ANÁLISIS DEL RENDIMIENTO DEL CAFÉ BASADO EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

JUAN DAVID RINCÓN PATIÑO

Tesis de Maestría en Ingeniería Telemática

Director
Juan Carlos Corrales Muñoz
Doctor en Ciencias de la Computación

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de investigación e-@mbiente
Popayán, 2020

JUAN DAVID RINCÓN PATIÑO

ANÁLISIS DEL RENDIMIENTO DEL CAFÉ BASADO EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Tesis presentada a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la
Universidad del Cauca para la obtención del
Título de

Magíster en:
Ingeniería Telemática

Director:
Juan Carlos Corrales Muñoz
Doctor en Ciencias de la Computación

Popayán
2020

A Dios, pilar de mi vida

A mis padres y hermanos, mi fuente de inspiración

A mi director, por sus sabios consejos

A mis amigos, por su infinito apoyo

Agradecimiento especial:

Al proyecto InnovAcción Cauca y a Ecotecma S.A.S., financiadores de la beca de maestría en la que se enmarca la presente investigación

Resumen estructurado

Antecedentes: la agricultura cumple un papel fundamental en América Latina y el Caribe, impactando no solamente la producción de los países, sino también el empleo, las exportaciones y, en general, la vida de millones de personas. En Colombia, el cultivo de café ha construido un tejido social de extremo valor en las zonas rurales del país, siendo la principal fuente de ingresos para más de la cuarta parte de la población rural. No obstante, enfrenta diversos retos, como los crecientes costos de producción, la degradación del suelo, la escasez de agua y el cambio climático, que terminan incidiendo en las tasas de rendimiento del cultivo. En consecuencia, el desarrollo de nuevas herramientas tecnológicas, pueden traer consigo el aumento del bienestar de los caficultores, mejorando las condiciones socioeconómicas de las familias rurales, evitando costos innecesarios y mejorando la productividad de sus cultivos.

Objetivo: analizar el rendimiento del cultivo del café a partir de métodos de aprendizaje automático.

Métodos: se propone una herramienta de apoyo para que los caficultores mejoren las tasas de rendimiento de su cultivo. Mediante un algoritmo de aprendizaje automático se estima el rendimiento potencial del cultivo de café con seis meses de anticipación a la cosecha. Dicha estimación, además de ser presentada a los agricultores, es utilizada como insumo principal para un sistema de recomendaciones que sugiere buenas prácticas de manejo del cultivo encaminadas a mitigar condiciones climáticas adversas e incrementar el rendimiento del cultivo.

Resultados: un conjunto de datos que comprende registros históricos meteorológicos y de rendimiento del cultivo de café en Colombia. Un conjunto de modelos que permiten la estimación de la producción y rendimiento del cultivo de café en Colombia. Un sistema de recomendaciones basado en contenido para la sugerencia de actividades de manejo del cultivo de café. Finalmente, un prototipo orientado al usuario que condensa los productos anteriores.

Palabras clave: Aprendizaje automático, aprendizaje supervisado, regresión, café, rendimiento de cultivos.

Structured abstract

Background: Agriculture has played a fundamental role in Latin America and the Caribbean since it has impacted not only the production of many countries but also employment, exports and the lives of millions of people in general. In Colombia, coffee growing has built an extremely valuable social fabric in the rural areas of the country being the main source of incomes for more than a quarter of the rural population. However, it faces some challenges such as: the rising production costs, land degradation, water shortage and climate changes which influence the crop yield rates. Consequently, the development of new technological tools might bring about the increase in coffee growers' welfare, improving their socioeconomic conditions, avoiding unnecessary costs and making the crop productivity better.

Objective: To analyze the yield of coffee growing using automatic learning methods.

Methods: A support tool is proposed to help the coffee growers improve their crop yield rates. The potential yield of coffee growing is estimated by an automatic learning algorithm six months in advance of harvest. This estimate, in addition to being presented to farmers, is used as the main input for a recommender system suggesting good crop management practices aimed at mitigating adverse climatic conditions and increasing crop yield.

Results: A data set that comprises historical meteorological and yield records of coffee growing in Colombia. A set of models that allow the estimation of the production and yield of coffee growing in Colombia. A content-based recommendation system for suggestions for coffee crop management activities. Finally, a user-oriented prototype that joins the previous products.

Keywords: Machine learning, supervised learning, regression, coffee, crop yield.

Contenido

Lista de figuras

Lista de tablas

1 Introducción

1.1 Planteamiento del problema

La agricultura cumple un papel fundamental en la economía y en el tejido social de América Latina y el Caribe. En 2012 el sector contribuyó 5% del producto interno bruto (PIB) de la región. Además, existe un potencial importante para aumentos futuros en la productividad de pequeños y medianos agricultores quienes aún pueden obtener mejoras significativas en el sistema de producción [1]. En Colombia, tradicionalmente el sector agrícola ha tenido una gran importancia para la economía, teniendo en cuenta su contribución al PIB, al empleo y a las exportaciones. El peso de la agricultura primaria en el PIB fue del 5,2% en 2013, y en lo que se refiere al empleo, su participación fue del 17,5%. En la actualidad, los productos agrícolas representan aproximadamente el 11% del total de las exportaciones de Colombia, entre las cuales han predominado productos tradicionales como el plátano, el azúcar y el café [2], este último de particular importancia para el país. Es por esto que la inversión continua en investigación, desarrollo y servicios de extensión es fundamental para lograr las ganancias de productividad tan necesarias en el país [3].

En cuanto a la investigación y desarrollo en el sector agrícola, se ha enfocado principalmente en la producción de los cultivos, que se basa en tres fuentes principales de crecimiento: aumento de la tierra cultivada, incremento de la frecuencia de las cosechas (a menudo mediante regadío) y aumento del rendimiento. En particular, el crecimiento del rendimiento es el factor subyacente que permitirá los incrementos en la producción de cultivos en el futuro, representando aproximadamente un 70% hasta el año 2030 en los países en desarrollo [4].

Por su parte, la producción del sector agrícola enfrenta diferentes retos: los crecientes costos de producción, como el aumento de los precios de insumos y mano de obra, restricciones de recursos como la degradación del suelo, la escasez de agua y el aumento de las presiones ambientales [3]. De igual manera, el cambio climático tiene efectos diversos sobre la actividad agrícola, algunas zonas cultivadas se hacen inadecuadas para el cultivo, mientras otras pueden hacerse cada vez más áridas. Con el incremento de la temperatura, aumenta también la gama de insectos dañinos para la agricultura, además de la capacidad de supervivencia de las plagas durante el invierno [4]. En cuanto a los cambios físicos relevantes a la actividad agrícola que se visualizan bajo los escenarios climáticos más comunes figuran: aumento en la temperatura atmosférica y del suelo, alteraciones en las concentraciones de CO2 en la atmósfera, alza del nivel del mar, cambios en el ciclo hidrológico, así como en la calidad del agua y su disponibilidad, intensificación y aumento de eventos climáticos extremos (entre ellos sequías e inundaciones), y modificaciones en el nivel altitudinal de los puntos de rocío, entre otros [1]. Por otro lado, las plagas de insectos

son otra grave amenaza para la productividad, pueden devastar el rendimiento de los cultivos y transmitir enfermedades tanto a estos como al ganado. Además, se manifiesta la preocupación de que la dependencia de los plaguicidas para mantener el rendimiento no sólo tiene repercusiones negativas en el medio ambiente, sino que también propicia la resistencia de los insectos a los propios plaguicidas [3].

La información anterior evidencia la necesidad de esfuerzo continuo en investigación y desarrollo de nuevas tecnologías o en la mejora de herramientas ya existentes que permitan la estabilidad o crecimiento de las tasas de producción agrícola. Dichos esfuerzos deben estar enfocados hacia la reducción de costos de producción, incremento del rendimiento del cultivo, eficiencia en la aplicación de nutrientes y la gestión del riesgo a pérdidas debido a afectaciones por enfermedades, plagas o cambios climáticos extremos. En este orden de ideas, la adopción de las tecnologías de la información y las comunicaciones (TIC) en las áreas rurales promete traer grandes beneficios. En la agricultura, las ciencias de la computación (un área específica de las TIC) tienen el potencial de incrementar el acceso de los agricultores a información pública y privada, mejorar el alcance a servicios financieros, evitar el uso innecesario de algunos insumos e incrementar el rendimiento de los cultivos, entre otros aspectos [5]. El aprendizaje automático (ML, *machine learning*) es uno de los campos de las ciencias de la computación que actualmente aporta mejoras importantes a la agricultura. Algunas de las soluciones llevadas a cabo en este campo facilitan la recolección, categorización y entendimiento de información de diferentes fuentes (sensores, satélites, repositorios remotos, etc.), para generar modelos que sirvan como herramientas de apoyo para los agricultores [6].

Teniendo en cuenta las consideraciones mencionadas anteriormente, con el fin de aprovechar las tecnologías de la información y las comunicaciones para impulsar la producción del sector agrícola, en la presente propuesta se plantea la siguiente pregunta de investigación:

¿Cómo apoyar la estabilidad de las tasas de rendimiento del cultivo de café a través del análisis de datos?

1.2 Escenario de motivación

En Colombia, el café sigue siendo el cultivo que mayor peso tiene dentro de la canasta de productos agrícolas, manteniendo un valor sobre la producción nacional superior al 15% en el anterior quinquenio, a pesar de la disminución circunstancial de los últimos años, explicada por la caída en la producción del grano por cuenta de las alteraciones climáticas asociadas a eventos extremos. Así mismo, alrededor del cultivo se ha construido un tejido social de extremo valor en las zonas rurales del país, siendo la principal fuente de ingresos para más de medio millón de productores, cuyas familias están compuestas por el 25% de toda la población rural colombiana, cerca de tres millones de personas. De la misma forma,

el café aporta la tercera parte del empleo rural del país y es la actividad que más contribuye con la redistribución del ingreso en el campo. En 2012 el valor de la cosecha llegó a $3,4 billones, distribuidos en el 60% de los departamentos del país, un flujo importante de recursos que aportan a la economía regional. De ahí la importancia de esta actividad como eje para la disminución de la pobreza y potencial generador de condiciones de paz en Colombia [7].

En cuanto a las personas involucradas en la producción agrícola, los pequeños productores juegan un papel fundamental. En América Latina y el Caribe, el 81% de la actividad agrícola es llevada a cabo por pequeños agricultores, agrupando a una población de alrededor de 60 millones de personas, creando entre el 57% y el 77% del empleo agrícola en la región y generando, a nivel país, hasta el 67% del total de la producción alimentaria. Los pequeños agricultores no sólo producen la mayor parte de los alimentos que consumen los países de la región, sino que también desarrollan actividades agrícolas diversificadas, con las que juegan un papel fundamental a la hora de aportar a la sostenibilidad del medio ambiente y la conservación de la biodiversidad [8]. Sin embargo, en la actualidad, los pequeños productores enfrentan diversos retos asociados a la sostenibilidad de su actividad y el diseño de estrategias que permitan enfrentar los riesgos asociados al cambio climático. La producción agrícola por su naturaleza está expuesta a condiciones cambiantes de clima, que no siempre son favorables a la producción y la productividad. Ejemplo de ello son los fenómenos climáticos extremos como el fenómeno de La Niña, donde se desarrollan condiciones atmosféricas adversas para el crecimiento y desarrollo del cultivo, y a la vez favorables para la incidencia de enfermedades [7], [9].

Teniendo en cuenta los retos que enfrentan los pequeños productores, específicamente los pequeños caficultores, en el país están dispuestos cerca de 1500 técnicos del servicio de extensión que buscan atender sus necesidades. Aún así, dicho número de profesionales puede no ser suficiente para apoyar a los caficultores, más de 563mil personas, en la búsqueda de la rentabilidad de sus cultivos. Por consiguiente, ofrecer a los productores, en especial a los pequeños, diferentes herramientas para acceder a información acerca de buenas prácticas de manejo del cultivo, de una forma más fácil y rápida, puede ser de gran ayuda para mejorar las tasas de rendimiento de sus cultivos. En consecuencia, se pueden obtener importantes logros en el aumento del bienestar de los pequeños propietarios, mejorando las condiciones socioeconómicas de las familias rurales, evitando costos innecesarios y mejorando la productividad de sus cultivos [7], [10].

En este orden de ideas, en la presente investigación, se analiza el rendimiento del cultivo de café, para determinar su comportamiento de acuerdo a la variabilidad climática del país y con ello generar una herramienta de apoyo a los pequeños caficultores, que les permita conocer buenas prácticas de manejo encaminadas al incremento de las tasas de rendimiento de sus cultivos.

1.3 Objetivos

A continuación, se expone el objetivo general de la presente investigación, así como los objetivos específicos mediante los que se aborda la solución de la problemática formulada en la sección 1.1.

1.3.1 Objetivo general

Analizar el rendimiento del cultivo del café a partir de métodos de aprendizaje automático.

1.3.2 Objetivos específicos

- Construir un conjunto de datos que describa el rendimiento del cultivo de café.
- Implementar un modelo para el análisis predictivo del rendimiento del cultivo de café a partir de los factores influyentes.
- Sugerir diferentes actividades de manejo del cultivo de café orientadas a mantener o mejorar el rendimiento.
- Evaluar un prototipo orientado al usuario para mantener o mejorar el rendimiento del cultivo de café.

1.4 Partes de la memoria

La presente monografía se encuentra dividida en los siguientes seis capítulos que condensan la investigación realizada:

- ***Capítulo 1***: presenta la introducción, el planteamiento del problema de investigación y la estructura general del trabajo realizado.
- ***Capítulo 2***: denominado "Estado actual del conocimiento", hace referencia a las tecnologías y conceptos en los que se fundamenta la presente investigación, además de las experiencias previas llevadas a cabo en otras investigaciones de aprendizaje automático y agricultura que se relacionan con la expuesta en el presente trabajo.
- ***Capítulo 3***: denominado "Conjunto de datos sobre el rendimiento del café", presenta el proceso de construcción de un conjunto de datos meteorológicos y del rendimiento del cultivo de café en Colombia. El proceso descrito en el capítulo comprende la adquisición, selección, estructuración, verificación, limpieza e integración de los datos.
- ***Capítulo 4***: denominado "Modelo de regresión para la estimación del rendimiento del café", expone el uso de los datos presentados en el capítulo 3 para llevar a cabo

un proceso de entrenamiento utilizando distintas técnicas de aprendizaje automático para la generación de un modelo que permite la estimación de la producción y rendimiento del cultivo de café en Colombia.

- ***Capítulo 5***: denominado "Recomendaciones para el manejo del cultivo de café", presenta la construcción de un sistema de recomendaciones basado en contenido que permite sugerir diferentes actividades para el manejo del cultivo de café a partir de las condiciones climáticas del municipio.
- ***Capítulo 6***: denominado "Prototipo y experimentación", muestra la construcción y evaluación de un prototipo orientado al usuario que permite estimar el rendimiento del cultivo de café en Colombia y con ello observar diferentes actividades de manejo del cultivo para mejorar o mantener el registro esperado.
- ***Capítulo 7***: presenta la síntesis de los resultados de la presente investigación, así como las principales contribuciones y elementos a tener en cuenta en el desarrollo de trabajos futuros.

2 Estado actual del conocimiento

En este capítulo se expone la generación de la base conceptual siguiendo las fases propuestas por el "Modelo para la investigación documental" [11]. En este orden de ideas, se recopilan los conceptos y tecnologías en que se fundamenta la presente investigación. De igual manera, se describen investigaciones recientes y/o representativas que se han desarrollado entorno a la aplicación de algoritmos de aprendizaje automático en el campo de la agricultura.

El capítulo se encuentra dividido en los siguientes apartados:

- **Conceptos y definiciones fundamentales**: se definen las bases teóricas de la investigación llevada a cabo, exponiendo en detalle los conceptos que toman relevancia en los resultados obtenidos.
- **Trabajos relacionados**: presenta la exploración del estado actual del conocimiento llevada a cabo con el objetivo de determinar las investigaciones relacionadas con la presente, además de las futuras direcciones en el campo de la agricultura y el aprendizaje automático.

Adicionalmente, se llevó a cabo un análisis bibliométrico sobre la producción científica alrededor del aprendizaje automático, con el fin de obtener una visión general de los desarrollos generados durante la última década en esta área y mostrar tendencias que podrían ser la base para la presente investigación y futuros desarrollos en este campo. Dicho análisis puede observarse en el anexo A.

2.1 Conceptos y definiciones fundamentales

A continuación, se exponen los conceptos y definiciones que sirven como base para la consecución de los resultados de la presente investigación. Entre los conceptos mencionados se encuentran el rendimiento de un cultivo, aprendizaje automático, aprendizaje supervisado y algunas de las técnicas más utilizadas en dicho campo (árboles de decisión, redes bayesianas, redes neuronales artificiales y máquinas de vectores de soporte).

2.1.1 Rendimiento de un cultivo

El rendimiento de los cultivos es esencial en la agricultura, al igual que el aumento de este a través de nuevas tecnologías para la seguridad alimentaria mundial. La cifra más importante de rendimiento en la agricultura es el rendimiento promedio en una finca, municipio, departamento o todo el país. Esta cifra comúnmente es registrada en kilogramos o toneladas métricas por hectárea (kg o t/ha) y se reporta a partir de las

mediciones de rendimiento de los agricultores en diferentes encuestas y estadísticas locales o nacionales [12]. Por su parte, el rendimiento de un cultivo específico se refiere a la masa recolectada del producto al final de la cosecha y está estrechamente relacionado con el medio ambiente, involucrando la ubicación, estación del año, suministro de agua y variables climáticas como la radiación solar, la humedad y la temperatura. Comúnmente, se pueden mencionar diferentes conceptos relacionados con el rendimiento de un cultivo. Por una parte, se destaca el "rendimiento potencial", que se refiere al máximo rendimiento que puede ser alcanzado por un cultivo bajo unas condiciones preestablecidas, como el que se determina mediante modelos de simulación con supuestos fisiológicos y agronómicos [13]. Por otra parte, la "brecha de rendimiento" también es un concepto relevante, que se estima mediante la diferencia entre el "rendimiento potencial" y el "rendimiento efectivo" logrado por los agricultores en una escala espacial y temporal específica. Este último rendimiento, normalmente es inevitablemente menor que el rendimiento potencial, ya que se ve afectado por factores diversos del suelo, prácticas agrícolas variadas y limitaciones abióticas y de recursos naturales no renovables [14].

2.1.2 Aprendizaje automático

Este campo de las ciencias de la computación estudia la creación de programas computarizados que aprenden automáticamente de experiencias realizadas. Desde un punto de vista generalizado, un algoritmo de aprendizaje automático puede sacar provecho de datos para generar conocimiento a partir de la distribución de probabilidad de estos, debido a que los datos pueden ser vistos como ejemplos de las relaciones que existen entre diferentes variables de interés. Uno de los enfoques principales en este campo, es el de aprender automáticamente a reconocer patrones en los datos y tomar decisiones a partir de estos. El aprendizaje automático está estrechamente relacionado con otros campos de las ciencias de la computación, como lo son la inteligencia artificial, minería de datos y la teoría de probabilidad. Además, incluye el estudio de diferentes algoritmos, entre los que se encuentran las redes neuronales, máquinas de vectores de soporte, árboles de decisión, entre otros. Existen muchas aplicaciones exitosas en el campo del aprendizaje automático que analizan datos para generar herramientas que van desde optimizar el comportamiento de un robot o mejorar la producción de un cultivo, hasta predecir el comportamiento de las personas o reconocer rostros, entre otras [15].

2.1.3 Aprendizaje supervisado

En el campo del aprendizaje automático es posible distinguir dos categorías fundamentales: aprendizaje supervisado y aprendizaje no supervisado. El objetivo del aprendizaje supervisado es la búsqueda de algoritmos que razonen a partir de observaciones provistas externamente (llamadas instancias de entrenamiento), que contienen una variable dependiente ejemplo del resultado, comúnmente llamada clase y

variables independientes, normalmente conocidas como atributos, para generar un modelo que luego hace predicciones sobre instancias futuras [16]. Las instancias de entrenamiento pueden tomar diferentes valores, ya sean discretos o continuos, es por esto que se ha creado una convención de nomenclatura para la tarea de predicción: regresión, cuando se predicen valores continuos (o cuantitativos) y clasificación, cuando se predicen valores discretos (o cualitativos) [17]. Existen una gran variedad de técnicas de aprendizaje supervisado, entre las que se destacan las redes neuronales artificiales, regresión lineal, máquinas de vectores de soporte, árboles de decisiones, entre otras [18].

2.1.4 Árboles de decisión

Los árboles de decisión son una forma de dividir recursivamente datos en estructuras secuenciales y jerárquicas para representar reglas básicas sobre estos. Esta técnica de aprendizaje automático se utiliza para aproximar funciones objetivo de valores discretos, en las que la función aprendida se representa por un árbol de decisión o un conjunto de reglas de la forma "sí-luego". En particular, un árbol de decisión para un problema de clasificación se puede representar en forma de una estructura de árbol, donde cada uno de sus nodos puede ser una decisión final (hoja) o un nodo de decisión. Un nodo de hoja indica el valor de la clase (atributo de destino), mientras que un nodo de decisión especifica alguna prueba que se realizará en una única característica de la observación disponible, con una rama para cada resultado posible. El proceso de clasificación de una instancia dada por medio de un árbol de decisión comienza con la evaluación de la prueba contenida en el nodo raíz (el primer nodo de decisión) y se mueve a través del árbol hasta un nodo hoja, que proporciona la clasificación de la instancia [16].

2.1.5 Redes bayesianas

Una red bayesiana es un modelo gráfico probabilístico para representar un conjunto de variables aleatorias y sus relaciones condicionales a través del uso de grafos acíclicos dirigidos. En dicha red, los nodos representan variables, que pueden ser los componentes de un sistema, mientras que las aristas hacen referencia a las dependencias condicionales entre las variables. A cada nodo se le asigna una distribución de probabilidad en función de los estados de las variables que hacen parte del nodo. A través de esta técnica, se puede estimar la probabilidad a posteriori de las variables no conocidas, con base en las variables conocidas. La flexibilidad de su estructura y su motor de razonamiento probabilístico, hacen que las redes bayesianas sean un método importante para el modelamiento de largos y complejos sistemas [19].

2.1.6 Redes neuronales artificiales

Las redes neuronales artificiales se desarrollaron inicialmente sobre la base de que los sistemas biológicos de aprendizaje se construyen a partir de complejas redes de neuronas interconectadas. Existe una gran variedad de redes neuronales artificiales en la literatura, pero todas comparten la misma estructura principal, aquella en la que la red se compone de unidades (también llamadas neuronas) y conexiones entre ellas, que juntas determinan el comportamiento de la red. La elección del tipo de red depende del problema a resolver, pero las más populares son las redes basadas en propagación hacia atrás. Dicha red está conformada por tres o más capas de neuronas: una capa de entrada, una de salida y al menos una oculta. Todas las neuronas de cada capa (excepto las que pertenecen a la de salida) están conectadas por un axón a cada neurona de la siguiente capa. El entrenamiento de las redes neuronales artificiales se realiza generalmente mediante una actualización iterativa de los pesos en cada neurona en función de una señal de error. Este error se calcula en la capa de salida como la diferencia entre la clase real y los valores de salida actuales, multiplicada por la pendiente de una función de activación. Luego, la señal de error se propaga hacia atrás a las capas inferiores. Siguiendo este proceso, la propagación hacia atrás intenta minimizar el error en cada iteración. Los pesos de la red se ajustan mediante el algoritmo de aprendizaje, de manera que el error va disminuyendo. Tradicionalmente dos parámetros, llamados tasa de aprendizaje y factor de momento, se utilizan para controlar el ajuste del peso a lo largo de la propagación hacia atrás [16].

2.1.7 Máquinas de vectores de soporte

Son un conjunto de métodos de aprendizaje supervisado que se utilizan para problemas tanto de clasificación, como de regresión. En el caso de una tarea de clasificación, dado un conjunto de puntos (datos) en un espacio n-dimensional, cada uno perteneciente a una de las posibles clases, las máquinas de vectores de soporte (SVM, *support vector machines*) tienen como objetivo encontrar los hiperplanos separadores que maximizan el margen entre conjuntos de datos. Para calcular el margen entre los datos que pertenecen a dos clases diferentes (una tarea de clasificación binaria), se construyen dos hiperplanos paralelos (uno a cada lado del hiperplano de separación) y se "presionan" contra los dos conjuntos de datos lo máximo posible; cuanto mayor sea el margen, menor será el error de generalización del clasificador [16]. Para el caso de una tarea de regresión, se basa en el cálculo de una función de regresión lineal en un espacio de características de alta dimensión donde los datos de entrada se asignan a través de una función no lineal [20].

2.2 Trabajos relacionados

Alrededor del rendimiento en los cultivos agrícolas existen diversos trabajos que se apoyan en técnicas matemáticas y estadísticas para conocer la producción que se espera registrar

en un futuro dadas ciertas condiciones, y con ello tomar decisiones con antelación, que permitan controlar, incrementar u optimizar la cantidad producida. Con el objetivo de conocer algunas de las aproximaciones existentes que se relacionan con el problema de investigación presentado, se realizó un mapeo sistemático en diferentes bases de datos de revistas y conferencias académicas en línea, entre las que fueron seleccionadas: ScienceDirect, IEEEXplore, SpringerLink y ACM Digital Library. En esta sección se presentan algunos de los trabajos encontrados mediante el mapeo sistemático y que se relacionan con la presente investigación. Dichas aproximaciones están orientadas, por una parte, hacia el análisis del rendimiento de un cultivo agrícola en general y el uso de aprendizaje automático para predecirlo y, por otra parte, hacia la aplicación de técnicas de aprendizaje supervisado y no supervisado para estimar el rendimiento del cultivo de café como tal.

2.2.1 Modelos matemáticos y estadísticos para determinar el rendimiento de cultivos

Existe gran variedad de investigaciones dedicadas a la estimación o predicción del rendimiento de un cultivo. En un primer trabajo, presentando en [21], los investigadores estimaron el rendimiento del cultivo de maíz en Estados Unidos mediante el cálculo del parámetro de profundidad óptica de vegetación (VOD, *vegetation optical depth*), que se basa en las condiciones hídricas del cultivo y la biomasa disponible en el suelo. Los resultados presentados en la investigación demuestran que la región en la que se encuentra el cultivo es determinante para la elección de las métricas utilizadas en el cálculo del rendimiento. En el mismo sentido, en [22] propusieron dos modelos empíricos para el cálculo del rendimiento del cultivo de maíz en México; uno, con una precisión del 86%, basado en el índice de área foliar (LAF, *leaf area index*) y otro, con una precisión del 97%, basado en el índice de vegetación de diferencia normalizada (NDVI, *normalized difference vegetation index*). En el caso de [23], implementaron un modelo (CERES) para estimar el rendimiento del mismo cultivo en Pakistán. La calibración del modelo fue realizada mediante el uso de imágenes satelitales Landsat-8, series temporales de NDVI y datos de la temperatura de la superficie del terreno (LST, *land surface temperature*). En la investigación se destaca que, para predecir rendimientos regionales, el uso de imágenes satelitales es una herramienta útil, ya que permite incrementar el número de datos sobre el terreno para calibrar el modelo final.

Así mismo, en la investigación presentada en [24], se expone un sistema para la predicción del rendimiento del cultivo de manzanas en Corea del Sur construido utilizando una técnica estadística llamada Kernel Smoothing y datos climáticos mensuales, entre los que se encuentran la temperatura máxima, mínima y promedio, cantidad de lluvia y horas de sol. En un segundo trabajo, expuesto en [25], evaluaron tres métodos estadísticos (regresión, media móvil y alisado exponencial) para la predicción de la producción de café

en Filipinas. Para la comparación fueron utilizados datos de la producción trimestral de café durante cinco años en cinco provincias del país mencionado. Los resultados arrojaron errores de precisión entre el 9% y el 14%. Por su parte, en [26], se presentan cuatro modelos de regresión lineal múltiple para simular el rendimiento de trigo, maíz, canola y girasol en Hungría. Se utilizaron datos meteorológicos (temperatura y precipitación) y de cantidad de fertilizantes, así como registros del contenido de agua en el suelo e índices de vegetación basados en teledetección para el pronóstico del rendimiento. Los modelos construidos en la investigación predicen el rendimiento de los cultivos con una precisión del 67% para el trigo, 76% para la canola, 81% para el maíz y 68.5% para el girasol. De forma similar, en [27], desarrollaron un método para estimar la producción de cultivos de algodón, maíz, mijo, maní y sorgo a nivel de finca utilizando series de tiempo de alta resolución Sentinel-2 y datos terrestres en Mali. El modelo construido se basa en un enfoque Monte Carlo que combina índices espectrales y de área foliar. Los resultados muestran que la producción de los cultivos se puede explicar a partir de datos satelitales Sentinel-2 con una precisión general del 80%. Además, demuestran que la incorporación de la ubicación de las parcelas incrementa la precisión del modelo en un 5%.

Por otra parte, en la búsqueda llevada a cabo aparece el trigo como un cultivo ampliamente estudiado y para el que se han desarrollado varios modelos de predicción del rendimiento. En un primer trabajo, presentado en [28], realizaron un análisis del rendimiento de dicho cultivo tomando como base la radiación, precipitación y biomasa. Para esto, se estudian diferentes modelos matemáticos que permiten predecir el rendimiento del cultivo y a partir de datos existentes, se evalúan para conocer cuál tiene el mejor comportamiento en el campo de interés de dicha investigación. Como resultado del trabajo, se propone un modelo matemático que permite predecir el rendimiento del cultivo de trigo a partir de unas pocas variables y que puede ser insumo para la toma de decisiones de los agricultores. Por su parte, en [29], utilizaron el software AGROMETSHELL desarrollado por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, por sus siglas en inglés) para calcular los parámetros del índice de satisfacción de agua (WSI, *water satisfaction index*). Dichos resultados fueron correlacionados con estadísticas históricas (temperatura, humedad, precipitación, velocidad del viento, radiación y horas de sol) para producir pronósticos de rendimiento del cultivo de trigo en Turquía antes del tiempo de cosecha. Una de las conclusiones relevantes de la investigación hace énfasis en la imposibilidad de crear un único modelo que explique el rendimiento de un cultivo en general, por el contrario, existe la necesidad de desarrollar diferentes métodos de pronóstico del rendimiento en las regiones que presentan diversas condiciones agroecológicas. De manera similar, en [30], los investigadores combinaron imágenes satelitales con dos modelos matemáticos, uno de asimilación de carbono y otro de acumulación y rotación del mismo elemento químico, para predecir el rendimiento de canola y trigo a una escala de campo en Australia. Como

resultado del trabajo realizado se obtuvo un modelo basado en imágenes satélites llamado C-Crop para predecir el rendimiento con errores del orden del 32%.

Continuando con el trigo, en [31], compararon dos métodos para la predicción de su rendimiento en Uruguay a nivel país o de grandes áreas agrícolas. El primero, que presentó mejores resultados, es un método de regresión simple entre diferentes índices de vegetación. El segundo, es un modelo de cultivo basado en la optimización de dos parámetros (nitrógeno foliar y biomasa aérea inicial) utilizando series temporales. Los índices de vegetación utilizados en los dos modelos fueron obtenidos a partir de imágenes de alta resolución de la superficie terrestre capturadas por satélites Landsat. Entre los resultados de la investigación se resalta la estrecha relación entre el índice de vegetación de diferencia normalizada y el rendimiento de un cultivo. Un segundo trabajo, expuesto en [32], evalúa un modelo biofísico basado en parámetros de evapotranspiración real y contenido de agua en el suelo de la zona radicular, llamado Cultivo RS-Met, que permite generar predicciones tempranas de rendimiento de trigo bajo condiciones áridas y semiáridas en Israel. Los hallazgos de la investigación sugieren que el contenido y el suministro de agua de un cultivo están directamente relacionados con el rendimiento del mismo. En un tercer trabajo, presentado en [33], se utilizaron imágenes satelitales y de sensores de campo para la construcción de un modelo matemático que permite estimar el rendimiento del mismo cultivo en India. Las imágenes mencionadas fueron la base para el cálculo del NDVI y el índice de vegetación ajustada al suelo (SAVI, *soil adjusted vegetation index*). Dichos índices fueron correlacionados con datos reales de rendimiento del cultivo de trigo para construir el modelo de predicción. Con un propósito similar, en [34], índices de vegetación extraídos de imágenes Landsat-5 fueron utilizados para la construcción de varios modelos matemáticos para el cálculo del rendimiento de trigo en Beijing. Entre los índices empleados se encuentran: NDVI, índice de relación simple (SR, *simple ratio index*), índice de relación de absorción de clorofila (TCARI, *transformed chlorophyll absorption ratio index*), SAVI, índice de agua de diferencia normalizada (NDWI, *normalized difference water index*). Los modelos construidos fueron luego integrados utilizando el método de combinación de optimización de peso (WOC, *weight optimization combination*) para mejorar la precisión de la estimación del rendimiento.

Por último, en [35], se presenta un modelo matemático, llamado QUEFTS (*QUantitativeEvaluation of the Fertility of Tropical Soils*), que permite evaluar cuantitativamente la fertilidad de un suelo tropical. Dicha evaluación consiste de cuatro pasos que, en conjunto, permiten calcular los suministros potenciales de fósforo, nitrógeno y potasio, conocer la absorción real de cada nutriente y establecer un posible rendimiento del cultivo estudiado. El modelo generado, además de haber sido evaluado en un cultivo de maíz, puede ser aplicado a una variedad de cultivos que cumplan con unas condiciones listadas en la publicación, entre los que el café es un importante candidato. En [36], el modelo QUEFTS es mejorado. Para esto, se incluyen nuevas variables en el cálculo del rendimiento del cultivo, entre las que se encuentran la temperatura ambiente y el

contenido de arcilla del suelo. Con esto, la evaluación que se realiza del suelo en el que se tiene un cultivo puede ser aún más precisa. Además, en el mismo trabajo se presentan los pasos para calibrar el modelo, y así poder aplicarlo a un cultivo diferente al del maíz. Dada la utilidad para evaluar el posible rendimiento de un cultivo en un suelo tropical del modelo matemático mencionado, algunas investigaciones lo han tomado como base para generar diversas aproximaciones. Ese es el caso del trabajo presentando en [37], en el que investigadores desarrollan un modelo cuantitativo para estimar el rendimiento del café y la cantidad de nutrientes necesarios para un óptimo desarrollo de dicho cultivo. Para generar el modelo, tomaron dos insumos importantes: por una parte, datos de la producción anual de café en diversos distritos del norte de Tanzania y, por otra parte, el modelo QUEFTS, que al haber sido aplicado en un cultivo que crece bajo condiciones similares a las del café fue posible replicarlo en el cafeto arábico, una de las numerosas variedades de café existentes en el mundo. Como resultado de la investigación mencionada, obtuvieron una evolución del modelo QUEFTS que puede ser aplicado al café y que tiene en cuenta nuevas variables, tales como parámetros de manejo del cultivo y la densidad de plantas.

2.2.2 Aprendizaje automático para el análisis del rendimiento de cultivos

Lograr el máximo rendimiento de los cultivos (registrando los menores costos posibles) es uno de los principales objetivos de la producción agrícola. La detección y el manejo temprano de los problemas asociados al rendimiento de los cultivos pueden ayudar a incrementar la producción y las ganancias subsiguientes. Para esto, el aprendizaje automático ha sido una herramienta exitosa. A continuación, se presentan diferentes trabajos de investigación realizados en dicha área y que tienen como dominio de aplicación el rendimiento de cultivos.

En un primer trabajo presentado en [38], llevaron a cabo una revisión de diferentes aplicaciones basadas en aprendizaje automático y construidas en los últimos 15 años para la predicción del rendimiento agrícola. Los resultados de la investigación destacan el aporte de las técnicas de aprendizaje automático para construir soluciones rentables e integrales para una mejor estimación del rendimiento de los cultivos y la toma de decisiones. Todo esto debido a que pueden resolver de forma autónoma grandes problemas no lineales utilizando conjuntos de datos de múltiples fuentes (potencialmente interconectadas). Las conclusiones del trabajo destacan k-vecinos más cercanos, redes neuronales, regresión de vectores de soporte y árboles de regresión (en especial el algoritmo *M5*), como las técnicas de aprendizaje automático más exitosas para la estimación del rendimiento de cultivos. Con el mismo objetivo, en [39] se estudiaron varias técnicas de aprendizaje automático para la predicción del rendimiento de diferentes cultivos, entre los que se encuentran arroz, soya, trigo, caña de azúcar, algodón, maíz y papa. Algunas de las técnicas evaluadas fueron: regresión lineal múltiple, árboles de decisiones, redes neuronales, k-vecinos más

cercanos, entre otras más. Para la construcción de los modelos llevaron a cabo un proceso de minería de datos complejo, que comprende limpieza, integración, selección y transformación de datos, entre otros pasos. Gran parte de las técnicas utilizadas registraron una alta precisión, siendo regresión lineal múltiple y árboles de decisiones (para el caso del arroz), redes neuronales (en maíz) y regresión lineal (para el caso del trigo y la papa), las que obtuvieron los mejores resultados, teniendo una precisión entre el 95% y el 100%. Además de las mencionadas, los resultados de la investigación también destacan la regresión de vectores de soporte (*support vector regression*) como una técnica de aprendizaje automático que permite predecir el rendimiento de cualquier cultivo con una alta precisión. En un tercer trabajo, presentado en [40], evaluaron el comportamiento de varias técnicas de regresión al intentar predecir el rendimiento de diferentes cultivos agrícolas. Entre las variables que fueron insumo para el entrenamiento de los algoritmos, se encuentran la cantidad de fertilizante aplicado al cultivo, el índice de vegetación, la conductividad del suelo y el rendimiento. Como resultado de la investigación se tiene que las máquinas de vectores de soporte tienen los mejores resultados al intentar predecir el rendimiento de algunos cultivos.

Así mismo, en [41], utilizaron varias técnicas de aprendizaje automático, entre las que se encuentran *boosted regression trees*, *random forest*, *support vector regression* y *gaussian process regression*, para la predicción del rendimiento del maíz en Irán. Como datos para el entrenamiento de los modelos fueron utilizados registros de rendimiento de tres años y series temporales del NDVI obtenido a través de imágenes satelitales Landsat. Los resultados arrojaron que los árboles de regresión tienen un buen desempeño a la hora de estimar el rendimiento de un cultivo, consiguiendo una correlación de 0.87 para el caso específico de la investigación. Técnicas similares fueron utilizadas en [42], donde compararon dos algoritmos, *boosted regression trees* y *support vector machines*, para predecir el rendimiento de trigo en la provincia de Henán en China. Los modelos construidos tomaron como base tres tipos de predictores: NDVI único, incremental y dirigido. Las conclusiones de la investigación destacan que, para el problema específico abordado, la mejor técnica (de las dos evaluadas) es *boosted regression trees*, pero que los algoritmos de aprendizaje automático en general son una importante herramienta para la predicción del rendimiento agrícola. Otra técnica interesante para la estimación del rendimiento de un cultivo es el aprendizaje profundo; en [43], se describe el uso de dicha técnica y una regresión de vectores de soporte para construir un modelo que permita predecir el rendimiento del maíz en Illinois. La estimación del rendimiento se hace a nivel de condado y toma como datos base el índice de vegetación mejorado (*enhanced vegetation index*) y registros climáticos (temperatura, evaporación potencial y presión). Los mejores resultados de la investigación se registraron con el algoritmo de aprendizaje profundo, teniendo un coeficiente de correlación de 0.81, que lo destaca como una técnica interesante para la estimación del rendimiento agrícola.

Por otra parte, un cultivo para el que se han construido varios modelos de predicción del rendimiento es el de maíz. En [44], compararon el rendimiento de varios algoritmos de aprendizaje automático en la construcción de un modelo que permita predecir el rendimiento de dicho cultivo en Ohio, Estados Unidos. El entrenamiento fue realizado basado en propiedades del suelo (materia orgánica, capacidad de intercambio catiónico, magnesio, potasio y pH) e imágenes aéreas multiespectrales y utilizando diversos algoritmos: *linear regression* (que tuvo la precisión más baja)*, random forest, neural network, support vector machine, gradient boosting model* y *cubist*. El estudio sugiere que integrar datos de sensores remotos y algoritmos de aprendizaje automático puede ser útil para conocer el posible rendimiento e implementar prácticas agrícolas específicas para un sitio determinado. También en [45], estimaron el rendimiento del mismo cultivo en Iowa, Estados Unidos, a través del uso de datos climáticos (humedad, temperatura y precipitación) y tres algoritmos de aprendizaje automático basados en regresión (*multivariate polynomial regression*, *support vector machine regression* y *random forest*). El análisis llevado a cabo en la investigación sugiere que las máquinas vectores de soporte permiten crear modelos con una alta precisión en la estimación del rendimiento agrícola. Dicho algoritmo registró un coeficiente de determinación de 0.968 en el trabajo mencionado. De manera similar, en [46] calcularon el rendimiento del cultivo de maíz, y otros más, en India. Para esto, se utilizaron datos climáticos y características del cultivo, entre las que se encuentran el área cultivada, tipo de suelo, pH, presencia de enfermedades, temperatura y precipitación, entre otras. Los resultados de la investigación permitieron predecir la incidencia de enfermedades y el rendimiento potencial del cultivo. En dicho estudio, los autores enfatizan en la posibilidad de incrementar el rendimiento de los cultivos y las ganancias de los agricultores, si se logra conocer con anticipación el total del producto que se podría cosechar.

Otro cultivo ampliamente estudiado es el de arroz. En [47], cuantificaron la relación entre la variabilidad del clima (evaporación, precipitación, radiación solar, velocidad del viento, temperatura y humedad relativa) y el rendimiento del mencionado cultivo en Nigeria mediante varios algoritmos de aprendizaje automático, entre los que se encuentran regresión lineal múltiple, análisis de componentes principales y máquina de vectores de soporte. Los resultados de la investigación resaltan la relación directa que existe entre el rendimiento del cultivo y el clima de la región. En un segundo trabajo, expuesto en [48], los investigadores evaluaron la eficiencia de cuatro técnicas de aprendizaje automático (regresión lineal, regresión no lineal, árboles de regresión y redes neuronales artificiales) para predecir el rendimiento del arroz en Bangladesh específicamente. Así mismo, propusieron algunas variables relacionadas con el rendimiento del cultivo, entre las que se encuentran los parámetros climáticos, nutrientes del suelo y manejo del cultivo. Por otra parte, en [49] los investigadores utilizaron máquinas de vectores de soporte para intentar predecir el rendimiento del cultivo de arroz en veintisiete zonas de India a partir de registros adquiridos desde el año 1998 hasta el 2002. Como insumo para el entrenamiento

del algoritmo, los autores utilizaron datos de temperatura, precipitación, evapotranspiración y registros históricos de área sembrada y cosechada. En un cuarto trabajo, expuesto en [50], estimaron el rendimiento del arroz, papa y trigo en Bangladesh, mediante el uso de diferentes técnicas de aprendizaje automático (regresión lineal, k-vecinos más cercanos y redes neuronales). La investigación considera algunos parámetros climáticos (temperatura, humedad, cantidad de luz solar y precipitación) y del suelo (pH y salinidad) como factores relacionados directamente al rendimiento de los cultivos. Los resultados de la investigación resaltan que los modelos para la estimación del rendimiento permiten a los agricultores la oportunidad de aumentar sus ganancias y aumentar la producción global del país.

Además de los cultivos mencionados anteriormente, se han utilizado algoritmos de aprendizaje automático para intentar predecir el rendimiento de la soya, la caña de azúcar, el té, el algodón, entre otros. En primer lugar, en [51] utilizaron técnicas de aprendizaje profundo para construir un modelo que permite predecir el rendimiento del cultivo de soya en Argentina y Brasil. Para llevar a cabo la tarea de predicción, los investigadores tomaron imágenes satelitales MODIS (*moderate resolution imaging spectroradiometer*) que permiten extraer datos de reflectancia, temperatura diurna y nocturna y máscaras de cobertura terrestre. Por su parte, en [52] se presenta un análisis que utiliza una regresión lineal múltiple (MLR, *multiple linear regression*) para estimar la predicción del rendimiento del cultivo de té en India tomando como base el cambio climático observado por 30 años (1977-2006). Las variables observadas en el estudio fueron la cantidad de lluvia, temperatura, humedad relativa, evaporación y luz solar. El modelo construido registró un coeficiente de correlación máximo de 0.82 y demuestra, junto a otros resultados del trabajo mencionado, que la producción de los diferentes cultivos (té, en el caso de la investigación) depende significativamente de los registros climáticos en la región. Así mismo, en [53] se construye un modelo basado en aprendizaje automático para predecir el rendimiento de la caña de azúcar en Tailandia. Los datos de entrenamiento utilizados comprendieron tipo de suelo, área cultivada, variedad de la caña, esquema de cultivo, método de riego, método de control de plagas, tipo y fórmula de fertilizante, cantidad de lluvia y registros históricos de rendimiento. Fueron utilizados dos técnicas: *random forest* y *gradient boosting tree*, siendo la primera la que arrojó mejores resultados, registrando una precisión del 71.83%. En la investigación desarrollada en [54], se construye un modelo que permite estimar el rendimiento del cultivo de algodón en Tennessee, Estados Unidos. NDVI, SR, infrarrojo cercano (NIR, *near infra-red*) e índices de verdor, humedad y brillo del suelo, son algunos de los parámetros capturados a través de imágenes Landsat para el entrenamiento del modelo basado en redes neuronales artificiales. Los resultados obtenidos en la investigación permiten predecir el rendimiento con un error cercano al 8%. Por último, en [55] se utiliza una regresión lineal múltiple, una red neuronal bayesiana (BNN, *bayesian neural network*) y un particionamiento recursivo basado en modelos (MOB, *model-based recursive partitioning*) para estimar el rendimiento de los cultivos de cebada,

canola y trigo en Canadá. Los datos de entrenamiento utilizados en la investigación, consistieron en registros históricos (desde el año 2000 hasta el 2011) del rendimiento de los cultivos en tres provincias canadienses y diferentes índices calculados a partir de imágenes satelitales. Los resultados obtenidos fueron semejantes con las tres técnicas de aprendizaje automático, mostrando un desempeño parecido entre los métodos lineales y los no lineales.

2.2.3 Aprendizaje automático para el análisis del rendimiento del cultivo de café

En cuanto al rendimiento del café, son pocas las investigaciones disponibles en la literatura que han empleado algoritmos de aprendizaje automático para la estimación de su rendimiento. En el mapeo sistemático llevado a cabo fue posible encontrar dos de ellas. En un primer trabajo, presentado en [56], examinaroon las repercusiones del cambio climático en la producción mundial de dos variedades de café, arábica y robusta. Usando datos históricos de 19 variables bioclimáticas, en el estudio entrenaron tres algoritmos de aprendizaje automático: *support vector machines*, *random forest* y *MaxEnt*. Mediante el entrenamiento llevado a cabo se construyeron más de cien modelos que permiten estimar el rendimiento mundial de café. El análisis llevado a cabo en la investigación demuestra que el rendimiento del cultivo de café está estrechamente relacionado con la variabilidad climática y sugiere que temperaturas altas pueden reducir la producción de dicho cultivo a nivel global. Otros resultados importantes mencionan que el cambio climático reducirá el área global adecuada para el café en aproximadamente un 50% en un futuro y que los impactos de dicho fenómeno son más altos en latitudes y altitudes bajas. Además, se muestra que la migración altitudinal de la producción de café probablemente será una tendencia global y se sugiere que en áreas donde la producción de café sea todavía factible, los sistemas de producción tendrán que ser adaptados, planteando así desafíos sustanciales para los pequeños agricultores. Por último, en el trabajo expuesto en [57], compararon tres técnicas de aprendizaje automático (*multiple linear regression*, *random forest* y *extreme learning machine*) para estimar el rendimiento de la variedad robusta de café. Los modelos fueron entrenados utilizando datos de suelo, entre los que se encuentran: materia orgánica, potasio, boro, azufre, zinc, fósforo, nitrógeno, calcio, magnesio y pH. El análisis de la investigación muestra que la técnica con la mayor precisión fue *extreme learning machine* (con un error cuadrático medio de 496.35 kg ha^{-1}), superando los resultados obtenidos con los algoritmos *multiple linear regression* (1072.09 kg ha^{-1}) y *random forest* (1087.35 kg ha^{-1}). Además, en el estudio se muestra la utilidad del uso de diversas técnicas de aprendizaje automático y modelos de cultivos biofísicos para mejorar el rendimiento en granjas de pequeños agricultores.

2.2.4 Aportes y brechas de los trabajos relacionados

En los numerales anteriores se pudo establecer que existen diversas investigaciones realizadas para mejorar la producción agrícola. Algunos trabajos han modelado matemática y estadísticamente la relación entre diferentes variables para observar cómo influyen en el rendimiento de un cultivo, mientras otros han utilizado técnicas de aprendizaje automático para predecir la producción que se puede llegar a tener en diferentes cultivos y en casos específicos, en el café. Con esto, fue posible identificar los posibles aportes de los trabajos encontrados a la presente investigación y las brechas existentes entre estos, tal como se muestra en la Tabla 1.

Sección - trabajos	Aportes	Brechas
Modelos matemáticos y estadísticos para determinar el rendimiento de cultivos	Proponen un modelo matemático para el cálculo del rendimiento de un cultivo. Presentan ejemplos de calibración de modelos matemáticos para el cálculo del rendimiento. Realizan un modelamiento de las relaciones existentes entre diferentes variables incidentes en el rendimiento. Entre las variables se encuentran climáticas, edáficas y de manejo del cultivo.	Encuentran relaciones entre variables para conocer el comportamiento de otra. En la presente investigación se construyó un modelo que además de encontrar relaciones entre variables, permite predecir nuevos valores aprendiendo de datos existentes. Se basan en suposiciones sobre los datos estudiados, lo que limita el modelo estadístico generado.
Aplicaciones de aprendizaje automático para el análisis del rendimiento de cultivos	Facilitan el análisis de la relación entre variables intervinientes en el rendimiento de un cultivo. Estudian la precisión de diferentes técnicas de aprendizaje automático al predecir el rendimiento. Proponen un marco de referencia para el entrenamiento de un modelo para la predicción de la producción.	Llegan hasta la generación y evaluación de un modelo de predicción, sin construir alguna aplicación que utilice el modelo desarrollado. Los estudios son realizados a nivel global o en países europeos, asiáticos o norteamericanos mayoritariamente. La presente investigación está enfocada en Colombia.

	Exploran el uso de algoritmos de aprendizaje automático complejos o novedosos, como aprendizaje profundo y aprendizaje extremo.	
Aprendizaje automático para el análisis del rendimiento del cultivo de café	Utilizan diversos algoritmos de aprendizaje automático para intentar estimar el rendimiento del café. Además, analizan técnicas de regresión y clasificación, para conocer cuál de los dos enfoques genera mejores resultados. Demuestran la estrecha relación que existe entre la variabilidad climática y el rendimiento del café. Analizan la producción del cultivo de café a grandes escalas, demostrando que es posible realizar estimaciones a nivel país y región.	

Tabla 1. Aportes y brechas de los trabajos relacionados. Fuente propia.

La identificación de los aportes y brechas presentados en la Tabla 1, son una muestra de la importancia que ha tenido, y continúa teniendo, el apoyo a la agricultura mundial mediante el uso y aplicación de las ciencias de la computación. Específicamente, temas de interés actual, como *big data, data mining* y *machine learning*, juegan un importante papel a la hora de generar nuevas soluciones a los problemas que enfrenta la agricultura. Teniendo en cuenta esto, la presente investigación integra las disciplinas del conocimiento planteadas en los numerales anteriores, para desarrollar una aproximación que permite incrementar, o en su defecto controlar, el rendimiento del cultivo del café colombiano.

2.3 Conclusiones acerca del estado actual del conocimiento

El presente capítulo presenta, además de los conceptos y definiciones que sirven como base para la consecución de los resultados del presente trabajo, un mapeo sistemático construido para reconocer los trabajos existentes que guardan relación con la problemática abordada en la presente investigación. En este sentido, tomando como base el análisis y los resultados presentados en el capítulo, se concluye:

- Existe gran variedad de investigaciones dedicadas a la estimación o predicción del rendimiento de un cultivo. El enfoque utilizado para abordar dicha problemática va desde el modelamiento matemático y estadístico, hasta la aplicación de técnicas avanzadas de aprendizaje automático. Los cultivos estudiados comprenden el trigo, avena, caña de azúcar, arroz, soya, té, manzana, maíz, girasol, algodón, café, papa, entre muchos más.
- Mediante el uso de algoritmos de aprendizaje automático es posible construir soluciones rentables e integrales para una mejor estimación del rendimiento de los cultivos y la toma de decisiones, ya sea a nivel de finca, región o país. Para ello, pueden ser utilizados datos obtenidos de diversas fuentes, tales como imágenes satelitales, sensores remotos, estaciones meteorológicas, etc.
- Máquinas de vectores de soporte, redes neuronales, regresión lineal multiple y árboles de regresión, son algunas de las técnicas más utilizadas para la estimación del rendimiento agrícola. Otros algoritmos empleados en menor medida, pero con interesantes resultados, son los de aprendizaje profundo y aprendizaje reforzado, entre otros.
- El rendimiento de los cultivos, entre ellos el de café, está estrechamente relacionado con la variabilidad climática. Otras variables que tienen prelación en los modelos para la estimación de la producción agrícola son los índices de vegetación, propiedades del suelo (materia orgánica, calcio, magnesio y pH, etc.) y el manejo del cultivo (cantidad y tipo de fertilizante, por ejemplo).
- Aún cuando existe un amplio número de investigaciones alrededor del rendimiento de cultivos, la presente investigación propone un nuevo enfoque que puede aportar no solamente a la agricultura colombiana, sino también al campo ingenieril. Utilizando una menor cantidad de variables independientes y analizando diferentes algoritmos de aprendizaje automático, en la presente investigación se logra estimar el rendimiento del cultivo de café con seis meses de anticipación a la cosecha. Así mismo, se propone la utilización de sistemas de recomendaciones para sacar provecho de las predicciones realizadas gracias a la inteligencia artificial.

3 Conjunto de datos sobre el rendimiento del café

En este capítulo se presenta el proceso llevado a cabo para la obtención del conjunto de datos meteorológicos y de producción de café, que luego fue utilizado como base del modelo de predicción del rendimiento del mencionado cultivo. Tanto la construcción del conjunto de datos, como el entrenamiento del modelo (proceso presentado en el capítulo 1), fueron realizados siguiendo el marco de trabajo de CRISP-DM (*Cross Industry Standard Process for Data Mining*) [58]. La **Figura 1** muestra las fases propuestas por dicha metodología que fueron implementadas en la presente investigación.

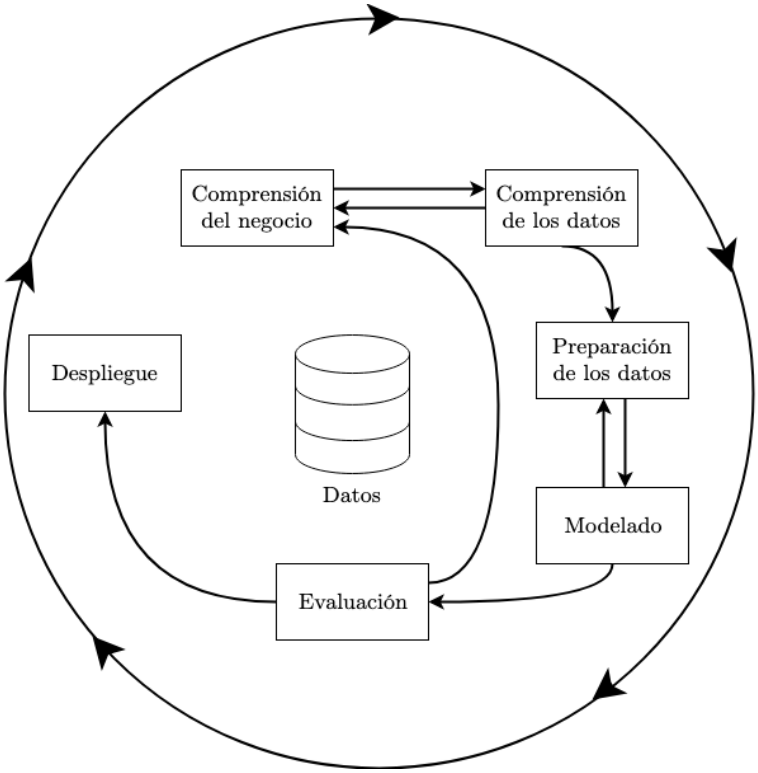


Figura 1. Fases propuestas por el modelo CRISP-DM. Adaptado de [58].

El modelo CRISP-DM ofrece un resumen del ciclo de vida de un proyecto de minería de datos. Este contiene las fases del proyecto, así como sus respectivas tareas y resultados. El ciclo de vida se desglosa en seis fases interconectadas entre sí por flechas que expresan las relaciones más relevantes y frecuentes: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Las tres primeras fases fueron la base para la consecución del conjunto de datos meteorológicos y de rendimiento del café. En este orden de ideas, el capítulo se encuentra dividido en los siguientes apartados:

- **Comprensión del negocio**: se exponen los objetivos que se buscan con la construcción del conjunto de datos y el entrenamiento del modelo, para tener claridad de los registros que se esperan obtener. Como parte de esto, se estudia el rendimiento de un cultivo agrícola en general y el del café, así como las variables que inciden en este.
- **Comprensión de los datos**: se presenta el proceso de recolección de datos iniciales y la descripción de estos.
- **Preparación de los datos**: expone las fases de selección, estructuración, limpieza e integración de los registros recolectados mediante el proceso mencionado en la comprensión de los datos.

3.1 Comprensión del negocio

Se busca construir un conjunto de datos que describa el rendimiento del cultivo de café, con el objetivo de ser utilizado luego como insumo principal para el entrenamiento de un modelo que permita estimar dicha variable. Es por esto que, en la presente investigación, es necesario entender el rendimiento de un cultivo agrícola en general y específicamente el del café, así como las variables que pueden incidir en él, ya sea de una forma positiva o negativa.

El rendimiento agrícola se entiende como la producción alcanzada por unidad de superficie cosechada, expresado comúnmente en toneladas por hectárea (t/ha). En este orden de ideas, el estudio del rendimiento de un cultivo puede abordarse mediante el entendimiento de la producción del mismo. Este último concepto, se entiende como la parte utilizable de la planta, normalmente medida como la cantidad de grano cosechado o de materia seca y expresada en toneladas (t). Cuando dicha producción se relaciona con los recursos usados para su obtención se utiliza el concepto de productividad. Se pueden registrar niveles de productividad agrícola variados, ya que existen una gran cantidad de factores limitantes. Cuando se presentan condiciones ideales, se obtiene la máxima producción posible o producción potencial, que corresponde a la producción registrada bajo varios aspectos: un ambiente físico propicio para la interacción de los factores determinantes, un manejo de cultivo ideal y con el mejor nivel de tecnología y material biológico [59].

La productividad de los diferentes cultivos agrícolas depende de diversas condiciones. La variabilidad climática y factores como el contenido de materia orgánica, topografía, especie y evapotranspiración, entre otros, influyen directamente en la producción agrícola [59]. La variabilidad climática hace referencia a las variaciones en los valores promedios del clima a escala temporal y espacial, más allá de los eventos individuales del tiempo. Como ejemplos de variabilidad climática se tienen sequías extendidas, inundaciones y condiciones resultantes de eventos como El Niño y La Niña [60]. En efecto, el agua siempre ha sido el principal factor que limita la producción en gran parte del mundo, donde la precipitación

no es suficiente para satisfacer la demanda de los cultivos; el estrés hídrico y la disminución de la transpiración tienen como resultado una reducción en la producción de biomasa, lo cual generalmente también reduce el rendimiento [61]. De la misma forma, la productividad del café se ve afectada por un gran número de factores, que pueden ser separados en tres grandes grupos [59]:

- ***Factores climáticos***: precipitación, temperatura atmosférica, humedad relativa, luz solar, velocidad y dirección del viento, concentración de CO_2, altitud y latitud, entre otros.
- ***Factores del suelo***: materia orgánica, capacidad de intercambio catiónico, pendiente y topografía, temperatura del suelo, fertilidad, disponibilidad de nitrógeno, potasio, etc.
- ***Factores del cultivo***: calidad de la semilla, especie y variedad del cultivo, fecha de siembra, densidad de siembra, evapotranspiración y actividades del manejo del cultivo, entre otros.

A continuación, se abordan en detalle algunos de los factores mencionados anteriormente, haciendo énfasis en las características que presentan la mayoría de las regiones cafeteras colombianas.

Disponibilidad de agua. El exceso de agua por un periodo prolongado reduce la floración del café, siendo más marcada en las épocas en que las demás condiciones climatológicas son favorables para la floración [62]. En todas las zonas cafeteras colombianas pueden existir diversas condiciones de suelo y de clima que conducen a niveles críticos de déficit o exceso hídrico. La falta de agua es más frecuente en las regiones con una inadecuada distribución de las lluvias y con texturas del suelo muy arenosas, pedregosas, cascajosas y suelos poco profundos [59].

Características físicas y químicas del suelo. Dichas características influyen directamente sobre la existencia de nutrientes, oxígeno y agua para el cafeto [63]. Por una parte, entre las características químicas del suelo de mayor importancia para el crecimiento de la planta, se encuentran la materia orgánica, el pH y los nutrimentos. La materia orgánica es considerada como uno de los principales indicadores de la productividad del suelo. Entre las funciones que desempeña se pueden destacar las siguientes: es fuente de nitrógeno, fósforo, azufre, boro y zinc, entre otros, incrementa la capacidad de intercambio de cationes, suministra energía para la actividad de los microorganismos, mejora la estructura de las partículas del suelo, capacidad de retención de agua y aireación [59]. Por otra parte, las características físicas del suelo tienen un papel importante en el vigor del cultivo, además, la producción depende en gran parte de la calidad de la relación suelo-aire-agua-temperatura. Dichos factores físicos, combinados con el estado y la cantidad de materia orgánica del suelo, afectan el desarrollo radical de la planta y con ello la capacidad de

absorción de nutrimentos, la colonización de la raíz por organismos benéficos y los procesos fisiológicos de la planta [59].

Temperatura del suelo. Es un factor tan importante como la disponibilidad de agua para el crecimiento óptimo de la planta del café. Entre 10°C y 40°C es el rango adecuado de temperatura del suelo para el crecimiento de las plantas cultivadas, aunque esta varía con la especie, la variedad, la edad del cultivo, el estado de desarrollo y el tiempo de exposición al sol [59].

Especie y variedad. La productividad del cultivo de café es afectada directamente por la integración de factores como la densidad de siembra, variedad sembrada, edad del cultivo, la competencia de plantas acompañantes, beneficio, cosecha y la ubicación donde está establecido el cultivo [63].

Variabilidad climática. Factores como la temperatura atmosférica, la lluvia y la radiación solar son los elementos climáticos de mayor importancia en la producción de café. Las deficiencias hídricas son un requisito primordial para la floración, pero si se prolongan por grandes espacios de tiempo no permiten la apertura floral, limitando así el crecimiento vegetativo y el llenado de los frutos. Por otra parte, los excesos de agua disminuyen la inducción floral y la formación de estructuras reproductivas, favoreciendo la aparición de enfermedades en el cultivo y promoviendo el lavado de nutrientes del suelo y las pérdidas por erosión [64]. Además, altas temperaturas atmosféricas pueden detener la fotosíntesis y evitar la fertilización de los óvulos del cafeto, así como llevar a una deshidratación de la planta [65]. Una de las características principales de las zonas tropicales, típicas colombianas, es que la variabilidad de diversos factores meteorológicos es más pronunciada a nivel diario que a nivel estacional, por lo tanto, los cambios que se registran diariamente de dichos elementos son los que más influyen en la respuesta fenológica del cultivo de café en estas zonas [60].

En cuanto a las zonas cafeteras de Colombia, la variabilidad climática asociada al Fenómeno de El Niño, se ha caracterizado en gran medida por el aumento del número de días con deficiencia hídrica, incrementos en el brillo solar y en la temperatura atmosférica y del suelo, al igual que diferencias diarias de estas últimas variables más pronunciadas, debido, entre otras cosas, a la disminución de la nubosidad, traduciéndose así en condiciones favorables para la floración. Por el contrario, la variabilidad climática asociada con el Fenómeno de La Niña, se ha caracterizado por la marcada reducción o desaparición del déficit hídrico y la aparición de excesos en este factor por encima de lo normal, disminución del brillo solar y la temperatura del suelo, al igual que cambios menos drásticos de la temperatura atmosférica entre el día y la noche, lo que se traduce en disminución del número de botones florales en café y con ello la reducción de la productividad de todo el cultivo [60].

3.2 Comprensión de los datos

La segunda fase de CRISP-DM es la comprensión de los datos. En esta se lleva a cabo la recolección de los datos base para el entrenamiento del modelo, además de realizarse una exploración y descripción de los mismos.

3.2.1 Análisis de fuentes de datos

Como primera medida, se realizó un análisis de posibles fuentes de datos. Como resultado de la búsqueda se obtuvieron las siguientes opciones: el Departamento Administrativo Nacional de Estadística (DANE), el Centro Nacional de Investigaciones de Café (Cenicafé) y las plataformas digitales colombianas Datos Abiertos del Ministerio de Tecnologías de la Información y las Comunicaciones (www.datos.gov.co), Plataforma Agroclimática Cafetera (AgroClima) y Agronet del Ministerio de Agricultura y Desarrollo Rural.

En las fuentes de datos mencionadas anteriormente, se buscaron registros del rendimiento y producción de café en Colombia, así como las variables que influyen en los mismos, mencionadas en la sección 3.1. Al finalizar el proceso de búsqueda, se concluyó que podrían obtenerse registros de rendimiento y producción de café en Colombia, pero en cuanto a las variables que inciden en él, su disponibilidad es mínima, por lo tanto, sería viable la obtención únicamente de registros meteorológicos. En este orden de ideas, se eligieron como fuentes de datos finales la plataforma digital Datos Abiertos y la página web oficial de Cenicafé. Por una parte, en la primera fuente de datos (Datos Abiertos), se encontraron registros históricos de rendimiento y producción de café en diferentes municipios de Colombia, que incluyen datos desde el año 2007 hasta el 2015. Por otra parte, en la página web de Cenicafé, se hallaron los anuarios meteorológicos cafeteros, documentos que registran el clima en los municipios cafeteros de Colombia desde el año 2006. Además de esto, se utilizaron también como fuentes de datos dos documentos de Cenicafé que presentan las zonas de distribución de lluvia en Colombia y los meses en los que se registra la cosecha principal de café en los municipios cafeteros del país, además registran la productividad en las diferentes regiones cafeteras colombianas de acuerdo a la variabilidad climática en las mismas [63], [66].

3.2.2 Recolección de los datos iniciales

La recolección de los datos se llevó a cabo en dos etapas. La primera etapa fue la obtención de los registros de rendimiento y producción anual de café en los municipios cafeteros de Colombia a través de la plataforma digital Datos Abiertos del Ministerio de Tecnologías de la Información y las Comunicaciones (www.datos.gov.co). La segunda etapa fue la extracción de los datos climáticos mensuales de los mismos municipios para los que se obtuvieron los registros de rendimiento, desde los anuarios meteorológicos cafeteros de

Cenicafé (disponibles en www.cenicafe.org). Se llevó a cabo el proceso presentado en la Figura 2 para la extracción de los datos climáticos necesarios para el proceso de modelamiento.

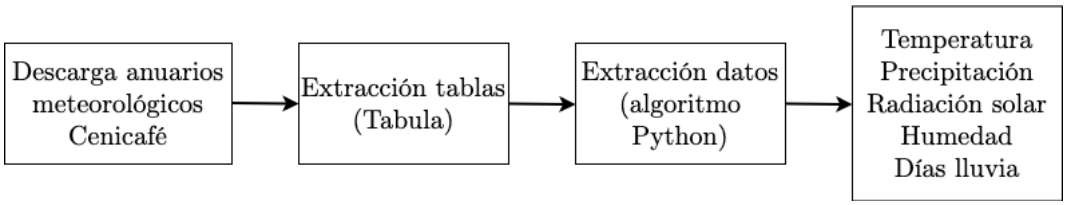


Figura 2. Proceso de extracción de datos anuarios meteorológicos. Fuente propia.

Como se expone en el proceso de la Figura 2, para la recolección de los datos meteorológicos, se descargaron los archivos PDF con los anuarios meteorológicos desde la plataforma digital de Cenicafé. Dichos ficheros contenían registros climáticos mensuales de los municipios cafeteros colombianos desde el año 2006 hasta el año 2016, organizados en tablas. Para extraer entonces los datos de interés desde las tablas en las que se encontraban, fue necesario el uso de Tabula, una herramienta de código abierto y gratuita para extraer contenido tabular desde archivos PDF [67]. En este sentido, utilizando la herramienta mencionada, se extrajeron las tablas con los registros climáticos mensuales necesarios, que comprendían desde el año 2007 hasta el 2015, para cada uno de los municipios de los que ya se tenían datos de rendimiento y producción. Puesto que las tablas extraídas contenían datos innecesarios, se implementó un algoritmo usando el lenguaje de programación Python, que permitió extraer los datos de interés para la investigación. Así, se tuvieron como resultado dos conjuntos de datos, uno con instancias de rendimiento y producción y otro con registros climáticos compuesto por las siguientes variables: temperatura (máxima, mínima, promedio, máxima absoluta y mínima absoluta), humedad relativa, precipitación, días de lluvia y horas de brillo solar.

3.3 Preparación de los datos

La fase de preparación de los datos está compuesta por varias tareas, entre las que se encuentran: la selección de los datos que serán utilizados para el modelamiento y la estructuración de los mismos, así como la verificación de su calidad, limpieza e integración en un solo conjunto de datos.

Conociendo que, por una parte, se tenían datos de rendimiento y producción de café anual (2007 al 2015) de un total de 53 municipios cafeteros y, por otra parte, se tenían los datos climáticos (temperatura, humedad relativa, precipitación, días de lluvia y horas de brillo solar) mensuales para las mismas ubicaciones y en los mismos años, como primera medida se desarrolló un algoritmo utilizando el lenguaje de programación Python para que ambos

conjuntos de datos tuviesen una misma temporalidad en sus registros. Dado que los datos de producción y rendimiento eran anuales, el algoritmo mencionado fue desarrollado para que cada instancia de los registros meteorológicos correspondiese al clima presentado en los doce meses de un año en un mismo municipio, tal como se muestra en la Figura 3.

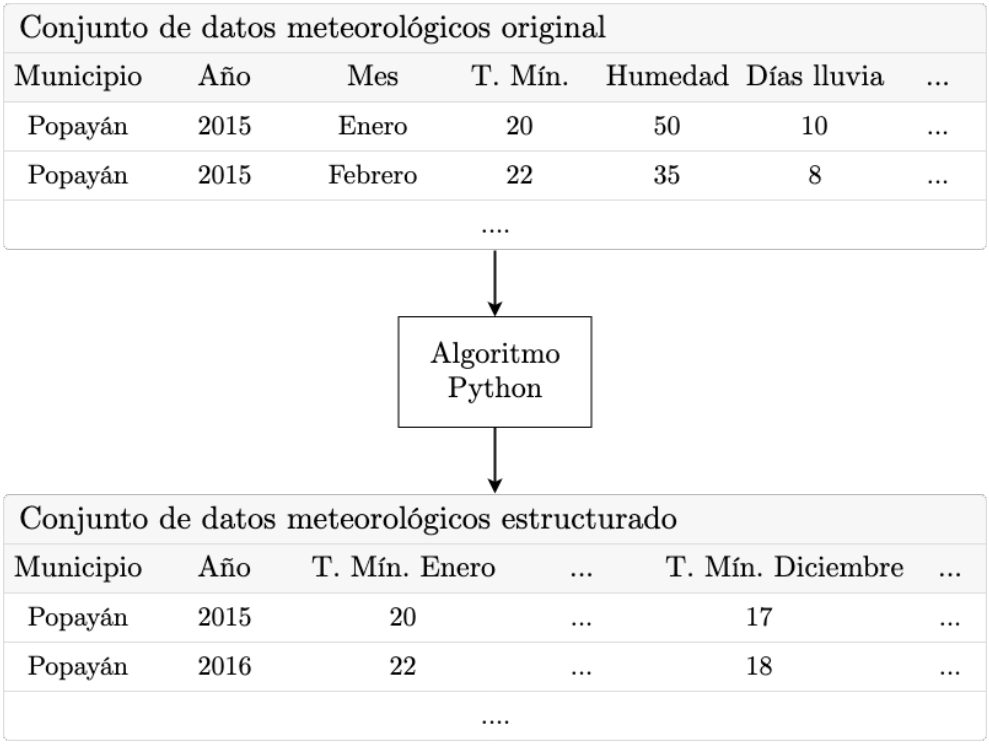


Figura 3. Estructuración del conjunto de datos meteorológicos. Fuente propia.

Con el proceso presentando en la Figura 3, se obtuvieron los dos conjuntos de datos (rendimiento y clima) en la misma temporalidad. En un segundo paso, se llevó a cabo un diagnóstico de la calidad de los datos [68], además de realizar la limpieza de los mismos. En la Figura 4, se presentan las tareas de limpieza de datos llevadas a cabo.

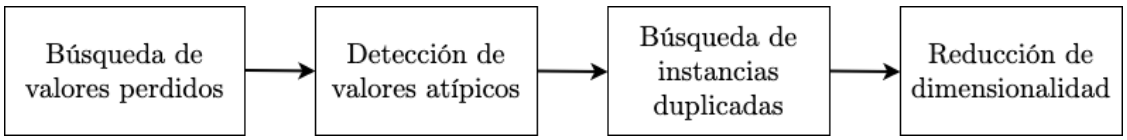


Figura 4. Proceso de limpieza de datos. Adaptado de [69] y [70].

- ***Búsqueda de valores perdidos***: espacios en blanco, palabras como "NaN" o "null" y caracteres especiales como "*" o "?" fueron buscados para verificar que no existiesen valores perdidos en los conjuntos de datos. Se encontraron algunas instancias con un gran número de valores perdidos (más de la mitad de atributos

faltantes), por lo que fueron eliminadas completamente de los conjuntos de datos. Por otra parte, no se eliminaron algunas instancias con pocos atributos faltantes, teniendo como objetivo principal que el modelo entrenado pueda generar resultados precisos con base en nuevas instancias que no necesariamente tengan registros de todos sus atributos. Finalmente, en el mismo orden de ideas, no se llevó a cabo un proceso de imputación de datos.

- ***Detección de valores atípicos***: para detectar si existían valores que se desviaran significativamente de los demás, se utilizaron dos algoritmos, DBSCAN (*Density-based spatial clustering of applications with noise*) y LOF (*Local Outlier Factor*). Los valores extremos y atípicos encontrados con los dos algoritmos mencionados fueron eliminados de los dos conjuntos de datos.
- ***Búsqueda de instancias duplicadas***: se utilizó un filtro para detectar si existían casos con esta anomalía. El filtro no encontró instancias con dicho problema, dado que los dos conjuntos de datos fueron construidos cuidadosamente y no se implementaron algoritmos de imputación en los pasos anteriores.
- ***Reducción de dimensionalidad***: se analizaron los atributos que mayor grado de correlación tenían con la clase objetivo. Mayor información sobre el proceso de selección de atributos, se expone en la sección 4.2.3.

Luego de los procesos mencionados anteriormente, se obtuvieron como resultado dos conjuntos de datos procesados. Por una parte, los registros anuales (2006 al 2015) de rendimiento y producción de café en 53 municipios cafeteros de Colombia, conformados por los siguientes atributos: área sembrada, área cosechada, producción y rendimiento. Por otra parte, datos históricos meteorológicos para los mismos municipios, conformados por los siguientes atributos: temperatura (máxima, mínima, promedio, máxima absoluta y mínima absoluta), precipitación, horas de brillo solar, humedad relativa y días de lluvia, para cada mes del año. En un cuarto paso, los dos conjuntos de datos mencionados fueron unidos en uno. Para esto, utilizando los mapas cafeteros y las tablas de meses de floración y cosecha disponibles en [63] y [66] se definieron cuáles eran los meses que conformaban un año cafetero en cada municipio, en este orden de ideas, cada instancia quedó compuesta por los registros de producción y rendimiento en un año, así como las condiciones climáticas durante el respectivo año cafetero (el mes en que se registró la cosecha y los once inmediatamente anteriores), tal como se muestra en la Figura 5.

Municipio	Año	Área sembrada	Producción	Rendimiento	...	T. Mín. Mes 1	...	Brillo Sol Mes 12
Popayán	2015	3214	2488	1.1	...	20	...	13
				...				

Figura 5. Ejemplo de una instancia construida con registros de rendimiento y variables climáticas. Fuente propia.

En el mismo orden de ideas, tomando como base [63], se agregó un nuevo atributo: una variable lógica que representa si en el municipio existe una mitaca (cosecha secundaria con producción mucho más baja que la principal, conocida en Colombia también como traviesa). Además, siguiendo los mapas (véase anexo B) expuestos en [66], [71], se determinó la región cafetera y de distribución de lluvia en Colombia a la que pertenece el municipio del registro, generándose así dos atributos nuevos (región cafetera y región distribución de lluvia). Por último, dado que se detectó heterocedasticidad en los datos (la varianza de los errores no era constante en todas las observaciones) y su distribución de probabilidad era altamente asimétrica, se normalizaron los dos atributos objetivos que se tenían en el conjunto de datos (producción y rendimiento) utilizando la función logaritmo natural [72]–[74]. Con ello, se construyeron entonces otros dos atributos: el logaritmo natural de la producción y del rendimiento, expresados como Ln(Producción) y Ln(Rendimiento) respectivamente. Al finalizar todo el proceso de construcción, se generó un conjunto de datos con 410 instancias y 123 atributos, que se refieren a los siguientes metadatos: año del registro, departamento, municipio, si existe (o no) una mitaca en el año, área sembrada con café, área cosechada, producción, rendimiento y el registro de las variables climáticas de cada uno de los doce meses que hacen parte del año cafetero, contando como doceavo el mes en el que se registra la cosecha. En la Tabla 2, se presentan en detalle cada uno de los atributos que componen el conjunto de datos base para el entrenamiento del modelo.

Atributo	**Descripción**	**Tipo**	**Unidad**
Año	Año (2007 - 2015) para el que se tiene el registro	Nominal	-
Departamento	Departamento al que corresponde la instancia	Nominal	-
Municipio	Municipio al que corresponde la instancia	Nominal	-
Altitud	Altura sobre el nivel del mar en la que está ubicada el municipio	Numérico	m
Floración	Mes en el que se da la floración en el municipio correspondiente	Nominal	-
Cosecha	Mes en el que se da la cosecha	Nominal	-
Mitaca	Existe (o no) una mitaca en el municipio	Nominal	Si/No
Región cafetera	Región cafetera colombiana (determinada por la Federación Nacional de Cafeteros - FNC) a la que pertenece el municipio	Nominal	-
Región distribución de lluvia	Región de distribución de lluvia (determinada por la FNC) a la que pertenece el municipio	Nominal	-

Área sembrada	Área que se registró como sembrada en el año correspondiente	Numérico	ha
Área cosechada	Área que se cosechó en el año correspondiente	Numérico	ha
Producción	Cantidad de café producida en el municipio en el año correspondiente	Numérico	t
Rendimiento	Rendimiento registrado en el municipio en el año respectivo	Numérico	t/ha
Ln(Producción)	Logaritmo natural de la cantidad de café producida en el municipio	Numérico	-
Ln(Rendimiento)	Logaritmo natural del rendimiento registrado en el municipio	Numérico	-
T_Min_Med_MX	Media mensual de temperaturas mínimas diarias. **NOTA** (aplica para todas las filas siguientes): "X" hace referencia al mes del año cafetero, entre 1 y 12, donde 12 es el mes en que se registró la cosecha.	Numérico	°C
T_Max_Med_MX	Media mensual de temperaturas máximas diarias	Numérico	°C
T_Med_MX	Media mensual de temperaturas medias diarias	Numérico	°C
T_Max_Abs_MX	Temperatura máxima absoluta en el mes	Numérico	°C
T_Min_Abs_MX	Temperatura mínima absoluta en el mes	Numérico	°C
Hum_Rel_MX	Media mensual de humedad relativa media diaria	Numérico	%
Pre_Tot_MX	Media mensual de precipitaciones medias diarias	Numérico	mm
Dias_Llu_MX	Número de días lluviosos (precipitación acumulada >= 1mm) en el mes	Numérico	días
Bri_Sol_MX	Media mensual de horas diarias con brillo solar	Numérico	horas

Tabla 2. Variables del conjunto de datos base para el modelamiento. Fuente propia.

En la Tabla 2, se exponen las variables que hacen parte del conjunto de datos que agrupa registros de producción y rendimiento del cultivo de café en municipios colombianos, así como datos meteorológicos históricos de los mismos. Dicho conjunto de datos es la base

para el entrenamiento de un modelo que permita predecir el rendimiento del cultivo de café esperado en los municipios cafeteros de Colombia, que se presenta en el siguiente capítulo.

3.4 Conclusiones acerca del conjunto de datos sobre el rendimiento del café

El presente capítulo expone las tres primeras fases de CRISP-DM (comprensión del negocio, comprensión de los datos y preparación de los datos) seguidas para la construcción de un conjunto de datos sobre el rendimiento del café. Utilizando registros meteorológicos e históricos de rendimiento y producción de dicho cultivo en diferentes municipios cafeteros de Colombia, se construyó un conjunto de datos que es insumo principal para el entrenamiento de un modelo que permite estimar el rendimiento del cultivo de café, tal como se presenta en el siguiente capítulo. En este orden de ideas, tomando como base los procesos y resultados expuestos en el presente capítulo, se concluye:

- Siguiendo la metodología CRISP-DM se llevaron a cabo varios procesos iterativos que permitieron el entendimiento del problema, la recolección de datos y la construcción de un conjunto de datos que describe el rendimiento del cultivo de café en 53 municipios colombianos. Dicho conjunto de datos, al encontrarse organizado y efectivamente descrito, puede ser utilizado como insumo principal para investigaciones alrededor de la variabilidad climática y la producción de café en Colombia.
- La productividad de los diferentes cultivos agrícolas, entre ellos el del café, guarda una estrecha relación con diversas condiciones climáticas, físicas y químicas. La variabilidad climática y factores como la fecha de siembra, disponibilidad de materia orgánica y nitrógeno, edad del cultivo, calidad de la semilla y la evapotranspiración, entre otros, influyen directamente en la producción agrícola.
- Aunque el rendimiento del cultivo de café depende de varios factores edáficos, fenológicos y de manejo del mismo, las variables climáticas como la temperatura, la lluvia y la radiación solar, cumplen un papel determinante sobre este. Las deficiencias hídricas prolongadas, los excesos de agua y las altas temperaturas atmosféricas, por ejemplo, traen consigo cambios drásticos en el rendimiento del café.
- En Colombia, el acceso a los datos agrícolas es limitado. Específicamente, la construcción de un conjunto de datos del rendimiento del cultivo de café es una tarea compleja, debido a que existen pocos datos disponibles y el acceso a estos presenta una gran limitación.

4 Modelo de regresión para la estimación del rendimiento del café

El presente capítulo se centra en la construcción de un modelo de aprendizaje automático que permita la estimación del rendimiento del cultivo de café. Para el entrenamiento del mencionado modelo se tomó como base CRISP-DM, específicamente la fase de modelado de dicho marco de trabajo. A continuación, en la Figura 6, se exponen las tres actividades realizadas como parte del proceso de modelamiento.

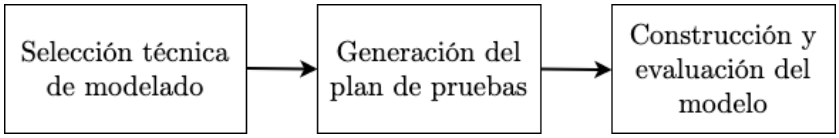


Figura 6. Etapas que conforman la fase de modelado. Adaptado de [58].

Siguiendo las etapas que se exponen en la Figura 6, el presente capítulo se encuentra dividido en los siguientes apartados:

- **Selección de la técnica de modelado**: en esta primera etapa se analizaron diversos algoritmos de aprendizaje automático para encontrar el que presente el mejor comportamiento con base en el conjunto de datos expuesto en el capítulo 1. Tomando como punto de partida diferentes investigaciones en las que se realizó una evaluación de técnicas de aprendizaje automático, en el presente proyecto fueron analizadas seis diferentes: *linear regression, multilayer perceptron, context-sensitive multilayer perceptron, support vector machine for regression, multivariate regression prediction model* y *random forest*.
- **Generación del plan de pruebas**: se describen los parámetros de desempeño que permiten validar y entender los resultados obtenidos a través del entrenamiento del modelo de aprendizaje automático.
- **Construcción y evaluación del modelo**: se exponen todas las pruebas realizadas durante el proceso de entrenamiento del modelo para la estimación del rendimiento del cultivo de café.

Tanto la selección de la técnica de modelado como la construcción y evaluación del modelo fueron realizadas utilizando la versión 3.8.2 del software Weka (*Waikato Environment for Knowledge Analysis*), que provee un conjunto de algoritmos implementados y herramientas para la consecución de proyectos de aprendizaje automático, incluyendo tanto la preparación de los datos de entrada como el entrenamiento y la evaluación de los

modelos [75]. Dicho software incluye algoritmos para tareas de regresión, clasificación, agrupamiento, asociación y selección de atributos que, combinados con técnicas de análisis estadístico y visualización de resultados, permiten construir modelos basados en aprendizaje automático y el seguimiento de metodologías como CRISP-DM [76].

4.1 Selección de la técnica de modelado

Con el objetivo de seleccionar la técnica adecuada para generar un modelo que permita la estimación del rendimiento del cultivo de café, se tomaron como punto de partida los resultados obtenidos a partir de la generación del estado actual del conocimiento expuesta en el capítulo 1.

Por una parte, un primer grupo de investigaciones, presentadas en [26], [31], [39], [56], [57], lograron predecir el rendimiento de diversos cultivos agrícolas haciendo uso de modelos de regresión lineal (simple y múltiple). Igualmente, un segundo grupo de investigaciones, expuestas en [38]–[40], [45], [56], [57], expusieron las máquinas de vectores de soporte como uno de los algoritmos más utilizados para la estimación del rendimiento agrícola. De manera similar, en [38], [39] lograron generar diversos modelos mediante el uso de redes neuronales. Por último, en [38], [39], [41], [42], [53], [56], [57] implementaron árboles de regresión para modelar el comportamiento de cultivos como el maíz, el trigo y el arroz.

Por otra parte, en una actividad previa, desarrollada en el marco de la presente investigación y expuesta en [69], fueron evaluadas cuatro técnicas de aprendizaje automático para analizar la fluctuación del mercado del aguacate en Estados Unidos. Entre las técnicas evaluadas se encuentran: regresión lineal, redes neuronales (*multilayer perceptron*), máquinas vectores de soporte para regresión y árboles de regresión (*multivariate regression prediction model*), siendo las dos últimas técnicas las que arrojaron mejores resultados en el modelamiento.

Con base en las publicaciones mencionadas anteriormente, se eligieron las siguientes seis técnicas de aprendizaje supervisado para ser evaluadas: *linear regression, multilayer perceptron, context-sensitive multilayer perceptron, support vector machine for regression* (implementación *SMOreg* de Weka)*, multivariate regression prediction model* (implementación *M5P* de Weka) y *random forest*. Como primera medida, teniendo en cuenta que el comportamiento de una u otra técnica de aprendizaje automático al intentar modelar una variable objetivo depende del dominio de aplicación y los tipos de datos, se evaluó la precisión de las mismas mediante el entrenamiento de un modelo basado en el conjunto de datos presentado en el capítulo 1. Para ello, se analizó la correlación de las diferentes técnicas al predecir el rendimiento del cultivo a partir del municipio, área sembrada y condiciones climáticas del año cafetero. En ese orden de ideas, la Tabla 3

muestra tres parámetros (descritos en la sección 4.2.2), el coeficiente de correlación, el error medio absoluto y el error cuadrático medio, resultados del entrenamiento del modelo.

	Linear regression	Multilayer perceptron	CS M. perceptron	SMOreg	M5P	Random forest
Coeficiente de correlación	0.57	0.48	0.47	0.57	0.64	0.65
Error medio absoluto	0.22	0.27	0.27	0.21	0.19	0.19
Error cuadrático medio	0.28	0.34	0.33	0.28	0.24	0.24

Tabla 3. Precisión de las técnicas de aprendizaje automático pre-seleccionadas. Fuente propia.

Los parámetros presentados en la Tabla 3 muestran que los dos algoritmos con el mejor desempeño son *random forest* y *multivariate regression prediction model* (implementación *M5P* de Weka), seguidos por *linear regression* y *support vector machine for regression* (implementación *SMOreg* de Weka), mientras que *multilayer perceptron* y *context-sensitive multilayer perceptron* presentan la menor precisión.

A partir del análisis de precisión llevado a cabo, los resultados de las investigaciones mencionadas previamente y el modelo de evaluación expuesto en [77], fue posible realizar una comparación de las seis técnicas de aprendizaje automático estudiadas, para así seleccionar una, que es la base para el siguiente paso, el entrenamiento del modelo que permite estimar el rendimiento del cultivo de café. Para ello, se evaluaron cinco métricas: el coeficiente de correlación resultado de los modelos (etiquetado como precisión), la tolerancia al ruido (el nivel de afectación en el proceso de entrenamiento por parte de datos categorizados como ruido), la tolerancia a valores perdidos (el nivel de afectación en el proceso de entrenamiento por parte de datos faltantes), la velocidad de entrenamiento del modelo (velocidad de aprendizaje) y la velocidad con la cual el modelo es capaz de predecir un nuevo valor de la clase (velocidad de clasificación).

La Tabla 4, presenta la evaluación de las técnicas a través de las diferentes métricas mencionadas y que pueden tomar valores entre 1 y 4, siendo 4 el mejor rendimiento y 1 el peor. La precisión, la velocidad de aprendizaje y la velocidad de clasificación, fueron determinadas utilizando el modelo mencionado en la construcción de la Tabla 3. Por su parte, la tolerancia al ruido y a valores perdidos se tomaron a partir de las investigaciones expuestas en [77], [78].

	Linear regression	Multilayer perceptron	CS M. perceptron	SMOreg	M5P	Random forest
Precisión	3	2	2	3	4	4
T. ruido	1	2	2	2	3	2
T. val. perdidos	2	1	2	2	3	3
V. aprendizaje	3	1	1	2	4	4
V. clasificación	4	3	4	4	4	4

Tabla 4 . Comparación de las técnicas de aprendizaje automático pre-seleccionadas. Fuente propia.

Con el propósito de determinar la técnica adecuada para la fase de modelamiento, los resultados obtenidos en la Tabla 4 se resumen en la Figura 7, donde se exponen las métricas de desempeño de las seis técnicas de aprendizaje automático pre-seleccionadas.

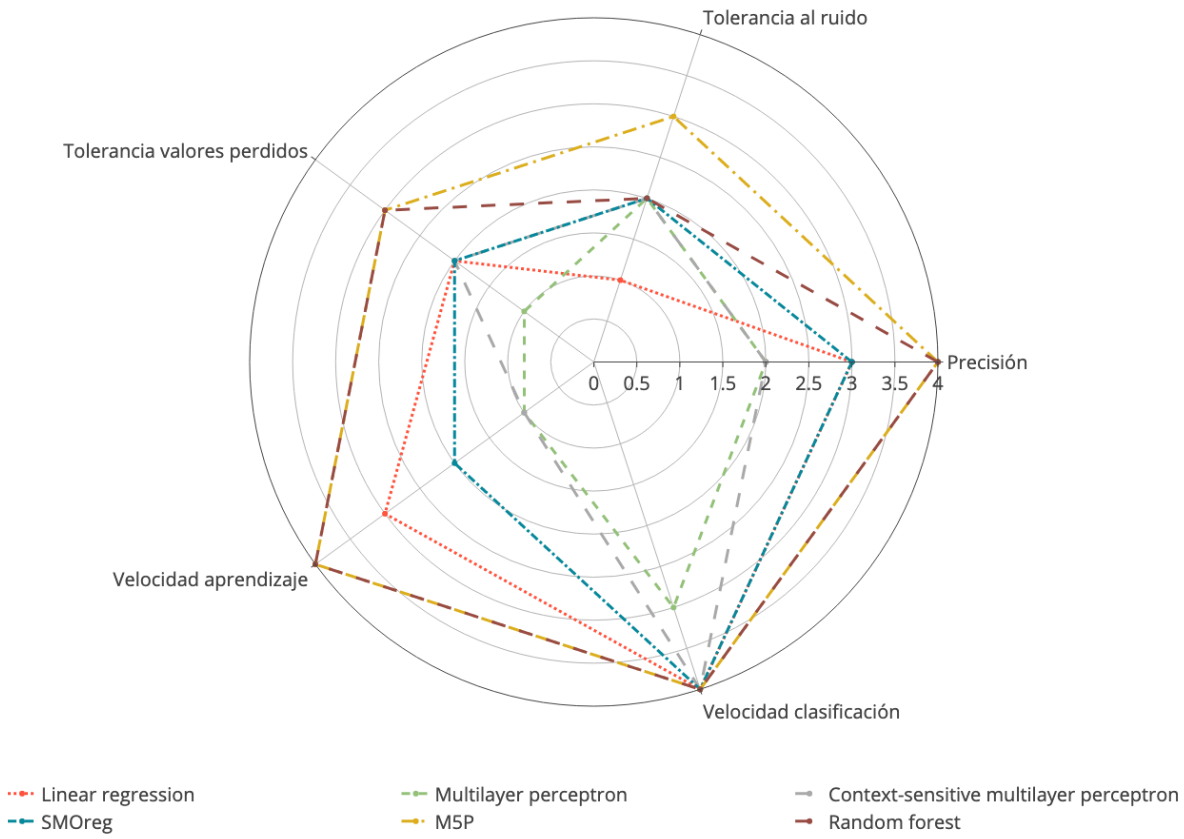


Figura 7. Evaluación de las técnicas de aprendizaje automático pre-seleccionadas. Fuente propia.

De la Figura 7, puede inferirse que los algoritmos *multivariate regression prediction model* (implementación *M5P* de Weka) y *random forest* son los que presentan el mejor

desempeño respecto al problema abordado en la presente investigación. En vista de que el algoritmo *multivariate regression prediction model* (implementación *M5P* de Weka) presenta la capacidad de manejar eficientemente conjuntos de datos con una gran cantidad de atributos y es robusto cuando se trata de datos faltantes [79] (dos características del conjunto de datos construido en el marco de la presente investigación), es seleccionado como el algoritmo de aprendizaje automático para llevar a cabo el entrenamiento del modelo que permita la estimación del rendimiento del cultivo de café.

M5P es la implementación en Weka del algoritmo *M5* propuesto originalmente por J. R. Quinlan [80], que introducía árboles de decisiones para predecir variables continuas, en este caso llamados árboles de regresión. A diferencia de otros árboles de regresión, el algoritmo *M5* no tiene valores en sus hojas, sino modelos lineales de múltiples variables que permiten obtener una mejor precisión. Posteriormente, dicho algoritmo fue modificado por Wang y Witten [81] para tener un especial manejo con atributos continuos y valores perdidos, mediante cambios en la función de reducción de desviación estándar (SDR, *standard deviation reduction*) [82].

Una vez seleccionada la técnica de aprendizaje automático base para el modelamiento, en la siguiente sección, se procede a revisar las diferentes métricas que serán evaluadas para determinar el modelo que presente el mejor desempeño al estimar el rendimiento del cultivo de café.

4.2 Generación del plan de pruebas

El objetivo del presente capítulo es la construcción de un modelo que permita predecir el rendimiento del cultivo de café. Para este fin, se debe realizar una evaluación del modelo después de construido, esto se puede lograr mediante tres enfoques: entrenar y probar con el mismo conjunto de datos; dividirlo en dos diferentes, uno de prueba y uno de entrenamiento; o dividirlo en varios distintos para luego promediar los resultados. Este último enfoque tiene el nombre de validación cruzada y es la base para la evaluación del modelo de la presente investigación. Así mismo, en esta sección, se describen las diferentes métricas base de la evaluación del modelo generado. Por otra parte, pocas veces el primer modelo entrenado, usando el conjunto de datos base en su totalidad (todas las instancias y atributos), presenta una alta precisión, de modo que se hace necesario el entrenamiento de diferentes modelos que permitan, mediante la selección de conjuntos más pequeños de diferentes atributos y "ensayo y error", terminar en la construcción de un último modelo que registre la precisión esperada. Por consiguiente, la presente sección también estudia las técnicas de selección de atributos que permitan acotar las características más importantes del conjunto de datos base.

4.2.1 Validación cruzada

Existen varios enfoques que permiten entrenar y evaluar un modelo de regresión. El primero de ellos consiste en realizar los dos procesos mencionados utilizando el mismo conjunto de datos. En este, como primera medida, se utilizan todas las instancias para el entrenamiento y construcción del modelo; luego, se elige una pequeña porción de ellas como el conjunto de pruebas. Las instancias de dicho conjunto, sin sus respectivas clases objetivo, son utilizadas para predecir nuevos valores mediante el modelo ya construido. Finalmente, se comparan los valores predichos por el modelo con los valores reales en el conjunto de prueba y así conocer la precisión del mismo. Sin embargo, este enfoque presenta una limitación; debido a que el modelo es evaluado utilizando una porción del mismo conjunto de datos con el que fue construido, es probable que se obtenga una alta precisión de entrenamiento (*in-sample accuracy*), pero no una alta precisión fuera de la muestra (*out-of-sample accuracy*).

La precisión de entrenamiento es el porcentaje de predicciones correctas que hace el modelo cuando se utiliza el conjunto de pruebas para evaluarlo. No obstante, registrar una alta precisión de entrenamiento no es necesariamente algo positivo, ya que puede resultar en un ajuste excesivo de los datos, teniéndose así un modelo altamente capacitado para el conjunto de datos, pero no para nuevos registros, es decir, un modelo no generalizado. Por el contrario, la precisión fuera de la muestra es el porcentaje de predicciones correctas que el modelo hace a partir de datos desconocidos; un buen modelo de aprendizaje automático debe registrar entonces una alta precisión fuera de la muestra [83], [84].

Un segundo enfoque, que permite mejorar la precisión fuera de la muestra, consiste en dividir el conjunto de datos en dos diferentes, uno de entrenamiento y uno de prueba, mutuamente excluyentes. Esto proporciona una evaluación más precisa, ya que el conjunto de datos con el que se realiza la evaluación no hace parte del entrenamiento del modelo, siendo así una solución más realista. Sin embargo, al dividir el conjunto de datos en dos, los valores que pertenecen al conjunto de prueba ahora no harán parte del modelamiento, perdiéndose así datos valiosos que podrían mejorar la precisión del mismo [85].

Buscando utilizar todos los datos tanto en el entrenamiento como la evaluación y obtener la mayor precisión fuera de la muestra, aparece un tercer enfoque, llamado validación cruzada (*cross validation*). Este enfoque consiste en dividir el conjunto de datos en k subconjuntos para luego tomar uno de ellos como el conjunto de pruebas, mientras que los restantes *k-1* se convierten en el conjunto de entrenamiento. Esta tarea se repite k veces, durante las cuales cada uno de los subconjuntos es utilizado una única vez como conjunto de prueba. Finalmente, se promedian todos los resultados para generar así el modelo final [86]. En el caso de la presente investigación, se utiliza un $k = 10$, es decir, se divide el conjunto de datos en 10 subconjuntos, donde cada uno de ellos se utiliza una vez como conjunto de prueba y el restante 90% de los datos como conjunto de entrenamiento.

4.2.2 Métricas de evaluación del modelo

Para determinar la precisión del modelo de regresión entrenado con el objetivo de predecir el rendimiento del cultivo de café, se deben evaluar los datos estimados frente a los reales mediante el cálculo de diferentes métricas estadísticas. A continuación, se describen las métricas utilizadas para la evaluación del desempeño del modelo expuesto en la sección 4.3.

- ***Coeficiente de correlación (CC)***: es una de las métricas estadísticas más utilizadas para la evaluación de diferentes resultados de investigaciones. El CC permite medir qué tan fuerte es la relación existente entre dos o más variables [87]. Para el caso de la presente investigación, el cálculo del CC proporciona el grado de relación entre los datos reales y los generados por el modelo. Esta métrica se define mediante la siguiente ecuación [41]:

$$CC = \frac{cov(r,e)}{\sigma_r \sigma_e} = \frac{\sum_{i=1}^{n}[(r_i - \bar{r})(e_i - \bar{e})]}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2 \sum_{i=1}^{n}(e_i - \bar{e})^2}}$$

  Donde $cov(r,e)$ es la covarianza de los datos reales y los estimados, σ_r y σ_e se refieren a la desviación estándar de los mismos, n es el número de datos estudiados, r_i indica el valor real en la posición i, $\bar{r}$ es la media de los datos reales, e_i es el i^o valor estimado y $\bar{e}$ es la media de los datos estimados. Tanto $\bar{r}$ como $\bar{e}$ se pueden calcular mediante las siguientes ecuaciones:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad\qquad \bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$$

  El CC puede tomar valores entre -1 y 1, de la siguiente forma: un resultado igual a 1 indica una correlación positiva perfecta, mientras que al ser igual a -1 indica una correlación negativa perfecta. Un CC igual a 0 significa que no existe ninguna relación entre las variables estudiadas.

- ***Error medio absoluto (MAE, por sus siglas en inglés)***: mide la magnitud promedio de los errores presentes en un conjunto de valores estimados, sin tener en cuenta la dirección de cada uno de ellos. Comúnmente, se calcula como el promedio de las diferencias absolutas entre los valores reales y los estimados, a través de la siguiente ecuación [57]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|r_i - e_i|$$

  Donde r_i indica el valor real en la posición i, e_i es el i^o valor estimado y n es el número de datos estudiados.

- ***Error cuadrático medio (RMSE, por sus siglas en inglés)***: es una métrica estadística, que al igual que el MAE, también permite estimar la magnitud promedio del error. Elevar al cuadrado los errores antes de promediarlos, otorga un peso relativamente alto a los mayores de ellos, siendo así mucho más útil el uso del RMSE cuando no se desea que el modelo pueda generar errores grandes. Esta métrica se calcula mediante la siguiente ecuación [88]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - e_i)^2}$$

Donde r_i indica el valor real en la posición i, e_i es el i^o valor estimado y n es el número de datos estudiados.

4.2.3 Selección de atributos

En la actualidad, es común encontrar conjuntos de datos que contienen una gran cantidad de atributos, donde varios de ellos pueden ser irrelevantes, redundantes o, peor aún, desfavorables para el proceso de modelamiento. Por consiguiente, llevar a cabo una selección de atributos puede traer importantes mejoras, tales como: aumentar la precisión del modelo, reducir tiempo y costo computacional y permitir una mejor visualización o comprensión de los datos. En este orden de ideas, existen dos opciones principales para llevar a cabo una selección de atributos: conocimiento de experto o un método tradicional de selección [70].

Considerando lo mencionado anteriormente y que el conjunto de datos base de la presente investigación contiene 123 atributos, se eligen dos métodos tradicionales de selección de atributos para evaluar cuáles de ellos inciden directamente en la variable objetivo:

- ***CfsSubsetEval***: permite elegir un subconjunto de atributos que tengan la mayor capacidad predictiva individual (correlación alta con la clase) y el menor grado de redundancia entre ellos (correlación baja con las demás características) [89].
- ***CorrelationAttributeEval***: evalúa la importancia de cada atributo mediante la obtención de la correlación entre él y la clase [90].

4.3 Construcción y evaluación del modelo

Tal como se expuso en las secciones 4.1 y 4.2, la construcción del modelo de predicción del rendimiento es realizada mediante la implementación *M5P* en Weka, utilizando una validación cruzada con diez subconjuntos. Así mismo, buscando mejorar los resultados obtenidos, mediante la selección de diferentes combinaciones de datos (instancias y atributos) se construyeron varios modelos. En general, se eligieron diferentes atributos,

así como determinados conjuntos de instancias, para entrenar diferentes modelos hasta mejorar los resultados de los mismos. A continuación, en la Figura 8, se muestran las combinaciones de atributos e instancias utilizadas en el proceso de modelamiento.

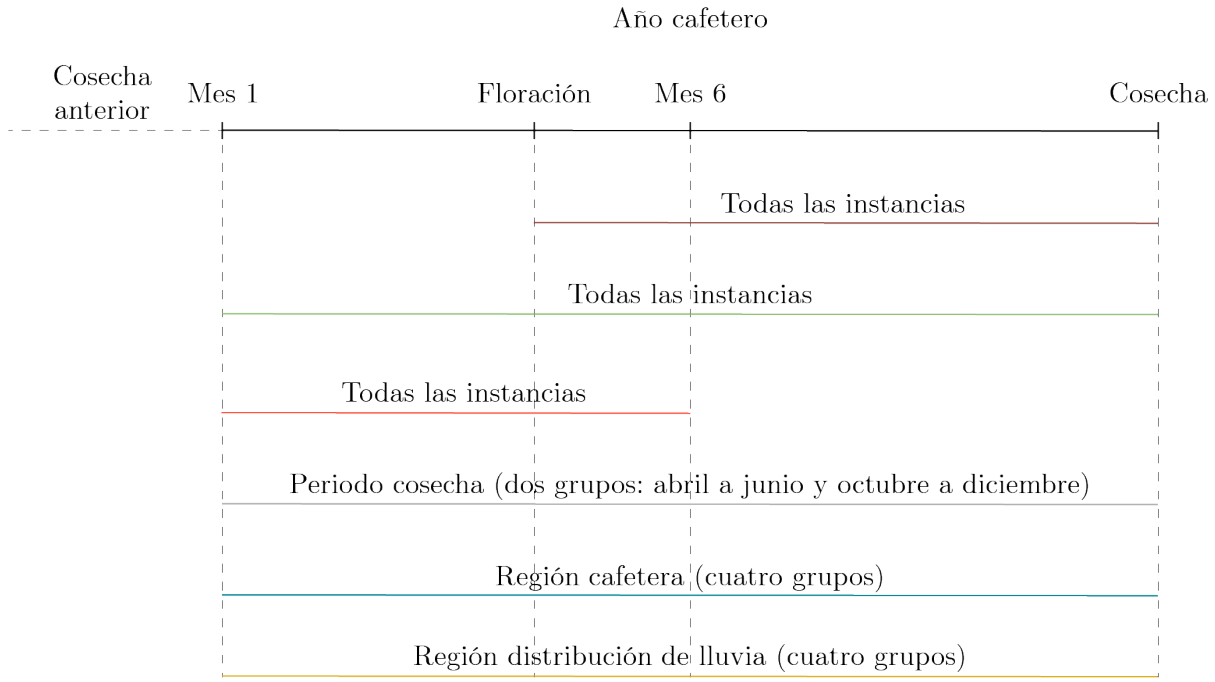


Figura 8. Subconjuntos de datos utilizados en el proceso de modelamiento. Fuente propia.

En la Figura 8 se muestran los subconjuntos de datos (atributos e instancias) utilizados para llevar a cabo el proceso de modelamiento. En cuanto a los atributos, considerando que la floración es el estado fenológico determinante de la producción del café y está estrechamente ligada a condiciones climáticas como la disponibilidad hídrica, la temperatura atmosférica y el brillo solar, entre otras [60], [62], se eligieron las variables meteorológicas que hacen parte de diferentes periodos del año cafetero, siempre teniendo como eje principal el mes en que se presenta dicho estado fenológico en cada municipio.

Dado que la disponibilidad de agua en el suelo durante los meses comprendidos entre la floración y la cosecha pueden llevar a un desarrollo normal del fruto del cafeto [59], una primera evaluación se basó en las variables climáticas de dicho periodo. De igual forma, conociendo que las condiciones previas a la floración pueden también limitar o favorecer la aparición de dicho estado y con esto incidir directamente en el rendimiento del cultivo [60], [64], se utilizaron como atributos las variables climáticas de otros dos periodos: todos los meses que hacen parte del año cafetero y los primeros seis meses del mismo.

Respecto a las instancias, se realizaron experimentos con diferentes grupos, tales como: los registros que pertenecen a un mismo periodo de cosecha (abril a junio u octubre a

diciembre), los que pertenecen a una misma región cafetera o a una misma región de distribución de lluvia.

En primer lugar, se eligieron como variables predictivas el área sembrada, municipio, departamento, altitud, floración, cosecha, mitaca y los datos climáticos de los meses comprendidos entre la floración y la cosecha, con ellos y todas las instancias, se entrenó un modelo para estimar el rendimiento que arrojó los siguientes resultados: CC = 0.6437, MAE = 0.1943, RMSE = 0.2438. Con el objetivo de mejorar la correlación del modelo mencionado, se realizó un proceso de selección de atributos (implementaciones *CfsSubsetEval* y *CorrelationAttributeEval*), que arrojó una estrecha relación entre el área sembrada, el municipio (que a su vez está ligado a las propiedades del suelo), algunos datos climáticos (brillo solar del cuarto mes, precipitación total del quinto mes y temperatura máxima media del noveno mes, específicamente) y el rendimiento del cultivo, atributos con los que se entrenaron un par de nuevos modelos, obteniéndose como mejor resultado un CC = 0.663, un MAE = 0.188 y un RMSE = 0.2348.

Dado que el desempeño de los modelos entrenados no es tan bueno como se esperaba, se consideraron varios aspectos que podrían generar mejores resultados: por una parte, la clase (el rendimiento del cultivo de café) es el resultado de una operación matemática entre otros dos registros del mismo conjunto de datos, la producción y el área cosechada, siendo este último un dato estimado a simple vista por los agricultores, existiendo así la posibilidad de ser poco preciso y trayendo consigo un rendimiento relativo. Por otra parte, el rendimiento agrícola puede entenderse como la producción alcanzada por unidad de superficie cosechada, por lo tanto, si se logra construir un modelo que estime con una alta correlación la producción del cultivo de café, sería insumo amplio y suficiente para el cálculo del rendimiento. Dicho lo anterior, se elige la producción como una segunda clase a ser estudiada en el proceso de modelamiento. Sin embargo, dicha clase presenta una alta varianza y su distribución de probabilidad es asimétrica, por lo que se utilizó una tercera y última clase para el entrenamiento de los nuevos modelos, el logaritmo natural de la producción.

En ese orden de ideas y considerando que los meses anteriores a la floración son determinantes para la misma, en segunda instancia, se entrenaron nuevos modelos utilizando los doce meses del año cafetero para estimar las tres clases objetivo. El mejor resultado se obtuvo utilizando el área sembrada y el clima, con un CC = 0.9404, un MAE = 0.27 y un RMSE = 0.39, para estimar el Ln(Producción). Es necesario aclarar que, tanto en este caso como en los experimentos que serán expuestos posteriormente, se tuvieron resultados ligeramente más precisos (del orden de milésimas) agregando el municipio entre los atributos, pero teniendo en cuenta que se busca generar el mejor modelo para ser llevado en un trabajo futuro a una aplicación real, no se tomaron en cuenta. Agregar el municipio no solamente incrementa el número de variables predictivas, sino que también limita la aplicación del modelo a registros que se obtengan únicamente de los municipios incluidos en el conjunto de datos base.

En tercer lugar, conociendo que el modelo previo presenta una alta correlación, pero que requiere como insumos principales las variables climáticas de doce meses completos para poder estimar el rendimiento (escenario poco adecuado para una aplicación real), se realizaron nuevos experimentos buscando acotar las variables predictivas. Tal como se mencionó en la sección 3.1, tanto los meses inmediatamente anteriores como los posteriores a la floración son determinantes para el rendimiento del cultivo, por lo tanto, se evaluaron nuevos modelos que toman como base el clima de los primeros seis meses del año cafetero. En esta ocasión, se obtuvo un modelo con una correlación semejante al mencionado anteriormente pero que requiere muchos menos atributos, el área sembrada y datos climáticos de seis meses específicamente, para predecir un nuevo registro. Tanto las métricas arrojadas por el modelo (CC = 0.9404, MAE = 0.2711, RMSE = 0.3903) como la posibilidad de generar un valor de producción seis meses antes de la cosecha, lo hacen un candidato importante a ser el modelo final de la presente investigación.

En cuarto lugar, se realizaron pruebas dividiendo el conjunto de datos en diferentes subconjuntos con registros que guardasen relación. Por una parte, se dividió en dos, los datos de los municipios en los que la cosecha se registra entre abril y junio y en los que se da entre octubre y noviembre. Por otra parte, se dividió en cuatro, separando los registros según la región cafetera en la que se ubica geográficamente el municipio. Por último, fue dividido en otros cuatro subconjuntos de acuerdo a la región de distribución de lluvia. Aunque se obtuvieron modelos con una alta correlación (CC $\approx$ 0.95), se obtuvieron otros con una baja (CC $\approx$ 0.79), disminuyendo así la posibilidad de permitir la estimación de la producción del cultivo de café en todo el territorio colombiano.

Teniendo en cuenta que el mejor modelo hasta el momento continúa siendo el compuesto por el área sembrada y el clima de los primeros seis meses del año cafetero, en quinta instancia, se llevó a cabo un proceso de selección de atributos sobre este, aplicando las implementaciones *CfsSubsetEval* y *CorrelationAttributeEval*. Utilizando las variables predictivas elegidas por estos, se construyeron dos modelos que incrementaron levemente la correlación, arrojando un CC = 0.9434 y un CC = 0.9453, respectivamente. Sin embargo, dicha mejora en las métricas de los modelos no compensa el hecho de omitir atributos que son importantes para el dominio de aplicación, así como determinantes a la hora de generar recomendaciones para un manejo adecuado del cultivo. En un último paso, se llevó a cabo un proceso de limpieza de datos perdidos sobre el conjunto de datos candidato, para verificar si existe un cambio en el desempeño del mismo. El nuevo modelo obtuvo como resultado un CC = 0.9418, incremento que no es significativo para la investigación, puesto que afecta la capacidad del modelo para estimar con una alta correlación el rendimiento del cultivo aún en casos donde no existan registros de alguna de las variables predictivas.

Al finalizar el proceso de modelamiento descrito hasta el momento, se obtuvieron un total de 78 modelos. Tanto los atributos e instancias como las métricas de desempeño de cada uno de ellos se presentan en el anexo C. El modelo seleccionado es el entrenado mediante

el algoritmo *multivariate regression prediction model* – implementación *M5P*, que utiliza como atributos el área sembrada y variables climáticas de los primeros seis meses del año cafetero para estimar la producción (el logaritmo natural de la producción, específicamente) del cultivo de café con hasta seis meses de antelación. Las características principales de dicho modelo se presentan a continuación:

- Un CC de 0.9404, que indica un alto grado de correlación positiva entre los datos reales y los generados por el modelo.
- Un MAE de 0.27, que muestra la diferencia promedio entre un valor estimado y uno real.
- Un RMSE de 0.39, indicando la diferencia promedio entre valores reales y estimados por el modelo, teniendo un mayor peso los errores más altos.

Una vez seleccionado el modelo apropiado para la presente investigación, fue utilizado para estimar nuevos valores a partir de datos conocidos, con el objetivo de observar la correlación del mismo en un ambiente real. En este orden de ideas, se estimó la producción de café en municipios de diversos departamentos de Colombia utilizando el área sembrada y datos climáticos de los primeros seis meses del año cafetero, para así comparar los resultados con los registros reales. Las Figuras Figura **9** y Figura **10** muestran la comparación realizada para los municipios del departamento de Antioquia, área en la que existe un alto grado de relación entre los datos reales y los estimados por el modelo (CC = 0.9735).

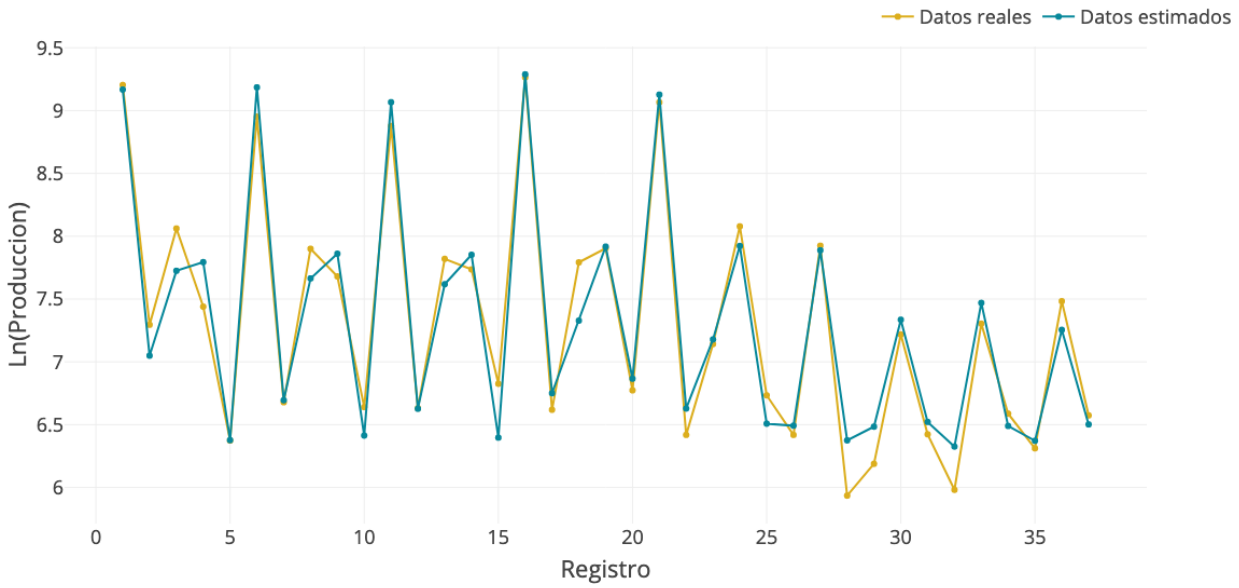


Figura 9. Comparación entre el valor estimado y el valor real de la producción en municipios del departamento de Antioquia. Fuente propia.

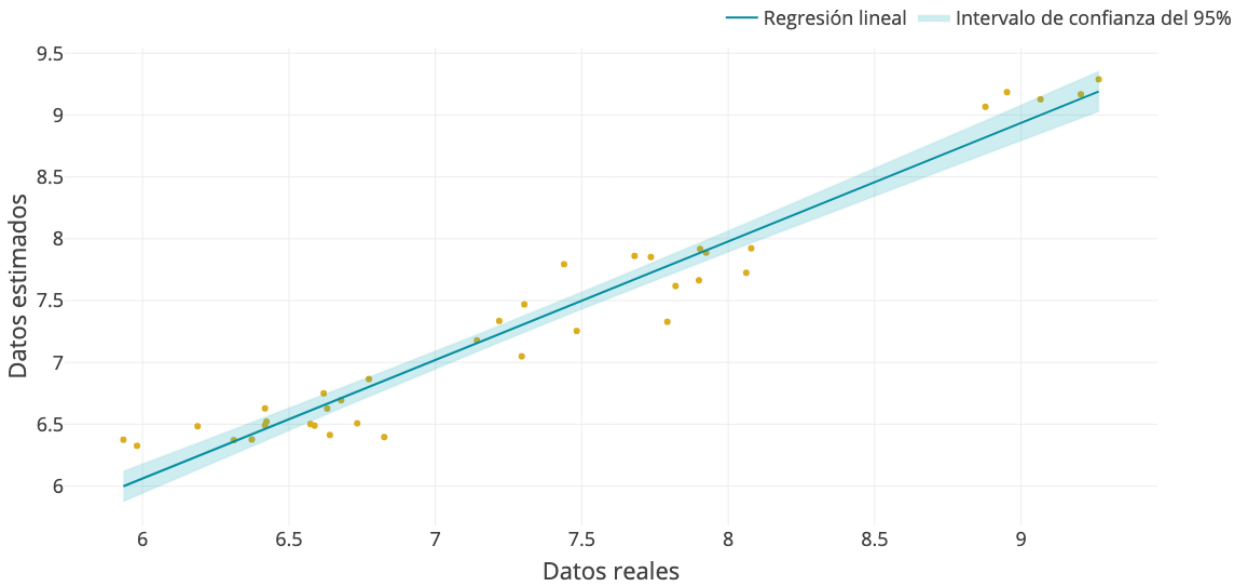


Figura 10. Correlación entre datos reales y estimados de la producción en municipios del departamento de Antioquia. Fuente propia.

En el mismo orden de ideas, con un CC = 0.9646, Quindío es otro de los departamentos evaluados y que registran una alta correlación entre los datos estimados y los reales. Las Figuras Figura **11** y Figura **12** muestran los resultados obtenidos para dicho departamento.

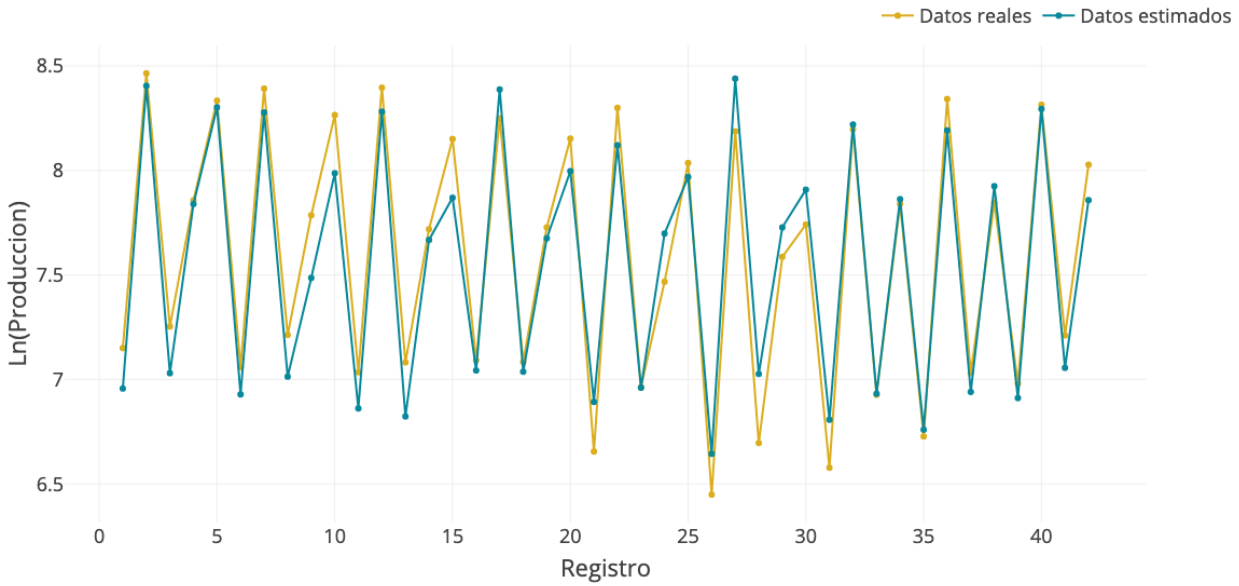


Figura 11. Comparación entre el valor estimado y el valor real de la producción en municipios del departamento de Quindío. Fuente propia.

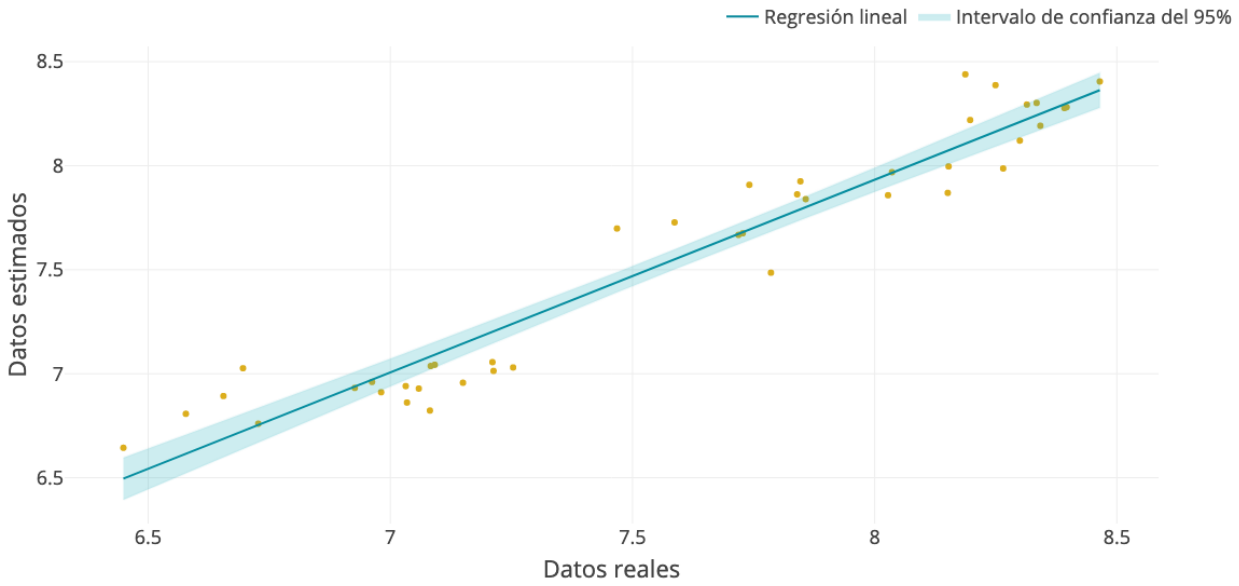


Figura 12. Correlación entre datos reales y estimados de la producción en municipios del departamento de Quindío. Fuente propia.

En ese mismo orden de ideas, se evaluaron registros estimados para los municipios del departamento de Nariño, obteniéndose así un CC = 0.923. La comparación realizada y el análisis de correlación, pueden observarse en las Figuras Figura **13** Figura 14.

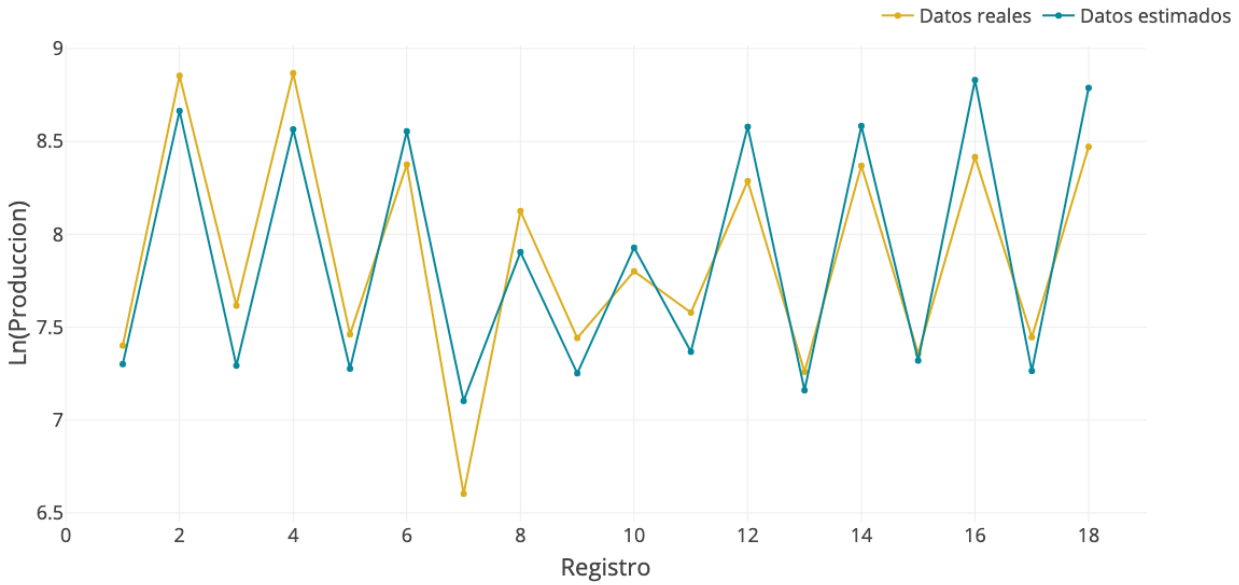


Figura 13. Comparación entre el valor estimado y el valor real de la producción en municipios del departamento de Nariño. Fuente propia.

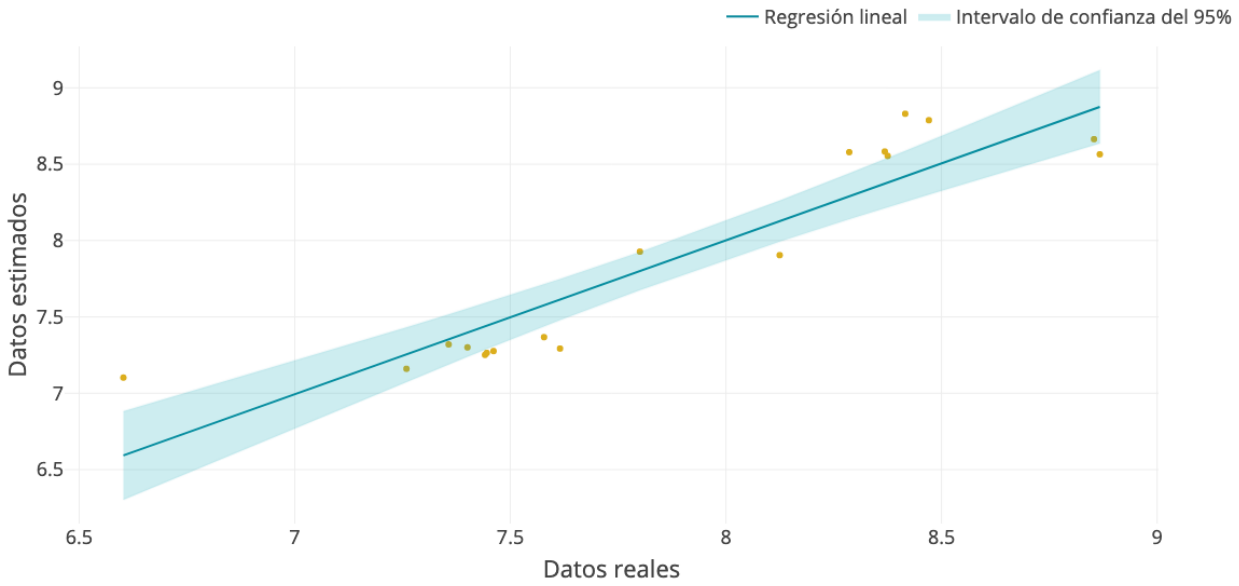


Figura 14. Correlación entre datos reales y estimados de la producción en municipios del departamento de Nariño. Fuente propia.

Finalmente, tal como se muestra en las Figuras Figura **15**Figura 16, se utilizaron cincuenta registros aleatorios para analizar el comportamiento del modelo, obteniéndose como resultado un CC = 0.9842.

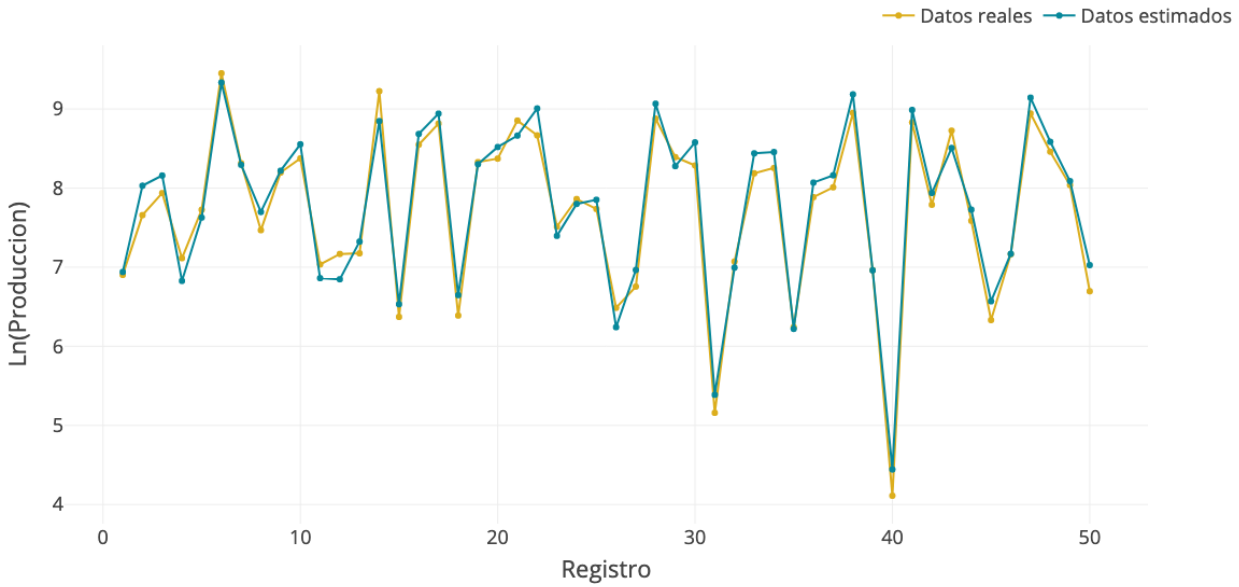


Figura 15. Comparación entre el valor estimado y el valor real de la producción en cincuenta municipios de Colombia. Fuente propia.

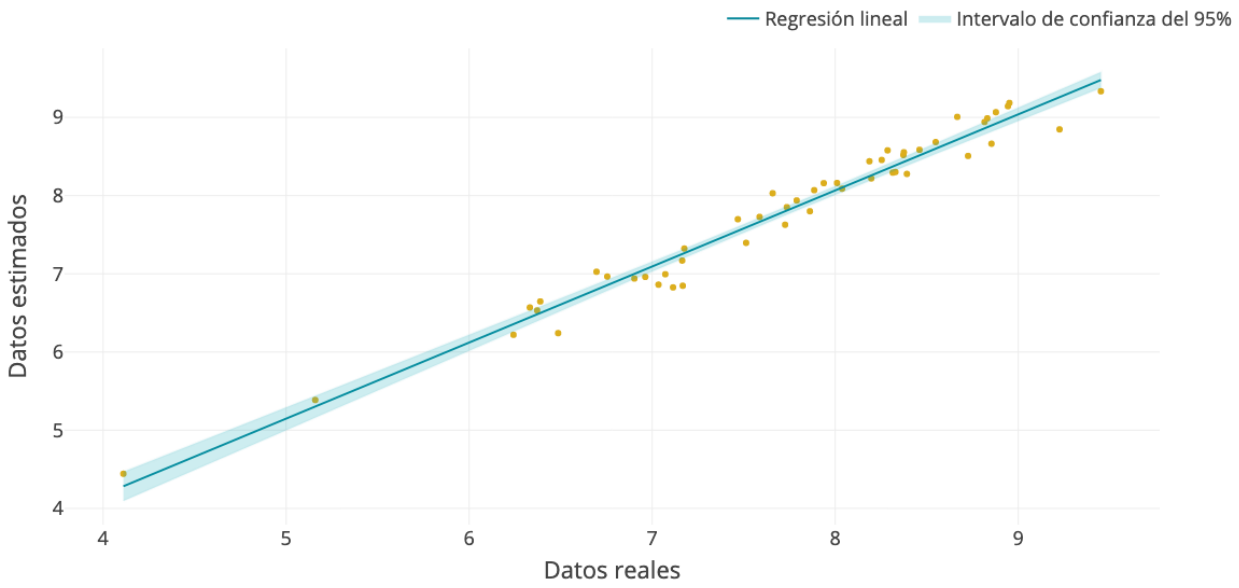


Figura 16. Correlación entre datos reales y estimados de la producción en cincuenta municipios de Colombia. Fuente propia.

En las Figuras Figura **9**Figura 16, se realiza una comparación entre registros reales de producción de café en Colombia y datos estimados mediante el modelo construido en la presente investigación. Los resultados obtenidos muestran la correlación del modelo al estimar la producción y el rendimiento del cultivo de café en Colombia hasta seis meses antes de que se registre la cosecha. Por lo tanto, se propone como una herramienta interesante para conocer la fluctuación de la producción del café colombiano, mediante el uso del área sembrada y registros climáticos de los primeros seis meses del año cafetero como variables predictivas. Cabe destacar que aún cuando el modelo está construido para predecir el logaritmo natural de la producción de café, es el insumo principal para el análisis de la producción y el rendimiento de dicho cultivo a nivel municipal, puesto que el análisis matemático de dicho dato y el área sembrada, permite conocer las mencionadas variables objetivo, tal como se expuso en la sección 3.1.

4.4 Conclusiones acerca del modelo de regresión para la estimación del rendimiento del café

En este capítulo se expuso el proceso de construcción del modelo base para la presente investigación. Como primera medida, se llevó a cabo un análisis del desempeño de seis técnicas de aprendizaje automático frente al conjunto de datos de entrenamiento. Por otra parte, se eligieron tanto los enfoques de entrenamiento y evaluación como los métodos de selección de atributos necesarios para el proceso de modelamiento. Por último, se construyeron y evaluaron setenta y ocho modelos de aprendizaje automático, de los cuales

fue seleccionado uno para la estimación de la producción de café en los municipios cafeteros colombianos. Como resultado de los procesos mencionados, se concluye lo siguiente:

- Aún cuando existen diferentes factores influyentes en el rendimiento del café, se determinó que a partir de pocas variables meteorológicas es posible estimar el rendimiento del cultivo con una correlación mayor al 90%. Así mismo, se encontró que el periodo determinante para alcanzar ese grado de correlación corresponde a los seis primeros meses del año cafetero.
- Se encontró que el algoritmo *multivariate regression prediction model* (implementación M5P de Weka) es el más adecuado para el conjunto de datos utilizado. En ese sentido, al analizar el rendimiento del cultivo de café a partir de registros meteorológicos, el mencionado algoritmo presenta los mejores resultados, por encima de las otras cinco técnicas evaluadas.
- Las condiciones climáticas de los primeros seis meses del año cafetero son las que mayor incidencia tienen en la estimación del rendimiento potencial del café. No obstante, existe una alta correlación entre el municipio y el rendimiento del cultivo, esto refuerza la afirmación de que el suelo juega un papel determinante en las tasas de producción del cultivo de café. Variables como la disponibilidad de nutrientes, materia orgánica, entre otras, determinan en gran medida el rendimiento que se pueda registrar en la cosecha de café.
- Para el caso específico del conjunto de datos base de la presente investigación, la predicción del rendimiento del cultivo de café resulta en una baja correlación, esto debido a una carencia de fiabilidad en dicha variable, puesto que es determinada a partir del área cosechada, medida que nace de fuentes que normalmente registran algunos errores (fotografías, imágenes satelitales o conocimiento empírico del agricultor, entre otras). En contraste, los modelos construidos para estimar la producción presentan una mejor correlación, siendo así una herramienta más confiable para la toma de medidas respecto del manejo del cultivo.
- El conjunto de datos utilizado en este proyecto presentó problemas de heterogeneidad y heterocedasticidad, que fueron abordados en la fase de pre-procesamiento de los datos, mediante la limpieza de los mismos y el uso de la función logaritmo natural para su normalización. Dicha etapa de pre-procesamiento no solamente permitió eliminar extremos y outliers, sino que también trajo consigo un incremento considerable en el desempeño de los modelos construidos. Así mismo, se determinó que la utilización de instancias con valores perdidos puede terminar con resultados benéficos en el proceso de entrenamiento de un modelo de aprendizaje automático preciso.
- Las seis técnicas de aprendizaje automático analizadas presentaron buenos resultados, siendo los algoritmos *multivariate regression prediction model* – implementación *M5P* y *random forest* los que registraron el mejor desempeño

respecto a la predicción de la producción y el rendimiento del cultivo de café. No obstante, esto no excluye la posibilidad de que otros algoritmos o enfoques sean analizados en trabajos futuros. Por ejemplo, el problema de investigación puede ser abordado mediante series de tiempo, con lo cual enfoques como los algoritmos metaheurísticos, aprendizaje profundo o algoritmos genéticos, podrían ser utilizados para estimar el rendimiento del cultivo de café y así generar una comparación con los resultados de la presente investigación.

- El modelo construido permite estimar la producción del cultivo de café con una alta correlación (CC = 0.9404, MAE = 0.27 y RMSE = 0.39). En este orden de ideas, se propone como una herramienta interesante para conocer la fluctuación de la producción del café colombiano con meses de anticipación, mediante el uso del área sembrada y registros climáticos de los primeros seis meses del año cafetero como variables predictivas. Esta información puede ser base para la toma de decisiones de agricultores, cooperativas, federaciones y, en general, todas las personas, empresas y organizaciones involucradas en la cadena de valor del café.

5 Recomendaciones para el manejo del cultivo de café

En el capítulo anterior se expuso la construcción de un modelo basado en aprendizaje automático que permite conocer el rendimiento potencial en diferentes municipios en Colombia teniendo en cuenta sus condiciones climáticas. En el caso de que dicho rendimiento sea bajo, es posible proponer un conjunto de recomendaciones para el manejo del cultivo que permitan mejorar la productividad del mismo. En este orden de ideas, el presente capítulo expone el proceso de construcción de un sistema de recomendaciones basado en contenido que permite conocer diversas sugerencias encaminadas al incremento del rendimiento del cultivo de café a partir de las condiciones climáticas del municipio. De esta manera, el presente capítulo se encuentra dividido en los siguientes apartados:

- **Clasificación del rendimiento potencial**: expone el análisis estadístico llevado a cabo para conocer cuándo un rendimiento potencial puede ser catalogado como bajo, medio o alto, según el comportamiento a nivel nacional.
- **Definición de recomendaciones**: presenta la construcción de un conjunto de recomendaciones para el manejo del cultivo de café a partir de variables climáticas y rangos preseleccionados.
- **Construcción del sistema de recomendaciones**: muestra la construcción de un sistema de recomendaciones basado en contenido que genera sugerencias para mejorar el rendimiento del cultivo de café considerando las condiciones climáticas del municipio.
- **Recomendaciones generales**: expone la generación de un conjunto de recomendaciones generales que, sin importar las condiciones climáticas o el rendimiento potencial, pueden traer consigo el mejoramiento de la productividad del cultivo de café. Así, sirven de complemento al sistema de recomendaciones presentado en la sección anterior y de apoyo a los caficultores del país para conocer buenas prácticas de manejo del cultivo.

5.1 Clasificación del rendimiento potencial

Con el objetivo de determinar el nivel del rendimiento que se espera que se registre en un municipio, tomando como base el promedio nacional, se llevó a cabo un análisis estadístico de cuantiles. Los cuantiles permiten describir aspectos claves de una distribución como lo son la tendencia central y la propagación de los datos alrededor de esta. Cuantil es un término general que incluye cuartiles, deciles y percentiles y que se refiere a un valor de una variable que tiene un porcentaje específico de la distribución debajo de él. Los cuartiles dividen el conjunto de datos en cuatro partes iguales, los deciles en diez partes y los percentiles en cien. En cuanto a los cuartiles, el primero de estos, también conocido como

Q1 o percentil 25, tiene el 25% de los datos por debajo y el 75% de los mismos por encima. El segundo cuartil, conocido como Q2, mediana o percentil 50, es el punto en los datos que los divide en dos partes iguales. El tercer cuartil, también expresado como Q3 o percentil 75, tiene el 75% de los datos bajo él. Por último, el cuarto cuartil o Q4 no es más que el 100% de los datos [91], [92].

Dicho lo anterior, se utilizaron datos de rendimiento de café a nivel municipal en Colombia, registrados durante los años 2007 y 2018, para determinar su distribución y los diferentes cuartiles, permitiendo así clasificar el rendimiento potencial en tres grupos diferentes (bajo, medio y alto). La Figura 17 muestra la distribución de los registros de rendimiento del café en el país y sus respectivos cuartiles.

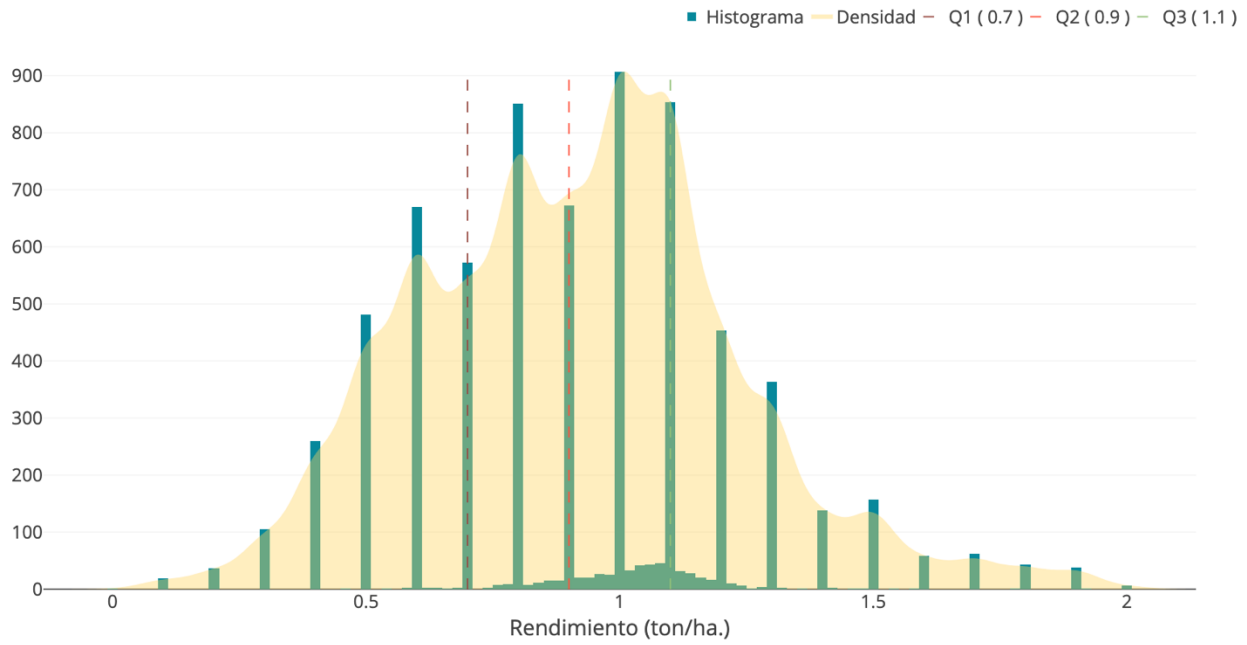


Figura 17. Distribución del rendimiento del café en Colombia durante el 2007 y el 2018. Fuente propia.

A partir de la distribución y los cuartiles expuestos en la Figura 17, se determinaron tres niveles de rendimiento potencial, de la siguiente forma:

- **Bajo**: menor a 0.7 t/ha.
- **Medio**: entre 0.7 t/ha y 1.1 t/ha.
- **Alto**: mayor a 1.1 t/ha.

En caso de que el rendimiento potencial calculado por el modelo expuesto en el capítulo 1 sea bajo, es decir, sea menor a 0.7 t/ha, mediante la ejecución de un sistema de recomendaciones se presentan al usuario una serie de sugerencias, basadas en las condiciones climáticas del municipio, encaminadas a mejorarlo. Tanto el conjunto general

de sugerencias y la construcción del sistema de recomendaciones, se exponen en las secciones 5.2 y 5.3, respectivamente.

5.2 Definición de recomendaciones

Tal como se expuso en la sección 3.1, las condiciones climáticas del sitio del cultivo son determinantes para la producción del café, factores como la temperatura atmosférica, la lluvia y la radiación solar, pueden influir directamente en la floración, el crecimiento vegetativo, el llenado de los frutos y la aparición de enfermedades. En ese orden de ideas, mediante una exploración en la literatura [7], [59], [60], [62]–[64], [93]–[95] y con el acompañamiento de un especialista en el cultivo de café, se analizaron las variables base en la construcción del modelo para determinar los rangos que afectan o favorecen el rendimiento potencial. Así mismo, se construyó un conjunto de recomendaciones de manejo del cultivo que, basadas en diferentes condiciones climáticas, buscan mejorar el rendimiento potencial. A continuación, en la Tabla 5, se exponen las variables seleccionadas y los rangos determinados. Es importante resaltar que a cada condición de las variables seleccionadas le fue asignada una etiqueta, con el objetivo de facilitar el uso de las mismas en la construcción del sistema de recomendaciones.

Variable	Rango	Condición
Temperatura media mensual	$T < 13\ ^{o}C$	T1: mínima crítica
	$13\ ^{o}C \geq T > 19\ ^{o}C$ $21\ ^{o}C > T \geq 32\ ^{o}C$	T2: normal
	$19\ ^{o}C \geq T \geq 21\ ^{o}C$	T3: óptima
	$T > 32\ ^{o}C$	T4: máxima crítica
Precipitación total anual	$P < 1400$ mm	P1: déficit
	1400 mm $\geq P \geq 2900$ mm	P2: normal
	$P > 2900$ mm	P3: exceso
Cantidad anual de horas de brillo solar	$H < 1300$ horas	H1: muy baja
	1300 horas $\geq H \geq 1500$ horas	H2: baja
	1500 horas $> H \geq 1700$ horas	H3: media
	1700 horas $> H \geq 1900$ horas	H4: alta
	$H > 1900$ horas	H5: muy alta

Tabla 5. Rangos de las variables seleccionadas que afectan o favorecen el rendimiento potencial del café. Adaptado de [7], [59], [93].

Utilizando los rangos expuestos en la Tabla 5, se llevó a cabo la construcción de un conjunto inicial de veinte (20) recomendaciones de manejo del cultivo de café que tienen como objetivo la mejora del rendimiento potencial. En concreto, se analizaron diferentes condiciones climáticas que pueden afectar el rendimiento potencial del café y se plantearon una serie de recomendaciones del manejo del cultivo, basadas en la literatura mencionada anteriormente y el conocimiento de un especialista en el cultivo de café, que permitan hacerle frente a las mismas. En la Tabla 6, se presentan algunas de las condiciones climáticas analizadas y el conjunto de recomendaciones construido para las mismas (las veinte recomendaciones construidas pueden observarse en el anexo D).

Condiciones climáticas	Recomendaciones
T1, P1, H1	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta
T1, P1, H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío
T1, P3, H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros
T1, P3, H2	Se recomienda establecer un cultivo a libre exposición y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros
T1, P3, H4	Se recomienda establecer cultivos a libre exposición e implementar bajas densidades de siembra para evitar una posible proliferación de enfermedades
T1, P3, H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua
T4, P1, H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y

	poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta
T4, P1, H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta
T4, P3, H4	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo
T4, P3, H5	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo y reducir la distancia de siembra para evitar proliferación de entes patógenos

Tabla 6. Conjunto inicial de recomendaciones para el mejoramiento del rendimiento potencial del café. Fuente propia.

Es importante resaltar que el conjunto de recomendaciones expuesto en la Tabla 6, es una propuesta inicial para la construcción y puesta en marcha del sistema de recomendaciones expuesto en la sección 5.3. No obstante, la ampliación y mejoramiento del mismo se proponen como un trabajo futuro interesante.

5.3 Construcción del sistema de recomendaciones

Los sistemas de recomendaciones (RS, *recommender systems*) son herramientas software que usan diferentes tecnologías analíticas con el objetivo de sugerir ítems de interés para usuarios específicos. Dichas sugerencias se relacionan normalmente con diferentes procesos de toma de decisiones, como a qué evento asistir, cuales hábitos saludables seguir, qué película ver, qué lugares visitar, entre otros [96], [97]. Una de las clasificaciones más comunes de sistemas de recomendaciones está dividida en cuatro, de la siguiente forma: basados en contenido, filtrado colaborativo, basados en conocimiento y sistemas híbridos, tal como se exponen a continuación:

- **RS basados en contenido**: sugieren diferentes ítems a partir del contenido de los mismos y los intereses de cada uno de los usuarios. Así, la similitud de las sugerencias es calculada en función de sus características y las de los usuarios [96], [97].

- **RS de filtrado colaborativo**: entregan recomendaciones basados en las calificaciones hechas con anterioridad por usuarios con intereses similares. En este sentido, la similitud de los gustos de dos usuarios es calculada en función de su historial de calificaciones [96], [97].
- **RS basados en conocimiento**: entregan recomendaciones a partir de conocimiento en un campo específico, entendiendo con antelación cómo ciertas características de un ítem satisfacen las necesidades y preferencias de los usuarios. En concreto, pueden definir qué tan útil es una sugerencia para un usuario, tomando como base conocimiento adquirido con anterioridad [97].
- **RS híbridos**: combinan una o varias de las técnicas anteriores con el objetivo de entregar recomendaciones más precisas, siendo así el tipo de sistemas de recomendaciones más utilizado en el campo empresarial. Un RS híbrido combina dos técnicas, buscando siempre utilizar las ventajas de una para corregir o mejorar las desventajas de la otra [96], [97].

Con el objetivo de llevar a cabo su función principal, la de identificar sugerencias útiles para un usuario específico, un RS debe predecir qué tan exacta es la recomendación que espera entregar. En ese sentido, el sistema de recomendaciones compara la pertinencia de los diferentes ítems y luego decide cuál sugerir. La similitud entre la sugerencia y los usuarios puede ser abordada mediante diferentes técnicas. A continuación, se exponen las utilizadas en la presente investigación [98]:

- Una de las más comunes medidas de similitud es la **distancia euclidiana**:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

  Donde n es el número de atributos o dimensiones y x_k y y_k son los atributos número k de los objetos x y y, respectivamente.

- La **distancia de Mahalanobis** se define como:

$$d(x,y) = \sqrt{(x-y)\,\sigma^{-1}\,(x-y)^T}$$

  Donde σ hace referencia a la matriz de covarianza de los datos.

- La **similitud coseno** es otro enfoque muy común, en el que se consideran los ítems como vectores n-dimensionales y se calcula su similitud como el coseno del ángulo que forman. En ese sentido, la similitud coseno se define como:

$$\cos(x,y) = \frac{(x \cdot y)}{\|x\|\|y\|}$$

Donde $\cdot$ se refiere al producto escalar y $\|x\|$ y $\|y\|$ son la norma de los vectores x y y, respectivamente.

Teniendo en cuenta que se busca la construcción de un sistema de recomendaciones que permita generar sugerencias para mejorar el rendimiento potencial en los diferentes municipios analizados, aún cuando las condiciones climáticas del mismo no coincidan totalmente con las listadas en la Tabla 6, se eligen los basados en contenido como el tipo de RS a construir. El proceso básico de un RS basado en contenido consiste en relacionar las preferencias e intereses de un usuario con los atributos de un ítem, obteniendo como resultado el nivel de similitud, es decir, el nivel de interés del usuario en ese ítem. Si la correlación entre las características del usuario y las del ítem es alta, es entonces una sugerencia que se puede generar [99]. La aplicación de RS basados en contenidos es amplia, va desde su uso en buscadores web para analizar el comportamiento de los usuarios y recomendarles ítems de su interés [100], hasta su aplicación para la generación de programas de vida saludable para pacientes con enfermedades crónicas [101] o recomendaciones de cursos en línea [102]. En cuanto al campo de la agricultura, algunas de sus aplicaciones incluyen sistemas para la elección de los cultivos adecuados para diferentes áreas [103] o las mejores fechas de cultivo para obtener buenos resultados en la etapa de producción [104] y la generación de recomendaciones de libros sobre manejo del cultivo para los agricultores [105], entre otros.

En este orden de ideas, tomando como base la teoría de los RS basados en contenido, los municipios para los que se buscan generar las sugerencias pueden ser vistos como "usuarios" del sistema de recomendaciones que tienen un perfil compuesto por las condiciones climáticas del mismo, que a su vez se relaciona con las características para las que fueron construidas cada una de las sugerencias. Este enfoque propone las siguientes ventajas [99]:

- **Independencia de históricos**: la generación de sugerencias puede hacerse a través de las condiciones climáticas del municipio, siendo innecesario un histórico de recomendaciones que hayan sido entregadas para municipios con perfiles similares, tal como es requisito en los RS de filtrado colaborativo.
- **Transparencia**: es posible permitir que los usuarios conozcan el nivel de similitud que existe entre la sugerencia y el municipio de interés, así como las condiciones que llevaron a que un elemento haya sido elegido como la recomendación final. Por el contrario, otros tipos de RS funcionan como cajas negras, donde solo es posible explicar que la recomendación se da basada en sugerencias entregadas anteriormente a municipios con un perfil similar.
- **Crecimiento del conjunto de recomendaciones**: El RS basado en contenido permite recomendar ítems que no han sido sugeridos antes. De este modo, no se presenta un problema común en otros tipos de RS, que son incapaces de recomendar nuevos ítems hasta tanto se tenga una calificación histórica de los mismos.

Tal como se mencionó anteriormente, el RS propuesto en la presente investigación busca generar recomendaciones para los usuarios comparando el perfil del municipio de interés con los atributos de los ítems que hacen parte del conjunto de sugerencias expuesto en la Tabla 6. En ese sentido, se calcula el nivel de similitud entre el municipio y las sugerencias, resultando en una lista clasificada de recomendaciones potencialmente interesantes. Por último, se entrega al usuario la recomendación que tiene el mayor nivel de similitud con el perfil del municipio. A continuación, se expone el pseudocódigo del sistema de recomendaciones construido:

Algoritmo sistema de recomendaciones

```
1: Procedimiento GenerarRecomendacion(condicionesClimaticas)
2:     #Obtener el csv que tiene el conjunto de recomendaciones construido
3:     recomendaciones ← leerCSV("recomendacionesRS.csv")
4:     #Calcular el nivel de similitud de las condiciones climáticas del municipio analizado con
5:     #las características del conjunto de recomendaciones
6:     matrizRecomendaciones ← codificacionOneHot(variables en recomendaciones)
7:     similitudMatriz ← calcularSimilitud(matrizRecomendaciones)
8:     similitudMunicipio ← similitudMatriz[condicionesClimaticas]
9:     similitudMunicipio ← organizarOrdenDescendente(similitudMunicipio)
10:    #Obtener la recomendación con el nivel más alto de similitud
11:    porcentajeSimilitud ← 0
12:    textoRecomendacion ← nulo
13:    Para similitud en similitudMunicipio hacer
14:        indiceRecomendacion ← similitud[0]
15:        textoRecomendacion ← recomendaciones[indiceRecomendacion]
16:        Si textoRecomendacion no está vació
17:            porcentajeSimilitud ← similitud[1]
18:            Fin Para
19:        Fin Si
20:    Fin Para
21:    retorna porcentajeSimilitud y textoRecomendacion
22: Fin procedimiento
```

El sistema de recomendaciones recibe como entrada las condiciones de las tres variables climáticas (temperatura media mensual, precipitación total anual y cantidad anual de horas de brillo solar) durante el último año en el municipio para el que se desea generar la recomendación. El procedimiento inicia con la creación de la variable "recomendaciones", a la cual le es asignada la lectura del archivo donde se encuentra el conjunto de sugerencias construido y que está compuesto por dos columnas; la primera, llamada "variables", contiene todas las combinaciones posibles de condiciones climáticas; la segunda, nombrada "recomendación", está compuesta por las sugerencias creadas para unas condiciones

específicas, tal como se muestra en la Tabla 6. Como paso siguiente, mediante la codificación One-Hot, se convierte la columna "variables" en doce diferentes, que hacen referencia a las condiciones expuestas en la Tabla 5, donde su valor equivale a uno cuando la condición existe y cero cuando no. Dicha transformación se realiza con el objetivo de facilitar el cálculo de la similitud, puesto que no es posible realizarla sobre variables categóricas. En el siguiente paso, se calcula la similitud entre cada una de las sugerencias y las diferentes combinaciones de condiciones climáticas, utilizando tres medidas diferentes, similitud coseno, distancia euclidiana y distancia Mahalanobis. Con ello, de forma seguida, se elige la recomendación que tiene el mayor nivel de similitud con las condiciones del municipio de interés y es entregada como salida del RS junto con su respectivo porcentaje de similitud.

La evaluación de la precisión de las diferentes medidas de similitud utilizadas (similitud coseno, distancia euclidiana y distancia Mahalanobis), se llevó a cabo mediante un estudio de usuario, involucrando a un especialista en el cultivo de café. En ese sentido, el especialista evaluó el porcentaje de similitud existente entre veinte sugerencias y las respectivas condiciones climáticas para las que fueron generadas. Finalmente, se compararon los porcentajes de similitud determinados por el sistema de recomendaciones y el usuario, para calcular el MAE y el RMSE del sistema construido con cada una de las medidas de similitud. En el anexo D se exponen las condiciones climáticas, recomendaciones y porcentajes de similitud mencionados. Además, en la **Tabla 7** se presentan los resultados de la evaluación de las tres medidas de similitud.

Similitud coseno		**Distancia euclidiana**		**Distancia Mahalanobis**	
MAE	RMSE	MAE	RMSE	MAE	RMSE
29.583	39.3083	37.932	48.3474	29.61	39.6556

Tabla 7. MAE y RMSE de las medidas de similitud utilizadas en la construcción del sistema de recomendaciones. Fuente propia.

Tomando como base los resultados presentados en la Tabla 7, fue seleccionada la similitud coseno como la medida de similitud para el sistema de recomendaciones final. No obstante, la distancia Mahalanobis presentó un desempeño similar. Igualmente, todas las medidas de similitud analizadas presentan errores altos, lo que puede explicarse debido al bajo número de sugerencias iniciales con las que fue construido el sistema de recomendaciones. En este orden de ideas, se propone como trabajo futuro el incremento del número de sugerencias, así como el mejoramiento de las mismas al disminuir su nivel de generalidad.

Finalmente, el sistema de recomendaciones propuesto se convierte en módulo esencial del prototipo expuesto en el capítulo 6 para la generación de recomendaciones encaminadas a mejorar el rendimiento potencial en diferentes municipios de Colombia. Sin embargo,

conociendo los resultados de la evaluación y teniendo en cuenta que existe la posibilidad de encontrarse municipios para los que el RS sería incapaz de generar recomendaciones que guarden un alto nivel de similitud, se construyeron diversas sugerencias generales de manejo del cultivo que pueden servir de apoyo a los caficultores para, directa o indirectamente, mejorar el rendimiento de su cultivo, tal como se expone en la sección 5.4.

5.4 Recomendaciones generales

Ante la posibilidad de que el sistema de recomendaciones propuesto en la sección 5.3 genere sugerencias con un nivel de similitud bajo para municipios con condiciones climáticas totalmente desconocidas para él, cobra relevancia la construcción de un conjunto de recomendaciones generales que, aún cuando no estén construidas para un perfil de municipio específico, puedan influir positivamente en el rendimiento potencial del cultivo de café.

Tal como se expuso en la sección 3.1, el rendimiento potencial del cultivo de café no se ve afectado o favorecido únicamente por las condiciones climáticas del sitio. La eficiencia del proceso productivo se puede ver influenciado también por las prácticas de manejo relacionadas con el control de arvenses, las plagas y enfermedades, así como por el suministro de los nutrientes esenciales. Por último, las buenas prácticas de cosecha y beneficio complementan el proceso productivo y llevan a mejoras en la producción obtenida [59]. En este orden de ideas, mediante una exploración en la literatura y con el acompañamiento de un especialista, se construyó un conjunto de cincuenta y siete (57) recomendaciones generales que pueden influir en el rendimiento potencial del cultivo de café para el incremento del mismo. A continuación, en la Tabla 8, se presentan algunas de las sugerencias construidas, mientras que en el anexo E se expone la totalidad de las mismas.

Recomendación	Tema
Para que el desarrollo del fruto del cafeto sea normal se requiere disponibilidad de agua en el suelo durante los ocho meses comprendidos entre la floración y la cosecha, con un período crítico entre las semanas 8 y 16, en el cual se define el tamaño del fruto. Por esto, se recomienda implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua.	Suelo
Las arvenses ejercen una cobertura que protege el suelo de los impactos directos de las gotas de lluvia, disminuyendo la erosión superficial, por lo cual se deben realizar cortes altos con machete o guadaña. Nunca se debe hacer uso del azadón, debido a que se descubre el suelo y promueve la erosión.	Arvenses

De manera permanente, sea durante los meses secos o los húmedos, es necesaria la revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros.	Fitosanitario
La aplicación de materia orgánica (compost, bocashi, vermicompost o lombrinaza, biofertilizantes, biofermentos y abono verde) favorece las condiciones de calidad del suelo y reduce los requerimientos de fertilizantes inorgánicos, lo que puede generar un mejor rendimiento del cultivo y reducir los impactos negativos sobre el agroecosistema.	Nutrición
Aumentar la materia orgánica, obras de conservación y cosecha de aguas en la finca, aplicaciones foliares más frecuentes durante sequías y sistemas de riego, pueden traer grandes beneficios cuando las lluvias en la zona son irregulares.	Generalidades

Tabla 8. Conjunto de recomendaciones generales para el manejo del cultivo de café. Adaptado de [9], [59], [113], [63], [106]–[112].

5.5 Conclusiones acerca de las recomendaciones para el manejo del cultivo de café

En el presente capítulo se expuso el proceso de construcción de un sistema de recomendaciones basado en contenido que entrega sugerencias para favorecer el rendimiento potencial del cultivo de café. Dicho sistema toma como base las condiciones climáticas del último año del municipio de interés y, mediante la similitud coseno, encuentra la sugerencia adecuada. Así mismo, se construyó un conjunto de recomendaciones generales para el manejo del cultivo de café que sirven como complemento al sistema de recomendaciones mencionado. En este orden de ideas, a partir de los procesos y resultados expuestos en el presente capítulo, se concluye:

- El rendimiento potencial del cultivo de café es influido por una gran cantidad de variables climatológicas, edáficas, fenológicas, entre otras. En el caso de la presente investigación, pudieron establecerse diversos rangos en cuanto a la temperatura, precipitación y horas de brillo solar. No obstante, tanto dichos rangos como nuevas variables pueden analizarse con mayor profundidad, lo que abre las puertas a sugerencias más precisas y un sistema de recomendaciones más complejo.
- Las recomendaciones expuestas en el presente capítulo constituyen una propuesta inicial, construida a partir de literatura disponible y la guía de un especialista, que puede apoyar la estabilidad de las tasas de rendimiento del cultivo de café y con esto contribuir a mejorar la calidad de vida de los caficultores. No obstante, el sistema de recomendaciones construido presentó un RMSE cercano al 40%, lo que abre una ventana de trabajo para el planteamiento de mayores y más precisas sugerencias.

- El enfoque basado en contenido permitió construir un sistema de recomendaciones que trata los municipios como "usuarios" a los que se les entrega sugerencias a partir de su perfil. En ese sentido, es posible utilizar las condiciones climáticas del municipio para generar recomendaciones que permitan mejorar el rendimiento potencial del cultivo de café. Si bien el sistema de recomendaciones presenta el funcionamiento esperado, se propone como un trabajo futuro el desarrollo de un sistema de recomendaciones híbrido, que pueda realimentarse de los resultados obtenidos por los caficultores al implementar las prácticas sugeridas por el sistema.
- La utilización de la similitud coseno trajo consigo mejores resultados que las otras medidas analizadas, distancia euclidiana y distancia Mahalanobis. Sin embargo, se hace interesante la evaluación de otras medidas de similitud y/o correlación, como la correlación de Pearson o la distancia Minkowski.
- El sistema de recomendaciones basado en contenido propone varias ventajas. Por una parte, hace que sea innecesario un histórico de recomendaciones que hayan sido entregadas para municipios con perfiles similares, tal como sucede con otros tipos de sistemas. Por otra parte, permite que los usuarios conozcan el nivel de similitud que existe entre la sugerencia y el municipio de interés, así como las condiciones que llevaron a que un elemento haya sido elegido como la recomendación final. Por último, tiene la posibilidad de recomendar ítems que no han sido sugeridos antes, por lo que el conjunto de sugerencias puede ser mejorado en cualquier momento.
- Si bien los sistemas de recomendaciones basados en contenidos traen consigo varias ventajas, también tienen algunas debilidades. En primer lugar, las recomendaciones dependen de qué tanto conocimiento se tenga del perfil de los municipios, siendo crucial la existencia de una correlación (por mínima que sea) entre el perfil de un municipio y los atributos de una sugerencia. En segundo lugar, con el tiempo el sistema puede producir recomendaciones poco novedosas que ya hayan sido presentadas a los agricultores. En último lugar, dado que el conjunto de sugerencias está compuesto por un número limitado de ítems, es probable que el sistema no sepa cuál recomendar para un perfil de municipio totalmente nuevo, es decir, en caso de que las características del municipio no guarden relación alguna con los atributos de las sugerencias. En este sentido, se propone como trabajo futuro la utilización de otros enfoques de sistemas de recomendaciones, como filtrado colaborativo o híbridos.

6 Prototipo y experimentación

En el capítulo 1 se expuso el entrenamiento de un modelo de aprendizaje automático que permite predecir el rendimiento del cultivo de café en municipios de Colombia. Por su parte, en el capítulo 5 se presentó la construcción de un sistema de recomendaciones encaminadas a mejorar el rendimiento potencial del cultivo de café, en el caso de que el mismo pueda ser bajo. En este orden de ideas, en el presente capítulo se expone el proceso de diseño, implementación y evaluación de un prototipo que hace uso del modelo de aprendizaje automático y el sistema de recomendaciones mencionado, para servir de herramienta de apoyo a los agricultores colombianos para la toma de decisiones en búsqueda de la estabilidad o mejora de las tasas de rendimiento y producción de sus cultivos de café.

El capítulo se encuentra dividido en los siguientes apartados:

- **Desarrollo del prototipo**: se exponen las funcionalidades, arquitectura y diseño del prototipo propuesto.
- **Experimentación**: se describen todos los pasos llevados a cabo para la realización de una encuesta a especialistas en ciencias de la computación, ambientales y agrícolas para la evaluación de la pertinencia y diseño del prototipo propuesto en el numeral anterior.
- **Resultados**: en este último apartado se analizan los resultados obtenidos a través de la encuesta llevada a cabo.

6.1 Desarrollo del prototipo

Utilizando el algoritmo *multivariate regression prediction model* de aprendizaje automático, en el capítulo 1 se construyó un modelo que permite predecir el rendimiento potencial del cultivo de café en diferentes municipios colombianos. Para ello, toma como base el área sembrada y las condiciones climáticas de los primeros seis meses del año cafetero inmediatamente anterior. Posteriormente, el capítulo 5 se construyó un sistema de recomendaciones basado en contenido que permite generar sugerencias para enfrentar las condiciones climáticas recientes de los municipios y con ello buscar el incremento del rendimiento potencial del cultivo de café. Como resultado, utilizando tanto el modelo de aprendizaje automático como el sistema de recomendaciones, se diseñó y construyó un prototipo que busca orientar a los caficultores colombianos para que implementen buenas prácticas de cultivo y busquen con ello obtener las tasas de rendimiento potencial esperadas.

En las siguientes secciones se presentan las funcionalidades, arquitectura e interfaz del prototipo propuesto.

6.1.1 Funcionalidades y arquitectura

El prototipo propuesto busca ser una herramienta de apoyo para las personas involucradas en la cadena productiva del café en Colombia. Por medio del prototipo se puede estimar el rendimiento potencial del cultivo en diferentes zonas colombianas y su comportamiento con relación al promedio nacional; en caso de que el rendimiento potencial sea bajo, permite el acceso a una serie de recomendaciones específicas para la zona que, basadas en las condiciones climáticas, están encaminadas a mejorarlo. Sumado a esto, es posible observar en él diferentes recomendaciones generales que pueden ayudar a mejorar la productividad del cultivo y una serie de preguntas frecuentes sobre el manejo del mismo. Sintetizando, las funcionalidades propuestas en el prototipo son las siguientes:

- **Estimación del rendimiento potencial**: mediante el modelo de aprendizaje automático construido en el capítulo 1, se estima la producción esperada y el rendimiento potencial del cultivo de café en la siguiente cosecha en Colombia. Dichos datos pueden ser generados para diferentes áreas geográficas, como lo son municipios, departamentos o zonas cafeteras colombianas.
- **Visualización de históricos**: se exponen gráficas de los históricos de producción y rendimiento registrados en la zona de interés del usuario (municipio, departamento o zona cafetera).
- **Generación de recomendaciones específicas**: en caso de que el rendimiento potencial calculado para la zona de interés del usuario sea bajo, a partir del sistema de recomendaciones expuesto en el capítulo 5, se generan sugerencias basadas en las condiciones climáticas del sitio para buscar la mejora de las tasas de producción y rendimiento.
- **Acceso a recomendaciones generales de manejo del cultivo**: permite al usuario el acceso a una serie de recomendaciones generales en temas como nutrición, fitosanitario, arvenses y suelo, encaminadas al incremento del rendimiento potencial del cultivo de café.
- **Preguntas frecuentes**: se expone un conjunto de preguntas frecuentes acerca del cultivo de café que pueden tener las personas involucradas en la cadena productiva del café, tales como: ¿cómo almacenar el fruto?, ¿qué variedad plantar?, ¿cuándo y cómo revisar la existencia de plagas y enfermedades?, entre otras.

De acuerdo a las funcionalidades expuesta anteriormente, se plantea el siguiente diagrama de alto nivel de la arquitectura del prototipo.

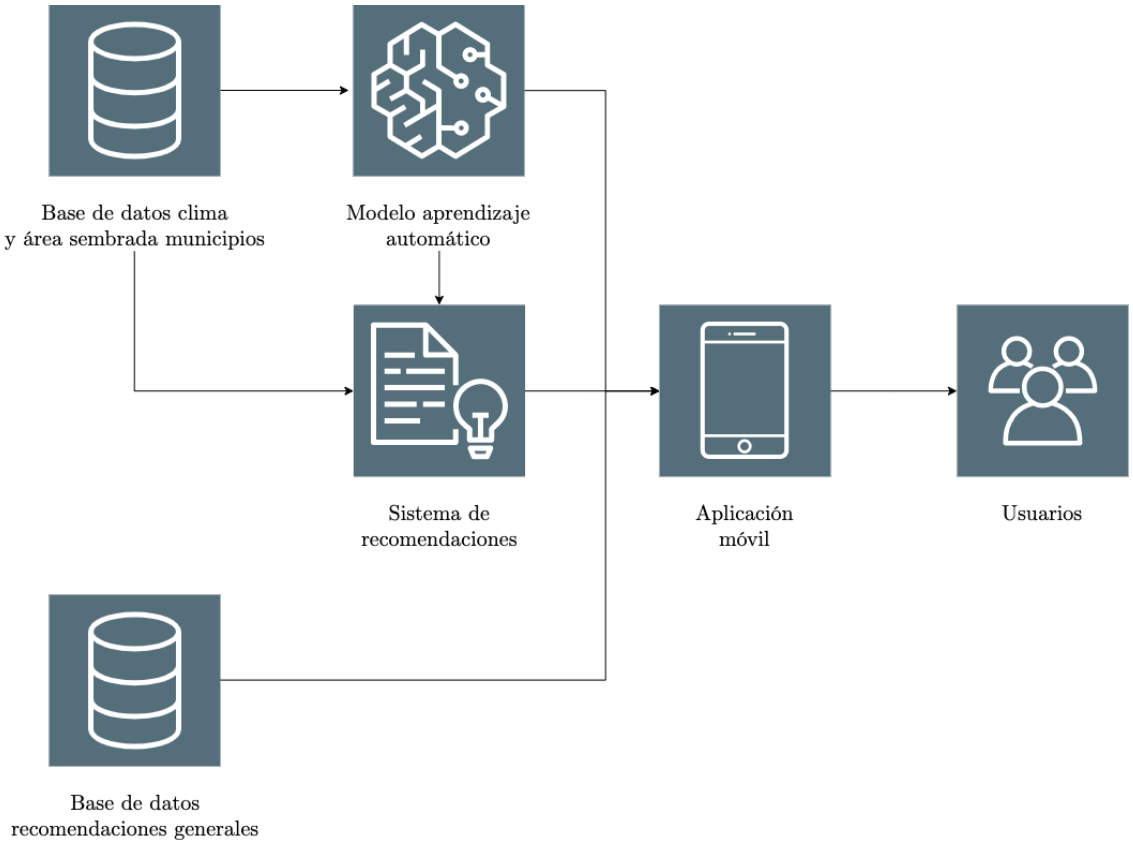


Figura 18. Diagrama de alto nivel de la arquitectura del prototipo. Fuente propia.

La Figura 18 muestra el diagrama de alto nivel de la arquitectura del prototipo, la cual permite la prestación de las funcionalidades mencionadas anteriormente. En dicha arquitectura se encuentran los usuarios que interactúan con la aplicación móvil y la información disponible en la misma. En la aplicación, puede observarse el rendimiento potencial del cultivo de café, estimado por el modelo de aprendizaje automático que a su vez toma como insumo principal registros históricos de clima y área sembrada del cultivo. De manera semejante, dado el escenario en que el rendimiento potencial calculado por el modelo sea bajo, el usuario puede ver una serie de sugerencias específicas para su sitio de interés, generadas por un sistema de recomendaciones basado en contenido. De igual forma, los usuarios tienen acceso a una base de datos de recomendaciones generales sobre el manejo del cultivo de café, que pueden ayudarles a mantener o mejorar sus tasas de rendimiento.

Con el objetivo de profundizar sobre la arquitectura y conocer el comportamiento de los componentes de la misma, en la Figura 19 se expone la vista lógica del prototipo. Dicha vista describe y modela las partes que componen el sistema, además explica cómo interactúan estas entre sí. Para ello, se divide en tres capas: aplicación (componentes que implementan las funcionalidades del prototipo), mediación (elementos que permiten la

comunicación entre los componentes de las otras dos capas) y almacenamiento (componentes básicos de almacenamiento del sistema) [114].

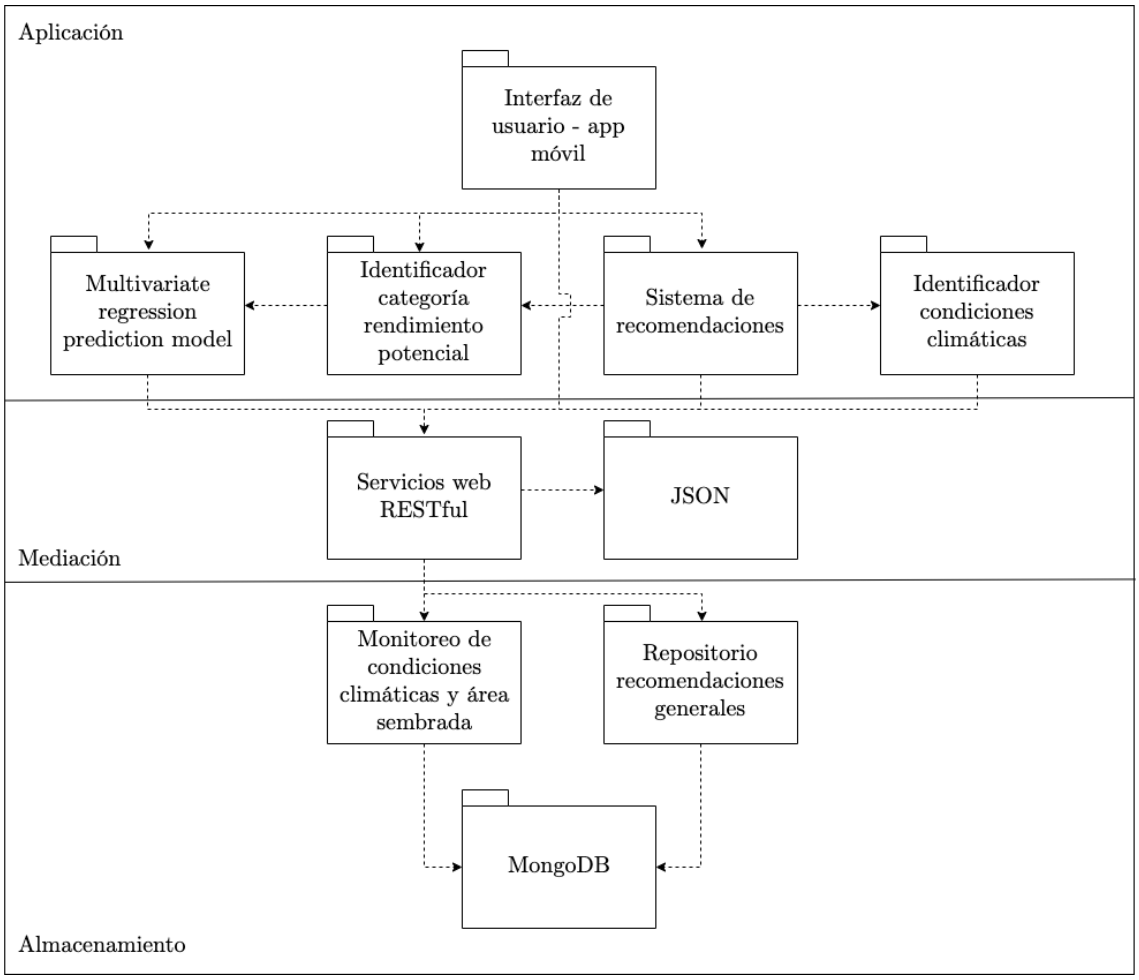


Figura 19. Vista lógica del prototipo. Fuente propia.

A continuación, se describe cada uno de los componentes de la vista lógica presentada en la Figura 19:

- **MongoDB**: es un sistema de gestión de bases de datos NoSQL que se caracteriza por el almacenamiento de los datos en documentos de tipo JSON. Algunas de sus ventajas incluyen uso de pocos recursos y fácil escalabilidad horizontal. Sobre MongoDB están soportados los dos componentes siguientes.
- **Monitoreo de condiciones climáticas y área sembrada**: en una base de datos MongoDB se guardan registros diarios de las variables climáticas que requiere el modelo *multivariate regression prediction model* (temperatura, humedad, precipitación y brillo solar). Así mismo, se guardan los históricos de área sembrada

de café en municipios de Colombia. Todos estos datos son utilizados por el modelo para realizar la estimación del rendimiento potencial en los municipios.

- **Repositorio de recomendaciones generales**: en una base de datos MongoDB se guarda y actualiza el conjunto de recomendaciones generales presentado en la Tabla 8 que, aún cuando no están construidas para unas condiciones climáticas específicas, son presentadas a los caficultores en la búsqueda de mejorar las tasas de rendimiento de sus cultivos.
- **Servicios web RESTful**: es un conjunto de servicios web que se encargan del acceso a las bases de datos mencionadas anteriormente. Mediante dichos servicios se consultan los datos climáticos, los registros de área sembrada y las recomendaciones generales cuando el usuario interactúa con la aplicación.
- **JSON**: es el formato utilizado por los servicios web para el intercambio de todos los datos.
- **Multivariate regression prediction model**: es el modelo de aprendizaje automático que permite predecir la producción y el rendimiento potencial del café en los diferentes municipios soportados en el prototipo.
- **Identificador categoría rendimiento potencial**: este componente se encarga de identificar a cuál de las tres categorías propuestas en esta investigación (bajo, medio y alto) corresponde el rendimiento potencial calculado por el modelo de aprendizaje automático. Por una parte, dicha categoría es presentada al usuario con fines informativos. Por otra parte, es insumo principal para conocer si deben generarse sugerencias a través del sistema de recomendaciones (si se registra un rendimiento potencial bajo).
- **Sistema de recomendaciones**: se encarga de la generación de recomendaciones encaminadas al incremento del rendimiento potencial en un municipio específico, de acuerdo a las condiciones climáticas que se han presentado en él en el último año.
- **Identificador condiciones climáticas**: este componente tiene como tarea la identificación y etiquetamiento de las condiciones climáticas que serán la entrada del sistema de recomendaciones para la generación de nuevas sugerencias. Para ello, analiza la temperatura media, precipitación total y cantidad de horas de brillo solar durante el último año y las etiqueta según las condiciones presentadas en la Tabla 5.
- **Interfaz de usuario**: es el medio por el que el usuario se comunica con los demás componentes del sistema. Mediante la interfaz, expuesta en la sección 6.1.2, el usuario puede consultar los datos generados por cuatro de los componentes anteriores: el modelo de aprendizaje automático, el identificador de la categoría del

rendimiento potencial, el sistema de recomendaciones y el repositorio de recomendaciones generales.

6.1.2 Diseño

Teniendo en cuenta las funcionalidades propuestas en la sección 6.1.1, se llevó a cabo el diseño y construcción de la interfaz de usuario del prototipo. En la primera pantalla, presentada en la Figura 20a, el usuario tiene acceso a todas las funcionalidades de la aplicación, que incluyen estimación del rendimiento potencial del café, recomendaciones y preguntas frecuentes. En cuanto a la primera función, estimación del rendimiento, el usuario puede elegir diferentes áreas geográficas que desee analizar, desde municipios hasta zonas cafeteras colombianas. En caso tal que acceda a una de las áreas geográficas disponibles, se despliega la pantalla presentada en la Figura 20b, donde se expone el rendimiento potencial y la producción estimada para la próxima cosecha en el municipio, departamento o zona cafetera. De igual modo, en dicha pantalla se puede observar una gráfica con los históricos de rendimiento y producción en el área seleccionada. Por otra parte, si el rendimiento potencial es bajo entonces se habilita la opción de ingresar a la pantalla expuesta en la Figura 20c, donde pueden conocerse algunas sugerencias, basadas en las condiciones climáticas del municipio, que podrían ayudar a mejorarlo.

Figura 20. De izquierda a derecha: (a) pantalla principal de la aplicación; (b) rendimiento potencial e históricos en un municipio; (c) recomendaciones para un municipio con un rendimiento potencial bajo. Fuente propia.

La segunda funcionalidad de la aplicación permite a los usuarios acceder a una serie de recomendaciones generales encaminadas a mejorar el rendimiento del cultivo de café; tal

como se expone en la Figura 21a, dichas recomendaciones están asociadas en cinco grupos según el aspecto al que hagan referencia: nutrición, arvenses, suelo, fitosanitario o actividades generales. En la Figura 21b y la Figura 21c, se exponen dos ejemplos de sugerencias a las que pueden acceder los usuarios al utilizar la aplicación. Finalmente, la aplicación tiene una pantalla, presentada en la Figura 22a, que permite a los usuarios acceder a un conjunto de preguntas frecuentes que pueden guiarlos en temas del cultivo de café, sin estar ligadas necesariamente al rendimiento del mismo. En la Figura 22b, se presenta un ejemplo de la pantalla con el detalle de una pregunta frecuente.

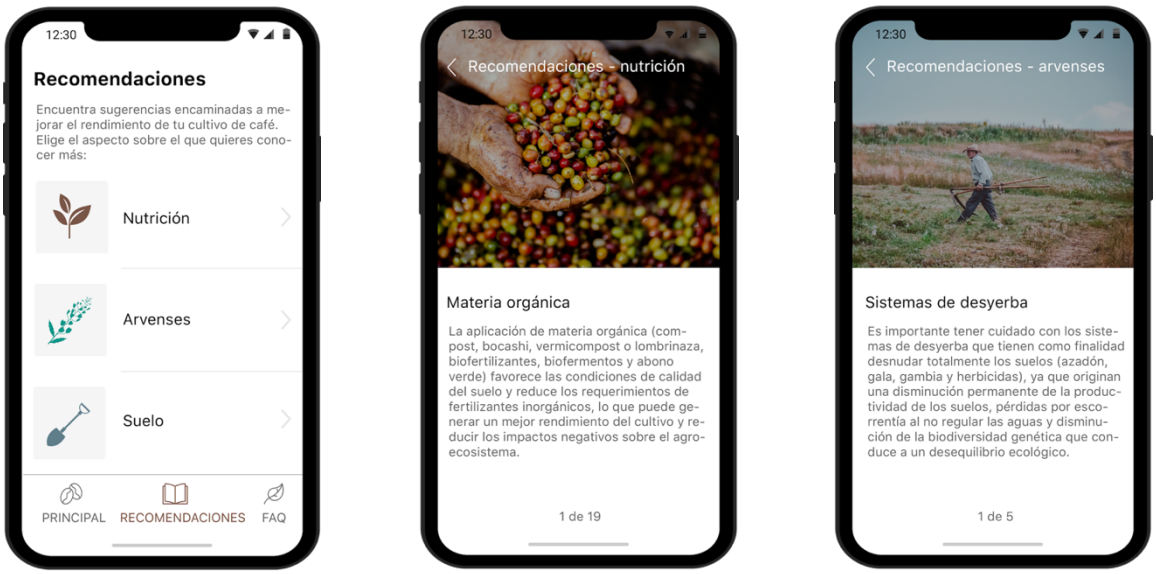


Figura 21. De izquierda a derecha: (a) pantalla principal de los grupos de recomendaciones generales; (b) recomendaciones en el campo de nutrición; (c) recomendaciones en el campo de arvenses. Fuente propia.

Figura 22. De izquierda a derecha: (a) pantalla principal de las preguntas frecuentes; (b) detalle de una pregunta frecuente. Fuente propia.

Para la etapa de evaluación del prototipo, todas las pantallas presentadas anteriormente fueron agrupadas en una aplicación web que muestra el flujo entre ellas, permitiendo de esta manera que diferentes usuarios puedan navegar en la misma y tener una idea general del funcionamiento de la aplicación móvil en caso de ser llevada a producción. De este modo, tanto la versión de producción de la aplicación como el montaje del sistema propuesto, se plantean como un trabajo futuro.

6.2 Experimentación

El objetivo principal de este apartado es llevar a cabo una evaluación del prototipo propuesto. Para esto, siguiendo la metodología de Experimentación en ingeniería de software [115], se ejecutan una serie de pasos que permiten determinar la satisfacción general de los usuarios al interactuar con el prototipo y la usabilidad del mismo. Tal como se presenta en la Figura 23, el proceso de experimentación está dividido en seis pasos. En primer lugar, se definen los objetivos de la experimentación. Luego, se determina el ambiente en el que el experimento será ejecutado. En un tercer paso, se eligen las personas que realizarán la evaluación del prototipo. Posteriormente, se plantean los instrumentos necesarios para el quinto paso, la ejecución de la prueba. Finalmente, se lleva a cabo el análisis e interpretación de los resultados. En la presente sección se profundiza en los primeros cinco pasos, mientras que el último de ellos es expuesto en la sección 6.3.

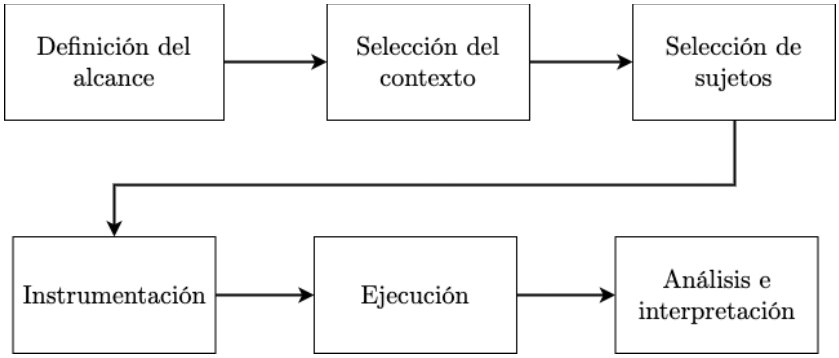


Figura 23. Descripción general del proceso de experimentación. Adaptado de [115].

6.2.1 Alcance

En la sección 6.1 se propuso un prototipo que busca ser una herramienta de apoyo para las personas involucradas en la cadena productiva del café colombiano. El prototipo permite estimar el rendimiento potencial del cultivo en diferentes áreas geográficas del país (municipios, departamentos y zonas cafeteras), así como su comportamiento con relación al promedio nacional. Sumado a esto, en caso de que el rendimiento potencial sea

bajo, el prototipo presenta una serie de recomendaciones específicas para la zona que, basadas en las condiciones climáticas, están encaminadas a mejorarlo. En este orden de ideas, a continuación, se definen los objetivos y alcance de la fase de experimentación:

Analizar las características del prototipo con el objetivo de evaluarlas respecto a la usabilidad y satisfacción general desde el punto de vista de los usuarios en el contexto de especialistas en las ciencias agrícolas, ambientales y de la computación.

6.2.2 Selección del contexto

La evaluación se llevó a cabo en línea con especialistas en ciencias agrícolas, ambientales y de la computación. Para ello, se utilizó una encuesta descriptiva y exploratoria. Por una parte, la naturaleza descriptiva permite generar afirmaciones acerca de las características del prototipo. Por otra parte, la naturaleza exploratoria sirve como estudio previo a una versión de producción de la aplicación, de esta forma pueden identificarse problemas iniciales, así como funcionalidades que podrían ser agregadas en un nuevo desarrollo.

La naturaleza del contexto implica que los usuarios tengan información clara sobre los objetivos de la investigación, el manejo de los datos y el funcionamiento del prototipo. En ese sentido, los individuos encuestados tienen acceso a dos artículos esenciales, presentados en la sección 6.2.4, además de la encuesta en sí: información sobre el alcance de la prueba y el funcionamiento del prototipo e información sobre el manejo de datos y confidencialidad de los mismos.

6.2.3 Selección de sujetos

Con el objetivo de obtener una visión general del aporte que puede brindar el prototipo propuesto para todas las personas vinculadas, de una u otra forma, a la cadena productiva del café en Colombia, se eligen especialistas en ciencias agrícolas, ciencias ambientales y ciencias de la computación como los individuos a ser encuestados. Entre los perfiles de los sujetos se encuentran profesionales e investigadores, todos con estudios posgraduales, ya sea en maestría o doctorado. Igualmente, todos los encuestados registran conocimiento y experiencia respecto al campo de la agricultura, algunos habiendo realizado estancias de investigación en institutos agronómicos, como lo son el Instituto Nacional de la Investigación Agronómica de Francia y el Centro Agronómico Tropical de Investigación y Enseñanza de Costa Rica. Así mismo, todas las personas seleccionadas han realizado investigaciones o trabajado en temas del cultivo de café.

Por otra parte, aunque es válida la postura de que es necesario un gran número de individuos para llevar a cabo una evaluación completa de la usabilidad de una aplicación, algunos estudios demuestran que pueden generarse excelentes resultados provenientes de pruebas con cerca de cinco o diez usuarios [116], [117]. En ese sentido, en la presente investigación fueron seleccionados nueve individuos para la evaluación.

6.2.4 Instrumentación

Teniendo en cuenta que son necesarias pautas para guiar a los participantes en el experimento, así como el prototipo en sí, se plantean los siguientes ítems como parte de la instrumentación base para llevar a cabo la evaluación:

- **Prototipo**: aplicación en línea que permite ver e interactuar con las pantallas propuestas en la sección 6.1.2. El bosquejo del prototipo se encuentra disponible en http://bit.ly/prototipoCafe.
- **Consentimiento informado**: expone la información necesaria para que los usuarios puedan tomar la decisión de participar o no en la evaluación. En ese sentido, en el consentimiento se informan varias características importantes de la prueba. En primer lugar, se exponen los objetivos de la misma. Igualmente, se informa que tendrán acceso en cualquier momento a información relevante del estudio, así como a los resultados finales. Finalmente, se comunica que no es capturada ninguna información confidencial o sensible para los usuarios, limitándose única y exclusivamente a las respuestas que sean brindadas. El consentimiento presentado a los usuarios puede verse en el anexo F.
- **Encuesta**: busca obtener, de manera sistemática y ordenada, información sobre la satisfacción general de los usuarios al utilizar el prototipo, así como su opinión frente a la usabilidad del mismo. En este contexto, con base en las propuestas de diferentes investigaciones para analizar la aceptación y usabilidad de un sistema informático [118]–[120], se creó un cuestionario compuesto de veinte preguntas encaminadas a evaluar los siguientes aspectos sobre el prototipo: reacción general frente al mismo, organización de las pantallas, terminología e información y conocimiento previo requerido. Así mismo, entre las preguntas mencionadas, se incluyeron algunas que buscan conocer características interesantes para los usuarios que pudiesen mejorar una nueva versión de la aplicación, así como funcionalidades que pueden ser eliminadas de la misma. La encuesta construida está disponible en http://bit.ly/evaluacionPrototipoCafe o puede verse en el anexo G.
- **Correo electrónico de invitación a la prueba**: tiene como objetivo invitar a la evaluación a los individuos seleccionados, incluyendo los dos enlaces en los que se encuentran disponibles el prototipo y la encuesta. Igualmente, contiene información sobre la naturaleza de la prueba y las características principales para entender el funcionamiento del prototipo. En el anexo H, se expone el formato del correo electrónico enviado a cada uno de los participantes.

6.2.5 Ejecución de la prueba

La fase de ejecución se inició con el envío del correo de invitación, expuesto en la sección 6.2.4, a cada uno de los nueve individuos seleccionados. Posteriormente, se hizo seguimiento de los accesos a la encuesta y el prototipo, para determinar finalmente cuántos individuos decidieron llevar a cabo la evaluación. Igualmente, durante toda la fase de experimentación, se vigiló la disponibilidad y correcto funcionamiento tanto del prototipo como de la encuesta en línea, con el objetivo de garantizar el acceso a los usuarios. Al finalizar la etapa de experimentación, se tuvo que los nueve individuos seleccionados interactuaron con el prototipo, aceptaron el consentimiento informado y diligenciaron totalmente el cuestionario.

6.3 Resultados

Una vez completada la encuesta por los nueve usuarios seleccionados, se generaron diversos resultados. En primera medida, tomando como base las respuestas a la primera, segunda y tercera pregunta, se tuvo que seis de ellos hacen parte de la academia, mientras que tres trabajan o administran alguna empresa. En cuanto al rol de los encuestados, cinco de ellos fungen como investigadores, tres como profesionales y un estudiante. Por otra parte, por lo que concierne al campo en el que se enmarca la profesión de los usuarios evaluados, se tienen siete en las ciencias de la computación, uno en ciencias ambientales y uno en ciencias agrícolas.

En segunda instancia, la cuarta y quinta pregunta de la encuesta fueron encaminadas a evaluar la reacción general frente al uso del prototipo. En la Figura 24 se exponen los resultados de las dos preguntas mencionadas. Por una parte, en cuanto a la estética general del prototipo, más de la mitad de los encuestados piensan que es “totalmente agradable”, mientras que el resto de los mismos se encuentran divididos entre “moderadamente agradable” y “ni agradable, ni desagradable”. Por otra parte, todos los usuarios encuestados piensan que el manejo de la aplicación es como mínimo “moderadamente fácil”. Dichas respuestas permiten concluir, que aún cuando la estética general de la aplicación es agradable y el manejo de la misma es fácil, existen diversas características que podrían mejorarse respecto a esos dos aspectos. Para ello, al final de la presente sección, en la Tabla 9, se hace un análisis de los comentarios entregados por lo usuarios en las últimas preguntas de la encuesta, buscando conocer qué aspectos presentan falencias y pueden mejorarse en una versión de producción de la aplicación y cuáles características fueron del agrado de los usuarios y podrían ser potenciadas en futuros desarrollos.

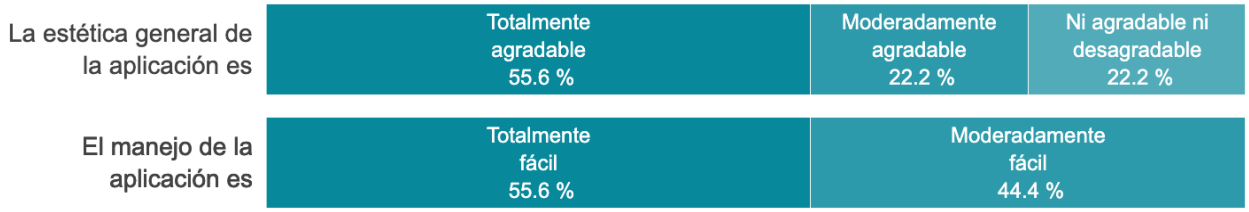


Figura 24. Resultados del cuestionario acerca de la reacción general de los usuarios frente al uso del prototipo. Fuente propia.

En un tercer grupo de preguntas, que incluyen desde la sexta hasta la novena, se evaluó la organización de las pantallas. Los resultados obtenidos, expuestos en la Figura 25, sugieren que la organización de las pantallas del prototipo es adecuada, destacándose la clara secuencia de las mismas (todos los encuestados piensan que la secuencia es como mínimo "moderadamente clara") y la integración con las funciones de la aplicación (más del 75% de los encuestados opinan que las funciones se encuentran "totalmente integradas"). Ahora bien, en futuros desarrollos, pueden mejorarse características como la organización de la información en las pantallas y la facilidad de lectura de las mismas, intentando disminuir la cantidad de palabras utilizadas e incrementar el tamaño de la fuente empleada.

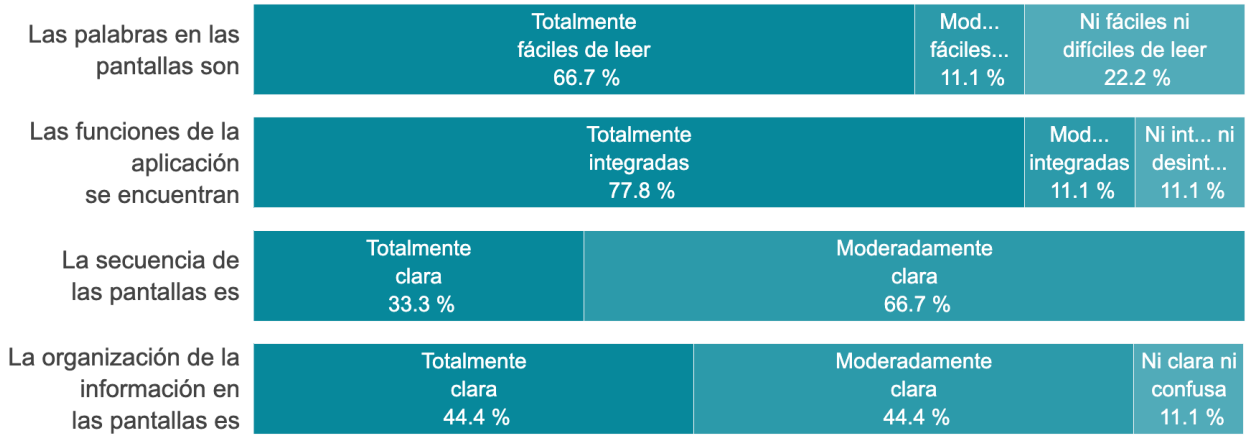


Figura 25. Resultados del cuestionario acerca de la organización de las pantallas del prototipo. Fuente propia.

La décima y undécima pregunta fueron construidas para evaluar la terminología e información utilizada en la aplicación. El porcentaje de las respuestas entregadas por los encuestados se presenta en la Figura 26. Los resultados sugieren que el uso de los términos es consistente, donde el 66.7% de los encuestados opina que es "totalmente consistente", mientras que el porcentaje restante afirma que es "moderadamente consistente". Por otra parte, en cuanto al lenguaje utilizado en la aplicación, las respuestas de los encuestados sugieren que es fácil de leer. Aún así, es un aspecto que puede mejorarse en futuros

desarrollos, puesto que el 22.2% de ellos respondieron que el lenguaje utilizado es “ni fácil, ni difícil de leer”.

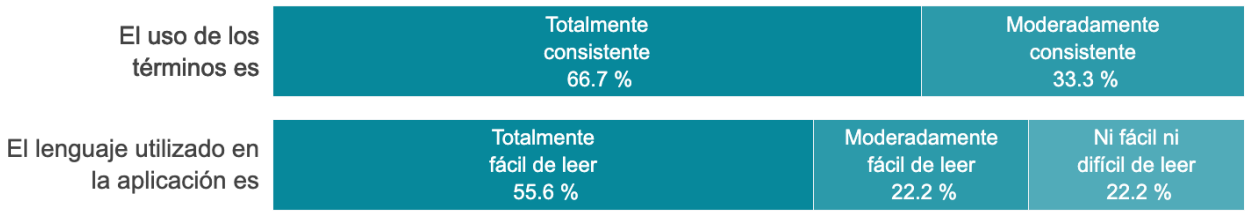


Figura 26. Resultados del cuestionario acerca de la terminología e información en el prototipo. Fuente propia.

En quinta instancia, mediante la duodécima y decimotercera pregunta, se evaluó el nivel de conocimiento previo sobre el prototipo y el manejo de aplicaciones móviles en general que deberían tener los usuarios para poder utilizar la aplicación propuesta. Los resultados, expuestos en la Figura 27, muestran que más de la mitad de los encuestados (55.6%) afirman que podrían utilizar la aplicación sin instrucciones previas, sin embargo, el porcentaje restante (44.4%) no está totalmente de acuerdo. Por otra parte, tan solo el 44.4% de los encuestados afirma que no es necesaria experiencia previa en el manejo de aplicaciones móviles. En este orden de ideas, los resultados permiten concluir que se hace necesaria la construcción de una guía básica o tutorial que conduzca a los usuarios a través de las funciones más importantes de la aplicación, con el objetivo de facilitarles el manejo y entendimiento de la misma.

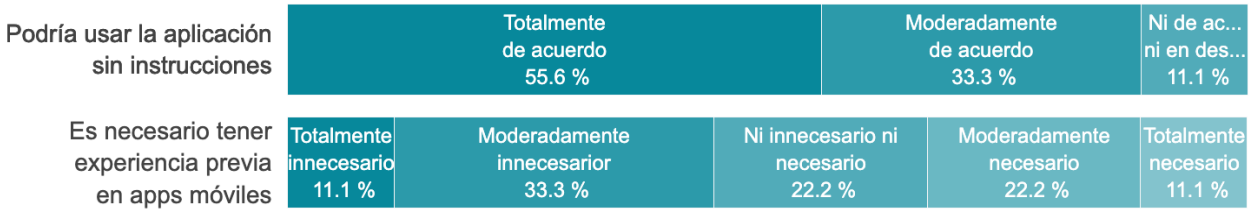


Figura 27. Resultados del cuestionario acerca del conocimiento previo requerido para el uso del prototipo. Fuente propia.

Un último grupo de preguntas, que incluye desde la decimocuarta hasta la decimoctava, fue construido para conocer la opinión general de los encuestados acerca de la aplicación propuesta, así como aspectos que no fueron analizados en preguntas anteriores. La Figura 28 expone los resultados registrados para la decimocuarta y decimoquinta pregunta. En ellas se destaca la importancia de la predicción del rendimiento del cultivo, donde todos los encuestados opinaron que puede ayudar a los caficultores en la búsqueda del incremento de la rentabilidad de sus cultivos (77.8% respondió que predecir el rendimiento puede ayudarlos siempre, mientras que el restante 22.2% respondió que puede hacerlo

muchas veces). Por otra parte, en cuanto a la pregunta de si la idea plasmada en el prototipo es novedosa o convencional, las respuestas fueron variadas, lo que sugiere que pueden potenciarse algunas características de la aplicación que permitan diferenciarla de modelos matemáticos existentes en la literatura científica para la predicción del rendimiento de cultivos.

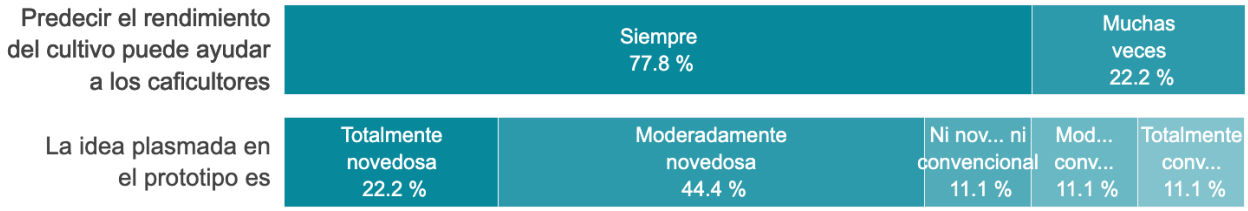


Figura 28. Resultados del cuestionario acerca de la opinión general de los usuarios. Fuente propia.

Finalmente, en la Tabla 9 se reflejan las respuestas entregadas por los encuestados en las últimas preguntas. Dichas preguntas se encontraban dirigidas a colectar la opinión general de los usuarios, investigando sobre los aspectos que más y menos llamaron su atención, así como características que, desde su perspectiva, deberían ser agregadas en una nueva versión. En ese orden de ideas, los elementos plasmados en la Tabla 9 pueden ser la base esencial para la generación de una versión de producción, así como trabajos futuros alrededor de la presente investigación.

	Aspectos destacados	Aspectos a mejorar
Análisis del rendimiento del café	- Estimación del rendimiento potencial. - Gráficas de históricos de producción y rendimiento. - Análisis por zonas.	- Relación entre el área geográfica analizada y las recomendaciones generadas no es muy evidente. - Estimación del rendimiento potencial a nivel de finca utilizando información específica del sitio (variedad, características fisicoquímicas, etc.). - Captura automática de la ubicación del dispositivo móvil. - Error del rendimiento estimado no es evidente.
Estética general de la aplicación	- Texto utilizado para la entrega de las recomendaciones. - Colores y simplicidad del diseño de las pantallas.	- Los valores estimados no resaltan sobre otros datos de la aplicación, aún cuando son el eje central de la misma.

		- Cantidad de texto utilizado en algunas pantallas es alto. - Párrafos secundarios utilizan colores opacos. - Algunos textos secundarios son pequeños, mientras que algunos primarios son demasiado grandes. - Pantalla con el municipio con mejor y peor rendimiento potencial. - Creación de un logotipo que identifique la aplicación.
Usabilidad de la aplicación	- Facilidad de acceso a cada una de las funciones. - Secuencia e integración de las pantallas.	- Guía o tutorial que permita conocer la aplicación previamente. - Flujo entre pantallas relacionadas con las recomendaciones para zonas específicas podría ser más claro.

Tabla 9. Síntesis de aspectos destacados y a mejorar de la aplicación, según la opinión de los encuestados. Fuente propia.

6.4 Conclusiones acerca del prototipo y la experimentación

En el presente capítulo se expuso el diseño y evaluación de un prototipo que busca ser una herramienta de apoyo para los caficultores colombianos en la búsqueda de mejorar las tasas de rendimiento de sus cultivos. Dicho prototipo utiliza el modelo de aprendizaje automático y el sistema de recomendaciones, expuestos en los capítulos 1 y 5 respectivamente, para predecir el rendimiento potencial del cultivo de café y generar sugerencias, basadas en condiciones climáticas históricas, que permitan mejorarlo. Por consiguiente, a partir de los procesos y resultados expuestos en el presente capítulo, se concluye:

- Aún cuando el modelo de aprendizaje automático construido presenta una alta correlación en la predicción de la producción de café, por sí solo no es herramienta amplia y suficiente para el apoyo de las personas involucradas en la cadena productiva de dicho cultivo. Es en ese sentido que cobra relevancia la construcción del prototipo propuesto en el presente capítulo, puesto que aprovecha las características del modelo y el sistema de recomendaciones, para brindar

información de valor a los usuarios. Es el conjunto de todas las características del prototipo, que puede servir como herramienta de apoyo para que los caficultores busquen las tasas de rendimiento deseadas.

- En ingeniería de software, dependiendo del propósito de la evaluación y las condiciones de la investigación, existen tres tipos diferentes de estrategias que pueden llevarse a cabo: encuesta, estudio de caso y experimento. En el caso de la investigación en curso, el uso de la encuesta permitió no solamente evaluar diferentes aspectos del prototipo (organización, terminología, conocimiento previo requerido, entre otros), sino también obtener una visión general de especialistas en el campo que puede servir como base para el perfeccionamiento de la aplicación, el sistema de recomendaciones y el modelo de aprendizaje propuestos en la presente investigación.
- De acuerdo a la evaluación de la satisfacción general de los usuarios frente al prototipo, la estética general de la aplicación es agradable y el manejo de la misma es fácil. Esto puede tener como razón de ser, el diseño simple y conciso propuesto. Sin embargo, aún quedan aspectos por mejorar respecto a dicho tema; la cantidad de palabras utilizadas, así como el tamaño de algunos textos, son algunos de los aspectos que pueden replantearse en un trabajo futuro.
- En general, la usabilidad del prototipo propuesto es adecuada. Preguntas dirigidas a evaluar la integración de las funciones en la aplicación, la secuencia de las pantallas, la organización de la información y el nivel de conocimiento previo necesario, tuvieron todas respuestas positivas. Características como la facilidad de acceso a las diversas funciones de la aplicación, la clara conexión entre pantallas y las gráficas de históricos presentadas, permitieron a los usuarios estar satisfechos con la usabilidad del prototipo.
- Características como la fácil estimación del rendimiento potencial, el análisis por diferentes áreas geográficas, los diferentes colores y simplicidad del diseño, la existencia de gráficas de datos históricos y la secuencia e integración de las pantallas, son aspectos destacados por parte de los especialistas seleccionados para la evaluación del prototipo, que pueden ser potenciados en aras de facilitar el manejo de la aplicación por parte de los caficultores. Por otra parte, la estimación del rendimiento potencial a nivel de finca utilizando información específica del sitio (variedad, características fisicoquímicas, etc.), la existencia de una guía o tutorial que permita conocer la aplicación previamente y la evidencia del error del rendimiento estimado, son nuevas características que podrían ser añadidas en trabajos futuros, con el objetivo de brindar mayores y mejores herramientas a los agricultores colombianos para la mejora de las tasas de rendimiento de sus cultivos.

7 Conclusiones y trabajo futuro

En este capítulo se resaltan las conclusiones más importantes obtenidas a través de la consecución de cada uno de los objetivos de la presente investigación. Así mismo, se plasman las contribuciones logradas con el trabajo desarrollado y las publicaciones generadas a través de la diseminación de los resultados obtenidos. Finalmente, se proponen diferentes trabajos futuros que pueden mejorar los resultados obtenidos en la presente investigación y con ello aportar a la calidad de vida de los caficultores colombianos.

7.1 Conclusiones

En la presente investigación se analizó el rendimiento del cultivo del café a partir de métodos de aprendizaje automático, en busca de generar una herramienta que permitiese apoyar la estabilidad de las tasas de rendimiento del cultivo y con ello contribuir a mejorar la calidad de vida de todas las personas involucradas en la cadena productiva del café en Colombia, especialmente la de los pequeños agricultores. En ese orden de ideas, se construyó un modelo de aprendizaje automático que permite calcular el rendimiento potencial del café. Posteriormente, se construyó un sistema de recomendaciones que toma las condiciones climáticas de un área específica para generar sugerencias de manejo del cultivo. Tanto el modelo como el sistema de recomendaciones constituyen un prototipo que busca ser una herramienta de apoyo para los caficultores colombianos. Acorde con el trabajo realizado y como resultado de los procesos mencionados, en cada capítulo se exponen diversas conclusiones específicas de los mismos. Por su parte, en la presente sección se proponen las siguientes conclusiones generales de la investigación:

- En la literatura científica existen diversas investigaciones con propuestas para la estimación o predicción del rendimiento de un cultivo. Dichas propuestas abarcan desde el modelamiento matemático y estadístico, hasta la aplicación de técnicas avanzadas de aprendizaje automático. Así mismo, la mayor parte de los estudios están dirigidos a cultivos extensivos, comprendiendo el trigo, arroz, caña de azúcar, soya y algodón, entre otros. Por su parte, en cuanto a la producción del café en Colombia, actualmente se basa en una metodología de muestreo en cafetales que se hace sobre los frutos. En consecuencia, la presente investigación cobra relevancia en el campo de la agricultura colombiana, generando una propuesta que, basada en algoritmos de aprendizaje automático, permite calcular el rendimiento potencial de la siguiente cosecha en el país.
- La productividad y rendimiento del cultivo de café se ven influidos directamente por diversas condiciones climáticas, físicas y químicas. Por una parte, variables climáticas como la temperatura, la lluvia y las horas de brillo solar, cumplen un papel determinante. Las deficiencias hídricas prolongadas, los excesos de agua y las

altas temperaturas atmosféricas, por ejemplo, traen consigo cambios drásticos en el rendimiento del café. Por otra parte, factores como la fecha de siembra, disponibilidad de materia orgánica y nitrógeno, edad del cultivo, calidad de la semilla y la evapotranspiración, entre otros, influyen directamente en la producción. Muestra de todo esto fue la alta correlación encontrada entre el rendimiento del cultivo de café, variables climáticas (temperatura, precipitación y horas de brillo solar) y el área geográfica analizada.

- En Colombia el acceso a datos sobre el sector agrícola es limitado. Registros acerca de los cultivos, como variedades, densidades de siembra y fechas de cosecha, entre otros, son escasos. Igualmente, datos sobre suelo, como porcentaje de materia orgánica, macro y micronutrientes, son casi nulos. En ese sentido, las investigaciones en el campo de la agricultura pueden verse limitadas y, con ello, generar resultados perfectibles. Es así que se hace necesario el incentivo de la política de suministro de información y datos abiertos por parte de las entidades gubernamentales, permitiendo con ello la generación de más y mejores investigaciones que ayuden a mejorar la calidad de vida de los colombianos.
- Se evaluó la precisión de diferentes algoritmos de aprendizaje automático, teniendo los mejores resultados con las técnicas *multivariate regression prediction model* y *support vector machine for regression*, con un coeficiente de correlación de 0.94. Todos los algoritmos evaluados fueron seleccionados con base en estudios previos. Sin embargo, esto no excluye la posibilidad de analizar otros algoritmos que podrían arrojar resultados exitosos. De manera semejante, podrían ser evaluados otros enfoques para analizar su desempeño en el área de interés de la investigación. Por ejemplo, si el rendimiento del cultivo de café y los datos climáticos se ordenan cronológicamente, podrían abordarse como un problema de series de tiempo. Por tanto, enfoques como algoritmos metaheurísticos, algoritmos genéticos y aprendizaje profundo, podrían usarse para estimar el rendimiento del cultivo y compararlo con el enfoque presentado en esta investigación.
- El sistema de recomendaciones construido permite utilizar las condiciones climáticas de un municipio para generar recomendaciones que guíen a los caficultores en aras de mejorar el rendimiento potencial de sus cultivos. En ese sentido, se propone como una herramienta interesante para apoyarlos en la búsqueda de altas tasas de rendimiento. Adicionalmente, al utilizar el enfoque basado en contenido para la construcción del sistema, permite incrementar el número de recomendaciones fácilmente en trabajos futuros y así mejorar la precisión del mismo.
- Tanto el modelo de aprendizaje automático como el sistema de recomendaciones propuestos, presentan un funcionamiento adecuado. Aún así, de forma individual pueden no ser una herramienta amplia y suficiente para el apoyo de las personas

involucradas en la cadena productiva del cultivo de café. En ese orden de ideas, cobra relevancia el prototipo propuesto, puesto que, aprovechando las características de diferentes módulos, se brinda información de valor a los agricultores colombianos. Es el conjunto de todas las características del prototipo que puede servir como herramienta de apoyo para que los caficultores busquen las tasas de rendimiento deseadas.

- La evaluación del prototipo propuesto fue exitosa. Por una parte, se contó con la participación de nueve especialistas en los campos de las ciencias agrícolas, ambientales y de la computación, que entregaron su visión general sobre la usabilidad y pertinencia de la aplicación propuesta. Por otra parte, preguntas dirigidas a evaluar el diseño, integración y organización, entre otros aspectos del prototipo, tuvieron respuestas positivas. Características como la facilidad de acceso a las diversas funciones de la aplicación, la clara conexión entre pantallas y las gráficas de históricos presentadas, permitieron a los encuestados estar satisfechos con la aplicación propuesta.

7.2 Contribuciones y publicaciones

Las principales contribuciones de este trabajo se dan en el entorno investigativo del aprendizaje automático aplicado a la agricultura y se presentan a continuación:

- Un conjunto de datos que comprende registros históricos meteorológicos y de rendimiento del cultivo de café en Colombia.
- Un conjunto de modelos que permiten la estimación de la producción y rendimiento del cultivo de café en Colombia.
- Un sistema de recomendaciones basado en contenido para la sugerencia de actividades de manejo del cultivo de café.
- Un prototipo orientado al usuario que permite estimar la producción y el rendimiento del cultivo de café en Colombia, además de observar diferentes actividades de manejo del cultivo para mejorar o mantener el rendimiento.
- Un artículo titulado "*Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data*", publicado en octubre de 2018 en la revista *Sustainability*, en la edición número 10 del volumen 10, clasificada Q2 JCR-SJR y A2 Publindex-Colciencias, en el que se expone un proceso piloto de la investigación llevado a cabo usando como cultivo principal el aguacate. Ver anexo I.
- Un artículo titulado "*Exploring machine learning: A bibliometric general approach using Citespace*", publicado en agosto de 2018 en la revista *F1000Research*, en el volumen 7, clasificada Q1 SJR y A1 Publindex-Colciencias, en el que se presenta

el mapeo sistemático realizado en la herramienta Citespace para conocer el estado actual del conocimiento alrededor del aprendizaje automático. Ver anexo J.

- Un artículo titulado "*Exploring machine learning: A bibliometric general approach using SciMAT*", publicado en agosto de 2018 en la revista *F1000Research*, en el volumen 7, clasificada Q1 SJR y A1 Publindex-Colciencias, en el que se presenta el mapeo sistemático realizado en la herramienta SciMAT para conocer el estado actual del conocimiento alrededor del aprendizaje automático. Ver anexo K.

7.3 Trabajos futuros

En la presente investigación se propuso una herramienta que busca ayudar a los caficultores colombianos a mejorar las tasas de rendimiento de sus cultivos. Para ello, se estima el rendimiento potencial y se generan sugerencias, basadas en las condiciones climáticas del sitio, encaminadas a mejorarlo. En ese sentido, con relación a los resultados obtenidos en la presente investigación, se proponen los siguientes trabajos futuros:

- **Utilizar información adicional para mejorar la correlación y aplicabilidad del modelo de aprendizaje automático**: tanto en la presente investigación como en la literatura científica, se demuestra que existe una alta correlación entre las condiciones específicas del sitio del cultivo y el rendimiento del mismo. Se propone entonces utilizar variables adicionales con el objetivo de completar un sistema que pueda predecir con una alta correlación el rendimiento del cultivo de café. Variables como la disponibilidad de nutrientes, materia orgánica, temperatura del suelo, entre otras, deberían ser analizadas. Así mismo, se propone como un trabajo interesante el uso de las mismas variables para el análisis de clases objetivo diferentes, como la incidencia de enfermedades o el requerimiento de nutrientes. Finalmente, es posible utilizar registros climáticos y de rendimiento del cultivo de años actuales, para evaluar y mejorar el modelo propuesto en la presente investigación.
- **Analizar diferentes técnicas de aprendizaje automático y enfoques de análisis de datos**: las seis técnicas de aprendizaje automático analizadas presentaron buenos resultados. No obstante, se hace interesante la posibilidad de analizar otros algoritmos y enfoques en trabajos futuros. Por ejemplo, el problema de investigación puede ser abordado mediante series de tiempo; con ello, algoritmos metaheurísticos, algoritmos genéticos o métodos avanzados de aprendizaje profundo, podrían ser utilizados para generar una comparación con los resultados de la presente investigación. De manera similar, podrían usarse métodos de ensamble para combinar varios modelos que analicen variables demográficas, edáficas y fenológicas para mejorar la precisión al estimar el rendimiento potencial del cultivo de café.

- **Construir un conjunto de recomendaciones sobre el manejo del cultivo de café más grande y preciso**: las sugerencias expuestas en la presente investigación constituyen una propuesta inicial, construida a partir de literatura disponible y la guía de un especialista, que puede apoyar la estabilidad de las tasas de rendimiento del cultivo de café. Sin embargo, el número de recomendaciones es bajo y fueron construidas a partir de tres variables climáticas (temperatura, precipitación y horas de brillo solar). En ese sentido, se propone como trabajo futuro el incremento de las recomendaciones propuestas, así como también la utilización de un número mayor de variables que permitan generar sugerencias más precisas y acordes con todas las propiedades del cultivo. Por último, es necesaria la inclusión de un número mayor de especialistas que lleven a cabo la evaluación del sistema de recomendaciones propuesto.
- **Desarrollar un sistema de recomendaciones híbrido para mejorar las tasas de rendimiento del cultivo**: si bien el sistema de recomendaciones basado en contenido presenta el funcionamiento esperado, generando sugerencias de acuerdo al perfil del municipio de interés, se propone como un trabajo futuro el desarrollo de un sistema de recomendaciones híbrido, que pueda realimentarse de los resultados obtenidos por los caficultores al implementar las prácticas sugeridas por el mismo. Igualmente, aún cuando el cálculo de la similitud coseno permitió la generación exitosa de recomendaciones, se hace interesante la inclusión o evaluación de otras medidas como la correlación de Pearson.

8 Referencias bibliográficas

[1] W. Vergara, A. R. Rios, P. Trapido, and H. Malarín, "Agricultura y clima futuro en América Latina y el Caribe: impactos sistémicos y posibles respuestas," 2014.

[2] OECD, "OECD Review of Agricultural Policies: Colombia 2015," pp. 21–23, 2015.

[3] OECD/FAO, "OCDE-FAO Perspectivas Agrícolas 2014 - 2023," pp. 23–68, 2014.

[4] FAO, *Agricultura mundial, hacia Los Anos 2015/2030: Informe resumido*. FAO, 2002.

[5] J. C. Aker, I. Ghosh, and J. Burrell, "The promise (and pitfalls) of ICT for agriculture initiatives," *Agric. Econ.*, vol. 47, no. S1, pp. 35–48, 2016.

[6] M. R. Bendre, R. C. Thool, and V. R. Thool, "Big data in precision agriculture: Weather forecasting for future farming," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015, pp. 744–750.

[7] F. Gast *et al.*, *Manual del cafetero colombiano: investigación y tecnología para la sostenibilidad de la caficultura - Tomo 1*. FNC-Cenicafé, 2013.

[8] S. Salcedo and L. Guzmán, *Agricultura familiar en América Latina y el Caribe: recomendaciones de política*. FAO, 2014.

[9] Á. Jaramillo-Robledo and J. Arcila-Pulgarín, "Variabilidad climática en la zona cafetera colombiana asociada al evento de la niña y su efecto en la caficultura," in *Avances Técnicos Cenicafé*, no. 389, FNC, 2009, pp. 1–9.

[10] M. G. de la Cadena, J. L. Saltijeral Giles, and S. M. Sosa Clavijo, *Guía para el desarrollo de mercados de productores: Proyecto "Creación de Cadenas Cortas Agroalimentarias en la Ciudad de México."* FAO, 2017.

[11] C. H. Botero, *Un modelo para investigación documental: guia teórico-práctica sobre construcción de estados del arte con importantes reflexiones sobre la investigación.* Señal Editora, 2000.

[12] R. A. Fischer, "Definitions and determination of crop yield, yield gaps, and of rates of change," *F. Crop. Res.*, vol. 182, pp. 9–18, 2015.

[13] L. T. Evans and R. A. Fischer, "Yield Potential: Its Definition, Measurement, and Significance," in *Crop science*, 1999, vol. 39, no. 6, pp. 1544–1551.

[14] D. B. Lobell, K. G. Cassman, and C. B. Field, "Crop Yield Gaps: Their Importance, Magnitudes, and Causes," *Annu. Rev. Environ. Resour.*, vol. 34, no. 1, pp. 179–204, 2009.

[15] A. A. L. Selvakumar and G. M. Nazer, "An implementation of expert system in garlic using (ABC) Algorithm," in *2011 3rd International Conference on Electronics Computer Technology*, 2011, vol. 1, pp. 45–48.

[16] M. Castelli, L. Vanneschi, and Á. R. Largo, "Supervised Learning: Classification," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 342–349.

[17] D. C. Corrales, J. C. Corrales, and A. Figueroa-Casas, "Towards Detecting Crop Diseases and Pest by Supervised Learning," *Ing. y Univ.*, vol. 19, no. 1, pp. 207–228, 2015.

[18] T. Hastie, J. Friedman, and R. Tibshirani, "Overview of Supervised Learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer New York, 2001, pp. 9–40.

[19] E. Zarei, A. Azadeh, N. Khakzad, M. M. Aliabadi, and I. Mohammadfam, "Dynamic safety assessment of natural gas stations using Bayesian network," *J. Hazard. Mater.*, vol. 321, pp. 830–840, 2016.

[20] D. Basak, S. Pal, and D. Chandra Patranabis, "Support Vector Regression," *Neural Inf. Process. – Lett. Rev.*, vol. 11, no. 10, pp. 203–224, 2007.

[21] D. Chaparro *et al.*, "L-Band Vegetation Optical Depth for Crop Phenology Monitoring and Crop Yield Assessment," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 8225–8227.

[22] Y. M. Fernandez-Ordoñez and J. Soria-Ruiz, "Maize crop yield estimation with remote sensing and empirical models," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 3035–3038.

[23] I. Ahmad *et al.*, "Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan," *J. Indian Soc. Remote Sens.*, vol. 46, no. 10, pp. 1701–1711, Oct. 2018.

[24] H. Lee and A. Moon, "Development of yield prediction system based on real-time agricultural meteorological information," in *16th International Conference on Advanced Communication Technology*, 2014, pp. 1292–1295.

[25] T. R. Tolentino and A. A. Hernandez, "Assessment of Predictive Models for Coffee Production in the Philippines," in *2018 16th International Conference on ICT and Knowledge Engineering (ICT KE)*, 2018, pp. 1–6.

[26] A. Kern *et al.*, "Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices," *Agric. For. Meteorol.*, vol. 260–261, pp. 300–320, 2018.

[27] M.-J. Lambert, P. C. S. Traoré, X. Blaes, P. Baret, and P. Defourny, "Estimating smallholder crops production at village level from Sentinel-2 time series in Mali's cotton belt," *Remote Sens. Environ.*, vol. 216, pp. 647–657, 2018.

[28] A. Shastry, S. H. A, and M. Hegde, "A parameter based ANFIS model for crop yield prediction," in *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 253–257.

[29] H. Yildiz, A. Mermer, M. Aydoğdu, and O. Şïmşek, "Forecasting of winter wheat yield for Turkey using Water Balance Model," in *2015 Fourth International Conference on Agro-Geoinformatics (Agro-geoinformatics)*, 2015, pp. 352–356.

[30] R. J. Donohue, R. A. Lawes, G. Mata, D. Gobbett, and J. Ouzman, "Towards a national, remote-sensing-based model for predicting field-scale crop yield," *F. Crop. Res.*, vol. 227, pp. 79–90, 2018.

[31] D. V Gaso, A. G. Berger, and V. S. Ciganda, "Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images," *Comput. Electron. Agric.*, vol. 159, pp. 75–83, 2019.

[32] D. Helman, I. M. Lensky, and D. J. Bonfil, "Early prediction of wheat grain yield production from root-zone soil water content at heading using Crop RS-Met," *F. Crop. Res.*, vol. 232, pp. 11–23, 2019.

[33] N. K. Gontia and K. N. Tiwari, "Yield Estimation Model and Water Productivity of Wheat Crop (Triticum aestivum) in an Irrigation Command Using Remote Sensing and GIS," *J. Indian Soc. Remote Sens.*, vol. 39, no. 1, pp. 27–37, Mar. 2011.

[34] X. Xu *et al.*, "Winter Wheat Yield Estimation Coupling Weight Optimization Combination Method with Remote Sensing Data from Landsat5 TM," in *Computer and Computing Technologies in Agriculture V*, 2012, pp. 284–292.

[35] B. H. Janssen, F. C. T. Guiking, D. van der Eijk, E. M. A. Smaling, J. Wolf, and H. van Reuler, "A system for quantitative evaluation of the fertility of tropical soils (QUEFTS)," *Geoderma*, vol. 46, no. 4, pp. 299–318, 1990.

[36] E. M. A. Smaling and B. H. Janssen, "Calibration of quefts, a model predicting nutrient uptake and yields from chemical soil fertility indices," *Geoderma*, vol. 59, no. 1, pp. 21–44, 1993.

[37] G. P. Maro, J. P. Mrema, B. M. Msanya, B. H. Janssen, and J. M. Teri, "Developing a Coffee Yield Prediction and Integrated Soil Fertility Management Recommendation Model for Northern Tanzania," *Int. J. Plant Soil Sci.*, vol. 3, no. 4, pp. 380–396, 2014.

[38] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Comput. Electron. Agric.*, vol. 151, pp. 61–69, 2018.

[39] Y. Gandge, "A study on various data mining techniques for crop yield prediction," in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017, pp. 420–423.

[40] G. Rub, "Data Mining of Agricultural Yield Data: A Comparison of Regression Models," in *Advances in Data Mining. Applications and Theoretical Aspects*, 2009, pp. 24–37.

[41] H. Aghighi, M. Azadbakht, D. Ashourloo, H. S. Shahrabi, and S. Radiom, "Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 12, pp. 4563–4577, Dec. 2018.

[42] M. Stas, J. Van Orshoven, Q. Dong, S. Heremans, and B. Zhang, "A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT," in *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 2016, pp. 1–5.

[43] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and

remotely sensed data," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 858–861.

[44] S. Khanal, J. Fulton, A. Klopfenstein, N. Douridas, and S. Shearer, "Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield," *Comput. Electron. Agric.*, vol. 153, pp. 213–225, 2018.

[45] A. Shah, A. Dubey, V. Hemnani, D. Gala, and D. R. Kalbande, "Smart Farming System: Crop Yield Prediction Using Regression Techniques," in *Proceedings of International Conference on Wireless Communication*, 2018, pp. 49–56.

[46] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, 2016, pp. 1–4.

[47] P. G. Oguntunde, G. Lischeid, and O. Dietrich, "Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis," *Int. J. Biometeorol.*, vol. 62, no. 3, pp. 459–469, Mar. 2018.

[48] M. M. Rahman, N. Haq, and R. M. Rahman, "Machine Learning Facilitated Rice Prediction in Bangladesh," in *2014 Annual Global Online Conference on Information and Computer Technology*, 2014, pp. 1–4.

[49] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp. 1–5.

[50] A. T. M. S. Ahamed *et al.*, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," in *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2015, pp. 1–6.

[51] A. X. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon, "Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018, pp. 50:1--50:5.

[52] R. D. Baruah, S. Roy, R. M. Bhagat, and L. N. Sethi, "Use of Data Mining Technique for Prediction of Tea Yield in the Face of Climate Change of Assam, India," in *2016 International Conference on Information Technology (ICIT)*, 2016, pp. 265–269.

[53] P. Charoen-Ung and P. Mittrapiyanuruk, "Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques," in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2018, pp. 1–6.

[54] A. Haghverdi, R. A. Washington-Allen, and B. G. Leib, "Prediction of cotton lint

yield from phenology of crop indices using artificial neural networks," *Comput. Electron. Agric.*, vol. 152, pp. 186–197, 2018.

[55] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods," *Agric. For. Meteorol.*, vol. 218–219, pp. 74–84, 2016.

[56] C. Bunn, P. Läderach, O. Ovalle Rivera, and D. Kirschke, "A bitter cup: climate change profile of global production of Arabica and Robusta coffee," *Clim. Change*, vol. 129, no. 1, pp. 89–101, Mar. 2015.

[57] L. Kouadio, R. C. Deo, V. Byrareddy, J. F. Adamowski, S. Mushtaq, and V. P. Nguyen, "Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties," *Comput. Electron. Agric.*, vol. 155, pp. 324–338, 2018.

[58] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000.

[59] J. A. Pulgarín, "Factores que determinan la productividad del cafetal," in *Sistemas de producción de café en Colombia*, 1st ed., vol. 1, FNC-Cenicafé, 2007, pp. 61–86.

[60] V. H. Ramírez, J. Arcila, Á. Jaramillo, J. R. Rendón, G. Cuesta, and J. C. García, "Variabilidad climática y la floración del café en Colombia," p. 8, 2011.

[61] P. Steduto, T. C. Hsiao, E. Fereres, and D. Raes, *Respuesta del rendimiento de los cultivos al agua*, no. 66. FAO, 2012.

[62] V. H. Ramírez, J. Arcila, Á. Jaramillo, J. R. Rendón, G. Cuesta, and H. Menza, "Floración del café en Colombia y su relación con la disponibilidad hídrica, térmica y de brillo solar," vol. 2, no. 61, pp. 132–158, 2010.

[63] Á. J. Robledo, V. H. Ramírez, and J. Arcila Pulgarín, *Distribución de la lluvia: Clave para planificar las labores en el cultivo del café en Colombia*. Cenicafé, 2011.

[64] C. Bustamante González, A. Pérez Díaz, R. Rivera Espinosa, G. M. Martín Alonso, and R. Viñals Nuñez, "Influencia de las precipitaciones en el rendimiento de Coffea Canephora Pierre ex Froehner cultivado en suelos pardos de la región oriental de Cuba," *Cultiv. Trop.*, vol. 36, no. 4, pp. 21–27, 2015.

[65] M. E. Fernández, "Efectos del cambio climático en el rendimiento de tres cultivos mediante el uso del Modelo AquaCrop," 2013.

[66] Á. J. Robledo, "Épocas recomendadas para la siembra del café en Colombia," p. 12, 2016.

[67] A. S. Corrêa and P.-O. Zander, "Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools," in *Proceedings of the 18th Annual International Conference on Digital Government Research*, 2017, pp. 54–63.

[68] D. C. Corrales, A. Ledezma, and J. C. Corrales, "A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal," *J. Comput.*, vol. 10, no. 6, pp. 396–405, 2015.

[69] J. Rincon-Patino, E. Lasso, and J. C. Corrales, "Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data," *Sustainability*, vol. 10, no. 10, 2018.

[70] D. Corrales, E. Lasso, A. Ledezma Espino, and J. Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?," *J. Intell. Fuzzy Syst.*, vol. 34, pp. 1–11, 2018.

[71] F. N. de Cafeteros de Colombia, "Nuestras regiones cafeteras," 2010. [Online]. Available: http://www.cafedecolombia.com/particulares/es/la_tierra_del_cafe/regiones_cafeteras/. [Accessed: 30-Jul-2019].

[72] J. E. Clark, J. W. Osborne, P. Gallagher, and S. Watson, "A simple method for optimising transformation of non-parametric data: an illustration by reference to cortisol assays," *Hum. Psychopharmacol.*, vol. 31, no. 4, pp. 259–267, 2016.

[73] G. H. Skrepnek, "Regression methods in the empiric analysis of health care data," *J. Manag. Care Pharmac*, vol. 11, no. 3, pp. 240–251, 2005.

[74] K. Rasouli, W. W. Hsieh, and A. J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *J. Hydrol.*, vol. 414–415, pp. 284–293, 2012.

[75] E. Frank *et al.*, "Weka-A Machine Learning Workbench for Data Mining," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2010, pp. 1269–1277.

[76] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[77] D. C. Corrales, "Multiclasificador para la detección de la roya en cultivos de café en Colombia," Universidad del Cauca, 2014.

[78] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, 2007, pp. 3–24.

[79] A. Behnood, V. Behnood, M. M. Gharehveran, and K. E. Alyamac, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," *Constr. Build. Mater.*, vol. 142, pp. 199–207, 2017.

[80] J. R. Quinlan, "Learning With Continuous Classes," in *Proceedings of the Australian Joint Conference on Artificial Intelligence*, 1992, pp. 343–348.

[81] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," in *Proc of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague*, 1997.

[82] C. Zhan, A. Gan, and M. Hadi, "Prediction of Lane Clearance Time of Freeway

Incidents Using the M5P Tree Algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1549–1557, Dec. 2011.

[83] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *Eur. J. Oper. Res.*, vol. 116, no. 1, pp. 16–32, 1999.

[84] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *Int. J. Forecast.*, vol. 16, no. 4, pp. 437–450, 2000.

[85] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci. (Ny).*, vol. 191, pp. 192–213, 2012.

[86] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," in *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit*, 2007, pp. 60–69.

[87] R. Taylor, "Interpretation of the Correlation Coefficient: A Basic Review," *J. Diagnostic Med. Sonogr.*, vol. 6, no. 1, pp. 35–39, 1990.

[88] E. G. Lasso, "Sistema experto basado en emparejamiento de patrones," Universidad del Cauca, 2016.

[89] S. S. Rathore and A. Gupta, "A Comparative Study of Feature-ranking and Feature-subset Selection Techniques for Improved Fault Prediction," in *Proceedings of the 7th India Software Engineering Conference*, 2014, pp. 7:1--7:10.

[90] D. Ferreira, H. Peixoto, J. Machado, and A. Abelha, "Predictive Data Mining in Nutrition Therapy," in *2018 13th APCA International Conference on Automatic Control and Soft Computing (CONTROLO)*, 2018, pp. 137–142.

[91] S. Chakraborti and J. Li, "Confidence Interval Estimation of a Normal Percentile," *Am. Stat.*, vol. 61, no. 4, pp. 331–336, 2007.

[92] U. R. Hodeghatta and U. Nayak, "Introduction to descriptive analytics," in *Business Analytics Using R - A Practical Approach*, Berkeley, CA: Apress, 2017, pp. 59–89.

[93] A. Peña, V. Ramírez, J. Valencia, and Á. Jaramillo, *La lluvia como factor de amenaza para el cultivo del café en Colombia*. Cenicafé, 2012.

[94] F. Gast *et al.*, *Manual del cafetero colombiano: investigación y tecnología para la sostenibilidad de la caficultura - Tomo 2*. FNC-Cenicafé, 2013.

[95] F. Gast *et al.*, *Manual del cafetero colombiano: investigación y tecnología para la sostenibilidad de la caficultura - Tomo 3*. FNC-Cenicafé, 2013.

[96] C. Figueroa, I. Vagliano, O. R. Rocha, and M. Morisio, "A Systematic Literature Review of Linked Data-based Recommender Systems," *Concurr. Comput. Pr. Exper.*, vol. 27, no. 17, pp. 4659–4684, Dec. 2015.

[97] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 1–35.

[98] X. Amatriain, A. Jaimes*, N. Oliver, and J. M. Pujol, "Data Mining Methods for Recommender Systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 39–71.

[99] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based Recommender Systems: State of the Art and Trends," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 73–105.

[100] H. Lieberman, "Letizia: An Agent That Assists Web Browsing," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, 1995, pp. 924–929.

[101] Y. Zang, Y. An, and X. T. Hu, "Automatically recommending healthy living programs to patients with chronic diseases through hybrid content-based and collaborative filtering," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 578–582.

[102] A. A. Neamah and A. S. El-Ameer, "Design and Evaluation of a Course Recommender System Using Content-Based Approach," in *2018 International Conference on Advanced Science and Engineering (ICOASE)*, 2018, pp. 1–6.

[103] M. J. Mokarrama and M. S. Arefin, "RSF: A recommendation system for farmers," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 843–850.

[104] M. A. Salam *et al.*, "Climate Recommender System for Wheat Cultivation in North Egyptian Sinai Peninsula," in *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014*, 2014, pp. 121–130.

[105] B. Li, "Recommendation System of Crop Planting Books Based on Big Data," in *Revista de la Facultad de Agronomia de la Universidad del Zulia*, 2019, vol. 36, no. 4, pp. 1166–1174.

[106] G. I. Puerta Quintero, "Buenas prácticas agrícolas para el café," in *Avances Técnicos Cenicafé*, no. 349, FNC, 2006, pp. 1–12.

[107] G. I. Puerta Quintero, "Buenas prácticas para la prevención de los defectos de la calidad del café: fermento, reposado, fenólico y mohoso," in *Avances Técnicos Cenicafé*, no. 461, S. M. Marín López, Ed. FNC, 2015, pp. 1–12.

[108] F. Farfán Valencia, "Las buenas prácticas agrícolas en la caficultura," in *Sistemas de producción de cafés en Colombia*, 1st ed., vol. 1, FNC-Cenicafé, 2011, pp. 276–309.

[109] J. Pulgarín and F. Farfán Valencia, "Consideraciones sobrela nutrición mineral y orgánica en la producción de la finca," in *Sistemas de producción de cafés en Colombia*, 1st ed., vol. 1, FNC-Cenicafé, 2011, pp. 201–232.

[110] F. Farfán, "Evaluación de la vulnerabilidad del suelo en el cultivo del café a la variabilidad climática," 2018.

[111] F. Farfán, "Percepción de los caficultores de los municipios de Salamina (Caldas), Santuario y Balboa (Risaralda), frente a la variabilidad climática," 2017.

[112] F. Bustamante, C. H. Isaza, N. van Heeren, G. Torres, and R. Romero, *Buenas prácticas para la producción de café*. Solidaridad - Coffee Support Network, 2009.

[113] P. Descamps, *Técnicas para la producción sostenible de café frente al cambio climático*. Instituto Nacional de Innovación y Transferencia en Tecnología Agropecuaria, 2017.
[114] R. Miles and K. Hamilton, *Learning UML 2.0*. O'Reilly Media, 2006.
[115] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*, 1st ed. Springer-Verlag Berlin Heidelberg, 2012.
[116] J. Nielsen and T. K. Landauer, "A Mathematical Model of the Finding of Usability Problems," in *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 1993, pp. 206–213.
[117] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Behav. Res. Methods, Instruments, Comput.*, vol. 35, no. 3, pp. 379–383, Aug. 2003.
[118] J. Chin, V. Diehl, and K. Norman, "Development of a Tool Measuring User Satisfaction of the Human-Computer Interface," 1988.
[119] J. Nielsen and R. Molich, "Heuristic Evaluation of User Interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 249–256.
[120] J. Nielsen, "Enhancing the Explanatory Power of Usability Heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994, pp. 152–158.
[121] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach*. Springer Berlin Heidelberg, 2013.
[122] C. Crisci, B. Ghattas, and G. Perera, "A review of supervised machine learning algorithms and their applications to ecological data," *Ecol. Modell.*, vol. 240, pp. 113–122, 2012.
[123] I. D. López, A. Figueroa, and J. C. Corrales, "Adaptive Prediction of Water Quality Using Computational Intelligence Techniques," in *Computational Science and Its Applications -- ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part II*, Cham: Springer International Publishing, 2017, pp. 45–59.
[124] I. Smeureanu, G. Ruxanda, and L. M. Badea, "Customer segmentation in private banking sector using machine learning techniques," *J. Bus. Econ. Manag.*, vol. 14, no. 5, pp. 923–939, 2013.
[125] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.
[126] D. C. Corrales, A. F. Casas, A. Ledezma, and J. C. Corrales, "Two-Level Classifier Ensembles for Coffee Rust Estimation in Colombian Crops," *Int. J. Agric. Environ. Inf. Syst.*, vol. 7, no. 3, pp. 41–59, 2016.
[127] E. Lasso, Ó. Valencia, and J. C. Corrales, "Decision Support System for Coffee Rust Control Based on Expert Knowledge and Value-Added Services," in *Computational*

Science and Its Applications -- ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part II, Cham: Springer International Publishing, 2017, pp. 70–83.

[128] J. A. Moral-Muñoz, M. J. Cobo, E. Peis, M. Arroyo-Morales, and E. Herrera-Viedma, "Analyzing the research in Integrative & Complementary Medicine by means of science mapping," *Complement. Ther. Med.*, vol. 22, no. 2, pp. 409–418, 2014.

[129] M. A. Martínez, M. J. Cobo, M. Herrera, and E. Herrera-Viedma, "Analyzing the Scientific Evolution of Social Work Using Science Mapping," *Res. Soc. Work Pract.*, vol. 25, no. 2, pp. 257–277, 2015.

[130] C. Chen, Z. Hu, S. Liu, and H. Tseng, "Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace," *Expert Opin. Biol. Ther.*, vol. 5, no. 12, pp. 593–608, 2012.

[131] M. J. Cobo, F. Chiclana, A. Collop, J. de Ona, and E. Herrera-Viedma, "A Bibliometric Analysis of the Intelligent Transportation Systems Research Based on Science Mapping," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 901–908, 2014.

[132] C. Chen, *Information Visualization: Beyond the Horizon*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[133] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 359–377, 2006.

[134] C. Chen, "Searching for intellectual turning points: Progressive knowledge domain visualization," *Proc. Natl. Acad. Sci.*, vol. 101, no. suppl 1, pp. 5303–5310, 2004.

[135] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "SciMAT: A new science mapping analysis software tool," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1609–1630, 2012.

[136] K. W. McCain, "Cocited author mapping as a valid representation of intellectual structure," *JASIS*, vol. 37, pp. 111–122, 1986.

[137] A. Rip and J. Courtial, "Co-word maps of biotechnology: An example of cognitive scientometrics," *Scientometrics*, vol. 6, no. 6, pp. 381–400, 1984.

[138] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field," *J. Informetr.*, vol. 5, no. 1, pp. 146–166, 2011.

[139] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 7, pp. 1382–1402, 2011.

[140] E. Garfield, "Scientography: Mapping the tracks of science," *Curr. Contents Soc. Behav. Sci.*, vol. 7, no. 45, pp. 5–10, 1994.

[141] M. Callon, J. Courtial, and F. Laville, "Co-word analysis as a tool for describing the

network of interactions between basic and technological research: The case of polymer chemsitry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, 1991.

[142] J. Rincon-Patino, G. Ramirez-Gonzalez, and J. C. Corrales, "Exploring machine learning: A bibliometric general approach using Citespace," *F1000Research*, vol. 7, no. 1240, 2018.

[143] J. Rincon-Patino, G. Ramirez-Gonzalez, and J. C. Corrales, "Exploring machine learning: A bibliometric general approach using SciMAT," *F1000Research*, vol. 7, no. 1210, 2018.

ANÁLISIS DEL RENDIMIENTO DEL CAFÉ BASADO EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

ANEXOS

JUAN DAVID RINCÓN PATIÑO

Tesis de Maestría en Ingeniería Telemática

Director
Juan Carlos Corrales Muñoz
Doctor en Ciencias de la Computación

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de investigación e-@mbiente
Popayán, 2019

A Análisis bibliométrico en aprendizaje automático

En el campo de aprendizaje automático se enmarcan investigaciones acerca de los diferentes procesos de aprendizaje humano, el análisis teórico de posibles algoritmos de aprendizaje y métodos aplicados en distintos dominios [121]. Estudios basados en aprendizaje automático han permitido a investigadores y compañías analizar eventos de mortalidad masiva [122], predecir la calidad del agua [123], segmentar clientes de entidades bancarias [124], clasificar texto automáticamente [125] y predecir el desarrollo de enfermedades en cultivos, como la roya en el café [126], [127], entre otras aplicaciones.

Considerando el creciente interés de la comunidad científica en los retos que el aprendizaje automático presenta, es importante realizar un análisis general de dicho campo de investigación. Un enfoque interesante para ese propósito es el análisis bibliométrico, ya que permite generar una forma diferente de visualizar la información y así familiarizarse rápidamente con el estado actual del conocimiento alrededor del campo de interés de una investigación [128]. Realizar un análisis bibliométrico puede contribuir al progreso de un estudio en diferentes maneras, dado que permite identificar fuentes confiables de publicaciones científicas, establecer bases académicas para evaluar nuevos desarrollos, identificar autores relevantes y crear una visión general de los trabajos realizados alrededor de un campo de investigación [129]. Este enfoque ha sido utilizado ya en diversas investigaciones, [128] y [130] llevan a cabo un análisis bibliométrico en el campo de la medicina, mientras que [129] y [131] lo hacen para conocer los trabajos realizados en el área de comportamiento social y sistemas de transporte inteligentes, respectivamente.

A continuación, se presentan la metodología utilizada y los resultados obtenidos a través del análisis bibliométrico llevado a cabo para conocer el estado actual del conocimiento en el campo de aprendizaje automático.

Metodología

En esta sección se presenta el proceso llevado a cabo en la investigación para construir una visión general de los trabajos realizados alrededor del campo de aprendizaje automático. Para esto, se describe el conjunto de datos utilizado, los parámetros configurados en Citespace y SciMAT y, finalmente, el desarrollo del análisis bibliométrico como tal.

I. Datos utilizados para el análisis

Scopus y Web of Science (WoS) fueron tomadas como las dos bases de datos bibliográficas primarias para la obtención de los artículos publicados en revistas científicas y que luego serían la base para el análisis bibliométrico realizado. En las bases de datos mencionadas, se buscaron artículos y conferencias publicadas alrededor del campo de aprendizaje

automático en el lapso comprendido entre los años 2007 y 2017. Se obtuvieron 41 962 registros en WoS y 61 475 en Scopus. La Figura anexos 1 muestra el número de artículos publicados por año en las dos bases de datos y la tendencia de crecimiento en las investigaciones relacionadas con el aprendizaje automático.

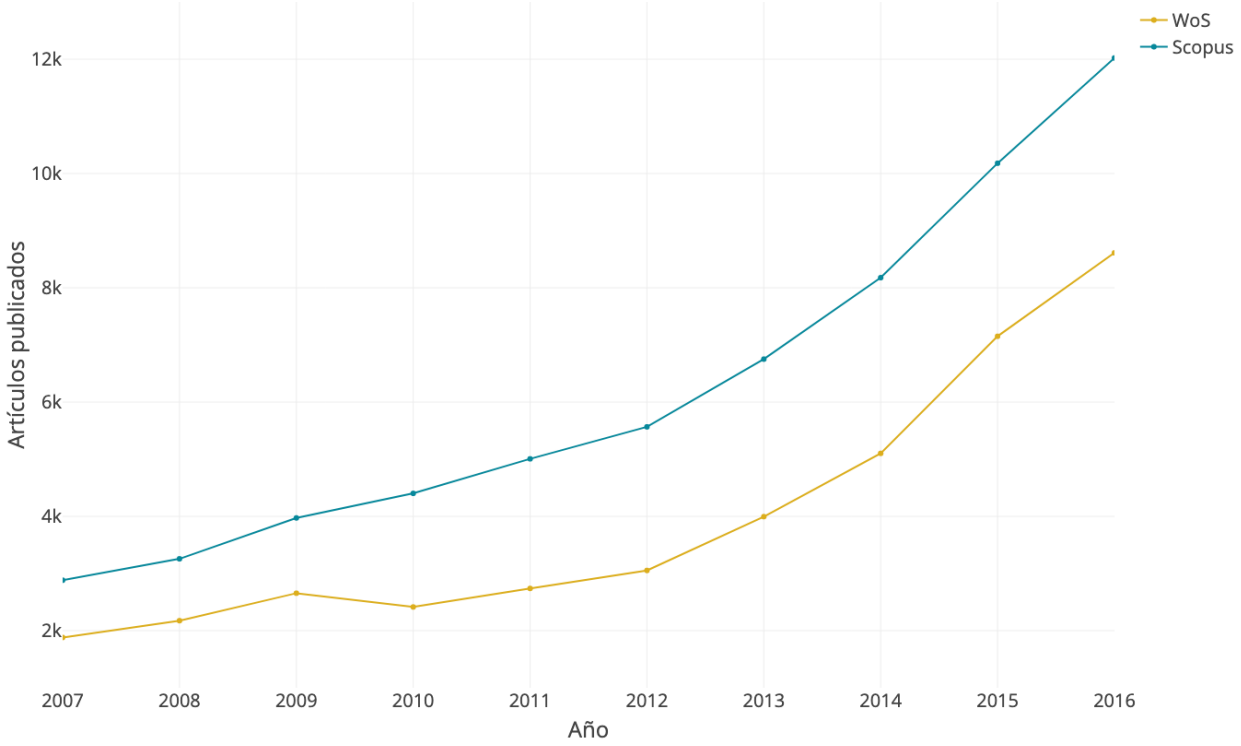


Figura anexos 1. Número de artículos obtenidos de Scopus y WoS con su respectivo año. Fuente propia.

II. Configuración de las herramientas

El análisis bibliométrico fue llevado a cabo utilizando las herramientas Citespace [132]–[134] y SciMAT [135]. Como primera medida, se realizó la configuración de Citespace. Para esta herramienta fueron tomados los registros de WoS como insumo principal y se agruparon en cortes de tiempo de un año. Además, el título, resumen y palabras clave de los artículos, fueron seleccionados como fuentes de términos para el análisis. Por último, se cambió el tamaño de las redes generadas, con el fin de observar con más detalle los términos relevantes. El número de nodos utilizado en cada red se puede observar en su respectivo análisis.

En segundo lugar, se utilizaron los registros obtenidos de Scopus para llevar a cabo el análisis con SciMAT. Las palabras clave fueron tomadas como unidad de análisis y procesadas para eliminar sinónimos y duplicados. Además, se dividieron los artículos en seis diferentes grupos: 2007-2009, 2010-2012, 2013-2014, 2015, 2016 y 2017, con el objetivo de tener grupos más pequeños y con un número similar de registros que permitiesen tener

una visión más detallada de los temas que han cobrado relevancia en las investigaciones alrededor del aprendizaje automático en los últimos años. Finalmente, se llevó a cabo el análisis con la siguiente configuración: inclusión de todos los periodos, palabras de los autores como unidad de análisis, co-ocurrencia como el tipo de red, fuerza de asociación como la medida de normalización, máximo 7 y mínimo 5 nodos para la generación de redes e índice h y suma de citaciones como las medidas de calidad.

III. Desarrollo del análisis

Citespace es una herramienta que permite visualizar patrones transitorios y tendencias en la literatura científica [133]. En el análisis llevado a cabo en la presente investigación, se exploraron tres técnicas bibliométricas disponibles entre sus características. La primera, análisis de autores, para investigar los principales escritores alrededor de un tema en específico [136] y que utiliza sus nombres, países de afiliación e instituciones como unidades de análisis. En segundo lugar, el análisis de palabras para establecer vínculos entre documentos [137], a través de co-ocurrencias de palabras clave y categorías. Por último, el análisis de citas, que proporciona como resultado el autor citado, las citas y las co-ocurrencias de revistas citadas.

SciMAT, por su parte, permite generar una representación espacial de la relación entre disciplinas, campos, especialidades, autores y documentos [129] a través de la implementación de un marco de referencia longitudinal propuesto por [138], [139], que toma como base el índice h de los autores y las co-ocurrencias de palabras clave en los diferentes artículos alrededor de un tema específico. En SciMAT, el análisis de co-ocurrencias de palabras proporciona en primer lugar información sobre los temas de un campo de investigación y, en segundo lugar, permite analizar y seguir la evolución de un campo de estudio a lo largo de períodos de tiempo consecutivos [140]. El índice h se utiliza para medir el impacto de los diferentes autores y áreas temáticas identificadas [129]. Siguiendo el trabajo propuesto en [138], se crearon diagramas estratégicos para ver la importancia de diferentes temas en las investigaciones alrededor del campo de aprendizaje automático y se construyeron clústeres que muestran la relación entre los mismos. Por una parte, cobran relevancia los conceptos de centralidad y densidad [141]. La centralidad mide la intensidad de la interacción de un tema con los otros; si la relación es fuerte, entonces el vínculo que los une en el clúster será más grande. La densidad mide la intensidad de los enlaces internos dentro del grupo; si un tema tiene mayor densidad, entonces el nodo que lo representa tendrá un diámetro mayor. Por otra parte, tomando como referencia los dos conceptos mencionados anteriormente, los temas que aparecen en los diagramas estratégicos pueden ser clasificados en cuatro grupos, tal como se observa en la Figura anexos 2.

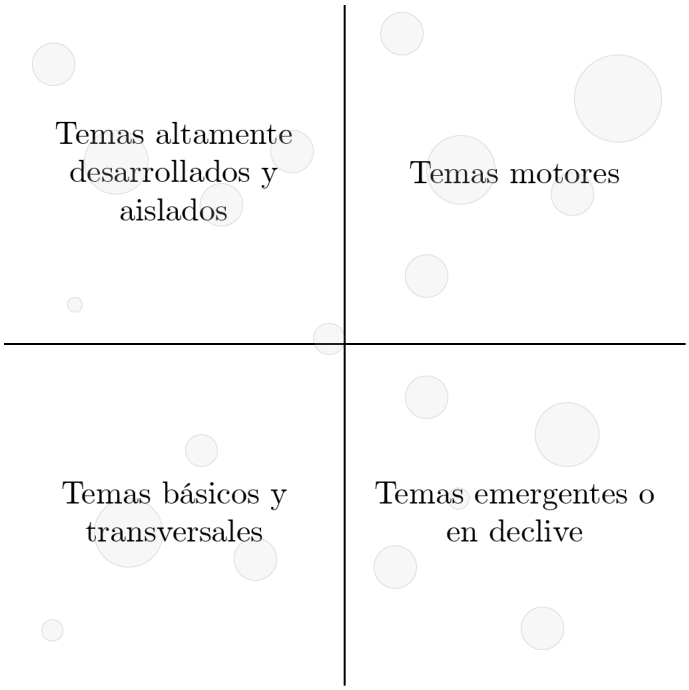


Figura anexos 2. Disposición de los temas en los diagramas estratégicos de SciMAT. Adaptado de [129].

En la Figura anexos 2, se presentan los cuatro grupos posibles en los que pueden ser clasificados los temas que hacen parte de los diagramas estratégicos generados en SciMAT [129], [138]:

- ***Motores***: aquí se agrupan los temas que se encuentran bien desarrollados y que son cruciales para el campo de investigación.
- ***Básicos y transversales***: temas que no están desarrollados suficientemente pero que son significativos para el campo de investigación.
- ***Emergentes o en declive***: se refieren a los temas que apenas empiezan a ser desarrollados en el campo estudiado o que ya han sido desarrollados y están empezando a perder importancia para el mismo.
- ***Altamente desarrollados y aislados***: temas que se encuentran estudiados en gran medida y tienen una relación débil con el campo de investigación.

Resultados

En la presente sección se exponen los resultados más relevantes del análisis bibliométrico llevado a cabo con las publicaciones obtenidas alrededor del aprendizaje automático. Es importante resaltar que, tanto el análisis como los resultados alcanzados, son más extensos que lo plasmado en esta; diferentes redes y diagramas estratégicos, así como una discusión más profunda sobre los resultados obtenidos, pueden encontrarse en [142] para Citespace

y [143] para SciMAT. Con el objetivo de diferenciar los resultados obtenidos en cada herramienta, la presente sección se encuentra dividida en dos subsecciones.

I. Resultados del análisis con Citespace

Con el objetivo de encontrar los principales temas que se abordan a través del aprendizaje automático, se llevó a cabo un análisis inicial de las categorías de las que hacen parte todos los registros. La **Figura anexos 3** muestra la red construida con las co-ocurrencias de categorías en los documentos, configurando el número máximos de nodos igual a 50.

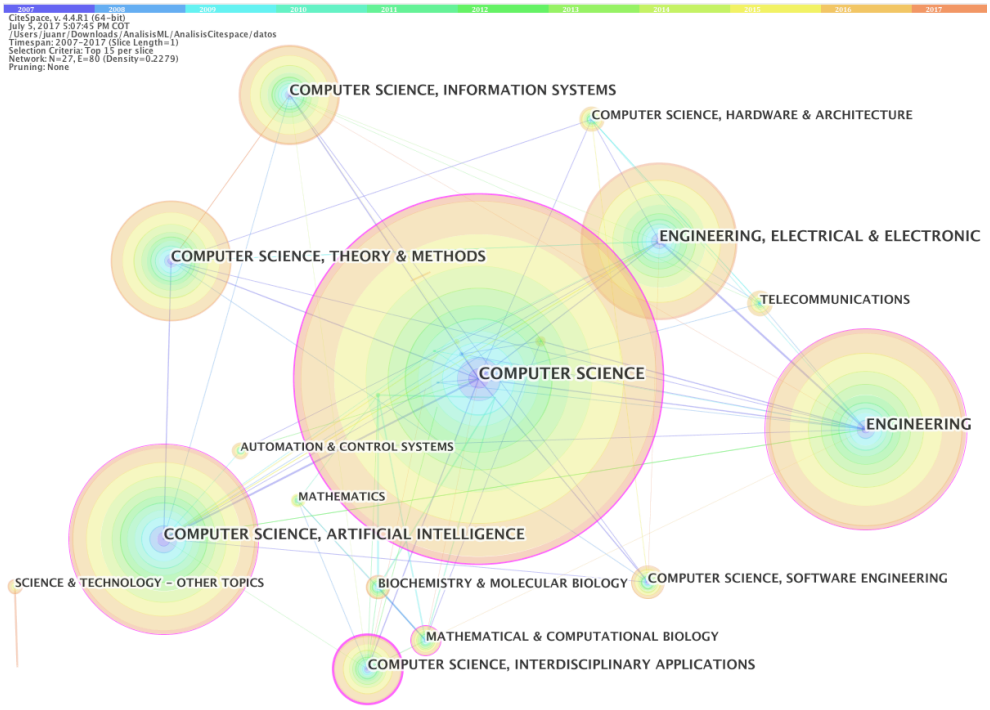


Figura anexos 3. Red construida a partir de las co-ocurrencias de categorías en Citespace. Fuente propia.

La red presentada en la **Figura anexos 3** está conformada por 27 nodos que representan las categorías más comunes en las que están enmarcados todos los registros analizados. Como era de esperarse, aparecen categorías transversales que se relacionan con toda investigación realizada alrededor del campo de aprendizaje automático; ciencias de la computación, métodos y teorías en dicho campo, ingeniería, matemáticas y sistemas de información, son algunos de los temas generales que pueden verse en la red. Entre los nodos que hacen parte de la red generada, se destacan "*Computer science – interdisciplinary applications*", "*Mathematical and Computational biology*" y "*Biochemistry and Molecular biology*", con una centralidad de 0.47, 0.18 y 0.12 respectivamente, demostrando que las técnicas de aprendizaje automático pueden ser utilizadas para crear diversas soluciones en un gran número de campos de aplicación, resaltando biología y

bioquímica como dos de las categorías en las que más publicaciones han sido realizadas durante el periodo analizado en la presente investigación.

Con el objetivo de complementar el análisis mencionado anteriormente (las categorías de las que hacen parte los registros obtenidos), se llevó acabo uno de palabras clave. Dicho análisis permite observar las tendencias emergentes, proporcionando información sobre el contenido de los artículos publicados sobre aprendizaje automático. La **Figura anexos 4** muestra la red generada a partir del análisis de palabras clave, configurando el número máximos de nodos igual a 50.

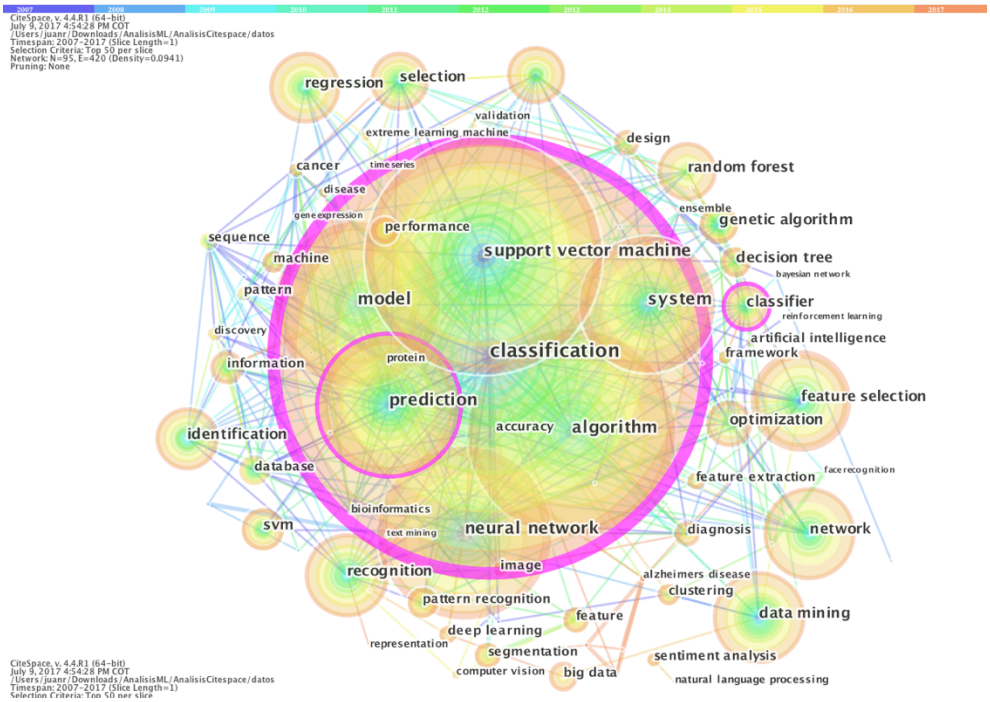


Figura anexos 4. Red construida a partir de las co-ocurrencias de palabras clave en Citespace. Fuente propia.

La red que se observa en la **Figura anexos 4** muestra la relevancia de diferentes palabras clave en todas las publicaciones encontradas alrededor de aprendizaje automático. La palabra clave con el mayor número de citaciones que aparece en la red es "*classification*", con un total de 5546, seguida por "*support vector machine*" (3347), "*algorithm*" (2681), "*neural network*" (2450), "*model*" (2253), "*system*" (1898), "*prediction*" (1893), "*feature selection*" (1559), "*data mining*" (1335), "*network*" (1304), "*recognition*" (1283), "*regression*" (1110), "*artificial neural network*" (1048) y "*random forest*" (971). A partir de estas palabras clave, se puede observar que los algoritmos de clasificación, como máquinas de vectores de soporte y algunos tipos de redes neuronales artificiales, se han estudiado ampliamente y representan un importante punto de inflexión, actuando como puentes que vinculan conceptos en diferentes períodos. Otros algoritmos relevantes son los utilizados

para fines de regresión, como las redes neuronales o los árboles de regresión y los utilizados para propósitos de agrupación, como k vecinos más cercanos.

Por último, tomando como base la red de co-ocurrencias de palabras clave presentada en la **Figura anexos 4**, se aplicó un filtro para eliminar los conceptos y temas inherentes a toda investigación relacionada con el aprendizaje automático; nodos como "*data*", "*information*", "*classification*" o " *random forest*", fueron filtrados para poder ver las palabras clave más importantes en las investigaciones analizadas y que no necesariamente se relacionan con un proyecto de aprendizaje automático. La **Figura anexos 5** muestra la red obtenida a partir del filtrado mencionado.

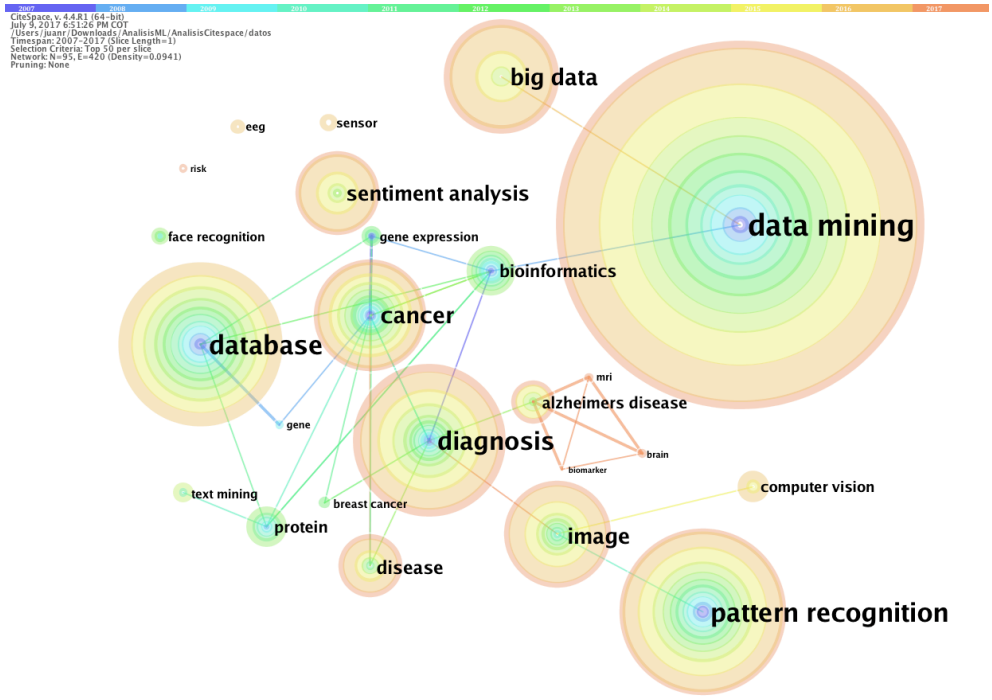


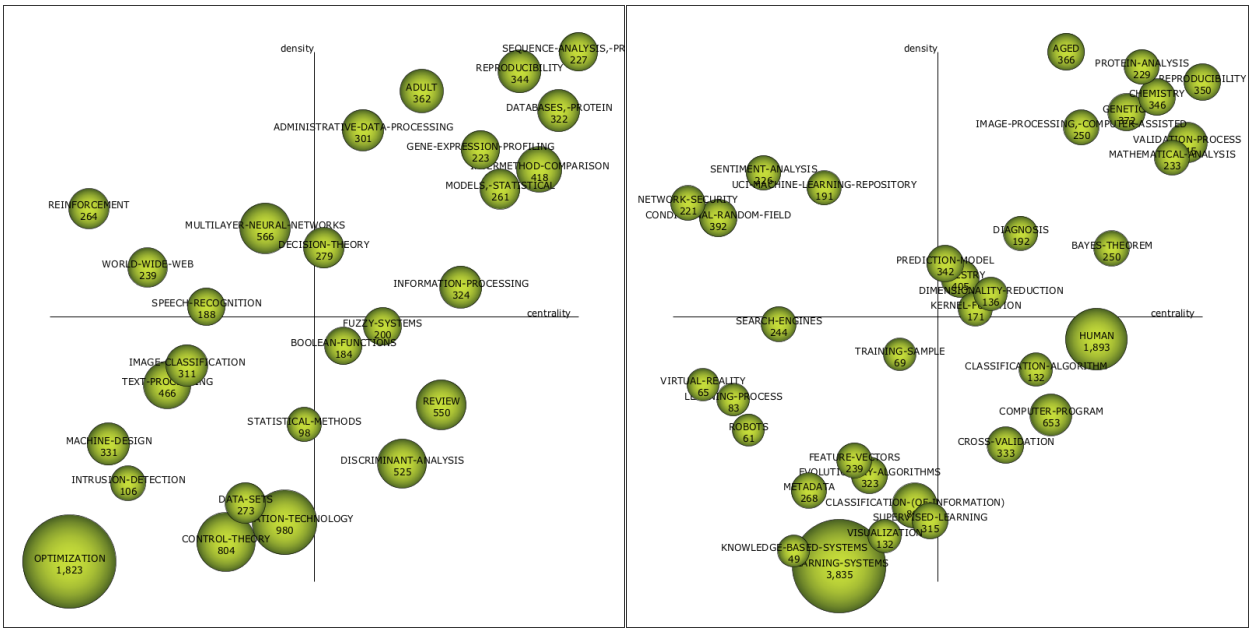
Figura anexos 5. Red construida a partir del filtrado de palabras clave en Citespace. Fuente propia.

Las palabras clave más importantes que aparecen en la red observada en la **Figura anexos 5**, tomando como criterio el número de citaciones, son "*data mining*" (1335), seguida de "*pattern recognition*" (652), "*database*" (624), "*diagnosis*" (599), "*cancer*" (449) y "*big data*" (420). Otras palabras clave relevantes son "image" (414), "*sentiment analysis*" (325), "*disease*" (240), "*bioinformatics*" (209), "*Alzheimer's disease*" (188), "*protein*" (170) and "*computer vision*" (131). Dicha distribución de los nodos permite observar que la minería de datos es un tema importante en los trabajos publicados y que el aprendizaje automático está cobrando cada vez más importancia en el campo de la salud, sobretodo para el diagnóstico de enfermedades como el cáncer o el Alzheimer, mediante el uso de bases de datos construidas a partir de diferentes fuentes, tales como electroencefalogramas o múltiples sensores. Además de esto, el aprendizaje automático ha sido (y continúa siendo)

una herramienta relevante para la extracción de información en diferentes campos de aplicación a través del análisis de imágenes.

II. Resultados del análisis con SciMAT

En busca de identificar los temas de investigación que han sido determinantes en los trabajos realizados alrededor del aprendizaje automático durante la última década, se realizó un análisis de diagramas estratégicos en SciMAT. Como se mencionó en la sección 0, el total de los registros que hacían parte del conjunto de datos fueron divididos en seis periodos con un número similar de investigaciones; los grupos resultantes de la división realizada son los siguientes: 2007-2009, 2010-2012, 2013-2014, 2015, 2016 y 2017. Los diagramas estratégicos construidos para los periodos mencionados se dividen en cuatro cuadrantes, tal como se muestra en la sección 0: el cuadrante superior derecho presenta los temas motores, el cuadrante superior izquierdo muestra los temas altamente desarrollados y aislados, el inferior derecho los temas básicos y transversales, mientras que en la parte inferior izquierda se ubican los temas emergentes o en declive. La Figura anexos 6 muestra los diagramas estratégicos construidos en SciMAT para los seis periodos indicados anteriormente.



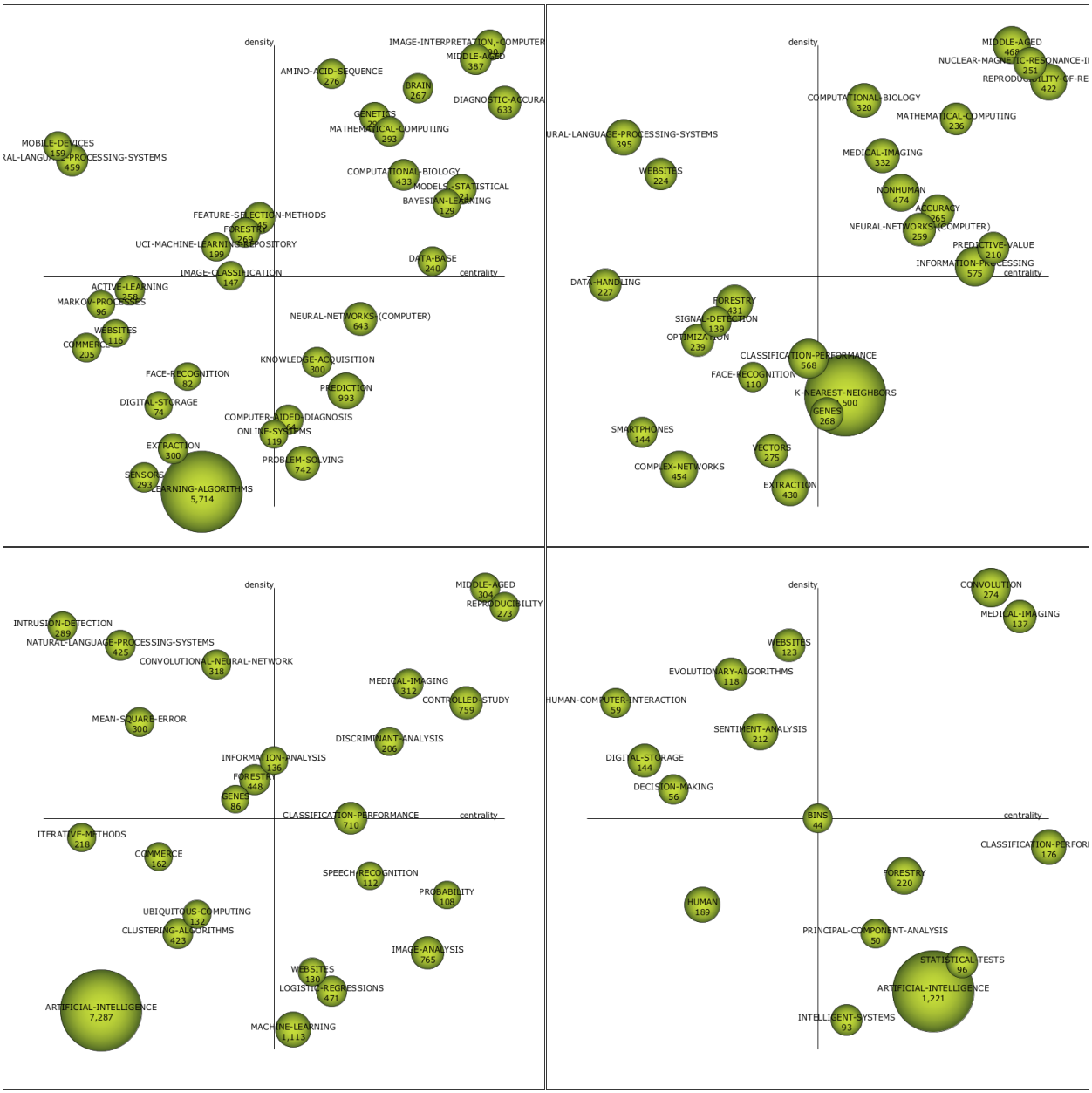


Figura anexos 6. Diagramas estratégicos construidos para los periodos 2007-2009, 2010-2012, 2013-2014, 2015, 2016 y 2017 en SciMAT. Fuente propia.

Los diagramas estratégicos presentados en la Figura anexos 6 muestran los temas de investigación más relevantes durante los años comprendidos entre el 2007 y el 2017. El análisis de los diagramas permite identificar que temas relacionados con sitios web y el almacenamiento de datos (metadatos, bases de datos, gestión de datos, entre otros) fueron básicos (hasta finales del 2014) en las investigaciones en aprendizaje automático pero ya se encuentran en una etapa de alto desarrollo, por lo que han empezado a perder relevancia en nuevas publicaciones; esto no quiere decir que, los conjuntos de datos por ejemplo, no sean necesarios para desarrollar un proyecto de aprendizaje automático, sino que son

temas poco destacados en las publicaciones actuales. Por otra parte, los resultados obtenidos en los diagramas estratégicos permiten destacar la evolución de las redes neuronales artificiales, que fueron altamente desarrolladas en el periodo 2007-2009 y tema motor en el año 2015, pero en los dos últimos años las investigaciones se centraron en un tipo específico de estas, las redes neuronales artificiales convolucionales, que fueron altamente desarrolladas en el 2016 y un tema motor en el 2017. En el mismo sentido, se resalta el uso de aprendizaje automático en temas relacionados con el procesamiento, clasificación e interpretación de imágenes, que fueron básicos en las investigaciones publicadas en el lapso 2007-2009 y motores desde el 2010 hasta el 2014, pero desde dicho año cobró importancia un campo específico del procesamiento de imágenes, las imágenes médicas, que fueron un tema motor durante los últimos tres años (2015-2017). En contraste, el procesamiento de texto y los sistemas de lógica difusa, fueron temas altamente desarrollados entre el 2007 y el 2009, pero disminuyeron su importancia y no aparecen como ejes centrales de las investigaciones publicadas durante todos los demás periodos. Lo mismo sucede con temas relacionados con genética y el análisis de proteínas y aminoácidos, que fueron motores durante varios años (desde el 2007 hasta el 2014) pero no hacen parte de los focos de las publicaciones en aprendizaje automático de los años analizados. Por último, se destacan diversos temas como comercio electrónico, reconocimiento facial, realidad virtual y computación ubicua, además de varias técnicas de aprendizaje supervisado como *decision trees, random forest* y *adaptive boosting*, que han sido básicos y transversales en gran parte de las investigaciones en aprendizaje automático en los últimos años (2014 en adelante) y que seguirán siendo tendencia en un futuro cercano.

B Mapas región cafetera y distribución de lluvia

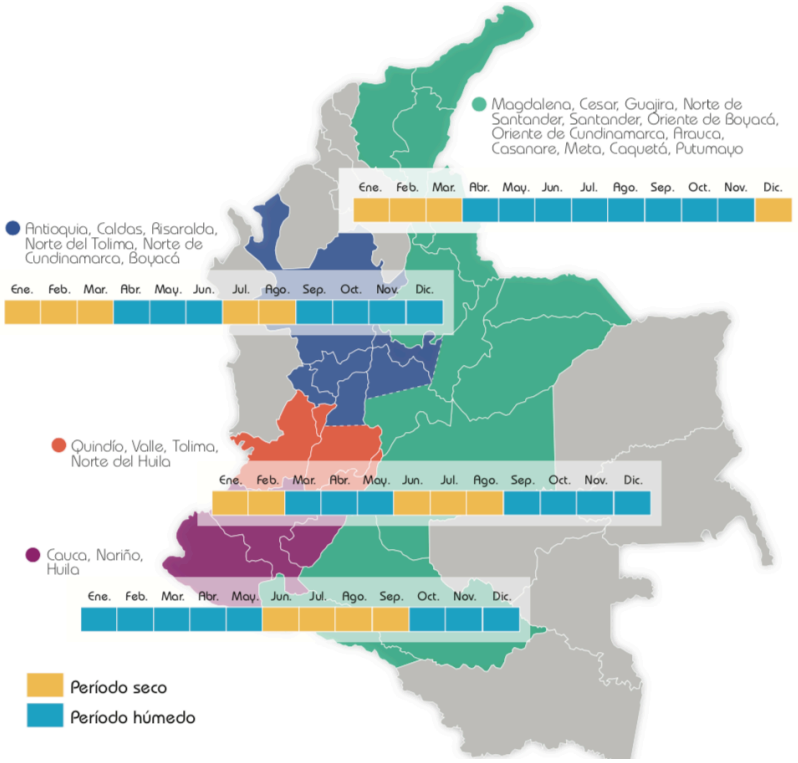


Figura anexos 7. Mapa de distribución de lluvia en Colombia. Obtenido de [66].

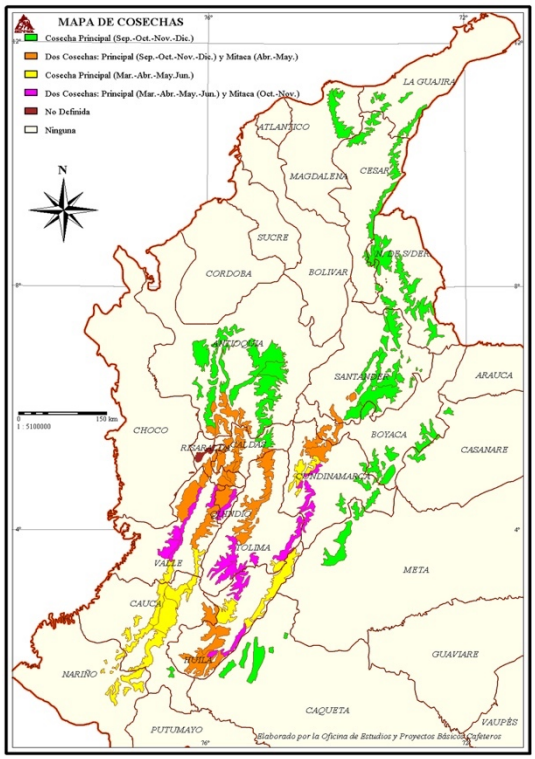


Figura anexos 8. Mapa regiones cafeteras colombianas. Obtenido de [71].

C Métricas de desempeño de los modelos entrenados

Nº	Instancias	Atributos	Rendimiento			Producción			Ln(Producción)		
			CC	MAE	RMSE	CC	MAE	RMSE	CC	MAE	RMSE
1	Todas	Área sembrada, municipio, departamento, altitud, floración, cosecha, mitaca, clima - floración a cosecha	0.6437	0.1943	0.2438	-	-	-	-	-	-
2	Todas	Área sembrada, clima - floración a cosecha	0.4406	0.224	0.2887	-	-	-	-	-	-
3	Todas	Municipio, clima - floración a cosecha	0.6343	0.1947	0.2453	-	-	-	-	-	-
4	Todas	Área sembrada, clima - floración a cosecha (Bri_Sol_h_M4, Pre_Tot_mm_M5, T_Max_Med_M9, específicamente)	0.2977	0.2325	0.299	-	-	-	-	-	-
5	Todas	Municipio, clima - floración a cosecha (Bri_Sol_h_M4, Pre_Tot_mm_M5, T_Max_Med_M9, específicamente)	0.663	0.188	0.2348	-	-	-	-	-	-
6	Todas	Clima - año cafetero	0.4819	0.2182	0.2791	0.3969	2171.16	2874.160	0.5259	0.775	0.9986
7	Todas	Área sembrada, clima - año cafetero	0.5091	0.2113	0.2725	0.8706	953.139	1503.116	0.9404	0.27	0.39
8	Todas	Municipio, clima - año cafetero	0.6633	0.1888	0.2362	0.8571	990.682	1588.741	0.9133	0.3121	0.4669
9	Todas	Área sembrada, municipio, clima - año cafetero	0.6533	0.1873	0.24	0.8808	922.955	1440.879	0.944	0.2717	0.3771
10	Todas	Clima - primeros seis (6) meses del año cafetero	0.4688	0.218	0.2777	0.3818	2199.71	2877.728	0.4896	0.7711	1.0104
11	Todas	Área sembrada, clima - primeros seis (6) meses del año cafetero	0.5146	0.2132	0.2684	0.8845	915.972	1421.178	0.9404	0.2711	0.3903
12	Todas	Municipio, clima - primeros seis (6) meses del año cafetero	0.672	0.1884	0.2322	0.8701	1005.87	1505.702	0.9204	0.2907	0.4478
13	Todas	Área sembrada, municipio, clima - primeros seis (6) meses del año cafetero	0.6636	0.1901	0.2353	0.8912	896.328	1381.920	0.9473	0.2616	0.3665
14	Periodo cosecha - Oct -> Dic	Área sembrada, clima - año cafetero	0.6423	0.2209	0.2853	0.9105	753.038	1200.019	0.9505	0.2869	0.3778
15	Periodo cosecha - Abr -> Jun	Área sembrada, clima - año cafetero	0.2305	0.1995	0.2602	0.8739	976.836	1495.865	0.8806	0.2941	0.4756
16	Región mapa cafetero (verde)	Área sembrada, clima - año cafetero	0.4082	0.2416	0.3048	0.9406	504.529	750.2183	0.9352	0.3444	0.4351
17	Región mapa cafetero (naranja)	Área sembrada, clima - año cafetero	0.2996	0.226	0.3008	0.8237	1212.12	1790.003	0.9093	0.2571	0.3459

18	Región mapa cafetero (amarilla)	Área sembrada, clima - año cafetero	0.2361	0.1753	0.2501	0.841	906.195	1177.497	0.8796	0.285	0.3691
19	Región mapa cafetero (morada)	Área sembrada, clima - año cafetero	0.558	0.1655	0.2043	0.9659	500.582	766.2367	0.7981	0.416	0.6638
20	Región distribución de lluvia (azul)	Área sembrada, clima - año cafetero	0.272	0.2514	0.3303	0.885	1017.20	1527.387	0.9304	0.2934	0.3704
21	Región distribución de lluvia (morada)	Área sembrada, clima - año cafetero	0.2595	0.1429	0.2044	0.7591	1249.87	1851.029	0.8385	0.233	0.3237
22	Región distribución de lluvia (verde)	Área sembrada, clima - año cafetero	0.2723	0.1716	0.222	0.9416	368.131	588.7998	0.9176	0.3134	0.5196
23	Región distribución de lluvia (naranja)	Área sembrada, clima - año cafetero	0.1247	0.1989	0.2406	0.9168	785.394	1103.855	0.9305	0.2294	0.2876
24	Todas	Área sembrada, clima - primeros seis (6) meses del año cafetero (CfsSubsetEval: Hum_Rel_M1, Pre_Tot_M1, Dias_Llu_M1, Hum_Rel_M2, T_Max_Med_M3, Hum_Rel_M3, Bri_Sol_M3, Hum_Rel_M4, Bri_Sol_M4, Hum_Rel_M5, Pre_Tot_M5, Dias_Llu_M5, Bri_Sol_M5, Hum_Rel_M6, Pre_Tot_M6, Dias_Llu_M6, Bri_Sol_M6)	0.3921	0.2243	0.2896	0.8796	862.077	1446.657	0.9434	0.2748	0.3806
25	Todas	Área sembrada, clima - primeros seis (6) meses del año cafetero (CorrelationAttributeEval: Hum_Rel_M1, Dias_Llu_M1, Hum_Rel_M2, T_Min_Med_M3, Hum_Rel_M3, Bri_Sol_M3, T_Min_Med_M4, Hum_Rel_M4, T_Min_Med_M5, Hum_Rel_M5, Pre_Tot_M5, Dias_Llu_M5, T_Max_Med_M6, T_Med_M6, Pre_Tot_M6,	0.4057	0.2219	0.2862	0.8778	879.831	1459.706	0.9453	0.2686	0.3739

		Dias_Llu_M6, Bri_Sol_M6)									
26	Todas, exceptuando las instancias con valores perdidos (total = 363)	Área sembrada, clima - primeros seis (6) meses del año cafetero	0.397	0.2317	0.2959	0.8699	915.129	1496.089	0.9418	0.274	0.3846

Tabla anexos 1. Atributos, instancias y métricas de desempeño de los modelos entrenados. Fuente propia.

D Porcentajes de similitud del sistema de recomendaciones

Condición evaluada	Recomendación	Porcentaje especialista	Porcentaje sistema
T1 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	100	100
T3 P3 H3	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	50	33.33
T1 P3 H3	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	66.67
T2 P2 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	100	33.33
T1 P3 H4	Se recomienda establecer cultivos a libre exposición e implementar bajas densidades de siembra, para evitar una posible proliferación de enfermedades	100	100
T2 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	66.67
T3 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	66.67
T4 P1 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	75	100
T3 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	75	66.67
T1 P2 H3	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	33.33
T4 P3 H3	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo	25	66.67
T1 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H1	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	100

T1 P3 H2	Se recomienda establecer un cultivo a libre exposición y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	50	100
T4 P3 H5	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo y reducir la distancia de siembra para evitar proliferación de entes patógenos	25	100
T1 P1 H5	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	66.67
T4 P2 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	66.67
T4 P3 H4	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo	25	100
T4 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	100

Tabla anexos 2. Porcentaje de similitud entre las condiciones climáticas y las sugerencias generadas utilizando la similitud coseno. Fuente propia.

Condición evaluada	Recomendación	Porcentaje especialista	Porcentaje sistema
T1 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	100	100
T3 P3 H3	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	50	18.35
T1 P3 H3	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	42.26
T2 P2 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	100	18.35
T1 P3 H4	Se recomienda establecer cultivos a libre exposición e implementar bajas densidades de siembra, para evitar una posible proliferación de enfermedades	100	100
T2 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	42.26
T3 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de	100	42.26

	transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta		
T4 P1 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	75	100
T3 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	75	42.26
T1 P2 H3	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	18.35
T4 P3 H3	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo	25	42.26
T1 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H1	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	100
T1 P3 H2	Se recomienda establecer un cultivo a libre exposición y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	50	100
T4 P3 H5	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo y reducir la distancia de siembra para evitar proliferación de entes patógenos	25	100
T1 P1 H5	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	42.26
T4 P2 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	42.26
T4 P3 H4	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo	25	100
T4 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	100

Tabla anexos 3. Porcentaje de similitud entre las condiciones climáticas y las sugerencias generadas utilizando la distancia euclidiana. Fuente propia.

Condición evaluada	Recomendación	Porcentaje especialista	Porcentaje sistema
T1 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	100	100
T3 P3 H3	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	100	38.76
T1 P3 H3	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	50	64.64
T2 P2 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	100	40.59
T1 P3 H4	Se recomienda establecer cultivos a libre exposición e implementar bajas densidades de siembra, para evitar una posible proliferación de enfermedades	100	100
T2 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	56.7
T3 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	46.55
T4 P1 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	75	100
T3 P3 H5	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	75	62.2
T1 P2 H3	Se recomienda implementar cultivos a libre exposición, establecer bajas densidades de siembra para evitar una posible proliferación de enfermedades y evaluar el grado de drenaje del suelo para evitar sobresaturaciones de agua	25	42.31
T4 P3 H3	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo y reducir la distancia de siembra para evitar proliferación de entes patógenos	50	63.49
T1 P3 H1	Se recomienda establecer un cultivo a libre exposición, prestar atención a las condiciones de drenaje del suelo, si observa sobresaturación de agua, es necesario realizar adaptaciones del terreno con el fin de mejorar el flujo de agua y evitar problemas de exceso hídrico para las plantas y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H1	Se recomienda establecer el cultivo a libre exposición, considerando que una baja temperatura reduce la pérdida potencial por evapotranspiración, de esta forma se facilitaría un mayor aprovechamiento de la energía solar para el crecimiento vegetativo o producción de la planta	100	100
T1 P3 H2	Se recomienda establecer un cultivo a libre exposición y hacer una revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros	100	100
T1 P1 H2	Se recomienda establecer un sistema productivo a libre exposición o con bajo porcentaje de sombrío	50	100

T4 P3 H5	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo y reducir la distancia de siembra para evitar proliferación de entes patógenos	25	100
T1 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	25	73.27
T4 P2 H4	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo) y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	56.7
T4 P3 H4	Se recomienda evaluar el grado de pendiente del terreno, en caso de ser muy elevada, establecer taludes o estructuras de contención para evitar la erosión del suelo	25	100
T4 P1 H5	Se recomienda implementar un arreglo productivo agroforestal, mantener la cobertura vegetal del suelo, establecer distancias de siembra menores para reducir la evapotranspiración (densidades muy altas pueden incrementar el proceso de transpiración de las plantas, ya que el agua puede quedar retenida en sus hojas y no llegar al suelo), implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua y poner atención al control de arvenses, ya que estas incrementan el consumo de agua limitando la disponibilidad para la planta	100	100

Tabla anexos 4. Porcentaje de similitud entre las condiciones climáticas y las sugerencias generadas utilizando la distancia Mahalanobis. Fuente propia.

E Recomendaciones generales para el manejo del cultivo de café

Recomendación	Tema
Para que el desarrollo del fruto del cafeto sea normal se requiere disponibilidad de agua en el suelo durante los ocho meses comprendidos entre la floración y la cosecha, con un período crítico entre las semanas 8 y 16, en el cual se define el tamaño del fruto. Por esto, se recomienda implementar soluciones que permitan tener una disponibilidad de agua constante durante los meses secos, tales como captación de agua lluvia, reservorios de agua y sistemas de cosecha de agua.	Suelo
Durante los meses húmedos, para disminuir el efecto del exceso de lluvia se recomienda tomar acciones relacionadas con la conservación del suelo y prácticas para drenar los excedentes de agua del suelo. Los lotes de la finca que presenten encharcamientos deben drenarse por medio de zanjas y utilizando estructuras de bioingeniería, como filtros vivos.	Suelo
Las arvenses ejercen una cobertura que protege el suelo de los impactos directos de las gotas de lluvia, disminuyendo la erosión superficial, por lo cual se deben realizar cortes altos con machete o guadaña. Nunca se debe hacer uso del azadón, debido a que se descubre el suelo y promueve la erosión.	Arvenses
El sombrío en los cafetales favorece la conservación del agua, especialmente en los períodos de sequía prolongados.	Generalidades
Aunque el sombrío es de gran ayuda para el cultivo de café, es necesario regularlo, puesto que si el cafetal está bajo sombrío excesivo puede generarse una humedad sobreabundante que favorece el incremento de enfermedades como el mal rosado, roya y gotera, entre otras.	Fitosanitario
De manera permanente, sea durante los meses secos o los húmedos, es necesaria la revisión periódica de los cafetales para evaluar la presencia de broca, roya, mal rosado y llagas radicales, entre otros.	Fitosanitario
Durante la época seca se deben controlar las arvenses, especialmente en el plato del árbol, debido a que éstas consumen gran cantidad de agua del suelo.	Arvenses
Para el café recién sembrado se recomienda cubrir el plato del árbol con coberturas provenientes de las desyerbas, podas o de otros residuos, para conservar la humedad en el suelo.	Generalidades
Las labores de soqueo se recomienda realizarlas durante los meses secos.	Generalidades
La fertilización debe realizarse cuando se generalicen las lluvias y cuando el suelo se encuentre bien húmedo, ya que bajo esta condición las raíces pueden interceptar mayor cantidad de nutrimentos.	Nutrición
Para un buen manejo de la broca del café deben cosecharse los frutos secos y maduros de los árboles, y si es necesario del suelo, una vez hayan finalizado los períodos de cosecha principal y mitaca. El repase se debe realizar después de 2 a 3 semanas de realizado el último pase de cosecha, que generalmente coincide con el período crítico del ataque de la broca.	Fitosanitario

El café cosechado debe empacarse en sacos o recipientes limpios, libres de contaminaciones, protegerse de daños físicos y de altas temperaturas y entregarse pronto al beneficiadero, de tal forma que no se sobrefermente, se humedezca o se contamine.	Generalidades
Para los frutos y granos recogidos durante el repase debe implementarse un sistema para su manejo en la finca de tal forma que no se junten con el café cosechado, sino que se empaquen por separado utilizando bolsas o cualquier recipiente.	Generalidades
Es importante realizar la cosecha selectiva de frutos, con menos de 2% de verdes y más de 80% de maduros. El patrón de corte es un elemento clave para garantizar una excelente recolección.	Generalidades
Para evitar la contaminación y pérdida del café, es importante hacer un control riguroso de las sustancias, dosis, época y forma de aplicación de fungicidas, herbicidas e insecticidas.	Fitosanitario y arvenses
Debe establecerse un sistema de monitoreo y control de plagas y enfermedades en los almácigos, y registrar el insecto o patógeno observado, el sistema de control, el nombre del producto, la dosis aplicada para su manejo y la fecha de aplicación del producto.	Fitosanitario
En cuanto a la erosión del suelo, deben adoptarse técnicas de cultivo adecuadas, tales como: la aplicación de coberturas muertas o 'mulch', las siembras a través de la pendiente, la construcción de drenajes, el manejo de coberturas vivas, la aplicación de abonos orgánicos y el establecimiento de árboles y arbustos en las orillas de caminos, carreteras y fuentes de agua, entre otras.	Suelo
Debe velarse por el mantenimiento de la fertilidad del suelo mediante la aplicación de fertilizantes (orgánicos e inorgánicos). Sin embargo, la cantidad suministrada no debe exceder las necesidades del cultivo.	Nutrición
Durante eventos climáticos como La Niña, el exceso de humedad en el suelo y en el ambiente favorece la presencia de enfermedades como el mal rosado y la roya del café e incremento en la ocurrencia de llagas radicales, por esto es importante llevar a cabo buenas prácticas de drenaje durante eventos adversos como el mencionado.	Fitosanitario
En las regiones con períodos secos muy prolongados no conviene establecer cultivos en altas densidades de siembra, debido a que en los cortos periodos de lluvia se aumenta la cantidad de agua interceptada por la parte aérea de las plantas, disminuyendo la cantidad que llega a la superficie del suelo, afectando significativamente el rendimiento.	Generalidades
En la etapa de germinación no se requiere la adición de nutrimentos ya que la semilla contiene todas las sustancias necesarias para su desarrollo y solamente necesita condiciones adecuadas de humedad, oscuridad y temperatura.	Nutrición
El empleo de pulpa descompuesta en las bolsas del almácigo permite obtener plantas vigorosas y sanas.	Nutrición
No es recomendable el uso de fertilizantes granulados o foliares durante la etapa de almácigo, ya que estos no sustituyen los efectos benéficos de la pulpa y se corre el riesgo de intoxicar las plantas.	Nutrición
Durante el crecimiento vegetativo (levante), hay una demanda alta de nitrógeno y fósforo principalmente, aunque otros nutrimentos pueden ser también necesarios, dependiendo de las características del suelo. Las cantidades y fuentes de fertilizante a aplicar dependen de cada tipo de suelo y la mejor forma de determinarlas es mediante el análisis de suelos.	Nutrición

Durante la etapa de producción la planta también debe producir nuevo crecimiento de ramas y hojas para las cosechas futuras, razón por la cual es necesario un buen manejo nutricional. Además del nitrógeno (N) y el fósforo (P) se necesita potasio (K), calcio (Ca), magnesio(Mg), azufre (S) y microelementos como el boro, zinc, manganeso, cobre, hierro y molibdeno. Las cantidades y fuentes de fertilizante que deben aplicarse dependerán de cada tipo de suelo y la mejor forma de determinarlas es mediante el análisis de suelos.	Nutrición
La frecuencia recomendada para aplicar los fertilizantes es dos meses antes del comienzo de la cosecha en cada semestre, es decir dos aplicaciones por año. Puede variar según características del suelo.	Nutrición
En general, no es necesaria la aplicación rutinaria de micronutrientes. Estas deben hacerse solamente en casos muy específicos.	Nutrición
La fertilización foliar no es una buena opción para el campo de los macronutrientes, debido a que para que la planta los aproveche, éstos deben penetrar la cutícula o las estomas de la hoja y luego entrar hasta las células para ser incorporados al metabolismo de la planta. Aunque este método de suministro permite una rápida utilización de los nutrimentos y la corrección de deficiencias en menos tiempo que con las aplicaciones al suelo, la respuesta es solo temporal y serían necesarias muchas aplicaciones.	Nutrición
Para el suministro de los elementos mayores, la fertilización foliar se debe considerar como un complemento y no como sustituto de la fertilización edáfica.	Nutrición
La aplicación de materia orgánica (compost, bocashi, vermicompost o lombrinaza, biofertilizantes, biofermentos y abono verde) favorece las condiciones de calidad del suelo y reduce los requerimientos de fertilizantes inorgánicos, lo que puede generar un mejor rendimiento del cultivo y reducir los impactos negativos sobre el agroecosistema.	Nutrición
En la etapa de almácigo, los principales problemas presentados son: nutrición, mancha de hierro, nematodos, entre otros. Algunos de estos problemas pueden ser controlados con la utilización de materia orgánica para el llenado de bolsas en los almácigos, siendo la principal fuente de esta la pulpa de café descompuesta.	Nutrición
La gallinaza constituye un excelente abono orgánico para almácigos de café, las plantas provenientes de almácigos con gallinaza presentan mejor vigor y desarrollo que las que provienen de almácigos hechos en suelo sin abono.	Nutrición
Algunas estrategias de adaptación de los sistemas de producción a la variabilidad climática, son: establecimiento de árboles para la protección de fuentes de agua, implementación y mantenimiento de reservas naturales, entre otros.	Generalidades
Evitar la aplicación de enmiendas (cales o yeso) de manera generalizada; esto puede generar problemas de alcalinidad para el café y desbalances nutricionales.	Nutrición
El establecimiento de semilleros debe realizarse con suficiente anticipación (7 a 12 meses), de forma tal que la siembra definitiva en campo coincida con el inicio de la época de lluvias.	Generalidades
En cuanto a la siembra, seleccionar plantas sanas y bien nutridas. Estas deben tener de 3 a 4 cruces (ramas laterales). El tamaño de la planta en el momento de trasplante depende del tamaño de bolsa empleada. En algunas zonas, se usan bolsas muy pequeñas	Generalidades

por lo cual el trasplante debe ser rápido. Plantas enfermas, atrasadas y con arquitectura indeseable se deben desechar.	
No se debe usar como fertilizante excremento humano o aguas negras, esto debido al riesgo que implica para la salud y por que se contamina el suelo.	Nutrición
La creación de compostaje puede realizarse con estiércol de animales (porcinos, vacunos, equinos, avícolas, entre otros), pero siempre conociendo la legislación internacional respecto a dicha práctica, puesto que algunos países no permiten la importación de productos que hayan sido fertilizados con esa clase de compostaje.	Nutrición
El uso de trampas para monitoreo (detección y seguimiento) de plagas es una excelente alternativa puesto que, conociendo la plaga y su nivel de incidencia y severidad, se puede llevar a cabo un control efectivo de la misma.	Fitosanitario
Buena labor de preparación del suelo, selección de variedades resistentes, fertilización oportuna y adecuada, manejo adecuado de arvenses y buenas prácticas de recolección de la cosecha, son algunas de las medidas eficaces para el manejo integrado de las plagas.	Fitosanitario
Re-Re es una práctica muy importante que consiste en recoger los frutos de café del suelo y recolectar los que se han quedado en los cafetales una vez finalizada la cosecha. Esta práctica debe ser periódica y permanente y puede reducir en un alto porcentaje las poblaciones de broca en el cultivo al eliminar las fuentes de propagación y alimento del insecto.	Fitosanitario
Hongos entomopatógenos como el Beauveria bassiana, que tiene buen comportamiento en condiciones de alta humedad y temperatura e insectos parasitoides como Cephalonomia stephanoderis, Prorops nasuta y Phymastichus coffea, que son liberados después de realizado el Re-Re, son opciones adecuadas para el control biológico de la broca.	Fitosanitario
Si la temperatura ha aumentado en los últimos años, establecer sombra en el cafetal y diversificar la producción, pueden ser medidas efectivas.	General
Aumentar la materia orgánica, obras de conservación y cosecha de aguas en la finca, aplicaciones foliares más frecuentes durante sequías y sistemas de riego, pueden traer grandes beneficios cuando las lluvias en la zona son irregulares.	General
Si la fertilidad disminuye, puede abordarse mediante la programación de fertilización con base en análisis de suelo, el encalamiento (si el análisis de suelo lo sugiere), la plantación de leguminosas en el cafetal y la aplicación de materia orgánica.	Nutrición
Los sistemas de desyerba que tienen como finalidad desnudar totalmente los suelos (azadón, gala, gambia y herbicidas), originan una disminución permanente de la productividad de los suelos, pérdidas por escorrentía al no regular las aguas y disminución de la biodiversidad genética que conduce a un desequilibrio ecológico y a una agricultura insostenible.	Arvenses
Entre 65 y 75% del control de la broca se hace a partir del Control Cultural, conocido como RE-RE, que consiste en recoger todos los frutos maduros de la plantación y repasar para recoger aquellos que se hayan quedado; la recolección oportuna debe dirigirse a granos maduros, sobre maduros y secos en el árbol y en el suelo, con el fin de romper su ciclo biológico.	Fitosanitario

No se debe hacer la práctica del soqueo durante la época lluviosa, ya que puede incidir en la generación de enfermedades.	Fitosanitario
Más de 35 días sin lluvia pueden ocasionar un estrés hídrico muy alto, lo que traería consigo daños en el cultivo. En caso de presentarse este escenario, es importante tomar medidas que permitan retener la humedad en el suelo.	Generalidades
Si se sobrepasan 20 días por trimestre con un índice de humedad del suelo mayor a 0.5, se generará una menor floración en el cultivo, trayendo consigo una disminución del rendimiento del mismo.	Suelo
Durante la germinación, es importante verificar que las plantas que están en el suelo a utilizar como sustrato no están afectadas por nematodos (abultamientos en las raíces).	Generalidades
Para el manejo de arvenses, como control preventivo, es importante aplicar herbicida a las bolsas con el sustrato antes de sembrar la chapola.	Arvenses
Antes del trasplante al almácigo, remojar la semilla en una solución de 2 gr por litro de MicosPlag (hongo antagonista).	Generalidades
Durante el crecimiento vegetativo, se recomienda revisar presencia de nematodos y cochinillas en las raíces de las plantas cada mes (1 raíz por cada 100 plantas).	Fitosanitario
Aplicar fósforo (materia orgánica o fertilizante químico) durante el crecimiento vegetativo (entre el día 133 y 196 después de la germinación), permite evitar mancha de hierro.	Nutrición
En caso de encontrarse cochinillas aplicar insecticida y repetir dentro de 15 días.	Fitosanitario
Durante la etapa productiva, si la materia orgánica es menor o igual a 8% aplicar 7 gr por planta de nitrógeno, si es mayor a 8% aplicar 5 gr por planta.	Nutrición

Tabla anexos 5. Conjunto de recomendaciones generales para el manejo del cultivo de café. Adaptado de [9], [59], [113], [63], [106]–[112].

F Consentimiento informado

Como una persona experta en computación, agricultura o medioambiente, queremos solicitar tu ayuda para evaluar el prototipo propuesto en la tesis de maestría "Análisis del rendimiento del café basado en técnicas de aprendizaje automático" de la Universidad del Cauca. Esta encuesta está construida con fines de investigación, por lo que la información aquí registrada solo será utilizada en el marco de la tesis mencionada anteriormente. La información capturada en esta encuesta se limita exclusivamente a las respuestas dadas por los encuestados y en ningún momento se captura información confidencial (nombres, documento de identidad, etc.) o datos que no hayan sido brindados por ellos. Así mismo, los encuestados podrán hacer preguntas sobre el proyecto en cualquier momento y conocer los resultados del mismo en futuros artículos. En este orden de ideas, ¿aceptas hacer parte de la evaluación del trabajo de grado "Análisis del rendimiento del café basado en técnicas de aprendizaje automático" de la Universidad del Cauca?

□ Acepto

□ No acepto

G Cuestionario de evaluación del prototipo

1. ¿Cuál es tu campo de trabajo?

   □ Industria

   □ Academia

2. ¿Cuál es tu rol primario?

   □ Investigador

   □ Administrador

   □ Profesional

   □ Estudiante

   □ Otro: _____

3. ¿En qué campo se enmarca tu profesión?

   □ Ciencias exactas

   □ Ciencias de la computación

   □ Ciencias ambientales

   □ Ciencias agrícolas

   □ Otro: _____

Reacción general frente al prototipo

4. El manejo de la aplicación es:

Totalmente complicado	Moderadamente complicado	Ni fácil ni complicado	Moderadamente fácil	Totalmente fácil

5. La estética general de la aplicación es:

Totalmente desagradable	Moderadamente desagradable	Ni agradable ni desagradable	Moderadamente agradable	Totalmente agradable

Organización de las pantallas

6. Las palabras en las pantallas son:

Totalmente difíciles de leer	Moderadamente difíciles de leer	Ni fáciles ni difíciles de leer	Moderadamente fáciles de leer	Totalmente fáciles de leer

7. La organización de la información en las pantallas es:

Totalmente confusa	Moderadamente confusa	Ni clara ni confusa	Moderadamente clara	Totalmente clara

8. La secuencia de las pantallas es:

Totalmente confusa	Moderadamente confusa	Ni clara ni confusa	Moderadamente clara	Totalmente clara

9. Las funciones de la aplicación se encuentran:

Totalmente desintegradas	Moderadamente desintegradas	Ni integradas ni desintegradas	Moderadamente integradas	Totalmente integradas

Terminología e información

10. El uso de los términos es:

Totalmente inconsistente	Moderadamente inconsistente	Ni consistente ni inconsistente	Moderadamente consistente	Totalmente consistente

11. El lenguaje utilizado en la aplicación es:

Totalmente difícil de entender	Moderadamente difícil de entender	Ni fácil ni difícil de entender	Moderadamente fácil de entender	Totalmente fácil de entender

Conocimiento previo requerido

12. Podría usar la aplicación sin instrucciones:

Totalmente en desacuerdo	Moderadamente en desacuerdo	Ni de acuerdo ni en desacuerdo	Moderadamente de acuerdo	Totalmente de acuerdo

13. Es necesario que los usuarios tengan experiencia previa en el manejo de aplicaciones móviles:

Totalmente necesario	Moderadamente necesario	Ni innecesario ni necesario	Moderadamente innecesario	Totalmente innecesario

Opinión general sobre la aplicación

14. Predecir el rendimiento del cultivo puede ayudar a los caficultores:

Nunca	Pocas veces	Algunas veces	Muchas veces	Siempre

15. La idea plasmada en el prototipo es:

Totalmente convencional	Moderadamente convencional	Ni novedosa ni convencional	Moderadamente novedosa	Totalmente novedosa

16. ¿Qué es lo que más te gustó de la aplicación?

Entendemos que las preguntas que exigen un texto como respuesta pueden ser tediosas, pero no te preocupes, solo serán tres como esta y puede ser un texto corto. ¡Tu opinión será la base para desarrollos futuros!

Respuesta: _____

17. Casi terminamos... ¿Qué es lo que menos te gustó de la aplicación?

    Respuesta: _____

18. ¿Cómo mejorarías la aplicación? Algo que te gustaría que tuviese.

    Respuesta: _____

19. Antes de que te vayas, ¿te importaría dejarnos tu correo electrónico?

    Esperamos no usarlo, pero en el peor de los casos será para contactarte en caso de que algo no nos quede claro.

    Respuesta: _____

H Correo de invitación a la evaluación del prototipo

Hola <primer nombre persona encuestada>.

En el marco de la tesis de maestría "Análisis del rendimiento del café basado en técnicas de aprendizaje automático" se construyó un modelo basado en aprendizaje automático para predecir el rendimiento del cultivo de café con seis meses de anticipación en diferentes municipios de Colombia, esto a partir de las condiciones climáticas. Dicho modelo es insumo principal para un prototipo que busca ser una herramienta de apoyo para las personas involucradas en la cadena productiva del café en Colombia. Por medio del prototipo se puede observar el rendimiento del cultivo y su comportamiento con relación al promedio nacional; en caso de que el rendimiento esperado sea bajo, se tiene acceso a una serie de recomendaciones específicas para la zona encaminadas a mejorarlo. Sumado a esto, es posible observar en él diferentes recomendaciones generales que pueden ayudar a mejorar la productividad del cultivo y una serie de preguntas frecuentes sobre el manejo del mismo.

En ese orden de ideas, se quiere evaluar el prototipo construido desde la perspectiva de expertos y para ello fueron seleccionadas nueve personas que tienen amplios conocimientos en aprendizaje automático, aplicaciones móviles y/o agricultura.

Si es posible, queremos pedirte que por favor ingreses desde tu computador al primer link que se encuentra al final de este correo para conocer el prototipo y luego accedas al segundo link para responder una encuesta acerca del mismo. El prototipo muestra el flujo de pantallas que tendría la aplicación móvil y la encuesta por su parte está constituida por 19 preguntas, por lo que todo este proceso no debería tomar mucho tiempo. Valoramos infinitamente tu participación en este proceso.

Link 1 (prototipo): http://bit.ly/prototipoCafe

Link 2 (encuesta): http://bit.ly/evaluacionPrototipoCafe

En caso de quedar congelado el prototipo, recomendamos volver a cargar la página porque muy seguramente puede deberse a un fallo en la red.

Estamos atentos a cualquier inquietud.

Saludos.

I Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data

sustainability

MDPI

Article

Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data

Juan Rincon-Patino *, Emmanuel Lasso and Juan Carlos Corrales

Grupo de ingeniería Telemática, Universidad del Cauca, Campus Tulcán, Popayán 190002, Colombia; eglasso@unicauca.edu.co (E.L.); jcorral@unicauca.edu.co (J.C.C.)

* Correspondence: juanrincon@unicauca.edu.co; Tel.: +57-28209800 (ext. 2129)

Received: 20 August 2018; Accepted: 26 September 2018; Published: 29 September 2018

check for updates

Abstract: *Persea americana*, commonly known as avocado, is becoming increasingly important in global agriculture. There are dozens of avocado varieties, but more than 85% of the avocados harvested and sold in the world are of the Hass one. Furthermore, information on the market of agricultural products is valuable for decision-making; this has made researchers try to determine the behavior of the avocado market, based on data that might affect it one way or another. In this paper, a machine learning approach for estimating the number of units sold monthly and the total sales of Hass avocados in several cities in the United States, using weather data and historical sales records, is presented. For that purpose, four algorithms were evaluated: Linear Regression, Multilayer Perceptron, Support Vector Machine for Regression and Multivariate Regression Prediction Model. The last two showed the best accuracy, with a correlation coefficient of 0.995 and 0.996, and a Relative Absolute Error of 7.971 and 7.812, respectively. Using the Multivariate Regression Prediction Model, an application that allows avocado producers and sellers to plan sales through the estimation of the profits in dollars and the number of avocados that could be sold in the United States was created.

Keywords: avocado; weather; regression model; machine learning; mobile application

1. Introduction

Persea americana, commonly known as avocado, first appeared in Mexico thousands of years ago, but it was not until 1871 that it was brought to California, United States. By the 1950s, there were dozens of varieties being sold in the markets of the country, with Fuerte being the most consumed variety. About twenty years later, this situation changed, and the Hass avocado started to be the most consumed variety in the country and in the world. At present, avocado consumption happens not only due to its flavor, but also due to its healthy contribution to people's diets [1].

Currently, 85% of the avocados produced and sold in the world are of the Hass variety. This variety grows almost the whole year round and in different regions. Some of the leading producing countries are Mexico, United States, Chile, Australia, South Africa and Israel, with Mexico being the largest producer in the world, representing about a third of the worldwide production [2]. The United States is the leading country concerning imports, and has evolved towards a market of almost a million tons of avocados [3]. The avocado market has grown 16% every year since 2008 in the United States, and this trend is expected to continue, at least in the medium term. States like Florida, California and Hawaii are producers of avocado in this country, but the production does not meet the market demands, so avocados are imported from Mexico, Chile, Peru, New Zealand and the Dominican Republic, among other countries. However, avocado consumption is not uniform across the country. For example, about 90% of the families in California consume avocados, in a proportion of more than three units per month. However, in some states of the Great Plains, only a little more than half of the families consume this fruit, in a proportion of no more than two units per month [3].

Figura anexos 9. Portada artículo "Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data". Fuente propia.

J Exploring machine learning: A bibliometric general approach using Citespace

F1000Research F1000Research 2018, 7:1240 Last updated: 19 DEC 2018

Check for updates

RESEARCH ARTICLE

Exploring machine learning: A bibliometric general approach using Citespace [version 1; referees: 1 approved, 1 approved with reservations]

Juan Rincon-Patino, Gustavo Ramirez-Gonzalez, Juan Carlos Corrales

Telematic Engineering Department, University of Cauca, Popayán, Cauca, 190001, Colombia

v1 **First published:** 10 Aug 2018, **7**:1240 (https://doi.org/10.12688/f1000research.15619.1)

Latest published: 10 Aug 2018, **7**:1240 (https://doi.org/10.12688/f1000research.15619.1)

Abstract

Background: Machine learning researches algorithms that allow a machine to learn about resolving problems in different application domains. Due to the wide number of machine learning applications, it is necessary for newcomers to the field to have alternatives to explore this field faster.

Methods: In this paper, we present a science mapping analysis on the machine learning research in the period 2007-2017. This study was develop using the CiteSpace tool based on results from Clarivate Web of Science. This analysis shows how the field has evolved, by highlighting the most notable authors, institutions, keywords, countries, categories, and journals.

Results: The results provide information on trends and possibilities in the near future, particularly in areas such as health, biology and banking, where machine learning is a valuable tool to generate solutions.

Conclusions: Machine learning is being widely studied, and several institutions in countries like the USA and China constantly generate machine learning based solutions. Diseases, such as cancer or Alzheimer's disease, studies in biology, such as the protein molecule, virtual reality, commerce, smartphones, and ubiquitous computing, are all fields where machine learning contributes to resolving problems.

Keywords

machine learning, science mapping, bibliometrics, topic analysis, citeSpace

This article is included in the Science Policy Research gateway.

Open Peer Review

Referee Status: ✓ ?

	Invited Referees 1	2
version 1 published 10 Aug 2018	✓ report	? report

1 **Sally Ellingson**, University of Kentucky, USA

2 **Chaomei Chen**, Drexel University, USA

Discuss this article

Comments (0)

Figura anexos 10. Portada artículo "Exploring machine learning: A bibliometric general approach using Citespace". Fuente propia.

K Exploring machine learning: A bibliometric general approach using SciMAT

F1000Research F1000Research 2018, 7:1210 Last updated: 15 AUG 2018

Check for updates

RESEARCH ARTICLE

Exploring machine learning: A bibliometric general approach using SciMAT [version 1; referees: 2 approved with reservations]

Juan Rincon-Patino, Gustavo Ramirez-Gonzalez, Juan Carlos Corrales

Telematic Engineering Department, University of Cauca, Popayán, Cauca, 190001, Colombia

v1 **First published:** 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)
Latest published: 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

Abstract

Background: Machine learning is becoming increasingly important for companies and the scientific community. In this study, we perform a bibliometric analysis on machine learning research, in order to provide an overview of the scientific work during the period 2007-2017 in this area and to show trends that could be the basis for future developments in the field.
Methods: This study is carried out using the SciMAT tool based on results extracted from Scopus. This analysis shows the strategic diagrams of evolution and a set of thematic networks. The results provide information on broad tendencies of machine learning.
Results: The results show that SciMAT is a useful tool to carry out a science mapping analysis, and emphasizes the premise that machine learning has boundless applications and will continue to be an interesting research field in the future.
Conclusions: Some of the conclusions exposed in this study show that classification algorithms have been widely studied and represent a relevant tool for generating different machine learning applications. Nonetheless, regression algorithms are becoming increasingly important in the scientific community, allowing the generation of solutions to predict diseases, sales, and yields, for example.

Keywords
machine learning, science mapping, bibliometrics, topic analysis, SciMAT

Open Peer Review

Referee Status: ? ?

	Invited Referees 1	2
version 1 published 07 Aug 2018	? report	? report

1 **Rajesh Kumar Tiwari**, Amity University Uttar Pradesh, India

2 **Mikhail G. Dozmorov**, Virginia Commonwealth University, USA

Discuss this article

Comments (0)

Corresponding author: Gustavo Ramirez-Gonzalez (gramirez@unicauca.edu.co)

Author roles: Rincon-Patino J: Conceptualization, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Ramirez-Gonzalez G**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Corrales JC**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The authors are grateful to the Telematics Engineering Group (GIT) of the University of Cauca for scientific support and Innovacción Cauca project for master's scholarship granted to J. Rincon-Patino.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Rincon-Patino J, Ramirez-Gonzalez G and Corrales JC. **Exploring machine learning: A bibliometric general approach using SciMAT [version 1; referees: 2 approved with reservations]** *F1000Research* 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

First published: 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

Figura anexos 11. Portada artículo "Exploring machine learning: A bibliometric general approach using SciMAT". Fuente propia.