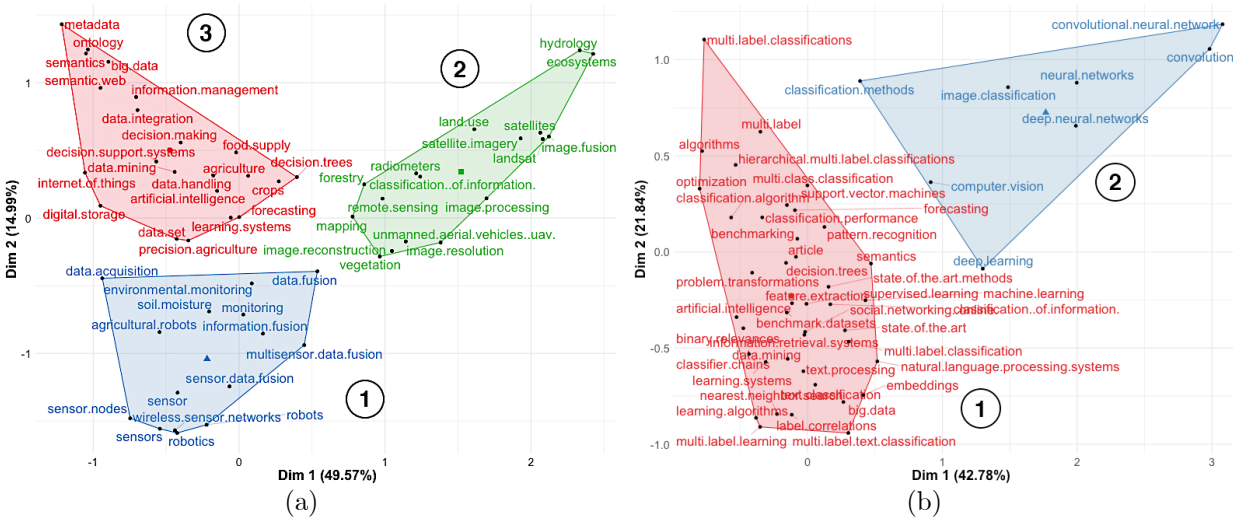



**Figure 2.6.** Conceptual structured map for (a) data fusion and (b) multi-label classification in agriculture.

These statements are supported by the current research trends, which were determined considering a time window from 2011 to 2020 based on the topics previously analyzed as shown in Figure 2.7 and Figure 2.8. These figures show the evolution of the research topics through annual trends. In both, we can see that DF and MLC correspond to recent topics with a high frequency in scientific publications.

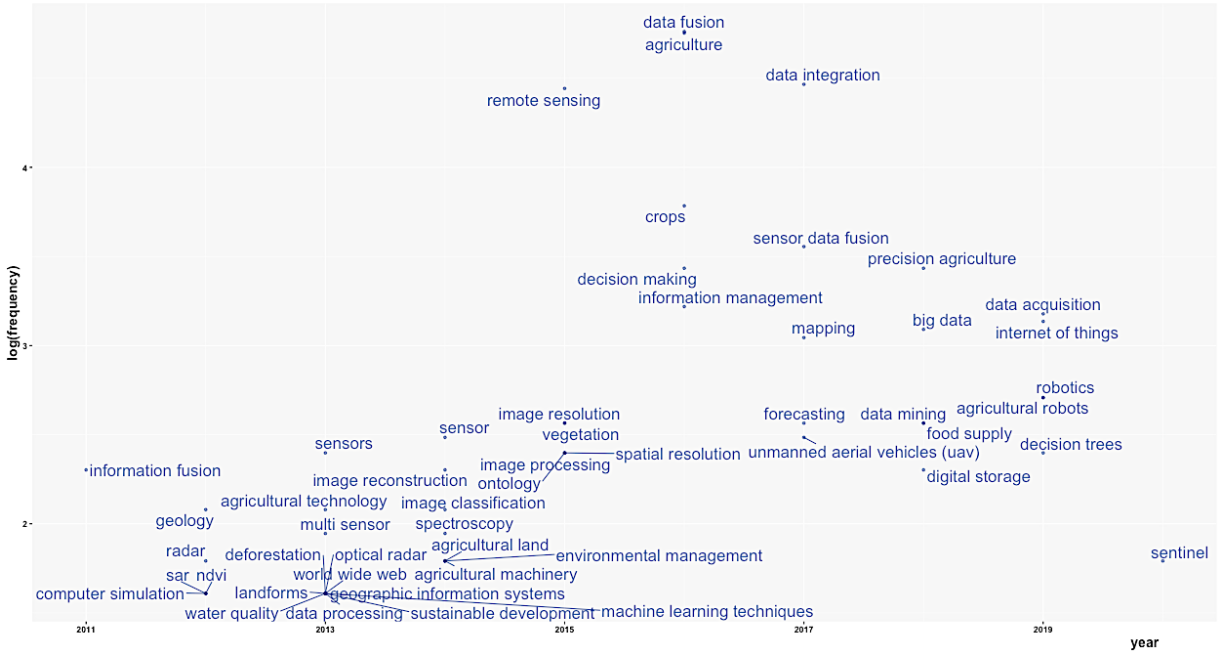



**Figure 2.7.** Trend topics around data fusion in agriculture.

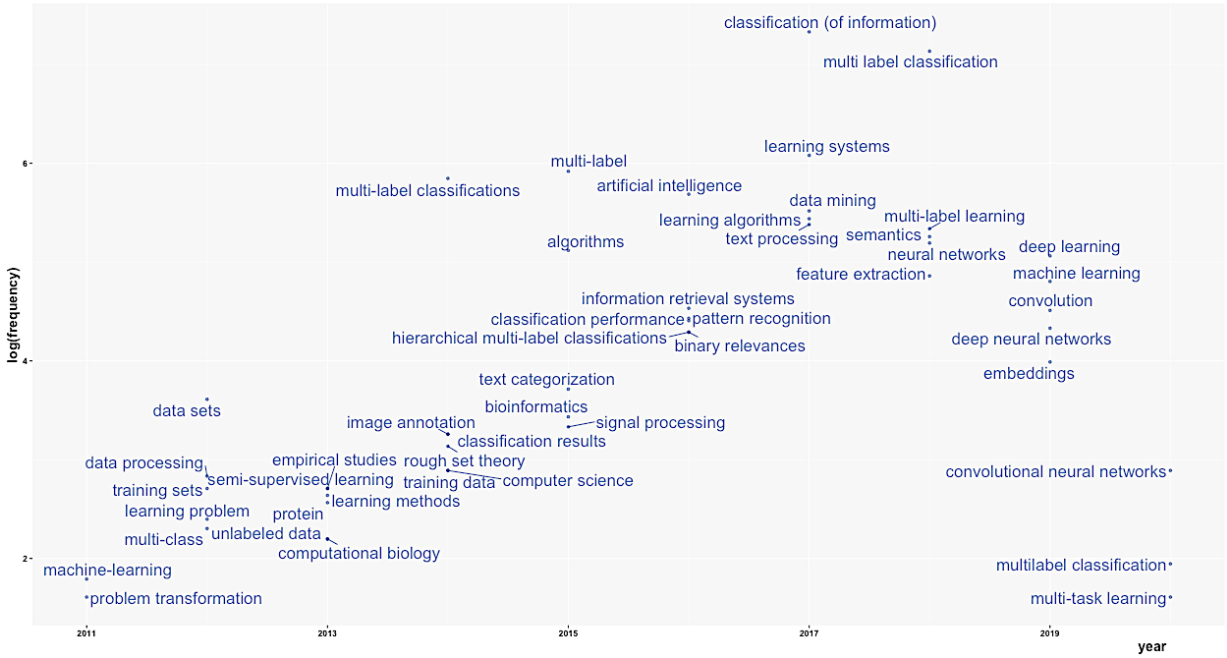



**Figure 2.8.** Trend topics around data fusion in agriculture.

### 2.2.2 Systematic Literature Review

In this review, we investigate two core topics such as *data fusion* and *multi-label classification* techniques in agricultural contexts and climate vulnerability assessments. Planning, conducting, and reporting phases are presented below.

***Planning***

This phase comprises three parts: protocol, quality assessment, and data extraction, which, in turn, include different sections. These parts and their sections are described below.

- *Protocol.* This part describes the objectives and methods of the systematic review. It should be prepared before starting the review and used as a guide to develop it. The protocol comprises the following sections:
  - The review objectives.
  - The elements of the PICOC method [57] for describing a searchable question such as:
    - The *population* in which the evidence is collected (Who?).
    - The *intervention* applied in the empirical study (What or How?).
    - The *comparison* to which the intervention is confronted (Compared to what?).
    - The *outcomes* of the experiment (What are you trying to accomplish or improve?).
    - The *context* of the study (In what kind of organization or circumstances?).
  - The research questions.
  - The keywords and their synonyms.
  - The search strings.
  - The search sources.
  - The selection criteria for inclusion or exclusion of papers.

Table 2.3 and Table 2.4 present the protocols for data fusion and multi-label classification, respectively.

| Topic | |
|---|---|
| Data fusion in agricultural contexts and climate vulnerability assessments | |
| Objectives | |
| To describe approaches of data fusion applied to agricultural contexts, specifically in climate vulnerability assessments. | |
| PICOC | |
| Population | Official open data, restricted data with access permission, results of agricultural vulnerability analysis |
| Intervention | Data Fusion, Data Integration |
| Comparison | A comparison of methods, algorithms, data fusion models, datasets |
| Outcome | A systematic literature survey report including synthesis of most relevant articles published |

| | | |
|---|---|---|
| | on data fusion for crop prediction | |
| Context | A systematic investigation to consolidate a peer-reviewed and academic research, classification and comparison, trends and future research directions | |
| Research Questions | | |
| RQ1.1. | Which Data Fusion (DF) approaches have been applied in Climate Vulnerability Assessments (CVAs)? | |
| RQ1.2. | Which Data Fusion (DF) approaches have been applied in other agricultural contexts? | |
| Keywords and Synonyms | | |
| Keyword | Synonyms | Related PICOC |
| climate vulnerability assessments | agricultural vulnerability, agricultural vulnerability analysis, CVA, methodologies for agricultural vulnerability | Population |
| data fusion | data integration | Intervention |
| open data | data, dataset | Population |
| Search String | | |
| ("climate vulnerability assessments" OR "agricultural vulnerability" OR "agricultural vulnerability analysis" OR "CVA" OR "methodologies for agricultural vulnerability") AND ("open data" OR "data" OR "dataset") AND ("data fusion" OR "data integration") | | |
| Sources | | |
| Google Scholar | https://scholar.google.com/ | |
| IEEE Digital Library | http://ieeexplore.ieee.org | |
| Science@Direct | http://www.sciencedirect.com | |
| Scopus | http://www.scopus.com | |
| Springer Link | http://link.springer.com | |
| Selection Criteria | | |
| Inclusion Criteria | Exclusion Criteria | |
| – Papers about DF applied to CVAs<br>– Papers about DF applied to agricultural contexts<br>– Papers about DF applied to crop production<br>– Papers whose objectives are similar to research, but applying other techniques | – It is not an article<br>– Papers about DF applied to domains other than agriculture<br>– Papers not published in the last ten years<br>– Papers outside the area of computer science<br>– Papers that not included the keywords<br>– Techniques other than DF applied to agriculture | |

**Table 2.3.** Systematic review protocol around data fusion in agricultural contexts and climate vulnerability assessments.

| | |
|---|---|
| Topic | |
| Multi-label classification in agricultural contexts and climate vulnerability assessments | |
| Objectives | |
| To describe approaches of multi-label classification applied to agricultural contexts, specifically in climate vulnerability assessments. | |
| PICOC | |
| Population | Official open data, restricted data with access permission, results of agricultural vulnerability analysis |
| Intervention | Multi-Label Classification |
| Comparison | A comparison of methods, algorithms, MLC models, datasets |
| Outcome | A systematic literature survey report including synthesis of most relevant articles published on multi-label classification for crop prediction |
| Context | A systematic investigation to consolidate a peer-reviewed and academic research, classification and comparison, trends and future research directions |
| Research Questions | |

| | | |
|---|---|---|
| RQ2.1. | Which Multi-Label Classification (MLC) approaches have been applied in agriculture? | |
| RQ2.2. | Which Multi-Label Classification (MLC) approaches have been applied for crop prediction? | |
| Keywords and Synonyms | | |
| Keyword | Synonyms | Related PICOC |
| climate vulnerability assessments | agricultural vulnerability, agricultural vulnerability analysis, CVA, methodologies for agricultural vulnerability | Population |
| algorithm | approach, method, technique | Comparison |
| machine learning | artificial intelligence | Intervention |
| multi-label classification | mlc, multi-label, multi-label learning | Intervention |
| multi-label dataset | mld, multi-label data | Population |
| open data | data, dataset | Population |
| Search String | | |
| ("agricultural vulnerability analysis" OR "agricultural vulnerability" OR "methodologies for agricultural vulnerability" OR "multi-label dataset" OR "mld" OR "multi-label data" OR "open data" OR "data" OR "dataset") AND ("machine learning" OR "artificial intelligence" OR "multi-label classification" OR "mlc" OR "multi-label" OR "multi-label learning") AND ("algorithm" OR "approach" OR "method" OR "technique") | | |
| Sources | | |
| Google Scholar | https://scholar.google.com/ | |
| IEEE Digital Library | http://ieeexplore.ieee.org | |
| Science@Direct | http://www.sciencedirect.com | |
| Scopus | http://www.scopus.com | |
| Springer Link | http://link.springer.com | |
| Selection Criteria | | |
| Inclusion Criteria | Exclusion Criteria | |
| – Papers about MLC applied to agriculture<br>– Papers from the area of computer science<br>– Papers published in the last ten years<br>– Papers that include one or more keywords<br>– Papers whose objectives are similar to research, but applying other techniques | – It is not an article<br>– Papers about MLC applied to domains other than agriculture<br>– Papers not published in the last ten years<br>– Papers outside the area of computer science<br>– Papers that not included the keywords<br>– Techniques other than MLC applied to agriculture | |

**Table 2.4.** Systematic review protocol around MLC in agricultural contexts and climate vulnerability assessments.

– *Quality Assessment*. This part builds a form to evaluate the quality of the papers selected in the review. The quality assessment is presented in Table 2.5, and it comprises the following sections:
  - Formulation of questions and answers to evaluate research quality.
  - Assignment of weights to each answer.
  - Setting of a maximum score (calculated based on the number of questions and on the answer of greater weight), and a cutoff score (to select only those items exceeding this value, which continue to the data extraction phase).

| Questions | |
|---|---|
| – Is the paper based on research or is it merely a report based on expert opinion?<br>– Is there a clear statement of the aims of the research?<br>– Is there a description of the context in which the research was carried out?<br>– Was the data collected in a way that addressed the research issue?<br>– Was the data analysis sufficiently rigorous?<br>– Is there a clear statement of findings?<br>– Is the study of value for research or practice? | |
| Answers | |
| Description | Weight |
| Yes | 1.0 |
| Partly | 0.5 |
| No | 0.0 |
| Quality Assessment Scores | |
| Max Score | 7.0 |
| Cutoff Score | 4.5 |

**Table 2.5.** Checklist to assess research quality of systematic review around data fusion and multi-label classification in agricultural contexts and climate vulnerability assessments.

– *Data Extraction.* This part is focused on designing a form for extracting data from articles that pass the quality assessment. The data extraction form is presented in Table 2.6, and it comprises the following sections:
  - Description of those aspects that need to be extracted from the items.
  - Assignment of values to each aspect identified in the previous point.

| Description | Type | Values |
|---|---|---|
| Has any DF/MLC technique been applied in agricultural contexts? | Boolean field | True / False |
| If the previous answer is yes, any DF/MLC technique been applied in CVAs? | Boolean field | True / False |
| Is there any methodology or method to apply DF/MLC? | Boolean field | True / False |
| Is there any methodology to evaluate the DF/MLC performance? | Boolean field | True / False |
| Is any experiment carried out? | Boolean field | True / False |
| If the previous answer is yes, choose the evaluation method used | Select one field | – Case study<br>– Experimental or user evaluation<br>– Performance test<br>– User surveys<br>– User testing |

**Table 2.6.** Data extraction form of systematic review around Data Fusion (DF) and Multi-Label Classification (MLC) in agricultural contexts and climate vulnerability assessments.

### *Conducting*

The aim of this phase is to find all relevant studies on the research topic; therefore, the bibliographic search must be exhaustive. This phase comprises six parts: search,

import studies, study selection, quality assessment, data extraction, and data analysis. These parts and their sections are described below.

– *Search.* In this part, we specified the search strings for each of the sources selected in the planning phase (Table 2.7). These search strings are formed from the base string defined by the PICOC method and are used to start the search for papers in each bibliographic database.

| Data fusion in agricultural contexts and climate vulnerability assessments | |
|---|---|
| Source | Search String |
| Google Scholar | (“data fusion” OR “data integration”) AND (“crop production”) |
| IEEE Digital Library | ((“All Metadata”: “data fusion” OR “data integration”) AND “All Metadata”: “crop production”) |
| Science@Direct | (“data fusion” OR “data integration”) AND (“crop vulnerability” OR “crop production”) |
| Scopus | ( TITLE-ABS-KEY ( “data fusion” OR “data integration” ) AND TITLE-ABS-KEY ( “crop production” ) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) OR LIMIT-TO ( PUBYEAR , 2014 ) OR LIMIT-TO ( PUBYEAR , 2013 ) OR LIMIT-TO ( PUBYEAR , 2012 ) OR LIMIT-TO ( PUBYEAR , 2011 ) ) AND ( LIMIT-TO ( DOCTYPE , “ar” ) OR LIMIT-TO ( DOCTYPE , “cp” ) OR LIMIT-TO ( DOCTYPE , “re” ) ) AND ( LIMIT-TO ( SUBJAREA , “AGRI” ) OR LIMIT-TO ( SUBJAREA , “COMP” ) ) AND ( LIMIT-TO ( LANGUAGE , “English” ) ) |
| Springer Link | (“data fusion” OR “data integration”) AND (“crop vulnerability” OR “crop production”) |
| Multi-label classification in agricultural contexts and climate vulnerability assessments | |
| Source | Search String |
| Google Scholar | (“multi-label classification” AND agriculture) |
| IEEE Digital Library | (("All Metadata":"multi-label classification") AND "All Metadata":agriculture) |
| Science@Direct | (“multi-label classification” OR “multi-label” OR “multi-label learning”) AND (“agriculture” OR “crop” OR “agricultural”) |
| Scopus | ( TITLE-ABS-KEY ( “multi-label classification” OR “multi-label” OR “multi-label learning” ) AND TITLE-ABS-KEY ( “agriculture” OR “crop” OR “agricultural” ) ) AND ( LIMIT-TO ( SUBJAREA , “COMP” ) ) AND ( EXCLUDE ( DOCTYPE , “cr” ) OR EXCLUDE ( DOCTYPE , “ed” ) OR EXCLUDE ( DOCTYPE , “re” ) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) OR LIMIT-TO ( PUBYEAR , 2014 ) OR LIMIT-TO ( PUBYEAR , 2013 ) OR LIMIT-TO ( PUBYEAR , 2012 ) OR LIMIT-TO ( PUBYEAR , 2011 ) ) |
| Springer Link | (“multi-label classification” OR “multi-label” OR “multi-label learning”) AND (“agriculture” OR “crop” OR “agricultural”) |

**Table 2.7.** Search strings for each of the sources selected in the planning phase.

– *Import Studies.* In this second part, we collected meta-data from the studies found in the different sources (Appendix A). The number of studies per topic and source are presented in Figure 2.9.

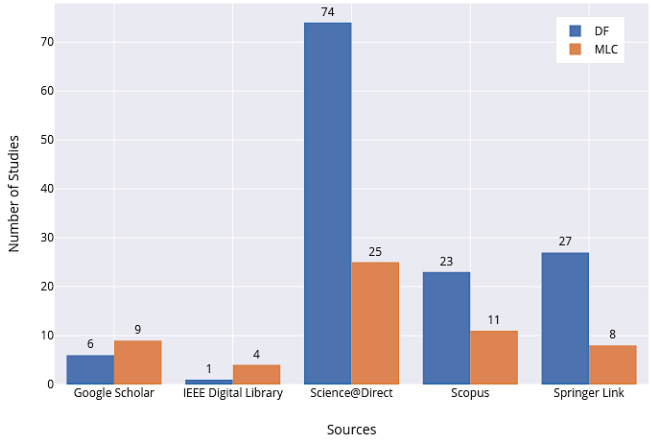



**Figure 2.9.** Number of studies collected around DF and MLC per source.

– *Study Selection.* In this part, we obtained a first selection of relevant studies for the research questions (primary studies). First, we eliminated duplicate studies, and then, we reviewed the articles by title and abstract. In this way, we classified all studies into the category of accepted (fulfill one of the inclusion criteria) or rejected (fulfill one of the exclusion criteria). A detailed list of these studies can be found in Appendix A.

– *Quality Assessment.* Here, we evaluated the quality of the accepted papers according to the parameters defined in the planning phase. We reviewed the articles in full and established which of these were of higher quality, eliminating those that did not satisfy a minimum threshold. Table 2.8 summarizes the scores for each accepted article and the details of this evaluation can be found in Appendix A.

| Data fusion in agricultural contexts and climate vulnerability assessments | | |
|---|---|---|
| Title | Reference | Quality Score |
| A generic ontological network for Agri-food experiment integration – Application to viticulture and winemaking | [58] | 6.5 |
| Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean | [59] | 6.5 |
| Sensor data fusion for soil health assessment | [60] | 6.0 |
| Progresses on data fusion technology of crop growth model and multi-source observation information | [61] | 6.5 |
| Plant localization and discrimination using 2D+3D computer vision for robotic intra-row weed control | [62] | 6.5 |
| PAID: Predictive agriculture analysis of data integration in India | [63] | 5.0 |
| Scalable pixel-based crop classification combining Sentinel-2 and Landsat-8 data time series: Case study of the Duero river basin | [64] | 7.0 |
| Multi-sensor Fusion of Remote Sensing Data for Crop Disease Detection | [65] | 6.0 |
| Multiple on-line soil sensors and data fusion approach for delineation of water holding capacity zones for site specific irrigation | [66] | 6.0 |

| | | |
|---|---|---|
| Monitoring oil palm plantations in Malaysia | [67] | 5.5 |
| Integration of in situ measured soil status and remotely sensed hyperspectral data to improve plant production system monitoring: Concept, perspectives and limitations | [68] | 6.0 |
| Improved maize cultivated area estimation over a large-scale combining MODIS-EVI time series data and crop phenological information | [69] | 6.5 |
| Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation | [70] | 6.5 |
| Field partition by proximal and remote sensing data fusion | [71] | 6.5 |
| Data integration for climate vulnerability mapping in West Africa | [72] | 7.0 |
| Assessment of the effectiveness of spatiotemporal fusion of multi-source satellite images for cotton yield estimation | [73] | 6.0 |
| A smart farming approach in automatic detection of favorable conditions for planting and crop production in the upper basin of Cauca River | [74] | 5.0 |
| AgriFuture: A New Theory of Change Approach to Building Climate-Resilient Agriculture | [75] | 5.0 |
| Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study | [76] | 6.5 |
| Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach | [77] | 6.5 |
| Data Fusion of Proximal Soil Sensing and Remote Crop Sensing for the Delineation of Management Zones in Arable Crop Precision Farming | [78] | 5.0 |
| Inverse Problems and Data Fusion for Crop Production Applications Targeting Optimal Growth - Fertilization | [79] | 5.0 |
| Soybean yield prediction from UAV using multimodal data fusion and deep learning | [80] | 6.5 |
| Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images | [81] | 6.5 |
| Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine | [82] | 6.5 |
| Large-scale data integration reveals colocalization of gene functional groups with meta-QTL for multiple disease resistance in barley | [83] | 6.0 |
| Multi-label classification in agricultural contexts and climate vulnerability assessments | | |
| Title | Reference | Quality Score |
| Multi-label learning for crop leaf diseases recognition and severity estimation based on convolutional neural networks | [84] | 6.0 |
| Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions | [85] | 6.5 |
| Machine Learning for Apple Fruit Diseases Classification System | [86] | 5.0 |
| Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based Land Transformation Model: an integrated ML-CA-LTM modeling framework | [87] | 7.0 |
| Multi-label class assignment in land-use modelling | [88] | 6.5 |
| Plant recommender system based on multi-label classification | [89] | 5.5 |
| Research on deep learning in apple leaf disease recognition | [90] | 6.0 |
| Deep Learning - a New Approach for Multi-Label Scene Classification in Planetscope and Sentinel-2 Imagery | [91] | 5.0 |
| A comparative study of land classification using remotely sensed data | [92] | 3.0 |
| AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms | [93] | 5.5 |
| Categorizing videos using a personalized category catalogue | [94] | 4.5 |
| Multi-label Classification of Big NCDC Weather Data Using Deep Learning Model | [95] | 4.5 |

**Table 2.8.** Quality assessment for each accepted article around DF and MLC topics. The rows highlighted in gray represents articles lower or equal to the cutoff score (minimum quality score = 5.0), which were not considered in the final phase of the review.

– *Data Extraction.* The aim of this part is to extract and collect, from each selected article, information to answer the research questions. The extraction

process must be homogeneous; therefore, we used the form created in the planning phase. Full data extraction form can be found in Appendix A.

– *Data Analysis.* This part presents the interpreted results obtained through the data extraction form. These analyses should be guided according to the research questions to be solved. First, using Figure 2.10, we identify the final articles by year since 2011 (after study selection and quality assessment). We reaffirmed the trend obtained through systematic mapping, with a predominance of studies corresponding to data fusion. Likewise, we evidenced the relevance that multi-label classification in agriculture has gained in recent years.

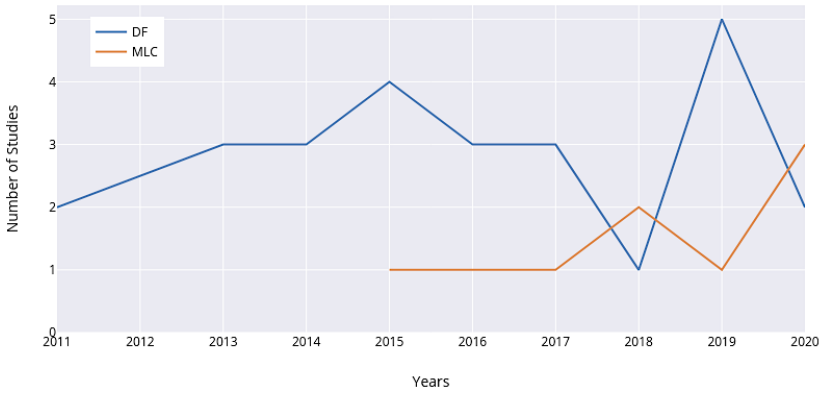



**Figure 2.10.** Number of articles selected by year of publication.

On the other hand, Figure 2.11 presents the percentage of affirmative and negative responses for each question on the data extraction form. Analyzing the number of DF and MLC studies applied in agriculture, we identified a low proportion using these techniques to support agricultural vulnerability assessments (14.3%). Likewise, although we found methodologies to apply DF or MLC, only 22.9% of the papers evaluate the performance of such methodology. The above implies focusing efforts to evaluate and validate the results of our data fusion strategy.

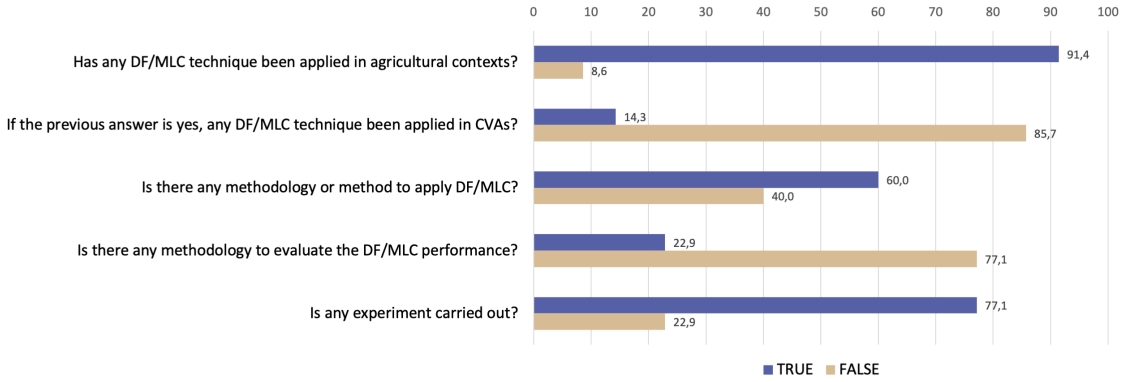



**Figure 2.11.** Percentage of affirmative and negative responses for each question on the data extraction form.

Finally, Figure 2.12 shows the percentage of studies according to experimental evaluation method. We identified the experimental evaluation methods such as case studies, performance testing, and experimental or user evaluation. The latter corresponds to the predominant method for developing experiments.

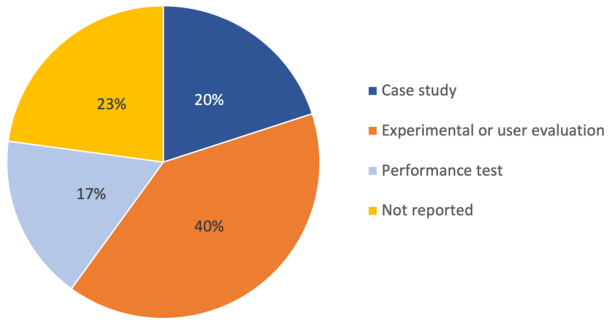



**Figure 2.12.** Percentage of studies according to experimental evaluation method.

After analyzing the previous results, we complemented this exploration with a synthesis of the most relevant works for our study around the DF and MLC topics.

*Synthesis of Data Fusion Studies*

Agricultural Data Fusion groups different approaches depending on the type of data sources, the crops involved, and the integration objectives. Most methods combine different types of satellite imagery (including Landsat-8, Sentinel-2, STRM, MODIS-EVI, and MODIS-NDVI) considering integration objectives such as possible planting areas for rice, soybeans, and corn [82]; estimation of cultivated areas for corn [69]; estimation of yield for corn, soybeans, and cotton [59], [73]; and classification of large crop areas [64]. Satellite images are also

combined with an in-situ, survey, or multi-sensor data to identify areas of climate vulnerability [72], determine variations in wheat production [77], detect areas suitable for tomato cultivation [71], and characterize the spatial complexity of soils [76].

Other approaches integrate multi-sensor and in-situ data for predicting wheat and other crop yields [61], [70], planning and monitoring of oil palm and barley plantations [67], [79], detecting crop diseases for fungicide applications [65], and determining soil health indicators [60]. Likewise, some works are focused on integrating historical data around crops such as production, yield, diseases, among others, to solve problems such as detection of genomic regions of pathogens in crops [83], management of viticulture and winemaking processes [58], production and yield estimation of sugarcane crops [63], and identification of crop management areas for application of agricultural inputs [78].

Finally, other studies integrate multispectral images from Unmanned Aerial Vehicles (UAVs) to determine soybean crop yield and improve crop production monitoring [68], [80]. In turn, Gai et al. [62] propose the integration of 2D and 3D images to recognize plants in different stages of growth. Likewise, we highlight two theoretical works closely related to our research. These studies propose to integrate climate, environmental, social, economic, cultural, political, and institutional data for decision making in smart farming contexts using Big Data technologies [74], [75].

*Synthesis of Multi-Label Classification Studies*

Research works around the MLC paradigm in agriculture are focused on classification for two issues such as land uses and crop diseases. Regarding land use classification, conventional models assign a single land use label to each spatial unit. Therefore, several approaches classify this coverage unit with several labels simultaneously (mixed land use). Shendryk et al. [91] propose combining deep learning models with MLC to classify atmospheric conditions and land use coverage from satellite images of the Amazon forest. In the same research line, Omrani et al. [87] developed an integrated modeling framework (multi-label learning, cellular automata, and land transformation models) to classify land uses in Luxembourg. Using data from this country, the same

authors [88] proposed to solve the same classification problem but using the K-Nearest Neighbors (KNN) technique in the MLC paradigm.

Crop diseases classification is another widely addressed problem. Convolutional Neural Networks (CNN) combined with MLC algorithms are used to detect simultaneous diseases in crops [84], [85], [90]. Likewise, Abd El-aziz et al. [86] detect diseases in apple fruit using the Multi-Label KNN (ML-KNN). This algorithm is also used to identify plants from different characteristics detected in photos [89]. Finally, we highlight the approach of Doshi et al. [93], which is the most related to our proposal. This approach presents *AgroConsultant*, an intelligent system to assist Indian farmers on which crop to plant depending on the planting season, geographic location, soil characteristics, and environmental factors such as temperature and rainfall. Algorithms such as Decision Trees (DT), KNN, Random Forest (RF), and Neural Networks (NN) were used for prediction tasks, selecting NN as the best technique with 91% accuracy.

After analyzing the advantages and contributions of previous works around DF and MLC in agricultural contexts and agricultural vulnerability assessments, the shortcomings found from the systematic literature review are mentioned below.

*Shortcomings*

- Despite the different techniques employed in data fusion, the reviewed works do not consider an integrated multi-label approach to solve several issues around agricultural vulnerability assessments.
- Some works consider integrating information from different dimensions such as economic, social, political, among others, to solve different issues in agriculture. However, these approaches are limited only to a research proposal, establishing their implementation for future work.
- Although the reviewed approaches apply different experimental evaluation methods, most papers do not perform a comprehensive evaluation of data fusion methodologies.

***Reporting***

In this part, we prepared a final report with all the detailed steps corresponding to the previous systematic review. For more details, this report can be found in Appendix A.

## 2.3. Summary

In this chapter, we defined the main concepts that guide this doctoral thesis and analyzed the related works through a systematic mapping and a systematic literature review. Initially, we established the scenario for Climate-Smart Agriculture (CSA) and Agricultural Vulnerability Assessments (CVA), where Data Fusion (DF) and Multi-Label Classification (MLC) became the fundamental thematic cores for our research. Subsequently, we answered the research questions formulated both in the systematic mapping and in the systematic literature review, and finally we obtained the research gaps, which justified the development of our proposal.

# Chapter 3

# Data Fusion Strategy

This chapter describes the main components of the proposed data fusion strategy. In each component, we proposed different techniques to cover three essential aspects such as data preparation, data integration, and data analysis. In data preparation, we established a process for cleaning up the collected data sources. In data fusion, we define a scheme to combine data sources from different dimensions. Finally, in data analysis, we apply several multi-label classification algorithms to determine crop adaptability. In addition, we describe other complementary components such as process refinement, database management, and human/computer interface.

## 3.1. Data Fusion Strategy Overview

The proposed strategy is based on the Joint Directors of Laboratories (JDL) data fusion model [38], one of the most widely used models for data fusion tasks. We selected this model considering its functional approach, as opposed to other process-oriented models. JDL facilitates understanding data fusion techniques and communication between stakeholders to achieve common objectives. This premise implies that the data fusion strategy's actions do not always follow a strict or canonical order to achieve the final goal. In this sense, JDL categorizes the data fusion functions by levels according to different types of problems. *Level 0 - Data Assessment* (estimation of signal or object-observable states), *Level 1 - Object Assessment*. (estimation of entity states from inferences about observations), *Level 2 - Situation Assessment* (estimation of entity states based on relations among entities), *Level 3 - Impact Assessment* (estimation of effects on situations of planned actions,

*Level 4 - Process Refinement* (adaptive data acquisition and processing to support mission objectives).

Figure 3.1 presents an overview of the data fusion strategy. Adaptations and contributions in each level are described below.

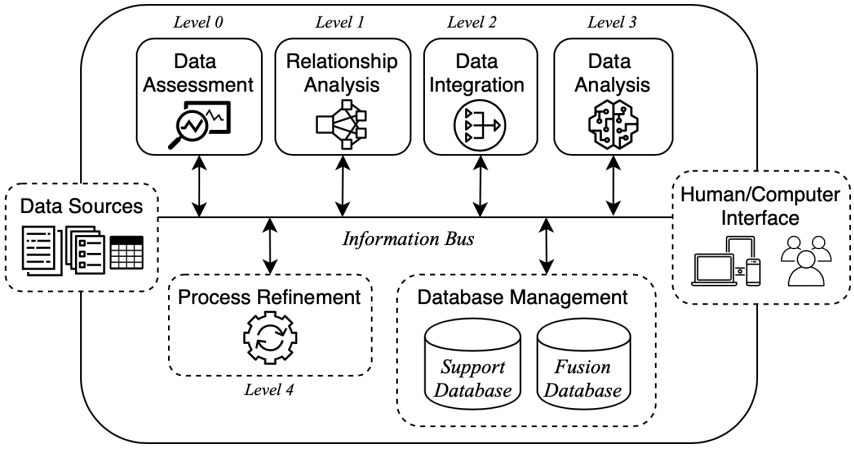



**Figure 3.1.** Overview of the proposed data fusion strategy. Adapted from JDL data fusion model [37].

- *Level 0 - Data Assessment.* Evaluation of gathered data sources to establish an overview of data quality.
- *Level 1 - Relationship Analysis.* Identification of implicit relationships between data sources through analysis of temporal and spatial scales.
- *Level 2 - Data Integration.* Production of new data sets (combined data sources) with more synthesized and reliable added value.
- *Level 3 - Data Analysis.* Implementation of different multi-label learning techniques or algorithms to train a set of models and estimate one or more target variables.
- *Level 4 - Process Refinement.* Cross-level monitoring of data fusion strategy levels.

From this strategy, we highlight different contributions such as i) a formal process for data preparation in climate-smart agriculture environments, ii) a method for associating and integrating data sources from different dimensions of agricultural vulnerability, iii) a training and evaluation scheme for multi-label classification models, and finally, iv) the implementation of a data lake for storage management as an improvement to data fusion models.

## 3.2. Data Assessment (Level 0)

Data assessment or Level 0 evaluates the gathered data sources and defines a preparation process needed to improve data quality. This level's components present a modular scheme where the results obtained in a module allow feedback to the next one. Level 0 is composed of three phases: data sources evaluation, data sources preprocessing, and variables prioritization. These phases represent the data cleaning process and involve refinement through collaborative assessments between data experts and other areas of knowledge. Figure 3.2 presents the task flow required for data preparation. The phases and its respective modules are described in the following paragraphs.

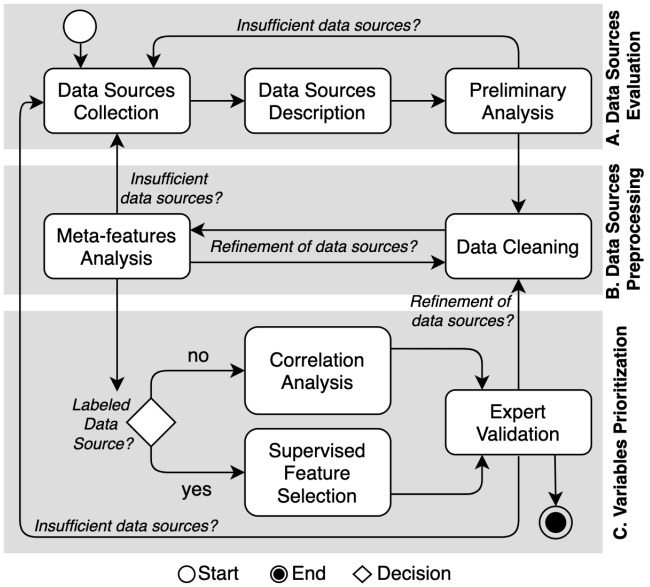



**Figure 3.2.** Phases and modules of Data Assessment (Level 0) in the data fusion strategy.

### 3.2.1 Data Sources Evaluation

The main objective of this phase is to determine the availability of official information around the study area. Here, data are collected from sources with three fundamental characteristics: open, official, and public. These must be classified into groups corresponding to vulnerability dimensions. Subsequently, all sources should be described considering the provider entity and the main theme of the dataset. In addition, basic metadata is extracted, such as size, number of attributes, number of instances, time window, among others. Having consolidated these metadata, a preliminary analysis of all data sources is finally developed. These tasks are described below.

***Data Sources Collection***

The use of open data is a key aspect of this study. In this sense, searching for data sources is not only a task oriented to the web scrapping of official organizations on their websites, but also to the results of climate vulnerability assessments previously developed in the study area. Raw datasets must be collected and subsequently, as the result of this step, classified in vulnerability dimensions. These dimensions must be defined according to the topics related to each dataset, or in some cases, these are already defined in the CVAs.

***Data Sources Description***

After the data collection, these sources and their providers must be described. The idea in this step is to collect information about the sources and entities that provided this data. It is important to describe, in general, features such as data source identifier, data source name, entity name, mission, vision, area of influence, target population, data provided, among others. A document with the description of all data sources must be generated at the end of this step.

***Preliminary Analysis***

This step is intended to provide an overview of all data sources. Each dataset has specific metadata that must be consolidated in a descriptive table. The metadata considered for this preliminary analysis were the following: identifier (acronym previously assigned in the description of data sources), format (original file type of the data source), time window (the total sample period), number of instances (number of records contained in the dataset), number of attributes (number of variables sampled per dataset), time scale (sampling interval like daily, monthly, annual, among others), spatial scale (relative length, area, distance, and size of the sampling area), and size of the data source (number of bytes in the stored file). When the table is consolidated, comparative charts can be obtained from the collected metadata. In this study, we propose to use the number of instances and the number of attributes to determine the amount of data available. If there is not enough information at the end of this phase, the analyst can return to the data collection step.

### 3.2.2 Data Sources Preprocessing

This phase focuses on identifying data quality problems and applying different cleaning steps to solve them. Data sources preprocessing includes an initial stage to define the steps to follow during the cleaning process. Similarly, identify the possible

problems for each data source and apply the necessary steps depending on the problem to be addressed. Finally, the results of this phase are the pre-processed datasets and a series of meta-features to determine the level of cleanliness in the data sources (before and after the preprocessing). This analysis allows refining the data preparation process through a new iteration. If necessary, collect new data sources to feedback the process. The steps of this phase are described below.

***Data Cleaning***

This module aims to describe the possible cleaning steps required by each raw data source. To pre-process a data source, the steps presented in Figure 3.3 must be executed according to the task to be performed (classification or regression respectively) [96]. These tasks must be executed depending on the dataset characteristics. The above implies that if the data source is labeled with a target class, the entire task flow is applied. While an unlabeled dataset only the imputation, outlier detection, removal of duplicate instances, and dimensionality reduction steps are considered. Next, the cleaning task flow is described.

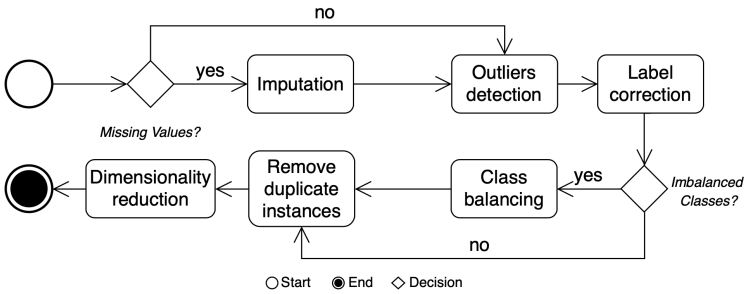



**Figure 3.3.** Data cleaning process for classification models. Quoted from [96].

- *Check missing values.* Inspection of missing values in an attribute indicating a problem in the data source, e.g., defective measurements, sensor failures, incomplete surveys, or data transfer problems [97].
- *Outlier detection.* An observation can be univariate or multivariate. This is called atypical value or outlier when it seems to be inconsistent for the remaining observations, i.e., it deviates noticeably from others [98], [99].
- *Label correction.* It consists in detecting those self-contradictory instances, i.e., duplicate samples which have different class labels. This implies a lack of harmony among dataset elements [100].

- *Class balancing.* This task is applied when data distribution is unequal among classes or target variables. If a dataset is unbalanced, accuracy and quality of learning tend to decrease [101].
- *Removal of duplicate instances.* Identification of instances and redundant attributes which can negatively affect the performance of the models [102].
- *Dimensionality reduction.* If a data source contains a large number of features, a future predictive model will decrease its performance by overfitting. By reducing the dimensionality, the number of features decreases without affecting the model performance [103].

As mentioned earlier, the previous cleaning steps are applied to raw data sources. These tasks are selected according to the quality of each data source. Here, software tools to support this process must be defined. There are several software tools with different features according to the user's expertise. From a spreadsheet in Excel to statistical environments and specialized programming languages such as R, Python, among others.

***Meta-Features Analysis***

After applying the cleaning steps, checking data quality is an essential aspect. In this sense, different metrics must be used to determine adequate data distribution. These metrics allow us to establish a comparison of data quality before and after the cleaning process. These should be summarized in a table for all data sets. At this point, the analyst has three options. First, in case of no improvement in data distribution, reapply the cleaning steps. Second, in case of insufficient data sources, return to the data collection step. If neither of the above two options is used, continue with the next and last phase (variables prioritization). Metrics used in this module are described below.

- *Data dimensionality.* This metric refers to the ratio between the number of attributes and instances of a dataset [104].
- *Missing values ratio.* Represents the proportion of missing values in all dataset values (attributes x instances) [105].
- *Duplicate instances ratio.* This refers to the proportion of duplicate instances of the total instances in a dataset [106].
- *Mean absolute skewness.* Skewness (S) measures the symmetry of data distribution. If $S > 0$, the distribution curve is tilted to the left. If $S < 0$, the

slope is to the left. Whereas if S $\approx$ 0, it approaches a normal distribution. Mean Absolute Skewness represents the average skewness of all numeric attributes in the dataset [107].

- *Mean absolute kurtosis.* Kurtosis (K) is a way to measure the flatness of data distribution. If K $>$ 3, the distribution curve is considered Leptokurtic (wide/fat tails). If K $<$ 3, it is Platykurtic (thin tails). Whereas if K $\approx$ 0, it is called Mesokurtic (normal tails). Mean Absolute Kurtosis represents the average kurtosis of all numerical attributes in the dataset [107].
- *Mean attribute entropy.* Entropy is also known as a measure of impurity in a set of elements. Impurity refers to the distribution of the decision categories in the set. This set is pure if it is composed of the same class tag, and is considered impure if it contains different tags in the same proportion. Mean Attribute Entropy represents the average entropy of all nominal attributes in the dataset [107]. In this study, the maximum likelihood was the method used to estimate entropy.

### 3.2.3 Variables Prioritization

The focus of this last phase is to determine the most important variables or attributes in each data source. This task is established through a ranking and the method to determine the importance of variables depends on the type of data source (labeled or unlabeled). In the last step, the ranking is validated through expert knowledge around the subject of each dataset. Through this latest expert evaluation, the datasets can be pre-processed again or new datasets can be collected. The two modules of this phase are described below.

#### *Feature Selection*

This step determines the prioritization process to be used considering whether a dataset is labeled or not. In other words, if there is a response variable (often called a label), in each data source. If the dataset is labeled, a supervised feature selection process is applied, otherwise, a correlation analysis approach is used. The two approaches to feature selection used in this study are mentioned below.

- *Supervised approach.* It is the most common practice and there are three main categories of feature selection algorithms: wrappers, filters, and embedded methods. In this study, we propose the use of the Random Forest (RF) and Logistic Regression (LR) methods to select the most representative variables

[108]. Both algorithms are included in the embedded strategies which combine the advantages of filter and wrapping methods. We selected these methods considering their popularity around machine learning algorithms. In addition, these are very successful due to high predictive performance, low overfitting, and easy interpretation. This last aspect is characterized by calculating the contribution of each variable to the decision in a simple way [109].

– *Correlation Analysis.* These methods represent a greater challenge to determine the importance of variables, however, this approach can also establish some advantages. Expert criteria are not necessary to tag datasets and these methods can perform better even when no prior knowledge is available [109]. In the present research work, the process of selecting attributes in this analysis is guided by the correlation between variables in a dataset. Those variables with the highest correlation coefficient are considered more relevant.

***Expert Validation***

From the perspective of the application domain, it is important to corroborate the results with expert knowledge. This would reduce the uncertainty of the predictive models applied in the future. In this sense, the process described in this chapter refines the set of prioritized variables. It should be noted that no variable is discarded considering that, in the future, they may be useful for other analyzes.

## 3.3. Relationship Analysis (Level 1)

In level 1, we identify the implicit relationships between data sources by analyzing the temporal and spatial scales. For this purpose, we establish a spatio-temporal characterization process where all possible relationships are consolidated and analyzed to verify their relevance in the data fusion strategy. Also, we build a relationship scheme to guide the data sources integration at level 2. The components of level 1 are presented in detail below.

### 3.3.1 Spatio-Temporal Characterization of Data Sources

In this component, we identify possible spatio-temporal relationships between data sources. Meta-features determine the temporal relationships in the data sources such as time window, temporal scale, and sampling intervals. On the other hand, spatial relationships are focused on the area covered by a territorial division. The spatial scales handle these divisions, including farms, villages, municipalities, states, regions,

and countries. Similarly, we also identify other types of relationships inherent to the data fusion objective. In our case, crops in a specific area could be an indication of these possible additional relationships. Finally, all the Spatio-Temporal Meta-Features (STMF) should be consolidated in a summary table for further analysis.

### 3.3.2 Data Source Relationship Scheme

The data source relationship scheme corresponds to a matrix that establishes the strength of the possible relationships. This matrix is generated by comparing the STMF consolidated in the previous component. The data source relationship scheme is described below and presented in Figure 3.4.

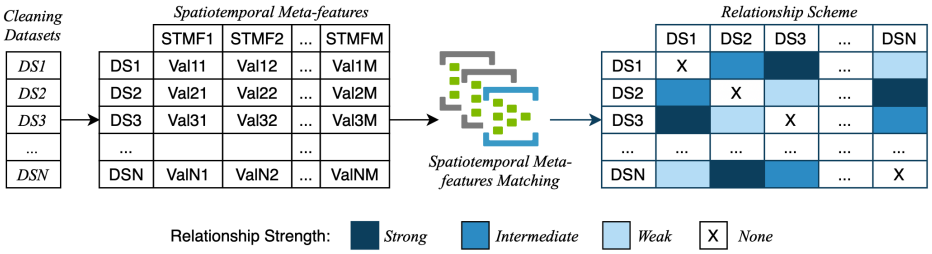



**Figure 3.4.** Relationship Scheme. DS: Data Source, STMF: Spatial-Temporal Meta-Feature, Val: Attribute Value.

#### *Creating the Relationship Scheme*

On the left side of Figure 3.4, the consolidated table of STMF is shown. The rows represent the pre-processed data sources of size N. The columns correspond to the STMF of size M. We transform the table into a matrix of N x N dimensions, where the color of each cell represents the relationship strength between the data sources (right side of Figure 3.4). To obtain the strength of a relationship, we compare the STMF of one data source with the others using a similarity function. This function returns a value between 0 and 1. The values near 1 represent a high similarity, and the values near 0 a low similarity. We define the following ranges and colors to assign the strengths of the relationships: 0 to 0.25 (light blue, weak relationship), 0.25 to 0.75 (blue, intermediate relationship), and 0.75 to 1 (dark blue, strong relationship).

#### *Relationship Scheme Analysis*

After obtaining the relationship scheme, we can integrate a table with guidelines about the possible relationships identified and how the data sources are involved. These guidelines indicate which STMFs we should consider at level 2 (data

integration), for example, the resulting time windows, the regularity of sampling intervals, and specific attributes that would facilitate integration.

## 3.4. Data Integration (Level 2)

Level 2 corresponds to data integration, a reduction process to generate new data sets (combined data sources) with a more synthesized and reliable added value. At this level, we select a data integration approach and apply a method to combine the data sources based on Entity Matching [110]. Finally, we label the resulting datasets according to the final data fusion strategy's objective. These steps are described in detail below.

### 3.4.1 Selecting the Integration State

Depending on the nature of the datasets and the statistical problem to be solved, data from different sources can be integrated into three different states: early, intermediate, or late [111]–[113]. These states of data integration are described below and presented in Figure 3.5.

- *Early Integration.* In this state, a single feature space groups the data sources' attributes without changing their format and nature. However, a disadvantage lies in the increased dimensionality of the combined datasets.
- *Intermediate Integration.* Before being combined, intermediate integration transforms the attributes of all data sources into a common feature space. Then, a model learns a joint representation of several data sets and merges them during a later stage of analysis.
- *Late Integration.* Each dataset trains one or more models separately, and an assembly method combines the final results. This state has advantages such as the free choice of the best algorithm and the parallel analysis of each dataset.

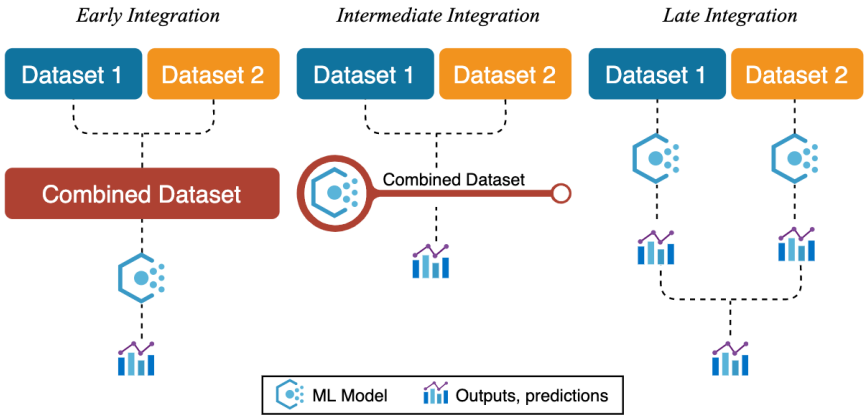



**Figure 3.5.** Stages of data integration. Adapted from [111].

### 3.4.2 Integrating Data Sources

The basic idea is to ensure that the matching attributes have the same structure and their content follows the same formats. Next, we apply indexing or blocking to reduce the computational effort when comparing record pairs, i.e., using highly tolerant similarity measures to filter out those record pairs that are "obvious" non-matches. After that, only those not pruned by the indexing or blocking step are inspected in record pair comparison. Finally, we apply a function to combine the records that have been matched.

### 3.4.3 Labeling Combined Data Sources

Each record in the combined datasets is finally labeled with one (single label) or more (multi-label) target variables or classes. This process depends on the specific problem addressed in the data fusion strategy and must be guided by expert knowledge. In this sense, we can apply labeling techniques such as manual or automatic labeling. Manual labeling requires supervision of one or more experts in the specific field, who assign the labels to the respective samples. This process can also be supported by reviewing the literature in the knowledge areas around the data sources. On the other hand, in automatic labeling, different clustering algorithms can be applied to find groups representing common labels or classes.

### 3.4.4 Exploratory Analysis in Multi-Label Datasets

After combining and labeling, data sources require exploratory analysis to determine the effectiveness of data integration before to the last level of the data fusion strategy (level 3 - data analysis). In the case of Multi-Label Datasets (MLD), the exploratory analysis is based on [44] and [41], which define a set of measures to describe the

combined data sources. These measures provide information about the data distribution and the possible behavior of a classification algorithm or a preprocessing technique. These metrics are described below.

- *Discarded Attributes.* The number of irrelevant attributes discarded from the MLD.
- *Number of Attributes.* The total number of attributes in the dataset (not including discarded attributes).
- *Number of Instances.* The total number of samples/records/rows contained in the MLD.
- *Number of Inputs.* The number of predictive attributes or independent variables.
- *Number of Labels.* The number of target variables or classes.
- *Number of Labelsets.* The number of label combinations or sets of labels assigned to each row.
- *Number of Single Labelsets.* The total number of unique labelsets in the MLD.
- *Maximum Frequency.* The number of appearances of the most common label in the MLD.
- *Cardinality.* The average number of relevant or active labels per instance [41]. From the Eq. (3.1), let $D$ be an MLD, $n$ the number of instances, $k$ the number of labels, and $Y_i$ the labelset corresponding to the *i-th* data sample.

$$Card(D) = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \tag{3.1}$$

- *Density.* The average number of relevant or active labels per instance regarding the total number of labels [41] as shown in Eq. (3.2).

$$Dens(D) = \frac{1}{k}\frac{1}{n}\sum_{i=1}^{n}|Y_i| \tag{3.2}$$

- *IRLbl.* The Imbalance Ratio of a Label, i.e., the higher the IRLbl, the larger the imbalance. This measure allows identifying which labels are minority or

majority [114]. In Eq. (3.3), let $D$ be an MLD, $L$ the full set of labels in it, $y$ the label being analyzed, and $Y_i$ the labelset of *i-th* instance in $D$.

$$IRLbl(y) = \frac{\max_{y' \in L}(\sum_{i=1}^{|D|} [\![y' \in Y_i]\!])}{\sum_{i=1}^{|D|} [\![y \in Y_i]\!]} \tag{3.3}$$

– *MeanIR.* Average IRLbl for an MLD. This measure estimates the global imbalance level of an MLD [114] as shown in Eq. (3.4).

$$MeanIR = \frac{1}{|L|} \sum_{y \in L} IRLbl(y) \tag{3.4}$$

– *SCUMBLE.* The Score of ConcUrrence among iMBalanced LabEls. The lower the SCUMBLE, the lower the relationships between the imbalanced labels in the MLD [115]. Eq. (3.5) shows the SCUMBLE per instance, while Eq. (3.6) presents the global SCUMBLE for the MLD.

$$SCUMBLE_{ins} = 1 - \frac{1}{\overline{IRLbl_i}} \left( \prod_{l=1}^{|L|} IRLbl_{il} \right)^{(1/|L|)} \tag{3.5}$$

$$SCUMBLE(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} SCUMBLE_{ins}(i) \tag{3.6}$$

– *SCUMBLE.CV.* Coefficient of Variation of the SCUMBLE value. A high SCUMBLE.CV value represents a major difference between label relationships [115]. Eq. (3.8) presents the SCUMBLE.CV, which is based on Eq. (3.7).

$$SCUMBLE_{\sigma} = \sqrt{\sum_{i=1}^{|D|} \frac{(SCUMBLE_{ins}(i) - SCUMBLE)^2}{|D| - 1}} \tag{3.7}$$

$$SCUMBLE.CV = \frac{SCUMBLE_{\sigma}}{SCUMBLE} \tag{3.8}$$

– *TCS.* Theoretical Complexity Score evaluates the theoretical complexity of an MLD. The higher the TCS values, the more challenging to train a predictive model from the MLD [116]. As shown in Eq. (3.9), let $f$ be the number of input features, $k$ the number of labels, and $ls$ the number of distinct labelsets. The logarithm of the product of these three factors provides the TCS of an MLD.

$$TCS = log(f \times k \times ls) \tag{3.9}$$

## 3.5. Data Analysis (Level 3)

In data analysis or Level 3, we apply several techniques or machine learning algorithms to train a set of models and estimate one or more target variables. In this sense, we use different metrics to evaluate the trained models' performance and select the best ones. We apply an Analysis of Variance (ANOVA) [117], which identifies statistically significant differences among model performances. Finally, we validate the models' results with actual data, i.e., data from a real scenario.

### 3.5.1 Model Training Scheme

In this component, we generate several predictive models and train them from both the Combined Data Sources (CDS) at level 2 and a set of variations of those sources. These variations correspond to modified versions of a specific CDS, and these are mentioned below.

– *Original CDS.* An initial version of the combined data source without modifications.
– *Decoupled CDS.* A majority label frequently appears in instances, while a minority label appears rarely. When the majority and minority labels coincide in the same instance, the minority labels are more difficult to classify due to the majority's bias. To separate the labels, we apply the REMEDIAL algorithm (REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels) [118] to the CDS obtaining a new version of this source (Decoupled CDS). The number of instances increases according to the proportion of

instances containing both majority and minority labels through this technique. REMEDIAL is recommended for datasets with a high SCUMBLE value [119].

- *Infrequent Positive Label Removal (IPLR).* Combined data sources excluding infrequent positive labels (labels with value 1), we must define the number of excluded labels according to the frequency distribution.
- *Skewness Labels Removal (SLR).* Combined data source excluding skewness labels, i.e., majority and minority positive labels at a defined threshold according to the frequency distribution.

After obtaining the CDS variations, we apply a combination of a multi-label classification strategy plus a machine learning algorithm to each CDS. Through this combination, we generate a set of trained models to estimate the values of one (single label) or several target variables (multiple labels). MLC strategies transform a dataset to apply a base algorithm. These strategies include Binary Relevance (BR) [49], Binary Relevance Plus (BRPLUS) [50], Ensemble of Classifier Chains (ECC) [51], Label Powerset(LP) [48], Hierarchy Of Multi-label classifiER (HOMER) [52], and Random k-labelsets (RAKEL) [53]. On the other hand, Random Forest (RF), Support Vector Machines (SVM), K Nearest Neighbor (KNN), Sequential Minimal Optimization (SMO), C5.0 Decision Trees (C5.0), Naive Bayes (NB), eXtreme Gradient Boosting (XGB), Classification and Regression Trees (CART), Majority Class Prediction (MAJORITY), and Random Prediction (RANDOM) represent the base algorithms [43]. Figure 3.6 presents the generation and training of $S$ models from the combined data sources (1 to $N$ CDS), the CDS variations (1 to $M$ variations), the MLC strategies (1 to $P$ strategies), and the basic machine learning algorithms (1 to $Q$ algorithms), where the number of models is represented by $S = N*M*P*Q$.

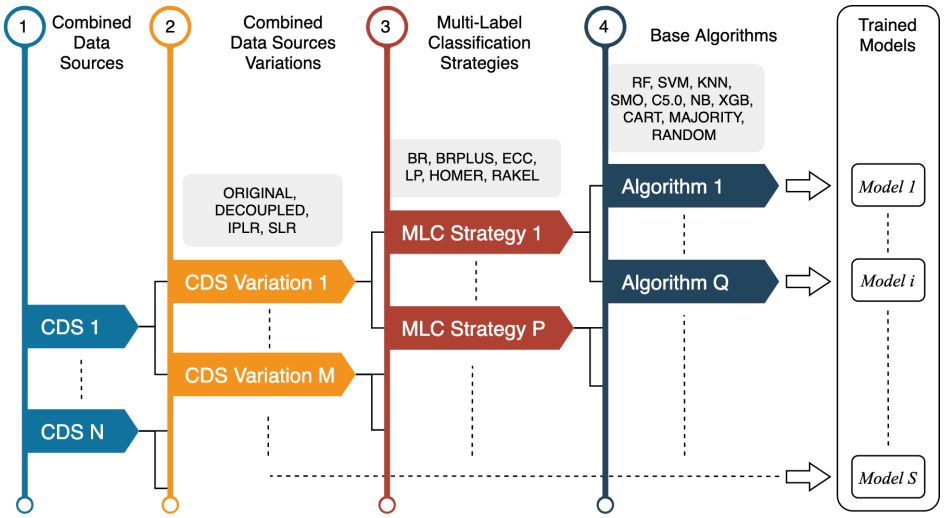



**Figure 3.6.** Model generation and training scheme for combined data source variations applying different multi-label classification strategies and base algorithms.

### 3.5.2 Model Performance Evaluation

To evaluate model performance with test data, we use different metrics for both traditional supervised learning and multi-label learning [42], [120]. Metrics such as Accuracy, Recall, Precision, F1-score, Area Under the ROC Curve (AUC), among others, are included in supervised learning. While multi-label learning involves metrics such as Hamming-Loss, Ranking-Loss, One-Error, among others. On the other hand, the above metrics could eventually generate very approximate results among models. To address this issue, we apply an Analysis of Variance (ANOVA) [117], which allows us to determine possible significant differences in model evaluation results. We considered several aspects such as metrics distribution (normality test), homogeneity of variance across groups (homoscedasticity), paired tests, and post-hoc comparisons to apply the ANOVA model.

### 3.5.3 Model Validation

To determine if the data fusion strategy is applicable in a real agricultural environment, we validate the models' best estimates with actual data. These data can be obtained in the first stage of the fusion strategy (data collection) or later, depending on the mentioned strategy's objectives. Finally, different techniques for comparing results must be defined in this validation component, again emphasizing the fusion strategy's objective.

## 3.6. Process Refinement (Level 4)

Process refinement is a transversal level that monitors each of the components or levels in the data fusion strategy. Level 4 represents the planning, control, and allocation of resources to tasks. This refinement also includes experts in different knowledge domains, which provide constant feedback to validate the data fusion strategy. In this sense, we can also use CVAs' results in this validation process as a benchmark.

## 3.7. Database Management

This component manages two databases, the support database and the fusion database. The first one stores the raw data sources and the second one stores the combined data sources. The storage also supplies additional information about the datasets, dimensional characteristics, characterization of the agricultural area, and model calibration. Similarly, this component manages the data ingestion, i.e., the process of flowing data from its origin to one or more datastores such as a data lake, relational databases, search engines, among others. This study is oriented to the development of an innovative data fusion strategy using new technologies; therefore, we managed the databases by implementing a Data Lake.

A data lake refers to a massively scalable storage repository which contains a large amount of raw data that remains stored until needed [121]. These repositories handle the Schema on Read following the ELT steps (Extract, Load, and Transform), where the schema and data requirements are defined once they are consulted [122]. Our data lake implementation is composed of three steps. The first one is data ingestion, or the process of flowing data from its source to one or more data stores. The second is data storage, where each item, stored in the data lake, is indexed and tagged with a unique identifier, known as a Managed Information Object (MIO) [123]. In the last step, the previously stored MIOs are transformed from queries to obtain the datasets used in the data fusion strategy.

## 3.8. Human/Computer Interface

This component presents the information obtained from the analysis of the different datasets. This software tool allows the user to consult the results according to the objective, which has been initially defined in the data fusion strategy. The information

must be presented considering each user type as a farmer, technician, extensionist, researcher, or decision-maker. We must orient the tool towards how the user can take advantage of all the information visualized in the respective knowledge domain.

## 3.9. Summary

In this chapter, we presented an overview of the proposed data fusion strategy. We described the main aspects concerning the levels and components of this strategy. These levels covered the three fundamental pillars of our proposal as data preparation, integration, and analysis. The following chapters will present the results through a case study at the upper Cauca river basin guided by this data fusion strategy.

# Chapter 4

# Data Assessment

This chapter presents the results of data assessment process through a case study. The steps mentioned in the previous chapter for data preparation (Level 0) were applied to the southwestern region of Colombia in the upper Cauca river basin, Cauca zone. Initially, we describe the data sources used in this research, the study area, its characteristics, and a previous CVA to contextualize the subsequent data analyses. Finally, we show the evaluation and the most significant results in each module.

## 4.1. Study Area

The Cauca River is the primary water source in the western region of Colombia. This river extends from the area called *"Macizo Colombiano"* (approximately 3,200 meters above sea level - m.a.s.l.) to the Magdalena River in the north of the country, covering about 1,360 km through nine regions from south to north. The upper Cauca river basin (Figure 4.1) has approximately 23,000 km$^2$, and 32% corresponds to Cauca, 47% to Valle del Cauca, 13% to Risaralda, and 8% to Quindío. In this research, we focused specifically on the Cauca zone, where the altitude varies between 4,700 m.a.s.l. at the summit of the Puracé volcano, and 950 m.a.s.l. in the Cauca River's alluvial valley (approximate area of 7,368 km$^2$). In this subbasin, agriculture represents a significant percentage of the Colombian economy and benefits about 23 municipalities. Even the country's food security, in a certain percentage, is directly affected by agricultural production in this area. The food demand of urban centers is largely supplied by small-scale commercial farms of coffee, beans, corn, cassava, fruit trees, vegetables, medicinal plants, livestock, and fish farming. In social and economic

terms, the rural economy is significant and a primary source of food security both in the region and in the country [21].

Productivity zones (agricultural, livestock, industrial, and forestry) that consume environmental goods and services, cause an imbalance in this basin, risking permanence over time and soil productivity. Additionally, its growth outside a fundamental ecological matrix at the regional level intensifies problems related to flood threats. On the other hand, the lack of timely and accurate information for decision-making in agricultural contexts, as well as the few monitoring systems, is one of the main barriers to properly assessing agro-climatic vulnerability. In this sense, climate vulnerability assessments play an essential role in managing potential risks to food security in this area. The AVA methodology (Agriculture, Vulnerability, and Adaptation) [124] quantified and analyzed productive systems and land use to support integrated planning of productive processes and sustainability.

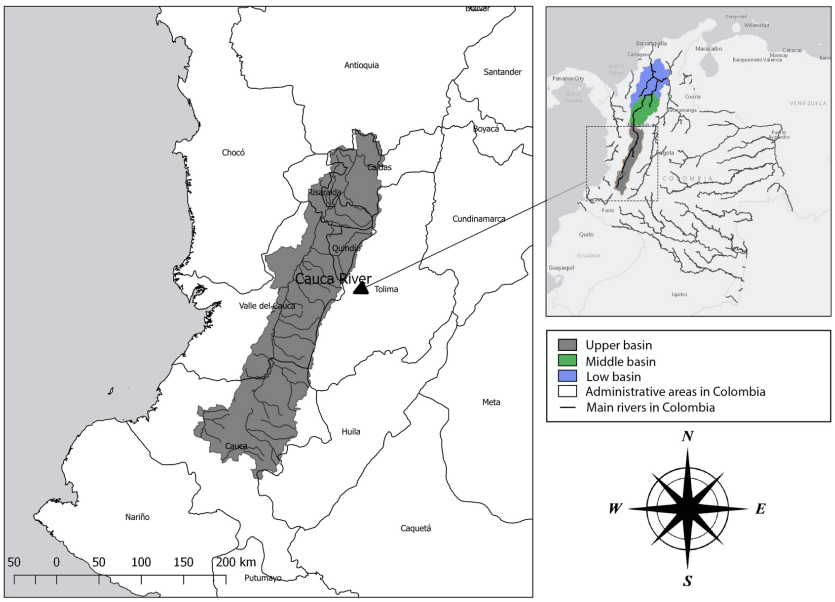



**Figure 4.1.** Coverage zone of the upper Cauca river basin. Quoted from [38].

The AVA approach was to measure the vulnerability of both the principal crops and the municipalities in the study area in the face of climate change's new reality. In AVA, the authors defined a set of indicators to measure climate vulnerability at UCRB. Different actors (farmers, productive associations, universities, government institutions, research centers, among others) identified four agro-climatic vulnerability dimensions as biophysical, economic-productive, sociocultural, and political-

institutional. Indicators were associated with these four dimensions to finally calculate the vulnerability of cacao, bean, potato, sugarcane, banana, and coffee crops. These crops were selected, considering that they cover about 1,067,000 hectares of the UCRB, and more information was available for such calculations [124]. The four dimensions of AVA are described below.

- *Biophysics.* It refers to a set of structures and relationships of the environmental context. It includes physicochemical and water quality analysis of lotic and lentic ecosystems existing in the UCRB, geomorphology, soils, vegetation cover, water bodies, ecosystems, and agro-ecosystems services corresponding to the previous elements.
- *Economic-Productive.* Information related to agricultural production and cultivated areas, agricultural systems productivity, municipal income that depends on the agricultural activity, and processes of agricultural resources optimization (results of national agricultural census).
- *Sociocultural.* Characteristics of a household, for example, those with female heads of household and older adults. Unsatisfied basic needs of communities, literacy levels, and the risk to suffer a natural catastrophe in a population (semi-structured surveys).
- *Political-Institutional.* Evaluation of the institutional, administrative, and fiscal management capacity of public institutions at the local, regional, and national levels. Besides, the priorities and percentage of investment of the budget of the territorial entities, in subjects related to environmental sustainability and the agricultural sector.

## 4.2. Data Sources Evaluation

This section describes the first phase of the data preparation process (Figure 4.2), applied to the case study in the upper Cauca river basin, Cauca zone.

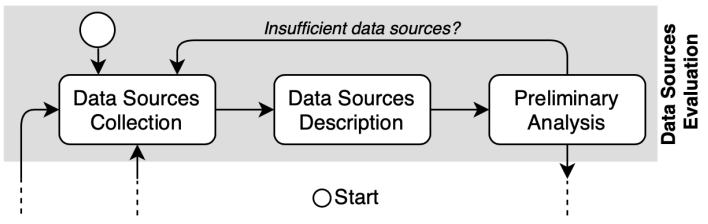



**Figure 4.2.** First phase of the data preparation process.

### 4.2.1 Data Sources Collection

Considering the upper Cauca river basin as the study area of this research work, different data sources with public access were required as a fundamental input. After a first search process using a secondary data collection method, 16 data sources were identified and the respective datasets were downloaded to a local repository. Finally, the four dimensions defined in the case study were complemented with some of the SAFA indicators (Sustainability Assessment of Food and Agriculture systems) proposed by the Food and Agriculture Organization of the United Nations (FAO) [125]. Figure 4.3 presents these data sources classified by dimension.

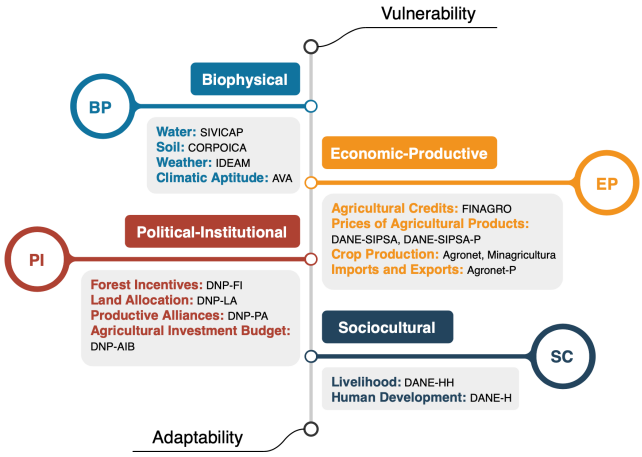



**Figure 4.3.** Classification of datasets according to the four dimensions identified in the AVA methodology.

The main source to find information initially pointed to the Web. Additionally, complementary information was extracted from the results of different vulnerability analyzes previously developed in the study area, in this case from the AVA methodology. We consulted the websites of official entities to extract data. In some cases, data were available directly from the websites, in others, these sites contained links to more specialized web portals where information could be found. *Datos Abiertos Colombia* [126] and *Archivo Nacional de Datos (ANDA)* [127] are governmental web platforms that collect public data in Colombia from different sectors, without restrictions and for specific use in research projects. These sources highlight the use of official open data to support this study. The Open Data Barometer [128] is an indicator that aims to measure the prevalence and impact of open data initiatives in the world. In this sense, Colombia assumes an important role in the implementation of this type of policy; 12th in the ranking proposed by this organization and in third place in Latin America, below Mexico and Uruguay.

Considering the classification presented in Figure 4.3, the data sources are described in the next step.

### 4.2.2 Data Sources Description

Once the data collection step has been developed, we proceeded to describe them. Organizations and data sources by vulnerability dimension are presented below as a summary description. These data sources can be downloaded from the following repository: *https://dataverse.harvard.edu/dataverse/dfcropp*. More details about data dictionary and metadata can be found in Appendix B.

***Biophysical Dimension Data Sources***

- *SIVICAP (bp_sivicap).* "Sistema de Información de la Vigilancia de la Calidad del Agua para Consumo Humano" by its acronym in Spanish. It is an information system for monitoring water quality developed by the National Institute of Health in Colombia. This dataset was extracted from the report of physical, chemical, and microbiological characteristics of water quality in Colombian municipalities [129].
- *CORPOICA (bp_corpoica).* "Corporación Colombiana de Investigación Agropecuaria" by its acronym in Spanish [130], currently called AGROSAVIA, is a decentralized public entity responsible for generating scientific knowledge and technological solutions through research, innovation, technology transfer and training of researchers, for the benefit of the Colombian agricultural sector. The data was obtained from the soil analysis service of CORPOICA for the agricultural sector. It focuses on evaluating soil fertility, salinity, and parameters to develop fertilization plans, application of amendments, and land adequacy to achieve profitable production.
- *IDEAM (bp_ideam).* "Instituto de Hidrología, Meteorología y Estudios Ambientales" by its acronym in Spanish. It is a public institution of technical and scientific support to the National Environmental System, which generates knowledge, produces reliable, consistent and timely information on the state and dynamics of natural resources and the environment, facilitating the definition and adjustment of environmental policies and decision-making by the public and private sectors, and citizens in general. The respective data set was obtained from a request to the IDEAM about the data of the climatic stations installed in the study area [131].

- *AVA (bp_ava).* Interinstitutional and multisectoral analysis of vulnerability and adaptation to climate change for the agricultural sector of the upper Cauca River basin impacting adaptation policies. It is a project from the Climate and Development Knowledge Network (CDKN) [124]. The available data represents the percentage of municipal areas with climatic aptitude per crop.

***Economic-Productive Dimension Data Sources***

- *FINAGRO (ep_finagro).* "Fondo para el Financiamiento del Sector Agropecuario" by its acronym in Spanish [132], is an entity that promotes the development of the Colombian rural sector, with financing instruments and rural development, that stimulate investment. FINAGRO grants resources to financial entities to encourage loans to productive projects. Data were obtained from the Agronet portal and these represent the total value of disbursements and the number of credits granted by FINAGRO and Banco Agrario to different types of producers for each department by the municipality during the selected analysis period.
- *DANE-SIPSA-P (ep_dane_sipsa_p).* One of the functions of DANE ("Departamento Administrativo Nacional de Estadística" by its acronym in Spanish) [127] is to provide basic information for decision-making in all sectors of the economy. SIPSA ("Sistema de Información de Precios" by its acronym in Spanish) [133] presents the wholesale prices of agri-food products that are marketed in Colombia. Additionally, information on inputs and factors associated with agricultural production and the level of the food supply in cities. Data about the sown and harvested areas for different crops, as well as data on production, yield, costs, prices, gross income, profits, profitability, among others.
- *Agronet (ep_agronet).* Agronet [134] is the Information and Communication Network of the Agricultural Sector of Colombia, led by the Ministry of Agriculture and Rural Development and supported by the United Nations Organization for Food and Agriculture (FAO). This platform stores information from several official sources linked to the agricultural sector in a centralized way. By having data from official sources, Agronet provides reliable and updated information, according to the frequency of the investigations. Data collected represents the evolution of the harvested area, production, and yield in a period in a selected municipality.

- *Minagricultura (ep_minagricultura).* The Ministry of Agriculture and Rural Development (Minagricultura) [135] is a Colombian Ministry whose main objectives are the formulation, coordination, and adoption of policies, plans, programs, and projects in the agricultural, fishing, and rural development sectors. Data were obtained from the Agronet portal and corresponds to the evolution of the planted and harvested areas, as well as the production and yield of different crops in a period in a selected municipality.
- *Agronet-P (ep_agronet_p).* The data provider was Agronet, as described previously. Data represent the prices of different agricultural products and their respective variations in the international stock market [136].
- *DANE-SIPSA (ep_dane_sipsa).* The provider of this dataset is the same as *DANE-SIPSA-P* previously described. In this case, data were obtained from historical series of the wholesale quotations of the main agricultural products in the food basket [133].

***Political-Institutional Dimension Data Sources***

- *DNP-AIB (pi_dnp_aib).* "Departamento Nacional de Planeación" (DNP for its acronym in Spanish) [137] is an entity that defines the implementation of a strategic vision of the country in the social, economic, and environmental fields. This dataset represents the Colombian climate finance strategy and methodology for the identification of budget executions in initiatives compatible with climate change [138].
- *DNP-FI (pi_dnp_fi).* The data provider was DNP and these data are related to the forest incentive certificate as a recognition of the National Government through the Ministry of Agriculture and Rural Development to the positive externalities of reforestation in its commercial component [139]. This recognition is made by the Ministry through FINAGRO by delegation.
- *DNP-LA (pi_dnp_la).* DNP was the data provider and the dataset refers to the progressive access to the land ownership of the agrarian workers and the use of the nation's uncultivated lands, giving preference to the allocation of low-income farmers and establishing farmer reserve zones for the promotion of rural property [140].
- *DNP-PA (pi_dnp_pa).* The data provider was DNP. Data correspond to the access of small rural producers to factors of production such as land and labor, enhancing their use and complementing investment capacity through the direct

support of profitable productive initiatives with a contribution from the project, a resource called Modular Incentive [141].

***Sociocultural Dimension Data Sources***

- *DANE-HH (sc_dane_hh).* The National Administrative Department of Statistics [127] applies different censuses to obtain data from several aspects of the population in Colombia, among them, the national agricultural census. Data refers to the number of households within each dwelling and the perception of subjective poverty and displacement directed at the head of the household or his spouse [142].
- *DANE-H (sc_dane_h).* DANE was the data provider as previously described. This data source contains the number of dwellings, the occupation status of the same, the availability of public services, and the condition of the predominant materials [143].

### 4.2.3 Preliminary Analysis

This step is intended to determine the amount of information available for each indicator in the previously established dimensions. Therefore, the analyst can determine if more information is required to continue the process or move on to the second phase. Once the data sources were identified and classified, their main characteristics were extracted following the methodology proposed in [96]. Table 4.1 consolidates the following metadata: data source identifier, file format, time window, number of instances, number of attributes, temporal scale, spatial scale, and data source size.

| Data source identifier | File Format | Time Window | Number of Instances | Number of Attributes | Temporal Scale | Spatial Scale | Data Source Size |
|---|---|---|---|---|---|---|---|
| *bp_sivicap* | CSV | 2015 | 1,019 | 47 | Annual | Municipality | 238 KB |
| *bp_corpoica* | CSV | 2013 – 2016 | 24,179 | 31 | Annual | Municipality | 6.8 MB |
| *bp_ideam* | CSV | 2012 – 2019 | 2,042 | 8 | Monthly | Municipality | 109 KB |
| *bp_ava* | PDF | 2007 – 2011 | 99 | 8 | Annual | Municipality | 5 KB |
| *ep_finagro* | XLS | 2004 – 2019 | 3,993 | 5 | Annual | Municipality | 232 KB |
| *ep_dane_sipsa_p* | CSV | 2004 – 2019 | 130 | 13 | Annual | Municipality | 12 KB |
| *ep_agronet* | XLS | 2007 – 2016 | 3,789 | 8 | Annual | Municipality | 119.2 MB |
| *ep_minagricultura* | XLS | 2007 – 2015 | 5,911 | 16 | Annual | Municipality | 14 MB |
| *ep_agronet_p* | XLS | 2007 – 2016 | 3,789 | 8 | Annual | Municipality | 119.2 MB |
| *ep_dane_sipsa* | XLS | 2013 – 2017 | 170,118 | 5 | Monthly | Municipality | 11.7 MB |
| *pi_dnp_aib* | XLS | 2005 – 2013 | 19,286 | 8 | Annual | Municipality | 8.3 MB |
| *pi_dnp_fi* | XLS | 1995 – 2014 | 494 | 5 | Annual | Department | 2.2 MB |
| *pi_dnp_la* | XLS | 2010 – 2014 | 160 | 4 | Annual | Department | 2.2 MB |
| *pi_dnp_pa* | XLS | 2002 – 2013 | 783 | 9 | Annual | Municipality | 2.3 MB |
| *sc_dane_hh* | CSV | 2014 | 128,332 | 21 | Annual | Municipality | 8.9 MB |
| *sc_dane_h* | CSV | 2014 | 462,649 | 16 | Annual | Municipality | 8.9 MB |

**Table 4.1.** Data sources metadata.

As seen in Table 4.1, Excel's XLS format predominates in most files. This facilitated the initial conversion to CSV files which was the common format to start the cleaning process. Additionally, the predominant temporal scale was annual and the spatial scale was mostly at the municipality level. As mentioned above, each of these datasets was associated with a sustainability indicator. We selected these indicators considering the type of information from the data sources (Figure 4.4).

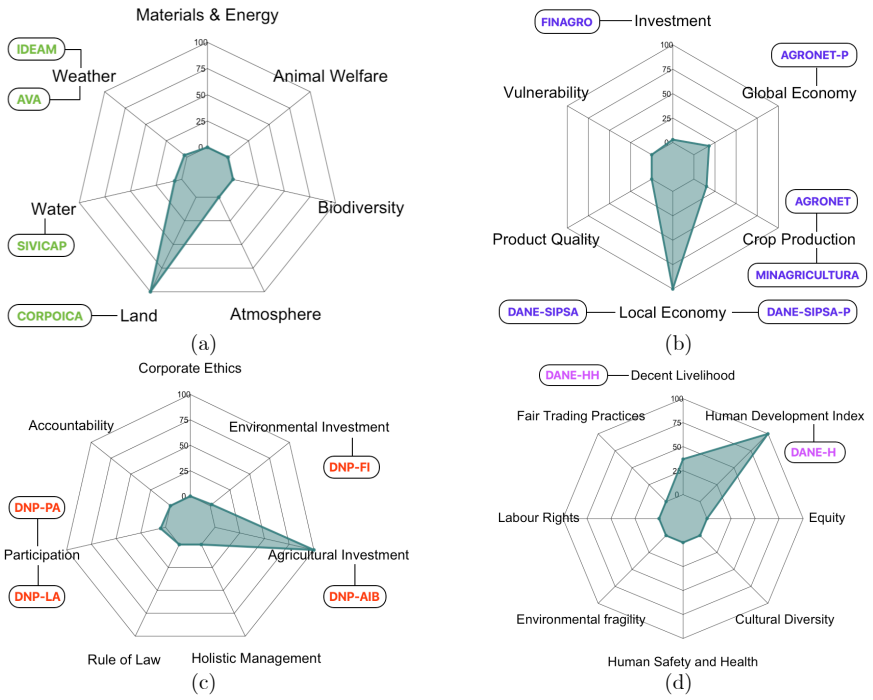



**Figure 4.4.** Availability of open data in the analyzed vulnerability dimensions. (a) Biophysical (data sources in green), (b) Economic-Productive (data sources in blue), (c) Political-Institutional (data sources in red), (d) Sociocultural (data sources in violet).

Figure 4.4 presents an overview of the amount of data available per indicator in each dimension, i.e., the percentage represented by the number of instances and the number of attributes. In this way, the radar chart establishes a starting point for data preparation that will be described in the following sections. It is worth mentioning that collecting different data sources for all the indicators is a complex task, therefore this was supported by the judgment of several experts in each of the indicators. In this case, a greater amount of data was found for the following indicators: Land (Biophysical), Local Economy (Economic-Productive), Agricultural Investment (Political-Institutional), and Human Development Index (Sociocultural). However, other indicators highlight the need to return to the step of collecting data sources. The above suggests an exhaustive search process to increase the percentage of data

around these indicators. For this case study, despite having conducted a deep search for information sources, only some additional data were found which did not significantly increase the amount of data.

## 4.3. Data Sources Preprocessing

In this section, we present the second phase of the data preparation process applied to the study case (Figure 4.5). The aim of this phase is to establish an evaluation of the main meta-features to determine the data quality once the cleaning process is finished.

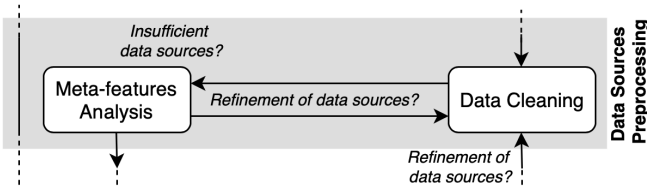



**Figure 4.5.** Second phase of the data preparation process.

### 4.3.1 Data Cleaning

In any data-oriented process, there is a risk of finding duplicate information, missing values, data at different spatial-temporal scales, noise, among others. These considerations must be taken into account to achieve a successful data integration process. Different inconsistencies were identified in all data sources following the methodology proposed in [96] and consigned in Appendix C. Therefore, the cleaning steps were applied to each of the 16 data sources which are presented in Table 4.2.

| Data Source | Issues | Applied Solutions |
|---|---|---|
| *bp-sivicap* | MV, O, HD | MR, TM, LR and RF |
| *bp-corpoica* | MV, O, HD | MR, TM, LR and RF |
| *bp-ideam* | MV | MR |
| *bp-ava* | None | None |
| *ep-finagro* | MV, O | MR, TM |
| *ep-dane-sipsa-p* | MV, O, HD | MR, TM, LR and RF |
| *ep-agronet* | MV, O, HD | MR, TM, CM |
| *ep-minagricultura* | MV, O | MR, TM |
| *ep-agronet-p* | MV, O | MR, TM |
| *ep-dane-sipsa* | MV, O | MR, TM |
| *pi-dnp-aib* | MV, O, DI | MR, TM, DF |
| *pi-dnp-fi* | MV, O | MR, TM |
| *pi-dnp-la* | MV, O | MR, TM |
| *pi-dnp-pa* | MV, O, DI | MR, TM, DF |
| *sc-dane-hh* | MV | MR |
| *sc-dane-h* | MV | MR |

**Table 4.2.** Cleaning steps applied to the 16 data sources. MV: Missing Values, O: Outliers, HD: High Dimensionality DI: Duplicate Instances, MR: Mean Replacement, TM: Tukey's Method, LR and RF: Logistic Regression and Random Forest, CM: Correlation Matrix, DF: Duplicated Function.

To illustrate a specific example, we present the SIVICAP dataset, corresponding to water quality, which is a very important indicator of the biophysical dimension. The SIVICAP data source collects information about physicochemical and microbiological variables obtained from water samples from different municipalities in the UCRB, Cauca zone. The cleaning steps for SIVICAP are presented below.

***Check Missing Values***

Table 4.3 presents all the attributes of the SIVICAP data source and the respective check of missing values. This cleaning task was applied to each of the attributes along with a summary of the main measures of central tendency. Those attributes with a high percentage of missing values have a greater probability of being initially discarded (in this case we have chosen a percentage higher than 70%). In Table 4.3, those attributes with this characteristic were discarded. In contrast, attributes with low percentages of missing values were selected to continue with the cleaning process. The imputation method used was a direct replacement for the mean value of each attribute.

| Attribute | Mean | Median | MV | PMV |
|---|---|---|---|---|
| *Total Samples* | 45.84 | 26 | 0 | 0 |
| *Apparent Color* | 21.58 | 6.11 | 61 | 5.98 |
| *Turbidity* | 2.87 | 1.17 | 21 | 2.06 |
| *pH* | 7.32 | 7.27 | 18 | 1.76 |
| *Free Residual Chlorine* | 0.98 | 0.85 | 83 | 8.14 |
| *Total Alkalinity* | 47.48 | 34.56 | 96 | 9.42 |
| *Calcium* | 17.41 | 13.64 | 524 | 51.42 |
| *Phosphates* | 0.35 | 0.10 | 458 | 44.94 |
| *Manganese* | 0.41 | 0 | 854 | 83.80 |
| *Molybdenum* | 0.17 | 0 | 908 | 89.10 |
| *Magnesium* | 8.93 | 3.90 | 498 | 48.87 |
| *Zinc* | 1.93 | 0 | 887 | 87.04 |
| *Total Hardness* | 61.47 | 45.60 | 76 | 7.45 |
| *Sulphates* | 19.67 | 10.67 | 317 | 31.10 |
| *Total Iron* | 0.17 | 0.08 | 245 | 24.04 |
| *Chloride* | 12.27 | 5.90 | 190 | 18.64 |
| *Nitrates* | 1.41 | 1.15 | 755 | 74.09 |
| *Nitrites* | 0.03 | 0 | 371 | 36.40 |
| *Aluminum* | 2.69 | 0.05 | 612 | 60.05 |
| *Fluorides* | 0.20 | 0.03 | 559 | 54.85 |
| *COT* | 4.08 | 1.25 | 823 | 80.76 |
| *Total Coliforms* | 457.12 | 22.87 | 3 | 0.29 |
| *E Coli* | 47.92 | 0.54 | 3 | 0.29 |
| *Antimony* | 0 | 0 | 907 | 89 |
| *Arsenic* | 0.13 | 0 | 908 | 89.1 |
| *Barium* | 0 | 0 | 991 | 97.25 |
| *Cadmium* | 0 | 0 | 891 | 87.43 |
| *Free and Dissociable Cyanide* | 0 | 0 | 915 | 89.79 |
| *Copper* | 0.01 | 0 | 867 | 85.08 |
| *Total Chrome* | 0 | 0 | 896 | 87.92 |
| *Mercury* | 0 | 0 | 880 | 86.35 |
| *Nickel* | 0 | 0 | 959 | 94.11 |

| | | | | |
|---|---|---|---|---|
| *Lead* | 0 | 0 | 965 | 94.70 |
| *Selenium* | 0 | 0 | 912 | 89.49 |
| *Total Trihalomethanes* | 0 | 0 | 926 | 90.87 |
| *Polycyclic Aromatic Hydrocarbons* | 0 | 0 | 950 | 93.22 |
| *Giardia* | 0.08 | 0 | 953 | 93.52 |
| *Cryptosporidium* | 0.10 | 0 | 953 | 93.52 |
| *Total Pesticides* | 0 | 0 | 970 | 95.19 |
| *Organophosphorates and Carbamates* | 0 | 0 | 1015 | 99.60 |
| *Mesophiles* | 1625 | 65 | 882 | 86.55 |
| *IRCA Average* | 23.79 | 17.27 | 0 | 0 |
| *IRCA Base Average* | 23.72 | 17.27 | 0 | 0 |
| *Average Risk Level* | - | - | 0 | 0 |

**Table 4.3.** Summary of missing values and measures of central tendency for the attributes of the SIVICAP data source. MV: Missing Values, PMV: Percentage of Missing Values.

### *Outlier Detection*

There are different methods to detect outliers that include the Standard Regression (SR) and the Tukey's Method (TM) [144]. In this work, we used the last method considering its non-dependence on the data distribution. Additionally, it ignores the mean and the standard deviation, which are influenced by extreme values. This method is based on quartiles: Q1 (first quartile, value $\geq 1/4$ of the data), Q2 (second quartile or median, value $\geq 1/2$ of the data), and Q3 (third quartile, value $\geq 3/4$ of the data). The interquartile range (IQR) corresponds to Q3 - Q1 and according to TM, outliers are values more than 1.5 times the IQR of the quartiles (below Q1 - 1.5IQR or above Q3 + 1.5IQR). Box plots are used to detect these outliers, where the median and the lower and upper quartiles allow visualizing the distribution of the data. To apply the Tukey's method, we used an R script implemented by Dhana [145]. In this example, we only show the detection process for the pH attribute (Figure 4.6). The outliers detected were replaced by "NA" values and the process of replacing missing values was applied again. The other attributes of this dataset (and of all datasets) followed the same process.

### *Label Correction, Class Balancing, and Removal of Duplicate Instances*

For this data set, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced acceptably. Finally, no duplicate instances were found.

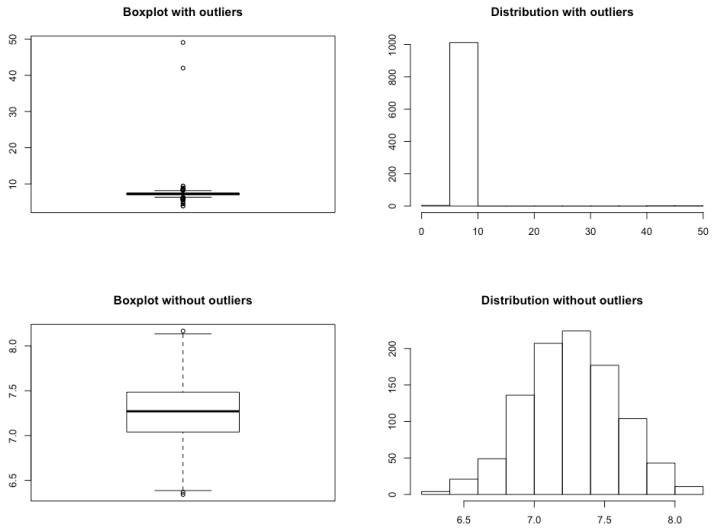



**Figure 4.6.** Data distribution before and after the outlier detection process for the SIVICAP data source. Attribute: pH, Outliers identified: 43, Proportion (%) of outliers: 4.4, Mean of the outliers: 8.51, Mean without removing outliers: 7.32, Mean removing outliers: 7.26.

### *Dimensionality Reduction*

We considered two cases to find the attributes that best represent the data sources: labeled and unlabeled data. In the first case, the importance of variables was determined by the respective correlation matrix. To show these correlations, Figure 4.7 presents a correlogram (or autocorrelation plot), which is an image of correlation statistics that highlights 20 attributes with the highest correlations among themselves. In this case, the dark blue tone represents a direct autocorrelation between variables, while dark brown shows an inverse-self. For example, in Figure 4.7, we observed a high direct correlation between the variables *Total Alkalinity* and *Total Hardness*. In the same way, there is an inverse correlation between *Free Residual Chlorine* and the *Average Water Quality Risk Index* (*IRCA* for its acronym in Spanish).

Furthermore, Table 4.4 shows the eigenvalues for the correlation matrix. These values represent the explained variance of the analyzed attributes. The explained percentage and the total accumulated percentage of the variance are also presented. Based on these values, we established the number of attributes that best explain the total variance. In this example, we could select between 8 (explain 64.13% of the total variance) and 14 (explain 91.09% of the total variance) attributes.

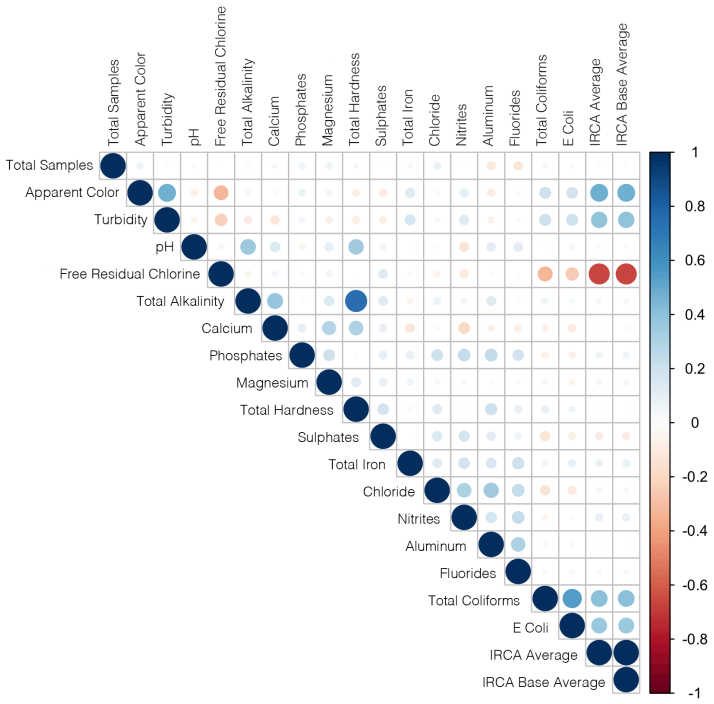



**Figure 4.7.** Correlated attributes for the SIVICAP data source.

| Order | Explained Variance | Explained Percentage | Accumulated Percentage |
|---|---|---|---|
| 1 | 2.8 | 14.09 | 14.09 |
| 2 | 2.53 | 12.73 | 26.82 |
| 3 | 1.92 | 9.66 | 36.48 |
| 4 | 1.3 | 6.54 | 43.02 |
| 5 | 1.11 | 5.58 | 48.60 |
| 6 | 1.07 | 5.38 | 53.98 |
| 7 | 1.01 | 5.08 | 59.07 |
| 8 | 1 | 5.06 | 64.13 |
| 9 | 0.99 | 4.98 | 69.11 |
| 10 | 0.97 | 4.88 | 73.99 |
| 11 | 0.93 | 4.68 | 78.67 |
| 12 | 0.88 | 4.43 | 83.09 |
| 13 | 0.8 | 4.02 | 87.12 |
| 14 | 0.79 | 3.97 | 91.09 |
| 15 | 0.69 | 3.47 | 94.56 |
| 16 | 0.54 | 2.72 | 97.28 |
| 17 | 0.28 | 1.41 | 98.69 |
| 18 | 0.18 | 0.91 | 99.60 |
| 19 | 0.08 | 0.40 | 99.99 |
| 20 | 0 | 0 | 100 |

**Table 4.4.** Eigenvalues of the SIVICAP dataset correlation matrix.

On the other hand, the second case corresponds to the labeled datasets. Here, the importance of variables was calculated using two methods: Logistic Regression (LR) and Random Forest (RF). In LR, large positive values of overall signify higher importance of the feature in the prediction of positive class [108]. On the other hand,

in RF, the more the Gini Index decreases for a feature, the more important it is. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient [109]. A comparison of the results obtained with both methods is presented in Table 4.5.

| Attribute | Logistic Regression - Overall | Random Forest - Mean Decrease Gini |
|---|---|---|
| *Total Samples* | 0.42 | 9.41 |
| *Apparent Color* | 2.05 | 20.79 |
| *Turbidity* | 0.48 | 17.04 |
| *pH* | 0.83 | 4.73 |
| *Free Residual Chlorine* | 3.91 | 44.33 |
| *Total Alkalinity* | 0.44 | 4.75 |
| *Calcium* | 0.88 | 2.96 |
| *Phosphates* | 0.12 | 3.45 |
| *Magnesium* | 0.64 | 2.57 |
| *Total Hardness* | 0.19 | 5.80 |
| *Sulphates* | 1.57 | 5.01 |
| *Total Iron* | 1.13 | 3.98 |
| *Chloride* | 1.35 | 4.92 |
| *Nitrites* | 0.99 | 3.53 |
| *Aluminum* | 2.51 | 3.85 |
| *Fluorides* | 0.38 | 2.35 |
| *Total Coliforms* | 2.37 | 36.95 |
| *E Coli* | 0.65 | 50.62 |
| *IRCA Average* | 0.42 | 272.24 |
| *IRCA Base Average* | 0.22 | 264.94 |

**Table 4.5.** Importance of variables using the LR and RF methods for the SIVICAP data source.

Once the cleaning steps have been applied, two versions of the datasets are obtained. The former in its raw format, and the last one already processed. With the next step (meta-features analysis) it is intended to corroborate whether their quality improved significantly or not, and thus continue with the last phase of the process.

### 4.3.2 Meta-features Analysis

This step established an overview about data quality characteristics, evaluating whether the new processed datasets present an improvement over the original ones. In this sense, Table 4.6 presents the meta-features previously described in section 3.2.2, and details can be consulted in Appendix D. We established a data quality improvement in most of the datasets mainly influenced by the processes of imputation and removal of outliers. Skewness (values closer to zero) and kurtosis (values closer to three) demonstrated a better distribution of data after the application of the cleaning steps. For example, skewness and kurtosis values decreased significantly for the *ep-agronet-p* and *pi-dnp-aib* datasets (datasets with more outliers), where they were initially high. On the other hand, the number of missing values and duplicate instances were reduced to zero for all datasets. At this point, we obtained the clean

data sources by improving their quality in order to proceed to the last phase of the process, which is presented in the following section.

| Data source | Instances | | Attributes | | Missing Values | | Duplicate Instances | | Data Dimensionality | | Missing Values Ratio | | Duplicate Instances Ratio | | Mean Absolute Skewness | | Mean Absolute Kurtosis | | Mean Attribute Entropy | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P |
| *bp-sivicap* | 1,019 | 1,019 | 47 | 24 | 16,576 | 0 | 0 | 0 | 0.04 | 0.02 | 0.34 | 0 | 0 | 0 | 9.83 | 1.08 | 176.96 | 18.78 | 2.80 | 2.80 |
| *bp-corpoica* | 24,179 | 24,179 | 31 | 28 | 68,187 | 0 | 0 | 0 | 0 | 0 | 0.09 | 0 | 0 | 0 | 16.23 | 1.24 | 1,774.18 | 4.02 | 2.89 | 2.89 |
| *bp-ideam* | 2,042 | 2,042 | 8 | 8 | 547 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0 | 0 | 0.48 | 0.22 | 3.12 | 3.40 | 1.92 | 1.92 |
| *bp-ava* | 99 | 99 | 8 | 8 | 0 | 0 | 0 | 0 | 0.08 | 0.08 | 0 | 0 | 0 | 0 | 0.54 | 0.54 | 3.06 | 3.06 | 3.06 | 3.06 |
| *ep-finagro* | 3,993 | 3,993 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13.74 | 1.82 | 252.70 | 5.82 | 2.17 | 2.17 |
| *ep-dane-sipsa-p* | 130 | 130 | 13 | 13 | 0 | 0 | 0 | 0 | 0.10 | 0.10 | 0 | 0 | 0 | 0 | 3.22 | 3.22 | 17.55 | 17.55 | 2.50 | 2.50 |
| *ep-agronet* | 3,789 | 3,789 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4.95 | 2.01 | 31.48 | 7.23 | 2.39 | 2.39 |
| *ep-minagricultura* | 5,911 | 5,911 | 16 | 16 | 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6.41 | 2.02 | 54.47 | 7.13 | 2.73 | 2.73 |
| *ep-agronet-p* | 25,228 | 25,228 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48.17 | 0.60 | 5,953.73 | 4.24 | 3.23 | 3.23 |
| *ep-dane-sipsa* | 170,118 | 170,118 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.92 | 1.02 | 8.20 | 3.37 | 3.56 | 3.56 |
| *pi-dnp-aib* | 19,286 | 19,286 | 8 | 8 | 0 | 0 | 32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50.10 | 1.93 | 3,347.86 | 6.89 | 3.09 | 3.09 |
| *pi-dnp-fi* | 494 | 494 | 5 | 5 | 594 | 0 | 0 | 0 | 0.01 | 0 | 0.24 | 0 | 0 | 0 | 6.05 | 0.12 | 58.12 | 2.46 | 3.10 | 3.10 |
| *pi-dnp-la* | 160 | 160 | 4 | 4 | 17 | 0 | 0 | 0 | 0.02 | 0.02 | 0.02 | 0 | 0 | 0 | 6.03 | 1.55 | 49.50 | 4.94 | 2.54 | 2.54 |
| *pi-dnp-pa* | 783 | 775 | 9 | 9 | 77 | 0 | 0 | 0 | 0.01 | 0.01 | 0.01 | 0 | 0.01 | 0 | 5.05 | 0.89 | 73.51 | 3.33 | 4.23 | 3.99 |
| *sc-dane-hh* | 128,332 | 128,332 | 21 | 21 | 12,546 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *NNA* | *NNA* | *NNA* | *NNA* | 2.01 | 1.65 |
| *sc-dane-h* | 462,649 | 462,649 | 16 | 16 | 35,691 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *NNA* | *NNA* | *NNA* | *NNA* | 2.09 | 2.27 |

**Table 4.6.** Results of datasets preprocessing. R: Raw dataset, P: Pre-processed dataset. NNA: No Numeric Attributes.

## 4.4. Variables Prioritization

This section presents the third and final phase of the data preparation process (Figure 4.8). Here, the most important attributes of each data source are identified and contrasted with the criteria of several experts to validate its relevance. This final phase and its modules applied to the case study are presented below.

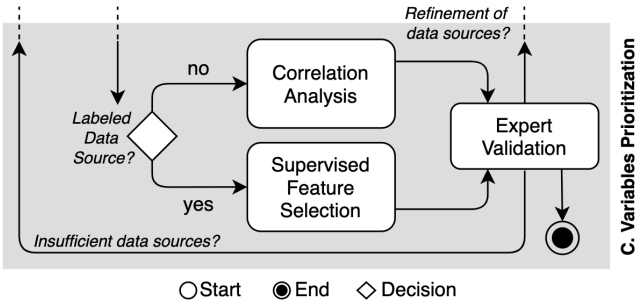



**Figure 4.8.** Third phase of the data preparation process.

### 4.4.1 Feature Selection

As observed in the example described in section 4.3.1 (data cleaning - dimensionality reduction) and the other datasets reported in Appendix C, we calculated the importance of variables using different techniques applied according to the type of

dataset: labeled (supervised approach) or unlabeled (correlation analysis approach). These approaches are presented below.

***Supervised Approach***

In the case of labeled datasets, the attributes were ranked through two automatic methods such as Logistic Regression and Random Forest as mentioned in section 3.2.3. Some aspects considered to apply these techniques are mentioned below.

- In logistic regression, the equation to describe the model to be fitted to the SIVICAP dataset is presented in Equation (4.1). Where $ARL$ is defined as the Average Risk Level of water contamination, which is expressed as the linear combination among the 20 attributes of the dataset.

$$\begin{aligned} ARL = {} & (-0.0006)\,Total\ Samples + (-0.051)Apparent\ Color + (-0.063)\,Turbidity \\ & + (-0.368)pH + (1.83)Free\ Residual\ Chlorine \\ & + (0.003)Total\ Alkalinity + (-0.028)\,Calcium \\ & + (0.415)Phosphates + (0.092)Magnesium \\ & + (0.001)Total\ Hardness + (-0.031)Sulphates \\ & + (1.649)Total\ Iron + (-0.059)Chloride + (14.063)Nitrites \\ & + (0.050)Aluminum + (0.143)Fluorides \\ & + (-0.001)Total\ Coliforms + (-0.006)E\ Coli \\ & + (-0.074)IRCA\ Average + (-0.039)IRCA\ Base\ Average \end{aligned} \tag{4.1}$$

- In random forest, we used the *Breiman's* algorithm for feature selection (implementation in R software). This ensemble method is based on randomization to increase the diversity of decision trees (basic learners). RF uses Bagging to generate a training set for each random tree, which divides each node by gaining information.

***Correlation Analysis Approach***

On the other hand, we generated a correlation matrix for unlabeled datasets in order to select the most important attributes (those with the highest correlation). A correlogram is shown in Figure 4.9 to better visualize the correlation degree between the variables in the AVA data source. In this example, variables correspond to the Climate Adaptability (CA) of some crops analyzed in the AVA project. This correlogram indicates a high direct correlation between the CA of the banana crop and the CA of the coffee and cocoa crops. It also highlights a high inverse correlation between the CA of the potato crop and the CA of the two crops mentioned above (coffee and cocoa).

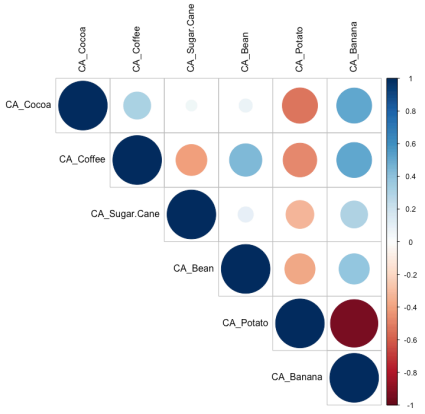



**Figure 4.9.** Correlated attributes for the AVA data source.

From the previous correlation matrix, we were able to establish the most important variables in this dataset. These variables were the CAs of banana, potato, Coffee, Cocoa, Bean, and Sugar Cane respectively. The ranking of variables obtained after applying such techniques was validated by a group of experts. The description of this experiment and their respective results are presented in the next section.

### 4.4.2 Expert Validation

In the final step of the process, an experiment was developed in both labeled and unlabeled datasets to validate the relevance of the results in the ranking of variables. In this way, four experts in domains such as agribusiness engineering, environmental engineering, and chemical engineering, had the role of raters. Each rater identified the most important variables or attributes in each data source according to their knowledge area (Appendix E). Hence, different rankings were obtained from the perspective of each rater. Finally, these were compared with the rankings of the automatic methods (LR, RF in Table 4.5) using the *Fleiss's Kappa* index [146] for assessing the reliability of agreement between a fixed number of raters (Appendix F).

For a better visualization, the results of the ranking process are presented using two of the 16 data sources (SIVICAP, the labeled data source, and AVA, the unlabeled data source). It should be noted that the same procedure was applied for each of the 16 datasets. In the first instance, Table 4.7 shows the rankings by Logistic Regression, Random Forest, and expert criteria for the SIVICAP data source (labeled data), where the most important variable is represented by the ranking number 1. Furthermore, Table 4.8 shows the ranking obtained through the correlation matrix and the ranking determined by the experts for the AVA data source.

| Ranking | LR | RF | Expert 1 | Expert 2 | Expert 3 | Expert 4 |
|---|---|---|---|---|---|---|
| 1 | Free Residual Chlorine | IRCA Average | Free Residual Chlorine | pH | Sulphates | pH |
| 2 | Aluminum | IRCA Base Average | Sulphates | Aluminum | Phosphates | Turbidity |
| 3 | Total Coliforms | E Coli | Total Coliforms | Total Iron | Turbidity | Total Coliforms |
| 4 | Apparent Color | Free Residual Chlorine | Apparent Color | Fluorides | pH | Apparent Color |
| 5 | Sulphates | Total Coliforms | Aluminum | Sulphates | Nitrites | Sulphates |
| 6 | Chloride | Apparent Color | pH | Chloride | Chloride | Chloride |
| 7 | Total Iron | Turbidity | Calcium | Turbidity | Total Iron | Total Iron |
| 8 | Nitrites | Total Samples | Total Iron | Nitrites | Free Residual Chlorine | Calcium |
| 9 | Calcium | Total Hardness | Chloride | Calcium | Calcium | Total Hardness |
| 10 | pH | Sulphates | Turbidity | Free Residual Chlorine | Aluminum | Phosphates |
| 11 | E Coli | Chloride | E Coli | E Coli | E Coli | E Coli |
| 12 | Magnesium | Total Alkalinity | Total Alkalinity | Magnesium | Magnesium | Total Alkalinity |
| 13 | Turbidity | pH | Phosphates | Total Coliforms | Total Coliforms | Nitrites |
| 14 | Total Alkalinity | Total Iron | Magnesium | Total Alkalinity | Total Alkalinity | Free Residual Chlorine |
| 15 | IRCA Average | Aluminum | IRCA Average | Apparent Color | Apparent Color | IRCA Average |
| 16 | Total Samples | Nitrites | IRCA Base Average | Total Hardness | Total Hardness | Aluminum |
| 17 | Fluorides | Phosphates | Fluorides | Phosphates | Fluorides | Fluorides |
| 18 | IRCA Base Average | Calcium | Nitrites | IRCA Average | IRCA Average | IRCA Base Average |
| 19 | Total Hardness | Magnesium | Total Hardness | IRCA Base Average | IRCA Base Average | Magnesium |
| 20 | Phosphates | Fluorides | Total Samples | Total Samples | Total Samples | Total Samples |

**Table 4.7.** Importance of variables for the SIVICAP data source (Labeled Data – Supervised Approach).

| Ranking | Corr. Matrix | Expert 1 | Expert 2 | Expert 3 | Expert 4 |
|---|---|---|---|---|---|
| 1 | CA Banana | CA Banana | CA Banana | CA Banana | CA Banana |
| 2 | CA Potato | CA Potato | CA Potato | CA Potato | CA Potato |
| 3 | CA Coffee | CA Cocoa | CA Coffee | CA Cocoa | CA Coffee |
| 4 | CA Cocoa | CA Coffee | CA Cocoa | CA Sugar Cane | CA Cocoa |
| 5 | CA Sugar Cane | CA Sugar Cane | CA Sugar Cane | CA Bean | CA Sugar Cane |
| 6 | CA Bean | CA Bean | CA Bean | CA Coffee | CA Bean |

**Table 4.8.** Importance of variables for the AVA data source (Unlabeled Data – Correlation Analysis Approach).

Subsequently, to determine the performance of the data preparation process for all datasets, we evaluated the level of agreement between automatic rankings and expert raters. Table 4.9 presents the results of *Fleiss's Kappa*. We used the following ranges to interpret this measure: $< 0$ = poor agreement; [0.01 - 0.20] = slight agreement; [0.21 - 0.40] = fair agreement; [0.41 - 0.60] = moderate agreement; [0.61 - 0.80] = substantial agreement; [0.81 - 1.00] = almost perfect agreement. The results of this evaluation presented an acceptable level of agreement for all raters and most data sources. Similarly, p-values remained at the threshold of statistical significance (typically less than 0.05), which indicates that the agreement between raters is significantly better than would be expected by chance. It should be noted that for all labeled datasets, we observed a better concordance with the logistic regression method.

| Data source | Fleiss's Kappa | P-value | Interpretation |
|---|---|---|---|
| *bp-sivicap* | 0.40 / 0.30 * | 0 | Fair agreement |
| *bp-corpoica* | 0.43 / 0.26 * | 0 / 0 * | Fair agreement |
| *bp-ideam* | 0.65 | 0 | Substantial agreement |
| *bp-ava* | 0.70 | 0 | Substantial agreement |
| *ep-finagro* | 0.32 | 4.3e-6 | Fair agreement |
| *ep-dane-sipsa-p* | 0.16 / 0.09 * | 4.6e-11 / 2e-4 * | Slight agreement |

| | | | |
|---|---|---|---|
| *ep-agronet* | 0.32 | 7.5e-15 | Fair agreement |
| *ep-minagricultura* | 0.43 | 0 | Moderate agreement |
| *ep-agronet-p* | 0.66 | 0 | Substantial agreement |
| *ep-dane-sipsa* | 0.30 | 2.2e-5 | Fair agreement |
| *pi-dnp-aib* | 0.78 | 0 | Substantial agreement |
| *pi-dnp-fi* | 0.7 | 0 | Substantial agreement |
| *pi-dnp-la* | 0.73 | 8.8e-16 | Substantial agreement |
| *pi-dnp-pa* | 0.56 | 0 | Moderate agreement |
| *sc-dane-hh* | 0.59 | 0 | Moderate agreement |
| *sc-dane-h* | 0.88 | 0 | Almost perfect agreement |

**Table 4.9.** Results of Fleiss's kappa agreement measure for all data sources.

Finally, Figure 4.10 shows the three most relevant attributes for each data source grouped into the respective agricultural vulnerability dimension. From these attributes, we can establish a starting point for data integration that covers levels 1 and 2.

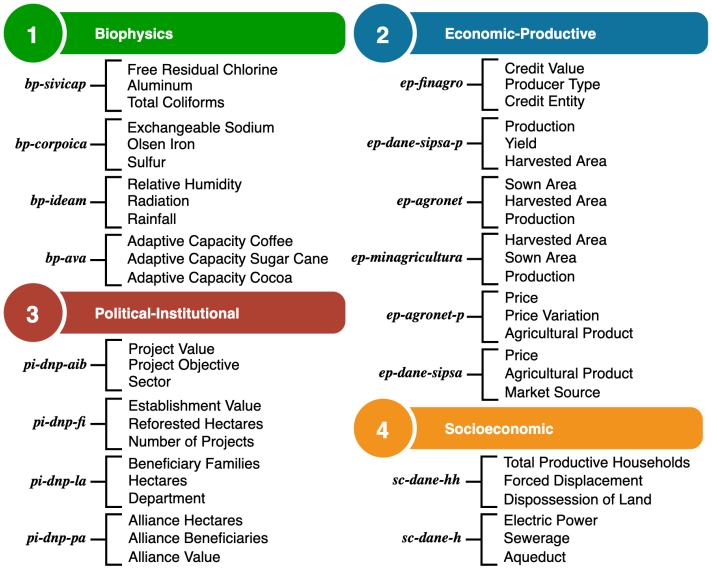



**Figure 4.10.** Most relevant attributes for each data source obtained at level 0 (data assessment).

## 4.5. Summary

In this chapter, we described Level 0 or Data Evaluation by implementing a case study in the upper Cauca river basin, Cauca zone. In addition to describing the details of the study area, we developed a data preparation process to support climate vulnerability assessments. We detailed how this process was applied using data sources from different dimensions of agricultural vulnerability in the study area. Additionally, we described the three main phases of the process such as data sources evaluation, data sources preprocessing, and variables prioritization.

# Chapter 5

# Data Integration and Analysis

In this chapter, we continue to develop the case study on levels 1, 2, and 3 of the data fusion strategy. In level 1, we identified the implicit relationships between data sources by analyzing the temporal and spatial scales. For this purpose, we established a spatio-temporal characterization process, where all possible relationships were consolidated and analyzed to verify their relevance in the data fusion strategy. Also, we built a relationship scheme to guide the data sources integration at Level 2. In this second level, we selected a data integration approach and applied a method to combine the data sources based on *Entity Matching* [110]. We labeled the resulting datasets according to the final objective of the data fusion strategy. Finally, in Level 3, we evaluated and validated several MLC models for predicting crops in the upper Cauca river basin, Cauca zone. The results of levels 1, 2, and 3 are presented below.

## 5.1. Relationship Analysis (Level 1)

In this section, we present the relationship analysis steps (Level 1) applied in the case study in the upper Cauca river basin, Cauca zone.

### 5.1.1 Spatio-Temporal Characterization of Data Sources

We identified meta-features at the temporal and spatial level, such as the time window, the temporal scale, the sampling interval, and the spatial scale. We also extracted additional meta-features that can guide the data integration process, in this case, information about the crops such as the spatial units where they are grown, cultivated, produced, or commercialized. Table 5.1 summarizes the above meta-features for each of the 16 data sources.

| Dimension | Data Source | Temporal Meta-features | | | Spatial Meta-features | Other Meta-features | Units |
|---|---|---|---|---|---|---|---|
| | | Time Window | Temporal Scale | Sampling Intervals | Spatial Scale | | |
| Biophysical | bp-sivicap | 2015 | A | Regular | De, Mu | - | - |
| | bp-corpoica | 2013 - 2016 | A | Irregular | De, Mu | Crop | Point |
| | bp-ideam | 2012 - 2019 | A, M | Regular | De, Mu | - | - |
| | bp-ava | 2007 - 2011 | - | - | De, Mu | Crop | Point |
| Economic-Productive | ep-finagro | 2004 - 2019 | A | Regular | De | - | - |
| | ep-dane-sipsa-p | 2004 - 2019 | A | Regular | De | Crop | Area (t/Ha) |
| | ep-agronet | 2007 - 2016 | A | Regular | De | Crop | Area (t/Ha) |
| | ep-minagricultura | 2007 - 2015 | A, BA | Regular | De | Group, Subgroup, Crop | Area (t/Ha) |
| | ep-agronet-p | 2007 - 2016 | A, M, D | Irregular | St | Crop | Point |
| | ep-dane-sipsa | 2013 - 2017 | A, M | Regular | De, Mu, Ma | Group, Crop | Point |
| Political-Institutional | pi-dnp-aib | 2005 - 2013 | A | Regular | De, Co | - | - |
| | pi-dnp-fi | 1995 - 2014 | A | Irregular | De | - | - |
| | pi-dnp-la | 2010 - 2014 | A | Irregular | De | - | - |
| | pi-dnp-pa | 2002 - 2013 | A | Regular | De, Mu | Crop | Area (Ha/Alliance) |
| Sociocultural | sc-dane-hh | 2014 | - | - | De, Mu | - | - |
| | sc-dane-h | 2014 | - | - | De, Mu | - | - |

**Table 5.1.** Spatio-temporal meta-features of data sources. A: Annual, BA: Biannual, M: Monthly, Mu: Municipality, De: Department, Ma: Market, Co: Corporation. Group and Subgroup represent hierarchical crop categories such as fruit trees, vegetables, among others.

### 5.1.2 Data Source Relationship Scheme

From Table 5.1, we compare the meta-features of one data source with the other 15 to obtain the strength of each relationship. We use a function to compare, one by one, the temporal and spatial attributes among two data sources. The similarity intervals used were: 0 - 0.25 (weak relationship), 0.25 - 0.75 (intermediate relationship), 0.75 - 1 (strong relationship). Subsequently, we averaged the strength of all spatio-temporal attributes for each compared tuple of datasets. For example, if we compare the attribute "time scale" in the datasets *bp_sivicap* and *bp_corpoica*, the strength of this relationship will be 1 (100%) considering the same scale (annual) in both datasets. While the strength will be 0.5 (50%) if the compared datasets are *bp_sivicap* (annual) and *bp_ideam* (annual, monthly). Finally, we generated the *Data Source Relationship Matrix* shown in Figure 5.1. According to these relationships' strengths, we established the following best combinations among the datasets.

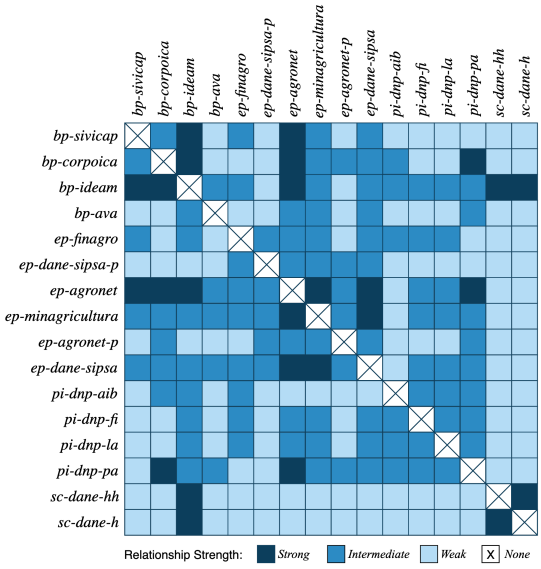



**Figure 5.1.** Data Source Relationship Matrix.

- *bp-ideam + (bp-sivicap or bp-corpoica)*
- *ep-agronet + (bp-sivicap, bp-corpoica, or bp-ideam)*
- *ep-minagricultura + ep-agronet*
- *ep-dane-sipsa + (ep-agronet or ep-minagricultura)*
- *pi-dnp-pa + (bp-corpoica or ep-agronet)*
- *sc-dane-hh + bp-ideam*
- *sc-dane-h + (bp-ideam or sc-dane-hh)*

## 5.2. Data Integration (Level 2)

This section presents the results obtained in Level 2 or Data Integration applied in the case study in the upper Cauca river basin, Cauca zone. The main finding at this level was data sources combined and labeled with multiple crops.

### 5.2.1 Selecting the Integration State

We selected the early integration approach considering an increase in the added value of a combined dataset. Combining data from different dimensions of agricultural vulnerability allows this increase in value. Different combined datasets were enriched with more features to train subsequent predictive models. The ability to determine, in advance, the theoretical complexity of a combined dataset for a machine learning model was a key aspect of this approach.

### 5.2.2 Integrating Data Sources

As mentioned in section 3.4, we generate new datasets with a more synthesized and reliable added value through a reduction process. To develop this integration, we first identify the most related data sources to the data fusion strategy objective. In our case, considering attributes on production and crop yield per municipality, we identified the *ep-agronet* as the central data source. We applied entity matching to associate the selected data sources based on *ep-agronet* and the relationship matrix information (Figure 5.1). We used the *Jaro-Winkler* similarity function [147] to identify the matches between municipality names in each tuple of data sources. We selected this similarity algorithm by considering two key aspects such as good performance comparing short strings and its name-comparison oriented design. The Jaro-Winkler similarity function is presented in Eq. (5.1). Let $s_1$ and $s_2$ the strings to be compared, $c$ the number of agreeing characters, and $t$ the number of transpositions.

$$sim_{jaro}(s_1, s_2) = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_1|} + \frac{c-t}{c}\right) \tag{5.1}$$

Once this similarity function was applied, we obtained four combined data sources, which are mentioned below.

- *CDS1: bp-ideam + ep-agronet* → (climate + crop production)
- *CDS2: bp-corpoica + ep-agronet* → (soil + crop production)
- *CDS3: bp-sivicap + ep-agronet* → (water quality + crop production)
- *CDS4: pi-dnp-pa +ep-agronet* → (productive alliances + crop production)

We discarded other combinations considering the low number of resulting instances (between 4 and 15 instances). Furthermore, it was possible to combine five data sources to obtain a Combined Global Dataset (CGD), which represents the combination of the maximum possible number of data sources. This combination is mentioned below.

- *CGD: bp-ideam + bp-corpoica + bp-sivicap + pi-dnp-pa + ep-agronet* → (climate + soil + water quality + productive alliances + crop production)

These combined data sources can be downloaded from the following repository: *https://dataverse.harvard.edu/dataverse/dfcropp.*

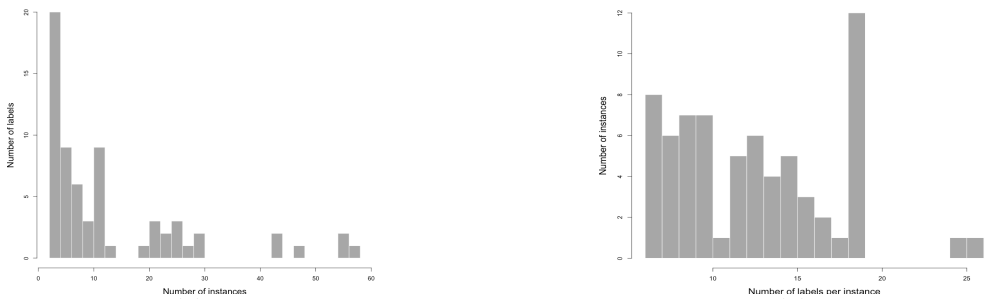
### 5.2.3 Labeling Combined Data Sources

Subsequently, we labeled each of the previous five combined datasets according to the data fusion strategy's objective, i.e., to predict the crops that can best adapt in a specific territory in the short term. This objective induces a multi-label classification problem, where a label represents a crop. Based on crop yield data in the *ep-agronet* dataset, we identify the crops in a municipality by assigning binary labels to each instance in the combined datasets. Therefore, if a municipality produced a crop in a specific year, the value one was assigned, and 0 otherwise.

### 5.2.4 Exploratory Analysis in Multi-Label Datasets

To determine the quality of data integration, we applied a multi-label exploratory analysis [44]. This analysis was composed of different plots and metrics, which allowed us to identify whether the resulting combined datasets were suitable for training a machine learning model. The detailed results of the exploratory analysis are presented in Appendix G. As an example, we present the exploratory analysis for the specific case of the IDEAM-AGRONET dataset (CDS1).

Initially, the Labels Histogram (LH), presented in Figure 5.2 (a), shows the number of labels according to the number of instances in which they are active. This plot visualizes the label dispersion per instance, i.e., if the graph accumulates on the left, the majority of labels appear in a few instances (high dispersion), while the graph accumulates on the right otherwise. In this case, the labels tend to be more dispersed as they accumulate to the left. On the other hand, the Cardinality Histogram (CH) presented in in Figure 5.2 (b), determines the number of instances depending on the cardinality, i.e., an accumulation on the left indicates a large number of instances with few labels, and an accumulation on the right represents many instances with many labels. In our example dataset, we identified a number of instances with a moderate percentage of labels.




(a) (b)

**Figure 5.2.** (a) labels and (b) cardinality histogram for the IDEAM-AGRONET MLD.

One of the most relevant plots in this exploratory analysis was the Label Bar (LB) diagram, which indicates how many instances contain a label (Figure 5.3). In this research, we added three thresholds, B1, B2, and B3, to establish the most frequent crops or labels. These thresholds allowed us to generate different variations of the original data sources as shown in section 5.3.1. In the case of the CDS1 dataset, the most relevant crops in the B3 threshold were coffee, beans, and cassava; in the B2 threshold were the three previous crops plus corn, banana, and tomato; and finally, in the B1 threshold, we found all the previous crops plus pumpkin, peas, cocoa, sugar cane, onion, coriander, fique, string beans, lulo fruit, technified corn, blackberry, orange, and potato.

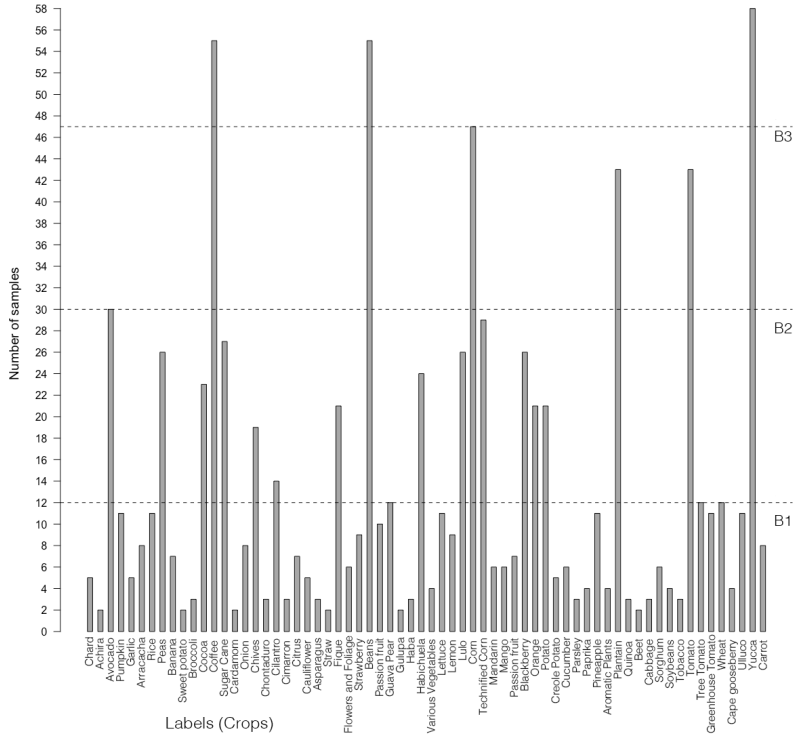



**Figure 5.3.** Labels bar diagram for the IDEAM-AGRONET MLD.

The Label Concurrence (LC) diagram shows the co-occurrences between labels for each instance. This circular diagram is divided into arcs, where each arc represents a label. The width of an arc is proportional to the number of instances in which the corresponding label appears. Different bands start from each arc, and their width is proportional to the number of instances in which the two connected labels appear together. Label concurrency is important in the study of MLDs to determine the success of a preprocessing technique [115]. Figure 5.4 shows a greater interaction among the most frequent crops previously mentioned. Nevertheless, we identified

interactions involving minority labels, as in the case of cardamom, asparagus, and achira. These interactions are mainly established with majority labels and are presented in Table 5.2, where we mention the SCUMBLE values and the number of interactions for each minority label.

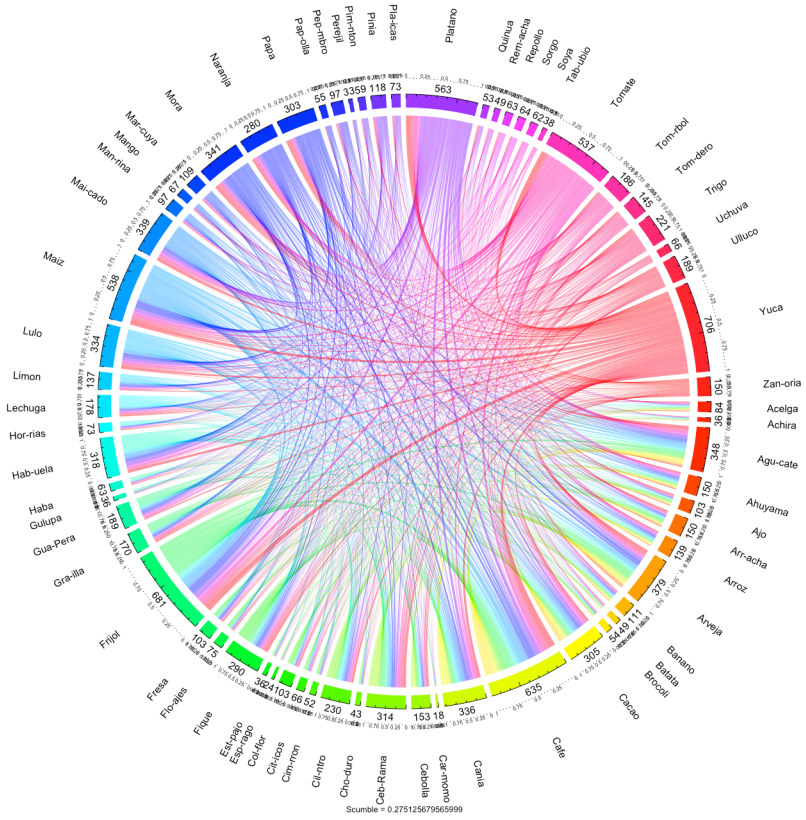



**Figure 5.4.** Concurrence among labels in IDEAM-AGRONET MLD.

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Cardamom | Asparagus | Achira |
| Aguacate | 0.26 | - | 3 | - |
| Arracacha | 0.33 | - | - | 2 |
| Arveja | 0.27 | - | - | - |
| Cafe | 0.26 | 2 | 3 | 2 |
| Cebolla | 0.35 | - | - | 2 |
| Cebolla.Rama | 0.31 | 2 | - | 2 |
| Fique | 0.34 | - | 3 | 2 |
| Frijol | 0.28 | 2 | 3 | 2 |
| Granadilla | 0.30 | - | - | 2 |
| Lulo | 0.29 | 2 | - | 2 |
| Maiz | 0.26 | 2 | 3 | 2 |
| Naranja | 0.28 | 2 | - | - |
| Platano | 0.27 | 2 | - | - |
| Tomate | 0.30 | 2 | 3 | - |
| Yuca | 0.27 | 2 | 3 | - |

**Table 5.2.** Analysis of minority labels for IDEAM-AGRONET MLD.

On the other hand, Table 5.3 presents the exploratory analysis metrics for the IDEAM-AGRONET MLD and its variations. These variations were determined following the definition established in section 3.5.1. MLD corresponds to the original dataset; RD is the decoupled dataset applying the REMEDIAL algorithm; RIPL are the datasets with the labels selected at thresholds B1, B2, and B3 respectively; and RSL are the datasets with the skewness labels respectively (majority and minority positive labels at a defined threshold according to the frequency distribution).

| Metric | MLD | RD | RIPL | | | RSL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | B1 | B2 | B3 | S1 | S2 | S3 | S4 |
| Discarded attributes | 3 | 0 | 47 | 60 | 63 | 24 | 38 | 51 | 52 |
| Number of attributes | 72 | 72 | 25 | 12 | 9 | 48 | 34 | 21 | 20 |
| Number of instances | 69 | 103 | 69 | 69 | 69 | 69 | 69 | 69 | 69 |
| Number of inputs | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| Number of labels | 66 | 66 | 19 | 6 | 3 | 42 | 28 | 15 | 14 |
| Number of labelsets | 48 | 63 | 39 | 16 | 6 | 45 | 41 | 33 | 32 |
| Number of single labelsets | 31 | 34 | 19 | 3 | 0 | 26 | 22 | 12 | 12 |
| Maximum Frequency | 3 | 5 | 3 | 12 | 40 | 3 | 3 | 3 | 6 |
| Cardinality | 12.92 | 8.66 | 8.81 | 4.36 | 2.43 | 11.78 | 10.28 | 6.17 | 5.89 |
| Density | 0.19 | 0.13 | 0.46 | 0.72 | 0.81 | 0.28 | 0.36 | 0.41 | 0.42 |
| Mean IR | 10.1 | 10.1 | 2.11 | 1.17 | 1.03 | 4.76 | 3.08 | 1.78 | 1.73 |
| SCUMBLE | 0.27 | 0.13 | 0.07 | 0.006 | 0.0002 | 0.18 | 0.12 | 0.03 | 0.03 |
| SCUMBLE CV | 0.39 | 0.77 | 0.29 | 0.33 | 0.6 | 0.38 | 0.34 | 0.26 | 0.25 |
| TCS | 9.85 | 10.12 | 8.39 | 6.35 | 4.68 | 9.33 | 8.83 | 7.99 | 7.89 |

**Table 5.3.** Metrics of exploratory analysis for the IDEAM-AGRONET MLD and its variations.

From Table 5.3 and Figure 5.5, we discarded the use of RD variation because although it reduced the SCUMBLE value (decoupling between majority and minority labels), it increased the theoretical complexity of the dataset. We also discarded the RIPL and RSL variations because although they decreased the SCUMBLE value and the theoretical complexity, the difference in the latter value was not significant considering the TCS ranges identified in Table 5.5. Therefore, we decided to keep as many labels as possible to train subsequent predictive models covering a wider variety of agricultural crops, i.e., we used the MLD with all agricultural crops.

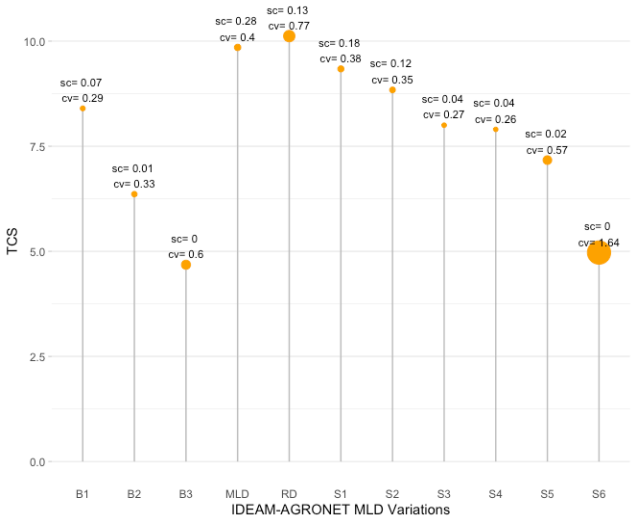



**Figure 5.5.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the IDEAM-AGRONET MLD and its variations.

The above exploratory analysis was similarly applied to the five combined datasets. In Table 5.4, we used the metrics previously described in section 3.4.4 to identify the most appropriate datasets to apply the MLC algorithms used in Level 3.

| Metric | CDS1 | CDS2 | CDS3 | CDS4 | CGD |
|---|---|---|---|---|---|
| Discarded attributes | 3 | 3 | 3 | 4 | 3 |
| Number of attributes | 72 | 87 | 80 | 62 | 125 |
| Number of instances | 69 | 80 | 39 | 55 | 73 |
| Number of inputs | 6 | 25 | 14 | 7 | 47 |
| Number of labels | 66 | 62 | 66 | 55 | 78 |
| Number of labelsets | 48 | 44 | 39 | 43 | 68 |
| Number of single labelsets | 31 | 27 | 39 | 36 | 63 |
| Maximum Frequency | 3 | 8 | 1 | 5 | 2 |
| Cardinality | 12.92 | 10.21 | 12.12 | 9.9 | 11.21 |
| Density | 0.19 | 0.16 | 0.18 | 0.18 | 0.14 |
| Mean IR | 10.1 | 18.21 | 11.41 | 8.78 | 19.56 |
| SCUMBLE | 0.27 | 0.23 | 0.28 | 0.21 | 0.29 |
| SCUMBLE.CV | 0.39 | 0.68 | 0.49 | 0.59 | 0.58 |
| TCS | 9.85 | 11.13 | 10.49 | 9.71 | 12.42 |

**Table 5.4.** Exploratory analysis metrics for the five combined data sources. *CDS1: bp-ideam + ep-agronet, CDS2: bp-corpoica +ep-agronet, CDS3: bp-sivicap + ep-agronet, CDS4: pi-dnp-pa+ ep-agronet, CGD: bp-ideam + bp-corpoica + bp-sivicap +pi-dnp-pa + ep-agronet.*

The results presented in Table 5.4 indicate an acceptable quality in all five combined data sources for the next level of the data fusion strategy (Level 3). This finding is based on three key metrics, such as TCS, SCUMBLE, and Mean IR. The first corresponds to a low value in Theoretical Complexity Score (TCS) compared to more complex MLDs used in different studies [116]. The TCS values for the combined data sources were between 9.71 and 12.42, indicating less complexity in

learning a predictive model. The second refers to the global level of unbalanced labels (Mean IR), with values between 8.78 and 19.56, which indicates an acceptable average level of imbalance in all combined data sources compared to other well-known MLDs [148]. Finally, the third metric indicates the Score of ConcUrrence among iMBal-anced LabEls (SCUMBLE) with values between 0.21 and 0.29, which indicates a low concurrence among minority and majority labels, considering that this measure is in the range [0 - 1] [119].

## 5.3. Data Analysis (Level 3)

In this section, we explore different MLC strategies and machine learning algorithms to generate models for predicting one or more target variables. Following the case study at upper Cauca river basin, we trained several predictive models with the combined data sources obtained in Level 2. The data fusion strategy's objective was to predict crops produced at a specific site in the short term. We evaluated the models with test data by applying different multi-label metrics. From these tests, we select the best models considering both crop classification and ranking. Finally, we validated these models with actual crop production data considering ordered and unordered crops in the ranking.

### 5.3.1 Model Training Scheme

We trained 2430 models following the model generation scheme previously defined in Figure 3.6. The number of models was obtained by S = N*M*P*Q, where N = 5 corresponds to the number of data sources, M = 9 to the number of data sources variations, P = 6 to the number of MLC strategies, and Q = 9 to the number of ML algorithms; therefore, S = 5*9*6*9 = 2430. Regarding the variations of the data sources, we used the original data source; the variation decoupled using the REMEDIAL algorithm; the variation with 28, 14, and 5 infrequent labels removed; and the variation with 64, 54, 48, and 36 skewness labels removed (most frequent and infrequent labels).

### 5.3.2 Model Performance Evaluation

We evaluated the above 2430 models (Appendix H) through the metrics presented in section 3.5.2. Table 5.5 summarizes the best models for each MLC strategy (MLCS) applied to the combined data sources. We did not consider the results of variations in the combined data sources, since they did not exceed the results of the original combined data sources. In the case of accuracy, precision, recall, and F1-

score metrics, we have highlighted the highest values in each combined data source. On the other hand, for hamming-loss, ranking-loss, and one-error metrics, we have highlighted the lowest values considering that these are loss functions, i.e., the best results correspond to the lowest values.

| CDS | MLCS | Accuracy | | Precision | | Recall | | F1-Score | | Hamming-Loss | | Ranking-Loss | | One-Error | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Value | Alg. | Value | Alg. | Value | Alg. | Value | Alg. | Value | Alg. | Value | Alg. | Value | Alg. |
| CDS1 | BR | 0.6 | RF | 0.83 | RF | 0.73 | NB | 0.74 | RF | 0.09 | RF | 0.04 | RF | 0.04 | SVM |
| | BRPLUS | 0.56 | RF | 0.82 | RF | 0.67 | NB | 0.69 | RF | 0.11 | RF | 0.05 | RF | 0.07 | C5.0 |
| | ECC | 0.51 | RF | 0.69 | C5.0 | 0.7 | RF | 0.66 | RF | 0.13 | RF | 0.1 | RF | 0.05 | RF |
| | HOMER | 0.12 | RF | 0.49 | RF | 0.13 | RF | 0.2 | RF | 0.19 | RF | 0.49 | SMO | 0.44 | RF |
| | LP | 0.77 | C5.0 | 0.88 | C5.0 | 0.84 | C5.0 | 0.85 | C5.0 | 0.05 | C5.0 | 0.09 | C5.0 | 0.11 | C5.0 |
| | RAKEL | 0.67 | C5.0 | 0.82 | RF | 0.79 | C5.0 | 0.79 | C5.0 | 0.08 | C5.0 | 0.07 | C5.0 | 0.04 | XGB |
| CDS2 | BR | 0.63 | SMO | 0.84 | RF | 0.78 | C5.0 | 0.74 | SMO | 0.08 | RF | 0.05 | RF | 0.03 | RF |
| | BRPLUS | 0.66 | SMO | 0.87 | RF | 0.76 | C5.0 | 0.75 | SMO | 0.09 | RF | 0.05 | RF | 0.01 | RF |
| | ECC | 0.61 | SMO | 0.73 | SMO | 0.78 | SMO | 0.72 | SMO | 0.1 | RF | 0.07 | C5.0 | 0.04 | C5.0 |
| | HOMER | 0.17 | RF | 0.47 | RF | 0.19 | RF | 0.25 | RF | 0.17 | RF | 0.46 | SMO | 0.49 | RF |
| | LP | 0.69 | SMO | 0.79 | C5.0 | 0.78 | SMO | 0.77 | SMO | 0.09 | C5.0 | 0.14 | SMO | 0.23 | C5.0 |
| | RAKEL | 0.68 | C5.0 | 0.83 | RF | 0.81 | C5.0 | 0.78 | C5.0 | 0.08 | RF | 0.07 | C5.0 | 0.09 | C5.0 |
| CDS3 | BR | 0.38 | RF | 0.75 | SVM | 0.58 | NB | 0.54 | RF | 0.14 | RF | 0.17 | RF | 0.13 | CART |
| | BRPLUS | 0.36 | RF | 0.75 | RF | 0.56 | NB | 0.52 | RF | 0.14 | RF | 0.17 | RF | 0.13 | C5.0 |
| | ECC | 0.38 | XGB | 0.57 | CART | 0.6 | XGB | 0.54 | XGB | 0.17 | RF | 0.19 | C5.0 | 0.15 | SVM |
| | HOMER | 0.06 | SMO | 0.24 | SMO | 0.07 | SMO | 0.1 | SMO | 0.22 | RF | 0.56 | SMO | 0.75 | C5.0 |
| | LP | 0.32 | C5.0 | 0.75 | XGB | 0.5 | RF | 0.47 | C5.0 | 0.16 | CART | 0.34 | RF | 0.2 | CART |
| | RAKEL | 0.39 | RF | 0.77 | SVM | 0.59 | NB | 0.55 | RF | 0.14 | RF | 0.23 | C5.0 | 0.18 | RF |
| CDS4 | BR | 0.38 | RF | 0.66 | SVM | 0.51 | RF | 0.51 | RF | 0.14 | RF | 0.11 | RF | 0.18 | RF |
| | BRPLUS | 0.39 | RF | 0.67 | RF | 0.49 | XGB | 0.51 | RF | 0.14 | RF | 0.11 | RF | 0.16 | RF |
| | ECC | 0.36 | CART | 0.56 | C5.0 | 0.54 | XGB | 0.49 | CART | 0.15 | C5.0 | 0.13 | NB | 0.22 | SVM |
| | HOMER | 0.07 | RF | 0.26 | C5.0 | 0.09 | RF | 0.11 | RF | 0.21 | C5.0 | 0.45 | SMO | 0.67 | C5.0 |
| | LP | 0.44 | NB | 0.57 | NB | 0.56 | NB | 0.54 | NB | 0.13 | NB | 0.21 | NB | 0.43 | NB |
| | RAKEL | 0.38 | XGB | 0.69 | SVM | 0.48 | NB | 0.5 | XGB | 0.13 | RF | 0.17 | XGB | 0.25 | XGB |
| CGD | BR | 0.61 | RF | 0.83 | RF | 0.74 | NB | 0.74 | RF | 0.09 | RF | 0.04 | RF | 0.04 | SVM |
| | BRPLUS | 0.56 | RF | 0.83 | RF | 0.68 | NB | 0.7 | RF | 0.11 | RF | 0.05 | RF | 0.07 | C5.0 |
| | ECC | 0.52 | RF | 0.7 | C5.0 | 0.7 | RF | 0.67 | RF | 0.13 | RF | 0.1 | RF | 0.05 | RF |
| | HOMER | 0.12 | RF | 0.49 | RF | 0.14 | RF | 0.2 | RF | 0.19 | RF | 0.49 | SMO | 0.44 | RF |
| | LP | 0.78 | C5.0 | 0.89 | C5.0 | 0.85 | C5.0 | 0.86 | C5.0 | 0.05 | C5.0 | 0.09 | C5.0 | 0.11 | C5.0 |
| | RAKEL | 0.68 | C5.0 | 0.83 | RF | 0.79 | C5.0 | 0.79 | C5.0 | 0.08 | C5.0 | 0.07 | C5.0 | 0.04 | XGB |

**Table 5.5.** Performance metrics for the best predictive models in the MLC approach. CDS: Combined Data Source, MLCS: Multi-Label Classification Strategy, Alg: Machine Learning Algorithm.

These results showed two relevant findings in the performance of predictive models. From the point of view of conventional performance measurements (Accuracy, Precision, Recall, and F1-measure), the Label Powerset (LP) strategy obtained the best results when combined with the C5.0 and Naïve Bayes algorithms in the CDS1, CDS4, and CGD datasets, while the RAKEL strategy performed well with a broader set of algorithms such as RF, C5.0, NB, and SVM. On the other hand, from the perspective of MLC-oriented metrics (hamming-loss, ranking-loss, and one-error), we only observed a similar behavior of hamming-loss concerning conventional metrics for the same combined datasets. However, for specialized label ranking evaluation metrics such as ranking-loss and one-error, the best

performances were obtained using the BR and BRPLUS strategies predominantly in conjunction with the Random Forest algorithm.

Although previous results showed better performance in some MLC strategies, it is impossible to establish a significant difference among methods at first sight. Considering the above, we performed a statistical significance test (details can be found in Appendix J) using an Analysis of Variance (ANOVA) [117]. We identified the total variance from the variance among sample groups (in this case, the groups correspond to each applied MLC strategy). We evaluated the hamming-loss metric for classification tasks and the ranking-loss metric for ranking tasks. We selected hamming-loss considering relevant aspects such as its behavior similar to conventional metrics and also because it is a metric oriented to MLC approaches. Also, ranking-loss was selected because it is the main MLC-oriented metric for evaluating label rankings. Furthermore, we also applied several a priori and a posteriori ANOVA tests to determine those groups with significant differences. The first of these analyses was the normality test, where we checked an adequate metrics distribution in each group. We used the *Shapiro-Wilk* normality test [149] because it is the most used, efficient, and useful when samples are small. This test showed high levels of normality in the group distributions for both hamming-loss and ranking-loss metrics. After checking the normality, we verified the homogeneity of variances by applying a homoscedasticity test. One of the most used is *Levene's* test [150]. We identified variations in the sample groups through this test, considering a p-value lower than 0.05 in the CDS1, CD2, and CGD datasets. After these two analyses, we applied linear ANOVA models for each combined data source. Table 5.6 summarizes the results of normality, homoscedasticity, and ANOVA analyses.

| CDS | Normality Test | | Homoscedasticity Test | | ANOVA | |
|---|---|---|---|---|---|---|
| | Hamming-Loss | Ranking-Loss | Hamming-Loss | Ranking-Loss | Hamming-Loss | Ranking-Loss |
| CDS1 | 94.4% | 92.3% | $6.5 * 10^{-15}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ |
| CDS2 | 98.1% | 90.1% | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ |
| CDS3 | 100% | 99.6% | 0.17 | 0.71 | $1.5 * 10^{-11}$ | $2.2 * 10^{-16}$ |
| CDS4 | 100% | 95.3% | 0.9 | 0.05 | $9.9 * 10^{-12}$ | $2.2 * 10^{-16}$ |
| CDS5 | 96.2% | 95.3% | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ | $2.2 * 10^{-16}$ |

**Table 5.6.** Results of normality (percentage of groups with a normal distribution), homoscedasticity(p-value), and ANOVA (p-value) analyses.

Although the homoscedasticity test showed variations in the CDS1, CD2, and CGD datasets, Table 5.6 indicates variations in all datasets, considering that p-values in ANOVA were less than 0.05. To accurately determine such variations, we analyzed

each group or model independently in each combined data source. For this purpose, we used the ANOVA results and two additional analyses, such as the *Pairwise t-test* [151] and the *Tukey's test* [152]. From these tests, we identified those specific groups (MLC strategies) where variations occurred (p-value $< 0.05$ for both hamming-loss and ranking-loss metrics.), which are summarized in Table 5.7.

| CDS | ANOVA | Pairwise T-Test | Tukey's Test |
|---|---|---|---|
| CDS1 | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with CART, NB, SMO, and SVM; LP with CART, SMO, and SVM | All MLC strategies combined with MAJORITY and RANDOM algorithms | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms |
| CDS2 | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms; LP and RAKEL with NB | All MLC strategies combined with MAJORITY and RANDOM algorithms | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms |
| CDS3 | HOMER with NB and SVM | HOMER with all algorithms; BR with NB | HOMER with all algorithms |
| CDS4 | HOMER with all algorithms | HOMER with all algorithms | HOMER with all algorithms |
| CDS5 | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms; BR, BRPLUS, and RAKEL with NB | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms | All MLC strategies combined with MAJORITY and RANDOM algorithms; HOMER with all algorithms |

**Table 5.7.** Tests to identify MLC strategies (groups) with significant differences in each combined data source.

Table 5.7 presents the groups with significant differences, of which we identified two predominant patterns. The former was the combination of the MAJORITY and RANDOM algorithms with all the tested MLC strategies. The latter was the combination of the HOMER strategy with all the applied algorithms. Contrasting these results with those presented in Table 5.7, the significant differences correspond mainly to the worst-performing MLC strategies but not to the best models. Therefore, there were no significant differences among the best performing MLC strategies. At this point, we could not establish which models were the most suitable for crop prediction. For this reason, we validated the crop rankings predicted by all models except those with significant differences previously mentioned.

### 5.3.3 Model Validation

We validated the MLC models with actual data from the Colombian Agriculture Ministry through the Agronet web platform. The predictive models were trained with data from 2012 to 2015 and validated with data from 2016 to 2018. These data were obtained from the production trends and yield of different crops by the site in Colombia. To transform the validation data into actual rankings comparable to the predicted ones, we converted crop yield data into annual growth rates from 2016 to 2018. We then averaged the growth rates by crop and ranked these

averages from highest to lowest in each municipality. To compare rankings, we initially focused on conventional correlation indices such as Pearson, Spearman, and Kendall. However, these indexes have some disadvantages for our work. For example, these only compare rankings of the same length, and they also work with an homogeneous weight distribution regardless of the items position.

Considering the above, we used two similarity measures for comparing indefinite ranked and unranked lists. For unranked lists, we used a simple similarity measure that indicates the percentage of items from the actual ranking contained in the predicted one. We refer to this measure as *Unranked Lists' Similarity (ULS)*. Furthermore, for ranked lists, we used the *Rank Biased Overlap (RBO)* [153]. This similarity measure assumes that the top rank is more important than the bottom rank. In other words, exchanges or perturbations in the top rank are more significant and more strongly penalized than those in the bottom rank. The RBO range varies from 0 to 1, where 1 corresponds to identical rankings, and 0 represents disjointed rankings. We can also adjust weight ranking positions through the $p$ parameter (between 0 and 1). For RBO, a low $p$ represents a high weighting in the top-ranking items (top-weighted). On the other hand, when $p$ is equal to 0, only the top-k items are considered ($k$ is the evaluation depth parameter). Finally, when $p$ is close to 1, weights are arbitrarily flat, and the evaluation is arbitrarily deeper in the rankings.

To determine similarity between the predicted and actual crop rankings, we extracted the ULS and RBO values for each model applied in each municipality, and then, averaged them to obtain an overall value. To display these values for a particular dataset, Figure 5.6 shows the global ULSs and RBOs of each MLC model for the CDS1 dataset.

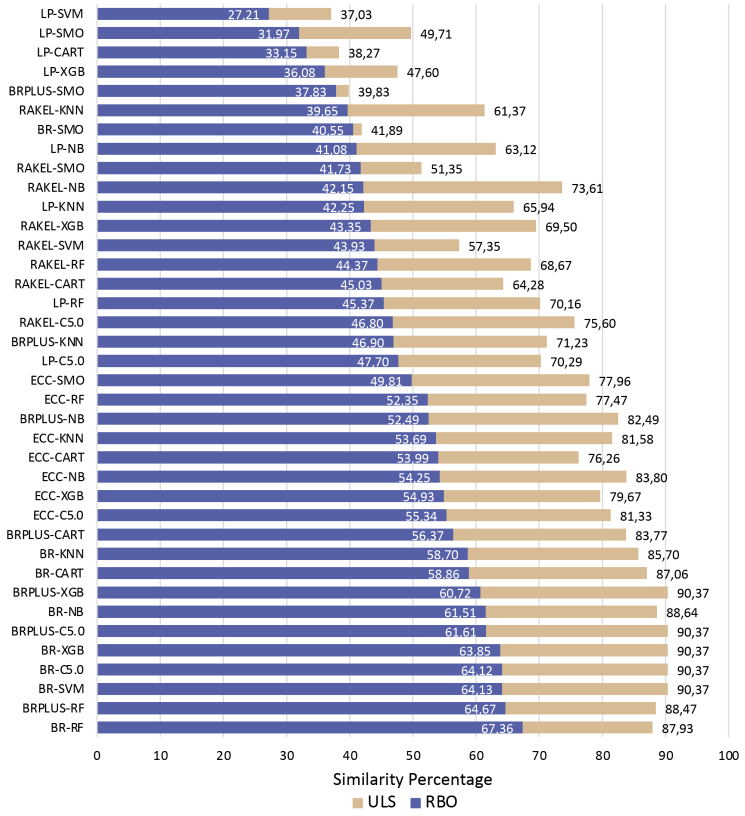



**Figure 5.6.** Global similarities (percentage values) for each MLC model in the CDS1 dataset using RBO (Rank Biased Overlap) and ULS (Unranked Lists' Similarity) metrics.

In the same line, Table 5.8 summarizes the best MLC models validated with actual crop rankings. We obtained ULS values above 90% for most of the combined data sources. These results indicate a good performance of the predictive models without considering the crop ranking. However, we prioritized the RBO measure for being more exhaustive in evaluating the position of each element within the ranking. The maximum average RBO value was 0.67 for the CDS1 dataset with the BR-RF model. On the other hand, CDS3 obtained a maximum average of 0.56 using the BR-CART model. At first sight, we could consider these values at a low level of similarity concerning to the actual rankings; how-ever, these values could be acceptable considering that the test data correspond to data unknown for the MLC models. Furthermore, although the difference was not significant be-tween these models, BR-RF obtained a high and more generalized performance in the 5 datasets. This finding allows considering the application of this unique model in the crop prediction contemplated in our proposal. As an example of the rankings predicted by the BR-RF model, Figure 5.7 presents a comparison of the actual and

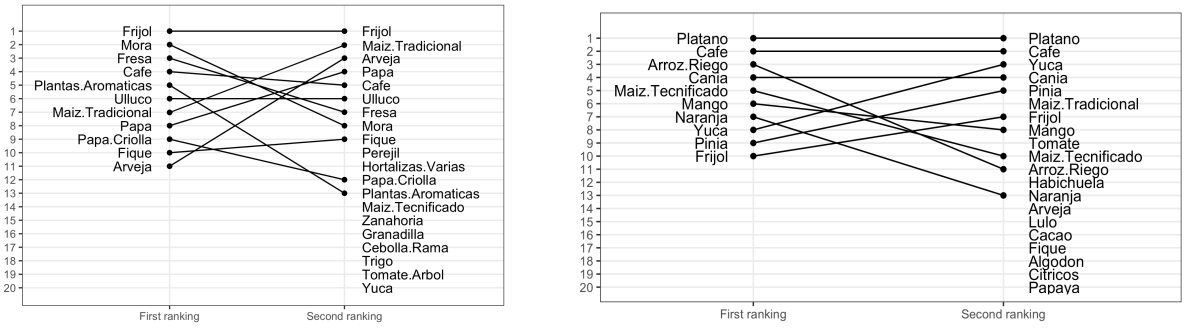
predicted rankings for the municipalities of Totoro (Figure 5.7a) and Santander de Quilichao (Figure 5.7b) in Cauca, Colombia.

| CDS | MLC Model | RBO | ULS |
|---|---|---|---|
| CDS1 | BR-RF | 0.67 | 0.87 |
| | BRPLUS-RF | 0.64 | 0.88 |
| | BR-SVM | 0.64 | 0.9 |
| | BR-C5.0 | 0.64 | 0.9 |
| | BR-XGB | 0.63 | 0.9 |
| CDS2 | BR-RF | 0.61 | 0.95 |
| | BR-XGB | 0.61 | 0.96 |
| | BRPLUS-XGB | 0.61 | 0.96 |
| | BRPLUS-RF | 0.61 | 0.96 |
| | BRPLUS-C5.0 | 0.6 | 0.96 |
| CDS3 | BR-CART | 0.56 | 0.94 |
| | BR-RF | 0.56 | 0.96 |
| | BRPLUS-CART | 0.56 | 0.93 |
| | BR-XGB | 0.55 | 0.96 |
| | BRPLUS-C5.0 | 0.54 | 0.96 |
| CDS4 | BR-RF | 0.61 | 0.94 |
| | BR-XGB | 0.59 | 0.94 |
| | BRPLUS-RF | 0.58 | 0.94 |
| | BR-NB | 0.58 | 0.91 |
| | BRPLUS-XGB | 0.58 | 0.94 |
| CGD | BRPLUS-RF | 0.65 | 0.97 |
| | BR-RF | 0.64 | 0.98 |
| | BR-XGB | 0.61 | 0.98 |
| | BRPLUS-XGB | 0.61 | 0.98 |
| | ECC-RF | 0.57 | 0.81 |

**Table 5.8.** Five highest overall RBO (Rank Biased Overlap) and its respective ULS (Unranked Lists' Similarity) values for each MLC model across all combined data sources.

As shown in the examples in Figure 5.7, the selected model is correct for most of the crops in the top-rank. Although these crops' order was not the same as the actual order (which is difficult to obtain in practical terms), the model matched in the first or second position. Furthermore, the predicted ranking gets additional crops in the bottom-rank, which could be considered by experts for further analysis of new crops' adaptation in a territory. We obtained similar results with the other municipalities considered in this study. On the other hand, we performed a final test comparing RBO values with both training and actual data. We identify similarities in the predicted crops using known and unknown data for the model through this test. Figure 5.8 presents the RBO values obtained in all municipalities using training and actual data in the CGD dataset. The other datasets showed similar behavior, with most municipalities retaining the same trend according to the results presented in Table 5.9. These results indicate a low difference between the RBO obtained with training data and actual data (Training Data RBO - Actual Data RBO). Likewise, Pearson's correlation coefficient reaffirms a high level

of agreement between these two trends (correlation coefficient close to 1 and p-value $< 0.05$). Details about the validation process can be found in Appendix K.




(a) (b)

**Figure 5.7.** Comparison of crop rankings by municipality in the (a) CDS1 and (b) CGD datasets applying the BR-RF model. The first ranking corresponds to the actual ranking, while the second is the predicted ranking.

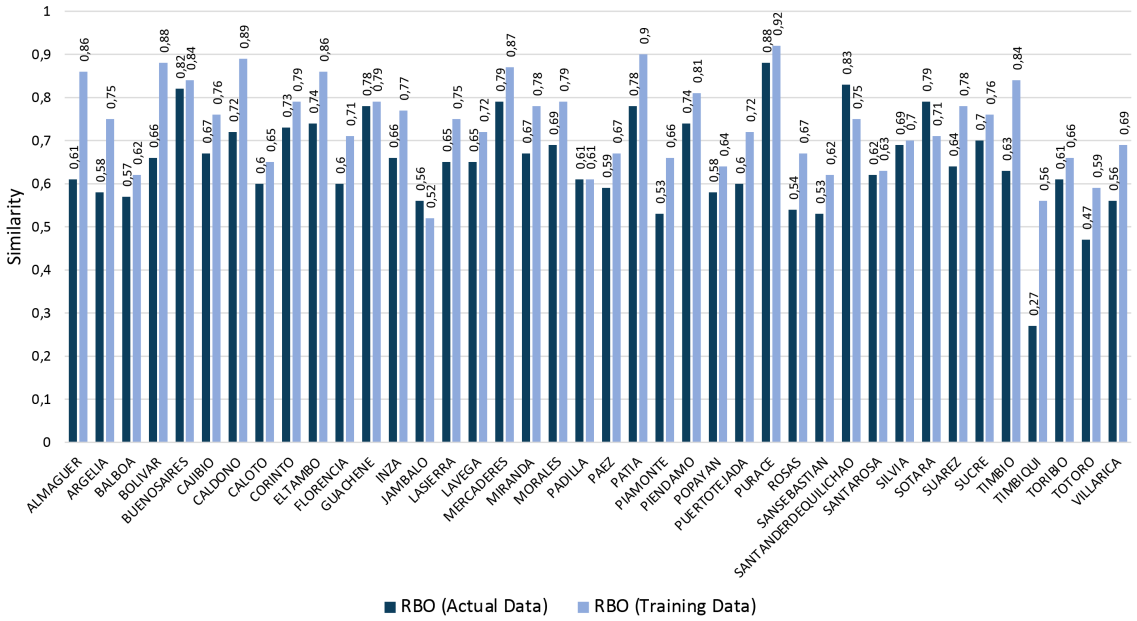



**Figure 5.8.** Global RBO values for all municipalities using the BR-RF model in the CDS1 dataset.

| CDS | Average RBO Variation | Pearson's Correlation Coefficient | p-value |
|---|---|---|---|
| CDS1 | 6% | 0.71 | $1.2 * 10^{-4}$ |
| CDS2 | 9% | 0.67 | $2.8 * 10^{-3}$ |
| CDS3 | 7% | 0.6 | $3.7 * 10^{-5}$ |
| CDS4 | 9% | 0.72 | $7.1 * 10^{-5}$ |
| CDS5 | 9% | 0.73 | $7.9 * 10^{-8}$ |

**Table 5.9.** Variation and correlation between RBOs obtained with training and actual data in all combined data sources.

## 5.4. Summary

In this chapter, we presented levels 1, 2, and 3 of the data fusion strategy applied to a case study at UCRB. In Level 1 or Relationship Analysis, we identified the most closely related data sources through a spatio-temporal meta-characteristic analysis. In Level 2 or Data Integration, we combined the previously identified data sources using an Entity Matching approach. The quality of these combined data sources was assessed using an exploratory MLC-oriented analysis. Finally, we trained and evaluated a set of MLC models from these combined data sources to predict the most probable crops to be produced in the short term in the UCRB (Cauca zone) municipalities.

# Chapter 6

# Process Refinement and User Interface

This chapter presents the fundamentals of Level 4 in the data fusion strategy. This level corresponds to the refinement of the process, which we have developed in conjunction with database management through the Data Lake concept. We describe stages such as data ingestion, data storage, data transformation and processing, and data governance. Furthermore, we present the main aspects of the software prototype developed from the data fusion strategy. In this sense, we first show the graphical user interface, and then, we describe the sequence and deployment diagrams. The former was used to show interactions between the web application objects when querying the predicted crops; and the latter to visualize the physical distribution of the prototype components. Finally, we provide the software and data availability through implementation details and the respective URL to access the data sources.

## 6.1. Process Refinement (Level 4)

We used a data lake as a bridge to connect process refinement (Level 4) with database management. This choice was based on the key principle of Level 4 mentioned in the JDL model: *"adaptive data acquisition and processing to support mission objectives."* [37]. Considering this conceptualization, a data lake offers great flexibility in data handling compared to a data warehouse. However, the arrival of the data lake does not displace the data warehouse, but rather serves to expand the organization's data management capabilities [154]. Furthermore, to reinforce this choice, other differential aspects between data lake and data warehouse are described below.

- *Storage.* Data lakes retain all data, allowing previous analysis processes to be performed again and retain data that may be useful in the future.
- *Data Types.* Data lakes support all data types, which are processed only when necessary (schema on read).
- *Users.* Although a data warehouse is user-friendly for its proper structuring, usability, and easy understanding for operational users, in data lakes, users need advanced analytical tools for predictive modeling and statistical analysis.
- *Adaptation to Changes.* Data lakes easily adapt to changes in organizational analysis objectives.
- *Data Views.* Data lakes provide faster data views based on meta-data rather than rigid tables.

Figure 6.1 presents the general scheme of data fusion managed by the implementation of a data lake. The following elements around the data can be identified: ingestion, storage, transformation, processing, and governance. These elements and their operation are described below.

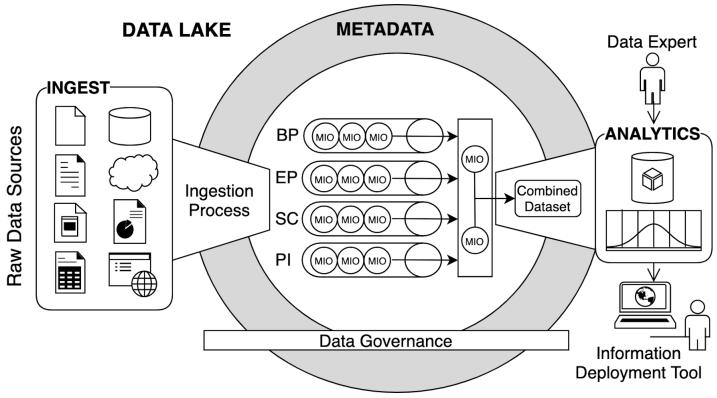



**Figure 6.1.** General outline of the storage management in the data fusion strategy. Vulnerability dimensions are represented by: Biophysics (BP), Economic-Productive (EP), Sociocultural (SC), and Political-Institutional (IP).

### 6.2.1 Data Ingestion

Data Ingestion can be described as the process of flowing data from its origin to one or more data stores, such as a data lake, databases, search engines, among others. In this case, the data were stored in the first one. To develop this process, we consider some key aspects as mentioned below.

- Initially, the data sources shown in Table 4.1 were in different formats (structured, semi-structured, and unstructured). There are two ways to transfer raw data to the data lake, by batches or data streams.
- Data lakes ecosystems defines a common format that everyone understood. The formats used to store sources are *Avro*, *SequenceFile*, and *JSON (JavaScript Object Notation)*, which support data exchange. We selected the latter because it is self-descriptive, simple, easy to understand, fast in parsing and processing, and is lightweight (bytes) in transfers [155].
- In the early stages of the data life cycle, a filtering process is performed through simple field manipulations, JSON analysis, de-duplication, and masking functions.
- Routing data from source to data stores using rules based on data attributes and automatic conversion of data types and formats [156].

Initially in Figure 6.2, we searched and collected different raw data sources. These sources can be obtained from official and private websites (with prior authorization to use the data), as well as from the results of previous climate vulnerability assessments. The data sources were in different formats such as CSV, XML, TXT, and PDF, and these were classified in each of the four dimensions as follows: Biophysical (SIVICAP, CORPOICA, IDEAM, AVA), Economic-Productive (FINAGRO, DANE-SIPSA, AGRONET, MINAGRICULTURA, AGRONET-P, DANE-SIPSA), Political-Institutional (DNP-AIB, DNP-FI, DNP-LA, DNP-PA), and Sociocultural (DANE-HH and DANE-H). The data lake was implemented on Hortonworks Data Platform (HDP), Sandbox version 2.6.5. Also, the data sources were ingested and stored in the Hadoop Distributed File System (HDFS). Data experts performed the ingestion process in the data lake (batch ingestion). This process involved developing ingest pipelines to connect a source to multiple storage destinations, therefore, IT personnel played an important role. Methods like Sqoop, Flume, and Kafka are widely used for this purpose [157]. These pipelines must be continuously monitored in order to guarantee completeness of data over time.

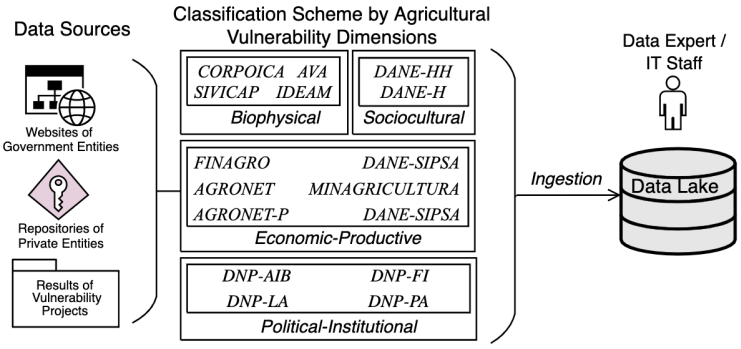



**Figure 6.2.** Initial process of data collection and grouping of data sources by dimension to be later ingested into the data lake.

### 6.2.2 Data Storage

To store data, data lakes handle a plane structure, unlike data warehouses, which use a file or directory structure. In this process, each element stored in the data lake is indexed and labeled with a unique identifier, known as an extended meta-data tag. These elements are known in the scientific literature as Managed Information Object (MIO) [123]. In many technological components, data are considered as passive entities that cannot protect themselves. Therefore, this approach is required to transform the passive data into an active entity. This entity controls or minimizes the risk of unauthorized disclosure of data. A MIO corresponds to data assets ingested in the data lake and defines a series of specific usability criteria. This object consists of *(i)* data/content, which represents confidential data (digital content) that must be protected against privacy violations, unauthorized dissemination, data leaks, among others, and *(ii)* meta-data (in XML format) to describe the content and specify various policies on its use and protection. In other words, content may have different levels of security and applicable policies to determine the specific data that can be shared with a particular organization. Two types of policies are associated with meta-data, security, and performance policies. The first involves access control, integrity control, leak control, diffusion control, audit control, etc., and the second one encapsulates performance policies such as data discovery, data mining, and quality of service. Finally, to reinforce the behavior of MIO, data and meta-data are encapsulated in an external layer responsible for data policies and control their dissemination. The main characteristics of a MIO are described below.

– *Traceable.* Each state of the object life cycle is saved to establish how it is moved, transformed, combined, and integrated.

- *Profiled.* Clear metrics to evaluate the quality of a MIO.
- *Understandable.* A MIO has the same meaning for a specific context and all stakeholders must understand it in the same way.
- *Searchable.* At the time of data ingestion, a MIO must be indexed, which allows it to be easily found through a search procedure.
- *Secure.* Methods of protection such as masking, encryption, access control lists, among others, must be defined.

Each data source is stored in the data lake as a MIO, which offers a common format for all sources. In this example, Agronet and DNP-PA MIOs are detailed. For the Agronet MIO, the following attributes were identified: crop, year, state, municipality, harvested area, sown area, production, and yield. Similarly, for the DNP-PA MIO: year, state, municipality, crop, name, alliance value, IM alliance value, Number of Beneficiaries Approved by Alliance (NBAA), and Number of Hectares per Alliance (NHA). The other stored objects have a similar structure to the two MIOs presented in Figure 6.3 and its respective attributes identified in Appendix B. The attributes Crop, Year, State, and Municipality are common in both MIOs; therefore, these were highlighted to indicate merge possibilities. Additionally, it was tagged and indexed through meta-data that identifies it for a future search process. This meta-data can group policies of provenance (origin and traceability of MIO), integrity (algorithm for checking integrity of sensitive data), privacy (access control policies for sensitive data), dissemination (authorized personnel for the dissemination of MIOs under specific conditions), life span (removal date and time of MIOs), and security (algorithms for encryption and decryption of MIOs).

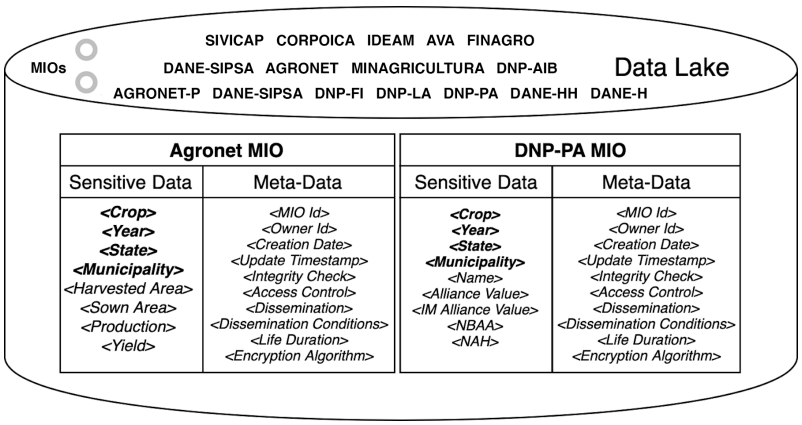



**Figure 6.3.** Storage and basic structure of MIOs in the data lake.

### 6.2.3 Data Transformation and Processing

In this process, the previously stored MIOs are available in a catalog for different user types. Consumers can request future or past information, the former using a subscription and the latter using query predicates. XPath or XQuery [158], [159] allow queries about MIOs and their respective meta-data. In this research, the users are the data analysts or data experts, who have access to these objects through information queries. This way of interacting with the stored data is based on the Data and Analytics as a Service model (supply-demand model) [160], where users browse the data catalog and select those of interest to answer their information needs (similar to a shopping cart in a supermarket). The final output of this process corresponds to a subset of data preprocessed and ready for the analysis procedures to be applied. The above is executed semi-automatically considering that data experts generate and publish analytical models, as well as submit refined or transformed datasets to the data lake to share with other stakeholders.

### 6.2.4 Data Governance

Data governance is a defined data management process, which an organization uses to ensure high data quality, and at the same time, be available throughout its life cycle [161]. Although this proposal does not focus on a data quality problem, it is important to provide the same value for a data lake as for traditional relational systems. Data governance is transversal in the implementation of data lake, offering the benefits mentioned below.

- Identification of the owner of a dataset and who to consult when there are questions about the data.
- Management of standard definitions that allow the user to know the correct values for a data element.
- Rapid evaluation of the possible use of data within a specific business process.
- Improved security of confidential data around access permissions.
- Improved data traceability and full understanding of the life cycle of the information stored in the data lake, including meta-data management.

## 6.2. Human/Computer Interface

We developed IoT-Agro (*https://www.iot-agro.com/servicios/modelocultivo*), a software prototype to deploy the predicted crops in a municipality from all the

previous results. This tool is designed to address planting and crop production processes in the municipalities of the upper Cauca river basin, Cauca zone, and some neighboring municipalities outside this basin. A user is enabled to consult the crops with a higher probability of being produced in a municipality in the coming years, considering data from different dimensions such as biophysical, economic-productive, socio-cultural, and political-institutional. We generated five predictive models (climate, soil, water quality, productive alliances, and global model) from the combination of publicly available data corresponding to these dimensions. This module of IoT-Agro allows the user to select the municipality, the combined data source, and its related variables (attributes) to consult the crop's planting probability in the short term. We used a combination of three programming languages to implement this tool. The R language was used to generate the predictive models, Java for the deployment of web services, and PHP to build the web site. Figure 6.4 shows the graphical user interface implemented in the crop prediction module of the IoT-Agro platform.

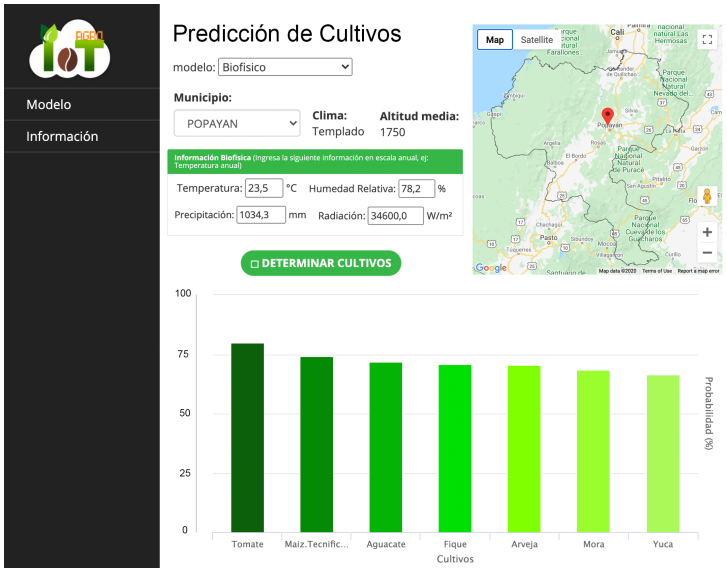



**Figure 6.4.** IoT-Agro web interface.

Figure 6.5 presents the sequence diagram to query the predicted crops. This diagram integrates elements from different architectural views by using a small set of relevant scenarios or instances of the most general use cases. We present the sequence of interactions between objects and processes as an abstraction of the most important requirements [162]. In this case, the user queries the crops to be predicted by the MLC models. The user interface collects a set of attributes, which correspond to

different conditions of a given agricultural vulnerability dimension. Subsequently, the application obtains a trained model from the combined data. Based on the conditions entered by the user, this model returns a list of the main crops and their respective probability values for the selected municipality.

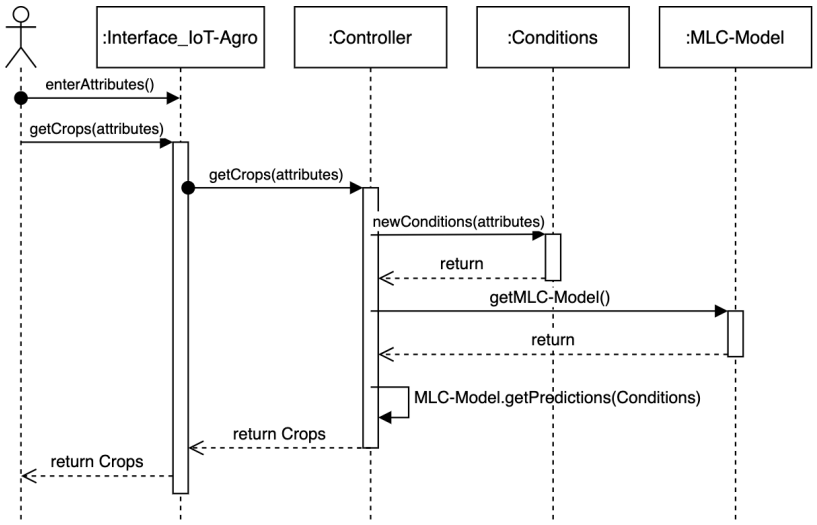



**Figure 6.5.** Sequence diagram for consulting predicted crops.

On the other hand, all the application components are visualized in the deploy diagram presented in Figure 6.6. We used a web server to deploy the IoT-Agro web application and host the web services that query the MLC models. These models were hosted on another database server, and simultaneously, a Data Lake was instantiated using the Hadoop Distributed File System (HDFS). We emphasize that the models were generated using the combined data previously stored in the Data Lake.

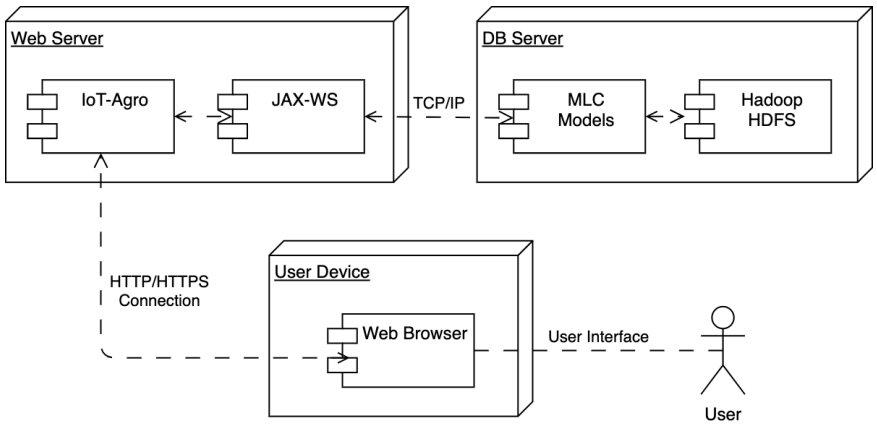



**Figure 6.6.** Deployment diagram for the crop prediction module in the IoT-Agro web application.

### 6.3.1 Software and Data Availability

The data sources used in this study were published with persistent DOIs in the *Harvard Dataverse* repository. We also present the URL and the software tools to develop the IoT-Agro web platform.

***IoT-Agro Web Platform***

- URL: `https://www.iot-agro.com/servicios/modelocultivo`
- Predictive Models: R 3.6.2 and RStudio 1.1.46
- Web Services: Java (JAX-WS)
- Web Site: CakePHP 3.6.4

***Original Data Sources***

- Climate
  - Provider: IDEAM
  - Size of Archive: 116.6 KB
  - DOI: `https://doi.org/10.7910/DVN/QQN1LG`
- Soil
  - Provider: CORPOICA (AGROSAVIA)
  - Size of Archive: 7 MB
  - DOI: `https://doi.org/10.7910/DVN/HTWRNV`
- Water Quality
  - Provider: SIVICAP
  - Size of Archive: 237.6 KB
  - DOI: `https://doi.org/10.7910/DVN/WXX7NG`
- Productive Alliances
  - Provider: DNP
  - Size of Archive: 74.2 KB
  - DOI: `https://doi.org/10.7910/DVN/PN1OTH`
- Crop Production
  - Provider: Agronet
  - Size of Archive: 219.1 KB
  - DOI: `https://doi.org/10.7910/DVN/UY7QLL`

***Combined Data Sources***

- Climate and Crop Production
  - Provider: IDEAM, Agronet
  - Size of Archive: 13.5 KB
  - DOI: `https://doi.org/10.7910/DVN/ONUDCA`

- Soil and Crop Production
  - Provider: CORPOICA (AGROSAVIA), Agronet
  - Size of Archive: 30.6 KB
  - DOI: https://doi.org/10.7910/DVN/OQFYF7
- Water Quality and Crop Production
  - Provider: SIVICAP, Agronet
  - Size of Archive: 9.3 KB
  - DOI: https://doi.org/10.7910/DVN/AWMXXL
- Productive Alliances and Crop Production
  - Provider: DNP, Agronet
  - Size of Archive: 12.1 KB
  - DOI: https://doi.org/10.7910/DVN/KAKSHN
- Climate, Soil, Water Quality, Productive Alliances, and Crop Production
  - Provider: IDEAM, CORPOICA (AGROSAVIA), SIVICAP, DNP, Agronet
  - Size of Archive: 36.7 KB
  - DOI: https://doi.org/10.7910/DVN/TOSGCM

## 6.3. Summary

This chapter presented the process refinement or Level 4 by applying the UCRB case study. We emphasized the need to implement a data lake to connect process refinement with database management. We described the main components of the data lake implementation from a data fusion strategy perspective, and how this technology enables constant refinement of resources. Finally, we presented the software prototype resulting from the data fusion strategy, the sequence and deployment diagrams, and the software and data availability through the respective access links.

# Chapter 7

# Conclusions and Future Work

This chapter describes the main conclusions obtained from the development of this doctoral thesis. Subsequently, we present the recommendations, and finally we propose future work, which can be implemented from this study.

## 7.1. Conclusions

In this study, we proposed a data fusion strategy to support Climate Vulnerability Assessments (CVA). This strategy tries to cover different issues around Climate-Smart Agri-culture (CSA), in our case, predicting the crops that could be produced in a territory in the short term. Our study area was the upper Cauca river basin in Colombia, specifically, the municipalities in the Cauca zone and some neighboring ones outside this basin. From the CSA perspective, our approach offers three main contributions. The first is a multi-dimensional data preparation process specifically oriented towards CVAs. The second corresponds to an adaptation of the JDL data fusion model to define a strategy for merging data from different agricultural vulnerability dimensions. Finally, the third contribution refers to the modeling and implementing a multi-label classification approach for crop prediction. All three are part of an overall data fusion strategy from which different stakeholders can benefit such as researchers, agricultural associations, farmers, policy-makers, decision-makers, among others.

Data processing to generate information is an important aspect of any decision-making process. Therefore, stakeholders need to reduce the uncertainty generated in complex systems. In these types of systems, environmental variables and social interaction are determinants to establish in a safe and accurate way the investment

strategies in the territories. In these areas, it is intended to improve living conditions by combining quantitative and qualitative data and reducing uncertainty in data quality. In this way, it is possible to establish decisions in public policy about risks and vulnerabilities in a region with an acceptable level of certainty.

Several studies highlight the need to implement this type of solutions [124], [163]–[165], where data from different domains can support the decision-making process in an articulated way between different organizations to focus on the needs of small, medium, and large producers. The above could allow small producers to have a direct channel with experts, who can refine their information needs. In this sense, information requirements are best understood by facilitating the refinement of analytical models applied to data sources. On the side of large business producers, this could have an impact on a better use of resources for agricultural production in the short term. It also greatly expands the vision for planting and crop production by providing new options to address potential crop shortfalls caused by climate change and variability.

Therefore, methodologies for the analysis of agricultural vulnerability in climate change contexts constitute a valuable tool to analyze and define adaptation strategies. Similarly, vulnerability assessments can guide processes of change in both land use and crops located in areas not optimal for growth and production. Vulnerability analysis and information derived from future climate scenarios provide valuable inputs for a productive reorganization of a given territory, contributing to the territorial management plans. On the other hand, experts recommend performing these analyzes periodically, as well as considering future climate scenarios, to ensure the effectiveness of adaptation and mitigation measures, responding to changing realities and needs. However, replicating these studies requires systematic processes that guarantee reliability in the new results. Therefore, the data fusion strategy proposed in this research is a guide for those experts in the area who need to replicate relevant analyzes from the countries or regions that constantly apply them. Indeed, the process of preparing multi-dimensional data provides quality information, free access, permanently updated, and at different climatic scales (national, regional, departmental, and municipal), productive systems (crop phenology, abiotic factors, crop management, etc.) and other specific and relevant socio-economic and environmental data for vulnerability analysis.

From a data driven perspective, evaluating the meta-features of a data source, before and after preprocessing, was a key aspect in this research work. An important advantage corresponds to the reduction of time and effort in the implementation and training of a large number of predictive models, which in many cases are not applied in a real or production environment. By consolidating meta-features, we were able to establish an overview around the quality of each data source. In the same way, strategies to improve this quality can be defined by analyzing each meta-feature together with the expert staff.

Furthermore, prioritizing the importance of variables is a mixed task. Different statistical or machine learning models can establish which are the most relevant attributes in a data source. However, expert knowledge assumes a leading role that cannot be ignored. The expert's point of view refined the prioritization results considering the percentage of error inherent in the automatic methods. Likewise, it should be noted that the prioritization methods are not always the same for all data sources. For labeled datasets (data with target variables or classes) it is easy to calculate a score of importance with methods such as logistic regression and random forest, among others. On the other hand, for unlabeled datasets, the previous methods cannot be applied and this prioritization must be obtained through the values of the correlation matrix. Those attributes with higher correlation values were considered more relevant in the final ranking. The ranking of each of the above methods was validated through the criteria of several experts. The degree of agreement between the automatic ranking and the ranking of raters was determined by the Fleiss's Kappa measure. Having a large number of variables and raters, it is difficult to obtain a high level of agreement. However, the results of this statistical measure obtained acceptable values for most of the data sources and, at a glance, a certain coincidence can be observed between groups of variables. Likewise, for the labeled datasets, a better concordance was found for the ranking of the logistic regression method.

On the other hand, through a relationship analysis, we identified the main combined data sources. For this purpose, we used Spatio-Temporal Meta-Features (STMF) for guiding the integration process. A key finding was to identify a central dataset for labeling all Combined Data Sources (CDS). It allowed us to adjust the CDS to our crop prediction objective, considering that the Agronet dataset (the central dataset) was related to information about production and crop yield per municipality. Therefore, the labeling process should be done in parallel with the integration,

considering the target variables. In our case, such target variables corresponded to the crops associated with a municipality. These questions involve a single target variable and a set of them, so we used the Multi-Label Learning (MLL) approach. In this way, assessing the CDS quality became a relevant factor for the subsequent stage of generating predictive models. We performed this evaluation using an exploratory analysis oriented to Multi-Label Datasets (MLD). This analysis provided key metrics such as Theoretical Complexity Score (TCS) to identify those CDS that might be most appropriate to train subsequent predictive models.

Furthermore, we selected the BR-RF (Binary Relevance - Random Forest) model to perform the crop prediction by evaluating and validating the predictive models. This model comprises a Multi-Label Classification (MLC) strategy and a Machine Learning (ML) algorithm. We selected this model considering two important findings. The first one was related to prediction and ranking tasks. Although we found models with better prediction performance, such as Label Powerset (LP) and Random k-label sets (RAKEL), the results were poor in crop probability ranking. The second finding derived from the previous one, when we tried to balance the prediction and ranking performances. We applied statistical significance tests to check that there were no significant differences in the prediction task with those better models for ranking. These tests and the validation with actual crop production data allowed us to select the BR-RF model as the best performance for our final prediction objective. While the ULS (Unranked Lists' Similarity) exceeded 90% regardless of the order of elements in both the predicted and actual rankings, RBO (Rank Biased Overlap) similarity reached a maximum of 67% strictly considering the order. Nevertheless, these results indicates that for more exhaustive ranking comparisons, the last similarity percentage is acceptable, considering the difficulty in comparing the position of each element within the ranking. In this sense, we can use the predicted rankings to provide crop recommendations at the same level of relevance, i.e., which crops could be produced in the short term without considering the probability (ranking positions). On the other hand, if we provide a ranking of crops, we require a strategy to improve the RBO similarity in the predictive models. This strategy should be oriented towards the collection and processing of new data sources, that allow us to feed back the models by better approximating real-world conditions.

Finally, the use of intelligent decision-making tools represents an advantage in different domains, particularly in agriculture. Although developing mechanisms to

merge heterogeneous data sources involves a high degree of complexity, current technological advances allow these tasks to be implemented more reliably. Data lakes represent a promising alternative around different aspects such as flexibility, reuse, scalability, and access protocols. These aspects allow linking expert personnel in data analysis, updating and reusing analytical models that support the information needs of different users. On the other hand, it is essential to ensure cleanliness and order in a data lake, otherwise there is a risk of having an unmanageable data swamp for all stakeholders. A data swamp is a data lake where there is a broken ingestion process, thereby data do not include descriptive meta-data and a mechanism to maintain it, and ultimately a big pile of effectively unusable data is obtained.

## 7.2. Future Work

Although the data fusion strategy was validated in the context of crop prediction, we intend to implement it not only for another objective of agricultural vulnerability assessments, but also in other knowledge domains. This implies strengthening the strategy through a generalizable approach, which applies it transversally without depending on the application domain. Currently, a proposal is being developed to implement this data fusion strategy in the environmental domain, specifically for water resource management. Likewise, implementing this strategy in a Climate-Smart Village is a relevant aspect to complement the validation process using a test environment and subsequently be implemented in a real farm context.

This data fusion strategy has many components running automatically, however, there are processes that are still developed with the assistance of the data analysis expert. One of these processes is the labeling of MLDs, which requires expert supervision to extract key information to determine the respective labels of an instance. With a more automated strategy in mind, it is essential to implement an automatic labeling process for MLDs, specifically for this case study (crop prediction).

In order to improve the data source integration process, it is essential to optimize the similarity function used in Entity Matching. In this research, we used a basic similarity function based on the Jaro-Winkler approach. However, there are more specialized similarity functions (natural language processing), which not only compare two text strings, but also perform a semantic analysis of the context to determine if two entities are the same.

The success of data-driven approaches lies not only in a large amount of data and information, but also in the confidence that good data quality provides. Refining the fusion

strategy with more quality data sources would improve the performance of predictive models to decrease the gap between predicted and actual crops. The data gathering process requires greater exhaustiveness and involving more stakeholders by sharing relevant data.

Our strategy should be compared with other approaches, for example, Multi-View Multi-Label Learning. This new approach allows modeling each instance from multiple views. A view describes the same object from another perspective. In our case, we could model a view from the attributes of an agricultural vulnerability dimension. Currently, several multi-view multi-label algorithms have been implemented and could be compared to the algorithms used in this dissertation.

# References


[1] FAO, The State of the World's Land and Water Resources for Food and Agriculture: Managing Systems at Risk. Food and Agriculture Organization of the United Nations, Rome and Earthscan, London, 2011.

[2] P. Gut, D. Ackerknecht, and S. K. für A. T. am ILE, Climate Responsive Building: Appropriate Building Construction in Tropical and Subtropical Regions. SKAT, 1993.

[3] United-Nations, "World Population Prospects: The 2008 Revision," Popul. Dev. Rev., vol. 36, no. 4, pp. 854–855, 2010, doi: 10.1111/j.1728-4457.2010.00368.x.

[4] H. C. J. Godfray et al., "Food Security: The Challenge of Feeding 9 Billion People," Science, vol. 327, no. 5967, pp. 812–818, Feb. 2010, doi: 10.1126/science.1185383.

[5] A. D. Tripathi, R. Mishra, K. K. Maurya, R. B. Singh, and D. W. Wilson, "Estimates for World Population and Global Food Availability for Global Health," in The Role of Functional Food Security in Global Health, Elsevier, 2019, pp. 3–24. doi: 10.1016/B978-0-12-813148-0.00001-3.

[6] TheWorldBank, Agriculture and Rural Development. 2019. Accessed: Feb. 13, 2019. [Online]. Available: https://data.worldbank.org/topic/agriculture-and-rural-development?view=chart

[7] FAO, The State of Food and Agriculture - Leveraging Food Systems for Inclusive Rural Transformation, 1st ed. Food and Agriculture Organization of the United Nations, 2017.

[8] D. K. Ray, P. C. West, M. Clark, J. S. Gerber, A. V. Prishchepov, and S. Chatterjee, "Climate change has likely already affected global food production," PLOS ONE, vol. 14, no. 5, pp. 1–18, 2019, doi: 10.1371/journal.pone.0217148.

[9] R. B Singh et al., "Functional Food Security and the Heart," J. Cardiol. Ther., vol. 4, no. 1, pp. 599–607, 2017, doi: 10.17554/j.issn.2309-6861.2017.04.125.

[10] USAID, Climate-Resilient Development: A Framework for Understanding and Addressing Climate Change, U.S. Agency for International Development. 2014. Accessed: Nov. 07, 2020. [Online]. Available: http://www.usaid.gov/sites/default/files/documents/1865/climate-resilient-developmentframework.pdf

[11] J. J. McCarthy and IPCC, Eds., Climate change 2001: impacts, adaptation, and vulnerability: contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge, UK ; New York: Cambridge University Press, 2001.

[12] S. Tao, Y. Xu, K. Liu, J. Pan, and S. Gou, "Research Progress in Agricultural Vulnerability to Climate Change," Adv. Clim. Change Res., vol. 2, no. 4, pp. 203–210, 2011, doi: 10.3724/SP.J.1248.2011.00203.

[13] USAID, Designing Climate Vulnerability Assessments, U.S. Agency for International Development. 2018. Accessed: Nov. 07, 2020. [Online]. Available: https://www.climatelinks.org/sites/default/files/asset/document/2018_USAID-ATLAS-Project_Designing-Climate-Vulnerability-Assessments.pdf

[14] FAO, Climate Smart Agriculture Sourcebook, Climate-smart crop production practices and technologies. 2012. Accessed: Oct. 07, 2020. [Online]. Available: http://www.fao.org/climate-smart-agriculture-sourcebook/production-resources/module-b1-crops/chapter-b1-2/en/

[15] D. Cash, W. C. Clark, R. Corell, N. Dickson, J. M. Hall, and E. Parson, "Assessing Vulnerability to Global Environmental Risks," in Report of the Workshop on Vulnerability to Global Environmental Change - Challenges for Research, Assessment and Decision Making, Airlie House, Warrenton, Virginia, Sep. 2000, p. 7.

[16] D. Schröter, C. Polsky, and A. G. Patt, "Assessing vulnerabilities to the effects of global change: an eight step approach," Mitig. Adapt. Strateg. Glob. Change, vol. 10, no. 4, pp. 573–595, Oct. 2005, doi: 10.1007/s11027-005-6135-9.

[17] C. H. Ramírez, J. B. Valencia, and C. F. O. Paniagua, "Modelos de Vulnerabilidad Agrícola ante los efectos del cambio climático," CIMEXUS, vol. 9, no. 2, pp. 31-48–48, Jan. 2015.

[18] CGIAR, "Encuesta rural sobre intervenciones climaticas inteligentes," 2016. http://cac.foodsecurityportal.org/regional-sub-portal-blog-entry/latin- america/886/riesgo-y-resiliencia (accessed Oct. 17, 2018).

[19] IFPRI, "The international model for policy analysis of agricultural commodities and trade (impact): Model description for version 3," 2015. http://www.ifpri.org/publication/international-model-policy-analysis- agricultural-commodities-and-trade-impact-model-0 (accessed Oct. 18, 2018).

[20] CGIAR and CCAFS, Apoyando a Honduras en la construcción de una estrategia más robusta de adaptación al clima. 2014. Accessed: Oct. 17, 2018. [Online]. Available: https://ccafs.cgiar.org/es/blog/apoyando-honduras-en-la-construccion-de-una-estrategia-mas-robusta-de-adaptacion-al-clima

[21] CDKN, "Agricultura, Vulnerabilidad y Adaptación: metodología para medir la vulnerabilidad del sector agrícola - Climate and Development Knowledge Network," 2011. http://cdkn.org/ project/agricultura-vulnerabilidad-adaptacion-cuenca-alta-cauca/ (accessed Jul. 10, 2018).

[22] FAO, Climate Smart Agriculture Sourcebook, Conducting assessments and appraisals. 2012. Accessed: Oct. 07, 2020. [Online]. Available: http://www.fao.org/climate-smart-agriculture-sourcebook/enabling-frameworks/module-c8-impact-assessments/chapter-c8-2/en/

[23] USAID, Climate Change & Development Strategy: 2012-2016, U.S. Agency for International Development. Washington, DC, 2012. Accessed: Nov. 07, 2020. [Online]. Available: http://pdf.usaid.gov/pdf_docs/PDACS780.pdf

[24] TheWorldBank, Open Data + Agriculture Can Transform How Farmers Respond to Looming Crises. 2013. Accessed: Jul. 13, 2020. [Online]. Available: https://www.worldbank.org/en/news/feature/2013/04/26/open-data-can-transform-farmers-response-to-crisis

[25] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," in Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, Swinton, UK, UK, 2008, pp. 68–77. Accessed: Feb. 04, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2227115.2227123

[26] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Keele University and Durham University Joint Report, UK, EBSE 2007-001, 2007. Accessed: May 05, 2011. [Online]. Available: http://www.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf

[27] L. Lipper et al., "Climate-smart agriculture for food security," Nat. Clim. Change, vol. 4, p. 1068 EP-, Nov. 2014.

[28] E. Sharma, "Chapter 2 Climate Change and its Impacts in the Hindu Kush-Himalayas: An Introduction," in Community, Environment and Disaster Risk Management, vol. 11, A. Lamadrid and I. Kelman, Eds. Emerald Group Publishing Limited, 2012, pp. 17–32. doi: 10.1108/S2040-7262(2012)0000011008.

[29] USAID, "Climate Vulnerability Assessment - An Annex to the USAID Clomate-Resilient Development Framework," U.S. Agency for International Development, Washington, DC, Technical, 2016. Accessed: Oct. 13, 2020. [Online]. Available: https://pdf.usaid.gov/pdf_docs/PA00KZ84.pdf

[30] J.-M. Faurès, M. Bernardi, and R. Gommes, "There Is No Such Thing as an Average: How Farmers Manage Uncertainty Related to Climate and Other Factors," Int. J. Water Resour. Dev., vol. 26, no. 4, pp. 523–542, Dec. 2010, doi: 10.1080/07900627.2010.519515.

[31] A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012. doi: 10.1016/C2011-0-06130-6.

[32] X. L. Dong and D. Srivastava, Big Data Integration. Morgan & Claypool Publishers, 2015.

[33] M. Cocchi, "Chapter 1 - Introduction: Ways and Means to Deal With Data From Multiple Sources," in Data Fusion Methodology and Applications, vol. 31, M. Cocchi, Ed. Elsevier, 2019, pp. 1–26. doi: 10.1016/B978-0-444-63984-4.00001-6.

[34] J. Esteban, A. Starr, R. Willetts, P. Hannah, and P. Bryanston-Cross, "A Review of Data Fusion Models and Architectures: Towards Engineering Guidelines," Neural Comput. Appl., vol. 14, no. 4, pp. 273–281, Dec. 2005, doi: 10.1007/s00521-004-0463-7.

[35] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition," IEEE Trans. Inf. Forensics Secur., vol. 11, no. 9, pp. 1984–1996, Sep. 2016, doi: 10.1109/TIFS.2016.2569061.

[36] O. Sidek and S. A. Quadri, "A review of data fusion models and systems," Int. J. Image Data Fusion, vol. 3, no. 1, pp. 3–21, Mar. 2012, doi: 10.1080/19479832.2011.645888.

[37] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," in Sensor Fusion: Architectures, Algorithms, and Applications III, 1999, vol. 3719, pp. 430–441. doi: 10.1117/12.341367.

[38] I. D. López, J. F. Grass, A. Figueroa, and J. C. Corrales, "A proposal for a multi-domain data fusion strategy in a climate-smart agriculture context," Int. Trans. Oper. Res., pp. 1–22, 2020, doi: 10.1111/itor.12899.

[39] G. Parmigiani, "Decision Theory: Bayesian," in International Encyclopedia of the Social and Behavioral Sciences, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 3327–3334. doi: 10.1016/B0-08-043076-7/00403-4.

[40] M. Sugiyama, "Chapter 26 - Least Squares Classification," in Introduction to Statistical Machine Learning, M. Sugiyama, Ed. Boston: Morgan Kaufmann, 2016, pp. 295–302. doi: 10.1016/B978-0-12-802121-7.00037-6.

[41] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus, Multilabel Classification - Problem Analysis, Metrics and Techniques. Springer International Publishing, 2016. doi: 10.1007/978-3-319-41111-8.

[42] E. Gibaja and S. Ventura, "A Tutorial on Multilabel Learning," ACM Comput. Surv., vol. 47, no. 3, Apr. 2015, doi: 10.1145/2716262.

[43] A. Rivolli and A. C. de Carvalho, "The utiml Package: Multi-label Classification in R," R J., vol. 10, no. 2, pp. 24–37, 2018, doi: 10.32614/RJ-2018-041.

[44] D. Charte and F. Charte, "mldr: Paquete R para Exploración de Datos Multietiqueta," in XVI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2015), Albacete, Spain, 2015, pp. 695–704.

[45] F. Charte and D. Charte, "Working with Multilabel Datasets in R: The mldr Package," R J., vol. 7, pp. 149–162, 2015, doi: 10.32614/RJ-2015-027.

[46] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A Lazy Learning Approach to Multi-Label Learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.

[47] E. Gibaja and S. Ventura, "Multi-label learning: a review of the state of the art and ongoing research," WIREs Data Min. Knowl. Discov., vol. 4, no. 6, pp. 411–444, 2014, doi: 10.1002/widm.1139.

[48] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," Int. J. Data Warehous. Min. IJDWM, vol. 2007, pp. 1–13, 2007, doi: 10.4018/jdwm.2007070101.

[49] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-label Data," in Data Mining and Knowledge Discovery Handbook, Boston, MA: Springer US, 2010, pp. 667–685. doi: 10.1007/978-0-387-09823-4_34.

[50] E. Alvares-Cherman, J. Metz, and M. C. Monard, "Incorporating label dependency into the binary relevance framework for multi-label classification," Expert Syst. Appl., vol. 39, no. 2, pp. 1647–1655, 2012, doi: 10.1016/j.eswa.2011.06.056.

[51] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier Chains for Multi-label Classification," in Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, 2009, pp. 254–269. doi: 10.1007/978-3-642-04174-7_17.

[52] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and Efficient Multilabel Classification in Domains with Large Number of Labels," in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery, Workshop on Mining Multidimensional Data, Antwerp, Belgium, 2008, pp. 30–44.

[53] G. Tsoumakas and I. Vlahavas, "Random k-Labelsets: An Ensemble Method for Multilabel Classification," in Machine Learning: ECML 2007, Berlin, Heidelberg, 2007, pp. 406–417. doi: 10.1007/978-3-540-74958-5_38.

[54] M. Aria and C. Cuccurullo, "bibliometrix : An R-tool for comprehensive science mapping analysis," J. Informetr., vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[55] F. J. Rodríguez-Sedano, "Uso de herramienta on-line Parsifal para la elaboración de una revisión sistemática de la literatura (SLR)," Mar. 2019, doi: 10.5281/ZENODO.2603914.

[56] M. Greenacre and J. Blasius, Eds., Multiple Correspondence Analysis and Related Methods, 0 ed. Chapman and Hall/CRC, 2006. doi: 10.1201/9781420011319.

[57] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in Software Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. Accessed: Nov. 05, 2015. [Online]. Available: http://link.springer.com/10.1007/978-3-642-29044-2

[58] A. R. Muljarto, J.-M. Salmon, B. Charnomordic, P. Buche, A. Tireau, and P. Neveu, "A generic ontological network for Agri-food experiment integration – Application to viticulture and winemaking," Comput. Electron. Agric., vol. 140, pp. 433–442, 2017, doi: 10.1016/j.compag.2017.06.020.

[59] C. Liao, J. Wang, T. Dong, J. Shang, J. Liu, and Y. Song, "Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean," Sci. Total Environ., vol. 650, pp. 1707–1721, 2019, doi: 10.1016/j.scitotenv.2018.09.308.

[60] K. S. Veum, K. A. Sudduth, R. J. Kremer, and N. R. Kitchen, "Sensor data fusion for soil health assessment," Geoderma, vol. 305, pp. 53–61, 2017, doi: 10.1016/j.geoderma.2017.05.031.

[61] W. Jing and L. Xin, "Progresses on data fusion technology of crop growth model and multi-source observation information," Remote Sens. Technol. Appl., vol. 30, no. 2, pp. 209–219, 2015, doi: 10.11873/j.issn.1004-0323.2015.2.0209.

[62] J. Gai, L. Tang, and B. Steward, "Plant localization and discrimination using 2D+3D computer vision for robotic intra-row weed control," in 2016 American Society of Agricultural and Biological Engineers Annual International Meeting, ASABE 2016, 2016, pp. 1–15. doi: 10.13031/aim.20162460814.

[63] P. Grover and R. Johari, "PAID: Predictive agriculture analysis of data integration in India," in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 184–188.

[64] L. Piedelobo et al., "Scalable pixel-based crop classification combining Sentinel-2 and Landsat-8 data time series: Case study of the Duero river basin," Agric. Syst., vol. 171, pp. 36–50, 2019, doi: 10.1016/j.agsy.2019.01.005.

[65] D. Moshou, I. Gravalos, D. K. C. Bravo, R. Oberti, J. S. West, and H. Ramon, "Multisensor Fusion of Remote Sensing Data for Crop Disease Detection," in Geospatial Techniques for Managing Environmental Resources, J. K. Thakur, S. K. Singh, A. Ramanathan, M. B. K. Prasad, and W. Gossel, Eds. Springer Netherlands, 2011, pp. 201–219. doi: 10.1007/978-94-007-1858-6_13.

[66] A. M. Mouazen, S. A. Alhwaimel, B. Kuang, and T. Waine, "Multiple on-line soil sensors and data fusion approach for delineation of water holding capacity zones for site specific irrigation," Soil Tillage Res., vol. 143, pp. 95–105, 2014, doi: 10.1016/j.still.2014.06.003.

[67] C. Pohl, K. D. Kanniah, and C. K. Loong, "Monitoring oil palm plantations in Malaysia," in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Jul. 2016, pp. 2556–2559. doi: 10.1109/IGARSS.2016.7729660.

[68] L. Tits, B. Somers, J. Stuckens, J. Farifteh, and P. Coppin, "Integration of in situ measured soil status and remotely sensed hyperspectral data to improve plant production system monitoring: Concept, perspectives and limitations," Remote Sens. Environ., vol. 128, pp. 197–211, 2013, doi: 10.1016/j.rse.2012.10.006.

[69] J. Zhang, L. Feng, and F. Yao, "Improved maize cultivated area estimation over a large scale combining MODIS-EVI time series data and crop phenological information," ISPRS J. Photogramm. Remote Sens., vol. 94, pp. 102–113, 2014, doi: 10.1016/j.isprsjprs.2014.04.023.

[70] J. Meng, X. Du, and B. Wu, "Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation," Int. J. Digit. Earth, vol. 6, no. 3, pp. 203–218, 2013, doi: 10.1080/17538947.2011.623189.

[71] D. De Benedetto, A. Castrignano, M. Diacono, M. Rinaldi, S. Ruggieri, and R. Tamborrino, "Field partition by proximal and remote sensing data fusion," Biosyst. Eng., vol. 114, no. 4, pp. 372–383, 2013, doi: 10.1016/j.biosystemseng.2012.12.001.

[72] A. De Sherbinin et al., "Data integration for climate vulnerability mapping in West Africa," ISPRS Int. J. Geo-Inf., vol. 4, no. 4, pp. 2561–2582, 2015, doi: 10.3390/ijgi4042561.

[73] L. Meng et al., "Assessment of the effectiveness of spatiotemporal fusion of multi-source satellite images for cotton yield estimation," Comput. Electron. Agric., vol. 162, pp. 44–52, 2019, doi: 10.1016/j.compag.2019.04.001.

[74] I. D. López and J. C. Corrales, "A Smart Farming Approach in Automatic Detection of Favorable Conditions for Planting and Crop Production in the Upper Basin of Cauca River," in Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change, vol. 687, P. Angelov, J. A. Iglesias, and J. C. Corrales, Eds. Cham: Springer International Publishing, 2018, pp. 223–233. doi: 10.1007/978-3-319-70187-5_17.

[75] H. Mousannif and J. Zahir, "AgriFuture: A New Theory of Change Approach to Building Climate-Resilient Agriculture," in Advanced Intelligent Systems for Sustainable Development (AI2SD'2018), vol. 911, M. Ezziyyani, Ed. Cham: Springer International Publishing, 2019, pp. 88–97. doi: 10.1007/978-3-030-11878-5_10.

[76] W. Ji et al., "Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study," Geoderma, vol. 341, pp. 111–128, 2019, doi: 10.1016/j.geoderma.2019.01.006.

[77] M. L. Mann and J. M. Warner, "Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach," Field Crops Res., vol. 201, pp. 60–74, 2017, doi: 10.1016/j.fcr.2016.10.014.

[78] X. E. Pantazi, D. Moshou, A. M. Mouazen, T. Alexandridis, and B. Kuang, "Data Fusion of Proximal Soil Sensing and Remote Crop Sensing for the Delineation of Management Zones in Arable Crop Precision Farming," in HAICTA, 2015, pp. 765–776.

[79] B. Kaur and R. K. A. Owusu, "Inverse Problems and Data Fusion for Crop Production Applications Targeting Optimal Growth - Fertilization," in 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), Sep. 2015, pp. 108–114. doi: 10.1109/DEXA.2015.39.

[80] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, "Soybean yield prediction from UAV using multimodal data fusion and deep learning," Remote Sens. Environ., vol. 237, p. 111599, 2020, doi: 10.1016/j.rse.2019.111599.

[81] L. Wang, Y. Tian, X. Yao, Y. Zhu, and W. Cao, "Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images," Field Crops Res., vol. 164, pp. 178–188, 2014, doi: https://doi.org/10.1016/j.fcr.2014.05.001.

[82] N. You and J. Dong, "Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine," ISPRS J. Photogramm. Remote Sens., vol. 161, pp. 109–123, 2020, doi: 10.1016/j.isprsjprs.2020.01.001.

[83] P. Schweizer and N. Stein, "Large-scale data integration reveals colocalization of gene functional groups with meta-QTL for multiple disease resistance in barley," Mol. Plant. Microbe Interact., vol. 24, no. 12, pp. 1492–1501, 2011, doi: 10.1094/MPMI-05-11-0107.

[84] M. Ji, K. Zhang, Q. Wu, and Z. Deng, "Multi-label learning for crop leaf diseases recognition and severity estimation based on convolutional neural networks," Soft Comput., 2020, doi: 10.1007/s00500-020-04866-z.

[85] A. Picon, M. Seitz, A. Alvarez-Gila, P. Mohnke, A. Ortiz-Barredo, and J. Echazarra, "Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions," Comput. Electron. Agric., vol. 167, 2019, doi: 10.1016/j.compag.2019.105093.

[86] A. A. Abd El-aziz, A. Darwish, D. Oliva, and A. E. Hassanien, "Machine Learning for Apple Fruit Diseases Classification System," Adv. Intell. Syst. Comput., vol. 1153 AISC, pp. 16–25, 2020, doi: 10.1007/978-3-030-44289-7_2.

[87] H. Omrani, A. Tayyebi, and B. Pijanowski, "Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based Land Transformation Model: an integrated ML-CA-LTM modeling framework," GIScience Remote Sens., vol. 54, no. 3, pp. 283–304, 2017, doi: 10.1080/15481603.2016.1265706.

[88] H. Omrani, F. Abdallah, O. Charif, and N. T. Longford, "Multi-label class assignment in land-use modelling," Int. J. Geogr. Inf. Sci., vol. 29, no. 6, pp. 1023–1041, 2015, doi: 10.1080/13658816.2015.1008004.

[89] A. Tharwat, H. Mahdi, and A. E. Hassanien, "Plant recommender system based on multi-label classification," in International Conference on Advanced Intelligent Systems and Informatics, 2016, pp. 825–835. doi: 10.1007/978-3-319-48308-5_79.

[90] Y. Zhong and M. Zhao, "Research on deep learning in apple leaf disease recognition," Comput. Electron. Agric., vol. 168, p. 105146, 2020, doi: 10.1016/j.compag.2019.105146.

[91] I. Shendryk, Y. Rist, R. Lucas, P. Thorburn, and C. Ticehurst, "Deep Learning - a New Approach for Multi-Label Scene Classification in Planetscope and Sentinel-2 Imagery," in IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 1116–1119. doi: 10.1109/IGARSS.2018.8517499.

[92] K. Kulkarni and P. A. Vijaya, "A comparitive study of land classification using remotely sensed data," in 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 36–41. doi: 10.1109/ICCMC.2017.8282720.

[93] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms," in Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697349.

[94] R. B. Bairi, A. Vani, P. Ahuja, and G. Ramakrishnan, "Categorising Videos Using a Personalised Category Catalogue," in Proceedings of the Second ACM IKDD Conference on Data Sciences, New York, NY, USA, 2015, pp. 49–58. doi: 10.1145/2732587.2732594.

[95] Doreswamy, I. Gad, and B. R. Manjunatha, "Multi-label Classification of Big NCDC Weather Data Using Deep Learning Model," in Soft Computing Systems, Singapore, 2018, pp. 232–241.

[96] D. C. Corrales, A. Ledezma, and J. C. Corrales, "From Theory to Practice: A Data Quality Framework for Classification Tasks," Symmetry, vol. 10, no. 7, 2018, doi: 10.3390/sym10070248.

[97] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Inf. Sci., vol. 233, pp. 25–35, 2013, doi: https://doi.org/10.1016/j.ins.2013.01.021.

[98] D. M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.

[99] V. Barnett and T. Lewis, Outliers in statistical data. Wiley, 1978.

[100] B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: A Survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 5, pp. 845–869, May 2014, doi: 10.1109/TNNLS.2013.2292894.

[101] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans Knowl Data Eng, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[102] M. Bosu and S. Macdonell, "A Taxonomy of Data Quality Challenges in Empirical Software Engineering," in Proceedings of the 22nd Australian Software Engineering Conference (ASWEC2013), Melbourne, Australia, 2013, pp. 97–106. doi: 10.1109/ASWEC.2013.21.

[103] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in 2014 Science and Information Conference, Aug. 2014, pp. 372–378. doi: 10.1109/SAI.2014.6918213.

[104] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, and Y. Zhou, "A Feature Subset Selection Algorithm Automatic Recommendation Method," CoRR, vol. abs/1402.0570, 2014.

[105] G. Lindner and R. Studer, "AST: Support for Algorithm Selection with a CBR Approach," in Principles of Data Mining and Knowledge Discovery, Berlin, Heidelberg, 1999, pp. 418–423.

[106] R. Engels and C. Theusinger, "Using a Data Metric for Preprocessing Advice for Data Mining Applications," in In Proceedings of the European Conference on Artificial Intelligence (ECAI-98), 1998, pp. 430–434.

[107] M. Reif, F. Shafait, and A. Dengel, "Meta 2-Features: Providing Meta-Learners More Information," 2012.

[108] R. Zakharov and P. Dupont, "Ensemble Logistic Regression for Feature Selection," in Pattern Recognition in Bioinformatics, 2011, pp. 133–144.

[109] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 13, no. 5, pp. 971–989, Sep. 2016, doi: 10.1109/TCBB.2015.2478454.

[110] P. Christen, "Concepts and techniques for record linkage, entity resolution, and duplicate detection," in Data Matching, 2012, 1st ed., p. 272. doi: 10.1007/978-3-642-31164-2.

[111] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," Inf. Fusion, vol. 50, pp. 71–91, 2019, doi: 10.1016/j.inffus.2018.09.012.

[112] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. T. Greenwood, and J. Beyene, "Data Integration in Genetics and Genomics: Methods and Challenges," Hum. Genomics Proteomics, vol. 1, no. 1, Jan. 2009, doi: 10.4061/2009/869093.

[113] A. Serra, M. Fratello, D. Greco, and R. Tagliaferri, "Data integration in genomics and systems biology," in 2016 IEEE Congress on Evolutionary Computation (CEC), 2016, pp. 1272–1279.

[114] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "A First Approach to Deal with Imbalance in Multi-label Datasets," in Hybrid Artificial Intelligent Systems, Berlin, Heidelberg, 2013, pp. 150–160. doi: 10.1007/978-3-642-40846-5_16.

[115] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms," in Hybrid Artificial Intelligence Systems, Cham, 2014, pp. 110–121. doi: 10.1007/978-3-319-07617-1_10.

[116] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "On the Impact of Dataset Complexity and Sampling Strategy in Multilabel Classifiers Performance," in Hybrid Artificial Intelligent Systems, Cham, 2016, pp. 500–511. doi: 10.1007/978-3-319-32034-2_42.

[117] H. Scheffé, The Analysis of Variance. Wiley, 1999.

[118] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "Resampling Multilabel Datasets by Decoupling Highly Imbalanced Labels," in Hybrid Artificial Intelligent Systems, Cham, 2015, pp. 489–501. doi: 10.1007/978-3-319-19644-2_41.

[119] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Dealing with difficult minority labels in imbalanced mutilabel data sets," Neurocomputing, vol. 326–327, pp. 39–53, 2019, doi: 10.1016/j.neucom.2016.08.158.

[120] M. Zhang and Z. Zhou, "A Review on Multi-Label Learning Algorithms," IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, 2014, doi: 10.1109/TKDE.2013.39.

[121] B. Inmon, Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump. Technics Publications, 2016.
[122] V. C. Storey and I.-Y. Song, "Big data technologies and Management: What conceptual modeling can do," Data Knowl. Eng., vol. 108, pp. 50–67, 2017, doi: 10.1016/j.datak.2017.01.001.
[123] J. P. Loyall et al., "QoS enabled dissemination of managed information objects in a publish-subscribe-query information broker," in Proceedings of Society of Photo-Optical Instrumentation Engineers - SPIE, 2009, vol. 7350, pp. 7350-7350–12. doi: 10.1117/12.818744.
[124] C. Peterson et al., "Analysing vulnerability: a multi-dimensional approach from Colombia's upper Cauca river basin," Policy Brief. Cali, Colombia, Climate and Development Knowledge Network (CDKN), 2012.
[125] FAO, "SAFA Indicators - Sustainability Assessment of Food and Agriculture Systems," Food and Agriculture Organization of the United Nations, Rome, Italy, \notype, 2013.
[126] MinTIC, "Datos Abiertos Colombia - Plataforma de datos abiertos del gobierno Colombiano," 2017. https://www.datos.gov.co/ (accessed Aug. 07, 2017).
[127] DANE, "ANDA - Archivo Nacional de Datos," 2006. https://sitios.dane.gov.co/visor-anda/ (accessed Jul. 07, 2017).
[128] OpenDataBarometer, "The Open Data Barometer," 2008. https://opendatabarometer.org (accessed Sep. 07, 2019).
[129] SIVICAP, "Características Calidad del Agua - Instituto Nacional de Salud," 2017. https://www.datos.gov.co/Salud-y-Protecci-n-Social/Caracter-sticas-Calidad-del-Agua-SIVICAP/jjzc-8w82 (accessed May 29, 2019).
[130] CORPOICA, "Resultados de Análisis de Laboratorio Suelos en Colombia," 2017. https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Resultados-de-An-lisis-de-Laboratorio-Suelos-en-Co/ch4u-f3i5 (accessed Apr. 18, 2019).
[131] IDEAM, "Solicitud de información - Red de estaciones climáticas IDEAM," 2014. http://www.ideam.gov.co/solicitud-de-informacion (accessed May 07, 2018).
[132] FINAGRO, "Fondo para el Financiamiento del Sector Agropecuario," 2013. https://www.finagro.com.co/quienes-somos/informacion-institucional (accessed Sep. 06, 2019).
[133] DANE, "Sistema de información de precios SIPSA," 2019. https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/sipsa (accessed Jul. 05, 2019).
[134] MinAgricultura, "Estadísticas Agropecuarias - Agronet," 2016. http://www.agronet.gov.co/estadistica/Paginas/default.aspx (accessed May 07, 2017).
[135] Minagricultura, "Ministerio de Agricultura y Desarrollo Rural," 2019. https://www.minagricultura.gov.co/paginas/default.aspx (accessed May 06, 2019).
[136] Minagricultura, "Precios internacionales de productos," 2019. https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=17 (accessed Jun. 06, 2019).
[137] DNP, "Datos y Estadisticas DNP," 2015. https://www.dnp.gov.co/programas/inversiones-y-finanzas-publicas/Datos-y-Estadisticas/Paginas/Datos-y-Estadisticas.aspx (accessed Aug. 07, 2017).
[138] DNP, "Infografía Financiamiento Climático en Colombia," 2017. https://mrvapp.dnp.gov.co/General/InfografiaGeneral/ (accessed Sep. 06, 2019).
[139] FINAGRO, "Certificado de Incentivos Forestales," 2013. https://www.finagro.com.co/productos-y-servicios/incentivo-forestal (accessed Sep. 06, 2019).
[140] DNP, "Dirección de Desarrollo Rural Sostenible," 2018. https://www.dnp.gov.co/DNP/organigrama/subdireccion-sectorial/Paginas/direccion-de-desarrollo-rural-sostenible.aspx (accessed Sep. 06, 2019).
[141] DNP, "Proyecto apoyo a alianzas productivas - PAAP," 2018. https://www.minagricultura.gov.co/tramites-servicios/desarrollo-rural/Paginas/v1/Proyecto-apoyo-a-alianzas-productivas-PAAP.aspx (accessed Sep. 06, 2019).
[142] DANE, "COLOMBIA - Tercer Censo Nacional Agropecuario - 2014 - 3er CNA," 2014. http://microdatos.dane.gov.co/index.php/catalog/513/data_dictionary#page=F11&tab=data-dictionary (accessed Oct. 06, 2019).
[143] DANE, "COLOMBIA - Tercer Censo Nacional Agropecuario - 2014 - 3er CNA," 2014. http://microdatos.dane.gov.co/index.php/catalog/513/datafile/F9 (accessed Oct. 06, 2019).
[144] D. C. Hoaglin, B. Iglewicz, and J. W. Tukey, "Performance of Some Resistant Rules for Outlier Labeling," J. Am. Stat. Assoc., vol. 81, no. 396, pp. 991–999, 1986.
[145] K. Dhana, "Data Management in R - Identify, describe, plot, and remove the outliers from the dataset," 2016. https://datascienceplus.com/identify-describe-plot-and-removing-the-outliers-from-the-dataset/ (accessed Sep. 02, 2019).
[146] J. L. Fleiss and others, "Measuring nominal scale agreement among many raters," Psychol. Bull., vol. 76, no. 5, pp. 378–382, 1971.
[147] M. A. Jaro, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," J. Am. Stat. Assoc., vol. 84, no. 406, pp. 414–420, 1989.
[148] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," Neurocomputing, vol. 163, pp. 3–16, 2015, doi: 10.1016/j.neucom.2014.08.091.
[149] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)†," Biometrika, vol. 52, no. 3–4, pp. 591–611, 1965, doi: 10.1093/biomet/52.3-4.591.
[150] J. L. Gastwirth, Y. R. Gel, and W. Miao, "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice," Stat. Sci., vol. 24, no. 3, pp. 343–360, 2009.

[151] J. F. Box, “Guinness, Gosset, Fisher, and Small Samples,” Stat. Sci., vol. 2, no. 1, pp. 45–52, 1987, doi: 10.1214/ss/1177013437.

[152] J. W. Tukey, “Comparing Individual Means in the Analysis of Variance,” Biometrics, vol. 5, no. 2, pp. 99–114, 1949.

[153] W. Webber, A. Moffat, and J. Zobel, “A Similarity Measure for Indefinite Rankings,” ACM Trans. Inf. Syst., vol. 28, no. 4, Nov. 2010, doi: 10.1145/1852102.1852106.

[154] H. Fang, “Managing data lakes in big data era: What’s a data lake and why has it became popular in data management ecosystem,” 2015, pp. 820–824. doi: 10.1109/CYBER.2015.7288049.

[155] V. N. Kumar and P. Shindgikar, Modern Big Data Processing with Hadoop: Expert Techniques for Architecting End-To-end Big Data Solutions to Get Valuable Insights. Packt Publishing, 2018.

[156] R. Hai, S. Geisler, and C. Quix, “Constance: An intelligent data lake system,” 2016, vol. 26-June-2016, pp. 2097–2100. doi: 10.1145/2882903.2899389.

[157] O. Mendelevitch, C. Stella, and D. Eadline, Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale, 1st ed. Addison-Wesley Professional, 2016.

[158] D. Chamberlin, “XQuery: A Query Language for XML,” in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2003, p. 682.

[159] J. Melton and S. Buxton, Querying XML: XQuery, XPath, and SQL/XML in context. Morgan Kaufmann, 2011.

[160] D. Delen and H. Demirkan, “Data, information and analytics as services,” Decis. Support Syst., vol. 55, no. 1, pp. 359–363, 2013, doi: 10.1016/j.dss.2012.05.044.

[161] C. Madera and A. Laurent, “The next information architecture evolution: The data lake wave,” 2016, pp. 174–180. doi: 10.1145/3012071.3012077.

[162] K. S. Rubin and A. Goldberg, “Object behavior analysis,” Commun. ACM, vol. 35, no. 9, pp. 48–62, Sep. 1992, doi: 10.1145/130994.130996.

[163] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, “Big Data in Smart Farming – A review,” Agric. Syst., vol. 153, pp. 69–80, May 2017, doi: 10.1016/j.agsy.2017.01.023.

[164] J. L. Hatfield et al., “Indicators of climate change in agricultural systems,” Clim. Change, Jun. 2018, doi: 10.1007/s10584-018-2222-2.

[165] L. Bizikova, P. Larkin, S. Mitchell, and R. Waldick, “An indicator set to track resilience to climate change in agriculture: A policy-maker’s perspective,” Land Use Policy, vol. 82, pp. 444–456, Mar. 2019, doi: 10.1016/j.landusepol.2018.11.057.

# Appendices

# Appendix A

# Literature Review Report

This appendix presents the products resulting from the mapping and systematic literature review processes. Initially, we show the main analyses of the systematic mapping around the works related to Data Fusion and Multi-Label Classification. Finally, we present the respective systematic review reports for these two topics.

## A.1 Systematic Mapping around Data Fusion

| Description | Results |
|---|---|
| MAIN INFORMATION ABOUT DATA | |
| Timespan | 2011:2020 |
| Sources (Journals, Books, etc) | 126 |
| Documents | 275 |
| Average years from publication | 5.13 |
| Average citations per documents | 6.011 |
| Average citations per year per doc | 0.8941 |
| References | 6399 |
| DOCUMENT TYPES | |
| article | 72 |
| book chapter | 6 |
| conference paper | 196 |
| review | 1 |
| DOCUMENT CONTENTS | |
| Keywords Plus (ID) | 2416 |
| Author's Keywords (DE) | 874 |
| AUTHORS | |
| Authors | 1057 |
| Author Appearances | 1209 |
| Authors of single-authored documents | 10 |
| Authors of multi-authored documents | 1047 |
| AUTHORS COLLABORATION | |
| Single-authored documents | 11 |
| Documents per Author | 0.26 |
| Authors per Document | 3.84 |
| Co-Authors per Documents | 4.4 |
| Collaboration Index | 3.97 |

**Table A. 1.** Main Information about the collection.

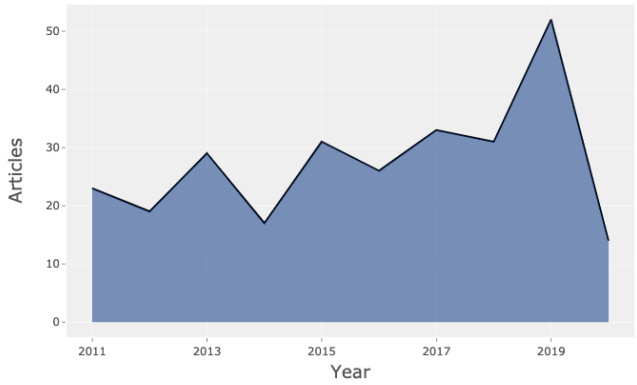



**Figure A. 1.** Annual Scientific Production (Annual Growth Rate: -5.37%).

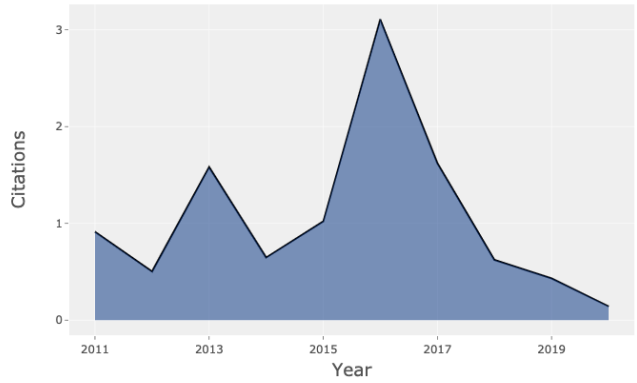



**Figure A. 2.** Average Citations per Year.

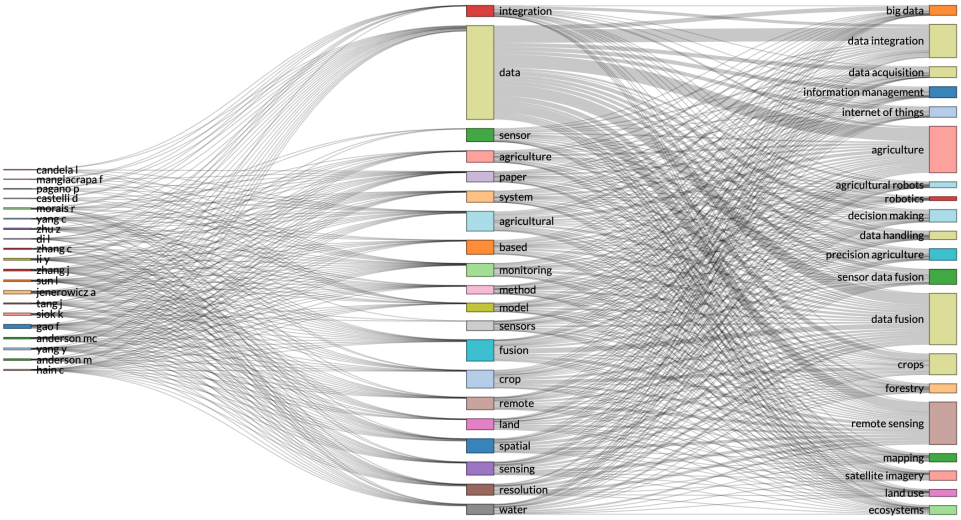



**Figure A. 3.** Three-Fields Plot (Author, Abstract, and Keywords).

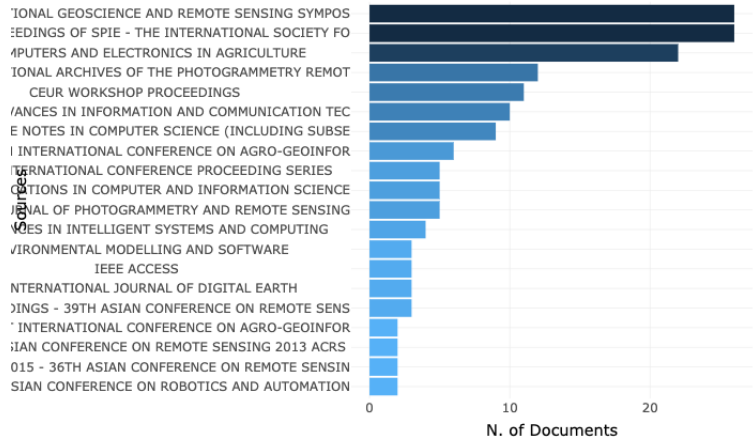



**Figure A. 4.** Most Relevant Sources.

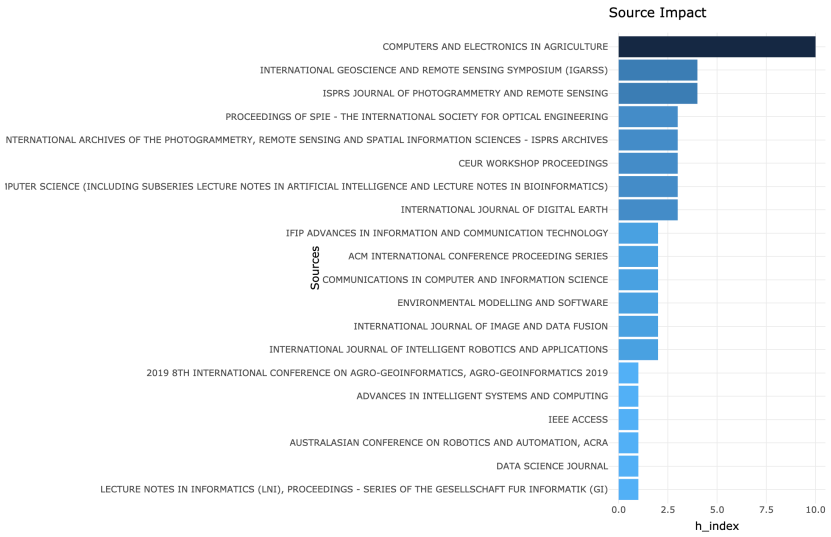



**Figure A. 5.** Source Impact.

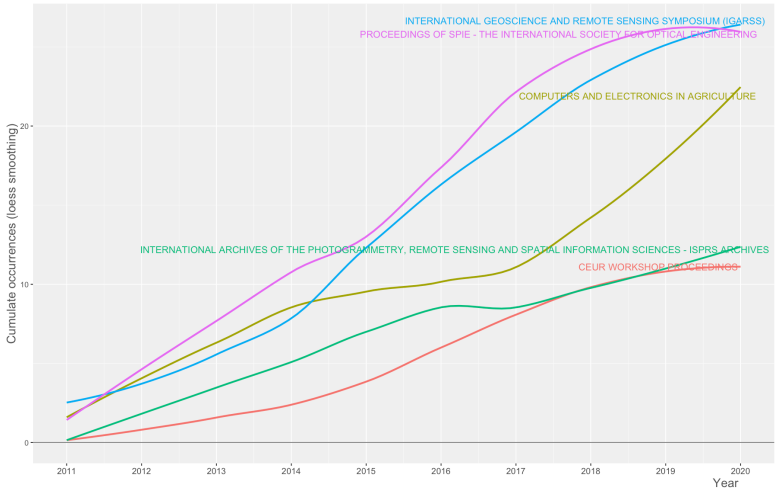



**Figure A. 6.** Source Dynamics (Cumulate).

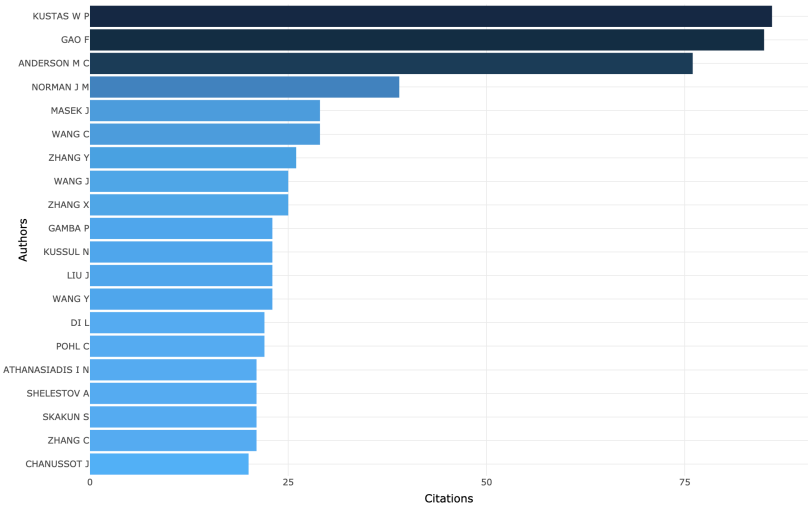



**Figure A. 7.** Most Local Cited Authors.

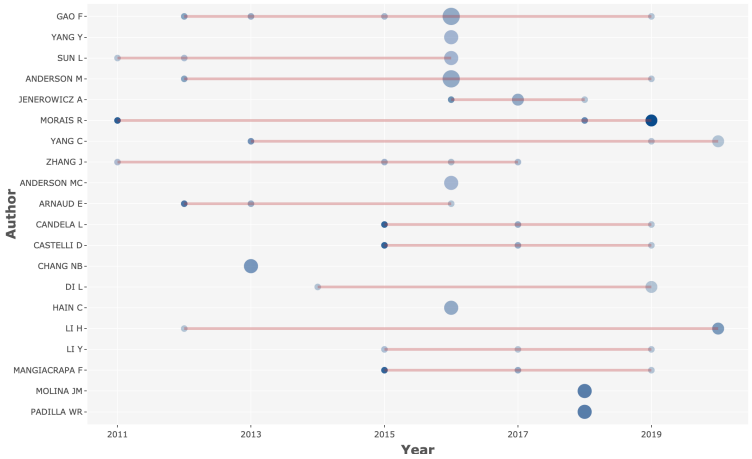



**Figure A. 8.** Authors' Production over Time.

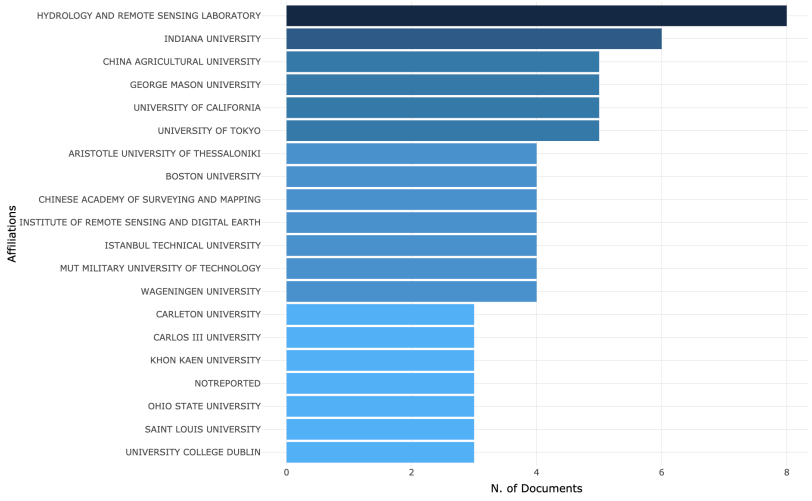



**Figure A. 9.** Most Relevant Affiliations.

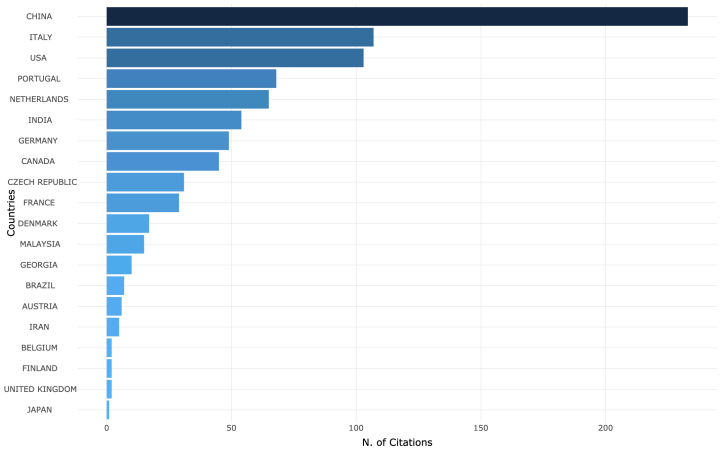



**Figure A. 10.** Most Cited Countries.

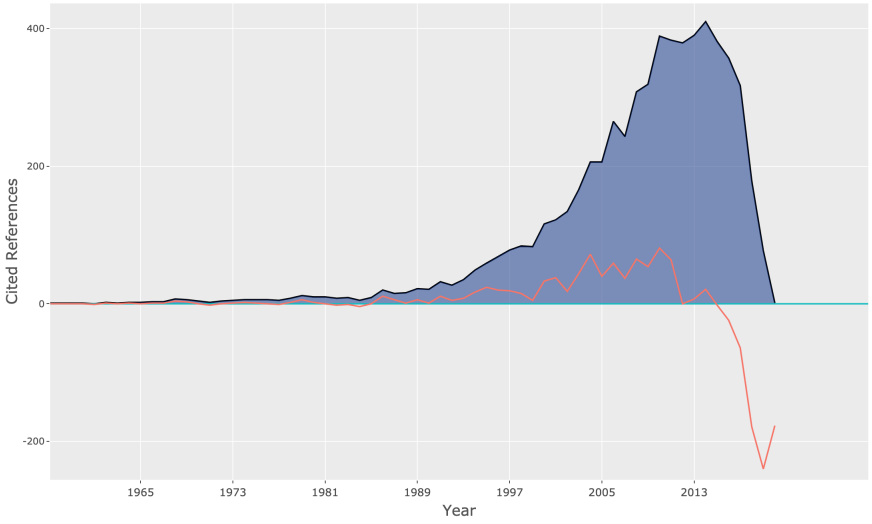



**Figure A. 11.** Reference Publication Year Spectroscopy.

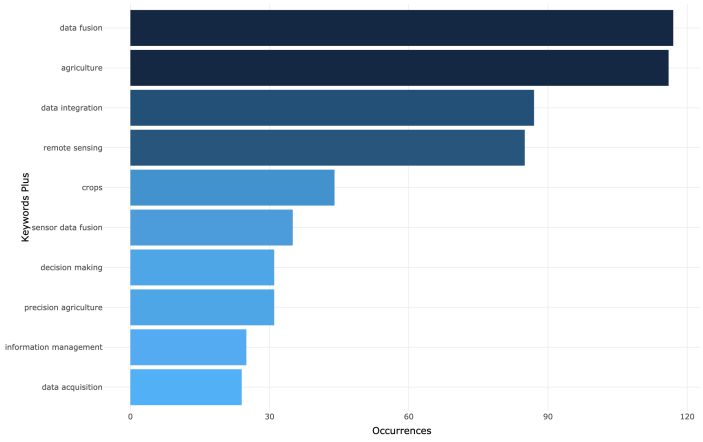



**Figure A. 12.** Most Frequent Words.

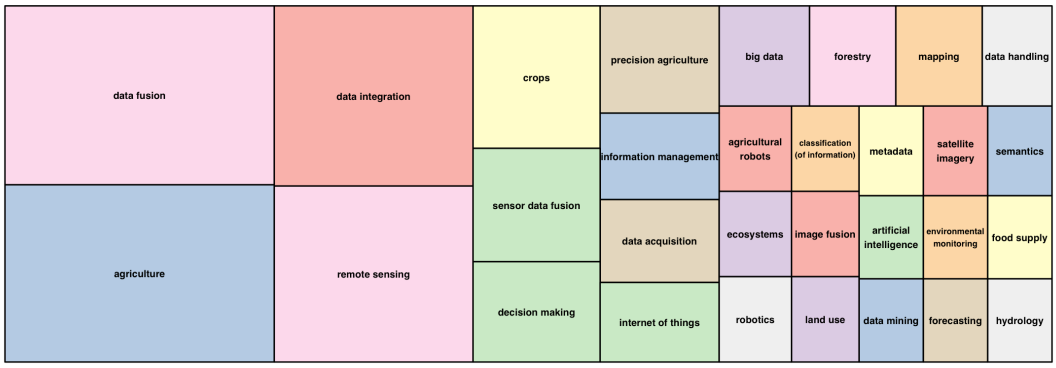



**Figure A. 13.** Word Tree Map.

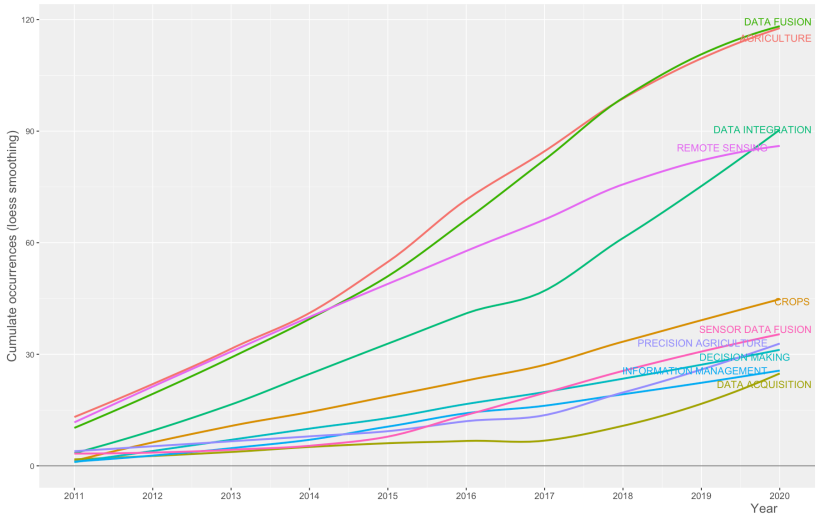



**Figure A. 14.** Word Dynamics (Cumulate).

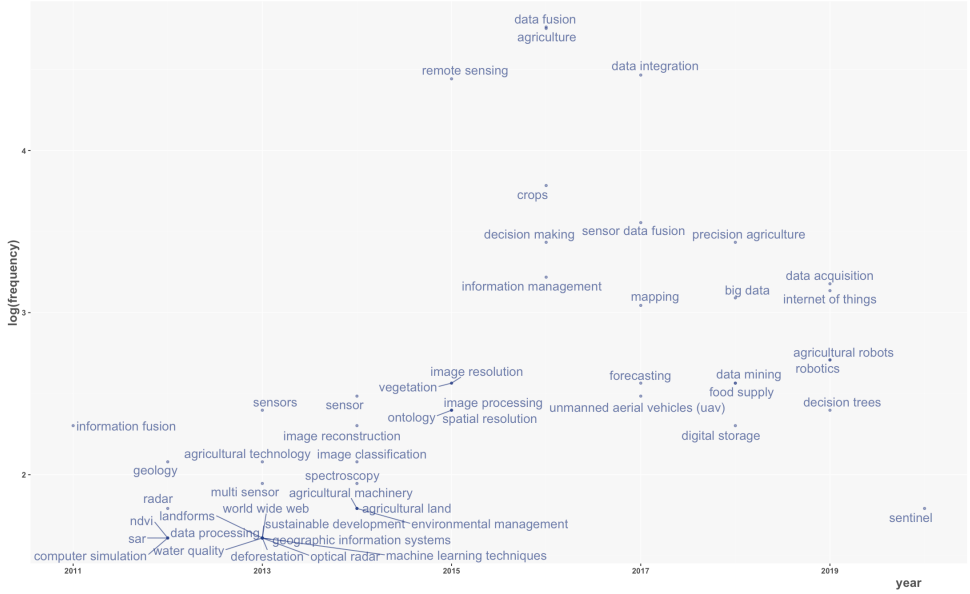



**Figure A. 15.** Trend Topics

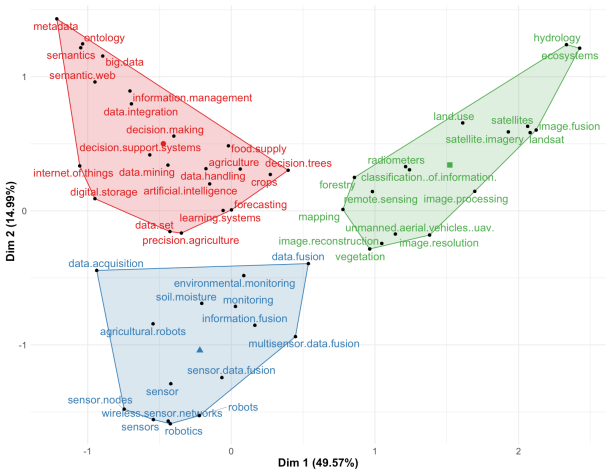



**Figure A. 16.** Factorial Analysis (Conceptual Structured Map, method MCA).

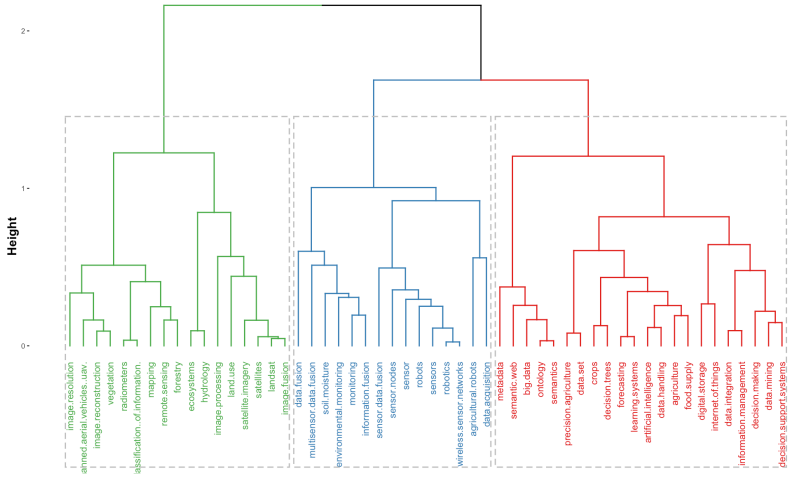



**Figure A. 17.** Topic Dendrogram.

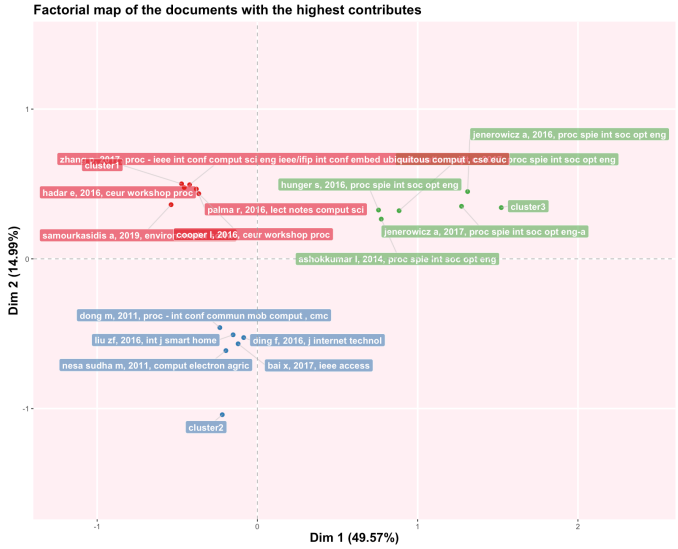



**Figure A. 18.** Factorial map of the documents with the highest contributes.

## A.2 Systematic Mapping around Multi-Label Classification

| Description | Results |
|---|---|
| MAIN INFORMATION ABOUT DATA | |
| Timespan | 2011:2020 |
| Sources (Journals, Books, etc) | 688 |
| Documents | 1953 |
| Average years from publication | 4.43 |
| Average citations per documents | 10.26 |
| Average citations per year per doc | 1.545 |
| References | 50046 |
| DOCUMENT TYPES | |
| article | 663 |
| article in press | 4 |
| book | 1 |
| book chapter | 17 |
| conference paper | 1264 |
| review | 4 |
| DOCUMENT CONTENTS | |
| Keywords Plus (ID) | 7045 |
| Author's Keywords (DE) | 3348 |
| AUTHORS | |
| Authors | 3903 |
| Author Appearances | 6855 |
| Authors of single-authored documents | 43 |
| Authors of multi-authored documents | 3860 |
| AUTHORS COLLABORATION | |
| Single-authored documents | 61 |
| Documents per Author | 0.5 |
| Authors per Document | 2 |
| Co-Authors per Documents | 3.51 |
| Collaboration Index | 2.04 |

**Table A. 2.** Main Information about the collection.

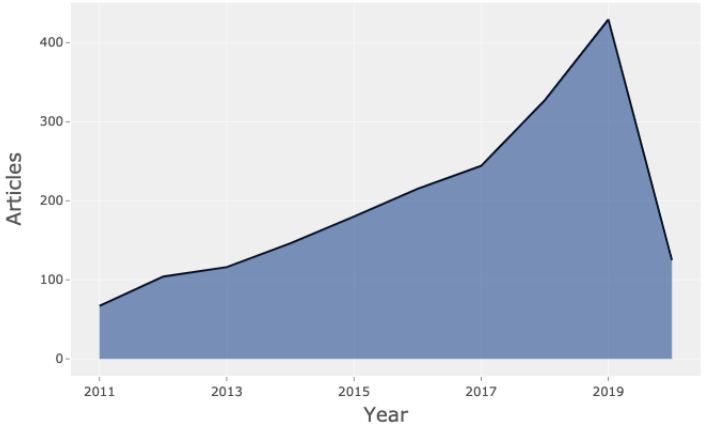



**Figure A. 19.** Annual Scientific Production (Annual Growth Rate: 7.17%).

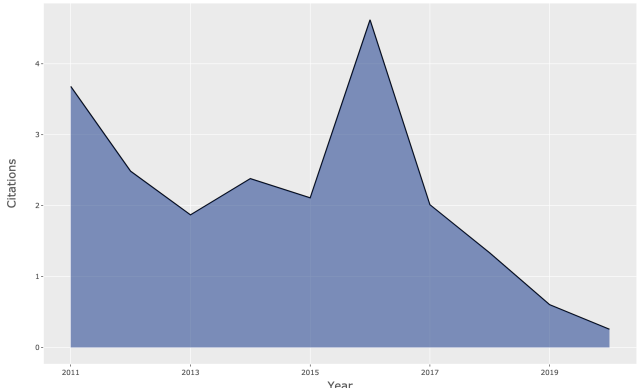



**Figure A. 20.** Average Citations per Year.

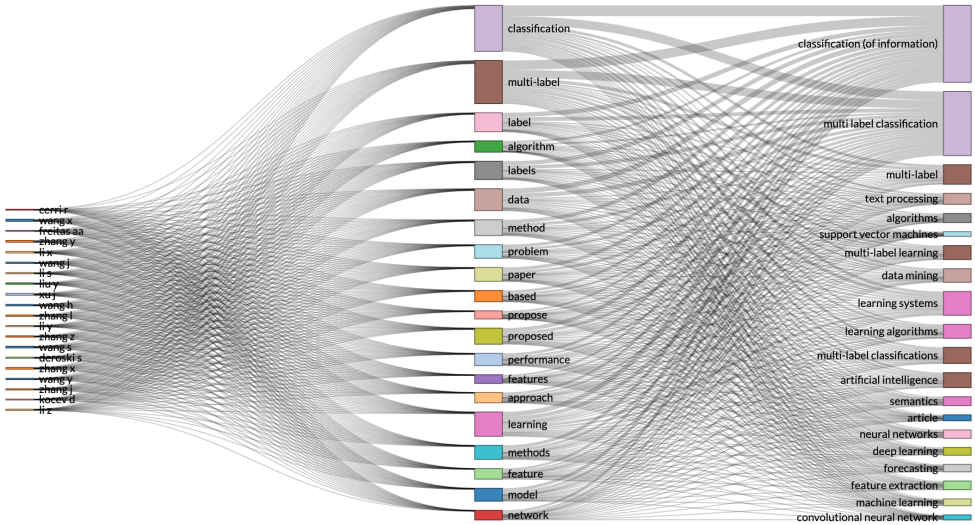



**Figure A. 21.** Three-Fields Plot (Author, Abstract, and Keywords).

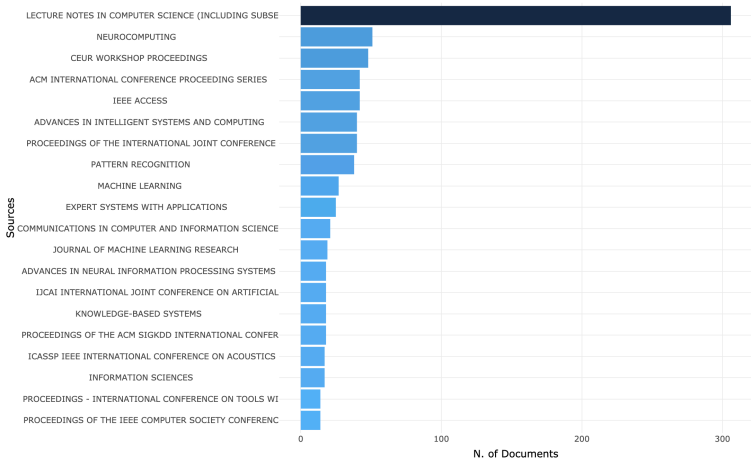



**Figure A. 22.** Most Relevant Sources.

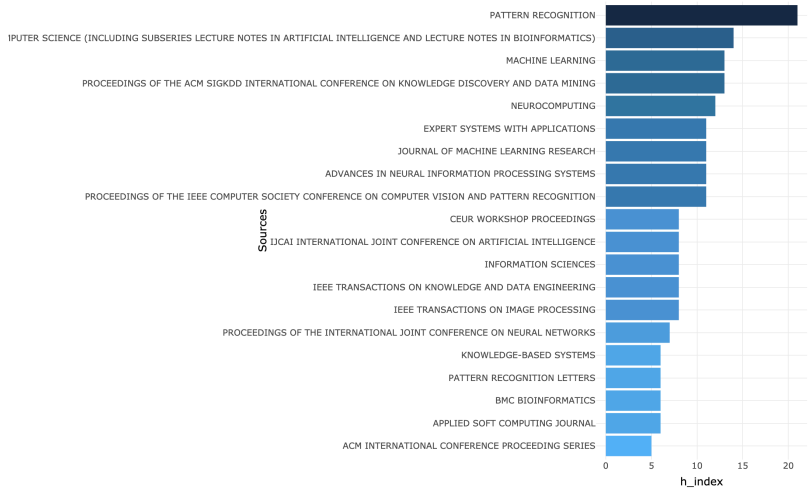



**Figure A. 23.** Source Impact.

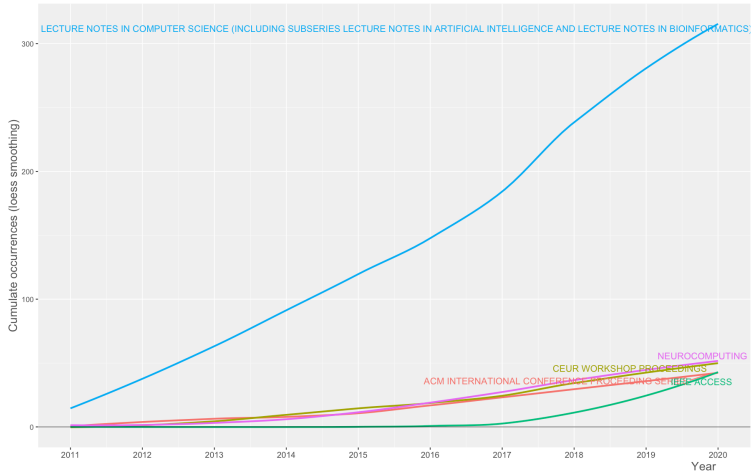



**Figure A. 24.** Source Dynamics (Cumulate).

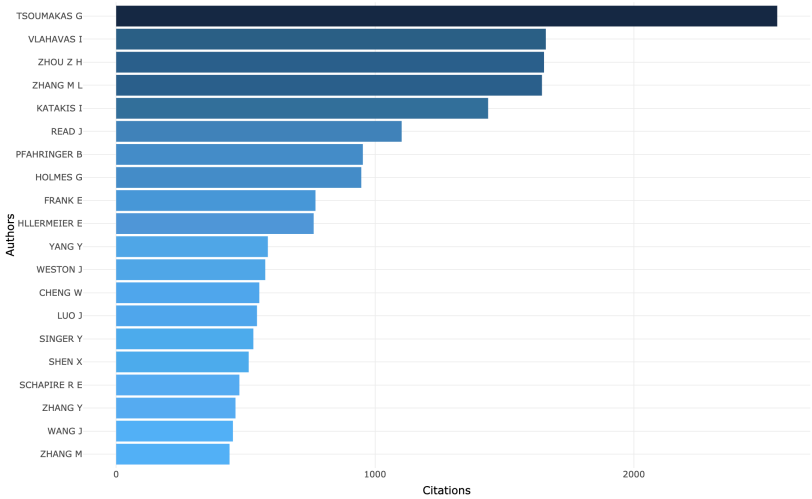



**Figure A. 25.** Most Local Cited Authors.

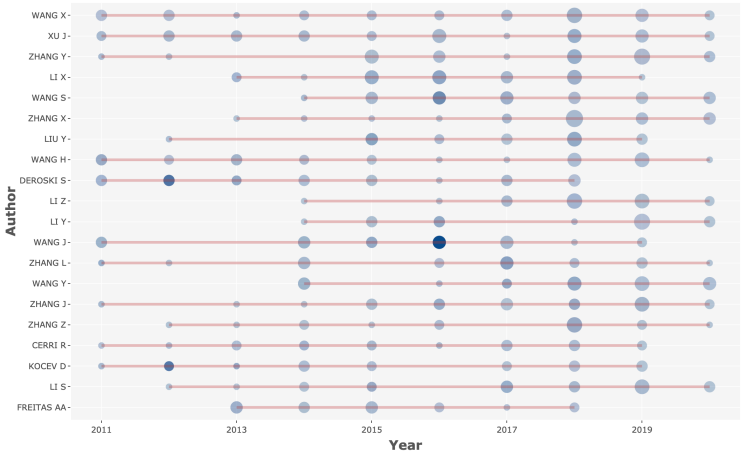



**Figure A. 26.** Authors' Production over Time.

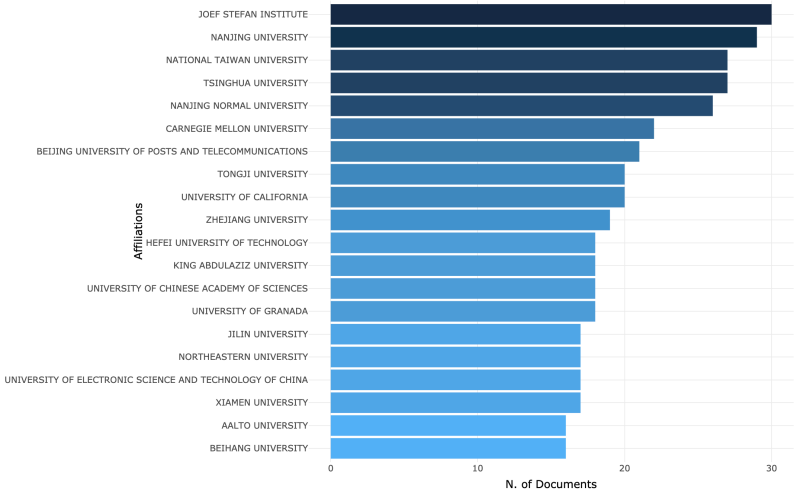



**Figure A. 27.** Most Relevant Affiliations.

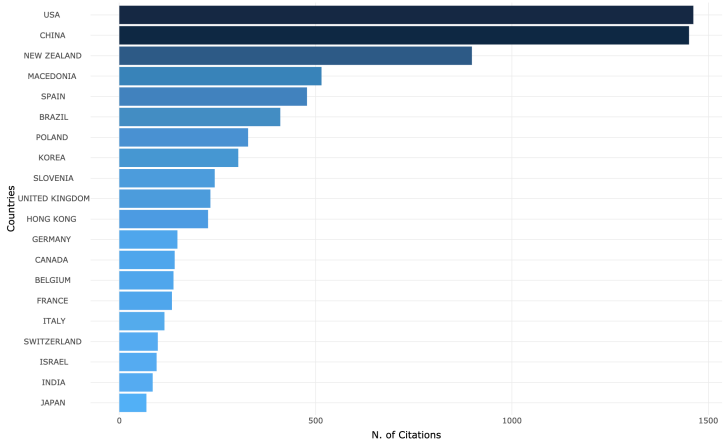



**Figure A. 28.** Most Cited Countries.

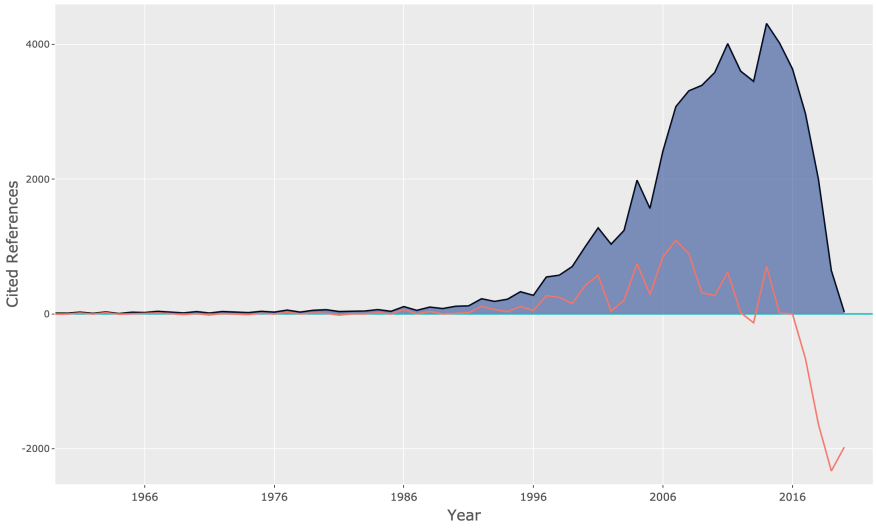



**Figure A. 29.** Reference Publication Year Spectroscopy.

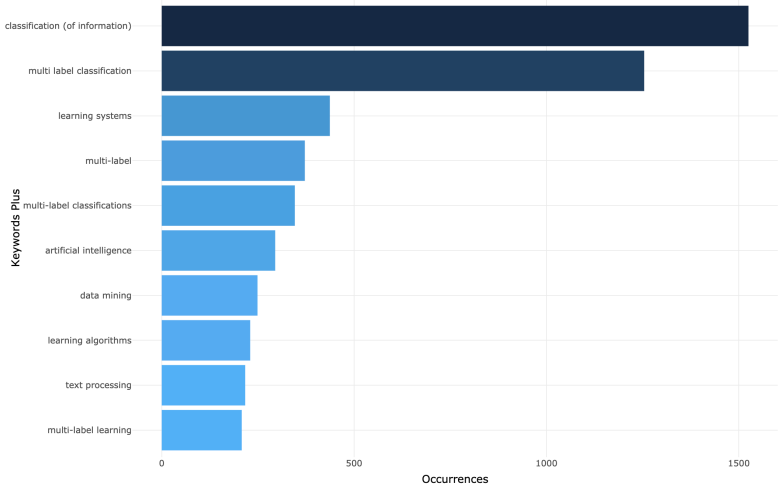



**Figure A. 30.** Most Frequent Words.

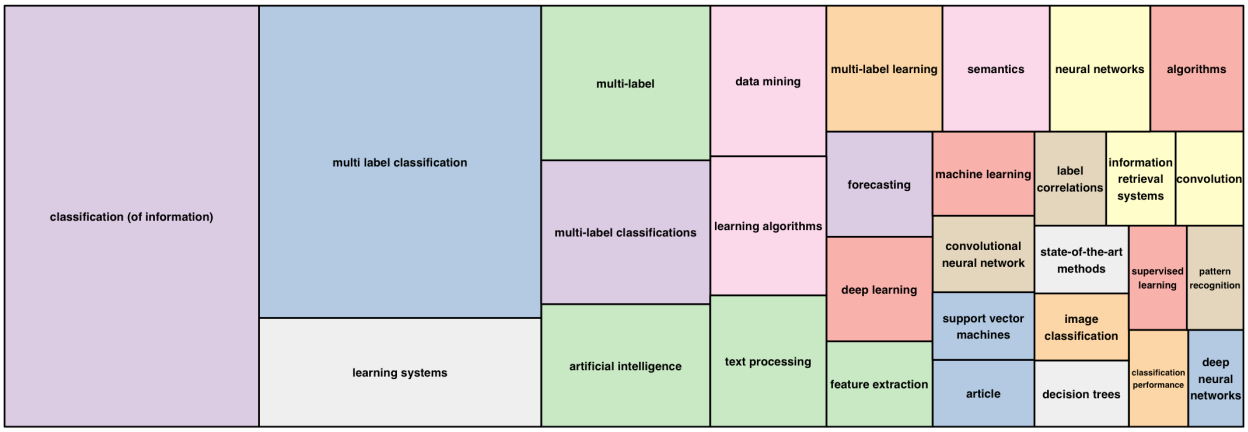



**Figure A. 31.** Word Tree Map.

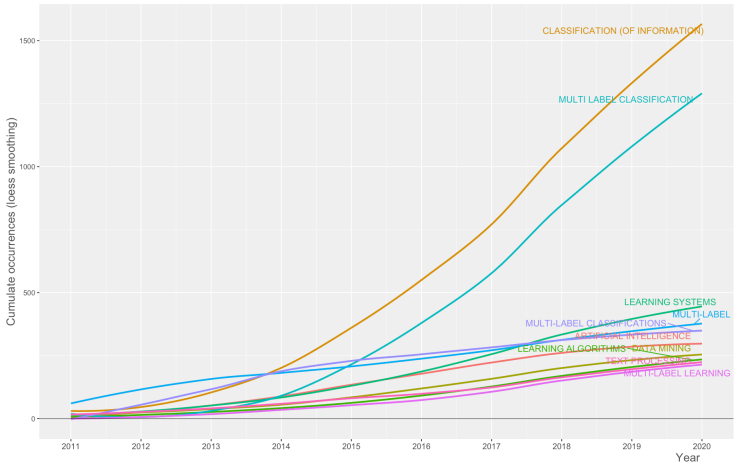



**Figure A. 32.** Word Dynamics (Cumulate).

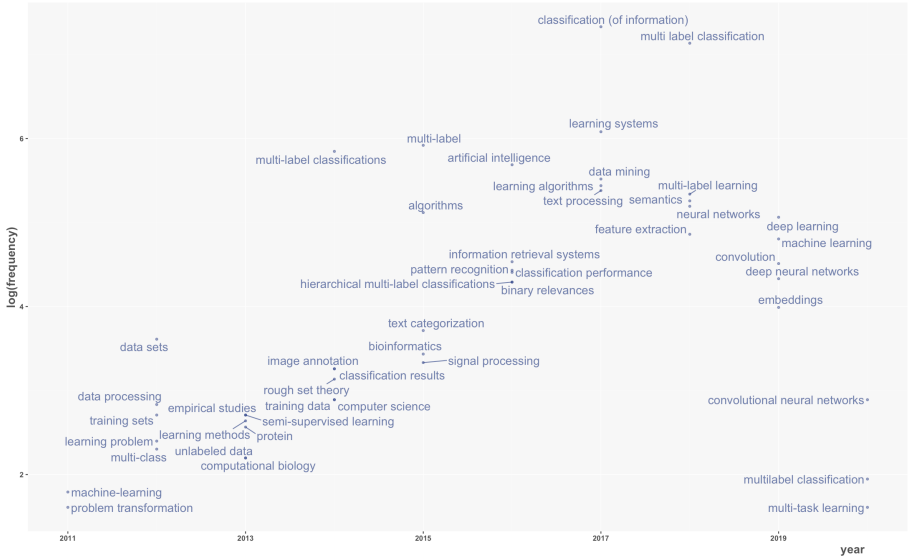



**Figure A. 33.** Trend Topics


**Figure A. 34.** Factorial Analysis (Conceptual Structured Map, method MCA).

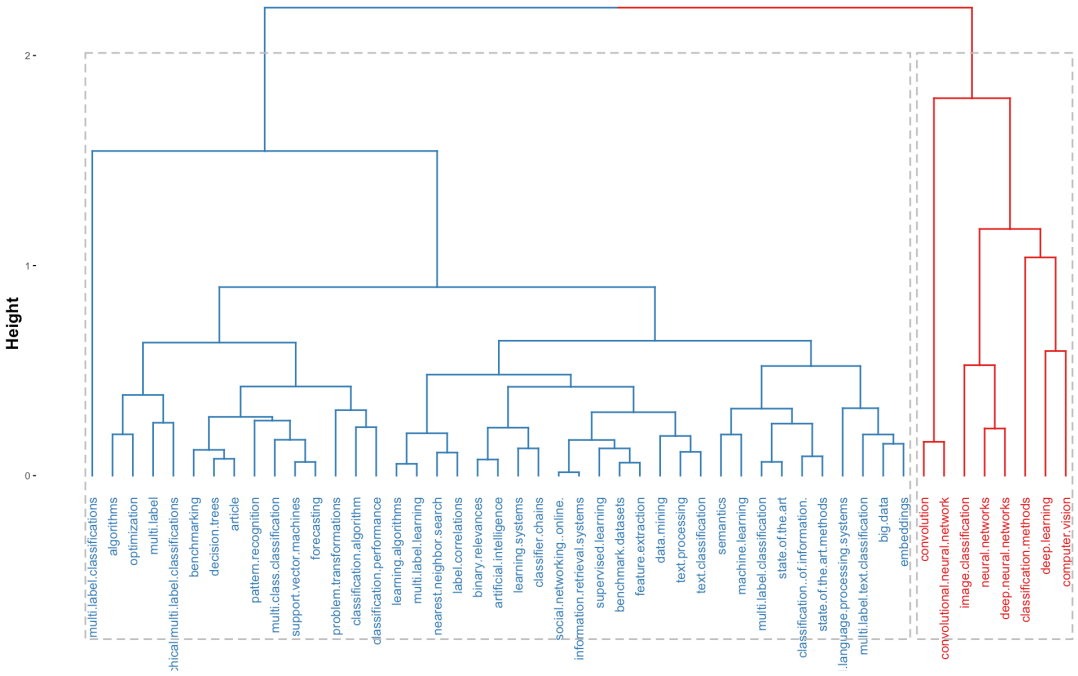



**Figure A. 35.** Topic Dendrogram.

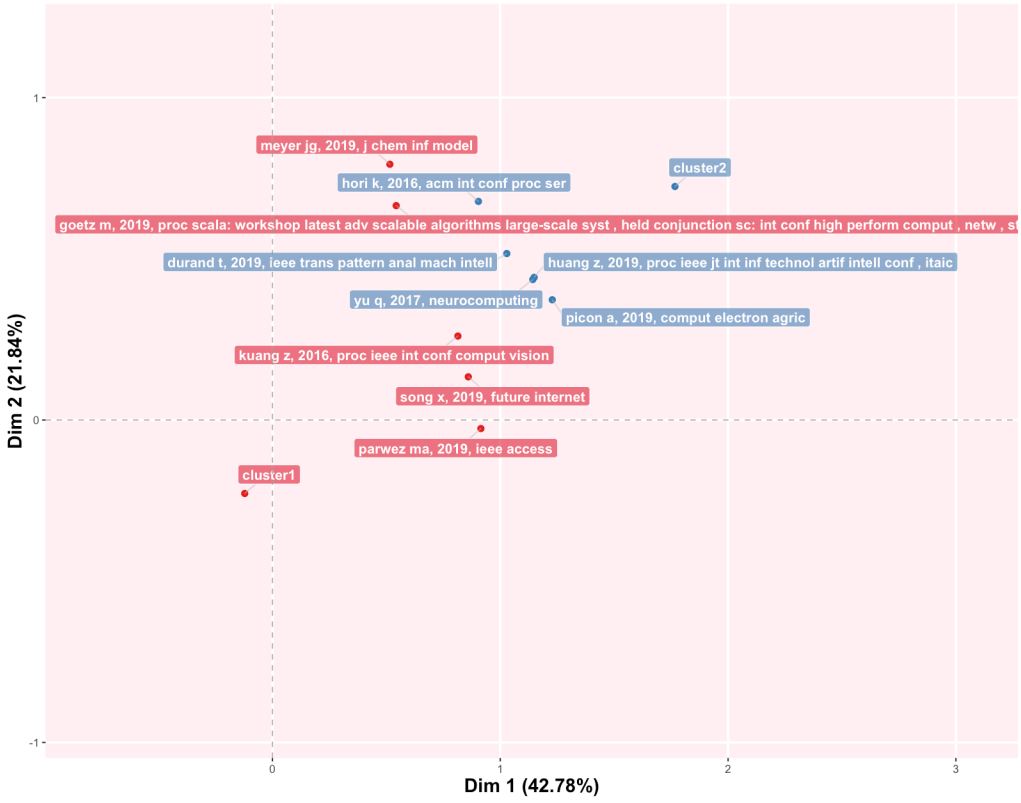



**Figure A. 36.** Factorial map of the documents with the highest contributes.

## A.3 Systematic Review Report around Data Fusion

Techniques for integrating data in agricultural contexts and climate vulnerability assessments.

***Planning***

To describe approaches of data fusion applied to agricultural contexts, specifically in climate vulnerability assessments.

**PICOC**

- Population: Official open data, restricted data with access permission, results of agricultural vulnerability analysis
- Intervention: Data Fusion, Data Integration
- Comparison: A comparison of methods, algorithms, data fusion models, datasets
- Outcome: A systematic literature survey report including synthesis of most relevant articles published on data fusion for crop prediction
- Context: A systematic investigation to consolidate a peer-reviewed and academic research, classification and comparison, trends and future research directions

**Research Questions**

RQ.1. What kinds of Data Fusion (DF) approaches have been applied in Climate Vulnerability Assessments (CVAs)?

RQ.2. What kinds of Data Fusion (DF) approaches have been applied in other agricultural contexts?

**Keywords and Synonyms**

| Keyword | Synonyms |
|---|---|
| climate vulnerability assessments | agricultural vulnerability, agricultural vulnerability analysis, CVA, methodologies for agricultural vulnerability |
| data fusion | data integration |
| open data | data, dataset |

**Search String**

("climate vulnerability assessments" OR "agricultural vulnerability" OR "agricultural vulnerability analysis" OR "CVA" OR "methodologies for agricultural vulnerability" OR "open data" OR "data" OR "dataset") AND ("data fusion" OR "data integration")

**Sources**

- Google Scholar (https://scholar.google.com/)
- IEEE Digital Library (http://ieeexplore.ieee.org)
- Science@Direct (http://www.sciencedirect.com)
- Scopus (http://www.scopus.com)
- Springer Link (http://link.springer.com)

**Selection Criteria**

Inclusion Criteria:

- Papers about DF applied in CVAs
- Papers about DF applied to agricultural contexts
- Papers about DF applied to crop production
- Papers whose objectives are similar to research, but but applying other approaches

Exclusion Criteria:

- It is not an article
- Papers about DF applied to domains other than agriculture
- Papers not published in the last ten years
- Papers outside the area of computer science
- Papers that not included the keywords
- Techniques other than DF applied to agriculture

**Quality Assessment Checklist**

Questions:

- Is the paper based on research or is it merely a report based on expert opinion?
- Is there a clear statement of the aims of the research?
- Is there a description of the context in which the research was carried out?
- Was the data collected in a way that addressed the research issue?
- Was the data analysis sufficiently rigorous?
- Is there a clear statement of findings?
- Is the study of value for research or practice?

Answers:

- Yes
- Partly
- No

**Data Extraction Form**

- Has any DF technique been applied in agricultural contexts?

- If the previous answer is yes, any DF technique been applied in CVAs?
- Is there any methodology or method to apply DF?
- Is there any methodology to evaluate the DF performance?
- Is any experiment carried out?
- If the previous answer is yes, choose the evaluation method used

***Conducting***

**Digital Libraries Search Strings**

Google Scholar:
("data fusion" OR "data integration") AND ("crop production")

IEEE Digital Library:
(("All Metadata": "data fusion" OR "data integration") AND "All Metadata": "crop production")

Science@Direct:
("data fusion" OR "data integration") AND ("crop vulnerability" OR "crop production")

Scopus:
( TITLE-ABS-KEY ( "data fusion" OR "data integration" ) AND TITLE-ABS-KEY ( "crop production" ) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) OR LIMIT-TO ( PUBYEAR , 2014 ) OR LIMIT-TO ( PUBYEAR , 2013 ) OR LIMIT-TO ( PUBYEAR , 2012 ) OR LIMIT-TO ( PUBYEAR , 2011 ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" ) OR LIMIT-TO ( DOCTYPE , "re" ) ) AND ( LIMIT-TO ( SUBJAREA , "AGRI" ) OR LIMIT-TO ( SUBJAREA , "COMP" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )

Springer Link:
("data fusion" OR "data integration") AND ("crop vulnerability" OR "crop production")

**Imported Studies**

- Google Scholar: 6
- IEEE Digital Library: 1
- Science@Direct: 74

- Scopus: 23
- Springer Link: 27

## Related Works

| Title | Author | Journal | Year | Status |
|---|---|---|---|---|
| A generic ontological network for Agri-food experiment integration – Application to viticulture and winemaking | Aunur Rofiq Muljarto and Jean-Michel Salmon and Brigitte Charnomordic and Patrice Buche and Anne Tireau and Pascal Neveu | Computers and Electronics in Agriculture | 2017 | Accepted |
| Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean | Chunhua Liao and Jinfei Wang and Taifeng Dong and Jiali Shang and Jiangui Liu and Yang Song | Science of The Total Environment | 2019 | Accepted |
| Sensor data fusion for soil health assessment | Veum, Kristen S and Sudduth, Kenneth A and Kremer, Robert J and Kitchen, Newell R | Geoderma | 2017 | Accepted |
| Progresses on data fusion technology of crop growth model and multi-source observation information | Jing, Wang and Xin, Li | Remote Sensing Technology and Application | 2015 | Accepted |
| Plant localization and discrimination using 2D+3D computer vision for robotic intra-row weed control | Gai, J. and Tang, L. and Steward, B. | 2016 American Society of Agricultural and Biological Engineers Annual International Meeting, ASABE 2016 | 2016 | Accepted |
| PAID: Predictive agriculture analysis of data integration in India | Grover, Purva and Johari, Rahul | | 2016 | Accepted |
| Scalable pixel-based crop classification combining Sentinel-2 and Landsat-8 data time series: Case study of the Duero river basin | Laura Piedelobo and David Hernández-López and Rocío Ballesteros and Amal Chakhar and Susana [Del Pozo] and Diego González-Aguilera and Miguel A. Moreno | Agricultural Systems | 2019 | Accepted |
| Multisensor Fusion of Remote Sensing Data for Crop Disease Detection | Moshou, Dimitrios and Gravalos, Ioannis and Bravo, Dimitrios Kateris Cedric and Oberti, Roberto and West, Jon S. and Ramon, Herman | | 2011 | Accepted |
| Multiple on-line soil sensors and data fusion approach for delineation of water holding capacity zones for site specific irrigation | Mouazen, A.M. and Alhwaimel, S.A. and Kuang, B. and Waine, T. | Soil and Tillage Research | 2014 | Accepted |
| Monitoring oil palm plantations in Malaysia | Pohl, C. and Kanniah, K. D. and Loong, C. K. | | 2016 | Accepted |
| Integration of in situ measured soil status and remotely sensed hyperspectral data to improve plant production system monitoring: Concept, perspectives and limitations | Laurent Tits and Ben Somers and Jan Stuckens and Jamshid Farifteh and Pol Coppin | Remote Sensing of Environment | 2013 | Accepted |
| Improved maize cultivated area estimation over a large scale combining MODIS-EVI time series data and crop phenological information | Zhang, J. and Feng, L. and Yao, F. | ISPRS Journal of Photogrammetry and Remote Sensing | 2014 | Accepted |
| Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation | Meng, J. and Du, X. and Wu, B. | International Journal of Digital Earth | 2013 | Accepted |
| Field partition by proximal and remote sensing data fusion | De Benedetto, Daniela and Castrignano, Annamaria and Diacono, Mariangela and Rinaldi, Michele and Ruggieri, Sergio and Tamborrino, Rosanna | Biosystems engineering | 2013 | Accepted |
| Data integration for climate vulnerability mapping in West Africa | De Sherbinin, A. and Chai-Onn, T. and Jaiteh, M. and Mara, V. and Pistolesi, L. and Schnarr, E. and Trzaska, S. | ISPRS International Journal of Geo-Information | 2015 | Accepted |
| Assessment of the effectiveness of spatiotemporal fusion of multi-source satellite images for cotton yield estimation | Linghua Meng and Huanjun Liu and Xinle Zhang and Chunying Ren and Susan Ustin and Zhengchao Qiu and Mengyuan Xu and Dong Guo | Computers and Electronics in Agriculture | 2019 | Accepted |
| A smart farming approach in automatic detection of favorable conditions for planting and crop production in the upper basin of Cauca River | López, I.D. and Corrales, J.C. | Advances in Intelligent Systems and Computing | 2018 | Accepted |
| AgriFuture: A New Theory of Change Approach to Building Climate-Resilient Agriculture | Mousannif, Hajar and Zahir, Jihad | | 2019 | Accepted |
| Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study | Ji, W. and Adamchuk, V.I. and Chen, S. and Mat Su, A.S. and Ismail, A. and Gan, Q. and Shi, Z. and Biswas, A. | Geoderma | 2019 | Accepted |
| Spatial management strategies for nitrogen in | Eleonora Cordero and Louis | Science of The Total Environment | 2019 | Rejected |

| | | | | |
|---|---|---|---|---|
| maize production based on soil and crop data | Longchamps and Raj Khosla and Dario Sacco | | | |
| Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia | Puyu Feng and Bin Wang and De Li Liu and Qiang Yu | Agricultural Systems | 2019 | Rejected |
| Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach | Michael L. Mann and James M. Warner | Field Crops Research | 2017 | Accepted |
| Extent of vulnerability in wheat producing agro-ecologies of India: Tracking from indicators of cross-section and multi-dimension data | Sendhil, R. and Jha, A. and Kumar, A. and Singh, S. | Ecological Indicators | 2018 | Rejected |
| Data Fusion of Proximal Soil Sensing and Remote Crop Sensing for the Delineation of Management Zones in Arable Crop Precision Farming. | Pantazi, Xanthoula Eirini and Moshou, Dimitrios and Mouazen, Abdul Mounem and Alexandridis, Thomas and Kuang, Boyan | | 2015 | Accepted |
| Inverse Problems and Data Fusion for Crop Production Applications Targeting Optimal Growth - Fertilization | B. Kaur and R. K. A. Owusu | | 2015 | Accepted |
| Brief Probe into the Key Factors that Influence Beijing Agricultural Drought Vulnerability | Huang, Lingmiao and Yang, Peiling and Ren, Shumei | | 2014 | Rejected |
| Understanding an Ontology through Divergent Exploration | Kozaki, Kouji and Hirota, Takeru and Mizoguchi, Riichiro | | 2011 | Rejected |
| Ontology for Seamless Integration of Agricultural Data and Models | Athanasiadis, Ioannis N. and Rizzoli, Andrea-Emilio and Janssen, Sander and Andersen, Erling and Villa, Ferdinando | | 2009 | Rejected |
| A Tool for Classification of Cacao Production in Colombia Based on Multiple Classifier Systems | Plazas, Julián Eduardo and López, Iván Darío and Corrales, Juan Carlos | | 2017 | Rejected |
| Database Construction of Soybean Biology | Jiang, Gui-fang and Zhong, Nan | | 2011 | Rejected |
| Explanatory Analysis of Data from a Distributed Solar Collector Field | Berenguel, Manuel and Klempous, Ryszard and Maciejewski, Henryk and Nikodem, Jan and Nikodem, Maciej and Valenzuela, Loreto | | 2005 | Rejected |
| A Review and Impact of Data Mining and Image Processing Techniques for Aerial Plant Pathology | Pudumalar, S. and Muthuramalingam, S. and Shanmugapriyan, R. | | 2020 | Rejected |
| An Agricultural Habitat Information Acquisition and Remote Intelligent Decision System Based on the Internet of Things | Hu, Ze Lin and Gao, Yi and Li, Miao and Li, Hua Long and Yang, Xuan Jiang and Ma, Zhi Run | | 2019 | Rejected |
| The Impact of Climate Change on the Potential Suitable Distribution of Major Crops in Zambia and the Countermeasures | Wang, Yanqin and Tan, Zhen and Sun, Guojun | | 2015 | Rejected |
| Application of Information Technology on Traceability System for Agro-Food Quality and Safety | Xia, Xue and Qiu, Yun and Hu, Lin and Zhou, Guomin | | 2015 | Rejected |
| Internet of Things-Based Hardware and Software for Smart Agriculture: A Review | Sharma, Brij Bhushan and Kumar, Nagesh | | 2020 | Rejected |
| Simulation in Production and Logistics: Trends, Solutions and Applications | Wenzel, Sigrid and Boyaci, Pinar and Jessen, Ulrich | | 2010 | Rejected |
| Agricultural Intensification, Population Growth and Forest Cover Change: Evidence from Spatially Explicit Land Use Modeling in the Central Highlands of Vietnam | Müller, Daniel and Zeller, Manfred | | 2004 | Rejected |
| Representing Ecological Niches in a Conceptual Model | Semwayo, Daniel T. and Berman, Sonia | | 2004 | Rejected |
| Prediction Based Agro Advisory System for Crop Protection | Mangala, R. Ruba and Padmapriya, A. | | 2019 | Rejected |
| A Brief Review of Big Data in the Agriculture Domain | Bazán-Vera, William and Bermeo-Almeida, Oscar and Cardenas-Rodriguez, Mario and Ferruzola-Gómez, Enrique | | 2019 | Rejected |
| A Flexible and Reliable Wireless Sensor Network Architecture for Precision Agriculture in a Tomato Greenhouse | Ramdoo, Vimla Devi and Khedo, Kavi Kumar and Bhoyroo, Vishwakalyan | | 2019 | Rejected |
| Design and Development of the Agricultural Model: A Way to Connect Farmer Community to Agriculture Market for Betterment of Rural Development | Ghadiyali, Tejas and Lad, Kalpesh and Dhodiya, Jayesh | | 2018 | Rejected |
| A Collaborative Approach to Build a KBS for Crop Selection: Combining Experts Knowledge and Machine Learning Knowledge Discovery | Anley, Mulualem Bitew and Tesema, Tibebe Beshah | | 2019 | Rejected |
| Contextual Soft Classification Approaches for Crops Identification Using Multi-sensory Remote Sensing Data: Machine Learning Perspective for | Khobragade, Anand N. and Raghuwanshi, Mukesh M. | | 2015 | Rejected |

| | | | | |
|---|---|---|---|---|
| Satellite Images | | | | |
| Construction and Reuse of Linked Agriculture Data: An Experience of Taiwan Government Open Data | Deng, Dongpo and Mai, Guan-Shuo and Shiau, Steven | | 2018 | Rejected |
| IVIP – A Scientific Workflow System to Support Experts in Spatial Planning of Crop Production | Tuot, Christopher J. and Sintek, Michael and Dengel, Andreas R. | | 2008 | Rejected |
| Environmental reporting and accounting in Australia: Progress, prospects and research priorities | Albert [van Dijk] and Richard Mount and Philip Gibbons and Michael Vardon and Pep Canadell | Science of The Total Environment | 2014 | Rejected |
| Remote sensing of dryland ecosystem structure and function: Progress, challenges, and opportunities | William K. Smith and Matthew P. Dannenberg and Dong Yan and Stefanie Herrmann and Mallory L. Barnes and Greg A. Barron-Gafford and Joel A. Biederman and Scott Ferrenberg and Andrew M. Fox and Amy Hudson and John F. Knowles and Natasha MacBean and David J.P. Moore and Pamela L. Nagler and Sasha C. Reed and William A. Rutherford and Russell L. Scott and Xian Wang and Julia Yang | Remote Sensing of Environment | 2019 | Rejected |
| Challenges and prospects in connectivity analysis in agricultural systems: Actions to implement policies on land management and carbon storage at EU level | A.P. Fernández-Getino and J.L. Alonso-Prados and M.I. Santín-Montanyá | Land Use Policy | 2018 | Rejected |
| The road towards plant phenotyping via WSNs: An overview | Fadi Al-Turjman | Computers and Electronics in Agriculture | 2019 | Rejected |
| Targeting management practices for rice yield gains in stress-prone environments of Myanmar | A.M. Radanielson and Yoichiro Kato and Leo Kris Palao and Gudina Feyisa and Arelene Julia Malabayabas and Jorrel K. Aunario and Cornelia Garcia and Jane G. Balanza and Khin Thawda Win and Rakesh K. Singh and Chenie Zamora and Daw Tin Tin Myint and David E. Johnson | Field Crops Research | 2019 | Rejected |
| Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping | Patrick Griffiths and Claas Nendel and Patrick Hostert | Remote Sensing of Environment | 2019 | Rejected |
| Forecasting yield by integrating agrarian factors and machine learning models: A survey | Dhivya Elavarasan and Durai Raj Vincent and Vishal Sharma and Albert Y. Zomaya and Kathiravan Srinivasan | Computers and Electronics in Agriculture | 2018 | Rejected |
| Multi-level automation of farm management information systems | Dimitris S. Paraforos and Vangelis Vassiliadis and Dietrich Kortenbruck and Kostas Stamkopoulos and Vasileios Ziogas and Athanasios A. Sapounas and Hans W. Griepentrog | Computers and Electronics in Agriculture | 2017 | Rejected |
| Implementation of a magnetorheological damper on a no-till seeding assembly for optimising seeding depth | Galibjon M. Sharipov and Dimitris S. Paraforos and Hans W. Griepentrog | Computers and Electronics in Agriculture | 2018 | Rejected |
| Capturing transformation of flood hazard over a large River Basin under changing climate using a top-down approach | A. Gusain and M.P. Mohanty and S. Ghosh and C. Chatterjee and S. Karmakar | Science of The Total Environment | 2020 | Rejected |
| An IoT-based cognitive monitoring system for early plant disease forecast | Ahmed Khattab and Serag E.D. Habib and Haythem Ismail and Sahar Zayan and Yasmine Fahmy and Mohamed M. Khairy | Computers and Electronics in Agriculture | 2019 | Rejected |
| Spatiotemporal patterns of evapotranspiration, gross primary productivity, and water use efficiency of cropland in agroecosystems and their relation to the water-saving project in the Shiyang River Basin of Northwestern China | Fei Tian and Yu Zhang | Computers and Electronics in Agriculture | 2020 | Rejected |
| Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive | P.V. Potapov and S.A. Turubanova and A. Tyukavina and A.M. Krylov and J.L. McCarty and V.C. Radeloff and M.C. Hansen | Remote Sensing of Environment | 2015 | Rejected |
| Optimize the spatial distribution of crop water consumption based on a cellular automata model: A case study of the middle Heihe River basin, China | Liuyue He and Jianxia Bao and Andre Daccache and Sufen Wang and Ping Guo | Science of The Total Environment | 2020 | Rejected |
| Impacts of reduced model complexity and driver resolution on cropland ecosystem photosynthesis estimates | Andrew Revill and A. Anthony Bloom and Mathew Williams | Field Crops Research | 2016 | Rejected |
| Energy efficient data transmission in automatic irrigation system using wireless sensor networks | M. [Nesa Sudha] and M.L. Valarmathi and Anni Susan Babu | Computers and Electronics in Agriculture | 2011 | Rejected |

| | | | | |
|---|---|---|---|---|
| Scenario farmland protection zoning based on production potential: A case study in China | Yilun Liu and Luo Liu and A-Xing Zhu and Chunhua Lao and Guohua Hu and Yueming Hu | Land Use Policy | 2020 | Rejected |
| Mapping soil moisture with the OPtical TRApezoid Model (OPTRAM) based on long-term MODIS observations | Ebrahim Babaeian and Morteza Sadeghi and Trenton E. Franz and Scott Jones and Markus Tuller | Remote Sensing of Environment | 2018 | Rejected |
| Inter-comparison of satellite-retrieved and Global Land Data Assimilation System-simulated soil moisture datasets for global drought analysis | Yongwei Liu and Yuanbo Liu and Wen Wang | Remote Sensing of Environment | 2019 | Rejected |
| Intensify production, transform biomass to energy and novel goods and protect soils in Europe—A vision how to mobilize marginal lands | P. Schröder and B. Beckers and S. Daniels and F. Gnädinger and E. Maestri and N. Marmiroli and M. Mench and R. Millan and M.M. Obermeier and N. Oustriere and T. Persson and C. Poschenrieder and F. Rineau and B. Rutkowska and T. Schmid and W. Szulc and N. Witters and A. Sæbø | Science of The Total Environment | 2018 | Rejected |
| An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States | David M. Johnson | Remote Sensing of Environment | 2014 | Rejected |
| Monitoring plant diseases and pests through remote sensing technology: A review | Jingcheng Zhang and Yanbo Huang and Ruiliang Pu and Pablo Gonzalez-Moreno and Lin Yuan and Kaihua Wu and Wenjiang Huang | Computers and Electronics in Agriculture | 2019 | Rejected |
| Evaluation and uncertainty estimation of the impact of air quality modelling on crop yields and premature deaths using a multi-model ensemble | Efisio Solazzo and Angelo Riccio and Rita [Van Dingenen] and Luana Valentini and Stefano Galmarini | Science of The Total Environment | 2018 | Rejected |
| Development of an integrated Cropland and Soil Data Management system for cropping system applications | Yubin Yang and Lloyd Ted Wilson and Jing Wang and Xiaobao Li | Computers and Electronics in Agriculture | 2011 | Rejected |
| Integration of bio-physical and economic models to analyze management intensity and landscape structure effects at farm and landscape level | Martin Schönhart and Thomas Schauppenlehner and Erwin Schmid and Andreas Muhar | Agricultural Systems | 2011 | Rejected |
| Land susceptibility to water and wind erosion risks in the East Africa region | Ayele Almaw Fenta and Atsushi Tsunekawa and Nigussie Haregeweyn and Jean Poesen and Mitsuru Tsubo and Pasquale Borrelli and Panos Panagos and Matthias Vanmaercke and Jente Broeckx and Hiroshi Yasuda and Takayuki Kawai and Yasunori Kurosaki | Science of The Total Environment | 2020 | Rejected |
| A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies | Antonio [Rafael Braga] and Danielo [G. Gomes] and Richard Rogers and Edgar [E. Hassler] and Breno [M. Freitas] and Joseph [A. Cazier] | Computers and Electronics in Agriculture | 2020 | Rejected |
| Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review | Anna Chlingaryan and Salah Sukkarieh and Brett Whelan | Computers and Electronics in Agriculture | 2018 | Rejected |
| A methodology to assess the impact of climate variability and change on water resources, food security and economic welfare | Abdelaziz A. Gohar and Adrian Cashman | Agricultural Systems | 2016 | Rejected |
| Mapping pine plantations in the southeastern U.S. using structural, spectral, and temporal remote sensing data | M.E. Fagan and D.C. Morton and B.D. Cook and J. Masek and F. Zhao and R.F. Nelson and C. Huang | Remote Sensing of Environment | 2018 | Rejected |
| Modelling spatial and temporal variability of water quality from different monitoring stations using mixed effects model theory | Davor Romić and Annamaria Castrignanò and Marija Romić and Gabriele Buttafuoco and Marina [Bubalo Kovačić] and Gabrijel Ondrašek and Monika Zovko | Science of The Total Environment | 2020 | Rejected |
| Integrated sensing of soil moisture at the field-scale: Measuring, modeling and sharing for improved agricultural decision support | Andrew J. Phillips and Nathaniel K. Newlands and Steve H.L. Liang and Benjamin H. Ellert | Computers and Electronics in Agriculture | 2014 | Rejected |
| Soybean yield prediction from UAV using multimodal data fusion and deep learning | Maitiniyazi Maimaitijiang and Vasit Sagan and Paheding Sidike and Sean Hartling and Flavio Esposito and Felix B. Fritschi | Remote Sensing of Environment | 2020 | Accepted |
| Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia | Omid Rahmati and Fatemeh Falah and Kavina Shaanu Dayal and Ravinesh C. Deo and Farnoush Mohammadi and Trent Biggs and Davoud Davoudi Moghaddam and Seyed Amir Naghibi and Dieu Tien Bui | Science of The Total Environment | 2020 | Rejected |
| Exploring the characteristics and utilisation of Farm Management Information Systems (FMIS) | Jana Munz and Nicola Gindele and Reiner Doluschitz | Computers and Electronics in Agriculture | 2020 | Rejected |

| | | | | |
|---|---|---|---|---|
| in Germany | | | | |
| Integration of in situ and satellite data for top-down mapping of Ambrosia infection level | Predrag Lugonja and Sanja Brdar and Isidora Simović and Gordan Mimić and Yuliia Palamarchuk and Mikhail Sofiev and Branko Šikoparija | Remote Sensing of Environment | 2019 | Rejected |
| Towards globally customizable ecosystem service models | Javier Martínez-López and Kenneth J. Bagstad and Stefano Balbi and Ainhoa Magrach and Brian Voigt and Ioannis Athanasiadis and Marta Pascual and Simon Willcock and Ferdinando Villa | Science of The Total Environment | 2019 | Rejected |
| The Evaporative Stress Index as an indicator of agricultural drought in Brazil: An assessment based on crop yield impacts | Martha C. Anderson and Cornelio A. Zolin and Paulo C. Sentelhas and Christopher R. Hain and Kathryn Semmens and M. [Tugrul Yilmaz] and Feng Gao and Jason A. Otkin and Robert Tetrault | Remote Sensing of Environment | 2016 | Rejected |
| Deeply synergistic optical and SAR time series for crop dynamic monitoring | Wenzhi Zhao and Yang Qu and Jiage Chen and Zhanliang Yuan | Remote Sensing of Environment | 2020 | Rejected |
| Contribution of ecological policies to vegetation restoration: A case study from Wuqi County in Shaanxi Province, China | Daojun Zhang and Qiqi Jia and Xin Xu and Shunbo Yao and Haibin Chen and Xianhui Hou | Land Use Policy | 2018 | Rejected |
| Optoelectronic proximal sensing vehicle-mounted technologies in precision agriculture: A review | Federico Pallottino and Francesca Antonucci and Corrado Costa and Carlo Bisaglia and Simone Figorilli and Paolo Menesatti | Computers and Electronics in Agriculture | 2019 | Rejected |
| Next generation data systems and knowledge products to support agricultural producers and science-based policy decision making | Susan M. Capalbo and John M. Antle and Clark Seavert | Agricultural Systems | 2017 | Rejected |
| Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine | Luo Liu and Xiangming Xiao and Yuanwei Qin and Jie Wang and Xinliang Xu and Yueming Hu and Zhi Qiao | Remote Sensing of Environment | 2020 | Rejected |
| Scale effect on spatial patterns of ecosystem services and associations among them in semi-arid area: A case study in Ningxia Hui Autonomous Region, China | Shuna Xu and Yanfang Liu and Xia Wang and Guangxia Zhang | Science of The Total Environment | 2017 | Rejected |
| Complex water management in modern agriculture: Trends in the water-energy-food nexus over the High Plains Aquifer | Samuel J. Smidt and Erin M.K. Haacker and Anthony D. Kendall and Jillian M. Deines and Lisi Pei and Kayla A. Cotterman and Haoyang Li and Xiao Liu and Bruno Basso and David W. Hyndman | Science of The Total Environment | 2016 | Rejected |
| Detecting irrigation extent, frequency, and timing in a heterogeneous arid agricultural region using MODIS time series, Landsat imagery, and ancillary data | Yaoliang Chen and Dengsheng Lu and Lifeng Luo and Yadu Pokhrel and Kalyanmoy Deb and Jingfeng Huang and Youhua Ran | Remote Sensing of Environment | 2018 | Rejected |
| Sustainable agricultural development in a rural area in the Netherlands? Assessing impacts of climate and socio-economic change at farm and landscape level | Pytrik Reidsma and Martha M. Bakker and Argyris Kanellopoulos and Shah J. Alam and Wim Paas and Johannes Kros and Wim [de Vries] | Agricultural Systems | 2015 | Rejected |
| Socio-economic and biophysical determinants of land degradation in Vietnam: An integrated causal analysis at the national level | Quyet Manh Vu and Quang Bao Le and Emmanuel Frossard and Paul L.G. Vlek | Land Use Policy | 2014 | Rejected |
| Field-based phenomics for plant genetics research | Jeffrey W. White and Pedro Andrade-Sanchez and Michael A. Gore and Kevin F. Bronson and Terry A. Coffelt and Matthew M. Conley and Kenneth A. Feldmann and Andrew N. French and John T. Heun and Douglas J. Hunsaker and Matthew A. Jenks and Bruce A. Kimball and Robert L. Roth and Robert J. Strand and Kelly R. Thorp and Gerard W. Wall and Guangyao Wang | Field Crops Research | 2012 | Rejected |
| Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology | Sander J.C. Janssen and Cheryl H. Porter and Andrew D. Moore and Ioannis N. Athanasiadis and Ian Foster and James W. Jones and John M. Antle | Agricultural Systems | 2017 | Rejected |
| Farm ponds in southern China: Challenges and solutions for conserving a neglected wetland ecosystem | Wenjun Chen and Bin He and Daniel Nover and Haiming Lu and Jian Liu and Wei Sun and Wen Chen | Science of The Total Environment | 2019 | Rejected |
| A review on the practice of big data analysis in agriculture | Andreas Kamilaris and Andreas Kartakoullis and Francesc X. Prenafeta-Boldú | Computers and Electronics in Agriculture | 2017 | Rejected |

| | | | | |
|---|---|---|---|---|
| Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods | François Waldner and Yang Chen and Roger Lawes and Zvi Hochman | Remote Sensing of Environment | 2019 | Rejected |
| AgroPortal: A vocabulary and ontology repository for agronomy | Clément Jonquet and Anne Toulet and Elizabeth Arnaud and Sophie Aubin and Esther [Dzalé Yeumo] and Vincent Emonet and John Graybeal and Marie-Angélique Laporte and Mark A. Musen and Valeria Pesce and Pierre Larmande | Computers and Electronics in Agriculture | 2018 | Rejected |
| Improving arable farm enterprise integration – Review of existing technologies and practices from a farmer's perspective | J.W. Kruize and R.M. Robbemond and H. Scholten and J. Wolfert and A.J.M. Beulens | Computers and Electronics in Agriculture | 2013 | Rejected |
| Evaluating the GHG mitigation-potential of alternate wetting and drying in rice through life cycle assessment | Cara Fertitta-Roberts and Patricia Y. Oikawa and G. [Darrel Jenerette] | Science of The Total Environment | 2019 | Rejected |
| Modeling vulnerability of groundwater to pollution under future scenarios of climate change and biofuels-related land use change: A case study in North Dakota, USA | Ruopu Li and James W. Merchant | Science of The Total Environment | 2013 | Rejected |
| Assessing lost cultural heritage. A case study of the eastern coast of Las Palmas de Gran Canaria city (Spain) | Eva Pérez-Hernández and Carolina Peña-Alonso and Luis Hernández-Calvento | Land Use Policy | 2020 | Rejected |
| Improving remotely-sensed crop monitoring by NDVI-based crop phenology estimators for corn and soybeans in Iowa and Illinois, USA | Bumsuk Seo and Jihye Lee and Kyung-Do Lee and Sukyoung Hong and Sinkyu Kang | Field Crops Research | 2019 | Rejected |
| A mechanical-dielectric-high frequency acoustic sensor fusion for soil physical characterization | Mojtaba Naderi-Boldaji and Mehari Z. Tekeste and Richard A. Nordstorm and Daniel J. Barnard and Stuart J. Birrell | Computers and Electronics in Agriculture | 2019 | Rejected |
| Coupling analysis of greenhouse-led farmland transition and rural transformation development in China's traditional farming area: A case of Qingzhou City | Dazhuan Ge and Zhihua Wang and Shuangshuang Tu and Hualou Long and Huili Yan and Dongqi Sun and Weifeng Qiao | Land Use Policy | 2019 | Rejected |
| The use of satellite data for crop yield gap analysis | David B. Lobell | Field Crops Research | 2013 | Rejected |
| A reference architecture for Farm Software Ecosystems | J.W. Kruize and J. Wolfert and H. Scholten and C.N. Verdouw and A. Kassahun and A.J.M. Beulens | Computers and Electronics in Agriculture | 2016 | Rejected |
| Evaluating management zone maps for variable rate fungicide application and selective harvest | Rebecca L. Whetton and Toby W. Waine and Abdul M. Mouazen | Computers and Electronics in Agriculture | 2018 | Rejected |
| A comparative assessment of land suitability evaluation methods for agricultural land use planning at village level | Duraisamy Vasu and Rajeev Srivastava and Nitin G. Patil and Pramod Tiwary and Padikkal Chandran and Surendra [Kumar Singh] | Land Use Policy | 2018 | Rejected |
| A smart multiple spatial and temporal resolution system to support precision agriculture from satellite images: Proof of concept on Aglianico vineyard | A. Brook and V. [De Micco] and G. Battipaglia and A. Erbaggio and G. Ludeno and I. Catapano and A. Bonfante | Remote Sensing of Environment | 2020 | Rejected |
| Estimating wheat yield by integrating the WheatGrow and PROSAIL models | Zhang, L. and Guo, C.L. and Zhao, L.Y. and Zhu, Y. and Cao, W.X. and Tian, Y.C. and Cheng, T. and Wang, X. | Field Crops Research | 2016 | Rejected |
| Dynamic within-season irrigation scheduling for maize production in Northwest China: A Method Based on Weather Data Fusion and yield prediction by DSSAT | Chen, S. and Jiang, T. and Ma, H. and He, C. and Xu, F. and Malone, R.W. and Feng, H. and Yu, Q. and Siddique, K.H.M. and Dong, Q. and He, J. | Agricultural and Forest Meteorology | 2020 | Rejected |
| Sentinel-1&2 for near real time cropping pattern monitoring in drought prone areas. Application to irrigation water needs in Telangana, South-India | Ferrant, S. and Selles, A. and Le Page, M. and AlBitar, A. and Mermoz, S. and Gascoin, S. and Bouvet, A. and Ahmed, S. and Kerr, Y. | International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives | 2019 | Rejected |
| Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images | Wang, L. and Tian, Y. and Yao, X. and Zhu, Y. and Cao, W. | Field Crops Research | 2014 | Accepted |
| Establishing crop productivity using RADARSAT-2 | McNairn, H. and Shang, J. and Jiao, X. and Deschamps, B. | International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives | 2012 | Rejected |
| Bioinformatics resources for pollen | Ambrosino, L. and Bostan, H. and Ruggieri, V. and Chiusano, M.L. | Plant Reproduction | 2016 | Rejected |
| Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils | Rodrigues, F.A. and Bramley, R.G.V. and Gobbett, D.L. | Geoderma | 2015 | Rejected |
| Performance of optimized hyperspectral reflectance indices and partial least squares regression for estimating the chlorophyll | El-Hendawy, S. and Al-Suhaibani, N. and Elsayed, S. and Alotaibi, M. and Hassan, W. and Schmidhalter, U. | Plant Physiology and Biochemistry | 2019 | Rejected |

| | | | | |
|---|---|---|---|---|
| fluorescence and grain yield of wheat grown in simulated saline field conditions | | | | |
| Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine | You, N. and Dong, J. | ISPRS Journal of Photogrammetry and Remote Sensing | 2020 | Accepted |
| Batch-processing of AquaCrop plug-in for rainfed maize using satellite derived Fractional Vegetation Cover data | Mohamed Sallah, A.-H. and Tychon, B. and Piccard, I. and Gobin, A. and Van Hoolst, R. and Djaby, B. and Wellens, J. | Agricultural Water Management | 2019 | Rejected |
| Large-scale data integration reveals colocalization of gene functional groups with meta-QTL for multiple disease resistance in barley | Schweizer, P. and Stein, N. | Molecular Plant-Microbe Interactions | 2011 | Accepted |
| A categorization framework for common computer vulnerabilities and exposures | Chen, Z. and Zhang, Y. and Chen, Z. | Computer Journal | 2010 | Rejected |
| FuzzyFusion TM, an application architecture for multisource information fusion | Fox, K.L. and Henning, R.R. | Proceedings of SPIE - The International Society for Optical Engineering | 2009 | Rejected |
| Resilient Communication, Object Classification and Data Fusion in Unmanned Aerial Systems | Allison, J.A. and Ptucha, R. and Lyshevski, S.E. | 2018 International Conference on Unmanned Aircraft Systems, ICUAS 2018 | 2018 | Rejected |
| Inverse Problems and Data Fusion for Crop Production Applications Targeting Optimal Growth - Fertilization | Kaur, B. and Owusu, R. K. A. | None | 2015 | Duplicated |
| A Smart Farming Approach in Automatic Detection of Favorable Conditions for Planting and Crop Production in the Upper Basin of Cauca River | López, Iván Darío and Corrales, Juan Carlos | None | 2018 | Duplicated |
| Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach | Mann, M.L. and Warner, J.M. | Field Crops Research | 2017 | Duplicated |
| Inverse Problems and Data Fusion for crop production applications targeting optimal growth-Fertilization | Kaur, Bipjeet and Owusu, Robert KA | None | 2015 | Duplicated |

## A.4 Systematic Review Report around Multi-Label Classification

Algorithms and techniques for crop prediction from MLC approach.

***Planning***

Describe approaches that have used Multi-Label Classification (MLC) in agricultural contexts and specifically for crop prediction.

**PICOC**

- Population: Official open data, restricted data with access permission, results of agricultural vulnerability analysis
- Intervention: Multi-Label Classification
- Comparison: A comparison of methods, algorithms, datasets
- Outcome: A systematic literature survey report including synthesis of most relevant articles published on multi-label classification for crop prediction
- Context: A systematic investigation to consolidate a peer-reviewed and academic research, classification and comparison, trends and future research directions

**Research Questions**

RQ.1. What Multi-Label Classification (MLC) approaches have been applied in agriculture?

RQ.2. What Multi-Label Classification (MLC) approaches have been applied for crop prediction?

**Keywords and Synonyms**

| Keyword | Synonyms |
|---|---|
| agricultural vulnerability analysis | agricultural vulnerability, methodologies for agricultural vulnerability |
| algorithm | approach, method, technique |
| machine learning | artificial intelligence |
| multi-label classification | mlc, multi-label, multi-label learning |
| multi-label dataset | mld, multi-label data |
| open data | data, dataset |

**Search String**

("agricultural vulnerability analysis" OR "agricultural vulnerability" OR "methodologies for agricultural vulnerability" OR "multi-label dataset" OR "mld" OR "multi-label data" OR "open data" OR "data" OR "dataset") AND ("machine learning" OR "artificial intelligence" OR "multi-label classification" OR "mlc" OR "multi-label" OR "multi-label learning") AND ("algorithm" OR "approach" OR "method" OR "technique")

**Sources**

- Google Scholar (https://scholar.google.com/)
- IEEE Digital Library (http://ieeexplore.ieee.org)
- Science@Direct (http://www.sciencedirect.com)
- Scopus (http://www.scopus.com)
- Springer Link (http://link.springer.com)

**Selection Criteria**

Inclusion Criteria:

- Papers about MLC applied to agriculture
- Papers from the area of computer science
- Papers published in the last ten years
- Papers that include one or more keywords
- Papers whose objectives are similar to research, but applying other techniques

Exclusion Criteria:
- It is not an article
- Papers about MLC applied to domains other than agriculture
- Papers not published in the last ten years
- Papers outside the area of computer science
- Papers that not included the keywords
- Techniques other than MLC applied to agriculture

**Quality Assessment Checklist**

Questions:
- Is the study of value for research or practice?
- Is there a clear statement of findings?
- Was the data analysis sufficiently rigorous?
- Was the data collected in a way that addressed the research issue?
- Is there a description of the context in which the research was carried out?
- Is there a clear statement of the aims of the research?
- Is the paper based on research or is it merely a report based on expert opinion?

Answers:
- Yes
- Partly
- No

**Data Extraction Form**
- Has any MLC technique been applied in the area of agriculture?
- If the previous answer is yes, any MLC technique been applied in crop prediction?
- Is there any methodology or method to apply MLC?
- Is there any methodology to evaluate the MLC performance?
- Is any experiment carried out?
- If the previous answer is yes, choose the evaluation method used

***Conducting***

**Digital Libraries Search Strings**

IEEE Digital Library:
(("All Metadata": "multi-label classification") AND "All Metadata":agriculture)

Science@Direct:
("multi-label classification" OR "multi-label" OR "multi-label learning") AND
("agriculture" OR "crop" OR "agricultural")

Scopus:
( TITLE-ABS-KEY ( “multi-label classification” OR “multi-label” OR “multi-label learning” ) AND TITLE-ABS-KEY ( “agriculture” OR “crop” OR “agricultural” ) ) AND ( LIMIT-TO ( SUBJAREA , “COMP” ) ) AND ( EXCLUDE ( DOCTYPE , “cr” ) OR EXCLUDE ( DOCTYPE , “ed” ) OR EXCLUDE ( DOCTYPE , “re” ) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) OR LIMIT-TO ( PUBYEAR , 2014 ) OR LIMIT-TO ( PUBYEAR , 2013 ) OR LIMIT-TO ( PUBYEAR , 2012 ) OR LIMIT-TO ( PUBYEAR , 2011 ) )

Springer Link:
(“multi-label classification” OR “multi-label” OR “multi-label learning”) AND (“agriculture” OR “crop” OR “agricultural”)

## Imported Studies

- Google Scholar: 9
- IEEE Digital Library: 4
- Science@Direct: 25
- Scopus: 11
- Springer Link: 8

## Related Works

| Title | Author | Journal | Year | Status |
|---|---|---|---|---|
| Multi-label learning for crop leaf diseases recognition and severity estimation based on convolutional neural networks | Ji, M. and Zhang, K. and Wu, Q. and Deng, Z. | Soft Computing | 2020 | Accepted |
| Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions | Picon, A. and Seitz, M. and Alvarez-Gila, A. and Mohnke, P. and Ortiz-Barredo, A. and Echazarra, J. | Computers and Electronics in Agriculture | 2019 | Accepted |
| Machine Learning for Apple Fruit Diseases Classification System | Abd El-aziz, A.A. and Darwish, A. and Oliva, D. and Hassanien, A.E. | Advances in Intelligent Systems and Computing | 2020 | Accepted |
| Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based Land Transformation Model: an integrated ML-CA-LTM modeling framework | Omrani, Hichem and Tayyebi, Amin and Pijanowski, Bryan | GIScience \& Remote Sensing | 2017 | Accepted |
| Multi-label class assignment in land-use modelling | Omrani, Hichem and Abdallah, Fahed and Charif, Omar and Longford, Nicholas T | International Journal of Geographical Information Science | 2015 | Accepted |
| Plant recommender system based on multi-label classification | Tharwat, Alaa and Mahdi, Hani and Hassanien, Aboul Ella | | 2016 | Accepted |
| Research on deep learning in apple leaf disease recognition | Yong Zhong and Ming Zhao | Computers and Electronics in Agriculture | 2020 | Accepted |
| CUPID: consistent unlabeled probability of identical distribution for image classification | Zhonglong Zheng and Jianshu Zhang and Suhang Zhu and Changbing Tang and Feilong Lin and Hui Lan and Zhongyu Chen and Jie Yang | Knowledge-Based Systems | 2017 | Rejected |
| Mutual information-based label distribution feature selection for multi-label learning | Wenbin Qian and Jintao Huang and Yinglong Wang and Wenhao Shu | Knowledge-Based Systems | 2020 | Rejected |
| A procedural texture generation framework based on semantic descriptions | Junyu Dong and Lina Wang and Jun Liu and Ying Gao and Lin Qi and Xin | Knowledge-Based Systems | 2019 | Rejected |

| | | | | |
|---|---|---|---|---|
| | Sun | | | |
| A novel image measurement algorithm for common mushroom caps based on convolutional neural network | Chuan-Pin Lu and Jiun-Jian Liaw | Computers and Electronics in Agriculture | 2020 | Rejected |
| Weighted feature selection via discriminative sparse multi-view learning | Jing Zhong and Nan Wang and Qiang Lin and Ping Zhong | Knowledge-Based Systems | 2019 | Rejected |
| A constrained least squares regression model | Haoliang Yuan and Junjie Zheng and Loi Lei Lai and Yuan Yan Tang | Information Sciences | 2018 | Rejected |
| Automatic image annotation via compact graph based semi-supervised learning | Mingbo Zhao and Tommy W.S. Chow and Zhao Zhang and Bing Li | Knowledge-Based Systems | 2015 | Rejected |
| Wikipedia-based cross-language text classification | Marcos Antonio [Mouriño García] and Roberto [Pérez Rodríguez] and Luis [Anido Rifón] | Information Sciences | 2017 | Rejected |
| Performance evaluation of deep feature learning for RGB-D image/video classification | Ling Shao and Ziyun Cai and Li Liu and Ke Lu | Information Sciences | 2017 | Rejected |
| Joint correntropy metric weighting and block diagonal regularizer for robust multiple kernel subspace clustering | Chao Yang and Zhenwen Ren and Quansen Sun and Mingna Wu and Maowei Yin and Yuan Sun | Information Sciences | 2019 | Rejected |
| Robust $\ell 2-$Hypergraph and its applications | Taisong Jin and Zhengtao Yu and Yue Gao and Shengxiang Gao and Xiaoshuai Sun and Cuihua Li | Information Sciences | 2019 | Rejected |
| YAKE! Keyword extraction from single documents using multiple local features | Ricardo Campos and Vítor Mangaravite and Arian Pasquali and Alípio Jorge and Célia Nunes and Adam Jatowt | Information Sciences | 2020 | Rejected |
| Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening | Tao Li and Yingqi Gao and Kai Wang and Song Guo and Hanruo Liu and Hong Kang | Information Sciences | 2019 | Rejected |
| Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association | Thiago T. Santos and Leonardo L. [de Souza] and Andreza A. [dos Santos] and Sandra Avila | Computers and Electronics in Agriculture | 2020 | Rejected |
| FLYOLOv3 deep learning for key parts of dairy cow body detection | Bo Jiang and Qian Wu and Xuqiang Yin and Dihua Wu and Huaibo Song and Dongjian He | Computers and Electronics in Agriculture | 2019 | Rejected |
| Sparse feature space representation: A unified framework for semi-supervised and domain adaptation learning | Long Liu and Lechao Yang and Bin Zhu | Knowledge-Based Systems | 2018 | Rejected |
| Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine | Xinye Wang and Yi Yang and Yitian Xu and Qian Chen and Hongmei Wang and Huafang Gao | Knowledge-Based Systems | 2020 | Rejected |
| Nondestructive detection of chilled mutton freshness based on multi-label information fusion and adaptive BP neural network | Jiang Xinhua and Xue Heru and Zhang Lina and Gao Xiaojing and Wu Guodong and Bai Jie | Computers and Electronics in Agriculture | 2018 | Rejected |
| QuMinS: Fast and scalable querying, mining and summarizing multi-modal databases | Robson L.F. Cordeiro and Fan Guo and Donna S. Haverkamp and James H. Horne and Ellen K. Hughes and Gunhee Kim and Luciana A.S. Romani and Priscila P. Coltri and Tamires T. Souza and Agma J.M. Traina and Caetano Traina and Christos Faloutsos | Information Sciences | 2014 | Rejected |
| Visual features based automated identification of fish species using deep convolutional neural networks | Hafiz Tayyab Rauf and M. Ikram Ullah Lali and Saliha Zahoor and Syed Zakir Hussain Shah and Abd Ur Rehman and Syed Ahmad Chan Bukhari | Computers and Electronics in Agriculture | 2019 | Rejected |
| Deep model with neighborhood-awareness for text tagging | Shaowei Qin and Hao Wu and Rencan Nie and Jun He | Knowledge-Based Systems | 2020 | Rejected |
| Classification of breast cancer histology images using incremental boosting convolution networks | Duc My Vo and Ngoc-Quang Nguyen and Sang-Woong Lee | Information Sciences | 2019 | Rejected |
| Semi-supervised learning with convolutional neural networks for UAV images automatic recognition | Willian Paraguassu Amorim and Everton Castelão Tetila and Hemerson Pistori and João Paulo Papa | Computers and Electronics in Agriculture | 2019 | Rejected |
| Feature Selection for Multi-label Learning Using Mutual Information and GA | Yu, Yingand Wang, Yinglong | | 2014 | Rejected |
| Multi-label Classification Using Rough Sets | Yu, Yingand Miao, Duoqianand Zhang, Zhifeiand Wang, Lei | | 2013 | Rejected |
| Combining Dimensionality Reduction with Random Forests for Multi-label Classification Under Interactivity Constraints | Nair-Benrekia, Noureddine-Yassineand Kuntz, Pascaleand Meyer, Frank | | 2017 | Rejected |
| Low-Shot Multi-label Incremental Learning for Thoracic Diseases Diagnosis | Wang, Qingfengand Cheng, Jie-Zhiand Zhou, Yingand Zhuang, Hangand Li, Changlongand Chen, Boand Liu, Zhiqinand Huang, Junand Wang, Chaoand Zhou, Xuehai | | 2018 | Rejected |
| IPC Multi-label Classification Based on the Field | Lim, Soraand Kwon, YongJin | | 2016 | Rejected |

| | | | | |
|---|---|---|---|---|
| Functionality of Patent Documents | | | | |
| A Double Weighted Naive Bayes for Multi-label Classification | Yan, Xuesongand Li, Weiand Wu, Qinghuaand Sheng, Victor S. | | 2016 | Rejected |
| Single-label and multi-label conceptor classifiers in pre-trained neural networks | Guangwu Qian and Lei Zhang and Yan Wang | Neural Computing and Applications | 2018 | Rejected |
| Systematic approach of multi-label classification for production scheduling | Muñoz, Edrisi and Capón-García, Elisabet | Computers \& Chemical Engineering | 2019 | Rejected |
| Deep Learning-a New Approach for Multi-Label Scene Classification in Planetscope and Sentinel-2 Imagery | Shendryk, Iurii and Rist, Yannik and Lucas, Rob and Thorburn, Peter and Ticehurst, Catherine | | 2018 | Accepted |
| A comparitive study of land classification using remotely sensed data | K. {Kulkarni} and P. A. {Vijaya} | | 2017 | Accepted |
| AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms | Doshi, Z. and Nadkarni, S. and Agrawal, R. and Shah, N. | Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018 | 2018 | Accepted |
| AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms | Z. {Doshi} and S. {Nadkarni} and R. {Agrawal} and N. {Shah} | | 2018 | Duplicated |
| Fast Multi-Label Low-Rank Linearized SVM Classification Algorithm Based on Approximate Extreme Points | Z. {Sun} and K. {Hu} and T. {Hu} and J. {Liu} and K. {Zhu} | IEEE Access | 2018 | Rejected |
| A Digital Signage Audience Classification Model Based on the Huff Model and Backpropagation Neural Network | X. {Zhang} and X. {Xie} and Y. {Wang} and X. {Zhang} and D. {Jiang} and C. {Yu} and Y. {Liang} | IEEE Access | 2020 | Rejected |
| Weakly-supervised learning approach for potato defects segmentation | Marino, S. and Beauseroy, P. and Smolarz, A. | Engineering Applications of Artificial Intelligence | 2019 | Rejected |
| Multiple-Instance ranking based deep hashing for multi-Label image retrieval | Chen, G. and Cheng, X. and Su, S. and Tang, C. | Neurocomputing | 2020 | Rejected |
| Beyond object proposals: Random crop pooling for multi-label image recognition | Wang, M. and Luo, C. and Hong, R. and Tang, J. and Feng, J. | IEEE Transactions on Image Processing | 2016 | Rejected |
| A big remote sensing data analysis using deep learning framework | Balti, H. and Chebbi, I. and Mellouli, N. and Farah, I.R. and Lamolle, M. | Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019 | 2019 | Rejected |
| Categorising videos using a personalised category catalogue | Bairi, R.B. and Vani, A. and Ahuja, P. and Ramakrishnan, G. | ACM International Conference Proceeding Series | 2015 | Accepted |
| Multi-label Classification of Big NCDC Weather Data Using Deep Learning Model | Doreswamy and Gad, I. and Manjunatha, B.R. | Communications in Computer and Information Science | 2018 | Accepted |
| Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions | Artzai Picon and Maximiliam Seitz and Aitor Alvarez-Gila and Patrick Mohnke and Amaia Ortiz-Barredo and Jone Echazarra | Computers and Electronics in Agriculture | 2019 | Duplicated |
| Multi-label classification using rough sets | Yu, Ying and Miao, Duoqian and Zhang, Zhifei and Wang, Lei | None | 2013 | Duplicated |
| Fast multi-label low-rank linearized SVM classification algorithm based on approximate extreme points | Sun, Zhongwei and Hu, Keyong and Hu, Tong and Liu, Jing and Zhu, Kai | IEEE Access | 2018 | Duplicated |
| Nondestructive detection of chilled mutton freshness based on multi-label information fusion and adaptive bp neural network | Xinhua, Jiang and Heru, Xue and Lina, Zhang and Xiaojing, Gao and Guodong, Wu and Jie, Bai | Computers and electronics in agriculture | 2018 | Duplicated |
| Multi-label Classification of Big NCDC Weather Data Using Deep Learning Model | Doreswamy and Gad, Ibrahim and Manjunatha, B. R. | None | 2018 | Duplicated |
| Beyond object proposals: Random crop pooling for multi-label image recognition | Wang, Meng and Luo, Changzhi and Hong, Richang and Tang, Jinhui and Feng, Jiashi | IEEE Transactions on Image Processing | 2016 | Duplicated |
| A comparitive study of land classification using remotely sensed data | Kulkarni, K. and Vijaya, P.A. | Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017 | 2018 | Duplicated |

# Appendix B

# Data Sources Description

This appendix presents the data dictionary, which compiles the most relevant characteristics about the 16 initial data sources used in this research. Data description is shown from Table B. 1 to Table B. 16.

## B.1 SIVICAP Data Source

| | | | |
|---|---|---|---|
| **Source name** | bp_sivicap | **File format** | CSV |
| **Dimension** | Biophysical | | |
| **Indicator** | Water | | |
| **Data provider information** | SIVICAP is an information system for monitoring the quality of water for human consumption. Developed by the National Institute of Health (INS for its acronym in Spanish) which is a public scientific-technical public health, national coverage, which contributes to the protection of health in Colombia through knowledge management, the monitoring of the health status of the population and the provision of assets and services of interest in public health. | | |
| **Data source overview** | Report of the physical, chemical and microbiological characteristics of water quality in different municipalities. | | |
| **URL** | https://www.datos.gov.co/Salud-y-Protecci-n-Social/Caracter-sticas-Calidad-del-Agua-SIVICAP/jjzc-8w82 | | |
| **Language** | Spanish | **Time window** | 2015 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 1,019 | **Number of attributes** | 47 |
| **Number of files** | 1 | **Data source size** | 238 KB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| ANIO | Year of sampling | Integer |
| DEPARTAMENTO | Territorial units of first level in Colombia. | Nominal |

| | | |
|---|---|---|
| MUNICIPO | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| TOTAL.MUESTRAS | Total Samples | Numeric |
| RESULTADO.COLOR.APARENTE | Apparent Color | Numeric |
| RESULTADO.TURBIEDAD | Turbidity | Numeric |
| RESULTADO.PH | pH | Numeric |
| RESULTADO.CLORO.RESIDUAL.LIBRE | Free Residual Chlorine | Numeric |
| RESULTADO.ALCALINIDAD.TOTAL | Total Alkalinity | Numeric |
| RESULTADO.CALCIO | Calcium | Numeric |
| RESULTADO.FOSFATOS | Phosphates | Numeric |
| RESULTADO.MANGANESO | Manganese | Numeric |
| RESULTADO.MOLIBDENO | Molybdenum | Numeric |
| RESULTADO.MAGNESIO | Magnesium | Numeric |
| RESULTADO.ZINC | Zinc | Numeric |
| RESULTADO.DUREZA.TOTAL | Total Hardness | Numeric |
| RESULTADO.SULFATOS | Sulphates | Numeric |
| RESULTADO.HIERRO.TOTAL | Total Iron | Numeric |
| RESULTADO.CLORUROS | Chloride | Numeric |
| RESULTADO.NITRATOS | Nitrates | Numeric |
| RESULTADO.NITRITOS | Nitrites | Numeric |
| RESULTADO.ALUMINIO | Aluminum | Numeric |
| RESULTADO.FLORUROS | Fluorides | Numeric |
| RESULTADO.COT | COT | Numeric |
| RESULTADO.COLIFORMES.TOTALES | Total Coliforms | Numeric |
| RESULTADO.E.COLI | E Coli | Numeric |
| RESULTADO.ANTIMONIO | Antimony | Numeric |
| RESULTADO.ARSÉNICO | Arsenic | Numeric |
| RESULTADO.BARIO | Barium | Numeric |
| RESULTADO.CADMIO | Cadmium | Numeric |
| RESULTADO.CIANURO.LIBRE.Y.DIASOCIABLE | Free and Dissociable Cyanide | Numeric |
| RESULTADO.COBRE | Copper | Numeric |
| RESULTADO.Cromo.total | Total Chrome | Numeric |
| RESULTADO.MERCURIO | Mercury | Numeric |
| RESULTADO.NIQUEL | Nickel | Numeric |
| RESULTADO.PLOMO | Lead | Numeric |
| RESULTADO.SELENIO | Selenium | Numeric |
| RESULTADO.TRIHALOMETANOS.TOTALES | Total Trihalomethanes | Numeric |
| RESULTADO.HIDROCARBUROS.AROMÁTICOS .POLICICLICOS | Polycyclic Aromatic Hydrocarbons | Numeric |
| RESULTADO.GIARDIA | Giardia | Numeric |
| RESULTADO.CRYPTOSPORIDIUM | Cryptosporidium | Integer |
| RESULTADO.PLAGUICIDAS.TOTALES | Total Pesticides | Integer |
| RESULTADO.ORGANOFOSFORADOS.Y.CARBA MATOS | Organophosphorates and Carbamates | Integer |
| RESULTADO.MESÓFILOS | Mesophiles | Numeric |
| IRCA.PROMEDIO | IRCA Average | Numeric |
| IRCA.BASE.PROMEDIO | IRCA Base Average | Numeric |
| NIVEL.DE.RIESGO.PROMEDIO | Average Risk Level | Nominal |

**Table B. 1.** Metadata for SIVICAP data source.

# B.2 CORPOICA Data Source

| | | | |
|---|---|---|---|
| **Source name** | bp_corpoica | **File format** | CSV |
| **Dimension** | Biophysical | | |
| **Indicator** | Soil | | |

| | | | |
|---|---|---|---|
| **Data provider information** | The Colombian Agricultural Research Corporation (CORPOICA for its acronym in Spanish), is a decentralized public entity responsible for generating scientific knowledge and technological solutions through research, innovation, technology transfer and training of researchers, for the benefit of the Colombian agricultural sector. | | |
| **Data source overview** | Soil analysis service of CORPOICA for the agricultural sector. It focuses on evaluating soil fertility, salinity, and parameters to develop fertilization plans, application of amendments, and land adequacy to achieve profitable production. | | |
| **URL** | https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Resultados-de-An-lisis-de-Laboratorio-Suelos-en-Co/ch4u-f3i5 | | |
| **Language** | Spanish | **Time window** | 2013 - 2016 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 24,179 | **Number of attributes** | 31 |
| **Number of files** | 1 | **Data source size** | 6.8 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Departamento | Territorial units of first level in Colombia. | Nominal |
| Municipio | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| Cultivo | Crop | Nominal |
| Estado | Crop state | Nominal |
| Tiempo Establecimiento | Crop establishment time | Nominal |
| Topografia | Topography | Nominal |
| Drenaje | Sewer system | Nominal |
| Riego | Irrigation | Nominal |
| Fertilizantes aplicados | Applied fertilizers | Nominal |
| Fecha de análisis | Date of analysis | Date |
| pH | Water pH: soil 2.5: 1.0 | Numeric |
| Materia organica | Organic matter (MO)% | Numeric |
| Fosforo | Phosphorus (P) Bray II mg / kg | Numeric |
| Azufre | Sulfur (S) Monocalcium Phosphate mg / kg | Numeric |
| Acidez | Acidity (Al + H) KCL cmol (+) / kg | Numeric |
| Aluminio intercambiable | Aluminum (Al) interchangeable cmol (+) / kg | Numeric |
| Calcio intercambiable | Calcium (Ca) exchangeable cmol (+) / kg | Numeric |
| Magnesio intercambiable | Magnesium (Mg) exchangeable cmol (+) / kg | Numeric |
| Potasio intercambiable | Exchangeable potassium (K) cmol (+) / kg | Numeric |
| Sodio intercambiable | Sodium (Na) exchangeable cmol (+) / kg | Numeric |
| capacidad de intercambio cationico | cation exchange capacity (CICE) sum of bases cmol (+) / kg | Numeric |
| Conductividad electrica | Electrical conductivity (EC) ratio 2.5: 1.0 dS / m | Numeric |
| Hierro olsen | Iron (Fe) available olsen mg / kg | Numeric |
| Cobre | Copper (Cu) available mg / kg | Nominal |
| Manganeso | Manganese (Mn) available Olsen mg / kg | Nominal |
| Zinc olsen | Zinc (Zn) available Olsen mg / kg | Nominal |
| Boro | Boron (B) available mg / kg | Numeric |
| Hierro | Iron (Fe) available double acid mg / kg | Numeric |
| Cobre doble acido | Copper (Cu) available double acid mg / kg | Numeric |
| Manganeso | Manganese (Mn) available double acid mg / kg | Numeric |
| Zinc | Zinc (Zn) available double acid mg / kg | Nominal |

**Table B. 2.** Metadata for CORPOICA data source.

## B.3 IDEAM Data Source

| **Source name** | bp_ideam | **File format** | CSV |
|---|---|---|---|
| **Dimension** | Biophysical | | |
| **Indicator** | Weather | | |
| **Data provider information** | This is a public institution of technical and scientific support to the National Environmental System, which generates knowledge, produces reliable, consistent and timely information on the state and dynamics of natural resources and the environment, facilitating the definition and adjustment of environmental policies and decision-making by the public and private sectors, and citizens in general. | | |
| **Data source overview** | Request to the IDEAM about the data of the climatic stations installed in the study area. | | |
| **URL** | http://www.ideam.gov.co/solicitud-de-informacion | | |
| **Language** | Spanish | **Time window** | 2012 - 2019 |
| **Temporal scale** | Monthly | **Spatial scale** | Municipality |
| **Number of instances** | 2,042 | **Number of attributes** | 8 |
| **Number of files** | 23 | **Data source size** | 109 KB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Departamento | Territorial units of first level in Colombia. | Nominal |
| Municipio | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| Anio | Year of sampling | Integer |
| Mes | Month of sampling | Integer |
| Temperatura | Temperature in Celsius degrees | Numeric |
| Precipitacion | Precipitation in millimeters per month | Numeric |
| Humedad relativa | Relative humidity (%) | Numeric |
| Radiacion | Radiation (w/m$^2$) | Numeric |

**Table B. 3.** Metadata for IDEAM data source.

## B.4 AVA Data Source

| **Source name** | bp_ava | **File format** | PDF |
|---|---|---|---|
| **Dimension** | Biophysical | | |
| **Indicator** | Climatic Aptitude | | |
| **Data provider information** | Agriculture, Vulnerability, and Adaptation (AVA) is an interinstitutional and multisectoral analysis of vulnerability and adaptation to climate change for the agricultural sector of the upper Cauca River basin impacting adaptation policies. | | |
| **Data source overview** | Percentage of municipal areas with climatic aptitude per crop. | | |
| **URL** | http://cdkn.org/project/agricultura-vulnerabilidad-adaptacion-cuenca-alta-cauca/ | | |
| **Language** | English | **Time window** | 2007 - 2011 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 99 | **Number of attributes** | 8 |
| **Number of files** | 1 | **Data source size** | 5 KB |

| Attributes | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Department | Territorial units of first level in Colombia. | Nominal |
| Municipality | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| CA_Cocoa | Climatic aptitude of cocoa by municipality in the basin. | Numeric |
| CA_Coffee | Climatic aptitude of coffee by municipality in the basin. | Numeric |
| CA_Sugar Cane | Climatic aptitude of sugar cane by municipality in the basin. | Numeric |
| CA_Bean | Climatic aptitude of bean by municipality in the basin. | Numeric |
| CA_Potato | Climatic aptitude of potato by municipality in the basin. | Numeric |
| CA_Banana | Climatic aptitude of banana by municipality in the basin. | Numeric |

**Table B. 4.** Metadata for AVA data source.

# B.5 FINAGRO Data Source

| | | | |
|---|---|---|---|
| **Source name** | ep_finagro | **File format** | XLS |
| **Dimension** | Economic-Productive | | |
| **Indicator** | Agricultural Credits | | |
| **Data provider information** | The Fund for the Financing of the Agricultural Sector, is an entity that promotes the development of the Colombian rural sector, with financing instruments and rural development, that stimulate investment. FINAGRO grants resources to financial entities to encourage loans to productive projects. It also facilitates access to financing instruments and manages the development of an agricultural project. | | |
| **Data source overview** | Total value of disbursements and number of credits granted by FINAGRO and Banco Agrario to different types of producers for each department by municipality during the selected analysis period. | | |
| **URL** | https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=44 | | |
| **Language** | Spanish | **Time window** | 2004 - 2019 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 3,993 | **Number of attributes** | 5 |
| **Number of files** | 16 | **Data source size** | 232 KB |

| Attributes | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Entidad | Financing organization. | Nominal |
| Tipo productor | Type of producer (small, medium, large, strategic alliances, associations, among others). | Nominal |
| Departamento | Territorial units of first level in Colombia. | Nominal |
| Anio | Year of sampling | Integer |
| Valor | Value of credit granted. | Numeric |

**Table B. 5.** Metadata for FINAGRO data source.

# B.6 DANE-SIPSA-P Data Source

| | | | |
|---|---|---|---|
| **Source name** | ep_dane-sipsa-p | **File format** | CSV |
| **Dimension** | Economic-Productive | | |
| **Indicator** | Prices of Agricultural Products | | |

| | | | |
|---|---|---|---|
| **Data provider information** | One of the functions of the National Administrative Department of Statistics (DANE) is to provide basic information for decision-making in all sectors of the economy. SIPSA presents the wholesale prices of agri-food products that are marketed in Colombia. Additionally, information on inputs and factors associated with agricultural production and the level of food supply in cities. | | |
| **Data source overview** | Data about the area sown and harvested for different crops, as well as data on production, yield, costs, prices, gross income, profits, profitability, among others. | | |
| **URL** | https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/sipsa | | |
| **Language** | Spanish | **Time window** | 2004 - 2019 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 130 | **Number of attributes** | 13 |
| **Number of files** | 1 | **Data source size** | 12 KB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| ANIO | Year of sampling | Integer |
| CULTIVOS | Crop or agricultural product. | Nominal |
| SUPERFICIE SEMBRADA | Sown area or the extension of a surface in which a crop is planted. | Numeric |
| SUPERFICIE COSECHADA | Hectares harvested on a certain surface. | Numeric |
| PRODUCCION | Crop production in tons. | Numeric |
| RENDIMIENTO | Crop yield in tons per hectare. | Numeric |
| PRECIO AL PRODUCTOR KG | Price for the producer per kilogram. | Numeric |
| PRECIO AL PRODUCTOR TON | Price for the producer per ton. | Numeric |
| COSTO PRODUCCION | Production cost. | Numeric |
| INGRESO BRUTO PRODUCCION | Gross production income. | Numeric |
| COSTO TOTAL PRODUCCION | Total cost of production. | Numeric |
| UTILIDAD | Utility | Numeric |
| RENTABILIDAD | Cost effectiveness | Numeric |

**Table B. 6.** Metadata for DANE-SIPSA-P data source.

# B.7 Agronet Data Source

| | | | |
|---|---|---|---|
| **Source name** | ep_agronet | **File format** | XLS |
| **Dimension** | Economic-Productive | | |
| **Indicator** | Crop Production | | |
| **Data provider information** | Agronet is the Information and Communication Network of the Agricultural Sector of Colombia, led by the Ministry of Agriculture and Rural Development and supported by the United Nations Organization for Food and Agriculture (FAO). This platform stores information from various official sources linked to the agricultural sector in a centralized way. By having data from official sources, Agronet provides reliable and updated information, according to the periodicity of the investigations. To collect the information to be disclosed, Agronet makes alliances with various institutions at the national, regional, governmental, institutional, and other levels. | | |
| **Data source overview** | Evolution of the harvested area, production, and yield of a crop in a period of time in a selected municipality. | | |

| **URL** | https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=4 | | |
|---|---|---|---|
| **Language** | Spanish | **Time window** | 2007 - 2016 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 3,789 | **Number of attributes** | 8 |
| **Number of files** | 240 | **Data source size** | 119.2 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Departamento | Territorial units of first level in Colombia. | |
| Cultivo | Crop or agricultural product. | |
| Anio | Year of sampling | Integer |
| Municipio | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| Area Sem. (has) | Sown area or the extension of a surface in which a crop is planted. | Numeric |
| Area Cos. (has) | Hectares harvested on a certain surface. | Numeric |
| Producción (Ton) | Crop production in tons. | Numeric |
| Rendimiento (ton/ha) | Crop yield in tons per hectare. | Numeric |

**Table B. 7.** Metadata for Agronet data source.

# B.8 Minagricultura Data Source

| **Source name** | ep_minagricultura | **File format** | XLS |
|---|---|---|---|
| **Dimension** | Economic-Productive | | |
| **Indicator** | Crop Production | | |
| **Data provider information** | The Ministry of Agriculture and Rural Development (Minagricultura) is a Ministry of the Colombian Republic whose main objectives are the formulation, coordination and adoption of policies, plans, programs and projects in the agricultural, fishing and rural development sectors. | | |
| **Data source overview** | Evolution of the harvested area, production, and yield of a crop in a period of time in a selected municipality. | | |
| **URL** | https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=105 | | |
| **Language** | Spanish | **Time window** | 2007 - 2015 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 5,911 | **Number of attributes** | 16 |
| **Number of files** | 1 | **Data source size** | 14 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| COD DEP | Department code | Integer |
| DEPARTAMENTO | Territorial units of first level in Colombia. | Nominal |
| COD MUN | Municipality code | Integer |
| MUNICIPIO | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| GRUPO CULTIVO | General group of the crop. | Nominal |
| SUBGRUPO CULTIVO | Subgroup of the crop | Nominal |
| CULTIVO | Crop | Nominal |
| SISTEMA PRODUCTIVO | Regional disaggregation and/or productive system. | Nominal |
| COD CULTIVO | Crop code | Integer |
| NOMBRE CIENTIFICO | Scientific name of the crop | Nominal |
| PERIODO | Date of measurement | Nominal |
| AREA SEMBRADA (has) | Sown area or the extension of a surface in which a crop is planted. | Numeric |
| AREA COSECHADA (has) | Hectares harvested on a certain surface. | Numeric |

| | | |
|---|---|---|
| PRODUCCION (ton) | Crop production in tons. | Numeric |
| RENDIMIENTO (ton/ha) | Crop yield in tons per hectare. | Numeric |
| ESTADO FISICO PRODUCCION | Physical state of production (green paddy, dry grain, fresh fruit, among others) | Nominal |

**Table B. 8.** Metadata for Minagricultura data source.

# B.9 Agronet-P Data Source

| | | | |
|---|---|---|---|
| **Source name** | ep_agronet_p | **File format** | XLS |
| **Dimension** | Economic-Productive | | |
| **Indicator** | Imports and Exports | | |
| **Data provider information** | Agronet is the Information and Communication Network of the Agricultural Sector of Colombia, led by the Ministry of Agriculture and Rural Development and supported by the United Nations Organization for Food and Agriculture (FAO). This platform stores information from various official sources linked to the agricultural sector in a centralized way. By having data from official sources, Agronet provides reliable and updated information, according to the periodicity of the investigations. To collect the information to be disclosed, Agronet makes alliances with various institutions at the national, regional, governmental, institutional, and other levels. | | |
| **Data source overview** | Prices of the different agricultural products and their respective variations in the international stock market. | | |
| **URL** | https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=4 | | |
| **Language** | Spanish | **Time window** | 2007 - 2016 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 3,789 | **Number of attributes** | 8 |
| **Number of files** | 240 | **Data source size** | 119.2 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Fecha | Export date. | Date |
| Producto | Crop or agricultural product. | Nominal |
| Precio | Export price. | Numeric |
| Unidad | Export unit, e.g., US/Ton. | Nominal |
| Variacion | Price variation. | Numeric |
| Fuente | Product quotation market. | Nominal |

**Table B. 9.** Metadata for Agronet-P data source.

# B.10 DANE-SIPSA Data Source

| | | | |
|---|---|---|---|
| **Source name** | ep_dane_sipsa | **File format** | XLS |
| **Dimension** | Economic-Productive | | |
| **Indicator** | Prices of Agricultural Products | | |
| **Data provider information** | One of the functions of the National Administrative Department of Statistics (DANE) is to provide basic information for decision-making in all sectors of the economy. SIPSA presents the wholesale prices of agri-food products that are marketed in Colombia. Additionally, information | | |

| | | | |
|---|---|---|---|
| | on inputs and factors associated with agricultural production and the level of food supply in cities. | | |
| **Data source overview** | Historical series of the wholesale quotations of the main agricultural products that make up the food basket. | | |
| **URL** | https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/sipsa | | |
| **Language** | Spanish | **Time window** | 2013 - 2017 |
| **Temporal scale** | Monthly | **Spatial scale** | Municipality |
| **Number of instances** | 170,118 | **Number of attributes** | 5 |
| **Number of files** | 1 | **Data source size** | 11.7 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| Fecha | Data collection date. | Date |
| Grupo | Category of agricultural product. | Nominal |
| Producto | Agricultural product. | Nominal |
| Fuente | Municipality where the product is marketed. | Nominal |
| Precio | Price of the agricultural product. | Numeric |

**Table B. 10.** Metadata for DANE-SIPSA data source.

# B.11 DNP-AIB Data Source

| **Source name** | pi_dnp_aib | **File format** | XLS |
|---|---|---|---|
| **Dimension** | Political-Institutional | | |
| **Indicator** | Agricultural Investment Budget | | |
| **Data provider information** | The National Planning Department is an entity that defines the implementation of a strategic vision of the country in the social, economic and environmental fields. | | |
| **Data source overview** | Colombian climate finance strategy and methodology for the identification of budget executions in initiatives compatible with climate change. | | |
| **URL** | https://mrvapp.dnp.gov.co/General/InfografiaGeneral/ | | |
| **Language** | Spanish | **Time window** | 2005 - 2013 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 19,286 | **Number of attributes** | 8 |
| **Number of files** | 1 | **Data source size** | 8.3 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| ANIO | Year of budget execution. | Integer |
| TIPO_ENTIDAD | Type of territorial entity. | Nominal |
| CODIGO_ENTIDAD | DANE code for territorial entities. | Nominal |
| ENTIDAD | Name of the national entity, CAR, department or municipality. | Nominal |
| CUENTA-PROYECTO | Name of the project or account in the original source. | Nominal |
| SECTOR | Sector name in the database of the original source. | Nominal |
| OBJETIVO | Strategy against climate change (adaptation, mitigation, mixed). | Nominal |
| Valor (Miles de $) | Investment value. | Numeric |

**Table B. 11.** Metadata for DNP-AIB data source.

## B.12 DNP-FI Data Source

| **Source name** | pi_dnp_fi | **File format** | XLS |
|---|---|---|---|
| **Dimension** | Political-Institutional | | |
| **Indicator** | Forest Incentives | | |
| **Data provider information** | The National Planning Department is an entity that defines the implementation of a strategic vision of the country in the social, economic and environmental fields. | | |
| **Data source overview** | The forest incentive certificate is a recognition of the National Government through the Ministry of Agriculture and Rural Development to the positive externalities of reforestation in its commercial component. This recognition is made by the Ministry through FINAGRO by delegation. | | |
| **URL** | https://www.finagro.com.co/productos-y-servicios/incentivo-forestal | | |
| **Language** | Spanish | **Time window** | 1995 - 2014 |
| **Temporal scale** | Annual | **Spatial scale** | Department |
| **Number of instances** | 494 | **Number of attributes** | 5 |
| **Number of files** | 1 | **Data source size** | 2.2 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| ANIO | Start year forestry project. | Integer |
| DEPARTAMENTO | Territorial units of first level in Colombia. | Nominal |
| NUMERO PROYECTOS | Number of forest projects per year. | Nominal |
| HECTAREAS REFORESTADAS | Number of hectares reforested per year. | Nominal |
| VALOR ESTABLECIMIENTO | Name of the project or account in the original source. | Numeric |

**Table B. 12.** Metadata for DNP-FI data source.

## B.13 DNP-LA Data Source

| **Source name** | pi_dnp_la | **File format** | XLS |
|---|---|---|---|
| **Dimension** | Political-Institutional | | |
| **Indicator** | Land Allocation | | |
| **Data provider information** | The National Planning Department is an entity that defines the implementation of a strategic vision of the country in the social, economic and environmental fields. | | |
| **Data source overview** | Progressive access to the land ownership of the agrarian workers and the use of the nation's uncultivated lands, giving preference to the allocation of low-income peasants and establishing peasant reserve zones for the promotion of rural property. | | |
| **URL** | https://www.dnp.gov.co/DNP/organigrama/subdireccion-sectorial/Paginas/direccion-de-desarrollo-rural-sostenible.aspx | | |
| **Language** | Spanish | **Time window** | 2010 - 2014 |
| **Temporal scale** | Annual | **Spatial scale** | Department |
| **Number of instances** | 160 | **Number of attributes** | 4 |
| **Number of files** | 1 | **Data source size** | 2.2 MB |

| **Attributes** |
|---|

| Name | Short Description | Type |
|---|---|---|
| ANIO | Year of land allocation. | Integer |
| DEPARTAMENTO | Territorial units of first level in Colombia. | Nominal |
| HECTAREAS | Number of hectares allocated per year. | Nominal |
| FAMILIAS | Number of beneficiary families. | Numeric |

**Table B. 13.** Metadata for DNP-LA data source.

# B.14 DNP-PA Data Source

| **Source name** | pi_dnp_pa | **File format** | XLS |
|---|---|---|---|
| **Dimension** | Political-Institutional | | |
| **Indicator** | Productive Alliances | | |
| **Data provider information** | The National Planning Department is an entity that defines the implementation of a strategic vision of the country in the social, economic and environmental fields. | | |
| **Data source overview** | The data correspond to the access of small rural producers to factors of production such as land and labor, enhancing their use and complementing investment capacity through the direct support of profitable productive initiatives with a contribution from the project, a resource called Modular Incentive. | | |
| **URL** | https://www.minagricultura.gov.co/tramites-servicios/desarrollo-rural/Paginas/v1/Proyecto-apoyo-a-alianzas-productivas-PAAP.aspx | | |
| **Language** | Spanish | **Time window** | 2002 - 2013 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 783 | **Number of attributes** | 9 |
| **Number of files** | 1 | **Data source size** | 2.3 MB |

**Attributes**

| Name | Short Description | Type |
|---|---|---|
| Departamento | Territorial units of first level in Colombia. | Nominal |
| Año | Start year of the productive alliance. | Integer |
| Producto principal de la alianza | Main product of the alliance. | Nominal |
| Nombre abreviado de la alianza | Abbreviated name of the alliance. | Nominal |
| Municipios | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| Valor de la alianza ($ millones) | Value of the alliance | Numeric |
| Valor IM de la alianza ($ millones) | Value of the Modular Incentive (IM) of the alliance | Numeric |
| No. beneficiarios aprobados alianza | Number of beneficiaries approved in the alliance. | Integer |
| Número de hectareas en la alianza | Number of hectares in the alliance. | Numeric |

**Table B. 14.** Metadata for DNP-PA data source.

# B.15 DANE-HH Data Source

| **Source name** | sc_dane_hh | **File format** | CSV |
|---|---|---|---|
| **Dimension** | Sociocultural | | |
| **Indicator** | Livelihood | | |

| **Data provider information** | The National Administrative Department of Statistics (DANE for its acronym in Spanish) applies different censuses to obtain data from several aspects of the population in Colombia, among them, the national agricultural census. | | |
|---|---|---|---|
| **Data source overview** | The data refers to the number of households within each dwelling and the perception of subjective poverty and displacement directed at the head of the household or his spouse. | | |
| **URL** | http://microdatos.dane.gov.co/index.php/catalog/513/data_dictionary#page=F11&tab=data-dictionary | | |
| **Language** | Spanish | **Time window** | 2014 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 128,332 | **Number of attributes** | 21 |
| **Number of files** | 1 | **Data source size** | 8.9 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| TIPO_REG | Type of region | Nominal |
| PAIS | Country | Nominal |
| P_DEPTO | Territorial units of first level in Colombia. | Nominal |
| P_MUNIC | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| UC_UO | Coverage unit and observation unit. | Nominal |
| ENCUESTA | Survey number | Nominal |
| COD_VEREDA | Village code | Nominal |
| ITER_HG | Household type | Nominal |
| P_S15P165A | House order number | Nominal |
| P_S15P165B | Household order number | Nominal |
| P_S15P177 | Do you consider yourself poor? | Nominal |
| P_S15P178 | You think that the current standard of living of your home, compared to the one you were 5 years ago, is: Better, Equal, Worse, No Information. | Nominal |
| P_S15P179_SP1 | You or a member of your household has experienced any of the following situations: Forced displacement. | Nominal |
| P_S15P179A | Year of forced displacement | Nominal |
| P_S15P179_SP2 | You or a member of your household has experienced any of the following situations: Land dispossession. | Nominal |
| P_S15P179B | Year of land dispossession | Nominal |
| P_S15P179_SP3 | You or a member of your household has experienced any of the following situations: Forced abandonment of land | Nominal |
| P_S15P179C | Year of forced abandonment of land | Nominal |
| P_S15P179_SP4 | You or a member of your household has experienced any of the following situations: None. | Nominal |
| P_S15P180 | You have been able to return to your property. | Nominal |
| TOT_PROD_HOGAR | Total productive people in the home. | Nominal |

**Table B. 15.** Metadata for DANE-HH data source.

# B.16 DANE-H Data Source

| **Source name** | sc_dane_h | **File format** | CSV |
|---|---|---|---|
| **Dimension** | Sociocultural | | |
| **Indicator** | Human Development | | |

| **Data provider information** | The National Administrative Department of Statistics (DANE for its acronym in Spanish) applies different censuses to obtain data from several aspects of the population in Colombia, among them, the national agricultural census. | | |
|---|---|---|---|
| **Data source overview** | The data source contains the number of dwellings, the occupation status of the same, the availability of public services and the condition of the predominant materials. | | |
| **URL** | http://microdatos.dane.gov.co/index.php/catalog/513/datafile/F9 | | |
| **Language** | Spanish | **Time window** | 2014 |
| **Temporal scale** | Annual | **Spatial scale** | Municipality |
| **Number of instances** | 462,649 | **Number of attributes** | 16 |
| **Number of files** | 1 | **Data source size** | 8.9 MB |

| **Attributes** | | |
|---|---|---|
| **Name** | **Short Description** | **Type** |
| TIPO_REG | Type of region | Nominal |
| PAIS | Country | Nominal |
| P_DEPTO | Territorial units of first level in Colombia. | Nominal |
| P_MUNIC | Territorial units of second level in Colombia. A department consists of municipalities. | Nominal |
| UC_UO | Coverage unit and observation unit. | Nominal |
| ENCUESTA | Survey number | Nominal |
| COD_VEREDA | Village code | Nominal |
| P_S15P159 | House order number | Nominal |
| P_S15P160 | Occupation of housing? | Nominal |
| P_S15P161 | How many groups of people cook their food separately and usually reside in this house? | Nominal |
| P_S15P162 | Predominant material of exterior walls. | Nominal |
| P_S15P163 | Predominant material of the floors. | Nominal |
| P_S15P164_SP1 | Which of the following public, private or communal services does the dwelling have? Electric power | Nominal |
| P_S15P164_SP2 | Which of the following public, private or communal services does the dwelling have? Sewerage | Nominal |
| P_S15P164_SP3 | Which of the following public, private or communal services does the dwelling have? Aqueduct | Nominal |
| P_S15P164_SP4 | Which of the following public, private or communal services does the dwelling have? None | Nominal |

**Table B. 16.** Metadata for DANE-H data source.

# Appendix C

# Results of Data Sources Preprocessing

This appendix presents the data cleaning steps for all 16 initial data sources. We show several steps corresponding to check missing values, remove duplicate instances, detect outliers, reduce dimensionality, and identify importance of variables.

## C.1. SIVICAP Data Source

### *Check Missing Values*

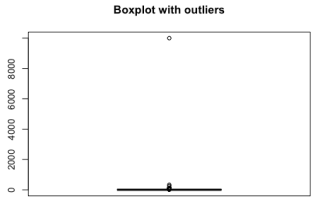
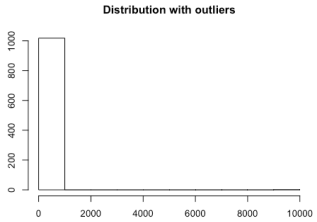
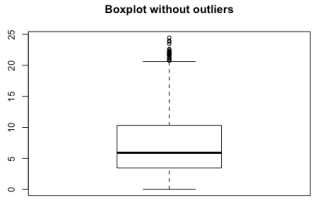
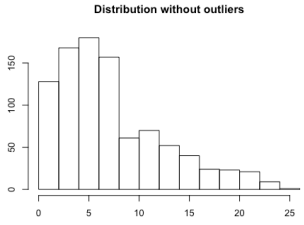
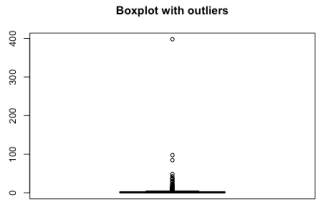
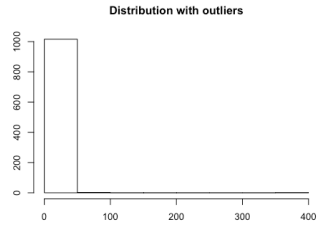
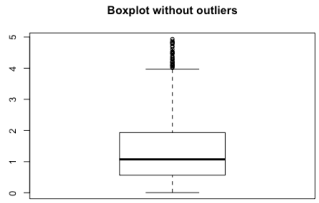
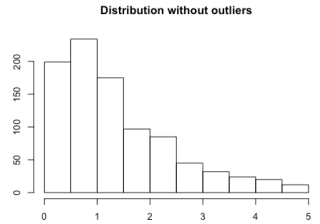
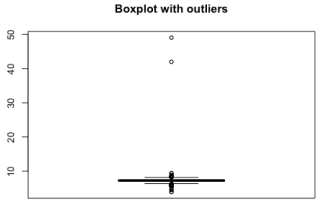
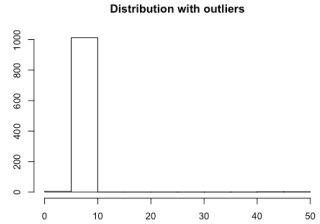
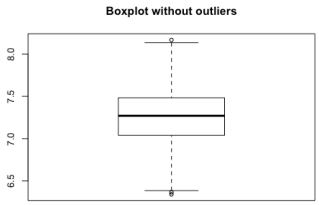
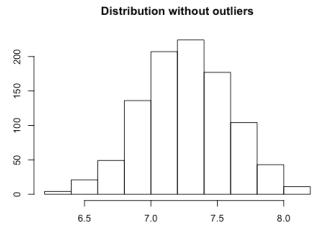
```
   ANIO           DEPARTAMENTO       MUNICIPO   TOTAL.MUESTRAS    RESULTADO.COLOR.APARENTE RESULTADO.TURBIEDAD
Min.   :2015   Antioquia   :123   Buenavista:  4   Min.   :   1.00   Min.   :    0.000        Min.   :  0.0047
1st Qu.:2015   Boyacá      :123   La Unión  :  4   1st Qu.:  13.00   1st Qu.:    3.525        1st Qu.:  0.5878
Median :2015   Cundinamarca:116   Villanueva:  4   Median :  26.00   Median :    6.111        Median :  1.1724
Mean   :2015   Santander   : 87   Albania   :  3   Mean   :  45.84   Mean   :   21.589        Mean   :  2.8763
3rd Qu.:2015   Nariño      : 64   Argelia   :  3   3rd Qu.:  49.50   3rd Qu.:   12.547        3rd Qu.:  2.3621
Max.   :2015   Bolívar     : 45   Bolívar   :  3   Max.   :2546.00   Max.   :10000.000        Max.   :397.8059
               (Other)     :461   (Other)   :998                     NA's   :61               NA's   :21

 RESULTADO.PH    RESULTADO.CLORO.RESIDUAL.LIBRE RESULTADO.ALCALINIDAD.TOTAL RESULTADO.CALCIO  RESULTADO.FOSFATOS
Min.   : 3.900   Min.   : 0.0000                Min.   :  1.445            Min.   :  0.000   Min.   : 0.0000
1st Qu.: 7.021   1st Qu.: 0.4783                1st Qu.: 16.585            1st Qu.:  7.564   1st Qu.: 0.0386
Median : 7.274   Median : 0.8542                Median : 34.560            Median : 13.645   Median : 0.1000
Mean   : 7.320   Mean   : 0.9873                Mean   : 47.488            Mean   : 17.419   Mean   : 0.3580
3rd Qu.: 7.499   3rd Qu.: 1.1512                3rd Qu.: 63.611            3rd Qu.: 21.072   3rd Qu.: 0.1917
Max.   :49.055   Max.   :83.0000                Max.   :334.650            Max.   :207.000   Max.   :69.0000
NA's   :18       NA's   :83                     NA's   :96                 NA's   :524       NA's   :458

RESULTADO.MANGANESO RESULTADO.MOLIBDENO RESULTADO.MAGNESIO RESULTADO.ZINC    RESULTADO.DUREZA.TOTAL RESULTADO.SULFATOS
Min.   : 0.0000     Min.   :0.0000     Min.   :  0.000    Min.   :  0.000   Min.   :  2.75         Min.   :  0.000
1st Qu.: 0.0000     1st Qu.:0.0000     1st Qu.:  2.028    1st Qu.:  0.000   1st Qu.: 26.20         1st Qu.:  4.944
Median : 0.0000     Median :0.0050     Median :  3.909    Median :  0.000   Median : 45.60         Median : 10.678
Mean   : 0.4175     Mean   :0.1742     Mean   :  8.930    Mean   :  1.938   Mean   : 61.47         Mean   : 19.674
3rd Qu.: 0.0360     3rd Qu.:0.0200     3rd Qu.:  9.500    3rd Qu.:  0.000   3rd Qu.: 74.36         3rd Qu.: 24.124
Max.   :32.2000     Max.   :9.9000     Max.   :526.393    Max.   :177.320   Max.   :623.86         Max.   :342.100
NA's   :854         NA's   :908        NA's   :498        NA's   :887       NA's   :76             NA's   :317

RESULTADO.HIERRO.TOTAL RESULTADO.CLORUROS RESULTADO.NITRATOS RESULTADO.NITRITOS RESULTADO.ALUMINIO RESULTADO.FLORUROS
Min.   : 0.00000       Min.   :  0.3333   Min.   :0.000      Min.   :0.0000     Min.   :  0.0000   Min.   : 0.0000
1st Qu.: 0.04800       1st Qu.:  3.4750   1st Qu.:0.550      1st Qu.:0.0000     1st Qu.:  0.0309   1st Qu.: 0.0000
Median : 0.08612       Median :  5.9000   Median :1.155      Median :0.0073     Median :  0.0590   Median : 0.0390
Mean   : 0.17539       Mean   : 12.2777   Mean   :1.410      Mean   :0.0390     Mean   :  2.6968   Mean   : 0.2029
3rd Qu.: 0.15650       3rd Qu.: 10.7000   3rd Qu.:1.857      3rd Qu.:0.0174     3rd Qu.:  0.1232   3rd Qu.: 0.1034
Max.   :17.14750       Max.   :616.3275   Max.   :6.914      Max.   :7.7897     Max.   :806.6247   Max.   :31.6050
NA's   :245            NA's   :190        NA's   :755        NA's   :371        NA's   :612        NA's   :559

RESULTADO.COT    RESULTADO.COLIFORMES.TOTALES RESULTADO.E.COLI   RESULTADO.ANTIMONIO RESULTADO.ARSÉNICO RESULTADO.BARIO
Min.   :  0.000  Min.   :    0.000            Min.   :   0.000   Min.   :0e+00       Min.   : 0.0000    Min.   :0.000
1st Qu.:  0.500  1st Qu.:    0.365            1st Qu.:   0.000   1st Qu.:0e+00       1st Qu.: 0.0000    1st Qu.:0.000
Median :  1.256  Median :   22.874            Median :   0.549   Median :0e+00       Median : 0.0000    Median :0.000
Mean   :  4.089  Mean   :  457.123            Mean   :  47.925   Mean   :2e-04       Mean   : 0.1319    Mean   :0.008
3rd Qu.:  2.789  3rd Qu.:  452.748            3rd Qu.:  24.425   3rd Qu.:0e+00       3rd Qu.: 0.0000    3rd Qu.:0.000
Max.   :175.143  Max.   :18025.182            Max.   :7176.429   Max.   :1e-02       Max.   :14.5942    Max.   :0.200
```

```
NA's   :823       NA's   :3                    NA's   :3          NA's   :907         NA's   :908         NA's   :991

 RESULTADO.CADMIO RESULTADO.CIANURO.LIBRE.Y.DIASOCIABLE RESULTADO.COBRE  RESULTADO.Cromo.total RESULTADO.MERCURIO RESULTADO.NIQUEL
Min.   :0e+00    Min.   :0.0000                        Min.   :0.0000   Min.   :0.0000        Min.   :0.0000     Min.   :0.0000
1st Qu.:0e+00    1st Qu.:0.0000                        1st Qu.:0.0000   1st Qu.:0.0000        1st Qu.:0.0000     1st Qu.:0.0000
Median :0e+00    Median :0.0000                        Median :0.0000   Median :0.0000        Median :0.0000     Median :0.0000
Mean   :1e-04    Mean   :0.0044                        Mean   :0.0151   Mean   :0.0038        Mean   :0.0001     Mean   :0.0034
3rd Qu.:0e+00    3rd Qu.:0.0000                        3rd Qu.:0.0300   3rd Qu.:0.0000        3rd Qu.:0.0000     3rd Qu.:0.0050
Max.   :5e-03    Max.   :0.0500                        Max.   :0.1100   Max.   :0.0300        Max.   :0.0026     Max.   :0.0223
NA's   :891      NA's   :915                           NA's   :867      NA's   :896           NA's   :880        NA's   :959

 RESULTADO.PLOMO  RESULTADO.SELENIO RESULTADO.TRIHALOMETANOS.TOTALES RESULTADO.HIDROCARBUROS.AROMÁTICOS.POLICICLICOS
Min.   :0.0000   Min.   :0e+00     Min.   :0.0000                   Min.   :0e+00
1st Qu.:0.0000   1st Qu.:0e+00     1st Qu.:0.0000                   1st Qu.:0e+00
Median :0.0000   Median :0e+00     Median :0.0000                   Median :0e+00
Mean   :0.0002   Mean   :1e-04     Mean   :0.0061                   Mean   :1e-04
3rd Qu.:0.0000   3rd Qu.:0e+00     3rd Qu.:0.0100                   3rd Qu.:0e+00
Max.   :0.0074   Max.   :1e-02     Max.   :0.0800                   Max.   :2e-03
NA's   :965      NA's   :912       NA's   :926                      NA's   :950

 RESULTADO.GIARDIA RESULTADO.CRYPTOSPORIDIUM RESULTADO.PLAGUICIDAS.TOTALES RESULTADO.ORGANOFOSFORADOS.Y.CARBAMATOS
Min.   :0.0000    Min.   :0.0000            Min.   :0                     Min.   :0
1st Qu.:0.0000    1st Qu.:0.0000            1st Qu.:0                     1st Qu.:0
Median :0.0000    Median :0.0000            Median :0                     Median :0
Mean   :0.0884    Mean   :0.1061            Mean   :0                     Mean   :0
3rd Qu.:0.0000    3rd Qu.:0.0000            3rd Qu.:0                     3rd Qu.:0
Max.   :2.0000    Max.   :2.0000            Max.   :0                     Max.   :0
NA's   :953       NA's   :953               NA's   :970                   NA's   :1015

 RESULTADO.MESÓFILOS IRCA.PROMEDIO     IRCA.BASE.PROMEDIO   NIVEL.DE.RIESGO.PROMEDIO
Min.   :    0.0     Min.   :  0.000   Min.   :  0.000    ALTO                   :282
1st Qu.:    0.0     1st Qu.:  4.466   1st Qu.:  4.441    BAJO                   :177
Median :   65.0     Median : 17.271   Median : 17.271    INVIABLE SANITARIAMENTE: 17
Mean   : 1625.0     Mean   : 23.795   Mean   : 23.728    MEDIO                  :262
3rd Qu.:  453.9     3rd Qu.: 39.962   3rd Qu.: 39.563    SIN RIESGO             :281
Max.   :54260.8     Max.   :100.000   Max.   :100.000
NA's   :882
```

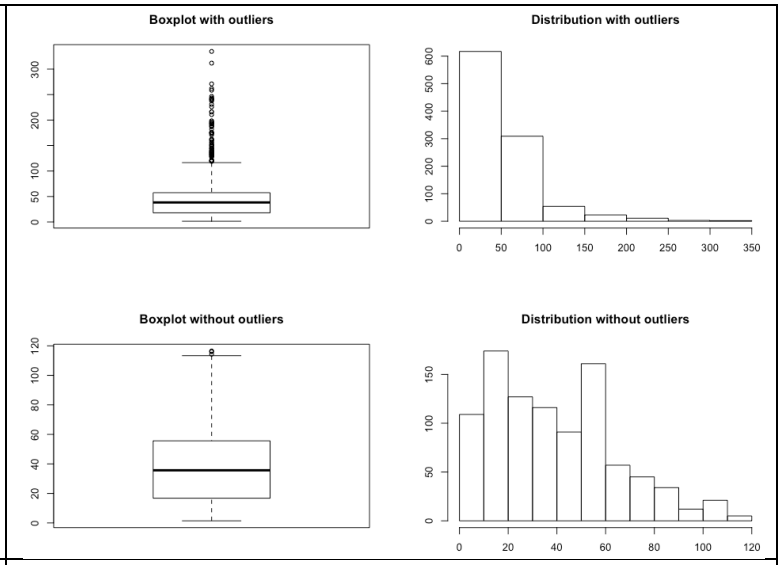
Those attributes with a high percentage of missing values (NA's $>$ 70% or NA's $>$ 713) have a greater probability of being initially discarded. In the previous summary, those attributes with this characteristic are highlighted. In contrast, attributes with low percentages of missing values are selected to continue with the cleaning process and these are listed below.

```
ANIO
DEPARTAMENTO
MUNICIPO
TOTAL.MUESTRAS
RESULTADO.COLOR.APARENTE
RESULTADO.TURBIEDAD
RESULTADO.PH
RESULTADO.CLORO.RESIDUAL.LIBRE
RESULTADO.ALCALINIDAD.TOTAL
RESULTADO.CALCIO
RESULTADO.FOSFATOS
RESULTADO.MAGNESIO
RESULTADO.DUREZA.TOTAL
RESULTADO.SULFATOS
RESULTADO.HIERRO.TOTAL
RESULTADO.CLORUROS
RESULTADO.NITRITOS
RESULTADO.ALUMINIO
RESULTADO.FLORUROS
RESULTADO.COLIFORMES.TOTALES
RESULTADO.E.COLI
IRCA.PROMEDIO
IRCA.BASE.PROMEDIO
NIVEL.DE.RIESGO.PROMEDIO
```

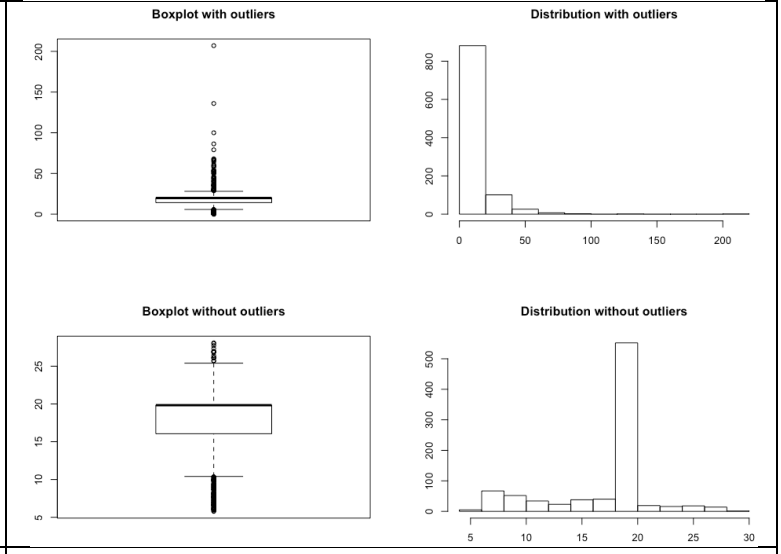
## *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **RESULTADO.COLOR.APARENTE**<br>Outliers identified: 85<br>Proportion (%) of outliers: 9.1<br>Mean of the outliers: 168.87<br>Mean without removing outliers: 20.71<br>Mean if we remove outliers: 7.23 |     |
| **RESULTADO.TURBIEDAD**<br>Outliers identified: 96<br>Proportion (%) of outliers: 10.4<br>Mean of the outliers: 17.17<br>Mean without removing outliers: 2.87<br>Mean if we remove outliers: 1.38 |     |
| **RESULTADO.PH**<br>Outliers identified: 43<br>Proportion (%) of outliers: 4.4<br>Mean of the outliers: 8.51<br>Mean without removing outliers: 7.32<br>Mean if we remove outliers: 7.26 |     |

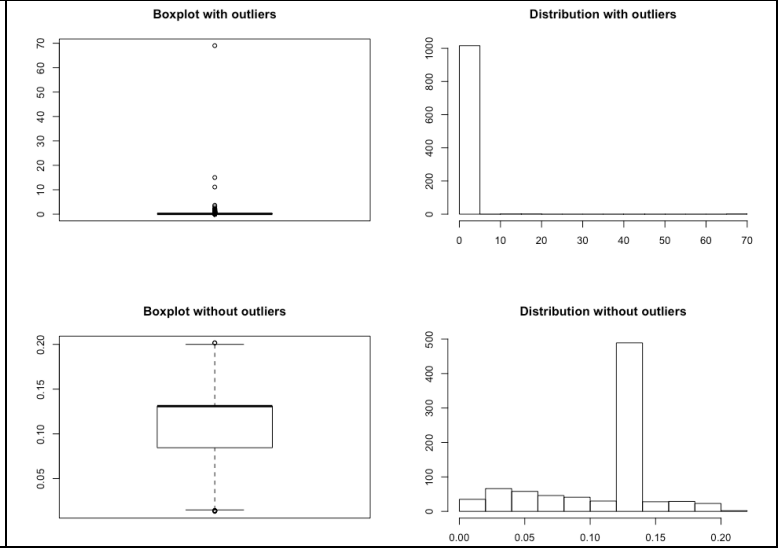
**RESULTADO.ALCALINIDAD.TOTAL**
Outliers identified: 67
Proportion (%) of outliers: 7
Mean of the outliers: 175.95
Mean without removing outliers: 48.26
Mean if we remove outliers: 39.27



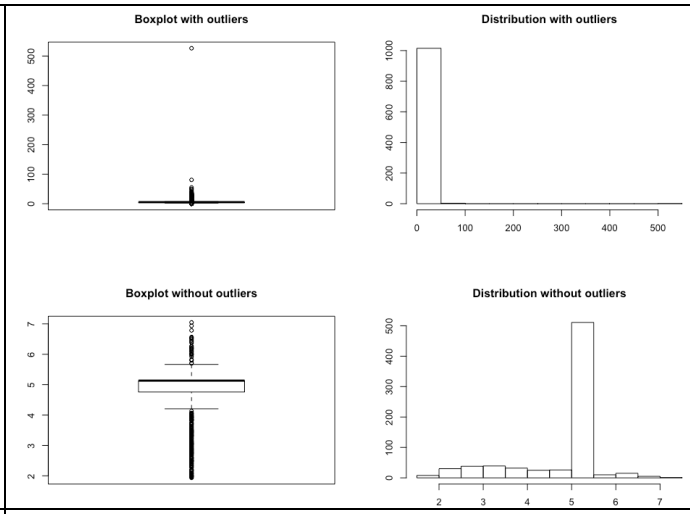

**RESULTADO.CALCIO**
Outliers identified: 140
Proportion (%) of outliers: 15.9
Mean of the outliers: 25.65
Mean without removing outliers: 18.66
Mean if we remove outliers: 17.55



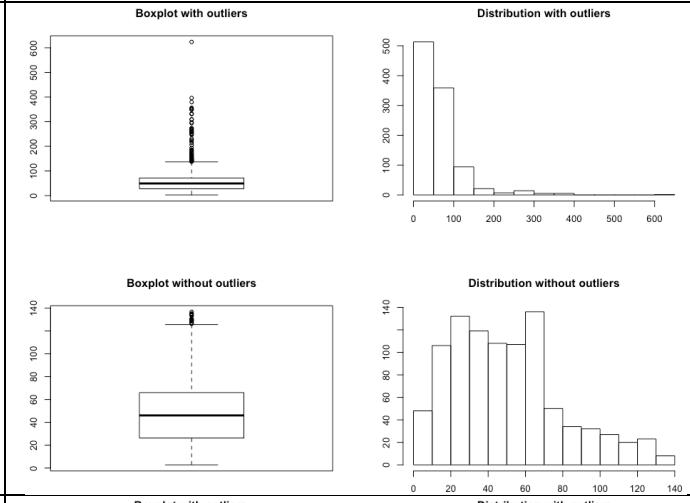

**RESULTADO.FOSFATOS**
Outliers identified: 172
Proportion (%) of outliers: 20.3
Mean of the outliers: 0.97
Mean without removing outliers: 0.26
Mean if we remove outliers: 0.11



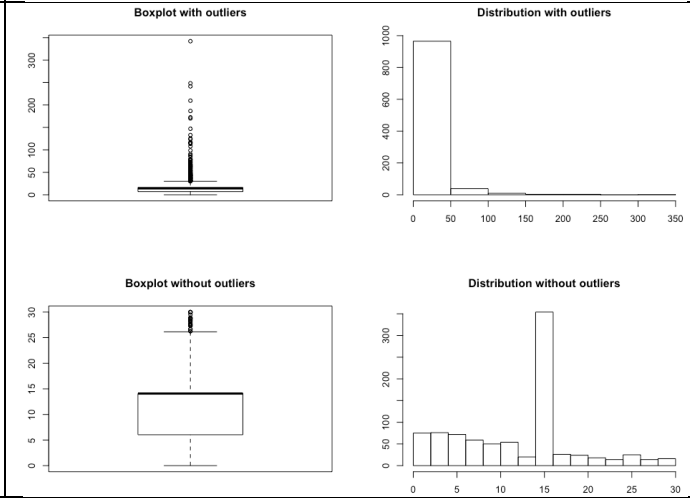
**RESULTADO.MAGNESIO**
Outliers identified: 279
Proportion (%) of outliers: 37.7
Mean of the outliers: 13.34
Mean without removing outliers: 7.07
Mean if we remove outliers: 4.71



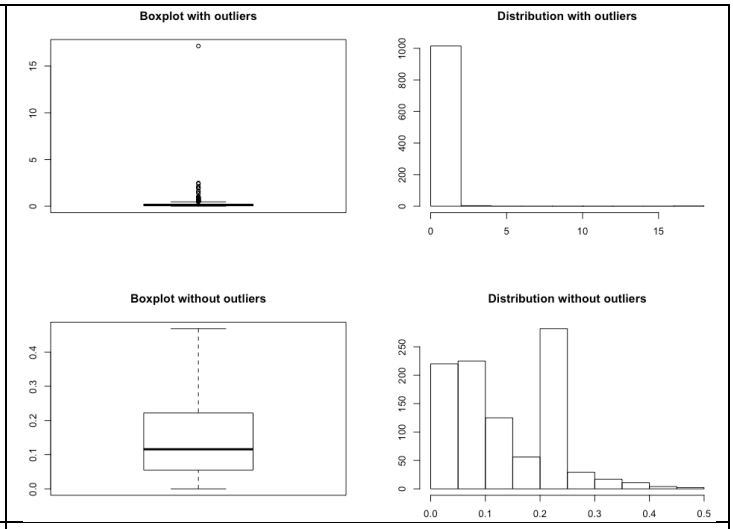

**RESULTADO.DUREZA.TOTAL**
Outliers identified: 69
Proportion (%) of outliers: 7.3
Mean of the outliers: 221.19
Mean without removing outliers: 61.81
Mean if we remove outliers: 50.23



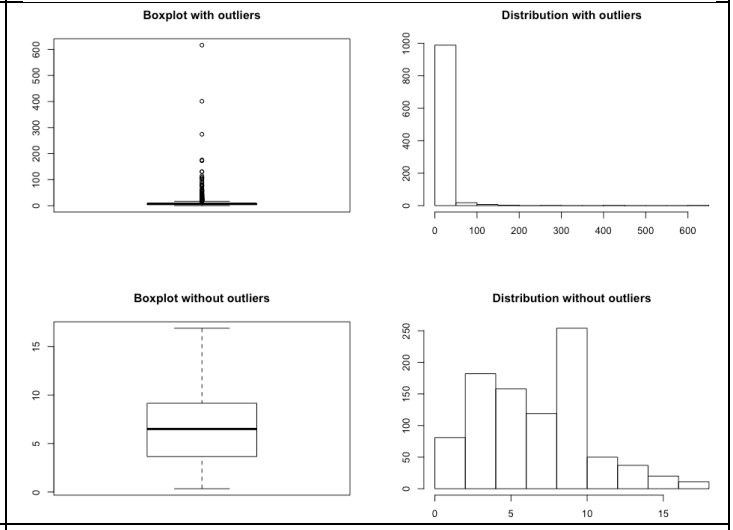

**RESULTADO.SULFATOS**
Outliers identified: 122
Proportion (%) of outliers: 13.6
Mean of the outliers: 63.94
Mean without removing outliers: 17.93
Mean if we remove outliers: 11.68



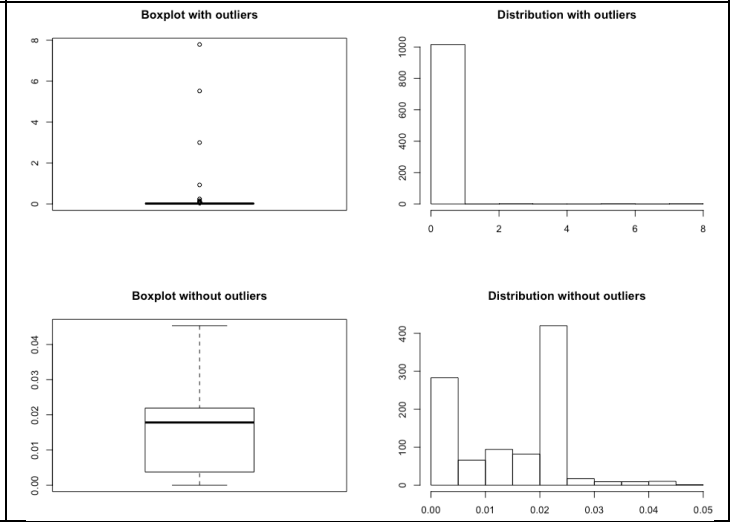
**RESULTADO.HIERRO.TOTAL**
Outliers identified: 48
Proportion (%) of outliers: 4.9
Mean of the outliers: 1.24
Mean without removing outliers: 0.19
Mean if we remove outliers: 0.13



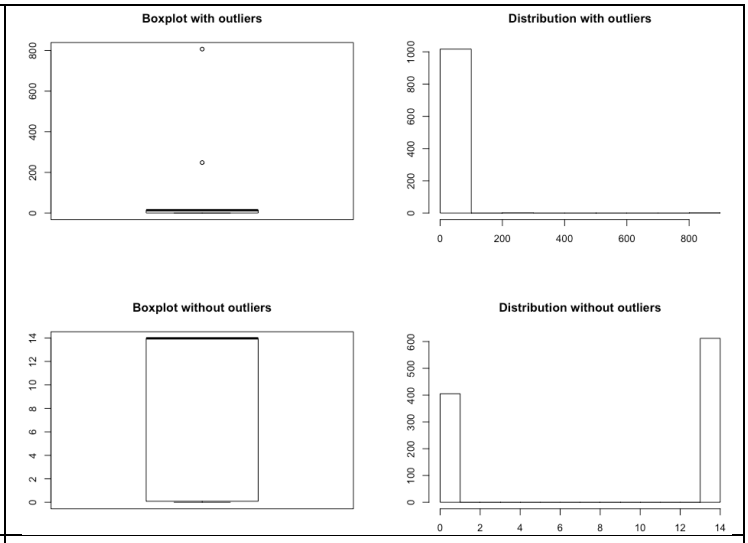

**RESULTADO.CLORUROS**
Outliers identified: 107
Proportion (%) of outliers: 11.7
Mean of the outliers: 54.32
Mean without removing outliers: 11.7
Mean if we remove outliers: 6.7



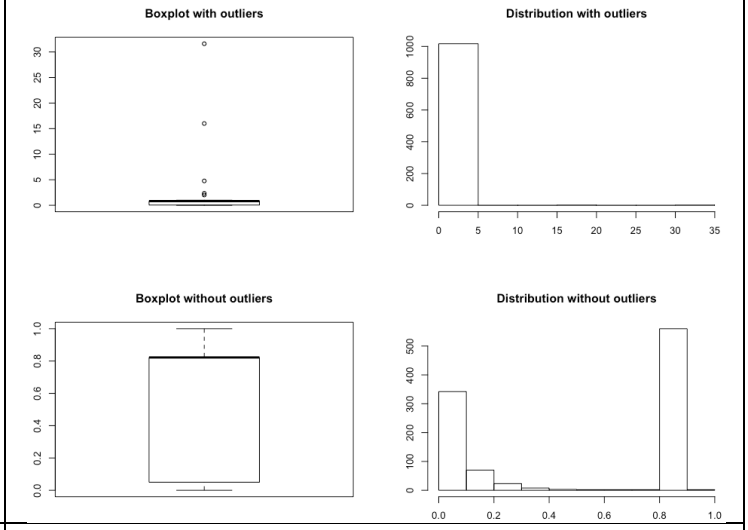

**RESULTADO.NITRITOS**
Outliers identified: 28
Proportion (%) of outliers: 2.8
Mean of the outliers: 0.69
Mean without removing outliers: 0.03
Mean if we remove outliers: 0.01



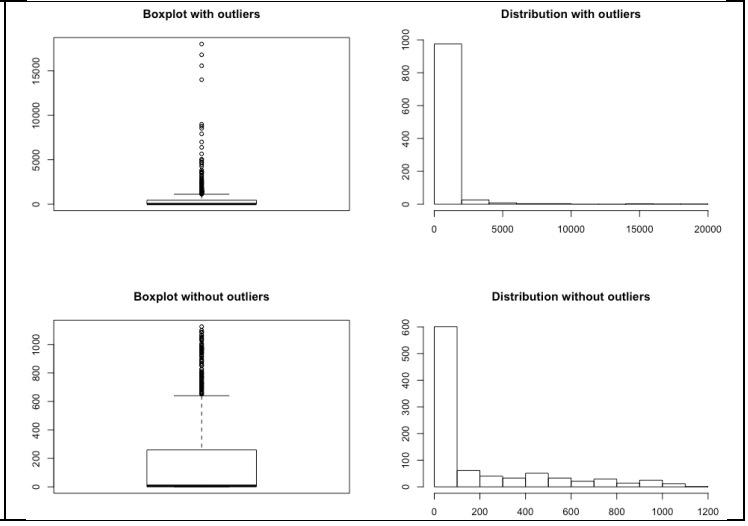
**RESULTADO.ALUMINIO**
Outliers identified: 2
Proportion (%) of outliers: 0.2
Mean of the outliers: 527.45
Mean without removing outliers: 9.47
Mean if we remove outliers: 8.45




**RESULTADO.FLORUROS**
Outliers identified: 5
Proportion (%) of outliers: 0.5
Mean of the outliers: 11.35
Mean without removing outliers: 0.54
Mean if we remove outliers: 0.49




**RESULTADO.COLIFORMES.TOTALES**
Outliers identified: 98
Proportion (%) of outliers: 10.6
Mean of the outliers: 3085.21
Mean without removing outliers: 456.97
Mean if we remove outliers: 177.31



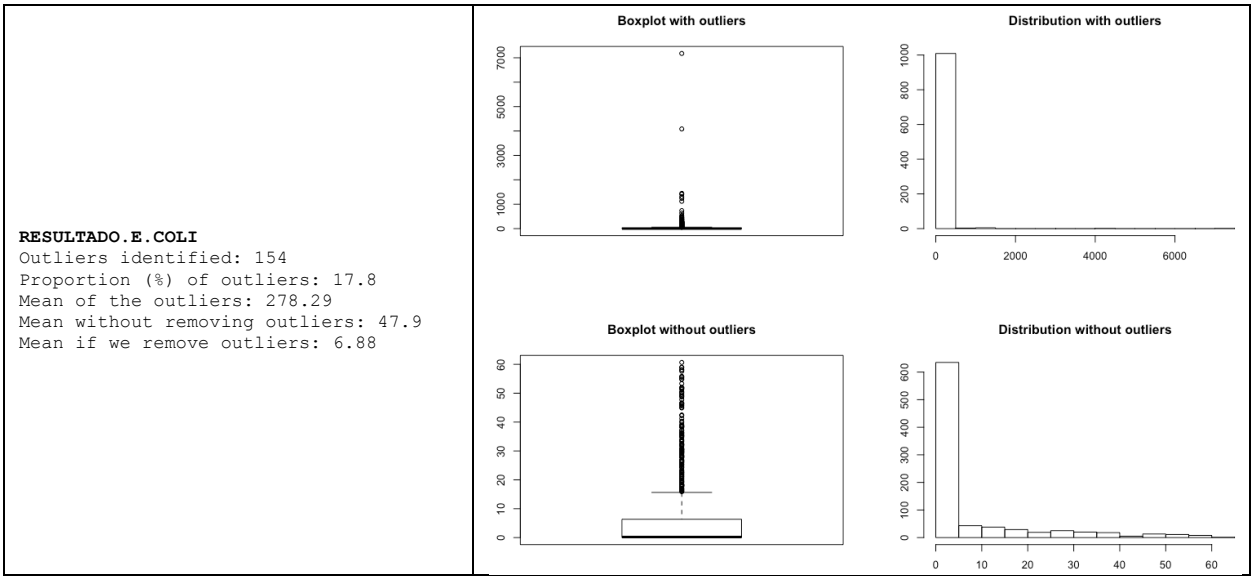



**Table C. 1.** Results of the outlier detection process for each of the SIVICAP dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. This method also adds an X attribute which represents an instance identifier. Later, we obtained the following summary of measures.

```
       X               ANIO           DEPARTAMENTO       MUNICIPO   TOTAL.MUESTRAS    RESULTADO.COLOR.APARENTE
 Min.   :   1.0   Min.   :2015   Antioquia   :123   Buenavista:  4   Min.   :   1.00   Min.   : 0.000
 1st Qu.: 255.5   1st Qu.:2015   Boyacá      :123   La Unión  :  4   1st Qu.:  13.00   1st Qu.: 3.688
 Median : 510.0   Median :2015   Cundinamarca:116   Villanueva:  4   Median :  26.00   Median : 6.598
 Mean   : 510.0   Mean   :2015   Santander   : 87   Albania   :  3   Mean   :  45.84   Mean   : 7.245
 3rd Qu.: 764.5   3rd Qu.:2015   Nariño      : 64   Argelia   :  3   3rd Qu.:  49.50   3rd Qu.: 9.758
 Max.   :1019.0   Max.   :2015   Bolívar     : 45   Bolívar   :  3   Max.   :2546.00   Max.   :24.457
                                 (Other)     :461   (Other)   :998
 RESULTADO.TURBIEDAD  RESULTADO.PH   RESULTADO.CLORO.RESIDUAL.LIBRE RESULTADO.ALCALINIDAD.TOTAL RESULTADO.CALCIO
 Min.   :0.004667    Min.   :6.342   Min.   :0.0000                 Min.   :  1.445            Min.   : 5.782
 1st Qu.:0.596441    1st Qu.:7.052   1st Qu.:0.5121                 1st Qu.: 17.997            1st Qu.:17.710
 Median :1.193333    Median :7.250   Median :0.7910                 Median : 38.333            Median :19.836
 Mean   :1.362457    Mean   :7.262   Mean   :0.8005                 Mean   : 39.265            Mean   :17.590
 3rd Qu.:1.776612    3rd Qu.:7.471   3rd Qu.:1.0995                 3rd Qu.: 55.634            3rd Qu.:19.836
 Max.   :4.934433    Max.   :8.166   Max.   :2.0320                 Max.   :116.495            Max.   :28.083

 RESULTADO.FOSFATOS RESULTADO.MAGNESIO RESULTADO.DUREZA.TOTAL RESULTADO.SULFATOS RESULTADO.HIERRO.TOTAL
 Min.   :0.01339    Min.   :1.933      Min.   :  2.75         Min.   : 0.000     Min.   :0.00000
 1st Qu.:0.10708    1st Qu.:4.727      1st Qu.: 27.75         1st Qu.: 7.045     1st Qu.:0.05594
 Median :0.13098    Median :5.133      Median : 49.30         Median :12.053     Median :0.12500
 Mean   :0.11139    Mean   :4.715      Mean   : 50.17         Mean   :11.709     Mean   :0.13498
 3rd Qu.:0.13098    3rd Qu.:5.133      3rd Qu.: 66.08         3rd Qu.:14.083     3rd Qu.:0.22186
 Max.   :0.20167    Max.   :7.048      Max.   :136.78         Max.   :30.000     Max.   :0.46864

 RESULTADO.CLORUROS RESULTADO.NITRITOS RESULTADO.ALUMINIO RESULTADO.FLORUROS RESULTADO.COLIFORMES.TOTALES
 Min.   : 0.3333    Min.   :0.00000    Min.   : 0.00000   Min.   :0.00000    Min.   :   0.000
 1st Qu.: 3.9571    1st Qu.:0.00400    1st Qu.: 0.08592   1st Qu.:0.05083    1st Qu.:   0.367
 Median : 7.3485    Median :0.01733    Median :13.96952   Median :0.82222    Median :  23.354
 Mean   : 6.7778    Mean   :0.01436    Mean   : 8.44917   Mean   :0.48932    Mean   : 172.934
 3rd Qu.: 9.1585    3rd Qu.:0.02193    3rd Qu.:13.96952   3rd Qu.:0.82222    3rd Qu.: 211.458
 Max.   :16.9000    Max.   :0.04533    Max.   :13.96952   Max.   :1.00000    Max.   :1125.673

 RESULTADO.E.COLI  IRCA.PROMEDIO     IRCA.BASE.PROMEDIO   NIVEL.DE.RIESGO.PROMEDIO
 Min.   : 0.0000   Min.   :  0.000   Min.   :  0.000    ALTO                   :282
 1st Qu.: 0.0000   1st Qu.:  4.466   1st Qu.:  4.441    BAJO                   :177
 Median : 0.5667   Median : 17.271    Median : 17.271    INVIABLE SANITARIAMENTE: 17
 Mean   : 6.7924   Mean   : 23.795   Mean   : 23.728    MEDIO                  :262
 3rd Qu.: 6.3062   3rd Qu.: 39.962   3rd Qu.: 39.563    SIN RIESGO             :281
```

For this data set, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found.

### *Dimensionality Reduction*

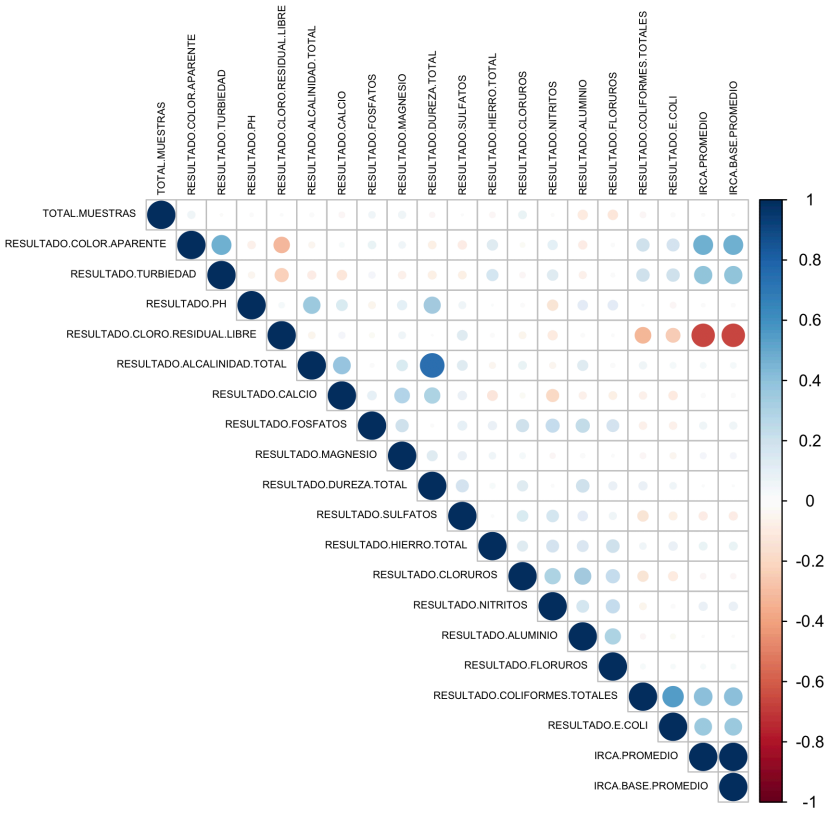
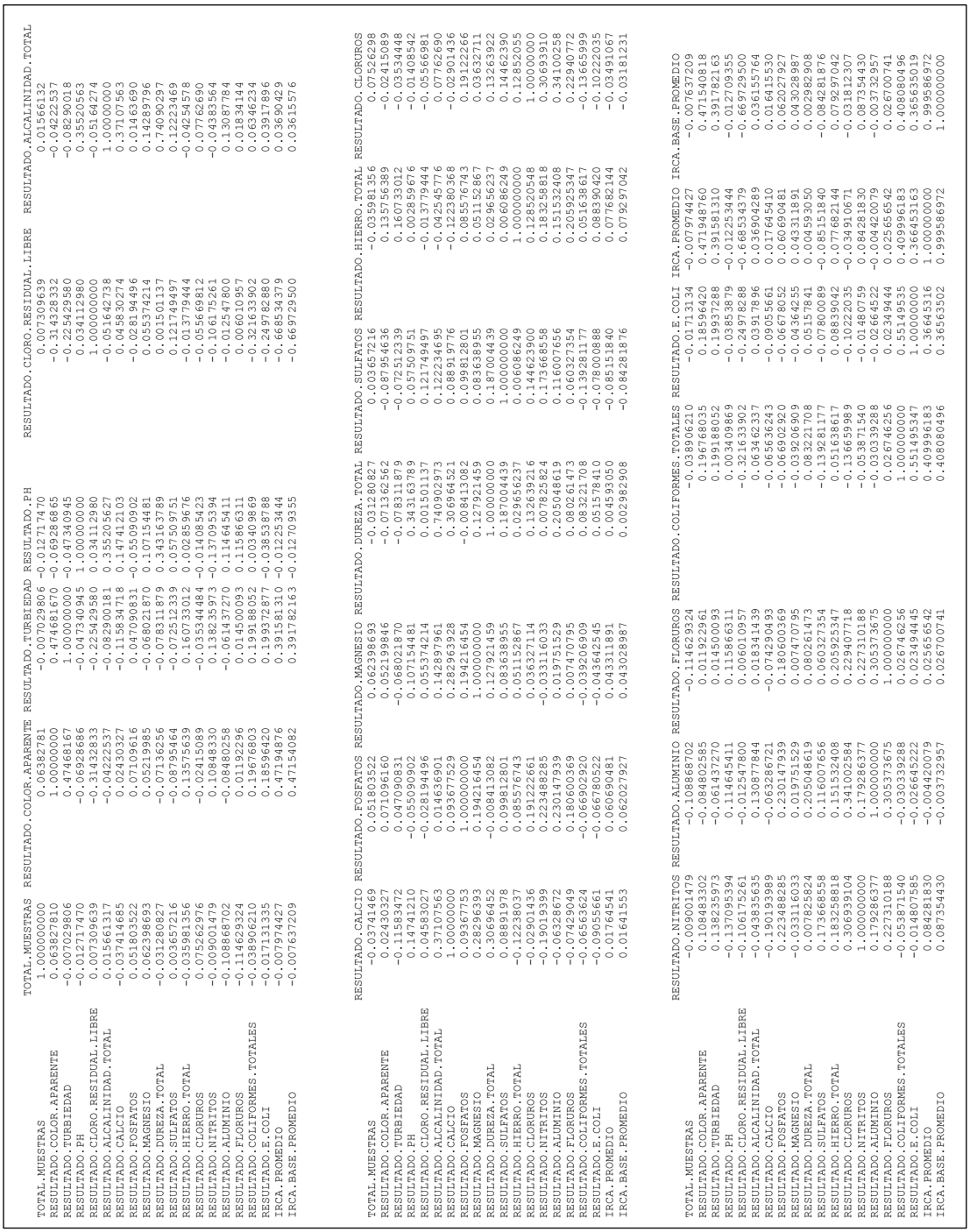



**Figure C. 1.** Correlated attributes for the SIVICAP dataset.



| | TOTAL.MUESTRAS | RESULTADO.COLOR.APARENTE | RESULTADO.TURBIEDAD | RESULTADO.PH | RESULTADO.CLORO.RESIDUAL.LIBRE | RESULTADO.ALCALINIDAD.TOTAL |
|---|---|---|---|---|---|---|
| TOTAL.MUESTRAS | 1.000000000 | 0.063827781 | -0.007029806 | -0.012717470 | 0.007309639 | 0.015661132 |
| RESULTADO.COLOR.APARENTE | 0.063827810 | 1.000000000 | 0.474681670 | -0.069286865 | -0.314328332 | -0.042225370 |
| RESULTADO.TURBIEDAD | -0.007029806 | 0.474681670 | 1.000000000 | -0.047340945 | -0.225429580 | -0.082900180 |
| RESULTADO.PH | -0.012717470 | -0.069286865 | -0.047340945 | 1.000000000 | 0.034112980 | 0.355205627 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 0.007309639 | -0.314328332 | -0.225429580 | 0.034112980 | 1.000000000 | 0.051642738 |
| RESULTADO.ALCALINIDAD.TOTAL | 0.015661317 | -0.042225370 | -0.082900181 | 0.355205627 | -0.051642738 | 1.000000000 |
| RESULTADO.CALCIO | -0.037414685 | 0.024303270 | -0.115834718 | 0.147412103 | 0.045830274 | 0.371075630 |
| RESULTADO.FOSFATOS | 0.051803522 | 0.071096160 | 0.047090831 | -0.055090902 | -0.028194496 | 0.014636900 |
| RESULTADO.MAGNESIO | 0.062398693 | 0.052199846 | -0.068021870 | 0.107154481 | 0.055374214 | 0.142897960 |
| RESULTADO.DUREZA.TOTAL | -0.031280827 | -0.071362562 | -0.078311879 | 0.343163789 | 0.001501137 | 0.740902973 |
| RESULTADO.HIERRO.TOTAL | -0.035981356 | -0.135756389 | 0.160733012 | 0.002859676 | 0.013779444 | -0.054554578 |
| RESULTADO.SULFATOS | 0.003657216 | -0.087954636 | 0.072512339 | 0.057509751 | 0.121749497 | -0.076626900 |
| RESULTADO.CLORUROS | -0.075262976 | -0.135756389 | -0.160733012 | 0.002859676 | -0.013779444 | 0.076626900 |
| RESULTADO.NITRITOS | -0.009001479 | -0.108483302 | -0.138235973 | -0.013709539 | -0.014645411 | -0.106175261 |
| RESULTADO.ALUMINIO | -0.108868702 | -0.084802585 | -0.061437270 | 0.014500093 | 0.011586631 | 0.006010957 |
| RESULTADO.FLOROUROS | -0.114629324 | -0.011922961 | -0.014500093 | 1.000000000 | -0.321633902 | -0.032194496 |
| RESULTADO.COLIFORMES.TOTALES | -0.038906210 | -0.196768035 | -0.199188052 | 0.003409869 | 0.088919776 | -0.122380368 |
| RESULTADO.E.COLI | -0.017131335 | -0.185960420 | -0.199372877 | -0.038538788 | -0.249782880 | 0.063462340 |
| IRCA.PROMEDIO | -0.007974427 | 0.471948760 | 0.391581310 | -0.012253444 | -0.668534379 | 0.036904290 |
| IRCA.BASE.PROMEDIO | -0.007637209 | 0.471540818 | 0.391782163 | -0.012709355 | -0.669729500 | 0.036155764 |

| | RESULTADO.CALCIO | RESULTADO.FOSFATOS | RESULTADO.MAGNESIO | RESULTADO.DUREZA.TOTAL | RESULTADO.SULFATOS | RESULTADO.HIERRO.TOTAL | RESULTADO.CLORUROS |
|---|---|---|---|---|---|---|---|
| TOTAL.MUESTRAS | -0.037414690 | 0.051803522 | 0.062398693 | -0.031280827 | 0.003657216 | -0.035981356 | 0.075262976 |
| RESULTADO.COLOR.APARENTE | 0.024303270 | 0.071096160 | 0.052199846 | -0.071362562 | -0.087954636 | 0.135756389 | -0.024150890 |
| RESULTADO.TURBIEDAD | -0.115834720 | 0.047090831 | -0.068021870 | -0.078311879 | -0.072512339 | 0.160733012 | -0.035344480 |
| RESULTADO.PH | 0.147412100 | -0.055090902 | 0.107154481 | 0.343163789 | 0.057509751 | 0.002859676 | -0.014085420 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | -0.028194496 | 0.055374214 | 0.055374214 | 0.001501137 | 0.121749497 | -0.013779444 | 0.055669810 |
| RESULTADO.ALCALINIDAD.TOTAL | 0.371075630 | 0.014636901 | 0.142897961 | 0.740902973 | 0.122234695 | -0.042545776 | 0.077626900 |
| RESULTADO.CALCIO | 1.000000000 | 0.093677529 | 0.282963928 | 0.306964521 | 0.088919776 | -0.122380368 | -0.029014360 |
| RESULTADO.FOSFATOS | 0.093677529 | 1.000000000 | 0.194216454 | -0.008413082 | 0.099812801 | 0.085576743 | 0.191222660 |
| RESULTADO.MAGNESIO | 0.282963928 | 0.194216454 | 1.000000000 | 0.127921459 | 0.083638955 | 0.051152867 | 0.036327110 |
| RESULTADO.DUREZA.TOTAL | 0.306964521 | -0.008413082 | 0.127921459 | 1.000000000 | 0.187004439 | 0.029656237 | 0.132639220 |
| RESULTADO.SULFATOS | 0.088919776 | 0.099812801 | 0.083638955 | 0.187004439 | 1.000000000 | 0.006086249 | 0.144623900 |
| RESULTADO.HIERRO.TOTAL | -0.122380368 | 0.085576743 | 0.051152867 | 0.029656237 | 0.006086249 | 1.000000000 | 0.128520548 |
| RESULTADO.CLORUROS | -0.029014360 | 0.191222661 | 0.036327110 | 0.132639216 | 0.144623900 | 0.128520548 | 1.000000000 |
| RESULTADO.NITRITOS | -0.190193989 | 0.223488285 | -0.033116033 | 0.007825824 | 0.117366558 | 0.183259818 | 0.306939104 |
| RESULTADO.ALUMINIO | -0.063286721 | 0.230147939 | 0.019751529 | 0.205048619 | 0.116007656 | 0.151532408 | 0.341002584 |
| RESULTADO.FLOROUROS | -0.074290493 | 0.180600369 | 0.007470795 | 0.080261473 | 0.060327354 | 0.205925347 | 0.227402772 |
| RESULTADO.COLIFORMES.TOTALES | -0.065636243 | -0.066902920 | -0.039206909 | 0.083221708 | -0.139281177 | 0.051638617 | -0.136659989 |
| RESULTADO.E.COLI | -0.090556610 | -0.066780522 | -0.043642545 | 0.051578410 | -0.078000888 | 0.088390420 | -0.102220350 |
| IRCA.PROMEDIO | 0.017645410 | 0.060690481 | 0.043311891 | 0.004593050 | -0.085151840 | 0.077682144 | -0.034910671 |
| IRCA.BASE.PROMEDIO | 0.016415530 | 0.062027927 | 0.043028987 | 0.002982908 | -0.084281876 | 0.079297042 | -0.031812307 |

| | RESULTADO.NITRITOS | RESULTADO.ALUMINIO | RESULTADO.FLOROUROS | RESULTADO.COLIFORMES.TOTALES | RESULTADO.E.COLI | IRCA.PROMEDIO | IRCA.BASE.PROMEDIO |
|---|---|---|---|---|---|---|---|
| TOTAL.MUESTRAS | -0.009001479 | -0.108868702 | -0.114629324 | -0.038906210 | -0.017131334 | -0.007974427 | -0.007637209 |
| RESULTADO.COLOR.APARENTE | 0.108483302 | -0.084802585 | 0.011922961 | 0.196768035 | 0.185960420 | 0.471948760 | 0.471540818 |
| RESULTADO.TURBIEDAD | 0.138235973 | -0.061437270 | 0.014500093 | 0.199188052 | 0.199372877 | 0.391581310 | 0.391782163 |
| RESULTADO.PH | -0.137095394 | 0.114645411 | 0.115866311 | 0.003409869 | -0.038538788 | -0.012253444 | -0.012709355 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | -0.106175261 | -0.012547800 | 0.006010957 | -0.321633902 | -0.249782880 | -0.668534379 | -0.669729500 |
| RESULTADO.ALCALINIDAD.TOTAL | -0.043835635 | 0.130877844 | 0.018341439 | 0.063462337 | 0.039178796 | 0.036904289 | 0.036155764 |
| RESULTADO.CALCIO | -0.190193989 | -0.063286721 | -0.074290493 | -0.065636243 | -0.090556610 | 0.017645410 | 0.016415530 |
| RESULTADO.FOSFATOS | 0.223488285 | 0.230147939 | 0.180600369 | -0.066902920 | -0.066780522 | 0.060690481 | 0.062027927 |
| RESULTADO.MAGNESIO | -0.033116033 | 0.019751529 | 0.007470795 | -0.039206909 | -0.043642545 | 0.043311891 | 0.043028987 |
| RESULTADO.DUREZA.TOTAL | 0.007825824 | 0.205048619 | 0.080261473 | 0.083221708 | 0.051578410 | 0.004593050 | 0.002982908 |
| RESULTADO.SULFATOS | 0.173668558 | 0.116007656 | 0.060327354 | -0.139281177 | -0.078000889 | -0.085151840 | -0.084281876 |
| RESULTADO.HIERRO.TOTAL | 0.183258818 | 0.151532408 | 0.205925347 | 0.051638617 | 0.088390042 | 0.077682144 | 0.079297042 |
| RESULTADO.CLORUROS | 0.306939104 | 0.341002584 | 0.227402772 | -0.136659989 | -0.102220035 | -0.034910671 | -0.031812307 |
| RESULTADO.NITRITOS | 1.000000000 | 0.179286377 | 0.227310188 | -0.053871540 | -0.014807594 | 0.084281830 | 0.087354430 |
| RESULTADO.ALUMINIO | 0.179286377 | 1.000000000 | 0.305373675 | -0.030339288 | -0.026646236 | -0.004420079 | -0.003732957 |
| RESULTADO.FLOROUROS | 0.227310188 | 0.305373675 | 1.000000000 | 0.026746256 | 0.023494445 | 0.025656542 | 0.026700741 |
| RESULTADO.COLIFORMES.TOTALES | -0.053871540 | -0.030339288 | 0.026746256 | 1.000000000 | 0.551495347 | 0.409996183 | 0.408080496 |
| RESULTADO.E.COLI | -0.014807585 | -0.026645222 | 0.023494445 | 0.551495347 | 1.000000000 | 0.366453163 | 0.365635019 |
| IRCA.PROMEDIO | 0.084281830 | -0.004420079 | 0.025656542 | 0.409996183 | 0.366453163 | 1.000000000 | 0.999586972 |
| IRCA.BASE.PROMEDIO | 0.087354430 | -0.003732957 | 0.026700741 | 0.408080496 | 0.365635019 | 0.999586972 | 1.000000000 |

**Table C. 2.** Correlation matrix for all numerical attributes in the SIVICAP data set.

**Significance Levels**

| | Estimate | Std. Error | z-value | Pr(>\|z\|) | Significance |
|---|---|---|---|---|---|
| (Intercept) | 6.6895969 | 3.2027568 | 2.089 | 0.0367 | * |
| TOTAL.MUESTRAS | -0.0006021 | 0.0014071 | -0.428 | 0.6687 | |
| RESULTADO.COLOR.APARENTE | -0.0516198 | 0.0250881 | -2.058 | 0.0396 | * |
| RESULTADO.TURBIEDAD | -0.0630833 | 0.1293756 | -0.488 | 0.6258 | |
| RESULTADO.PH | -0.3687104 | 0.4415066 | -0.835 | 0.4037 | |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 1.8326622 | 0.4686733 | 3.910 | 9.22e-05 | *** |
| RESULTADO.ALCALINIDAD.TOTAL | 0.0033185 | 0.0074478 | 0.446 | 0.6559 | |
| RESULTADO.CALCIO | -0.0288041 | 0.0324573 | -0.887 | 0.3748 | |
| RESULTADO.FOSFATOS | 0.4157325 | 3.2684862 | 0.127 | 0.8988 | |
| RESULTADO.MAGNESIO | 0.0926019 | 0.1442445 | 0.642 | 0.5209 | |
| RESULTADO.DUREZA.TOTAL | 0.0012046 | 0.0062400 | 0.193 | 0.8469 | |
| RESULTADO.SULFATOS | -0.0314414 | 0.0200245 | -1.570 | 0.1164 | |
| RESULTADO.HIERRO.TOTAL | 1.6498693 | 1.4490480 | 1.139 | 0.2549 | |
| RESULTADO.CLORUROS | -0.0596607 | 0.0439708 | -1.357 | 0.1748 | |
| RESULTADO.NITRITOS | 14.0639482 | 14.1833024 | 0.992 | 0.3214 | |
| RESULTADO.ALUMINIO | 0.0500850 | 0.0199138 | 2.515 | 0.0119 | * |
| RESULTADO.FLORUROS | 0.1438011 | 0.3716002 | 0.387 | 0.6988 | |
| RESULTADO.COLIFORMES.TOTALES | -0.0010786 | 0.0004536 | -2.378 | 0.0174 | * |
| RESULTADO.E.COLI | -0.0061070 | 0.0092940 | -0.657 | 0.5111 | |
| IRCA.PROMEDIO | -0.0744167 | 0.1737160 | -0.428 | 0.6684 | |
| IRCA.BASE.PROMEDIO | -0.0395580 | 0.1737084 | -0.228 | 0.8199 | |

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Logistic Regression**

| | Overall |
|---|---|
| TOTAL.MUESTRAS | 0.4278971 |
| RESULTADO.COLOR.APARENTE | 2.0575381 |
| RESULTADO.TURBIEDAD | 0.4875981 |
| RESULTADO.PH | 0.8351187 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 3.9103195 |
| RESULTADO.ALCALINIDAD.TOTAL | 0.4455744 |
| RESULTADO.CALCIO | 0.8874469 |
| RESULTADO.FOSFATOS | 0.1271942 |
| RESULTADO.MAGNESIO | 0.6419792 |
| RESULTADO.DUREZA.TOTAL | 0.1930414 |
| RESULTADO.SULFATOS | 1.5701485 |
| RESULTADO.HIERRO.TOTAL | 1.1385884 |
| RESULTADO.CLORUROS | 1.3568251 |
| RESULTADO.NITRITOS | 0.9915849 |
| RESULTADO.ALUMINIO | 2.5150928 |
| RESULTADO.FLORUROS | 0.3869781 |
| RESULTADO.COLIFORMES.TOTALES | 2.3779949 |
| RESULTADO.E.COLI | 0.6570878 |
| IRCA.PROMEDIO | 0.4283811 |
| IRCA.BASE.PROMEDIO | 0.2277265 |

**Random Forest**

| | MeanDecreaseGini |
|---|---|
| TOTAL.MUESTRAS | 9.417735 |
| RESULTADO.COLOR.APARENTE | 20.791546 |
| RESULTADO.TURBIEDAD | 17.047379 |
| RESULTADO.PH | 4.738698 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 44.332178 |
| RESULTADO.ALCALINIDAD.TOTAL | 4.759860 |
| RESULTADO.CALCIO | 2.963708 |
| RESULTADO.FOSFATOS | 3.451186 |
| RESULTADO.MAGNESIO | 2.571204 |
| RESULTADO.DUREZA.TOTAL | 5.804322 |
| RESULTADO.SULFATOS | 5.018895 |
| RESULTADO.HIERRO.TOTAL | 3.988846 |
| RESULTADO.CLORUROS | 4.923597 |
| RESULTADO.NITRITOS | 3.532562 |
| RESULTADO.ALUMINIO | 3.857195 |
| RESULTADO.FLORUROS | 2.358447 |
| RESULTADO.COLIFORMES.TOTALES | 36.959588 |
| RESULTADO.E.COLI | 50.627683 |
| IRCA.PROMEDIO | 272.243588 |
| IRCA.BASE.PROMEDIO | 264.946801 |

**Table C. 3.** Importance of variables using the Logistic Regression and Random Forest methods for the SIVICAP data set.

# C.2. CORPOICA Data Source

### *Check Missing Values*

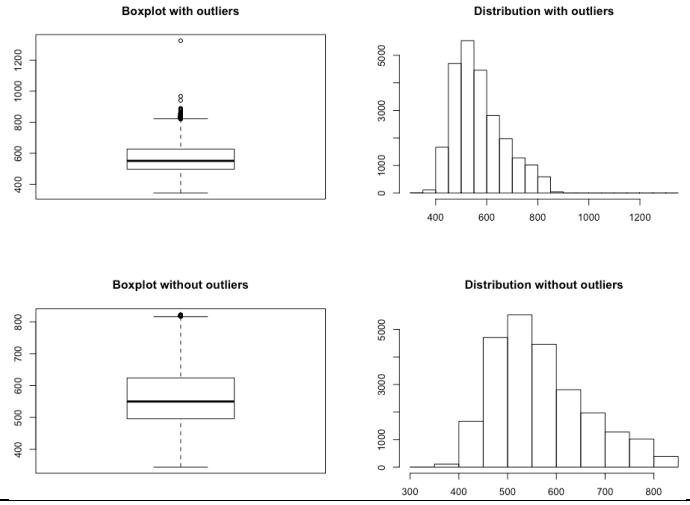
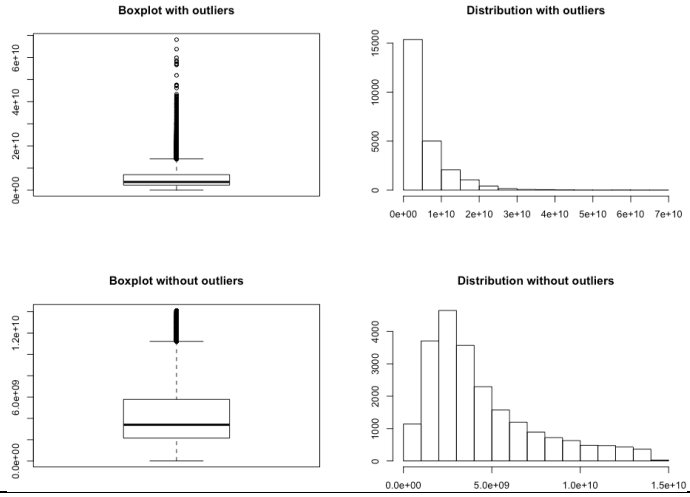
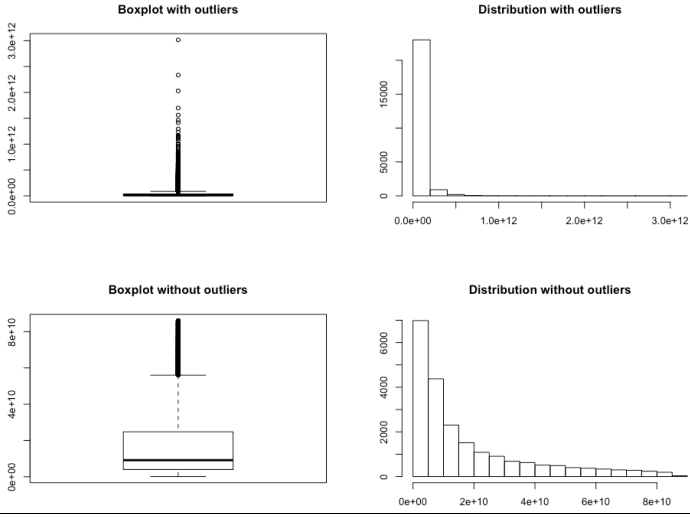
```
      Departamento          Municipio                    Cultivo              Estado         Tiempo.Establecimiento
Cundinamarca      :5584   C£cuta       :  742   Pastos             : 3874   Establecido   :13714   de 0 a 1 a¤o  : 2543
Valle             :5213   Villavicencio:  733   No indica          : 2407   No indica     : 3634   de 1 a 5 a¤os : 4895
Meta              :2093   Ceret,       :  595   Aguacate           : 2030   Por establecer: 6831   de 5 a 10 a¤os: 1564
Boyaca            :2040   No Indica    :  595   Ca¤a panelera/azucar: 1676                          mas de 10 a¤os: 2241
Antioquia         :1251   Espinal      :  501   Cacao              : 1572                          no indica     :    3
Norte De Santander: 940   Palmira      :  463   Caf,               : 1237                          no Indica     :  300
(Other)           :7058   (Other)      :20550   (Other)            :11383                          No indica     :12633

    Topografia        Drenaje                      Riego          Fertilizantes.aplicados             Fecha.de.análisis
No indica:3834   Bueno    :13852   No cuenta con riego:11758   No indica         :11548                        :15295
Ondulado :6246   Malo     :  336   No indica          : 8997   No aplica         : 5988      09/12/2015 12:00:00 AM:  494
Pendiente:5820   No indica: 6549   Goteo              : 1055   N-P-K             : 4372      08/01/2016 12:00:00 AM:  450
Plano    :8279   Regular  : 3442   Aspersi¢n          :  979   Abono org nico    :  499      12/04/2014 12:00:00 AM:  441
                                   Gravedad           :  960   N                 :  441      12/12/2014 12:00:00 AM:  344
                                   Manguera           :  380   N-P-K, m s menores:  440      05/12/2014 12:00:00 AM:  342
                                   (Other)            :   50   (Other)           :  891      (Other)                : 6813

     pH         Materia.organica       Fosforo             Azufre             Acidez          Aluminio.intercambiable
Min.   : 344.0   Min.   :1.147e+07   Min.   :8.804e+07   Min.   :1.143e+08   Min.   :0.000e+00   Min.   :0.000e+00
1st Qu.: 497.0   1st Qu.:2.263e+09   1st Qu.:4.325e+09   1st Qu.:3.765e+09   1st Qu.:0.000e+00   1st Qu.:0.000e+00
```

```
 Median : 551.0   Median :3.683e+09   Median :1.129e+10   Median :6.013e+09   Median :0.000e+00   Median :0.000e+00
 Mean   : 570.3   Mean   :5.689e+09   Mean   :4.000e+10   Mean   :1.266e+10   Mean   :9.120e+08   Mean   :6.966e+08
 3rd Qu.: 627.0   3rd Qu.:7.007e+09   3rd Qu.:3.704e+10   3rd Qu.:1.066e+10   3rd Qu.:1.337e+09   3rd Qu.:9.876e+08
 Max.   :1325.0   Max.   :6.809e+10   Max.   :3.016e+12   Max.   :4.863e+12   Max.   :2.366e+10   Max.   :1.799e+10
 NA's   :2                            NA's   :3           NA's   :2           NA's   :1           NA's   :1

Calcio.intercambiable Magnesio.intercambiable Potasio.intercambiable Sodio.intercambiable capacidad.de.intercambio.cationico
 Min.   :2.844e+07     Min.   :6.155e+06       Min.   :1.688e+05      Min.   :0.000e+00    Min.   :1.961e+08
 1st Qu.:1.903e+09     1st Qu.:5.037e+08       1st Qu.:1.384e+08      1st Qu.:4.896e+07    1st Qu.:4.571e+09
 Median :4.755e+09     Median :1.219e+09       Median :2.686e+08      Median :8.985e+07    Median :7.980e+09
 Mean   :6.959e+09     Mean   :2.144e+09       Mean   :4.086e+08      Mean   :1.797e+08    Mean   :1.060e+10
 3rd Qu.:9.580e+09     3rd Qu.:2.569e+09       3rd Qu.:5.121e+08      3rd Qu.:1.904e+08    3rd Qu.:1.399e+10
 Max.   :1.125e+11     Max.   :3.066e+10       Max.   :1.536e+10      Max.   :1.197e+10    Max.   :1.244e+11

Conductividad.electrica  Hierro.olsen              Cobre              Manganeso           Zinc.olsen         Boro
 Min.   :5.284e+05       Min.   :1.000e+08              : 2302             : 2302              : 2302   Min.   :1.349e+05
 1st Qu.:1.680e+08       1st Qu.:4.945e+10   < 1,00     : 1798   < 1,00    :  832   < 1,00     : 1970   1st Qu.:1.477e+08
 Median :2.626e+08       Median :1.350e+11   2,400000000:  328   1,600000000:  224   1,000000000:  482   Median :2.276e+08
 Mean   :5.519e+08       Mean   :3.018e+11   2,000000000:  318   1,400000000:  216   0,600000000:  468   Mean   :3.023e+08
 3rd Qu.:4.439e+08       3rd Qu.:4.020e+11   2,300000000:  302   2,000000000:  205   0,700000000:  456   3rd Qu.:3.459e+08
 Max.   :1.852e+12       Max.   :6.947e+12   2,100000000:  299   1,800000000:  197   0,500000000:  425   Max.   :1.537e+10
 NA's   :63              NA's   :2371        (Other)    :18832   (Other)    :20203   (Other)    :18076   NA's   :3

    Hierro         Cobre.doble.acido  Manganeso.1           Zinc
 Min.   :   12400   Min.   : 1600     Min.   :    800          :21878
 1st Qu.:  234000   1st Qu.: 8400     1st Qu.:  12400   < 0,40 :  177
 Median :  380808   Median :12288     Median :  43800   0,2000 :   77
 Mean   :  646255   Mean   :14002     Mean   : 118328   0,1600 :   72
 3rd Qu.:  864000   3rd Qu.:17600     3rd Qu.: 147700   0,3600 :   71
 Max.   :12632128   Max.   :98520     Max.   :2088000   0,3200 :   69
 NA's   :21878      NA's   :21944     NA's   :21919     (Other): 1835
```

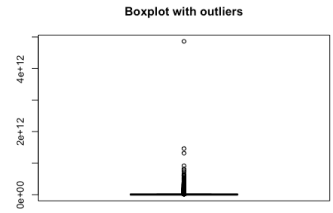
Attributes with a high percentage of lost values (NA's $>$ 70% or NA's $>$ 16,925) have a greater probability of being initially discarded. In the previous summary, those attributes with this characteristic are highlighted. In contrast, attributes with low percentages of missing values are selected to continue with the cleaning process and these are listed below.

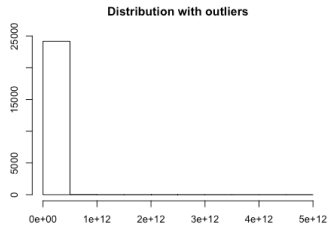
```
Departamento
Municipio
Cultivo
Estado
Tiempo.Establecimiento
Topografia
Drenaje
Riego
Fertilizantes.aplicados
Fecha.de.analisis
pH
Materia.organica
Fosforo
Azufre
Acidez
Aluminio.intercambiable
Calcio.intercambiable
Magnesio.intercambiable
Potasio.intercambiable
Sodio.intercambiable
Capacidad.de.intercambio.cationico
Conductividad.electrica
Hierro.olsen
Cobre
Manganeso
Zinc.olsen
Boro
Hierro
Cobre.doble.acido
Manganeso
Zinc
```

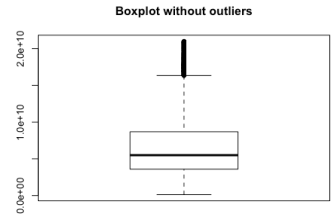
## *Outliers Detection*

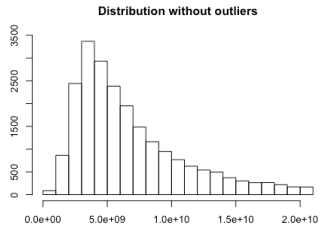
| Attribute Metrics | Distribution Charts |
|---|---|
| **pH**<br>Outliers identified: 245<br>Proportion (%) of outliers: 1<br>Mean of the outliers: 839.77<br>Mean without removing outliers: 570.29<br>Mean if we remove outliers: 567.53 |  |
| **Materia.organica**<br>Outliers identified: 2031<br>Proportion (%) of outliers: 9.2<br>Mean of the outliers: 19925339955<br>Mean without removing outliers: 5688671782<br>Mean if we remove outliers: 4383151055 |  |
| **Fosforo**<br>Outliers identified: 2551<br>Proportion (%) of outliers: 11.8<br>Mean of the outliers: 2.27514e+11<br>Mean without removing outliers: 40003679404<br>Mean if we remove outliers: 17887032973 |  |

**Azufre**
Outliers identified: 2369
Proportion (%) of outliers: 10.9
Mean of the outliers: 67240790658
Mean without removing outliers: 12656906444
Mean if we remove outliers: 6728010446

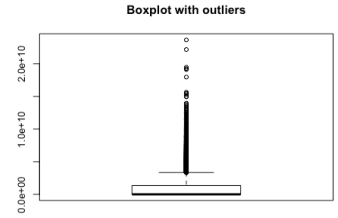


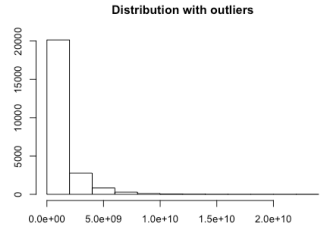



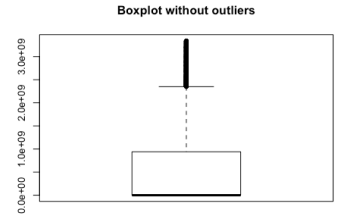



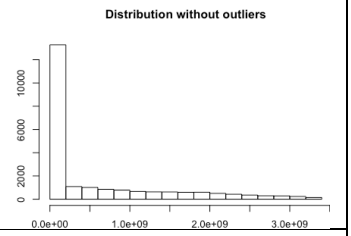




**Acidez**
Outliers identified: 1795
Proportion (%) of outliers: 8
Mean of the outliers: 5265197581
Mean without removing outliers: 911994334
Mean if we remove outliers: 562905707

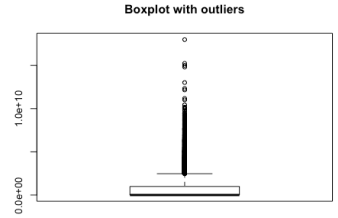


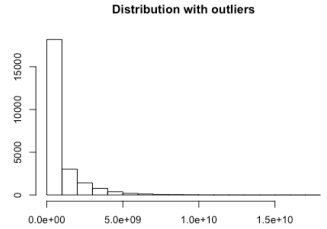



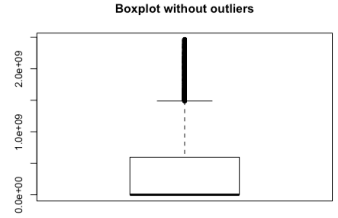



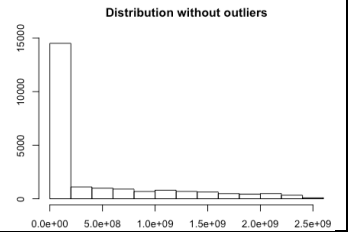




**Aluminio.intercambiable**
Outliers identified: 2067
Proportion (%) of outliers: 9.3
Mean of the outliers: 4032539128
Mean without removing outliers: 696633054
Mean if we remove outliers: 384797044

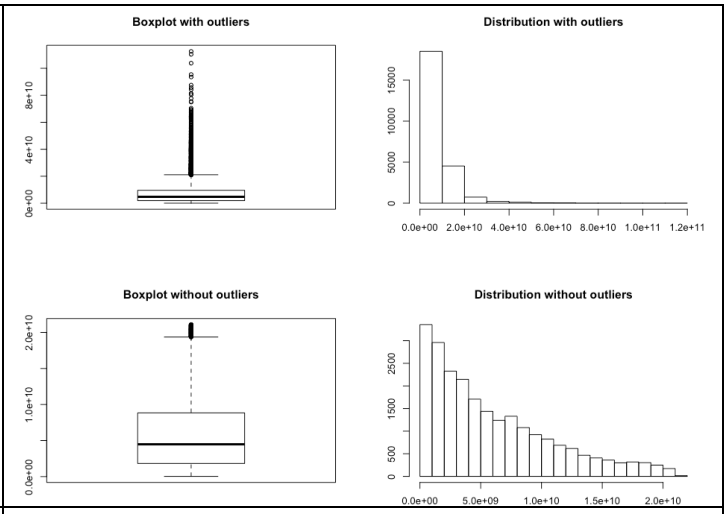


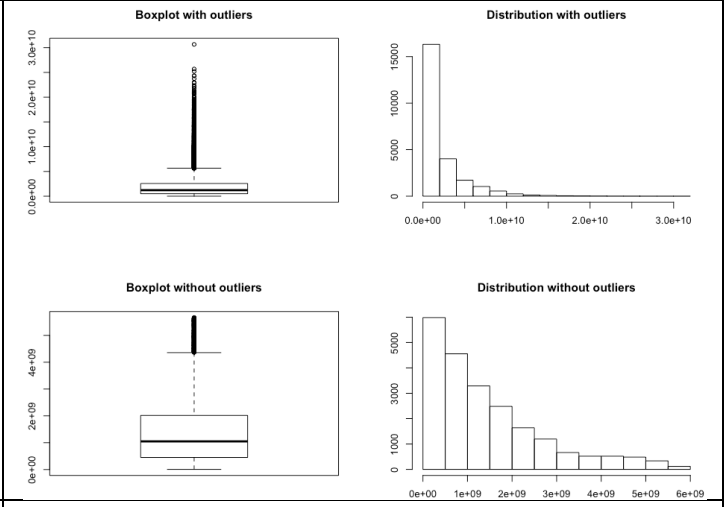



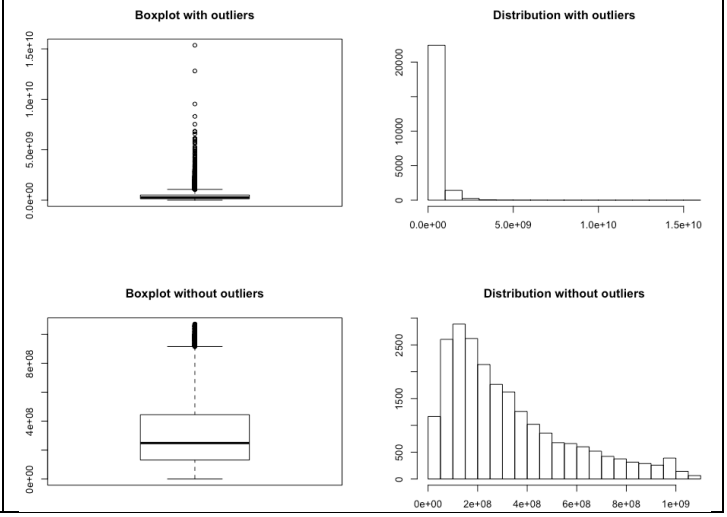
**Calcio.intercambiable**
Outliers identified: 952
Proportion (%) of outliers: 4.1
Mean of the outliers: 32040393795
Mean without removing outliers: 6958790290
Mean if we remove outliers: 5930776060



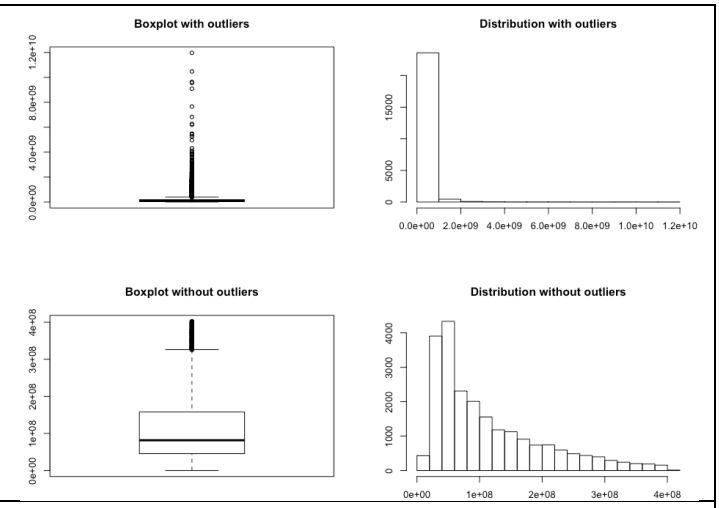

**Magnesio.intercambiable**
Outliers identified: 2389
Proportion (%) of outliers: 11
Mean of the outliers: 8629441813
Mean without removing outliers: 2144042636
Mean if we remove outliers: 1.433e+09



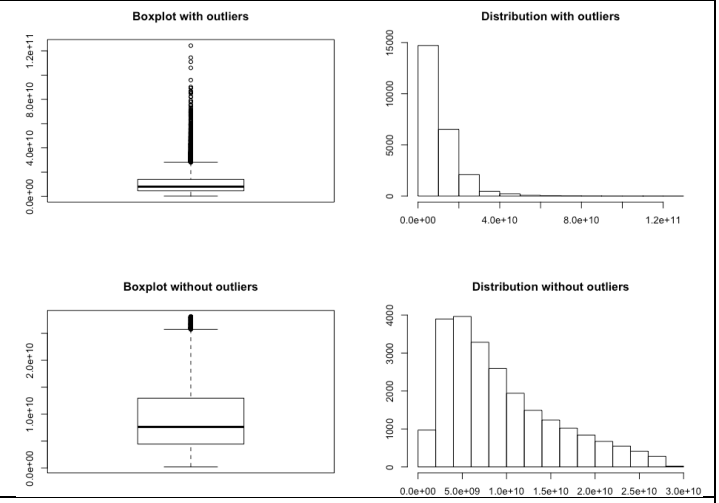

**Potasio.intercambiable**
Outliers identified: 1534
Proportion (%) of outliers: 6.8
Mean of the outliers: 1716174806
Mean without removing outliers: 408558703
Mean if we remove outliers: 319979189



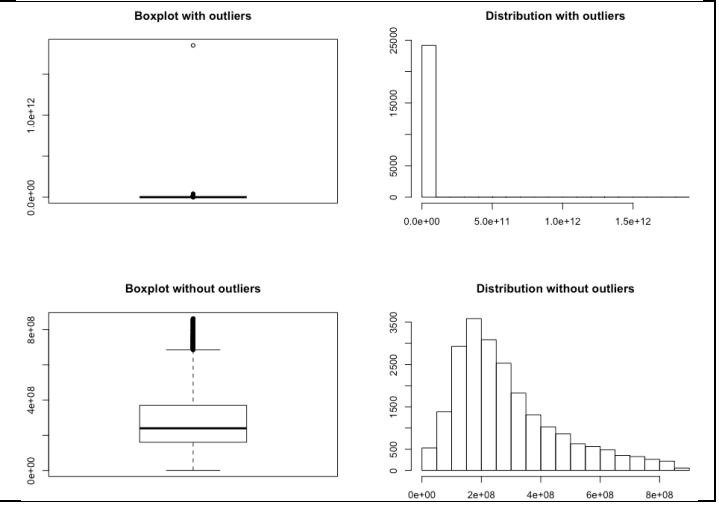
**Sodio.intercambiable**
Outliers identified: 1923
Proportion (%) of outliers: 8.6
Mean of the outliers: 957175611
Mean without removing outliers: 179668304
Mean if we remove outliers: 112488822



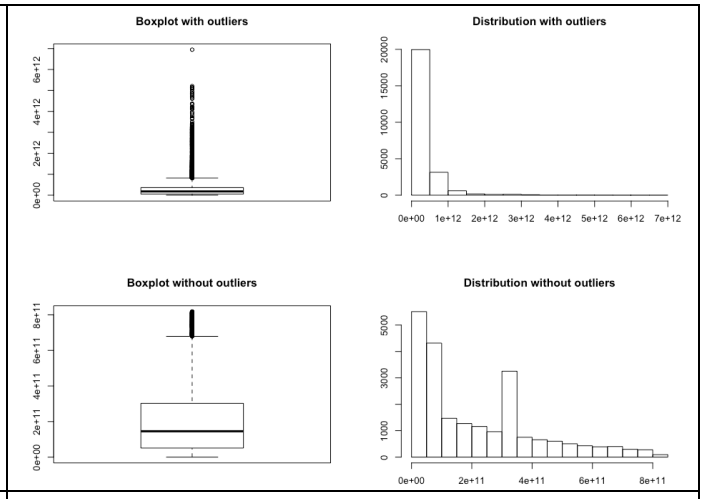

**capacidad.de.intercambio.cationico**
Outliers identified: 1009
Proportion (%) of outliers: 4.4
Mean of the outliers: 39504260664
Mean without removing outliers: 10602790192
Mean if we remove outliers: 9344197887




**Conductividad.electrica**
Outliers identified: 2227
Proportion (%) of outliers: 10.1
Mean of the outliers: 3162991899
Mean without removing outliers: 551716577
Mean if we remove outliers: 286806311



**Hierro.olsen**
Outliers identified: 1833
Proportion (%) of outliers: 8.2
Mean of the outliers: 1.381154e+12
Mean without removing outliers: 301815156852
Mean if we remove outliers: 2.13279e+11




**Boro**
Outliers identified: 1738
Proportion (%) of outliers: 7.7
Mean of the outliers: 1087054992
Mean without removing outliers: 302344412
Mean if we remove outliers: 241570517

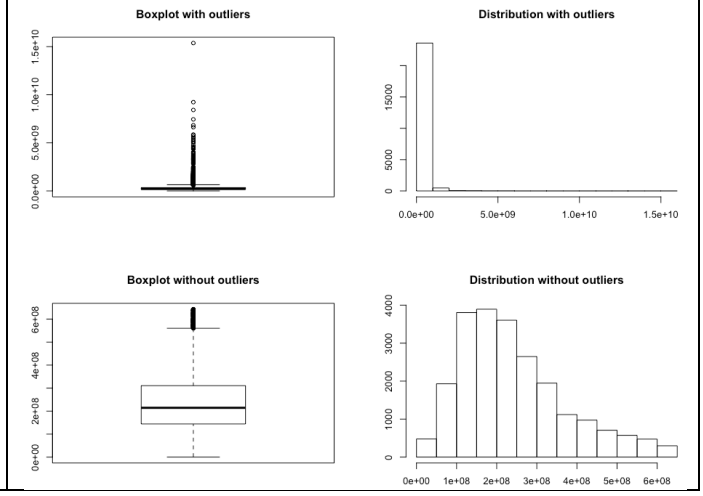



**Table C. 4.** Results of the outlier detection process for each of the CORPOICA dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
     Departamento            Municipio                Cultivo                Estado          Tiempo.Establecimiento     Topografia      Drenaje
Cundinamarca :5584   CÃcuta       :  742   Pastos             : 3874   Establecido  :13714   de 0 a 1 aÃ±o  : 2543   No indica:3834   Bueno    :13852
Valle        :5213   Villavicencio:  733   No indica          : 2407   No indica    : 3634   de 1 a 5 aÃ±os : 4895   Ondulado :6246   Malo     :  336
Meta         :2093   Ceret,       :  595   Aguacate           : 2030   Por establecer: 6831   de 5 a 10 aÃ±os: 1564   Pendiente:5820   No indica: 6549
Boyaca       :2040   No Indica    :  595   CaÃ±a panelera/azucar: 1676                          mas de 10 aÃ±os: 2241   Plano    :8279   Regular  : 3442
Antioquia    :1251   Espinal      :  501   Cacao              : 1572                          no indica     :    3
Norte De Santander: 940   Palmira      :  463   Caf,               : 1237                          no Indica     :  300
(Other)      :7058   (Other)      :20550   (Other)            :11383                          No indica     :12633

           Riego        Fertilizantes.aplicados            Fecha.de.análisis          pH          Materia.organica      Fosforo             Azufre
No cuenta con riego:11758   No indica         :11548                    :15295   Min.   :344.0   Min.   :1.147e+07   Min.   :8.804e+07   Min.   :1.143e+08
No indica          : 8997   No aplica         : 5988   09/12/2015 12:00:00 AM:  494   1st Qu.:497.0   1st Qu.:2.153e+09   1st Qu.:4.327e+09   1st Qu.:3.765e+09
Goteo              : 1055   N-P-K             : 4372   08/01/2016 12:00:00 AM:  450   Median :551.0   Median :3.404e+09   Median :1.129e+10   Median :5.973e+09
Aspersiï¿½n         :  979   Abono org nico    :  499   12/04/2014 12:00:00 AM:  441   Mean   :567.5   Mean   :4.383e+09   Mean   :1.760e+10   Mean   :6.654e+09
Gravedad           :  960   N                 :  441   12/12/2014 12:00:00 AM:  344   3rd Qu.:623.0   3rd Qu.:5.781e+09   3rd Qu.:2.161e+10   3rd Qu.:8.176e+09
Manguera           :  380   N-P-K, m s menores:  440   05/12/2014 12:00:00 AM:  342   Max.   :822.0   Max.   :1.412e+10   Max.   :8.612e+10   Max.   :2.100e+10
(Other)            :   50   (Other)           :  891   (Other)                : 6813                   NA's   :2031

   Acidez          Aluminio.intercambiable Calcio.intercambiable Magnesio.intercambiable Potasio.intercambiable Sodio.intercambiable capacidad.de.intercambio.cationico
Min.   :0.000e+00   Min.   :0.000e+00       Min.   :2.844e+07     Min.   :6.155e+06       Min.   :1.688e+05      Min.   :        0    Min.   :1.961e+08
1st Qu.:0.000e+00   1st Qu.:0.000e+00       1st Qu.:1.903e+09     1st Qu.:5.037e+08       1st Qu.:1.384e+08      1st Qu.: 48960000    1st Qu.:4.571e+09
Median :0.000e+00   Median :0.000e+00       Median :4.755e+09     Median :1.219e+09       Median :2.616e+08      Median : 89852910    Median :7.380e+09
Mean   :5.641e+08   Mean   :3.881e+08       Mean   :5.895e+09     Mean   :1.432e+09       Mean   :3.163e+08      Mean   :111290718    Mean   :9.262e+09
3rd Qu.:8.290e+08   3rd Qu.:4.975e+08       3rd Qu.:8.611e+09     3rd Qu.:1.850e+09       3rd Qu.:4.238e+08      3rd Qu.:147936740    3rd Qu.:1.256e+10
Max.   :3.342e+09   Max.   :2.467e+09       Max.   :2.109e+10     Max.   :5.663e+09       Max.   :1.071e+09      Max.   :402376450    Max.   :2.812e+10

Conductividad.electrica  Hierro.olsen             Cobre                Manganeso            Zinc.olsen             Boro              Zinc
Min.   :   528400       Min.   :1.000e+08              : 2302                : 2302                : 2302   Min.   :   134934          :21878
1st Qu.:168559600       1st Qu.:5.660e+10   < 1,00     : 1798   < 1,00     :  832   < 1,00     : 1970   1st Qu.:147736121   < 0,40 :  177
Median :237070368       Median :1.810e+11   2,400000000:  328   1,600000000:  224   1,000000000:  482   Median :211818887   0,2000 :   77
Mean   :282225396       Mean   :2.136e+11   2,000000000:  318   1,400000000:  216   0,600000000:  468   Mean   :239431953   0,1600 :   72
3rd Qu.:351180500       3rd Qu.:3.023e+11   2,300000000:  302   2,000000000:  205   0,700000000:  456   3rd Qu.:300902927   0,3600 :   71
Max.   :862348800       Max.   :8.183e+11   2,100000000:  299   1,800000000:  197   0,500000000:  425   Max.   :642977149   0,3200 :   69
                                            (Other)    :18832   (Other)    :20203   (Other)    :18076                       (Other): 1835
```

For this data set, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found.

### *Dimensionality Reduction*

| | pH | Materia.organica | Fosforo | Azufre | Acidez | Aluminio.intercambiable | Calcio.intercambiable |
|---|---|---|---|---|---|---|---|
| pH | 1.0000000000 | -0.21692851 | 0.208440275 | 0.0003637371 | -0.61890659 | -0.5835978 | 0.58902444 |
| Materia.organica | -0.2169285124 | 1.00000000 | 0.021136026 | 0.1991387082 | 0.13823731 | 0.1173553 | -0.10713148 |
| Fosforo | 0.2084402748 | 0.02113603 | 1.000000000 | 0.1390930486 | -0.10107150 | -0.1076478 | 0.18194908 |
| Azufre | 0.0003637371 | 0.19913871 | 0.139093049 | 1.0000000000 | 0.03389969 | 0.0268740 | 0.02684553 |
| Acidez | -0.6189065863 | 0.13823731 | -0.101071500 | 0.0338996882 | 1.00000000 | 0.8980912 | -0.47216744 |
| Aluminio.intercambiable | -0.5835978221 | 0.11735534 | -0.107647837 | 0.0268740031 | 0.89809116 | 1.0000000 | -0.45437429 |
| Calcio.intercambiable | 0.5890244352 | -0.10713148 | 0.181949075 | 0.0268455288 | -0.47216744 | -0.4543743 | 1.00000000 |
| Magnesio.intercambiable | 0.4571256213 | -0.12928653 | 0.111125993 | 0.0159109492 | -0.41266222 | -0.3982696 | 0.54011209 |
| Potasio.intercambiable | 0.2213271739 | 0.15369123 | 0.215929279 | 0.1745445297 | -0.17103966 | -0.1764888 | 0.29526553 |
| Sodio.intercambiable | 0.1883642290 | -0.03805374 | 0.126473956 | 0.0855114477 | -0.12205074 | -0.1236563 | 0.25731570 |
| capacidad.de.intercambio.cationico | 0.5154367273 | -0.09727420 | 0.173374364 | 0.0473682129 | -0.38531085 | -0.3726912 | 0.85620280 |
| Conductividad.electrica | 0.2456621704 | 0.14654752 | 0.216429773 | 0.2912083946 | -0.14850567 | -0.1563259 | 0.28229293 |
| Hierro.olsen | -0.5294985411 | 0.20579787 | 0.003785986 | 0.0514558593 | 0.41226982 | 0.3834673 | -0.33466076 |
| Boro | 0.2662454345 | 0.01421306 | 0.219943010 | 0.1641352646 | -0.13999513 | -0.1355542 | 0.26676630 |

| | Magnesio.intercambiable | Potasio.intercambiable | Sodio.intercambiable | capacidad.de.intercambio.cationico | Conductividad.electrica |
|---|---|---|---|---|---|
| pH | 0.45712562 | 0.22132717 | 0.188364229 | 0.51543673 | 0.2456622 |
| Materia.organica | -0.12928653 | 0.15369123 | -0.038053736 | -0.09727420 | 0.1465475 |
| Fosforo | 0.11112599 | 0.21592928 | 0.126473956 | 0.17337436 | 0.2164298 |
| Azufre | 0.01591095 | 0.17454453 | 0.085511448 | 0.04736821 | 0.2912084 |
| Acidez | -0.41266222 | -0.17103966 | -0.122050742 | -0.38531085 | -0.1485057 |
| Aluminio.intercambiable | -0.39826964 | -0.17648879 | -0.123656263 | -0.37269122 | -0.1563259 |
| Calcio.intercambiable | 0.54011209 | 0.29526553 | 0.257315698 | 0.85620280 | 0.2822929 |
| Magnesio.intercambiable | 1.00000000 | 0.28194289 | 0.189758636 | 0.56053819 | 0.1847873 |
| Potasio.intercambiable | 0.28194289 | 1.00000000 | 0.158453585 | 0.31405563 | 0.3094685 |
| Sodio.intercambiable | 0.18975864 | 0.15845358 | 1.000000000 | 0.27415421 | 0.2120372 |
| capacidad.de.intercambio.cationico | 0.56053819 | 0.31405563 | 0.274154207 | 1.00000000 | 0.2874299 |
| Conductividad.electrica | 0.18478730 | 0.30946854 | 0.212037175 | 0.28742994 | 1.0000000 |
| Hierro.olsen | -0.29408525 | -0.07409805 | -0.003134633 | -0.28011168 | -0.1103359 |
| Boro | 0.17669205 | 0.24309115 | 0.214926808 | 0.27092567 | 0.2838664 |

| | Hierro.olsen | Boro |
|---|---|---|
| pH | -0.529498541 | 0.26624543 |
| Materia.organica | 0.205797866 | 0.01421306 |
| Fosforo | 0.003785986 | 0.21994301 |
| Azufre | 0.051455859 | 0.16413526 |
| Acidez | 0.412269817 | -0.13999513 |
| Aluminio.intercambiable | 0.383467287 | -0.13555421 |
| Calcio.intercambiable | -0.334660764 | 0.26676630 |
| Magnesio.intercambiable | -0.294085252 | 0.17669205 |
| Potasio.intercambiable | -0.074098055 | 0.24309115 |
| Sodio.intercambiable | -0.003134633 | 0.21492681 |
| capacidad.de.intercambio.cationico | -0.280111684 | 0.27092567 |
| Conductividad.electrica | -0.110335892 | 0.28386641 |
| Hierro.olsen | 1.000000000 | -0.07777140 |
| Boro | -0.077771404 | 1.00000000 |

**Table C. 5.** Correlation matrix for all numerical attributes in the CORPOICA data set.

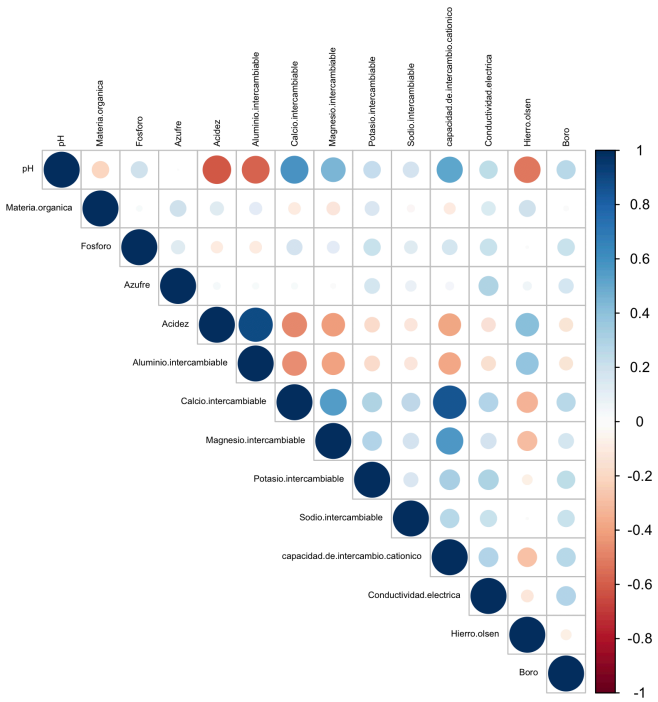



**Figure C. 2.** Correlated attributes for the CORPOICA dataset.

**Significance Levels**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Significance |
|---|---|---|---|---|---|
| (Intercept) | 7.223e+00 | 1.459e+00 | 4.952 | 7.35e-07 | *** |
| pH | -4.321e-04 | 2.150e-03 | -0.201 | 0.840719 | |
| Materia.organica | -3.435e-11 | 6.055e-11 | -0.567 | 0.570470 | |
| Fosforo | 2.750e-11 | 1.013e-11 | 2.716 | 0.006605 | ** |
| Azufre | 1.620e-10 | 5.265e-11 | 3.078 | 0.002085 | ** |
| Acidez | 1.110e-10 | 1.224e-09 | 0.091 | 0.927718 | |
| Aluminio.intercambiable | 1.279e-10 | 1.674e-09 | 0.076 | 0.939077 | |
| Calcio.intercambiable | 1.105e-11 | 3.678e-11 | 0.300 | 0.763807 | |
| Magnesio.intercambiable | -2.432e-10 | 9.200e-11 | -2.644 | 0.008197 | ** |
| Potasio.intercambiable | -1.484e-09 | 5.268e-10 | -2.817 | 0.004841 | ** |
| Sodio.intercambiable | -7.829e-09 | 1.246e-09 | -6.283 | 3.32e-10 | *** |
| capacidad.de.intercambio.cationico | -4.487e-11 | 2.834e-11 | -1.583 | 0.113386 | |
| Conductividad.electrica | 4.219e-10 | 9.076e-10 | 0.465 | 0.642033 | |
| Hierro.olsen | 9.240e-12 | 2.736e-12 | 3.377 | 0.000734 | *** |
| Boro | -1.951e-09 | 1.026e-09 | -1.902 | 0.057154 | . |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Logistic Regression**

| | Overall |
|---|---|
| pH | 0.20097340 |
| Materia.organica | 0.56735889 |
| Fosforo | 2.71613000 |
| Azufre | 3.07790742 |
| Acidez | 0.09071644 |
| Aluminio.intercambiable | 0.07642984 |
| Calcio.intercambiable | 0.30048520 |
| Magnesio.intercambiable | 2.64384295 |
| Potasio.intercambiable | 2.81741672 |
| Sodio.intercambiable | 6.28284677 |
| capacidad.de.intercambio.cationico | 1.58315467 |
| Conductividad.electrica | 0.46485788 |
| Hierro.olsen | 3.37656780 |
| Boro | 1.90213371 |

**Random Forest**

| | MeanDecreaseGini |
|---|---|
| pH | 1916.3444 |
| Materia.organica | 2009.5893 |
| Fosforo | 1805.4638 |
| Azufre | 1616.0617 |
| Acidez | 703.0391 |
| Aluminio.intercambiable | 580.0169 |
| Calcio.intercambiable | 1728.9131 |
| Magnesio.intercambiable | 1666.8800 |
| Potasio.intercambiable | 1728.8113 |
| Sodio.intercambiable | 1797.2777 |
| capacidad.de.intercambio.cationico | 1734.8684 |
| Conductividad.electrica | 1575.6146 |
| Hierro.olsen | 2150.7251 |
| Boro | 1614.7282 |

**Table C. 6.** Importance of variables using the logistic regression and Random Forest methods for the CORPOICA data set.

# C.3. IDEAM Data Source

### *Check Missing Values*

```
Departamento    Municipio         Anio          Mes       Temperatura    Precipitacion    Humedad.relativa   Radiacion
 Cauca:2042   El Tambo:  89   Min.   :2012   Abril  :184   Min.   :11.53   Min.   :  2.70   Min.   :64.02    Min.   : 34492
              Guachene:  89   1st Qu.:2013   Febrero:184   1st Qu.:16.60   1st Qu.: 89.77   1st Qu.:80.50    1st Qu.: 78588
              Jambalo :  89   Median :2015   Marzo  :184   Median :18.72   Median :165.25   Median :84.89    Median : 96329
              Miranda :  89   Mean   :2015   Enero  :183   Mean   :18.48   Mean   :196.75   Mean   :84.14    Mean   : 98836
              Morales :  89   3rd Qu.:2017   Mayo   :179   3rd Qu.:20.93   3rd Qu.:271.67   3rd Qu.:88.16    3rd Qu.:116543
              Padilla :  89   Max.   :2019   Agosto :161   Max.   :24.44   Max.   :826.89   Max.   :95.69    Max.   :197421
              (Other) :1508                  (Other):967                                                    NA's   :547

Departamento    Municipio         Anio          Mes       Temperatura    Precipitacion    Humedad.relativa   Radiacion
 Cauca:2042   El Tambo:  89   Min.   :2012   Abril  :184   Min.   :11.53   Min.   :  2.70   Min.   :64.02    Min.   : 34492
              Guachene:  89   1st Qu.:2013   Febrero:184   1st Qu.:16.60   1st Qu.: 89.77   1st Qu.:80.50    1st Qu.: 84766
              Jambalo :  89   Median :2015   Marzo  :184   Median :18.72   Median :165.25   Median :84.89    Median : 98836
              Miranda :  89   Mean   :2015   Enero  :183   Mean   :18.48   Mean   :196.75   Mean   :84.14    Mean   : 98836
              Morales :  89   3rd Qu.:2017   Mayo   :179   3rd Qu.:20.93   3rd Qu.:271.67   3rd Qu.:88.16    3rd Qu.:107502
              Padilla :  89   Max.   :2019   Agosto :161   Max.   :24.44   Max.   :826.89   Max.   :95.69    Max.   :197421
              (Other) :1508                  (Other):967
```

### *Outliers Detection*

Radiation was the only variable that presented outliers, however, these were not eliminated considering that the extreme values agree with the real behavior of this variable. On the other hand, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, no duplicate instances were found.

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes in this data set. Correlation matrix and correlogram are presented below.

| | Temperatura | Precipitacion | Humedad.relativa | Radiacion |
|---|---|---|---|---|
| Temperatura | 1.00000000 | -0.08023149 | -0.1161774 | 0.07657454 |
| Precipitacion | -0.08023149 | 1.00000000 | 0.3930722 | -0.42395653 |
| Humedad.relativa | -0.11617737 | 0.39307218 | 1.0000000 | -0.73151762 |
| Radiacion | 0.07657454 | -0.42395653 | -0.7315176 | 1.00000000 |

**Table C. 7.** Correlation matrix for all numerical attributes in the IDEAM data set.

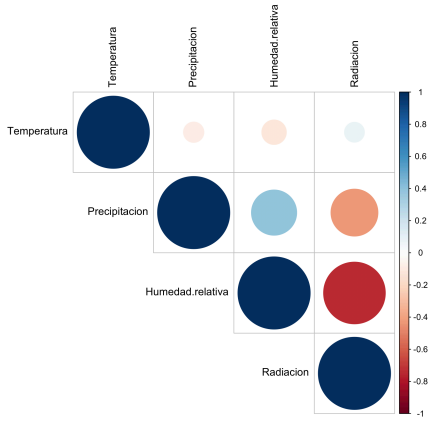



**Figure C. 3.** Correlated attributes for the IDEAM dataset.

# C.4. FINAGRO Data Source

### *Check Missing Values*

```
        Entidad                          Tipo.productor          Departamento          Anio          Valor
Banco Agrario:1842   Pequenos Productores                :1043   SANTANDER          : 178   Min.   :2004   Min.   :      2
Finagro      :2151   Medianos Productores                :1001   NORTE DE SANTANDER: 176   1st Qu.:2007   1st Qu.:   1186
                     Grandes Productores                 : 661   BOLIVAR            : 165   Median :2011   Median :   5777
                     Agremiaciones de Pequenos productores: 464   CESAR              : 164   Mean   :2011   Mean   :  32188
                     Agremiaciones medianos y grandes    : 351   NARI\x84O          : 164   3rd Qu.:2015   3rd Qu.:  27786
                     Otros Productores                   : 342   TOLIMA             : 159   Max.   :2019   Max.   :2765919
                     (Other)                             : 131   (Other)           :2987
```

### *Outliers Detection*

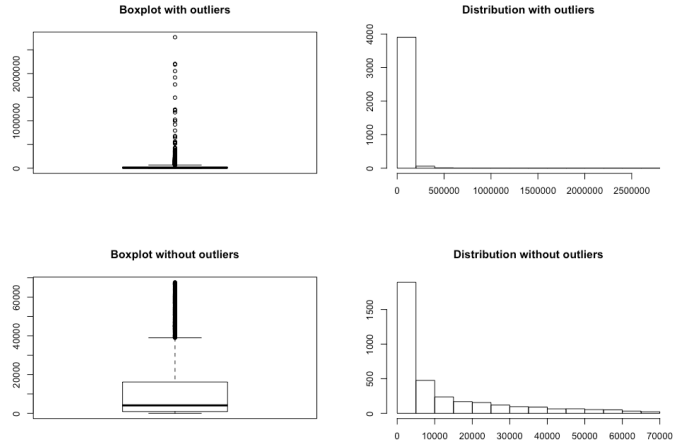
| Attribute Metrics | Distribution Charts |
|---|---|
| **Valor**<br>Outliers identified: 472<br>Proportion (%) of outliers: 13.4<br>Mean of the outliers: 187110.6<br>Mean without removing outliers: 32187.49<br>Mean if we remove outliers: 11419.62 |  |

**Table C. 8.** Results of the outlier detection process for each of the FINAGRO dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
        Entidad                          Tipo.productor          Departamento          Anio          Valor
Banco Agrario:1842   Pequenos Productores                :1043   SANTANDER          : 178   Min.   :2004   Min.   :    2.0
Finagro      :2151   Medianos Productores                :1001   NORTE DE SANTANDER: 176   1st Qu.:2007   1st Qu.:  915.1
                     Grandes Productores                 : 661   BOLIVAR            : 165   Median :2011   Median : 4077.3
                     Agremiaciones de Pequenos productores: 464   CESAR              : 164   Mean   :2011   Mean   :11419.6
                     Agremiaciones medianos y grandes    : 351   NARI\x84O          : 164   3rd Qu.:2015   3rd Qu.:16168.2
                     Otros Productores                   : 342   TOLIMA             : 159   Max.   :2019   Max.   :67654.4
                     (Other)                             : 131   (Other)           :2987                  NA's   :472
```

```
        Entidad                          Tipo.productor          Departamento          Anio          Valor
Banco Agrario:1842   Pequenos Productores                :1043   SANTANDER          : 178   Min.   :2004   Min.   :    2
Finagro      :2151   Medianos Productores                :1001   NORTE DE SANTANDER: 176   1st Qu.:2007   1st Qu.: 1186
                     Grandes Productores                 : 661   BOLIVAR            : 165   Median :2011   Median : 5777
                     Agremiaciones de Pequenos productores: 464   CESAR              : 164   Mean   :2011   Mean   :11420
                     Agremiaciones medianos y grandes    : 351   NARI\x84O          : 164   3rd Qu.:2015   3rd Qu.:12628
                     Otros Productores                   : 342   TOLIMA             : 159   Max.   :2019   Max.   :67654
                     (Other)                             : 131   (Other)           :2987
```

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes in this data set.

# C.5. AVA Data Source

### *Check Missing Values*

Missing values were not found in this data set.

```
          Department       Municipality    CA_Cocoa        CA_Coffee     CA_Sugar.Cane      CA_Bean         CA_Potato      CA_Banana
 CALDAS          :19   AGUADAS      : 1   Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.0   Min.   :  0.20
 CAUCA           :23   ALCALA       : 1   1st Qu.:  2.90   1st Qu.:17.90   1st Qu.:  0.00   1st Qu.: 12.95   1st Qu.: 0.0   1st Qu.: 66.60
 QUINDIO         :12   ANDALUCIA    : 1   Median : 17.40   Median :34.30   Median :  0.00   Median : 33.30   Median : 6.8   Median : 90.50
 RISARALDA       :12   ANSERMA      : 1   Mean   : 28.57   Mean   :35.94   Mean   : 17.89   Mean   : 39.82   Mean   :14.9   Mean   : 79.73
 VALLE DEL CAUCA:33   ANSERMANUEVO: 1   3rd Qu.: 50.20   3rd Qu.:49.60   3rd Qu.: 31.85   3rd Qu.: 65.55   3rd Qu.:25.4   3rd Qu.:100.00
                      APIA         : 1   Max.   :100.00   Max.   :98.90   Max.   :100.00   Max.   :100.00   Max.   :64.9   Max.   :100.00
                      (Other)      :93
```

### *Outliers Detection*

Outliers were not found in this data set.

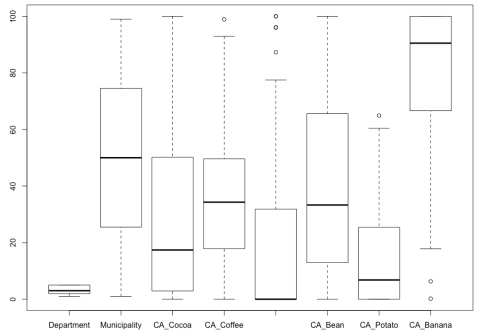



**Figure C. 4.** Boxplots for attributes in the AVA data set.

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes and instances in this data set.

| | CA_Cocoa | CA_Coffee | CA_Sugar.Cane | CA_Bean | CA_Potato | CA_Banana |
|---|---|---|---|---|---|---|
| CA_Cocoa | 1.00000000 | 0.3104672 | 0.05031136 | 0.07341722 | -0.5245350 | 0.5176518 |
| CA_Coffee | 0.31046719 | 1.0000000 | -0.40416649 | 0.44238637 | -0.4785407 | 0.5150858 |
| CA_Sugar.Cane | 0.05031136 | -0.4041665 | 1.00000000 | 0.09805982 | -0.3287699 | 0.3061735 |
| CA_Bean | 0.07341722 | 0.4423864 | 0.09805982 | 1.00000000 | -0.3868465 | 0.3873873 |
| CA_Potato | -0.52453504 | -0.4785407 | -0.32876992 | -0.38684648 | 1.0000000 | -0.9461025 |
| CA_Banana | 0.51765177 | 0.5150858 | 0.30617354 | 0.38738727 | -0.9461025 | 1.0000000 |

**Table C. 9.** Correlation matrix for all numerical attributes in the AVA data set.

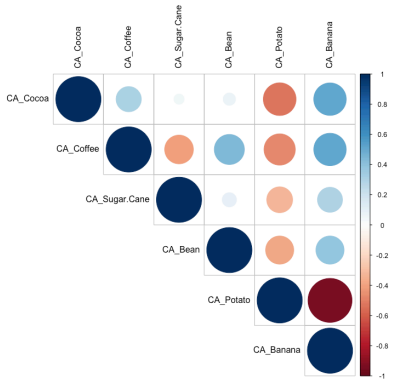



**Figure C. 5.** Correlated attributes for the AVA dataset.

# C.6. DANE-SIPSA-P Data Source

### *Check Missing Values*

```
      ANIO           CULTIVOS    SUPERFICIE.SEMBRADA SUPERFICIE.COSECHADA    PRODUCCION        RENDIMIENTO     PRECIO.AL.PRODUCTOR.KG
 Min.   :2011   AGUACATE :  5   Min.   :    4.00    Min.   :    4.0      Min.   :     2.0   Min.   :   2.0   Min.   : 150.0
 1st Qu.:2012   AHUYAMA  :  5   1st Qu.:   41.25    1st Qu.:   31.5      1st Qu.:   157.5   1st Qu.:  18.0   1st Qu.: 525.4
 Median :2013   AJI DULCE:  5   Median :  294.50    Median :  198.8      Median :   900.0   Median : 162.5   Median :1018.5
 Mean   :2013   BATATA   :  5   Mean   : 1197.76    Mean   :  987.1      Mean   :  9451.1   Mean   : 405.8   Mean   :1331.8
 3rd Qu.:2014   BERENJENA:  5   3rd Qu.: 1289.75    3rd Qu.:  930.5      3rd Qu.:  6690.3   3rd Qu.: 588.2   3rd Qu.:1700.0
 Max.   :2015   CACAO    :  5   Max.   :13541.00    Max.   :12837.0      Max.   :115533.0   Max.   :3932.0   Max.   :5864.0
                (Other)  :100

PRECIO.AL.PRODUCTOR.TON COSTO.PRODUCCION   INGRESO.BRUTO.PRODUCCION COSTO.TOTAL.PRODUCCION    UTILIDAD          RENTABILIDAD
 Min.   : 150000        Min.   : 1144400   Min.   :3.000e+06        Min.   :7.304e+06       Min.   :-6.861e+09   Min.   : -9496.0
 1st Qu.: 525417        1st Qu.: 2515358   1st Qu.:2.033e+08        1st Qu.:1.174e+08       1st Qu.: 1.574e+07   1st Qu.:   357.5
 Median :1018452        Median : 2987032   Median :1.083e+09        Median :6.332e+08       Median : 3.187e+08   Median :  5891.0
 Mean   :1331812        Mean   : 4007807   Mean   :6.113e+09        Mean   :2.589e+09       Mean   : 3.526e+09   Mean   : 12410.5
 3rd Qu.:1700000        3rd Qu.: 4626926   3rd Qu.:7.300e+09        3rd Qu.:2.855e+09       3rd Qu.: 3.106e+09   3rd Qu.: 20169.8
 Max.   :5864000        Max.   :40150000   Max.   :8.107e+10        Max.   :2.315e+10       Max.   : 5.792e+10   Max.   :107679.0
```

### *Outliers Detection*

Although the outlier detection step was fully executed, the distribution of the values of each attribute was analyzed in detail. It was observed that extreme values agree with the reality of agricultural production considering the great variety of crops and their different variations in their value chain. For this reason, these values were not discarded. In the same sense, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found.

### *Dimensionality Reduction*

| | SUPERFICIE.SEMBRADA | SUPERFICIE.COSECHADA | PRODUCCION | RENDIMIENTO | PRECIO.AL.PRODUCTOR.KG | PRECIO.AL.PRODUCTOR.TON | COSTO.PRODUCCION |
|---|---|---|---|---|---|---|---|
| SUPERFICIE.SEMBRADA | 1.00000000 | 0.9837627 | 0.9360272 | 0.1570235 | -0.21099414 | -0.21099416 | -0.15045370 |
| SUPERFICIE.COSECHADA | 0.98376271 | 1.0000000 | 0.9605785 | 0.1331964 | -0.21646680 | -0.21646680 | -0.15669211 |
| PRODUCCION | 0.93602717 | 0.9605785 | 1.0000000 | 0.2121702 | -0.28326026 | -0.28326024 | -0.15012883 |
| RENDIMIENTO | 0.15702353 | 0.1331964 | 0.2121702 | 1.0000000 | -0.12318093 | -0.12318092 | 0.34515722 |
| PRECIO.AL.PRODUCTOR.KG | -0.21099414 | -0.2164668 | -0.2832603 | -0.1231809 | 1.00000000 | 1.00000000 | 0.08083861 |
| PRECIO.AL.PRODUCTOR.TON | -0.21099416 | -0.2164668 | -0.2832602 | -0.1231809 | 1.00000000 | 1.00000000 | 0.08083858 |
| COSTO.PRODUCCION | -0.15045370 | -0.1566921 | -0.1501288 | 0.3451572 | 0.08083861 | 0.08083858 | 1.00000000 |
| INGRESO.BRUTO.PRODUCCION | 0.92052468 | 0.9008564 | 0.8460917 | 0.2028894 | -0.10595133 | -0.10595135 | -0.13953285 |
| COSTO.TOTAL.PRODUCCION | 0.96421073 | 0.9546736 | 0.9288406 | 0.1875389 | -0.21249703 | -0.21249706 | -0.11900691 |
| UTILIDAD | 0.81921646 | 0.7954463 | 0.7288204 | 0.1945821 | -0.03695427 | -0.03695428 | -0.13963420 |
| RENTABILIDAD | 0.05263003 | 0.0437932 | 0.0274576 | 0.1521889 | 0.31251143 | 0.31251134 | -0.17776179 |

| | INGRESO.BRUTO.PRODUCCION | COSTO.TOTAL.PRODUCCION | UTILIDAD | RENTABILIDAD |
|---|---|---|---|---|
| SUPERFICIE.SEMBRADA | 0.9205247 | 0.96421073 | 0.81921646 | 0.05263003 |
| SUPERFICIE.COSECHADA | 0.9008564 | 0.95467359 | 0.79544634 | 0.04379320 |
| PRODUCCION | 0.8460917 | 0.92884056 | 0.72882040 | 0.02745760 |
| RENDIMIENTO | 0.2028894 | 0.18753892 | 0.19458212 | 0.15218893 |
| PRECIO.AL.PRODUCTOR.KG | -0.1059513 | -0.21249703 | -0.03695427 | 0.31251143 |
| PRECIO.AL.PRODUCTOR.TON | -0.1059514 | -0.21249706 | -0.03695428 | 0.31251134 |
| COSTO.PRODUCCION | -0.1395328 | -0.11900691 | -0.13963420 | -0.17776179 |
| INGRESO.BRUTO.PRODUCCION | 1.0000000 | 0.90533055 | 0.97054826 | 0.21605791 |
| COSTO.TOTAL.PRODUCCION | 0.9053306 | 1.00000000 | 0.77635566 | 0.02764654 |
| UTILIDAD | 0.9705483 | 0.77635566 | 1.00000000 | 0.30485375 |
| RENTABILIDAD | 0.2160579 | 0.02764654 | 0.30485375 | 1.00000000 |

**Table C. 10.** Correlation matrix for all numerical attributes in the DANE-SIPSA-P data set.

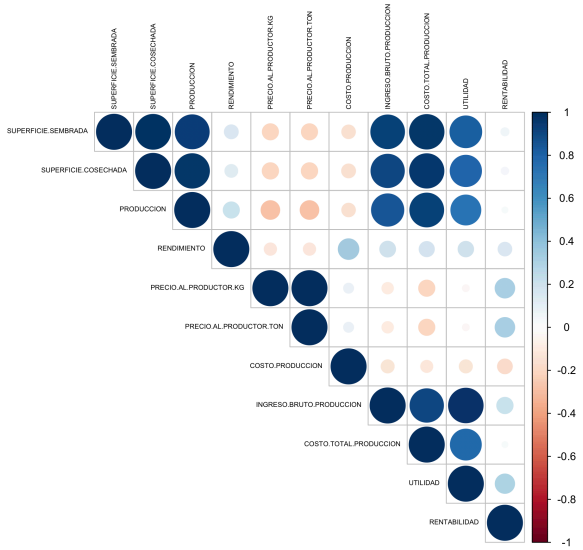



**Figure C. 6.** Correlated attributes for the DANE-SIPSA-P dataset.

**Significance Levels**

```
                           Estimate Std. Error z value Pr(>|z|) Significance
(Intercept)                1.126e+00  2.155e+00   0.522    0.601
SUPERFICIE.SEMBRADA        8.871e-03  9.366e-03   0.947    0.344
SUPERFICIE.COSECHADA      -1.510e-02  1.130e-02  -1.336    0.182
PRODUCCION                 1.477e-03  7.043e-04   2.096    0.036 *
RENDIMIENTO               -2.071e-03  1.542e-03  -1.343    0.179
PRECIO.AL.PRODUCTOR.KG    -2.512e+02  5.430e+02  -0.463    0.644
PRECIO.AL.PRODUCTOR.TON    2.512e-01  5.430e-01   0.463    0.644
COSTO.PRODUCCION           3.360e-07  7.588e-07   0.443    0.658
INGRESO.BRUTO.PRODUCCION  -7.469e-07  1.656e-04  -0.005    0.996
COSTO.TOTAL.PRODUCCION     7.463e-07  1.656e-04   0.005    0.996
UTILIDAD                   7.459e-07  1.656e-04   0.005    0.996
RENTABILIDAD              -4.304e-05  3.571e-05  -1.206    0.228
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Logistic Regression**

| | Overall |
|---|---|
| SUPERFICIE.SEMBRADA | 0.947161199 |
| SUPERFICIE.COSECHADA | 1.335997912 |
| PRODUCCION | 2.096474227 |
| RENDIMIENTO | 1.343297555 |
| PRECIO.AL.PRODUCTOR.KG | 0.462591049 |
| PRECIO.AL.PRODUCTOR.TON | 0.462596291 |
| COSTO.PRODUCCION | 0.442748015 |
| INGRESO.BRUTO.PRODUCCION | 0.004511052 |
| COSTO.TOTAL.PRODUCCION | 0.004507531 |
| UTILIDAD | 0.004504702 |
| RENTABILIDAD | 1.205558209 |

**Random Forest**

| | MeanDecreaseGini |
|---|---|
| SUPERFICIE.SEMBRADA | 16.669002 |
| SUPERFICIE.COSECHADA | 11.688769 |
| PRODUCCION | 11.328599 |
| RENDIMIENTO | 7.142747 |
| PRECIO.AL.PRODUCTOR.KG | 12.526532 |
| PRECIO.AL.PRODUCTOR.TON | 11.977799 |
| COSTO.PRODUCCION | 18.946370 |
| INGRESO.BRUTO.PRODUCCION | 9.056522 |
| COSTO.TOTAL.PRODUCCION | 11.884720 |
| UTILIDAD | 7.575780 |
| RENTABILIDAD | 5.657407 |

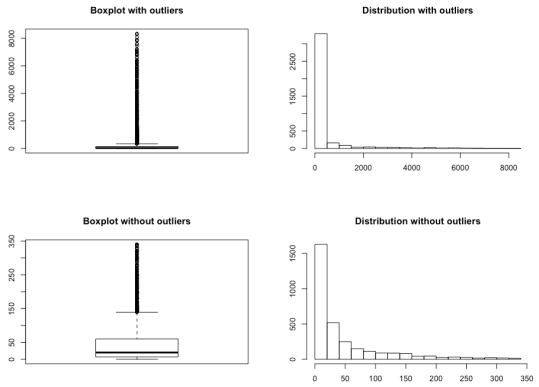
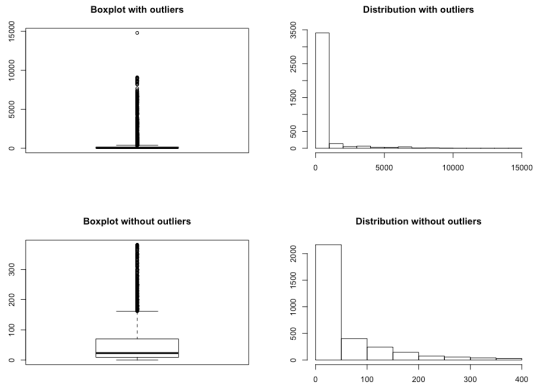
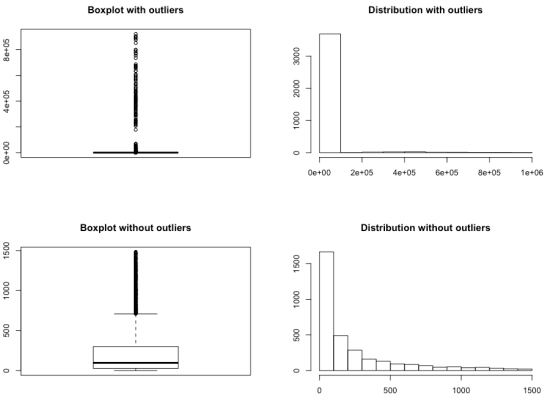
**Table C. 11.** Importance of variables using the logistic regression and Random Forest methods for the DANE-SIPSA-P data set.

# C.7. Agronet Data Source

### *Check Missing Values*

```
Departamento          Cultivo          Anio          Municipio    Area.Cosechada    Area.Sembrada      Produccion      Rendimiento
 Cauca:3789   Maiz Tradicional: 282   Min.   :2002   Jambaló: 160   Min.   :   0.0   Min.   :   0.0   Min.   :     0   Min.   :  0.00
              Frijol          : 261   1st Qu.:2009   Miranda: 150   1st Qu.:   9.0   1st Qu.:  10.0   1st Qu.:    35   1st Qu.:  1.40
              Cafe            : 255   Median :2011   Patía  : 148   Median :  29.5   Median :  34.0   Median :   140   Median :  4.00
              Yuca            : 250   Mean   :2011   Sotará : 145   Mean   : 365.5   Mean   : 433.6   Mean   : 13538   Mean   : 10.72
              Platano         : 248   3rd Qu.:2013   Toribio: 143   3rd Qu.: 141.8   3rd Qu.: 159.0   3rd Qu.:   615   3rd Qu.: 10.00
              Tomate          : 145   Max.   :2014   Silvia : 142   Max.   :8343.0   Max.   :14795.0   Max.   :921841   Max.   :171.40
              (Other)         :2348                  (Other):2901
```

## *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **Area.Cosechada**<br>Outliers identified: 629<br>Proportion (%) of outliers: 19.9<br>Mean of the outliers: 1957.69<br>Mean without removing outliers: 365.51<br>Mean if we remove outliers: 48.58 |  |
| **Area.Sembrada**<br>Outliers identified: 628<br>Proportion (%) of outliers: 19.9<br>Mean of the outliers: 2333.65<br>Mean without removing outliers: 433.62<br>Mean if we remove outliers: 56.14 |  |
| **Produccion**<br>Outliers identified: 546<br>Proportion (%) of outliers: 16.8<br>Mean of the outliers: 92547.72<br>Mean without removing outliers: 13537.48<br>Mean if we remove outliers: 235.1 |  |

**Rendimiento**
Outliers identified: 305
Proportion (%) of outliers: 8.8
Mean of the outliers: 72.23
Mean without removing outliers: 10.72
Mean if we remove outliers: 5.34

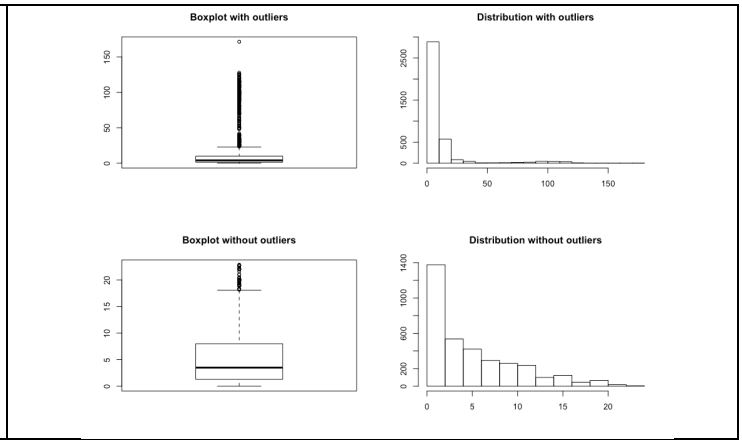



**Table C. 12.** Results of the outlier detection process for each of the Agronet dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
Departamento            Cultivo         Anio         Municipio    Area.Cosechada    Area.Sembrada      Produccion      Rendimiento
 Cauca:3789   Maiz Tradicional: 282   Min.   :2002   Jambaló: 160   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0   Min.   : 0.000
              Frijol          : 261   1st Qu.:2009   Miranda: 150   1st Qu.:  7.00   1st Qu.:  9.00   1st Qu.:  27.0   1st Qu.: 1.280
              Cafe            : 255   Median :2011   Patía  : 148   Median : 20.00   Median : 23.00   Median :  96.0   Median : 3.500
              Yuca            : 250   Mean   :2011   Sotará : 145   Mean   : 48.58   Mean   : 56.14   Mean   : 235.1   Mean   : 5.337
              Platano         : 248   3rd Qu.:2013   Toribio: 143   3rd Qu.: 60.00   3rd Qu.: 70.00   3rd Qu.: 300.0   3rd Qu.: 8.000
              Tomate          : 145   Max.   :2014   Silvia : 142   Max.   :340.00   Max.   :381.90   Max.   :1484.0   Max.   :22.850
              (Other)         :2348                  (Other):2901   NA's   :629      NA's   :628      NA's   :546      NA's   :305

Departamento            Cultivo         Anio         Municipio    Area.Cosechada    Area.Sembrada      Produccion      Rendimiento
 Cauca:3789   Maiz Tradicional: 282   Min.   :2002   Jambaló: 160   Min.   :  0.00   Min.   :  0.0    Min.   :   0.0   Min.   : 0.000
              Frijol          : 261   1st Qu.:2009   Miranda: 150   1st Qu.:  9.00   1st Qu.: 10.0    1st Qu.:  35.0   1st Qu.: 1.400
              Cafe            : 255   Median :2011   Patía  : 148   Median : 29.50   Median : 34.0    Median : 140.0   Median : 4.000
              Yuca            : 250   Mean   :2011   Sotará : 145   Mean   : 47.68   Mean   : 55.0    Mean   : 227.2   Mean   : 5.348
              Platano         : 248   3rd Qu.:2013   Toribio: 143   3rd Qu.: 46.70   3rd Qu.: 55.0    3rd Qu.: 245.0   3rd Qu.: 7.950
              Tomate          : 145   Max.   :2014   Silvia : 142   Max.   :340.00   Max.   :381.9    Max.   :1484.0   Max.   :22.850
              (Other)         :2348                  (Other):2901
```

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes and instances in this data set. In the same sense, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found. Correlation matrix and correlogram are presented below.

| | Area.Cosechada | Area.Sembrada | Produccion | Rendimiento |
|---|---|---|---|---|
| Area.Cosechada | 1.0000000 | 0.85777472 | 0.3307419 | -0.05106310 |
| Area.Sembrada | 0.8577747 | 1.00000000 | 0.3588835 | -0.04345484 |
| Produccion | 0.3307419 | 0.35888346 | 1.0000000 | 0.17370238 |
| Rendimiento | -0.0510631 | -0.04345484 | 0.1737024 | 1.00000000 |

**Table C. 13.** Correlation matrix for all numerical attributes in the Agronet data set.

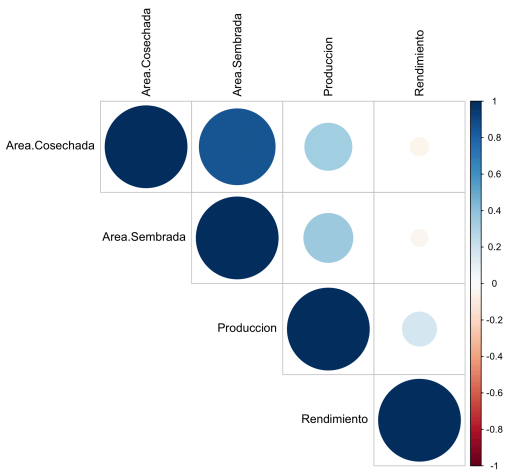


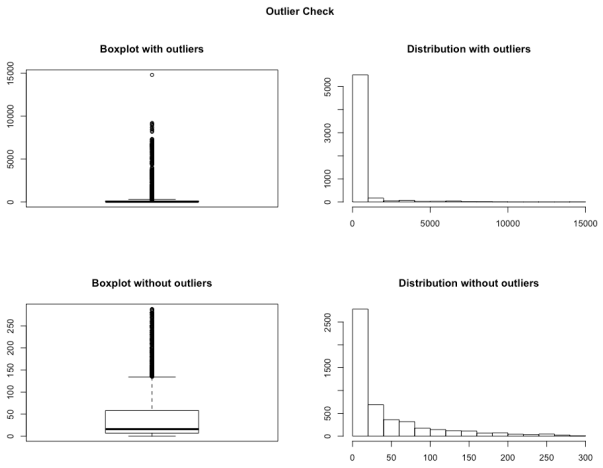
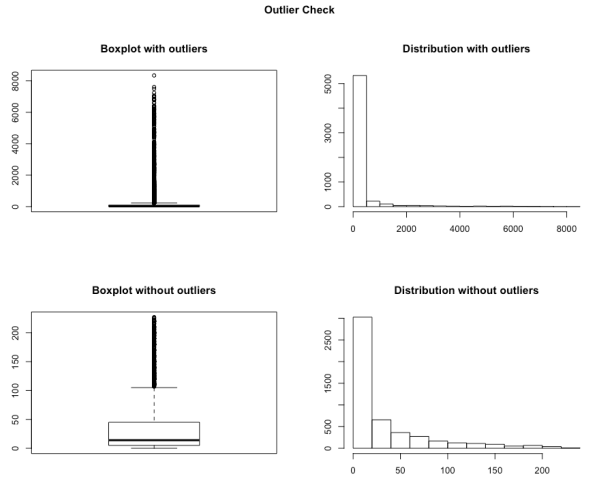
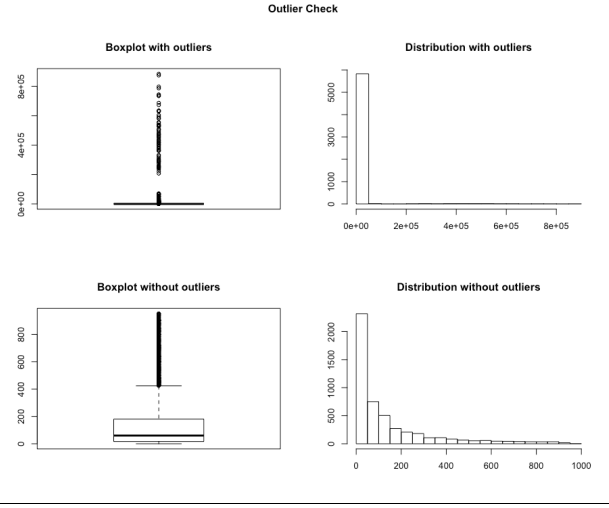

**Figure C. 7.** Correlated attributes for the Agronet dataset.

# C.8. Minagricultura Data Source

### *Check Missing Values*

```
   COD.DEP    DEPARTAMENTO     COD.MUN         MUNICIPIO                    GRUPO.CULTIVO  SUBGRUPO.CULTIVO
Min.   :19   CAUCA:5911    Min.   :19001   PATIA  : 240   FRUTALES              :1138   MAIZ    : 734
1st Qu.:19                 1st Qu.:19256   SILVIA : 240   CEREALES              :1015   FRIJOL  : 528
Median :19                 Median :19455   MIRANDA: 230   HORTALIZAS            :1012   TOMATE  : 402
Mean   :19                 Mean   :19462   TORIBIO: 230   LEGUMINOSAS           : 915   CAÑA    : 354
3rd Qu.:19                 3rd Qu.:19743   JAMBALO: 229   TUBERCULOS Y PLATANOS: 868   CAFE    : 286
Max.   :19                 Max.   :19845   SOTARA : 223   OTROS PERMANENTES     : 776   PLATANO: 278
                                           (Other):4519   (Other)               : 187   (Other):3329

   CULTIVO            SISTEMA.PRODUCTIVO  COD.CULTIVO                      NOMBRE..CIENTIFICO     PERIODO
MAIZ    : 734   MAIZ TRADICIONAL: 555     Min.   :1.110e+11   ZEA MAYS               : 734     2014   : 310
FRIJOL  : 528   FRIJOL          : 528     1st Qu.:1.110e+11   PHASEOLUS VULGARIS     : 528     2015   : 310
TOMATE  : 402   CAFE            : 286     Median :1.110e+11   LYCOPERSICUM ESCULETUM : 402     2013   : 302
CAFE    : 286   TOMATE          : 285     Mean   :1.115e+11   SACCHARUM OFFICINARUM  : 354     2011   : 293
PLATANO: 278    PLATANO         : 278     3rd Qu.:1.120e+11   COFFEA ARABICA         : 286     2012   : 288
YUCA    : 274   YUCA            : 274     Max.   :1.130e+11   MUSA X PARADISIACA     : 278     2010   : 287
(Other):3409    (Other)         :3705                         (Other)                :3329     (Other):4121

AREA.SEMBRADA      AREA.COSECHADA     PRODUCCION        RENDIMIENTO          ESTADO.FISICO.PRODUCCION
Min.   :    0.09   Min.   :   0.0   Min.   :     0.0   Min.   :  0.060   GRANO SECO      :1745
1st Qu.:    8.00   1st Qu.:   6.0   1st Qu.:    22.5   1st Qu.:  1.400   FRUTO FRESCO    :1434
Median :   25.00   Median :  20.0   Median :    95.2   Median :  3.771   HORTALIZA FRESCA: 994
Mean   :  305.20   Mean   : 253.7   Mean   :  6437.8   Mean   :  8.840   TUBERCULO FRESCO: 591
3rd Qu.:  120.00   3rd Qu.:  95.1   3rd Qu.:   394.0   3rd Qu.:  9.000   PERGAMINO SECO  : 286
Max.   :14795.00   Max.   :8343.0   Max.   :885190.0   Max.   :171.400   PANELA          : 269
                                                      NA's   :49        (Other)         : 592
```

## *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **AREA.SEMBRADA**<br>Outliers identified: 918<br>Proportion (%) of outliers: 18.4<br>Mean of the outliers: 1732.83<br>Mean without removing outliers: 305.2<br>Mean if we remove outliers: 42.72 |  |
| **AREA.COSECHADA**<br>Outliers identified: 932<br>Proportion (%) of outliers: 18.7<br>Mean of the outliers: 1425.56<br>Mean without removing outliers: 253.69<br>Mean if we remove outliers: 34.33 |  |
| **PRODUCCION**<br>Outliers identified: 944<br>Proportion (%) of outliers: 19<br>Mean of the outliers: 39562.21<br>Mean without removing outliers: 6437.81<br>Mean if we remove outliers: 142.37 |  |

**RENDIMIENTO**
Outliers identified: 419
Proportion (%) of outliers: 7.6
Mean of the outliers: 59.07
Mean without removing outliers: 8.84
Mean if we remove outliers: 5.01

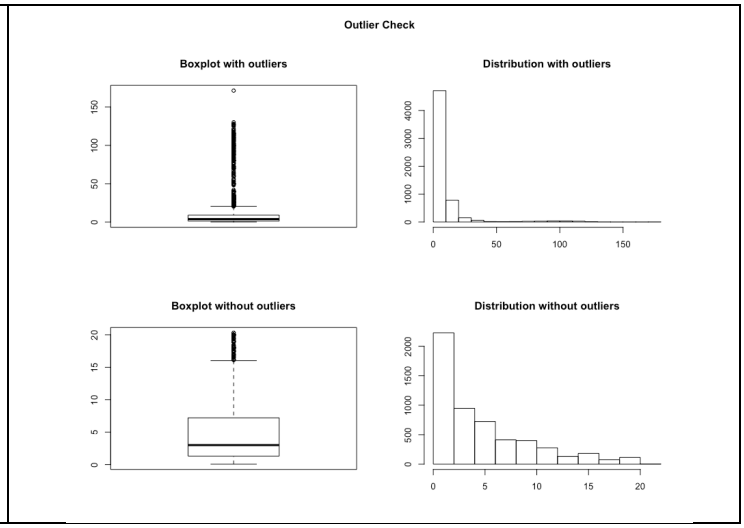



**Table C. 14.** Results of the outlier detection process for each of the Minagricultura dataset attributes.

```
   COD.DEP    DEPARTAMENTO     COD.MUN         MUNICIPIO                  GRUPO.CULTIVO  SUBGRUPO.CULTIVO
Min.   :19   CAUCA:5911    Min.   :19001   PATIA  : 240   FRUTALES              :1138   MAIZ    : 734
1st Qu.:19                 1st Qu.:19256   SILVIA : 240   CEREALES              :1015   FRIJOL  : 528
Median :19                 Median :19455   MIRANDA: 230   HORTALIZAS            :1012   TOMATE  : 402
Mean   :19                 Mean   :19462   TORIBIO: 230   LEGUMINOSAS           : 915   CAÑA    : 354
3rd Qu.:19                 3rd Qu.:19743   JAMBALO: 229   TUBERCULOS Y PLATANOS: 868   CAFE    : 286
Max.   :19                 Max.   :19845   SOTARA : 223   OTROS PERMANENTES     : 776   PLATANO: 278
                                           (Other):4519   (Other)               : 187   (Other):3329

   CULTIVO            SISTEMA.PRODUCTIVO  COD.CULTIVO                    NOMBRE..CIENTIFICO     PERIODO
MAIZ   : 734   MAIZ TRADICIONAL: 555     Min.   :1.110e+11   ZEA MAYS               : 734     2014   : 310
FRIJOL : 528   FRIJOL          : 528     1st Qu.:1.110e+11   PHASEOLUS VULGARIS     : 528     2015   : 310
TOMATE : 402   CAFE            : 286     Median :1.110e+11   LYCOPERSICUM ESCULETUM : 402     2013   : 302
CAFE   : 286   TOMATE          : 285     Mean   :1.115e+11   SACCHARUM OFFICINARUM  : 354     2011   : 293
PLATANO: 278   PLATANO         : 278     3rd Qu.:1.120e+11   COFFEA ARABICA         : 286     2012   : 288
YUCA   : 274   YUCA            : 274     Max.   :1.130e+11   MUSA X PARADISIACA     : 278     2010   : 287
(Other):3409   (Other)         :3705                         (Other)                :3329     (Other):4121

AREA.SEMBRADA    AREA.COSECHADA     PRODUCCION     RENDIMIENTO        ESTADO.FISICO.PRODUCCION
Min.   :  0.09   Min.   :  0.00   Min.   :  0.0   Min.   : 0.060   GRANO SECO      :1745
1st Qu.:  7.00   1st Qu.:  5.00   1st Qu.: 16.8   1st Qu.: 1.305   FRUTO FRESCO    :1434
Median : 16.00   Median : 14.00   Median : 60.0   Median : 3.000   HORTALIZA FRESCA: 994
Mean   : 42.72   Mean   : 34.33   Mean   :142.4   Mean   : 5.008   TUBERCULO FRESCO: 591
3rd Qu.: 58.00   3rd Qu.: 45.00   3rd Qu.:180.0   3rd Qu.: 7.200   PERGAMINO SECO  : 286
Max.   :288.00   Max.   :227.00   Max.   :951.0   Max.   :20.312   PANELA          : 269
NA's   :918      NA's   :932      NA's   :944     NA's   :419      (Other)         : 592
```

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
   COD.DEP    DEPARTAMENTO     COD.MUN         MUNICIPIO                  GRUPO.CULTIVO  SUBGRUPO.CULTIVO
Min.   :19   CAUCA:5911    Min.   :19001   PATIA  : 240   FRUTALES              :1138   MAIZ    : 734
1st Qu.:19                 1st Qu.:19256   SILVIA : 240   CEREALES              :1015   FRIJOL  : 528
Median :19                 Median :19455   MIRANDA: 230   HORTALIZAS            :1012   TOMATE  : 402
Mean   :19                 Mean   :19462   TORIBIO: 230   LEGUMINOSAS           : 915   CAÑA    : 354
3rd Qu.:19                 3rd Qu.:19743   JAMBALO: 229   TUBERCULOS Y PLATANOS: 868   CAFE    : 286
Max.   :19                 Max.   :19845   SOTARA : 223   OTROS PERMANENTES     : 776   PLATANO: 278
                                           (Other):4519   (Other)               : 187   (Other):3329
   CULTIVO            SISTEMA.PRODUCTIVO  COD.CULTIVO                    NOMBRE..CIENTIFICO     PERIODO
MAIZ   : 734   MAIZ TRADICIONAL: 555     Min.   :1.110e+11   ZEA MAYS               : 734     2014   : 310
FRIJOL : 528   FRIJOL          : 528     1st Qu.:1.110e+11   PHASEOLUS VULGARIS     : 528     2015   : 310
TOMATE : 402   CAFE            : 286     Median :1.110e+11   LYCOPERSICUM ESCULETUM : 402     2013   : 302
CAFE   : 286   TOMATE          : 285     Mean   :1.115e+11   SACCHARUM OFFICINARUM  : 354     2011   : 293
PLATANO: 278   PLATANO         : 278     3rd Qu.:1.120e+11   COFFEA ARABICA         : 286     2012   : 288
YUCA   : 274   YUCA            : 274     Max.   :1.130e+11   MUSA X PARADISIACA     : 278     2010   : 287
(Other):3409   (Other)         :3705                         (Other)                :3329     (Other):4121
```

```
 AREA.SEMBRADA     AREA.COSECHADA      PRODUCCION      RENDIMIENTO         ESTADO.FISICO.PRODUCCION
 Min.   :  0.09   Min.   :  0.00   Min.   :  0.0   Min.   : 0.060   GRANO SECO      :1745
 1st Qu.:  8.00   1st Qu.:  6.00   1st Qu.: 22.5   1st Qu.: 1.429   FRUTO FRESCO    :1434
 Median : 25.00   Median : 20.00   Median : 95.2   Median : 3.900   HORTALIZA FRESCA: 994
 Mean   : 41.86   Mean   : 33.73   Mean   :137.4   Mean   : 5.003   TUBERCULO FRESCO: 591
 3rd Qu.: 45.00   3rd Qu.: 34.00   3rd Qu.:144.0   3rd Qu.: 7.000   PERGAMINO SECO  : 286
 Max.   :288.00   Max.   :227.00   Max.   :951.0   Max.   :20.312   PANELA          : 269
                                                                    (Other)         : 592
```

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes and instances in this data set. In the same sense, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found. Correlation matrix and correlogram are presented below.

| | AREA.SEMBRADA | AREA.COSECHADA | PRODUCCION | RENDIMIENTO |
|---|---|---|---|---|
| AREA.SEMBRADA | 1.00000000 | 0.83529036 | 0.3450326 | -0.05408425 |
| AREA.COSECHADA | 0.83529036 | 1.00000000 | 0.3567647 | -0.06693261 |
| PRODUCCION | 0.34503258 | 0.35676471 | 1.0000000 | 0.18009309 |
| RENDIMIENTO | -0.05408425 | -0.06693261 | 0.1800931 | 1.00000000 |

**Table C. 15.** Correlation matrix for all numerical attributes in the Minagricultura data set.

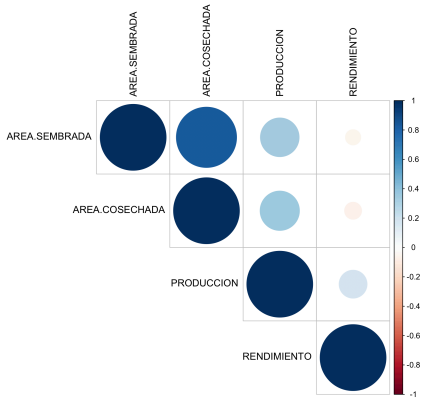



**Figure C. 8.** Correlated attributes for the Minagricultura dataset.

## C.9. Agronet-P Data Source

### *Check Missing Values*

```
                Fecha                 Producto         Precio                                  Unidad
 Abril 30 de 2007  :   10   Algodón     :3995   Min.   :  12.91   Centavos de dólar por bushel:  306
 Agosto 31 de 2009 :   10   Cacao       :3246   1st Qu.: 141.90   Ctvs US/lb                  : 6985
 Enero 31 de 2007  :   10   Café        :2990   Median : 187.50   Dólares por tonelada        :  266
 Enero 31 de 2008  :   10   Trigo HRW   :2950   Mean   : 503.18   US/ton                      : 3246
 Enero 31 de 2010  :   10   Soya        :2944   3rd Qu.: 319.43   US/Ton                      :14425
 Febrero 28 de 2007:   10   Maíz Blanco:2939   Max.   :3730.25
 (Other)           :25168   (Other)     :6164
   Variacion                                  Fuente
 Min.   : -94.0000   Bolsa de Chicago           : 2657
 1st Qu.:  -1.0000   Bolsa de Chicago via Reuters: 5879
 Median :   0.0000   Bolsa de Kansas via Reuters : 2950
 Mean   :   0.1146   Bolsa de Nueva York        :10231
 3rd Qu.:   1.0000   Cerrd Rice co Inc          :  266
 Max.   :1018.0000   USDA via Reuters           : 3245
```

### *Outliers Detection*

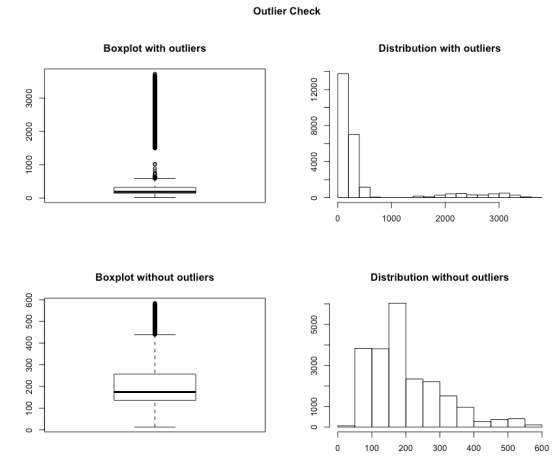
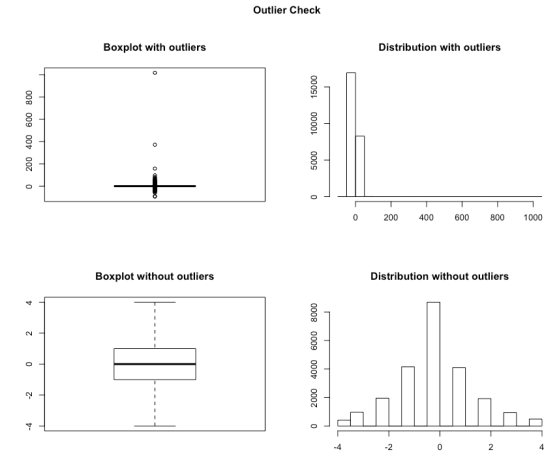
| Attribute Metrics | Distribution Charts |
|---|---|
| **Precio**<br>Outliers identified: 3321<br>Proportion (%) of outliers: 15.2<br>Mean of the outliers: 2508.69<br>Mean without removing outliers: 503.18<br>Mean if we remove outliers: 199.16 |  |
| **Variacion**<br>Outliers identified: 1609<br>Proportion (%) of outliers: 6.8<br>Mean of the outliers: 1.72<br>Mean without removing outliers: 0.11<br>Mean if we remove outliers: 0.01 |  |

**Table C. 16.** Results of the outlier detection process for each of the Agronet-P dataset attributes.

In the same sense, it was not necessary to perform the label correction process, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since the classes were balanced in an acceptable way. Finally, no duplicate instances were found.

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes in this data set.

## C.10. DANE-SIPSA Data Source

### *Check Missing Values*

```
   Fecha                                  Grupo                      Producto
dic-16 :  2986   FRUTAS                      :71589   Pimentón            :  2586
nov-16 :  2977   GRANOS Y CEREALES           :18463   Tomate de árbol     :  2573
mar-17 :  2910   TUBERCULOS RAICES Y PLATANOS:23121   Habichuela          :  2476
oct-16 :  2902   VERDURAS Y HORTALIZAS       :56945   Mora de Castilla    :  2442
jun-17 :  2894                                        Chócolo mazorca     :  2322
abr-17 :  2893                                        Plátano hartón verde:  2310
(Other):152556                                        (Other)             :155409

                                    Fuente           Precio
Medellín Central Mayorista de Antioquia:  7734   Min.   :   82
Bogotá DC Corabastos                   :  6982   1st Qu.:  964
Bucaramanga Centroabastos              :  5788   Median : 1542
Tunja Complejo de Servicios del Sur    :  5519   Mean   : 2038
Cúcuta Cenabastos                      :  4982   3rd Qu.: 2650
Neiva Surabastos                       :  4918   Max.   :18667
(Other)                                :134195
```

### *Outliers Detection*

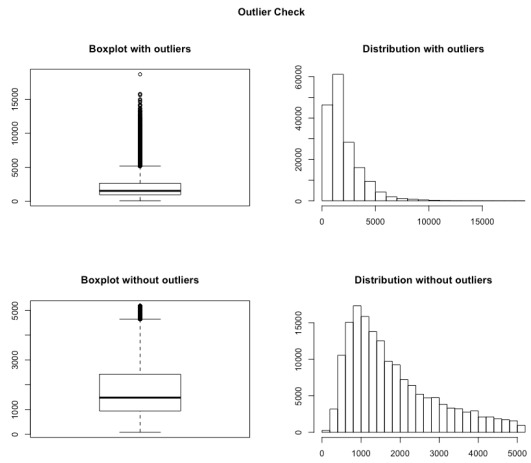
| Attribute Metrics | Distribution Charts |
|---|---|
| **Precio**<br>Outliers identified: 7904<br>Proportion (%) of outliers: 4.9<br>Mean of the outliers: 6783.05<br>Mean without removing outliers: 2038.27<br>Mean if we remove outliers: 1807.07 |  |

**Table C. 17.** Results of the outlier detection process for each of the DANE-SIPSA dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
    Fecha                                  Grupo                      Producto
dic-16 :  2986   FRUTAS                      :71589   Pimentón            :  2586
nov-16 :  2977   GRANOS Y CEREALES           :18463   Tomate de árbol     :  2573
mar-17 :  2910   TUBERCULOS RAICES Y PLATANOS:23121   Habichuela          :  2476
oct-16 :  2902   VERDURAS Y HORTALIZAS       :56945   Mora de Castilla    :  2442
jun-17 :  2894                                        Chócolo mazorca     :  2322
abr-17 :  2893                                        Plátano hartón verde:  2310
(Other):152556                                        (Other)             :155409

                                    Fuente           Precio
Medellín Central Mayorista de Antioquia:  7734   Min.   :  82
Bogotá DC Corabastos                   :  6982   1st Qu.: 964
Bucaramanga Centroabastos              :  5788   Median :1542
Tunja Complejo de Servicios del Sur    :  5519   Mean   :1807
Cúcuta Cenabastos                      :  4982   3rd Qu.:2354
Neiva Surabastos                       :  4918   Max.   :5179
(Other)                                :134195
```

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering that this data set only has an attribute of numerical type.

# C.11. DNP-AIB Data Source

### *Check Missing Values*

```
      ANIO                          TIPO.ENTIDAD    COD.ENTIDAD            ENTIDAD
Min.   :2005   CAR (Fuente MADS)        : 1696   CVC000 :  255   CVC          :  255
1st Qu.:2008   DEPARTAMENTO (Fuente FUT):  426   321600 :   94   CORPONARIÑO  :   94
Median :2009   MUNICIPIO (Fuente FUT)   :17016   322200 :   79   Rionegro     :   93
Mean   :2010   NACIONAL (Fuente SIIF)   :  148   CORPOU :   74   CORPAMAG     :   79
3rd Qu.:2011                                     05615  :   70   CORPOURABA   :   74
Max.   :2013                                     320101 :   70   San Francisco:   74
                                                 (Other):18644   (Other)      :18617
                                                                                                                                                  CUENTA.PROYECTO
11014  CONSERVACIÓN PROTECCIÓN RESTAURACIÓN Y APROVECHAMIENTO DE RECURSOS NATURALES Y DEL MEDIO AMBIENTE                                                 :2320
11012  EDUCACIÓN AMBIENTAL NO FORMAL                                                                                                                    :1697
11011  CONSERVACIÓN DE MICROCUENCAS QUE ABASTECEN EL ACUEDUCTO PROTECCIÓN DE FUENTES Y REFORESTACIÓN DE DICHAS CUENCAS:1509
1104  REFORESTACIÓN Y CONTROL DE EROSIÓN                                                                                                                :1448
11215  CONTRATOS CELEBRADOS CON CUERPOS DE BOMBEROS PARA LA PREVENCIÓN Y CONTROL DE INCENDIOS                                                            :1311
11210  FORTALECIMIENTO DE LOS COMITÉS DE PREVENCIÓN Y ATENCIÓN DE DESASTRES                                                                              :1103
(Other)                                                                                                                                                 :9898
                                SECTOR                                      OBJETIVO        VALOR
AGUA POTABLE Y SANEAMIENTO BÁSICO:   78   ADAPTACIÓN                           :7586   Min.   :        1
AMBIENTE Y DESARROLLO SOSTENIBLE :12435   ADAPTACIÓN - DESARROLLO              :2454   1st Qu.:     6360
GESTIÓN DEL RIESGO               : 6766   MITIGACIÓN                          : 952   Median :    16787
TRANSPORTE                       :    3   MITIGACIÓN - ADAPTACIÓN             :7505   Mean   :   252086
VIVIENDA                         :    4   MITIGACIÓN - ADAPTACIÓN - DESARROLLO: 760   3rd Qu.:    51033
                                          MITIGACIÓN - DESARROLLO              :  29   Max.   :241561456
```

### *Outliers Detection*

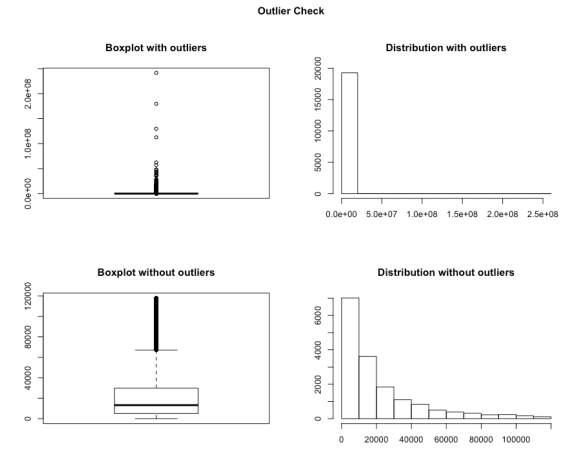
| Attribute Metrics | Distribution Charts |
|---|---|
| **VALOR**<br>Outliers identified: 2927<br>Proportion (%) of outliers: 17.9<br>Mean of the outliers: 1537110<br>Mean without removing outliers: 252085.7<br>Mean if we remove outliers: 22165.48 |  |

**Table C. 18.** Results of the outlier detection process for each of the DNP-AIB dataset attributes.

### *Removal of Duplicate Instances*

In this data set, 32 duplicate instances were found which were removed. Through this step, the data set remained with 19286 instances of the initial 19253.

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes in this data set. On the other hand, correlation matrix and

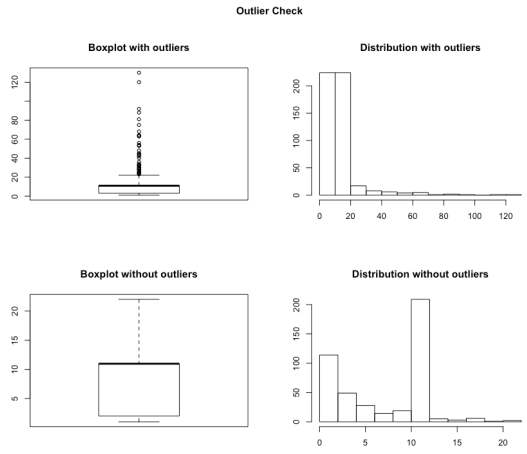
correlogram are not presented considering that there is only a single numeric attribute.

# C.12. DNP-FI Data Source

### *Check Missing Values*

```
 ANIO          DEPARTAMENTO NUMERO.PROYECTOS HECTAREAS.REFORESTADAS VALOR.ESTABLECIMIENTO
Min.   :1995   Antioquia: 19   Min.   :  1.00   Min.   :    3.4        Min.   :2.999e+06
1st Qu.:1999   Arauca   : 19   1st Qu.:  2.00   1st Qu.:   57.0        1st Qu.:4.632e+07
Median :2005   Atlántico: 19   Median :  4.00   Median :  210.8        Median :1.553e+08
Mean   :2005   Bolivar  : 19   Mean   : 10.93   Mean   :  830.6        Mean   :6.976e+08
3rd Qu.:2010   Boyacá   : 19   3rd Qu.: 10.00   3rd Qu.:  643.5        3rd Qu.:5.073e+08
Max.   :2014   Caldas   : 19   Max.   :130.00   Max.   :26760.0        Max.   :2.321e+10
          (Other)  :380   NA's   :198      NA's   :199            NA's   :198

 ANIO          DEPARTAMENTO NUMERO.PROYECTOS HECTAREAS.REFORESTADAS VALOR.ESTABLECIMIENTO
Min.   :1995   Antioquia: 19   Min.   :  1.00   Min.   :    3.4        Min.   :2.999e+06
1st Qu.:1999   Arauca   : 19   1st Qu.:  3.00   1st Qu.:  144.3        1st Qu.:1.158e+08
Median :2005   Atlántico: 19   Median : 10.96   Median :  830.6        Median :6.998e+08
Mean   :2005   Bolivar  : 19   Mean   : 10.94   Mean   :  830.6        Mean   :6.985e+08
3rd Qu.:2010   Boyacá   : 19   3rd Qu.: 10.96   3rd Qu.:  830.6        3rd Qu.:6.998e+08
Max.   :2014   Caldas   : 19   Max.   :130.00   Max.   :26760.0        Max.   :2.321e+10
                               (Other)  :380
```

### *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **NUMERO.PROYECTOS**<br>Outliers identified: 44<br>Proportion (%) of outliers: 9.8<br>Mean of the outliers: 46.82<br>Mean without removing outliers: 10.94<br>Mean if we remove outliers: 7.43 |  |

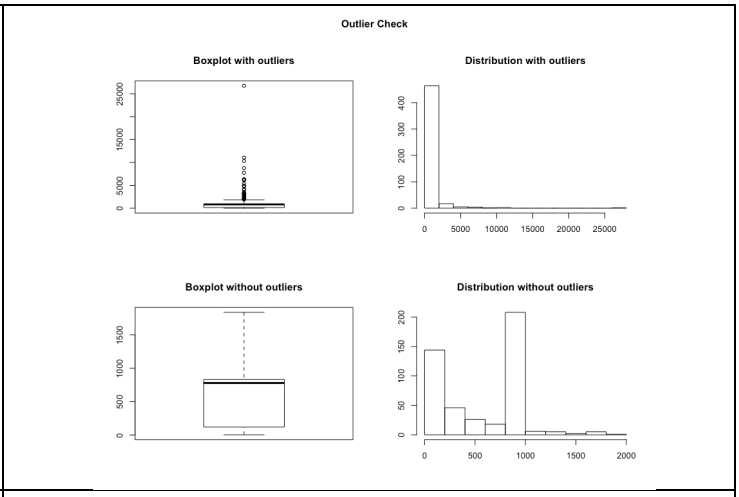
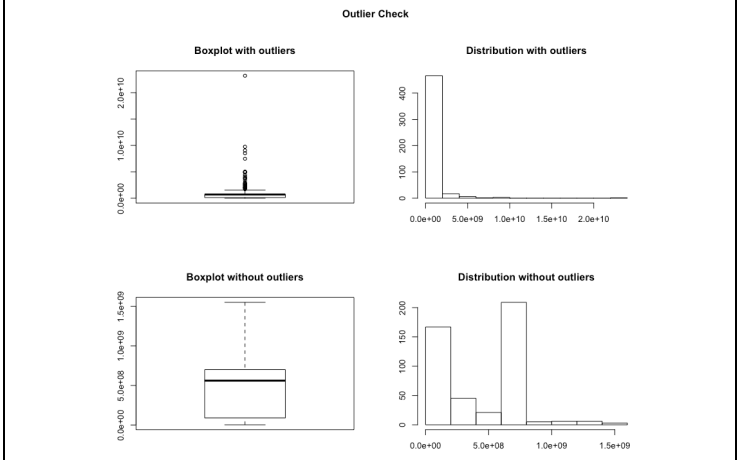
| | |
|---|---|
| **HECTAREAS.REFORESTADAS**<br>Outliers identified: 33<br>Proportion (%) of outliers: 7.2<br>Mean of the outliers: 4886.18<br>Mean without removing outliers: 830.56<br>Mean if we remove outliers: 540.24 |  |
| **VALOR.ESTABLECIMIENTO**<br>Outliers identified: 32<br>Proportion (%) of outliers: 6.9<br>Mean of the outliers: 4377843254<br>Mean without removing outliers: 698460791<br>Mean if we remove outliers: 443611789 |  |

**Table C. 19.** Results of the outlier detection process for each of the DNP-FI dataset attributes.

Therefore, the label correction process was not necessary considering the no contradictory instances in this dataset. Additionally, class balancing was not performed since this dataset does not contain classes. Finally, no duplicate instances were found.

### *Dimensionality Reduction*

It was not necessary to carry out the dimensionality reduction process considering the small number of attributes in this data set. Correlation matrix and correlogram are presented below.

| | NUMERO.PROYECTOS | HECTAREAS.REFORESTADAS | VALOR.ESTABLECIMIENTO |
|---|---|---|---|
| NUMERO.PROYECTOS | 1.0000000 | 0.7093533 | 0.6894912 |
| HECTAREAS.REFORESTADAS | 0.7093533 | 1.0000000 | 0.9875196 |
| VALOR.ESTABLECIMIENTO | 0.6894912 | 0.9875196 | 1.0000000 |

**Table C. 20.** Correlation matrix for all numerical attributes in the DNP-FI data set.

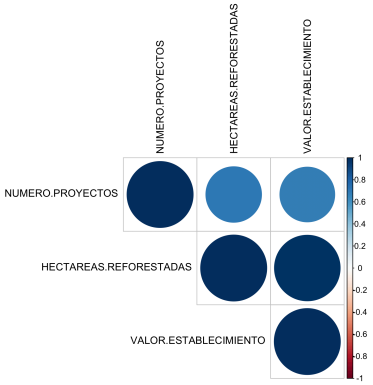



**Figure C. 9.** Correlated attributes for the DNP-FI dataset.

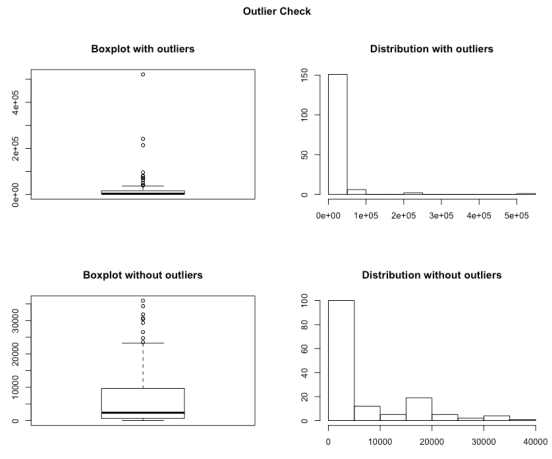
# C.13. DNP-LA Data Source

### *Check Missing Values*

```
     ANIO          DEPARTAMENTO   HECTAREAS           FAMILIAS
Min.   :2010   Amazonas :  5   Min.   :     2.0   Min.   :   0.00
1st Qu.:2011   Antioquia:  5   1st Qu.:   604.5   1st Qu.:  62.75
Median :2012   Arauca   :  5   Median :  2293.4   Median : 198.00
Mean   :2012   Atlantico:  5   Mean   : 15605.7   Mean   : 456.46
3rd Qu.:2013   Bolivar  :  5   3rd Qu.:  7421.7   3rd Qu.: 569.00
Max.   :2014   Boyaca   :  5   Max.   :521065.0   Max.   :6981.00
                    (Other)  :130   NA's   :17
```

```
     ANIO          DEPARTAMENTO   HECTAREAS           FAMILIAS
Min.   :2010   Amazonas :  5   Min.   :     2.0   Min.   :   0.00
1st Qu.:2011   Antioquia:  5   1st Qu.:   778.8   1st Qu.:  62.75
Median :2012   Arauca   :  5   Median :  2939.4   Median : 198.00
Mean   :2012   Atlantico:  5   Mean   : 15605.7   Mean   : 456.46
3rd Qu.:2013   Bolivar  :  5   3rd Qu.: 15605.7   3rd Qu.: 569.00
Max.   :2014   Boyaca   :  5   Max.   :521065.0   Max.   :6981.00
                    (Other)  :130
```

### *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **HECTAREAS**<br>Outliers identified: 12<br>Proportion (%) of outliers: 8.1<br>Mean of the outliers: 129898.9<br>Mean without removing outliers: 15605.66<br>Mean if we remove outliers: 6338.64 |  |

**FAMILIAS**
Outliers identified: 10
Proportion (%) of outliers: 6.7
Mean of the outliers: 2775.9
Mean without removing outliers: 456.46
Mean if we remove outliers: 301.83

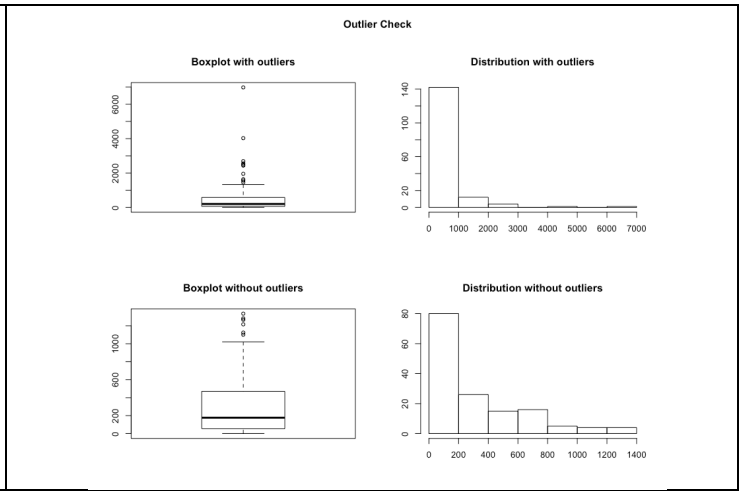



**Table C. 21.** Results of the outlier detection process for each of the DNP-LA dataset attributes.

### *Dimensionality Reduction*

The label correction process was not necessary, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since this dataset does not contain classes. Finally, no duplicate instances were found. On the other hand, correlation matrix and the scatter plot for HECTAREAS and FAMILIAS attributes is presented below.

| | HECTAREAS | FAMILIAS |
|---|---|---|
| HECTAREAS | 1.00000000 | 0.05014406 |
| FAMILIAS | 0.05014406 | 1.00000000 |

**Table C. 22.** Correlation matrix for all numerical attributes in the DNP-LA data set.

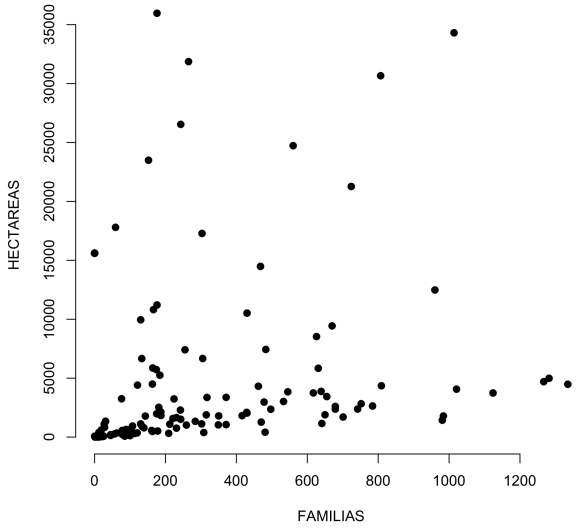


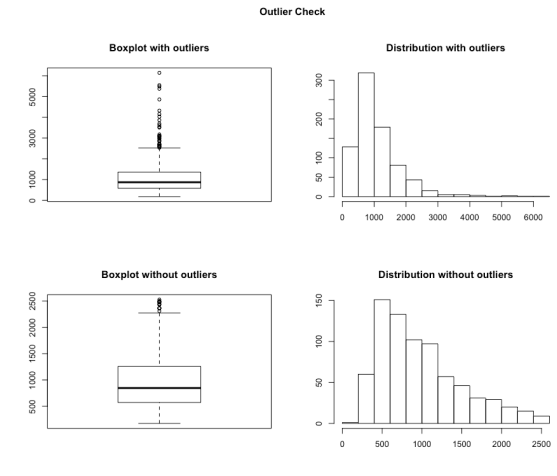

**Figure C. 10.** Scatter plot for HECTAREAS and FAMILIAS attributes.

# C.14. DNP-PA Data Source

## *Check Missing Values*

```
 DEPARTAMENTO      ANIO         PRODUCTO.ALIANZA                    NOMBRE.ALIANZA                        MUNICIPIO
Antioquia: 60   Min.   :2002   Cacao        :151                                                :  9                :  9
Huila    : 53   1st Qu.:2009   Leche        :105    Achiote Riohacha COPROCOJUME                :  1   Vistahermosa:  9
Bolivar  : 46   Median :2012   Café especial: 84    Acuícola Tadó (ACUACH)                      :  1   Puerto Rico :  8
Meta     : 46   Mean   :2011   Plátano      : 50    Agroforestal. Zambrano. Bolívar ASOEAT I:  1   Tumaco      :  7
Cauca    : 45   3rd Qu.:2013   Caucho       : 39    Aguacate Aguadas (AGROAGUADAS)              :  1   Pitalito    :  6
Santander: 42   Max.   :2013   Mora         : 35    Aguacate Alvarado (APROAGUACATE)            :  1   Riosucio    :  6
(Other)  :491   NA's   :9      (Other)      :319    (Other)                                     :769   (Other)     :738

 VALOR.ALIANZA    VALOR.IM.ALIANZA BENEFICIARIOS.ALIANZA HECTAREAS.ALIANZA
 Min.   : 174.9   Min.   :  61.9   Min.   : 15.00        Min.   :   0.4
 1st Qu.: 576.6   1st Qu.: 147.6   1st Qu.: 38.00        1st Qu.:  48.0
 Median : 864.3   Median : 208.0   Median : 51.00        Median :  82.5
 Mean   :1083.7   Mean   : 249.3   Mean   : 63.26        Mean   : 124.3
 3rd Qu.:1363.0   3rd Qu.: 303.9   3rd Qu.: 78.00        3rd Qu.: 144.0
 Max.   :6148.3   Max.   :1260.0   Max.   :400.00        Max.   :4300.0
 NA's   :9        NA's   :9        NA's   :9             NA's   :86

    DEPARTAMENTO      ANIO           PRODUCTO.ALIANZA                                        NOMBRE.ALIANZA        MUNICIPIO
 Antioquia: 60   Min.   : 217   Cacao        :151                                                :  9                :  9
 Huila    : 53   1st Qu.:2009   Leche        :105    Achiote Riohacha COPROCOJUME                :  1   Vistahermosa:  9
 Bolivar  : 46   Median :2012   Café especial: 84    Acuícola Tadó (ACUACH)                      :  1   Puerto Rico :  8
 Meta     : 46   Mean   :1990   Plátano      : 50    Agroforestal. Zambrano. Bolívar ASOEAT I:  1   Tumaco      :  7
 Cauca    : 45   3rd Qu.:2013   Caucho       : 39    Aguacate Aguadas (AGROAGUADAS)              :  1   Pitalito    :  6
 Santander: 42   Max.   :2013   Mora         : 35    Aguacate Alvarado (APROAGUACATE)            :  1   Riosucio    :  6
 (Other)  :491                  (Other)      :319    (Other)                                     :769   (Other)     :738

VALOR.ALIANZA    VALOR.IM.ALIANZA BENEFICIARIOS.ALIANZA HECTAREAS.ALIANZA
 Min.   : 174.9   Min.   :  61.9   Min.   : 15.00        Min.   :   0.4
 1st Qu.: 579.1   1st Qu.: 148.4   1st Qu.: 38.00        1st Qu.:  51.0
 Median : 867.6   Median : 210.1   Median : 51.00        Median :  95.0
 Mean   :1083.8   Mean   : 249.4   Mean   : 63.26        Mean   : 124.3
 3rd Qu.:1358.8   3rd Qu.: 300.0   3rd Qu.: 77.50        3rd Qu.: 134.7
 Max.   :6148.3   Max.   :1260.0   Max.   :400.00        Max.   :4300.0
```

## *Outliers Detection*

| Attribute Metrics | Distribution Charts |
|---|---|
| **VALOR.ALIANZA**<br>Outliers identified: 32<br>Proportion (%) of outliers: 4.3<br>Mean of the outliers: 3500.77<br>Mean without removing outliers: 1083.83<br>Mean if we remove outliers: 980.84 |  |

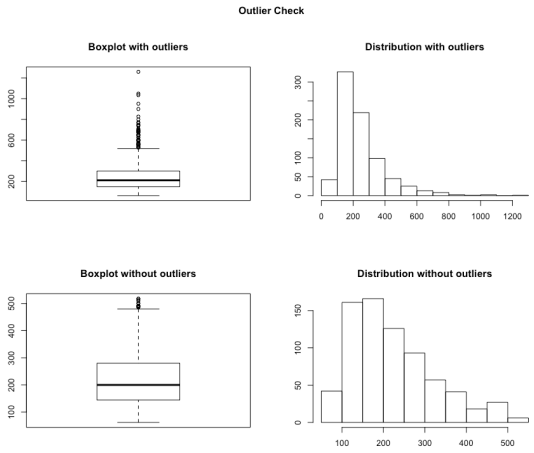
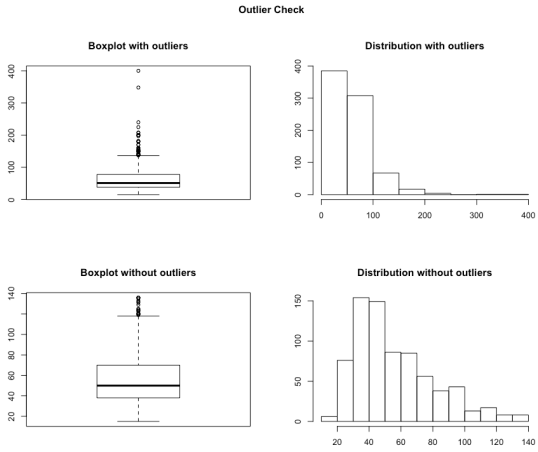
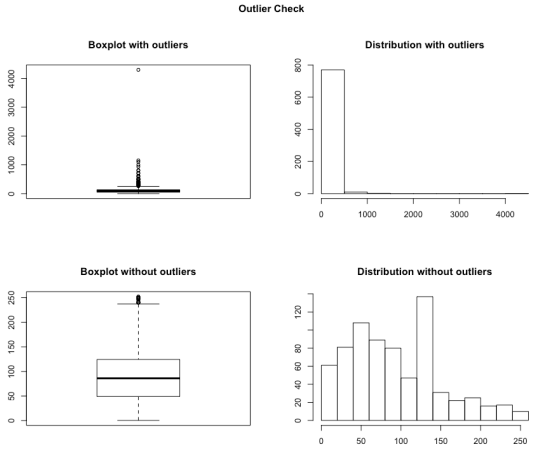
| | |
|---|---|
| **VALOR.IM.ALIANZA**<br>Outliers identified: 46<br>Proportion (%) of outliers: 6.2<br>Mean of the outliers: 681.5<br>Mean without removing outliers: 249.36<br>Mean if we remove outliers: 222.39 |  |
| **BENEFICIARIOS.ALIANZA**<br>Outliers identified: 44<br>Proportion (%) of outliers: 6<br>Mean of the outliers: 174.52<br>Mean without removing outliers: 63.26<br>Mean if we remove outliers: 56.64 |  |
| **HECTAREAS.ALIANZA**<br>Outliers identified: 59<br>Proportion (%) of outliers: 8.1<br>Mean of the outliers: 500.65<br>Mean without removing outliers: 124.34<br>Mean if we remove outliers: 93.67 |  |

**Table C. 23.** Results of the outlier detection process for each of the DNP-PA dataset attributes.

The outliers detected by the previous method were replaced by NA values and the process of replacing lost values was applied again. Later, we obtained the following summary of measures.

```
   DEPARTAMENTO      ANIO           PRODUCTO.ALIANZA                                    NOMBRE.ALIANZA        MUNICIPIO
Antioquia: 60   Min.   : 217   Cacao        :151                                              :  1    Vistahermosa:  9
Huila    : 53   1st Qu.:2009   Leche        :105     Achiote Riohacha COPROCOJUME             :  1    Puerto Rico :  8
Bolivar  : 46   Median :2012   Café especial: 84     Acuícola Tadó (ACUACH)                   :  1    Tumaco      :  7
Meta     : 46   Mean   :2008   Plátano      : 50     Agroforestal. Zambrano. Bolívar ASOEAT I:  1    Pitalito    :  6
Cauca    : 45   3rd Qu.:2013   Caucho       : 39     Aguacate Aguadas (AGROAGUADAS)           :  1    Riosucio    :  6
Santander: 42   Max.   :2013   Mora         : 35     Aguacate Alvarado (APROAGUACATE)         :  1    Unguía      :  6
(Other)  :483                  (Other)      :311     (Other)                                  :769    (Other)     :733
VALOR.ALIANZA    VALOR.IM.ALIANZA BENEFICIARIOS.ALIANZA HECTAREAS.ALIANZA
Min.   : 174.9   Min.   : 61.9    Min.   : 15.00        Min.   :  0.40
1st Qu.: 577.2   1st Qu.:147.7    1st Qu.: 38.00        1st Qu.: 51.00
Median : 866.1   Median :208.0    Median : 51.00        Median : 89.36
Mean   : 976.9   Mean   :221.6    Mean   : 56.46        Mean   : 93.03
3rd Qu.:1241.5   3rd Qu.:271.8    3rd Qu.: 70.00        3rd Qu.:124.34
Max.   :2525.4   Max.   :517.9    Max.   :136.00        Max.   :252.00
```

### *Removal of Duplicate Instances*

In this data set, 8 duplicate instances were found which were removed. Through this step, the data set remained with 775 instances of the initial 783.

### *Dimensionality Reduction*

The label correction process was not necessary, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since this dataset does not contain classes. Correlation matrix and correlogram are presented below.

| | VALOR.ALIANZA | VALOR.IM.ALIANZA | BENEFICIARIOS.ALIANZA | HECTAREAS.ALIANZA |
|---|---|---|---|---|
| VALOR.ALIANZA | 1.0000000 | 0.6468547 | 0.5393530 | 0.4936599 |
| VALOR.IM.ALIANZA | 0.6468547 | 1.0000000 | 0.7906273 | 0.4203510 |
| BENEFICIARIOS.ALIANZA | 0.5393530 | 0.7906273 | 1.0000000 | 0.4480550 |
| HECTAREAS.ALIANZA | 0.4936599 | 0.4203510 | 0.4480550 | 1.0000000 |

**Table C. 24.** Correlation matrix for all numerical attributes in the DNP-PA data set.

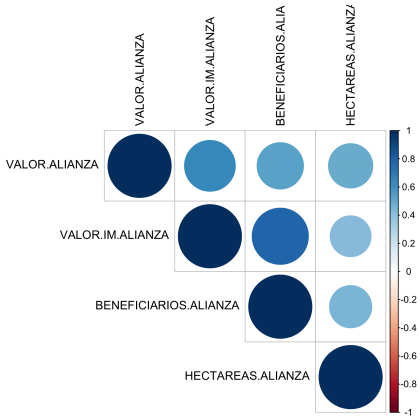



**Figure C. 11.** Correlated attributes for the DNP-PA dataset.

## C.15. DANE-HH Data Source

### *Check Missing Values*

```
TIPO_REG    PAIS         P_DEPTO        P_MUNIC           UC_UO            ENCUESTA         COD_VEREDA        ITER_HG
 N:128332   170:128332   19:128332   19256  : 8315   10000   :  9036   22351888:    23   19256011:  1770   101    :118434
                                     19001  : 7658   99990000:  2310   81387178:    18   19001999:  1333   102    :  4275
                                     19698  : 6662   0       :  1510   88025016:    13   19809008:  1145   201    :  2871
                                     19100  : 6232   20000   :  1245   81568493:    11   19807001:  1140   103    :   897
                                     19743  : 5560   680000  :  1147   60023592:     9   19256013:  1118   301    :   605
                                     19137  : 5321   30000   :  1066   81707835:     9   19100012:  1072   2101   :   530
                                     (Other):88584   (Other) :112018   (Other) :128249   (Other) :120754   (Other):   720

  P_S15P165A       P_S15P165B     P_S15P177 P_S15P178 P_S15P179_SP1   P_S15P179A     P_S15P179_SP2   P_S15P179B     P_S15P179_SP3
1      :123897   1      :122715   1:94935   1:42383   1   : 17405   2000   :  1919   1   :  1255   2000   :    97   1   :  1539
2      :  2972   2      :  4403   2:30516   2:56738   NA's:110927   2001   :  1851   NA's:127077   2013   :    85   NA's:126793
3      :   642   3      :   919   9: 2881   3:26202                 2012   :  1505                 9999   :    82
21     :   536   4      :   218             9: 3009                 2008   :  1426                 2001   :    76
4      :   139   5      :    48                                     2011   :  1404                 2011   :    69
5      :    59   6      :    14                                     (Other):  9300                 (Other):   846
(Other):    87   (Other):    15                                     NA's   :110927                 NA's   :127077

  P_S15P179C     P_S15P179_SP4 P_S15P180     TOT_PROD_HOGAR
2001   :   202   1   :108914   1   :  9714   Min.   :1.0
9999   :   122   NA's: 19418   2   :  9362   1st Qu.:1.0
2000   :    98                 9   :   342   Median :1.0
2013   :    89                 NA's:108914   Mean   :1.3
2012   :    80                               3rd Qu.:2.0
(Other):   948                               Max.   :6.0
NA's   :126793                               NA's   :63868
```

After identifying missing values in the data set, we proceeded to replace them taking into account the following considerations.

- Missing values of *P_S15P179_SP1*, *P_S15P179_SP2*, *P_S15P179_SP3*, and *P_S15P179_SP4* attributes were replaced by the value 2, adding this new category to the respective distributions. This new category represents the "*NO*" responses of the respondents.
- The values of the *P_S15P179A*, *P_S15P179B*, and *P_S15P179C* attributes correspond to a year, therefore they were replaced by a category "*9999*", which indicates the absence of this value.
- The *P_S15P180* attribute contains three categories: *1 (Yes)*, *2 (No)*, and *9 (No information)*. The missing values of this attribute were replaced with the value 9.
- The missing values of the TOT_PROD_HOGAR attribute were replaced by the mode value of this distribution, considering that they can only be integer values when dealing with a number of people.

Considering the above, the new data set summary is presented below.

```
TIPO_REG    PAIS         P_DEPTO        P_MUNIC           UC_UO            ENCUESTA         COD_VEREDA        ITER_HG
 N:128332   170:128332   19:128332   19256  : 8315   10000   :  9036   22351888:    23   19256011:  1770   101    :118434
                                     19001  : 7658   99990000:  2310   81387178:    18   19001999:  1333   102    :  4275
                                     19698  : 6662   0       :  1510   88025016:    13   19809008:  1145   201    :  2871
                                     19100  : 6232   20000   :  1245   81568493:    11   19807001:  1140   103    :   897
                                     19743  : 5560   680000  :  1147   60023592:     9   19256013:  1118   301    :   605
                                     19137  : 5321   30000   :  1066   81707835:     9   19100012:  1072   2101   :   530
                                     (Other):88584   (Other) :112018   (Other) :128249   (Other) :120754   (Other):   720

  P_S15P165A       P_S15P165B     P_S15P177 P_S15P178 P_S15P179_SP1   P_S15P179A     P_S15P179_SP2   P_S15P179B     P_S15P179_SP3
1      :123897   1      :122715   1:94935   1:42383   1: 17405      9999   :111530   1:  1255      9999   :127159   1:  1539
2      :  2972   2      :  4403   2:30516   2:56738   2:110927      2000   :  1919   2:127077      2000   :    97   2:126793
3      :   642   3      :   919   9: 2881   3:26202                 2001   :  1851                 2013   :    85
21     :   536   4      :   218             9: 3009                 2012   :  1505                 2001   :    76
4      :   139   5      :    48                                     2008   :  1426                 2011   :    69
5      :    59   6      :    14                                     2011   :  1404                 2007   :    66
(Other):    87   (Other):    15                                     (Other):  8697                 (Other):   780
```

```
  P_S15P179C     P_S15P179_SP4 P_S15P180  TOT_PROD_HOGAR
 9999   :126915   1:108914      1:  9714   Min.   :1.00
 2001   :   202   2: 19418      2:  9362   1st Qu.:1.00
 2000   :    98                 9:109256   Median :1.00
 2013   :    89                            Mean   :1.15
 2012   :    80                            3rd Qu.:1.00
 2009   :    76                            Max.   :6.00
 (Other):   872
```

Furthermore, the label correction process was not necessary, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since this dataset does not contain classes and no duplicate instances were found. Finally, correlation matrix and correlogram are not presented considering that there is only a single numeric attribute.

# C.16. DANE-P Data Source

***Check Missing Values***

```
TIPO_REG    PAIS         P_DEPTO        P_MUNIC            UC_UO             ENCUESTA          COD_VEREDA        P_S15P159
 M:148957   170:148957   19:148957   19256  :  9852   10000   :  9620   22355441:    32   19256011:  2150   1      :140352
                                     19001  :  8527   99990000:  2788   82128944:    32   19001999:  1417   2      :  6238
                                     19100  :  8106   0       :  1633   22351888:    25   19807001:  1416   3      :  1241
                                     19698  :  7011   20000   :  1350   88099280:    22   19100012:  1342   21     :   562
                                     19137  :  6272   680000  :  1315   81387178:    18   19256013:  1251   4      :   237
                                     19743  :  6073   30000   :  1186   60010489:    15   19809008:  1158   5      :   108
                                     (Other):103116   (Other) :131065   (Other) :148813   (Other) :140223   (Other):   219

P_S15P160    P_S15P161        P_S15P162     P_S15P163    P_S15P164_SP1 P_S15P164_SP2 P_S15P164_SP3 P_S15P164_SP4
 1:122715   1      :118312   2      :52927   1   :  974   1   :97930    1   : 5092    1   :37974    2   : 23655
 3: 10196   2      :  3484   1      :42695   2   : 8551   2   : 1069    2   :93177    2   :60295    NA's:125302
 4: 16046   3      :   701   3      :19201   3   :46352   9   :   61    9   :   66    9   :   66
            4      :   170   5      : 4214   4   :14003   NA's:49897    NA's:50622    NA's:50622
            5      :    34   9      : 1953   5   :50730
            (Other):    14   (Other): 1725   9   : 2105
            NA's   : 26242   NA's   :26242   NA's:26242
```

After identifying missing values in the data set, we proceeded to replace them taking into account the following considerations.

- The *P_S15P164_SP1, P_S15P164_SP2, P_S15P164_SP3* attributes contains three categories: *1 (Yes)*, *2 (No)*, and *9 (No information)*. The missing values of these attributes were replaced with the value 9.
- A new category was added to *P_S15P164_SP4*, in this case the value 9. The missing values of these attributes were replaced with the value 9.
- The missing values of the *P_S15P161*, *P_S15P162*, and *P_S15P163* attributes were replaced with the value 9.

Considering the above, the new data set summary is presented below.

```
TIPO_REG    PAIS         P_DEPTO        P_MUNIC            UC_UO             ENCUESTA          COD_VEREDA        P_S15P159
 M:148957   170:148957   19:148957   19256  :  9852   10000   :  9620   22355441:    32   19256011:  2150   1      :140352
                                     19001  :  8527   99990000:  2788   82128944:    32   19001999:  1417   2      :  6238
                                     19100  :  8106   0       :  1633   22351888:    25   19807001:  1416   3      :  1241
                                     19698  :  7011   20000   :  1350   88099280:    22   19100012:  1342   21     :   562
                                     19137  :  6272   680000  :  1315   81387178:    18   19256013:  1251   4      :   237
                                     19743  :  6073   30000   :  1186   60010489:    15   19809008:  1158   5      :   108
                                     (Other):103116   (Other) :131065   (Other) :148813   (Other) :140223   (Other):   219
```

```
P_S15P160    P_S15P161        P_S15P162     P_S15P163 P_S15P164_SP1 P_S15P164_SP2 P_S15P164_SP3 P_S15P164_SP4
1:122715   1      :118312   2      :52927   1:  974   1:97930       1: 5092       1:37974       2: 23655
3: 10196   9      : 26244   1      :42695   2: 8551   2: 1069       2:93177       2:60295       9:125302
4: 16046   2      :  3484   9      :28195   3:46352   9:49958       9:50688       9:50688
           3      :   701   3      :19201   4:14003
           4      :   170   5      : 4214   5:50730
           5      :    34   4      :  879   9:28347
           (Other):    12   (Other):  846
```

Furthermore, the label correction process was not necessary, considering that the dataset has no contradictory instances. Additionally, class balancing was not performed since this dataset does not contain classes and no duplicate instances were found. Finally, correlation matrix and correlogram are not presented considering that there are no numerical attributes.

# Appendix D

# Dataset Meta-Features

In this appendix, we present the Skewness, Kurtosis, and Entropy values for each attribute of the 16 initial data sources. We also show these three values once the datasets were cleaned.

## D.1 SIVICAP Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| ANIO | | | 0,00 | | | 0,00 |
| DEPARTAMENTO | | | 2,97 | | | 2,97 |
| MUNICIPIO | | | 6,81 | | | 6,81 |
| TOTAL.MUESTRAS | 15,23 | 309,89 | | 15,23 | 309,89 | |
| RESULTADO.COLOR.APARENTE | 30,74 | 949,38 | | 1,04 | 3,75 | |
| RESULTADO.TURBIEDAD | 24,44 | 686,40 | | 1,31 | 4,36 | |
| RESULTADO.PH | 20,33 | 448,82 | | 0,03 | 2,90 | |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 26,37 | 757,56 | | 0,17 | 2,46 | |
| RESULTADO.ALCALINIDAD.TOTAL | 2,31 | 9,94 | | 0,69 | 3,02 | |
| RESULTADO.CALCIO | 4,36 | 37,74 | | -1,20 | 3,86 | |
| RESULTADO.FOSFATOS | 21,21 | 476,69 | | -0,91 | 3,47 | |
| RESULTADO.MANGANESO | 10,02 | 112,56 | | | | |
| RESULTADO.MOLIBDENO | 7,61 | 63,98 | | | | |
| RESULTADO.MAGNESIO | 17,55 | 363,50 | | -1,65 | 5,85 | |
| RESULTADO.ZINC | 10,41 | 114,19 | | | | |
| RESULTADO.DUREZA.TOTAL | 3,12 | 18,39 | | 0,76 | 3,24 | |
| RESULTADO.SULFATOS | 4,91 | 38,26 | | 0,30 | 3,36 | |
| RESULTADO.HIERRO.TOTAL | 22,50 | 575,32 | | 0,47 | 2,69 | |
| RESULTADO.CLORUROS | 11,95 | 193,07 | | 0,34 | 2,88 | |
| RESULTADO.NITRATOS | 1,67 | 6,44 | | | | |
| RESULTADO.NITRITOS | 16,63 | 295,20 | | -0,08 | 2,30 | |
| RESULTADO.ALUMINIO | 18,05 | 340,82 | | -0,42 | 1,18 | |
| RESULTADO.FLORUROS | 16,43 | 291,23 | | -0,29 | 1,15 | |
| RESULTADO.COT | 7,97 | 67,32 | | | | |
| RESULTADO.COLIFORMES.TOTALES | 7,88 | 83,52 | | 1,74 | 5,07 | |
| RESULTADO.E.COLI | 19,27 | 444,80 | | 2,39 | 8,32 | |
| RESULTADO.ANTIMONIO | 7,36 | 57,55 | | | | |
| RESULTADO.ARSÉNICO | 10,39 | 109,01 | | | | |
| RESULTADO.BARIO | 4,89 | 25,24 | | | | |
| RESULTADO.CADMIO | 6,57 | 54,75 | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| RESULTADO.CIANURO.LIBRE.Y.DIASOCIABLE | 2,91 | 10,02 | | | | |
| RESULTADO.COBRE | 1,73 | 6,18 | | | | |
| RESULTADO.Cromo.total | 2,02 | 6,01 | | | | |
| RESULTADO.MERCURIO | 5,24 | 34,11 | | | | |
| RESULTADO.NIQUEL | 1,81 | 5,22 | | | | |
| RESULTADO.PLOMO | 6,25 | 42,92 | | | | |
| RESULTADO.SELENIO | 8,92 | 85,47 | | | | |
| RESULTADO.TRIHALOMETANOS.TOTALES | 3,70 | 21,96 | | | | |
| RESULTADO.HIDROCARBUROS.AROMÁTICOS.POLICICLICOS | 3,98 | 18,21 | | | | |
| RESULTADO.GIARDIA | 4,07 | 18,60 | | | | |
| RESULTADO.CRYPTOSPORIDIUM | 3,98 | 17,24 | | | | |
| RESULTADO.PLAGUICIDAS.TOTALES | | | | | | |
| RESULTADO.ORGANOFOSFORADOS.Y.CARBAMATOS | | | | | | |
| RESULTADO.MESÓFILOS | 6,64 | 52,10 | | | | |
| IRCA.PROMEDIO | 0,85 | 2,91 | | 0,85 | 2,91 | |
| IRCA.BASE.PROMEDIO | 0,86 | 2,92 | | 0,86 | 2,92 | |
| NIVEL.DE.RIESGO.PROMEDIO | | | 1,43 | | | 1,43 |
| Mean | 9,83 | 176,96 | 2,80 | 1,08 | 18,78 | 2,80 |

**Table D. 1.** Skewness, Kurtosis, and Entropy values for SIVICAP data source.

# D.2 CORPOICA Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Departamento | | | 2,53 | | | 2,53 |
| Municipio | | | 5,39 | | | 5,39 |
| Cultivo | | | 3,28 | | | 3,28 |
| Estado | | | 0,96 | | | 0,96 |
| Tiempo.Establecimiento | | | 1,35 | | | 1,35 |
| Topografia | | | 1,35 | | | 1,35 |
| Drenaje | | | 1,01 | | | 1,01 |
| Riego | | | 1,19 | | | 1,19 |
| Fertilizantes.aplicados | | | 1,44 | | | 1,44 |
| Fecha.de.análisis | | | 2,05 | | | 2,05 |
| Cobre | | | 6,16 | | | 6,16 |
| Manganeso | | | 6,82 | | | 6,82 |
| Zinc.olsen | | | 6,11 | | | 6,11 |
| Zinc | | | 0,81 | | | 0,81 |
| pH | 0,75 | 3,09 | | 0,68 | 2,88 | |
| Materia.organica | 2,51 | 12,71 | | 1,29 | 4,17 | |
| Fosforo | 7,62 | 129,89 | | 1,69 | 5,23 | |
| Azufre | 56,05 | 5472,25 | | 1,33 | 4,48 | |
| Acidez | 3,04 | 18,23 | | 1,51 | 4,31 | |
| Aluminio.intercambiable | 2,90 | 15,48 | | 1,65 | 4,75 | |
| Calcio.intercambiable | 3,07 | 21,07 | | 1,03 | 3,35 | |
| Magnesio.intercambiable | 2,66 | 12,72 | | 1,37 | 4,59 | |
| Potasio.intercambiable | 5,80 | 94,01 | | 1,15 | 3,69 | |
| Sodio.intercambiable | 10,59 | 224,93 | | 1,31 | 4,14 | |
| capacidad.de.intercambio.cationico | 2,47 | 14,26 | | 1,00 | 3,28 | |
| Conductividad.electrica | 153,49 | 23742,05 | | 1,22 | 4,19 | |
| Hierro.olsen | 3,86 | 26,57 | | 1,10 | 3,64 | |
| Boro | 10,70 | 280,16 | | 0,97 | 3,61 | |
| Hierro | 4,98 | 60,50 | | | | |
| Cobre.doble.acido | 2,33 | 14,69 | | | | |
| Manganeso.1 | 3,14 | 18,47 | | | | |
| Mean | 16,23 | 1774,18 | 2,89 | 1,24 | 4,02 | 2,89 |

**Table D. 2.** Skewness, Kurtosis, and Entropy values for CORPOICA data source.

## D.3 IDEAM Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Departamento | | | 0,00 | | | 0,00 |
| Municipio | | | 3,14 | | | 3,14 |
| Anio | | | 2,05 | | | 2,05 |
| Mes | | | 2,49 | | | 2,49 |
| Temperatura | -0,35 | 2,05 | | -0,35 | 2,05 | |
| Precipitacion | 1,19 | 4,41 | | 1,19 | 4,41 | |
| Humedad.relativa | 0,56 | 2,92 | | -0,56 | 2,92 | |
| Radiacion | 0,51 | 3,10 | | 0,60 | 4,24 | |
| Mean | 0,48 | 3,12 | 1,92 | 0,22 | 3,40 | 1,92 |

**Table D. 3.** Skewness, Kurtosis, and Entropy values for IDEAM data source.

## D.4 AVA Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Department | | | 4,60 | | | 4,60 |
| Municipality | | | 1,53 | | | 1,53 |
| CA_Cocoa | 0,93 | 2,60 | | 0,93 | 2,60 | |
| CA_Coffee | 0,50 | 2,69 | | 0,50 | 2,69 | |
| CA_Sugar.Cane | 1,53 | 4,24 | | 1,53 | 4,24 | |
| CA_Bean | 0,36 | 2,00 | | 0,36 | 2,00 | |
| CA_Potato | 1,14 | 3,24 | | 1,14 | 3,24 | |
| CA_Banana | -1,22 | 3,58 | | -1,22 | 3,58 | |
| Mean | 0,54 | 3,06 | 3,06 | 0,54 | 3,06 | 3,06 |

**Table D. 4.** Skewness, Kurtosis, and Entropy values for AVA data source.

## D.5 FINAGRO Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Entidad | | | 0,69 | | | 0,69 |
| Tipo.productor | | | 1,80 | | | 1,80 |
| Departamento | | | 3,44 | | | 3,44 |
| Anio | | | 2,76 | | | 2,76 |
| Valor | 13,74 | 252,70 | | 1,82 | 5,82 | |
| Mean | 13,74 | 252,70 | 2,17 | 1,82 | 5,82 | 2,17 |

**Table D. 5.** Skewness, Kurtosis, and Entropy values for FINAGRO data source.

## D.6 DANE-SIPSA-P Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| ANIO | | | 1,61 | | | 1,61 |
| CULTIVOS | | | 3,39 | | | 3,39 |
| SUPERFICIE.SEMBRADA | 3,45 | 16,20 | | 3,45 | 16,20 | |
| SUPERFICIE.COSECHADA | 3,95 | 20,59 | | 3,95 | 20,59 | |
| PRODUCCION | 3,24 | 14,42 | | 3,24 | 14,42 | |
| RENDIMIENTO | 2,75 | 15,54 | | 2,75 | 15,54 | |
| PRECIO.AL.PRODUCTOR.KG | 1,79 | 6,67 | | 1,79 | 6,67 | |
| PRECIO.AL.PRODUCTOR.TON | 1,79 | 6,67 | | 1,79 | 6,67 | |
| COSTO.PRODUCCION | 6,68 | 52,94 | | 6,68 | 52,94 | |
| INGRESO.BRUTO.PRODUCCION | 3,40 | 17,98 | | 3,40 | 17,98 | |
| COSTO.TOTAL.PRODUCCION | 2,85 | 12,13 | | 2,85 | 12,13 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| UTILIDAD | 3,63 | 21,57 | | 3,63 | 21,57 | |
| RENTABILIDAD | 1,85 | 8,34 | | 1,85 | 8,34 | |
| Mean | 3,22 | 17,55 | 2,50 | 3,22 | 17,55 | 2,50 |

**Table D. 6.** Skewness, Kurtosis, and Entropy values for DANE-SIPSA-P data source.

# D.7 Agronet Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Departamento | | | 0,00 | | | 0,00 |
| Cultivo | | | 3,78 | | | 3,78 |
| Anio | | | 2,13 | | | 2,13 |
| Municipio | | | 3,66 | | | 3,66 |
| Area.Cosechada | 4,35 | 23,57 | | 2,36 | 8,86 | |
| Area.Sembrada | 4,42 | 25,52 | | 2,35 | 8,84 | |
| Produccion | 7,28 | 59,87 | | 2,12 | 7,23 | |
| Rendimiento | 3,73 | 16,96 | | 1,23 | 4,00 | |
| Mean | 4,95 | 31,48 | 2,39 | 2,01 | 7,23 | 2,39 |

**Table D. 7.** Skewness, Kurtosis, and Entropy values for Agronet data source.

# D.8 Minagricultura Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| COD.DEP | | | 0,00 | | | 0,00 |
| DEPARTAMENTO | | | 0,00 | | | 0,00 |
| COD.MUN | | | 3,66 | | | 3,66 |
| MUNICIPIO | | | 3,66 | | | 3,66 |
| GRUPO.CULTIVO | | | 1,91 | | | 1,91 |
| SUBGRUPO.CULTIVO | | | 3,53 | | | 3,53 |
| CULTIVO | | | 3,64 | | | 3,64 |
| SISTEMA.PRODUCTIVO | | | 3,79 | | | 3,79 |
| COD.CULTIVO | | | 3,79 | | | 3,79 |
| NOMBRE.CIENTIFICO | | | 3,59 | | | 3,59 |
| PERIODO | | | 3,27 | | | 3,27 |
| ESTADO.FISICO.PRODUCCION | | | 1,93 | | | 1,93 |
| AREA.SEMBRADA | 5,46 | 38,95 | | 2,21 | 7,95 | |
| AREA.COSECHADA | 5,30 | 34,49 | | 2,21 | 7,88 | |
| PRODUCCION | 10,41 | 119,59 | | 2,31 | 8,36 | |
| RENDIMIENTO | 4,48 | 24,87 | | 1,37 | 4,34 | |
| Mean | 6,41 | 54,47 | 2,73 | 2,02 | 7,13 | 2,73 |

**Table D. 8.** Skewness, Kurtosis, and Entropy values for Minagricultura data source.

# D.9 Agronet-P Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Fecha | | | 8,23 | | | 8,23 |
| Producto | | | 2,15 | | | 2,15 |
| Unidad | | | 1,04 | | | 1,04 |
| Fuente | | | 1,51 | | | 1,51 |
| Precio | 2,38 | 7,18 | | 1,16 | 4,72 | |
| Variacion | 93,96 | 11900,28 | | 0,04 | 3,75 | |
| Mean | 48,17 | 5953,73 | 3,23 | 0,60 | 4,24 | 3,23 |

**Table D. 9.** Skewness, Kurtosis, and Entropy values for Agronet-P data source.

## D.10 DANE-SIPSA Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| Fecha | | | 4,09 | | | 4,09 |
| Grupo | | | 1,24 | | | 1,24 |
| Producto | | | 5,03 | | | 5,03 |
| Fuente | | | 3,88 | | | 3,88 |
| Precio | 1,92 | 8,20 | | 1,02 | 3,37 | |
| Mean | 1,92 | 8,20 | 3,56 | 1,02 | 3,37 | 3,56 |

**Table D. 10.** Skewness, Kurtosis, and Entropy values for DANE-SIPSA data source.

## D.11 DNP-AIB Data Source

| | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| Attribute | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| ANIO | | | 1,97 | | | 1,97 |
| TIPO.ENTIDAD | | | 0,45 | | | 0,45 |
| COD.ENTIDAD | | | 6,88 | | | 6,88 |
| ENTIDAD | | | 6,78 | | | 6,78 |
| CUENTA.PROYECTO | | | 3,60 | | | 3,60 |
| SECTOR | | | 0,68 | | | 0,68 |
| OBJETIVO | | | 1,28 | | | 1,28 |
| VALOR | 50,10 | 3347,86 | | 1,93 | 6,89 | |
| Mean | 50,10 | 3347,86 | 3,09 | 1,93 | 6,89 | 3,09 |

**Table D. 11.** Skewness, Kurtosis, and Entropy values for DNP-AIB data source.

## D.12 DNP-FI Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| ANIO | | | 2,94 | | | 2,94 |
| DEPARTAMENTO | | | 3,26 | | | 3,26 |
| NUMERO.PROYECTOS | 3,29 | 16,03 | | -0,07 | 2,17 | |
| HECTAREAS.REFORESTADAS | 7,47 | 80,45 | | 0,19 | 2,62 | |
| VALOR.ESTABLECIMIENTO | 7,37 | 77,87 | | 0,26 | 2,60 | |
| Mean | 6,05 | 58,12 | 3,10 | 0,12 | 2,46 | 3,10 |

**Table D. 12.** Skewness, Kurtosis, and Entropy values for DNP-FI data source.

## D.13 DNP-LA Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| ANIO | | | 1,61 | | | 1,61 |
| DEPARTAMENTO | | | 3,47 | | | 3,47 |
| HECTAREAS | 7,26 | 64,29 | | 1,74 | 5,53 | |
| FAMILIAS | 4,80 | 34,72 | | 1,36 | 4,35 | |
| Mean | 6,03 | 49,50 | 2,54 | 1,55 | 4,94 | 2,54 |

**Table D. 13.** Skewness, Kurtosis, and Entropy values for DNP-LA data source.

# D.14 DNP-PA Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| DEPARTAMENTO | | | 3,22 | | | 2,01 |
| ANIO | | | 2,00 | | | 2,01 |
| PRODUCTO.ALIANZA | | | 3,20 | | | 3,17 |
| NOMBRE.ALIANZA | | | 6,64 | | | 6,65 |
| MUNICIPIO | | | 6,10 | | | 6,11 |
| VALOR.ALIANZA | 2,32 | 11,60 | | 0,95 | 3,27 | |
| VALOR.IM.ALIANZA | 2,02 | 9,02 | | 0,92 | 3,34 | |
| BENEFICIARIOS.ALIANZA | 2,60 | 15,47 | | 1,00 | 3,58 | |
| HECTAREAS.ALIANZA | 13,28 | 257,94 | | 0,68 | 3,14 | |
| Mean | 5,05 | 73,51 | 4,23 | 0,89 | 3,33 | 3,99 |

**Table D. 14.** Skewness, Kurtosis, and Entropy values for DNP-PA data source.

# D.15 DANE-HH Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| TIPO_REG | | | 0,00 | | | 0,00 |
| PAIS | | | 0,00 | | | 0,00 |
| P_DEPTO | | | 0,00 | | | 0,00 |
| P_MUNIC | | | 3,55 | | | 3,55 |
| UC_UO | | | 6,83 | | | 6,83 |
| ENCUESTA | | | 11,65 | | | 11,65 |
| COD_VEREDA | | | 6,52 | | | 6,52 |
| ITER_HG | | | 0,40 | | | 0,40 |
| P_S15P165A | | | 0,19 | | | 0,19 |
| P_S15P165B | | | 0,21 | | | 0,21 |
| P_S15P177 | | | 0,65 | | | 0,65 |
| P_S15P178 | | | 1,14 | | | 1,14 |
| P_S15P179_SP1 | | | 0,00 | | | 0,40 |
| P_S15P179A | | | 2,97 | | | 0,77 |
| P_S15P179_SP2 | | | 0,00 | | | 0,05 |
| P_S15P179B | | | 3,38 | | | 0,08 |
| P_S15P179_SP3 | | | 0,00 | | | 0,06 |
| P_S15P179C | | | 3,22 | | | 0,10 |
| P_S15P179_SP4 | | | 0,00 | | | 0,42 |
| P_S15P180 | | | 0,77 | | | 0,52 |
| TOT_PROD_HOGAR | | | 0,68 | | | 1,03 |
| Mean | | | 2,01 | | | 1,65 |

**Table D. 15.** Skewness, Kurtosis, and Entropy values for DANE-HH data source.

# D.16 DANE-H Data Source

| Attribute | Raw Dataset | | | Clean Dataset | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | Entropy | Skewness | Kurtosis | Entropy |
| TIPO_REG | | | 0,00 | | | 0,00 |
| PAIS | | | 0,00 | | | 0,00 |
| P_DEPTO | | | 0,00 | | | 0,00 |
| P_MUNIC | | | 3,56 | | | 3,56 |
| UC_UO | | | 6,96 | | | 6,96 |
| ENCUESTA | | | 11,82 | | | 11,82 |
| COD_VEREDA | | | 6,54 | | | 6,54 |
| P_S15P159 | | | 0,28 | | | 0,28 |
| P_S15P160 | | | 0,58 | | | 0,58 |
| P_S15P161 | | | 0,18 | | | 0,61 |
| P_S15P162 | | | 1,28 | | | 1,47 |

| | | | | | | |
|---|---|---|---|---|---|---|
| P_S15P163 | | | 1,27 | | | 1,47 |
| P_S15P164_SP1 | | | 0,06 | | | 0,68 |
| P_S15P164_SP2 | | | 0,21 | | | 0,78 |
| P_S15P164_SP3 | | | 0,67 | | | 1,08 |
| P_S15P164_SP4 | | | 0,00 | | | 0,44 |
| Mean | | | 2,09 | | | 2,27 |

**Table D. 16.** Skewness, Kurtosis, and Entropy values for DANE-H data source.

# Appendix E

## Ranking of Variables by Experts

Each attribute of the 16 data sources was ranked according to its level of importance. In this appendix, we compiled these rankings according to the classification performed by four experts in the domains of environmental sciences, agricultural sciences, chemistry, among others. These data were obtained through interviews with each of them. On the other hand, we also present the rankings by Logistic Regression and Random Forest techniques, in cases with labeled datasets. For unlabeled datasets, the rankings were addressed by the respective correlation matrix.

| Dimension | Indicator | Data Source | Ranking by Logistic Regression or Correlation Matrix | Ranking by Random Forest | Ranking of Rater 1 | Ranking of Rater 2 | Ranking of Rater 3 | Ranking of Rater 4 |
|---|---|---|---|---|---|---|---|---|
| Biophysical | Water | SIVICAP | RESULTADO.CLORO.RESIDUAL.LIBRE | IRCA.PROMEDIO | RESULTADO.CLORO.RESIDUAL.LIBRE | RESULTADO.PH | RESULTADO.SULFATOS | RESULTADO.PH |
| | | | RESULTADO.ALUMINIO | IRCA.BASE.PROMEDIO | RESULTADO.SULFATOS | RESULTADO.ALUMINIO | RESULTADO.FOSFATOS | RESULTADO.TURBIEDAD |
| | | | RESULTADO.COLIFORMES.TOTALES | RESULTADO.E.COLI | RESULTADO.COLIFORMES.TOTALES | RESULTADO.HIERRO.TOTAL | RESULTADO.TURBIEDAD | RESULTADO.COLIFORMES.TOTALES |
| | | | RESULTADO.COLOR.APARENTE | RESULTADO.CLORO.RESIDUAL.LIBRE | RESULTADO.COLOR.APARENTE | RESULTADO.FLORUROS | RESULTADO.PH | RESULTADO.COLOR.APARENTE |
| | | | RESULTADO.SULFATOS | RESULTADO.COLIFORMES.TOTALES | RESULTADO.ALUMINIO | RESULTADO.SULFATOS | RESULTADO.NITRITOS | RESULTADO.SULFATOS |
| | | | RESULTADO.CLORUROS | RESULTADO.COLOR.APARENTE | RESULTADO.PH | RESULTADO.CLORUROS | RESULTADO.CLORUROS | RESULTADO.CLORUROS |
| | | | RESULTADO.HIERRO.TOTAL | RESULTADO.TURBIEDAD | RESULTADO.CALCIO | RESULTADO.TURBIEDAD | RESULTADO.HIERRO.TOTAL | RESULTADO.HIERRO.TOTAL |
| | | | RESULTADO.NITRITOS | TOTAL.MUESTRAS | RESULTADO.HIERRO.TOTAL | RESULTADO.NITRITOS | RESULTADO.CLORO.RESIDUAL.LIBRE | RESULTADO.CALCIO |
| | | | RESULTADO.CALCIO | RESULTADO.DUREZA.TOTAL | RESULTADO.CLORUROS | RESULTADO.CALCIO | RESULTADO.CALCIO | RESULTADO.DUREZA.TOTAL |
| | | | RESULTADO.PH | RESULTADO.SULFATOS | RESULTADO.TURBIEDAD | RESULTADO.CLORO.RESIDUAL.LIBRE | RESULTADO.ALUMINIO | RESULTADO.FOSFATOS |
| | | | RESULTADO.E.COLI | RESULTADO.CLORUROS | RESULTADO.E.COLI | RESULTADO.E.COLI | RESULTADO.E.COLI | RESULTADO.E.COLI |
| | | | RESULTADO.MAGNESIO | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.MAGNESIO | RESULTADO.MAGNESIO | RESULTADO.ALCALINIDAD.TOTAL |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | RESULTADO.TURBIEDAD | RESULTADO.PH | RESULTADO.FOSFATOS | RESULTADO.COLIFORMES.TOTALES | RESULTADO.COLIFORMES.TOTALES | RESULTADO.NITRITOS |
| | | | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.HIERRO.TOTAL | RESULTADO.MAGNESIO | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.CLORO.RESIDUAL.LIBRE |
| | | | IRCA.PROMEDIO | RESULTADO.ALUMINIO | IRCA.PROMEDIO | RESULTADO.COLOR.APARENTE | RESULTADO.COLOR.APARENTE | IRCA.PROMEDIO |
| | | | TOTAL.MUESTRAS | RESULTADO.NITRITOS | IRCA.BASE.PROMEDIO | RESULTADO.DUREZA.TOTAL | RESULTADO.DUREZA.TOTAL | RESULTADO.ALUMINIO |
| | | | RESULTADO.FLORUROS | RESULTADO.FOSFATOS | RESULTADO.FLORUROS | RESULTADO.FOSFATOS | RESULTADO.FLORUROS | RESULTADO.FLORUROS |
| | | | IRCA.BASE.PROMEDIO | RESULTADO.CALCIO | RESULTADO.NITRITOS | IRCA.PROMEDIO | IRCA.PROMEDIO | IRCA.BASE.PROMEDIO |
| | | | RESULTADO.DUREZA.TOTAL | RESULTADO.MAGNESIO | RESULTADO.DUREZA.TOTAL | IRCA.BASE.PROMEDIO | IRCA.BASE.PROMEDIO | RESULTADO.MAGNESIO |
| | | | RESULTADO.FOSFATOS | RESULTADO.FLORUROS | TOTAL.MUESTRAS | TOTAL.MUESTRAS | TOTAL.MUESTRAS | TOTAL.MUESTRAS |
| | | | MUNICIPIO | MUNICIPIO | MUNICIPIO | MUNICIPIO | MUNICIPIO | MUNICIPIO |
| | | | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO |
| | | | ANIO | ANIO | ANIO | ANIO | ANIO | ANIO |
| | | | NIVEL.DE.RIESGO.PROMEDIO | NIVEL.DE.RIESGO.PROMEDIO | NIVEL.DE.RIESGO.PROMEDIO | NIVEL.DE.RIESGO.PROMEDIO | NIVEL.DE.RIESGO.PROMEDIO | NIVEL.DE.RIESGO.PROMEDIO |
| | Soil | CORPOICA | Sodio.intercambiable | Hierro.olsen | Potasio.intercambiable | Sodio.intercambiable | Sodio.intercambiable | pH |
| | | | Hierro.olsen | Materia.organica | Hierro.olsen | Hierro.olsen | Hierro.olsen | Hierro.olsen |
| | | | Azufre | pH | Calcio.intercambiable | Azufre | Azufre | Azufre |
| | | | Potasio.intercambiable | Fosforo | Sodio.intercambiable | Potasio.intercambiable | Potasio.intercambiable | Fosforo |
| | | | Fosforo | Sodio.intercambiable | Fosforo | Fosforo | Fosforo | Potasio.intercambiable |
| | | | Magnesio.intercambiable | capacidad.de.intercambio.cationico | Magnesio.intercambiable | Magnesio.intercambiable | Magnesio.intercambiable | Magnesio.intercambiable |
| | | | Boro | Calcio.intercambiable | pH | Boro | Boro | Materia.organica |
| | | | capacidad.de.intercambio.cationico | Potasio.intercambiable | capacidad.de.intercambio.cationico | capacidad.de.intercambio.cationico | capacidad.de.intercambio.cationico | capacidad.de.intercambio.cationico |
| | | | Materia.organica | Magnesio.intercambiable | Materia.organica | Materia.organica | Materia.organica | Boro |
| | | | Conductividad.electrica | Azufre | Conductividad.electrica | Conductividad.electrica | Conductividad.electrica | Aluminio.intercambiable |
| | | | Calcio.intercambiable | Boro | Azufre | Calcio.intercambiable | Calcio.intercambiable | Calcio.intercambiable |
| | | | pH | Conductividad.electrica | Boro | pH | pH | Sodio.intercambiable |
| | | | Acidez | Acidez | Acidez | Acidez | Acidez | Acidez |
| | | | Aluminio.intercambiable | Aluminio.intercambiable | Aluminio.intercambiable | Aluminio.intercambiable | Aluminio.intercambiable | Conductividad.electrica |
| | | | Cobre | Cobre | Cobre | Cobre | Cobre | Cobre |
| | | | Manganeso | Manganeso | Manganeso | Manganeso | Manganeso | Manganeso |
| | | | Zinc | Zinc | Zinc | Zinc | Zinc | Zinc |
| | | | Zinc.olsen | Zinc.olsen | Zinc.olsen | Zinc.olsen | Zinc.olsen | Zinc.olsen |
| | | | Topografia | Topografia | Topografia | Topografia | Topografia | Topografia |
| | | | Fertilizantes.aplicados | Fertilizantes.aplicados | Fertilizantes.aplicados | Fertilizantes.aplicados | Fertilizantes.aplicados | Fertilizantes.aplicados |
| | | | Riego | Riego | Riego | Riego | Riego | Riego |
| | | | Drenaje | Drenaje | Drenaje | Drenaje | Drenaje | Drenaje |
| | | | Tiempo.Establecimiento | Tiempo.Establecimiento | Tiempo.Establecimiento | Tiempo.Establecimiento | Tiempo.Establecimiento | Tiempo.Establecimiento |
| | | | Estado | Estado | Estado | Estado | Estado | Estado |
| | | | Municipio | Municipio | Municipio | Municipio | Municipio | Municipio |
| | | | Departamento | Departamento | Departamento | Departamento | Departamento | Departamento |
| | | | Fecha.de.análisis | Fecha.de.análisis | Fecha.de.análisis | Fecha.de.análisis | Fecha.de.análisis | Fecha.de.análisis |
| | | | Cultivo | Cultivo | Cultivo | Cultivo | Cultivo | Cultivo |
| | Weather | IDEAM | Humedad.relativa | | Precipitacion | Precipitacion | Precipitacion | Precipitacion |
| | | | Radiacion | | Temperatura | Temperatura | Temperatura | Temperatura |
| | | | Precipitacion | | Radiacion | Humedad.relativa | Radiacion | Humedad.relativa |
| | | | Temperatura | | Humedad.relativa | Radiacion | Humedad.relativa | Radiacion |
| | | | Municipio | | Municipio | Municipio | Municipio | Municipio |
| | | | Departamento | | Departamento | Departamento | Departamento | Departamento |
| | | | Anio | | Anio | Anio | Anio | Anio |
| | | | Mes | | Mes | Mes | Mes | Mes |
| Economic-Productive | Agricultural Credits | FINAGRO | Valor | | Tipo.productor | Valor | Tipo.productor | Valor |
| | | | Tipo.productor | | Entidad | Tipo.productor | Valor | Tipo.productor |
| | | | Departamento | | Valor | Departamento | Entidad | Entidad |
| | | | Entidad | | Departamento | Entidad | Departamento | Departamento |
| | | | Anio | | Anio | Anio | Anio | Anio |
| | Crop Vulnerability | AVA | CA_Coffee | | CA_Coffee | CA_Coffee | CA_Coffee | CA_Coffee |
| | | | CA_Sugar.Cane | | CA_Sugar.Cane | CA_Sugar.Cane | CA_Sugar.Cane | CA_Sugar.Cane |
| | | | CA_Cocoa | | CA_Cocoa | CA_Banana | CA_Cocoa | CA_Banana |
| | | | CA_Banana | | CA_Banana | CA_Cocoa | CA_Banana | CA_Potato |
| | | | CA_Potato | | CA_Potato | CA_Potato | CA_Potato | CA_Bean |
| | | | CA_Bean | | CA_Bean | CA_Bean | CA_Bean | CA_Cocoa |
| | | | Municipality | | Municipality | Municipality | Municipality | Municipality |
| | | | Department | | Department | Department | Department | Department |
| | Prices of Agricultural Products | DANE-SIPSA-P | PRODUCCION | COSTO.PRODUCCION | UTILIDAD | PRODUCCION | RENDIMIENTO | RENTABILIDAD |
| | | | RENDIMIENTO | SUPERFICIE.SEMBRADA | RENTABILIDAD | RENDIMIENTO | PRODUCCION | UTILIDAD |
| | | | SUPERFICIE.COSECHADA | PRECIO.AL.PRODUCTOR.KG | RENDIMIENTO | SUPERFICIE.COSECHADA | COSTO.TOTAL.PRODUCCION | COSTO.PRODUCCION |
| | | | RENTABILIDAD | PRECIO.AL.PRODUCTOR.TON | PRODUCCION | SUPERFICIE.SEMBRADA | UTILIDAD | RENDIMIENTO |
| | | | SUPERFICIE.SEMBRADA | COSTO.TOTAL.PRODUCCION | INGRESO.BRUTO.PRODUCCION | RENTABILIDAD | PRECIO.AL.PRODUCTOR.TON | PRODUCCION |
| | | | PRECIO.AL.PRODUCTOR.TON | SUPERFICIE.COSECHADA | PRECIO.AL.PRODUCTOR.KG | PRECIO.AL.PRODUCTOR.TON | PRECIO.AL.PRODUCTOR.KG | COSTO.TOTAL.PRODUCCION |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | PRECIO.AL.PRODUCTOR.KG | PRODUCCION | PRECIO.AL.PRODUCTOR.TON | PRECIO.AL.PRODUCTOR.KG | RENTABILIDAD | SUPERFICIE.COSECHADA |
| | | | INGRESO.BRUTO.PRODUCCION | INGRESO.BRUTO.PRODUCCION | CULTIVOS | INGRESO.BRUTO.PRODUCCION | INGRESO.BRUTO.PRODUCCION | INGRESO.BRUTO.PRODUCCION |
| | | | COSTO.PRODUCCION | UTILIDAD | SUPERFICIE.COSECHADA | COSTO.PRODUCCION | COSTO.PRODUCCION | PRECIO.AL.PRODUCTOR.TON |
| | | | COSTO.TOTAL.PRODUCCION | RENDIMIENTO | SUPERFICIE.SEMBRADA | COSTO.TOTAL.PRODUCCION | SUPERFICIE.COSECHADA | PRECIO.AL.PRODUCTOR.KG |
| | | | UTILIDAD | RENTABILIDAD | COSTO.TOTAL.PRODUCCION | UTILIDAD | SUPERFICIE.SEMBRADA | SUPERFICIE.SEMBRADA |
| | | | ANIO | ANIO | COSTO.PRODUCCION | ANIO | ANIO | ANIO |
| | | | CULTIVOS | CULTIVOS | ANIO | CULTIVOS | CULTIVOS | CULTIVOS |
| | Crop Production | Agronet | Area.Sembrada | | Produccion | Area.Sembrada | Area.Sembrada | Rendimiento |
| | | | Area.Cosechada | | Rendimiento | Area.Cosechada | Area.Cosechada | Produccion |
| | | | Produccion | | Cultivo | Rendimiento | Produccion | Cultivo |
| | | | Rendimiento | | Municipio | Produccion | Rendimiento | Municipio |
| | | | Cultivo | | Departamento | Cultivo | Cultivo | Area.Sembrada |
| | | | Municipio | | Area.Cosechada | Municipio | Municipio | Area.Cosechada |
| | | | Departamento | | Area.Sembrada | Departamento | Departamento | Departamento |
| | | | Anio | | Anio | Anio | Anio | Anio |
| | Crop Production | Minagricultura | AREA.COSECHADA | | PRODUCCION | AREA.COSECHADA | SISTEMA.PRODUCTIVO | SISTEMA.PRODUCTIVO |
| | | | AREA.SEMBRADA | | RENDIMIENTO | AREA.SEMBRADA | ESTADO.FISICO.PRODUCCION | ESTADO.FISICO.PRODUCCION |
| | | | PRODUCCION | | ESTADO.FISICO.PRODUCCION | PRODUCCION | PRODUCCION | RENDIMIENTO |
| | | | RENDIMIENTO | | CULTIVO | RENDIMIENTO | RENDIMIENTO | PRODUCCION |
| | | | CULTIVO | | NOMBRE..CIENTIFICO | SISTEMA.PRODUCTIVO | CULTIVO | CULTIVO |
| | | | SISTEMA.PRODUCTIVO | | COD.CULTIVO | CULTIVO | SISTEMA.PRODUCTIVO | AREA.COSECHADA |
| | | | ESTADO.FISICO.PRODUCCION | | GRUPO.CULTIVO | ESTADO.FISICO.PRODUCCION | ESTADO.FISICO.PRODUCCION | AREA.SEMBRADA |
| | | | SUBGRUPO.CULTIVO | | SUBGRUPO.CULTIVO | SUBGRUPO.CULTIVO | SUBGRUPO.CULTIVO | SUBGRUPO.CULTIVO |
| | | | GRUPO.CULTIVO | | MUNICIPIO | GRUPO.CULTIVO | GRUPO.CULTIVO | GRUPO.CULTIVO |
| | | | NOMBRE..CIENTIFICO | | COD.MUN | NOMBRE..CIENTIFICO | NOMBRE..CIENTIFICO | NOMBRE..CIENTIFICO |
| | | | MUNICIPIO | | DEPARTAMENTO | MUNICIPIO | MUNICIPIO | MUNICIPIO |
| | | | DEPARTAMENTO | | COD.DEP | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO |
| | | | PERIODO | | AREA.COSECHADA | PERIODO | PERIODO | PERIODO |
| | | | COD.CULTIVO | | AREA.SEMBRADA | COD.CULTIVO | COD.CULTIVO | COD.CULTIVO |
| | | | COD.MUN | | SISTEMA.PRODUCTIVO | COD.MUN | COD.MUN | COD.MUN |
| | | | COD.DEP | | PERIODO | COD.DEP | COD.DEP | COD.DEP |
| | Imports and Exports | Agronet-P | Precio | | Precio | Precio | Precio | Precio |
| | | | Variacion | | Producto | Variacion | Producto | Variacion |
| | | | Producto | | Fuente | Producto | Variacion | Producto |
| | | | Fuente | | Variacion | Fuente | Fuente | Fuente |
| | | | Unidad | | Unidad | Unidad | Unidad | Unidad |
| | | | Fecha | | Fecha | Fecha | Fecha | Fecha |
| | Prices of Agricultural Products | DANE-SIPSA | Precio | | Producto | Precio | Precio | Fuente |
| | | | Producto | | Grupo | Fuente | Producto | Precio |
| | | | Fuente | | Precio | Producto | Fuente | Producto |
| | | | Grupo | | Fuente | Grupo | Grupo | Grupo |
| | | | Fecha | | Fecha | Fecha | Fecha | Fecha |
| Socio-Cultural | Livelihood | DANE-HH | TOT_PROD_HOGAR | | TOT_PROD_HOGAR | TOT_PROD_HOGAR | P_S15P179_SP3 | TOT_PROD_HOGAR |
| | | | P_S15P179_SP1 | | P_S15P179_SP1 | P_S15P179_SP3 | TOT_PROD_HOGAR | ITER_HG |
| | | | P_S15P179_SP2 | | P_S15P179_SP2 | P_S15P179_SP2 | P_S15P179_SP2 | P_S15P179_SP1 |
| | | | P_S15P179_SP3 | | P_S15P179_SP3 | P_S15P179_SP1 | P_S15P179_SP1 | P_S15P179_SP3 |
| | | | P_S15P179_SP4 | | P_S15P179_SP4 | P_S15P179_SP4 | P_S15P179_SP4 | P_S15P179_SP4 |
| | | | P_S15P165A | | P_S15P179A | P_S15P165A | P_S15P180 | P_S15P179_SP2 |
| | | | P_S15P165B | | P_S15P179B | P_S15P165B | P_S15P165B | P_S15P165A |
| | | | P_S15P177 | | P_S15P179C | P_S15P177 | P_S15P177 | P_S15P177 |
| | | | P_S15P178 | | P_S15P180 | P_S15P178 | P_S15P178 | P_S15P178 |
| | | | P_S15P179A | | P_S15P165A | P_S15P180 | P_S15P179A | P_S15P179A |
| | | | P_S15P179B | | P_S15P165B | P_S15P179A | P_S15P179B | P_S15P179B |
| | | | P_S15P179C | | P_S15P177 | P_S15P179B | P_S15P179C | P_S15P179C |
| | | | P_S15P180 | | P_S15P178 | P_S15P179C | P_S15P165A | P_S15P180 |
| | | | P_MUNIC | | P_MUNIC | P_MUNIC | P_MUNIC | P_MUNIC |
| | | | COD_VEREDA | | COD_VEREDA | COD_VEREDA | COD_VEREDA | COD_VEREDA |
| | | | ITER_HG | | ITER_HG | ITER_HG | ITER_HG | P_S15P165B |
| | | | UC_UO | | UC_UO | UC_UO | UC_UO | UC_UO |
| | | | ENCUESTA | | ENCUESTA | ENCUESTA | ENCUESTA | ENCUESTA |
| | | | P_DEPTO | | P_DEPTO | P_DEPTO | P_DEPTO | P_DEPTO |
| | | | PAIS | | PAIS | PAIS | PAIS | PAIS |
| | | | TIPO_REG | | TIPO_REG | TIPO_REG | TIPO_REG | TIPO_REG |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Human Development | DANE-H | P_S15P164_SP1 | | P_S15P164_SP1 | P_S15P164_SP3 | P_S15P164_SP3 | P_S15P164_SP3 |
| | | | P_S15P164_SP2 | | P_S15P164_SP2 | P_S15P164_SP2 | P_S15P164_SP1 | P_S15P164_SP2 |
| | | | P_S15P164_SP3 | | P_S15P164_SP3 | P_S15P164_SP1 | P_S15P164_SP2 | P_S15P164_SP1 |
| | | | P_S15P164_SP4 | | P_S15P164_SP4 | P_S15P164_SP4 | P_S15P164_SP4 | P_S15P164_SP4 |
| | | | P_S15P161 | | P_S15P161 | P_S15P161 | P_S15P161 | P_S15P161 |
| | | | P_S15P162 | | P_S15P162 | P_S15P162 | P_S15P162 | P_S15P162 |
| | | | P_S15P163 | | P_S15P163 | P_S15P163 | P_S15P163 | P_S15P163 |
| | | | P_S15P159 | | P_S15P159 | P_S15P159 | P_S15P159 | P_S15P159 |
| | | | P_S15P160 | | P_S15P160 | P_S15P160 | P_S15P160 | P_S15P160 |
| | | | P_MUNIC | | P_MUNIC | P_MUNIC | P_MUNIC | P_MUNIC |
| | | | COD_VEREDA | | COD_VEREDA | COD_VEREDA | COD_VEREDA | COD_VEREDA |
| | | | UC_UO | | UC_UO | UC_UO | UC_UO | UC_UO |
| | | | ENCUESTA | | ENCUESTA | ENCUESTA | ENCUESTA | ENCUESTA |
| | | | P_DEPTO | | P_DEPTO | P_DEPTO | P_DEPTO | P_DEPTO |
| | | | PAIS | | PAIS | PAIS | PAIS | PAIS |
| | | | TIPO_REG | | TIPO_REG | TIPO_REG | TIPO_REG | TIPO_REG |
| Political-Institutional | Agricultural Investment Budget | DNP-AIB | VALOR | | VALOR | VALOR | VALOR | OBJETIVO |
| | | | OBJETIVO | | OBJETIVO | SECTOR | OBJETIVO | VALOR |
| | | | SECTOR | | SECTOR | OBJETIVO | SECTOR | SECTOR |
| | | | TIPO.ENTIDAD | | TIPO.ENTIDAD | TIPO.ENTIDAD | TIPO.ENTIDAD | TIPO.ENTIDAD |
| | | | ENTIDAD | | ENTIDAD | ENTIDAD | ENTIDAD | ENTIDAD |
| | | | CUENTA.PROYECTO | | CUENTA.PROYECTO | CUENTA.PROYECTO | CUENTA.PROYECTO | CUENTA.PROYECTO |
| | | | COD.ENTIDAD | | COD.ENTIDAD | COD.ENTIDAD | COD.ENTIDAD | COD.ENTIDAD |
| | | | ANIO | | ANIO | ANIO | ANIO | ANIO |
| | Forest Incentives | DNP-FI | VALOR.ESTABLECIMIENTO | | HECTAREAS.REFORESTADAS | VALOR.ESTABLECIMIENTO | HECTAREAS.REFORESTADAS | VALOR.ESTABLECIMIENTO |
| | | | HECTAREAS.REFORESTADAS | | VALOR.ESTABLECIMIENTO | HECTAREAS.REFORESTADAS | VALOR.ESTABLECIMIENTO | HECTAREAS.REFORESTADAS |
| | | | NUMERO.PROYECTOS | | NUMERO.PROYECTOS | NUMERO.PROYECTOS | NUMERO.PROYECTOS | NUMERO.PROYECTOS |
| | | | DEPARTAMENTO | | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO |
| | | | ANIO | | ANIO | ANIO | ANIO | ANIO |
| | Land Allocation | DNP-LA | FAMILIAS | | FAMILIAS | HECTAREAS | FAMILIAS | FAMILIAS |
| | | | HECTAREAS | | HECTAREAS | FAMILIAS | HECTAREAS | HECTAREAS |
| | | | DEPARTAMENTO | | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO |
| | | | ANIO | | ANIO | ANIO | ANIO | ANIO |
| | Productive Alliances | DNP-PA | HECTAREAS.ALIANZA | | NOMBRE.ALIANZA | HECTAREAS.ALIANZA | BENEFICIARIOS.ALIANZA | BENEFICIARIOS.ALIANZA |
| | | | BENEFICIARIOS.ALIANZA | | PRODUCTO.ALIANZA | BENEFICIARIOS.ALIANZA | HECTAREAS.ALIANZA | PRODUCTO.ALIANZA |
| | | | VALOR.ALIANZA | | BENEFICIARIOS.ALIANZA | VALOR.ALIANZA | VALOR.ALIANZA | VALOR.ALIANZA |
| | | | VALOR.IM.ALIANZA | | HECTAREAS.ALIANZA | VALOR.IM.ALIANZA | VALOR.IM.ALIANZA | VALOR.IM.ALIANZA |
| | | | PRODUCTO.ALIANZA | | VALOR.ALIANZA | PRODUCTO.ALIANZA | PRODUCTO.ALIANZA | HECTAREAS.ALIANZA |
| | | | NOMBRE.ALIANZA | | VALOR.IM.ALIANZA | NOMBRE.ALIANZA | NOMBRE.ALIANZA | NOMBRE.ALIANZA |
| | | | MUNICIPIO | | MUNICIPIO | MUNICIPIO | MUNICIPIO | MUNICIPIO |
| | | | DEPARTAMENTO | | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO | DEPARTAMENTO |
| | | | ANIO | | ANIO | ANIO | ANIO | ANIO |

# Appendix F

# Fleiss' Kappa Test Results

Fleiss' Kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters. In this test, we compared the variable rankings of four experts to determine the level of agreement among raters. The results for each data source are presented below.

## F.1 SIVICAP Data Source

| Etiquetas de fila | ANIO | DEPARTAMENTO | IRCA.BASE.PROMEDIO | IRCA.PROMEDIO | MUNICIPIO | NIVEL.DE.RIESGO.PROMEDIO | RESULTADO.ALCALINIDAD.TOTAL | RESULTADO.ALUMINIO | RESULTADO.CALCIO | RESULTADO.CLORO.RESIDUAL.LIB | RESULTADO.CLORUROS | RESULTADO.COLIFORMES.TOTALE | RESULTADO.COLOR.APARENTE | RESULTADO.DUREZA.TOTAL | RESULTADO.E.COLI | RESULTADO.FLORUROS | RESULTADO.FOSFATOS | RESULTADO.HIERRO.TOTAL | RESULTADO.MAGNESIO | RESULTADO.NITRITOS | RESULTADO.PH | RESULTADO.SULFATOS | RESULTADO.TURBIEDAD | TOTAL.MUESTRAS | Total general |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | 2 | | | | | | | | | | | 2 | 1 | | | 5 |
| 2 | | | | | | | | 2 | | | | | | | | | 1 | | | | | 1 | 1 | | 5 |
| 3 | | | | | | | | | | | | 3 | | | | | | 1 | | | | | 1 | | 5 |
| 4 | | | | | | | | | | | | | 3 | | | 1 | | | | | 1 | | | | 5 |
| 5 | | | | | | | | 1 | | | | | | | | | | | | 1 | | 3 | | | 5 |
| 6 | | | | | | | | | | | 4 | | | | | | | | | | 1 | | | | 5 |
| 7 | | | | | | | | | 1 | | | | | | | | | 3 | | | | | 1 | | 5 |
| 8 | | | | | | | | | 1 | 1 | | | | | | | | 1 | | 2 | | | | | 5 |
| 9 | | | | | | | | | 3 | | 1 | | | 1 | | | | | | | | | | | 5 |
| 10 | | | | | | | | 1 | | 1 | | | | | | | 1 | | | | 1 | | 1 | | 5 |
| 11 | | | | | | | | | | | | | | | 5 | | | | | | | | | | 5 |
| 12 | | | | | | | 2 | | | | | | | | | | | | 3 | | | | | | 5 |
| 13 | | | | | | | | | | | | 2 | | | | | 1 | | | 1 | | | 1 | | 5 |
| 14 | | | | | | | 3 | | | 1 | | | | | | | | | 1 | | | | | | 5 |
| 15 | | | | 3 | | | | | | | | | 2 | | | | | | | | | | | | 5 |
| 16 | | | 1 | | | | | 1 | | | | | | 2 | | | | | | | | | | 1 | 5 |

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | | | | | | | | | | | | | | | 4 | 1 | | | | | | | | 5 |
| 18 | | | 2 | 2 | | | | | | | | | | | | | | | 1 | | | | | 5 |
| 19 | | | 2 | | | | | | | | | | 2 | | | | | 1 | | | | | | 5 |
| 20 | | | | | | | | | | | | | | | | 1 | | | | | | | 4 | 5 |
| 21 | | | | | 5 | | | | | | | | | | | | | | | | | | | 5 |
| 22 | | 5 | | | | | | | | | | | | | | | | | | | | | | 5 |
| 23 | 5 | | | | | | | | | | | | | | | | | | | | | | | 5 |
| 24 | | | | | | 5 | | | | | | | | | | | | | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 120 |

**Table F. 1.** Matches among raters in the positions of variables for the SIVICAP dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | RESULTADO.CLORO.RESIDUAL.LIBRE | CM |
| 1 | RESULTADO.CLORO.RESIDUAL.LIBRE | J1 |
| 1 | RESULTADO.PH | J2 |
| 1 | RESULTADO.SULFATOS | J3 |
| 1 | RESULTADO.PH | J4 |
| 2 | RESULTADO.ALUMINIO | CM |
| 2 | RESULTADO.SULFATOS | J1 |
| 2 | RESULTADO.ALUMINIO | J2 |
| 2 | RESULTADO.FOSFATOS | J3 |
| 2 | RESULTADO.TURBIEDAD | J4 |
| 3 | RESULTADO.COLIFORMES.TOTALES | CM |
| 3 | RESULTADO.COLIFORMES.TOTALES | J1 |
| 3 | RESULTADO.HIERRO.TOTAL | J2 |
| 3 | RESULTADO.TURBIEDAD | J3 |
| 3 | RESULTADO.COLIFORMES.TOTALES | J4 |
| 4 | RESULTADO.COLOR.APARENTE | CM |
| 4 | RESULTADO.COLOR.APARENTE | J1 |
| 4 | RESULTADO.FLORUROS | J2 |
| 4 | RESULTADO.PH | J3 |
| 4 | RESULTADO.COLOR.APARENTE | J4 |
| 5 | RESULTADO.SULFATOS | CM |
| 5 | RESULTADO.ALUMINIO | J1 |
| 5 | RESULTADO.SULFATOS | J2 |
| 5 | RESULTADO.NITRITOS | J3 |
| 5 | RESULTADO.SULFATOS | J4 |
| 6 | RESULTADO.CLORUROS | CM |
| 6 | RESULTADO.PH | J1 |
| 6 | RESULTADO.CLORUROS | J2 |
| 6 | RESULTADO.CLORUROS | J3 |
| 6 | RESULTADO.CLORUROS | J4 |
| 7 | RESULTADO.HIERRO.TOTAL | CM |
| 7 | RESULTADO.CALCIO | J1 |
| 7 | RESULTADO.TURBIEDAD | J2 |
| 7 | RESULTADO.HIERRO.TOTAL | J3 |
| 7 | RESULTADO.HIERRO.TOTAL | J4 |
| 8 | RESULTADO.NITRITOS | CM |
| 8 | RESULTADO.HIERRO.TOTAL | J1 |
| 8 | RESULTADO.NITRITOS | J2 |
| 8 | RESULTADO.CLORO.RESIDUAL.LIBRE | J3 |
| 8 | RESULTADO.CALCIO | J4 |
| 9 | RESULTADO.CALCIO | CM |
| 9 | RESULTADO.CLORUROS | J1 |
| 9 | RESULTADO.CALCIO | J2 |
| 9 | RESULTADO.CALCIO | J3 |
| 9 | RESULTADO.DUREZA.TOTAL | J4 |
| 10 | RESULTADO.PH | CM |
| 10 | RESULTADO.TURBIEDAD | J1 |
| 10 | RESULTADO.CLORO.RESIDUAL.LIBRE | J2 |
| 10 | RESULTADO.ALUMINIO | J3 |
| 10 | RESULTADO.FOSFATOS | J4 |
| 11 | RESULTADO.E.COLI | CM |

| | | |
|---|---|---|
| 11 | RESULTADO.E.COLI | J1 |
| 11 | RESULTADO.E.COLI | J2 |
| 11 | RESULTADO.E.COLI | J3 |
| 11 | RESULTADO.E.COLI | J4 |
| 12 | RESULTADO.MAGNESIO | CM |
| 12 | RESULTADO.ALCALINIDAD.TOTAL | J1 |
| 12 | RESULTADO.MAGNESIO | J2 |
| 12 | RESULTADO.MAGNESIO | J3 |
| 12 | RESULTADO.ALCALINIDAD.TOTAL | J4 |
| 13 | RESULTADO.TURBIEDAD | CM |
| 13 | RESULTADO.FOSFATOS | J1 |
| 13 | RESULTADO.COLIFORMES.TOTALES | J2 |
| 13 | RESULTADO.COLIFORMES.TOTALES | J3 |
| 13 | RESULTADO.NITRITOS | J4 |
| 14 | RESULTADO.ALCALINIDAD.TOTAL | CM |
| 14 | RESULTADO.MAGNESIO | J1 |
| 14 | RESULTADO.ALCALINIDAD.TOTAL | J2 |
| 14 | RESULTADO.ALCALINIDAD.TOTAL | J3 |
| 14 | RESULTADO.CLORO.RESIDUAL.LIBRE | J4 |
| 15 | IRCA.PROMEDIO | CM |
| 15 | IRCA.PROMEDIO | J1 |
| 15 | RESULTADO.COLOR.APARENTE | J2 |
| 15 | RESULTADO.COLOR.APARENTE | J3 |
| 15 | IRCA.PROMEDIO | J4 |
| 16 | TOTAL.MUESTRAS | CM |
| 16 | IRCA.BASE.PROMEDIO | J1 |
| 16 | RESULTADO.DUREZA.TOTAL | J2 |
| 16 | RESULTADO.DUREZA.TOTAL | J3 |
| 16 | RESULTADO.ALUMINIO | J4 |
| 17 | RESULTADO.FLORUROS | CM |
| 17 | RESULTADO.FLORUROS | J1 |
| 17 | RESULTADO.FOSFATOS | J2 |
| 17 | RESULTADO.FLORUROS | J3 |
| 17 | RESULTADO.FLORUROS | J4 |
| 18 | IRCA.BASE.PROMEDIO | CM |
| 18 | RESULTADO.NITRITOS | J1 |
| 18 | IRCA.PROMEDIO | J2 |
| 18 | IRCA.PROMEDIO | J3 |
| 18 | IRCA.BASE.PROMEDIO | J4 |
| 19 | RESULTADO.DUREZA.TOTAL | CM |
| 19 | RESULTADO.DUREZA.TOTAL | J1 |
| 19 | IRCA.BASE.PROMEDIO | J2 |
| 19 | IRCA.BASE.PROMEDIO | J3 |
| 19 | RESULTADO.MAGNESIO | J4 |
| 20 | RESULTADO.FOSFATOS | CM |
| 20 | TOTAL.MUESTRAS | J1 |
| 20 | TOTAL.MUESTRAS | J2 |
| 20 | TOTAL.MUESTRAS | J3 |
| 20 | TOTAL.MUESTRAS | J4 |
| 21 | MUNICIPIO | CM |
| 21 | MUNICIPIO | J1 |
| 21 | MUNICIPIO | J2 |
| 21 | MUNICIPIO | J3 |
| 21 | MUNICIPIO | J4 |
| 22 | DEPARTAMENTO | CM |
| 22 | DEPARTAMENTO | J1 |
| 22 | DEPARTAMENTO | J2 |
| 22 | DEPARTAMENTO | J3 |
| 22 | DEPARTAMENTO | J4 |
| 23 | ANIO | CM |
| 23 | ANIO | J1 |
| 23 | ANIO | J2 |
| 23 | ANIO | J3 |
| 23 | ANIO | J4 |
| 24 | NIVEL.DE.RIESGO.PROMEDIO | CM |

| | | |
|---|---|---|
| 24 | NIVEL.DE.RIESGO.PROMEDIO | J1 |
| 24 | NIVEL.DE.RIESGO.PROMEDIO | J2 |
| 24 | NIVEL.DE.RIESGO.PROMEDIO | J3 |
| 24 | NIVEL.DE.RIESGO.PROMEDIO | J4 |

**Table F. 2.** Ranking of variables established by J1, J2, J3, and J4 raters for the SIVICAP dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,41 | 0,01 | 30,36 | 0,00 | 0,38 | 0,44 |
| ANIO | 1,00 | 0,06 | 15,49 | 0,00 | 0,87 | 1,13 |
| DEPARTAMENTO | 1,00 | 0,06 | 15,49 | 0,00 | 0,87 | 1,13 |
| IRCA.BASE.PROMEDIO | 0,17 | 0,06 | 2,56 | 0,01 | 0,04 | 0,29 |
| IRCA.PROMEDIO | 0,37 | 0,06 | 5,79 | 0,00 | 0,25 | 0,50 |
| MUNICIPIO | 1,00 | 0,06 | 15,49 | 0,00 | 0,87 | 1,13 |
| NIVEL.DE.RIESGO.PROMEDIO | 1,00 | 0,06 | 15,49 | 0,00 | 0,87 | 1,13 |
| RESULTADO.ALCALINIDAD.TOTAL | 0,37 | 0,06 | 5,79 | 0,00 | 0,25 | 0,50 |
| RESULTADO.ALUMINIO | 0,06 | 0,06 | 0,94 | 0,35 | -0,07 | 0,19 |
| RESULTADO.CALCIO | 0,27 | 0,06 | 4,18 | 0,00 | 0,14 | 0,40 |
| RESULTADO.CLORO.RESIDUAL.LIBRE | 0,06 | 0,06 | 0,94 | 0,35 | -0,07 | 0,19 |
| RESULTADO.CLORUROS | 0,58 | 0,06 | 9,03 | 0,00 | 0,46 | 0,71 |
| RESULTADO.COLIFORMES.TOTALES | 0,37 | 0,06 | 5,79 | 0,00 | 0,25 | 0,50 |
| RESULTADO.COLOR.APARENTE | 0,37 | 0,06 | 5,79 | 0,00 | 0,25 | 0,50 |
| RESULTADO.DUREZA.TOTAL | 0,17 | 0,06 | 2,56 | 0,01 | 0,04 | 0,29 |
| RESULTADO.E.COLI | 1,00 | 0,06 | 15,49 | 0,00 | 0,87 | 1,13 |
| RESULTADO.FLORUROS | 0,58 | 0,06 | 9,03 | 0,00 | 0,46 | 0,71 |
| RESULTADO.FOSFATOS | -0,04 | 0,06 | -0,67 | 0,50 | -0,17 | 0,08 |
| RESULTADO.HIERRO.TOTAL | 0,27 | 0,06 | 4,18 | 0,00 | 0,14 | 0,40 |
| RESULTADO.MAGNESIO | 0,27 | 0,06 | 4,18 | 0,00 | 0,14 | 0,40 |
| RESULTADO.NITRITOS | 0,06 | 0,06 | 0,94 | 0,35 | -0,07 | 0,19 |
| RESULTADO.PH | 0,06 | 0,06 | 0,94 | 0,35 | -0,07 | 0,19 |
| RESULTADO.SULFATOS | 0,27 | 0,06 | 4,18 | 0,00 | 0,14 | 0,40 |
| RESULTADO.TURBIEDAD | -0,04 | 0,06 | -0,67 | 0,50 | -0,17 | 0,08 |
| TOTAL.MUESTRAS | 0,58 | 0,06 | 9,03 | 0,00 | 0,46 | 0,71 |

**Table F. 3.** Fleiss' Kappa test metrics for the SIVICAP dataset.

# F.2 IDEAM Data Source

| Etiquetas de fila | ANIO | DEPARTAMENTO | HUMEDAD.RELATIVA | MES | MUNICIPIO | PRECIPITACION | RADIACION | TEMPERATURA | TOTAL GENERAL |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | 1 | | | 4 | | | 5 |
| 2 | | | | | | | 1 | 4 | 5 |
| 3 | | | 2 | | | 1 | 2 | | 5 |
| 4 | | | 2 | | | | 2 | 1 | 5 |
| 5 | | | | | 5 | | | | 5 |
| 6 | | 5 | | | | | | | 5 |
| 7 | 5 | | | | | | | | 5 |
| 8 | | | | 5 | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 40 |

**Table F. 4.** Matches among raters in the positions of variables for the IDEAM dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | HUMEDAD.RELATIVA | CM |
| 1 | PRECIPITACION | J1 |
| 1 | PRECIPITACION | J2 |

| | | |
|---|---|---|
| 1 | PRECIPITACION | J3 |
| 1 | PRECIPITACION | J4 |
| 2 | RADIACION | CM |
| 2 | TEMPERATURA | J1 |
| 2 | TEMPERATURA | J2 |
| 2 | TEMPERATURA | J3 |
| 2 | TEMPERATURA | J4 |
| 3 | PRECIPITACION | CM |
| 3 | RADIACION | J1 |
| 3 | HUMEDAD.RELATIVA | J2 |
| 3 | RADIACION | J3 |
| 3 | HUMEDAD.RELATIVA | J4 |
| 4 | TEMPERATURA | CM |
| 4 | HUMEDAD.RELATIVA | J1 |
| 4 | RADIACION | J2 |
| 4 | HUMEDAD.RELATIVA | J3 |
| 4 | RADIACION | J4 |
| 5 | MUNICIPIO | CM |
| 5 | MUNICIPIO | J1 |
| 5 | MUNICIPIO | J2 |
| 5 | MUNICIPIO | J3 |
| 5 | MUNICIPIO | J4 |
| 6 | DEPARTAMENTO | CM |
| 6 | DEPARTAMENTO | J1 |
| 6 | DEPARTAMENTO | J2 |
| 6 | DEPARTAMENTO | J3 |
| 6 | DEPARTAMENTO | J4 |
| 7 | ANIO | CM |
| 7 | ANIO | J1 |
| 7 | ANIO | J2 |
| 7 | ANIO | J3 |
| 7 | ANIO | J4 |
| 8 | MES | CM |
| 8 | MES | J1 |
| 8 | MES | J2 |
| 8 | MES | J3 |
| 8 | MES | J4 |

**Table F. 5.** Ranking of variables established by J1, J2, J3, and J4 raters for the IDEAM dataset

| | Kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,66 | 0,04 | 15,55 | 0,00 | 0,57 | 0,74 |
| Anio | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Departamento | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Humedad.relativa | 0,09 | 0,11 | 0,77 | 0,44 | -0,13 | 0,30 |
| Mes | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Municipio | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Precipitacion | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |
| Radiacion | 0,09 | 0,11 | 0,77 | 0,44 | -0,13 | 0,30 |
| Temperatura | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |

**Table F. 6.** Fleiss' Kappa test metrics for the IDEAM dataset.

# F.3 FINAGRO Data Source

| Etiquetas de fila | ANIO | DEPARTAMENTO | ENTIDAD | TIPO.PRODUCTOR | VALOR | TOTAL GENERAL |
|---|---|---|---|---|---|---|
| 1 | | | | 2 | 3 | 5 |
| 2 | | | 1 | 3 | 1 | 5 |
| 3 | | 2 | 2 | | 1 | 5 |
| 4 | | 3 | 2 | | | 5 |
| 5 | 5 | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 25 |

**Table F. 7.** Matches among raters in the positions of variables for the FINAGRO dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | VALOR | CM |
| 1 | TIPO.PRODUCTOR | J1 |
| 1 | VALOR | J2 |
| 1 | TIPO.PRODUCTOR | J3 |
| 1 | VALOR | J4 |
| 2 | TIPO.PRODUCTOR | CM |
| 2 | ENTIDAD | J1 |
| 2 | TIPO.PRODUCTOR | J2 |
| 2 | VALOR | J3 |
| 2 | TIPO.PRODUCTOR | J4 |
| 3 | DEPARTAMENTO | CM |
| 3 | VALOR | J1 |
| 3 | DEPARTAMENTO | J2 |
| 3 | ENTIDAD | J3 |
| 3 | ENTIDAD | J4 |
| 4 | ENTIDAD | CM |
| 4 | DEPARTAMENTO | J1 |
| 4 | ENTIDAD | J2 |
| 4 | DEPARTAMENTO | J3 |
| 4 | DEPARTAMENTO | J4 |
| 5 | ANIO | CM |
| 5 | ANIO | J1 |
| 5 | ANIO | J2 |
| 5 | ANIO | J3 |
| 5 | ANIO | J4 |

**Table F. 8.** Ranking of variables established by J1, J2, J3, and J4 raters for the FINAGRO dataset

| | kappa | s.e. | z-stat | p-value | lower | Upper |
|---|---|---|---|---|---|---|
| Total | 0,33 | 0,07 | 4,60 | 0,00 | 0,19 | 0,46 |
| Anio | 1,00 | 0,14 | 7,07 | 0,00 | 0,72 | 1,28 |
| Departamento | 0,25 | 0,14 | 1,77 | 0,08 | -0,03 | 0,53 |
| Entidad | 0,00 | 0,14 | 0,00 | 1,00 | -0,28 | 0,28 |
| Tipo.productor | 0,25 | 0,14 | 1,77 | 0,08 | -0,03 | 0,53 |
| Valor | 0,13 | 0,14 | 0,88 | 0,38 | -0,15 | 0,40 |

**Table F. 9.** Fleiss' Kappa test metrics for the FINAGRO dataset.

# F.4 AVA Data Source

| Etiquetas de fila | CA_BANANA | CA_BEAN | CA_COCOA | CA_COFFEE | CA_POTATO | CA_SUGAR.CANE | DEPARTMENT | MUNICIPALITY | TOTAL GENERAL |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | 5 | | | | | 5 |
| 2 | | | | | | 5 | | | 5 |
| 3 | 2 | | 3 | | | | | | 5 |
| 4 | 3 | | 1 | | 1 | | | | 5 |
| 5 | | 1 | | | 4 | | | | 5 |
| 6 | | 4 | 1 | | | | | | 5 |
| 7 | | | | | | | | 5 | 5 |
| 8 | | | | | | | 5 | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 40 |

**Table F. 10**. Matches among raters in the positions of variables for the AVA dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | CA_Coffee | CM |
| 1 | CA_Coffee | J1 |
| 1 | CA_Coffee | J2 |
| 1 | CA_Coffee | J3 |
| 1 | CA_Coffee | J4 |
| 2 | CA_Sugar.Cane | CM |
| 2 | CA_Sugar.Cane | J1 |
| 2 | CA_Sugar.Cane | J2 |
| 2 | CA_Sugar.Cane | J3 |
| 2 | CA_Sugar.Cane | J4 |
| 3 | CA_Cocoa | CM |
| 3 | CA_Cocoa | J1 |
| 3 | CA_Banana | J2 |
| 3 | CA_Cocoa | J3 |
| 3 | CA_Banana | J4 |
| 4 | CA_Banana | CM |
| 4 | CA_Banana | J1 |
| 4 | CA_Cocoa | J2 |
| 4 | CA_Banana | J3 |
| 4 | CA_Potato | J4 |
| 5 | CA_Potato | CM |
| 5 | CA_Potato | J1 |
| 5 | CA_Potato | J2 |
| 5 | CA_Potato | J3 |
| 5 | CA_Bean | J4 |
| 6 | CA_Bean | CM |
| 6 | CA_Bean | J1 |
| 6 | CA_Bean | J2 |
| 6 | CA_Bean | J3 |
| 6 | CA_Cocoa | J4 |
| 7 | Municipality | CM |
| 7 | Municipality | J1 |
| 7 | Municipality | J2 |
| 7 | Municipality | J3 |
| 7 | Municipality | J4 |
| 8 | Department | CM |
| 8 | Department | J1 |
| 8 | Department | J2 |
| 8 | Department | J3 |
| 8 | Department | J4 |

**Table F. 11.** Ranking of variables established by J1, J2, J3, and J4 raters for the AVA dataset

| | kappa | s.e. | z-stat | p-value | lower | Upper |
|---|---|---|---|---|---|---|
| Total | 0,70 | 0,04 | 16,57 | 0,00 | 0,62 | 0,78 |
| CA_Banana | 0,31 | 0,11 | 2,81 | 0,00 | 0,10 | 0,53 |
| CA_Bean | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |
| CA_Cocoa | 0,20 | 0,11 | 1,79 | 0,07 | -0,02 | 0,42 |
| CA_Coffee | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| CA_Potato | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |
| CA_Sugar.Cane | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Department | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Municipality | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |

**Table F. 12.** Fleiss' Kappa test metrics for the AVA dataset.

# F.5 DANE-SIPSA-P Data Source

| Etiquetas de fila | ANIO | COSTO.TOTAL.PRODUCCION | CULTIVOS | INGRESO.BRUTO.PRODUCCION | PRECIO.AL.PRODUCTOR.KG | PRECIO.AL.PRODUCTOR.TON | PRODUCCION | RENDIMIENTO | RENTABILIDAD | SUPERFICIE.COSECHADA | SUPERFICIE.SEMBRADA | UTILIDAD | COSTO.PRODUCCION | Total general |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | 2 | 1 | 1 | | | 1 | | 5 |
| 2 | | | | | | | 1 | 2 | 1 | | | 1 | | 5 |
| 3 | | 1 | | | | | | 1 | | 2 | | | 1 | 5 |
| 4 | | | | | | | 1 | 1 | 1 | | 1 | 1 | | 5 |
| 5 | | | | 1 | | 1 | 1 | | 1 | | 1 | | | 5 |
| 6 | | 1 | | | 2 | 2 | | | | | | | | 5 |
| 7 | | | | | 2 | 1 | | | 1 | 1 | | | | 5 |
| 8 | | | 1 | 4 | | | | | | | | | | 5 |
| 9 | | | | | | 1 | | | | 1 | | | 3 | 5 |
| 10 | | 2 | | | 1 | | | | | 1 | 1 | | | 5 |
| 11 | | 1 | | | | | | | | | 2 | 2 | | 5 |
| 12 | 4 | | | | | | | | | | | | 1 | 5 |
| 13 | 1 | | 4 | | | | | | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 65 |

**Table F. 13.** Matches among raters in the positions of variables for the DANE – SIPSA-P dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | PRODUCCION | CM |
| 1 | UTILIDAD | J1 |
| 1 | PRODUCCION | J2 |
| 1 | RENDIMIENTO | J3 |
| 1 | RENTABILIDAD | J4 |
| 2 | RENDIMIENTO | CM |
| 2 | RENTABILIDAD | J1 |
| 2 | RENDIMIENTO | J2 |
| 2 | PRODUCCION | J3 |
| 2 | UTILIDAD | J4 |
| 3 | SUPERFICIE.COSECHADA | CM |
| 3 | RENDIMIENTO | J1 |
| 3 | SUPERFICIE.COSECHADA | J2 |
| 3 | COSTO.TOTAL.PRODUCCION | J3 |
| 3 | COSTO.PRODUCCION | J4 |
| 4 | RENTABILIDAD | CM |
| 4 | PRODUCCION | J1 |
| 4 | SUPERFICIE.SEMBRADA | J2 |
| 4 | UTILIDAD | J3 |
| 4 | RENDIMIENTO | J4 |
| 5 | SUPERFICIE.SEMBRADA | CM |
| 5 | INGRESO.BRUTO.PRODUCCION | J1 |
| 5 | RENTABILIDAD | J2 |
| 5 | PRECIO.AL.PRODUCTOR.TON | J3 |
| 5 | PRODUCCION | J4 |
| 6 | PRECIO.AL.PRODUCTOR.TON | CM |
| 6 | PRECIO.AL.PRODUCTOR.KG | J1 |
| 6 | PRECIO.AL.PRODUCTOR.TON | J2 |
| 6 | PRECIO.AL.PRODUCTOR.KG | J3 |
| 6 | COSTO.TOTAL.PRODUCCION | J4 |
| 7 | PRECIO.AL.PRODUCTOR.KG | CM |
| 7 | PRECIO.AL.PRODUCTOR.TON | J1 |

| | | |
|---|---|---|
| 7 | PRECIO.AL.PRODUCTOR.KG | J2 |
| 7 | RENTABILIDAD | J3 |
| 7 | SUPERFICIE.COSECHADA | J4 |
| 8 | INGRESO.BRUTO.PRODUCCION | CM |
| 8 | CULTIVOS | J1 |
| 8 | INGRESO.BRUTO.PRODUCCION | J2 |
| 8 | INGRESO.BRUTO.PRODUCCION | J3 |
| 8 | INGRESO.BRUTO.PRODUCCION | J4 |
| 9 | COSTO.PRODUCCION | CM |
| 9 | SUPERFICIE.COSECHADA | J1 |
| 9 | COSTO.PRODUCCION | J2 |
| 9 | COSTO.PRODUCCION | J3 |
| 9 | PRECIO.AL.PRODUCTOR.TON | J4 |
| 10 | COSTO.TOTAL.PRODUCCION | CM |
| 10 | SUPERFICIE.SEMBRADA | J1 |
| 10 | COSTO.TOTAL.PRODUCCION | J2 |
| 10 | SUPERFICIE.COSECHADA | J3 |
| 10 | PRECIO.AL.PRODUCTOR.KG | J4 |
| 11 | UTILIDAD | CM |
| 11 | COSTO.TOTAL.PRODUCCION | J1 |
| 11 | UTILIDAD | J2 |
| 11 | SUPERFICIE.SEMBRADA | J3 |
| 11 | SUPERFICIE.SEMBRADA | J4 |
| 12 | ANIO | CM |
| 12 | COSTO.PRODUCCION | J1 |
| 12 | ANIO | J2 |
| 12 | ANIO | J3 |
| 12 | ANIO | J4 |
| 13 | CULTIVOS | CM |
| 13 | ANIO | J1 |
| 13 | CULTIVOS | J2 |
| 13 | CULTIVOS | J3 |
| 13 | CULTIVOS | J4 |

**Table F. 14.** Ranking of variables established by J1, J2, J3, and J4 raters for the DANE – SIPSA-P dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,17 | 0,03 | 6,58 | 0,00 | 0,12 | 0,22 |
| ANIO | 0,57 | 0,09 | 6,46 | 0,00 | 0,39 | 0,74 |
| COSTO.TOTAL.PRODUCCION | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| CULTIVOS | 0,57 | 0,09 | 6,46 | 0,00 | 0,39 | 0,74 |
| INGRESO.BRUTO.PRODUCCION | 0,57 | 0,09 | 6,46 | 0,00 | 0,39 | 0,74 |
| PRECIO.AL.PRODUCTOR.KG | 0,13 | 0,09 | 1,52 | 0,13 | -0,04 | 0,31 |
| PRECIO.AL.PRODUCTOR.TON | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| PRODUCCION | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| RENDIMIENTO | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| RENTABILIDAD | -0,08 | 0,09 | -0,95 | 0,34 | -0,26 | 0,09 |
| SUPERFICIE.COSECHADA | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| SUPERFICIE.SEMBRADA | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| UTILIDAD | 0,03 | 0,09 | 0,29 | 0,78 | -0,15 | 0,20 |
| COSTO.PRODUCCION | 0,24 | 0,09 | 2,76 | 0,01 | 0,07 | 0,41 |

**Table F. 15.** Fleiss' Kappa test metrics for the DANE-SIPSA-P dataset.

# F.6 Agronet Data Source

| Etiquetas de fila | Anio | Area.Cosechada | Area.Sembrada | Cultivo | Departamento | Municipio | Produccion | Rendimiento | Total general |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | 3 | | | | 1 | 1 | 5 |
| 2 | | 3 | | | | | 1 | 1 | 5 |
| 3 | | | | 2 | | | 2 | 1 | 5 |
| 4 | | | | | | 2 | 1 | 2 | 5 |
| 5 | | | 1 | 3 | 1 | | | | 5 |
| 6 | | 2 | | | | 3 | | | 5 |
| 7 | | | 1 | | 4 | | | | 5 |
| 8 | 5 | | | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 40 |

**Table F. 16.** Matches among raters in the positions of variables for the Agronet dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | Area.Sembrada | CM |
| 1 | Produccion | J1 |
| 1 | Area.Sembrada | J2 |
| 1 | Area.Sembrada | J3 |
| 1 | Rendimiento | J4 |
| 2 | Area.Cosechada | CM |
| 2 | Rendimiento | J1 |
| 2 | Area.Cosechada | J2 |
| 2 | Area.Cosechada | J3 |
| 2 | Produccion | J4 |
| 3 | Produccion | CM |
| 3 | Cultivo | J1 |
| 3 | Rendimiento | J2 |
| 3 | Produccion | J3 |
| 3 | Cultivo | J4 |
| 4 | Rendimiento | CM |
| 4 | Municipio | J1 |
| 4 | Produccion | J2 |
| 4 | Rendimiento | J3 |
| 4 | Municipio | J4 |
| 5 | Cultivo | CM |
| 5 | Departamento | J1 |
| 5 | Cultivo | J2 |
| 5 | Cultivo | J3 |
| 5 | Area.Sembrada | J4 |
| 6 | Municipio | CM |
| 6 | Area.Cosechada | J1 |
| 6 | Municipio | J2 |
| 6 | Municipio | J3 |
| 6 | Area.Cosechada | J4 |
| 7 | Departamento | CM |
| 7 | Area.Sembrada | J1 |
| 7 | Departamento | J2 |
| 7 | Departamento | J3 |
| 7 | Departamento | J4 |
| 8 | Anio | CM |
| 8 | Anio | J1 |
| 8 | Anio | J2 |
| 8 | Anio | J3 |
| 8 | Anio | J4 |

**Table F. 17.** Ranking of variables established by J1, J2, J3, and J4 raters for the Agronet dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,33 | 0,04 | 7,78 | 0,00 | 0,25 | 0,41 |
| Anio | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| Area.Cosechada | 0,31 | 0,11 | 2,81 | 0,00 | 0,10 | 0,53 |
| Area.Sembrada | 0,20 | 0,11 | 1,79 | 0,07 | -0,02 | 0,42 |
| Cultivo | 0,31 | 0,11 | 2,81 | 0,00 | 0,10 | 0,53 |
| Departamento | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |
| Municipio | 0,31 | 0,11 | 2,81 | 0,00 | 0,10 | 0,53 |
| Produccion | -0,03 | 0,11 | -0,26 | 0,80 | -0,25 | 0,19 |
| Rendimiento | -0,03 | 0,11 | -0,26 | 0,80 | -0,25 | 0,19 |

**Table F. 18.** Fleiss' Kappa test metrics for the Agronet dataset.

# F.7 Minagricultura Data Source

| Etiquetas de fila | AREA.COSECHADA | AREA.SEMBRADA | COD.CULTIVO | COD.DEP | COD.MUN | CULTIVO | DEPARTAMENTO | ESTADO.FISICO.PRODUCCION | GRUPO.CULTIVO | MUNICIPIO | NOMBRE..CIENTIFICO | PERIODO | PRODUCCION | RENDIMIENTO | SISTEMA.PRODUCTIVO | SUBGRUPO.CULTIVO | Total general |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | | | | | | | | | 1 | | 2 | | 5 |
| 2 | | 2 | | | | | | 2 | | | | | | 1 | | | 5 |
| 3 | | | | | | | | 1 | | | | | 3 | 1 | | | 5 |
| 4 | | | | | | 1 | | | | | | | 1 | 3 | | | 5 |
| 5 | | | | | | 3 | | | | | 1 | | | | 1 | | 5 |
| 6 | 1 | | 1 | | | 1 | | | | | | | | | 2 | | 5 |
| 7 | | 1 | | | | | | 3 | 1 | | | | | | | | 5 |
| 8 | | | | | | | | | | | | | | | | 5 | 5 |
| 9 | | | | | | | | | 4 | 1 | | | | | | | 5 |
| 10 | | | | | 1 | | | | | | 4 | | | | | | 5 |
| 11 | | | | | | | 1 | | | 4 | | | | | | | 5 |
| 12 | | | | 1 | | | 4 | | | | | | | | | | 5 |
| 13 | 1 | | | | | | | | | | | 4 | | | | | 5 |
| 14 | | 1 | 4 | | | | | | | | | | | | | | 5 |
| 15 | | | | | 4 | | | | | | | | | | 1 | | 5 |
| 16 | | | | 4 | | | | | | | | 1 | | | | | 5 |
| Total general | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 5 | 80 |

**Table F. 19.** Matches among raters in the positions of variables for the Minagricultura dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | AREA.COSECHADA | CM |
| 1 | PRODUCCION | J1 |
| 1 | AREA.COSECHADA | J2 |
| 1 | SISTEMA.PRODUCTIVO | J3 |
| 1 | SISTEMA.PRODUCTIVO | J4 |
| 2 | AREA.SEMBRADA | CM |
| 2 | RENDIMIENTO | J1 |
| 2 | AREA.SEMBRADA | J2 |
| 2 | ESTADO.FISICO.PRODUCCION | J3 |
| 2 | ESTADO.FISICO.PRODUCCION | J4 |
| 3 | PRODUCCION | CM |
| 3 | ESTADO.FISICO.PRODUCCION | J1 |
| 3 | PRODUCCION | J2 |
| 3 | PRODUCCION | J3 |

| | | |
|---|---|---|
| 3 | RENDIMIENTO | J4 |
| 4 | RENDIMIENTO | CM |
| 4 | CULTIVO | J1 |
| 4 | RENDIMIENTO | J2 |
| 4 | RENDIMIENTO | J3 |
| 4 | PRODUCCION | J4 |
| 5 | CULTIVO | CM |
| 5 | NOMBRE..CIENTIFICO | J1 |
| 5 | SISTEMA.PRODUCTIVO | J2 |
| 5 | CULTIVO | J3 |
| 5 | CULTIVO | J4 |
| 6 | SISTEMA.PRODUCTIVO | CM |
| 6 | COD.CULTIVO | J1 |
| 6 | CULTIVO | J2 |
| 6 | SISTEMA.PRODUCTIVO | J3 |
| 6 | AREA.COSECHADA | J4 |
| 7 | ESTADO.FISICO.PRODUCCION | CM |
| 7 | GRUPO.CULTIVO | J1 |
| 7 | ESTADO.FISICO.PRODUCCION | J2 |
| 7 | ESTADO.FISICO.PRODUCCION | J3 |
| 7 | AREA.SEMBRADA | J4 |
| 8 | SUBGRUPO.CULTIVO | CM |
| 8 | SUBGRUPO.CULTIVO | J1 |
| 8 | SUBGRUPO.CULTIVO | J2 |
| 8 | SUBGRUPO.CULTIVO | J3 |
| 8 | SUBGRUPO.CULTIVO | J4 |
| 9 | GRUPO.CULTIVO | CM |
| 9 | MUNICIPIO | J1 |
| 9 | GRUPO.CULTIVO | J2 |
| 9 | GRUPO.CULTIVO | J3 |
| 9 | GRUPO.CULTIVO | J4 |
| 10 | NOMBRE..CIENTIFICO | CM |
| 10 | COD.MUN | J1 |
| 10 | NOMBRE..CIENTIFICO | J2 |
| 10 | NOMBRE..CIENTIFICO | J3 |
| 10 | NOMBRE..CIENTIFICO | J4 |
| 11 | MUNICIPIO | CM |
| 11 | DEPARTAMENTO | J1 |
| 11 | MUNICIPIO | J2 |
| 11 | MUNICIPIO | J3 |
| 11 | MUNICIPIO | J4 |
| 12 | DEPARTAMENTO | CM |
| 12 | COD.DEP | J1 |
| 12 | DEPARTAMENTO | J2 |
| 12 | DEPARTAMENTO | J3 |
| 12 | DEPARTAMENTO | J4 |
| 13 | PERIODO | CM |
| 13 | AREA.COSECHADA | J1 |
| 13 | PERIODO | J2 |
| 13 | PERIODO | J3 |
| 13 | PERIODO | J4 |
| 14 | COD.CULTIVO | CM |
| 14 | AREA.SEMBRADA | J1 |
| 14 | COD.CULTIVO | J2 |
| 14 | COD.CULTIVO | J3 |
| 14 | COD.CULTIVO | J4 |
| 15 | COD.MUN | CM |
| 15 | SISTEMA.PRODUCTIVO | J1 |
| 15 | COD.MUN | J2 |
| 15 | COD.MUN | J3 |
| 15 | COD.MUN | J4 |
| 16 | COD.DEP | CM |
| 16 | PERIODO | J1 |
| 16 | COD.DEP | J2 |
| 16 | COD.DEP | J3 |

| | | |
|---|---|---|
| 16 | COD.DEP | J4 |

**Table F. 20.** Ranking of variables established by J1, J2, J3, and J4 raters for the Minagricultura dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,43 | 0,02 | 21,11 | 0,00 | 0,39 | 0,47 |
| AREA.COSECHADA | 0,08 | 0,08 | 1,00 | 0,32 | -0,08 | 0,23 |
| AREA.SEMBRADA | 0,08 | 0,08 | 1,00 | 0,32 | -0,08 | 0,23 |
| COD.CULTIVO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| COD.DEP | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| COD.MUN | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| CULTIVO | 0,25 | 0,08 | 3,20 | 0,00 | 0,10 | 0,41 |
| DEPARTAMENTO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| ESTADO.FISICO.PRODUCCION | 0,28 | 0,08 | 3,53 | 0,00 | 0,12 | 0,43 |
| GRUPO.CULTIVO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| MUNICIPIO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| NOMBRE..CIENTIFICO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| PERIODO | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| PRODUCCION | 0,25 | 0,08 | 3,20 | 0,00 | 0,10 | 0,41 |
| RENDIMIENTO | 0,25 | 0,08 | 3,20 | 0,00 | 0,10 | 0,41 |
| SISTEMA.PRODUCTIVO | 0,10 | 0,08 | 1,25 | 0,21 | -0,06 | 0,25 |
| SUBGRUPO.CULTIVO | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |

**Table F. 21.** Fleiss' Kappa test metrics for the Minagricultura dataset.

# F.8 Agronet-P Data Source

| Etiquetas de fila | FECHA | FUENTE | PRECIO | PRODUCTO | UNIDAD | VARIACION | TOTAL GENERAL |
|---|---|---|---|---|---|---|---|
| 1 | | | 5 | | | | 5 |
| 2 | | | | 2 | | 3 | 5 |
| 3 | | 1 | | 3 | | 1 | 5 |
| 4 | | 4 | | | | 1 | 5 |
| 5 | | | | | 5 | | 5 |
| 6 | 5 | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 30 |

**Table F. 22.** Matches among raters in the positions of variables for the Agronet-P dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | Precio | CM |
| 1 | Precio | J1 |
| 1 | Precio | J2 |
| 1 | Precio | J3 |
| 1 | Precio | J4 |
| 2 | Variacion | CM |
| 2 | Producto | J1 |
| 2 | Variacion | J2 |
| 2 | Producto | J3 |
| 2 | Variacion | J4 |
| 3 | Producto | CM |
| 3 | Fuente | J1 |
| 3 | Producto | J2 |
| 3 | Variacion | J3 |
| 3 | Producto | J4 |
| 4 | Fuente | CM |
| 4 | Variacion | J1 |
| 4 | Fuente | J2 |
| 4 | Fuente | J3 |
| 4 | Fuente | J4 |
| 5 | Unidad | CM |

| | | |
|---|---|---|
| 5 | Unidad | J1 |
| 5 | Unidad | J2 |
| 5 | Unidad | J3 |
| 5 | Unidad | J4 |
| 6 | Fecha | CM |
| 6 | Fecha | J1 |
| 6 | Fecha | J2 |
| 6 | Fecha | J3 |
| 6 | Fecha | J4 |

**Table F. 23**. Ranking of variables established by J1, J2, J3, and J4 raters for the Agronet-P dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,66 | 0,06 | 11,43 | 0,00 | 0,55 | 0,77 |
| Fecha | 1,00 | 0,13 | 7,75 | 0,00 | 0,75 | 1,25 |
| Fuente | 0,52 | 0,13 | 4,03 | 0,00 | 0,27 | 0,77 |
| Precio | 1,00 | 0,13 | 7,75 | 0,00 | 0,75 | 1,25 |
| Producto | 0,28 | 0,13 | 2,17 | 0,03 | 0,03 | 0,53 |
| Unidad | 1,00 | 0,13 | 7,75 | 0,00 | 0,75 | 1,25 |
| Variacion | 0,16 | 0,13 | 1,24 | 0,22 | -0,09 | 0,41 |

**Table F. 24**. Fleiss' Kappa test metrics for the Agronet-P dataset.

# F.9 DANE-SIPSA Data Source

| Etiquetas de fila | FECHA | FUENTE | GRUPO | PRECIO | PRODUCTO | TOTAL GENERAL |
|---|---|---|---|---|---|---|
| 1 | | 1 | | 3 | 1 | 5 |
| 2 | | 1 | 1 | 1 | 2 | 5 |
| 3 | | 2 | | 1 | 2 | 5 |
| 4 | | 1 | 4 | | | 5 |
| 5 | 5 | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 25 |

**Table F. 25.** Matches among raters in the positions of variables for the DANE- SIPSA dataset.

| **Ranking** | **Classification** | **Judge** |
|---|---|---|
| 1 | Precio | CM |
| 1 | Producto | J1 |
| 1 | Precio | J2 |
| 1 | Precio | J3 |
| 1 | Fuente | J4 |
| 2 | Producto | CM |
| 2 | Grupo | J1 |
| 2 | Fuente | J2 |
| 2 | Producto | J3 |
| 2 | Precio | J4 |
| 3 | Fuente | CM |
| 3 | Precio | J1 |
| 3 | Producto | J2 |
| 3 | Fuente | J3 |
| 3 | Producto | J4 |
| 4 | Grupo | CM |
| 4 | Fuente | J1 |
| 4 | Grupo | J2 |
| 4 | Grupo | J3 |
| 4 | Grupo | J4 |
| 5 | Fecha | CM |
| 5 | Fecha | J1 |
| 5 | Fecha | J2 |
| 5 | Fecha | J3 |
| 5 | Fecha | J4 |

**Table F. 26.** Ranking of variables established by J1, J2, J3, and J4 raters for the DANE- SIPSA dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,30 | 0,07 | 4,24 | 0,00 | 0,16 | 0,44 |
| Fecha | 1,00 | 0,14 | 7,07 | 0,00 | 0,72 | 1,28 |
| Fuente | -0,13 | 0,14 | -0,88 | 0,38 | -0,40 | 0,15 |
| Grupo | 0,50 | 0,14 | 3,54 | 0,00 | 0,22 | 0,78 |
| Precio | 0,13 | 0,14 | 0,88 | 0,38 | -0,15 | 0,40 |
| Producto | 0,00 | 0,14 | 0,00 | 1,00 | -0,28 | 0,28 |

**Table F. 27**. Fleiss' Kappa test metrics for the DANE- SIPSA dataset.

# F.10 DANE-HH Data Source

| Etiquetas de fila | COD_VEREDA | ENCUESTA | ITER_HG | P_DEPTO | P_MUNIC | P_S15P165A | P_S15P165B | P_S15P177 | P_S15P178 | P_S15P179_SP1 | P_S15P179_SP2 | P_S15P179_SP3 | P_S15P179_SP4 | P_S15P179A | P_S15P179B | P_S15P179C | P_S15P180 | PAIS | TIPO_REG | TOT_PROD_HOGAR | UC_UO | Total general |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | 1 | | | | | | | | 4 | | 5 |
| 2 | | | 1 | | | | | | | 2 | | 1 | | | | | | | | 1 | | 5 |
| 3 | | | | | | | | | | 1 | 4 | | | | | | | | | | | 5 |
| 4 | | | | | | | | | | 2 | | 3 | | | | | | | | | | 5 |
| 5 | | | | | | | | | | | | | 5 | | | | | | | | | 5 |
| 6 | | | | | | 2 | | | | | 1 | | | 1 | | | 1 | | | | | 5 |
| 7 | | | | | | 1 | 3 | | | | | | | | 1 | | | | | | | 5 |
| 8 | | | | | | | | 4 | | | | | | | | 1 | | | | | | 5 |
| 9 | | | | | | | | | 4 | | | | | | | | 1 | | | | | 5 |
| 10 | | | | | | 1 | | | | | | | | 3 | | | 1 | | | | | 5 |
| 11 | | | | | | | 1 | | | | | | | 1 | 3 | | | | | | | 5 |
| 12 | | | | | | | | 1 | | | | | | | 1 | 3 | | | | | | 5 |
| 13 | | | | | | 1 | | | 1 | | | | | | | 1 | 2 | | | | | 5 |
| 14 | | | | | 5 | | | | | | | | | | | | | | | | | 5 |
| 15 | 5 | | | | | | | | | | | | | | | | | | | | | 5 |
| 16 | | | 4 | | | | 1 | | | | | | | | | | | | | | | 5 |
| 17 | | | | | | | | | | | | | | | | | | | | | 5 | 5 |
| 18 | | 5 | | | | | | | | | | | | | | | | | | | | 5 |
| 19 | | | | 5 | | | | | | | | | | | | | | | | | | 5 |
| 20 | | | | | | | | | | | | | | | | | | 5 | | | | 5 |
| 21 | | | | | | | | | | | | | | | | | | | 5 | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 105 |

**Table F. 28.** Matches among raters in the positions of variables for the DANE-HH dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | TOT_PROD_HOGAR | CM |
| 1 | TOT_PROD_HOGAR | J1 |
| 1 | TOT_PROD_HOGAR | J2 |
| 1 | P_S15P179_SP3 | J3 |
| 1 | TOT_PROD_HOGAR | J4 |
| 2 | P_S15P179_SP1 | CM |
| 2 | P_S15P179_SP1 | J1 |
| 2 | P_S15P179_SP3 | J2 |
| 2 | TOT_PROD_HOGAR | J3 |
| 2 | ITER_HG | J4 |
| 3 | P_S15P179_SP2 | CM |
| 3 | P_S15P179_SP2 | J1 |

| | | |
|---|---|---|
| 3 | P_S15P179_SP2 | J2 |
| 3 | P_S15P179_SP2 | J3 |
| 3 | P_S15P179_SP1 | J4 |
| 4 | P_S15P179_SP3 | CM |
| 4 | P_S15P179_SP3 | J1 |
| 4 | P_S15P179_SP1 | J2 |
| 4 | P_S15P179_SP1 | J3 |
| 4 | P_S15P179_SP3 | J4 |
| 5 | P_S15P179_SP4 | CM |
| 5 | P_S15P179_SP4 | J1 |
| 5 | P_S15P179_SP4 | J2 |
| 5 | P_S15P179_SP4 | J3 |
| 5 | P_S15P179_SP4 | J4 |
| 6 | P_S15P165A | CM |
| 6 | P_S15P179A | J1 |
| 6 | P_S15P165A | J2 |
| 6 | P_S15P180 | J3 |
| 6 | P_S15P179_SP2 | J4 |
| 7 | P_S15P165B | CM |
| 7 | P_S15P179B | J1 |
| 7 | P_S15P165B | J2 |
| 7 | P_S15P165B | J3 |
| 7 | P_S15P165A | J4 |
| 8 | P_S15P177 | CM |
| 8 | P_S15P179C | J1 |
| 8 | P_S15P177 | J2 |
| 8 | P_S15P177 | J3 |
| 8 | P_S15P177 | J4 |
| 9 | P_S15P178 | CM |
| 9 | P_S15P180 | J1 |
| 9 | P_S15P178 | J2 |
| 9 | P_S15P178 | J3 |
| 9 | P_S15P178 | J4 |
| 10 | P_S15P179A | CM |
| 10 | P_S15P165A | J1 |
| 10 | P_S15P180 | J2 |
| 10 | P_S15P179A | J3 |
| 10 | P_S15P179A | J4 |
| 11 | P_S15P179B | CM |
| 11 | P_S15P165B | J1 |
| 11 | P_S15P179A | J2 |
| 11 | P_S15P179B | J3 |
| 11 | P_S15P179B | J4 |
| 12 | P_S15P179C | CM |
| 12 | P_S15P177 | J1 |
| 12 | P_S15P179B | J2 |
| 12 | P_S15P179C | J3 |
| 12 | P_S15P179C | J4 |
| 13 | P_S15P180 | CM |
| 13 | P_S15P178 | J1 |
| 13 | P_S15P179C | J2 |
| 13 | P_S15P165A | J3 |
| 13 | P_S15P180 | J4 |
| 14 | P_MUNIC | CM |
| 14 | P_MUNIC | J1 |
| 14 | P_MUNIC | J2 |
| 14 | P_MUNIC | J3 |
| 14 | P_MUNIC | J4 |
| 15 | COD_VEREDA | CM |
| 15 | COD_VEREDA | J1 |
| 15 | COD_VEREDA | J2 |
| 15 | COD_VEREDA | J3 |
| 15 | COD_VEREDA | J4 |
| 16 | ITER_HG | CM |
| 16 | ITER_HG | J1 |

| | | |
|---|---|---|
| 16 | ITER_HG | J2 |
| 16 | ITER_HG | J3 |
| 16 | P_S15P165B | J4 |
| 17 | UC_UO | CM |
| 17 | UC_UO | J1 |
| 17 | UC_UO | J2 |
| 17 | UC_UO | J3 |
| 17 | UC_UO | J4 |
| 18 | ENCUESTA | CM |
| 18 | ENCUESTA | J1 |
| 18 | ENCUESTA | J2 |
| 18 | ENCUESTA | J3 |
| 18 | ENCUESTA | J4 |
| 19 | P_DEPTO | CM |
| 19 | P_DEPTO | J1 |
| 19 | P_DEPTO | J2 |
| 19 | P_DEPTO | J3 |
| 19 | P_DEPTO | J4 |
| 20 | PAIS | CM |
| 20 | PAIS | J1 |
| 20 | PAIS | J2 |
| 20 | PAIS | J3 |
| 20 | PAIS | J4 |
| 21 | TIPO_REG | CM |
| 21 | TIPO_REG | J1 |
| 21 | TIPO_REG | J2 |
| 21 | TIPO_REG | J3 |
| 21 | TIPO_REG | J4 |

**Table F. 29.** Ranking of variables established by J1, J2, J3, and J4 raters for the DANE-HH dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,60 | 0,02 | 38,56 | 0,00 | 0,56 | 0,63 |
| COD_VEREDA | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| ENCUESTA | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| ITER_HG | 0,58 | 0,07 | 8,40 | 0,00 | 0,44 | 0,72 |
| P_DEPTO | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| P_MUNIC | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| P_S15P165A | 0,06 | 0,07 | 0,80 | 0,43 | -0,08 | 0,19 |
| P_S15P165B | 0,27 | 0,07 | 3,84 | 0,00 | 0,13 | 0,40 |
| P_S15P177 | 0,58 | 0,07 | 8,40 | 0,00 | 0,44 | 0,72 |
| P_S15P178 | 0,58 | 0,07 | 8,40 | 0,00 | 0,44 | 0,72 |
| P_S15P179_SP1 | 0,16 | 0,07 | 2,32 | 0,02 | 0,02 | 0,30 |
| P_S15P179_SP2 | 0,58 | 0,07 | 8,40 | 0,00 | 0,44 | 0,72 |
| P_S15P179_SP3 | 0,27 | 0,07 | 3,84 | 0,00 | 0,13 | 0,40 |
| P_S15P179_SP4 | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| P_S15P179A | 0,27 | 0,07 | 3,84 | 0,00 | 0,13 | 0,40 |
| P_S15P179B | 0,27 | 0,07 | 3,84 | 0,00 | 0,13 | 0,40 |
| P_S15P179C | 0,27 | 0,07 | 3,84 | 0,00 | 0,13 | 0,40 |
| P_S15P180 | 0,06 | 0,07 | 0,80 | 0,43 | -0,08 | 0,19 |
| PAIS | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| TIPO_REG | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |
| TOT_PROD_HOGAR | 0,58 | 0,07 | 8,40 | 0,00 | 0,44 | 0,72 |
| UC_UO | 1,00 | 0,07 | 14,49 | 0,00 | 0,86 | 1,14 |

**Table F. 30.** Fleiss' Kappa test metrics for the DANE-HH dataset.

## F.11 DANE-H Data Source

| Etiquetas de fila | COD_VEREDA | ENCUESTA | P_DEPTO | P_MUNIC | P_S15P159 | P_S15P160 | P_S15P161 | P_S15P162 | P_S15P163 | P_S15P164_SP1 | P_S15P164_SP2 | P_S15P164_SP3 | P_S15P164_SP4 | PAIS | TIPO_REG | UC_UO | Total general |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | 2 | | 3 | | | | | 5 |
| 2 | | | | | | | | | | 1 | 4 | | | | | | 5 |
| 3 | | | | | | | | | | 2 | 1 | 2 | | | | | 5 |
| 4 | | | | | | | | | | | | | 5 | | | | 5 |
| 5 | | | | | | | 5 | | | | | | | | | | 5 |
| 6 | | | | | | | | 5 | | | | | | | | | 5 |
| 7 | | | | | | | | | 5 | | | | | | | | 5 |
| 8 | | | | | 5 | | | | | | | | | | | | 5 |
| 9 | | | | | | 5 | | | | | | | | | | | 5 |
| 10 | | | | 5 | | | | | | | | | | | | | 5 |
| 11 | 5 | | | | | | | | | | | | | | | | 5 |
| 12 | | | | | | | | | | | | | | | | 5 | 5 |
| 13 | | 5 | | | | | | | | | | | | | | | 5 |
| 14 | | | 5 | | | | | | | | | | | | | | 5 |
| 15 | | | | | | | | | | | | | | 5 | | | 5 |
| 16 | | | | | | | | | | | | | | | 5 | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 80 |

**Table F. 31.** Matches among raters in the positions of variables for the DANE-H dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | P_S15P164_SP1 | CM |
| 1 | P_S15P164_SP1 | J1 |
| 1 | P_S15P164_SP3 | J2 |
| 1 | P_S15P164_SP3 | J3 |
| 1 | P_S15P164_SP3 | J4 |
| 2 | P_S15P164_SP2 | CM |
| 2 | P_S15P164_SP2 | J1 |
| 2 | P_S15P164_SP2 | J2 |
| 2 | P_S15P164_SP1 | J3 |
| 2 | P_S15P164_SP2 | J4 |
| 3 | P_S15P164_SP3 | CM |
| 3 | P_S15P164_SP3 | J1 |
| 3 | P_S15P164_SP1 | J2 |
| 3 | P_S15P164_SP2 | J3 |
| 3 | P_S15P164_SP1 | J4 |
| 4 | P_S15P164_SP4 | CM |
| 4 | P_S15P164_SP4 | J1 |
| 4 | P_S15P164_SP4 | J2 |
| 4 | P_S15P164_SP4 | J3 |
| 4 | P_S15P164_SP4 | J4 |
| 5 | P_S15P161 | CM |
| 5 | P_S15P161 | J1 |
| 5 | P_S15P161 | J2 |
| 5 | P_S15P161 | J3 |
| 5 | P_S15P161 | J4 |
| 6 | P_S15P162 | CM |
| 6 | P_S15P162 | J1 |
| 6 | P_S15P162 | J2 |
| 6 | P_S15P162 | J3 |
| 6 | P_S15P162 | J4 |
| 7 | P_S15P163 | CM |
| 7 | P_S15P163 | J1 |
| 7 | P_S15P163 | J2 |
| 7 | P_S15P163 | J3 |
| 7 | P_S15P163 | J4 |
| 8 | P_S15P159 | CM |
| 8 | P_S15P159 | J1 |
| 8 | P_S15P159 | J2 |

| | | |
|---|---|---|
| 8 | P_S15P159 | J3 |
| 8 | P_S15P159 | J4 |
| 9 | P_S15P160 | CM |
| 9 | P_S15P160 | J1 |
| 9 | P_S15P160 | J2 |
| 9 | P_S15P160 | J3 |
| 9 | P_S15P160 | J4 |
| 10 | P_MUNIC | CM |
| 10 | P_MUNIC | J1 |
| 10 | P_MUNIC | J2 |
| 10 | P_MUNIC | J3 |
| 10 | P_MUNIC | J4 |
| 11 | COD_VEREDA | CM |
| 11 | COD_VEREDA | J1 |
| 11 | COD_VEREDA | J2 |
| 11 | COD_VEREDA | J3 |
| 11 | COD_VEREDA | J4 |
| 12 | UC_UO | CM |
| 12 | UC_UO | J1 |
| 12 | UC_UO | J2 |
| 12 | UC_UO | J3 |
| 12 | UC_UO | J4 |
| 13 | ENCUESTA | CM |
| 13 | ENCUESTA | J1 |
| 13 | ENCUESTA | J2 |
| 13 | ENCUESTA | J3 |
| 13 | ENCUESTA | J4 |
| 14 | P_DEPTO | CM |
| 14 | P_DEPTO | J1 |
| 14 | P_DEPTO | J2 |
| 14 | P_DEPTO | J3 |
| 14 | P_DEPTO | J4 |
| 15 | PAIS | CM |
| 15 | PAIS | J1 |
| 15 | PAIS | J2 |
| 15 | PAIS | J3 |
| 15 | PAIS | J4 |
| 16 | TIPO_REG | CM |
| 16 | TIPO_REG | J1 |
| 16 | TIPO_REG | J2 |
| 16 | TIPO_REG | J3 |
| 16 | TIPO_REG | J4 |

**Table F. 32.** Ranking of variables established by J1, J2, J3, and J4 raters for the DANE-H dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,88 | 0,02 | 43,11 | 0,00 | 0,84 | 0,92 |
| COD_VEREDA | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| ENCUESTA | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_DEPTO | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_MUNIC | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P159 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P160 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P161 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P162 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P163 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| P_S15P164_SP1 | 0,15 | 0,08 | 1,86 | 0,06 | -0,01 | 0,30 |
| P_S15P164_SP2 | 0,57 | 0,08 | 7,25 | 0,00 | 0,42 | 0,73 |
| P_S15P164_SP3 | 0,36 | 0,08 | 4,55 | 0,00 | 0,21 | 0,51 |
| P_S15P164_SP4 | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| PAIS | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| TIPO_REG | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |
| UC_UO | 1,00 | 0,08 | 12,65 | 0,00 | 0,85 | 1,15 |

**Table F. 33.** Fleiss' Kappa test metrics for the DANE-H dataset.

## F.12 DNP-AIB Data Source

| Etiquetas de fila | ANIO | COD.ENTIDAD | CUENTA.PROYECTO | ENTIDAD | OBJETIVO | SECTOR | TIPO.ENTIDAD | VALOR | Total general |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | 1 | | | 4 | 5 |
| 2 | | | | | 3 | 1 | | 1 | 5 |
| 3 | | | | | 1 | 4 | | | 5 |
| 4 | | | | | | | 5 | | 5 |
| 5 | | | | 5 | | | | | 5 |
| 6 | | | 5 | | | | | | 5 |
| 7 | | 5 | | | | | | | 5 |
| 8 | 5 | | | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 40 |

**Table F. 34.** Matches among raters in the positions of variables for the DNP-AIB dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | VALOR | CM |
| 1 | VALOR | J1 |
| 1 | VALOR | J2 |
| 1 | VALOR | J3 |
| 1 | OBJETIVO | J4 |
| 2 | OBJETIVO | CM |
| 2 | OBJETIVO | J1 |
| 2 | SECTOR | J2 |
| 2 | OBJETIVO | J3 |
| 2 | VALOR | J4 |
| 3 | SECTOR | CM |
| 3 | SECTOR | J1 |
| 3 | OBJETIVO | J2 |
| 3 | SECTOR | J3 |
| 3 | SECTOR | J4 |
| 4 | TIPO.ENTIDAD | CM |
| 4 | TIPO.ENTIDAD | J1 |
| 4 | TIPO.ENTIDAD | J2 |
| 4 | TIPO.ENTIDAD | J3 |
| 4 | TIPO.ENTIDAD | J4 |
| 5 | ENTIDAD | CM |
| 5 | ENTIDAD | J1 |
| 5 | ENTIDAD | J2 |
| 5 | ENTIDAD | J3 |
| 5 | ENTIDAD | J4 |
| 6 | CUENTA.PROYECTO | CM |
| 6 | CUENTA.PROYECTO | J1 |
| 6 | CUENTA.PROYECTO | J2 |
| 6 | CUENTA.PROYECTO | J3 |
| 6 | CUENTA.PROYECTO | J4 |
| 7 | COD.ENTIDAD | CM |
| 7 | COD.ENTIDAD | J1 |
| 7 | COD.ENTIDAD | J2 |
| 7 | COD.ENTIDAD | J3 |
| 7 | COD.ENTIDAD | J4 |
| 8 | ANIO | CM |
| 8 | ANIO | J1 |
| 8 | ANIO | J2 |
| 8 | ANIO | J3 |
| 8 | ANIO | J4 |

**Table F. 35.** Ranking of variables established by J1, J2, J3, and J4 raters for the DNP-AIB dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,79 | 0,04 | 18,59 | 0,00 | 0,70 | 0,87 |
| ANIO | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| COD.ENTIDAD | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| CUENTA.PROYECTO | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| ENTIDAD | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| OBJETIVO | 0,20 | 0,11 | 1,79 | 0,07 | -0,02 | 0,42 |
| SECTOR | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |
| TIPO.ENTIDAD | 1,00 | 0,11 | 8,94 | 0,00 | 0,78 | 1,22 |
| VALOR | 0,54 | 0,11 | 4,86 | 0,00 | 0,32 | 0,76 |

**Table F. 36.** Fleiss' Kappa test metrics for the DNP-AIB dataset.

# F.13 DNP-FI Data Source

| Etiquetas de fila | ANIO | DEPARTAMENTO | HECTAREAS.REFORESTADAS | NUMERO.PROYECTOS | VALOR.ESTABLECIMIENTO | Total general |
|---|---|---|---|---|---|---|
| 1 | | | 2 | | 3 | 5 |
| 2 | | | 3 | | 2 | 5 |
| 3 | | | | 5 | | 5 |
| 4 | | 5 | | | | 5 |
| 5 | 5 | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 25 |

**Table F. 37.** Matches among raters in the positions of variables for the DNP-FI dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | VALOR.ESTABLECIMIENTO | CM |
| 1 | HECTAREAS.REFORESTADAS | J1 |
| 1 | VALOR.ESTABLECIMIENTO | J2 |
| 1 | HECTAREAS.REFORESTADAS | J3 |
| 1 | VALOR.ESTABLECIMIENTO | J4 |
| 2 | HECTAREAS.REFORESTADAS | CM |
| 2 | VALOR.ESTABLECIMIENTO | J1 |
| 2 | HECTAREAS.REFORESTADAS | J2 |
| 2 | VALOR.ESTABLECIMIENTO | J3 |
| 2 | HECTAREAS.REFORESTADAS | J4 |
| 3 | NUMERO.PROYECTOS | CM |
| 3 | NUMERO.PROYECTOS | J1 |
| 3 | NUMERO.PROYECTOS | J2 |
| 3 | NUMERO.PROYECTOS | J3 |
| 3 | NUMERO.PROYECTOS | J4 |
| 4 | DEPARTAMENTO | CM |
| 4 | DEPARTAMENTO | J1 |
| 4 | DEPARTAMENTO | J2 |
| 4 | DEPARTAMENTO | J3 |
| 4 | DEPARTAMENTO | J4 |
| 5 | ANIO | CM |
| 5 | ANIO | J1 |
| 5 | ANIO | J2 |

| | | |
|---|---|---|
| 5 | ANIO | J3 |
| 5 | ANIO | J4 |

**Table F. 38**. Ranking of variables established by J1, J2, J3, and J4 raters for the DNP-FI dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,70 | 0,07 | 9,90 | 0,00 | 0,56 | 0,84 |
| ANIO | 1,00 | 0,14 | 7,07 | 0,00 | 0,72 | 1,28 |
| DEPARTAMENTO | 1,00 | 0,14 | 7,07 | 0,00 | 0,72 | 1,28 |
| HECTAREAS.REFORESTADAS | 0,25 | 0,14 | 1,77 | 0,08 | -0,03 | 0,53 |
| NUMERO.PROYECTOS | 1,00 | 0,14 | 7,07 | 0,00 | 0,72 | 1,28 |
| VALOR.ESTABLECIMIENTO | 0,25 | 0,14 | 1,77 | 0,08 | -0,03 | 0,53 |

**Table F. 39.** Fleiss' Kappa test metrics for the DNP-FI dataset.

# F.14 DNP-LA Data Source

| Etiquetas de fila | ANIO | DEPARTAMENTO | FAMILIAS | HECTAREAS | Total general |
|---|---|---|---|---|---|
| 1 | | | 4 | 1 | 5 |
| 2 | | | 1 | 4 | 5 |
| 3 | | 5 | | | 5 |
| 4 | 5 | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 20 |

**Table F. 40.** Matches among raters in the positions of variables for the DNP-LA dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | FAMILIAS | CM |
| 1 | FAMILIAS | J1 |
| 1 | HECTAREAS | J2 |
| 1 | FAMILIAS | J3 |
| 1 | FAMILIAS | J4 |
| 2 | HECTAREAS | CM |
| 2 | HECTAREAS | J1 |
| 2 | FAMILIAS | J2 |
| 2 | HECTAREAS | J3 |
| 2 | HECTAREAS | J4 |
| 3 | DEPARTAMENTO | CM |
| 3 | DEPARTAMENTO | J1 |
| 3 | DEPARTAMENTO | J2 |
| 3 | DEPARTAMENTO | J3 |
| 3 | DEPARTAMENTO | J4 |
| 4 | ANIO | CM |
| 4 | ANIO | J1 |
| 4 | ANIO | J2 |
| 4 | ANIO | J3 |
| 4 | ANIO | J4 |

**Table F. 41.** Ranking of variables established by J1, J2, J3, and J4 raters for the DNP-LA dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,73 | 0,09 | 8,03 | 0,00 | 0,55 | 0,91 |
| ANIO | 1,00 | 0,16 | 6,32 | 0,00 | 0,69 | 1,31 |
| DEPARTAMENTO | 1,00 | 0,16 | 6,32 | 0,00 | 0,69 | 1,31 |
| FAMILIAS | 0,47 | 0,16 | 2,95 | 0,00 | 0,16 | 0,78 |
| HECTAREAS | 0,47 | 0,16 | 2,95 | 0,00 | 0,16 | 0,78 |

**Table F. 42.** Fleiss' Kappa test metrics for the DNP-LA dataset.

# F.15 DNP-PA Data Source

| Etiqnetas de fila | ANIO | BENEFICIARIOS.ALIANZA | DEPARTAMENTO | HECTAREAS.ALIANZA | MUNICIPIO | NOMBRE.ALIANZA | PRODUCTO.ALIANZA | VALOR.ALIANZA | VALOR.IM.ALIANZA | Total general |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 2 | | 1 | | | | 5 |
| 2 | | 2 | | 1 | | | 2 | | | 5 |
| 3 | | 1 | | | | | | 4 | | 5 |
| 4 | | | | 1 | | | | | 4 | 5 |
| 5 | | | | 1 | | | 3 | 1 | | 5 |
| 6 | | | | | | 4 | | | 1 | 5 |
| 7 | | | | | 5 | | | | | 5 |
| 8 | | | 5 | | | | | | | 5 |
| 9 | 5 | | | | | | | | | 5 |
| Total general | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 45 |

**Table F. 43.** Matches among raters in the positions of variables for the DNP-PA dataset.

| Ranking | Classification | Judge |
|---|---|---|
| 1 | HECTAREAS.ALIANZA | CM |
| 1 | NOMBRE.ALIANZA | J1 |
| 1 | HECTAREAS.ALIANZA | J2 |
| 1 | BENEFICIARIOS.ALIANZA | J3 |
| 1 | BENEFICIARIOS.ALIANZA | J4 |
| 2 | BENEFICIARIOS.ALIANZA | CM |
| 2 | PRODUCTO.ALIANZA | J1 |
| 2 | BENEFICIARIOS.ALIANZA | J2 |
| 2 | HECTAREAS.ALIANZA | J3 |
| 2 | PRODUCTO.ALIANZA | J4 |
| 3 | VALOR.ALIANZA | CM |
| 3 | BENEFICIARIOS.ALIANZA | J1 |
| 3 | VALOR.ALIANZA | J2 |
| 3 | VALOR.ALIANZA | J3 |
| 3 | VALOR.ALIANZA | J4 |
| 4 | VALOR.IM.ALIANZA | CM |
| 4 | HECTAREAS.ALIANZA | J1 |
| 4 | VALOR.IM.ALIANZA | J2 |
| 4 | VALOR.IM.ALIANZA | J3 |
| 4 | VALOR.IM.ALIANZA | J4 |
| 5 | PRODUCTO.ALIANZA | CM |
| 5 | VALOR.ALIANZA | J1 |
| 5 | PRODUCTO.ALIANZA | J2 |
| 5 | PRODUCTO.ALIANZA | J3 |
| 5 | HECTAREAS.ALIANZA | J4 |
| 6 | NOMBRE.ALIANZA | CM |
| 6 | VALOR.IM.ALIANZA | J1 |
| 6 | NOMBRE.ALIANZA | J2 |
| 6 | NOMBRE.ALIANZA | J3 |
| 6 | NOMBRE.ALIANZA | J4 |
| 7 | MUNICIPIO | CM |
| 7 | MUNICIPIO | J1 |
| 7 | MUNICIPIO | J2 |
| 7 | MUNICIPIO | J3 |
| 7 | MUNICIPIO | J4 |
| 8 | DEPARTAMENTO | CM |
| 8 | DEPARTAMENTO | J1 |

| | | |
|---|---|---|
| 8 | DEPARTAMENTO | J2 |
| 8 | DEPARTAMENTO | J3 |
| 8 | DEPARTAMENTO | J4 |
| 9 | ANIO | CM |
| 9 | ANIO | J1 |
| 9 | ANIO | J2 |
| 9 | ANIO | J3 |
| 9 | ANIO | J4 |

**Table F. 44.** Ranking of variables established by J1, J2, J3, and J4 raters for the DNP-PA dataset

| | kappa | s.e. | z-stat | p-value | lower | upper |
|---|---|---|---|---|---|---|
| Total | 0,56 | 0,04 | 15,09 | 0,00 | 0,49 | 0,64 |
| ANIO | 1,00 | 0,11 | 9,49 | 0,00 | 0,79 | 1,21 |
| BENEFICIARIOS.ALIANZA | 0,10 | 0,11 | 0,95 | 0,34 | -0,11 | 0,31 |
| DEPARTAMENTO | 1,00 | 0,11 | 9,49 | 0,00 | 0,79 | 1,21 |
| HECTAREAS.ALIANZA | -0,01 | 0,11 | -0,12 | 0,91 | -0,22 | 0,19 |
| MUNICIPIO | 1,00 | 0,11 | 9,49 | 0,00 | 0,79 | 1,21 |
| NOMBRE.ALIANZA | 0,55 | 0,11 | 5,22 | 0,00 | 0,34 | 0,76 |
| PRODUCTO.ALIANZA | 0,33 | 0,11 | 3,08 | 0,00 | 0,12 | 0,53 |
| VALOR.ALIANZA | 0,55 | 0,11 | 5,22 | 0,00 | 0,34 | 0,76 |
| VALOR.IM.ALIANZA | 0,55 | 0,11 | 5,22 | 0,00 | 0,34 | 0,76 |

**Table F. 45.** Fleiss' Kappa test metrics for the DNP-PA dataset.

# Appendix G

# Exploratory Analysis of Multi-Label Datasets

After applying the data integration process and obtained the Combined Data Sources (CDS), we performed an exploratory analysis for five Multi-Label Datasets (MLD). Through this analysis, we verify the quality of the combined data sources. All metrics and plots are presented below.

## G.1. MLD: IDEAM-AGRONET (CDS1)

### *Initial Exploratory Metrics*

| | | | |
|---|---|---|---|
| Discarded attributes | 3 | Maximum Frequency | 3 |
| Number of attributes | 72 | Cardinality | 12.92754 |
| Number of instances | 69 | Density | 0.1958718 |
| Number of inputs | 6 | Mean IR | 10.10454 |
| Number of labels | 66 | SCUMBLE | 0.2751257 |
| Number of labelsets | 48 | SCUMBLE CV | 0.3984465 |
| Number of single labelsets | 31 | TCS | 9.852615 |

**Table G. 1.** Metrics of exploratory analysis for the IDEAM-AGRONET MLD.

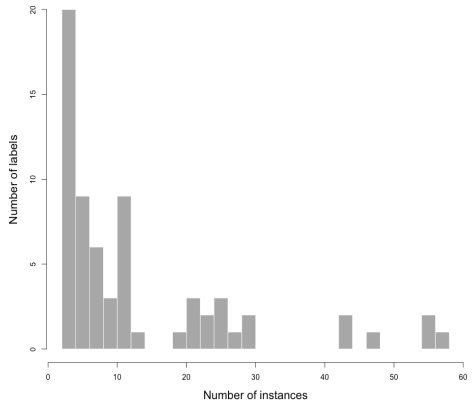



**Figure G. 1.** Labels histogram for the IDEAM-AGRONET MLD.

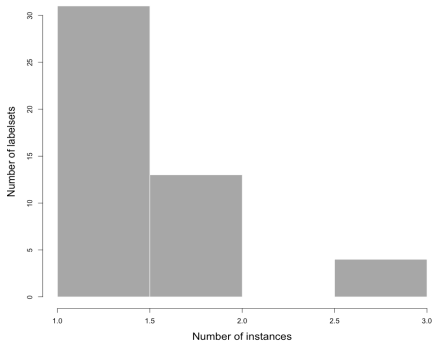



**Figure G. 2.** Labelsets histogram for the IDEAM-AGRONET MLD.

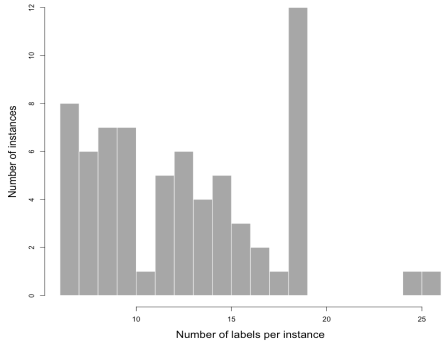



**Figure G. 3.** Cardinality histogram for the IDEAM-AGRONET MLD.

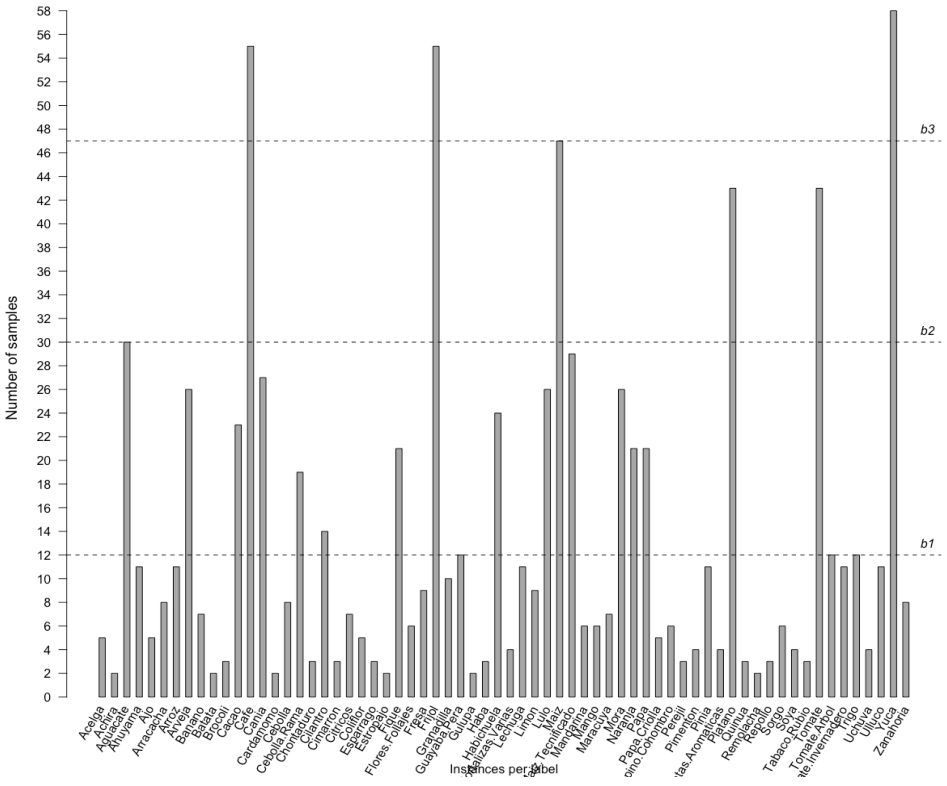



**Figure G. 4.** Labels bar diagram for the IDEAM-AGRONET MLD.

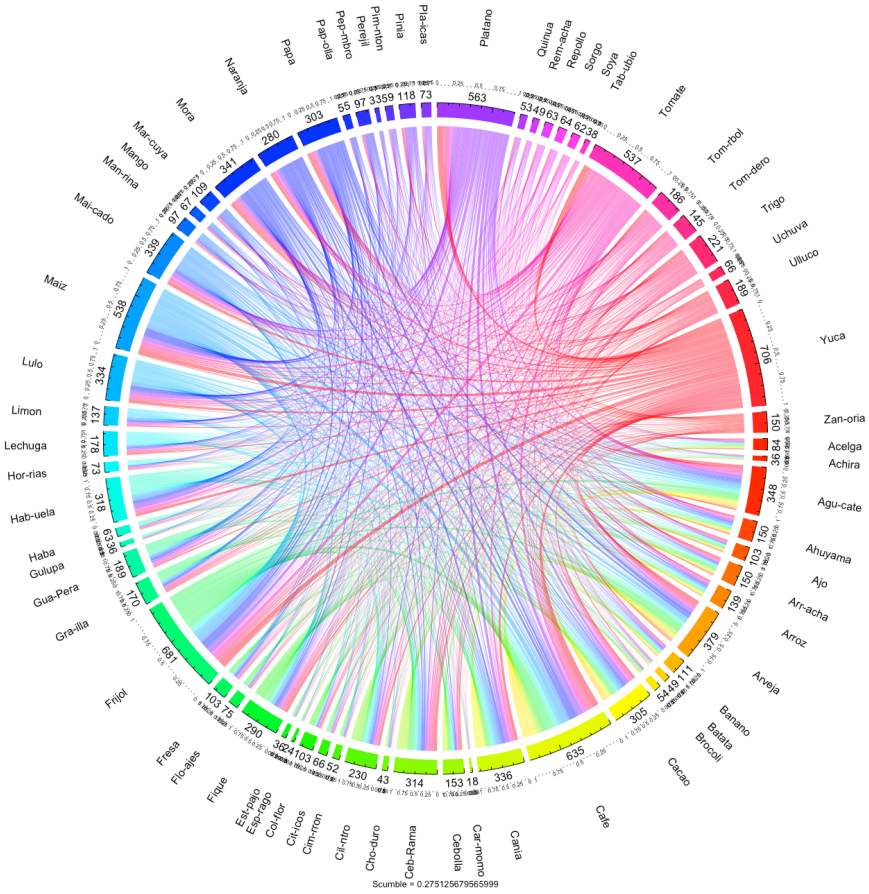



**Figure G. 5.** Concurrence among labels in IDEAM-AGRONET MLD.

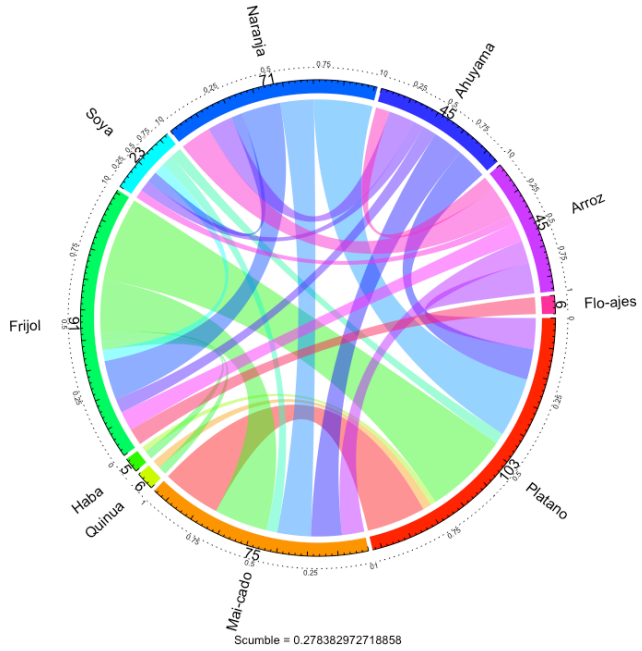



**Figure G. 6.** Concurrence among labels in IDEAM-AGRONET MLD (a 10-labels random subset).

## *Initial Exploratory Metrics by Label*

| | index | count | freq | IRLbl | SCUMBLE | SCUMBLE.CV |
|---|---|---|---|---|---|---|
| Acelga | 7 | 5 | 0.07246377 | 11.600000 | 0.3003219 | 0.193590963 |
| Achira | 8 | 2 | 0.02898551 | 29.000000 | 0.4128008 | 0.000000000 |
| Aguacate | 9 | 30 | 0.43478261 | 1.933333 | 0.2663858 | 0.462282826 |
| Ahuyama | 10 | 11 | 0.15942029 | 5.272727 | 0.3051228 | 0.220898034 |
| Ajo | 11 | 5 | 0.07246377 | 11.600000 | 0.3230106 | 0.123530454 |
| Arracacha | 12 | 8 | 0.11594203 | 7.250000 | 0.3357922 | 0.243761755 |
| Arroz | 13 | 11 | 0.15942029 | 5.272727 | 0.2913442 | 0.182287504 |
| Arveja | 14 | 26 | 0.37681159 | 2.230769 | 0.2750381 | 0.323191901 |
| Banano | 15 | 7 | 0.10144928 | 8.285714 | 0.2866018 | 0.305229161 |
| Batata | 16 | 2 | 0.02898551 | 29.000000 | 0.3636063 | 0.031396467 |
| Brocoli | 17 | 3 | 0.04347826 | 19.333333 | 0.2959469 | 0.065209995 |
| Cacao | 18 | 23 | 0.33333333 | 2.521739 | 0.2704844 | 0.334893824 |
| Cafe | 19 | 55 | 0.79710145 | 1.054545 | 0.2693031 | 0.443250883 |
| Cania | 20 | 27 | 0.39130435 | 2.148148 | 0.2547245 | 0.290446258 |
| Cardamomo | 21 | 2 | 0.02898551 | 29.000000 | 0.5332227 | 0.000000000 |
| Cebolla | 22 | 8 | 0.11594203 | 7.250000 | 0.3535450 | 0.148210869 |
| Cebolla.Rama | 23 | 19 | 0.27536232 | 3.052632 | 0.3138396 | 0.325217510 |
| Chontaduro | 24 | 3 | 0.04347826 | 19.333333 | 0.3777970 | 0.015296700 |
| Cilantro | 25 | 14 | 0.20289855 | 4.142857 | 0.2826734 | 0.248279955 |
| Cimarron | 26 | 3 | 0.04347826 | 19.333333 | 0.3640630 | 0.168109092 |
| Citricos | 27 | 7 | 0.10144928 | 8.285714 | 0.2670655 | 0.242975588 |
| Coliflor | 28 | 5 | 0.07246377 | 11.600000 | 0.3230106 | 0.123530454 |
| Esparrago | 29 | 3 | 0.04347826 | 19.333333 | 0.4479958 | 0.000000000 |
| Estropajo | 30 | 2 | 0.02898551 | 29.000000 | 0.3983248 | 0.053151777 |
| Fique | 31 | 21 | 0.30434783 | 2.761905 | 0.3401577 | 0.263080458 |
| Flores.Follajes | 32 | 6 | 0.08695652 | 9.666667 | 0.2995388 | 0.053135886 |
| Fresa | 33 | 9 | 0.13043478 | 6.444444 | 0.2853446 | 0.268032680 |
| Frijol | 34 | 55 | 0.79710145 | 1.054545 | 0.2830547 | 0.410763155 |
| Granadilla | 35 | 10 | 0.14492754 | 5.800000 | 0.3081129 | 0.218914265 |
| Guayaba.Pera | 36 | 12 | 0.17391304 | 4.833333 | 0.2841494 | 0.278606773 |
| Gulupa | 37 | 2 | 0.02898551 | 29.000000 | 0.4128008 | 0.000000000 |
| Haba | 38 | 3 | 0.04347826 | 19.333333 | 0.3716394 | 0.043283360 |
| Habichuela | 39 | 24 | 0.34782609 | 2.416667 | 0.2568264 | 0.286536349 |
| Hortalizas.Varias | 40 | 4 | 0.05797101 | 14.500000 | 0.3706683 | 0.028286978 |
| Lechuga | 41 | 11 | 0.15942029 | 5.272727 | 0.2853603 | 0.155241595 |
| Limon | 42 | 9 | 0.13043478 | 6.444444 | 0.2797756 | 0.258629002 |
| Lulo | 43 | 26 | 0.37681159 | 2.230769 | 0.2971979 | 0.442210879 |
| Maiz | 44 | 47 | 0.68115942 | 1.234043 | 0.2622750 | 0.474660968 |
| Maiz.Tecnificado | 45 | 29 | 0.42028986 | 2.000000 | 0.2689665 | 0.394493705 |
| Mandarina | 46 | 6 | 0.08695652 | 9.666667 | 0.2947445 | 0.294739081 |
| Mango | 47 | 6 | 0.08695652 | 9.666667 | 0.2497805 | 0.019447130 |
| Maracuya | 48 | 7 | 0.10144928 | 8.285714 | 0.2955449 | 0.268425706 |
| Mora | 49 | 26 | 0.37681159 | 2.230769 | 0.2572463 | 0.365679723 |
| Naranja | 50 | 21 | 0.30434783 | 2.761905 | 0.2873336 | 0.389620467 |
| Papa | 51 | 21 | 0.30434783 | 2.761905 | 0.2918299 | 0.312883144 |
| Papa.Criolla | 52 | 5 | 0.07246377 | 11.600000 | 0.3181811 | 0.210796228 |
| Pepino.Cohombro | 53 | 6 | 0.08695652 | 9.666667 | 0.2912512 | 0.325755055 |
| Perejil | 54 | 3 | 0.04347826 | 19.333333 | 0.3660326 | 0.055353024 |
| Pimenton | 55 | 4 | 0.05797101 | 14.500000 | 0.3470577 | 0.177663347 |
| Pinia | 56 | 11 | 0.15942029 | 5.272727 | 0.2234988 | 0.103696901 |
| Plantas.Aromaticas | 57 | 4 | 0.05797101 | 14.500000 | 0.3706683 | 0.028286978 |
| Platano | 58 | 43 | 0.62318841 | 1.348837 | 0.2758199 | 0.365734762 |
| Quinua | 59 | 3 | 0.04347826 | 19.333333 | 0.2970439 | 0.007946312 |
| Remolacha | 60 | 2 | 0.02898551 | 29.000000 | 0.3636063 | 0.031396467 |
| Repollo | 61 | 3 | 0.04347826 | 19.333333 | 0.3716394 | 0.043283360 |
| Sorgo | 62 | 6 | 0.08695652 | 9.666667 | 0.2722302 | 0.120455388 |
| Soya | 63 | 4 | 0.05797101 | 14.500000 | 0.3780487 | 0.069862961 |
| Tabaco.Rubio | 64 | 3 | 0.04347826 | 19.333333 | 0.3424031 | 0.077746981 |
| Tomate | 65 | 43 | 0.62318841 | 1.348837 | 0.3027153 | 0.328805929 |
| Tomate.Arbol | 66 | 12 | 0.17391304 | 4.833333 | 0.2495326 | 0.423389363 |
| Tomate.Invernadero | 67 | 11 | 0.15942029 | 5.272727 | 0.2959328 | 0.281402601 |
| Trigo | 68 | 12 | 0.17391304 | 4.833333 | 0.3246417 | 0.208155403 |
| Uchuva | 69 | 4 | 0.05797101 | 14.500000 | 0.2970470 | 0.053561209 |
| Ulluco | 70 | 11 | 0.15942029 | 5.272727 | 0.3276621 | 0.116366909 |
| Yuca | 71 | 58 | 0.84057971 | 1.000000 | 0.2778949 | 0.386266556 |
| Zanahoria | 72 | 8 | 0.11594203 | 7.250000 | 0.3385113 | 0.236484331 |

**Table G. 2.** Metrics of exploratory analysis for the IDEAM-AGRONET labels.

## *Concurrence Analysis*

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Cardamomo | Esparrago | Achira |
| Aguacate | 0.2663858 | - | 3 | - |
| Arracacha | 0.3357922 | - | - | 2 |
| Arveja | 0.2750381 | - | - | - |
| Cafe | 0.2693031 | 2 | 3 | 2 |
| Cebolla | 0.353545 | - | - | 2 |
| Cebolla.Rama | 0.3138396 | 2 | - | 2 |
| Fique | 0.3401577 | - | 3 | 2 |
| Frijol | 0.2830547 | 2 | 3 | 2 |
| Granadilla | 0.3081129 | - | - | 2 |
| Lulo | 0.2971979 | 2 | - | 2 |
| Maiz | 0.262275 | 2 | 3 | 2 |
| Naranja | 0.2873336 | 2 | - | - |
| Platano | 0.2758199 | 2 | - | - |
| Tomate | 0.3027153 | 2 | 3 | - |
| Yuca | 0.2778949 | 2 | 3 | - |

**Table G. 3.** Analysis of minority labels for IDEAM-AGRONET MLD.

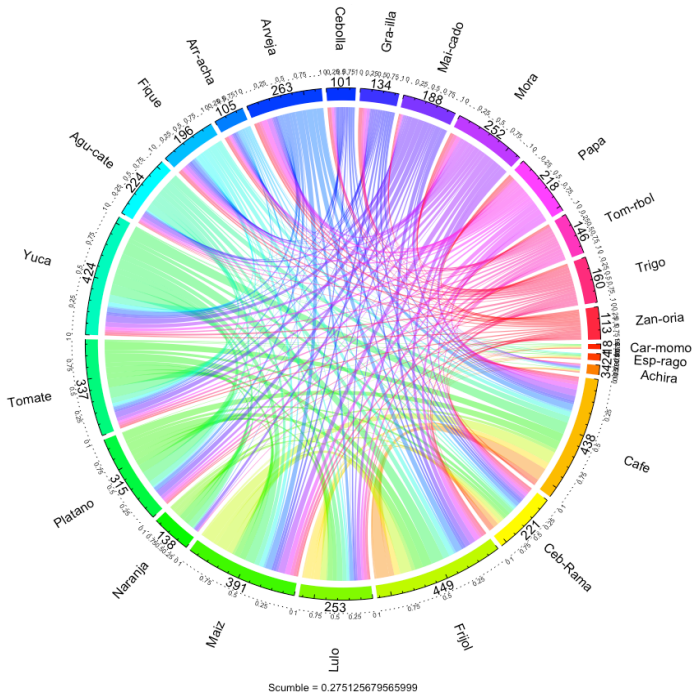



**Figure G. 7.** Concurrence among labels in IDEAM-AGRONET MLD (interactions of minority labels).

## *Metrics of exploratory analysis for the IDEAM-AGRONET*

| Metric | MLD | RD | RIPL | | | RSL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | B1 | B2 | B3 | S1 | S2 | S3 | S4 | S5 | S6 |
| Discarded attributes | 3 | 0 | 47 | 60 | 63 | 24 | 38 | 51 | 52 | 58 | 63 |
| Number of attributes | 72 | 72 | 25 | 12 | 9 | 48 | 34 | 21 | 20 | 14 | 9 |
| Number of instances | 69 | 103 | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 69 |
| Number of inputs | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| Number of labels | 66 | 66 | 19 | 6 | 3 | 42 | 28 | 15 | 14 | 8 | 3 |
| Number of labelsets | 48 | 63 | 39 | 16 | 6 | 45 | 41 | 33 | 32 | 27 | 8 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of single labelsets | 31 | 34 | 19 | 3 | 0 | 26 | 22 | 12 | 12 | 8 | 0 |
| Maximum Frequency | 3 | 5 | 3 | 12 | 40 | 3 | 3 | 3 | 6 | 8 | 16 |
| Cardinality | 12.92 | 8.66 | 8.81 | 4.36 | 2.43 | 11.78 | 10.28 | 6.17 | 5.89 | 3.62 | 1.24 |
| Density | 0.19 | 0.13 | 0.46 | 0.72 | 0.81 | 0.28 | 0.36 | 0.41 | 0.42 | 0.45 | 0.41 |
| Mean IR | 10.1 | 10.1 | 2.11 | 1.17 | 1.03 | 4.76 | 3.08 | 1.78 | 1.73 | 1.43 | 1.04 |
| SCUMBLE | 0.27 | 0.13 | 0.07 | 0.006 | 0.0002 | 0.18 | 0.12 | 0.03 | 0.03 | 0.01 | 0.0002 |
| SCUMBLE CV | 0.39 | 0.77 | 0.29 | 0.33 | 0.6 | 0.38 | 0.34 | 0.26 | 0.25 | 0.56 | 1.63 |
| TCS | 9.85 | 10.12 | 8.39 | 6.35 | 4.68 | 9.33 | 8.83 | 7.99 | 7.89 | 7.16 | 4.96 |

**Table G. 4.** Metrics of exploratory analysis for the IDEAM-AGRONET MLD and its variations.

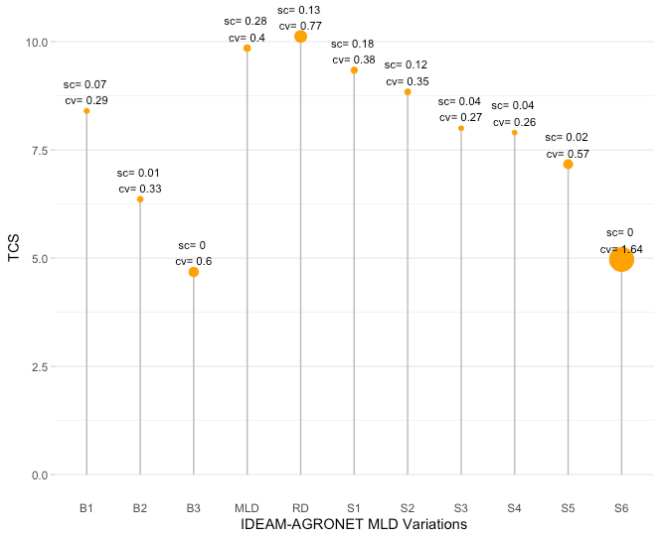



**Figure G. 8.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the IDEAM-AGRONET MLD and its variations.

# G.2. MLD: CORPOICA-AGRONET

### *Initial Exploratory Metrics*

| | | | |
|---|---|---|---|
| Discarded attributes | 3 | Maximum Frequency | 8 |
| Number of attributes | 87 | Cardinality | 10.2125 |
| Number of instances | 80 | Density | 0.1647177 |
| Number of inputs | 25 | Mean IR | 18.21208 |
| Number of labels | 62 | SCUMBLE | 0.2358326 |
| Number of labelsets | 44 | SCUMBLE CV | 0.6881884 |
| Number of single labelsets | 27 | TCS | 11.1302 |

**Table G. 5**. Metrics of exploratory analysis for the CORPOICA-AGRONET MLD.

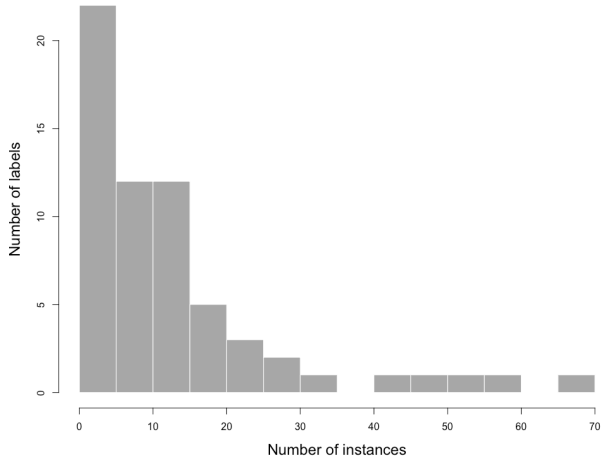



**Figure G. 9.** Labels histogram for the CORPOICA-AGRONET MLD.

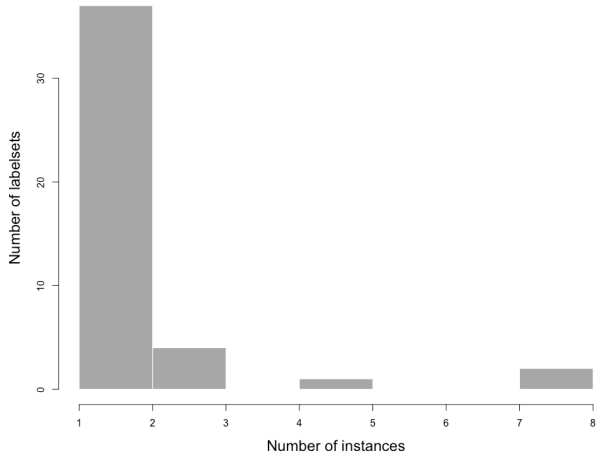



**Figure G. 10.** Labelsets histogram for the CORPOICA-AGRONET MLD.

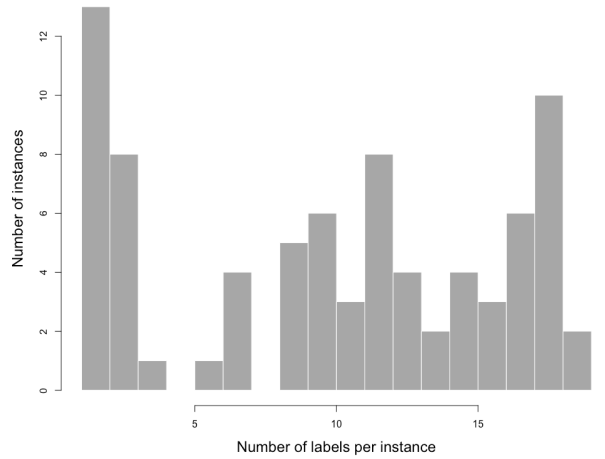



**Figure G. 11.** Cardinality histogram for the CORPOICA-AGRONET MLD.

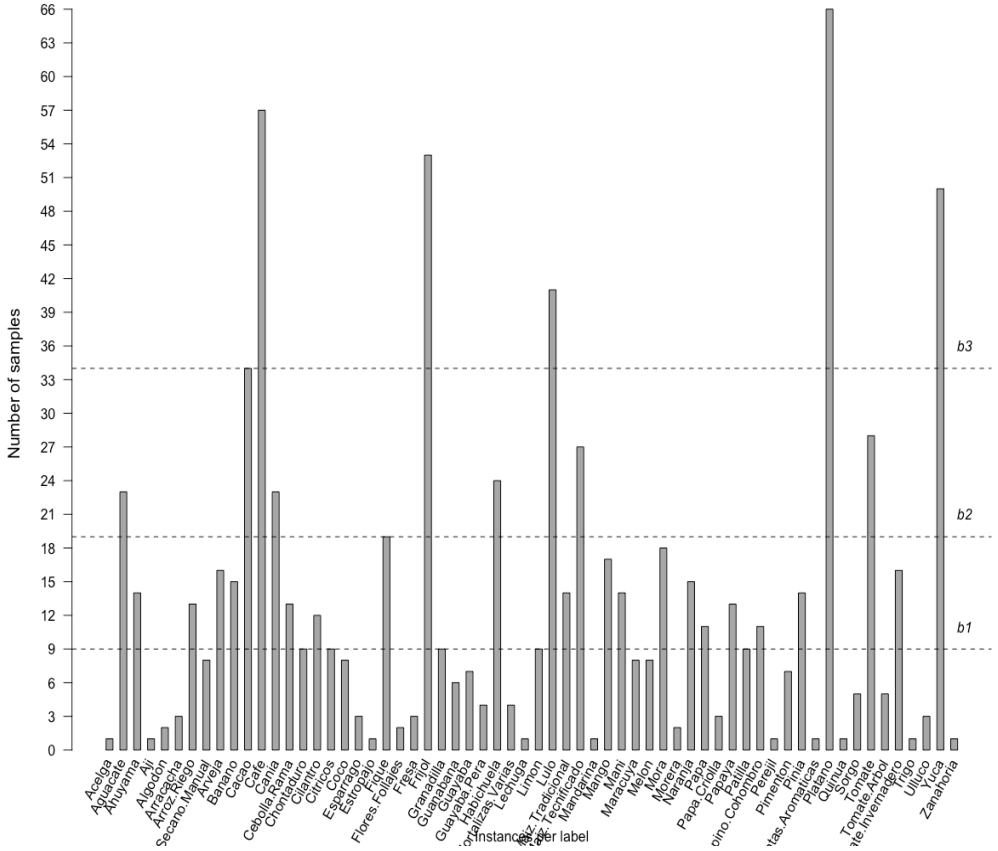



**Figure G. 12.** Labels bar diagram for the CORPOICA-AGRONET MLD.

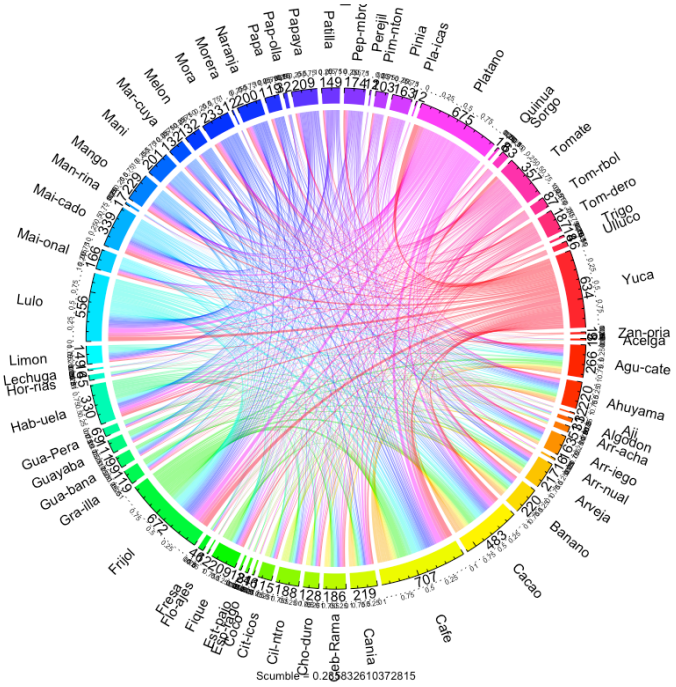



**Figure G. 13.** Concurrence among labels in CORPOICA-AGRONET MLD.

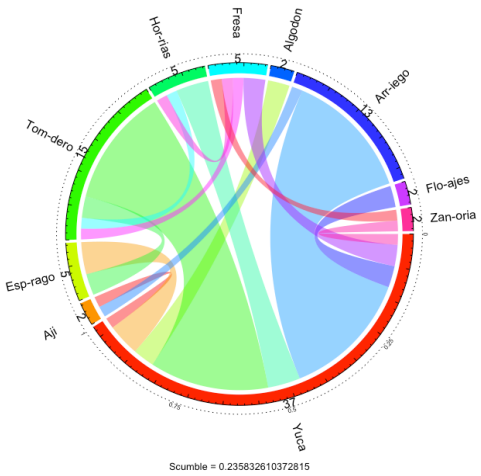



**Figure G. 14.** Concurrence among labels in CORPOICA-AGRONET MLD (a 10-labels random subset).

## *Initial exploratory metrics for labels*

| | index | count | freq | IRLbl | SCUMBLE | SCUMBLE.CV |
|---|---|---|---|---|---|---|
| Acelga | 26 | 1 | 0.0125 | 66.000000 | 0.6170428 | NA |
| Aguacate | 27 | 23 | 0.2875 | 2.869565 | 0.3303577 | 0.489548440 |
| Ahuyama | 28 | 14 | 0.1750 | 4.714286 | 0.2918635 | 0.369443513 |
| Aji | 29 | 1 | 0.0125 | 66.000000 | 0.5908586 | NA |
| Algodon | 30 | 2 | 0.0250 | 33.000000 | 0.3804345 | 0.007201167 |
| Arracacha | 31 | 3 | 0.0375 | 22.000000 | 0.3566862 | 0.345458073 |
| Arroz.Riego | 32 | 13 | 0.1625 | 5.076923 | 0.2073046 | 0.600415836 |
| Arroz.Secano.Manual | 33 | 8 | 0.1000 | 8.250000 | 0.3000734 | 0.000000000 |
| Arveja | 34 | 16 | 0.2000 | 4.125000 | 0.3192268 | 0.421015247 |
| Banano | 35 | 15 | 0.1875 | 4.400000 | 0.2260453 | 0.387678611 |
| Cacao | 36 | 34 | 0.4250 | 1.941176 | 0.2504673 | 0.337948835 |
| Cafe | 37 | 57 | 0.7125 | 1.157895 | 0.2783479 | 0.528601800 |
| Cania | 38 | 23 | 0.2875 | 2.869565 | 0.1854192 | 0.823474931 |
| Cebolla.Rama | 39 | 13 | 0.1625 | 5.076923 | 0.2923469 | 0.515919401 |
| Chontaduro | 40 | 9 | 0.1125 | 7.333333 | 0.2141812 | 0.200195008 |
| Cilantro | 41 | 12 | 0.1500 | 5.500000 | 0.2830327 | 0.502273829 |
| Citricos | 42 | 9 | 0.1125 | 7.333333 | 0.2430408 | 0.215119462 |
| Coco | 43 | 8 | 0.1000 | 8.250000 | 0.3000734 | 0.000000000 |
| Esparrago | 44 | 3 | 0.0375 | 22.000000 | 0.3971490 | 0.017215411 |
| Estropajo | 45 | 1 | 0.0125 | 66.000000 | 0.5105663 | NA |
| Fique | 46 | 19 | 0.2375 | 3.473684 | 0.2604835 | 0.530552964 |
| Flores.Follajes | 47 | 2 | 0.0250 | 33.000000 | 0.6346420 | 0.000000000 |
| Fresa | 48 | 3 | 0.0375 | 22.000000 | 0.5868776 | 0.089161776 |
| Frijol | 49 | 53 | 0.6625 | 1.245283 | 0.2891773 | 0.508338646 |
| Granadilla | 50 | 9 | 0.1125 | 7.333333 | 0.2624346 | 0.536317583 |
| Guanabana | 51 | 6 | 0.0750 | 11.000000 | 0.2386046 | 0.085205483 |
| Guayaba | 52 | 7 | 0.0875 | 9.428571 | 0.2693025 | 0.305717498 |
| Guayaba.Pera | 53 | 4 | 0.0500 | 16.500000 | 0.3330304 | 0.355394711 |
| Habichuela | 54 | 24 | 0.3000 | 2.750000 | 0.2749227 | 0.448557282 |
| Hortalizas.Varias | 55 | 4 | 0.0500 | 16.500000 | 0.3585603 | 0.315733960 |
| Lechuga | 56 | 1 | 0.0125 | 66.000000 | 0.6170428 | NA |
| Limon | 57 | 9 | 0.1125 | 7.333333 | 0.2990517 | 0.328508444 |
| Lulo | 58 | 41 | 0.5125 | 1.609756 | 0.2699130 | 0.373112856 |
| Maiz.Tradicional | 59 | 14 | 0.1750 | 4.714286 | 0.3044138 | 0.468596293 |
| Maiz.Tecnificado | 60 | 27 | 0.3375 | 2.444444 | 0.2357494 | 0.552274459 |
| Mandarina | 61 | 1 | 0.0125 | 66.000000 | 0.4989686 | NA |
| Mango | 62 | 17 | 0.2125 | 3.882353 | 0.2490002 | 0.555761902 |
| Mani | 63 | 14 | 0.1750 | 4.714286 | 0.2426548 | 0.339765187 |
| Maracuya | 64 | 8 | 0.1000 | 8.250000 | 0.2740621 | 0.247657610 |
| Melon | 65 | 8 | 0.1000 | 8.250000 | 0.2740621 | 0.247657610 |
| Mora | 66 | 18 | 0.2250 | 3.666667 | 0.3239653 | 0.474527992 |
| Morera | 67 | 2 | 0.0250 | 33.000000 | 0.6346420 | 0.000000000 |
| Naranja | 68 | 15 | 0.1875 | 4.400000 | 0.2546181 | 0.612272273 |
| Papa | 69 | 11 | 0.1375 | 6.000000 | 0.3536765 | 0.493074080 |
| Papa.Criolla | 70 | 3 | 0.0375 | 22.000000 | 0.4327300 | 0.187573649 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Papaya | 71 | 13 | 0.1625 | 5.076923 | 0.2821769 | 0.312618287 |
| Patilla | 72 | 9 | 0.1125 | 7.333333 | 0.2990517 | 0.328508444 |
| Pepino.Cohombro | 73 | 11 | 0.1375 | 6.000000 | 0.2763504 | 0.425486535 |
| Perejil | 74 | 1 | 0.0125 | 66.000000 | 0.5264556 | NA |
| Pimenton | 75 | 7 | 0.0875 | 9.428571 | 0.2499626 | 0.442138457 |
| Pinia | 76 | 14 | 0.1750 | 4.714286 | 0.1997940 | 0.620666312 |
| Plantas.Aromaticas | 77 | 1 | 0.0125 | 66.000000 | 0.5264556 | NA |
| Platano | 78 | 66 | 0.8250 | 1.000000 | 0.2353533 | 0.594688811 |
| Quinua | 79 | 1 | 0.0125 | 66.000000 | 0.6171345 | NA |
| Sorgo | 80 | 5 | 0.0625 | 13.200000 | 0.3278533 | 0.158328166 |
| Tomate | 81 | 28 | 0.3500 | 2.357143 | 0.3060005 | 0.600365817 |
| Tomate.Arbol | 82 | 5 | 0.0625 | 13.200000 | 0.3898513 | 0.418739937 |
| Tomate.Invernadero | 83 | 16 | 0.2000 | 4.125000 | 0.2395913 | 0.529891433 |
| Trigo | 84 | 1 | 0.0125 | 66.000000 | 0.6171345 | NA |
| Ulluco | 85 | 3 | 0.0375 | 22.000000 | 0.5868776 | 0.089161776 |
| Yuca | 86 | 50 | 0.6250 | 1.320000 | 0.2956784 | 0.479143332 |
| Zanahoria | 87 | 1 | 0.0125 | 66.000000 | 0.6171345 | NA |

**Table G. 6.** Metrics of exploratory analysis for the CORPOICA-AGRONET labels.

## *Concurrence Analysis*

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Flores.Follajes | Morera | Quinua |
| Aguacate | 0.3303577 | 2 | 2 | 1 |
| Arveja | 0.3192268 | - | - | 1 |
| Cafe | 0.2783479 | 2 | 2 | 1 |
| Cebolla.Rama | 0.2923469 | - | - | 1 |
| Frijol | 0.2891773 | 2 | 2 | 1 |
| Granadilla | 0.2624346 | - | - | 1 |
| Lulo | 0.269913 | - | - | 1 |
| Maiz | 0.3044138 | - | - | 1 |
| Mora | 0.3239653 | - | - | 1 |
| Papa | 0.3536765 | - | - | 1 |
| Tomate | 0.3060005 | 2 | 2 | - |
| Yuca | 0.2956784 | 2 | 2 | - |

**Table G. 7.** Analysis of minority labels for CORPOICA-AGRONET MLD.

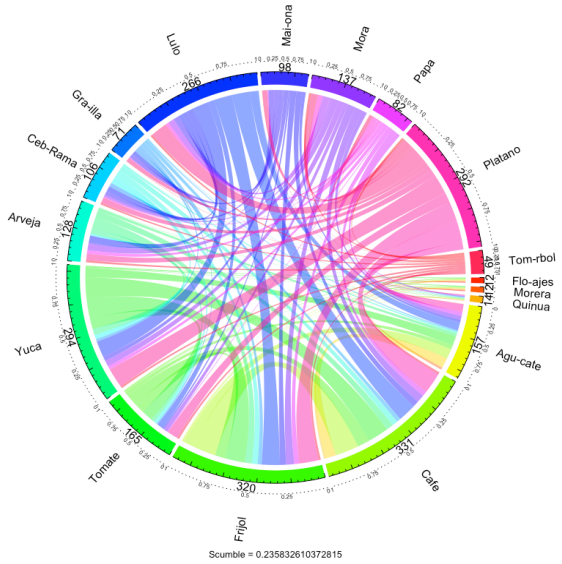



**Figure G. 15.** Concurrence among labels in CORPOICA-AGRONET MLD (interactions of minority labels).

### *Dataset Variations*

| Metric | MLD | RIPL | | | RSL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B1 | B2 | B3 | S1 | S2 | S3 | S4 | S5 | S6 |
| Discarded attributes | 3 | 34 | 51 | 57 | 23 | 34 | 47 | 52 | 56 | 59 |
| Number of attributes | 87 | 53 | 36 | 30 | 64 | 53 | 40 | 35 | 31 | 28 |
| Number of instances | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 |
| Number of inputs | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| Number of labels | 62 | 28 | 11 | 5 | 39 | 28 | 15 | 10 | 6 | 3 |
| Number of labelsets | 44 | 38 | 28 | 11 | 40 | 38 | 31 | 27 | 18 | 6 |
| Number of single labelsets | 27 | 21 | 12 | 1 | 22 | 21 | 16 | 10 | 6 | 0 |
| Maximum Frequency | 8 | 16 | 16 | 34 | 8 | 16 | 16 | 16 | 21 | 27 |
| Cardinality | 10.21 | 8.38 | 5.32 | 3.33 | 9.52 | 8.38 | 5.57 | 4.50 | 2.91 | 1.56 |
| Density | 0.16 | 0.29 | 0.48 | 0.66 | 0.24 | 0.29 | 0.37 | 0.45 | 0.48 | 0.52 |
| Mean IR | 18.21 | 3.61 | 1.96 | 1.26 | 4.86 | 3.61 | 2.29 | 1.77 | 1.46 | 1.23 |
| SCUMBLE | 0.23 | 0.10 | 0.04 | 0.007 | 0.15 | 0.10 | 0.06 | 0.03 | 0.01 | 0.005 |
| SCUMBLE CV | 0.68 | 0.63 | 0.69 | 0.77 | 0.55 | 0.63 | 0.68 | 0.70 | 0.85 | 1.05 |
| TCS | 11.13 | 10.18 | 8.94 | 7.22 | 10.57 | 10.18 | 9.36 | 8.81 | 7.90 | 6.10 |

**Table G. 8.** Metrics of exploratory analysis for the CORPOICA-AGRONET MLD and its variations.

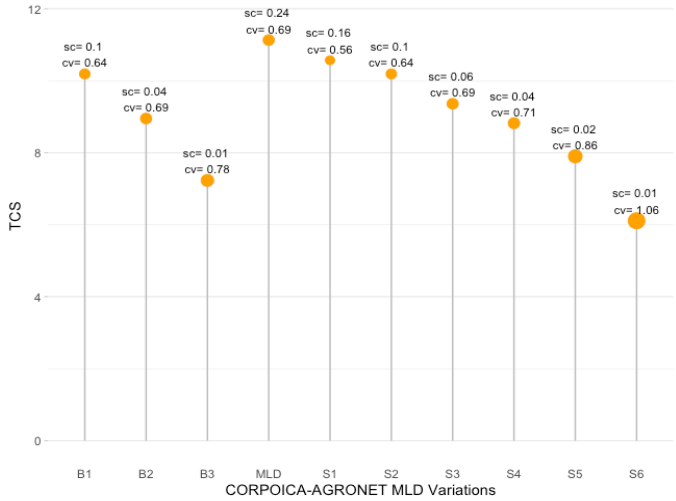



**Figure G. 16.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the CORPOICA-AGRONET MLD and its variations.

## G.3. MLD: SIVICAP-AGRONET

### *Initial Exploratory Metrics*

| | | | |
|---|---|---|---|
| Discarded attributes | 3 | Maximum Frequency | 1 |
| Number of attributes | 80 | Cardinality | 12.12821 |
| Number of instances | 39 | Density | 0.1837607 |
| Number of inputs | 14 | Mean IR | 11.41047 |
| Number of labels | 66 | SCUMBLE | 0.2898568 |
| Number of labelsets | 39 | SCUMBLE CV | 0.4957718 |
| Number of single labelsets | 39 | TCS | 10.49227 |

**Table G. 9.** Metrics of exploratory analysis for the SIVICAP-AGRONET MLD.

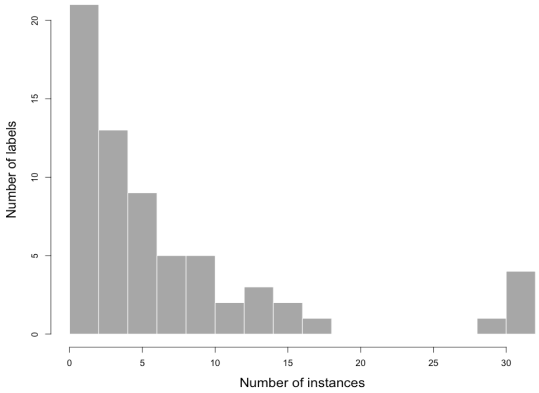



**Figure G. 17.** Labels histogram for the SIVICAP-AGRONET MLD.

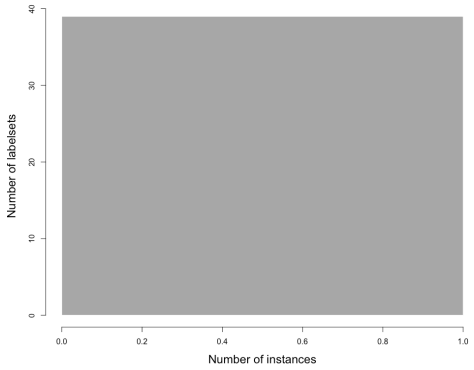



**Figure G. 18.** Labelsets histogram for the SIVICAP-AGRONET MLD.

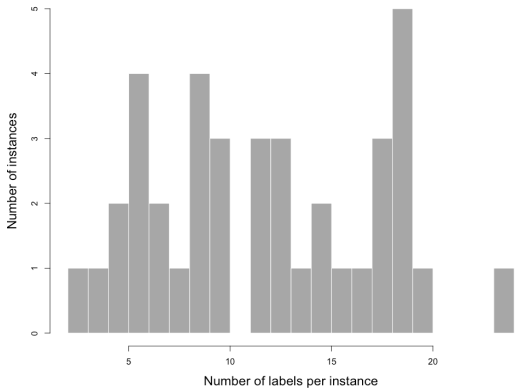



**Figure G. 19.** Cardinality histogram for the SIVICAP-AGRONET MLD.

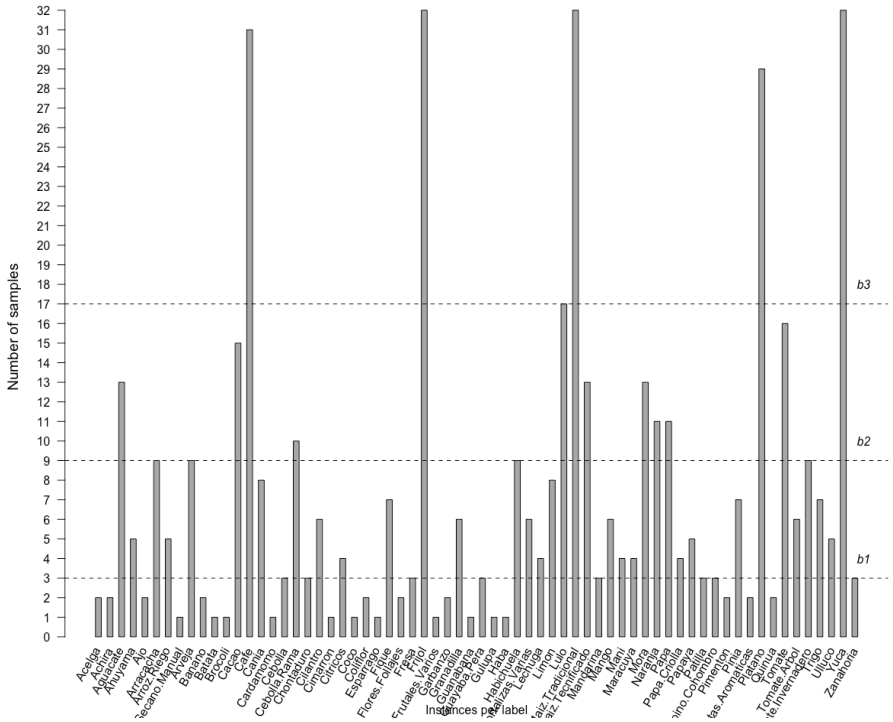



**Figure G. 20.** Labels bar diagram for the SIVICAP-AGRONET MLD.

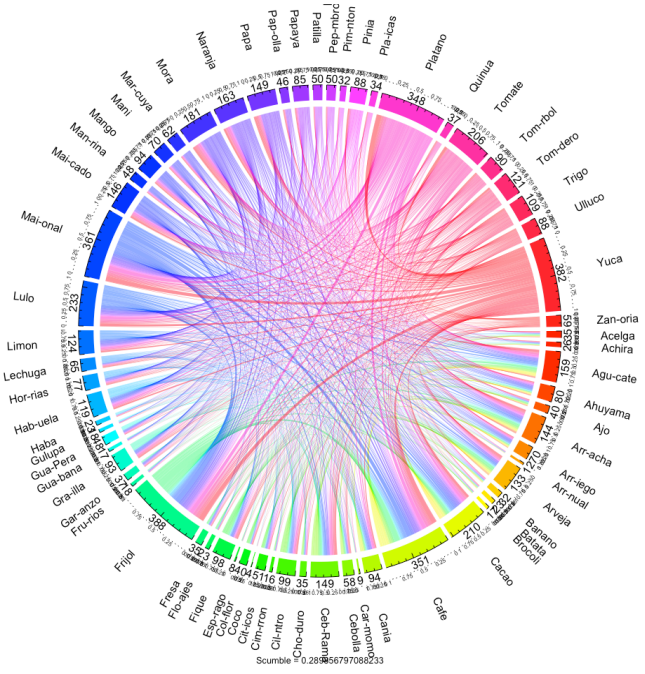



**Figure G. 21.** Concurrence among labels in SIVICAP-AGRONET MLD.

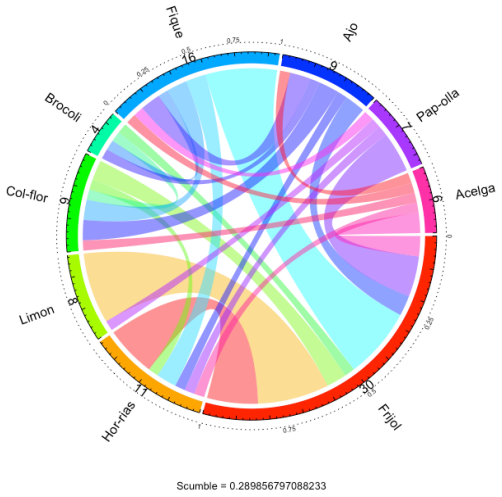



**Figure G. 22.** Concurrence among labels in SIVICAP-AGRONET MLD (a 10-labels random subset).

## *Initial Exploratory Metrics by Label*

| | index | count | freq | IRLbl | SCUMBLE | SCUMBLE.CV |
|---|---|---|---|---|---|---|
| Acelga | 15 | 2 | 0.05128205 | 16.000000 | 0.3558199 | 0.178189862 |
| Achira | 16 | 2 | 0.05128205 | 16.000000 | 0.3887534 | 0.066894726 |
| Aguacate | 17 | 13 | 0.33333333 | 2.461538 | 0.2930392 | 0.463118950 |
| Ahuyama | 18 | 5 | 0.12820513 | 6.400000 | 0.3120917 | 0.232885661 |
| Ajo | 19 | 2 | 0.05128205 | 16.000000 | 0.3837073 | 0.062455684 |
| Arracacha | 20 | 9 | 0.23076923 | 3.555556 | 0.3224962 | 0.329320419 |
| Arroz.Riego | 21 | 5 | 0.12820513 | 6.400000 | 0.2896054 | 0.308491797 |
| Arroz.Secano.Manual | 22 | 1 | 0.02564103 | 32.000000 | 0.4758830 | NA |
| Arveja | 23 | 9 | 0.23076923 | 3.555556 | 0.3096863 | 0.210437547 |
| Banano | 24 | 2 | 0.05128205 | 16.000000 | 0.3407836 | 0.148719529 |
| Batata | 25 | 1 | 0.02564103 | 32.000000 | 0.4006529 | NA |
| Brocoli | 26 | 1 | 0.02564103 | 32.000000 | 0.3667617 | NA |
| Cacao | 27 | 15 | 0.38461538 | 2.133333 | 0.2977868 | 0.299655581 |
| Cafe | 28 | 31 | 0.79487179 | 1.032258 | 0.2872400 | 0.466974309 |
| Cania | 29 | 8 | 0.20512821 | 4.000000 | 0.2368951 | 0.289339063 |
| Cardamomo | 30 | 1 | 0.02564103 | 32.000000 | 0.5661279 | NA |
| Cebolla | 31 | 3 | 0.07692308 | 10.666667 | 0.3792598 | 0.049080806 |
| Cebolla.Rama | 32 | 10 | 0.25641026 | 3.200000 | 0.3198099 | 0.374420667 |
| Chontaduro | 33 | 3 | 0.07692308 | 10.666667 | 0.4247220 | 0.465988281 |
| Cilantro | 34 | 6 | 0.15384615 | 5.333333 | 0.3107909 | 0.207099903 |
| Cimarron | 35 | 1 | 0.02564103 | 32.000000 | 0.3639330 | NA |
| Citricos | 36 | 4 | 0.10256410 | 8.000000 | 0.2734610 | 0.267328560 |
| Coco | 37 | 1 | 0.02564103 | 32.000000 | 0.6422545 | NA |
| Coliflor | 38 | 2 | 0.05128205 | 16.000000 | 0.3837073 | 0.062455684 |
| Esparrago | 39 | 1 | 0.02564103 | 32.000000 | 0.5576313 | NA |
| Fique | 40 | 7 | 0.17948718 | 4.571429 | 0.3676283 | 0.294433738 |
| Flores.Follajes | 41 | 2 | 0.05128205 | 16.000000 | 0.4155320 | 0.165983904 |
| Fresa | 42 | 3 | 0.07692308 | 10.666667 | 0.3057890 | 0.006538778 |
| Frijol | 43 | 32 | 0.82051282 | 1.000000 | 0.3035198 | 0.425245867 |
| Frutales.Varios | 44 | 1 | 0.02564103 | 32.000000 | 0.3813020 | NA |
| Garbanzo | 45 | 2 | 0.05128205 | 16.000000 | 0.3455654 | 0.146250650 |
| Granadilla | 46 | 6 | 0.15384615 | 5.333333 | 0.2956415 | 0.224215071 |
| Guanabana | 47 | 1 | 0.02564103 | 32.000000 | 0.3956142 | NA |
| Guayaba.Pera | 48 | 3 | 0.07692308 | 10.666667 | 0.2933555 | 0.262582572 |
| Gulupa | 49 | 1 | 0.02564103 | 32.000000 | 0.3703647 | NA |
| Haba | 50 | 1 | 0.02564103 | 32.000000 | 0.4006529 | NA |
| Habichuela | 51 | 9 | 0.23076923 | 3.555556 | 0.2867064 | 0.259175870 |
| Hortalizas.Varias | 52 | 6 | 0.15384615 | 5.333333 | 0.2671401 | 0.412762410 |
| Lechuga | 53 | 4 | 0.10256410 | 8.000000 | 0.3268341 | 0.229500076 |
| Limon | 54 | 8 | 0.20512821 | 4.000000 | 0.3218881 | 0.273140893 |
| Lulo | 55 | 17 | 0.43589744 | 1.882353 | 0.3229803 | 0.396337780 |
| Maiz.Tradicional | 56 | 32 | 0.82051282 | 1.000000 | 0.2984223 | 0.504304022 |
| Maiz.Tecnificado | 57 | 13 | 0.33333333 | 2.461538 | 0.2797159 | 0.400740712 |
| Mandarina | 58 | 3 | 0.07692308 | 10.666667 | 0.2961918 | 0.262744654 |
| Mango | 59 | 6 | 0.15384615 | 5.333333 | 0.3020686 | 0.275702973 |
| Mani | 60 | 4 | 0.10256410 | 8.000000 | 0.3364157 | 0.192852084 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Maracuya | 61 | 4 | 0.10256410 | 8.000000 | 0.3117506 | 0.269192547 |
| Mora | 62 | 13 | 0.33333333 | 2.461538 | 0.2951806 | 0.312139217 |
| Naranja | 63 | 11 | 0.28205128 | 2.909091 | 0.3092285 | 0.384365545 |
| Papa | 64 | 11 | 0.28205128 | 2.909091 | 0.2801624 | 0.311157167 |
| Papa.Criolla | 65 | 4 | 0.10256410 | 8.000000 | 0.2892915 | 0.204894441 |
| Papaya | 66 | 5 | 0.12820513 | 6.400000 | 0.3243863 | 0.192034121 |
| Patilla | 67 | 3 | 0.07692308 | 10.666667 | 0.3284462 | 0.185931924 |
| Pepino.Cohombro | 68 | 3 | 0.07692308 | 10.666667 | 0.3316743 | 0.118056934 |
| Pimenton | 69 | 2 | 0.05128205 | 16.000000 | 0.3450382 | 0.129447388 |
| Pinia | 70 | 7 | 0.17948718 | 4.571429 | 0.2754160 | 0.401015318 |
| Plantas.Aromaticas | 71 | 2 | 0.05128205 | 16.000000 | 0.3522790 | 0.194195715 |
| Platano | 72 | 29 | 0.74358974 | 1.103448 | 0.2895135 | 0.489352763 |
| Quinua | 73 | 2 | 0.05128205 | 16.000000 | 0.3077020 | 0.009775258 |
| Tomate | 74 | 16 | 0.41025641 | 2.000000 | 0.2857690 | 0.422259515 |
| Tomate.Arbol | 75 | 6 | 0.15384615 | 5.333333 | 0.2720323 | 0.307943937 |
| Tomate.Invernadero | 76 | 9 | 0.23076923 | 3.555556 | 0.3797233 | 0.271341472 |
| Trigo | 77 | 7 | 0.17948718 | 4.571429 | 0.3033592 | 0.289228961 |
| Ulluco | 78 | 5 | 0.12820513 | 6.400000 | 0.3373447 | 0.130547628 |
| Yuca | 79 | 32 | 0.82051282 | 1.000000 | 0.3101804 | 0.448882204 |
| Zanahoria | 80 | 3 | 0.07692308 | 10.666667 | 0.3588642 | 0.135346490 |

**Table G. 10.** Metrics of exploratory analysis for the SIVICAP-AGRONET labels.

## *Concurrence Analysis*

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Coco | Cardamomo | Esparrago |
| Aguacate | 0.2930392 | - | - | 1 |
| Cafe | 0.28724 | - | 1 | 1 |
| Chontaduro | 0.424722 | 1 | - | - |
| Cebolla.Rama | 0.3198099 | - | 1 | - |
| Fique | 0.3676283 | - | - | 1 |
| Frijol | 0.3035198 | - | 1 | 1 |
| Lulo | 0.3229803 | - | 1 | 1 |
| Maiz | 0.2984223 | 1 | 1 | 1 |
| Naranja | 0.3092285 | - | 1 | - |
| Platano | 0.2895135 | 1 | 1 | - |
| Tomate | 0.285769 | - | 1 | 1 |
| Yuca | 0.3101804 | 1 | 1 | 1 |

**Table G. 11.** Analysis of minority labels for SIVICAP-AGRONET MLD.

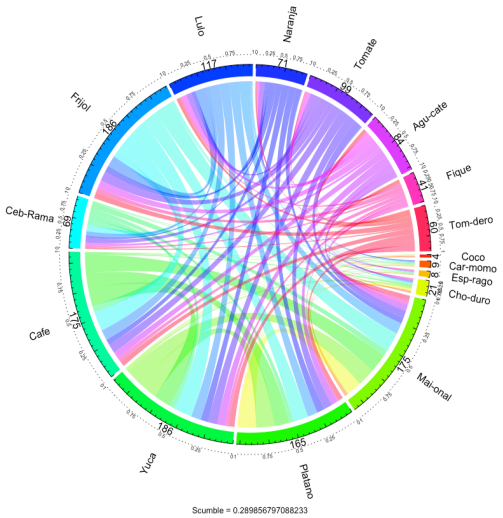



**Figure G. 23.** Concurrence among labels in SIVICAP-AGRONET MLD (interactions of minority labels).

### *Dataset Variations*

| Metric | MLD | RIPL | | | RSL | | |
|---|---|---|---|---|---|---|---|
| | | B1 | B2 | B3 | S1 | S2 | S3 |
| Discarded attributes | 3 | 29 | 52 | 61 | 11 | 21 | 29 |
| Number of attributes | 80 | 51 | 28 | 19 | 69 | 59 | 51 |
| Number of instances | 39 | 39 | 39 | 39 | 39 | 39 | 39 |
| Number of inputs | 14 | 14 | 14 | 14 | 14 | 14 | 14 |
| Number of labels | 66 | 37 | 14 | 5 | 55 | 45 | 37 |
| Number of labelsets | 39 | 39 | 38 | 14 | 39 | 39 | 39 |
| Number of single labelsets | 39 | 39 | 37 | 8 | 39 | 39 | 39 |
| Maximum Frequency | 1 | 1 | 2 | 18 | 1 | 1 | 1 |
| Cardinality | 12.12 | 10.71 | 7.05 | 4 | 11.84 | 11.33 | 10.71 |
| Density | 0.18 | 0.28 | 0.50 | 0.80 | 0.21 | 0.25 | 0.28 |
| Mean IR | 11.41 | 4.20 | 1.96 | 1.02 | 7.29 | 5.35 | 4.20 |
| SCUMBLE | 0.28 | 0.16 | 0.07 | 0.0006 | 0.23 | 0.20 | 0.16 |
| SCUMBLE CV | 0.49 | 0.42 | 0.41 | 0.54 | 0.44 | 0.42 | 0.42 |
| TCS | 10.49 | 9.91 | 8.91 | 6.88 | 10.30 | 10.10 | 9.91 |

**Table G. 12.** Metrics of exploratory analysis for the SIVICAP-AGRONET MLD and its variations.

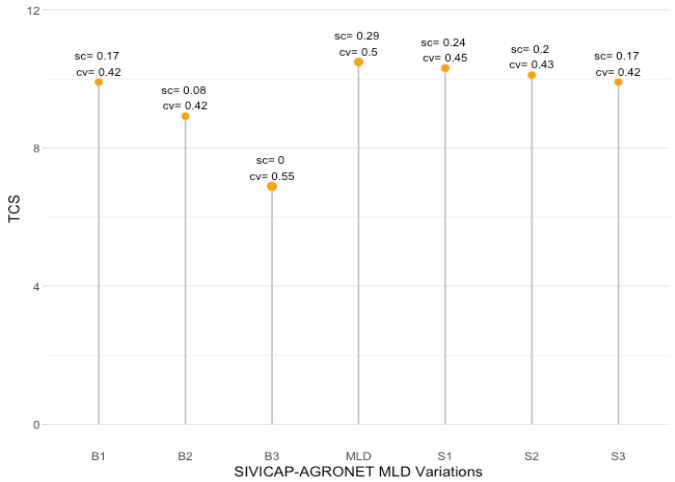



**Figure G. 24.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the SIVICAP-AGRONET MLD and its variations.

## G.4. MLD: DNP-PA-AGRONET

### *Initial Exploratory Metrics*

| | | | |
|---|---|---|---|
| Discarded attributes | 4 | Maximum Frequency | 5 |
| Number of attributes | 62 | Cardinality | 9.909091 |
| Number of instances | 55 | Density | 0.1801653 |
| Number of inputs | 7 | Mean IR | 8.789422 |
| Number of labels | 55 | SCUMBLE | 0.2117818 |
| Number of labelsets | 43 | SCUMBLE CV | 0.5952851 |
| Number of single labelsets | 36 | TCS | 9.714443 |

**Table G. 13.** Metrics of exploratory analysis for the DNP-PA-AGRONET MLD.

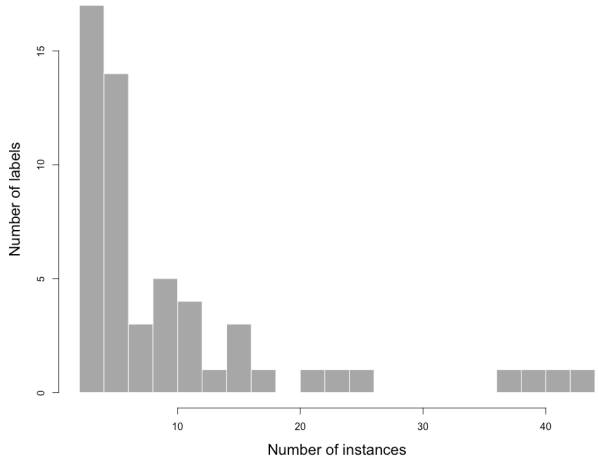



**Figure G. 25.** Labels histogram for the DNP-PA-AGRONET MLD.

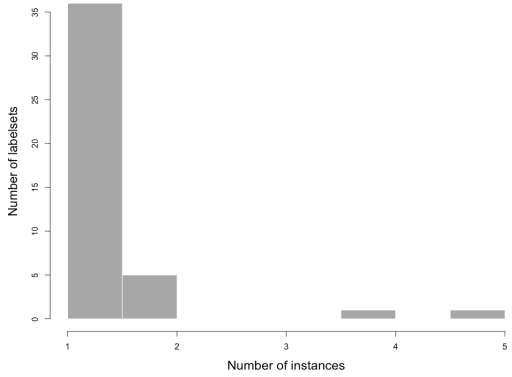



**Figure G. 26.** Labelsets histogram for the DNP-PA-AGRONET MLD.

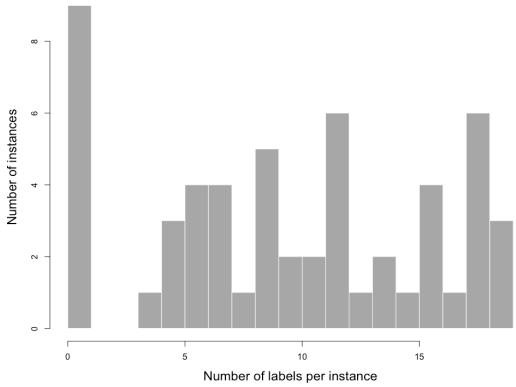



**Figure G. 27.** Cardinality histogram for the DNP-PA-AGRONET MLD.

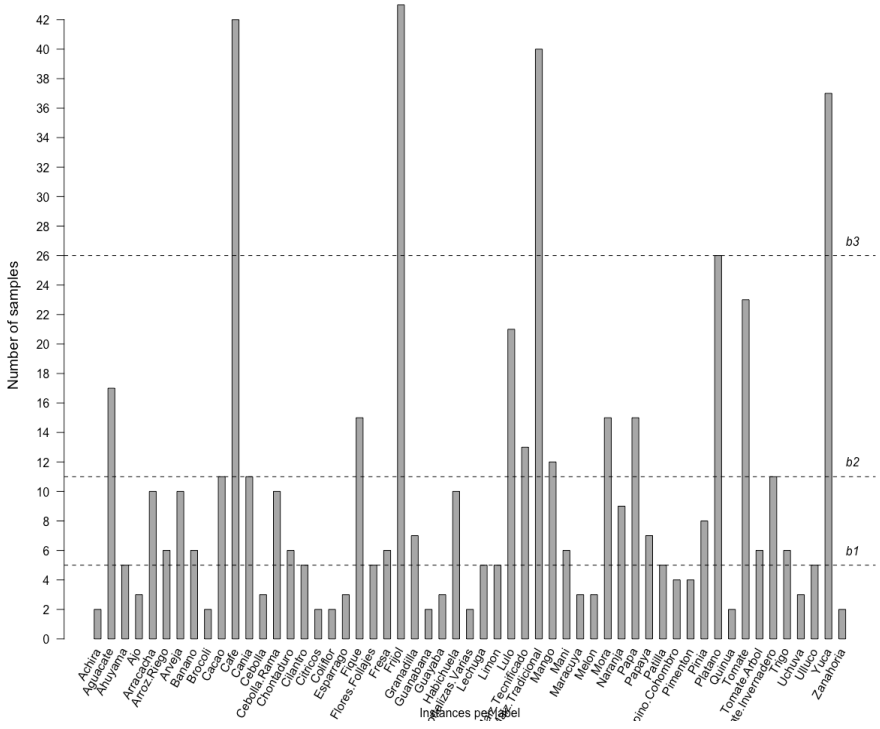



**Figure G. 28.** Labels bar diagram for the DNP-PA-AGRONET MLD.

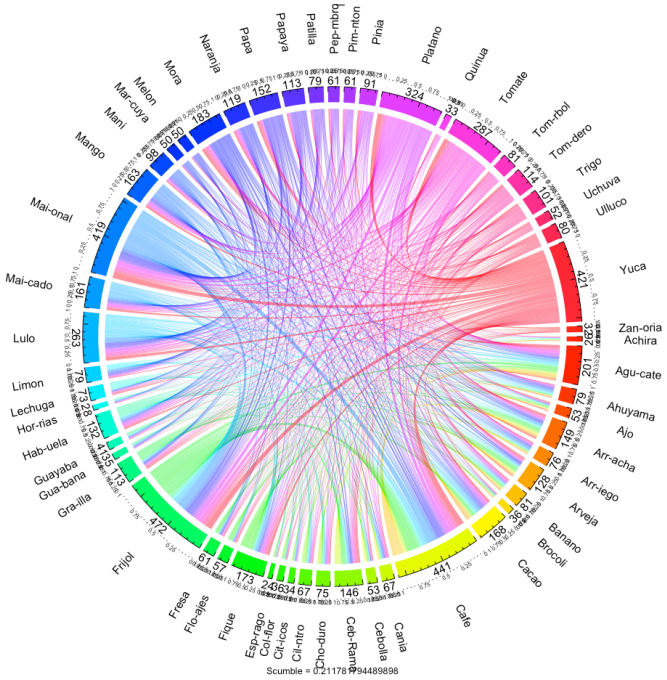


**Figure G. 29.** Concurrence among labels in DNP-PA-AGRONET MLD.

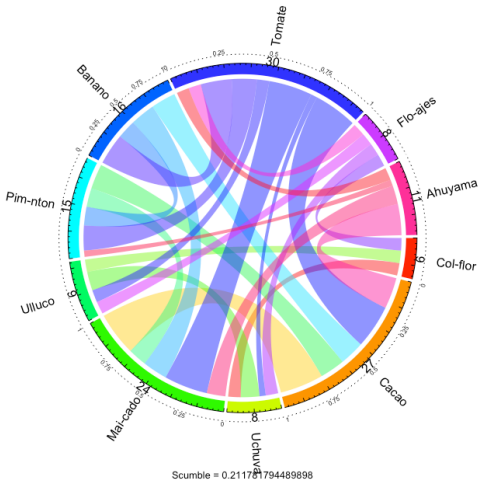



**Figure G. 30.** Concurrence among labels in DNP-PA-AGRONET MLD (a 10-labels random subset).

## *Initial Exploratory Metrics by Label*

| | index | count | freq | IRLbl | SCUMBLE | SCUMBLE.CV |
|---|---|---|---|---|---|---|
| Achira | 8 | 2 | 0.03636364 | 21.500000 | 0.3924445 | 0.17411004 |
| Aguacate | 9 | 17 | 0.30909091 | 2.529412 | 0.2741207 | 0.29186254 |
| Ahuyama | 10 | 5 | 0.09090909 | 8.600000 | 0.3105794 | 0.12346836 |
| Ajo | 11 | 3 | 0.05454545 | 14.333333 | 0.3229788 | 0.05671133 |
| Arracacha | 12 | 10 | 0.18181818 | 4.300000 | 0.2852929 | 0.22119165 |
| Arroz.Riego | 13 | 6 | 0.10909091 | 7.166667 | 0.2381650 | 0.31838787 |
| Arveja | 14 | 10 | 0.18181818 | 4.300000 | 0.2786351 | 0.26078900 |
| Banano | 15 | 6 | 0.10909091 | 7.166667 | 0.2500355 | 0.17512855 |
| Brocoli | 16 | 2 | 0.03636364 | 21.500000 | 0.3124038 | 0.00000000 |
| Cacao | 17 | 11 | 0.20000000 | 3.909091 | 0.2844283 | 0.20266427 |
| Cafe | 18 | 42 | 0.76363636 | 1.023810 | 0.2502068 | 0.37761870 |
| Cania | 19 | 11 | 0.20000000 | 3.909091 | 0.1236973 | 1.01413265 |
| Cebolla | 20 | 3 | 0.05454545 | 14.333333 | 0.3229788 | 0.05671133 |
| Cebolla.Rama | 21 | 10 | 0.18181818 | 4.300000 | 0.2909426 | 0.17749770 |
| Chontaduro | 22 | 6 | 0.10909091 | 7.166667 | 0.2493729 | 0.17267857 |
| Cilantro | 23 | 5 | 0.09090909 | 8.600000 | 0.3247199 | 0.23878981 |
| Citricos | 24 | 2 | 0.03636364 | 21.500000 | 0.3381011 | 0.13329877 |
| Coliflor | 25 | 2 | 0.03636364 | 21.500000 | 0.3124038 | 0.00000000 |
| Esparrago | 26 | 3 | 0.05454545 | 14.333333 | 0.3634434 | 0.05666115 |
| Fique | 27 | 15 | 0.27272727 | 2.866667 | 0.2506173 | 0.32002540 |
| Flores.Follajes | 28 | 5 | 0.09090909 | 8.600000 | 0.3128097 | 0.14635687 |
| Fresa | 29 | 6 | 0.10909091 | 7.166667 | 0.2682288 | 0.22011385 |
| Frijol | 30 | 43 | 0.78181818 | 1.000000 | 0.2558879 | 0.36290766 |
| Granadilla | 31 | 7 | 0.12727273 | 6.142857 | 0.2987071 | 0.14127543 |
| Guanabana | 32 | 2 | 0.03636364 | 21.500000 | 0.3344606 | 0.02814427 |
| Guayaba | 33 | 3 | 0.05454545 | 14.333333 | 0.3145315 | 0.07985844 |
| Habichuela | 34 | 10 | 0.18181818 | 4.300000 | 0.2794211 | 0.32010450 |
| Hortalizas.Varias | 35 | 2 | 0.03636364 | 21.500000 | 0.3746599 | 0.01770523 |
| Lechuga | 36 | 5 | 0.09090909 | 8.600000 | 0.2905504 | 0.12394118 |
| Limon | 37 | 5 | 0.09090909 | 8.600000 | 0.3105794 | 0.12346836 |
| Lulo | 38 | 21 | 0.38181818 | 2.047619 | 0.2853227 | 0.33534419 |
| Maiz.Tecnificado | 39 | 13 | 0.23636364 | 3.307692 | 0.2408282 | 0.28637571 |
| Maiz.Tradicional | 40 | 40 | 0.72727273 | 1.075000 | 0.2588377 | 0.37179437 |
| Mango | 41 | 12 | 0.21818182 | 3.583333 | 0.2736281 | 0.25191461 |
| Mani | 42 | 6 | 0.10909091 | 7.166667 | 0.3143510 | 0.12954731 |
| Maracuya | 43 | 3 | 0.05454545 | 14.333333 | 0.3373048 | 0.02454987 |
| Melon | 44 | 3 | 0.05454545 | 14.333333 | 0.3373048 | 0.02454987 |
| Mora | 45 | 15 | 0.27272727 | 2.866667 | 0.2573878 | 0.36400490 |
| Naranja | 46 | 9 | 0.16363636 | 4.777778 | 0.2572232 | 0.27613932 |
| Papa | 47 | 15 | 0.27272727 | 2.866667 | 0.2418599 | 0.37711981 |
| Papaya | 48 | 7 | 0.12727273 | 6.142857 | 0.3184427 | 0.12158962 |
| Patilla | 49 | 5 | 0.09090909 | 8.600000 | 0.3105794 | 0.12346836 |
| Pepino.Cohombro | 50 | 4 | 0.07272727 | 10.750000 | 0.2417888 | 0.05523220 |
| Pimenton | 51 | 4 | 0.07272727 | 10.750000 | 0.2417888 | 0.05523220 |
| Pinia | 52 | 8 | 0.14545455 | 5.375000 | 0.2217733 | 0.33893948 |
| Platano | 53 | 26 | 0.47272727 | 1.653846 | 0.2707802 | 0.34198699 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Quinua | 54 | 2 | 0.03636364 | 21.500000 | 0.3389368 | 0.09592197 |
| Tomate | 55 | 23 | 0.41818182 | 1.869565 | 0.2640456 | 0.24089594 |
| Tomate.Arbol | 56 | 6 | 0.10909091 | 7.166667 | 0.2954726 | 0.19420103 |
| Tomate.Invernadero | 57 | 11 | 0.20000000 | 3.909091 | 0.2641942 | 0.34041316 |
| Trigo | 58 | 6 | 0.10909091 | 7.166667 | 0.3227162 | 0.08031056 |
| Uchuva | 59 | 3 | 0.05454545 | 14.333333 | 0.3289111 | 0.08692802 |
| Ulluco | 60 | 5 | 0.09090909 | 8.600000 | 0.3026053 | 0.18400068 |
| Yuca | 61 | 37 | 0.67272727 | 1.162162 | 0.2606559 | 0.35608940 |
| Zanahoria | 62 | 2 | 0.03636364 | 21.500000 | 0.3168080 | 0.12195936 |

**Table G. 14.** Metrics of exploratory analysis for the DNP-PA-AGRONET labels.

## *Concurrence Analysis*

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Achira | Hortalizas.Varias | Esparrago |
| Aguacate | 0.2741207 | - | 2 | 2 |
| Arracacha | 0.2852929 | 1 | 1 | - |
| Arveja | 0.2786351 | 1 | 1 | - |
| Cafe | 0.2502068 | 2 | 2 | 3 |
| Cilantro | 0.3247199 | 2 | - | - |
| Cebolla.Rama | 0.2909426 | 1 | - | - |
| Fique | 0.2506173 | - | - | 3 |
| Flores.Follajes | 0.3128097 | - | - | 1 |
| Frijol | 0.2558879 | 2 | 2 | 3 |
| Granadilla | 0.2987071 | 1 | - | - |
| Habichuela | 0.2794211 | 2 | - | - |
| Lulo | 0.2853227 | 2 | 2 | 2 |
| Maiz | 0.2588377 | 2 | 2 | 3 |
| Mora | 0.2573878 | - | 2 | - |
| Platano | 0.2707802 | - | 2 | 1 |
| Tomate.Invernadero | 0.2641942 | - | - | 2 |
| Yuca | 0.2606559 | - | 2 | 3 |

**Table G. 15.** Analysis of minority labels for DNP-PA-AGRONET MLD.

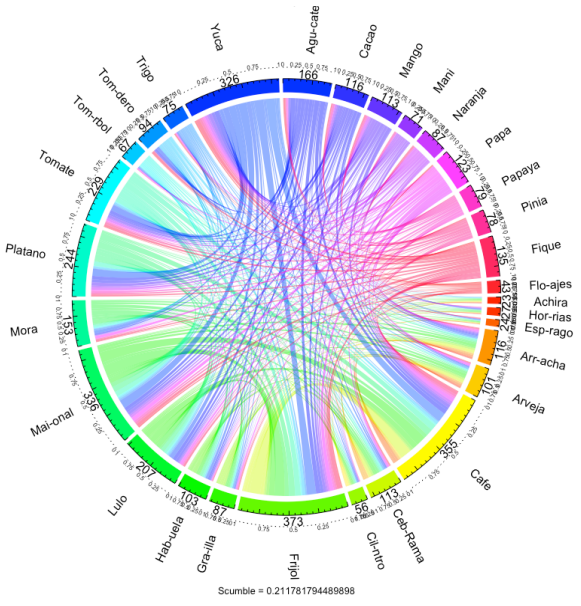



**Figure G. 31.** Concurrence among labels in DNP-PA-AGRONET MLD (interactions of minority labels).

### *Dataset Variations*

| Metric | MLD | RIPL | | | RSL | | |
|---|---|---|---|---|---|---|---|
| | | B1 | B2 | B3 | S1 | S2 | S3 |
| Discarded attributes | 4 | 25 | 42 | 51 | 8 | 15 | 31 |
| Number of attributes | 62 | 37 | 20 | 11 | 54 | 47 | 31 |
| Number of instances | 55 | 55 | 55 | 55 | 55 | 55 | 55 |
| Number of inputs | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| Number of labels | 55 | 30 | 13 | 4 | 47 | 40 | 24 |
| Number of labelsets | 43 | 42 | 33 | 9 | 42 | 42 | 40 |
| Number of single labelsets | 36 | 34 | 21 | 4 | 34 | 34 | 31 |
| Maximum Frequency | 5 | 5 | 9 | 27 | 5 | 5 | 5 |
| Cardinality | 9.90 | 8.34 | 5.80 | 2.94 | 9.61 | 9.23 | 7.69 |
| Density | 0.18 | 0.27 | 0.44 | 0.73 | 0.20 | 0.23 | 0.32 |
| Mean IR | 8.78 | 4.07 | 2.14 | 1.06 | 6.62 | 5.27 | 3.30 |
| SCUMBLE | 0.21 | 0.13 | 0.07 | 0.001 | 0.18 | 0.16 | 0.11 |
| SCUMBLE CV | 0.59 | 0.55 | 0.53 | 0.61 | 0.57 | 0.56 | 0.55 |
| TCS | 9.71 | 9.08 | 8 | 5.52 | 9.53 | 9.37 | 8.81 |

**Table G. 16.** Metrics of exploratory analysis for the DNP-PA-AGRONET MLD and its variations.

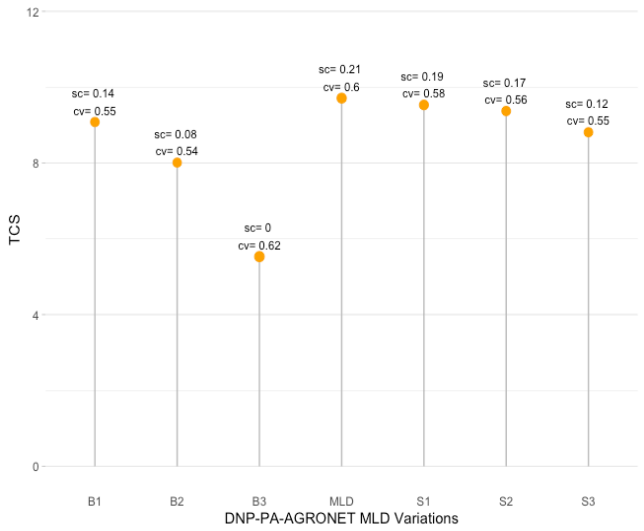



**Figure G. 32.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the DNP-PA-AGRONET MLD and its variations.

## G.5. CGD (COMBINED GLOBAL DATASET)

### *Initial Exploratory Metrics*

| | | | |
|---|---|---|---|
| Discarded attributes | 3 | Maximum Frequency | 2 |
| Number of attributes | 125 | Cardinality | 11.21918 |
| Number of instances | 73 | Density | 0.1438356 |
| Number of inputs | 47 | Mean IR | 19.56937 |
| Number of labels | 78 | SCUMBLE | 0.29878 |
| Number of labelsets | 68 | SCUMBLE CV | 0.5860287 |
| Number of single labelsets | 63 | TCS | 12.42636 |

**Table G. 17.** Metrics of exploratory analysis for the DNP-PA-AGRONET MLD.

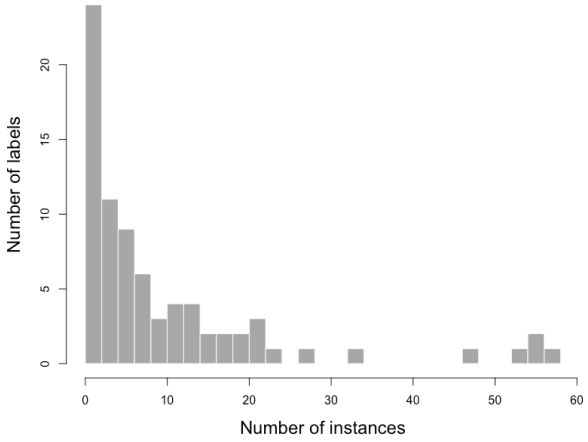



**Figure G. 33.** Labels histogram for the CGD MLD.

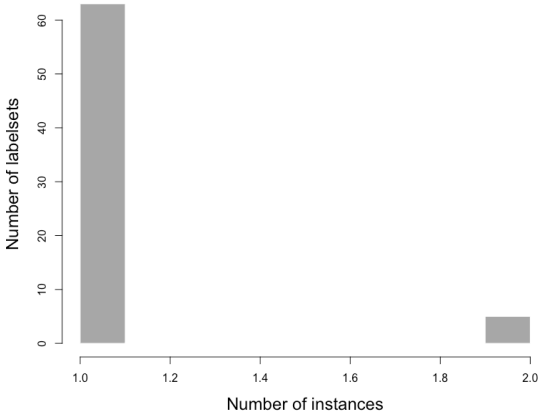



**Figure G. 34.** Labelsets histogram for the CGD MLD.

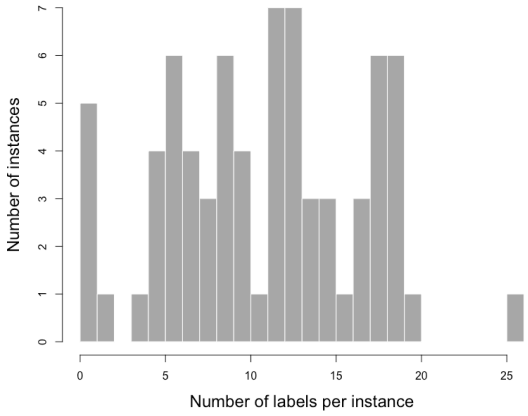



**Figure G. 35.** Cardinality histogram for the CGD MLD.

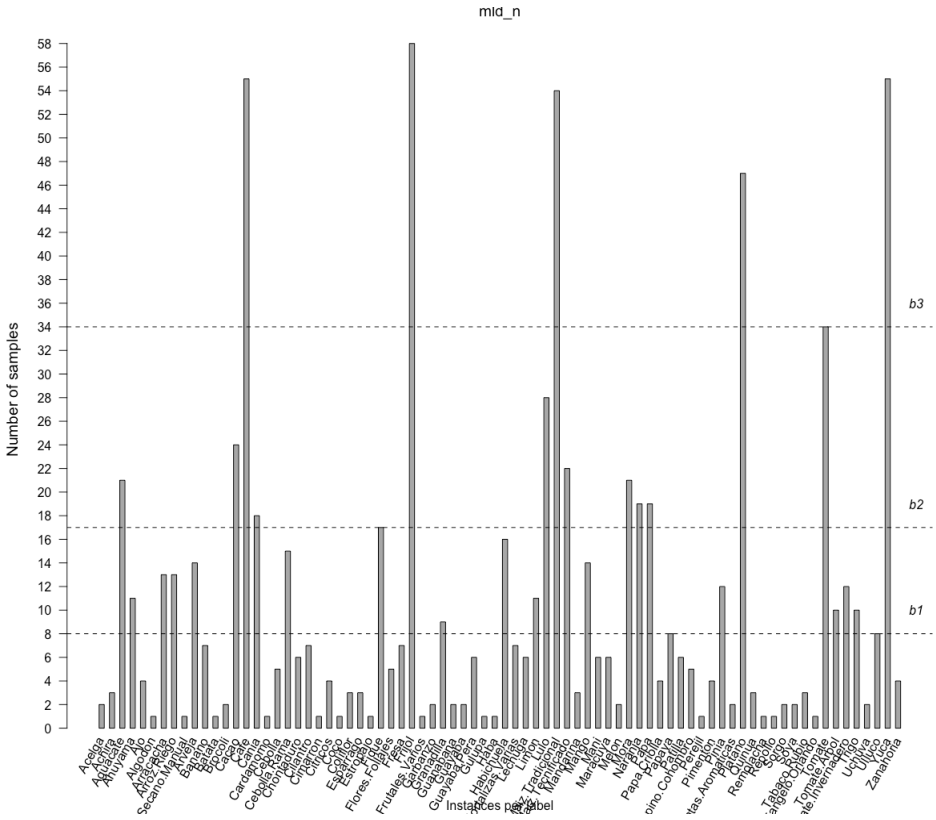



**Figure G. 36.** Labels bar diagram for the CGD MLD.

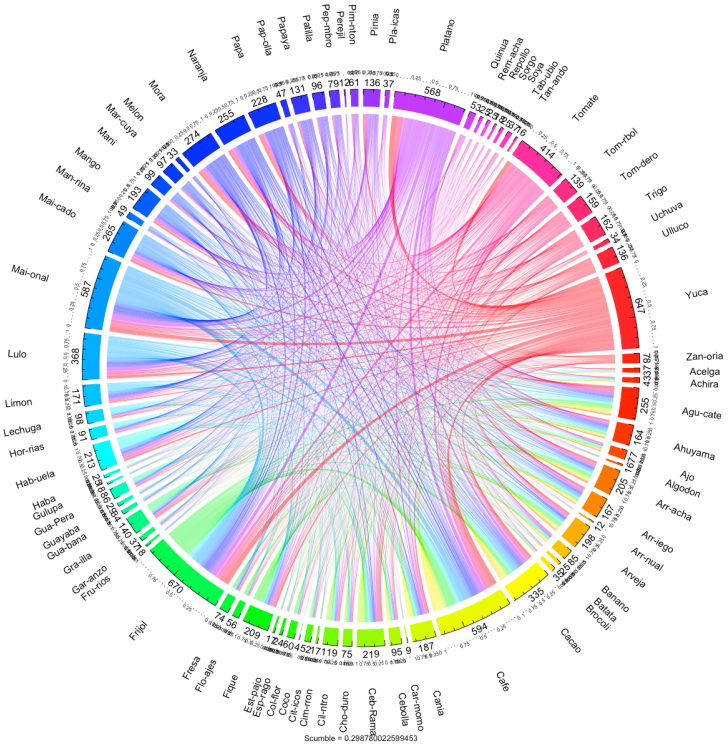



**Figure G. 37.** Concurrence among labels in CGD MLD.

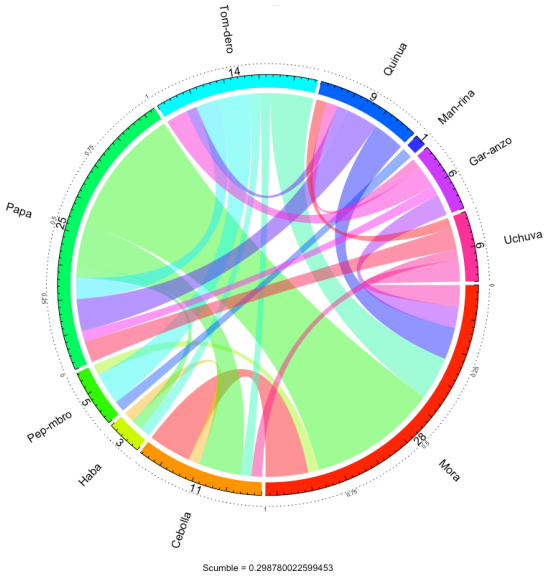



**Figure G. 38.** Concurrence among labels in CGD MLD (a 10-labels random subset).

## *Initial Exploratory Metrics by Label*

| | index | count | freq | IRLbl | SCUMBLE | SCUMBLE.CV |
|---|---|---|---|---|---|---|
| Acelga | 48 | 2 | 0.02739726 | 29.000000 | 0.4785363 | 0.1551120318 |
| Achira | 49 | 3 | 0.04109589 | 19.333333 | 0.4078811 | 0.2093569795 |
| Aguacate | 50 | 21 | 0.28767123 | 2.761905 | 0.3301768 | 0.3501881613 |
| Ahuyama | 51 | 11 | 0.15068493 | 5.272727 | 0.4192445 | 0.2338899879 |
| Ajo | 52 | 4 | 0.05479452 | 14.500000 | 0.3859178 | 0.2571869514 |
| Algodon | 53 | 1 | 0.01369863 | 58.000000 | 0.5636269 | NA |
| Arracacha | 54 | 13 | 0.17808219 | 4.461538 | 0.3696613 | 0.3474699776 |
| Arroz.Riego | 55 | 13 | 0.17808219 | 4.461538 | 0.3191088 | 0.4869005635 |
| Arroz.Secano.Manual | 56 | 1 | 0.01369863 | 58.000000 | 0.5958918 | NA |
| Arveja | 57 | 14 | 0.19178082 | 4.142857 | 0.3342869 | 0.3207010007 |
| Banano | 58 | 7 | 0.09589041 | 8.285714 | 0.2675029 | 0.4934509862 |
| Batata | 59 | 1 | 0.01369863 | 58.000000 | 0.5310226 | NA |
| Brocoli | 60 | 2 | 0.02739726 | 29.000000 | 0.3509985 | 0.0565778809 |
| Cacao | 61 | 24 | 0.32876712 | 2.416667 | 0.3775141 | 0.3126150557 |
| Cafe | 62 | 55 | 0.75342466 | 1.054545 | 0.3039287 | 0.5067257912 |
| Cania | 63 | 18 | 0.24657534 | 3.222222 | 0.2628897 | 0.6323777123 |
| Cardamomo | 64 | 1 | 0.01369863 | 58.000000 | 0.6942854 | NA |
| Cebolla | 65 | 5 | 0.06849315 | 11.600000 | 0.4028741 | 0.2331908875 |
| Cebolla.Rama | 66 | 15 | 0.20547945 | 3.866667 | 0.3559190 | 0.3815397815 |
| Chontaduro | 67 | 6 | 0.08219178 | 9.666667 | 0.3687155 | 0.4873982466 |
| Cilantro | 68 | 7 | 0.09589041 | 8.285714 | 0.3762790 | 0.3390494459 |
| Cimarron | 69 | 1 | 0.01369863 | 58.000000 | 0.5548572 | NA |
| Citricos | 70 | 4 | 0.05479452 | 14.500000 | 0.4039388 | 0.2869465908 |
| Coco | 71 | 1 | 0.01369863 | 58.000000 | 0.7330935 | NA |
| Coliflor | 72 | 3 | 0.04109589 | 19.333333 | 0.4110065 | 0.2551814925 |
| Esparrago | 73 | 3 | 0.04109589 | 19.333333 | 0.4522400 | 0.0875971386 |
| Estropajo | 74 | 1 | 0.01369863 | 58.000000 | 0.5548572 | NA |
| Fique | 75 | 17 | 0.23287671 | 3.411765 | 0.3292357 | 0.4121523680 |
| Flores.Follajes | 76 | 5 | 0.06849315 | 11.600000 | 0.3832976 | 0.1701623492 |
| Fresa | 77 | 7 | 0.09589041 | 8.285714 | 0.3247625 | 0.3593329829 |
| Frijol | 78 | 58 | 0.79452055 | 1.000000 | 0.3292985 | 0.4532600739 |
| Frutales.Varios | 79 | 1 | 0.01369863 | 58.000000 | 0.4992028 | NA |
| Garbanzo | 80 | 2 | 0.02739726 | 29.000000 | 0.4465128 | 0.1668817860 |
| Granadilla | 81 | 9 | 0.12328767 | 6.444444 | 0.3447454 | 0.2838305003 |
| Guanabana | 82 | 2 | 0.02739726 | 29.000000 | 0.4038457 | 0.1101613699 |
| Guayaba | 83 | 2 | 0.02739726 | 29.000000 | 0.4990481 | 0.1830048522 |
| Guayaba.Pera | 84 | 6 | 0.08219178 | 9.666667 | 0.3901867 | 0.2866488872 |
| Gulupa | 85 | 1 | 0.01369863 | 58.000000 | 0.4706995 | NA |
| Haba | 86 | 1 | 0.01369863 | 58.000000 | 0.5310226 | NA |
| Habichuela | 87 | 16 | 0.21917808 | 3.625000 | 0.3304003 | 0.2231127152 |
| Hortalizas.Varias | 88 | 7 | 0.09589041 | 8.285714 | 0.3532780 | 0.4487063814 |
| Lechuga | 89 | 6 | 0.08219178 | 9.666667 | 0.3906816 | 0.2456296008 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Limon | 90 | 11 | 0.15068493 | 5.272727 | 0.4174744 | 0.2990704407 |
| Lulo | 91 | 28 | 0.38356164 | 2.071429 | 0.3547400 | 0.3946842159 |
| Maiz.Tradicional | 92 | 54 | 0.73972603 | 1.074074 | 0.3196984 | 0.5195953915 |
| Maiz.Tecnificado | 93 | 22 | 0.30136986 | 2.636364 | 0.3315715 | 0.4411424029 |
| Mandarina | 94 | 3 | 0.04109589 | 19.333333 | 0.3866843 | 0.3879217967 |
| Mango | 95 | 14 | 0.19178082 | 4.142857 | 0.3103213 | 0.4674703030 |
| Mani | 96 | 6 | 0.08219178 | 9.666667 | 0.3745333 | 0.2115258664 |
| Maracuya | 97 | 6 | 0.08219178 | 9.666667 | 0.4521052 | 0.2590614761 |
| Melon | 98 | 2 | 0.02739726 | 29.000000 | 0.4994652 | 0.1816709400 |
| Mora | 99 | 21 | 0.28767123 | 2.761905 | 0.3259855 | 0.4279309495 |
| Naranja | 100 | 19 | 0.26027397 | 3.052632 | 0.3305107 | 0.4765143809 |
| Papa | 101 | 19 | 0.26027397 | 3.052632 | 0.3074457 | 0.3905165534 |
| Papa.Criolla | 102 | 4 | 0.05479452 | 14.500000 | 0.4309240 | 0.2431884508 |
| Papaya | 103 | 8 | 0.10958904 | 7.250000 | 0.4163241 | 0.2473966773 |
| Patilla | 104 | 6 | 0.08219178 | 9.666667 | 0.4217343 | 0.2462156993 |
| Pepino.Cohombro | 105 | 5 | 0.06849315 | 11.600000 | 0.3040702 | 0.0949709397 |
| Perejil | 106 | 1 | 0.01369863 | 58.000000 | 0.5311732 | NA |
| Pimenton | 107 | 4 | 0.05479452 | 14.500000 | 0.3151286 | 0.0546551866 |
| Pinia | 108 | 12 | 0.16438356 | 4.833333 | 0.2574153 | 0.5815312836 |
| Plantas.Aromaticas | 109 | 2 | 0.02739726 | 29.000000 | 0.5310979 | 0.0002005816 |
| Platano | 110 | 47 | 0.64383562 | 1.234043 | 0.3406917 | 0.5005890699 |
| Quinua | 111 | 3 | 0.04109589 | 19.333333 | 0.3834482 | 0.1349866620 |
| Remolacha | 112 | 1 | 0.01369863 | 58.000000 | 0.5310226 | NA |
| Repollo | 113 | 1 | 0.01369863 | 58.000000 | 0.5310226 | NA |
| Sorgo | 114 | 2 | 0.02739726 | 29.000000 | 0.4556736 | 0.1663505437 |
| Soya | 115 | 2 | 0.02739726 | 29.000000 | 0.4288314 | 0.0182351582 |
| Tabaco.Rubio | 116 | 3 | 0.04109589 | 19.333333 | 0.4199122 | 0.0390754299 |
| Tangelo.Orlando | 117 | 1 | 0.01369863 | 58.000000 | 0.5197660 | NA |
| Tomate | 118 | 34 | 0.46575342 | 1.705882 | 0.3273717 | 0.4225218982 |
| Tomate.Arbol | 119 | 10 | 0.13698630 | 5.800000 | 0.2954073 | 0.3728743233 |
| Tomate.Invernadero | 120 | 12 | 0.16438356 | 4.833333 | 0.3980069 | 0.2813665566 |
| Trigo | 121 | 10 | 0.13698630 | 5.800000 | 0.3624772 | 0.2686698380 |
| Uchuva | 122 | 2 | 0.02739726 | 29.000000 | 0.3971381 | 0.1142991558 |
| Ulluco | 123 | 8 | 0.10958904 | 7.250000 | 0.3894036 | 0.2822762010 |
| Yuca | 124 | 55 | 0.75342466 | 1.054545 | 0.3409699 | 0.4636640153 |
| Zanahoria | 125 | 4 | 0.05479452 | 14.500000 | 0.4099150 | 0.2636594487 |

**Table G. 18.** Metrics of exploratory analysis for the CGD labels.

## *Concurrence Analysis*

| Label | SCUMBLE | Minority Label Interactions | | |
|---|---|---|---|---|
| | | Coco | Cardamomo | Arroz.Secano.Manual |
| Arracacha | 0.3474699 | - | - | 1 |
| Cacao | 0.3126150 | - | - | 1 |
| Cafe | 0.5067257 | - | 1 | 1 |
| Chontaduro | 0.4873982 | 1 | - | - |
| Cebolla.Rama | 0.3815397 | - | 1 | - |
| Frijol | 0.4532600 | - | 1 | 1 |
| Limon | 0.2990704 | - | - | 1 |
| Lulo | 0.3946842 | - | 1 | 1 |
| Maiz | 0.5195953 | 1 | 1 | 1 |
| Mora | 0.4279309 | - | - | 1 |
| Naranja | 0.4765143 | - | 1 | - |
| Pinia | 0.5815312 | - | - | 1 |
| Platano | 0.5005890 | 1 | 1 | 1 |
| Tomate | 0.4225218 | - | 1 | - |
| Yuca | 0.4636640 | 1 | 1 | - |

**Table G. 19.** Analysis of minority labels for CGD MLD.

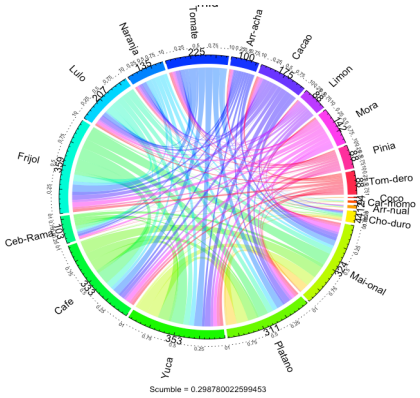



**Figure G. 39.** Concurrence among labels in CGD MLD (interactions of minority labels).

## *Dataset Variations*

| Metric | MLD | RIPL | | | RSL | | |
|---|---|---|---|---|---|---|---|
| | | B1 | B2 | B3 | S1 | S2 | S3 |
| Discarded attributes | 3 | 28 | 14 | 5 | 14 | 24 | 30 |
| Number of attributes | 125 | 97 | 111 | 120 | 111 | 101 | 95 |
| Number of instances | 73 | 73 | 73 | 73 | 73 | 73 | 73 |
| Number of inputs | 47 | 47 | 47 | 47 | 47 | 47 | 47 |
| Number of labels | 78 | 50 | 64 | 73 | 64 | 54 | 48 |
| Number of labelsets | 68 | 42 | 56 | 64 | 68 | 67 | 67 |
| Number of single labelsets | 63 | 34 | 49 | 59 | 63 | 61 | 61 |
| Maximum Frequency | 2 | 21 | 10 | 5 | 2 | 2 | 2 |
| Cardinality | 11.21 | 2.28 | 4.71 | 7.53 | 11.02 | 10.75 | 10.5 |
| Density | 0.14 | 0.04 | 0.07 | 0.1 | 0.17 | 0.19 | 0.21 |
| Mean IR | 19.56 | 3.94 | 6.85 | 12.21 | 11.16 | 7.85 | 6.42 |
| SCUMBLE | 0.29 | 0.05 | 0.13 | 0.19 | 0.26 | 0.23 | 0.21 |
| SCUMBLE CV | 0.58 | 1.66 | 1.03 | 0.81 | 0.53 | 0.5 | 0.49 |
| TCS | 12.42 | 11.49 | 12.03 | 12.29 | 12.22 | 12.04 | 11.92 |

**Table G. 20.** Metrics of exploratory analysis for the CGD MLD and its variations.

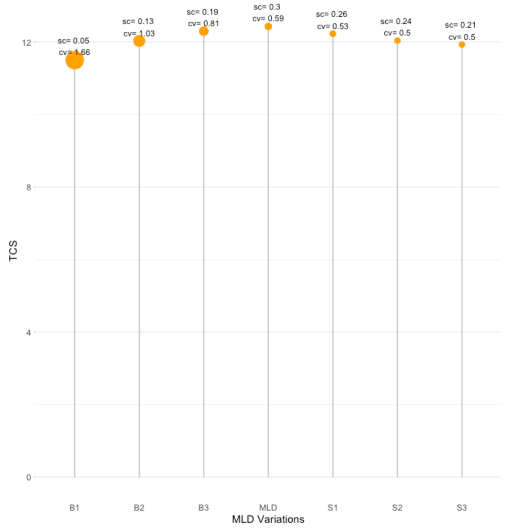



**Figure G. 40.** Comparison of SCUMBLE, SCUMBLE.CV, and TCS metrics for the CGD MLD and its variations.

# Appendix H

# Model Evaluation

In this appendix, we present the performance metrics of the models generated from the MLC strategies in each combined data source. These metrics correspond to Hamming-Loss, Precision, Recall, Accuracy, F1-Score, Ranking-Loss, and One-Error.

## H.1. Performance Measures for Models (CDS1)

| | mlc_methods | values_hl | values_pr | values_re | values_ac | values_f1 | values_rl | values_oe |
|---|---|---|---|---|---|---|---|---|
| 1 | BR-RF | 0.092 | 0.831 | 0.705 | 0.607 | 0.742 | 0.042 | 0.086 |
| 2 | BR-SVM | 0.116 | 0.803 | 0.592 | 0.503 | 0.658 | 0.096 | 0.043 |
| 3 | BR-SMO | 0.132 | 0.751 | 0.55 | 0.451 | 0.612 | 0.242 | 0.288 |
| 4 | BR-C5.0 | 0.099 | 0.786 | 0.731 | 0.594 | 0.736 | 0.074 | 0.1 |
| 5 | BR-NB | 0.141 | 0.645 | 0.735 | 0.511 | 0.661 | 0.083 | 0.09 |
| 6 | BR-XGB | 0.124 | 0.709 | 0.69 | 0.521 | 0.673 | 0.073 | 0.088 |
| 7 | BR-CART | 0.132 | 0.695 | 0.641 | 0.48 | 0.641 | 0.126 | 0.045 |
| 8 | BR-MAJORITY | 0.154 | 0.727 | 0.379 | 0.333 | 0.483 | 0.301 | 0.202 |
| 9 | BR-RANDOM | 0.49 | 0.195 | 0.493 | 0.16 | 0.272 | 0.492 | 0.886 |
| 10 | ECC-RF | 0.129 | 0.685 | 0.704 | 0.519 | 0.668 | 0.102 | 0.045 |
| 11 | ECC-SVM | 0.136 | 0.663 | 0.7 | 0.499 | 0.657 | 0.122 | 0.09 |
| 12 | ECC-SMO | 0.162 | 0.596 | 0.655 | 0.435 | 0.596 | 0.147 | 0.102 |
| 13 | ECC-C5.0 | 0.129 | 0.695 | 0.68 | 0.509 | 0.665 | 0.095 | 0.074 |
| 14 | ECC-NB | 0.142 | 0.655 | 0.659 | 0.475 | 0.633 | 0.109 | 0.076 |
| 15 | ECC-XGB | 0.134 | 0.675 | 0.689 | 0.498 | 0.654 | 0.103 | 0.088 |
| 16 | ECC-CART | 0.138 | 0.664 | 0.68 | 0.49 | 0.649 | 0.126 | 0.086 |
| 17 | ECC-MAJORITY | 0.165 | 0.629 | 0.44 | 0.342 | 0.496 | 0.293 | 0.202 |
| 18 | ECC-RANDOM | 0.275 | 0.286 | 0.185 | 0.118 | 0.202 | 0.469 | 0.681 |
| 19 | LP-RF | 0.086 | 0.81 | 0.796 | 0.69 | 0.783 | 0.127 | 0.245 |
| 20 | LP-SVM | 0.184 | 0.558 | 0.473 | 0.329 | 0.485 | 0.316 | 0.562 |
| 21 | LP-SMO | 0.168 | 0.61 | 0.565 | 0.406 | 0.556 | 0.272 | 0.519 |
| 22 | LP-C5.0 | 0.053 | 0.887 | 0.847 | 0.778 | 0.856 | 0.092 | 0.114 |
| 23 | LP-NB | 0.06 | 0.876 | 0.831 | 0.764 | 0.844 | 0.102 | 0.143 |
| 24 | LP-XGB | 0.196 | 0.551 | 0.532 | 0.354 | 0.504 | 0.298 | 0.576 |
| 25 | LP-CART | 0.176 | 0.603 | 0.519 | 0.372 | 0.52 | 0.3 | 0.536 |
| 26 | LP-MAJORITY | 0.239 | 0.416 | 0.547 | 0.302 | 0.453 | 0.315 | 0.752 |
| 27 | LP-RANDOM | 0.201 | 0.506 | 0.443 | 0.307 | 0.443 | 0.33 | 0.593 |
| 28 | HOMER-RF | 0.194 | 0.49 | 0.135 | 0.122 | 0.202 | 0.705 | 0.438 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 29 | HOMER-SVM | 0.219 | 0.278 | 0.098 | 0.082 | 0.139 | 0.713 | 0.655 |
| 30 | HOMER-SMO | 0.214 | 0.334 | 0.108 | 0.092 | 0.155 | 0.485 | 0.667 |
| 31 | HOMER-C5.0 | 0.201 | 0.441 | 0.135 | 0.119 | 0.197 | 0.708 | 0.452 |
| 32 | HOMER-NB | 0.211 | 0.365 | 0.111 | 0.095 | 0.161 | 0.697 | 0.681 |
| 33 | HOMER-XGB | 0.216 | 0.348 | 0.112 | 0.094 | 0.16 | 0.711 | 0.464 |
| 34 | HOMER-CART | 0.229 | 0.234 | 0.072 | 0.057 | 0.102 | 0.693 | 0.771 |
| 35 | HOMER-MAJORITY | 0.235 | 0.104 | 0.035 | 0.028 | 0.05 | 0.526 | 0.84 |
| 36 | HOMER-RANDOM | 0.248 | 0.086 | 0.028 | 0.021 | 0.039 | 0.757 | 0.912 |
| 37 | RAKEL-RF | 0.092 | 0.827 | 0.716 | 0.614 | 0.745 | 0.111 | 0.129 |
| 38 | RAKEL-SVM | 0.121 | 0.783 | 0.578 | 0.487 | 0.643 | 0.163 | 0.086 |
| 39 | RAKEL-SMO | 0.137 | 0.718 | 0.552 | 0.441 | 0.601 | 0.199 | 0.102 |
| 40 | RAKEL-C5.0 | 0.08 | 0.825 | 0.79 | 0.675 | 0.792 | 0.071 | 0.071 |
| 41 | RAKEL-NB | 0.132 | 0.656 | 0.744 | 0.534 | 0.677 | 0.106 | 0.257 |
| 42 | RAKEL-XGB | 0.099 | 0.783 | 0.72 | 0.596 | 0.73 | 0.093 | 0.043 |
| 43 | RAKEL-CART | 0.131 | 0.714 | 0.6 | 0.47 | 0.631 | 0.159 | 0.129 |
| 44 | RAKEL-MAJORITY | 0.158 | 0.691 | 0.392 | 0.332 | 0.484 | 0.273 | 0.202 |
| 45 | RAKEL-RANDOM | 0.437 | 0.237 | 0.575 | 0.197 | 0.323 | 0.409 | 0.769 |
| 46 | BRPLUS-RF | 0.105 | 0.827 | 0.64 | 0.563 | 0.695 | 0.052 | 0.086 |
| 47 | BRPLUS-SVM | 0.12 | 0.78 | 0.608 | 0.504 | 0.653 | 0.088 | 0.074 |
| 48 | BRPLUS-SMO | 0.147 | 0.707 | 0.51 | 0.405 | 0.559 | 0.266 | 0.202 |
| 49 | BRPLUS-C5.0 | 0.112 | 0.777 | 0.674 | 0.559 | 0.692 | 0.102 | 0.071 |
| 50 | BRPLUS-NB | 0.16 | 0.616 | 0.675 | 0.463 | 0.609 | 0.114 | 0.186 |
| 51 | BRPLUS-XGB | 0.132 | 0.706 | 0.651 | 0.492 | 0.645 | 0.095 | 0.1 |
| 52 | BRPLUS-CART | 0.146 | 0.655 | 0.635 | 0.459 | 0.618 | 0.158 | 0.193 |
| 53 | BRPLUS-MAJORITY | 0.154 | 0.727 | 0.379 | 0.333 | 0.483 | 0.301 | 0.202 |
| 54 | BRPLUS-RANDOM | 0.489 | 0.206 | 0.53 | 0.173 | 0.289 | 0.482 | 0.798 |
| 55 | BR-RF | 0.166 | 0.377 | 0.226 | 0.197 | 0.256 | 0.119 | 0.58 |
| 56 | BR-SVM | 0.126 | 0.576 | 0.199 | 0.19 | 0.28 | 0.314 | 0.405 |
| 57 | BR-SMO | 0.123 | 0.593 | 0.23 | 0.218 | 0.318 | 0.385 | 0.374 |
| 58 | BR-C5.0 | 0.122 | 0.567 | 0.303 | 0.275 | 0.371 | 0.273 | 0.436 |
| 59 | BR-NB | 0.16 | 0.448 | 0.544 | 0.314 | 0.433 | 0.146 | 0.493 |
| 60 | BR-XGB | 0.155 | 0.468 | 0.25 | 0.214 | 0.289 | 0.182 | 0.492 |
| 61 | BR-CART | 0.144 | 0.514 | 0.256 | 0.221 | 0.311 | 0.293 | 0.416 |
| 62 | BR-MAJORITY | 0.127 | 0.547 | 0.154 | 0.149 | 0.235 | 0.425 | 0.46 |
| 63 | BR-RANDOM | 0.502 | 0.129 | 0.461 | 0.11 | 0.189 | 0.492 | 0.844 |
| 64 | ECC-RF | 0.155 | 0.467 | 0.354 | 0.293 | 0.38 | 0.219 | 0.541 |
| 65 | ECC-SVM | 0.145 | 0.502 | 0.38 | 0.291 | 0.398 | 0.261 | 0.415 |
| 66 | ECC-SMO | 0.157 | 0.424 | 0.339 | 0.257 | 0.361 | 0.317 | 0.806 |
| 67 | ECC-C5.0 | 0.143 | 0.506 | 0.392 | 0.318 | 0.411 | 0.204 | 0.425 |
| 68 | ECC-NB | 0.158 | 0.474 | 0.321 | 0.241 | 0.335 | 0.165 | 0.417 |
| 69 | ECC-XGB | 0.147 | 0.483 | 0.354 | 0.297 | 0.39 | 0.214 | 0.434 |
| 70 | ECC-CART | 0.141 | 0.5 | 0.356 | 0.288 | 0.393 | 0.28 | 0.377 |
| 71 | ECC-MAJORITY | 0.141 | 0.456 | 0.254 | 0.214 | 0.316 | 0.37 | 0.445 |
| 72 | ECC-RANDOM | 0.232 | 0.177 | 0.119 | 0.069 | 0.12 | 0.465 | 0.862 |
| 73 | LP-RF | 0.172 | 0.262 | 0.257 | 0.217 | 0.25 | 0.405 | 0.785 |
| 74 | LP-SVM | 0.154 | 0.043 | 0.016 | 0.009 | 0.015 | 0.499 | 0.942 |
| 75 | LP-SMO | 0.163 | 0.111 | 0.094 | 0.063 | 0.093 | 0.466 | 0.925 |
| 76 | LP-C5.0 | 0.135 | 0.375 | 0.379 | 0.346 | 0.368 | 0.321 | 0.636 |
| 77 | LP-NB | 0.154 | 0.321 | 0.309 | 0.29 | 0.313 | 0.372 | 0.706 |
| 78 | LP-XGB | 0.183 | 0.151 | 0.142 | 0.096 | 0.137 | 0.457 | 0.905 |
| 79 | LP-CART | 0.177 | 0.178 | 0.179 | 0.129 | 0.169 | 0.433 | 0.901 |
| 80 | LP-MAJORITY | 0.145 | 0.049 | 0.004 | 0.004 | 0.008 | 0.496 | 0.951 |
| 81 | LP-RANDOM | 0.176 | 0.274 | 0.253 | 0.189 | 0.253 | 0.407 | 0.768 |
| 82 | HOMER-RF | 0.166 | 0.254 | 0.177 | 0.134 | 0.197 | 0.382 | 0.648 |
| 83 | HOMER-SVM | 0.181 | 0.338 | 0.27 | 0.196 | 0.291 | 0.455 | 0.524 |
| 84 | HOMER-SMO | 0.196 | 0.308 | 0.26 | 0.179 | 0.273 | 0.427 | 0.698 |
| 85 | HOMER-C5.0 | 0.144 | 0.451 | 0.263 | 0.222 | 0.32 | 0.454 | 0.426 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 86 | HOMER-NB | 0.163 | 0.349 | 0.224 | 0.165 | 0.253 | 0.477 | 0.535 |
| 87 | HOMER-XGB | 0.172 | 0.275 | 0.187 | 0.144 | 0.214 | 0.412 | 0.582 |
| 88 | HOMER-CART | 0.173 | 0.311 | 0.202 | 0.15 | 0.233 | 0.436 | 0.583 |
| 89 | HOMER-MAJORITY | 0.233 | 0.225 | 0.23 | 0.136 | 0.222 | 0.457 | 0.705 |
| 90 | HOMER-RANDOM | 0.191 | 0.132 | 0.051 | 0.034 | 0.061 | 0.539 | 0.872 |
| 91 | RAKEL-RF | 0.166 | 0.334 | 0.218 | 0.196 | 0.248 | 0.387 | 0.658 |
| 92 | RAKEL-SVM | 0.127 | 0.545 | 0.197 | 0.187 | 0.276 | 0.355 | 0.436 |
| 93 | RAKEL-SMO | 0.127 | 0.579 | 0.208 | 0.198 | 0.297 | 0.355 | 0.395 |
| 94 | RAKEL-C5.0 | 0.121 | 0.563 | 0.321 | 0.292 | 0.382 | 0.285 | 0.406 |
| 95 | RAKEL-NB | 0.156 | 0.455 | 0.581 | 0.33 | 0.451 | 0.193 | 0.555 |
| 96 | RAKEL-XGB | 0.16 | 0.346 | 0.234 | 0.21 | 0.267 | 0.354 | 0.581 |
| 97 | RAKEL-CART | 0.135 | 0.548 | 0.252 | 0.222 | 0.314 | 0.308 | 0.396 |
| 98 | RAKEL-MAJORITY | 0.133 | 0.522 | 0.092 | 0.09 | 0.15 | 0.419 | 0.436 |
| 99 | RAKEL-RANDOM | 0.377 | 0.136 | 0.33 | 0.103 | 0.174 | 0.485 | 0.923 |
| 100 | BRPLUS-RF | 0.17 | 0.362 | 0.189 | 0.17 | 0.224 | 0.241 | 0.589 |
| 101 | BRPLUS-SVM | 0.134 | 0.481 | 0.199 | 0.174 | 0.251 | 0.187 | 0.484 |
| 102 | BRPLUS-SMO | 0.139 | 0.498 | 0.313 | 0.267 | 0.367 | 0.368 | 0.514 |
| 103 | BRPLUS-C5.0 | 0.129 | 0.527 | 0.296 | 0.264 | 0.35 | 0.288 | 0.414 |
| 104 | BRPLUS-NB | 0.18 | 0.386 | 0.382 | 0.259 | 0.343 | 0.238 | 0.485 |
| 105 | BRPLUS-XGB | 0.153 | 0.403 | 0.269 | 0.23 | 0.298 | 0.24 | 0.521 |
| 106 | BRPLUS-CART | 0.154 | 0.437 | 0.26 | 0.209 | 0.291 | 0.361 | 0.554 |
| 107 | BRPLUS-MAJORITY | 0.127 | 0.547 | 0.154 | 0.149 | 0.235 | 0.425 | 0.46 |
| 108 | BRPLUS-RANDOM | 0.487 | 0.127 | 0.446 | 0.107 | 0.185 | 0.467 | 0.865 |
| 109 | BR-RF | 0.168 | 0.843 | 0.818 | 0.709 | 0.819 | 0.078 | 0.029 |
| 110 | BR-SVM | 0.202 | 0.828 | 0.741 | 0.637 | 0.77 | 0.122 | 0.029 |
| 111 | BR-SMO | 0.237 | 0.776 | 0.719 | 0.59 | 0.733 | 0.234 | 0.245 |
| 112 | BR-C5.0 | 0.165 | 0.827 | 0.835 | 0.708 | 0.822 | 0.1 | 0.1 |
| 113 | BR-NB | 0.222 | 0.765 | 0.787 | 0.63 | 0.762 | 0.135 | 0.06 |
| 114 | BR-XGB | 0.205 | 0.779 | 0.818 | 0.662 | 0.786 | 0.112 | 0.088 |
| 115 | BR-CART | 0.229 | 0.764 | 0.772 | 0.613 | 0.753 | 0.136 | 0.045 |
| 116 | BR-MAJORITY | 0.32 | 0.727 | 0.509 | 0.43 | 0.589 | 0.312 | 0.202 |
| 117 | BR-RANDOM | 0.441 | 0.527 | 0.503 | 0.345 | 0.503 | 0.474 | 0.526 |
| 118 | ECC-RF | 0.175 | 0.819 | 0.826 | 0.704 | 0.812 | 0.102 | 0.043 |
| 119 | ECC-SVM | 0.226 | 0.757 | 0.785 | 0.626 | 0.759 | 0.15 | 0.088 |
| 120 | ECC-SMO | 0.272 | 0.72 | 0.721 | 0.559 | 0.706 | 0.185 | 0.145 |
| 121 | ECC-C5.0 | 0.196 | 0.799 | 0.802 | 0.67 | 0.79 | 0.117 | 0.045 |
| 122 | ECC-NB | 0.254 | 0.735 | 0.753 | 0.596 | 0.732 | 0.173 | 0.119 |
| 123 | ECC-XGB | 0.201 | 0.792 | 0.797 | 0.657 | 0.783 | 0.122 | 0.071 |
| 124 | ECC-CART | 0.223 | 0.767 | 0.779 | 0.625 | 0.759 | 0.142 | 0.131 |
| 125 | ECC-MAJORITY | 0.37 | 0.604 | 0.631 | 0.442 | 0.604 | 0.317 | 0.202 |
| 126 | ECC-RANDOM | 0.485 | 0.47 | 0.392 | 0.273 | 0.412 | 0.494 | 0.505 |
| 127 | LP-RF | 0.168 | 0.838 | 0.845 | 0.739 | 0.828 | 0.166 | 0.174 |
| 128 | LP-SVM | 0.375 | 0.594 | 0.672 | 0.464 | 0.617 | 0.386 | 0.45 |
| 129 | LP-SMO | 0.312 | 0.681 | 0.719 | 0.528 | 0.677 | 0.322 | 0.348 |
| 130 | LP-C5.0 | 0.084 | 0.917 | 0.89 | 0.84 | 0.899 | 0.091 | 0.071 |
| 131 | LP-NB | 0.139 | 0.88 | 0.836 | 0.778 | 0.85 | 0.148 | 0.1 |
| 132 | LP-XGB | 0.31 | 0.678 | 0.737 | 0.547 | 0.684 | 0.319 | 0.348 |
| 133 | LP-CART | 0.332 | 0.67 | 0.665 | 0.49 | 0.641 | 0.331 | 0.381 |
| 134 | LP-MAJORITY | 0.407 | 0.565 | 0.518 | 0.381 | 0.528 | 0.426 | 0.202 |
| 135 | LP-RANDOM | 0.443 | 0.518 | 0.497 | 0.359 | 0.491 | 0.474 | 0.638 |
| 136 | HOMER-RF | 0.389 | 0.664 | 0.354 | 0.311 | 0.449 | 0.548 | 0.131 |
| 137 | HOMER-SVM | 0.42 | 0.596 | 0.346 | 0.289 | 0.426 | 0.558 | 0.174 |
| 138 | HOMER-SMO | 0.447 | 0.538 | 0.349 | 0.278 | 0.415 | 0.502 | 0.505 |
| 139 | HOMER-C5.0 | 0.356 | 0.699 | 0.359 | 0.323 | 0.463 | 0.553 | 0.19 |
| 140 | HOMER-NB | 0.407 | 0.644 | 0.295 | 0.261 | 0.391 | 0.561 | 0.186 |
| 141 | HOMER-XGB | 0.41 | 0.625 | 0.317 | 0.274 | 0.403 | 0.569 | 0.236 |
| 142 | HOMER-CART | 0.414 | 0.61 | 0.294 | 0.257 | 0.382 | 0.534 | 0.26 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 143 | HOMER-MAJORITY | 0.595 | 0.352 | 0.338 | 0.214 | 0.338 | 0.621 | 0.619 |
| 144 | HOMER-RANDOM | 0.502 | 0.368 | 0.124 | 0.105 | 0.177 | 0.647 | 0.664 |
| 145 | RAKEL-RF | 0.158 | 0.848 | 0.844 | 0.738 | 0.834 | 0.113 | 0.1 |
| 146 | RAKEL-SVM | 0.213 | 0.794 | 0.757 | 0.628 | 0.762 | 0.145 | 0.071 |
| 147 | RAKEL-SMO | 0.249 | 0.768 | 0.701 | 0.571 | 0.717 | 0.195 | 0.102 |
| 148 | RAKEL-C5.0 | 0.106 | 0.882 | 0.899 | 0.806 | 0.884 | 0.057 | 0.014 |
| 149 | RAKEL-NB | 0.213 | 0.77 | 0.802 | 0.647 | 0.775 | 0.142 | 0.086 |
| 150 | RAKEL-XGB | 0.168 | 0.818 | 0.845 | 0.717 | 0.822 | 0.104 | 0.029 |
| 151 | RAKEL-CART | 0.223 | 0.772 | 0.778 | 0.625 | 0.762 | 0.139 | 0.057 |
| 152 | RAKEL-MAJORITY | 0.34 | 0.682 | 0.511 | 0.414 | 0.573 | 0.33 | 0.202 |
| 153 | RAKEL-RANDOM | 0.501 | 0.468 | 0.576 | 0.347 | 0.504 | 0.508 | 0.693 |
| 154 | BRPLUS-RF | 0.177 | 0.835 | 0.809 | 0.706 | 0.809 | 0.093 | 0.029 |
| 155 | BRPLUS-SVM | 0.209 | 0.803 | 0.764 | 0.637 | 0.767 | 0.122 | 0.029 |
| 156 | BRPLUS-SMO | 0.237 | 0.75 | 0.76 | 0.607 | 0.742 | 0.241 | 0.202 |
| 157 | BRPLUS-C5.0 | 0.2 | 0.785 | 0.818 | 0.678 | 0.79 | 0.135 | 0.102 |
| 158 | BRPLUS-NB | 0.275 | 0.713 | 0.726 | 0.563 | 0.704 | 0.161 | 0.088 |
| 159 | BRPLUS-XGB | 0.209 | 0.782 | 0.796 | 0.648 | 0.775 | 0.127 | 0.102 |
| 160 | BRPLUS-CART | 0.243 | 0.772 | 0.72 | 0.587 | 0.727 | 0.155 | 0.074 |
| 161 | BRPLUS-MAJORITY | 0.32 | 0.727 | 0.509 | 0.43 | 0.589 | 0.312 | 0.202 |
| 162 | BRPLUS-RANDOM | 0.477 | 0.482 | 0.489 | 0.319 | 0.466 | 0.469 | 0.495 |
| 163 | BR-RF | 0.142 | 0.884 | 0.947 | 0.842 | 0.906 | 0.094 | 0.071 |
| 164 | BR-SVM | 0.167 | 0.852 | 0.945 | 0.81 | 0.886 | 0.116 | 0.071 |
| 165 | BR-SMO | 0.215 | 0.798 | 0.969 | 0.769 | 0.861 | 0.278 | 0.117 |
| 166 | BR-C5.0 | 0.154 | 0.864 | 0.951 | 0.827 | 0.896 | 0.138 | 0.1 |
| 167 | BR-NB | 0.214 | 0.851 | 0.878 | 0.763 | 0.849 | 0.129 | 0.06 |
| 168 | BR-XGB | 0.169 | 0.859 | 0.932 | 0.807 | 0.884 | 0.138 | 0.088 |
| 169 | BR-CART | 0.206 | 0.864 | 0.872 | 0.762 | 0.853 | 0.144 | 0.09 |
| 170 | BR-MAJORITY | 0.273 | 0.727 | 1 | 0.727 | 0.828 | 0.377 | 0.202 |
| 171 | BR-RANDOM | 0.51 | 0.735 | 0.478 | 0.398 | 0.539 | 0.4 | 0.16 |
| 172 | ECC-RF | 0.175 | 0.884 | 0.899 | 0.802 | 0.878 | 0.126 | 0.086 |
| 173 | ECC-SVM | 0.203 | 0.872 | 0.866 | 0.771 | 0.857 | 0.169 | 0.102 |
| 174 | ECC-SMO | 0.248 | 0.827 | 0.855 | 0.721 | 0.825 | 0.233 | 0.145 |
| 175 | ECC-C5.0 | 0.196 | 0.876 | 0.874 | 0.77 | 0.859 | 0.107 | 0.117 |
| 176 | ECC-NB | 0.224 | 0.861 | 0.85 | 0.751 | 0.84 | 0.162 | 0.088 |
| 177 | ECC-XGB | 0.187 | 0.874 | 0.882 | 0.78 | 0.864 | 0.128 | 0.071 |
| 178 | ECC-CART | 0.241 | 0.833 | 0.838 | 0.726 | 0.821 | 0.162 | 0.086 |
| 179 | ECC-MAJORITY | 0.352 | 0.71 | 0.845 | 0.638 | 0.757 | 0.388 | 0.202 |
| 180 | ECC-RANDOM | 0.371 | 0.725 | 0.773 | 0.595 | 0.722 | 0.373 | 0.245 |
| 181 | LP-RF | 0.18 | 0.866 | 0.905 | 0.794 | 0.872 | 0.196 | 0.1 |
| 182 | LP-SVM | 0.232 | 0.809 | 0.914 | 0.749 | 0.84 | 0.243 | 0.114 |
| 183 | LP-SMO | 0.254 | 0.79 | 0.919 | 0.726 | 0.829 | 0.296 | 0.131 |
| 184 | LP-C5.0 | 0.179 | 0.87 | 0.902 | 0.797 | 0.874 | 0.201 | 0.1 |
| 185 | LP-NB | 0.179 | 0.911 | 0.854 | 0.787 | 0.869 | 0.152 | 0.057 |
| 186 | LP-XGB | 0.204 | 0.845 | 0.909 | 0.78 | 0.862 | 0.234 | 0.129 |
| 187 | LP-CART | 0.311 | 0.762 | 0.809 | 0.672 | 0.768 | 0.363 | 0.231 |
| 188 | LP-MAJORITY | 0.273 | 0.727 | 1 | 0.727 | 0.828 | 0.377 | 0.202 |
| 189 | LP-RANDOM | 0.382 | 0.752 | 0.703 | 0.58 | 0.71 | 0.343 | 0.231 |
| 190 | HOMER-RF | 0.171 | 0.857 | 0.941 | 0.812 | 0.887 | 0.097 | 0.057 |
| 191 | HOMER-SVM | 0.169 | 0.84 | 0.959 | 0.811 | 0.885 | 0.154 | 0.133 |
| 192 | HOMER-SMO | 0.218 | 0.786 | 0.977 | 0.766 | 0.859 | 0.291 | 0.117 |
| 193 | HOMER-C5.0 | 0.158 | 0.871 | 0.939 | 0.824 | 0.893 | 0.141 | 0.088 |
| 194 | HOMER-NB | 0.245 | 0.848 | 0.833 | 0.722 | 0.824 | 0.163 | 0.114 |
| 195 | HOMER-XGB | 0.178 | 0.853 | 0.939 | 0.807 | 0.883 | 0.136 | 0.145 |
| 196 | HOMER-CART | 0.229 | 0.821 | 0.893 | 0.741 | 0.838 | 0.243 | 0.133 |
| 197 | HOMER-MAJORITY | 0.273 | 0.727 | 1 | 0.727 | 0.828 | 0.377 | 0.202 |
| 198 | HOMER-RANDOM | 0.554 | 0.691 | 0.378 | 0.344 | 0.465 | 0.428 | 0.281 |
| 199 | RAKEL-RF | 0.152 | 0.884 | 0.935 | 0.831 | 0.898 | 0.135 | 0.086 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 200 | RAKEL-SVM | 0.165 | 0.848 | 0.953 | 0.813 | 0.887 | 0.16 | 0.086 |
| 201 | RAKEL-SMO | 0.21 | 0.799 | 0.969 | 0.77 | 0.862 | 0.199 | 0.131 |
| 202 | RAKEL-C5.0 | 0.122 | 0.892 | 0.96 | 0.863 | 0.918 | 0.106 | 0.071 |
| 203 | RAKEL-NB | 0.205 | 0.881 | 0.857 | 0.763 | 0.854 | 0.126 | 0.057 |
| 204 | RAKEL-XGB | 0.163 | 0.875 | 0.924 | 0.817 | 0.888 | 0.161 | 0.086 |
| 205 | RAKEL-CART | 0.199 | 0.851 | 0.905 | 0.78 | 0.864 | 0.149 | 0.043 |
| 206 | RAKEL-MAJORITY | 0.273 | 0.727 | 1 | 0.727 | 0.828 | 0.344 | 0.202 |
| 207 | RAKEL-RANDOM | 0.454 | 0.8 | 0.51 | 0.453 | 0.599 | 0.35 | 0.19 |
| 208 | BRPLUS-RF | 0.142 | 0.889 | 0.942 | 0.841 | 0.905 | 0.091 | 0.043 |
| 209 | BRPLUS-SVM | 0.16 | 0.864 | 0.94 | 0.82 | 0.89 | 0.148 | 0.131 |
| 210 | BRPLUS-SMO | 0.204 | 0.812 | 0.956 | 0.777 | 0.864 | 0.264 | 0.131 |
| 211 | BRPLUS-C5.0 | 0.154 | 0.888 | 0.923 | 0.823 | 0.895 | 0.131 | 0.06 |
| 212 | BRPLUS-NB | 0.268 | 0.883 | 0.762 | 0.695 | 0.798 | 0.138 | 0.074 |
| 213 | BRPLUS-XGB | 0.172 | 0.864 | 0.927 | 0.804 | 0.883 | 0.125 | 0.06 |
| 214 | BRPLUS-CART | 0.22 | 0.852 | 0.869 | 0.748 | 0.845 | 0.168 | 0.105 |
| 215 | BRPLUS-MAJORITY | 0.273 | 0.727 | 1 | 0.727 | 0.828 | 0.377 | 0.202 |
| 216 | BRPLUS-RANDOM | 0.546 | 0.708 | 0.44 | 0.364 | 0.513 | 0.465 | 0.307 |
| 217 | BR-RF | 0.097 | 0.92 | 0.971 | 0.896 | 0.932 | 0.05 | 0.043 |
| 218 | BR-SVM | 0.14 | 0.867 | 0.976 | 0.848 | 0.9 | 0.1 | 0.1 |
| 219 | BR-SMO | 0.152 | 0.858 | 0.986 | 0.844 | 0.896 | 0.16 | 0.117 |
| 220 | BR-C5.0 | 0.111 | 0.895 | 0.974 | 0.869 | 0.916 | 0.094 | 0.074 |
| 221 | BR-NB | 0.149 | 0.91 | 0.905 | 0.839 | 0.889 | 0.058 | 0.06 |
| 222 | BR-XGB | 0.131 | 0.876 | 0.96 | 0.85 | 0.898 | 0.101 | 0.102 |
| 223 | BR-CART | 0.185 | 0.869 | 0.898 | 0.802 | 0.859 | 0.133 | 0.107 |
| 224 | BR-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 225 | BR-RANDOM | 0.541 | 0.802 | 0.45 | 0.422 | 0.546 | 0.204 | 0.176 |
| 226 | ECC-RF | 0.144 | 0.906 | 0.917 | 0.841 | 0.891 | 0.086 | 0.086 |
| 227 | ECC-SVM | 0.179 | 0.877 | 0.881 | 0.794 | 0.853 | 0.117 | 0.102 |
| 228 | ECC-SMO | 0.166 | 0.857 | 0.943 | 0.821 | 0.877 | 0.131 | 0.117 |
| 229 | ECC-C5.0 | 0.154 | 0.906 | 0.902 | 0.827 | 0.88 | 0.107 | 0.1 |
| 230 | ECC-NB | 0.192 | 0.887 | 0.876 | 0.787 | 0.853 | 0.101 | 0.102 |
| 231 | ECC-XGB | 0.144 | 0.917 | 0.919 | 0.846 | 0.898 | 0.064 | 0.043 |
| 232 | ECC-CART | 0.216 | 0.858 | 0.888 | 0.77 | 0.846 | 0.093 | 0.057 |
| 233 | ECC-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 234 | ECC-RANDOM | 0.336 | 0.781 | 0.792 | 0.652 | 0.734 | 0.257 | 0.236 |
| 235 | LP-RF | 0.13 | 0.906 | 0.95 | 0.86 | 0.907 | 0.121 | 0.086 |
| 236 | LP-SVM | 0.131 | 0.888 | 0.974 | 0.867 | 0.912 | 0.143 | 0.114 |
| 237 | LP-SMO | 0.152 | 0.853 | 0.986 | 0.839 | 0.894 | 0.167 | 0.131 |
| 238 | LP-C5.0 | 0.082 | 0.935 | 0.96 | 0.899 | 0.934 | 0.114 | 0.086 |
| 239 | LP-NB | 0.173 | 0.925 | 0.859 | 0.808 | 0.871 | 0.107 | 0.086 |
| 240 | LP-XGB | 0.129 | 0.901 | 0.926 | 0.861 | 0.897 | 0.129 | 0.129 |
| 241 | LP-CART | 0.193 | 0.819 | 0.89 | 0.8 | 0.839 | 0.221 | 0.2 |
| 242 | LP-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 243 | LP-RANDOM | 0.471 | 0.828 | 0.547 | 0.493 | 0.618 | 0.19 | 0.19 |
| 244 | HOMER-RF | 0.097 | 0.92 | 0.971 | 0.896 | 0.932 | 0.05 | 0.043 |
| 245 | HOMER-SVM | 0.14 | 0.867 | 0.976 | 0.848 | 0.9 | 0.1 | 0.1 |
| 246 | HOMER-SMO | 0.152 | 0.858 | 0.986 | 0.844 | 0.896 | 0.16 | 0.117 |
| 247 | HOMER-C5.0 | 0.111 | 0.895 | 0.974 | 0.869 | 0.916 | 0.094 | 0.074 |
| 248 | HOMER-NB | 0.149 | 0.91 | 0.905 | 0.839 | 0.889 | 0.058 | 0.06 |
| 249 | HOMER-XGB | 0.131 | 0.876 | 0.96 | 0.85 | 0.898 | 0.101 | 0.102 |
| 250 | HOMER-CART | 0.185 | 0.869 | 0.898 | 0.802 | 0.859 | 0.133 | 0.107 |
| 251 | HOMER-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 252 | HOMER-RANDOM | 0.541 | 0.802 | 0.45 | 0.422 | 0.546 | 0.204 | 0.176 |
| 253 | RAKEL-RF | 0.116 | 0.91 | 0.96 | 0.875 | 0.917 | 0.121 | 0.086 |
| 254 | RAKEL-SVM | 0.136 | 0.883 | 0.979 | 0.862 | 0.909 | 0.136 | 0.1 |
| 255 | RAKEL-SMO | 0.152 | 0.853 | 0.986 | 0.839 | 0.894 | 0.167 | 0.131 |
| 256 | RAKEL-C5.0 | 0.082 | 0.935 | 0.96 | 0.899 | 0.934 | 0.114 | 0.086 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 257 | RAKEL-NB | 0.173 | 0.925 | 0.859 | 0.808 | 0.871 | 0.107 | 0.086 |
| 258 | RAKEL-XGB | 0.129 | 0.901 | 0.926 | 0.861 | 0.897 | 0.129 | 0.129 |
| 259 | RAKEL-CART | 0.193 | 0.819 | 0.89 | 0.8 | 0.839 | 0.221 | 0.2 |
| 260 | RAKEL-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 261 | RAKEL-RANDOM | 0.471 | 0.828 | 0.547 | 0.493 | 0.618 | 0.19 | 0.19 |
| 262 | BRPLUS-RF | 0.111 | 0.915 | 0.96 | 0.879 | 0.918 | 0.043 | 0.014 |
| 263 | BRPLUS-SVM | 0.121 | 0.883 | 0.988 | 0.871 | 0.917 | 0.101 | 0.117 |
| 264 | BRPLUS-SMO | 0.137 | 0.87 | 0.986 | 0.856 | 0.904 | 0.16 | 0.117 |
| 265 | BRPLUS-C5.0 | 0.117 | 0.894 | 0.95 | 0.865 | 0.907 | 0.11 | 0.09 |
| 266 | BRPLUS-NB | 0.183 | 0.903 | 0.879 | 0.806 | 0.862 | 0.058 | 0.06 |
| 267 | BRPLUS-XGB | 0.131 | 0.876 | 0.96 | 0.85 | 0.898 | 0.087 | 0.102 |
| 268 | BRPLUS-CART | 0.185 | 0.869 | 0.898 | 0.802 | 0.859 | 0.133 | 0.107 |
| 269 | BRPLUS-MAJORITY | 0.19 | 0.81 | 1 | 0.81 | 0.871 | 0.231 | 0.202 |
| 270 | BRPLUS-RANDOM | 0.511 | 0.832 | 0.472 | 0.45 | 0.571 | 0.211 | 0.19 |
| 271 | BR-RF | 0.123 | 0.836 | 0.731 | 0.633 | 0.761 | 0.056 | 0.071 |
| 272 | BR-SVM | 0.158 | 0.799 | 0.627 | 0.53 | 0.68 | 0.102 | 0.029 |
| 273 | BR-SMO | 0.179 | 0.753 | 0.59 | 0.483 | 0.642 | 0.233 | 0.26 |
| 274 | BR-C5.0 | 0.124 | 0.806 | 0.761 | 0.635 | 0.769 | 0.084 | 0.1 |
| 275 | BR-NB | 0.187 | 0.666 | 0.745 | 0.532 | 0.679 | 0.108 | 0.06 |
| 276 | BR-XGB | 0.159 | 0.731 | 0.732 | 0.566 | 0.711 | 0.088 | 0.088 |
| 277 | BR-CART | 0.179 | 0.698 | 0.688 | 0.516 | 0.671 | 0.124 | 0.045 |
| 278 | BR-MAJORITY | 0.215 | 0.727 | 0.403 | 0.352 | 0.505 | 0.297 | 0.202 |
| 279 | BR-RANDOM | 0.483 | 0.284 | 0.479 | 0.215 | 0.345 | 0.508 | 0.795 |
| 280 | ECC-RF | 0.163 | 0.728 | 0.722 | 0.562 | 0.706 | 0.112 | 0.043 |
| 281 | ECC-SVM | 0.193 | 0.665 | 0.685 | 0.496 | 0.653 | 0.154 | 0.086 |
| 282 | ECC-SMO | 0.205 | 0.648 | 0.672 | 0.479 | 0.638 | 0.152 | 0.074 |
| 283 | ECC-C5.0 | 0.173 | 0.698 | 0.712 | 0.537 | 0.689 | 0.106 | 0.029 |
| 284 | ECC-NB | 0.211 | 0.646 | 0.642 | 0.47 | 0.619 | 0.156 | 0.131 |
| 285 | ECC-XGB | 0.179 | 0.696 | 0.701 | 0.526 | 0.679 | 0.118 | 0.057 |
| 286 | ECC-CART | 0.194 | 0.664 | 0.689 | 0.499 | 0.657 | 0.138 | 0.071 |
| 287 | ECC-MAJORITY | 0.244 | 0.585 | 0.496 | 0.36 | 0.518 | 0.284 | 0.202 |
| 288 | ECC-RANDOM | 0.415 | 0.324 | 0.361 | 0.196 | 0.317 | 0.474 | 0.664 |
| 289 | LP-RF | 0.121 | 0.804 | 0.796 | 0.687 | 0.784 | 0.142 | 0.231 |
| 290 | LP-SVM | 0.249 | 0.584 | 0.533 | 0.378 | 0.534 | 0.311 | 0.502 |
| 291 | LP-SMO | 0.23 | 0.627 | 0.586 | 0.432 | 0.582 | 0.29 | 0.505 |
| 292 | LP-C5.0 | 0.059 | 0.906 | 0.871 | 0.814 | 0.882 | 0.081 | 0.1 |
| 293 | LP-NB | 0.09 | 0.863 | 0.827 | 0.763 | 0.837 | 0.115 | 0.143 |
| 294 | LP-XGB | 0.257 | 0.582 | 0.592 | 0.408 | 0.557 | 0.302 | 0.55 |
| 295 | LP-CART | 0.256 | 0.595 | 0.571 | 0.402 | 0.545 | 0.311 | 0.567 |
| 296 | LP-MAJORITY | 0.346 | 0.423 | 0.583 | 0.317 | 0.47 | 0.367 | 0.752 |
| 297 | LP-RANDOM | 0.291 | 0.502 | 0.433 | 0.305 | 0.439 | 0.367 | 0.519 |
| 298 | HOMER-RF | 0.264 | 0.521 | 0.221 | 0.196 | 0.299 | 0.649 | 0.407 |
| 299 | HOMER-SVM | 0.301 | 0.468 | 0.189 | 0.153 | 0.251 | 0.661 | 0.407 |
| 300 | HOMER-SMO | 0.312 | 0.385 | 0.178 | 0.143 | 0.23 | 0.471 | 0.671 |
| 301 | HOMER-C5.0 | 0.264 | 0.533 | 0.216 | 0.193 | 0.296 | 0.653 | 0.421 |
| 302 | HOMER-NB | 0.286 | 0.473 | 0.178 | 0.149 | 0.242 | 0.665 | 0.45 |
| 303 | HOMER-XGB | 0.301 | 0.403 | 0.184 | 0.151 | 0.243 | 0.655 | 0.493 |
| 304 | HOMER-CART | 0.321 | 0.322 | 0.155 | 0.121 | 0.198 | 0.625 | 0.548 |
| 305 | HOMER-MAJORITY | 0.4 | 0.198 | 0.131 | 0.086 | 0.152 | 0.545 | 0.871 |
| 306 | HOMER-RANDOM | 0.347 | 0.119 | 0.03 | 0.027 | 0.048 | 0.705 | 0.883 |
| 307 | RAKEL-RF | 0.123 | 0.822 | 0.754 | 0.653 | 0.771 | 0.104 | 0.1 |
| 308 | RAKEL-SVM | 0.165 | 0.789 | 0.61 | 0.514 | 0.666 | 0.169 | 0.086 |
| 309 | RAKEL-SMO | 0.182 | 0.736 | 0.596 | 0.481 | 0.64 | 0.193 | 0.131 |
| 310 | RAKEL-C5.0 | 0.097 | 0.834 | 0.824 | 0.708 | 0.818 | 0.06 | 0.071 |
| 311 | RAKEL-NB | 0.17 | 0.69 | 0.777 | 0.577 | 0.713 | 0.116 | 0.143 |
| 312 | RAKEL-XGB | 0.136 | 0.782 | 0.741 | 0.613 | 0.744 | 0.095 | 0.057 |
| 313 | RAKEL-CART | 0.183 | 0.713 | 0.641 | 0.498 | 0.653 | 0.157 | 0.071 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 314 | RAKEL-MAJORITY | 0.217 | 0.718 | 0.404 | 0.35 | 0.504 | 0.285 | 0.202 |
| 315 | RAKEL-RANDOM | 0.5 | 0.294 | 0.581 | 0.241 | 0.377 | 0.481 | 0.75 |
| 316 | BRPLUS-RF | 0.139 | 0.826 | 0.677 | 0.591 | 0.721 | 0.067 | 0.045 |
| 317 | BRPLUS-SVM | 0.167 | 0.771 | 0.635 | 0.523 | 0.67 | 0.113 | 0.088 |
| 318 | BRPLUS-SMO | 0.2 | 0.696 | 0.566 | 0.452 | 0.602 | 0.255 | 0.202 |
| 319 | BRPLUS-C5.0 | 0.167 | 0.733 | 0.698 | 0.554 | 0.69 | 0.123 | 0.074 |
| 320 | BRPLUS-NB | 0.224 | 0.608 | 0.687 | 0.47 | 0.619 | 0.126 | 0.1 |
| 321 | BRPLUS-XGB | 0.174 | 0.713 | 0.697 | 0.533 | 0.68 | 0.117 | 0.086 |
| 322 | BRPLUS-CART | 0.199 | 0.659 | 0.684 | 0.493 | 0.648 | 0.154 | 0.193 |
| 323 | BRPLUS-MAJORITY | 0.215 | 0.727 | 0.403 | 0.352 | 0.505 | 0.297 | 0.202 |
| 324 | BRPLUS-RANDOM | 0.519 | 0.259 | 0.456 | 0.197 | 0.321 | 0.525 | 0.752 |
| 325 | BR-RF | 0.142 | 0.836 | 0.789 | 0.679 | 0.797 | 0.063 | 0.071 |
| 326 | BR-SVM | 0.181 | 0.822 | 0.683 | 0.586 | 0.727 | 0.108 | 0.043 |
| 327 | BR-SMO | 0.21 | 0.766 | 0.655 | 0.534 | 0.687 | 0.219 | 0.245 |
| 328 | BR-C5.0 | 0.147 | 0.817 | 0.797 | 0.671 | 0.795 | 0.089 | 0.1 |
| 329 | BR-NB | 0.211 | 0.709 | 0.769 | 0.579 | 0.719 | 0.125 | 0.06 |
| 330 | BR-XGB | 0.185 | 0.747 | 0.784 | 0.614 | 0.749 | 0.091 | 0.088 |
| 331 | BR-CART | 0.21 | 0.722 | 0.736 | 0.563 | 0.711 | 0.123 | 0.045 |
| 332 | BR-MAJORITY | 0.27 | 0.727 | 0.448 | 0.384 | 0.542 | 0.292 | 0.202 |
| 333 | BR-RANDOM | 0.526 | 0.341 | 0.479 | 0.253 | 0.39 | 0.54 | 0.652 |
| 334 | ECC-RF | 0.179 | 0.76 | 0.792 | 0.628 | 0.76 | 0.091 | 0.014 |
| 335 | ECC-SVM | 0.214 | 0.719 | 0.748 | 0.563 | 0.711 | 0.138 | 0.102 |
| 336 | ECC-SMO | 0.228 | 0.701 | 0.721 | 0.538 | 0.69 | 0.152 | 0.074 |
| 337 | ECC-C5.0 | 0.192 | 0.744 | 0.773 | 0.607 | 0.741 | 0.104 | 0.043 |
| 338 | ECC-NB | 0.252 | 0.668 | 0.682 | 0.514 | 0.658 | 0.164 | 0.102 |
| 339 | ECC-XGB | 0.201 | 0.737 | 0.753 | 0.585 | 0.728 | 0.115 | 0.086 |
| 340 | ECC-CART | 0.206 | 0.724 | 0.751 | 0.576 | 0.72 | 0.124 | 0.088 |
| 341 | ECC-MAJORITY | 0.312 | 0.575 | 0.603 | 0.41 | 0.575 | 0.281 | 0.202 |
| 342 | ECC-RANDOM | 0.468 | 0.379 | 0.408 | 0.243 | 0.379 | 0.488 | 0.567 |
| 343 | LP-RF | 0.143 | 0.825 | 0.82 | 0.712 | 0.807 | 0.141 | 0.174 |
| 344 | LP-SVM | 0.322 | 0.576 | 0.62 | 0.425 | 0.576 | 0.323 | 0.464 |
| 345 | LP-SMO | 0.277 | 0.644 | 0.679 | 0.483 | 0.635 | 0.283 | 0.493 |
| 346 | LP-C5.0 | 0.072 | 0.911 | 0.878 | 0.825 | 0.888 | 0.082 | 0.086 |
| 347 | LP-NB | 0.117 | 0.865 | 0.826 | 0.763 | 0.838 | 0.126 | 0.129 |
| 348 | LP-XGB | 0.294 | 0.622 | 0.679 | 0.474 | 0.621 | 0.294 | 0.493 |
| 349 | LP-CART | 0.334 | 0.585 | 0.604 | 0.403 | 0.558 | 0.343 | 0.507 |
| 350 | LP-MAJORITY | 0.419 | 0.449 | 0.637 | 0.359 | 0.514 | 0.404 | 0.624 |
| 351 | LP-RANDOM | 0.355 | 0.537 | 0.446 | 0.335 | 0.466 | 0.385 | 0.505 |
| 352 | HOMER-RF | 0.189 | 0.754 | 0.742 | 0.598 | 0.723 | 0.199 | 0.114 |
| 353 | HOMER-SVM | 0.27 | 0.613 | 0.717 | 0.496 | 0.64 | 0.254 | 0.205 |
| 354 | HOMER-SMO | 0.286 | 0.59 | 0.696 | 0.469 | 0.619 | 0.282 | 0.331 |
| 355 | HOMER-C5.0 | 0.186 | 0.744 | 0.745 | 0.597 | 0.723 | 0.219 | 0.088 |
| 356 | HOMER-NB | 0.255 | 0.656 | 0.652 | 0.485 | 0.63 | 0.275 | 0.202 |
| 357 | HOMER-XGB | 0.23 | 0.695 | 0.719 | 0.531 | 0.672 | 0.225 | 0.148 |
| 358 | HOMER-CART | 0.269 | 0.626 | 0.687 | 0.486 | 0.634 | 0.25 | 0.274 |
| 359 | HOMER-MAJORITY | 0.44 | 0.422 | 0.659 | 0.347 | 0.501 | 0.428 | 0.605 |
| 360 | HOMER-RANDOM | 0.424 | 0.329 | 0.14 | 0.11 | 0.183 | 0.526 | 0.74 |
| 361 | RAKEL-RF | 0.145 | 0.822 | 0.802 | 0.685 | 0.798 | 0.105 | 0.114 |
| 362 | RAKEL-SVM | 0.2 | 0.783 | 0.676 | 0.555 | 0.702 | 0.158 | 0.086 |
| 363 | RAKEL-SMO | 0.215 | 0.757 | 0.656 | 0.528 | 0.681 | 0.179 | 0.102 |
| 364 | RAKEL-C5.0 | 0.11 | 0.846 | 0.861 | 0.749 | 0.845 | 0.065 | 0.043 |
| 365 | RAKEL-NB | 0.191 | 0.723 | 0.817 | 0.625 | 0.751 | 0.111 | 0.131 |
| 366 | RAKEL-XGB | 0.16 | 0.781 | 0.802 | 0.66 | 0.778 | 0.1 | 0.057 |
| 367 | RAKEL-CART | 0.206 | 0.749 | 0.708 | 0.564 | 0.711 | 0.144 | 0.086 |
| 368 | RAKEL-MAJORITY | 0.276 | 0.711 | 0.446 | 0.379 | 0.536 | 0.285 | 0.202 |
| 369 | RAKEL-RANDOM | 0.52 | 0.362 | 0.557 | 0.284 | 0.427 | 0.511 | 0.645 |
| 370 | BRPLUS-RF | 0.157 | 0.826 | 0.759 | 0.656 | 0.772 | 0.071 | 0.029 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 371 | BRPLUS-SVM | 0.191 | 0.791 | 0.696 | 0.581 | 0.718 | 0.116 | 0.074 |
| 372 | BRPLUS-SMO | 0.228 | 0.717 | 0.665 | 0.524 | 0.669 | 0.237 | 0.29 |
| 373 | BRPLUS-C5.0 | 0.197 | 0.748 | 0.742 | 0.596 | 0.731 | 0.134 | 0.06 |
| 374 | BRPLUS-NB | 0.251 | 0.661 | 0.705 | 0.519 | 0.663 | 0.136 | 0.088 |
| 375 | BRPLUS-XGB | 0.198 | 0.736 | 0.747 | 0.587 | 0.723 | 0.121 | 0.043 |
| 376 | BRPLUS-CART | 0.227 | 0.697 | 0.728 | 0.545 | 0.693 | 0.153 | 0.148 |
| 377 | BRPLUS-MAJORITY | 0.27 | 0.727 | 0.448 | 0.384 | 0.542 | 0.292 | 0.202 |
| 378 | BRPLUS-RANDOM | 0.512 | 0.349 | 0.452 | 0.248 | 0.385 | 0.53 | 0.681 |
| 379 | BR-RF | 0.175 | 0.814 | 0.762 | 0.655 | 0.77 | 0.089 | 0.043 |
| 380 | BR-SVM | 0.22 | 0.777 | 0.667 | 0.559 | 0.696 | 0.134 | 0.014 |
| 381 | BR-SMO | 0.257 | 0.729 | 0.631 | 0.506 | 0.656 | 0.269 | 0.186 |
| 382 | BR-C5.0 | 0.177 | 0.791 | 0.788 | 0.652 | 0.774 | 0.107 | 0.071 |
| 383 | BR-NB | 0.236 | 0.707 | 0.752 | 0.575 | 0.708 | 0.149 | 0.071 |
| 384 | BR-XGB | 0.216 | 0.743 | 0.769 | 0.609 | 0.735 | 0.125 | 0.1 |
| 385 | BR-CART | 0.243 | 0.719 | 0.717 | 0.55 | 0.693 | 0.152 | 0.031 |
| 386 | BR-MAJORITY | 0.354 | 0.643 | 0.329 | 0.279 | 0.424 | 0.38 | 0.317 |
| 387 | BR-RANDOM | 0.474 | 0.427 | 0.428 | 0.271 | 0.413 | 0.487 | 0.517 |
| 388 | ECC-RF | 0.22 | 0.744 | 0.756 | 0.6 | 0.729 | 0.133 | 0.071 |
| 389 | ECC-SVM | 0.256 | 0.693 | 0.711 | 0.543 | 0.683 | 0.197 | 0.117 |
| 390 | ECC-SMO | 0.291 | 0.649 | 0.691 | 0.503 | 0.649 | 0.232 | 0.188 |
| 391 | ECC-C5.0 | 0.211 | 0.743 | 0.773 | 0.624 | 0.744 | 0.145 | 0.114 |
| 392 | ECC-NB | 0.268 | 0.68 | 0.691 | 0.532 | 0.66 | 0.211 | 0.174 |
| 393 | ECC-XGB | 0.245 | 0.718 | 0.711 | 0.553 | 0.69 | 0.164 | 0.1 |
| 394 | ECC-CART | 0.274 | 0.677 | 0.691 | 0.514 | 0.657 | 0.191 | 0.143 |
| 395 | ECC-MAJORITY | 0.405 | 0.508 | 0.516 | 0.336 | 0.493 | 0.376 | 0.317 |
| 396 | ECC-RANDOM | 0.479 | 0.388 | 0.33 | 0.225 | 0.335 | 0.544 | 0.707 |
| 397 | LP-RF | 0.15 | 0.828 | 0.832 | 0.735 | 0.814 | 0.157 | 0.171 |
| 398 | LP-SVM | 0.353 | 0.578 | 0.669 | 0.455 | 0.599 | 0.362 | 0.45 |
| 399 | LP-SMO | 0.334 | 0.606 | 0.644 | 0.454 | 0.597 | 0.363 | 0.36 |
| 400 | LP-C5.0 | 0.089 | 0.884 | 0.858 | 0.806 | 0.862 | 0.112 | 0.1 |
| 401 | LP-NB | 0.161 | 0.836 | 0.794 | 0.727 | 0.804 | 0.179 | 0.129 |
| 402 | LP-XGB | 0.3 | 0.648 | 0.7 | 0.524 | 0.65 | 0.312 | 0.386 |
| 403 | LP-CART | 0.353 | 0.592 | 0.581 | 0.411 | 0.554 | 0.379 | 0.307 |
| 404 | LP-MAJORITY | 0.423 | 0.481 | 0.416 | 0.315 | 0.435 | 0.464 | 0.607 |
| 405 | LP-RANDOM | 0.479 | 0.436 | 0.428 | 0.284 | 0.408 | 0.526 | 0.621 |
| 406 | HOMER-RF | 0.254 | 0.692 | 0.736 | 0.564 | 0.695 | 0.219 | 0.114 |
| 407 | HOMER-SVM | 0.306 | 0.624 | 0.761 | 0.522 | 0.666 | 0.256 | 0.171 |
| 408 | HOMER-SMO | 0.386 | 0.537 | 0.758 | 0.452 | 0.609 | 0.378 | 0.331 |
| 409 | HOMER-C5.0 | 0.232 | 0.732 | 0.737 | 0.587 | 0.716 | 0.219 | 0.086 |
| 410 | HOMER-NB | 0.306 | 0.683 | 0.569 | 0.454 | 0.59 | 0.296 | 0.071 |
| 411 | HOMER-XGB | 0.301 | 0.641 | 0.684 | 0.501 | 0.64 | 0.262 | 0.143 |
| 412 | HOMER-CART | 0.31 | 0.642 | 0.679 | 0.488 | 0.632 | 0.268 | 0.188 |
| 413 | HOMER-MAJORITY | 0.561 | 0.402 | 0.753 | 0.353 | 0.51 | 0.545 | 0.562 |
| 414 | HOMER-RANDOM | 0.445 | 0.453 | 0.216 | 0.173 | 0.273 | 0.476 | 0.526 |
| 415 | RAKEL-RF | 0.174 | 0.797 | 0.803 | 0.679 | 0.782 | 0.126 | 0.086 |
| 416 | RAKEL-SVM | 0.234 | 0.745 | 0.683 | 0.551 | 0.687 | 0.157 | 0.114 |
| 417 | RAKEL-SMO | 0.269 | 0.707 | 0.638 | 0.501 | 0.646 | 0.224 | 0.1 |
| 418 | RAKEL-C5.0 | 0.112 | 0.855 | 0.854 | 0.762 | 0.845 | 0.07 | 0.043 |
| 419 | RAKEL-NB | 0.212 | 0.713 | 0.816 | 0.623 | 0.746 | 0.134 | 0.086 |
| 420 | RAKEL-XGB | 0.19 | 0.77 | 0.806 | 0.652 | 0.768 | 0.126 | 0.1 |
| 421 | RAKEL-CART | 0.249 | 0.716 | 0.74 | 0.565 | 0.702 | 0.172 | 0.1 |
| 422 | RAKEL-MAJORITY | 0.38 | 0.571 | 0.371 | 0.287 | 0.434 | 0.389 | 0.521 |
| 423 | RAKEL-RANDOM | 0.498 | 0.423 | 0.6 | 0.335 | 0.481 | 0.494 | 0.662 |
| 424 | BRPLUS-RF | 0.185 | 0.791 | 0.761 | 0.651 | 0.758 | 0.106 | 0.043 |
| 425 | BRPLUS-SVM | 0.227 | 0.756 | 0.688 | 0.557 | 0.693 | 0.141 | 0.043 |
| 426 | BRPLUS-SMO | 0.252 | 0.714 | 0.688 | 0.535 | 0.677 | 0.261 | 0.229 |
| 427 | BRPLUS-C5.0 | 0.209 | 0.76 | 0.747 | 0.621 | 0.734 | 0.146 | 0.117 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 428 | BRPLUS-NB | 0.271 | 0.663 | 0.712 | 0.529 | 0.663 | 0.162 | 0.114 |
| 429 | BRPLUS-XGB | 0.22 | 0.744 | 0.741 | 0.589 | 0.718 | 0.145 | 0.129 |
| 430 | BRPLUS-CART | 0.239 | 0.738 | 0.663 | 0.54 | 0.676 | 0.16 | 0.074 |
| 431 | BRPLUS-MAJORITY | 0.354 | 0.643 | 0.329 | 0.279 | 0.424 | 0.38 | 0.317 |
| 432 | BRPLUS-RANDOM | 0.482 | 0.427 | 0.5 | 0.296 | 0.44 | 0.474 | 0.581 |
| 433 | BR-RF | 0.168 | 0.824 | 0.78 | 0.679 | 0.787 | 0.084 | 0.057 |
| 434 | BR-SVM | 0.211 | 0.79 | 0.696 | 0.588 | 0.722 | 0.13 | 0.014 |
| 435 | BR-SMO | 0.26 | 0.732 | 0.636 | 0.512 | 0.662 | 0.267 | 0.186 |
| 436 | BR-C5.0 | 0.181 | 0.791 | 0.787 | 0.653 | 0.774 | 0.108 | 0.071 |
| 437 | BR-NB | 0.233 | 0.72 | 0.76 | 0.592 | 0.72 | 0.149 | 0.071 |
| 438 | BR-XGB | 0.207 | 0.761 | 0.782 | 0.631 | 0.752 | 0.121 | 0.1 |
| 439 | BR-CART | 0.236 | 0.73 | 0.738 | 0.575 | 0.712 | 0.144 | 0.031 |
| 440 | BR-MAJORITY | 0.36 | 0.643 | 0.342 | 0.289 | 0.436 | 0.376 | 0.317 |
| 441 | BR-RANDOM | 0.503 | 0.422 | 0.511 | 0.306 | 0.448 | 0.5 | 0.612 |
| 442 | ECC-RF | 0.202 | 0.766 | 0.78 | 0.64 | 0.755 | 0.132 | 0.086 |
| 443 | ECC-SVM | 0.251 | 0.705 | 0.734 | 0.57 | 0.703 | 0.185 | 0.102 |
| 444 | ECC-SMO | 0.269 | 0.686 | 0.708 | 0.541 | 0.678 | 0.215 | 0.162 |
| 445 | ECC-C5.0 | 0.226 | 0.739 | 0.766 | 0.604 | 0.732 | 0.136 | 0.086 |
| 446 | ECC-NB | 0.265 | 0.703 | 0.702 | 0.544 | 0.673 | 0.209 | 0.202 |
| 447 | ECC-XGB | 0.234 | 0.744 | 0.718 | 0.579 | 0.712 | 0.151 | 0.114 |
| 448 | ECC-CART | 0.252 | 0.709 | 0.718 | 0.558 | 0.693 | 0.155 | 0.057 |
| 449 | ECC-MAJORITY | 0.415 | 0.51 | 0.501 | 0.332 | 0.487 | 0.38 | 0.317 |
| 450 | ECC-RANDOM | 0.488 | 0.422 | 0.4 | 0.261 | 0.393 | 0.513 | 0.562 |
| 451 | LP-RF | 0.146 | 0.84 | 0.831 | 0.744 | 0.82 | 0.15 | 0.143 |
| 452 | LP-SVM | 0.32 | 0.636 | 0.674 | 0.498 | 0.636 | 0.335 | 0.39 |
| 453 | LP-SMO | 0.326 | 0.622 | 0.671 | 0.479 | 0.621 | 0.356 | 0.374 |
| 454 | LP-C5.0 | 0.093 | 0.894 | 0.887 | 0.824 | 0.881 | 0.099 | 0.114 |
| 455 | LP-NB | 0.142 | 0.867 | 0.821 | 0.754 | 0.832 | 0.156 | 0.114 |
| 456 | LP-XGB | 0.265 | 0.695 | 0.741 | 0.581 | 0.696 | 0.275 | 0.343 |
| 457 | LP-CART | 0.333 | 0.63 | 0.592 | 0.446 | 0.581 | 0.353 | 0.286 |
| 458 | LP-MAJORITY | 0.487 | 0.408 | 0.374 | 0.293 | 0.38 | 0.534 | 0.624 |
| 459 | LP-RANDOM | 0.438 | 0.47 | 0.517 | 0.347 | 0.476 | 0.473 | 0.65 |
| 460 | HOMER-RF | 0.244 | 0.725 | 0.681 | 0.543 | 0.682 | 0.257 | 0.086 |
| 461 | HOMER-SVM | 0.28 | 0.669 | 0.674 | 0.503 | 0.649 | 0.331 | 0.114 |
| 462 | HOMER-SMO | 0.327 | 0.607 | 0.7 | 0.475 | 0.623 | 0.346 | 0.39 |
| 463 | HOMER-C5.0 | 0.233 | 0.743 | 0.659 | 0.548 | 0.682 | 0.27 | 0.133 |
| 464 | HOMER-NB | 0.299 | 0.675 | 0.602 | 0.472 | 0.611 | 0.352 | 0.214 |
| 465 | HOMER-XGB | 0.284 | 0.677 | 0.627 | 0.485 | 0.628 | 0.315 | 0.133 |
| 466 | HOMER-CART | 0.289 | 0.681 | 0.62 | 0.483 | 0.627 | 0.317 | 0.148 |
| 467 | HOMER-MAJORITY | 0.529 | 0.409 | 0.618 | 0.343 | 0.478 | 0.56 | 0.619 |
| 468 | HOMER-RANDOM | 0.442 | 0.461 | 0.168 | 0.143 | 0.228 | 0.512 | 0.512 |
| 469 | RAKEL-RF | 0.167 | 0.819 | 0.805 | 0.696 | 0.795 | 0.132 | 0.114 |
| 470 | RAKEL-SVM | 0.231 | 0.743 | 0.718 | 0.584 | 0.711 | 0.175 | 0.114 |
| 471 | RAKEL-SMO | 0.263 | 0.727 | 0.654 | 0.524 | 0.667 | 0.194 | 0.143 |
| 472 | RAKEL-C5.0 | 0.134 | 0.827 | 0.858 | 0.746 | 0.832 | 0.084 | 0.1 |
| 473 | RAKEL-NB | 0.214 | 0.726 | 0.81 | 0.634 | 0.751 | 0.136 | 0.129 |
| 474 | RAKEL-XGB | 0.2 | 0.762 | 0.801 | 0.657 | 0.766 | 0.136 | 0.1 |
| 475 | RAKEL-CART | 0.222 | 0.747 | 0.766 | 0.611 | 0.738 | 0.167 | 0.129 |
| 476 | RAKEL-MAJORITY | 0.385 | 0.57 | 0.387 | 0.3 | 0.448 | 0.407 | 0.479 |
| 477 | RAKEL-RANDOM | 0.509 | 0.426 | 0.625 | 0.34 | 0.488 | 0.489 | 0.571 |
| 478 | BRPLUS-RF | 0.177 | 0.799 | 0.786 | 0.68 | 0.778 | 0.1 | 0.057 |
| 479 | BRPLUS-SVM | 0.224 | 0.752 | 0.707 | 0.579 | 0.708 | 0.141 | 0.043 |
| 480 | BRPLUS-SMO | 0.264 | 0.7 | 0.692 | 0.534 | 0.672 | 0.274 | 0.229 |
| 481 | BRPLUS-C5.0 | 0.21 | 0.751 | 0.785 | 0.638 | 0.749 | 0.138 | 0.131 |
| 482 | BRPLUS-NB | 0.268 | 0.68 | 0.72 | 0.549 | 0.677 | 0.158 | 0.1 |
| 483 | BRPLUS-XGB | 0.218 | 0.75 | 0.754 | 0.604 | 0.73 | 0.14 | 0.129 |
| 484 | BRPLUS-CART | 0.239 | 0.74 | 0.675 | 0.552 | 0.686 | 0.152 | 0.074 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 485 | BRPLUS-MAJORITY | 0.36 | 0.643 | 0.342 | 0.289 | 0.436 | 0.376 | 0.317 |
| 486 | BRPLUS-RANDOM | 0.533 | 0.392 | 0.472 | 0.267 | 0.408 | 0.53 | 0.693 |
| 487 | BR-RF | 0.194 | 0.805 | 0.748 | 0.646 | 0.752 | 0.121 | 0.057 |
| 488 | BR-SVM | 0.236 | 0.789 | 0.649 | 0.571 | 0.688 | 0.174 | 0.06 |
| 489 | BR-SMO | 0.282 | 0.732 | 0.606 | 0.51 | 0.638 | 0.308 | 0.2 |
| 490 | BR-C5.0 | 0.206 | 0.794 | 0.751 | 0.626 | 0.743 | 0.135 | 0.043 |
| 491 | BR-NB | 0.249 | 0.734 | 0.745 | 0.6 | 0.713 | 0.192 | 0.1 |
| 492 | BR-XGB | 0.234 | 0.749 | 0.758 | 0.616 | 0.725 | 0.159 | 0.06 |
| 493 | BR-CART | 0.246 | 0.736 | 0.744 | 0.582 | 0.71 | 0.168 | 0.043 |
| 494 | BR-MAJORITY | 0.391 | 0.623 | 0.327 | 0.299 | 0.418 | 0.407 | 0.376 |
| 495 | BR-RANDOM | 0.467 | 0.468 | 0.468 | 0.317 | 0.449 | 0.473 | 0.452 |
| 496 | ECC-RF | 0.222 | 0.748 | 0.762 | 0.627 | 0.733 | 0.16 | 0.1 |
| 497 | ECC-SVM | 0.306 | 0.655 | 0.69 | 0.524 | 0.65 | 0.239 | 0.131 |
| 498 | ECC-SMO | 0.34 | 0.644 | 0.647 | 0.501 | 0.619 | 0.315 | 0.274 |
| 499 | ECC-C5.0 | 0.228 | 0.758 | 0.781 | 0.617 | 0.735 | 0.159 | 0.071 |
| 500 | ECC-NB | 0.277 | 0.699 | 0.705 | 0.57 | 0.683 | 0.231 | 0.145 |
| 501 | ECC-XGB | 0.258 | 0.718 | 0.748 | 0.578 | 0.701 | 0.193 | 0.071 |
| 502 | ECC-CART | 0.277 | 0.704 | 0.711 | 0.547 | 0.672 | 0.186 | 0.086 |
| 503 | ECC-MAJORITY | 0.462 | 0.5 | 0.449 | 0.322 | 0.454 | 0.441 | 0.407 |
| 504 | ECC-RANDOM | 0.479 | 0.485 | 0.385 | 0.268 | 0.391 | 0.492 | 0.519 |
| 505 | LP-RF | 0.174 | 0.819 | 0.807 | 0.716 | 0.793 | 0.19 | 0.117 |
| 506 | LP-SVM | 0.307 | 0.681 | 0.685 | 0.543 | 0.659 | 0.325 | 0.171 |
| 507 | LP-SMO | 0.338 | 0.633 | 0.65 | 0.499 | 0.617 | 0.389 | 0.302 |
| 508 | LP-C5.0 | 0.152 | 0.811 | 0.823 | 0.732 | 0.803 | 0.182 | 0.143 |
| 509 | LP-NB | 0.171 | 0.85 | 0.791 | 0.715 | 0.796 | 0.203 | 0.143 |
| 510 | LP-XGB | 0.281 | 0.708 | 0.727 | 0.598 | 0.697 | 0.291 | 0.245 |
| 511 | LP-CART | 0.34 | 0.651 | 0.655 | 0.51 | 0.623 | 0.35 | 0.362 |
| 512 | LP-MAJORITY | 0.439 | 0.514 | 0.539 | 0.409 | 0.512 | 0.481 | 0.607 |
| 513 | LP-RANDOM | 0.484 | 0.437 | 0.367 | 0.26 | 0.365 | 0.555 | 0.624 |
| 514 | HOMER-RF | 0.206 | 0.761 | 0.822 | 0.669 | 0.77 | 0.127 | 0.071 |
| 515 | HOMER-SVM | 0.289 | 0.654 | 0.813 | 0.57 | 0.692 | 0.19 | 0.145 |
| 516 | HOMER-SMO | 0.348 | 0.602 | 0.827 | 0.527 | 0.66 | 0.348 | 0.362 |
| 517 | HOMER-C5.0 | 0.176 | 0.798 | 0.848 | 0.703 | 0.797 | 0.122 | 0.088 |
| 518 | HOMER-NB | 0.271 | 0.714 | 0.758 | 0.581 | 0.699 | 0.189 | 0.1 |
| 519 | HOMER-XGB | 0.247 | 0.729 | 0.785 | 0.608 | 0.728 | 0.173 | 0.071 |
| 520 | HOMER-CART | 0.246 | 0.733 | 0.78 | 0.607 | 0.727 | 0.18 | 0.186 |
| 521 | HOMER-MAJORITY | 0.546 | 0.449 | 0.898 | 0.432 | 0.577 | 0.571 | 0.562 |
| 522 | HOMER-RANDOM | 0.474 | 0.446 | 0.399 | 0.278 | 0.384 | 0.506 | 0.624 |
| 523 | RAKEL-RF | 0.198 | 0.795 | 0.773 | 0.666 | 0.76 | 0.153 | 0.071 |
| 524 | RAKEL-SVM | 0.258 | 0.743 | 0.664 | 0.562 | 0.675 | 0.209 | 0.086 |
| 525 | RAKEL-SMO | 0.301 | 0.716 | 0.675 | 0.53 | 0.656 | 0.237 | 0.086 |
| 526 | RAKEL-C5.0 | 0.147 | 0.835 | 0.829 | 0.724 | 0.811 | 0.106 | 0.06 |
| 527 | RAKEL-NB | 0.223 | 0.745 | 0.804 | 0.64 | 0.749 | 0.158 | 0.086 |
| 528 | RAKEL-XGB | 0.199 | 0.785 | 0.803 | 0.667 | 0.765 | 0.15 | 0.1 |
| 529 | RAKEL-CART | 0.247 | 0.725 | 0.778 | 0.613 | 0.723 | 0.206 | 0.188 |
| 530 | RAKEL-MAJORITY | 0.444 | 0.518 | 0.522 | 0.395 | 0.504 | 0.456 | 0.526 |
| 531 | RAKEL-RANDOM | 0.514 | 0.446 | 0.573 | 0.332 | 0.474 | 0.525 | 0.59 |
| 532 | BRPLUS-RF | 0.203 | 0.783 | 0.769 | 0.657 | 0.754 | 0.124 | 0.029 |
| 533 | BRPLUS-SVM | 0.249 | 0.738 | 0.685 | 0.58 | 0.691 | 0.176 | 0.029 |
| 534 | BRPLUS-SMO | 0.302 | 0.664 | 0.694 | 0.536 | 0.657 | 0.325 | 0.305 |
| 535 | BRPLUS-C5.0 | 0.27 | 0.707 | 0.707 | 0.572 | 0.683 | 0.193 | 0.1 |
| 536 | BRPLUS-NB | 0.284 | 0.693 | 0.717 | 0.573 | 0.683 | 0.215 | 0.088 |
| 537 | BRPLUS-XGB | 0.304 | 0.687 | 0.673 | 0.521 | 0.647 | 0.2 | 0.06 |
| 538 | BRPLUS-CART | 0.285 | 0.7 | 0.68 | 0.536 | 0.663 | 0.194 | 0.1 |
| 539 | BRPLUS-MAJORITY | 0.391 | 0.623 | 0.327 | 0.299 | 0.418 | 0.407 | 0.376 |
| 540 | BRPLUS-RANDOM | 0.484 | 0.49 | 0.488 | 0.319 | 0.454 | 0.457 | 0.407 |
| 541 | BR-RF | 0.202 | 0.696 | 0.663 | 0.605 | 0.654 | 0.014 | 0.245 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 542 | BR-SVM | 0.193 | 0.74 | 0.646 | 0.617 | 0.666 | 0.021 | 0.245 |
| 543 | BR-SMO | 0.27 | 0.676 | 0.535 | 0.514 | 0.571 | 0.179 | 0.345 |
| 544 | BR-C5.0 | 0.202 | 0.704 | 0.646 | 0.595 | 0.647 | 0.021 | 0.245 |
| 545 | BR-NB | 0.249 | 0.652 | 0.637 | 0.568 | 0.619 | 0.036 | 0.274 |
| 546 | BR-XGB | 0.276 | 0.635 | 0.615 | 0.525 | 0.589 | 0.03 | 0.276 |
| 547 | BR-CART | 0.257 | 0.627 | 0.636 | 0.54 | 0.601 | 0.037 | 0.276 |
| 548 | BR-MAJORITY | 0.457 | 0.438 | 0.292 | 0.292 | 0.336 | 0.33 | 0.562 |
| 549 | BR-RANDOM | 0.555 | 0.351 | 0.462 | 0.282 | 0.366 | 0.419 | 0.667 |
| 550 | ECC-RF | 0.222 | 0.732 | 0.627 | 0.59 | 0.647 | 0.043 | 0.26 |
| 551 | ECC-SVM | 0.285 | 0.638 | 0.553 | 0.5 | 0.562 | 0.087 | 0.333 |
| 552 | ECC-SMO | 0.337 | 0.61 | 0.49 | 0.469 | 0.521 | 0.173 | 0.388 |
| 553 | ECC-C5.0 | 0.245 | 0.683 | 0.605 | 0.562 | 0.614 | 0.064 | 0.302 |
| 554 | ECC-NB | 0.279 | 0.651 | 0.567 | 0.511 | 0.575 | 0.071 | 0.302 |
| 555 | ECC-XGB | 0.275 | 0.661 | 0.588 | 0.537 | 0.594 | 0.043 | 0.288 |
| 556 | ECC-CART | 0.251 | 0.696 | 0.61 | 0.571 | 0.625 | 0.05 | 0.288 |
| 557 | ECC-MAJORITY | 0.457 | 0.438 | 0.292 | 0.292 | 0.336 | 0.339 | 0.562 |
| 558 | ECC-RANDOM | 0.525 | 0.394 | 0.352 | 0.299 | 0.349 | 0.361 | 0.576 |
| 559 | LP-RF | 0.193 | 0.689 | 0.668 | 0.617 | 0.656 | 0.064 | 0.288 |
| 560 | LP-SVM | 0.259 | 0.669 | 0.532 | 0.518 | 0.57 | 0.164 | 0.317 |
| 561 | LP-SMO | 0.26 | 0.683 | 0.538 | 0.523 | 0.577 | 0.15 | 0.302 |
| 562 | LP-C5.0 | 0.168 | 0.689 | 0.686 | 0.649 | 0.673 | 0.057 | 0.274 |
| 563 | LP-NB | 0.206 | 0.689 | 0.671 | 0.635 | 0.663 | 0.057 | 0.274 |
| 564 | LP-XGB | 0.245 | 0.675 | 0.63 | 0.579 | 0.626 | 0.107 | 0.331 |
| 565 | LP-CART | 0.276 | 0.623 | 0.567 | 0.509 | 0.567 | 0.137 | 0.319 |
| 566 | LP-MAJORITY | 0.457 | 0.438 | 0.292 | 0.292 | 0.336 | 0.33 | 0.562 |
| 567 | LP-RANDOM | 0.471 | 0.427 | 0.517 | 0.352 | 0.438 | 0.314 | 0.588 |
| 568 | HOMER-RF | 0.226 | 0.687 | 0.677 | 0.61 | 0.657 | 0.014 | 0.245 |
| 569 | HOMER-SVM | 0.292 | 0.664 | 0.665 | 0.565 | 0.634 | 0.03 | 0.245 |
| 570 | HOMER-SMO | 0.352 | 0.586 | 0.625 | 0.49 | 0.564 | 0.186 | 0.402 |
| 571 | HOMER-C5.0 | 0.23 | 0.675 | 0.645 | 0.594 | 0.637 | 0.05 | 0.302 |
| 572 | HOMER-NB | 0.25 | 0.661 | 0.64 | 0.56 | 0.622 | 0.036 | 0.274 |
| 573 | HOMER-XGB | 0.225 | 0.668 | 0.667 | 0.594 | 0.643 | 0.021 | 0.26 |
| 574 | HOMER-CART | 0.256 | 0.639 | 0.628 | 0.558 | 0.609 | 0.045 | 0.307 |
| 575 | HOMER-MAJORITY | 0.632 | 0.373 | 0.519 | 0.302 | 0.397 | 0.368 | 0.562 |
| 576 | HOMER-RANDOM | 0.555 | 0.351 | 0.462 | 0.282 | 0.366 | 0.419 | 0.667 |
| 577 | RAKEL-RF | 0.131 | 0.654 | 0.637 | 0.593 | 0.627 | 0.071 | 0.288 |
| 578 | RAKEL-SVM | 0.251 | 0.552 | 0.438 | 0.409 | 0.466 | 0.179 | 0.317 |
| 579 | RAKEL-SMO | 0.246 | 0.538 | 0.407 | 0.392 | 0.441 | 0.15 | 0.302 |
| 580 | RAKEL-C5.0 | 0.106 | 0.646 | 0.652 | 0.615 | 0.637 | 0.057 | 0.274 |
| 581 | RAKEL-NB | 0.174 | 0.644 | 0.626 | 0.589 | 0.618 | 0.057 | 0.274 |
| 582 | RAKEL-XGB | 0.187 | 0.632 | 0.606 | 0.555 | 0.597 | 0.107 | 0.331 |
| 583 | RAKEL-CART | 0.223 | 0.58 | 0.524 | 0.466 | 0.524 | 0.137 | 0.319 |
| 584 | RAKEL-MAJORITY | 0.416 | 0 | 0 | 0 | 0 | 0.33 | 0.562 |
| 585 | RAKEL-RANDOM | 0.437 | 0.413 | 0.51 | 0.344 | 0.428 | 0.314 | 0.588 |
| 586 | BRPLUS-RF | 0.207 | 0.696 | 0.668 | 0.61 | 0.657 | 0.014 | 0.245 |
| 587 | BRPLUS-SVM | 0.198 | 0.733 | 0.631 | 0.61 | 0.657 | 0.037 | 0.26 |
| 588 | BRPLUS-SMO | 0.29 | 0.648 | 0.552 | 0.488 | 0.556 | 0.15 | 0.317 |
| 589 | BRPLUS-C5.0 | 0.192 | 0.682 | 0.645 | 0.615 | 0.646 | 0.043 | 0.288 |
| 590 | BRPLUS-NB | 0.275 | 0.641 | 0.652 | 0.557 | 0.614 | 0.03 | 0.26 |
| 591 | BRPLUS-XGB | 0.276 | 0.635 | 0.615 | 0.525 | 0.589 | 0.03 | 0.276 |
| 592 | BRPLUS-CART | 0.257 | 0.627 | 0.636 | 0.54 | 0.601 | 0.037 | 0.276 |
| 593 | BRPLUS-MAJORITY | 0.457 | 0.438 | 0.292 | 0.292 | 0.336 | 0.33 | 0.562 |
| 594 | BRPLUS-RANDOM | 0.48 | 0.421 | 0.539 | 0.346 | 0.434 | 0.314 | 0.552 |

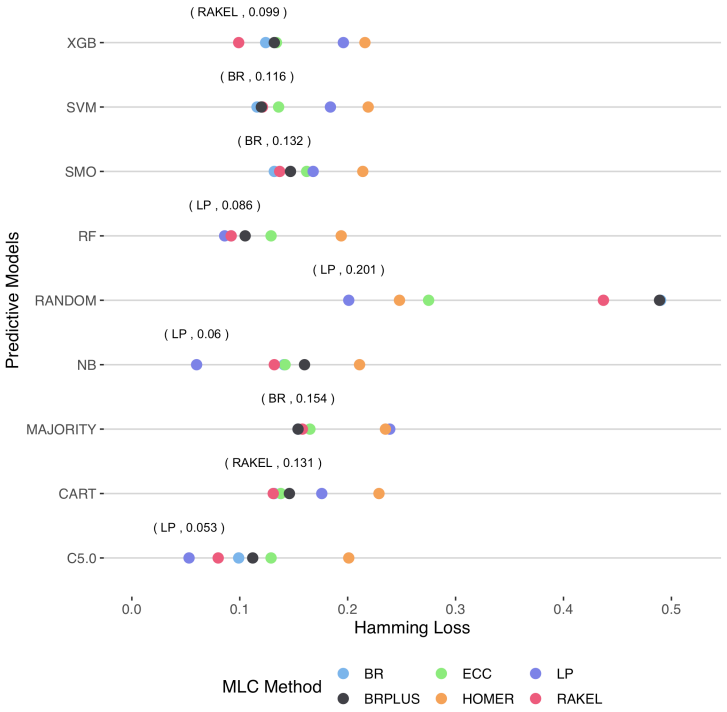



**Figure H. 1.** Hamming-Loss for all MLC strategies applied in CDS1.

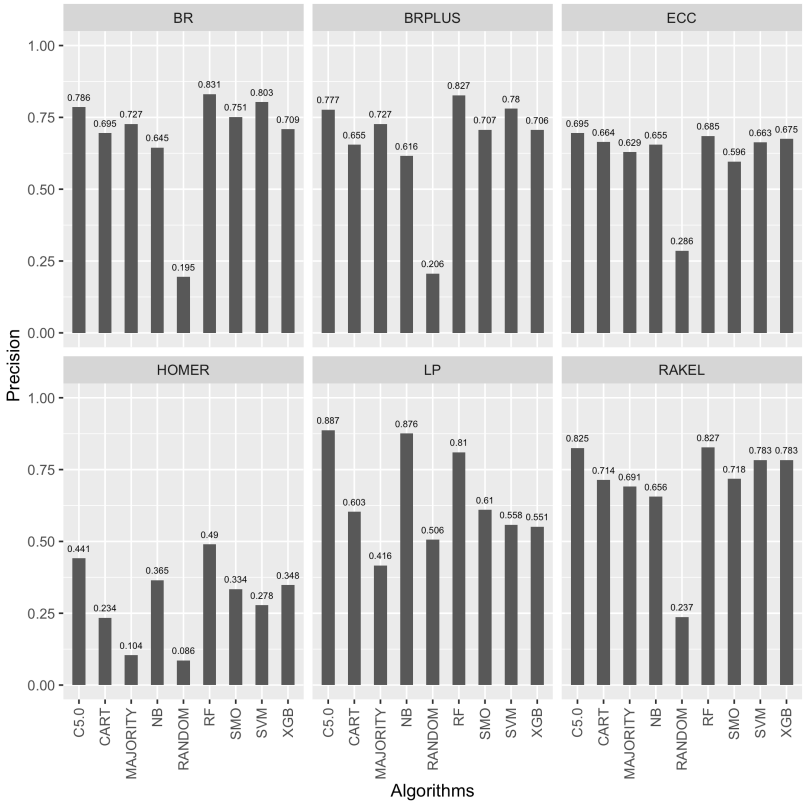



**Figure H. 2.** Precision for all MLC strategies applied in CDS1.

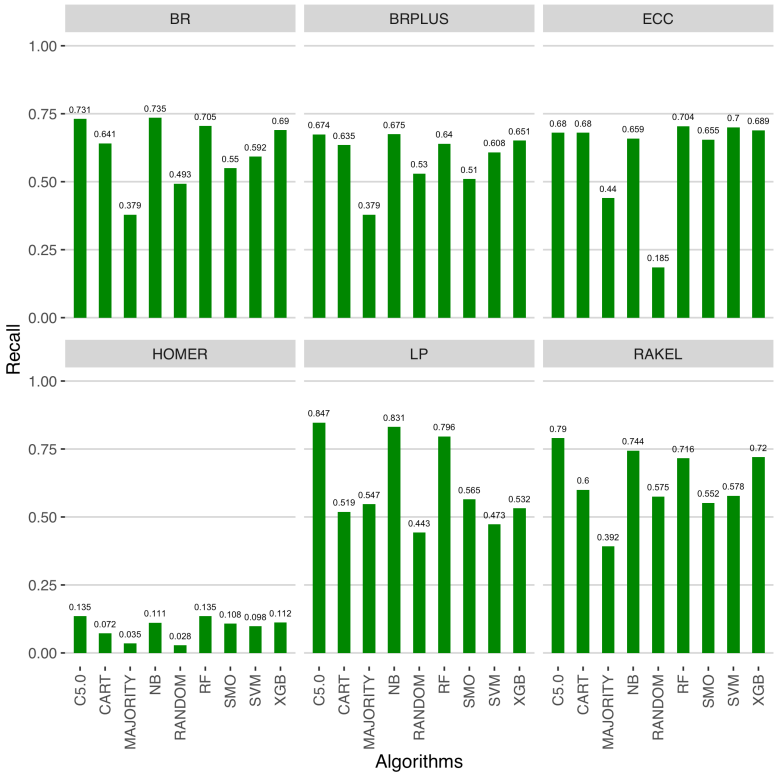



**Figure H. 3.** Recall for all MLC strategies applied in CDS1.

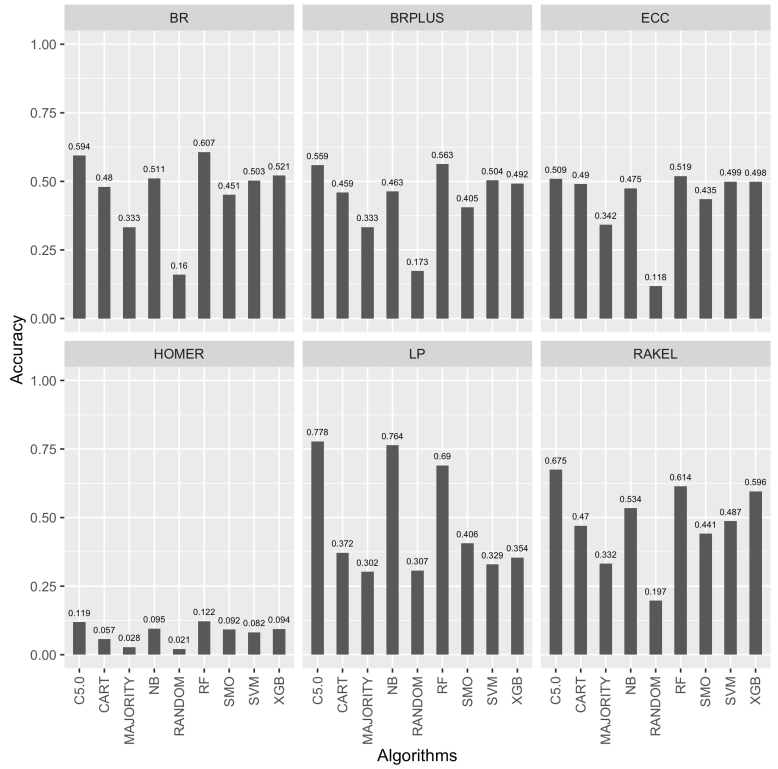



**Figure H. 4.** Accuracy for all MLC strategies applied in CDS1.

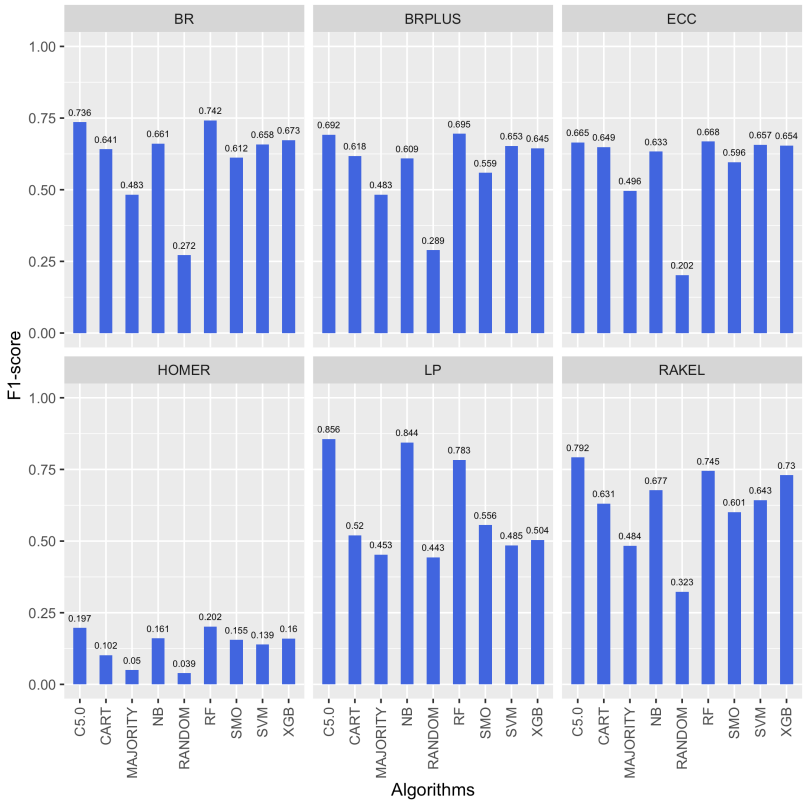



**Figure H. 5.** F1-Score for all MLC strategies applied in CDS1.

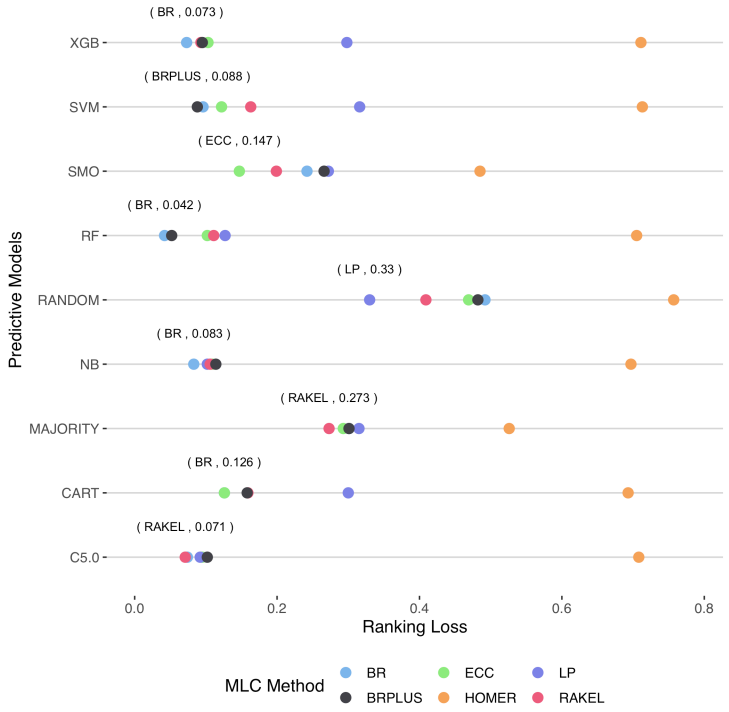



**Figure H. 6.** Ranking-Loss for all MLC strategies applied in CDS1.

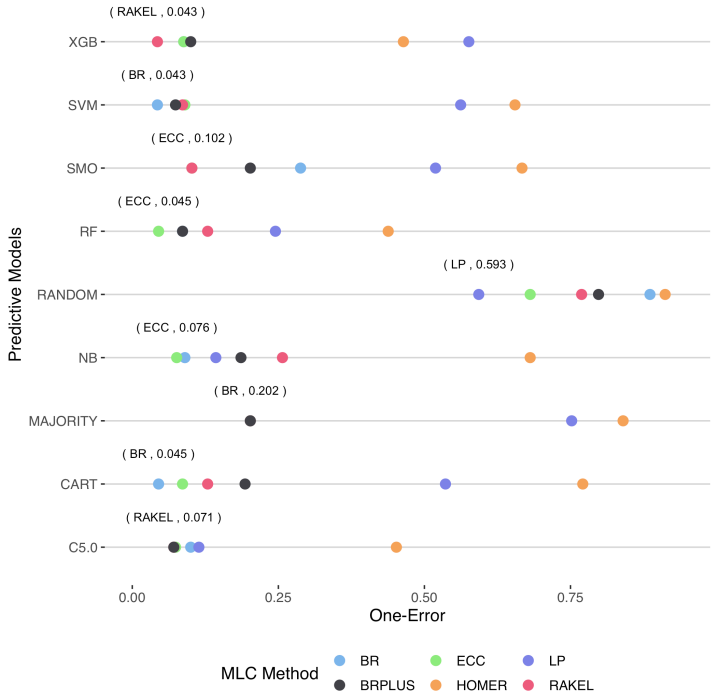



**Figure H. 7.** One-Error for all MLC strategies applied in CDS1.

# H.2. Performance Measures for Models (CDS2)

| | mlc_methods | values_hl | values_pr | values_re | values_ac | values_f1 | values_rl | values_oe |
|---|---|---|---|---|---|---|---|---|
| 1 | BR-RF | 0.084 | 0.848 | 0.69 | 0.601 | 0.718 | 0.053 | 0.025 |
| 2 | BR-SVM | 0.089 | 0.811 | 0.692 | 0.584 | 0.702 | 0.069 | 0.075 |
| 3 | BR-SMO | 0.089 | 0.773 | 0.756 | 0.637 | 0.741 | 0.151 | 0.225 |
| 4 | BR-C5.0 | 0.088 | 0.757 | 0.781 | 0.611 | 0.73 | 0.074 | 0.062 |
| 5 | BR-NB | 0.275 | 0.303 | 0.763 | 0.268 | 0.403 | 0.258 | 1 |
| 6 | BR-XGB | 0.096 | 0.762 | 0.737 | 0.593 | 0.709 | 0.072 | 0.025 |
| 7 | BR-CART | 0.103 | 0.705 | 0.75 | 0.556 | 0.682 | 0.079 | 0.075 |
| 8 | BR-MAJORITY | 0.14 | 0.674 | 0.361 | 0.249 | 0.388 | 0.285 | 0.288 |
| 9 | BR-RANDOM | 0.494 | 0.168 | 0.497 | 0.139 | 0.233 | 0.476 | 0.788 |
| 10 | ECC-RF | 0.101 | 0.684 | 0.749 | 0.532 | 0.656 | 0.115 | 0.1 |
| 11 | ECC-SVM | 0.098 | 0.698 | 0.734 | 0.555 | 0.672 | 0.09 | 0.125 |
| 12 | ECC-SMO | 0.096 | 0.737 | 0.782 | 0.616 | 0.726 | 0.08 | 0.1 |
| 13 | ECC-C5.0 | 0.1 | 0.705 | 0.746 | 0.552 | 0.674 | 0.073 | 0.038 |
| 14 | ECC-NB | 0.158 | 0.48 | 0.584 | 0.336 | 0.478 | 0.17 | 0.15 |
| 15 | ECC-XGB | 0.102 | 0.701 | 0.73 | 0.537 | 0.663 | 0.08 | 0.1 |
| 16 | ECC-CART | 0.119 | 0.656 | 0.695 | 0.491 | 0.622 | 0.098 | 0.1 |
| 17 | ECC-MAJORITY | 0.14 | 0.633 | 0.403 | 0.276 | 0.416 | 0.271 | 0.288 |
| 18 | ECC-RANDOM | 0.248 | 0.231 | 0.181 | 0.097 | 0.167 | 0.405 | 0.725 |
| 19 | LP-RF | 0.098 | 0.769 | 0.729 | 0.639 | 0.72 | 0.166 | 0.275 |
| 20 | LP-SVM | 0.173 | 0.557 | 0.489 | 0.385 | 0.465 | 0.29 | 0.575 |
| 21 | LP-SMO | 0.088 | 0.785 | 0.785 | 0.695 | 0.772 | 0.138 | 0.25 |
| 22 | LP-C5.0 | 0.088 | 0.794 | 0.765 | 0.672 | 0.752 | 0.146 | 0.225 |
| 23 | LP-NB | 0.135 | 0.564 | 0.686 | 0.452 | 0.551 | 0.197 | 0.475 |
| 24 | LP-XGB | 0.141 | 0.695 | 0.565 | 0.473 | 0.562 | 0.253 | 0.325 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 25 | LP-CART | 0.157 | 0.664 | 0.598 | 0.493 | 0.594 | 0.242 | 0.375 |
| 26 | LP-MAJORITY | 0.166 | 0.608 | 0.197 | 0.144 | 0.208 | 0.361 | 0.5 |
| 27 | LP-RANDOM | 0.212 | 0.37 | 0.361 | 0.216 | 0.316 | 0.386 | 0.75 |
| 28 | HOMER-RF | 0.167 | 0.474 | 0.192 | 0.172 | 0.253 | 0.679 | 0.488 |
| 29 | HOMER-SVM | 0.176 | 0.42 | 0.177 | 0.151 | 0.225 | 0.683 | 0.562 |
| 30 | HOMER-SMO | 0.177 | 0.414 | 0.173 | 0.154 | 0.226 | 0.461 | 0.588 |
| 31 | HOMER-C5.0 | 0.174 | 0.432 | 0.171 | 0.149 | 0.225 | 0.693 | 0.538 |
| 32 | HOMER-NB | 0.18 | 0.339 | 0.142 | 0.124 | 0.184 | 0.639 | 0.712 |
| 33 | HOMER-XGB | 0.18 | 0.425 | 0.167 | 0.148 | 0.223 | 0.682 | 0.575 |
| 34 | HOMER-CART | 0.191 | 0.325 | 0.124 | 0.103 | 0.165 | 0.664 | 0.662 |
| 35 | HOMER-MAJORITY | 0.216 | 0.155 | 0.07 | 0.052 | 0.091 | 0.513 | 0.9 |
| 36 | HOMER-RANDOM | 0.225 | 0.063 | 0.023 | 0.018 | 0.032 | 0.69 | 0.938 |
| 37 | RAKEL-RF | 0.078 | 0.839 | 0.727 | 0.633 | 0.742 | 0.113 | 0.112 |
| 38 | RAKEL-SVM | 0.091 | 0.813 | 0.675 | 0.571 | 0.691 | 0.147 | 0.138 |
| 39 | RAKEL-SMO | 0.086 | 0.792 | 0.765 | 0.66 | 0.755 | 0.103 | 0.15 |
| 40 | RAKEL-C5.0 | 0.076 | 0.808 | 0.817 | 0.684 | 0.786 | 0.069 | 0.088 |
| 41 | RAKEL-NB | 0.381 | 0.238 | 0.702 | 0.211 | 0.331 | 0.297 | 0.762 |
| 42 | RAKEL-XGB | 0.087 | 0.8 | 0.753 | 0.632 | 0.738 | 0.092 | 0.125 |
| 43 | RAKEL-CART | 0.102 | 0.715 | 0.719 | 0.542 | 0.671 | 0.105 | 0.125 |
| 44 | RAKEL-MAJORITY | 0.14 | 0.674 | 0.361 | 0.249 | 0.388 | 0.273 | 0.338 |
| 45 | RAKEL-RANDOM | 0.371 | 0.259 | 0.609 | 0.204 | 0.317 | 0.32 | 0.862 |
| 46 | BRPLUS-RF | 0.085 | 0.873 | 0.653 | 0.595 | 0.703 | 0.053 | 0.012 |
| 47 | BRPLUS-SVM | 0.091 | 0.809 | 0.72 | 0.609 | 0.718 | 0.064 | 0.05 |
| 48 | BRPLUS-SMO | 0.088 | 0.797 | 0.764 | 0.664 | 0.755 | 0.148 | 0.238 |
| 49 | BRPLUS-C5.0 | 0.098 | 0.743 | 0.767 | 0.597 | 0.707 | 0.094 | 0.1 |
| 50 | BRPLUS-NB | 0.263 | 0.314 | 0.685 | 0.272 | 0.404 | 0.241 | 0.962 |
| 51 | BRPLUS-XGB | 0.092 | 0.804 | 0.75 | 0.63 | 0.732 | 0.074 | 0.012 |
| 52 | BRPLUS-CART | 0.109 | 0.696 | 0.749 | 0.568 | 0.679 | 0.099 | 0.1 |
| 53 | BRPLUS-MAJORITY | 0.14 | 0.674 | 0.361 | 0.249 | 0.388 | 0.285 | 0.288 |
| 54 | BRPLUS-RANDOM | 0.497 | 0.158 | 0.446 | 0.128 | 0.218 | 0.526 | 0.912 |
| 55 | BR-RF | 0.139 | 0.845 | 0.749 | 0.643 | 0.754 | 0.064 | 0.025 |
| 56 | BR-SVM | 0.148 | 0.817 | 0.747 | 0.63 | 0.74 | 0.081 | 0.075 |
| 57 | BR-SMO | 0.144 | 0.798 | 0.793 | 0.669 | 0.771 | 0.159 | 0.188 |
| 58 | BR-C5.0 | 0.146 | 0.772 | 0.818 | 0.645 | 0.757 | 0.091 | 0.062 |
| 59 | BR-NB | 0.183 | 0.727 | 0.798 | 0.606 | 0.72 | 0.105 | 0.138 |
| 60 | BR-XGB | 0.16 | 0.778 | 0.775 | 0.622 | 0.737 | 0.097 | 0.025 |
| 61 | BR-CART | 0.177 | 0.714 | 0.793 | 0.586 | 0.709 | 0.102 | 0.075 |
| 62 | BR-MAJORITY | 0.244 | 0.674 | 0.469 | 0.292 | 0.439 | 0.283 | 0.288 |
| 63 | BR-RANDOM | 0.476 | 0.295 | 0.363 | 0.198 | 0.306 | 0.519 | 0.712 |
| 64 | ECC-RF | 0.168 | 0.748 | 0.796 | 0.614 | 0.723 | 0.112 | 0.075 |
| 65 | ECC-SVM | 0.158 | 0.759 | 0.787 | 0.623 | 0.725 | 0.102 | 0.112 |
| 66 | ECC-SMO | 0.156 | 0.788 | 0.784 | 0.658 | 0.756 | 0.103 | 0.062 |
| 67 | ECC-C5.0 | 0.171 | 0.75 | 0.781 | 0.608 | 0.72 | 0.107 | 0.05 |
| 68 | ECC-NB | 0.171 | 0.747 | 0.779 | 0.613 | 0.719 | 0.108 | 0.1 |
| 69 | ECC-XGB | 0.187 | 0.699 | 0.762 | 0.552 | 0.675 | 0.109 | 0.062 |
| 70 | ECC-CART | 0.193 | 0.684 | 0.745 | 0.537 | 0.662 | 0.132 | 0.075 |
| 71 | ECC-MAJORITY | 0.251 | 0.623 | 0.507 | 0.311 | 0.455 | 0.268 | 0.288 |
| 72 | ECC-RANDOM | 0.407 | 0.357 | 0.382 | 0.188 | 0.299 | 0.498 | 0.575 |
| 73 | LP-RF | 0.15 | 0.802 | 0.784 | 0.693 | 0.769 | 0.172 | 0.275 |
| 74 | LP-SVM | 0.288 | 0.647 | 0.59 | 0.462 | 0.565 | 0.32 | 0.475 |
| 75 | LP-SMO | 0.137 | 0.814 | 0.815 | 0.731 | 0.802 | 0.152 | 0.225 |
| 76 | LP-C5.0 | 0.117 | 0.858 | 0.847 | 0.756 | 0.828 | 0.126 | 0.188 |
| 77 | LP-NB | 0.24 | 0.572 | 0.724 | 0.455 | 0.564 | 0.26 | 0.462 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 78 | LP-XGB | 0.193 | 0.811 | 0.694 | 0.6 | 0.682 | 0.226 | 0.25 |
| 79 | LP-CART | 0.22 | 0.739 | 0.714 | 0.601 | 0.695 | 0.239 | 0.288 |
| 80 | LP-MAJORITY | 0.276 | 0.825 | 0.255 | 0.255 | 0.301 | 0.359 | 0.175 |
| 81 | LP-RANDOM | 0.367 | 0.388 | 0.483 | 0.241 | 0.352 | 0.446 | 0.738 |
| 82 | HOMER-RF | 0.172 | 0.728 | 0.847 | 0.651 | 0.744 | 0.146 | 0.175 |
| 83 | HOMER-SVM | 0.187 | 0.687 | 0.831 | 0.607 | 0.712 | 0.158 | 0.162 |
| 84 | HOMER-SMO | 0.162 | 0.786 | 0.794 | 0.677 | 0.768 | 0.174 | 0.238 |
| 85 | HOMER-C5.0 | 0.158 | 0.752 | 0.8 | 0.655 | 0.754 | 0.158 | 0.112 |
| 86 | HOMER-NB | 0.196 | 0.747 | 0.662 | 0.54 | 0.649 | 0.212 | 0.25 |
| 87 | HOMER-XGB | 0.174 | 0.719 | 0.793 | 0.62 | 0.727 | 0.17 | 0.15 |
| 88 | HOMER-CART | 0.2 | 0.683 | 0.75 | 0.569 | 0.679 | 0.182 | 0.162 |
| 89 | HOMER-MAJORITY | 0.373 | 0.418 | 0.643 | 0.289 | 0.421 | 0.354 | 0.762 |
| 90 | HOMER-RANDOM | 0.346 | 0.331 | 0.141 | 0.106 | 0.176 | 0.491 | 0.675 |
| 91 | RAKEL-RF | 0.137 | 0.824 | 0.779 | 0.666 | 0.768 | 0.107 | 0.088 |
| 92 | RAKEL-SVM | 0.15 | 0.827 | 0.728 | 0.617 | 0.729 | 0.142 | 0.088 |
| 93 | RAKEL-SMO | 0.139 | 0.82 | 0.806 | 0.706 | 0.796 | 0.117 | 0.112 |
| 94 | RAKEL-C5.0 | 0.124 | 0.812 | 0.852 | 0.71 | 0.806 | 0.079 | 0.062 |
| 95 | RAKEL-NB | 0.292 | 0.47 | 0.83 | 0.405 | 0.55 | 0.189 | 0.2 |
| 96 | RAKEL-XGB | 0.132 | 0.834 | 0.824 | 0.705 | 0.798 | 0.087 | 0.075 |
| 97 | RAKEL-CART | 0.147 | 0.781 | 0.806 | 0.644 | 0.752 | 0.097 | 0.062 |
| 98 | RAKEL-MAJORITY | 0.242 | 0.707 | 0.441 | 0.28 | 0.427 | 0.26 | 0.238 |
| 99 | RAKEL-RANDOM | 0.458 | 0.339 | 0.591 | 0.253 | 0.365 | 0.387 | 0.725 |
| 100 | BRPLUS-RF | 0.139 | 0.884 | 0.729 | 0.665 | 0.766 | 0.067 | 0.025 |
| 101 | BRPLUS-SVM | 0.154 | 0.816 | 0.767 | 0.637 | 0.735 | 0.075 | 0.05 |
| 102 | BRPLUS-SMO | 0.142 | 0.803 | 0.801 | 0.694 | 0.779 | 0.159 | 0.262 |
| 103 | BRPLUS-C5.0 | 0.16 | 0.768 | 0.827 | 0.651 | 0.756 | 0.108 | 0.038 |
| 104 | BRPLUS-NB | 0.177 | 0.72 | 0.755 | 0.602 | 0.698 | 0.091 | 0.1 |
| 105 | BRPLUS-XGB | 0.161 | 0.809 | 0.781 | 0.65 | 0.752 | 0.105 | 0.012 |
| 106 | BRPLUS-CART | 0.187 | 0.704 | 0.795 | 0.598 | 0.707 | 0.128 | 0.112 |
| 107 | BRPLUS-MAJORITY | 0.244 | 0.674 | 0.469 | 0.292 | 0.439 | 0.283 | 0.288 |
| 108 | BRPLUS-RANDOM | 0.475 | 0.306 | 0.46 | 0.214 | 0.33 | 0.495 | 0.65 |
| 109 | BR-RF | 0.144 | 0.866 | 0.888 | 0.775 | 0.852 | 0.067 | 0.025 |
| 110 | BR-SVM | 0.176 | 0.851 | 0.849 | 0.724 | 0.817 | 0.083 | 0.062 |
| 111 | BR-SMO | 0.165 | 0.844 | 0.867 | 0.747 | 0.834 | 0.157 | 0.088 |
| 112 | BR-C5.0 | 0.175 | 0.815 | 0.879 | 0.724 | 0.814 | 0.1 | 0.062 |
| 113 | BR-NB | 0.206 | 0.841 | 0.833 | 0.701 | 0.795 | 0.115 | 0.062 |
| 114 | BR-XGB | 0.169 | 0.846 | 0.869 | 0.742 | 0.828 | 0.078 | 0.025 |
| 115 | BR-CART | 0.195 | 0.791 | 0.867 | 0.695 | 0.792 | 0.099 | 0.05 |
| 116 | BR-MAJORITY | 0.343 | 0.674 | 0.63 | 0.426 | 0.568 | 0.292 | 0.288 |
| 117 | BR-RANDOM | 0.507 | 0.469 | 0.483 | 0.332 | 0.449 | 0.533 | 0.5 |
| 118 | ECC-RF | 0.166 | 0.834 | 0.848 | 0.734 | 0.811 | 0.093 | 0.075 |
| 119 | ECC-SVM | 0.193 | 0.825 | 0.841 | 0.702 | 0.79 | 0.106 | 0.125 |
| 120 | ECC-SMO | 0.17 | 0.85 | 0.869 | 0.755 | 0.837 | 0.102 | 0.062 |
| 121 | ECC-C5.0 | 0.176 | 0.823 | 0.847 | 0.724 | 0.811 | 0.094 | 0.062 |
| 122 | ECC-NB | 0.218 | 0.773 | 0.793 | 0.659 | 0.751 | 0.114 | 0.062 |
| 123 | ECC-XGB | 0.2 | 0.794 | 0.844 | 0.683 | 0.779 | 0.099 | 0.062 |
| 124 | ECC-CART | 0.207 | 0.785 | 0.811 | 0.657 | 0.759 | 0.11 | 0.088 |
| 125 | ECC-MAJORITY | 0.332 | 0.668 | 0.667 | 0.457 | 0.591 | 0.257 | 0.288 |
| 126 | ECC-RANDOM | 0.539 | 0.446 | 0.537 | 0.297 | 0.43 | 0.566 | 0.55 |
| 127 | LP-RF | 0.151 | 0.881 | 0.84 | 0.777 | 0.836 | 0.157 | 0.112 |
| 128 | LP-SVM | 0.242 | 0.795 | 0.761 | 0.662 | 0.739 | 0.247 | 0.2 |
| 129 | LP-SMO | 0.141 | 0.873 | 0.881 | 0.806 | 0.865 | 0.14 | 0.125 |
| 130 | LP-C5.0 | 0.131 | 0.902 | 0.891 | 0.816 | 0.87 | 0.113 | 0.062 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 131 | LP-NB | 0.36 | 0.609 | 0.781 | 0.488 | 0.605 | 0.335 | 0.412 |
| 132 | LP-XGB | 0.198 | 0.831 | 0.832 | 0.723 | 0.796 | 0.197 | 0.162 |
| 133 | LP-CART | 0.197 | 0.855 | 0.825 | 0.735 | 0.808 | 0.192 | 0.15 |
| 134 | LP-MAJORITY | 0.425 | 0.825 | 0.287 | 0.287 | 0.353 | 0.38 | 0.175 |
| 135 | LP-RANDOM | 0.427 | 0.581 | 0.604 | 0.405 | 0.513 | 0.448 | 0.5 |
| 136 | HOMER-RF | 0.16 | 0.814 | 0.954 | 0.78 | 0.849 | 0.102 | 0.088 |
| 137 | HOMER-SVM | 0.182 | 0.784 | 0.955 | 0.753 | 0.831 | 0.129 | 0.125 |
| 138 | HOMER-SMO | 0.142 | 0.865 | 0.912 | 0.805 | 0.873 | 0.131 | 0.05 |
| 139 | HOMER-C5.0 | 0.152 | 0.869 | 0.895 | 0.791 | 0.865 | 0.1 | 0.038 |
| 140 | HOMER-NB | 0.219 | 0.83 | 0.78 | 0.677 | 0.766 | 0.202 | 0.125 |
| 141 | HOMER-XGB | 0.169 | 0.831 | 0.903 | 0.754 | 0.836 | 0.11 | 0.038 |
| 142 | HOMER-CART | 0.164 | 0.851 | 0.873 | 0.748 | 0.832 | 0.118 | 0.088 |
| 143 | HOMER-MAJORITY | 0.493 | 0.495 | 0.923 | 0.462 | 0.587 | 0.543 | 0.738 |
| 144 | HOMER-RANDOM | 0.461 | 0.533 | 0.261 | 0.209 | 0.312 | 0.497 | 0.488 |
| 145 | RAKEL-RF | 0.136 | 0.879 | 0.91 | 0.807 | 0.869 | 0.088 | 0.1 |
| 146 | RAKEL-SVM | 0.18 | 0.832 | 0.869 | 0.725 | 0.811 | 0.104 | 0.062 |
| 147 | RAKEL-SMO | 0.14 | 0.887 | 0.874 | 0.802 | 0.866 | 0.102 | 0.075 |
| 148 | RAKEL-C5.0 | 0.131 | 0.875 | 0.916 | 0.809 | 0.876 | 0.081 | 0.05 |
| 149 | RAKEL-NB | 0.301 | 0.597 | 0.617 | 0.459 | 0.577 | 0.207 | 0.25 |
| 150 | RAKEL-XGB | 0.131 | 0.88 | 0.911 | 0.807 | 0.869 | 0.062 | 0.038 |
| 151 | RAKEL-CART | 0.173 | 0.82 | 0.919 | 0.759 | 0.835 | 0.083 | 0.062 |
| 152 | RAKEL-MAJORITY | 0.375 | 0.644 | 0.609 | 0.4 | 0.54 | 0.226 | 0.175 |
| 153 | RAKEL-RANDOM | 0.485 | 0.507 | 0.586 | 0.356 | 0.483 | 0.416 | 0.45 |
| 154 | BRPLUS-RF | 0.145 | 0.869 | 0.899 | 0.793 | 0.863 | 0.06 | 0.012 |
| 155 | BRPLUS-SVM | 0.178 | 0.85 | 0.857 | 0.726 | 0.81 | 0.084 | 0.012 |
| 156 | BRPLUS-SMO | 0.137 | 0.872 | 0.897 | 0.801 | 0.868 | 0.14 | 0.112 |
| 157 | BRPLUS-C5.0 | 0.158 | 0.841 | 0.91 | 0.768 | 0.843 | 0.086 | 0.05 |
| 158 | BRPLUS-NB | 0.193 | 0.851 | 0.847 | 0.726 | 0.808 | 0.081 | 0.05 |
| 159 | BRPLUS-XGB | 0.151 | 0.86 | 0.892 | 0.771 | 0.847 | 0.068 | 0.012 |
| 160 | BRPLUS-CART | 0.18 | 0.82 | 0.886 | 0.733 | 0.816 | 0.089 | 0.062 |
| 161 | BRPLUS-MAJORITY | 0.343 | 0.674 | 0.63 | 0.426 | 0.568 | 0.292 | 0.288 |
| 162 | BRPLUS-RANDOM | 0.493 | 0.49 | 0.521 | 0.341 | 0.464 | 0.497 | 0.425 |
| 163 | BR-RF | 0.13 | 0.856 | 0.909 | 0.828 | 0.867 | 0.051 | 0.075 |
| 164 | BR-SVM | 0.158 | 0.846 | 0.9 | 0.811 | 0.856 | 0.067 | 0.1 |
| 165 | BR-SMO | 0.115 | 0.866 | 0.881 | 0.814 | 0.858 | 0.057 | 0.088 |
| 166 | BR-C5.0 | 0.148 | 0.825 | 0.876 | 0.798 | 0.839 | 0.053 | 0.075 |
| 167 | BR-NB | 0.195 | 0.832 | 0.798 | 0.722 | 0.786 | 0.069 | 0.088 |
| 168 | BR-XGB | 0.158 | 0.837 | 0.857 | 0.777 | 0.828 | 0.046 | 0.075 |
| 169 | BR-CART | 0.18 | 0.828 | 0.838 | 0.745 | 0.809 | 0.046 | 0.088 |
| 170 | BR-MAJORITY | 0.358 | 0.674 | 0.883 | 0.624 | 0.711 | 0.245 | 0.288 |
| 171 | BR-RANDOM | 0.498 | 0.629 | 0.434 | 0.362 | 0.45 | 0.298 | 0.338 |
| 172 | ECC-RF | 0.2 | 0.799 | 0.805 | 0.739 | 0.782 | 0.05 | 0.088 |
| 173 | ECC-SVM | 0.195 | 0.812 | 0.811 | 0.731 | 0.786 | 0.071 | 0.088 |
| 174 | ECC-SMO | 0.15 | 0.837 | 0.846 | 0.776 | 0.826 | 0.059 | 0.075 |
| 175 | ECC-C5.0 | 0.158 | 0.834 | 0.846 | 0.779 | 0.827 | 0.056 | 0.075 |
| 176 | ECC-NB | 0.23 | 0.765 | 0.793 | 0.688 | 0.748 | 0.072 | 0.1 |
| 177 | ECC-XGB | 0.178 | 0.825 | 0.825 | 0.756 | 0.808 | 0.076 | 0.112 |
| 178 | ECC-CART | 0.215 | 0.841 | 0.803 | 0.733 | 0.798 | 0.088 | 0.088 |
| 179 | ECC-MAJORITY | 0.362 | 0.712 | 0.77 | 0.563 | 0.681 | 0.167 | 0.288 |
| 180 | ECC-RANDOM | 0.5 | 0.487 | 0.581 | 0.424 | 0.484 | 0.296 | 0.4 |
| 181 | LP-RF | 0.118 | 0.81 | 0.9 | 0.81 | 0.841 | 0.093 | 0.088 |
| 182 | LP-SVM | 0.178 | 0.775 | 0.888 | 0.775 | 0.81 | 0.128 | 0.112 |
| 183 | LP-SMO | 0.115 | 0.863 | 0.885 | 0.813 | 0.858 | 0.072 | 0.088 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 184 | LP-C5.0 | 0.152 | 0.819 | 0.89 | 0.796 | 0.838 | 0.093 | 0.088 |
| 185 | LP-NB | 0.45 | 0.748 | 0.547 | 0.431 | 0.566 | 0.152 | 0.262 |
| 186 | LP-XGB | 0.148 | 0.816 | 0.879 | 0.784 | 0.828 | 0.1 | 0.088 |
| 187 | LP-CART | 0.14 | 0.822 | 0.894 | 0.793 | 0.841 | 0.107 | 0.088 |
| 188 | LP-MAJORITY | 0.332 | 0.668 | 0.938 | 0.668 | 0.733 | 0.259 | 0.288 |
| 189 | LP-RANDOM | 0.442 | 0.618 | 0.557 | 0.444 | 0.537 | 0.233 | 0.288 |
| 190 | HOMER-RF | 0.268 | 0.823 | 0.76 | 0.698 | 0.781 | 0.127 | 0.112 |
| 191 | HOMER-SVM | 0.275 | 0.814 | 0.76 | 0.689 | 0.774 | 0.15 | 0.162 |
| 192 | HOMER-SMO | 0.265 | 0.847 | 0.743 | 0.693 | 0.777 | 0.114 | 0.125 |
| 193 | HOMER-C5.0 | 0.28 | 0.833 | 0.733 | 0.678 | 0.768 | 0.128 | 0.1 |
| 194 | HOMER-NB | 0.34 | 0.816 | 0.652 | 0.601 | 0.696 | 0.137 | 0.125 |
| 195 | HOMER-XGB | 0.285 | 0.828 | 0.737 | 0.68 | 0.769 | 0.121 | 0.088 |
| 196 | HOMER-CART | 0.3 | 0.828 | 0.72 | 0.663 | 0.756 | 0.142 | 0.1 |
| 197 | HOMER-MAJORITY | 0.418 | 0.656 | 0.76 | 0.538 | 0.653 | 0.294 | 0.338 |
| 198 | HOMER-RANDOM | 0.535 | 0.667 | 0.434 | 0.324 | 0.462 | 0.258 | 0.288 |
| 199 | RAKEL-RF | 0.138 | 0.835 | 0.912 | 0.822 | 0.857 | 0.064 | 0.088 |
| 200 | RAKEL-SVM | 0.158 | 0.834 | 0.917 | 0.814 | 0.854 | 0.076 | 0.088 |
| 201 | RAKEL-SMO | 0.11 | 0.86 | 0.895 | 0.818 | 0.862 | 0.049 | 0.075 |
| 202 | RAKEL-C5.0 | 0.138 | 0.837 | 0.897 | 0.811 | 0.851 | 0.054 | 0.1 |
| 203 | RAKEL-NB | 0.208 | 0.831 | 0.81 | 0.722 | 0.786 | 0.052 | 0.062 |
| 204 | RAKEL-XGB | 0.143 | 0.822 | 0.898 | 0.807 | 0.843 | 0.069 | 0.1 |
| 205 | RAKEL-CART | 0.162 | 0.829 | 0.874 | 0.786 | 0.831 | 0.083 | 0.1 |
| 206 | RAKEL-MAJORITY | 0.332 | 0.668 | 0.938 | 0.668 | 0.733 | 0.259 | 0.288 |
| 207 | RAKEL-RANDOM | 0.422 | 0.706 | 0.667 | 0.512 | 0.611 | 0.208 | 0.262 |
| 208 | BRPLUS-RF | 0.125 | 0.847 | 0.934 | 0.844 | 0.876 | 0.052 | 0.088 |
| 209 | BRPLUS-SVM | 0.16 | 0.832 | 0.89 | 0.797 | 0.841 | 0.061 | 0.1 |
| 210 | BRPLUS-SMO | 0.125 | 0.854 | 0.887 | 0.809 | 0.854 | 0.073 | 0.088 |
| 211 | BRPLUS-C5.0 | 0.16 | 0.825 | 0.86 | 0.781 | 0.827 | 0.058 | 0.075 |
| 212 | BRPLUS-NB | 0.19 | 0.792 | 0.801 | 0.718 | 0.771 | 0.047 | 0.112 |
| 213 | BRPLUS-XGB | 0.155 | 0.812 | 0.87 | 0.776 | 0.822 | 0.045 | 0.075 |
| 214 | BRPLUS-CART | 0.178 | 0.805 | 0.87 | 0.771 | 0.82 | 0.083 | 0.1 |
| 215 | BRPLUS-MAJORITY | 0.358 | 0.674 | 0.883 | 0.624 | 0.711 | 0.245 | 0.288 |
| 216 | BRPLUS-RANDOM | 0.478 | 0.659 | 0.487 | 0.386 | 0.485 | 0.241 | 0.312 |
| 217 | BR-RF | 0.117 | 0.842 | 0.714 | 0.615 | 0.731 | 0.054 | 0.025 |
| 218 | BR-SVM | 0.122 | 0.812 | 0.721 | 0.603 | 0.719 | 0.07 | 0.075 |
| 219 | BR-SMO | 0.117 | 0.788 | 0.781 | 0.659 | 0.761 | 0.156 | 0.175 |
| 220 | BR-C5.0 | 0.12 | 0.765 | 0.806 | 0.631 | 0.747 | 0.075 | 0.062 |
| 221 | BR-NB | 0.155 | 0.693 | 0.773 | 0.568 | 0.688 | 0.095 | 0.212 |
| 222 | BR-XGB | 0.131 | 0.773 | 0.763 | 0.612 | 0.728 | 0.078 | 0.025 |
| 223 | BR-CART | 0.142 | 0.708 | 0.782 | 0.576 | 0.7 | 0.083 | 0.075 |
| 224 | BR-MAJORITY | 0.201 | 0.674 | 0.378 | 0.263 | 0.403 | 0.302 | 0.288 |
| 225 | BR-RANDOM | 0.484 | 0.257 | 0.582 | 0.201 | 0.319 | 0.462 | 0.888 |
| 226 | ECC-RF | 0.133 | 0.732 | 0.772 | 0.581 | 0.7 | 0.108 | 0.062 |
| 227 | ECC-SVM | 0.137 | 0.719 | 0.772 | 0.588 | 0.701 | 0.096 | 0.1 |
| 228 | ECC-SMO | 0.122 | 0.77 | 0.793 | 0.657 | 0.754 | 0.074 | 0.05 |
| 229 | ECC-C5.0 | 0.134 | 0.746 | 0.748 | 0.584 | 0.699 | 0.077 | 0.062 |
| 230 | ECC-NB | 0.143 | 0.722 | 0.734 | 0.58 | 0.689 | 0.099 | 0.05 |
| 231 | ECC-XGB | 0.143 | 0.719 | 0.756 | 0.56 | 0.681 | 0.091 | 0.088 |
| 232 | ECC-CART | 0.16 | 0.645 | 0.696 | 0.483 | 0.616 | 0.115 | 0.088 |
| 233 | ECC-MAJORITY | 0.204 | 0.618 | 0.43 | 0.293 | 0.434 | 0.287 | 0.288 |
| 234 | ECC-RANDOM | 0.315 | 0.317 | 0.234 | 0.127 | 0.214 | 0.478 | 0.55 |
| 235 | LP-RF | 0.11 | 0.815 | 0.782 | 0.702 | 0.772 | 0.156 | 0.238 |
| 236 | LP-SVM | 0.24 | 0.611 | 0.54 | 0.419 | 0.521 | 0.309 | 0.55 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 237 | LP-SMO | 0.11 | 0.807 | 0.804 | 0.721 | 0.792 | 0.142 | 0.225 |
| 238 | LP-C5.0 | 0.102 | 0.839 | 0.829 | 0.739 | 0.813 | 0.127 | 0.2 |
| 239 | LP-NB | 0.192 | 0.56 | 0.672 | 0.45 | 0.552 | 0.237 | 0.45 |
| 240 | LP-XGB | 0.183 | 0.798 | 0.587 | 0.512 | 0.59 | 0.254 | 0.25 |
| 241 | LP-CART | 0.204 | 0.69 | 0.652 | 0.537 | 0.637 | 0.251 | 0.325 |
| 242 | LP-MAJORITY | 0.241 | 0.608 | 0.199 | 0.147 | 0.211 | 0.382 | 0.5 |
| 243 | LP-RANDOM | 0.289 | 0.403 | 0.376 | 0.223 | 0.327 | 0.429 | 0.65 |
| 244 | HOMER-RF | 0.219 | 0.55 | 0.411 | 0.329 | 0.443 | 0.537 | 0.375 |
| 245 | HOMER-SVM | 0.246 | 0.477 | 0.39 | 0.275 | 0.385 | 0.539 | 0.375 |
| 246 | HOMER-SMO | 0.209 | 0.659 | 0.466 | 0.409 | 0.519 | 0.338 | 0.338 |
| 247 | HOMER-C5.0 | 0.22 | 0.576 | 0.388 | 0.329 | 0.436 | 0.544 | 0.412 |
| 248 | HOMER-NB | 0.234 | 0.508 | 0.328 | 0.259 | 0.36 | 0.538 | 0.388 |
| 249 | HOMER-XGB | 0.242 | 0.512 | 0.362 | 0.273 | 0.383 | 0.534 | 0.375 |
| 250 | HOMER-CART | 0.256 | 0.45 | 0.331 | 0.251 | 0.353 | 0.462 | 0.5 |
| 251 | HOMER-MAJORITY | 0.33 | 0.263 | 0.244 | 0.133 | 0.215 | 0.489 | 0.8 |
| 252 | HOMER-RANDOM | 0.309 | 0.216 | 0.078 | 0.059 | 0.104 | 0.606 | 0.725 |
| 253 | RAKEL-RF | 0.103 | 0.853 | 0.763 | 0.669 | 0.771 | 0.108 | 0.088 |
| 254 | RAKEL-SVM | 0.125 | 0.829 | 0.695 | 0.591 | 0.704 | 0.15 | 0.138 |
| 255 | RAKEL-SMO | 0.112 | 0.801 | 0.793 | 0.678 | 0.772 | 0.112 | 0.138 |
| 256 | RAKEL-C5.0 | 0.107 | 0.812 | 0.836 | 0.701 | 0.797 | 0.072 | 0.1 |
| 257 | RAKEL-NB | 0.29 | 0.408 | 0.745 | 0.34 | 0.484 | 0.216 | 0.25 |
| 258 | RAKEL-XGB | 0.108 | 0.809 | 0.808 | 0.671 | 0.771 | 0.098 | 0.112 |
| 259 | RAKEL-CART | 0.133 | 0.718 | 0.779 | 0.576 | 0.705 | 0.091 | 0.112 |
| 260 | RAKEL-MAJORITY | 0.208 | 0.721 | 0.314 | 0.223 | 0.356 | 0.289 | 0.3 |
| 261 | RAKEL-RANDOM | 0.426 | 0.317 | 0.569 | 0.234 | 0.35 | 0.372 | 0.712 |
| 262 | BRPLUS-RF | 0.121 | 0.884 | 0.692 | 0.629 | 0.734 | 0.058 | 0.012 |
| 263 | BRPLUS-SVM | 0.126 | 0.816 | 0.742 | 0.625 | 0.731 | 0.067 | 0.038 |
| 264 | BRPLUS-SMO | 0.12 | 0.792 | 0.788 | 0.679 | 0.768 | 0.156 | 0.25 |
| 265 | BRPLUS-C5.0 | 0.138 | 0.747 | 0.812 | 0.628 | 0.734 | 0.087 | 0.038 |
| 266 | BRPLUS-NB | 0.143 | 0.758 | 0.769 | 0.614 | 0.716 | 0.075 | 0.088 |
| 267 | BRPLUS-XGB | 0.127 | 0.813 | 0.778 | 0.653 | 0.753 | 0.081 | 0.012 |
| 268 | BRPLUS-CART | 0.154 | 0.699 | 0.782 | 0.588 | 0.697 | 0.109 | 0.1 |
| 269 | BRPLUS-MAJORITY | 0.204 | 0.674 | 0.379 | 0.264 | 0.405 | 0.301 | 0.288 |
| 270 | BRPLUS-RANDOM | 0.49 | 0.247 | 0.45 | 0.186 | 0.299 | 0.521 | 0.712 |
| 271 | BR-RF | 0.139 | 0.845 | 0.749 | 0.643 | 0.754 | 0.064 | 0.025 |
| 272 | BR-SVM | 0.148 | 0.817 | 0.747 | 0.63 | 0.74 | 0.081 | 0.075 |
| 273 | BR-SMO | 0.144 | 0.798 | 0.793 | 0.669 | 0.771 | 0.159 | 0.188 |
| 274 | BR-C5.0 | 0.146 | 0.772 | 0.818 | 0.645 | 0.757 | 0.091 | 0.062 |
| 275 | BR-NB | 0.183 | 0.727 | 0.798 | 0.606 | 0.72 | 0.105 | 0.138 |
| 276 | BR-XGB | 0.16 | 0.778 | 0.775 | 0.622 | 0.737 | 0.097 | 0.025 |
| 277 | BR-CART | 0.177 | 0.714 | 0.793 | 0.586 | 0.709 | 0.102 | 0.075 |
| 278 | BR-MAJORITY | 0.244 | 0.674 | 0.469 | 0.292 | 0.439 | 0.283 | 0.288 |
| 279 | BR-RANDOM | 0.476 | 0.295 | 0.363 | 0.198 | 0.306 | 0.519 | 0.712 |
| 280 | ECC-RF | 0.168 | 0.748 | 0.796 | 0.614 | 0.723 | 0.112 | 0.075 |
| 281 | ECC-SVM | 0.158 | 0.759 | 0.787 | 0.623 | 0.725 | 0.102 | 0.112 |
| 282 | ECC-SMO | 0.156 | 0.788 | 0.784 | 0.658 | 0.756 | 0.103 | 0.062 |
| 283 | ECC-C5.0 | 0.171 | 0.75 | 0.781 | 0.608 | 0.72 | 0.107 | 0.05 |
| 284 | ECC-NB | 0.171 | 0.747 | 0.779 | 0.613 | 0.719 | 0.108 | 0.1 |
| 285 | ECC-XGB | 0.187 | 0.699 | 0.762 | 0.552 | 0.675 | 0.109 | 0.062 |
| 286 | ECC-CART | 0.193 | 0.684 | 0.745 | 0.537 | 0.662 | 0.132 | 0.075 |
| 287 | ECC-MAJORITY | 0.251 | 0.623 | 0.507 | 0.311 | 0.455 | 0.268 | 0.288 |
| 288 | ECC-RANDOM | 0.407 | 0.357 | 0.382 | 0.188 | 0.299 | 0.498 | 0.575 |
| 289 | LP-RF | 0.15 | 0.802 | 0.784 | 0.693 | 0.769 | 0.172 | 0.275 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 290 | LP-SVM | 0.288 | 0.647 | 0.59 | 0.462 | 0.565 | 0.32 | 0.475 |
| 291 | LP-SMO | 0.137 | 0.814 | 0.815 | 0.731 | 0.802 | 0.152 | 0.225 |
| 292 | LP-C5.0 | 0.117 | 0.858 | 0.847 | 0.756 | 0.828 | 0.126 | 0.188 |
| 293 | LP-NB | 0.24 | 0.572 | 0.724 | 0.455 | 0.564 | 0.26 | 0.462 |
| 294 | LP-XGB | 0.193 | 0.811 | 0.694 | 0.6 | 0.682 | 0.226 | 0.25 |
| 295 | LP-CART | 0.22 | 0.739 | 0.714 | 0.601 | 0.695 | 0.239 | 0.288 |
| 296 | LP-MAJORITY | 0.276 | 0.825 | 0.255 | 0.255 | 0.301 | 0.359 | 0.175 |
| 297 | LP-RANDOM | 0.367 | 0.388 | 0.483 | 0.241 | 0.352 | 0.446 | 0.738 |
| 298 | HOMER-RF | 0.172 | 0.728 | 0.847 | 0.651 | 0.744 | 0.146 | 0.175 |
| 299 | HOMER-SVM | 0.187 | 0.687 | 0.831 | 0.607 | 0.712 | 0.158 | 0.162 |
| 300 | HOMER-SMO | 0.162 | 0.786 | 0.794 | 0.677 | 0.768 | 0.174 | 0.238 |
| 301 | HOMER-C5.0 | 0.158 | 0.752 | 0.8 | 0.655 | 0.754 | 0.158 | 0.112 |
| 302 | HOMER-NB | 0.196 | 0.747 | 0.662 | 0.54 | 0.649 | 0.212 | 0.25 |
| 303 | HOMER-XGB | 0.174 | 0.719 | 0.793 | 0.62 | 0.727 | 0.17 | 0.15 |
| 304 | HOMER-CART | 0.2 | 0.683 | 0.75 | 0.569 | 0.679 | 0.182 | 0.162 |
| 305 | HOMER-MAJORITY | 0.373 | 0.418 | 0.643 | 0.289 | 0.421 | 0.354 | 0.762 |
| 306 | HOMER-RANDOM | 0.346 | 0.331 | 0.141 | 0.106 | 0.176 | 0.491 | 0.675 |
| 307 | RAKEL-RF | 0.137 | 0.824 | 0.779 | 0.666 | 0.768 | 0.107 | 0.088 |
| 308 | RAKEL-SVM | 0.15 | 0.827 | 0.728 | 0.617 | 0.729 | 0.142 | 0.088 |
| 309 | RAKEL-SMO | 0.139 | 0.82 | 0.806 | 0.706 | 0.796 | 0.117 | 0.112 |
| 310 | RAKEL-C5.0 | 0.124 | 0.812 | 0.852 | 0.71 | 0.806 | 0.079 | 0.062 |
| 311 | RAKEL-NB | 0.292 | 0.47 | 0.83 | 0.405 | 0.55 | 0.189 | 0.2 |
| 312 | RAKEL-XGB | 0.132 | 0.834 | 0.824 | 0.705 | 0.798 | 0.087 | 0.075 |
| 313 | RAKEL-CART | 0.147 | 0.781 | 0.806 | 0.644 | 0.752 | 0.097 | 0.062 |
| 314 | RAKEL-MAJORITY | 0.242 | 0.707 | 0.441 | 0.28 | 0.427 | 0.26 | 0.238 |
| 315 | RAKEL-RANDOM | 0.458 | 0.339 | 0.591 | 0.253 | 0.365 | 0.387 | 0.725 |
| 316 | BRPLUS-RF | 0.139 | 0.884 | 0.729 | 0.665 | 0.766 | 0.067 | 0.025 |
| 317 | BRPLUS-SVM | 0.154 | 0.816 | 0.767 | 0.637 | 0.735 | 0.075 | 0.05 |
| 318 | BRPLUS-SMO | 0.142 | 0.803 | 0.801 | 0.694 | 0.779 | 0.159 | 0.262 |
| 319 | BRPLUS-C5.0 | 0.16 | 0.768 | 0.827 | 0.651 | 0.756 | 0.108 | 0.038 |
| 320 | BRPLUS-NB | 0.177 | 0.72 | 0.755 | 0.602 | 0.698 | 0.091 | 0.1 |
| 321 | BRPLUS-XGB | 0.161 | 0.809 | 0.781 | 0.65 | 0.752 | 0.105 | 0.012 |
| 322 | BRPLUS-CART | 0.187 | 0.704 | 0.795 | 0.598 | 0.707 | 0.128 | 0.112 |
| 323 | BRPLUS-MAJORITY | 0.244 | 0.674 | 0.469 | 0.292 | 0.439 | 0.283 | 0.288 |
| 324 | BRPLUS-RANDOM | 0.475 | 0.306 | 0.46 | 0.214 | 0.33 | 0.495 | 0.65 |
| 325 | BR-RF | 0.149 | 0.667 | 0.6 | 0.509 | 0.604 | 0.081 | 0.262 |
| 326 | BR-SVM | 0.166 | 0.646 | 0.591 | 0.487 | 0.585 | 0.098 | 0.275 |
| 327 | BR-SMO | 0.159 | 0.635 | 0.618 | 0.52 | 0.607 | 0.152 | 0.262 |
| 328 | BR-C5.0 | 0.168 | 0.61 | 0.643 | 0.504 | 0.598 | 0.107 | 0.288 |
| 329 | BR-NB | 0.205 | 0.58 | 0.617 | 0.463 | 0.564 | 0.124 | 0.312 |
| 330 | BR-XGB | 0.161 | 0.632 | 0.628 | 0.511 | 0.605 | 0.096 | 0.25 |
| 331 | BR-CART | 0.186 | 0.594 | 0.624 | 0.483 | 0.582 | 0.108 | 0.275 |
| 332 | BR-MAJORITY | 0.312 | 0.633 | 0.306 | 0.287 | 0.406 | 0.255 | 0.288 |
| 333 | BR-RANDOM | 0.497 | 0.359 | 0.444 | 0.265 | 0.378 | 0.421 | 0.675 |
| 334 | ECC-RF | 0.167 | 0.63 | 0.623 | 0.507 | 0.6 | 0.096 | 0.275 |
| 335 | ECC-SVM | 0.178 | 0.622 | 0.625 | 0.496 | 0.592 | 0.102 | 0.288 |
| 336 | ECC-SMO | 0.162 | 0.616 | 0.632 | 0.517 | 0.604 | 0.106 | 0.212 |
| 337 | ECC-C5.0 | 0.166 | 0.634 | 0.631 | 0.508 | 0.601 | 0.098 | 0.3 |
| 338 | ECC-NB | 0.172 | 0.622 | 0.618 | 0.506 | 0.593 | 0.104 | 0.312 |
| 339 | ECC-XGB | 0.184 | 0.623 | 0.604 | 0.478 | 0.579 | 0.109 | 0.288 |
| 340 | ECC-CART | 0.198 | 0.592 | 0.591 | 0.466 | 0.567 | 0.117 | 0.3 |
| 341 | ECC-MAJORITY | 0.318 | 0.591 | 0.373 | 0.339 | 0.45 | 0.238 | 0.288 |
| 342 | ECC-RANDOM | 0.484 | 0.362 | 0.296 | 0.196 | 0.3 | 0.434 | 0.762 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 343 | LP-RF | 0.145 | 0.654 | 0.641 | 0.564 | 0.627 | 0.146 | 0.325 |
| 344 | LP-SVM | 0.247 | 0.527 | 0.497 | 0.397 | 0.483 | 0.263 | 0.525 |
| 345 | LP-SMO | 0.152 | 0.636 | 0.638 | 0.564 | 0.628 | 0.152 | 0.35 |
| 346 | LP-C5.0 | 0.112 | 0.681 | 0.67 | 0.606 | 0.66 | 0.105 | 0.275 |
| 347 | LP-NB | 0.338 | 0.525 | 0.509 | 0.401 | 0.496 | 0.24 | 0.425 |
| 348 | LP-XGB | 0.207 | 0.596 | 0.553 | 0.468 | 0.547 | 0.223 | 0.375 |
| 349 | LP-CART | 0.22 | 0.609 | 0.56 | 0.468 | 0.549 | 0.217 | 0.338 |
| 350 | LP-MAJORITY | 0.372 | 0 | 0 | 0 | 0 | 0.387 | 0.712 |
| 351 | LP-RANDOM | 0.381 | 0.44 | 0.406 | 0.316 | 0.399 | 0.316 | 0.588 |
| 352 | HOMER-RF | 0.338 | 0.544 | 0.473 | 0.359 | 0.478 | 0.319 | 0.475 |
| 353 | HOMER-SVM | 0.353 | 0.51 | 0.469 | 0.337 | 0.457 | 0.319 | 0.425 |
| 354 | HOMER-SMO | 0.334 | 0.57 | 0.439 | 0.369 | 0.477 | 0.263 | 0.338 |
| 355 | HOMER-C5.0 | 0.353 | 0.54 | 0.435 | 0.331 | 0.447 | 0.345 | 0.45 |
| 356 | HOMER-NB | 0.357 | 0.554 | 0.371 | 0.306 | 0.418 | 0.355 | 0.475 |
| 357 | HOMER-XGB | 0.328 | 0.554 | 0.419 | 0.333 | 0.448 | 0.338 | 0.45 |
| 358 | HOMER-CART | 0.328 | 0.57 | 0.421 | 0.333 | 0.449 | 0.311 | 0.412 |
| 359 | HOMER-MAJORITY | 0.578 | 0.332 | 0.402 | 0.238 | 0.351 | 0.441 | 0.712 |
| 360 | HOMER-RANDOM | 0.428 | 0.396 | 0.153 | 0.123 | 0.198 | 0.454 | 0.525 |
| 361 | RAKEL-RF | 0.135 | 0.67 | 0.639 | 0.557 | 0.631 | 0.101 | 0.312 |
| 362 | RAKEL-SVM | 0.175 | 0.661 | 0.558 | 0.476 | 0.571 | 0.13 | 0.262 |
| 363 | RAKEL-SMO | 0.147 | 0.657 | 0.622 | 0.558 | 0.627 | 0.124 | 0.262 |
| 364 | RAKEL-C5.0 | 0.142 | 0.666 | 0.654 | 0.565 | 0.643 | 0.091 | 0.238 |
| 365 | RAKEL-NB | 0.258 | 0.522 | 0.623 | 0.43 | 0.543 | 0.2 | 0.362 |
| 366 | RAKEL-XGB | 0.137 | 0.653 | 0.663 | 0.556 | 0.632 | 0.091 | 0.275 |
| 367 | RAKEL-CART | 0.152 | 0.635 | 0.647 | 0.525 | 0.613 | 0.092 | 0.262 |
| 368 | RAKEL-MAJORITY | 0.342 | 0.624 | 0.199 | 0.19 | 0.286 | 0.253 | 0.35 |
| 369 | RAKEL-RANDOM | 0.469 | 0.424 | 0.437 | 0.294 | 0.407 | 0.337 | 0.45 |
| 370 | BRPLUS-RF | 0.152 | 0.67 | 0.593 | 0.518 | 0.597 | 0.083 | 0.262 |
| 371 | BRPLUS-SVM | 0.17 | 0.628 | 0.611 | 0.494 | 0.58 | 0.1 | 0.262 |
| 372 | BRPLUS-SMO | 0.145 | 0.635 | 0.65 | 0.569 | 0.631 | 0.139 | 0.325 |
| 373 | BRPLUS-C5.0 | 0.178 | 0.606 | 0.665 | 0.515 | 0.605 | 0.114 | 0.275 |
| 374 | BRPLUS-NB | 0.188 | 0.615 | 0.63 | 0.493 | 0.58 | 0.104 | 0.288 |
| 375 | BRPLUS-XGB | 0.156 | 0.655 | 0.635 | 0.532 | 0.617 | 0.108 | 0.25 |
| 376 | BRPLUS-CART | 0.172 | 0.608 | 0.661 | 0.517 | 0.605 | 0.12 | 0.262 |
| 377 | BRPLUS-MAJORITY | 0.312 | 0.633 | 0.306 | 0.287 | 0.406 | 0.255 | 0.288 |
| 378 | BRPLUS-RANDOM | 0.475 | 0.396 | 0.439 | 0.283 | 0.399 | 0.373 | 0.488 |
| 379 | BR-RF | 0.149 | 0.664 | 0.672 | 0.56 | 0.641 | 0.072 | 0.275 |
| 380 | BR-SVM | 0.176 | 0.67 | 0.634 | 0.528 | 0.62 | 0.085 | 0.25 |
| 381 | BR-SMO | 0.174 | 0.649 | 0.651 | 0.543 | 0.629 | 0.162 | 0.262 |
| 382 | BR-C5.0 | 0.182 | 0.626 | 0.667 | 0.531 | 0.618 | 0.106 | 0.288 |
| 383 | BR-NB | 0.208 | 0.643 | 0.633 | 0.504 | 0.598 | 0.111 | 0.262 |
| 384 | BR-XGB | 0.17 | 0.651 | 0.658 | 0.541 | 0.627 | 0.083 | 0.25 |
| 385 | BR-CART | 0.205 | 0.612 | 0.654 | 0.51 | 0.603 | 0.103 | 0.25 |
| 386 | BR-MAJORITY | 0.36 | 0.633 | 0.385 | 0.351 | 0.467 | 0.25 | 0.288 |
| 387 | BR-RANDOM | 0.492 | 0.448 | 0.426 | 0.315 | 0.422 | 0.41 | 0.612 |
| 388 | ECC-RF | 0.179 | 0.645 | 0.635 | 0.523 | 0.608 | 0.101 | 0.288 |
| 389 | ECC-SVM | 0.196 | 0.599 | 0.606 | 0.498 | 0.576 | 0.114 | 0.312 |
| 390 | ECC-SMO | 0.182 | 0.628 | 0.618 | 0.518 | 0.602 | 0.108 | 0.3 |
| 391 | ECC-C5.0 | 0.19 | 0.641 | 0.645 | 0.523 | 0.613 | 0.091 | 0.238 |
| 392 | ECC-NB | 0.213 | 0.594 | 0.607 | 0.484 | 0.571 | 0.111 | 0.262 |
| 393 | ECC-XGB | 0.196 | 0.629 | 0.62 | 0.508 | 0.593 | 0.101 | 0.288 |
| 394 | ECC-CART | 0.201 | 0.628 | 0.619 | 0.503 | 0.594 | 0.107 | 0.288 |
| 395 | ECC-MAJORITY | 0.368 | 0.601 | 0.452 | 0.397 | 0.503 | 0.23 | 0.288 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 396 | ECC-RANDOM | 0.522 | 0.443 | 0.287 | 0.202 | 0.305 | 0.444 | 0.6 |
| 397 | LP-RF | 0.154 | 0.678 | 0.653 | 0.579 | 0.643 | 0.151 | 0.312 |
| 398 | LP-SVM | 0.238 | 0.582 | 0.604 | 0.489 | 0.567 | 0.224 | 0.425 |
| 399 | LP-SMO | 0.149 | 0.664 | 0.673 | 0.596 | 0.657 | 0.142 | 0.325 |
| 400 | LP-C5.0 | 0.135 | 0.684 | 0.682 | 0.605 | 0.662 | 0.116 | 0.262 |
| 401 | LP-NB | 0.375 | 0.571 | 0.564 | 0.44 | 0.537 | 0.257 | 0.4 |
| 402 | LP-XGB | 0.2 | 0.624 | 0.623 | 0.515 | 0.591 | 0.195 | 0.362 |
| 403 | LP-CART | 0.196 | 0.655 | 0.618 | 0.53 | 0.606 | 0.187 | 0.338 |
| 404 | LP-MAJORITY | 0.45 | 0 | 0 | 0 | 0 | 0.415 | 0.712 |
| 405 | LP-RANDOM | 0.442 | 0.437 | 0.413 | 0.312 | 0.405 | 0.373 | 0.588 |
| 406 | HOMER-RF | 0.372 | 0.563 | 0.426 | 0.359 | 0.467 | 0.357 | 0.388 |
| 407 | HOMER-SVM | 0.399 | 0.526 | 0.407 | 0.336 | 0.442 | 0.364 | 0.4 |
| 408 | HOMER-SMO | 0.392 | 0.563 | 0.389 | 0.34 | 0.438 | 0.288 | 0.375 |
| 409 | HOMER-C5.0 | 0.382 | 0.551 | 0.406 | 0.347 | 0.452 | 0.359 | 0.338 |
| 410 | HOMER-NB | 0.426 | 0.51 | 0.309 | 0.259 | 0.361 | 0.396 | 0.425 |
| 411 | HOMER-XGB | 0.388 | 0.553 | 0.397 | 0.334 | 0.44 | 0.366 | 0.35 |
| 412 | HOMER-CART | 0.4 | 0.527 | 0.386 | 0.32 | 0.427 | 0.325 | 0.325 |
| 413 | HOMER-MAJORITY | 0.538 | 0.406 | 0.358 | 0.263 | 0.367 | 0.448 | 0.65 |
| 414 | HOMER-RANDOM | 0.489 | 0.376 | 0.146 | 0.13 | 0.202 | 0.511 | 0.675 |
| 415 | RAKEL-RF | 0.139 | 0.677 | 0.689 | 0.588 | 0.659 | 0.087 | 0.275 |
| 416 | RAKEL-SVM | 0.181 | 0.657 | 0.621 | 0.52 | 0.603 | 0.121 | 0.3 |
| 417 | RAKEL-SMO | 0.145 | 0.684 | 0.665 | 0.591 | 0.656 | 0.123 | 0.288 |
| 418 | RAKEL-C5.0 | 0.148 | 0.669 | 0.703 | 0.595 | 0.666 | 0.074 | 0.238 |
| 419 | RAKEL-NB | 0.311 | 0.534 | 0.605 | 0.432 | 0.546 | 0.207 | 0.338 |
| 420 | RAKEL-XGB | 0.138 | 0.674 | 0.709 | 0.6 | 0.667 | 0.076 | 0.262 |
| 421 | RAKEL-CART | 0.169 | 0.634 | 0.691 | 0.549 | 0.631 | 0.081 | 0.238 |
| 422 | RAKEL-MAJORITY | 0.365 | 0.652 | 0.327 | 0.305 | 0.417 | 0.228 | 0.375 |
| 423 | RAKEL-RANDOM | 0.494 | 0.459 | 0.497 | 0.335 | 0.44 | 0.305 | 0.462 |
| 424 | BRPLUS-RF | 0.145 | 0.667 | 0.679 | 0.58 | 0.65 | 0.065 | 0.262 |
| 425 | BRPLUS-SVM | 0.179 | 0.653 | 0.639 | 0.524 | 0.607 | 0.086 | 0.238 |
| 426 | BRPLUS-SMO | 0.15 | 0.681 | 0.672 | 0.592 | 0.661 | 0.153 | 0.3 |
| 427 | BRPLUS-C5.0 | 0.185 | 0.624 | 0.693 | 0.543 | 0.623 | 0.114 | 0.275 |
| 428 | BRPLUS-NB | 0.192 | 0.651 | 0.649 | 0.529 | 0.611 | 0.086 | 0.262 |
| 429 | BRPLUS-XGB | 0.146 | 0.672 | 0.687 | 0.581 | 0.654 | 0.072 | 0.238 |
| 430 | BRPLUS-CART | 0.188 | 0.631 | 0.676 | 0.539 | 0.621 | 0.089 | 0.262 |
| 431 | BRPLUS-MAJORITY | 0.36 | 0.633 | 0.385 | 0.351 | 0.467 | 0.25 | 0.288 |
| 432 | BRPLUS-RANDOM | 0.495 | 0.456 | 0.436 | 0.304 | 0.421 | 0.404 | 0.488 |
| 433 | BR-RF | 0.175 | 0.631 | 0.592 | 0.51 | 0.58 | 0.076 | 0.312 |
| 434 | BR-SVM | 0.21 | 0.619 | 0.565 | 0.477 | 0.556 | 0.112 | 0.312 |
| 435 | BR-SMO | 0.198 | 0.609 | 0.573 | 0.481 | 0.559 | 0.185 | 0.325 |
| 436 | BR-C5.0 | 0.217 | 0.571 | 0.597 | 0.489 | 0.563 | 0.132 | 0.35 |
| 437 | BR-NB | 0.254 | 0.535 | 0.525 | 0.415 | 0.495 | 0.145 | 0.312 |
| 438 | BR-XGB | 0.221 | 0.584 | 0.563 | 0.47 | 0.548 | 0.127 | 0.325 |
| 439 | BR-CART | 0.244 | 0.584 | 0.571 | 0.467 | 0.549 | 0.133 | 0.362 |
| 440 | BR-MAJORITY | 0.406 | 0.606 | 0.409 | 0.362 | 0.466 | 0.168 | 0.338 |
| 441 | BR-RANDOM | 0.473 | 0.531 | 0.38 | 0.303 | 0.407 | 0.34 | 0.488 |
| 442 | ECC-RF | 0.188 | 0.582 | 0.591 | 0.501 | 0.567 | 0.11 | 0.312 |
| 443 | ECC-SVM | 0.221 | 0.573 | 0.576 | 0.476 | 0.551 | 0.127 | 0.338 |
| 444 | ECC-SMO | 0.217 | 0.589 | 0.571 | 0.478 | 0.556 | 0.123 | 0.3 |
| 445 | ECC-C5.0 | 0.208 | 0.587 | 0.583 | 0.486 | 0.558 | 0.122 | 0.312 |
| 446 | ECC-NB | 0.254 | 0.546 | 0.54 | 0.431 | 0.512 | 0.154 | 0.312 |
| 447 | ECC-XGB | 0.248 | 0.566 | 0.544 | 0.45 | 0.526 | 0.111 | 0.312 |
| 448 | ECC-CART | 0.233 | 0.576 | 0.558 | 0.459 | 0.54 | 0.129 | 0.362 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 449 | ECC-MAJORITY | 0.385 | 0.6 | 0.472 | 0.418 | 0.512 | 0.167 | 0.338 |
| 450 | ECC-RANDOM | 0.5 | 0.461 | 0.318 | 0.261 | 0.351 | 0.359 | 0.525 |
| 451 | LP-RF | 0.181 | 0.559 | 0.579 | 0.497 | 0.551 | 0.191 | 0.4 |
| 452 | LP-SVM | 0.221 | 0.5 | 0.576 | 0.456 | 0.517 | 0.218 | 0.5 |
| 453 | LP-SMO | 0.162 | 0.613 | 0.608 | 0.526 | 0.587 | 0.159 | 0.338 |
| 454 | LP-C5.0 | 0.223 | 0.534 | 0.549 | 0.462 | 0.522 | 0.232 | 0.438 |
| 455 | LP-NB | 0.471 | 0.529 | 0.456 | 0.376 | 0.474 | 0.308 | 0.4 |
| 456 | LP-XGB | 0.183 | 0.558 | 0.595 | 0.504 | 0.562 | 0.196 | 0.425 |
| 457 | LP-CART | 0.258 | 0.476 | 0.532 | 0.413 | 0.481 | 0.245 | 0.475 |
| 458 | LP-MAJORITY | 0.485 | 0 | 0 | 0 | 0 | 0.316 | 0.575 |
| 459 | LP-RANDOM | 0.558 | 0.366 | 0.362 | 0.264 | 0.345 | 0.402 | 0.562 |
| 460 | HOMER-RF | 0.462 | 0.535 | 0.505 | 0.408 | 0.505 | 0.281 | 0.375 |
| 461 | HOMER-SVM | 0.492 | 0.502 | 0.519 | 0.402 | 0.502 | 0.307 | 0.4 |
| 462 | HOMER-SMO | 0.433 | 0.557 | 0.468 | 0.399 | 0.496 | 0.277 | 0.5 |
| 463 | HOMER-C5.0 | 0.433 | 0.551 | 0.455 | 0.39 | 0.488 | 0.325 | 0.488 |
| 464 | HOMER-NB | 0.485 | 0.539 | 0.372 | 0.322 | 0.416 | 0.308 | 0.425 |
| 465 | HOMER-XGB | 0.444 | 0.545 | 0.464 | 0.387 | 0.484 | 0.298 | 0.375 |
| 466 | HOMER-CART | 0.44 | 0.541 | 0.451 | 0.377 | 0.479 | 0.31 | 0.4 |
| 467 | HOMER-MAJORITY | 0.569 | 0.45 | 0.545 | 0.382 | 0.486 | 0.446 | 0.575 |
| 468 | HOMER-RANDOM | 0.521 | 0.459 | 0.258 | 0.217 | 0.308 | 0.423 | 0.625 |
| 469 | RAKEL-RF | 0.16 | 0.603 | 0.61 | 0.531 | 0.586 | 0.113 | 0.325 |
| 470 | RAKEL-SVM | 0.202 | 0.553 | 0.57 | 0.485 | 0.542 | 0.168 | 0.362 |
| 471 | RAKEL-SMO | 0.179 | 0.588 | 0.578 | 0.497 | 0.562 | 0.14 | 0.338 |
| 472 | RAKEL-C5.0 | 0.185 | 0.605 | 0.621 | 0.52 | 0.587 | 0.118 | 0.3 |
| 473 | RAKEL-NB | 0.342 | 0.539 | 0.486 | 0.39 | 0.489 | 0.237 | 0.325 |
| 474 | RAKEL-XGB | 0.173 | 0.604 | 0.637 | 0.54 | 0.601 | 0.099 | 0.312 |
| 475 | RAKEL-CART | 0.215 | 0.556 | 0.599 | 0.491 | 0.557 | 0.125 | 0.325 |
| 476 | RAKEL-MAJORITY | 0.406 | 0.598 | 0.423 | 0.376 | 0.473 | 0.16 | 0.338 |
| 477 | RAKEL-RANDOM | 0.5 | 0.486 | 0.449 | 0.341 | 0.44 | 0.316 | 0.55 |
| 478 | BRPLUS-RF | 0.169 | 0.571 | 0.59 | 0.509 | 0.562 | 0.083 | 0.338 |
| 479 | BRPLUS-SVM | 0.225 | 0.535 | 0.554 | 0.454 | 0.522 | 0.121 | 0.312 |
| 480 | BRPLUS-SMO | 0.196 | 0.598 | 0.596 | 0.492 | 0.567 | 0.196 | 0.375 |
| 481 | BRPLUS-C5.0 | 0.229 | 0.54 | 0.588 | 0.48 | 0.545 | 0.132 | 0.325 |
| 482 | BRPLUS-NB | 0.262 | 0.477 | 0.512 | 0.394 | 0.467 | 0.136 | 0.312 |
| 483 | BRPLUS-XGB | 0.21 | 0.57 | 0.576 | 0.479 | 0.55 | 0.109 | 0.35 |
| 484 | BRPLUS-CART | 0.231 | 0.585 | 0.581 | 0.478 | 0.556 | 0.126 | 0.3 |
| 485 | BRPLUS-MAJORITY | 0.406 | 0.606 | 0.409 | 0.362 | 0.466 | 0.168 | 0.338 |
| 486 | BRPLUS-RANDOM | 0.46 | 0.53 | 0.411 | 0.322 | 0.421 | 0.356 | 0.488 |
| 487 | BR-RF | 0.167 | 0.619 | 0.612 | 0.556 | 0.598 | 0.044 | 0.362 |
| 488 | BR-SVM | 0.196 | 0.581 | 0.59 | 0.521 | 0.566 | 0.05 | 0.362 |
| 489 | BR-SMO | 0.162 | 0.585 | 0.554 | 0.515 | 0.553 | 0.131 | 0.412 |
| 490 | BR-C5.0 | 0.217 | 0.569 | 0.575 | 0.498 | 0.55 | 0.081 | 0.412 |
| 491 | BR-NB | 0.254 | 0.49 | 0.446 | 0.398 | 0.445 | 0.025 | 0.35 |
| 492 | BR-XGB | 0.208 | 0.583 | 0.548 | 0.502 | 0.548 | 0.056 | 0.362 |
| 493 | BR-CART | 0.275 | 0.515 | 0.488 | 0.435 | 0.482 | 0.05 | 0.362 |
| 494 | BR-MAJORITY | 0.471 | 0.575 | 0.419 | 0.356 | 0.445 | 0.131 | 0.475 |
| 495 | BR-RANDOM | 0.462 | 0.529 | 0.367 | 0.321 | 0.392 | 0.162 | 0.462 |
| 496 | ECC-RF | 0.229 | 0.54 | 0.508 | 0.448 | 0.497 | 0.031 | 0.35 |
| 497 | ECC-SVM | 0.233 | 0.544 | 0.533 | 0.465 | 0.516 | 0.012 | 0.325 |
| 498 | ECC-SMO | 0.2 | 0.577 | 0.55 | 0.494 | 0.541 | 0.031 | 0.362 |
| 499 | ECC-C5.0 | 0.221 | 0.583 | 0.55 | 0.483 | 0.539 | 0.019 | 0.325 |
| 500 | ECC-NB | 0.262 | 0.502 | 0.49 | 0.417 | 0.469 | 0.044 | 0.362 |
| 501 | ECC-XGB | 0.229 | 0.554 | 0.519 | 0.465 | 0.512 | 0.031 | 0.338 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 502 | ECC-CART | 0.221 | 0.56 | 0.535 | 0.488 | 0.529 | 0.062 | 0.388 |
| 503 | ECC-MAJORITY | 0.429 | 0.569 | 0.519 | 0.442 | 0.522 | 0.069 | 0.375 |
| 504 | ECC-RANDOM | 0.521 | 0.471 | 0.312 | 0.271 | 0.344 | 0.181 | 0.475 |
| 505 | LP-RF | 0.208 | 0.473 | 0.527 | 0.462 | 0.489 | 0.206 | 0.512 |
| 506 | LP-SVM | 0.233 | 0.425 | 0.496 | 0.421 | 0.448 | 0.256 | 0.562 |
| 507 | LP-SMO | 0.171 | 0.529 | 0.529 | 0.483 | 0.516 | 0.162 | 0.475 |
| 508 | LP-C5.0 | 0.212 | 0.521 | 0.535 | 0.456 | 0.504 | 0.188 | 0.488 |
| 509 | LP-NB | 0.267 | 0.54 | 0.498 | 0.442 | 0.501 | 0.131 | 0.462 |
| 510 | LP-XGB | 0.188 | 0.558 | 0.6 | 0.533 | 0.564 | 0.131 | 0.45 |
| 511 | LP-CART | 0.242 | 0.492 | 0.481 | 0.427 | 0.468 | 0.219 | 0.5 |
| 512 | LP-MAJORITY | 0.604 | 0.258 | 0.375 | 0.258 | 0.295 | 0.269 | 0.575 |
| 513 | LP-RANDOM | 0.558 | 0.423 | 0.36 | 0.317 | 0.372 | 0.25 | 0.538 |
| 514 | HOMER-RF | 0.292 | 0.59 | 0.656 | 0.571 | 0.607 | 0.031 | 0.35 |
| 515 | HOMER-SVM | 0.321 | 0.604 | 0.65 | 0.579 | 0.61 | 0.019 | 0.338 |
| 516 | HOMER-SMO | 0.362 | 0.606 | 0.629 | 0.56 | 0.6 | 0.106 | 0.425 |
| 517 | HOMER-C5.0 | 0.367 | 0.569 | 0.621 | 0.515 | 0.568 | 0.106 | 0.425 |
| 518 | HOMER-NB | 0.329 | 0.627 | 0.59 | 0.542 | 0.589 | 0.038 | 0.35 |
| 519 | HOMER-XGB | 0.317 | 0.592 | 0.627 | 0.548 | 0.594 | 0.044 | 0.35 |
| 520 | HOMER-CART | 0.337 | 0.581 | 0.6 | 0.506 | 0.562 | 0.075 | 0.388 |
| 521 | HOMER-MAJORITY | 0.479 | 0.521 | 0.675 | 0.521 | 0.57 | 0.269 | 0.575 |
| 522 | HOMER-RANDOM | 0.525 | 0.519 | 0.371 | 0.327 | 0.41 | 0.169 | 0.462 |
| 523 | RAKEL-RF | 0.225 | 0.473 | 0.515 | 0.45 | 0.48 | 0.206 | 0.512 |
| 524 | RAKEL-SVM | 0.238 | 0.421 | 0.496 | 0.417 | 0.445 | 0.269 | 0.575 |
| 525 | RAKEL-SMO | 0.171 | 0.529 | 0.529 | 0.483 | 0.516 | 0.162 | 0.475 |
| 526 | RAKEL-C5.0 | 0.212 | 0.521 | 0.535 | 0.456 | 0.504 | 0.188 | 0.488 |
| 527 | RAKEL-NB | 0.267 | 0.54 | 0.498 | 0.442 | 0.501 | 0.131 | 0.462 |
| 528 | RAKEL-XGB | 0.188 | 0.558 | 0.6 | 0.533 | 0.564 | 0.131 | 0.45 |
| 529 | RAKEL-CART | 0.242 | 0.492 | 0.481 | 0.427 | 0.468 | 0.219 | 0.5 |
| 530 | RAKEL-MAJORITY | 0.604 | 0.258 | 0.375 | 0.258 | 0.295 | 0.269 | 0.575 |
| 531 | RAKEL-RANDOM | 0.558 | 0.423 | 0.36 | 0.317 | 0.372 | 0.25 | 0.538 |
| 532 | BRPLUS-RF | 0.158 | 0.61 | 0.631 | 0.567 | 0.605 | 0.05 | 0.35 |
| 533 | BRPLUS-SVM | 0.208 | 0.55 | 0.567 | 0.492 | 0.538 | 0.05 | 0.362 |
| 534 | BRPLUS-SMO | 0.162 | 0.535 | 0.535 | 0.496 | 0.523 | 0.162 | 0.462 |
| 535 | BRPLUS-C5.0 | 0.171 | 0.552 | 0.621 | 0.535 | 0.57 | 0.106 | 0.412 |
| 536 | BRPLUS-NB | 0.254 | 0.452 | 0.435 | 0.388 | 0.426 | 0.031 | 0.35 |
| 537 | BRPLUS-XGB | 0.192 | 0.544 | 0.579 | 0.51 | 0.548 | 0.062 | 0.375 |
| 538 | BRPLUS-CART | 0.229 | 0.533 | 0.535 | 0.481 | 0.519 | 0.056 | 0.375 |
| 539 | BRPLUS-MAJORITY | 0.471 | 0.575 | 0.419 | 0.356 | 0.445 | 0.131 | 0.475 |
| 540 | BRPLUS-RANDOM | 0.462 | 0.529 | 0.367 | 0.321 | 0.392 | 0.162 | 0.462 |

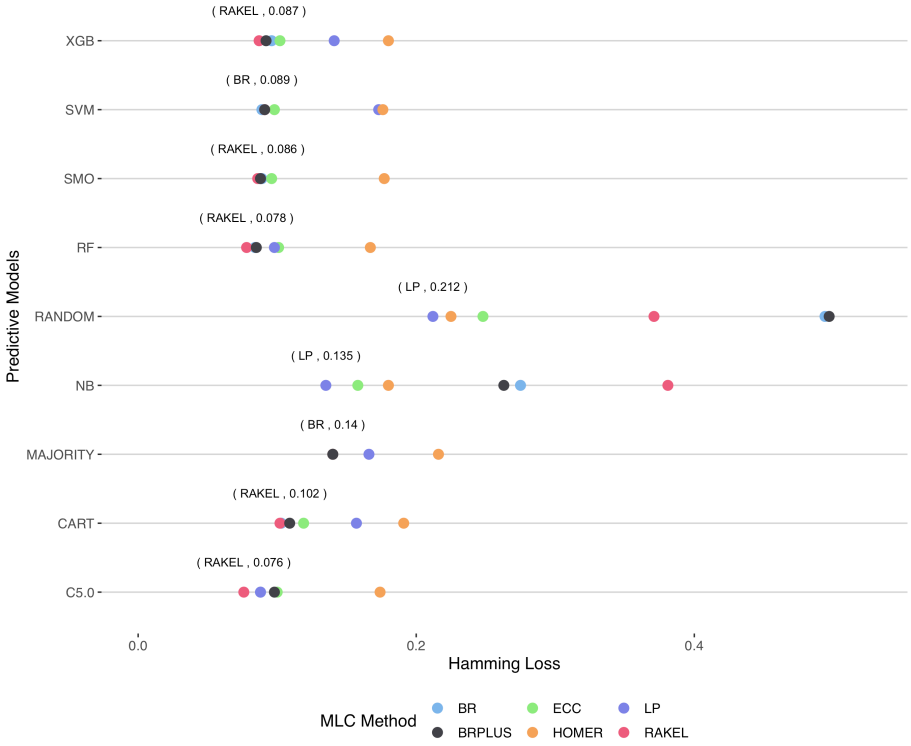



**Figure H. 8.** Hamming-Loss for all MLC strategies applied in CDS2.

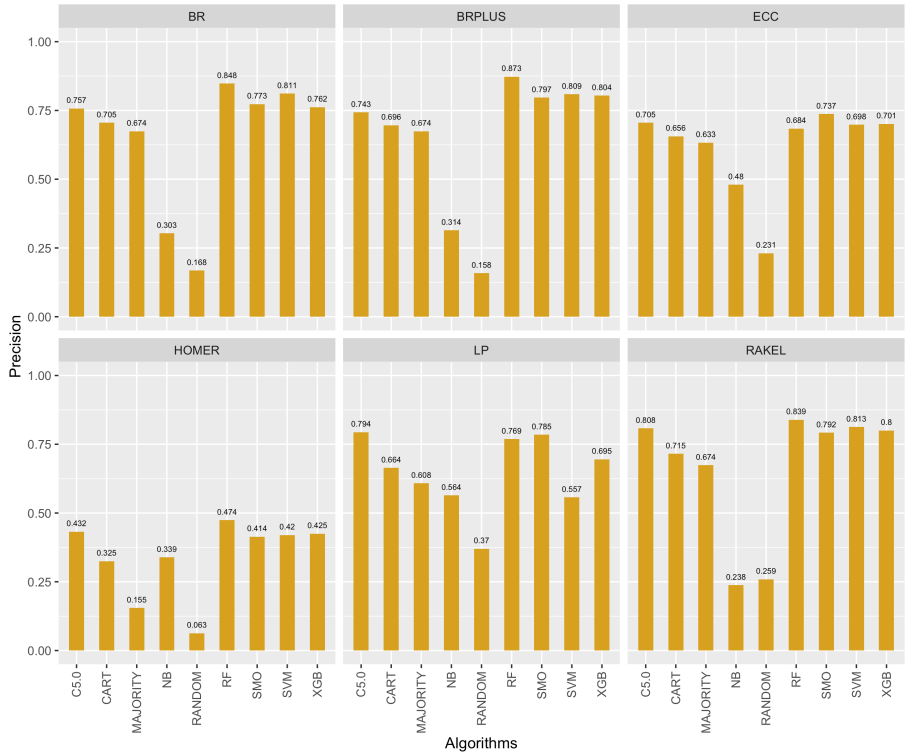



**Figure H. 9.** Precision for all MLC strategies applied in CDS2.

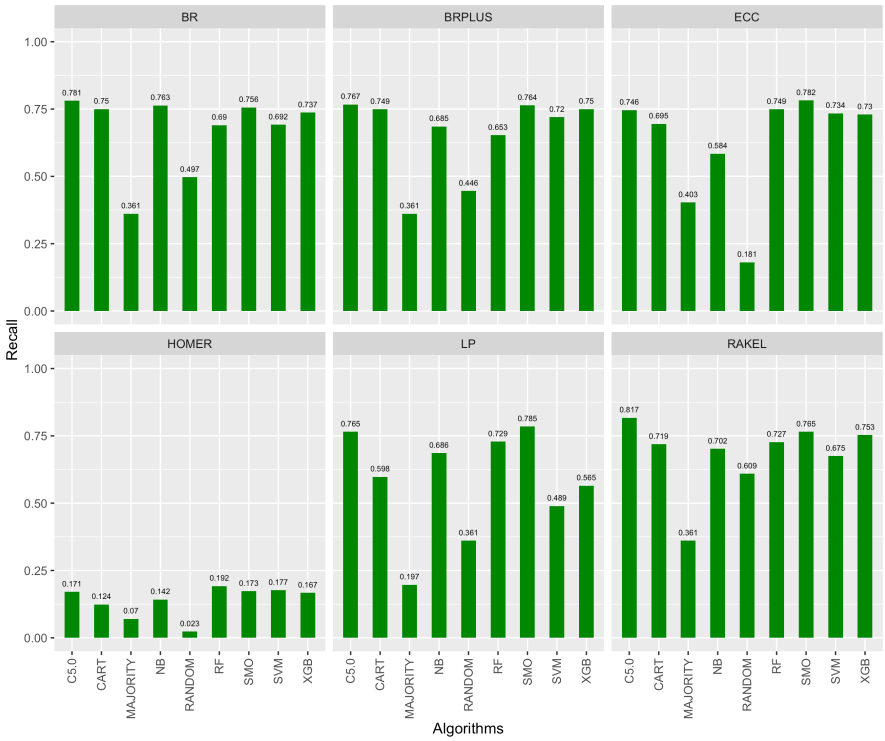



**Figure H. 10.** Recall for all MLC strategies applied in CDS2.

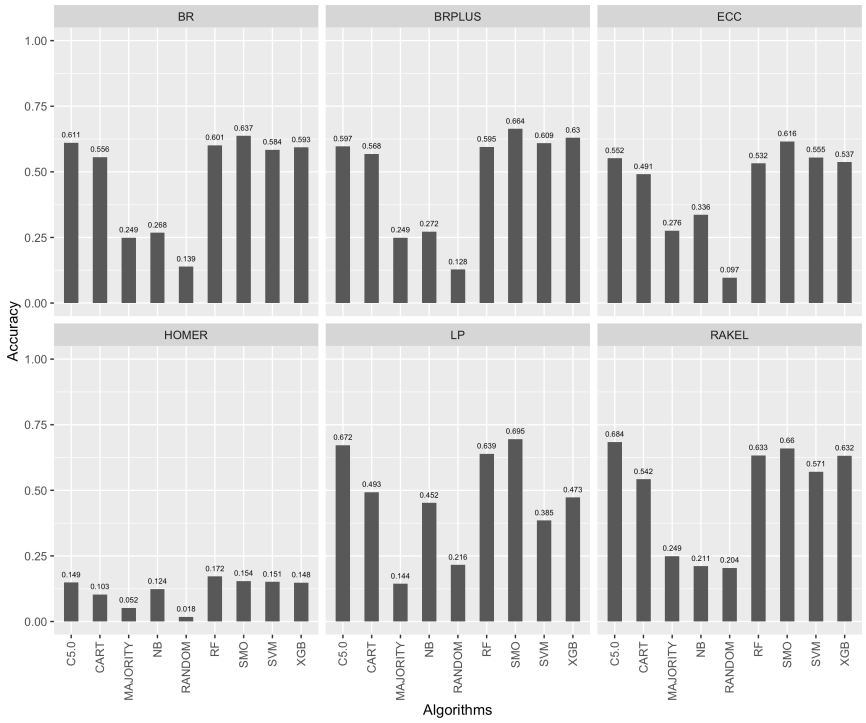



**Figure H. 11.** Accuracy for all MLC strategies applied in CDS2.

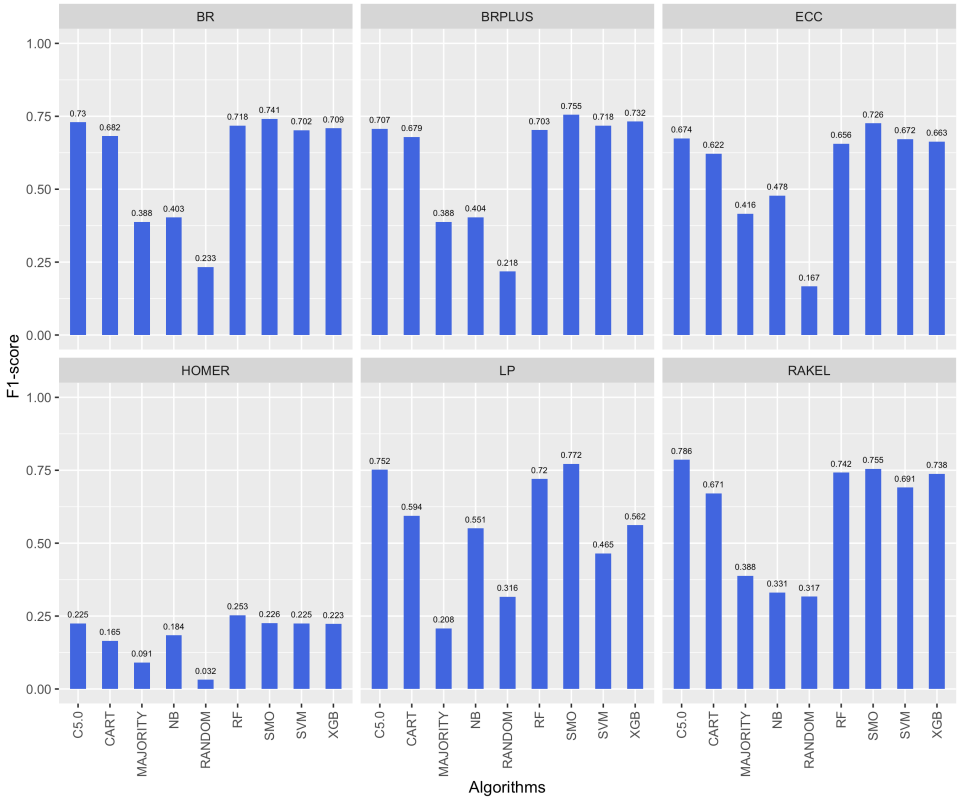



**Figure H. 12.** F1-Score for all MLC strategies applied in CDS2.

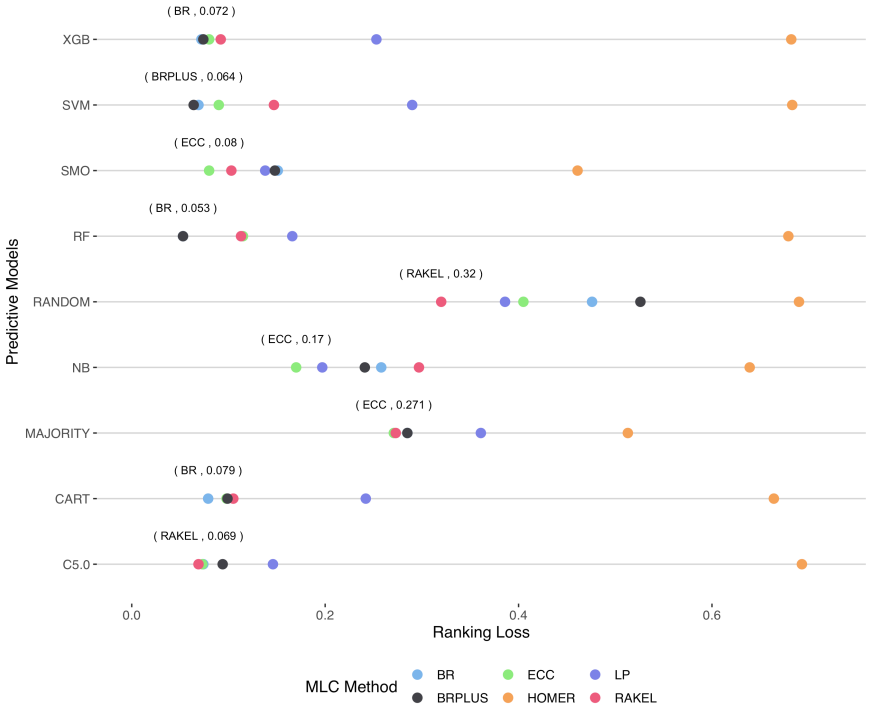



**Figure H. 13.** Ranking-Loss for all MLC strategies applied in CDS2.

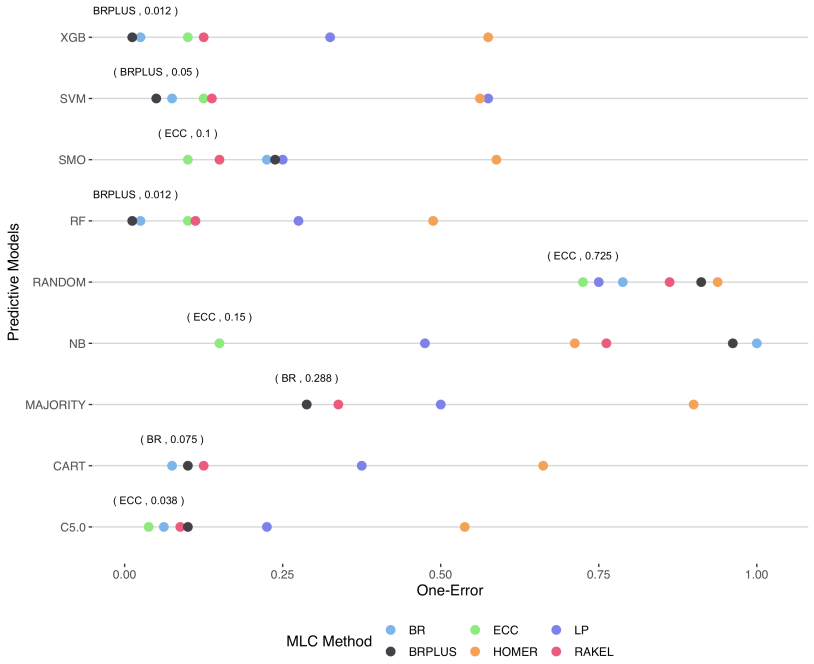



**Figure H. 14.** One-Error for all MLC strategies applied in CDS2.

# H.3. Performance Measures for Models (CDS3)

| | mlc_methods | values_hl | values_pr | values_re | values_ac | values_f1 | values_rl | values_oe |
|---|---|---|---|---|---|---|---|---|
| 1 | BR-RF | 0.144 | 0.716 | 0.476 | 0.387 | 0.543 | 0.171 | 0.15 |
| 2 | BR-SVM | 0.138 | 0.759 | 0.443 | 0.37 | 0.525 | 0.182 | 0.275 |
| 3 | BR-SMO | 0.149 | 0.674 | 0.485 | 0.369 | 0.525 | 0.314 | 0.392 |
| 4 | BR-C5.0 | 0.171 | 0.565 | 0.471 | 0.327 | 0.483 | 0.217 | 0.3 |
| 5 | BR-NB | 0.292 | 0.377 | 0.588 | 0.263 | 0.407 | 0.306 | 0.667 |
| 6 | BR-XGB | 0.167 | 0.567 | 0.504 | 0.347 | 0.506 | 0.195 | 0.275 |
| 7 | BR-CART | 0.156 | 0.608 | 0.511 | 0.362 | 0.524 | 0.201 | 0.125 |
| 8 | ECC-RF | 0.173 | 0.567 | 0.593 | 0.37 | 0.532 | 0.21 | 0.175 |
| 9 | ECC-SVM | 0.182 | 0.533 | 0.563 | 0.342 | 0.502 | 0.227 | 0.15 |
| 10 | ECC-SMO | 0.173 | 0.558 | 0.565 | 0.352 | 0.512 | 0.213 | 0.2 |
| 11 | ECC-C5.0 | 0.181 | 0.543 | 0.563 | 0.345 | 0.504 | 0.189 | 0.275 |
| 12 | ECC-NB | 0.181 | 0.546 | 0.517 | 0.328 | 0.486 | 0.206 | 0.308 |
| 13 | ECC-XGB | 0.171 | 0.556 | 0.605 | 0.38 | 0.542 | 0.198 | 0.15 |
| 14 | ECC-CART | 0.169 | 0.57 | 0.568 | 0.371 | 0.528 | 0.198 | 0.175 |
| 15 | LP-RF | 0.197 | 0.539 | 0.508 | 0.313 | 0.467 | 0.336 | 0.533 |
| 16 | LP-SVM | 0.187 | 0.575 | 0.244 | 0.173 | 0.284 | 0.419 | 0.683 |
| 17 | LP-SMO | 0.201 | 0.496 | 0.484 | 0.299 | 0.447 | 0.354 | 0.642 |
| 18 | LP-C5.0 | 0.175 | 0.582 | 0.476 | 0.325 | 0.479 | 0.335 | 0.375 |
| 19 | LP-NB | 0.204 | 0.47 | 0.457 | 0.275 | 0.419 | 0.368 | 0.633 |
| 20 | LP-XGB | 0.169 | 0.755 | 0.172 | 0.152 | 0.254 | 0.456 | 0.225 |
| 21 | LP-CART | 0.163 | 0.686 | 0.277 | 0.231 | 0.361 | 0.426 | 0.2 |
| 22 | HOMER-RF | 0.22 | 0.198 | 0.048 | 0.042 | 0.074 | 0.736 | 0.8 |
| 23 | HOMER-SVM | 0.227 | 0.134 | 0.045 | 0.036 | 0.065 | 0.745 | 0.867 |
| 24 | HOMER-SMO | 0.219 | 0.242 | 0.074 | 0.061 | 0.107 | 0.562 | 0.875 |
| 25 | HOMER-C5.0 | 0.236 | 0.165 | 0.045 | 0.036 | 0.066 | 0.747 | 0.75 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 26 | HOMER-NB | 0.246 | 0.099 | 0.048 | 0.035 | 0.063 | 0.712 | 0.925 |
| 27 | HOMER-XGB | 0.23 | 0.182 | 0.047 | 0.038 | 0.07 | 0.728 | 0.767 |
| 28 | HOMER-CART | 0.223 | 0.177 | 0.046 | 0.038 | 0.07 | 0.722 | 0.867 |
| 29 | RAKEL-RF | 0.142 | 0.714 | 0.492 | 0.392 | 0.551 | 0.275 | 0.175 |
| 30 | RAKEL-SVM | 0.139 | 0.771 | 0.425 | 0.357 | 0.511 | 0.299 | 0.275 |
| 31 | RAKEL-SMO | 0.146 | 0.687 | 0.503 | 0.384 | 0.541 | 0.28 | 0.258 |
| 32 | RAKEL-C5.0 | 0.161 | 0.608 | 0.479 | 0.355 | 0.509 | 0.233 | 0.25 |
| 33 | RAKEL-NB | 0.227 | 0.441 | 0.597 | 0.311 | 0.467 | 0.268 | 0.517 |
| 34 | RAKEL-XGB | 0.163 | 0.606 | 0.49 | 0.352 | 0.51 | 0.244 | 0.2 |
| 35 | RAKEL-CART | 0.144 | 0.666 | 0.522 | 0.39 | 0.55 | 0.246 | 0.25 |
| 36 | BRPLUS-RF | 0.141 | 0.75 | 0.43 | 0.368 | 0.522 | 0.167 | 0.175 |
| 37 | BRPLUS-SVM | 0.142 | 0.749 | 0.448 | 0.366 | 0.52 | 0.183 | 0.35 |
| 38 | BRPLUS-SMO | 0.158 | 0.671 | 0.468 | 0.344 | 0.502 | 0.33 | 0.333 |
| 39 | BRPLUS-C5.0 | 0.167 | 0.617 | 0.477 | 0.35 | 0.501 | 0.226 | 0.125 |
| 40 | BRPLUS-NB | 0.229 | 0.443 | 0.564 | 0.299 | 0.451 | 0.252 | 0.425 |
| 41 | BRPLUS-XGB | 0.169 | 0.579 | 0.508 | 0.347 | 0.507 | 0.189 | 0.175 |
| 42 | BRPLUS-CART | 0.165 | 0.58 | 0.511 | 0.35 | 0.51 | 0.2 | 0.2 |
| 43 | BR-RF | 0.215 | 0.72 | 0.523 | 0.426 | 0.582 | 0.194 | 0.15 |
| 44 | BR-SVM | 0.207 | 0.758 | 0.497 | 0.409 | 0.566 | 0.224 | 0.2 |
| 45 | BR-SMO | 0.223 | 0.685 | 0.528 | 0.411 | 0.567 | 0.292 | 0.367 |
| 46 | BR-C5.0 | 0.262 | 0.576 | 0.512 | 0.357 | 0.515 | 0.253 | 0.3 |
| 47 | BR-NB | 0.395 | 0.419 | 0.64 | 0.31 | 0.461 | 0.369 | 0.592 |
| 48 | BR-XGB | 0.258 | 0.573 | 0.551 | 0.377 | 0.537 | 0.228 | 0.275 |
| 49 | BR-CART | 0.24 | 0.608 | 0.561 | 0.389 | 0.554 | 0.248 | 0.125 |
| 50 | ECC-RF | 0.247 | 0.595 | 0.614 | 0.409 | 0.573 | 0.216 | 0.2 |
| 51 | ECC-SVM | 0.265 | 0.563 | 0.598 | 0.375 | 0.539 | 0.236 | 0.175 |
| 52 | ECC-SMO | 0.257 | 0.581 | 0.6 | 0.384 | 0.55 | 0.229 | 0.175 |
| 53 | ECC-C5.0 | 0.247 | 0.609 | 0.61 | 0.403 | 0.568 | 0.212 | 0.125 |
| 54 | ECC-NB | 0.264 | 0.579 | 0.572 | 0.37 | 0.531 | 0.226 | 0.225 |
| 55 | ECC-XGB | 0.252 | 0.595 | 0.614 | 0.405 | 0.569 | 0.221 | 0.15 |
| 56 | ECC-CART | 0.248 | 0.583 | 0.617 | 0.409 | 0.572 | 0.226 | 0.125 |
| 57 | LP-RF | 0.273 | 0.596 | 0.571 | 0.383 | 0.539 | 0.318 | 0.383 |
| 58 | LP-SVM | 0.268 | 0.74 | 0.265 | 0.201 | 0.323 | 0.373 | 0.3 |
| 59 | LP-SMO | 0.294 | 0.533 | 0.522 | 0.35 | 0.496 | 0.35 | 0.517 |
| 60 | LP-C5.0 | 0.266 | 0.603 | 0.507 | 0.361 | 0.515 | 0.332 | 0.3 |
| 61 | LP-NB | 0.296 | 0.511 | 0.492 | 0.314 | 0.462 | 0.364 | 0.533 |
| 62 | LP-XGB | 0.26 | 0.783 | 0.188 | 0.175 | 0.28 | 0.411 | 0.175 |
| 63 | LP-CART | 0.243 | 0.704 | 0.387 | 0.31 | 0.45 | 0.338 | 0.3 |
| 64 | HOMER-RF | 0.366 | 0.269 | 0.153 | 0.111 | 0.183 | 0.684 | 0.775 |
| 65 | HOMER-SVM | 0.407 | 0.221 | 0.149 | 0.098 | 0.164 | 0.707 | 0.817 |
| 66 | HOMER-SMO | 0.376 | 0.284 | 0.176 | 0.126 | 0.207 | 0.529 | 0.792 |
| 67 | HOMER-C5.0 | 0.356 | 0.303 | 0.154 | 0.114 | 0.188 | 0.676 | 0.608 |
| 68 | HOMER-NB | 0.373 | 0.233 | 0.117 | 0.083 | 0.142 | 0.668 | 0.792 |
| 69 | HOMER-XGB | 0.37 | 0.266 | 0.151 | 0.108 | 0.179 | 0.683 | 0.667 |
| 70 | HOMER-CART | 0.38 | 0.221 | 0.119 | 0.081 | 0.142 | 0.686 | 0.742 |
| 71 | RAKEL-RF | 0.212 | 0.722 | 0.544 | 0.436 | 0.593 | 0.248 | 0.175 |
| 72 | RAKEL-SVM | 0.212 | 0.79 | 0.441 | 0.377 | 0.53 | 0.277 | 0.2 |
| 73 | RAKEL-SMO | 0.227 | 0.698 | 0.526 | 0.411 | 0.564 | 0.271 | 0.175 |
| 74 | RAKEL-C5.0 | 0.229 | 0.636 | 0.575 | 0.413 | 0.574 | 0.228 | 0.158 |
| 75 | RAKEL-NB | 0.308 | 0.484 | 0.596 | 0.344 | 0.503 | 0.279 | 0.333 |
| 76 | RAKEL-XGB | 0.239 | 0.618 | 0.565 | 0.406 | 0.567 | 0.236 | 0.25 |
| 77 | RAKEL-CART | 0.229 | 0.667 | 0.506 | 0.381 | 0.541 | 0.247 | 0.3 |
| 78 | BRPLUS-RF | 0.211 | 0.754 | 0.491 | 0.411 | 0.567 | 0.195 | 0.15 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 79 | BRPLUS-SVM | 0.213 | 0.755 | 0.485 | 0.394 | 0.552 | 0.21 | 0.275 |
| 80 | BRPLUS-SMO | 0.228 | 0.702 | 0.527 | 0.409 | 0.565 | 0.295 | 0.25 |
| 81 | BRPLUS-C5.0 | 0.268 | 0.578 | 0.522 | 0.363 | 0.517 | 0.244 | 0.125 |
| 82 | BRPLUS-NB | 0.333 | 0.468 | 0.626 | 0.342 | 0.496 | 0.3 | 0.342 |
| 83 | BRPLUS-XGB | 0.258 | 0.592 | 0.559 | 0.379 | 0.541 | 0.224 | 0.175 |
| 84 | BRPLUS-CART | 0.257 | 0.58 | 0.56 | 0.376 | 0.538 | 0.238 | 0.2 |
| 85 | BR-RF | 0.281 | 0.745 | 0.715 | 0.566 | 0.71 | 0.199 | 0.15 |
| 86 | BR-SVM | 0.301 | 0.749 | 0.65 | 0.526 | 0.675 | 0.251 | 0.175 |
| 87 | BR-SMO | 0.286 | 0.744 | 0.708 | 0.561 | 0.705 | 0.281 | 0.292 |
| 88 | BR-C5.0 | 0.34 | 0.672 | 0.671 | 0.503 | 0.652 | 0.272 | 0.3 |
| 89 | BR-NB | 0.407 | 0.571 | 0.7 | 0.453 | 0.603 | 0.403 | 0.375 |
| 90 | BR-XGB | 0.318 | 0.686 | 0.703 | 0.532 | 0.679 | 0.246 | 0.25 |
| 91 | BR-CART | 0.307 | 0.691 | 0.752 | 0.55 | 0.698 | 0.236 | 0.125 |
| 92 | ECC-RF | 0.285 | 0.738 | 0.723 | 0.565 | 0.709 | 0.217 | 0.25 |
| 93 | ECC-SVM | 0.29 | 0.727 | 0.728 | 0.561 | 0.705 | 0.204 | 0.15 |
| 94 | ECC-SMO | 0.303 | 0.711 | 0.724 | 0.547 | 0.693 | 0.217 | 0.175 |
| 95 | ECC-C5.0 | 0.293 | 0.722 | 0.717 | 0.55 | 0.699 | 0.216 | 0.15 |
| 96 | ECC-NB | 0.343 | 0.676 | 0.664 | 0.501 | 0.646 | 0.291 | 0.175 |
| 97 | ECC-XGB | 0.289 | 0.73 | 0.735 | 0.56 | 0.708 | 0.22 | 0.15 |
| 98 | ECC-CART | 0.32 | 0.692 | 0.709 | 0.527 | 0.678 | 0.246 | 0.2 |
| 99 | LP-RF | 0.327 | 0.689 | 0.673 | 0.516 | 0.661 | 0.334 | 0.358 |
| 100 | LP-SVM | 0.424 | 0.753 | 0.362 | 0.281 | 0.427 | 0.374 | 0.3 |
| 101 | LP-SMO | 0.37 | 0.668 | 0.611 | 0.476 | 0.618 | 0.378 | 0.308 |
| 102 | LP-C5.0 | 0.318 | 0.698 | 0.677 | 0.522 | 0.669 | 0.323 | 0.208 |
| 103 | LP-NB | 0.351 | 0.663 | 0.634 | 0.487 | 0.634 | 0.366 | 0.233 |
| 104 | LP-XGB | 0.343 | 0.808 | 0.437 | 0.387 | 0.535 | 0.33 | 0.15 |
| 105 | LP-CART | 0.403 | 0.682 | 0.438 | 0.356 | 0.5 | 0.409 | 0.292 |
| 106 | HOMER-RF | 0.526 | 0.444 | 0.272 | 0.21 | 0.325 | 0.673 | 0.4 |
| 107 | HOMER-SVM | 0.575 | 0.396 | 0.265 | 0.198 | 0.307 | 0.695 | 0.567 |
| 108 | HOMER-SMO | 0.548 | 0.401 | 0.28 | 0.205 | 0.319 | 0.58 | 0.525 |
| 109 | HOMER-C5.0 | 0.555 | 0.404 | 0.204 | 0.16 | 0.259 | 0.711 | 0.567 |
| 110 | HOMER-NB | 0.55 | 0.424 | 0.232 | 0.179 | 0.288 | 0.65 | 0.408 |
| 111 | HOMER-XGB | 0.543 | 0.424 | 0.226 | 0.178 | 0.284 | 0.682 | 0.417 |
| 112 | HOMER-CART | 0.56 | 0.399 | 0.235 | 0.182 | 0.286 | 0.71 | 0.642 |
| 113 | RAKEL-RF | 0.289 | 0.739 | 0.721 | 0.565 | 0.711 | 0.212 | 0.125 |
| 114 | RAKEL-SVM | 0.287 | 0.787 | 0.635 | 0.533 | 0.68 | 0.253 | 0.2 |
| 115 | RAKEL-SMO | 0.289 | 0.738 | 0.706 | 0.558 | 0.703 | 0.254 | 0.225 |
| 116 | RAKEL-C5.0 | 0.296 | 0.706 | 0.758 | 0.572 | 0.713 | 0.223 | 0.15 |
| 117 | RAKEL-NB | 0.355 | 0.624 | 0.744 | 0.508 | 0.656 | 0.309 | 0.3 |
| 118 | RAKEL-XGB | 0.322 | 0.683 | 0.725 | 0.533 | 0.685 | 0.24 | 0.2 |
| 119 | RAKEL-CART | 0.314 | 0.687 | 0.706 | 0.53 | 0.679 | 0.245 | 0.225 |
| 120 | BRPLUS-RF | 0.262 | 0.775 | 0.716 | 0.584 | 0.724 | 0.198 | 0.15 |
| 121 | BRPLUS-SVM | 0.298 | 0.751 | 0.656 | 0.534 | 0.681 | 0.252 | 0.225 |
| 122 | BRPLUS-SMO | 0.301 | 0.713 | 0.714 | 0.55 | 0.693 | 0.298 | 0.217 |
| 123 | BRPLUS-C5.0 | 0.339 | 0.665 | 0.684 | 0.507 | 0.656 | 0.267 | 0.175 |
| 124 | BRPLUS-NB | 0.39 | 0.576 | 0.724 | 0.471 | 0.616 | 0.358 | 0.25 |
| 125 | BRPLUS-XGB | 0.323 | 0.692 | 0.707 | 0.532 | 0.678 | 0.239 | 0.25 |
| 126 | BRPLUS-CART | 0.321 | 0.686 | 0.737 | 0.536 | 0.684 | 0.239 | 0.175 |
| 127 | BR-RF | 0.22 | 0.819 | 0.948 | 0.773 | 0.854 | 0.19 | 0.125 |
| 128 | BR-SVM | 0.22 | 0.8 | 0.968 | 0.78 | 0.857 | 0.235 | 0.225 |
| 129 | BR-SMO | 0.245 | 0.81 | 0.907 | 0.748 | 0.831 | 0.252 | 0.175 |
| 130 | BR-C5.0 | 0.265 | 0.82 | 0.882 | 0.722 | 0.822 | 0.238 | 0.25 |
| 131 | BR-NB | 0.34 | 0.815 | 0.759 | 0.633 | 0.754 | 0.225 | 0.2 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 132 | BR-XGB | 0.262 | 0.809 | 0.899 | 0.729 | 0.823 | 0.275 | 0.225 |
| 133 | BR-CART | 0.217 | 0.82 | 0.941 | 0.771 | 0.855 | 0.192 | 0.075 |
| 134 | ECC-RF | 0.283 | 0.789 | 0.819 | 0.696 | 0.779 | 0.267 | 0.175 |
| 135 | ECC-SVM | 0.297 | 0.793 | 0.819 | 0.683 | 0.782 | 0.25 | 0.15 |
| 136 | ECC-SMO | 0.253 | 0.824 | 0.85 | 0.724 | 0.81 | 0.219 | 0.125 |
| 137 | ECC-C5.0 | 0.273 | 0.807 | 0.84 | 0.705 | 0.793 | 0.223 | 0.1 |
| 138 | ECC-NB | 0.333 | 0.784 | 0.8 | 0.652 | 0.76 | 0.308 | 0.2 |
| 139 | ECC-XGB | 0.287 | 0.808 | 0.842 | 0.695 | 0.798 | 0.269 | 0.125 |
| 140 | ECC-CART | 0.287 | 0.805 | 0.835 | 0.694 | 0.793 | 0.277 | 0.175 |
| 141 | LP-RF | 0.24 | 0.822 | 0.932 | 0.753 | 0.842 | 0.225 | 0.15 |
| 142 | LP-SVM | 0.195 | 0.805 | 1 | 0.805 | 0.874 | 0.258 | 0.2 |
| 143 | LP-SMO | 0.205 | 0.825 | 0.94 | 0.788 | 0.86 | 0.219 | 0.15 |
| 144 | LP-C5.0 | 0.242 | 0.827 | 0.903 | 0.747 | 0.838 | 0.225 | 0.15 |
| 145 | LP-NB | 0.245 | 0.813 | 0.901 | 0.742 | 0.831 | 0.24 | 0.175 |
| 146 | LP-XGB | 0.255 | 0.791 | 0.925 | 0.745 | 0.829 | 0.277 | 0.225 |
| 147 | LP-CART | 0.27 | 0.764 | 0.862 | 0.718 | 0.789 | 0.333 | 0.275 |
| 148 | HOMER-RF | 0.205 | 0.809 | 0.977 | 0.79 | 0.866 | 0.217 | 0.175 |
| 149 | HOMER-SVM | 0.2 | 0.809 | 0.982 | 0.797 | 0.87 | 0.25 | 0.225 |
| 150 | HOMER-SMO | 0.22 | 0.817 | 0.943 | 0.773 | 0.854 | 0.231 | 0.15 |
| 151 | HOMER-C5.0 | 0.248 | 0.808 | 0.893 | 0.737 | 0.825 | 0.248 | 0.2 |
| 152 | HOMER-NB | 0.385 | 0.831 | 0.697 | 0.584 | 0.715 | 0.198 | 0.125 |
| 153 | HOMER-XGB | 0.248 | 0.813 | 0.92 | 0.743 | 0.835 | 0.283 | 0.25 |
| 154 | HOMER-CART | 0.207 | 0.818 | 0.956 | 0.782 | 0.862 | 0.227 | 0.15 |
| 155 | RAKEL-RF | 0.21 | 0.822 | 0.962 | 0.783 | 0.864 | 0.225 | 0.15 |
| 156 | RAKEL-SVM | 0.2 | 0.805 | 0.995 | 0.8 | 0.871 | 0.258 | 0.2 |
| 157 | RAKEL-SMO | 0.21 | 0.815 | 0.952 | 0.783 | 0.859 | 0.267 | 0.225 |
| 158 | RAKEL-C5.0 | 0.225 | 0.805 | 0.949 | 0.77 | 0.851 | 0.229 | 0.15 |
| 159 | RAKEL-NB | 0.252 | 0.809 | 0.893 | 0.734 | 0.825 | 0.242 | 0.125 |
| 160 | RAKEL-XGB | 0.225 | 0.808 | 0.948 | 0.766 | 0.851 | 0.24 | 0.15 |
| 161 | RAKEL-CART | 0.22 | 0.799 | 0.964 | 0.78 | 0.855 | 0.269 | 0.2 |
| 162 | BRPLUS-RF | 0.22 | 0.819 | 0.948 | 0.773 | 0.854 | 0.206 | 0.125 |
| 163 | BRPLUS-SVM | 0.225 | 0.8 | 0.963 | 0.775 | 0.853 | 0.235 | 0.25 |
| 164 | BRPLUS-SMO | 0.21 | 0.822 | 0.941 | 0.778 | 0.857 | 0.221 | 0.15 |
| 165 | BRPLUS-C5.0 | 0.233 | 0.822 | 0.909 | 0.75 | 0.835 | 0.248 | 0.1 |
| 166 | BRPLUS-NB | 0.34 | 0.81 | 0.765 | 0.634 | 0.756 | 0.233 | 0.2 |
| 167 | BRPLUS-XGB | 0.267 | 0.8 | 0.903 | 0.727 | 0.821 | 0.235 | 0.25 |
| 168 | BRPLUS-CART | 0.217 | 0.82 | 0.941 | 0.771 | 0.854 | 0.215 | 0.075 |
| 169 | BR-RF | 0.164 | 0.723 | 0.493 | 0.404 | 0.561 | 0.18 | 0.1 |
| 170 | BR-SVM | 0.163 | 0.745 | 0.461 | 0.373 | 0.53 | 0.201 | 0.25 |
| 171 | BR-SMO | 0.173 | 0.676 | 0.496 | 0.379 | 0.535 | 0.298 | 0.392 |
| 172 | BR-C5.0 | 0.201 | 0.565 | 0.48 | 0.333 | 0.489 | 0.232 | 0.3 |
| 173 | BR-NB | 0.338 | 0.381 | 0.602 | 0.27 | 0.415 | 0.333 | 0.667 |
| 174 | BR-XGB | 0.195 | 0.567 | 0.515 | 0.352 | 0.512 | 0.206 | 0.275 |
| 175 | BR-CART | 0.182 | 0.608 | 0.522 | 0.368 | 0.53 | 0.223 | 0.125 |
| 176 | ECC-RF | 0.193 | 0.584 | 0.622 | 0.396 | 0.561 | 0.206 | 0.125 |
| 177 | ECC-SVM | 0.189 | 0.591 | 0.589 | 0.383 | 0.546 | 0.219 | 0.175 |
| 178 | ECC-SMO | 0.208 | 0.556 | 0.575 | 0.354 | 0.516 | 0.213 | 0.225 |
| 179 | ECC-C5.0 | 0.2 | 0.571 | 0.569 | 0.372 | 0.531 | 0.206 | 0.125 |
| 180 | ECC-NB | 0.203 | 0.551 | 0.543 | 0.348 | 0.507 | 0.215 | 0.258 |
| 181 | ECC-XGB | 0.203 | 0.549 | 0.59 | 0.367 | 0.532 | 0.199 | 0.1 |
| 182 | ECC-CART | 0.199 | 0.55 | 0.568 | 0.364 | 0.526 | 0.212 | 0.25 |
| 183 | LP-RF | 0.224 | 0.556 | 0.52 | 0.328 | 0.482 | 0.324 | 0.533 |
| 184 | LP-SVM | 0.21 | 0.642 | 0.251 | 0.183 | 0.298 | 0.392 | 0.683 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 185 | LP-SMO | 0.232 | 0.503 | 0.496 | 0.311 | 0.458 | 0.344 | 0.642 |
| 186 | LP-C5.0 | 0.202 | 0.593 | 0.486 | 0.338 | 0.493 | 0.323 | 0.375 |
| 187 | LP-NB | 0.235 | 0.479 | 0.467 | 0.286 | 0.431 | 0.358 | 0.633 |
| 188 | LP-XGB | 0.196 | 0.769 | 0.177 | 0.16 | 0.262 | 0.429 | 0.225 |
| 189 | LP-CART | 0.19 | 0.693 | 0.284 | 0.238 | 0.368 | 0.404 | 0.2 |
| 190 | HOMER-RF | 0.273 | 0.206 | 0.075 | 0.057 | 0.102 | 0.71 | 0.717 |
| 191 | HOMER-SVM | 0.285 | 0.178 | 0.074 | 0.053 | 0.097 | 0.714 | 0.85 |
| 192 | HOMER-SMO | 0.277 | 0.208 | 0.09 | 0.066 | 0.117 | 0.55 | 0.658 |
| 193 | HOMER-C5.0 | 0.27 | 0.179 | 0.064 | 0.051 | 0.09 | 0.717 | 0.875 |
| 194 | HOMER-NB | 0.289 | 0.215 | 0.082 | 0.061 | 0.108 | 0.699 | 0.742 |
| 195 | HOMER-XGB | 0.277 | 0.223 | 0.087 | 0.069 | 0.12 | 0.708 | 0.692 |
| 196 | HOMER-CART | 0.27 | 0.235 | 0.067 | 0.052 | 0.095 | 0.692 | 0.667 |
| 197 | RAKEL-RF | 0.162 | 0.721 | 0.5 | 0.403 | 0.562 | 0.26 | 0.25 |
| 198 | RAKEL-SVM | 0.16 | 0.784 | 0.439 | 0.373 | 0.525 | 0.271 | 0.275 |
| 199 | RAKEL-SMO | 0.172 | 0.686 | 0.509 | 0.388 | 0.546 | 0.26 | 0.225 |
| 200 | RAKEL-C5.0 | 0.187 | 0.613 | 0.515 | 0.364 | 0.524 | 0.236 | 0.233 |
| 201 | RAKEL-NB | 0.27 | 0.438 | 0.586 | 0.306 | 0.461 | 0.267 | 0.417 |
| 202 | RAKEL-XGB | 0.181 | 0.632 | 0.534 | 0.385 | 0.545 | 0.234 | 0.275 |
| 203 | RAKEL-CART | 0.17 | 0.659 | 0.528 | 0.387 | 0.55 | 0.231 | 0.35 |
| 204 | BRPLUS-RF | 0.162 | 0.756 | 0.447 | 0.383 | 0.538 | 0.174 | 0.125 |
| 205 | BRPLUS-SVM | 0.165 | 0.743 | 0.456 | 0.371 | 0.525 | 0.197 | 0.275 |
| 206 | BRPLUS-SMO | 0.185 | 0.665 | 0.488 | 0.357 | 0.516 | 0.307 | 0.333 |
| 207 | BRPLUS-C5.0 | 0.195 | 0.617 | 0.488 | 0.357 | 0.508 | 0.23 | 0.125 |
| 208 | BRPLUS-NB | 0.264 | 0.448 | 0.577 | 0.307 | 0.46 | 0.265 | 0.425 |
| 209 | BRPLUS-XGB | 0.198 | 0.579 | 0.521 | 0.352 | 0.513 | 0.195 | 0.175 |
| 210 | BRPLUS-CART | 0.193 | 0.58 | 0.522 | 0.355 | 0.516 | 0.22 | 0.2 |
| 211 | BR-RF | 0.189 | 0.72 | 0.51 | 0.414 | 0.571 | 0.187 | 0.125 |
| 212 | BR-SVM | 0.19 | 0.744 | 0.46 | 0.38 | 0.536 | 0.211 | 0.225 |
| 213 | BR-SMO | 0.199 | 0.679 | 0.511 | 0.392 | 0.548 | 0.287 | 0.392 |
| 214 | BR-C5.0 | 0.233 | 0.566 | 0.495 | 0.342 | 0.5 | 0.244 | 0.3 |
| 215 | BR-NB | 0.377 | 0.395 | 0.62 | 0.284 | 0.431 | 0.358 | 0.642 |
| 216 | BR-XGB | 0.227 | 0.567 | 0.532 | 0.363 | 0.523 | 0.217 | 0.275 |
| 217 | BR-CART | 0.211 | 0.608 | 0.537 | 0.377 | 0.54 | 0.239 | 0.125 |
| 218 | ECC-RF | 0.233 | 0.572 | 0.622 | 0.392 | 0.554 | 0.202 | 0.125 |
| 219 | ECC-SVM | 0.219 | 0.608 | 0.569 | 0.384 | 0.543 | 0.21 | 0.175 |
| 220 | ECC-SMO | 0.239 | 0.562 | 0.589 | 0.368 | 0.529 | 0.209 | 0.075 |
| 221 | ECC-C5.0 | 0.228 | 0.571 | 0.584 | 0.387 | 0.549 | 0.225 | 0.175 |
| 222 | ECC-NB | 0.244 | 0.534 | 0.575 | 0.353 | 0.513 | 0.221 | 0.1 |
| 223 | ECC-XGB | 0.233 | 0.565 | 0.575 | 0.369 | 0.531 | 0.209 | 0.175 |
| 224 | ECC-CART | 0.23 | 0.55 | 0.577 | 0.369 | 0.532 | 0.226 | 0.125 |
| 225 | LP-RF | 0.249 | 0.578 | 0.553 | 0.358 | 0.513 | 0.311 | 0.408 |
| 226 | LP-SVM | 0.244 | 0.644 | 0.257 | 0.189 | 0.306 | 0.371 | 0.683 |
| 227 | LP-SMO | 0.262 | 0.518 | 0.508 | 0.329 | 0.476 | 0.337 | 0.542 |
| 228 | LP-C5.0 | 0.236 | 0.593 | 0.493 | 0.343 | 0.497 | 0.321 | 0.325 |
| 229 | LP-NB | 0.265 | 0.493 | 0.478 | 0.298 | 0.445 | 0.35 | 0.583 |
| 230 | LP-XGB | 0.228 | 0.769 | 0.181 | 0.164 | 0.269 | 0.404 | 0.225 |
| 231 | LP-CART | 0.215 | 0.689 | 0.372 | 0.295 | 0.434 | 0.331 | 0.35 |
| 232 | HOMER-RF | 0.311 | 0.2 | 0.075 | 0.059 | 0.1 | 0.714 | 0.767 |
| 233 | HOMER-SVM | 0.356 | 0.126 | 0.067 | 0.047 | 0.082 | 0.715 | 0.925 |
| 234 | HOMER-SMO | 0.32 | 0.224 | 0.099 | 0.078 | 0.13 | 0.519 | 0.7 |
| 235 | HOMER-C5.0 | 0.316 | 0.186 | 0.074 | 0.056 | 0.097 | 0.721 | 0.817 |
| 236 | HOMER-NB | 0.334 | 0.171 | 0.066 | 0.048 | 0.085 | 0.684 | 0.875 |
| 237 | HOMER-XGB | 0.313 | 0.203 | 0.06 | 0.047 | 0.083 | 0.723 | 0.817 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 238 | HOMER-CART | 0.315 | 0.172 | 0.051 | 0.041 | 0.074 | 0.693 | 0.842 |
| 239 | RAKEL-RF | 0.187 | 0.703 | 0.52 | 0.412 | 0.573 | 0.238 | 0.175 |
| 240 | RAKEL-SVM | 0.189 | 0.764 | 0.443 | 0.368 | 0.524 | 0.272 | 0.225 |
| 241 | RAKEL-SMO | 0.2 | 0.679 | 0.527 | 0.396 | 0.554 | 0.249 | 0.233 |
| 242 | RAKEL-C5.0 | 0.221 | 0.588 | 0.544 | 0.373 | 0.532 | 0.246 | 0.233 |
| 243 | RAKEL-NB | 0.301 | 0.45 | 0.624 | 0.327 | 0.485 | 0.262 | 0.458 |
| 244 | RAKEL-XGB | 0.206 | 0.627 | 0.558 | 0.4 | 0.561 | 0.23 | 0.25 |
| 245 | RAKEL-CART | 0.197 | 0.655 | 0.534 | 0.394 | 0.556 | 0.23 | 0.275 |
| 246 | BRPLUS-RF | 0.185 | 0.76 | 0.467 | 0.397 | 0.554 | 0.184 | 0.1 |
| 247 | BRPLUS-SVM | 0.191 | 0.759 | 0.454 | 0.378 | 0.533 | 0.202 | 0.275 |
| 248 | BRPLUS-SMO | 0.21 | 0.679 | 0.513 | 0.382 | 0.542 | 0.295 | 0.3 |
| 249 | BRPLUS-C5.0 | 0.248 | 0.559 | 0.483 | 0.334 | 0.485 | 0.255 | 0.125 |
| 250 | BRPLUS-NB | 0.302 | 0.455 | 0.602 | 0.321 | 0.474 | 0.279 | 0.425 |
| 251 | BRPLUS-XGB | 0.23 | 0.581 | 0.538 | 0.365 | 0.526 | 0.206 | 0.175 |
| 252 | BRPLUS-CART | 0.224 | 0.58 | 0.538 | 0.365 | 0.526 | 0.232 | 0.2 |
| 253 | BR-RF | 0.215 | 0.72 | 0.523 | 0.426 | 0.582 | 0.194 | 0.15 |
| 254 | BR-SVM | 0.207 | 0.758 | 0.497 | 0.409 | 0.566 | 0.224 | 0.2 |
| 255 | BR-SMO | 0.223 | 0.685 | 0.528 | 0.411 | 0.567 | 0.292 | 0.367 |
| 256 | BR-C5.0 | 0.262 | 0.576 | 0.512 | 0.357 | 0.515 | 0.253 | 0.3 |
| 257 | BR-NB | 0.395 | 0.419 | 0.64 | 0.31 | 0.461 | 0.369 | 0.592 |
| 258 | BR-XGB | 0.258 | 0.573 | 0.551 | 0.377 | 0.537 | 0.228 | 0.275 |
| 259 | BR-CART | 0.24 | 0.608 | 0.561 | 0.389 | 0.554 | 0.248 | 0.125 |
| 260 | ECC-RF | 0.247 | 0.595 | 0.614 | 0.409 | 0.573 | 0.216 | 0.2 |
| 261 | ECC-SVM | 0.265 | 0.563 | 0.598 | 0.375 | 0.539 | 0.236 | 0.175 |
| 262 | ECC-SMO | 0.257 | 0.581 | 0.6 | 0.384 | 0.55 | 0.229 | 0.175 |
| 263 | ECC-C5.0 | 0.247 | 0.609 | 0.61 | 0.403 | 0.568 | 0.212 | 0.125 |
| 264 | ECC-NB | 0.264 | 0.579 | 0.572 | 0.37 | 0.531 | 0.226 | 0.225 |
| 265 | ECC-XGB | 0.252 | 0.595 | 0.614 | 0.405 | 0.569 | 0.221 | 0.15 |
| 266 | ECC-CART | 0.248 | 0.583 | 0.617 | 0.409 | 0.572 | 0.226 | 0.125 |
| 267 | LP-RF | 0.273 | 0.596 | 0.571 | 0.383 | 0.539 | 0.318 | 0.383 |
| 268 | LP-SVM | 0.268 | 0.74 | 0.265 | 0.201 | 0.323 | 0.373 | 0.3 |
| 269 | LP-SMO | 0.294 | 0.533 | 0.522 | 0.35 | 0.496 | 0.35 | 0.517 |
| 270 | LP-C5.0 | 0.266 | 0.603 | 0.507 | 0.361 | 0.515 | 0.332 | 0.3 |
| 271 | LP-NB | 0.296 | 0.511 | 0.492 | 0.314 | 0.462 | 0.364 | 0.533 |
| 272 | LP-XGB | 0.26 | 0.783 | 0.188 | 0.175 | 0.28 | 0.411 | 0.175 |
| 273 | LP-CART | 0.243 | 0.704 | 0.387 | 0.31 | 0.45 | 0.338 | 0.3 |
| 274 | HOMER-RF | 0.366 | 0.269 | 0.153 | 0.111 | 0.183 | 0.684 | 0.775 |
| 275 | HOMER-SVM | 0.407 | 0.221 | 0.149 | 0.098 | 0.164 | 0.707 | 0.817 |
| 276 | HOMER-SMO | 0.376 | 0.284 | 0.176 | 0.126 | 0.207 | 0.529 | 0.792 |
| 277 | HOMER-C5.0 | 0.356 | 0.303 | 0.154 | 0.114 | 0.188 | 0.676 | 0.608 |
| 278 | HOMER-NB | 0.373 | 0.233 | 0.117 | 0.083 | 0.142 | 0.668 | 0.792 |
| 279 | HOMER-XGB | 0.37 | 0.266 | 0.151 | 0.108 | 0.179 | 0.683 | 0.667 |
| 280 | HOMER-CART | 0.38 | 0.221 | 0.119 | 0.081 | 0.142 | 0.686 | 0.742 |
| 281 | RAKEL-RF | 0.212 | 0.722 | 0.544 | 0.436 | 0.593 | 0.248 | 0.175 |
| 282 | RAKEL-SVM | 0.212 | 0.79 | 0.441 | 0.377 | 0.53 | 0.277 | 0.2 |
| 283 | RAKEL-SMO | 0.227 | 0.698 | 0.526 | 0.411 | 0.564 | 0.271 | 0.175 |
| 284 | RAKEL-C5.0 | 0.229 | 0.636 | 0.575 | 0.413 | 0.574 | 0.228 | 0.158 |
| 285 | RAKEL-NB | 0.308 | 0.484 | 0.596 | 0.344 | 0.503 | 0.279 | 0.333 |
| 286 | RAKEL-XGB | 0.239 | 0.618 | 0.565 | 0.406 | 0.567 | 0.236 | 0.25 |
| 287 | RAKEL-CART | 0.229 | 0.667 | 0.506 | 0.381 | 0.541 | 0.247 | 0.3 |
| 288 | BRPLUS-RF | 0.211 | 0.754 | 0.491 | 0.411 | 0.567 | 0.195 | 0.15 |
| 289 | BRPLUS-SVM | 0.213 | 0.755 | 0.485 | 0.394 | 0.552 | 0.21 | 0.275 |
| 290 | BRPLUS-SMO | 0.228 | 0.702 | 0.527 | 0.409 | 0.565 | 0.295 | 0.25 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 291 | BRPLUS-C5.0 | 0.268 | 0.578 | 0.522 | 0.363 | 0.517 | 0.244 | 0.125 |
| 292 | BRPLUS-NB | 0.333 | 0.468 | 0.626 | 0.342 | 0.496 | 0.3 | 0.342 |
| 293 | BRPLUS-XGB | 0.258 | 0.592 | 0.559 | 0.379 | 0.541 | 0.224 | 0.175 |
| 294 | BRPLUS-CART | 0.257 | 0.58 | 0.56 | 0.376 | 0.538 | 0.238 | 0.2 |

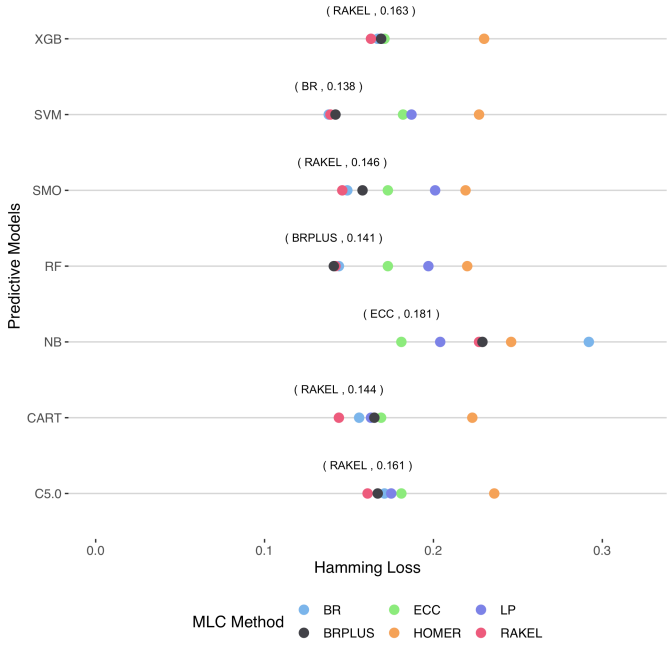



**Figure H. 15.** Hamming-Loss for all MLC strategies applied in CDS3.

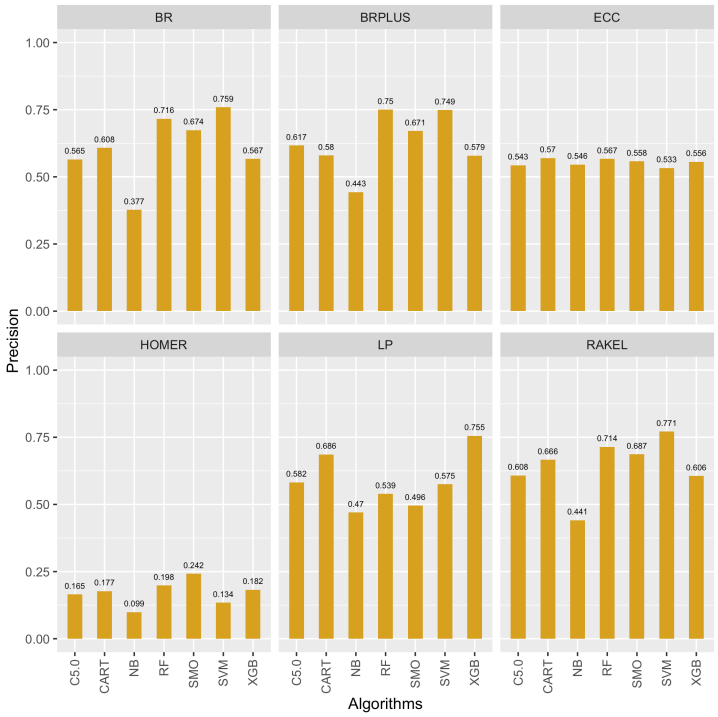



**Figure H. 16.** Precision for all MLC strategies applied in CDS3.

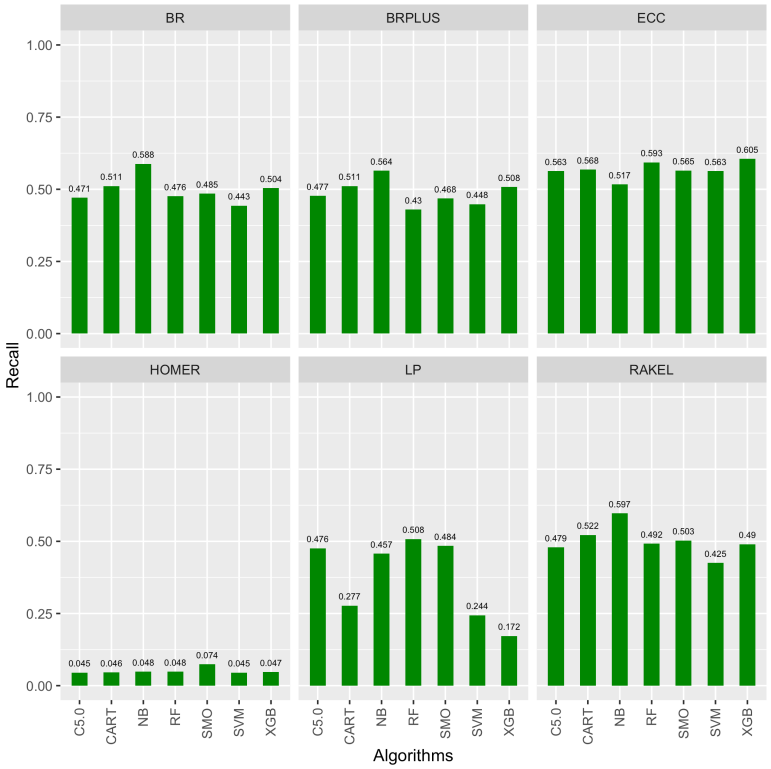



**Figure H. 17.** Recall for all MLC strategies applied in CDS3.

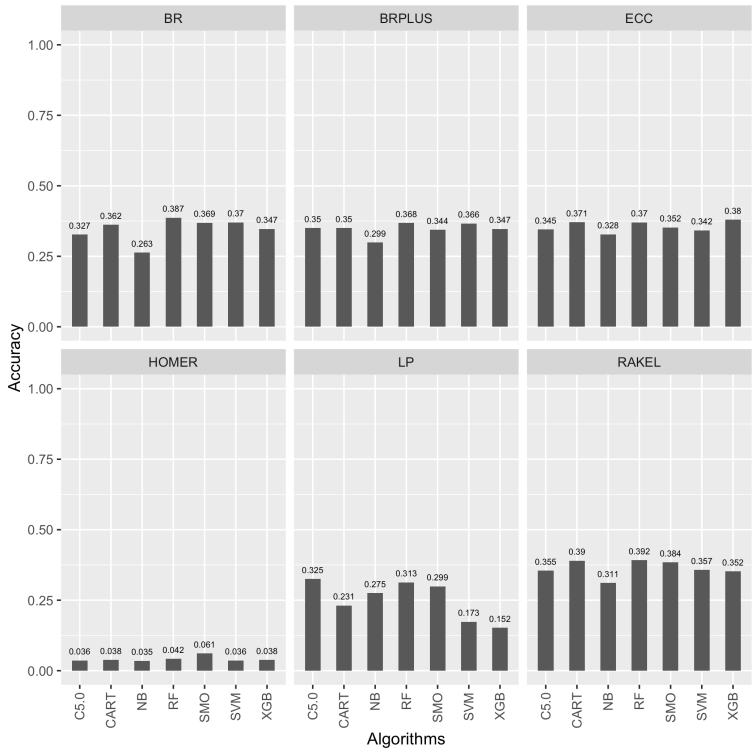



**Figure H. 18.** Accuracy for all MLC strategies applied in CDS3.

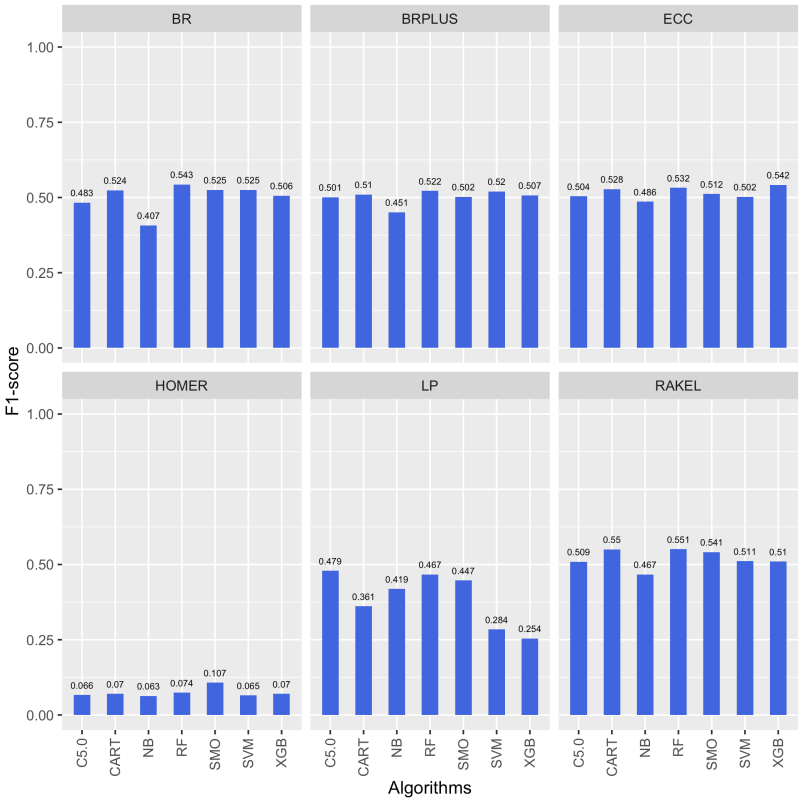



**Figure H. 19.** F1-Score for all MLC strategies applied in CDS3.

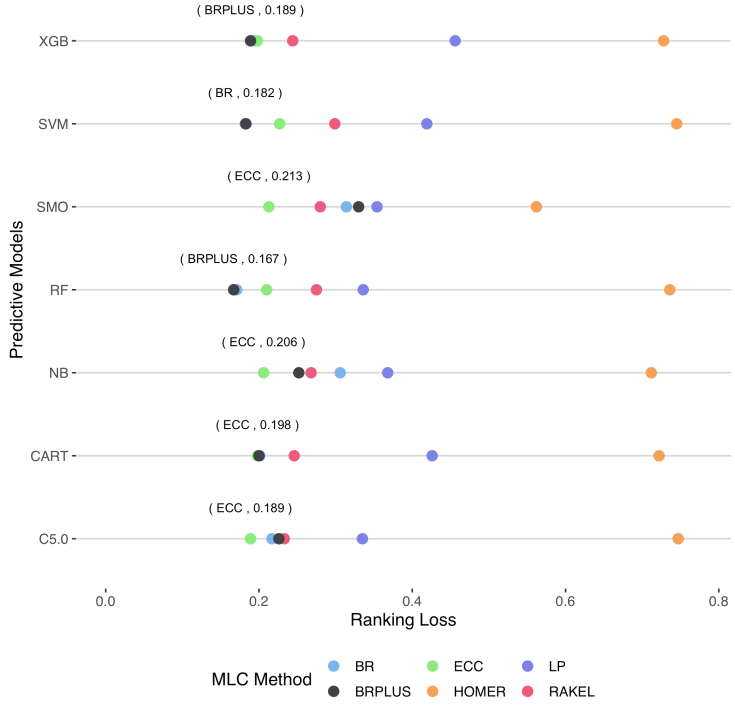



**Figure H. 20.** Ranking-Loss for all MLC strategies applied in CDS3.

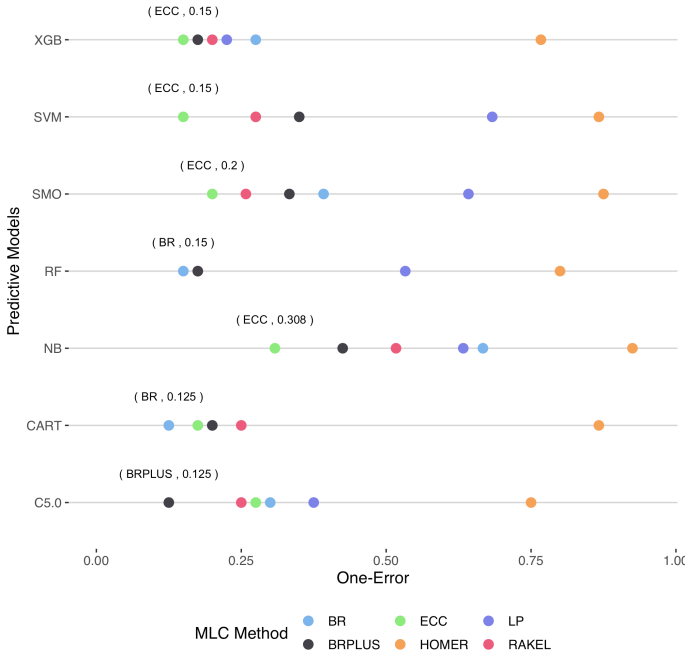



**Figure H. 21.** One-Error for all MLC strategies applied in CDS3.

# H.4. Performance Measures for Models (CDS4)

| | mlc_methods | values_hl | values_pr | values_re | values_ac | values_f1 | values_rl | values_oe |
|---|---|---|---|---|---|---|---|---|
| 1 | BR-RF | 0.143 | 0.582 | 0.515 | 0.386 | 0.514 | 0.106 | 0.18 |
| 2 | BR-SVM | 0.148 | 0.665 | 0.351 | 0.3 | 0.429 | 0.156 | 0.313 |
| 3 | BR-SMO | 0.138 | 0.65 | 0.412 | 0.34 | 0.469 | 0.263 | 0.333 |
| 4 | BR-C5.0 | 0.137 | 0.616 | 0.383 | 0.32 | 0.449 | 0.165 | 0.243 |
| 5 | BR-NB | 0.17 | 0.479 | 0.468 | 0.318 | 0.439 | 0.147 | 0.303 |
| 6 | BR-XGB | 0.141 | 0.588 | 0.505 | 0.378 | 0.508 | 0.114 | 0.223 |
| 7 | BR-CART | 0.155 | 0.526 | 0.395 | 0.292 | 0.42 | 0.234 | 0.287 |
| 8 | ECC-RF | 0.169 | 0.534 | 0.531 | 0.344 | 0.477 | 0.163 | 0.273 |
| 9 | ECC-SVM | 0.166 | 0.538 | 0.517 | 0.347 | 0.48 | 0.175 | 0.223 |
| 10 | ECC-SMO | 0.166 | 0.547 | 0.477 | 0.336 | 0.461 | 0.17 | 0.257 |
| 11 | ECC-C5.0 | 0.152 | 0.56 | 0.496 | 0.363 | 0.494 | 0.157 | 0.24 |
| 12 | ECC-NB | 0.163 | 0.511 | 0.502 | 0.334 | 0.462 | 0.133 | 0.31 |
| 13 | ECC-XGB | 0.164 | 0.55 | 0.544 | 0.352 | 0.492 | 0.15 | 0.227 |
| 14 | ECC-CART | 0.164 | 0.535 | 0.528 | 0.365 | 0.499 | 0.185 | 0.243 |
| 15 | LP-RF | 0.154 | 0.494 | 0.47 | 0.359 | 0.461 | 0.274 | 0.493 |
| 16 | LP-SVM | 0.184 | 0.21 | 0.106 | 0.106 | 0.114 | 0.441 | 0.773 |
| 17 | LP-SMO | 0.183 | 0.386 | 0.372 | 0.271 | 0.356 | 0.327 | 0.563 |
| 18 | LP-C5.0 | 0.147 | 0.46 | 0.4 | 0.333 | 0.411 | 0.297 | 0.493 |
| 19 | LP-NB | 0.132 | 0.573 | 0.562 | 0.447 | 0.546 | 0.208 | 0.427 |
| 20 | LP-XGB | 0.199 | 0.282 | 0.275 | 0.201 | 0.26 | 0.376 | 0.69 |
| 21 | LP-CART | 0.168 | 0.459 | 0.343 | 0.283 | 0.362 | 0.322 | 0.547 |
| 22 | HOMER-RF | 0.23 | 0.207 | 0.092 | 0.071 | 0.118 | 0.653 | 0.763 |
| 23 | HOMER-SVM | 0.22 | 0.214 | 0.078 | 0.063 | 0.107 | 0.664 | 0.69 |
| 24 | HOMER-SMO | 0.211 | 0.204 | 0.073 | 0.058 | 0.098 | 0.447 | 0.89 |
| 25 | HOMER-C5.0 | 0.211 | 0.261 | 0.074 | 0.063 | 0.107 | 0.653 | 0.67 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 26 | HOMER-NB | 0.217 | 0.228 | 0.074 | 0.063 | 0.107 | 0.653 | 0.717 |
| 27 | HOMER-XGB | 0.227 | 0.194 | 0.069 | 0.053 | 0.093 | 0.641 | 0.703 |
| 28 | HOMER-CART | 0.228 | 0.173 | 0.064 | 0.049 | 0.087 | 0.63 | 0.73 |
| 29 | RAKEL-RF | 0.133 | 0.616 | 0.468 | 0.372 | 0.495 | 0.205 | 0.267 |
| 30 | RAKEL-SVM | 0.142 | 0.691 | 0.345 | 0.306 | 0.431 | 0.248 | 0.273 |
| 31 | RAKEL-SMO | 0.138 | 0.649 | 0.393 | 0.327 | 0.453 | 0.234 | 0.293 |
| 32 | RAKEL-C5.0 | 0.136 | 0.638 | 0.398 | 0.339 | 0.468 | 0.208 | 0.29 |
| 33 | RAKEL-NB | 0.178 | 0.457 | 0.488 | 0.325 | 0.441 | 0.186 | 0.42 |
| 34 | RAKEL-XGB | 0.136 | 0.607 | 0.476 | 0.382 | 0.505 | 0.174 | 0.253 |
| 35 | RAKEL-CART | 0.148 | 0.565 | 0.387 | 0.304 | 0.428 | 0.21 | 0.37 |
| 36 | BRPLUS-RF | 0.14 | 0.671 | 0.476 | 0.39 | 0.513 | 0.106 | 0.16 |
| 37 | BRPLUS-SVM | 0.144 | 0.592 | 0.311 | 0.269 | 0.376 | 0.161 | 0.26 |
| 38 | BRPLUS-SMO | 0.143 | 0.618 | 0.417 | 0.333 | 0.46 | 0.28 | 0.407 |
| 39 | BRPLUS-C5.0 | 0.14 | 0.64 | 0.445 | 0.369 | 0.49 | 0.167 | 0.237 |
| 40 | BRPLUS-NB | 0.17 | 0.433 | 0.469 | 0.324 | 0.431 | 0.132 | 0.42 |
| 41 | BRPLUS-XGB | 0.144 | 0.602 | 0.497 | 0.382 | 0.507 | 0.132 | 0.187 |
| 42 | BRPLUS-CART | 0.159 | 0.52 | 0.441 | 0.313 | 0.443 | 0.211 | 0.35 |
| 43 | BR-RF | 0.2 | 0.596 | 0.559 | 0.425 | 0.55 | 0.131 | 0.18 |
| 44 | BR-SVM | 0.218 | 0.674 | 0.382 | 0.329 | 0.462 | 0.211 | 0.243 |
| 45 | BR-SMO | 0.205 | 0.657 | 0.442 | 0.367 | 0.497 | 0.267 | 0.35 |
| 46 | BR-C5.0 | 0.204 | 0.617 | 0.418 | 0.348 | 0.478 | 0.217 | 0.243 |
| 47 | BR-NB | 0.227 | 0.504 | 0.501 | 0.365 | 0.48 | 0.174 | 0.333 |
| 48 | BR-XGB | 0.212 | 0.594 | 0.535 | 0.403 | 0.532 | 0.141 | 0.223 |
| 49 | BR-CART | 0.237 | 0.526 | 0.431 | 0.317 | 0.447 | 0.276 | 0.287 |
| 50 | ECC-RF | 0.222 | 0.61 | 0.588 | 0.44 | 0.569 | 0.172 | 0.197 |
| 51 | ECC-SVM | 0.233 | 0.571 | 0.53 | 0.381 | 0.512 | 0.191 | 0.293 |
| 52 | ECC-SMO | 0.237 | 0.579 | 0.527 | 0.365 | 0.498 | 0.191 | 0.273 |
| 53 | ECC-C5.0 | 0.23 | 0.562 | 0.545 | 0.38 | 0.514 | 0.187 | 0.273 |
| 54 | ECC-NB | 0.233 | 0.565 | 0.586 | 0.397 | 0.527 | 0.169 | 0.237 |
| 55 | ECC-XGB | 0.242 | 0.559 | 0.539 | 0.373 | 0.512 | 0.186 | 0.24 |
| 56 | ECC-CART | 0.255 | 0.553 | 0.52 | 0.361 | 0.498 | 0.229 | 0.243 |
| 57 | LP-RF | 0.213 | 0.545 | 0.545 | 0.417 | 0.523 | 0.266 | 0.443 |
| 58 | LP-SVM | 0.286 | 0.21 | 0.107 | 0.107 | 0.115 | 0.441 | 0.69 |
| 59 | LP-SMO | 0.262 | 0.393 | 0.383 | 0.286 | 0.371 | 0.34 | 0.51 |
| 60 | LP-C5.0 | 0.224 | 0.464 | 0.425 | 0.352 | 0.431 | 0.306 | 0.477 |
| 61 | LP-NB | 0.194 | 0.599 | 0.569 | 0.464 | 0.567 | 0.228 | 0.427 |
| 62 | LP-XGB | 0.276 | 0.315 | 0.309 | 0.24 | 0.301 | 0.372 | 0.637 |
| 63 | LP-CART | 0.251 | 0.445 | 0.383 | 0.3 | 0.387 | 0.334 | 0.513 |
| 64 | HOMER-RF | 0.332 | 0.389 | 0.311 | 0.215 | 0.312 | 0.479 | 0.423 |
| 65 | HOMER-SVM | 0.371 | 0.308 | 0.224 | 0.144 | 0.234 | 0.525 | 0.57 |
| 66 | HOMER-SMO | 0.313 | 0.407 | 0.267 | 0.184 | 0.285 | 0.409 | 0.607 |
| 67 | HOMER-C5.0 | 0.329 | 0.376 | 0.282 | 0.174 | 0.277 | 0.498 | 0.597 |
| 68 | HOMER-NB | 0.303 | 0.389 | 0.228 | 0.167 | 0.26 | 0.505 | 0.637 |
| 69 | HOMER-XGB | 0.322 | 0.374 | 0.257 | 0.163 | 0.256 | 0.492 | 0.5 |
| 70 | HOMER-CART | 0.334 | 0.36 | 0.229 | 0.156 | 0.248 | 0.505 | 0.52 |
| 71 | RAKEL-RF | 0.191 | 0.641 | 0.516 | 0.423 | 0.546 | 0.188 | 0.307 |
| 72 | RAKEL-SVM | 0.211 | 0.698 | 0.38 | 0.338 | 0.469 | 0.231 | 0.273 |
| 73 | RAKEL-SMO | 0.207 | 0.63 | 0.44 | 0.358 | 0.485 | 0.223 | 0.29 |
| 74 | RAKEL-C5.0 | 0.198 | 0.625 | 0.461 | 0.385 | 0.514 | 0.189 | 0.293 |
| 75 | RAKEL-NB | 0.224 | 0.502 | 0.557 | 0.395 | 0.507 | 0.176 | 0.407 |
| 76 | RAKEL-XGB | 0.193 | 0.617 | 0.514 | 0.407 | 0.532 | 0.155 | 0.23 |
| 77 | RAKEL-CART | 0.204 | 0.604 | 0.445 | 0.356 | 0.486 | 0.219 | 0.207 |
| 78 | BRPLUS-RF | 0.2 | 0.664 | 0.548 | 0.438 | 0.563 | 0.117 | 0.213 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 79 | BRPLUS-SVM | 0.219 | 0.605 | 0.349 | 0.299 | 0.414 | 0.204 | 0.297 |
| 80 | BRPLUS-SMO | 0.21 | 0.636 | 0.457 | 0.366 | 0.495 | 0.278 | 0.463 |
| 81 | BRPLUS-C5.0 | 0.215 | 0.616 | 0.5 | 0.398 | 0.52 | 0.212 | 0.237 |
| 82 | BRPLUS-NB | 0.215 | 0.476 | 0.538 | 0.382 | 0.488 | 0.149 | 0.33 |
| 83 | BRPLUS-XGB | 0.212 | 0.608 | 0.549 | 0.424 | 0.546 | 0.17 | 0.187 |
| 84 | BRPLUS-CART | 0.243 | 0.523 | 0.488 | 0.345 | 0.474 | 0.236 | 0.35 |
| 85 | BR-RF | 0.239 | 0.625 | 0.641 | 0.508 | 0.613 | 0.136 | 0.207 |
| 86 | BR-SVM | 0.293 | 0.694 | 0.514 | 0.45 | 0.576 | 0.157 | 0.247 |
| 87 | BR-SMO | 0.269 | 0.683 | 0.553 | 0.477 | 0.59 | 0.195 | 0.277 |
| 88 | BR-C5.0 | 0.271 | 0.633 | 0.542 | 0.455 | 0.57 | 0.162 | 0.243 |
| 89 | BR-NB | 0.284 | 0.579 | 0.562 | 0.457 | 0.557 | 0.154 | 0.243 |
| 90 | BR-XGB | 0.258 | 0.615 | 0.588 | 0.482 | 0.583 | 0.142 | 0.21 |
| 91 | BR-CART | 0.303 | 0.615 | 0.525 | 0.424 | 0.54 | 0.2 | 0.287 |
| 92 | ECC-RF | 0.255 | 0.62 | 0.614 | 0.492 | 0.6 | 0.15 | 0.24 |
| 93 | ECC-SVM | 0.291 | 0.646 | 0.581 | 0.482 | 0.598 | 0.155 | 0.24 |
| 94 | ECC-SMO | 0.307 | 0.637 | 0.566 | 0.467 | 0.584 | 0.167 | 0.223 |
| 95 | ECC-C5.0 | 0.288 | 0.633 | 0.607 | 0.491 | 0.603 | 0.138 | 0.21 |
| 96 | ECC-NB | 0.286 | 0.603 | 0.564 | 0.457 | 0.563 | 0.164 | 0.257 |
| 97 | ECC-XGB | 0.293 | 0.62 | 0.583 | 0.469 | 0.585 | 0.155 | 0.243 |
| 98 | ECC-CART | 0.308 | 0.642 | 0.558 | 0.455 | 0.578 | 0.178 | 0.263 |
| 99 | LP-RF | 0.243 | 0.56 | 0.546 | 0.464 | 0.544 | 0.237 | 0.41 |
| 100 | LP-SVM | 0.395 | 0.254 | 0.227 | 0.183 | 0.235 | 0.359 | 0.48 |
| 101 | LP-SMO | 0.319 | 0.445 | 0.418 | 0.345 | 0.423 | 0.279 | 0.44 |
| 102 | LP-C5.0 | 0.3 | 0.435 | 0.416 | 0.345 | 0.417 | 0.3 | 0.45 |
| 103 | LP-NB | 0.249 | 0.668 | 0.658 | 0.566 | 0.654 | 0.187 | 0.317 |
| 104 | LP-XGB | 0.298 | 0.452 | 0.421 | 0.354 | 0.43 | 0.28 | 0.383 |
| 105 | LP-CART | 0.31 | 0.471 | 0.444 | 0.361 | 0.448 | 0.286 | 0.45 |
| 106 | HOMER-RF | 0.488 | 0.448 | 0.27 | 0.214 | 0.321 | 0.513 | 0.39 |
| 107 | HOMER-SVM | 0.518 | 0.411 | 0.272 | 0.204 | 0.317 | 0.531 | 0.637 |
| 108 | HOMER-SMO | 0.476 | 0.476 | 0.277 | 0.225 | 0.339 | 0.383 | 0.54 |
| 109 | HOMER-C5.0 | 0.478 | 0.452 | 0.242 | 0.198 | 0.305 | 0.509 | 0.457 |
| 110 | HOMER-NB | 0.467 | 0.487 | 0.207 | 0.183 | 0.278 | 0.521 | 0.44 |
| 111 | HOMER-XGB | 0.46 | 0.473 | 0.25 | 0.21 | 0.315 | 0.507 | 0.407 |
| 112 | HOMER-CART | 0.453 | 0.472 | 0.232 | 0.196 | 0.302 | 0.527 | 0.427 |
| 113 | RAKEL-RF | 0.213 | 0.652 | 0.6 | 0.516 | 0.612 | 0.168 | 0.31 |
| 114 | RAKEL-SVM | 0.303 | 0.685 | 0.494 | 0.432 | 0.548 | 0.173 | 0.26 |
| 115 | RAKEL-SMO | 0.276 | 0.665 | 0.535 | 0.46 | 0.564 | 0.161 | 0.257 |
| 116 | RAKEL-C5.0 | 0.249 | 0.65 | 0.57 | 0.486 | 0.593 | 0.178 | 0.26 |
| 117 | RAKEL-NB | 0.294 | 0.561 | 0.568 | 0.454 | 0.555 | 0.18 | 0.26 |
| 118 | RAKEL-XGB | 0.259 | 0.587 | 0.596 | 0.472 | 0.573 | 0.158 | 0.26 |
| 119 | RAKEL-CART | 0.287 | 0.594 | 0.541 | 0.444 | 0.552 | 0.19 | 0.277 |
| 120 | BRPLUS-RF | 0.229 | 0.641 | 0.636 | 0.52 | 0.619 | 0.131 | 0.207 |
| 121 | BRPLUS-SVM | 0.292 | 0.61 | 0.477 | 0.419 | 0.524 | 0.167 | 0.26 |
| 122 | BRPLUS-SMO | 0.266 | 0.661 | 0.595 | 0.502 | 0.606 | 0.197 | 0.39 |
| 123 | BRPLUS-C5.0 | 0.298 | 0.607 | 0.551 | 0.444 | 0.551 | 0.154 | 0.243 |
| 124 | BRPLUS-NB | 0.269 | 0.523 | 0.587 | 0.449 | 0.538 | 0.162 | 0.26 |
| 125 | BRPLUS-XGB | 0.265 | 0.618 | 0.619 | 0.493 | 0.598 | 0.146 | 0.227 |
| 126 | BRPLUS-CART | 0.297 | 0.615 | 0.56 | 0.445 | 0.56 | 0.187 | 0.26 |
| 127 | BR-RF | 0.138 | 0.733 | 0.783 | 0.708 | 0.749 | 0.075 | 0.19 |
| 128 | BR-SVM | 0.297 | 0.731 | 0.795 | 0.702 | 0.749 | 0.106 | 0.21 |
| 129 | BR-SMO | 0.223 | 0.734 | 0.789 | 0.702 | 0.75 | 0.11 | 0.24 |
| 130 | BR-C5.0 | 0.249 | 0.689 | 0.713 | 0.638 | 0.691 | 0.126 | 0.227 |
| 131 | BR-NB | 0.254 | 0.668 | 0.677 | 0.612 | 0.662 | 0.079 | 0.21 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 132 | BR-XGB | 0.267 | 0.693 | 0.705 | 0.631 | 0.684 | 0.096 | 0.21 |
| 133 | BR-CART | 0.327 | 0.68 | 0.666 | 0.592 | 0.657 | 0.14 | 0.283 |
| 134 | ECC-RF | 0.206 | 0.722 | 0.707 | 0.648 | 0.701 | 0.102 | 0.26 |
| 135 | ECC-SVM | 0.244 | 0.727 | 0.706 | 0.652 | 0.703 | 0.086 | 0.207 |
| 136 | ECC-SMO | 0.253 | 0.716 | 0.724 | 0.646 | 0.705 | 0.117 | 0.24 |
| 137 | ECC-C5.0 | 0.232 | 0.716 | 0.708 | 0.661 | 0.705 | 0.109 | 0.263 |
| 138 | ECC-NB | 0.277 | 0.665 | 0.669 | 0.619 | 0.657 | 0.102 | 0.207 |
| 139 | ECC-XGB | 0.272 | 0.704 | 0.677 | 0.631 | 0.678 | 0.096 | 0.243 |
| 140 | ECC-CART | 0.272 | 0.701 | 0.694 | 0.632 | 0.686 | 0.129 | 0.24 |
| 141 | LP-RF | 0.161 | 0.708 | 0.745 | 0.682 | 0.718 | 0.102 | 0.24 |
| 142 | LP-SVM | 0.313 | 0.689 | 0.767 | 0.682 | 0.72 | 0.11 | 0.24 |
| 143 | LP-SMO | 0.26 | 0.681 | 0.736 | 0.662 | 0.7 | 0.11 | 0.24 |
| 144 | LP-C5.0 | 0.296 | 0.673 | 0.688 | 0.629 | 0.673 | 0.105 | 0.24 |
| 145 | LP-NB | 0.246 | 0.669 | 0.658 | 0.599 | 0.653 | 0.076 | 0.207 |
| 146 | LP-XGB | 0.298 | 0.632 | 0.672 | 0.597 | 0.644 | 0.122 | 0.24 |
| 147 | LP-CART | 0.306 | 0.63 | 0.68 | 0.603 | 0.645 | 0.11 | 0.24 |
| 148 | HOMER-RF | 0.266 | 0.752 | 0.809 | 0.734 | 0.771 | 0.097 | 0.243 |
| 149 | HOMER-SVM | 0.272 | 0.728 | 0.827 | 0.728 | 0.768 | 0.122 | 0.227 |
| 150 | HOMER-SMO | 0.28 | 0.726 | 0.816 | 0.72 | 0.762 | 0.11 | 0.24 |
| 151 | HOMER-C5.0 | 0.277 | 0.742 | 0.808 | 0.723 | 0.765 | 0.126 | 0.223 |
| 152 | HOMER-NB | 0.312 | 0.757 | 0.689 | 0.633 | 0.705 | 0.101 | 0.243 |
| 153 | HOMER-XGB | 0.294 | 0.737 | 0.796 | 0.706 | 0.755 | 0.102 | 0.193 |
| 154 | HOMER-CART | 0.285 | 0.739 | 0.796 | 0.715 | 0.758 | 0.106 | 0.207 |
| 155 | RAKEL-RF | 0.14 | 0.726 | 0.767 | 0.702 | 0.738 | 0.102 | 0.24 |
| 156 | RAKEL-SVM | 0.297 | 0.716 | 0.796 | 0.702 | 0.745 | 0.11 | 0.24 |
| 157 | RAKEL-SMO | 0.24 | 0.71 | 0.761 | 0.682 | 0.725 | 0.11 | 0.24 |
| 158 | RAKEL-C5.0 | 0.252 | 0.701 | 0.705 | 0.649 | 0.691 | 0.095 | 0.223 |
| 159 | RAKEL-NB | 0.247 | 0.674 | 0.639 | 0.583 | 0.644 | 0.053 | 0.207 |
| 160 | RAKEL-XGB | 0.207 | 0.666 | 0.722 | 0.652 | 0.686 | 0.115 | 0.24 |
| 161 | RAKEL-CART | 0.318 | 0.665 | 0.693 | 0.618 | 0.667 | 0.11 | 0.24 |
| 162 | BRPLUS-RF | 0.132 | 0.739 | 0.783 | 0.714 | 0.752 | 0.096 | 0.223 |
| 163 | BRPLUS-SVM | 0.313 | 0.698 | 0.775 | 0.682 | 0.724 | 0.095 | 0.193 |
| 164 | BRPLUS-SMO | 0.207 | 0.726 | 0.816 | 0.72 | 0.762 | 0.11 | 0.24 |
| 165 | BRPLUS-C5.0 | 0.239 | 0.733 | 0.743 | 0.679 | 0.725 | 0.113 | 0.247 |
| 166 | BRPLUS-NB | 0.184 | 0.661 | 0.734 | 0.654 | 0.689 | 0.117 | 0.263 |
| 167 | BRPLUS-XGB | 0.263 | 0.689 | 0.74 | 0.656 | 0.704 | 0.108 | 0.227 |
| 168 | BRPLUS-CART | 0.322 | 0.668 | 0.71 | 0.623 | 0.674 | 0.128 | 0.28 |
| 169 | BR-RF | 0.157 | 0.591 | 0.523 | 0.397 | 0.524 | 0.115 | 0.18 |
| 170 | BR-SVM | 0.166 | 0.664 | 0.358 | 0.305 | 0.435 | 0.173 | 0.243 |
| 171 | BR-SMO | 0.154 | 0.652 | 0.419 | 0.347 | 0.476 | 0.259 | 0.333 |
| 172 | BR-C5.0 | 0.154 | 0.618 | 0.389 | 0.325 | 0.455 | 0.181 | 0.243 |
| 173 | BR-NB | 0.188 | 0.478 | 0.467 | 0.322 | 0.442 | 0.161 | 0.317 |
| 174 | BR-XGB | 0.158 | 0.589 | 0.513 | 0.384 | 0.513 | 0.123 | 0.223 |
| 175 | BR-CART | 0.176 | 0.526 | 0.401 | 0.297 | 0.425 | 0.256 | 0.287 |
| 176 | ECC-RF | 0.185 | 0.532 | 0.514 | 0.357 | 0.488 | 0.167 | 0.257 |
| 177 | ECC-SVM | 0.185 | 0.558 | 0.526 | 0.362 | 0.495 | 0.179 | 0.277 |
| 178 | ECC-SMO | 0.179 | 0.548 | 0.491 | 0.36 | 0.489 | 0.197 | 0.24 |
| 179 | ECC-C5.0 | 0.176 | 0.536 | 0.502 | 0.362 | 0.49 | 0.162 | 0.273 |
| 180 | ECC-NB | 0.18 | 0.563 | 0.48 | 0.352 | 0.48 | 0.146 | 0.24 |
| 181 | ECC-XGB | 0.192 | 0.531 | 0.521 | 0.342 | 0.48 | 0.172 | 0.297 |
| 182 | ECC-CART | 0.195 | 0.514 | 0.46 | 0.328 | 0.462 | 0.204 | 0.277 |
| 183 | LP-RF | 0.168 | 0.527 | 0.484 | 0.375 | 0.482 | 0.268 | 0.46 |
| 184 | LP-SVM | 0.209 | 0.21 | 0.106 | 0.106 | 0.114 | 0.434 | 0.69 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 185 | LP-SMO | 0.205 | 0.379 | 0.367 | 0.271 | 0.352 | 0.329 | 0.53 |
| 186 | LP-C5.0 | 0.171 | 0.445 | 0.413 | 0.339 | 0.417 | 0.294 | 0.497 |
| 187 | LP-NB | 0.151 | 0.577 | 0.549 | 0.439 | 0.542 | 0.216 | 0.427 |
| 188 | LP-XGB | 0.219 | 0.298 | 0.293 | 0.22 | 0.279 | 0.365 | 0.657 |
| 189 | LP-CART | 0.191 | 0.438 | 0.356 | 0.279 | 0.363 | 0.321 | 0.53 |
| 190 | HOMER-RF | 0.252 | 0.324 | 0.214 | 0.149 | 0.217 | 0.569 | 0.623 |
| 191 | HOMER-SVM | 0.272 | 0.244 | 0.145 | 0.091 | 0.153 | 0.595 | 0.723 |
| 192 | HOMER-SMO | 0.257 | 0.263 | 0.164 | 0.098 | 0.162 | 0.426 | 0.813 |
| 193 | HOMER-C5.0 | 0.248 | 0.237 | 0.172 | 0.094 | 0.152 | 0.589 | 0.593 |
| 194 | HOMER-NB | 0.233 | 0.311 | 0.17 | 0.132 | 0.192 | 0.584 | 0.597 |
| 195 | HOMER-XGB | 0.251 | 0.256 | 0.183 | 0.112 | 0.18 | 0.572 | 0.573 |
| 196 | HOMER-CART | 0.26 | 0.226 | 0.143 | 0.079 | 0.135 | 0.566 | 0.76 |
| 197 | RAKEL-RF | 0.144 | 0.629 | 0.472 | 0.384 | 0.508 | 0.201 | 0.267 |
| 198 | RAKEL-SVM | 0.163 | 0.702 | 0.338 | 0.305 | 0.432 | 0.253 | 0.313 |
| 199 | RAKEL-SMO | 0.154 | 0.677 | 0.4 | 0.338 | 0.465 | 0.226 | 0.31 |
| 200 | RAKEL-C5.0 | 0.15 | 0.652 | 0.399 | 0.339 | 0.47 | 0.205 | 0.343 |
| 201 | RAKEL-NB | 0.195 | 0.459 | 0.491 | 0.331 | 0.447 | 0.203 | 0.387 |
| 202 | RAKEL-XGB | 0.151 | 0.622 | 0.487 | 0.385 | 0.509 | 0.167 | 0.287 |
| 203 | RAKEL-CART | 0.163 | 0.562 | 0.393 | 0.307 | 0.432 | 0.21 | 0.293 |
| 204 | BRPLUS-RF | 0.154 | 0.678 | 0.513 | 0.423 | 0.543 | 0.115 | 0.16 |
| 205 | BRPLUS-SVM | 0.168 | 0.572 | 0.31 | 0.265 | 0.371 | 0.178 | 0.28 |
| 206 | BRPLUS-SMO | 0.162 | 0.616 | 0.425 | 0.338 | 0.467 | 0.276 | 0.43 |
| 207 | BRPLUS-C5.0 | 0.158 | 0.63 | 0.468 | 0.382 | 0.504 | 0.169 | 0.22 |
| 208 | BRPLUS-NB | 0.187 | 0.452 | 0.482 | 0.336 | 0.443 | 0.143 | 0.363 |
| 209 | BRPLUS-XGB | 0.162 | 0.602 | 0.504 | 0.387 | 0.513 | 0.145 | 0.187 |
| 210 | BRPLUS-CART | 0.18 | 0.52 | 0.449 | 0.319 | 0.448 | 0.227 | 0.35 |
| 211 | BR-RF | 0.177 | 0.591 | 0.535 | 0.402 | 0.53 | 0.125 | 0.163 |
| 212 | BR-SVM | 0.184 | 0.678 | 0.367 | 0.317 | 0.45 | 0.187 | 0.227 |
| 213 | BR-SMO | 0.174 | 0.652 | 0.425 | 0.352 | 0.481 | 0.26 | 0.333 |
| 214 | BR-C5.0 | 0.172 | 0.617 | 0.395 | 0.331 | 0.461 | 0.2 | 0.243 |
| 215 | BR-NB | 0.206 | 0.482 | 0.47 | 0.33 | 0.449 | 0.169 | 0.333 |
| 216 | BR-XGB | 0.178 | 0.591 | 0.521 | 0.39 | 0.52 | 0.134 | 0.223 |
| 217 | BR-CART | 0.197 | 0.526 | 0.411 | 0.304 | 0.432 | 0.267 | 0.287 |
| 218 | ECC-RF | 0.192 | 0.608 | 0.539 | 0.399 | 0.53 | 0.157 | 0.217 |
| 219 | ECC-SVM | 0.197 | 0.571 | 0.489 | 0.357 | 0.487 | 0.184 | 0.277 |
| 220 | ECC-SMO | 0.198 | 0.56 | 0.498 | 0.367 | 0.494 | 0.197 | 0.24 |
| 221 | ECC-C5.0 | 0.202 | 0.564 | 0.561 | 0.374 | 0.513 | 0.171 | 0.277 |
| 222 | ECC-NB | 0.191 | 0.542 | 0.527 | 0.37 | 0.499 | 0.149 | 0.273 |
| 223 | ECC-XGB | 0.201 | 0.555 | 0.57 | 0.369 | 0.513 | 0.156 | 0.26 |
| 224 | ECC-CART | 0.203 | 0.534 | 0.534 | 0.365 | 0.501 | 0.169 | 0.24 |
| 225 | LP-RF | 0.188 | 0.531 | 0.52 | 0.395 | 0.505 | 0.26 | 0.443 |
| 226 | LP-SVM | 0.237 | 0.21 | 0.106 | 0.106 | 0.115 | 0.432 | 0.69 |
| 227 | LP-SMO | 0.225 | 0.386 | 0.37 | 0.279 | 0.361 | 0.33 | 0.51 |
| 228 | LP-C5.0 | 0.19 | 0.452 | 0.417 | 0.345 | 0.422 | 0.294 | 0.477 |
| 229 | LP-NB | 0.168 | 0.583 | 0.554 | 0.446 | 0.549 | 0.219 | 0.41 |
| 230 | LP-XGB | 0.238 | 0.306 | 0.294 | 0.226 | 0.286 | 0.365 | 0.637 |
| 231 | LP-CART | 0.216 | 0.423 | 0.355 | 0.277 | 0.359 | 0.329 | 0.53 |
| 232 | HOMER-RF | 0.312 | 0.341 | 0.204 | 0.147 | 0.232 | 0.592 | 0.533 |
| 233 | HOMER-SVM | 0.296 | 0.291 | 0.159 | 0.119 | 0.196 | 0.604 | 0.66 |
| 234 | HOMER-SMO | 0.282 | 0.265 | 0.155 | 0.115 | 0.181 | 0.42 | 0.78 |
| 235 | HOMER-C5.0 | 0.269 | 0.34 | 0.167 | 0.13 | 0.208 | 0.615 | 0.59 |
| 236 | HOMER-NB | 0.275 | 0.305 | 0.148 | 0.12 | 0.189 | 0.596 | 0.653 |
| 237 | HOMER-XGB | 0.277 | 0.34 | 0.178 | 0.133 | 0.214 | 0.585 | 0.493 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 238 | HOMER-CART | 0.306 | 0.257 | 0.141 | 0.104 | 0.168 | 0.562 | 0.717 |
| 239 | RAKEL-RF | 0.164 | 0.631 | 0.504 | 0.409 | 0.531 | 0.193 | 0.25 |
| 240 | RAKEL-SVM | 0.183 | 0.679 | 0.356 | 0.315 | 0.447 | 0.249 | 0.273 |
| 241 | RAKEL-SMO | 0.171 | 0.672 | 0.411 | 0.348 | 0.476 | 0.228 | 0.293 |
| 242 | RAKEL-C5.0 | 0.17 | 0.635 | 0.412 | 0.345 | 0.477 | 0.187 | 0.29 |
| 243 | RAKEL-NB | 0.216 | 0.46 | 0.522 | 0.35 | 0.463 | 0.172 | 0.403 |
| 244 | RAKEL-XGB | 0.175 | 0.587 | 0.484 | 0.384 | 0.504 | 0.167 | 0.217 |
| 245 | RAKEL-CART | 0.186 | 0.595 | 0.435 | 0.336 | 0.465 | 0.21 | 0.273 |
| 246 | BRPLUS-RF | 0.176 | 0.662 | 0.514 | 0.412 | 0.537 | 0.121 | 0.18 |
| 247 | BRPLUS-SVM | 0.192 | 0.579 | 0.316 | 0.266 | 0.376 | 0.187 | 0.263 |
| 248 | BRPLUS-SMO | 0.187 | 0.591 | 0.445 | 0.346 | 0.475 | 0.273 | 0.517 |
| 249 | BRPLUS-C5.0 | 0.177 | 0.62 | 0.49 | 0.394 | 0.518 | 0.185 | 0.237 |
| 250 | BRPLUS-NB | 0.195 | 0.466 | 0.501 | 0.353 | 0.462 | 0.146 | 0.4 |
| 251 | BRPLUS-XGB | 0.18 | 0.605 | 0.518 | 0.399 | 0.524 | 0.157 | 0.187 |
| 252 | BRPLUS-CART | 0.202 | 0.52 | 0.461 | 0.327 | 0.457 | 0.23 | 0.35 |
| 253 | BR-RF | 0.214 | 0.614 | 0.594 | 0.456 | 0.578 | 0.134 | 0.163 |
| 254 | BR-SVM | 0.238 | 0.68 | 0.425 | 0.364 | 0.497 | 0.206 | 0.24 |
| 255 | BR-SMO | 0.224 | 0.667 | 0.473 | 0.398 | 0.524 | 0.261 | 0.277 |
| 256 | BR-C5.0 | 0.226 | 0.623 | 0.442 | 0.368 | 0.497 | 0.231 | 0.243 |
| 257 | BR-NB | 0.243 | 0.527 | 0.52 | 0.393 | 0.504 | 0.186 | 0.297 |
| 258 | BR-XGB | 0.236 | 0.595 | 0.561 | 0.42 | 0.549 | 0.149 | 0.223 |
| 259 | BR-CART | 0.26 | 0.54 | 0.461 | 0.345 | 0.473 | 0.273 | 0.287 |
| 260 | ECC-RF | 0.234 | 0.633 | 0.586 | 0.452 | 0.575 | 0.16 | 0.203 |
| 261 | ECC-SVM | 0.247 | 0.611 | 0.529 | 0.409 | 0.537 | 0.194 | 0.24 |
| 262 | ECC-SMO | 0.261 | 0.584 | 0.508 | 0.392 | 0.519 | 0.22 | 0.257 |
| 263 | ECC-C5.0 | 0.244 | 0.601 | 0.552 | 0.418 | 0.55 | 0.177 | 0.277 |
| 264 | ECC-NB | 0.249 | 0.571 | 0.502 | 0.393 | 0.51 | 0.203 | 0.257 |
| 265 | ECC-XGB | 0.244 | 0.594 | 0.539 | 0.41 | 0.537 | 0.194 | 0.277 |
| 266 | ECC-CART | 0.246 | 0.618 | 0.561 | 0.43 | 0.56 | 0.213 | 0.223 |
| 267 | LP-RF | 0.213 | 0.595 | 0.582 | 0.477 | 0.574 | 0.259 | 0.387 |
| 268 | LP-SVM | 0.33 | 0.261 | 0.148 | 0.136 | 0.174 | 0.432 | 0.62 |
| 269 | LP-SMO | 0.291 | 0.42 | 0.38 | 0.303 | 0.381 | 0.351 | 0.473 |
| 270 | LP-C5.0 | 0.256 | 0.455 | 0.419 | 0.349 | 0.425 | 0.325 | 0.46 |
| 271 | LP-NB | 0.219 | 0.606 | 0.584 | 0.479 | 0.579 | 0.238 | 0.41 |
| 272 | LP-XGB | 0.293 | 0.36 | 0.346 | 0.278 | 0.346 | 0.358 | 0.503 |
| 273 | LP-CART | 0.288 | 0.445 | 0.446 | 0.351 | 0.436 | 0.335 | 0.49 |
| 274 | HOMER-RF | 0.348 | 0.481 | 0.435 | 0.307 | 0.421 | 0.423 | 0.36 |
| 275 | HOMER-SVM | 0.397 | 0.381 | 0.361 | 0.233 | 0.35 | 0.49 | 0.473 |
| 276 | HOMER-SMO | 0.363 | 0.424 | 0.357 | 0.245 | 0.36 | 0.385 | 0.573 |
| 277 | HOMER-C5.0 | 0.331 | 0.459 | 0.403 | 0.255 | 0.378 | 0.442 | 0.397 |
| 278 | HOMER-NB | 0.32 | 0.422 | 0.321 | 0.23 | 0.335 | 0.467 | 0.5 |
| 279 | HOMER-XGB | 0.34 | 0.431 | 0.371 | 0.243 | 0.36 | 0.44 | 0.413 |
| 280 | HOMER-CART | 0.357 | 0.418 | 0.364 | 0.222 | 0.336 | 0.442 | 0.443 |
| 281 | RAKEL-RF | 0.202 | 0.676 | 0.579 | 0.483 | 0.601 | 0.188 | 0.307 |
| 282 | RAKEL-SVM | 0.242 | 0.705 | 0.391 | 0.343 | 0.475 | 0.251 | 0.257 |
| 283 | RAKEL-SMO | 0.227 | 0.656 | 0.442 | 0.373 | 0.496 | 0.235 | 0.257 |
| 284 | RAKEL-C5.0 | 0.229 | 0.626 | 0.478 | 0.397 | 0.515 | 0.192 | 0.29 |
| 285 | RAKEL-NB | 0.252 | 0.518 | 0.616 | 0.414 | 0.532 | 0.185 | 0.397 |
| 286 | RAKEL-XGB | 0.215 | 0.603 | 0.538 | 0.424 | 0.543 | 0.17 | 0.32 |
| 287 | RAKEL-CART | 0.222 | 0.604 | 0.487 | 0.401 | 0.519 | 0.216 | 0.297 |
| 288 | BRPLUS-RF | 0.226 | 0.67 | 0.585 | 0.467 | 0.59 | 0.127 | 0.16 |
| 289 | BRPLUS-SVM | 0.247 | 0.58 | 0.376 | 0.312 | 0.424 | 0.209 | 0.297 |
| 290 | BRPLUS-SMO | 0.243 | 0.625 | 0.482 | 0.385 | 0.512 | 0.277 | 0.41 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 291 | BRPLUS-C5.0 | 0.243 | 0.615 | 0.518 | 0.407 | 0.532 | 0.202 | 0.257 |
| 292 | BRPLUS-NB | 0.226 | 0.499 | 0.562 | 0.414 | 0.515 | 0.16 | 0.293 |
| 293 | BRPLUS-XGB | 0.23 | 0.624 | 0.577 | 0.452 | 0.571 | 0.169 | 0.187 |
| 294 | BRPLUS-CART | 0.266 | 0.534 | 0.516 | 0.37 | 0.498 | 0.246 | 0.35 |

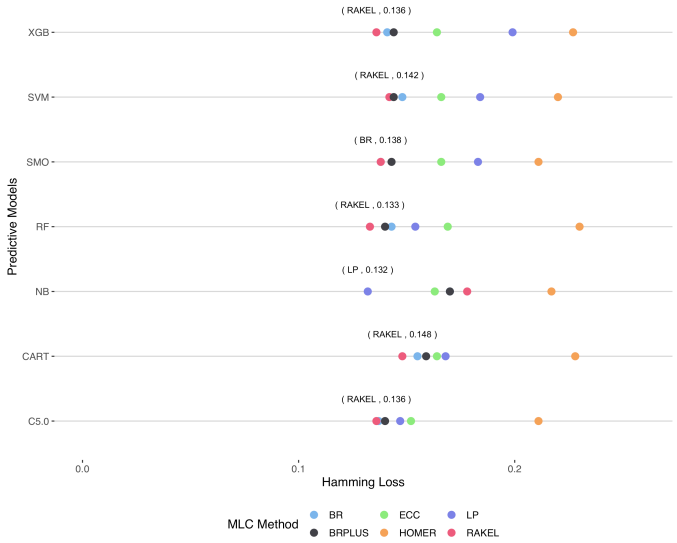



**Figure H. 22.** Hamming-Loss for all MLC strategies applied in CDS4.

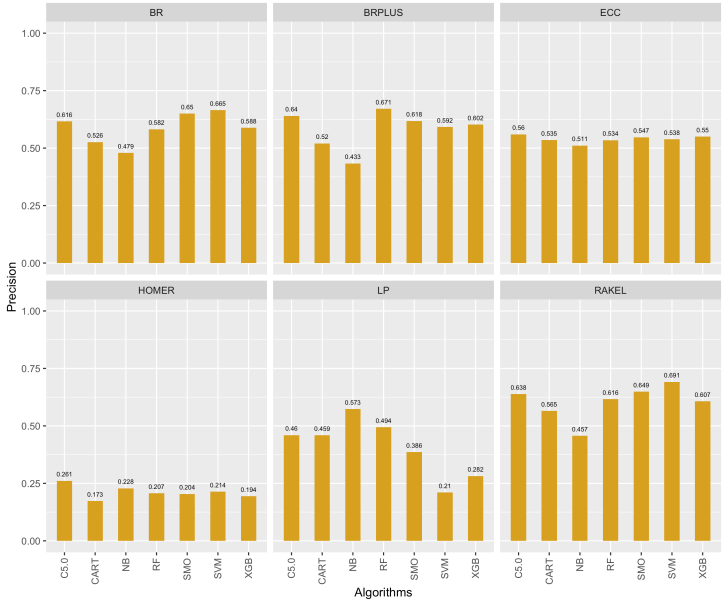



**Figure H. 23.** Precision for all MLC strategies applied in CDS4.

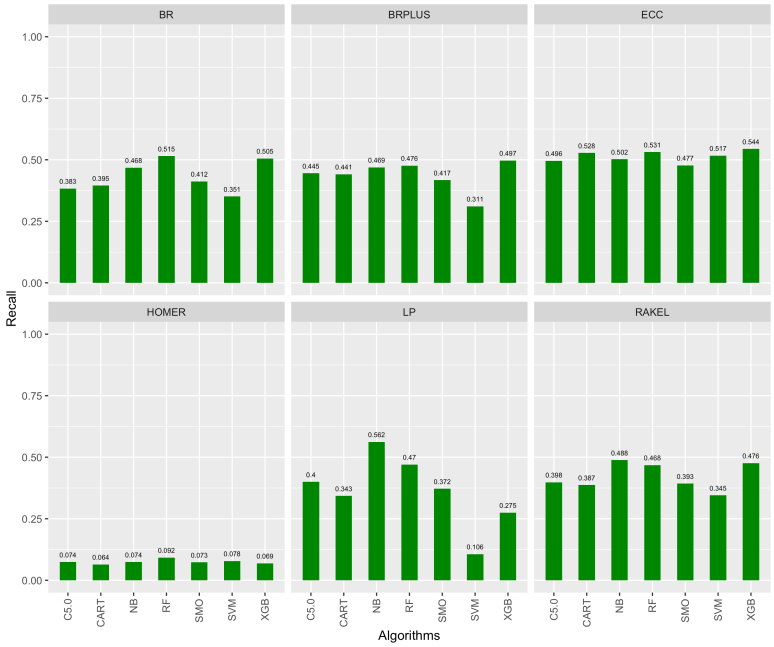



**Figure H. 24.** Recall for all MLC strategies applied in CDS4.

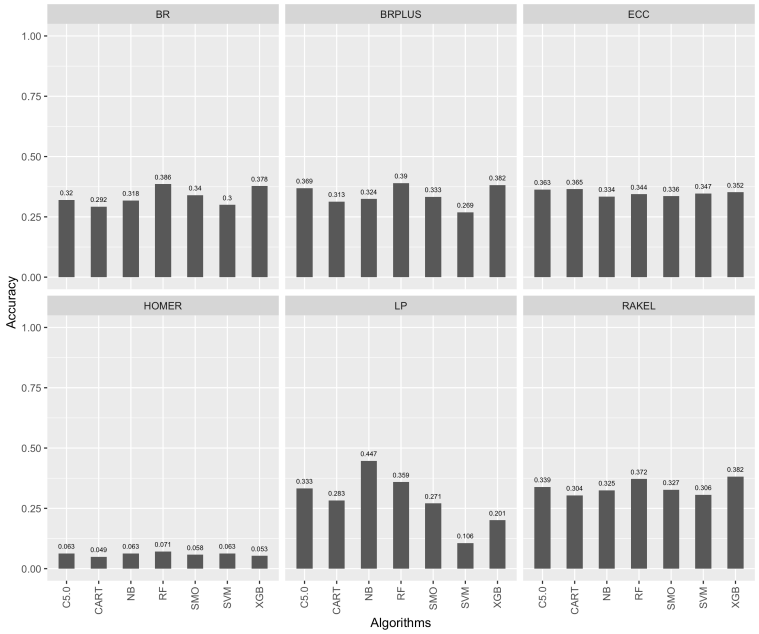



**Figure H. 25.** Accuracy for all MLC strategies applied in CDS4.

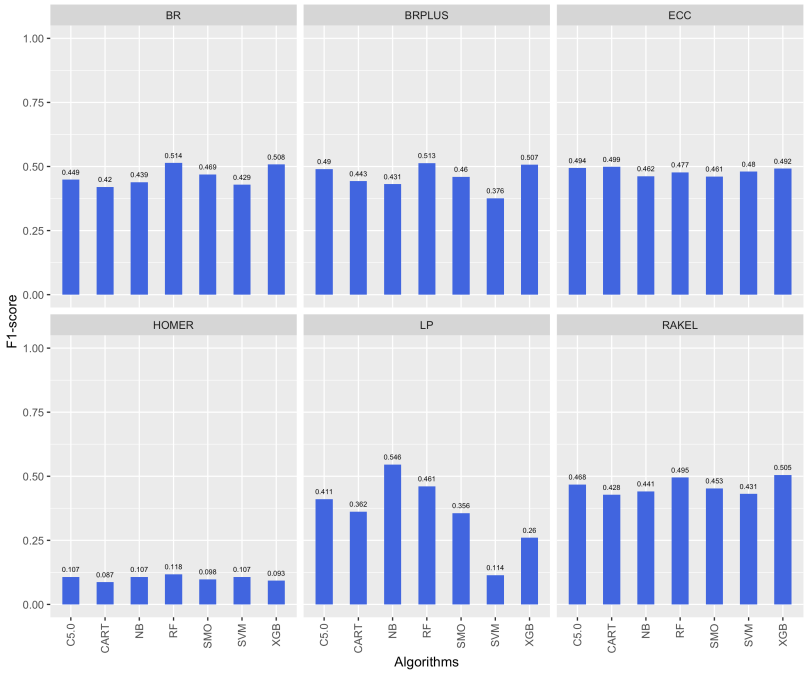



**Figure H. 26.** F1-Score for all MLC strategies applied in CDS4.

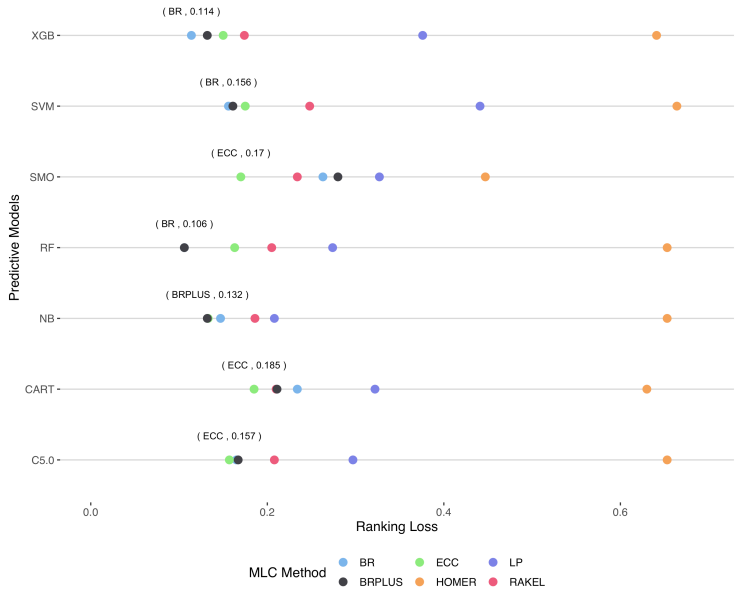



**Figure H. 27.** Ranking-Loss for all MLC strategies applied in CDS4.

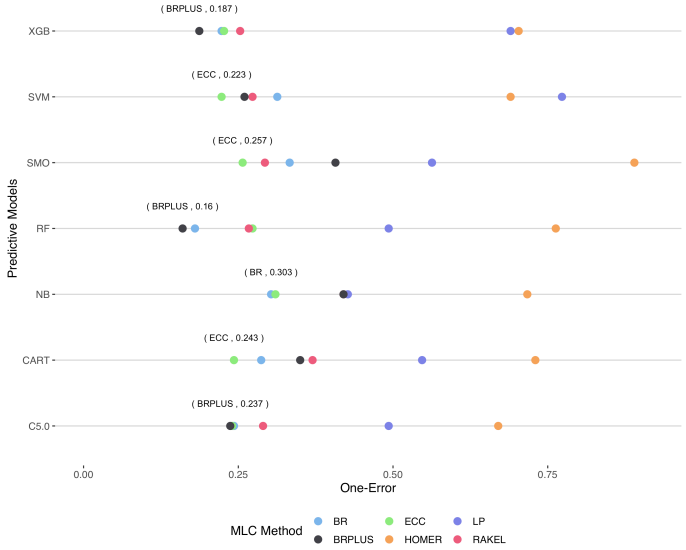



**Figure H. 28.** One-Error for all MLC strategies applied in CDS4.

# Appendix I

# Significance Tests

This appendix presents the statistical significance tests performed to identify differences in the results of different MLC models. Considering Hamming-Loss and Ranking-Loss as two of the most relevant metrics in the MLC approach, these are presented below using ANOVA tests.

## I.1. Normality Test

| MLC Method | Hamming-Loss Significance | | | | | Ranking-Loss Significance | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CDS1 | CDS2 | CDS3 | CDS4 | CGD | CDS1 | CDS2 | CDS3 | CDS4 | CGD |
| BR-RF | 0.48 | 0.48 | 0.71 | 0.66 | 0.52 | 0.90 | 0.86 | 0.58 | 0.12 | 0.54 |
| BR-SVM | 0.85 | 0.84 | 0.50 | 0.47 | 0.19 | 0.01 | 0.85 | 0.98 | 0.55 | 0.19 |
| BR-SMO | 0.33 | 0.88 | 0.97 | 0.85 | 0.41 | 0.69 | 0.00 | 0.46 | 0.00 | 0.41 |
| BR-C5.0 | 0.68 | 0.72 | 0.71 | 0.82 | 0.39 | 0.01 | 0.97 | 0.98 | 0.93 | 0.49 |
| BR-NB | 0.16 | 0.43 | 0.34 | 0.98 | 0.01 | 0.61 | 0.07 | 0.31 | 0.05 | 0.01 |
| BR-XGB | 0.88 | 0.36 | 0.79 | 0.58 | 0.41 | 1.00 | 0.81 | 0.77 | 0.55 | 0.41 |
| BR-CART | 0.36 | 0.75 | 0.81 | 0.77 | 0.85 | 0.01 | 0.31 | 0.20 | 0.10 | 0.85 |
| BR-MAJORITY | 0.91 | 0.95 | | | 0.64 | 0.66 | 0.02 | | | 0.64 |
| BR-RANDOM | 1.00 | 0.80 | | | 0.65 | 0.00 | 0.18 | | | 0.65 |
| ECC-RF | 0.38 | 0.60 | 0.47 | 0.96 | 0.74 | 0.55 | 0.00 | 0.00 | 0.01 | 0.74 |
| ECC-SVM | 0.93 | 0.80 | 0.27 | 0.82 | 0.76 | 0.85 | 0.01 | 0.89 | 0.01 | 0.78 |
| ECC-SMO | 0.39 | 0.92 | 0.71 | 0.79 | 0.42 | 0.10 | 0.17 | 0.24 | 0.39 | 0.42 |
| ECC-C5.0 | 0.78 | 0.75 | 0.91 | 0.94 | 0.97 | 0.40 | 0.35 | 0.39 | 0.55 | 0.97 |
| ECC-NB | 0.13 | 0.36 | 0.57 | 0.52 | 0.24 | 0.56 | 0.44 | 0.03 | 0.97 | 0.26 |
| ECC-XGB | 0.59 | 0.49 | 0.58 | 0.85 | 0.50 | 0.92 | 0.01 | 0.04 | 0.29 | 0.50 |
| ECC-CART | 0.27 | 0.11 | 0.96 | 0.93 | 0.94 | 0.64 | 0.26 | 0.45 | 0.88 | 0.94 |
| ECC-MAJORITY | 0.27 | 0.76 | | | 0.21 | 0.88 | 0.03 | | | 0.21 |
| ECC-RANDOM | 0.10 | 0.06 | | | 0.01 | 0.04 | 0.51 | | | 0.01 |
| LP-RF | 0.65 | 0.57 | 0.90 | 0.54 | 0.22 | 0.00 | 0.31 | 0.00 | 0.00 | 0.22 |
| LP-SVM | 0.65 | 0.24 | 0.05 | 0.91 | 0.55 | 0.45 | 0.23 | 0.01 | 0.00 | 0.55 |
| LP-SMO | 0.16 | 0.74 | 0.51 | 0.93 | 0.28 | 0.73 | 0.00 | 0.00 | 0.00 | 0.30 |
| LP-C5.0 | 0.15 | 0.06 | 0.93 | 0.55 | 0.19 | 0.01 | 0.04 | 0.00 | 0.00 | 0.19 |
| LP-NB | 0.64 | 0.82 | 0.92 | 0.57 | 0.11 | 0.01 | 0.67 | 0.00 | 0.01 | 0.11 |
| LP-XGB | 0.42 | 0.03 | 0.64 | 0.16 | 0.27 | 0.11 | 0.11 | 0.19 | 0.00 | 0.27 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| LP-CART | 0.03 | 0.59 | 0.19 | 0.43 | 0.21 | 0.14 | 0.01 | 0.03 | 0.00 | 0.21 |
| LP-MAJORITY | 0.19 | 0.96 | | | 0.60 | 0.96 | 0.15 | | | 0.58 |
| LP-RANDOM | 0.03 | 0.63 | | | 0.19 | 0.86 | 0.06 | | | 0.19 |
| HOMER-RF | 0.39 | 0.18 | 0.46 | 0.25 | 0.09 | 0.19 | 0.26 | 0.00 | 0.09 | 0.09 |
| HOMER-SVM | 0.44 | 0.18 | 0.66 | 0.47 | 0.08 | 0.48 | 0.33 | 0.00 | 0.02 | 0.08 |
| HOMER-SMO | 0.86 | 0.18 | 0.32 | 0.42 | 0.42 | 0.55 | 0.35 | 0.00 | 0.00 | 0.42 |
| HOMER-C5.0 | 0.63 | 0.10 | 0.09 | 0.24 | 0.10 | 0.13 | 0.09 | 0.00 | 0.08 | 0.10 |
| HOMER-NB | 0.66 | 0.22 | 0.34 | 0.19 | 0.29 | 0.53 | 0.76 | 0.00 | 0.02 | 0.29 |
| HOMER-XGB | 0.72 | 0.19 | 0.23 | 0.35 | 0.17 | 0.40 | 0.34 | 0.00 | 0.04 | 0.17 |
| HOMER-CART | 0.63 | 0.45 | 0.38 | 0.76 | 0.10 | 0.58 | 0.21 | 0.00 | 0.02 | 0.10 |
| HOMER-MAJORITY | 0.16 | 0.63 | | | 0.14 | 0.31 | 0.54 | | | 0.14 |
| HOMER-RANDOM | 0.13 | 0.29 | | | 0.05 | 0.61 | 0.40 | | | 0.05 |
| RAKEL-RF | 0.99 | 0.05 | 0.45 | 0.33 | 0.12 | 0.00 | 0.00 | 0.98 | 0.01 | 0.12 |
| RAKEL-SVM | 0.29 | 0.97 | 0.55 | 0.51 | 0.60 | 0.00 | 0.05 | 0.56 | 0.01 | 0.60 |
| RAKEL-SMO | 0.60 | 0.62 | 0.85 | 0.85 | 0.28 | 0.00 | 0.23 | 0.86 | 0.01 | 0.28 |
| RAKEL-C5.0 | 0.87 | 0.67 | 0.41 | 0.41 | 0.05 | 0.00 | 0.00 | 0.40 | 0.00 | 0.05 |
| RAKEL-NB | 0.37 | 0.83 | 0.86 | 0.95 | 0.65 | 0.73 | 0.22 | 0.65 | 0.00 | 0.65 |
| RAKEL-XGB | 0.41 | 0.47 | 0.49 | 0.95 | 0.22 | 0.00 | 0.40 | 0.91 | 0.01 | 0.22 |
| RAKEL-CART | 0.19 | 0.54 | 0.60 | 0.49 | 0.26 | 0.00 | 0.00 | 0.15 | 0.00 | 0.26 |
| RAKEL-MAJORITY | 0.53 | 0.51 | | | 0.42 | 0.82 | 0.04 | | | 0.42 |
| RAKEL-RANDOM | 0.09 | 0.95 | | | 0.51 | 0.01 | 0.64 | | | 0.51 |
| BRPLUS-RF | 0.58 | 0.32 | 0.90 | 0.39 | 0.78 | 0.06 | 0.18 | 0.59 | 0.86 | 0.78 |
| BRPLUS-SVM | 0.53 | 0.86 | 0.65 | 0.81 | 0.46 | 0.82 | 0.66 | 0.62 | 0.17 | 0.46 |
| BRPLUS-SMO | 0.41 | 0.69 | 0.47 | 0.94 | 0.21 | 0.35 | 0.02 | 0.02 | 0.00 | 0.21 |
| BRPLUS-C5.0 | 0.54 | 0.39 | 0.88 | 0.83 | 0.46 | 0.05 | 0.54 | 0.83 | 0.76 | 0.46 |
| BRPLUS-NB | 0.02 | 0.06 | 0.80 | 0.43 | 0.06 | 0.71 | 0.03 | 0.71 | 0.66 | 0.06 |
| BRPLUS-XGB | 0.47 | 0.55 | 0.88 | 0.54 | 0.33 | 0.55 | 0.20 | 0.45 | 0.53 | 0.33 |
| BRPLUS-CART | 0.68 | 0.25 | 0.87 | 0.81 | 0.67 | 0.00 | 0.28 | 0.16 | 0.09 | 0.67 |
| BRPLUS-MAJORITY | 0.91 | 0.95 | | | 0.64 | 0.66 | 0.02 | | | 0.64 |
| BRPLUS-RANDOM | 0.11 | 0.25 | | | 0.08 | 0.00 | 0.08 | | | 0.09 |

**Table I. 1.** Normality Test for Hamming-Loss and Ranking-Loss measures.

# I.2. Homoscedasticity Test

| | Hamming-Loss Levene's Test | | | | | Ranking-Loss Levene's Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CDS1 | CDS2 | CDS3 | CDS4 | CGD | CDS1 | CDS2 | CDS3 | CDS4 | CGD |
| F-value | 3,6949 | 5,7766 | 1,2273 | 0,7151 | 6,8098 | 4,2235 | 6,4444 | 0,8621 | 1,41 | 6,8098 |
| Pr(>F) | 6,53E-15 | 2,20E-16 | 0,1745 | 0,9018 | 2,20E-16 | 2,20E-16 | 2,20E-16 | 7,10E-01 | 5,98E-02 | 2,20E-16 |

**Table I. 2.** Homoscedasticity Test for Hamming-Loss and Ranking-Loss measures.

# I.3. ANOVA Test

### *Hamming-Loss (CDS1)*

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.238364 -0.033523  0.003727  0.039273  0.210727

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.153455   0.020441   7.507 2.51e-13 ***
datos$mlc_methodsBR-CART        0.052636   0.028907   1.821 0.069181 .
datos$mlc_methodsBR-MAJORITY    0.129364   0.028907   4.475 9.32e-06 ***
datos$mlc_methodsBR-NB          0.051182   0.028907   1.771 0.077200 .
```

```
datos$mlc_methodsBR-RANDOM        0.345818   0.028907  11.963  < 2e-16 ***
datos$mlc_methodsBR-RF           -0.001727   0.028907  -0.060 0.952375
datos$mlc_methodsBR-SMO           0.057182   0.028907   1.978 0.048425 *
datos$mlc_methodsBR-SVM           0.023818   0.028907   0.824 0.410332
datos$mlc_methodsBR-XGB           0.033909   0.028907   1.173 0.241302
datos$mlc_methodsBRPLUS-C5.0      0.024455   0.028907   0.846 0.397948
datos$mlc_methodsBRPLUS-CART      0.064182   0.028907   2.220 0.026816 *
datos$mlc_methodsBRPLUS-MAJORITY  0.129364   0.028907   4.475 9.32e-06 ***
datos$mlc_methodsBRPLUS-NB        0.086455   0.028907   2.991 0.002910 **
datos$mlc_methodsBRPLUS-RANDOM    0.348364   0.028907  12.051  < 2e-16 ***
datos$mlc_methodsBRPLUS-RF        0.007727   0.028907   0.267 0.789331
datos$mlc_methodsBRPLUS-SMO       0.064727   0.028907   2.239 0.025554 *
datos$mlc_methodsBRPLUS-SVM       0.028364   0.028907   0.981 0.326938
datos$mlc_methodsBRPLUS-XGB       0.045364   0.028907   1.569 0.117169
datos$mlc_methodsECC-C5.0         0.036818   0.028907   1.274 0.203331
datos$mlc_methodsECC-CART         0.065909   0.028907   2.280 0.022996 *
datos$mlc_methodsECC-MAJORITY     0.165909   0.028907   5.739 1.59e-08 ***
datos$mlc_methodsECC-NB           0.075818   0.028907   2.623 0.008967 **
datos$mlc_methodsECC-RANDOM       0.260455   0.028907   9.010  < 2e-16 ***
datos$mlc_methodsECC-RF           0.027091   0.028907   0.937 0.349094
datos$mlc_methodsECC-SMO          0.089727   0.028907   3.104 0.002010 **
datos$mlc_methodsECC-SVM          0.064182   0.028907   2.220 0.026816 *
datos$mlc_methodsECC-XGB          0.047000   0.028907   1.626 0.104558
datos$mlc_methodsHOMER-C5.0       0.054818   0.028907   1.896 0.058449 .
datos$mlc_methodsHOMER-CART       0.112091   0.028907   3.878 0.000118 ***
datos$mlc_methodsHOMER-MAJORITY   0.267818   0.028907   9.265  < 2e-16 ***
datos$mlc_methodsHOMER-NB         0.104909   0.028907   3.629 0.000311 ***
datos$mlc_methodsHOMER-RANDOM     0.275909   0.028907   9.545  < 2e-16 ***
datos$mlc_methodsHOMER-RF         0.064727   0.028907   2.239 0.025554 *
datos$mlc_methodsHOMER-SMO        0.140909   0.028907   4.875 1.44e-06 ***
datos$mlc_methodsHOMER-SVM        0.107182   0.028907   3.708 0.000231 ***
datos$mlc_methodsHOMER-XGB        0.091545   0.028907   3.167 0.001628 **
datos$mlc_methodsLP-C5.0         -0.047455   0.028907  -1.642 0.101254
datos$mlc_methodsLP-CART          0.126636   0.028907   4.381 1.42e-05 ***
datos$mlc_methodsLP-MAJORITY      0.194273   0.028907   6.721 4.61e-11 ***
datos$mlc_methodsLP-NB           -0.008727   0.028907  -0.302 0.762841
datos$mlc_methodsLP-RANDOM        0.227545   0.028907   7.872 1.93e-14 ***
datos$mlc_methodsLP-RF           -0.002273   0.028907  -0.079 0.937363
datos$mlc_methodsLP-SMO           0.102364   0.028907   3.541 0.000433 ***
datos$mlc_methodsLP-SVM           0.108909   0.028907   3.768 0.000183 ***
datos$mlc_methodsLP-XGB           0.088727   0.028907   3.069 0.002253 **
datos$mlc_methodsRAKEL-C5.0      -0.042818   0.028907  -1.481 0.139131
datos$mlc_methodsRAKEL-CART       0.047545   0.028907   1.645 0.100603
datos$mlc_methodsRAKEL-MAJORITY   0.138545   0.028907   4.793 2.13e-06 ***
datos$mlc_methodsRAKEL-NB         0.034091   0.028907   1.179 0.238792
datos$mlc_methodsRAKEL-RANDOM     0.320909   0.028907  11.101  < 2e-16 ***
datos$mlc_methodsRAKEL-RF        -0.006000   0.028907  -0.208 0.835651
datos$mlc_methodsRAKEL-SMO        0.060273   0.028907   2.085 0.037536 *
datos$mlc_methodsRAKEL-SVM        0.037545   0.028907   1.299 0.194560
datos$mlc_methodsRAKEL-XGB        0.009364   0.028907   0.324 0.746125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06779 on 540 degrees of freedom
Multiple R-squared:  0.673,   Adjusted R-squared:  0.6409
F-statistic: 20.97 on 53 and 540 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  53 5.1078 0.096373  20.969 < 2.2e-16 ***
Residuals         540 2.4818 0.004596
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

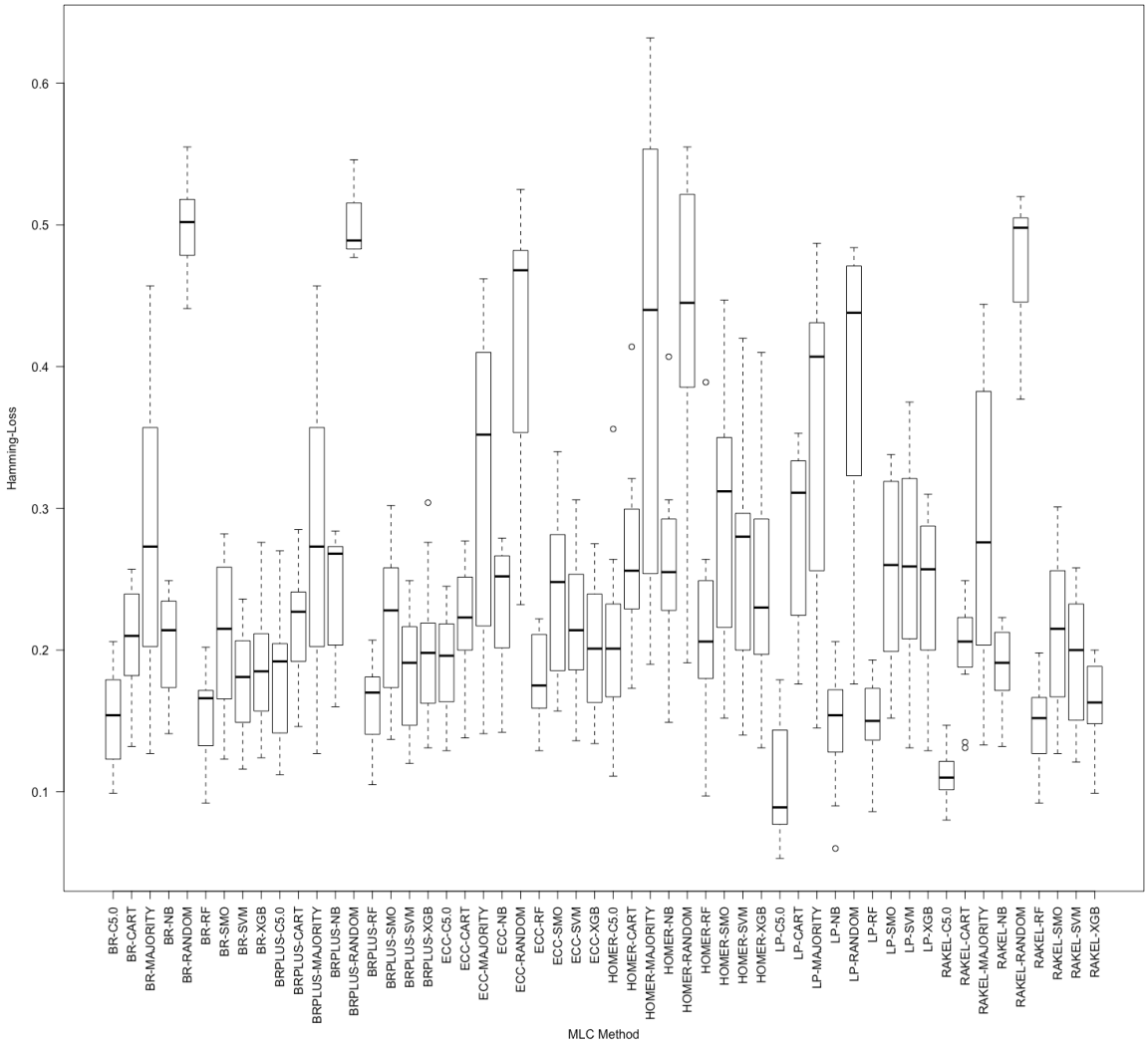



**Figure I. 1.** Significant differences among MLC models applied to CDS1 using ANOVA analysis on Hamming-Loss measure.

## *Hamming-Loss (CDS2)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.22070 -0.02995  0.00080  0.03120  0.27840

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.16070    0.02122   7.573 1.85e-13 ***
datos$mlc_methodsBR-CART         0.02770    0.03001   0.923 0.356417
datos$mlc_methodsBR-MAJORITY     0.14720    0.03001   4.905 1.27e-06 ***
datos$mlc_methodsBR-NB           0.05110    0.03001   1.703 0.089228 .
datos$mlc_methodsBR-RANDOM       0.32520    0.03001  10.837  < 2e-16 ***
datos$mlc_methodsBR-RF          -0.02140    0.03001  -0.713 0.476097
datos$mlc_methodsBR-SMO         -0.01400    0.03001  -0.467 0.641034
datos$mlc_methodsBR-SVM         -0.00180    0.03001  -0.060 0.952193
datos$mlc_methodsBR-XGB          0.00270    0.03001   0.090 0.928343
datos$mlc_methodsBRPLUS-C5.0     0.00300    0.03001   0.100 0.920406
datos$mlc_methodsBRPLUS-CART     0.02080    0.03001   0.693 0.488545
datos$mlc_methodsBRPLUS-MAJORITY 0.14750    0.03001   4.915 1.21e-06 ***
```

```
datos$mlc_methodsBRPLUS-NB         0.04320    0.03001   1.440 0.150617
datos$mlc_methodsBRPLUS-RANDOM     0.31930    0.03001  10.641  < 2e-16 ***
datos$mlc_methodsBRPLUS-RF        -0.02290    0.03001  -0.763 0.445754
datos$mlc_methodsBRPLUS-SMO       -0.02000    0.03001  -0.666 0.505412
datos$mlc_methodsBRPLUS-SVM        0.00380    0.03001   0.127 0.899283
datos$mlc_methodsBRPLUS-XGB       -0.00560    0.03001  -0.187 0.852038
datos$mlc_methodsECC-C5.0          0.00880    0.03001   0.293 0.769451
datos$mlc_methodsECC-CART          0.03330    0.03001   1.110 0.267672
datos$mlc_methodsECC-MAJORITY      0.14330    0.03001   4.775 2.38e-06 ***
datos$mlc_methodsECC-NB            0.03850    0.03001   1.283 0.200104
datos$mlc_methodsECC-RANDOM        0.28360    0.03001   9.451  < 2e-16 ***
datos$mlc_methodsECC-RF            0.00920    0.03001   0.307 0.759289
datos$mlc_methodsECC-SMO           0.00040    0.03001   0.013 0.989370
datos$mlc_methodsECC-SVM           0.01600    0.03001   0.533 0.594143
datos$mlc_methodsECC-XGB           0.02470    0.03001   0.823 0.410844
datos$mlc_methodsHOMER-C5.0        0.10700    0.03001   3.566 0.000399 ***
datos$mlc_methodsHOMER-CART        0.12090    0.03001   4.029 6.50e-05 ***
datos$mlc_methodsHOMER-MAJORITY    0.27600    0.03001   9.198  < 2e-16 ***
datos$mlc_methodsHOMER-NB          0.13550    0.03001   4.515 7.94e-06 ***
datos$mlc_methodsHOMER-RANDOM      0.25780    0.03001   8.591  < 2e-16 ***
datos$mlc_methodsHOMER-RF          0.10150    0.03001   3.382 0.000776 ***
datos$mlc_methodsHOMER-SMO         0.10310    0.03001   3.436 0.000642 ***
datos$mlc_methodsHOMER-SVM         0.12110    0.03001   4.036 6.32e-05 ***
datos$mlc_methodsHOMER-XGB         0.10940    0.03001   3.646 0.000295 ***
datos$mlc_methodsLP-C5.0          -0.02180    0.03001  -0.726 0.467896
datos$mlc_methodsLP-CART           0.04470    0.03001   1.490 0.136975
datos$mlc_methodsLP-MAJORITY       0.20200    0.03001   6.732 4.74e-11 ***
datos$mlc_methodsLP-NB             0.14610    0.03001   4.869 1.52e-06 ***
datos$mlc_methodsLP-RANDOM         0.24360    0.03001   8.118 3.93e-15 ***
datos$mlc_methodsLP-RF            -0.01420    0.03001  -0.473 0.636276
datos$mlc_methodsLP-SMO           -0.02450    0.03001  -0.816 0.414641
datos$mlc_methodsLP-SVM            0.07410    0.03001   2.469 0.013878 *
datos$mlc_methodsLP-XGB            0.02270    0.03001   0.756 0.449734
datos$mlc_methodsRAKEL-C5.0       -0.02200    0.03001  -0.733 0.463825
datos$mlc_methodsRAKEL-CART        0.00350    0.03001   0.117 0.907196
datos$mlc_methodsRAKEL-MAJORITY    0.16490    0.03001   5.495 6.31e-08 ***
datos$mlc_methodsRAKEL-NB          0.13350    0.03001   4.449 1.07e-05 ***
datos$mlc_methodsRAKEL-RANDOM      0.30340    0.03001  10.111  < 2e-16 ***
datos$mlc_methodsRAKEL-RF         -0.02190    0.03001  -0.730 0.465858
datos$mlc_methodsRAKEL-SMO        -0.02390    0.03001  -0.796 0.426154
datos$mlc_methodsRAKEL-SVM         0.00430    0.03001   0.143 0.886116
datos$mlc_methodsRAKEL-XGB        -0.02380    0.03001  -0.793 0.428091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0671 on 486 degrees of freedom
Multiple R-squared:  0.7185,  Adjusted R-squared:  0.6877
F-statistic:  23.4 on 53 and 486 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  53 5.5837 0.105353  23.399 < 2.2e-16 ***
Residuals         486 2.1881 0.004502
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

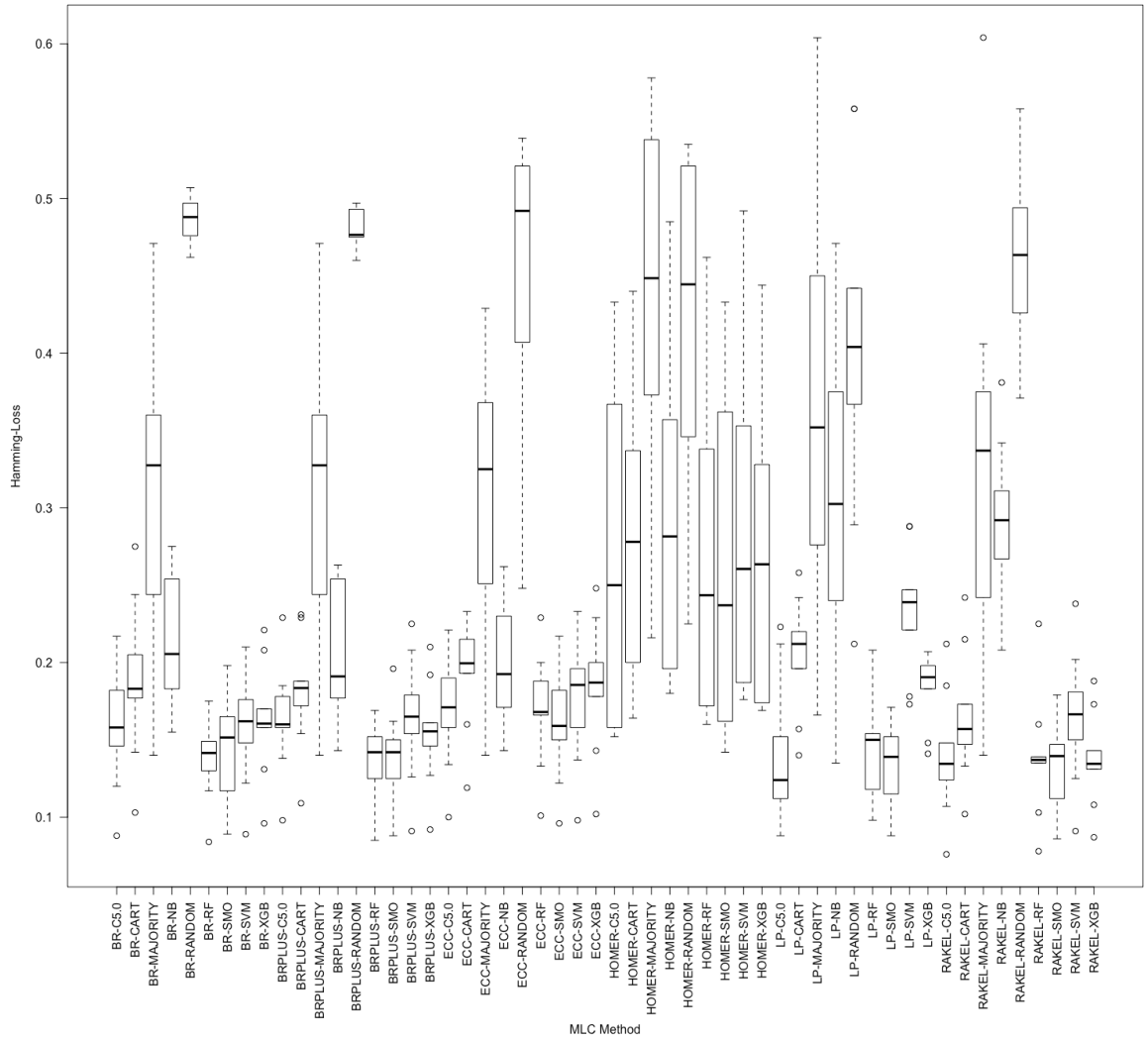



**Figure I. 2.** Significant differences among MLC models applied to CDS2 using ANOVA analysis on Hamming-Loss measure.

## *Hamming-Loss (CDS3)*

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.151000 -0.040929  0.002929  0.022357  0.226429

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.247714   0.024649  10.050  < 2e-16 ***
datos$mlc_methodsBR-CART     -0.025857   0.034859  -0.742 0.458926
datos$mlc_methodsBR-NB        0.115714   0.034859   3.319 0.001035 **
datos$mlc_methodsBR-RF       -0.043714   0.034859  -1.254 0.210997
datos$mlc_methodsBR-SMO      -0.033714   0.034859  -0.967 0.334395
datos$mlc_methodsBR-SVM      -0.044000   0.034859  -1.262 0.208038
datos$mlc_methodsBR-XGB      -0.007000   0.034859  -0.201 0.841011
datos$mlc_methodsBRPLUS-C5.0 -0.002286   0.034859  -0.066 0.947772
datos$mlc_methodsBRPLUS-CART -0.014286   0.034859  -0.410 0.682293
datos$mlc_methodsBRPLUS-NB    0.065286   0.034859   1.873 0.062249 .
datos$mlc_methodsBRPLUS-RF   -0.048857   0.034859  -1.402 0.162280
datos$mlc_methodsBRPLUS-SMO  -0.030571   0.034859  -0.877 0.381325
```

```
datos$mlc_methodsBRPLUS-SVM  -0.041000   0.034859  -1.176 0.240643
datos$mlc_methodsBRPLUS-XGB  -0.004429   0.034859  -0.127 0.899009
datos$mlc_methodsECC-C5.0    -0.009286   0.034859  -0.266 0.790167
datos$mlc_methodsECC-CART    -0.004714   0.034859  -0.135 0.892532
datos$mlc_methodsECC-NB       0.014000   0.034859   0.402 0.688309
datos$mlc_methodsECC-RF      -0.010429   0.034859  -0.299 0.765063
datos$mlc_methodsECC-SMO     -0.006286   0.034859  -0.180 0.857049
datos$mlc_methodsECC-SVM     -0.003857   0.034859  -0.111 0.911983
datos$mlc_methodsECC-XGB     -0.006714   0.034859  -0.193 0.847419
datos$mlc_methodsHOMER-C5.0   0.086143   0.034859   2.471 0.014130 *
datos$mlc_methodsHOMER-CART   0.085857   0.034859   2.463 0.014449 *
datos$mlc_methodsHOMER-NB     0.116571   0.034859   3.344 0.000951 ***
datos$mlc_methodsHOMER-RF     0.076143   0.034859   2.184 0.029863 *
datos$mlc_methodsHOMER-SMO    0.086000   0.034859   2.467 0.014289 *
datos$mlc_methodsHOMER-SVM    0.103286   0.034859   2.963 0.003339 **
datos$mlc_methodsHOMER-XGB    0.088143   0.034859   2.529 0.012066 *
datos$mlc_methodsLP-C5.0     -0.004143   0.034859  -0.119 0.905493
datos$mlc_methodsLP-CART     -0.001000   0.034859  -0.029 0.977137
datos$mlc_methodsLP-NB        0.022571   0.034859   0.647 0.517898
datos$mlc_methodsLP-RF        0.007000   0.034859   0.201 0.841011
datos$mlc_methodsLP-SMO       0.017714   0.034859   0.508 0.611782
datos$mlc_methodsLP-SVM       0.008857   0.034859   0.254 0.799639
datos$mlc_methodsLP-XGB      -0.003286   0.034859  -0.094 0.924980
datos$mlc_methodsRAKEL-C5.0  -0.026571   0.034859  -0.762 0.446626
datos$mlc_methodsRAKEL-CART  -0.033000   0.034859  -0.947 0.344718
datos$mlc_methodsRAKEL-NB     0.041000   0.034859   1.176 0.240643
datos$mlc_methodsRAKEL-RF    -0.045714   0.034859  -1.311 0.190919
datos$mlc_methodsRAKEL-SMO   -0.037571   0.034859  -1.078 0.282153
datos$mlc_methodsRAKEL-SVM   -0.047857   0.034859  -1.373 0.171015
datos$mlc_methodsRAKEL-XGB   -0.022714   0.034859  -0.652 0.515254
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06522 on 252 degrees of freedom
Multiple R-squared:  0.3866,  Adjusted R-squared:  0.2868
F-statistic: 3.874 on 41 and 252 DF,  p-value: 1.595e-11

Analysis of Variance Table

                   Df  Sum Sq   Mean Sq F value    Pr(>F)
datos$mlc_methods  41 0.67562 0.0164785  3.8744 1.595e-11 ***
Residuals         252 1.07178 0.0042531
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

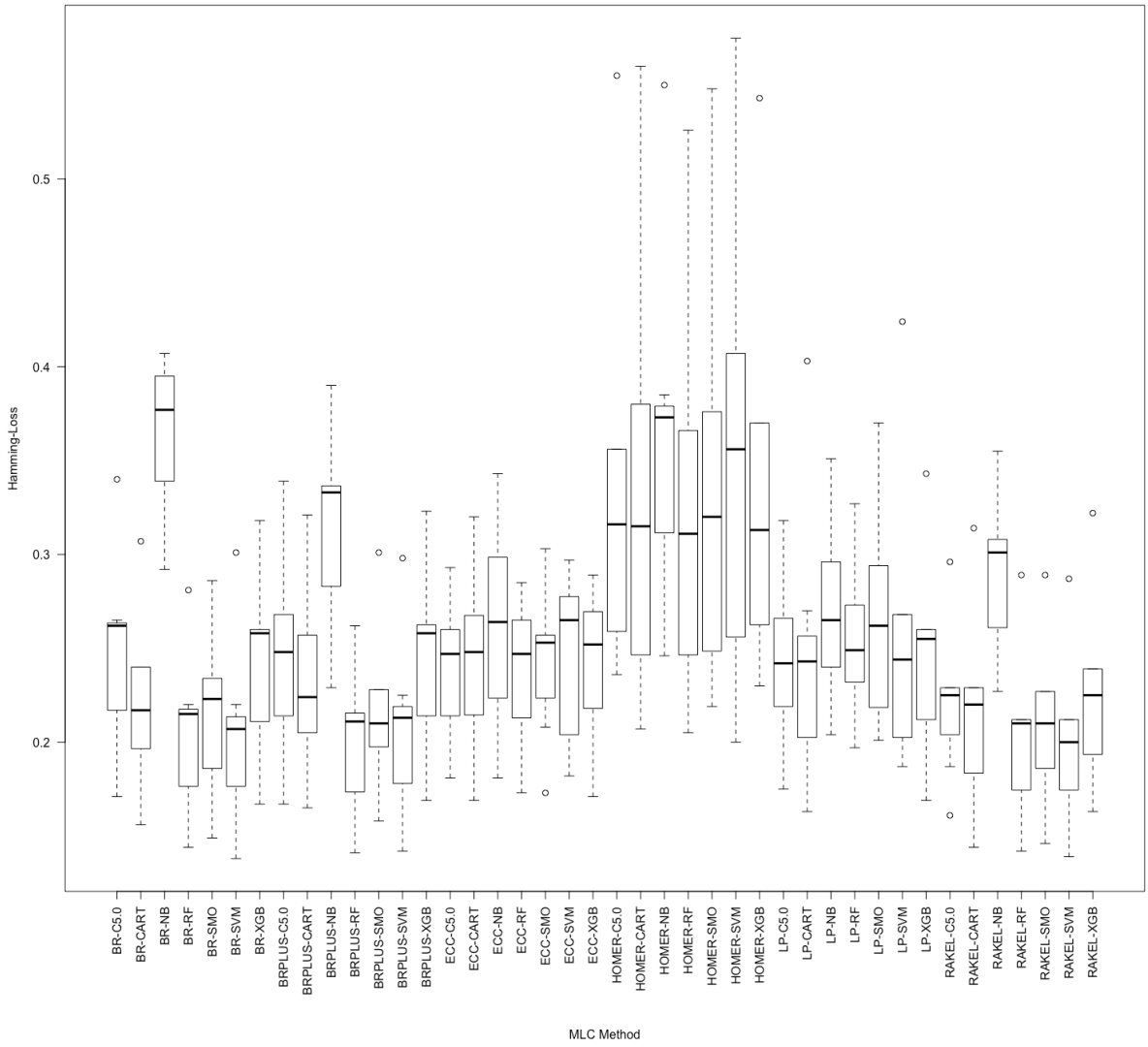



**Figure I. 3.** Significant differences among MLC models applied to CDS3 using ANOVA analysis on Hamming-Loss measure.

## *Hamming-Loss (CDS4)*

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.115143 -0.039679  0.000429  0.032393  0.182857

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.019e-01  2.158e-02   9.355  < 2e-16 ***
datos$mlc_methodsBR-CART    3.457e-02  3.051e-02   1.133 0.258297
datos$mlc_methodsBR-NB      2.271e-02  3.051e-02   0.744 0.457329
datos$mlc_methodsBR-RF     -2.071e-02  3.051e-02  -0.679 0.497853
datos$mlc_methodsBR-SMO    -3.714e-03  3.051e-02  -0.122 0.903213
datos$mlc_methodsBR-SVM     1.871e-02  3.051e-02   0.613 0.540224
datos$mlc_methodsBR-XGB     5.286e-03  3.051e-02   0.173 0.862614
datos$mlc_methodsBRPLUS-C5.0 8.143e-03  3.051e-02   0.267 0.789795
datos$mlc_methodsBRPLUS-CART 3.657e-02  3.051e-02   1.199 0.231837
datos$mlc_methodsBRPLUS-NB  4.714e-03  3.051e-02   0.154 0.877341
datos$mlc_methodsBRPLUS-RF -2.229e-02  3.051e-02  -0.730 0.465852
```

```
datos$mlc_methodsBRPLUS-SMO   7.143e-04  3.051e-02   0.023 0.981343
datos$mlc_methodsBRPLUS-SVM   2.314e-02  3.051e-02   0.758 0.448894
datos$mlc_methodsBRPLUS-XGB   6.143e-03  3.051e-02   0.201 0.840614
datos$mlc_methodsECC-C5.0     1.586e-02  3.051e-02   0.520 0.603746
datos$mlc_methodsECC-CART     3.286e-02  3.051e-02   1.077 0.282598
datos$mlc_methodsECC-NB       2.371e-02  3.051e-02   0.777 0.437786
datos$mlc_methodsECC-RF       7.143e-03  3.051e-02   0.234 0.815107
datos$mlc_methodsECC-SMO      2.686e-02  3.051e-02   0.880 0.379606
datos$mlc_methodsECC-SVM      2.143e-02  3.051e-02   0.702 0.483163
datos$mlc_methodsECC-XGB      2.786e-02  3.051e-02   0.913 0.362145
datos$mlc_methodsHOMER-C5.0   1.043e-01  3.051e-02   3.418 0.000736 ***
datos$mlc_methodsHOMER-CART   1.157e-01  3.051e-02   3.792 0.000187 ***
datos$mlc_methodsHOMER-NB     1.020e-01  3.051e-02   3.343 0.000955 ***
datos$mlc_methodsHOMER-RF     1.164e-01  3.051e-02   3.816 0.000171 ***
datos$mlc_methodsHOMER-SMO    1.099e-01  3.051e-02   3.600 0.000383 ***
datos$mlc_methodsHOMER-SVM    1.333e-01  3.051e-02   4.368 1.83e-05 ***
datos$mlc_methodsHOMER-XGB    1.083e-01  3.051e-02   3.549 0.000462 ***
datos$mlc_methodsLP-C5.0      2.443e-02  3.051e-02   0.801 0.424129
datos$mlc_methodsLP-CART      4.529e-02  3.051e-02   1.484 0.139028
datos$mlc_methodsLP-NB       -7.714e-03  3.051e-02  -0.253 0.800617
datos$mlc_methodsLP-RF       -1.043e-02  3.051e-02  -0.342 0.732810
datos$mlc_methodsLP-SMO       4.743e-02  3.051e-02   1.554 0.121358
datos$mlc_methodsLP-SVM       7.729e-02  3.051e-02   2.533 0.011922 *
datos$mlc_methodsLP-XGB       5.829e-02  3.051e-02   1.910 0.057248 .
datos$mlc_methodsRAKEL-C5.0  -4.143e-03  3.051e-02  -0.136 0.892111
datos$mlc_methodsRAKEL-CART   1.643e-02  3.051e-02   0.538 0.590775
datos$mlc_methodsRAKEL-NB     2.757e-02  3.051e-02   0.904 0.367082
datos$mlc_methodsRAKEL-RF    -3.229e-02  3.051e-02  -1.058 0.291034
datos$mlc_methodsRAKEL-SMO   -1.310e-16  3.051e-02   0.000 1.000000
datos$mlc_methodsRAKEL-SVM    1.829e-02  3.051e-02   0.599 0.549535
datos$mlc_methodsRAKEL-XGB   -1.100e-02  3.051e-02  -0.360 0.718779
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05709 on 252 degrees of freedom
Multiple R-squared:  0.3897,  Adjusted R-squared:  0.2904
F-statistic: 3.924 on 41 and 252 DF,  p-value: 9.933e-12

Analysis of Variance Table

                   Df  Sum Sq   Mean Sq F value    Pr(>F)
datos$mlc_methods  41 0.52428 0.0127874   3.924 9.933e-12 ***
Residuals         252 0.82121 0.0032588
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

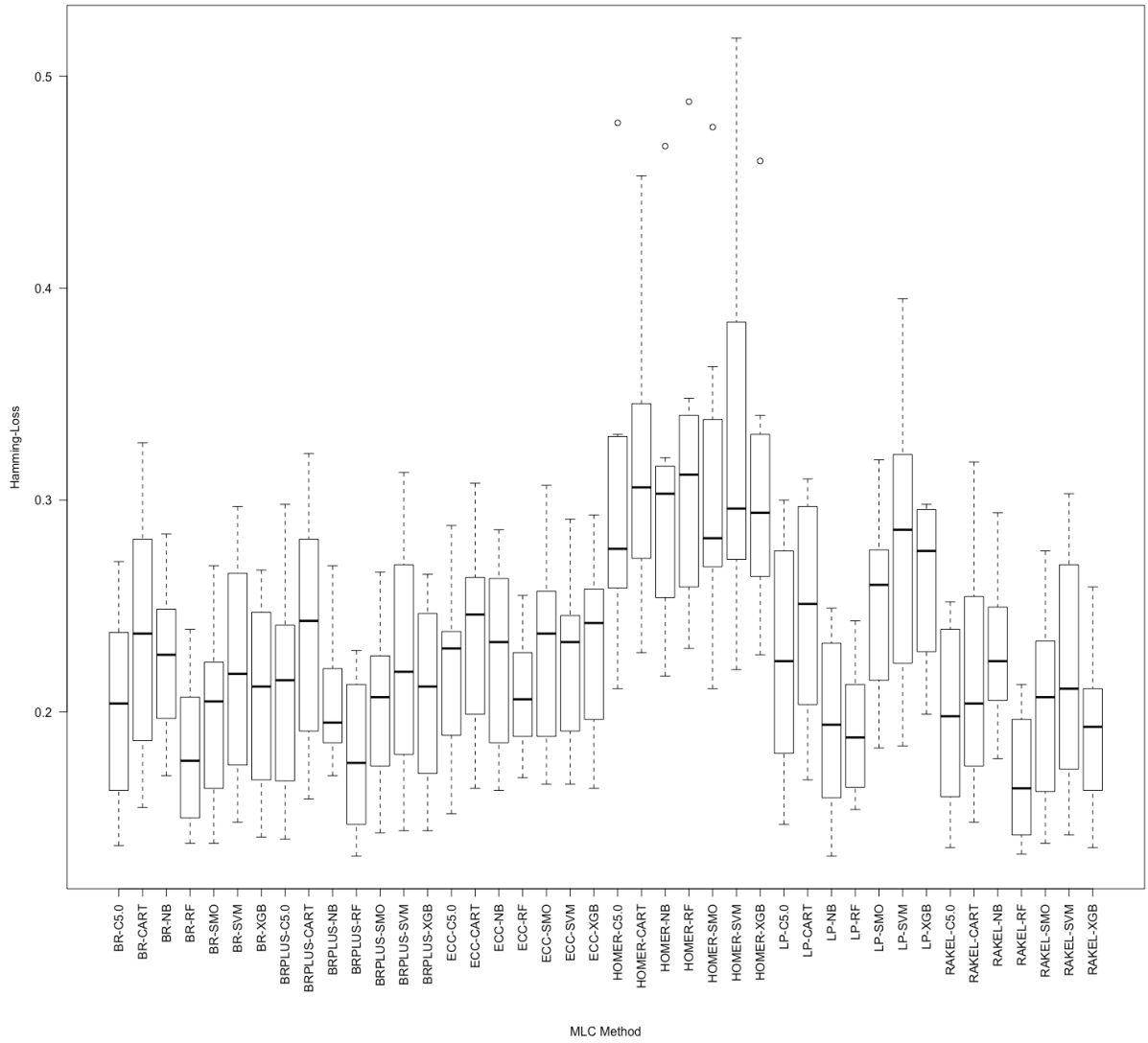



**Figure I. 4.** Significant differences among MLC models applied to CDS4 using ANOVA analysis on Hamming-Loss measure.

## *Hamming-Loss (CGD)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.23862 -0.04425 -0.00640  0.04121  0.29600

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.184057   0.011957  15.394  < 2e-16 ***
datos$mlc_methodsBR-CART        0.026200   0.016909   1.549 0.121461
datos$mlc_methodsBR-MAJORITY    0.110705   0.019525   5.670 1.68e-08 ***
datos$mlc_methodsBR-NB          0.058371   0.016909   3.452 0.000570 ***
datos$mlc_methodsBR-RANDOM      0.308848   0.019525  15.818  < 2e-16 ***
datos$mlc_methodsBR-RF         -0.019543   0.016909  -1.156 0.247945
datos$mlc_methodsBR-SMO         0.006486   0.016909   0.384 0.701351
datos$mlc_methodsBR-SVM         0.001914   0.016909   0.113 0.909877
datos$mlc_methodsBR-XGB         0.011086   0.016909   0.656 0.512168
datos$mlc_methodsBRPLUS-C5.0    0.009714   0.016909   0.575 0.565707
```

```
datos$mlc_methodsBRPLUS-CART      0.030571   0.016909   1.808 0.070789 .
datos$mlc_methodsBRPLUS-MAJORITY  0.110848   0.019525   5.677 1.61e-08 ***
datos$mlc_methodsBRPLUS-NB        0.053514   0.016909   3.165 0.001580 **
datos$mlc_methodsBRPLUS-RANDOM    0.307371   0.019525  15.742  < 2e-16 ***
datos$mlc_methodsBRPLUS-RF       -0.018343   0.016909  -1.085 0.278171
datos$mlc_methodsBRPLUS-SMO       0.008657   0.016909   0.512 0.608732
datos$mlc_methodsBRPLUS-SVM       0.006429   0.016909   0.380 0.703857
datos$mlc_methodsBRPLUS-XGB       0.013000   0.016909   0.769 0.442111
datos$mlc_methodsECC-C5.0         0.015400   0.016909   0.911 0.362558
datos$mlc_methodsECC-CART         0.035857   0.016909   2.121 0.034104 *
datos$mlc_methodsECC-MAJORITY     0.127990   0.019525   6.555 7.39e-11 ***
datos$mlc_methodsECC-NB           0.042371   0.016909   2.506 0.012311 *
datos$mlc_methodsECC-RANDOM       0.244324   0.019525  12.513  < 2e-16 ***
datos$mlc_methodsECC-RF           0.010486   0.016909   0.620 0.535262
datos$mlc_methodsECC-SMO          0.032429   0.016909   1.918 0.055305 .
datos$mlc_methodsECC-SVM          0.028257   0.016909   1.671 0.094886 .
datos$mlc_methodsECC-XGB          0.026057   0.016909   1.541 0.123503
datos$mlc_methodsHOMER-C5.0       0.085886   0.016909   5.079 4.21e-07 ***
datos$mlc_methodsHOMER-CART       0.110086   0.016909   6.510 9.89e-11 ***
datos$mlc_methodsHOMER-MAJORITY   0.244562   0.019525  12.526  < 2e-16 ***
datos$mlc_methodsHOMER-NB         0.115400   0.016909   6.825 1.23e-11 ***
datos$mlc_methodsHOMER-RANDOM     0.240133   0.019525  12.299  < 2e-16 ***
datos$mlc_methodsHOMER-RF         0.087857   0.016909   5.196 2.29e-07 ***
datos$mlc_methodsHOMER-SMO        0.112914   0.016909   6.678 3.30e-11 ***
datos$mlc_methodsHOMER-SVM        0.115600   0.016909   6.837 1.13e-11 ***
datos$mlc_methodsHOMER-XGB        0.099314   0.016909   5.873 5.14e-09 ***
datos$mlc_methodsLP-C5.0         -0.017086   0.016909  -1.010 0.312429
datos$mlc_methodsLP-CART          0.061429   0.016909   3.633 0.000289 ***
datos$mlc_methodsLP-MAJORITY      0.170800   0.019525   8.748  < 2e-16 ***
datos$mlc_methodsLP-NB            0.041971   0.016909   2.482 0.013156 *
datos$mlc_methodsLP-RANDOM        0.208038   0.019525  10.655  < 2e-16 ***
datos$mlc_methodsLP-RF           -0.005457   0.016909  -0.323 0.746937
datos$mlc_methodsLP-SMO           0.038200   0.016909   2.259 0.024003 *
datos$mlc_methodsLP-SVM           0.072629   0.016909   4.295 1.85e-05 ***
datos$mlc_methodsLP-XGB           0.045371   0.016909   2.683 0.007363 **
datos$mlc_methodsRAKEL-C5.0      -0.025886   0.016909  -1.531 0.125990
datos$mlc_methodsRAKEL-CART       0.012629   0.016909   0.747 0.455259
datos$mlc_methodsRAKEL-MAJORITY   0.123943   0.019525   6.348 2.81e-10 ***
datos$mlc_methodsRAKEL-NB         0.062571   0.016909   3.700 0.000222 ***
datos$mlc_methodsRAKEL-RANDOM     0.285419   0.019525  14.618  < 2e-16 ***
datos$mlc_methodsRAKEL-RF        -0.023743   0.016909  -1.404 0.160462
datos$mlc_methodsRAKEL-SMO        0.004600   0.016909   0.272 0.785623
datos$mlc_methodsRAKEL-SVM        0.007114   0.016909   0.421 0.674001
datos$mlc_methodsRAKEL-XGB       -0.010600   0.016909  -0.627 0.530823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07074 on 1668 degrees of freedom
Multiple R-squared:  0.5504,  Adjusted R-squared:  0.5361
F-statistic: 38.53 on 53 and 1668 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                    Df  Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods   53 10.2169 0.192772  38.527 < 2.2e-16 ***
Residuals         1668  8.3459 0.005004
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

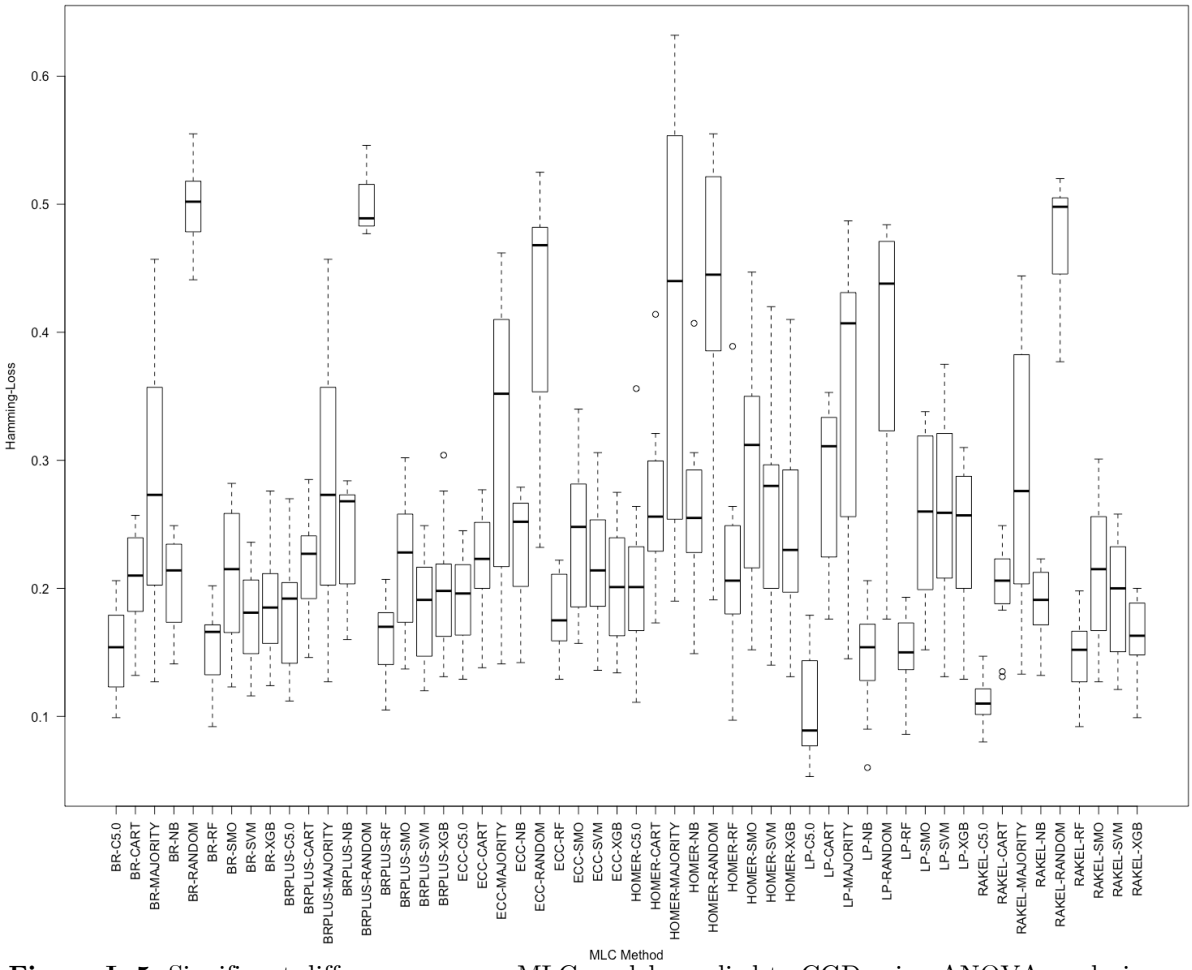



**Figure I. 5.** Significant differences among MLC models applied to CGD using ANOVA analysis on Hamming-Loss measure.

## *Ranking-Loss (CDS1)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.31591 -0.03909 -0.00750  0.03605  0.40982

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.1111818  0.0313528   3.546 0.000425 ***
datos$mlc_methodsBR-CART       0.0324545  0.0443396   0.732 0.464514
datos$mlc_methodsBR-MAJORITY   0.2277273  0.0443396   5.136 3.93e-07 ***
datos$mlc_methodsBR-NB         0.0079091  0.0443396   0.178 0.858495
datos$mlc_methodsBR-RANDOM     0.3423636  0.0443396   7.721 5.62e-14 ***
datos$mlc_methodsBR-RF        -0.0375455  0.0443396  -0.847 0.397498
datos$mlc_methodsBR-SMO        0.1410000  0.0443396   3.180 0.001557 **
datos$mlc_methodsBR-SVM        0.0176364  0.0443396   0.398 0.690967
datos$mlc_methodsBR-XGB       -0.0002727  0.0443396  -0.006 0.995095
datos$mlc_methodsBRPLUS-C5.0   0.0290909  0.0443396   0.656 0.512043
datos$mlc_methodsBRPLUS-CART   0.0547273  0.0443396   1.234 0.217637
datos$mlc_methodsBRPLUS-MAJORITY  0.2277273  0.0443396   5.136 3.93e-07 ***
datos$mlc_methodsBRPLUS-NB     0.0284545  0.0443396   0.642 0.521313
datos$mlc_methodsBRPLUS-RANDOM 0.3364545  0.0443396   7.588 1.43e-13 ***
datos$mlc_methodsBRPLUS-RF    -0.0200909  0.0443396  -0.453 0.650648
datos$mlc_methodsBRPLUS-SMO    0.1434545  0.0443396   3.235 0.001289 **
```

```
datos$mlc_methodsBRPLUS-SVM       0.0133636  0.0443396   0.301 0.763231
datos$mlc_methodsBRPLUS-XGB       0.0185455  0.0443396   0.418 0.675923
datos$mlc_methodsECC-C5.0         0.0110000  0.0443396   0.248 0.804163
datos$mlc_methodsECC-CART         0.0385455  0.0443396   0.869 0.385056
datos$mlc_methodsECC-MAJORITY     0.2251818  0.0443396   5.079 5.24e-07 ***
datos$mlc_methodsECC-NB           0.0480909  0.0443396   1.085 0.278581
datos$mlc_methodsECC-RANDOM       0.3370000  0.0443396   7.600 1.31e-13 ***
datos$mlc_methodsECC-RF           0.0075455  0.0443396   0.170 0.864937
datos$mlc_methodsECC-SMO          0.0935455  0.0443396   2.110 0.035339 *
datos$mlc_methodsECC-SVM          0.0541818  0.0443396   1.222 0.222250
datos$mlc_methodsECC-XGB          0.0174545  0.0443396   0.394 0.693990
datos$mlc_methodsHOMER-C5.0       0.2054545  0.0443396   4.634 4.51e-06 ***
datos$mlc_methodsHOMER-CART       0.2273636  0.0443396   5.128 4.09e-07 ***
datos$mlc_methodsHOMER-MAJORITY   0.3641818  0.0443396   8.213 1.60e-15 ***
datos$mlc_methodsHOMER-NB         0.2314545  0.0443396   5.220 2.56e-07 ***
datos$mlc_methodsHOMER-RANDOM     0.4087273  0.0443396   9.218  < 2e-16 ***
datos$mlc_methodsHOMER-RF         0.1840000  0.0443396   4.150 3.87e-05 ***
datos$mlc_methodsHOMER-SMO        0.2411818  0.0443396   5.439 8.11e-08 ***
datos$mlc_methodsHOMER-SVM        0.2253636  0.0443396   5.083 5.14e-07 ***
datos$mlc_methodsHOMER-XGB        0.2142727  0.0443396   4.833 1.76e-06 ***
datos$mlc_methodsLP-C5.0          0.0190000  0.0443396   0.429 0.668450
datos$mlc_methodsLP-CART          0.2089091  0.0443396   4.712 3.13e-06 ***
datos$mlc_methodsLP-MAJORITY      0.2910909  0.0443396   6.565 1.22e-10 ***
datos$mlc_methodsLP-NB            0.0449091  0.0443396   1.013 0.311588
datos$mlc_methodsLP-RANDOM        0.2855455  0.0443396   6.440 2.64e-10 ***
datos$mlc_methodsLP-RF            0.0578182  0.0443396   1.304 0.192794
datos$mlc_methodsLP-SMO           0.1937273  0.0443396   4.369 1.50e-05 ***
datos$mlc_methodsLP-SVM           0.1985455  0.0443396   4.478 9.21e-06 ***
datos$mlc_methodsLP-XGB           0.1631818  0.0443396   3.680 0.000256 ***
datos$mlc_methodsRAKEL-C5.0      -0.0134545  0.0443396  -0.303 0.761669
datos$mlc_methodsRAKEL-CART       0.0669091  0.0443396   1.509 0.131880
datos$mlc_methodsRAKEL-MAJORITY   0.2296364  0.0443396   5.179 3.15e-07 ***
datos$mlc_methodsRAKEL-NB         0.0148182  0.0443396   0.334 0.738360
datos$mlc_methodsRAKEL-RANDOM     0.3211818  0.0443396   7.244 1.51e-12 ***
datos$mlc_methodsRAKEL-RF         0.0304545  0.0443396   0.687 0.492473
datos$mlc_methodsRAKEL-SMO        0.0971818  0.0443396   2.192 0.028823 *
datos$mlc_methodsRAKEL-SVM        0.0711818  0.0443396   1.605 0.108995
datos$mlc_methodsRAKEL-XGB        0.0301818  0.0443396   0.681 0.496355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.104 on 540 degrees of freedom
Multiple R-squared:  0.5969,  Adjusted R-squared:  0.5573
F-statistic: 15.08 on 53 and 540 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  53 8.6448 0.163110  15.085 < 2.2e-16 ***
Residuals         540 5.8390 0.010813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

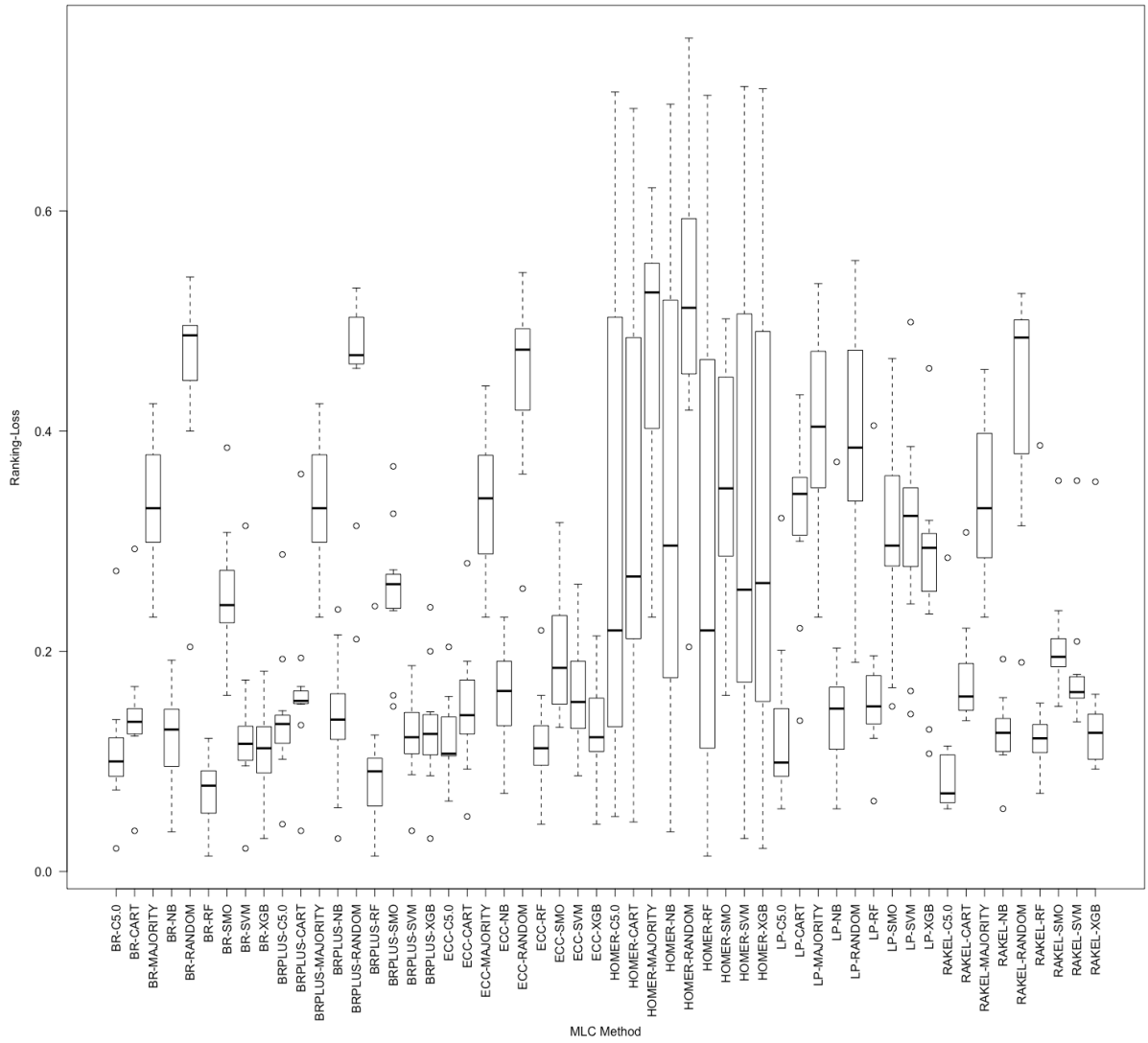



**Figure I. 6.** Significant differences among MLC models applied to CDS1 using ANOVA analysis on Ranking-Loss measure.

## *Ranking-Loss (CDS2)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.29000 -0.01982  0.00320  0.02520  0.40650

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.09100    0.02695   3.377 0.000792 ***
datos$mlc_methodsBR-CART      -0.00050    0.03811  -0.013 0.989538
datos$mlc_methodsBR-MAJORITY   0.15840    0.03811   4.156 3.82e-05 ***
datos$mlc_methodsBR-NB         0.02420    0.03811   0.635 0.525732
datos$mlc_methodsBR-RANDOM     0.32300    0.03811   8.475 2.82e-16 ***
datos$mlc_methodsBR-RF        -0.02840    0.03811  -0.745 0.456512
datos$mlc_methodsBR-SMO        0.05590    0.03811   1.467 0.143082
datos$mlc_methodsBR-SVM       -0.01140    0.03811  -0.299 0.764969
datos$mlc_methodsBR-XGB       -0.00800    0.03811  -0.210 0.833822
datos$mlc_methodsBRPLUS-C5.0   0.00970    0.03811   0.255 0.799200
datos$mlc_methodsBRPLUS-CART   0.01170    0.03811   0.307 0.758974
```

```
datos$mlc_methodsBRPLUS-MAJORITY  0.15830    0.03811   4.154 3.86e-05 ***
datos$mlc_methodsBRPLUS-NB        0.00730    0.03811   0.192 0.848177
datos$mlc_methodsBRPLUS-RANDOM    0.31600    0.03811   8.292 1.10e-15 ***
datos$mlc_methodsBRPLUS-RF       -0.02720    0.03811  -0.714 0.475748
datos$mlc_methodsBRPLUS-SMO       0.05750    0.03811   1.509 0.132009
datos$mlc_methodsBRPLUS-SVM      -0.01270    0.03811  -0.333 0.739097
datos$mlc_methodsBRPLUS-XGB      -0.00810    0.03811  -0.213 0.831776
datos$mlc_methodsECC-C5.0        -0.00660    0.03811  -0.173 0.862582
datos$mlc_methodsECC-CART         0.01800    0.03811   0.472 0.636918
datos$mlc_methodsECC-MAJORITY     0.13120    0.03811   3.443 0.000626 ***
datos$mlc_methodsECC-NB           0.01740    0.03811   0.457 0.648187
datos$mlc_methodsECC-RANDOM       0.32490    0.03811   8.525  < 2e-16 ***
datos$mlc_methodsECC-RF           0.00180    0.03811   0.047 0.962349
datos$mlc_methodsECC-SMO         -0.00210    0.03811  -0.055 0.956079
datos$mlc_methodsECC-SVM          0.00120    0.03811   0.031 0.974894
datos$mlc_methodsECC-XGB          0.00060    0.03811   0.016 0.987445
datos$mlc_methodsHOMER-C5.0       0.20060    0.03811   5.264 2.12e-07 ***
datos$mlc_methodsHOMER-CART       0.18610    0.03811   4.883 1.42e-06 ***
datos$mlc_methodsHOMER-MAJORITY   0.32410    0.03811   8.504 2.27e-16 ***
datos$mlc_methodsHOMER-NB         0.21270    0.03811   5.581 3.98e-08 ***
datos$mlc_methodsHOMER-RANDOM     0.36800    0.03811   9.656  < 2e-16 ***
datos$mlc_methodsHOMER-RF         0.18150    0.03811   4.762 2.53e-06 ***
datos$mlc_methodsHOMER-SMO        0.14160    0.03811   3.715 0.000226 ***
datos$mlc_methodsHOMER-SVM        0.19160    0.03811   5.027 7.00e-07 ***
datos$mlc_methodsHOMER-XGB        0.19230    0.03811   5.046 6.39e-07 ***
datos$mlc_methodsLP-C5.0          0.04620    0.03811   1.212 0.226002
datos$mlc_methodsLP-CART          0.12280    0.03811   3.222 0.001357 **
datos$mlc_methodsLP-MAJORITY      0.25770    0.03811   6.762 3.92e-11 ***
datos$mlc_methodsLP-NB            0.14670    0.03811   3.849 0.000134 ***
datos$mlc_methodsLP-RANDOM        0.28190    0.03811   7.397 6.16e-13 ***
datos$mlc_methodsLP-RF            0.07000    0.03811   1.837 0.066856 .
datos$mlc_methodsLP-SMO           0.05010    0.03811   1.315 0.189267
datos$mlc_methodsLP-SVM           0.16650    0.03811   4.369 1.53e-05 ***
datos$mlc_methodsLP-XGB           0.10910    0.03811   2.863 0.004382 **
datos$mlc_methodsRAKEL-C5.0      -0.00050    0.03811  -0.013 0.989538
datos$mlc_methodsRAKEL-CART       0.01630    0.03811   0.428 0.669057
datos$mlc_methodsRAKEL-MAJORITY   0.15670    0.03811   4.112 4.61e-05 ***
datos$mlc_methodsRAKEL-NB         0.10150    0.03811   2.663 0.007995 **
datos$mlc_methodsRAKEL-RANDOM     0.23880    0.03811   6.266 8.17e-10 ***
datos$mlc_methodsRAKEL-RF         0.01840    0.03811   0.483 0.629452
datos$mlc_methodsRAKEL-SMO        0.02390    0.03811   0.627 0.530874
datos$mlc_methodsRAKEL-SVM        0.05390    0.03811   1.414 0.157913
datos$mlc_methodsRAKEL-XGB       -0.00180    0.03811  -0.047 0.962349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08522 on 486 degrees of freedom
Multiple R-squared:  0.6544,  Adjusted R-squared:  0.6167
F-statistic: 17.36 on 53 and 486 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  53 6.6815 0.126067   17.36 < 2.2e-16 ***
Residuals         486 3.5294 0.007262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

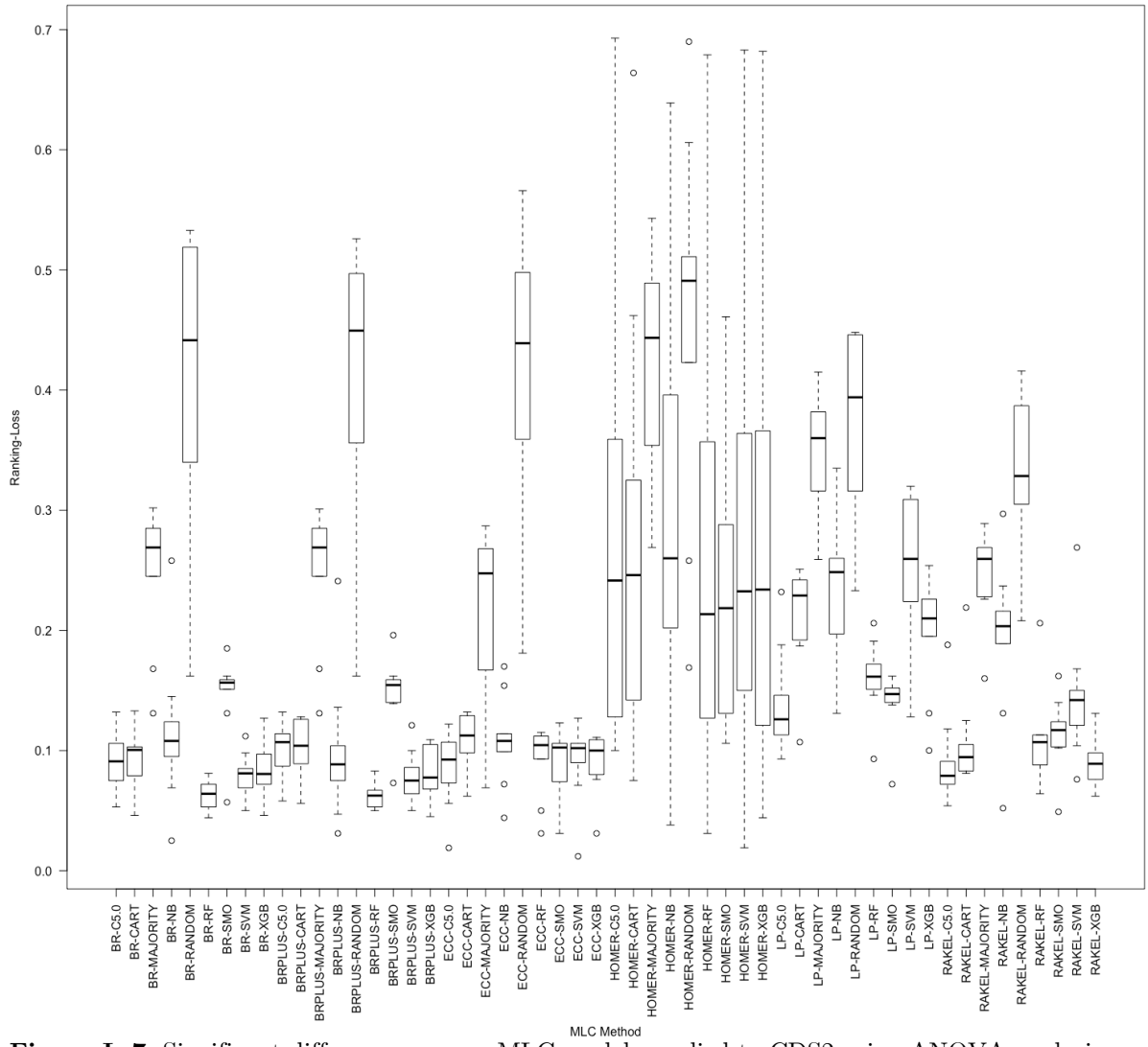



**Figure I. 7.** Significant differences among MLC models applied to CDS2 using ANOVA analysis on Ranking-Loss measure.

## *Ranking-Loss (CDS3)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.41414 -0.00907  0.00493  0.02271  0.10486

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.2441429  0.0281279   8.680 5.03e-16 ***
datos$mlc_methodsBR-CART    -0.0174286  0.0397789  -0.438 0.661663
datos$mlc_methodsBR-NB       0.0934286  0.0397789   2.349 0.019612 *
datos$mlc_methodsBR-RF      -0.0562857  0.0397789  -1.415 0.158314
datos$mlc_methodsBR-SMO      0.0438571  0.0397789   1.103 0.271286
datos$mlc_methodsBR-SVM     -0.0258571  0.0397789  -0.650 0.516271
datos$mlc_methodsBR-XGB     -0.0162857  0.0397789  -0.409 0.682590
datos$mlc_methodsBRPLUS-C5.0  0.0007143  0.0397789   0.018 0.985688
datos$mlc_methodsBRPLUS-CART -0.0181429  0.0397789  -0.456 0.648716
datos$mlc_methodsBRPLUS-NB    0.0397143  0.0397789   0.998 0.319055
datos$mlc_methodsBRPLUS-RF   -0.0557143  0.0397789  -1.401 0.162564
datos$mlc_methodsBRPLUS-SMO   0.0474286  0.0397789   1.192 0.234263
```

```
datos$mlc_methodsBRPLUS-SVM  -0.0314286  0.0397789  -0.790 0.430223
datos$mlc_methodsBRPLUS-XGB  -0.0281429  0.0397789  -0.707 0.479921
datos$mlc_methodsECC-C5.0    -0.0322857  0.0397789  -0.812 0.417771
datos$mlc_methodsECC-CART    -0.0140000  0.0397789  -0.352 0.725173
datos$mlc_methodsECC-NB      -0.0022857  0.0397789  -0.057 0.954224
datos$mlc_methodsECC-RF      -0.0250000  0.0397789  -0.628 0.530263
datos$mlc_methodsECC-SMO     -0.0257143  0.0397789  -0.646 0.518589
datos$mlc_methodsECC-SVM     -0.0181429  0.0397789  -0.456 0.648716
datos$mlc_methodsECC-XGB     -0.0245714  0.0397789  -0.618 0.537331
datos$mlc_methodsHOMER-C5.0   0.3981429  0.0397789  10.009  < 2e-16 ***
datos$mlc_methodsHOMER-CART   0.3867143  0.0397789   9.722  < 2e-16 ***
datos$mlc_methodsHOMER-NB     0.3671429  0.0397789   9.230  < 2e-16 ***
datos$mlc_methodsHOMER-RF     0.3870000  0.0397789   9.729  < 2e-16 ***
datos$mlc_methodsHOMER-SMO    0.2558571  0.0397789   6.432 6.29e-10 ***
datos$mlc_methodsHOMER-SVM    0.4034286  0.0397789  10.142  < 2e-16 ***
datos$mlc_methodsHOMER-XGB    0.3972857  0.0397789   9.987  < 2e-16 ***
datos$mlc_methodsLP-C5.0      0.0688571  0.0397789   1.731 0.084676 .
datos$mlc_methodsLP-CART      0.1242857  0.0397789   3.124 0.001990 **
datos$mlc_methodsLP-NB        0.1001429  0.0397789   2.517 0.012441 *
datos$mlc_methodsLP-RF        0.0652857  0.0397789   1.641 0.102000
datos$mlc_methodsLP-SMO       0.0890000  0.0397789   2.237 0.026137 *
datos$mlc_methodsLP-SVM       0.1215714  0.0397789   3.056 0.002483 **
datos$mlc_methodsLP-XGB       0.1441429  0.0397789   3.624 0.000351 ***
datos$mlc_methodsRAKEL-C5.0  -0.0122857  0.0397789  -0.309 0.757691
datos$mlc_methodsRAKEL-CART   0.0008571  0.0397789   0.022 0.982826
datos$mlc_methodsRAKEL-NB     0.0281429  0.0397789   0.707 0.479921
datos$mlc_methodsRAKEL-RF    -0.0004286  0.0397789  -0.011 0.991412
datos$mlc_methodsRAKEL-SMO    0.0204286  0.0397789   0.514 0.608015
datos$mlc_methodsRAKEL-SVM    0.0282857  0.0397789   0.711 0.477697
datos$mlc_methodsRAKEL-XGB   -0.0070000  0.0397789  -0.176 0.860457
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07442 on 252 degrees of freedom
Multiple R-squared:  0.8086,  Adjusted R-squared:  0.7775
F-statistic: 25.97 on 41 and 252 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  41 5.8974 0.143840  25.972 < 2.2e-16 ***
Residuals         252 1.3956 0.005538
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

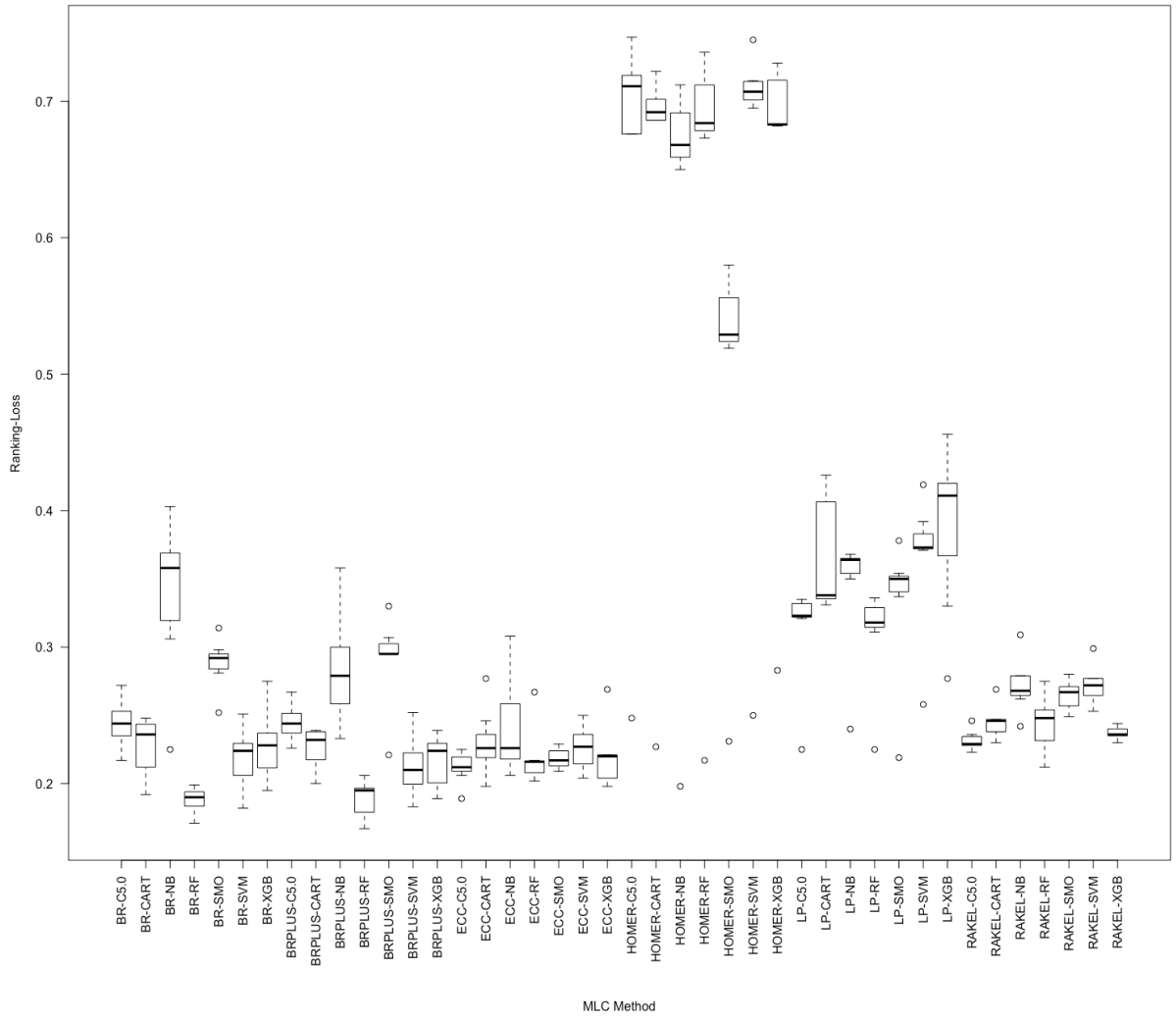



**Figure I. 8.** Significant differences among MLC models applied to CDS3 using ANOVA analysis on Ranking-Loss measure.

## *Ranking-Loss (CDS4)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.38857 -0.00986  0.01493  0.03157  0.17786

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.183143   0.031644   5.788 2.11e-08 ***
datos$mlc_methodsBR-CART      0.052000   0.044751   1.162  0.24634
datos$mlc_methodsBR-NB       -0.030286   0.044751  -0.677  0.49918
datos$mlc_methodsBR-RF       -0.065714   0.044751  -1.468  0.14323
datos$mlc_methodsBR-SMO       0.047571   0.044751   1.063  0.28879
datos$mlc_methodsBR-SVM      -0.012286   0.044751  -0.275  0.78390
datos$mlc_methodsBR-XGB      -0.054714   0.044751  -1.223  0.22261
datos$mlc_methodsBRPLUS-C5.0 -0.011429   0.044751  -0.255  0.79864
datos$mlc_methodsBRPLUS-CART  0.026143   0.044751   0.584  0.55962
datos$mlc_methodsBRPLUS-NB   -0.039000   0.044751  -0.871  0.38432
datos$mlc_methodsBRPLUS-RF   -0.067000   0.044751  -1.497  0.13560
datos$mlc_methodsBRPLUS-SMO   0.058429   0.044751   1.306  0.19287
datos$mlc_methodsBRPLUS-SVM  -0.011571   0.044751  -0.259  0.79618
```

```
datos$mlc_methodsBRPLUS-XGB  -0.036429   0.044751  -0.814  0.41640
datos$mlc_methodsECC-C5.0    -0.025857   0.044751  -0.578  0.56392
datos$mlc_methodsECC-CART     0.003571   0.044751   0.080  0.93645
datos$mlc_methodsECC-NB      -0.030857   0.044751  -0.690  0.49113
datos$mlc_methodsECC-RF      -0.030143   0.044751  -0.674  0.50120
datos$mlc_methodsECC-SMO     -0.003286   0.044751  -0.073  0.94153
datos$mlc_methodsECC-SVM     -0.016857   0.044751  -0.377  0.70672
datos$mlc_methodsECC-XGB     -0.024714   0.044751  -0.552  0.58126
datos$mlc_methodsHOMER-C5.0   0.307143   0.044751   6.863 5.20e-11 ***
datos$mlc_methodsHOMER-CART   0.293714   0.044751   6.563 2.98e-10 ***
datos$mlc_methodsHOMER-NB     0.306429   0.044751   6.847 5.71e-11 ***
datos$mlc_methodsHOMER-RF     0.292000   0.044751   6.525 3.71e-10 ***
datos$mlc_methodsHOMER-SMO    0.185429   0.044751   4.144 4.67e-05 ***
datos$mlc_methodsHOMER-SVM    0.321286   0.044751   7.179 7.84e-12 ***
datos$mlc_methodsHOMER-XGB    0.293857   0.044751   6.566 2.92e-10 ***
datos$mlc_methodsLP-C5.0      0.091286   0.044751   2.040  0.04241 *
datos$mlc_methodsLP-CART      0.107857   0.044751   2.410  0.01666 *
datos$mlc_methodsLP-NB        0.012857   0.044751   0.287  0.77412
datos$mlc_methodsLP-RF        0.054857   0.044751   1.226  0.22141
datos$mlc_methodsLP-SMO       0.112000   0.044751   2.503  0.01296 *
datos$mlc_methodsLP-SVM       0.195286   0.044751   4.364 1.87e-05 ***
datos$mlc_methodsLP-XGB       0.136571   0.044751   3.052  0.00252 **
datos$mlc_methodsRAKEL-C5.0  -0.004000   0.044751  -0.089  0.92885
datos$mlc_methodsRAKEL-CART   0.011857   0.044751   0.265  0.79126
datos$mlc_methodsRAKEL-NB    -0.018143   0.044751  -0.405  0.68552
datos$mlc_methodsRAKEL-RF    -0.005286   0.044751  -0.118  0.90607
datos$mlc_methodsRAKEL-SMO    0.019286   0.044751   0.431  0.66687
datos$mlc_methodsRAKEL-SVM    0.033286   0.044751   0.744  0.45769
datos$mlc_methodsRAKEL-XGB   -0.025143   0.044751  -0.562  0.57473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08372 on 252 degrees of freedom
Multiple R-squared:  0.691,   Adjusted R-squared:  0.6408
F-statistic: 13.75 on 41 and 252 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                   Df Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods  41 3.9508 0.096362  13.748 < 2.2e-16 ***
Residuals         252 1.7664 0.007009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

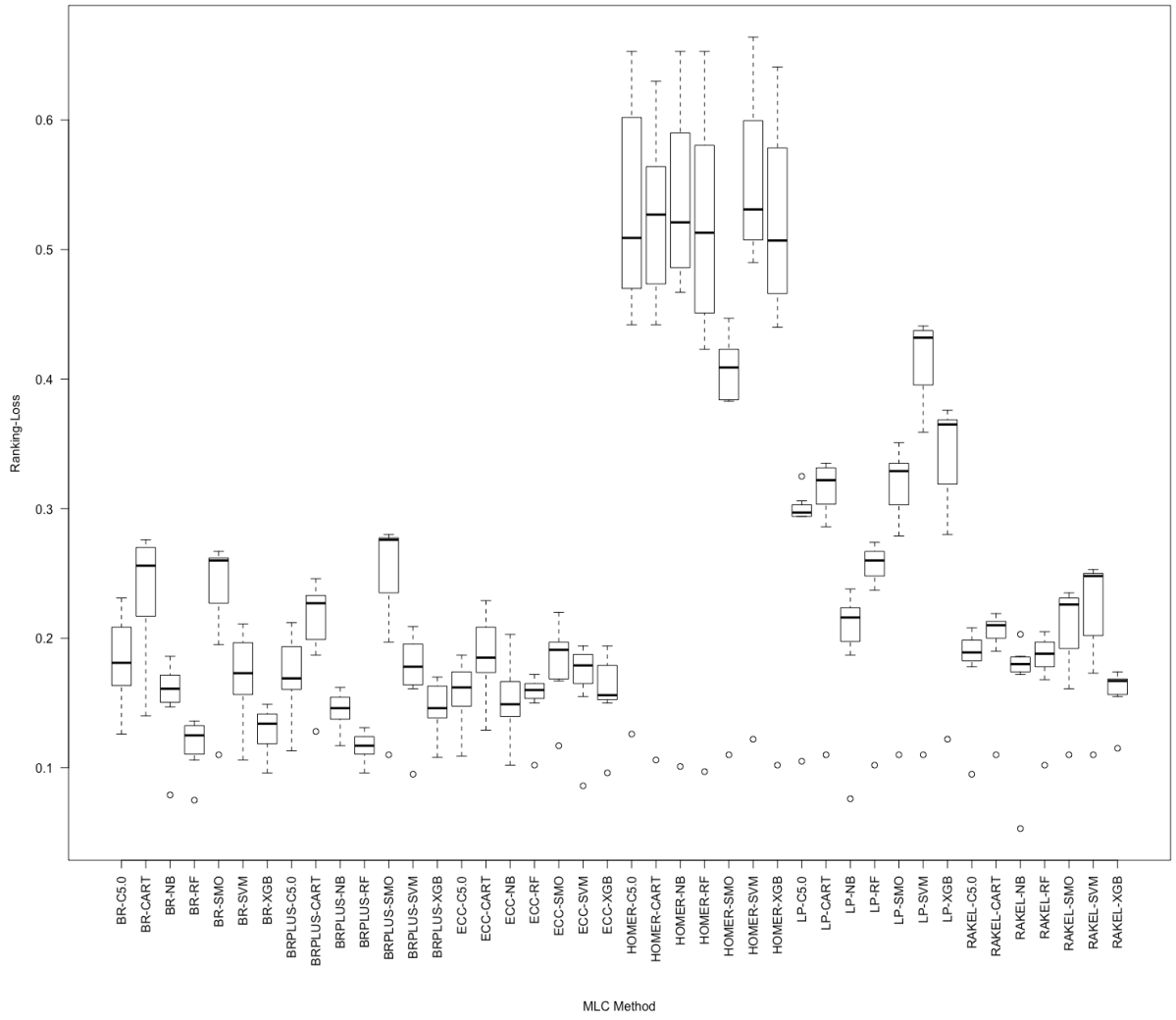



**Figure I. 9.** Significant differences among MLC models applied to CDS4 using ANOVA analysis on Ranking-Loss measure.

## *Ranking-Loss (CGD)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.23862 -0.04425 -0.00640  0.04121  0.29600

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.184057   0.011957  15.394  < 2e-16 ***
datos$mlc_methodsBR-CART        0.026200   0.016909   1.549 0.121461
datos$mlc_methodsBR-MAJORITY    0.110705   0.019525   5.670 1.68e-08 ***
datos$mlc_methodsBR-NB          0.058371   0.016909   3.452 0.000570 ***
datos$mlc_methodsBR-RANDOM      0.308848   0.019525  15.818  < 2e-16 ***
datos$mlc_methodsBR-RF         -0.019543   0.016909  -1.156 0.247945
datos$mlc_methodsBR-SMO         0.006486   0.016909   0.384 0.701351
datos$mlc_methodsBR-SVM         0.001914   0.016909   0.113 0.909877
datos$mlc_methodsBR-XGB         0.011086   0.016909   0.656 0.512168
datos$mlc_methodsBRPLUS-C5.0    0.009714   0.016909   0.575 0.565707
datos$mlc_methodsBRPLUS-CART    0.030571   0.016909   1.808 0.070789 .
datos$mlc_methodsBRPLUS-MAJORITY 0.110848  0.019525   5.677 1.61e-08 ***
datos$mlc_methodsBRPLUS-NB      0.053514   0.016909   3.165 0.001580 **
```

```
datos$mlc_methodsBRPLUS-RANDOM    0.307371   0.019525  15.742  < 2e-16 ***
datos$mlc_methodsBRPLUS-RF       -0.018343   0.016909  -1.085 0.278171
datos$mlc_methodsBRPLUS-SMO       0.008657   0.016909   0.512 0.608732
datos$mlc_methodsBRPLUS-SVM       0.006429   0.016909   0.380 0.703857
datos$mlc_methodsBRPLUS-XGB       0.013000   0.016909   0.769 0.442111
datos$mlc_methodsECC-C5.0         0.015400   0.016909   0.911 0.362558
datos$mlc_methodsECC-CART         0.035857   0.016909   2.121 0.034104 *
datos$mlc_methodsECC-MAJORITY     0.127990   0.019525   6.555 7.39e-11 ***
datos$mlc_methodsECC-NB           0.042371   0.016909   2.506 0.012311 *
datos$mlc_methodsECC-RANDOM       0.244324   0.019525  12.513  < 2e-16 ***
datos$mlc_methodsECC-RF           0.010486   0.016909   0.620 0.535262
datos$mlc_methodsECC-SMO          0.032429   0.016909   1.918 0.055305 .
datos$mlc_methodsECC-SVM          0.028257   0.016909   1.671 0.094886 .
datos$mlc_methodsECC-XGB          0.026057   0.016909   1.541 0.123503
datos$mlc_methodsHOMER-C5.0       0.085886   0.016909   5.079 4.21e-07 ***
datos$mlc_methodsHOMER-CART       0.110086   0.016909   6.510 9.89e-11 ***
datos$mlc_methodsHOMER-MAJORITY   0.244562   0.019525  12.526  < 2e-16 ***
datos$mlc_methodsHOMER-NB         0.115400   0.016909   6.825 1.23e-11 ***
datos$mlc_methodsHOMER-RANDOM     0.240133   0.019525  12.299  < 2e-16 ***
datos$mlc_methodsHOMER-RF         0.087857   0.016909   5.196 2.29e-07 ***
datos$mlc_methodsHOMER-SMO        0.112914   0.016909   6.678 3.30e-11 ***
datos$mlc_methodsHOMER-SVM        0.115600   0.016909   6.837 1.13e-11 ***
datos$mlc_methodsHOMER-XGB        0.099314   0.016909   5.873 5.14e-09 ***
datos$mlc_methodsLP-C5.0         -0.017086   0.016909  -1.010 0.312429
datos$mlc_methodsLP-CART          0.061429   0.016909   3.633 0.000289 ***
datos$mlc_methodsLP-MAJORITY      0.170800   0.019525   8.748  < 2e-16 ***
datos$mlc_methodsLP-NB            0.041971   0.016909   2.482 0.013156 *
datos$mlc_methodsLP-RANDOM        0.208038   0.019525  10.655  < 2e-16 ***
datos$mlc_methodsLP-RF           -0.005457   0.016909  -0.323 0.746937
datos$mlc_methodsLP-SMO           0.038200   0.016909   2.259 0.024003 *
datos$mlc_methodsLP-SVM           0.072629   0.016909   4.295 1.85e-05 ***
datos$mlc_methodsLP-XGB           0.045371   0.016909   2.683 0.007363 **
datos$mlc_methodsRAKEL-C5.0      -0.025886   0.016909  -1.531 0.125990
datos$mlc_methodsRAKEL-CART       0.012629   0.016909   0.747 0.455259
datos$mlc_methodsRAKEL-MAJORITY   0.123943   0.019525   6.348 2.81e-10 ***
datos$mlc_methodsRAKEL-NB         0.062571   0.016909   3.700 0.000222 ***
datos$mlc_methodsRAKEL-RANDOM     0.285419   0.019525  14.618  < 2e-16 ***
datos$mlc_methodsRAKEL-RF        -0.023743   0.016909  -1.404 0.160462
datos$mlc_methodsRAKEL-SMO        0.004600   0.016909   0.272 0.785623
datos$mlc_methodsRAKEL-SVM        0.007114   0.016909   0.421 0.674001
datos$mlc_methodsRAKEL-XGB       -0.010600   0.016909  -0.627 0.530823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07074 on 1668 degrees of freedom
Multiple R-squared:  0.5504,  Adjusted R-squared:  0.5361
F-statistic: 38.53 on 53 and 1668 DF,  p-value: < 2.2e-16

Analysis of Variance Table

                    Df  Sum Sq  Mean Sq F value    Pr(>F)
datos$mlc_methods   53 10.2169 0.192772  38.527 < 2.2e-16 ***
Residuals         1668  8.3459 0.005004
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

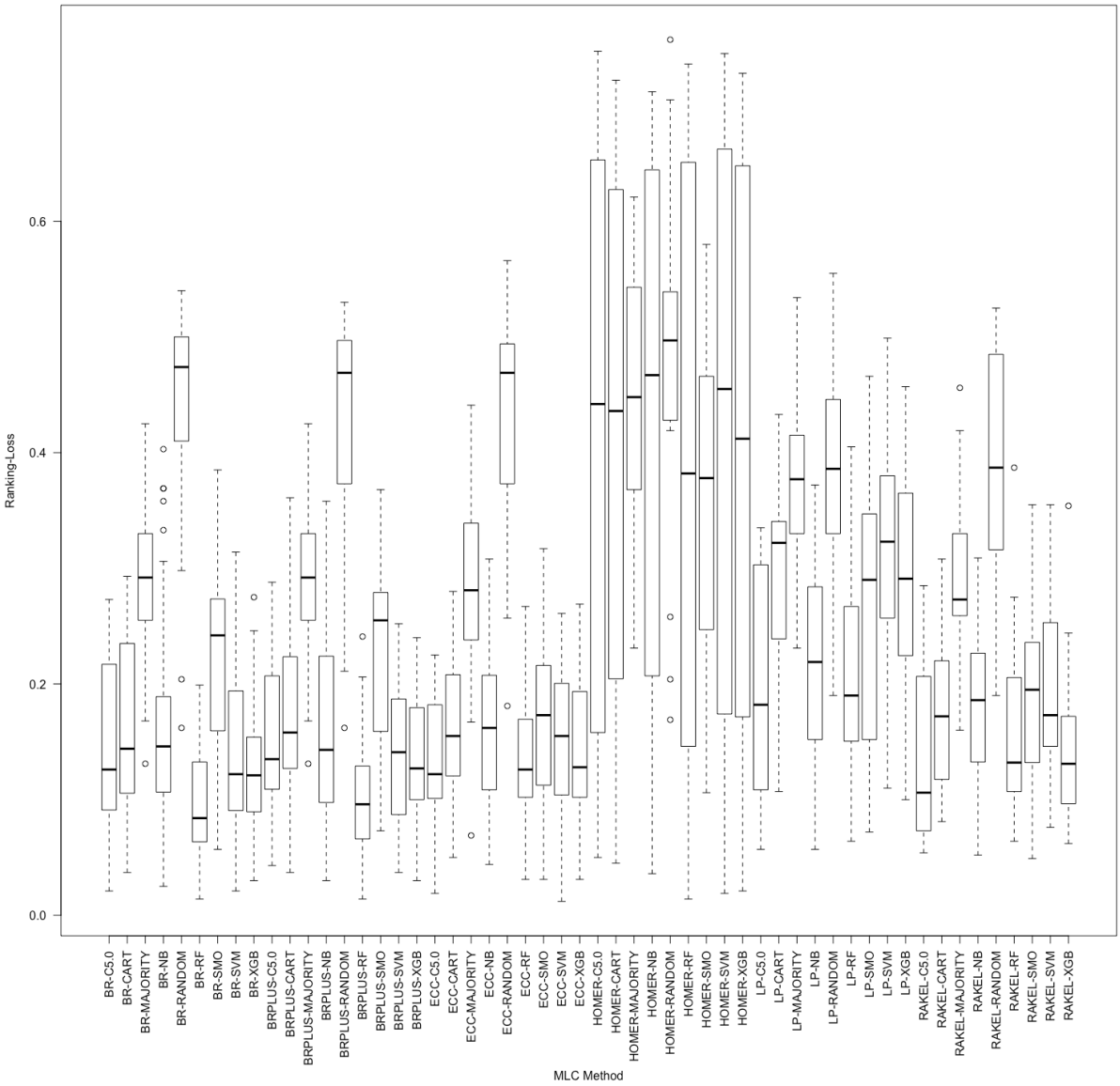



**Figure I. 10.** Significant differences among MLC models applied to CDS1 using ANOVA analysis on Ranking-Loss measure.

# Appendix J

# Model Validation

This appendix presents the tests with real agricultural crop production data to validate the MLC models. Initially, yesterday we found the crop production trends in each municipality and organized them as a ranking. Subsequently, we compare these actual rankings with the rankings predicted by the MLC models. For this purpose, we use the ULS (Unranked Lists' Similarity) and RBO (Rank Biased Overlap) similarity measures, the former omitting the order of the ranking elements, while the latter considers it strictly. The results of these tests are presented below.

## J.1. ULS and RBO Similarities in CDS1

| MLC-Models | RBO | ULS |
|---|---|---|
| BR-RF | 67.36 | 87.93 |
| BRPLUS-RF | 64.67 | 88.47 |
| BR-SVM | 64.13 | 90.37 |
| BR-C5.0 | 64.12 | 90.37 |
| BR-XGB | 63.85 | 90.37 |
| BRPLUS-C5.0 | 61.61 | 90.37 |
| BR-NB | 61.51 | 88.64 |
| BRPLUS-XGB | 60.72 | 90.37 |
| BR-CART | 58.86 | 87.06 |
| BR-KNN | 58.70 | 85.70 |
| BRPLUS-CART | 56.37 | 83.77 |
| ECC-C5.0 | 55.34 | 81.33 |
| ECC-XGB | 54.93 | 79.67 |
| ECC-NB | 54.25 | 83.80 |
| ECC-CART | 53.99 | 76.26 |
| ECC-KNN | 53.69 | 81.58 |
| BRPLUS-NB | 52.49 | 82.49 |
| ECC-RF | 52.35 | 77.47 |
| ECC-SMO | 49.81 | 77.96 |
| LP-C5.0 | 47.70 | 70.29 |
| BRPLUS-KNN | 46.90 | 71.23 |

| | | |
|---|---|---|
| RAKEL-C5.0 | 46.80 | 75.60 |
| LP-RF | 45.37 | 70.16 |
| RAKEL-CART | 45.03 | 64.28 |
| RAKEL-RF | 44.37 | 68.67 |
| RAKEL-SVM | 43.93 | 57.35 |
| RAKEL-XGB | 43.35 | 69.50 |
| LP-KNN | 42.25 | 65.94 |
| RAKEL-NB | 42.15 | 73.61 |
| RAKEL-SMO | 41.73 | 51.35 |
| LP-NB | 41.08 | 63.12 |
| BR-SMO | 40.55 | 41.89 |
| RAKEL-KNN | 39.65 | 61.37 |
| BRPLUS-SMO | 37.83 | 39.83 |
| LP-XGB | 36.08 | 47.60 |
| LP-CART | 33.15 | 38.27 |
| LP-SMO | 31.97 | 49.71 |
| LP-SVM | 27.21 | 37.03 |

**Table J. 1.** ULS and RBO Similarities in MLC models applied to CDS1.

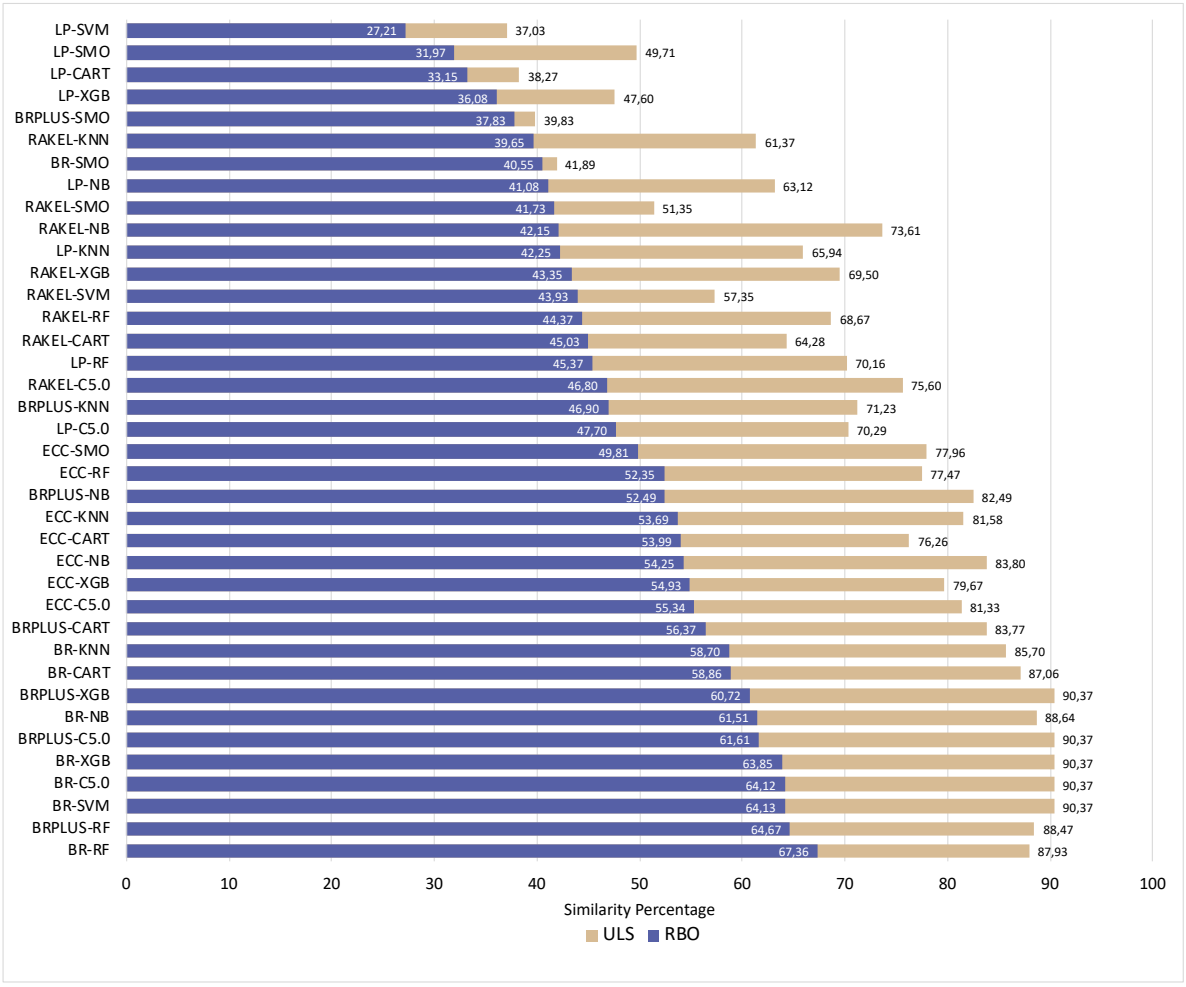



**Figure J. 1.** ULS and RBO Similarities in MLC models applied to CDS1.

| MUNICIPALITY | RBO (Actual Data) | RBO (Training Data) | Variation |
|---|---|---|---|
| BUENOSAIRES | 57% | 62% | 5% |
| CAJIBIO | 58% | 60% | 2% |
| CALDONO | 71% | 75% | 4% |
| CALOTO | 64% | 78% | 14% |
| CORINTO | 68% | 78% | 10% |
| ELTAMBO | 65% | 69% | 4% |
| GUACHENE | 71% | 85% | 14% |
| JAMBALO | 64% | 69% | 5% |
| MIRANDA | 69% | 79% | 10% |
| MORALES | 70% | 68% | -2% |
| PADILLA | 60% | 71% | 11% |
| PIENDAMO | 72% | 70% | -2% |
| POPAYAN | 50% | 62% | 12% |
| PUERTOTEJADA | 73% | 70% | -3% |
| PURACE | 64% | 68% | 4% |
| SANTANDERDEQUILICHAO | 67% | 72% | 5% |
| SILVIA | 65% | 77% | 12% |
| SOTARA | 76% | 78% | 2% |
| SUAREZ | 53% | 62% | 9% |
| TIMBIO | 69% | 66% | -3% |
| TORIBIO | 43% | 59% | 16% |
| TOTORO | 76% | 80% | 4% |
| VILLARICA | 64% | 78% | 14% |

**Table J. 2.** Variation in RBO similarities per municipality (actual vs. training data – CDS1).

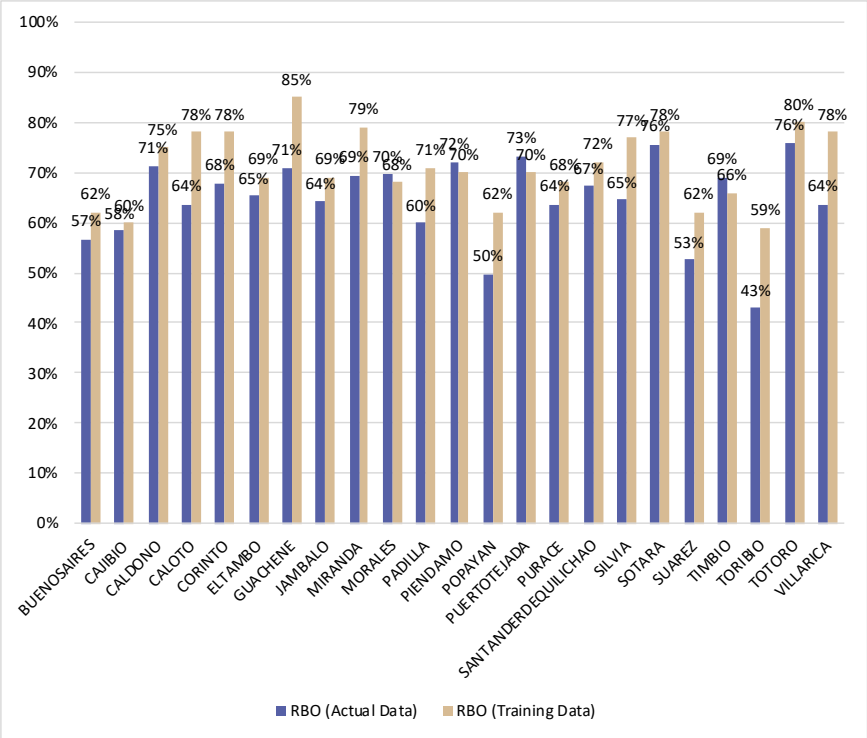



**Figure J. 2.** Variation in RBO similarities per municipality (actual vs. training data – CDS1).

## J.2. ULS and RBO Similarities in CDS2

| MLC-Models | RBO | ULS |
|---|---|---|
| BR-RF | 61.49 | 95.61 |
| BR-XGB | 61.19 | 96.87 |
| BRPLUS-XGB | 60.83 | 96.87 |
| BRPLUS-RF | 60.82 | 96.48 |
| BRPLUS-C5.0 | 60.41 | 96.87 |
| BR-C5.0 | 58.87 | 96.87 |
| BR-CART | 57.95 | 84.44 |
| BRPLUS-CART | 57.57 | 88.36 |
| ECC-SMO | 55.00 | 80.89 |
| ECC-C5.0 | 54.77 | 83.73 |
| BR-KNN | 53.13 | 75.49 |
| ECC-CART | 51.80 | 80.75 |
| ECC-KNN | 50.15 | 75.52 |
| BRPLUS-KNN | 48.91 | 67.19 |
| ECC-XGB | 47.63 | 76.81 |
| ECC-NB | 45.08 | 86.52 |
| RAKEL-C5.0 | 43.84 | 67.31 |
| RAKEL-RF | 43.39 | 54.84 |
| ECC-RF | 42.11 | 63.68 |
| RAKEL-SMO | 40.33 | 58.39 |
| BRPLUS-SMO | 39.23 | 52.99 |
| BR-SMO | 38.51 | 50.32 |
| LP-SMO | 37.17 | 46.98 |
| LP-C5.0 | 36.51 | 49.90 |
| RAKEL-CART | 36.08 | 59.65 |
| LP-CART | 35.80 | 48.26 |
| RAKEL-XGB | 35.55 | 50.23 |
| RAKEL-KNN | 33.08 | 46.13 |
| LP-KNN | 32.85 | 45.31 |
| LP-NB | 32.82 | 46.74 |
| LP-RF | 31.18 | 43.01 |
| RAKEL-NB | 29.36 | 85.62 |
| BRPLUS-NB | 28.91 | 74.55 |
| BR-NB | 28.88 | 88.99 |
| LP-XGB | 27.25 | 36.14 |
| HOMER-RF | 16.11 | 26.95 |
| HOMER-C5.0 | 15.33 | 27.94 |
| HOMER-CART | 14.94 | 25.47 |
| HOMER-XGB | 14.11 | 26.02 |
| HOMER-NB | 13.14 | 23.63 |
| HOMER-SMO | 12.42 | 7.11 |
| HOMER-KNN | 12.37 | 21.14 |

**Table J. 3.** ULS and RBO Similarities in MLC models applied to CDS2.

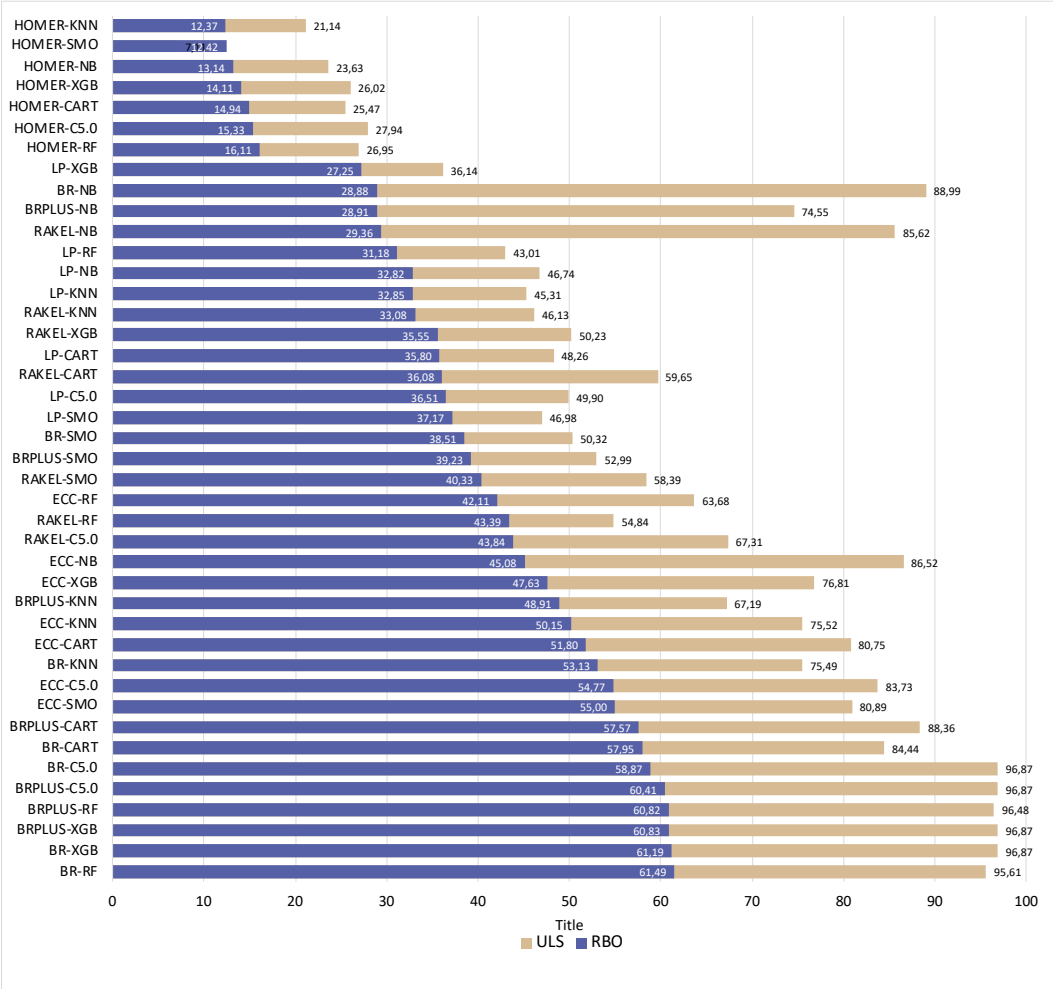



**Figure J. 3.** ULS and RBO Similarities in MLC models applied to CDS2.

| MUNICIPALITY | RBO (Actual Values) | RBO (Training Data) | Variation |
|---|---|---|---|
| ARGELIA | 78% | 72% | -6% |
| CAJIBIO | 69% | 70% | 1% |
| CALOTO | 65% | 72% | 7% |
| ELTAMBO | 71% | 74% | 3% |
| GUAPI | 64% | 72% | 8% |
| INZA | 51% | 71% | 20% |
| MERCADERES | 45% | 68% | 23% |
| MIRANDA | 84% | 75% | -9% |
| PATIA | 77% | 78% | 1% |
| PIAMONTE | 25% | 62% | 37% |
| PIENDAMO | 65% | 63% | -2% |
| POPAYAN | 55% | 60% | 5% |
| PUERTOTEJADA | 47% | 69% | 22% |
| SANTANDERDEQUILICHAO | 78% | 82% | 4% |
| SOTARA | 56% | 75% | 19% |
| TIMBIO | 61% | 65% | 4% |
| TOTORO | 43% | 58% | 15% |

**Table J. 4.** Variation in RBO similarities per municipality (actual vs. training data – CDS2).

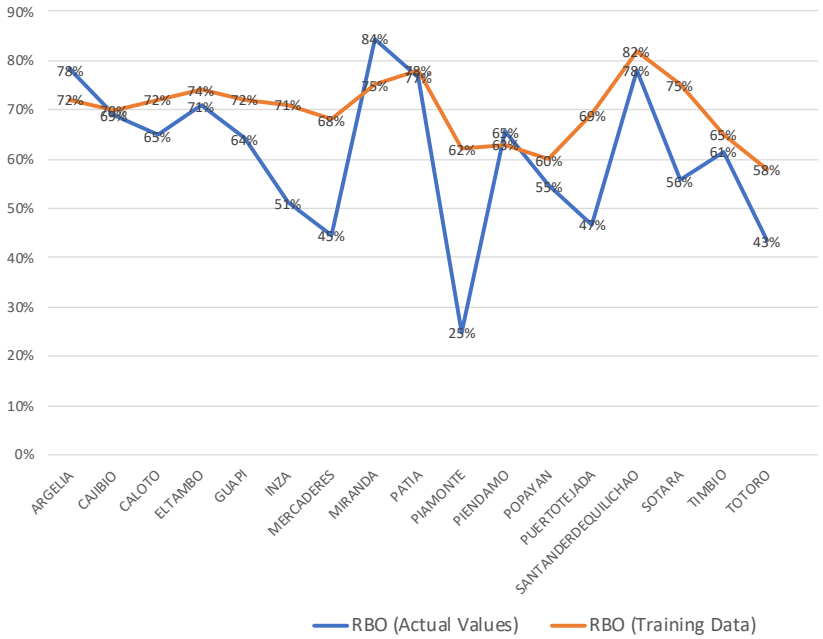



**Figure J. 4.** Variation in RBO similarities per municipality (actual vs. training data – CDS2).

# J.3. ULS and RBO Similarities in CDS3

| MLC-Models | RBO | ULS |
|---|---|---|
| BR-CART | 55.87 | 94.68 |
| BRPLUS-XGB | 55.67 | 96.54 |
| BRPLUS-CART | 55.63 | 93.60 |
| BR-XGB | 54.92 | 96.54 |
| BRPLUS-C5.0 | 53.84 | 96.54 |
| BR-C5.0 | 52.32 | 96.54 |
| ECC-C5.0 | 52.22 | 83.18 |
| ECC-CART | 51.97 | 80.75 |
| ECC-NB | 50.91 | 81.67 |
| ECC-XGB | 50.58 | 80.91 |
| ECC-SMO | 47.22 | 73.24 |
| RAKEL-C5.0 | 41.93 | 65.71 |
| RAKEL-CART | 39.08 | 60.79 |
| BRPLUS-NB | 39.06 | 73.71 |
| BR-NB | 38.54 | 81.82 |
| RAKEL-NB | 38.45 | 66.06 |
| RAKEL-XGB | 38.42 | 63.11 |
| RAKEL-SMO | 35.54 | 47.47 |
| BR-SMO | 34.83 | 39.05 |
| BRPLUS-SMO | 34.33 | 38.06 |
| LP-C5.0 | 30.63 | 40.92 |
| LP-SMO | 29.91 | 45.34 |
| LP-CART | 27.10 | 31.61 |
| LP-NB | 25.15 | 39.45 |
| HOMER-CART | 13.19 | 27.00 |
| LP-XGB | 13.13 | 12.90 |
| HOMER-C5.0 | 11.83 | 27.41 |

| | | |
|---|---|---|
| HOMER-XGB | 11.83 | 27.82 |
| HOMER-NB | 8.44 | 20.88 |
| HOMER-SMO | 2.58 | 5.82 |

**Table J. 5.** ULS and RBO Similarities in MLC models applied to CDS3.

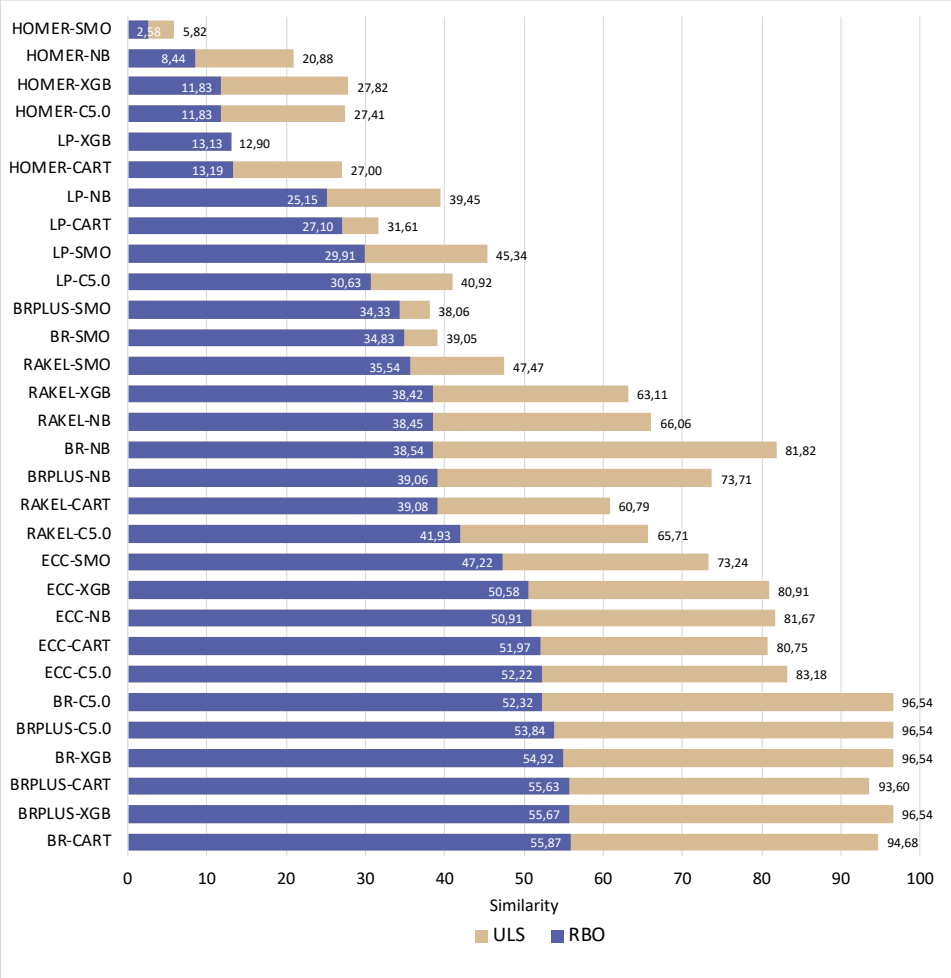



**Figure J. 5.** ULS and RBO Similarities in MLC models applied to CDS3.

| MUNICIPALITY | RBO (Actual Values) | RBO (Training Data) | Variation |
|---|---|---|---|
| ALMAGUER | 58% | 67% | 9% |
| ARGELIA | 60% | 63% | 3% |
| BALBOA | 57% | 62% | 5% |
| BOLIVAR | 62% | 65% | 3% |
| BUENOSAIRES | 63% | 66% | 3% |
| CAJIBIO | 55% | 63% | 8% |
| CALDONO | 67% | 65% | -2% |
| CALOTO | 56% | 69% | 13% |
| CORINTO | 69% | 63% | -6% |
| ELTAMBO | 55% | 60% | 5% |
| FLORENCIA | 53% | 65% | 12% |
| INZA | 65% | 67% | 2% |
| JAMBALO | 53% | 64% | 11% |

| | | | |
|---|---|---|---|
| LASIERRA | 56% | 64% | 8% |
| LAVEGA | 56% | 63% | 7% |
| MERCADERES | 49% | 58% | 9% |
| MIRANDA | 59% | 65% | 6% |
| MORALES | 68% | 64% | -4% |
| PADILLA | 55% | 65% | 10% |
| PAEZ | 53% | 59% | 6% |
| PATIA | 45% | 57% | 12% |
| PIAMONTE | 39% | 55% | 16% |
| PIENDAMO | 60% | 65% | 5% |
| POPAYAN | 46% | 66% | 20% |
| PUERTOTEJADA | 57% | 68% | 11% |
| PURACE | 53% | 61% | 8% |
| ROSAS | 52% | 60% | 8% |
| SANSEBASTIAN | 65% | 64% | -1% |
| SANTANDERDEQUILICHAO | 67% | 67% | 0% |
| SANTAROSA | 59% | 62% | 3% |
| SILVIA | 36% | 58% | 22% |
| SOTARA | 56% | 60% | 4% |
| SUAREZ | 57% | 65% | 8% |
| SUCRE | 66% | 69% | 3% |
| TIMBIO | 56% | 66% | 10% |
| TIMBIQUI | 28% | 59% | 31% |
| TORIBIO | 48% | 63% | 15% |
| TOTORO | 54% | 60% | 6% |
| VILLARICA | 56% | 58% | 2% |

**Table J. 6.** Variation in RBO similarities per municipality (actual vs. training data – CDS3).

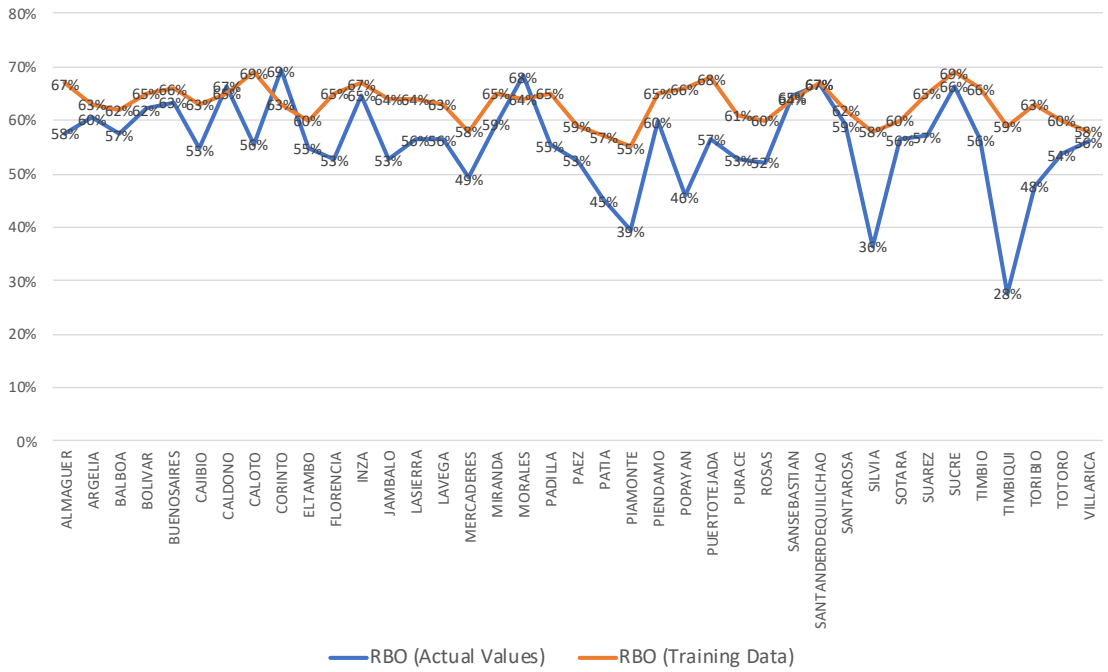



**Figure J. 6.** Variation in RBO similarities per municipality (actual vs. training data – CDS3).

## J.4. ULS and RBO Similarities in CDS4

| MLC-Models | RBO | ULS |
|---|---|---|
| BR-RF | 61.18 | 94.66 |
| BR-XGB | 59.33 | 94.66 |
| BRPLUS-RF | 58.49 | 94.66 |
| BR-NB | 58.08 | 91.97 |
| BRPLUS-XGB | 57.51 | 94.66 |
| ECC-RF | 55.91 | 87.50 |
| BRPLUS-NB | 50.91 | 84.42 |
| ECC-NB | 50.18 | 79.98 |
| ECC-KNN | 49.40 | 78.26 |
| ECC-XGB | 48.93 | 70.57 |
| ECC-SMO | 48.01 | 72.20 |
| BR-KNN | 41.56 | 67.31 |
| RAKEL-XGB | 39.27 | 53.56 |
| BRPLUS-KNN | 37.33 | 46.53 |
| BR-SMO | 36.04 | 48.61 |
| RAKEL-RF | 34.43 | 49.70 |
| RAKEL-SMO | 31.21 | 53.83 |
| RAKEL-NB | 30.27 | 57.70 |
| BRPLUS-SMO | 30.10 | 47.74 |
| LP-NB | 24.64 | 48.67 |
| LP-SMO | 20.65 | 39.96 |
| HOMER-RF | 17.86 | 27.78 |
| HOMER-KNN | 17.79 | 24.25 |
| HOMER-XGB | 17.67 | 30.81 |
| HOMER-NB | 17.37 | 29.33 |
| LP-RF | 16.80 | 36.89 |
| HOMER-SMO | 13.10 | 12.50 |
| LP-XGB | -8.15 | 18.72 |
| RAKEL-KNN | -17.10 | 18.08 |
| LP-KNN | -18.34 | 28.76 |

**Table J. 7.** ULS and RBO Similarities in MLC models applied to CDS4.

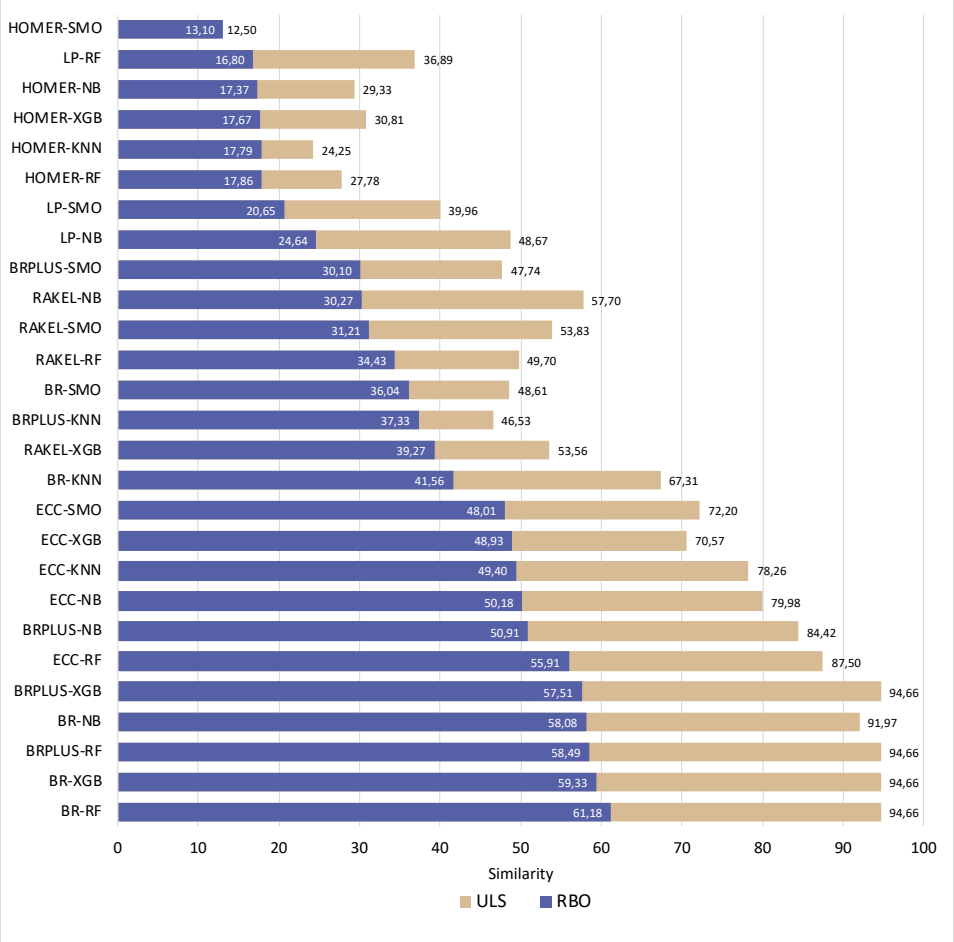



**Figure J. 7.** ULS and RBO Similarities in MLC models applied to CDS4.

| MUNICIPALITY | RBO (Actual Values) | RBO (Training Data) | Variation |
|---|---|---|---|
| ARGELIA | 56% | 65% | 9% |
| BUENOSAIRES | 74% | 76% | 2% |
| CAJIBIO | 63% | 72% | 9% |
| CALDONO | 67% | 74% | 7% |
| CALOTO | 50% | 68% | 18% |
| ELTAMBO | 73% | 75% | 2% |
| FLORENCIA | 48% | 65% | 17% |
| INZA | 60% | 62% | 2% |
| LASIERRA | 53% | 63% | 10% |
| MERCADERES | 71% | 69% | -2% |
| MORALES | 42% | 56% | 14% |
| PAEZ | 50% | 67% | 17% |
| PATIA | 80% | 75% | -5% |
| PIENDAMO | 69% | 71% | 2% |
| POPAYAN | 53% | 66% | 13% |
| PUERTOTEJADA | 38% | 61% | 23% |
| PURACE | 80% | 70% | -10% |
| ROSAS | 40% | 65% | 25% |
| SANTANDERDEQUILICHAO | 81% | 76% | -5% |
| SILVIA | 59% | 62% | 3% |

| | | | |
|---|---|---|---|
| SOTARA | 57% | 62% | 5% |
| TIMBIO | 61% | 68% | 7% |
| TIMBIQUI | 28% | 66% | 38% |
| TORIBIO | 50% | 57% | 7% |

**Table J. 8.** Variation in RBO similarities per municipality (actual vs. training data – CDS4).

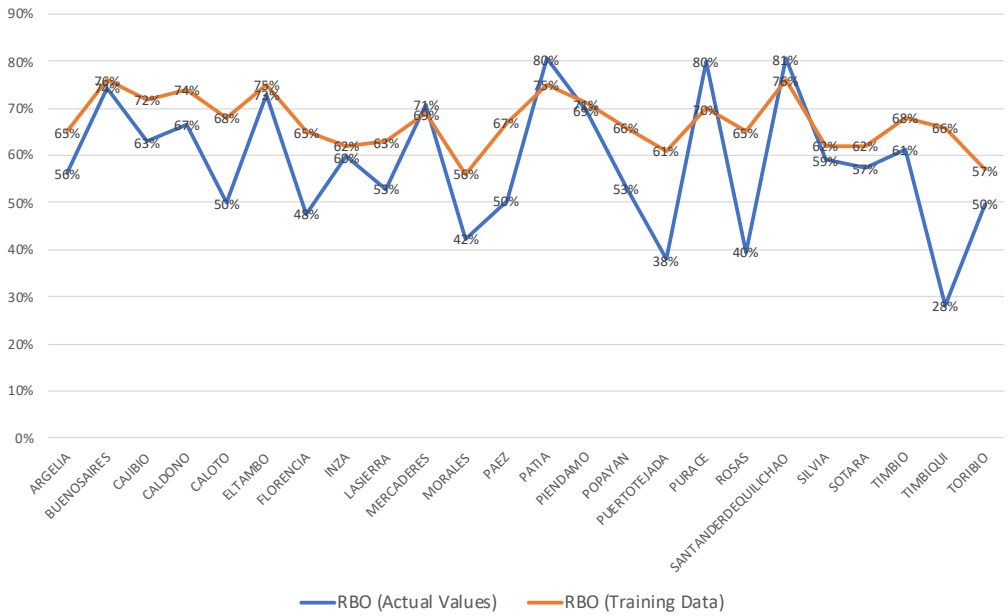



**Figure J. 8.** Variation in RBO similarities per municipality (actual vs. training data – CDS4).

# J.5. ULS and RBO Similarities in CGD

| MLC-Models | RBO | ULS |
|---|---|---|
| BRPLUS-RF | 64.85 | 97.88 |
| BR-RF | 64.00 | 98.06 |
| BR-XGB | 61.17 | 98.06 |
| BRPLUS-XGB | 60.73 | 98.06 |
| ECC-RF | 57.02 | 80.99 |
| HOMER-RF | 56.97 | 93.86 |
| BR-CART | 54.19 | 71.60 |
| BRPLUS-CART | 53.75 | 72.23 |
| HOMER-XGB | 53.08 | 94.75 |
| ECC-CART | 52.90 | 76.75 |
| ECC-KNN | 52.36 | 77.13 |
| ECC-SMO | 50.90 | 78.46 |
| HOMER-CART | 49.78 | 70.20 |
| HOMER-C5.0 | 49.50 | 94.88 |
| ECC-XGB | 49.47 | 78.01 |
| BR-KNN | 49.17 | 78.62 |
| RAKEL-XGB | 40.97 | 64.69 |
| BRPLUS-KNN | 40.51 | 66.44 |
| ECC-NB | 40.50 | 92.54 |
| RAKEL-SMO | 37.72 | 57.83 |
| BR-SMO | 35.47 | 48.64 |

| | | |
|---|---|---|
| RAKEL-RF | 34.92 | 57.45 |
| LP-RF | 34.35 | 53.80 |
| BRPLUS-SMO | 33.54 | 50.14 |
| LP-SMO | 32.51 | 50.43 |
| LP-NB | 32.50 | 51.67 |
| LP-KNN | 32.44 | 43.37 |
| RAKEL-KNN | 31.45 | 43.26 |
| HOMER-SMO | 29.85 | 50.82 |
| LP-XGB | 25.29 | 34.00 |
| BRPLUS-NB | 22.32 | 69.58 |
| RAKEL-NB | 22.21 | 92.87 |
| BR-NB | 12.35 | 78.91 |
| HOMER-NB | 8.28 | 16.14 |

**Table J. 9**. ULS and RBO Similarities in MLC models applied to CGD.

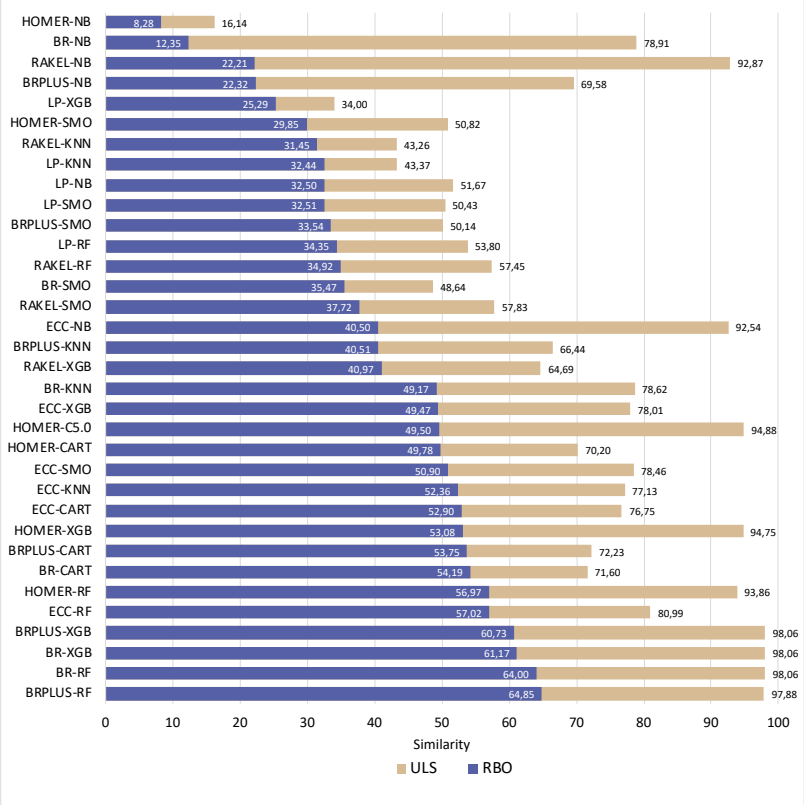



**Figure J. 9.** ULS and RBO Similarities in MLC models applied to CGD.

| **MUNICIPALITY** | **RBO (Actual Data)** | **RBO (Training Data)** | Variation |
|---|---|---|---|
| ALMAGUER | 61% | 86% | 25% |
| ARGELIA | 58% | 75% | 17% |
| BALBOA | 57% | 62% | 5% |
| BOLIVAR | 66% | 88% | 22% |
| BUENOSAIRES | 82% | 84% | 2% |
| CAJIBIO | 67% | 76% | 9% |
| CALDONO | 72% | 89% | 17% |
| CALOTO | 60% | 65% | 5% |
| CORINTO | 73% | 79% | 6% |

| | | | |
|---|---|---|---|
| ELTAMBO | 74% | 86% | 12% |
| FLORENCIA | 60% | 71% | 11% |
| GUACHENE | 78% | 79% | 1% |
| INZA | 66% | 77% | 11% |
| JAMBALO | 56% | 52% | -4% |
| LASIERRA | 65% | 75% | 10% |
| LAVEGA | 65% | 72% | 7% |
| MERCADERES | 79% | 87% | 8% |
| MIRANDA | 67% | 78% | 11% |
| MORALES | 69% | 79% | 10% |
| PADILLA | 61% | 61% | 0% |
| PAEZ | 59% | 67% | 8% |
| PATIA | 78% | 90% | 12% |
| PIAMONTE | 53% | 66% | 13% |
| PIENDAMO | 74% | 81% | 7% |
| POPAYAN | 58% | 64% | 6% |
| PUERTOTEJADA | 60% | 72% | 12% |
| PURACE | 88% | 92% | 4% |
| ROSAS | 54% | 67% | 13% |
| SANSEBASTIAN | 53% | 62% | 9% |
| SANTANDERDEQUILICHAO | 83% | 75% | -8% |
| SANTAROSA | 62% | 63% | 1% |
| SILVIA | 69% | 70% | 1% |
| SOTARA | 79% | 71% | -8% |
| SUAREZ | 64% | 78% | 14% |
| SUCRE | 70% | 76% | 6% |
| TIMBIO | 63% | 84% | 21% |
| TIMBIQUI | 27% | 56% | 29% |
| TORIBIO | 61% | 66% | 5% |
| TOTORO | 47% | 59% | 12% |
| VILLARICA | 56% | 69% | 13% |

**Table J. 10.** Variation in RBO similarities per municipality (actual vs. training data – CGD).

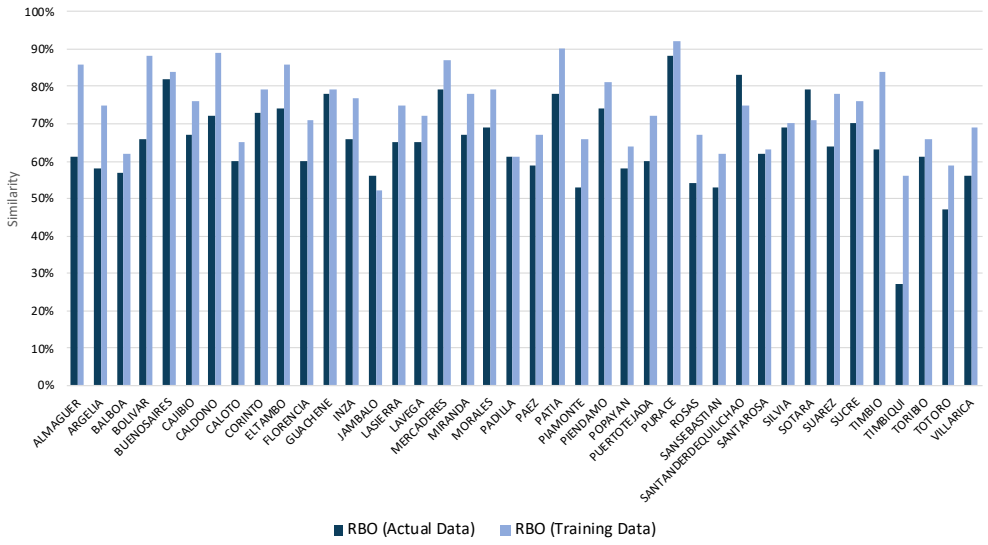



**Table J. 11.** Variation in RBO similarities per municipality (actual vs. training data – CGD).

# Appendix K

# MLC Model Training Time

This appendix presents the training times of different MLC strategies applied to the combined data sources. These tests can support decisions about the use of one model over another.

## K.1. Training Times for Models (CDS1)

| Strategy | Dataset Variations | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | mld_n | mld_d | mld_b1 | mld_b2 | mld_b3 | mld_s1 | mld_s2 | mld_s3 | mld_s4 | mld_s5 | mld_s6 |
| br | 45,80 | 52,63 | 14,56 | 5,52 | 3,42 | 31,86 | 23,66 | 13,89 | 11,99 | 7,73 | 3,52 |
| ecc | 1103,00 | 1121,78 | 175,75 | 42,83 | 21,30 | 563,02 | 342,11 | 146,95 | 128,65 | 65,90 | 22,93 |
| lp | 22,74 | 34,29 | 13,84 | 4,70 | 2,27 | 20,76 | 17,15 | 12,14 | 11,58 | 8,48 | 2,67 |
| homer | 91,77 | 94,96 | 28,90 | 12,03 | 3,65 | 72,87 | 50,57 | 29,02 | 27,25 | 13,66 | 4,30 |
| rakel | 144,57 | 163,51 | 50,88 | 18,08 | 2,48 | 113,96 | 81,49 | 45,36 | 39,96 | 26,61 | 2,97 |
| brplus | 239,65 | 249,16 | 39,79 | 11,37 | 5,16 | 131,79 | 74,47 | 34,01 | 28,93 | 15,73 | 5,74 |

**Table K. 1.** Training times for CDS1 variations.

## K.2. Training Times for Models (CDS2)

| Strategy | Dataset Variations | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mld_n | mld_b1 | mld_b2 | mld_b3 | mld_s1 | mld_s2 | mld_s3 | mld_s4 | mld_s5 | mld_s6 |
| br | 79.44 | 39.58 | 16.66 | 9.17 | 57.11 | 39.26 | 21.52 | 15.39 | 9.9 | 5.09 |
| ecc | 1139.43 | 393.12 | 121.97 | 53.33 | 636.4 | 394.88 | 171.44 | 106.9 | 61.64 | 30.91 |
| lp | 22.34 | 15.75 | 9.51 | 4.53 | 18.68 | 15.61 | 11.18 | 9.03 | 5.92 | 3.5 |
| homer | 120.26 | 66.65 | 27.31 | 12.46 | 86.35 | 66.73 | 37.09 | 22.9 | 14.85 | 5.95 |
| rakel | 231.55 | 116.92 | 50.24 | 21.54 | 151.65 | 114.04 | 65.19 | 44.63 | 26.86 | 3.72 |
| brplus | 300.48 | 103.77 | 35.51 | 16.53 | 153.49 | 102.83 | 49.78 | 31.58 | 18.89 | 10.94 |

**Table K. 2.** Training times for CDS2 variations.

## K.3. Training Times for Models (CDS3)

| Strategy | Dataset Variations | | | | | | |
|---|---|---|---|---|---|---|---|
| | mld_n | mld_b1 | mld_b2 | mld_b3 | mld_s1 | mld_s2 | mld_s3 |
| br | 51.54 | 27.74 | 11.44 | 4.26 | 39.63 | 37.88 | 29.38 |
| ecc | 1030.74 | 450.72 | 111.28 | 31.97 | 850.55 | 620.18 | 448.28 |
| lp | 13.79 | 12.62 | 10.74 | 2.89 | 16.23 | 14.69 | 12.67 |
| homer | 83.59 | 53.38 | 20.22 | 7.4 | 88.52 | 69.77 | 53.89 |
| rakel | 144.97 | 88.07 | 36.68 | 13.3 | 140.95 | 115.59 | 91.93 |
| brplus | 231.43 | 98.23 | 27.86 | 8.82 | 201.15 | 154.54 | 102.27 |

**Table K. 3.** Training times for CDS3 variations.

## K.4. Training Times for Models (CDS4)

| Strategy | Dataset Variations | | | | | | |
|---|---|---|---|---|---|---|---|
| | mld_n | mld_b1 | mld_b2 | mld_b3 | mld_s1 | mld_s2 | mld_s3 |
| br | 39.71 | 22.04 | 10.39 | 3.9 | 33.29 | 29.33 | 18.38 |
| ecc | 805.51 | 322.99 | 102.9 | 25.28 | 614.25 | 502.5 | 247.87 |
| lp | 26.18 | 23.68 | 17.13 | 5.59 | 24.81 | 24.81 | 22.49 |
| homer | 70.27 | 40.13 | 18.13 | 5.81 | 61.87 | 54.48 | 33.4 |
| rakel | 363.82 | 231.2 | 115.13 | 15.8 | 322.07 | 294.12 | 200.33 |
| brplus | 171.52 | 71.53 | 24.27 | 6.45 | 136.65 | 114.31 | 55.21 |

**Table K. 4.** Training times for CDS4 variations.