Método de conjunto de clasificadores para la detección de roya basado en datos sintéticos

Edwar Javier Girón Buitrón

Tesis de Maestría en Ingeniería Telemática

Director:
Dr. David Camilo Corrales Muñoz, PhD
Co-Directores:
Dr. José Antonio Iglesias Martínez, PhD
Departamento de informática, Universidad Carlos III de Madrid
Dr. Juan Carlos Corrales Muñoz, PhD

Universidad Del Cauca
Facultad de Ingeniería Electrónica y telecomunicaciones
Departamento de Telemática
e- @mbiente
Popayán, diciembre de 2019

Edwar Javier Girón Buitrón

Método de conjunto de clasificadores para la detección de roya basado en datos sintéticos

Tesis presentada a la
Facultadad de Ingeniería Electrónica y Telecomunicaciones
de la Universidad del Cauca, Colombia
para otorgar el grado académico de

Magíster en:
Ingeniería Telemática

Director:
Dr. David Camilo Corrales Muñoz, PhD
Co-directores:
Dr. José Antonio Iglesias Martínez PhD
Departamento de informática, Universidad Carlos III de Madrid
Dr. Juan Carlos Corrales Muñoz, PhD

Popayán
2019

Agradecimientos

Dedico este trabajo de grado a mis queridos padres: Marina Buitrón y Milton Girón, les agradezco cada palabra de aliento y cada ayuda brindada para el desarrollo de mi carrera profesional y de mi formación personal. Agradezco a mi hermano Fabián Girón por ser un amigo incondicional en todo momento, porque a pesar de los años, hablamos ahora como lo hacíamos de niños, con la misma confianza y sinceridad.

Agradezco a mis demás familiares, porque de todos ellos recibo siempre las mejores energías para seguir adelante con mis proyectos.

A mi novia Darcy Solarte por ser una compañera de vida, un apoyo y un impulso en mi diario vivir.

A mis tutores David Camilo Corrales, Juan Carlos Corrales y José Antonio Iglesias, personas que me brindaron su conocimiento desde sus diferentes áreas de especialidad, ustedes me dieron claridad, serenidad y empuje cuando fue necesario, les agradezco de todo corazón.

A mis amigos de universidad, compartir con ustedes los senderos universitarios hizo que cada obstáculo en el camino fuera más fácil de librar. Espero ser un apoyo para ustedes como cada uno de ustedes lo ha sido conmigo: Carolina, Julian, Juan Sebastian, Charic, Jessner, les deseo lo mejor.

A todos mis demás amigos, gracias por hacer parte de mi vida, siempre digo que en esta vida he tenido la fortuna de encontrar personas que valen mucho.

Un agradecimiento especial al grupo de investigación de Ingeniería Telemática de la Universidad del Cauca y al grupo Innovacción Cauca, gracias a su apoyo la propuesta de Maestría presentada en este documento es una realidad.

Finalmente, agradezco a Dios por todas las oportunidades que me brinda día a día para ser una mejor persona en este mundo. Gracias señor por tus desafíos, gracias por mi familia, amigos y tutores de vida.

Resumen estructurado

Antecedentes: la roya del café es una enfermedad que causa grandes pérdidas económicas alrededor del mundo. La severidad de esta enfermedad cambia en el tiempo, lo que ocasiona que los agricultores pierdan el control sobre la importancia económica que tiene la enfermedad en sus cultivos. Varios esfuerzos de investigación de aprendizaje supervisado, se han enfocado en crear modelos que permitan detectar la incidencia de roya en los cultivos. Con el objetivo de incrementar la precisión de los resultados, los esfuerzos de investigación en los últimos años se han centrado en la creación de modelos de mayor complejidad y en el aumento de muestras de roya en los conjuntos de datos.

Objetivos: establecer un método de conjunto para la detección de roya en el café basado en datos sintéticos.

Métodos: se propone en primer lugar, incrementar el número de muestras de roya en los conjuntos de datos de estudio, con el objetivo de ampliar el número de escenarios posibles del comportamiento de la enfermedad. Por otra parte, se propone un método de conjunto que permita realizar detecciones del porcentaje de incidencia de roya en cultivos de café colombianos, teniendo en cuenta los datos sintéticos construidos y el conocimiento experto sobre la enfermedad.

Resultados: el presente trabajo entrega como resultados tres conjuntos de datos con datos sintéticos de la incidencia de roya en Colombia, los cuales permiten ampliar el número de muestras de roya para la construcción de modelos de aprendizaje supervisado que permitan detectar el comportamiento de la enfermedad en distintas zonas de Colombia, Por otro lado, se entrega un método de conjunto capaz de realizar detecciones del porcentaje de incidencia de roya para las diferentes regiones cafeteras de Colombia.

Conclusiones: la creación de un mecanismo para la construcción de datos sintéticos que comprenda en sus procesos el comportamiento de algoritmos de interpolación combinado con conocimiento experto permite construir muestras mejor contextualizados, lo que ocasiona una mejor representación del comportamiento de la roya en los conjuntos de datos colombianos.

Palabras clave: roya, métodos de conjunto, datos sintéticos, Hemileia Vastatrix, detección de enfermedades, agricultura

ABSTRACT

Background: the coffee rust is a devastating disease that causes large economic losses across the world. The severity of this disease changes over time, so the farmers are not fully aware of the economic importance of the rust disease in the coffee crops. Several research works have created machine learning models to detect coffee rust. Several research works have created machine learning models to detect coffee rust. Nowadays, to increase the precision in the results, the research works have built more complex models, and they try increasing the number of coffee rust samples.

Objectives: establish an ensemble learning method for detection of coffee rust based on synthetic data.

Methods: first, it is proposed to increase the number of coffee rust samples in the datasets studied, to extend the number of scenes about coffee rust behavior. On the other hand, an ensemble learning method is proposed for the detection of the incidence coffee rust percentage in Colombian coffee crops, taking account of the synthetic data built and expert knowledge.

Results: this research work has the following results: three datasets with synthetic data about coffee rust incidence in Colombia that increase the number of coffee rust samples for construction of supervised learning models to detect coffee rust in the different Colombian regions. On the other hand, an ensemble method is delivered, which can detect the coffee rust incidence percentage to several Colombian regions.

Conclusions: the creation of a mechanism to create synthetic data that includes the behavior of interpolation algorithms and expert knowledge in their processes allows generating better samples and a better representation of the behavior of the coffee rust disease in Colombian datasets.

Keywords: coffee rust, ensemble methods, synthetic data, Hemileia Vastatrix, detection of diseases, agriculture.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

En este capítulo se describe el planteamiento del problema del proyecto, el escenario de motivación, el objetivo general y los objetivos específicos que guían el presente trabajo de maestría, y, finalmente se muestran las contribuciones realizadas a partir del desarrollo de este proyecto.

Introducción

1.1 Planteamiento del problema

La industria del café es uno de los objetos económicos a nivel mundial que desde 1990 ha incrementado su nivel de producción de una manera acelerada, de tal forma que al año 2015 su incremento sobrepasa el 50% [1], convirtiéndose en un producto de exportación importante en varios países del mundo, y en la principal fuente de recursos para varios agricultores de los países productores. Países como Brasil, Vietnam y Colombia ven la producción del café como una de sus principales fuentes económicas y sociales para el desarrollo de su país puesto que los tres países producen más del 50% del café del mundo.

La roya es el principal problema en los cultivos de café; el desarrollo fisiológico del cultivo, el nivel de producción de la planta y la distribución y cantidad de lluvia son los principales factores que contribuyen bajo ciertas condiciones a la aparición y rápida reproducción de esta enfermedad. Ya que entre los años 2008 y 2010 se han presentado abruptas alteraciones en las condiciones del clima, los agricultores no pueden realizar un control oportuno sobre sus cultivos, dando paso a que la roya aumente sus probabilidades de aparición, comprometiendo seriamente la cantidad y la calidad de la cosecha y así afectando de manera importante la producción de los países cafeteros como Colombia, El Salvador y gran parte de América Central [2].

Desde la aparición de roya en cultivos de café en Brasil en 1970, Colombia ha realizado investigaciones para crear variedades más resistentes a la enfermedad. Aunque las nuevas variedades de café han mostrado resistencia al hongo, todavía tres cuartas

partes de los cultivos son susceptibles a padecer de roya dependiendo de las condiciones climáticas y de la agronomía del cultivo [3].

Dado que la roya es una enfermedad que afecta a nivel mundial el sector cafetero y trae consigo notables pérdidas económicas, varios esfuerzos de investigación, especialmente desde el área de aprendizaje supervisado, se han enfocado en crear modelos que permitan detectar la incidencia de roya en los cultivos de café teniendo en cuenta variables meteorológicas (temperatura, pluviosidad, humedad, etc.) y agronómicas (distancia de siembra, carga de frutos, etc.), las cuales sirven para reducir la pérdida en las cosechas y evitar el uso masivo de fungicidas, aumentando así la calidad de los cultivos. Estos modelos realizan un proceso de aprendizaje a partir de un conjunto de datos con el objetivo de predecir, clasificar o detectar una nueva entrada. En [4], [5] los autores construyen conjuntos de datos para Colombia y Brasil respectivamente. A partir de estos conjuntos de datos diversas investigaciones obtienen modelos de clasificación de roya utilizando algoritmos de aprendizaje supervisado (árboles de decisión, máquinas de vectores de soporte, redes neuronales, redes bayesianas). Nuevas iniciativas como las propuestas en [5], [6] definen estructuras para combinar modelos de clasificación con el objetivo de aumentar la precisión en la detección, ya que teóricamente se ha demostrado que la combinación de un conjunto de modelos de clasificación genera soluciones más precisas que utilizar uno sólo. Este tipo de enfoque se conoce comúnmente como métodos de conjunto de clasificadores [7]–[9].

Aunque los avances desde la informática han contribuido en la detección de roya en cultivos de café, los conjuntos de datos utilizados en las investigaciones mencionadas anteriormente contienen baja cantidad de registros para realizar una correcta detección, esto debido a que la captura de datos es una actividad costosa para los agricultores y compleja para las personas que recolectan información. Este problema ocasiona que los modelos de clasificación poco precisos debido a la poca información recolectada [10].

En ese sentido, esfuerzos de investigación como los realizados en [10] empiezan a abordar la problemática de escasez de datos a partir del uso de algoritmos de interpolación que permiten incrementar las medidas del Porcentaje de Incidencia de Roya (PIR) en el café teniendo en cuenta los registros existentes en los actuales

conjuntos de datos; sin embargo, esta investigación se ve limitada a la construcción de nuevos datos considerando únicamente el comportamiento que el algoritmo pueda interpretar de los datos existentes, sin tener en cuenta otros factores que contribuyen al desarrollo de la enfermedad.

Teniendo en cuenta las anteriores consideraciones, es necesario aumentar el número de instancias del conjunto de datos actual de roya considerando factores que contribuyen al progreso de la enfermedad, con el objetivo de ofrecer modelos precisos. De la definición del problema se plantea la siguiente pregunta de investigación:

¿Cómo detectar el porcentaje de incidencia de la roya en cultivos de café teniendo en cuenta la escases de datos?

1.2 Escenario de motivación

Los cambios notables del calendario geo-climático están afectando los niveles de competitividad de los países exportadores de productos en los sistemas agrícolas, puesto que las características del suelo, los cambios extremos de temperatura, entre otros factores abióticos y bióticos producen la aparición de circunstancias propicias para el desarrollo de plagas y enfermedades, afectando la calidad de los cultivos y la productividad del mismo.

Según la Federación Nacional de Cafeteros de Colombia, más de 563.000[1] familias sin contar las familias independientes a esta organización dependen de la producción del café, los cambios geo-climáticos se convierten en un factor de riesgo alto para el buen desarrollo de los cultivos, así como para el adecuado desarrollo económico y social del país, considerando que al 2018 el café continua siendo uno de los principales productos de exportación de Colombia[2] aportando el 0,7% del PIB [11].

[1] http://www.cafedecolombia.com/particulares/es/la_tierra_del_cafe/la_gente_del_cafe/
[2] https://www.larepublica.co/analisis/sergio-clavijo-500041/panorama-cafetero-2018-2019-2797742

1.3 Objetivos

1.3.1 Objetivo general

- Establecer un método de conjunto para la detección de roya en el café basado en datos sintéticos

1.3.2 Objetivos específicos

- Construir un mecanismo para la generación de muestras sintéticas de incidencia de roya
- Adaptar un método de conjunto para la detección de roya con base en los conjuntos de datos construidos
- Evaluar las capacidades del método de conjunto para la detección de roya en el café

1.4 Contribuciones

Las principales contribuciones de este trabajo son:

- Tres conjuntos de datos correspondientes a las regiones de “Jazmín”, “Naranjal” y “Santagueda” con muestras sintéticas de la incidencia de roya para diferentes años de cultivo
- Un método de conjunto para la detección de roya basado en los conjuntos de datos con muestras sintéticas de la incidencia de roya
- Un mecanismo para la generación de muestras sintéticas de la incidencia de roya en cultivos de café colombianos
- Un prototipo web que implementa la API correspondiente al método de conjunto construido en este proyecto
- El siguiente artículo de investigación:
    - *“Rule-based expert system for detection of coffee rust warnings in Colombian crops”. Este artículo fue expuesto en la conferencia: “*6th International Symposium on Language & Knowledge Engineering” en la

ciudad de Puebla, México en los días 29, 30 y 31 de octubre del año 2018. La publicación del artículo se encuentra en *"Journal of Intelligent & Fuzzy Systems, vol. 36, no. 5, pp. 4765-4775, 2019"*. https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179025

1.5 Contenido de la monografía

La monografía se organiza en siete capítulos estructurados a partir de las fases de la metodología CRISP-DM (Cross Industry Process for Data Mining) [12], la cual ofrece una descripción del ciclo de vida en proyectos de minería de datos. La figura 1 muestra las seis fases de la metodología:

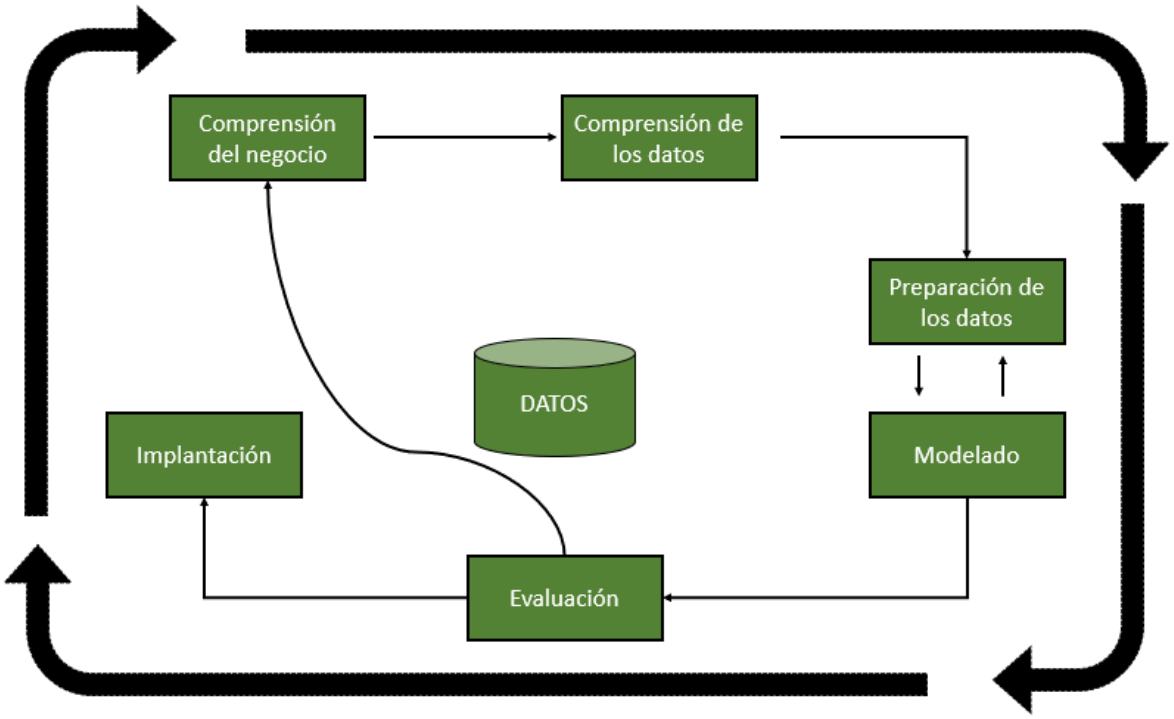


Figura 1. Fases de la metodología CRISP-DM [12]

Considerando las fases de la Figura 1, la monografía se encuentra estructurada de la siguiente manera:

Capítulo 2. Estado actual del conocimiento / comprensión del negocio. En este capítulo se describen los conceptos fundamentales del proyecto y los trabajos relacionados con el problema de investigación planteado.

Capítulo 3. Comprensión y preparación de los datos. En este capítulo se describen los conjuntos de datos utilizados en el proyecto y las actividades realizadas para tratar problemas de valores faltantes y selección de características encontrados en los conjuntos de datos.

Capítulo 4. Modelado. Este capítulo describe paso a paso la construcción del método de conjunto. Primero se realiza un análisis de series de tiempo en las variables de clima, con el objetivo de que el método de conjunto tenga la capacidad de realizar pronósticos del PIR cinco días en el futuro. Luego, fueron elegidos los algoritmos del método de conjunto, así como su estructura de funcionamiento.

Capítulo 5. Experimentación y evaluación. Este capítulo presenta el proceso de evaluación y las pruebas realizadas al modelo construido, con el objetivo de analizar las capacidades del método de conjunto de esta tesis de maestría.

Capítulo 6. Prototipo. En este capítulo se presenta el proceso de desarrollo para la construcción de un prototipo web que hace uso de la API que contiene el método de conjunto construido anteriormente para la detección de roya en cultivos de café colombianos.

Capítulo 7. Conclusiones y trabajos futuros. Finalmente, este capítulo analiza los resultados obtenidos en el presente proyecto de maestría, detallando las principales contribuciones obtenidas durante el desarrollo del trabajo, y son expuestas recomendaciones para trabajos futuros.

Capítulo 2: Estado actual del conocimiento / comprensión del negocio

En este capítulo, se presentan los precedentes teóricos que permiten comprender el contexto del presente trabajo de investigación, el cual propone un método de conjunto para la detección de la incidencia de roya en cultivos de café colombianos basado en datos sintéticos. Luego, se presentan los trabajos de investigación relacionados al planteamiento del problema de la tesis de maestría. Finalmente, se realiza un resumen que describe los principales aportes del capítulo.

2.1 Conceptos generales

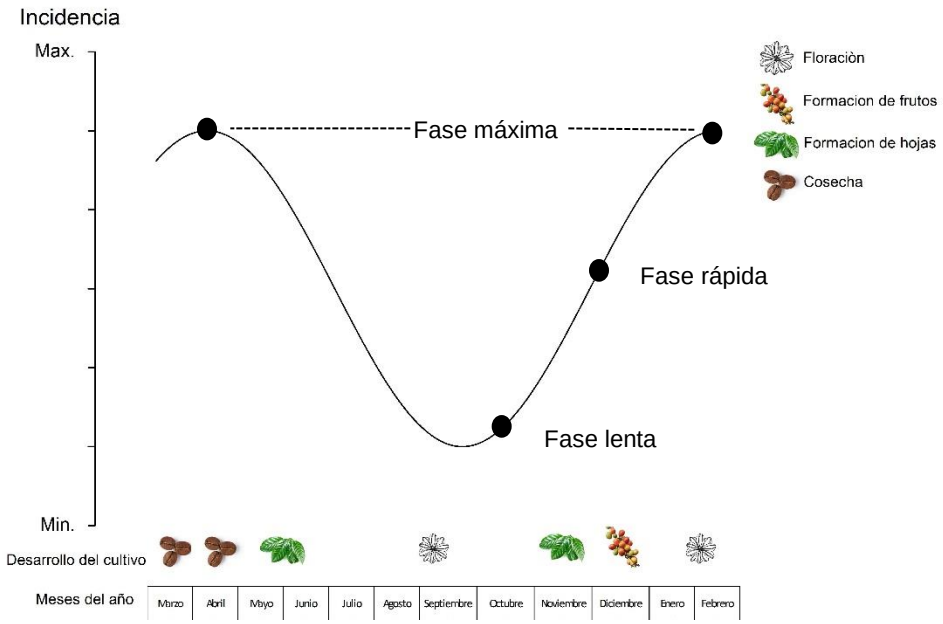
2.1.1 Roya en el café

La roya es una enfermedad que se presenta en algunos cultivos debido a la aparición de un hongo en las hojas de la planta; en la roya del cafeto es generada por el hongo Hemileia vastatrix que ocasiona pigmentación en las hojas de café y posteriormente la caída de ellas. El brote de esta enfermedad se concibe por propiedades del suelo, características del cultivo, manejo agronómico por parte del agricultor y las condiciones ambientales, siendo este último el factor que más variedad de condiciones presenta en el transcurso de los últimos años.

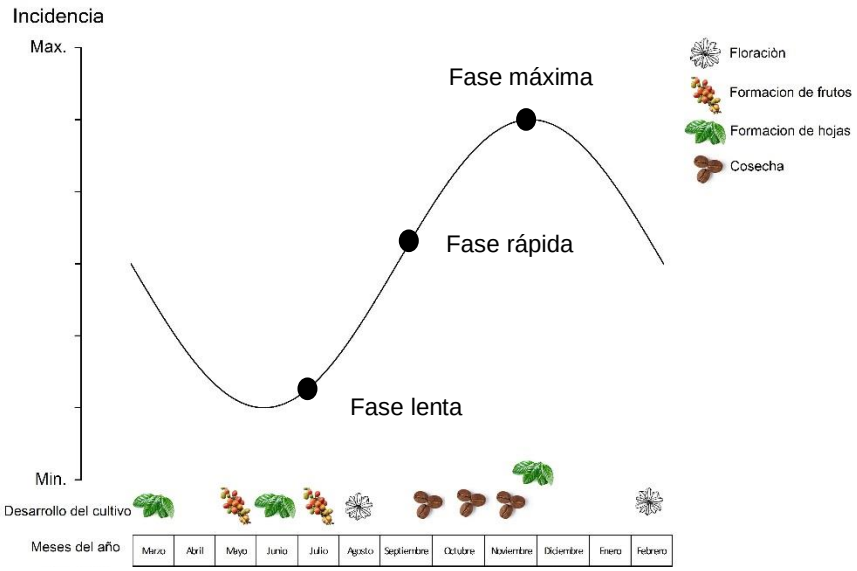
El ciclo de vida del hongo comienza con la diseminación de esporas en las primeras horas, que en grandes cantidades se visualiza como un polvo amarillo o naranja en las hojas de la planta. Una vez las esporas están en las hojas, inicia el proceso de germinación, la cual requiere ayuda de las condiciones ambientales para completar la infección de la hoja. Cuando la hoja se encuentra infectada, el hongo absorbe los nutrientes de la planta hasta el punto de poder reproducirse e iniciar el ciclo de diseminación de esporas de nuevo [13]. De esta manera, sin ningún tipo de control la enfermedad se convierte en una epidemia que avanza en el espacio y el tiempo. De acuerdo a las investigaciones, el progreso de la enfermedad sin un adecuado control se puede describir en tres etapas: fase lenta, fase rápida y fase terminal [13]. En la

fase lenta la epidemia inicia su proceso de infección en algunas hojas, repitiéndose el ciclo de infección muy lentamente. En la segunda fase, debido a la expansión de las esporas en la fase anterior, existe una gran cantidad de hojas infectadas al mismo tiempo, lo que aumenta la incidencia de la enfermedad. Finalmente, las hojas más afectadas mueren, reduciendo la población de hojas en las plantas, lo que ocasiona una reducción en el espacio de infección y por lo tanto la finalización de la epidemia.

En el contexto colombiano, teniendo en cuenta las fases de la enfermedad, autores [5], [6] definieron las curvas de progreso de la roya considerando la fenología del cultivo, la evaluación de la enfermedad en diferentes zonas cafeteras de Colombia, y la distribución de la cosecha. Los autores proponen dos curvas de crecimiento para explicar el aumento de la incidencia de la enfermedad en distintas zonas de Colombia. La primera curva aplica para las regiones de Quindío, Norte de Santander y Cauca, cuya cosecha principal es en el primer semestre del año (Figura 2a), y la segunda curva aplica para las regiones de Antioquia, Caldas y Risaralda cuya cosecha principal es en el segundo semestre del año (Figura 2b)



(a)



(b)

Figura 2. Curvas de progreso de la roya sin control químico [16]

La severidad de esta enfermedad varia año a año, ocasionando que los agricultores descuiden la importancia económica que representa la aparición de la enfermedad en sus cultivos; cuando el clima es favorable y la cosecha abundante, la roya causa grandes pérdidas en la producción [16].

2.1.2 Series de tiempo

Una serie de tiempo se define como una colección de datos que se recopilan, observan o registran en intervalos de tiempo regulares. Es así como una serie de tiempo puede representarse mediante una lista de tuplas $S = [(t_1, m_1), (t_2, m_2), \dots, (t_n, m_n)]$, donde *"m"* es una medición en el espacio y *"t"* el instante de tiempo al que corresponde la medición *"m".* Esta relación cronológica en los datos de la serie hace que el análisis de este tipo de fenómenos sea distinto a otros tipos de problemas [17].

Las series de tiempo se clasifican de manera general en series estacionarias y no estacionarias. Las series estacionarias son estables en el tiempo, es decir, la media y la varianza de los datos son constantes en el tiempo. Por su parte, las series no estacionarias tienen una tendencia que cambia en el tiempo. Los cambios en la media de los datos determinan un crecimiento o decrecimiento en su componente de

tendencia, por lo que la serie no oscila alrededor de un valor constante. El análisis de tendencia (cambio a largo plazo de la media), estacionalidad (variabilidad de los datos debido a estaciones en los mismos) y fenómenos aleatorios permite encontrar patrones en los datos, para de esta manera realizar tareas de pronóstico a través de algoritmos de series de tiempo y algoritmos de aprendizaje supervisado.

2.1.3 Método de conjunto homogéneo

Los métodos de conjunto o también llamados multiclasificadores son conjuntos de diferentes clasificadores que toman las predicciones individuales y las fusiona, dando como resultado la combinación de cada una de ellas [18]. Estos métodos se diferencian dependiendo las características que se deseen cubrir, algunas de ellas son: el número de clasificadores individuales acoplados, la estructura de acoplamiento utilizada para los clasificadores, el tipo de cada clasificador (redes neuronales, árboles de decisión, vecino más cercano, etc.), las características de los subconjuntos usados por cada clasificador del conjunto, la agregación de las decisiones particulares, el tamaño y la naturaleza de los conjuntos de datos de entrenamiento para los clasificadores y la combinación de clasificadores homogéneos o heterogéneos [18].

Los métodos de conjunto homogéneos, o también llamados métodos de ensamble se caracterizan por usar un solo algoritmo de aprendizaje supervisado dentro de sus clasificadores base. Por lo cual, la única diferencia existente entre los clasificadores internos del método de ensamble es la muestra de información que ingresa a cada clasificador en su etapa de entrenamiento.

2.1.4 Datos sintéticos

Los datos sintéticos se generan con el fin de satisfacer necesidades específicas que no se pueden encontrar en capturas originales de datos. El diccionario McGraw-Hill define a los datos sintéticos como “cualquier producción de datos aplicables a una situación que no se obtienen por medición directa” [19]. Estos datos sintéticos son generados para conocer ciertas condiciones que pueden o no ser encontrados en los datos originales, es decir, los datos sintéticos se convierten en valores teóricos que pueden medir indirectamente una situación particular que haya pasado o que aún no

ocurre en un contexto, con el fin de abarcar el mayor número de eventos posibles y así preparar a un sistema para que reaccione adecuadamente.

2.2 Trabajos relacionados

Esta sección hace una descripción de recientes investigaciones que relacionan el problema planteado en este proyecto. Inicialmente se exponen investigaciones orientadas al uso de aprendizaje supervisado en la detección de roya en el café, posteriormente son descritos trabajos relacionados con la construcción de datos sintéticos en distintos contextos de investigación. Finalmente, se describen trabajos acerca de métodos de conjuntos en diferentes contextos de aplicación.

2.2.1 Aprendizaje supervisado para la detección de roya en el café

Desde la informática, específicamente en aprendizaje supervisado, los esfuerzos de investigación orientados al cuidado de los cultivos de café se han incrementado exponencialmente debido a los cambios extremos presentes en la temperatura, humedad, pluviosidad y otros factores climáticos que contribuyen al desarrollo de plagas y enfermedades sobre cultivos de café, ocasionando grandes pérdidas económicas y ambientales a los agricultores [20]. A continuación, son presentados los trabajos de investigación relevantes que hacen uso de algoritmos de aprendizaje supervisado para la detección de plagas y enfermedades en cultivos de café.

En [21] realizan una revisión bibliográfica para la detección y predicción de plagas y enfermedades en diversos cultivos como: maíz, arroz, café, mango, maní y tomate, mediante el uso de algoritmos de aprendizaje supervisado (árboles de decisión, redes neuronales artificiales, redes bayesianas y máquinas de vectores de soporte). Los resultados obtenidos en la revisión sistemática, indican que los cultivos de café son los más abordados debido a la importancia para el desarrollo económico en países como Colombia, donde la producción de café es la principal actividad agrícola del país abarcando más de 360.000 familias dedicadas al cultivo del café. Estudios sobre la roya concluyen que las esporas que contienen la infección pueden propagarse a través de la lluvia y el viento, por lo que varios países como Colombia y Brasil, pueden ver afectados sus cultivos.

En este orden de ideas, investigadores colombianos detectan la tasa de infección de roya en cultivos de café a través de un conjunto de datos de entrenamiento que reúne 147 instancias, recolectadas entre los años 2011 y 2013. El conjunto de datos se compone de 21 atributos divididos en tres categorías: condiciones climáticas, propiedades físicas del cultivo y administración del cultivo [22]. La evaluación del dataset se realizó dividiendo en tres subconjuntos el conjunto de datos de entrenamiento (D1, D2, D3); D1 contiene los 21 atributos del dataset, D2 excluye atributos de la fertilidad del suelo y tres de las propiedades físicas de los cultivos, por su parte D3 reemplaza los atributos excluidos en D2 por una nueva variable. Los tres nuevos subconjuntos fueron aplicados en tres clasificadores de aprendizaje supervisado: SVR (Support Vector Regression), BPNN (Backpropagation Neuronal Network) y el árbol de regresión M5, de los cuales SVR obtiene los mejores resultados de desempeño en los tres subconjuntos de datos. Esta investigación concluye que: 1) La cantidad de instancias del dataset es aún pequeña y necesita aumentarse. 2) Un método de conjunto puede mejorar el rendimiento y la precisión de la detección de la roya, lo cual es propuesto como trabajo futuro.

En [6] se identifica que el uso de clasificadores simples no ofrece los resultados más precisos; el trabajo propone un método de conjunto en cascada para la detección de roya en el café, éste método consta de un clasificador de primer nivel (BPNN) y dos clasificadores de segundo nivel (M5 y SVR) que detectan la tasa de infección de roya mayor a 7,18% y menor a 7,18% respectivamente. En [23] proponen un sistema de alertas tempranas para la roya en el café basado en códigos de salida de corrección de error y SVM. De igual forma [24] propone reglas de detección de roya en el café basados en árboles de decisión y emparejamiento de grafos. Finalmente, autores en [25] construyeron un sistema experto basado en reglas que identifica los parámetros principales de clima en Colombia que contribuyen al desarrollo de la enfermedad. Los resultados muestran que la detección del progreso de la enfermedad debe realizarse considerando el estudio de series de tiempo, puesto que el nivel de incidencia de la enfermedad depende estrechamente de las condiciones ambientales de los 28 días anteriores a la medición.

Similarmente, investigadores brasileños detectan la roya en el café mediante un conjunto de datos que se caracteriza por ofrecer salidas binarias; si las tasas de infección son superiores a cinco puntos porcentuales, la salida de la clase es uno, de lo contrario la salida es cero. Una segunda opción plantea que, si la tasa de infección de roya es igual o superior a 10 puntos porcentuales entonces la salida de la clase es uno, de lo contrario la salida es cero. Este conjunto de datos fue recolectado durante 13 años (1998-2011) a través de una estación automática del clima ubicada en la granja experimental de la fundación PROCAFE en la ciudad de Varginha/MG (latitud sur 21º 34'00'', longitud 45° 24'22'' y altitud 940m) generando 612 instancias. El conjunto de datos se compone de 23 variables independientes que ofrecen información relacionada con la temperatura, humedad en las hojas y luminosidad en el cultivo [5]. Estas investigaciones hacen uso de técnicas de modelado como: árboles de decisión, redes neuronales artificiales, el método Random Forest y SVM. En [26], [27] utilizan este conjunto de datos sólo para los primeros ocho años de recolección, [26] propone un modelo a partir de dicho conjunto de datos haciendo uso de árboles de decisión difusos, [27] propone un caso de estudio de predicción de la roya en el café a partir del uso de redes bayesianas. En [28] se discute como aprender funciones que sean capaces de predecir si el valor de una variable objetivo supera un umbral específico, con el fin de que un predictor de roya en el café pueda determinar a partir de intervalos si la variable objetivo se encuentra en un estado de advertencia o de alerta. En este trabajo se presentan tres implementaciones basados en técnicas de aprendizaje supervisado: uno basado en regresión y dos en clasificación. En [29] se basan en SVM para formular predicciones respecto a variables continuas analizadas para identificar la fiabilidad del uso de sistemas de alertas tempranas en la roya del café.

En la Figura 3 son presentados los algoritmos de aprendizaje supervisado y sistemas expertos utilizados en el contexto de detección de roya en los cultivos de café para las investigaciones de Colombia y Brasil:

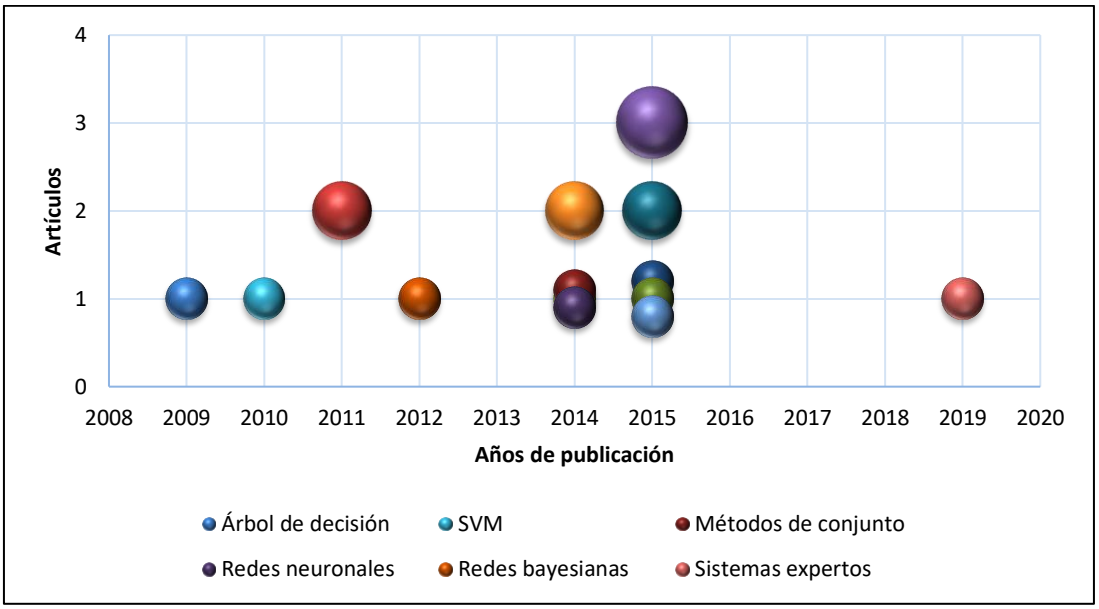


Figura 3. Algoritmos usados en investigaciones para la detección de roya en el café

De esta forma se puede observar en la Figura 3 que los algoritmos de caja negra como SVM y Redes Neuronales han sido utilizados desde 2010 y 2014 respectivamente hasta la actualidad dentro de 7 investigaciones de detección de roya. Lo anterior se da ya que son algoritmos de alta precisión y han demostrado en las investigaciones reducir la probabilidad de error en alertas para la detección de roya en el café [21]. Los algoritmos que basan su representación en grafos como los Árboles de Decisión y Redes Bayesianas de igual forma han sido utilizados para la detección de roya en el café, ya que su representación gráfica permite conocer las características del clasificador construido, por lo que estos algoritmos son muy utilizados en las investigaciones desde el 2009 hasta la actualidad. Finalmente los métodos de conjunto utilizados para la detección de roya en el café como Random Forest y basados en cascada, son nuevas iniciativas que obtienen mejor precisión que un clasificador simple, según afirmaciones de autores destacados en el medio [7]–[9]. Por este motivo se observa que desde el 2014 existen investigaciones que usan este tipo de métodos para la detección de roya en el café.

Los resultados actuales muestran que, tanto en el uso de clasificadores simples como en los métodos de conjunto, los datos de entrenamiento utilizados en las investigaciones cuentan con pocas instancias, lo que compromete la calidad de los

resultados generados por los clasificadores. Es por esta razón que se propone en el presente proyecto un mecanismo para la generación de muestras sintéticas de roya para cultivos de café colombianos, que permita ampliar el conocimiento de los algoritmos de aprendizaje supervisado, y, por lo tanto, que mejoren su precisión en la tarea de detección.

2.2.2 Datos sintéticos en distintos contextos de aplicación

Partiendo de la necesidad de incrementar el número de muestras de datos que permiten a los investigadores generalizar de mejor manera un problema específico cuando la información recolectada no es suficiente, varias investigaciones enfocadas en la construcción de datos virtuales que representen un contexto particular se han llevado a cabo. A continuación, son descritos algunos trabajos que hacen uso de los datos sintéticos para mejorar las capacidades de interpretación de problemas en distintos contextos de aplicación.

En [30]–[32] se han construido conjuntos de datos virtuales para disminuir el riesgo de la revelación de información privada y el control de correos no deseados (spam); estas investigaciones utilizan los datos sintéticos para crear un escenario a partir de los datos construidos, que permita hacer análisis estadístico para la descripción de una población sin revelar o utilizar los datos originales de la población analizada. Para lograr este objetivo, diversos métodos y algoritmos de construcción de datos han sido usados, por ejemplo, métodos estadísticos(IPSO, por sus siglas en inglés) [30], métodos de perturbación de datos (PEGS) [31] y algoritmos de balanceo de clases (SMOTE) [32]. Estas investigaciones indican que los datos sintéticos tienen un alto grado de confiabilidad en la simulación de escenarios.

Por otro lado, en el contexto de sistemas de manufactura donde la escasez de datos ocasiona que no se puedan realizar análisis robustos en etapas tempranas de funcionamiento de los sistemas, se utilizan los datos sintéticos para incrementar el tamaño de los conjuntos de datos de entrenamiento con el objetivo de mejorar la precisión de los análisis de dichos sistemas durante estas etapas. Este contexto hace uso principalmente de métodos de generación estadística de datos: DBSCAN [33], teoría de valores extremos [34].

En el área de la salud, se utilizan los datos sintéticos para mejorar la precisión de estudios médicos, haciendo uso de métodos estadísticos para la mejora de la precisión en los registros de series de tiempo de miocardio [35], y para el reconocimiento de expresiones faciales a partir de algoritmos de distribución lineal [36]. Por otra parte los algoritmos de interpolación se han usado en este contexto para construir datos a partir de la interpolación de imágenes que mejoren la precisión del estudio de imágenes cardiacas [37], y la resolución de imágenes satelitales a partir de la creación de nuevos puntos de datos [38]. En 2019, la interpolación de datos se ha utilizado en la biomédica para mejorar los objetos de imagen creados en simulaciones anatómicas para el estudio de enfermedades en el ser humano. El trabajo expuesto en [39] utiliza la técnica de interpolación "interpolation spoke" para mejorar las propiedades geométricas de un objeto y así construir imágenes médicas más precisas.

En el contexto de la agricultura, los algoritmos de interpolación han cumplido funciones en distintas áreas de la agricultura. En [40] se utiliza la interpolación para corregir valores atípicos en estudios de contaminación debido al uso de plaguicidas e insecticidas en cultivos de China. Por otra parte, en la detección de enfermedades en los cultivos de café en Colombia, en [10] proponen el uso del algoritmo de interpolación Cubic Spline para crear nuevas muestras del Porcentaje de Incidencia de Roya (PIR) y la construcción de tres conjuntos de datos basados en los datos construidos: meteorología diaria, variación meteorológica y meteorología previa. Los datos sintéticos en este trabajo fueron usados para entrenar varios algoritmos de aprendizaje supervisado (árbol de regresión M5, vectores de regresión SVR, perceptrón multicapa). Los resultados indican que la aproximación realizada con el conjunto de datos de meteorología previa y el algoritmo de perceptrón multicapa es la combinación con mejores resultados. Este trabajo resalta la importancia de trabajar con series de tiempo para el estudio de la enfermedad.

A continuación, en la Figura 4 son resumidos los enfoques de construcción de datos utilizados en los diferentes contextos de aplicación:

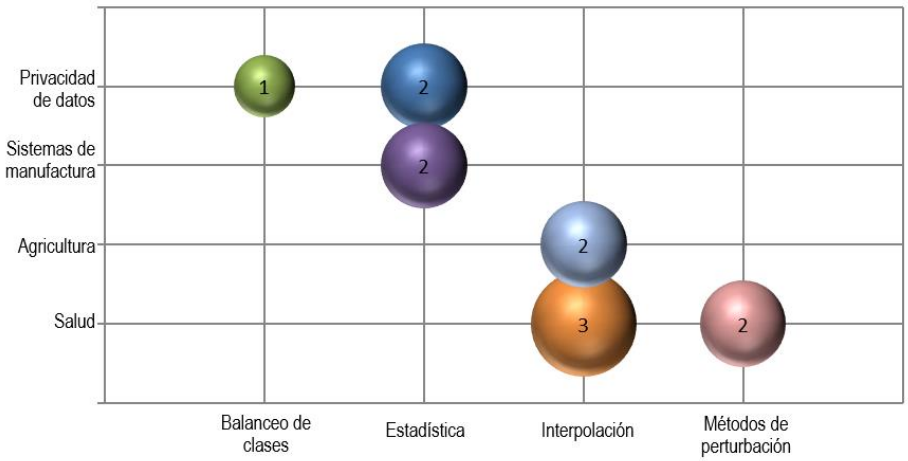


Figura 4. Métodos y algoritmos de construcción de datos sintéticos en distintos contextos de aplicación

Las investigaciones que hacen uso de algoritmos de interpolación son consideradas buenas alternativas para la construcción de datos, ya que las nuevas muestras son creadas a partir del comportamiento de los datos. Sin embargo, es necesario complementar la construcción de datos con factores adicionales del comportamiento de la enfermedad.

La investigación descrita en [10] tiene en cuenta únicamente el comportamiento del algoritmo de interpolación; para detectar apropiadamente la roya en el café es necesario crear datos sintéticos considerando los principales factores que contribuyen al desarrollo de la enfermedad. Adicional a esto, los resultados de este trabajo presentan valores atípicos al estimar valores negativos del porcentaje de incidencia de la roya.

2.2.3 Métodos de conjunto en distintos contextos de aplicación

Los métodos de conjunto se han utilizado en los últimos años dados sus buenos resultados en la detección y predicción de eventos en distintos contextos de aplicación. En [41], por ejemplo, se realiza una investigación que realiza varios experimentos, con el objetivo de evaluar si el uso de los métodos de conjunto son alternativas apropiadas para la predicción financiera con series de tiempo, los algoritmos Random subspace, stacking y bagging son puestos a prueba en los experimentos. Los resultados sugieren

que los algoritmos mejoran el rendimiento en las predicciones realizadas comparadas con clasificadores simples. De igual forma los trabajos en [42] y [43] realizan experimentos para adecuar algoritmos de clasificación simple y aumentar su precisión. En [42], la investigación presenta una serie de experimentos con redes neuronales para su uso en métodos de conjunto, puesto que este tipo de algoritmos se han empezado a utilizar en estos contexto. Por su parte en [43] se resalta que recientemente los SVM son usados para la predicción con series de tiempo, y el trabajo hace uso de los métodos de conjunto para mejorar el rendimiento de las predicciones.

En la industria eléctrica se han hecho investigaciones para comparar varios métodos de conjunto para el pronóstico de carga de energía con series de tiempo. Al comparar Random Forest y GBRT (Gradient Boosting Regression Trees), los resultados evidencian que el método GBRT es superior para el contexto de estudio. Por otra parte, estudios como los realizados en [44] y [45] proponen mejorar la precisión de los métodos de conjunto para su uso en el contexto de datos dados por series de tiempo. En [44] se utiliza el método boosting con redes neuronales recurrentes para la predicción con series de tiempo. El algoritmo propuesto en la investigación adiciona un nuevo parámetro que influencia las decisiones del método boosting. Los resultados demuestran que el uso de métodos de conjunto adaptados para las series de tiempo mejora considerablemente las predicciones. En [45] la investigación describe el uso de los métodos de ensamble para construir modelos apropiados para la predicción con series de tiempo.

Dentro del contexto cafetero, a partir del año 2014 se han venido utilizando los métodos de conjunto dado que según distintas afirmaciones [7]–[9], estos esquemas de clasificación muestran tener una mejor precisión que los clasificadores simples. En [6] investigadores colombianos proponen un método de conjunto en cascada de dos niveles para la detección de roya en el café, éste método consta de un clasificador de primer nivel (BPNN) y dos clasificadores de segundo nivel (M5 y SVR) que detectan el porcentaje de incidencia de roya mayor a 7,18% y menor a 7,18% respectivamente. Investigadores brasileños utilizan el método de conjunto Random Forest como fue detallado en secciones anteriores, para la detección de roya en los cultivos de café [5].

Dado que los métodos de conjunto son opciones viables para mejorar la precisión en la detección de enfermedades como la roya en el café, el presente trabajo de

investigación, tomará como punto de partida los trabajos antes mencionados para el desarrollo del método de conjunto para la detección de roya basado en datos sintéticos. Sin embargo, se debe tener en cuenta que, en la investigación de Colombia, la cantidad de registros por encima de 7,18% son pocos debido a la poca cantidad de datos utilizados en el método de conjunto. Además, esta investigación no considera los periodos de tiempo del desarrollo de la enfermedad. Este factor contribuye a la selección de los atributos de análisis que influyen en los clasificadores utilizados dentro del método de conjunto para la detección de la incidencia de la roya en el café.

2.3 Resumen

Este capítulo presenta los conceptos teóricos principales de la investigación, tales como: métodos de conjunto, datos sintéticos, series de tiempo, y roya en cultivos de café. Luego, fueron expuestos los trabajos de investigación que se relacionan con las temáticas del presente proyecto de Maestría. En el uso de aprendizaje supervisado en cultivos de café se encontraron 18 trabajos de investigación, donde a pesar del uso de algoritmos tradicionales y métodos de conjunto para la detección de roya en los cultivos cafeteros, se encuentra que los datos de entrenamiento cuentan con pocas instancias, comprometiendo los resultados de los modelos construidos. Para los trabajos que se relacionan con datos sintéticos, se describieron 12 trabajos de investigación que utilizan técnicas de generación de datos sintéticos para varios contextos de aplicación. La técnica de interpolación de datos muestra características que la convierten en una buena alternativa para el presente trabajo de grado. Finalmente, los 9 trabajos relacionados con métodos de conjunto muestran que el uso de clasificadores de este tipo mejora la precisión de clasificación. En Colombia se han utilizado métodos de conjunto para la detección de roya en el café, sin embargo, los resultados muestran que hacen falta incluir dentro del modelo características sobre el tiempo de desarrollo de la enfermedad en los cultivos.

Capítulo 3: comprensión y preparación de los datos

En este capítulo es descrito el proceso de comprensión de datos, en el cual se resumen las características de los conjuntos de datos utilizados. En segundo lugar, este capítulo presenta la preparación de los datos para adaptarlos al método de conjunto para la detección de roya en cultivos de café que será presentado en el capítulo 4.

3.1 Comprensión de los datos

Se utilizaron tres conjuntos de datos provenientes del Centro Nacional de Investigaciones del Café – CeniCafé[3] de diferentes regiones cafeteras de Colombia, tal como se muestra en la Figura 5. En estas regiones se recolectaron muestras de Incidencia de Roya (IR) en cultivos entre 1986 y 1988 en la región “El Jazmín”, 1985 en la región de “Santagueda”, y entre 1985-1987 en la región de “El Naranjal”. Estos conjuntos de datos contienen registros de roya en los cultivos durante epidemias presentadas en la región, lo que permite estudiar con mayor profundidad el comportamiento de la enfermedad en los cultivos de café colombianos.

Figura 5. Localización de regiones cafeteras en Google Maps

[3] https://www.cenicafe.org/

Además, se tomaron registros de estaciones meteorológicas de la temperatura media, mínima, y máxima (AvgTem, MaxTem, MinTem), punto de rocío, humedad relativa, presión de vapor, horas de sol y precipitación durante el día y la noche. Adicional a esto, a partir de las variables obtenidas, otras variables meteorológicas fueron calculadas, tales como, amplitud térmica (ΔT) y la precipitación total. La Tabla 1 presenta la descripción de las variables meteorológicas y de incidencia de roya en el café.

Atributo	Abreviación	Descripción
Temperatura mínima	MinTem	Media de los valores mínimos recolectados durante un día. Unidad de medida: °C.
Temperatura máxima	MaxTem	Media de los valores máximos recolectados durante un día. Unidad de medida: °C.
Temperatura promedio	AvgTem	Media de los valores medios recolectados durante un día. Unidad de medida: °C.
Amplitud térmica	ΔT	Diferencia entre los valores máximos y mínimos de temperatura. Unidad de medida: °C. Expresión matemática: $\Delta T^i = MaxTem^i - MinTem^i$ (3.1)
Humedad relativa	Hum	Cantidad de vapor de agua presente en el aire. Unidad de medida: % Expresión matemática: $Hum^i = \frac{Presión\ parcial^i}{Equilibrio\ del\ vapor\ de\ agua^i} x\ 100$ (3.2)
Presión de vapor	VPress	Presión a la cual el vapor de agua se encuentra en equilibrio termodinámico con su estado de condensación. Unidad de medida: milibares (mb)

Punto de rocío	PR	Temperatura en la cual el agua del aire comienza a enfriarse y a saturarse con el vapor de agua. Unidad de medida: °C.
Precipitación durante el día	Pd	Cantidad de precipitación en un lugar durante el día. Unidad de medida: milimetros (m.m.)
Precipitación durante la noche	Pn	Cantidad de precipitación en un lugar durante la noche. Unidad de medida: milimetros (m.m.)
Precipitación total	Pt	Cantidad de precipitación durante un día. Unidad de medida: milimetros (m.m.) Expresión matemática: $Pt = Pd + Pn$ (3.3)
Horas de sol	Sol	Número de horas en la cual la luz solar incide sobre un lugar. Unidad de medida: horas.
Muestras de Incidencia de Roya	IR	Número de hojas infectadas por el hongo *Hemileia vastatrix* dividido por el total de hojas disponibles en la planta. Unidad de medida: % Expresión matemática: $IR^i = \frac{Total\ de\ hojas\ infectadas\ en\ 60\ árboles^i}{Total\ de\ hojas\ en\ los\ 60\ árboles^i} x100$ (3.4)

Tabla 1. Descripción de los atributos de los conjuntos de datos de roya en el café

La Tabla 2 presenta el número de muestras recolectadas de IR y las variables de clima disponibles para cada región cafetera.

Atributos \ Región	**Jazmín**	**Santagueda**	**Naranjal**
AvgTem	x	x	X
MaxTem	x	x	X
MinTem	x	x	X
ΔT	x	x	X
Hum	x	x	X

VPress	x	x	
PR	x	x	X
Pd	x	x	X
Pn	x	x	X
Pt	x	x	X
Sol	x	x	X
Muestras de clima	**1096**	**365**	**1003**
IR	**43**	**26**	**48**

Tabla 2. Número de muestras recolectadas de roya y atributos de clima en las regiones “El Naranjal”, “Santagueda” y “El Jazmín”

Las tablas 3, 4 y 5 muestran las medidas estadísticas de las variables de los conjuntos de datos de las regiones “Jazmín”, “Santagueda” y “Naranjal” respectivamente, las cuales permitan brindar una idea sobre las propiedades de cada una de las variables de interés de esta propuesta. Las medidas estadísticas utilizadas para describir las variables son: porcentaje de valores perdidos, valor promedio, mediana, desviación estándar, porcentaje de ceros encontrados, valores mínimos y máximos.

Variable	# muestras	Valores perdidos (%)	Valor promedio	Desviación estándar	Ceros (%)	Valor mínimo	mediana	Valor máximo
MinTem	1096	0	15.46	1.01	0	11.8	15.5	18.5
MaxTem	1096	0	24.8	2.23	0	16.5	25	31
AvgTem	1096	0	19.72	1.5	0	15.18	19.65	24.75
ΔT	1096	0	9.34	2.28	0	1.3	9.4	15
Hum	1096	0	81.32	6.55	0	60.4	81.7	97.5
VPress	1096	0	18.54	1.14	0	14.8	18.6	21.91
PR	1096	0	16.26	0.98	0	12.9	16.3	19.6
Pd	1096	0	2.96	7.38	51.09	0	0	66.5
Pn	1096	0	4.66	11.35	55.47	0	0	108
Pt	1096	0	7.62	13.94	34.22	0	1.4	124.6
Sol	1096	0	3.78	2.55	4.56	0	3.5	10.3
IR	43	96.08	22	14	0	6	16	65

Tabla 3. Medidas de estadística descriptiva para las variables de la región “Jazmín”

Variable	# muestras	Valores perdidos (%)	Valor promedio	Desviación estándar	Ceros (%)	Valor mínimo	mediana	Valor máximo
MinTem	365	0	17	1.14	0	12.5	17	20
MaxTem	365	0	29.04	1.86	0	22	29.5	33.4
AvgTem	365	0	148.65	1699.25	0	19.35	22.8	23.7
ΔT	365	0	12.04	2.24	0	5.5	12.2	17.5
Hum	365	0	74.37	5.82	0	58.4	73.9	93.4
VPress	365	0	19.79	1.43	0	14	20	23.2
PR	365	0	17.2	1.25	0	11.4	17.4	19.9
Pd	365	0	1.06	4.08	0	0	0	48.2
Pn	365	0	4.12	8.55	0	0	0	60.3
Pt	365	0	5.19	9.8	0	0	0.4	60.3
Sol	365	0	5.96	2.81	0	0	6.5	10.8
IR	26	92.8	30	23	0.06	6	26	82

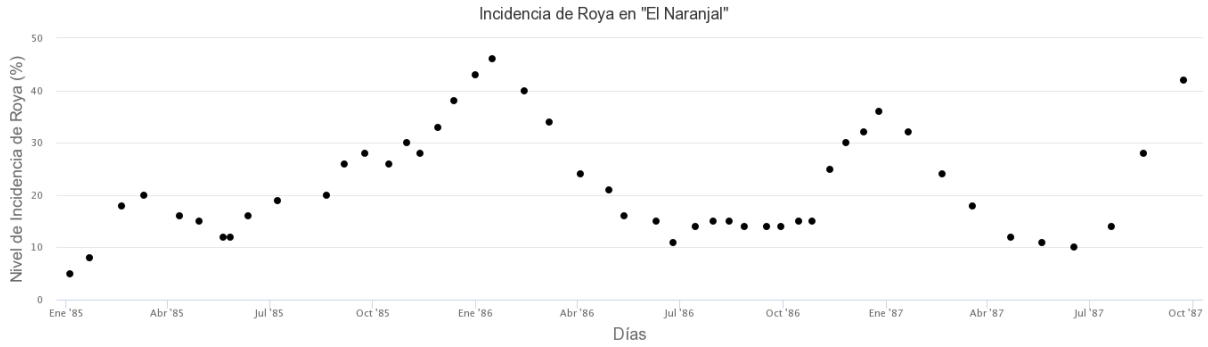
Tabla 4. Medidas de estadística descriptiva para las variables de la región “Santagueda”

Variable	# muestras	Valores perdidos (%)	Valor promedio	Desviación estándar	Ceros (%)	Valor mínimo	mediana	Valor máximo
MinTem	1003	0	16.25	1.03	0	11.4	16.4	18.8
MaxTem	1003	0	26.93	2.2	0	19.6	27.2	33.5
AvgTem	1003	0	21.01	1.34	0	17.2	21.1	25.4
ΔT	1003	0	10.68	2.38	0	1.5	10.9	16.5
Hum	1003	0	80.34	6.32	0	62.9	80.1	97.5
VPress	1003	0	4984.3	8704.05	0	10.1	20	23300
PR	1003	0	17.23	0.91	0	12.4	17.3	20
Pd	1003	0	1,72	5.5	63.21	0	0	64.7
Pn	1003	0	5.39	11.9	51.84	0	0	88.2
Pt	1003	0	7.11	13.16	36.79	0	0.8	88.2
Sol	1003	0	5.13	2.82	6.68	0	5	11
IR	48	95.21	22	10	0	5	19	46

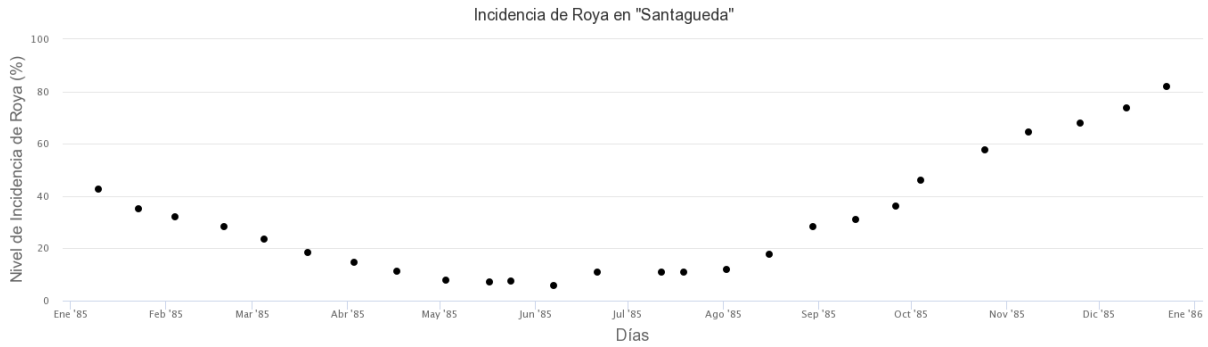
Tabla 5. Medidas de estadística descriptiva para las variables de la región “Naranjal”

Comparando el número de muestras obtenidas de las estaciones meteorológicas y las muestras de IR durante los mismos años de observación, la región de “Jazmín” cuenta con 43 muestras de la IR contra 1096 muestras de las distintas variables de clima mencionadas anteriormente, lo cual es equivalente a que la cantidad de valores no medidos para esta región durante los años de observación definidos es del 96%; para el caso de la región “Santagueda” el porcentaje es de 92,8% y en “El Naranjal” del

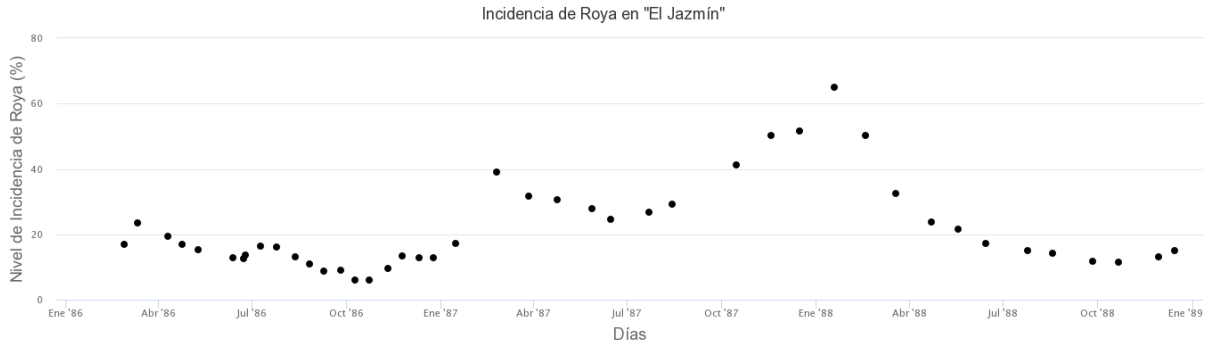
95,2%. La Figura 6 muestra la IR de las tres regiones cafeteras a través de la temporada cafetera anual durante varios años. Para el presente trabajo de investigación, a los valores de IR no medidos entre los días de un año también se le llamarán valores perdidos.



(a) Incidencia de roya en “El Naranjal”



(b) Incidencia de roya en “Santagueda”



(c) Incidencia de roya en “El Jazmín”

Figura 6. Incidencia de la roya en los cultivos de café colombianos: (a) Naranjal, (b) Santagueda y (c) Jazmín

En este sentido, el mínimo valor de incidencia de roya alcanzado en la región Naranjal (Figura 5a) fue de 5% y su valor máximo de 46%. Por su parte, en “Santagueda” (Figura 5b) sus valores mínimos y máximos llegaron al 6% y 82% respectivamente, siendo estos los valores más críticos de las regiones analizadas. Finalmente, para la región “Jazmín” (Figura 5c) el valor mínimo de incidencia alcanzado fue de 6% en octubre de 1986 y el máximo de 66% en enero de 1988.

3.2 Preparación de los datos

En esta sección se presentan las técnicas usadas para el procesamiento de los datos de clima y roya en el café. Con el objetivo de realizar una preparación de datos adecuada, se tomaron como base la metodología CRISP-DM [12] y actividades propuestas en [46]. De esta manera las tareas descritas en esta sección son: tratamiento de valores perdidos y selección de atributos.

3.2.1 Tratamiento de valores perdidos

Dentro de los distintos conjuntos de datos, se observa que la variable de incidencia de roya contiene pocas muestras en los tres conjuntos de datos (Tabla 2). Esto se debe a que la recolección de estos se realiza manualmente, analizando la cantidad de hojas infectadas en 60 árboles cada vez, por lo que es una actividad que requiere mucho esfuerzo; es por este motivo que no se hace constantemente.

Con el objetivo de aumentar el número de muestras de la IR, se construyó un mecanismo para crear datos sintéticos a partir de las muestras de roya en cultivos de café colombianos. El mecanismo se compone de dos módulos: el módulo de interpolación y el módulo de conocimiento experto. En el módulo de interpolación se encuentran los submódulos para la manipulación y análisis de datos: Pandas y Numpy, además del submódulo SciPy, que contiene el algoritmo de interpolación Cubic Spline. Dentro del módulo de conocimiento experto se encuentran dos submódulos: el submódulo de selección de años y el submódulo de suavizado. La Figura 7 presenta la arquitectura del mecanismo de creación de datos sintéticos en cultivos de café colombianos.

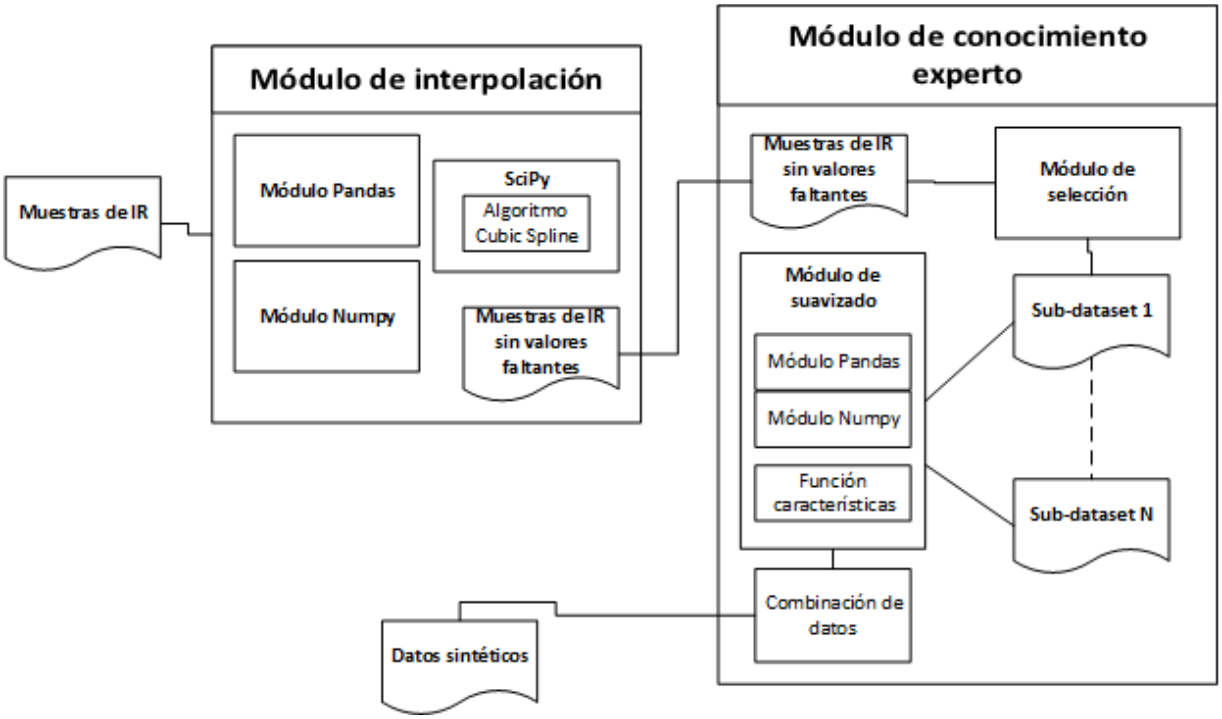


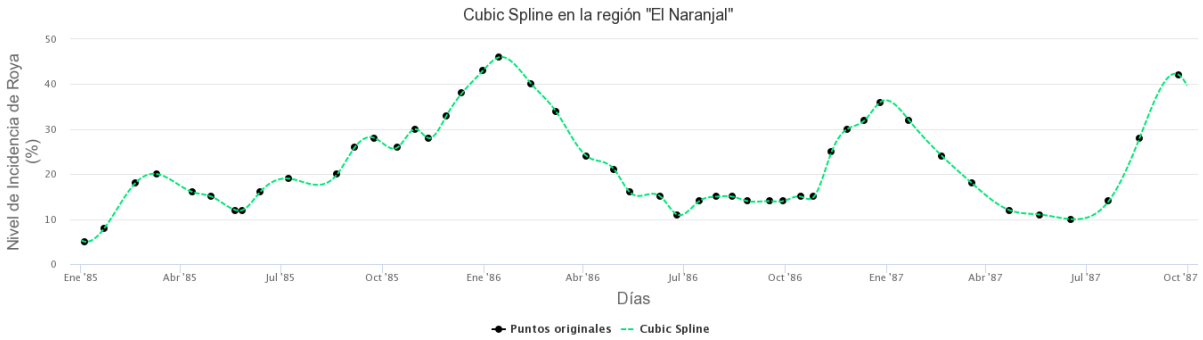
Figura 7. Mecanismo para la generación de datos sintéticos del porcentaje de incidencia de roya en cultivos de café colombianos

3.2.1.1 Módulo de interpolación

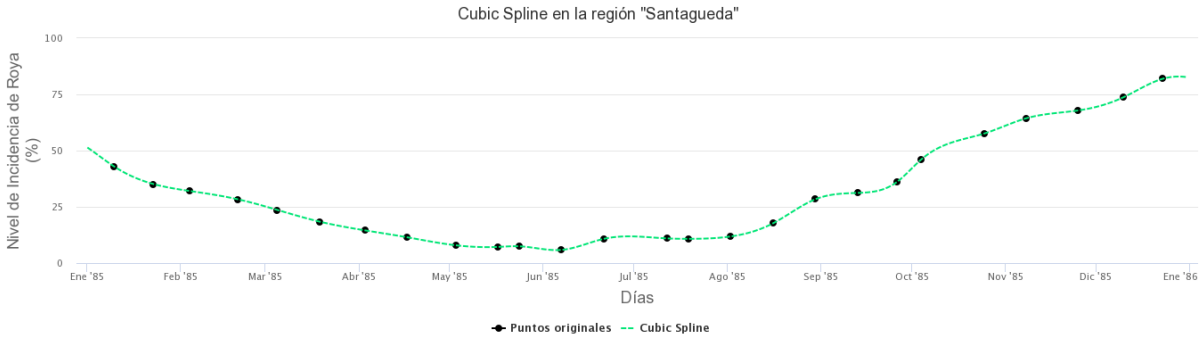
La interpolación permite crear nuevas muestras de la IR entre dos puntos cercanos conocidos. En este trabajo se utilizó la adaptación de la interpolación Cubic Spline propuesta en [10] para aumentar las muestras de IR. Este método consiste en la creación de curvas diferenciables segmentadas en porciones a través de polinomios

de tercer orden, las cuales pasan por un conjunto de puntos de control (muestras originales). Este método produce un sistema de tres diagonales que puede ser resuelto para obtener los coeficientes de los polinomios [47].

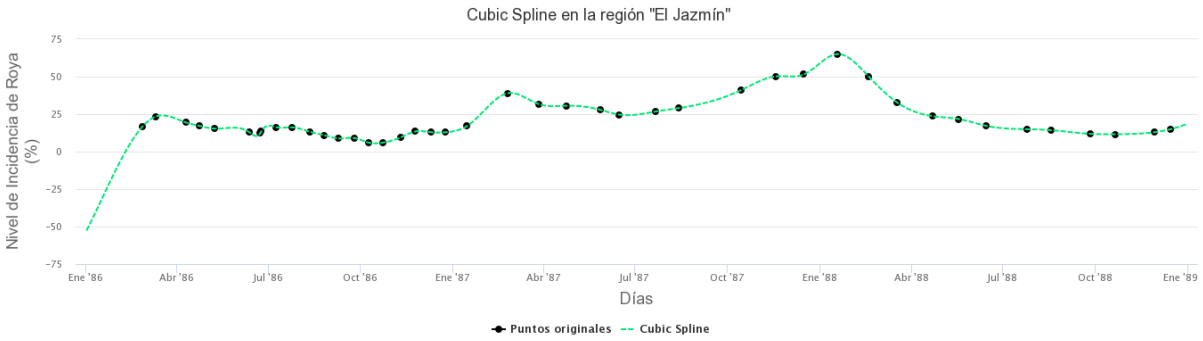
La Figura 8 muestra la Curva de Incidencia de Roya (CIR) creada de la interpolación cubic spline y las muestras originales de las regiones colombianas. La implementación del algoritmo se hizo a través de la librería SciPy de Python.



(a) Cubic Spline en la región “El Naranjal”



(b) Cubic Spline en la región “Santagueda”



(c) Cubic Spline en la región “El Jazmín”

Figura 8. Curva de incidencia de roya creada por el algoritmo de interpolación Cubic Spline en las regiones de café colombianos: (a) Naranjal, (b) Santagueda y (c) Jazmín

La Figura 8 muestra la forma en la que los puntos generados por el algoritmo Cubic Spline se adaptan a los puntos originales. Sin embargo, es posible observar que para la región “Jazmín” se crearon puntos de incidencia de roya negativos, lo cual no corresponde al comportamiento normal de esta curva. Es por esto que se propone el uso de un módulo adicional que tenga la capacidad de utilizar el conocimiento experto sobre el comportamiento de la roya para realizar ajustes a la curva de datos, y de esta manera obtener datos sintéticos mejor contextualizados.

3.2.1.2 Módulo de conocimiento experto

Una vez el módulo de interpolación crea la primera aproximación de la curva de incidencia, el módulo de conocimiento experto realiza un ajuste de la curva considerando el conocimiento experto encontrado en la literatura, tales como: artículos científicos, reportes técnicos y literatura relacionada con la incidencia de roya. Este procedimiento se realiza ya que el conjunto de datos original contiene pocas muestras de la IR, debido a este hecho, la primera aproximación del algoritmo Cubic Spline no es capaz por si solo de representar apropiadamente el comportamiento de la enfermedad en su curva de incidencia resultante.

La curva de incidencia resultante del primer módulo es ajustada basada en dos criterios: i) identificación de los parámetros para generar muestras sintéticas, ii) establecer el mecanismo de suavizado teniendo en cuenta el paso anterior.

- **Identificación de los parámetros para generar muestras sintéticas**

En el contexto colombiano, autores como [5], [6] definieron las curvas de progreso de la roya considerando la fenología del cultivo, la evaluación de la enfermedad en diferentes zonas cafeteras de Colombia, y la distribución de la cosecha (Figura 2). Considerando el comportamiento de las curvas de progreso de roya en Colombia propuestas en la Figura 2, se definieron los siguientes criterios para el ajuste de la curva de incidencia:

- El comportamiento de la enfermedad debe analizarse en periodos de un año. Esto debido a que las condiciones ambientales y del cultivo cambian a través del tiempo, alterando la intensidad con la cual se puede presentar la enfermedad. En la actualidad este fenómeno es evidente debido a la situación del cambio climático en los últimos años
- Las curvas de progreso pueden ser definidas a través de funciones polinomiales, ya que, al conocer la gráfica general del comportamiento de la roya para cada zona de Colombia y teniendo algunas muestras de roya en las distintas regiones analizadas, el mecanismo de creación de datos sintéticos puede identificar los parámetros necesarios para construir una función polinomial de la curva de progreso de la enfermedad que se adapte a las distintas zonas cafeteras de Colombia [48]. En este caso, se requieren los valores mínimos locales con el objetivo de establecer las raíces del polinomio, y el valor máximo global que sirve como punto de evaluación para hallar las constantes del polinomio
- Los puntos de la curva de progreso siempre son mayores o iguales a cero. Esto implica que el rango de los nuevos puntos sintéticos se encuentra entre 0% y 100%

Teniendo en cuenta los criterios establecidos, se construyeron dos submódulos dentro del módulo de conocimiento experto: i) el módulo de selección de años, que se encarga de dividir el conjunto de datos en periodos de un año ii) el módulo de suavizado, el cual recibe los subconjuntos entregados por el módulo de selección de años y se encarga de construir las funciones polinomiales características del progreso de la IR de acuerdo a las características de cada subconjunto de datos entregado por el módulo de selección.

- **Módulo de suavizado**

En este submódulo se construyó la expresión matemática que representa las curvas de progreso de la IR en cultivos de café colombianos. Para construir la función son necesarios:

1. Los puntos de datos entregados por el módulo de interpolación (puntos originales más los puntos sintéticos), los cuales brindan información sobre la incidencia de la roya en una región particular
2. Los criterios encontrados en la identificación general del comportamiento de la roya en Colombia, los cuales permiten limitar el espacio de creación de los nuevos puntos de roya, y filtrar los puntos de datos entregados por el módulo de interpolación que no se encuentren dentro de los criterios establecidos

La expresión matemática que representa la curva de incidencia de roya $\boldsymbol{f_{(x)}}$ es el producto del coeficiente principal ***"a"*** y las raíces de los polinomios representadas por los factores $(\boldsymbol{x - b_i})^{\boldsymbol{2}}$, donde ***"x"*** es el día del año en el cual el punto de IR "$\boldsymbol{b_i}$" es un mínimo relativo cercano a cero.

$$f(x) = a(x - b_1)^2 * \ (x - b_2)^2 * \ldots * (x - b_n)^2 \qquad (3.5)$$

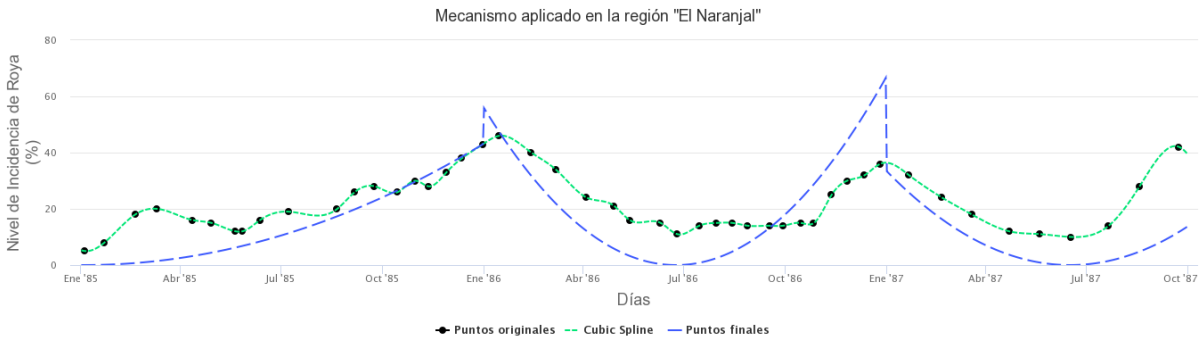
El pseudocódigo del módulo de suavizado se describe a continuación:

Algoritmo 1: mecanismo para crear datos sintéticos en cultivos de café colombianos

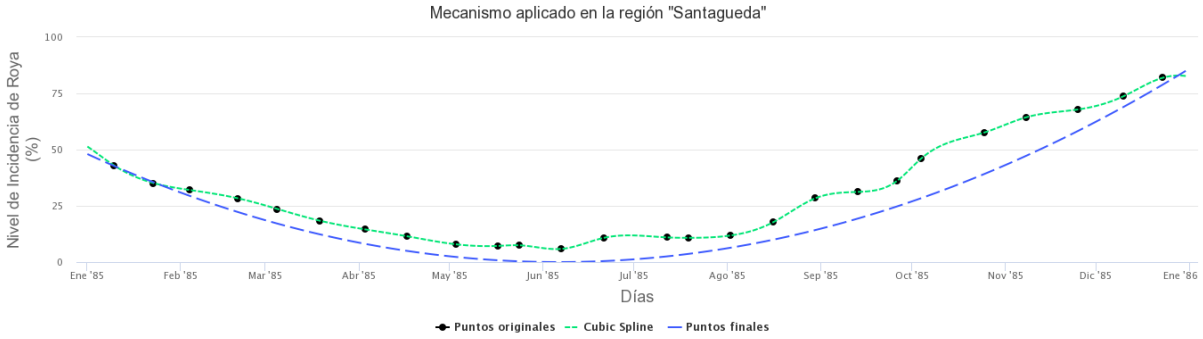
```
1     function (coffe_rust);
      INPUT: coffe_rust [] and factor_vector []
      OUTPUT: rust incidence curve f(x)
2     for day from 1 to 365
3     |   if coffee_rust [day] is a relative minimum
4     |   |   b = coffe_rust [day]
5     |   |   factor = (day – b)^2
6     |   |   factor_ vector [count] = factor
7     |   |   count = count + 1
8     |   end
9     |   a = max(coffee_rust) / factor_ vector [day where max (coffee_rust)]
10    |   f(x) = a*factor_ vector (day)
11    end
```

El mecanismo para la creación de datos sintéticos considera un vector con puntos de la IR y un vector que almacena las raíces de los polinomios encontradas en los datos. Para cada uno de los 365 días del año se analiza si la medición de roya en cada día corresponde a un valor mínimo relativo de las muestras. Cuando una medición de roya

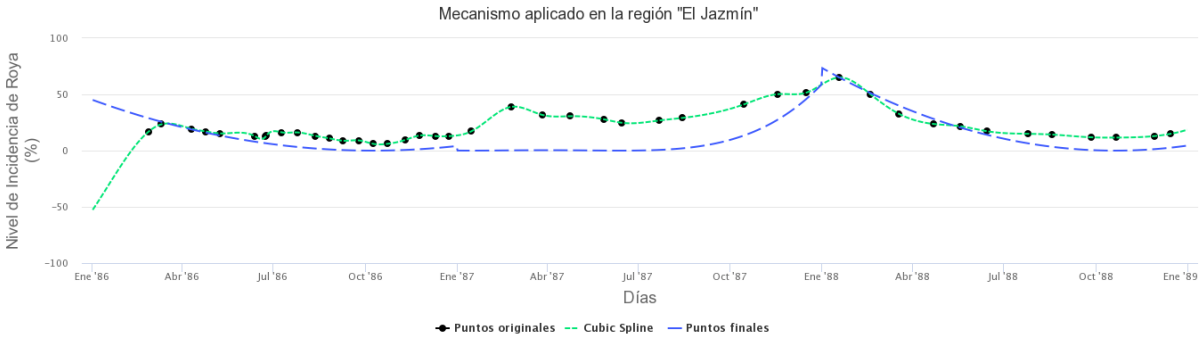
es un mínimo relativo se hallan las raíces del polinomio en ese punto y se almacenan en el vector “factor_vector”. Posteriormente se halla la constante del polinomio “a” utilizando como punto de referencia el valor máximo de roya en las muestras y el día en el que fue tomada esa medición. Finalmente, para cada día se calcula a través de la función característica el nuevo dato sintético. La Figura 9, muestra los resultados del módulo de suavizado para la generación de las nuevas muestras sintéticas que forman la curva final de progreso de la IR para las distintas regiones cafeteras analizadas. Esta curva grafica la tendencia de la incidencia de roya para cada región analizada.



(a) Mecanismo aplicado en la región “El Naranjal”



(b) Mecanismo aplicado en la región “Santagueda”



(c) Mecanismo aplicado en la región “El Jazmín”

Figura 9. Curvas de incidencia de roya creadas por la interpolación Cubic Spline y el conocimiento experto en regiones cafeteras colombianas: (a) Naranjal, (b) Santagueda y (c) Jazmín

Los resultados finales en la Figura 9 muestran que, el mecanismo propuesto utiliza el conocimiento experto y los datos sintéticos construidos en el módulo de interpolación para generalizar la curva de progreso de la incidencia de roya para cada una de las regiones cafeteras.

3.2.2 Selección de atributos

Con el objetivo de seleccionar los atributos meteorológicos apropiados para la estimación de la IR, se realizó un proceso de selección de características mediante el uso del algoritmo de eliminación recursiva de características (RFE) [49]. El algoritmo RFE consta de un proceso recursivo que otorga puntuaciones a las características dados por un estimador externo que asigna pesos a los atributos (Ej: Coeficientes de un modelo de regresión lineal). Una vez se tienen las puntuaciones de todas las características, se ordenan por orden de importancia y el usuario toma la decisión de eliminar las características con menor puntuación [8][9].

De esta manera, se utilizaron cuatro estimadores externos: Random Forest (RF), Regresión Lineal (Linear Model, LM), un Modelo de Adición Generalizada (Generalized Additive Model, GAM), y Árboles Bagging (Bagging trees, BT). Considerando los estimadores externos, se propusieron cinco experimentos utilizando los estimadores para las tres regiones cafeteras analizadas (Jazmín, Naranjal y Santagueda):

- **Experimento E-001:** utiliza todas las muestras del conjunto de datos con los atributos de clima recomendados por reportes técnicos y especialistas en cultivos [25]: temperatura mínima, temperatura máxima, amplitud de temperatura, humedad, precipitación durante el día y la noche, y horas de sol
- **Experimento E-002:** utiliza todas las muestras del conjunto de datos y todos los atributos de clima disponibles
- **Experimento E-003:** utiliza las muestras correspondientes a los periodos de floración dentro del conjunto de datos, todos los atributos de clima disponibles y tres subconjuntos de datos:
    - **Dataset DS-001:** la etapa de floración durante el primer semestre del año
    - **Dataset DS-002:** la etapa de floración durante el segundo semestre del año
    - **Dataset DS-003:** las etapas de floración durante ambos semestres del año
- **Experimento E-004:** utiliza las muestras correspondientes a los periodos de formación de hojas dentro del conjunto de datos, todos los atributos de clima disponibles, y tres subconjuntos de datos:
    - **Dataset DS-001:** la etapa de formación de hojas durante el primer semestre del año
    - **Dataset DS-002:** la etapa de formación de hojas durante el segundo semestre del año
    - **Dataset DS-003:** las etapas de formación de hojas durante ambos semestres del año
- **Experimento E-005:** utiliza las muestras correspondientes a los periodos de cosecha dentro del conjunto de datos, todos los atributos de clima disponibles, y tres subconjuntos de datos:
    - **Dataset DS-001:** la etapa de cosecha durante el primer semestre del año
    - **Dataset DS-002:** la etapa de cosecha durante el segundo semestre del año
    - **Dataset DS-003:** las etapas de cosecha durante ambos semestres del año

Los experimentos E-001 y E-002 consideran todas las muestras del conjunto de datos. Por otra parte, los experimentos E-003, 004 y 005 dividen el conjunto de datos total en tres subconjuntos de datos correspondientes a tres etapas de un cultivo de café:

floración, formación de hojas y cosecha. Considerando las tres regiones cafeteras, los experimentos y los estimadores externos utilizados, se construyeron 132 modelos de regresión.

La implementación del algoritmo RFE para la construcción de los distintos modelos de regresión proviene del paquete CARET del lenguaje estadístico R [51]; y la medida de evaluación usada para determinar los mejores modelos de regresión fue el error medio cuadrático (Root Mean Square Error, RMSE). El RMSE se define como la desviación estándar de los errores en la predicción, es decir que esta medida muestra la desviación de las diferencias entre los valores predichos y los valores observados [52]. La expresión matemática del RMSE está dada por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{pred\,i}\right)^2} \qquad (3.6)$$

Donde, *n* es el número de residuos y $\left(y_i - y_{pred\,i}\right)$ es el error en la predicción.

Los resultados de los experimentos para determinar los mejores modelos de regresión y los mejores atributos de clima para la detección de roya en cultivos de café colombianos se presentan en el Anexo A. En esta sección se realiza un estudio comparativo de los mejores modelos de regresión por experimento presentado.

3.2.2.1 Estudio comparativo

Considerando los resultados de los diferentes experimentos (E-001, 002, 003, 004 y 005), se realizó un estudio comparativo con el fin de seleccionar los mejores atributos para el método de conjunto para la detección de roya, y los mejores modelos de regresión que harán parte de la estructura del método de conjunto. Para la selección de los mejores experimentos, se tuvieron en cuenta los mejores RMSE teniendo en cuenta los siguientes criterios de satisfacción: 1) El RMSE promedio de las tres regiones cafeteras en cada experimento y 2) Los atributos seleccionados por el algoritmo RFE para cada experimento en las tres regiones cafeteras sea mayor o igual a tres atributos. El objetivo de elegir los experimentos basados en los criterios de satisfacción mencionados anteriormente es hallar los atributos y los modelos de

regresión que mejor se comporten en las tres regiones cafeteras de manera conjunta, dado que dichas regiones son cercanas entre sí.

Cada uno de los experimentos E-003, 004 y 005 involucran 3 sub-experimentos: i) para el primer semestre del año, ii) para el segundo semestre del año y iii) considerando todo el conjunto de datos. Los resultados mostrados a continuación corresponden a las estructuras finales analizadas de todos los experimentos. Los resultados completos de los experimentos E-001, 002, 003, 004 y 005 pueden encontrarse en el Anexo A.

- **Experimentos E-001 y E-002**

Para el experimento E-001, el cual comprende los atributos de clima recomendados por reportes técnicos y especialistas en café, los resultados indican que para la región "Jazmín" el mejor estimador externo es Random Forest con un RMSE de 0.17, para la región de "Naranjal" es GAM con un RMSE de 0.1472. Finalmente, para la región "Santagueda" el mejor estimador externo es Random Forest con un RMSE de 0.175. En promedio, para las condiciones del experimento E-001, Random Forest (RF) obtiene los mejores resultados (Tabla 6) y cumple con los demás criterios de satisfacción. Los atributos compartidos entre las tres regiones utilizando RF son: humedad, ΔT, temperatura máxima y temperatura mínima (ver Anexo A).

Por su parte, bajo las condiciones del experimento E-002, los resultados indican que, en promedio, el estimador externo que mejor se comporta en las tres regiones cafeteras es BT (Tabla 7). Los atributos compartidos por las regiones cafeteras durante este experimento utilizando BT son: presión de vapor, ΔT, humedad, temperatura media, temperatura máxima, temperatura mínima, y punto de rocío (Ver Anexo A).

Las tablas 6 y 7 muestran los resultados finales de los experimentos E-001 y E-002, resaltando en color verde los mejores estimadores externos.

Estimador externo	RMSE promedio	Atributos seleccionados

RF	0,17493973	4
LM	0,12928524	2
GAM	0,1731288	1
BT	0,17580007	4

Tabla 6. RMSE promedio y cantidad de atributos resultantes de la intersección de las regiones “Jazmín”, “Naranjal” y “Santagueda” para el experimento E-001

Estimador externo	RMSE promedio	Atributos seleccionados
RF	0,172030867	4
LM	0,172692933	6
GAM	0,1714155	2
BT	0,171690367	7

Tabla 7. RMSE promedio y cantidad de atributos resultantes de la intersección de las regiones “Jazmín”, “Naranjal” y “Santagueda” para el experimento E-002

- **Experimentos E-003, E-004 y E-005**

En el experimento E-003, el cuál utiliza únicamente los datos del periodo de floración del cultivo, el estimador externo que obtiene el menor RMSE en promedio y cumple con los demás criterios de satisfacción es RF (Tabla 8). Por otro lado, para el experimento E-004 que utiliza los datos correspondientes al periodo de formación de hojas, los resultados muestran que, para el primer semestre del año el estimador RF obtiene mejores resultados, mientras que en el segundo semestre del año el algoritmo BT obtiene el mejor RMSE promedio, con tres atributos seleccionados (Tabla 9). Finalmente, la Tabla 10 muestra que para el experimento E-005 que utiliza los datos durante el periodo de cosecha del cultivo, el mejor estimador externo para las tres regiones cafeteras es BT.

Estimador externo	Estructura del dataset	RMSE promedio	Atributos seleccionados
RF	Todo el conjunto de datos	0,0812838	5
LM		1,3045451	0
GAM		0,1629495	0
BT		0,0836244	3

Tabla 8. RMSE promedio y cantidad de atributos resultantes de la intersección de las regiones “Jazmín”, “Naranjal” y “Santagueda” para el experimento E-003

Estimador externo	Estructura del dataset	RMSE promedio semestre I	RMSE promedio semestre II	Atributos seleccionados Semestres I y II	
RF	Conjunto de datos separados en dos semestres	0,03209	0,1089353	6	7
LM		0,0330645	0,10584401	2	2
GAM		0,1085783	0,22333333	0	0
BT		0,0321838	0,10710797	5	3

Tabla 9. RMSE promedio y cantidad de atributos resultantes de la intersección de las regiones “Jazmín”, “Naranjal” y “Santagueda” para el experimento E-004

Estimador externo	Estructura del dataset	RMSE promedio	Atributos seleccionados
RF	Todo el conjunto de datos	0,0850004	2
LM		0,0856614	1
GAM		0,197516	0
BT		0,0846666	5

Tabla 10. RMSE promedio y cantidad de atributos resultantes de la intersección de las regiones “Jazmín”, “Naranjal” y “Santagueda” para el experimento E-005

Teniendo en cuenta los experimentos descritos anteriormente, los resultados muestran que el análisis de datos para la detección de roya en el café debe hacerse por las etapas del cultivo puesto que los resultados de los experimentos E-003, 004 y 005 son

mejores que en los experimentos E-001 y 002 que evalúan la totalidad de los datos. Además, se observa que los atributos seleccionados para el conjunto de datos que sirve de entrada al método de conjunto para la detección de roya en cultivos de café colombianos se organizan de la siguiente manera:

- Para la etapa de floración, el estimador externo seleccionado es random forest con un RMSE promedio de 0,0812838. Los atributos seleccionados por el algoritmo RFE en esta etapa son: temperaturas máxima, mínima y media, humedad, y amplitud térmica
- Para la etapa de formación de hojas, el análisis debe dividirse entre los dos semestres del año cafetero, por lo tanto, los estimadores seleccionados son random forest para el primer semestre del año (RMSE=0,03209), y bagging tree para el segundo semestre del año (RMSE=0,10710797). Los atributos seleccionados son:
    - Punto de rocío, temperaturas mínima y máxima, humedad, horas de sol y amplitud térmica (para muestras del primer semestre del año)
    - Temperaturas mínima, media y amplitud de temperatura (para muestras del segundo semestre del año)
- Para la etapa de cosecha, el estimador externo seleccionado es bagging tree con un RMSE promedio de 0,0846666. Los atributos seleccionados en esta etapa son: temperaturas mínima, máxima y media, humedad y amplitud térmica

Dado que la literatura científica afirma que la variable de precipitación es una característica fundamental para el estudio del desarrollo de la roya, se realizó un análisis profundo de los valores de precipitación de los distintos conjuntos de datos estudiados, con el objetivo de evaluar el impacto de dichas variables en el progreso de la enfermedad (Ver Anexo A). Los resultados muestran que, para los conjuntos de datos de las regiones Santagueda, Naranjal y Jazmín, los niveles de precipitación durante los años de análisis fueron en general constantes con bajos niveles de precipitación, lo cual de acuerdo al conocimiento experto es un ambiente ideal para el aumento de la roya en los cultivos [13][16][53]. Aunque los resultados del proceso de selección de características no consideran la precipitación como una variable de alto impacto en el desarrollo de la enfermedad, se incluye la precipitación total en el proceso de experimentación y evaluación de los siguientes capítulos debido a la

importancia que tiene la variable en el estudio de la roya según las consideraciones del conocimiento experto.

3.3 Resumen

Este capítulo describió los procesos realizados para la comprensión y posterior preparación de los conjuntos de datos que son utilizados para la construcción del método de conjunto para la detección de roya. En primer lugar, se encuentra la descripción básica de los conjuntos de datos pertenecientes a tres regiones cafeteras de Colombia. En segundo lugar, son descritos a profundidad los procesos realizados dentro de los conjuntos de datos para resolver problemas encontrados en los mismos (valores faltantes y selección de características) y así preparar los conjuntos de datos para la etapa de modelamiento del capítulo 4. Los resultados de este capítulo presentan un mecanismo para la generación de muestras sintéticas de roya en Colombia, y una selección de atributos donde los mejores experimentos son: experimento E-003 utilizando el dataset DS-003, experimento E-004 usando los datasets DS-001 y DS-002, y el experimento E-005 con el dataset DS-003. Los experimentos E-003, 004 y 005 corresponden a dividir el análisis de conjunto de datos por las tres etapas del cultivo estudiadas (formación de hojas, floración y cosecha).

Capítulo 4: modelado

En este capítulo se presenta el método de conjunto de clasificadores para la detección de roya en cultivos de café basado en datos sintéticos. En primer lugar, se realiza un análisis de series de tiempo para las variables climatológicas, con el objetivo de permitir al método de conjunto realizar además de una detección, una predicción temprana sobre el comportamiento de la IR, y de esta manera fijar un precedente del uso de técnicas de series de tiempo en características de clima para el estudio de la roya en el café en tareas de predicción. En segundo lugar, se presenta la estructura del método de conjunto de acuerdo a los resultados de los experimentos mostrados en el capítulo 3.

4.1 Análisis de series de tiempo

Con el objetivo de realizar un pronóstico de la IR de al menos cinco días, se realizaron tres experimentos en las diferentes variables de clima de entrada al método de conjunto, utilizando series de tiempo tradicionales que se adaptan al contexto de estudio y que han sido ampliamente utilizadas para tareas de pronóstico (Box-jenkins, Holt winters) [54][55] y redes neuronales artificiales (Artificial Neural Networks, ANN) como una técnica de pronóstico de series de tiempo no lineal [56].

Considerando los tres conjuntos de datos (Jazmín, Naranjal y Santagueda), ocho variables de clima (Temperaturas mínima, media, máxima, punto de rocío, humedad, brillo solar, precipitación de día y noche), y las tres técnicas de pronóstico de series de tiempo (Box-jenkins, Holt-winters, ANN), se realizaron 1056 casos experimentales. Los experimentos son descritos a continuación teniendo en cuenta que cada experimento se realizó para los tres conjuntos de datos disponibles, de los cuales se utilizó el 70% de los datos como datos de entrenamiento, y el 30% restante como datos de validación. Los experimentos ES-001 y ES-002 utilizan los métodos tradicionales para el pronóstico con series de tiempo. Por otra parte, el experimento ES-003 utiliza el algoritmo ANN con tres distintas configuraciones de entrada (5 atributos, 10 atributos y 15 atributos de entrada) y tres configuraciones en los parámetros de la red neuronal, por lo cual un experimento de red neuronal puede ser referenciado en este capítulo de la siguiente manera: ANN-2-10, donde ANN indica la red neuronal, ANN-2 la configuración de la red neuronal, y ANN-2-10 indica el número de atributos de entrada para la configuración 2 de la red neuronal.

- **Experimento ES-001:** utiliza el método Box-jenkins y cuatros configuraciones del conjunto de datos:
    - **Dataset DSS-001:** todo el conjunto de datos
    - **Dataset DSS-002:** todos los datos correspondientes a la etapa de formación de hojas
    - **Dataset DSS-003:** todos los datos correspondientes a la etapa de floración
    - **Dataset DSS-004:** todos los datos correspondientes a la etapa de cosecha
- **Experimento ES-002:** utiliza el método Holt-winters y cuatros configuraciones del conjunto de datos:
    - **Dataset DSS-001:** todo el conjunto de datos
    - **Dataset DSS-002:** todos los datos correspondientes a la etapa de formación de hojas
    - **Dataset DSS-003:** todos los datos correspondientes a la etapa de floración
    - **Dataset DSS-004:** todos los datos correspondientes a la etapa de cosecha

- **Experimento ES-003:** utiliza redes neuronales artificiales, cuatro configuraciones del conjunto de datos y los siguientes subexperimentos por cada subconjunto de datos:
    - **ES-003-1: "ANN-1"**

      Configuración de la red neuronal:
        - Learning rate: 0.3
        - Momentum: 0.2
        - Training time: 500
        - Capas ocultas de la red: (# atributos + clase) /2

      ES-003-1-1: "ANN-1-5"
        - DSS-001: 5 atributos, 1 clase, Jazmín
        - DSS-002: 5 atributos, 1 clase, Naranjal
        - DSS-003: 5 atributos, 1 clase, Santagueda

      ES-003-1-2: "ANN-1-10"
        - DSS-001: 10 atributos, 1 clase, Jazmín
        - DSS-002: 10 atributos, 1 clase, Naranjal
        - DSS-003: 10 atributos, 1 clase, Santagueda

      ES-003-1-3: "ANN-1-15"
        - DSS-001: 15 atributos, 1 clase, Jazmín
        - DSS-002: 15 atributos, 1 clase, Naranjal
        - DSS-003: 15 atributos, 1 clase, Santagueda

    - **ES-003-2: "ANN-2"**

      Configuración de la red neuronal:
        - Learning rate: 0.5
        - Momentum: 0.2
        - Training time: 500
        - Capas ocultas de la red: (# atributos+ clase) /2

      ES-003-2-1: "ANN-2-5"
        - DSS-001: 5 atributos, 1 clase, Jazmín
        - DSS-002: 5 atributos, 1 clase, Naranjal
        - DSS-003: 5 atributos, 1 clase, Santagueda

      ES-003-2-2: "ANN-2-10"

- DSS-001: 10 atributos, 1 clase, Jazmín
- DSS-002: 10 atributos, 1 clase, Naranjal
- DSS-003: 10 atributos, 1 clase, Santagueda

ES-003-2-3: “ANN-2-15”

- DSS-001: 15 atributos, 1 clase, Jazmín
- DSS-002: 15 atributos, 1 clase, Naranjal
- DSS-003: 15 atributos, 1 clase, Santagueda

- **ES-003-3: “ANN-3”**

  Configuración de la red neuronal:

  - Learning rate: 0.2
  - Momentum: 0.2
  - Training time: 500
  - Capas ocultas de la red: (# atributos+ clase) /2

  ES-003-3-1: “ANN-3-5”

  - DSS-001: 5 atributos, 1 clase, Jazmín
  - DSS-002: 5 atributos, 1 clase, Naranjal
  - DSS-003: 5 atributos, 1 clase, Santagueda

  ES-003-3-2: “ANN-3-10”

  - DSS-001: 10 atributos, 1 clase, Jazmín
  - DSS-002: 10 atributos, 1 clase, Naranjal
  - DSS-003: 10 atributos, 1 clase, Santagueda

  ES-003-3-3: “ANN-3-15”

  - DSS-001: 15 atributos, 1 clase, Jazmín
  - DSS-002: 15 atributos, 1 clase, Naranjal
  - DSS-003: 15 atributos, 1 clase, Santagueda

La implementación de los métodos Box-jenkins y Holt-winters provienen del paquete FOREIGN del lenguaje estadístico R. Por otro lado, las redes neuronales se implementaron con el paquete SCIKIT-LEARN del lenguaje de programación Python. Las medidas de evaluación utilizadas para la elección de los mejores experimentos fueron: el Error Relativo (ER) que se define como el cociente entre el valor absoluto de la predicción y el valor real [57], el valor absoluto medio (Mean Absolute Error, MAE)

que compara la diferencia de medida entre dos valores continuos que expresan el mismo fenómeno [58], y el coeficiente de correlación de Pearson (PCorr) el cual permite analizar la relación lineal existente entre el valor observado y el valor predicho; entre más cercano a 1 o -1 sea el coeficiente de correlación mejor es la correlación lineal [59]. Las expresiones matemáticas que representan las métricas de evaluación son las siguientes:

$$ER^i = \frac{|predicción^i - observado^i|}{predicción^i} \tag{4.1}$$

$$MAE = \frac{1}{n} * \sum_1^n (predicción - observado) \tag{4.2}$$

$$\text{PCorr} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{4.3}$$

Donde:

- n: número de observaciones
- σ_{XY}: covarianza
- σ_X: desviación estándar de X
- σ_Y: desviación estándar de Y

Los resultados de los experimentos para determinar las mejores técnicas de pronóstico de series de tiempo para las distintas variables de clima se organizan a continuación por las variables de clima de cada región. En este capítulo se presentan los resultados de la variable de temperatura mínima para los experimentos que abarcan el subconjunto de datos de “todos los datos disponibles”. Además, se realiza un estudio comparativo para determinar las mejores técnicas de series de tiempo en cada variable de clima. Los resultados completos de los experimentos de esta etapa se encuentran en el Anexo B.

4.1.1 Temperatura mínima – Todos los datos disponibles

A continuación, son presentados los resultados de la temperatura mínima en las tres regiones cafeteras (Figuras 10, 11 y 12) utilizando todos los datos disponibles y las tres técnicas de series de tiempo utilizadas. Las métricas de evaluación ER, MAE y

correlación de Pearson nos permiten realizar un estudio de la capacidad de predicción de series de tiempo de cada una de las técnicas mencionadas en los presentes experimentos.

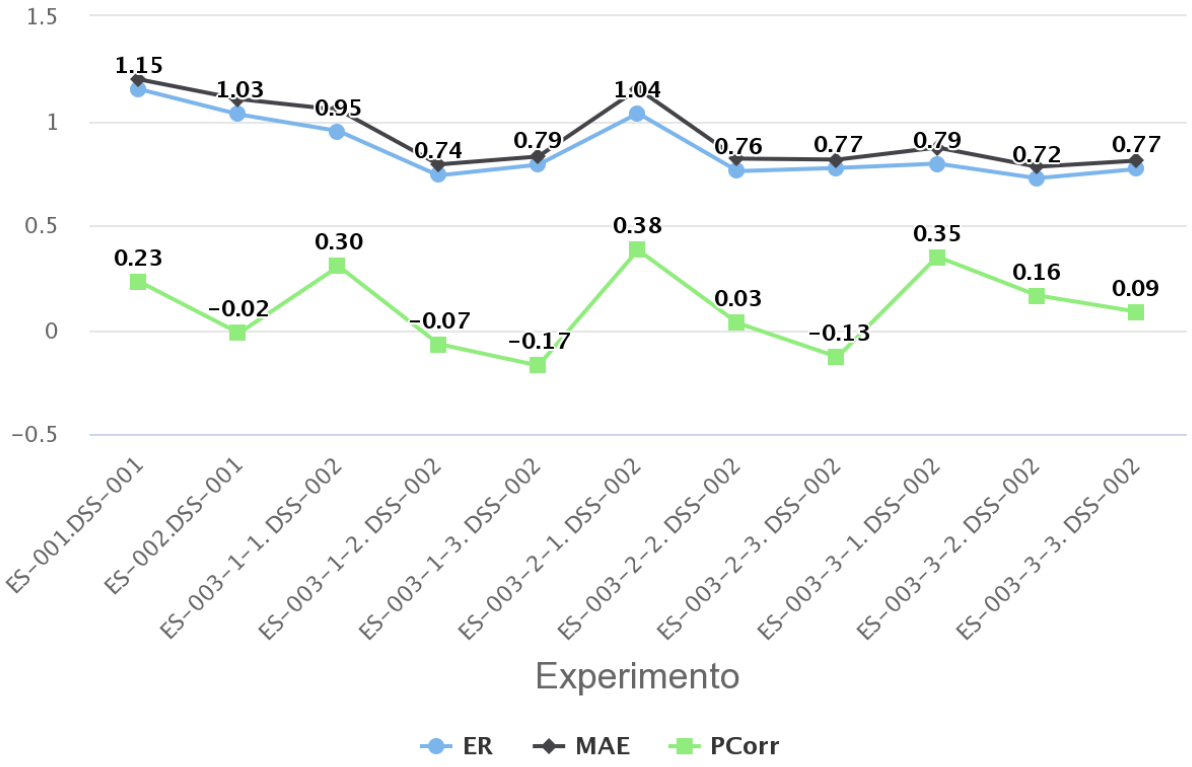


Figura 10. Resultados de los experimentos correspondientes a la intersección de la temperatura mínima y el subconjunto “Todos los datos disponibles” para la región “Naranjal”

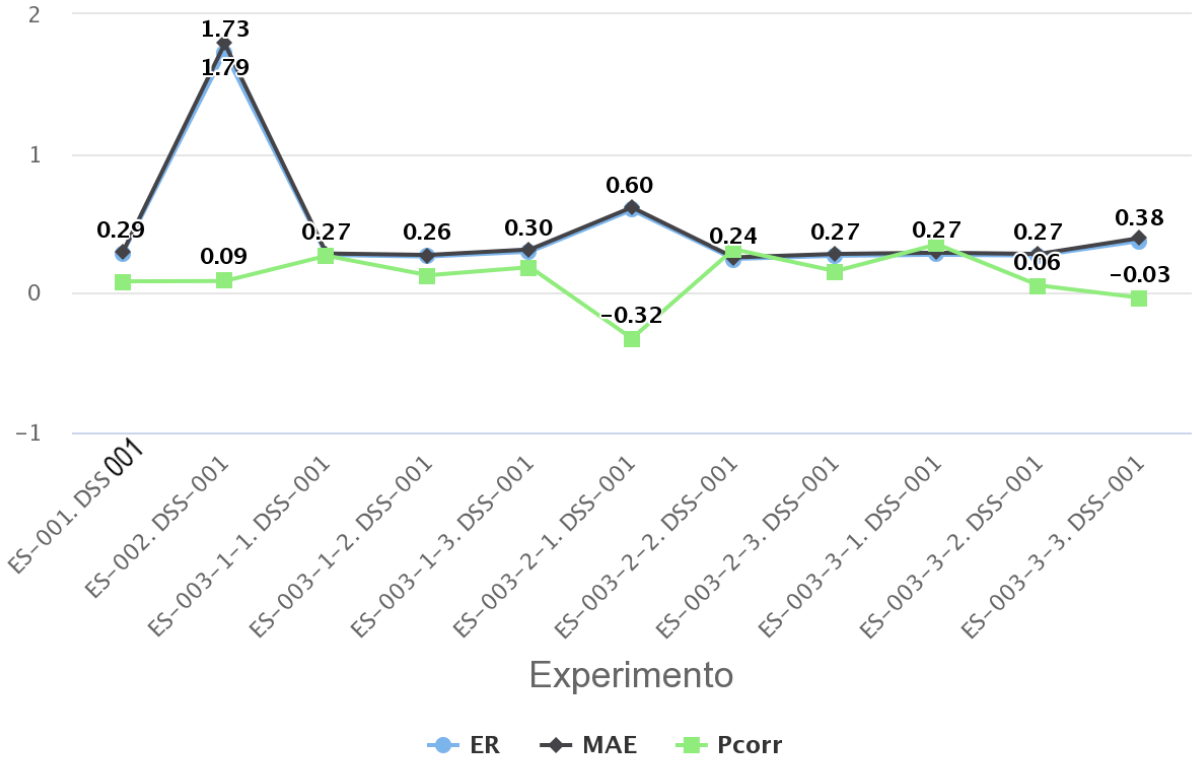


Figura 11. Resultados de los experimentos correspondientes a la intersección de la temperatura mínima y el subconjunto “Todos los datos disponibles” para la región “Jazmín”

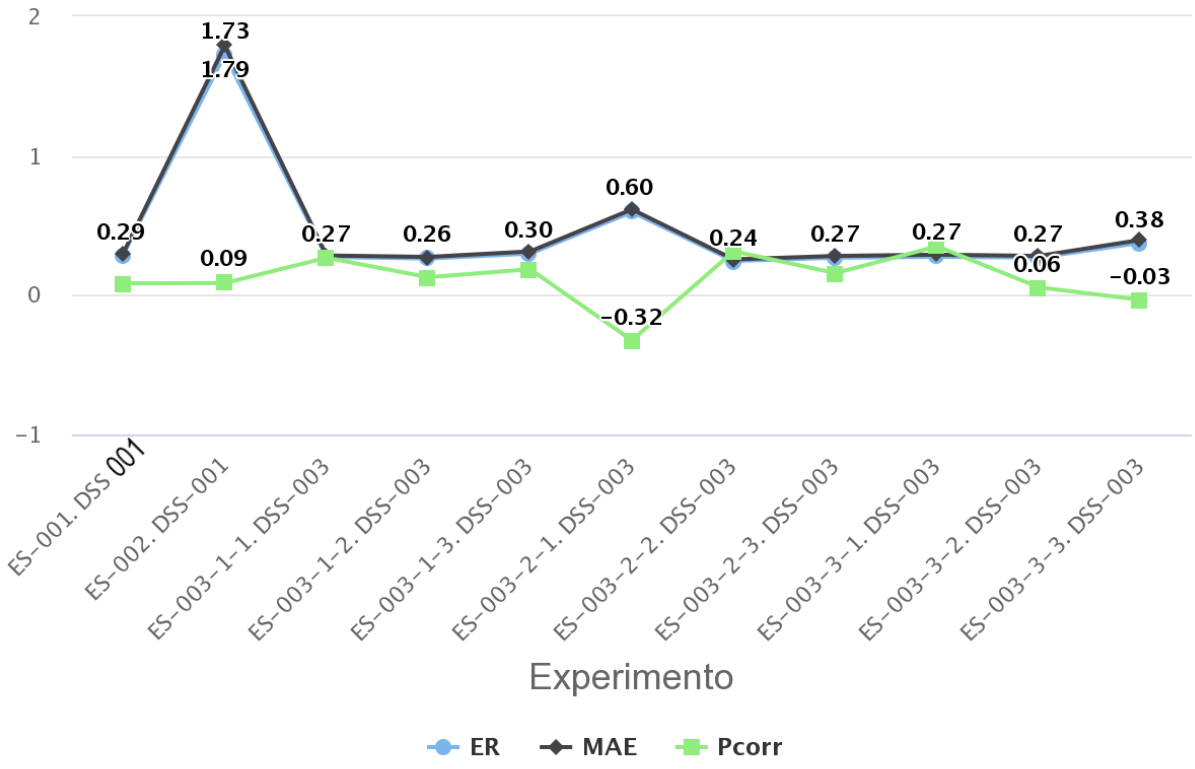


Figura 12. Resultados de los experimentos correspondientes a la intersección de la temperatura mínima y el subconjunto “Todos los datos disponibles” para la región “Santagueda”

Los resultados presentados en las Figuras 10, 11 y 12 muestran que los valores medidos de error son cercanos entre sí, sin embargo, los coeficientes de correlación presentan diferencias significativas. Para la región “Naranjal” el menor error absoluto medio (MAE) encontrado en la variable de temperatura mínima fue 0.71 y su mejor coeficiente de correlación se aproxima a 0.38. Para la región “Jazmín” el mejor resultado de error relativo es 0.245 y su mejor correlación de -0.325. Finalmente, para la región “Santagueda” los mejores resultados de MAE y correlación de Pearson son 0.94 y 0.65 respectivamente. Para la elección de los experimentos con los mejores resultados, se realiza a continuación un estudio comparativo sobre las métricas obtenidas en esta sección (Figuras 10, 11 y 12).

4.1.2 Estudio comparativo

En adición a las métricas calculadas anteriormente sobre cada variable climatológica, se utilizaron dos evaluaciones de significancia estadística: el test de Friedman y el T-test pareado, con el objetivo de elegir para cada variable de clima el experimento con mejores predicciones. El test de Friedman identifica la existencia de una significancia estadística siempre y cuando la distribución de los datos que se comparan es la misma [60]. Por otra parte, el T-test pareado compara la media de los datos en diferentes partes de una muestra [60], es decir, no requiere que los datos de comparación sean de distribuciones parecidas, pero si requiere que provengan de la misma población de datos. A continuación, se muestra en la Figura 13 las fases de evaluación de significancia estadística con la variable de temperatura mínima para la región "El Naranjal". Estas fases de evaluación aplican de igual forma para las demás variables de clima analizadas y para todos los conjuntos de datos.

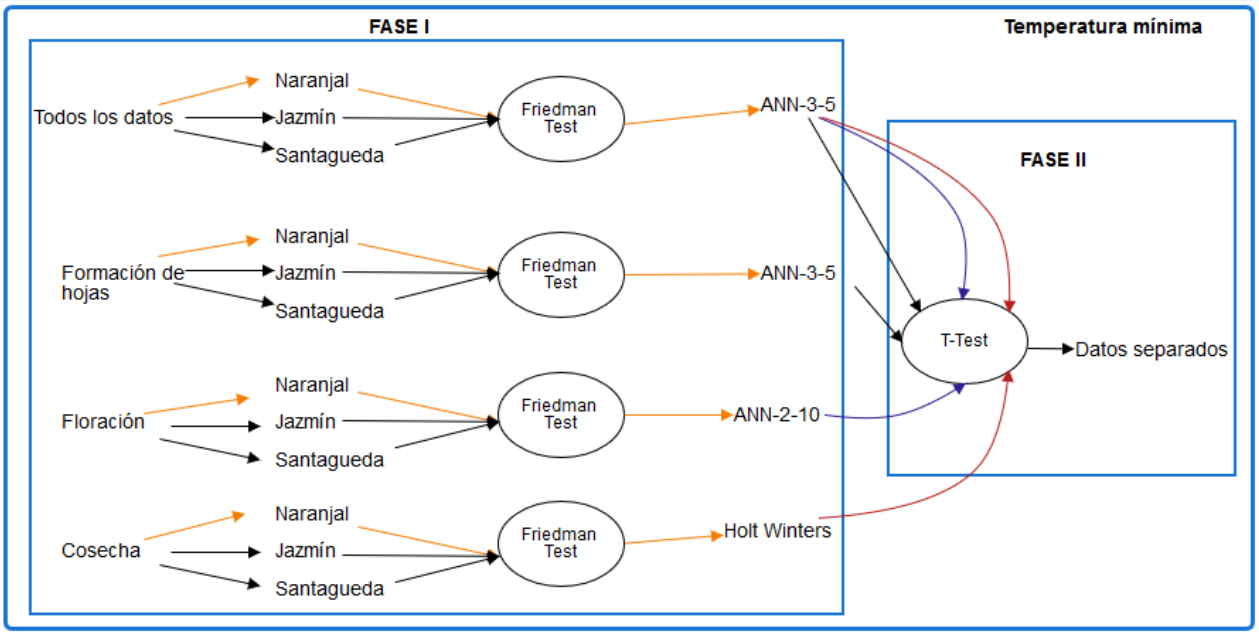


Figura 13. Fases de evaluación utilizando los test de Friedman y t-test pareado para la elección de métodos de pronóstico con series de tiempo para la variable de temperatura mínima de la región "El Naranjal"

Conforme a la primera fase de evaluación presentada en la Figura 13, el primer paso es hacer uso de la prueba de Friedman para determinar la existencia o no de significancia estadística en los resultados de los experimentos con series de tiempo

para la variable de temperatura mínima en cada una de las regiones (Figuras 10, 11 y 12), con el objetivo de determinar si algún experimento era significativamente mejor que los demás.

La prueba de Friedman consta de dos métricas de significancia estadística: el valor p y el puntaje o "score". El valor *p* muestra el nivel de significancia estadística de los resultados. Un valor p > 0.05 indica que no existe una diferencia estadística entre las muestras de datos. Por su parte, un valor p <= 0.05 muestra que existe una diferencia estadística significativa entre las muestras de datos. Finalmente, el puntaje o "score" es una métrica que dada una diferencia estadística significativa entre los datos (valor p <= 0.05), indica el nivel de la diferencia estadística, es decir, la cercanía o lejanía de un resultado con otro. Valores cercanos a cero del puntaje revelan una diferencia estadística pequeña entre las muestras de datos. Por otro lado, valores altos del puntaje, indican que existe una gran diferencia estadística entre los datos.

La tabla 11 muestra que, para la temperatura mínima de la región Naranjal en el conjunto de datos correspondiente a todos los datos disponibles, el valor p ($7.08576\,\mathrm{x}10^{-7}$) es menor a 0.05, lo cual indica dos cosas: i) Existe una diferencia estadística significativa entre elegir un experimento u otro, ii) El valor que brinda el "score" es confiable y determina que tanto se alejan los experimentos entre sí (score=47.685, alto). Para la elección de los mejores algoritmos de series de tiempo en la primera fase de evaluación se consideraron dos posibilidades:

i) Existe una diferencia significativa estadísticamente entre elegir uno u otro experimento, por lo cual se analiza el modelo que guarde una mejor relación entre los errores encontrados y el nivel de correlación
ii) No existe una diferencia entre elegir uno u otro experimento, en este punto se elige el modelo con mejor nivel de correlación en las predicciones.

En este caso, dado que para la temperatura mínima de la región "Naranjal" existe una diferencia estadística significativa entre los experimentos, se elige el modelo que guarde una mejor relación entre los errores y la correlación dado los resultados de la Figura 10, es decir, se busca en el modelo que tenga errores pequeños y una correlación lo más alta posible entre las predicciones y los datos de validación. En ese

sentido, el experimento que mejor cumple con las características mencionadas anteriormente para la temperatura mínima de “El Naranjal” es el experimento *ES-003-3-1-DSS-002*, correspondiente a la red neuronal ANN-3-5 (Figura 10). La tabla 8 muestra los resultados finales de esta fase para el conjunto de datos de “Todos los datos disponibles”, nombrando los modelos elegidos para la siguiente fase en la evaluación de algoritmos de series de tiempo. Los resultados de los conjuntos de datos restantes se encuentran en el Anexo B.

Región	Atributo	Score	p- value	Modelo
Jazmin	MinTem	39,559	2,02E-05	ANN-3-5
Jazmin	MaxTem	56,939	1,36E-08	ANN-1-10
Jazmin	AvgTem	84,352	7,00E-14	ANN-1-15
Jazmin	PR	39,384	2,17E-05	ANN-3-10
Jazmin	Hum	47,207	8,65E-07	Holt winters
Jazmin	Brillo solar	12,365	0,26	Holt winters
Jazmin	Pd	32,792	0,00	ANN-1-10
Jazmin	Pn	9,7135	0,46	ANN-1-10
Naranjal	MinTem	47,685	7,08E-07	ANN-3-5
Naranjal	MaxTem	64,240	5,66E-10	ANN-3-10
Naranjal	AvgTem	57,422	1,11E-08	ANN-1-15
Naranjal	PR	32,478	0,00	ANN-1-5
Naranjal	Hum	48,906	4,23E-07	ANN-1-10
Naranjal	Brillo solar	18,913	0,01	ANN-2-15
Naranjal	Pd	30,209	0,00	ANN-3-5
Naranjal	Pn	27,898	0,00	ANN-2-5
Santagueda	MinTem	56,029	2,02E-08	ANN-3-10
Santagueda	MaxTem	56,951	1,36E-08	Holt winters
Santagueda	AvgTem	48,371	2,17E-07	ANN-2-15
Santagueda	PR	48,629	4,76E-07	Box jenkins
Santagueda	Hum	58,981	5,64E-09	ANN-1-10
Santagueda	Brillo solar	59,635	4,24E-09	ANN-3-5
Santagueda	Pd	71,557	2,21E-11	Holt winters
Santagueda	Pn	55,269	2,81E-08	ANN-1-5

Tabla 11. Resultados de significancia estadística para las distintas regiones cafeteras (Jazmín, Naranjal, Santagueda) utilizando todos los datos disponibles y el test de Friedman

Una vez elegidos los mejores modelos para las distintas regiones cafeteras, se utiliza en la fase II el t-test pareado con el objetivo de determinar si es mejor utilizar para el pronóstico de series de tiempo todos los datos disponibles, o es más adecuado realizar

pronósticos por las etapas del cultivo analizadas (formación de hojas, floración y cosecha). La tabla 12 muestra los resultados de la fase II de evaluación para la temperatura mínima de la región Naranjal. Los resultados de la tabla muestran que no hay una diferencia significativa entre utilizar todo el conjunto de datos haciendo uso del algoritmo ANN-3-5 o dividirlo en varios subconjuntos utilizando el mismo algoritmo (valores p>0,05). En este caso, ya que los errores no son significativos, se opta por elegir los modelos que tengan una mejor correlación lineal al igual que en la fase anterior, por lo cual se decide realizar el pronóstico de series de tiempo dividiendo el dataset por las etapas del cultivo de café analizadas. Los algoritmos elegidos por guardar mejor correlación en la temperatura mínima del conjunto de datos “Todos los datos disponibles” son:

- Formación de hojas: ANN-3-5 (Ver Anexo B. Tabla B1)
- Floración: ANN-2-10 (Ver Anexo B. Tabla B2)
- Cosecha: Holt-Winters (Ver Anexo B. Tabla B3)

Temperatura mínima			
Conjuntos de datos	**Score**	**P-value**	**Decisión final**
Todos los datos vs Formación de hojas	-2.20	0,036	Datasets separados por las etapas del cultivo de café
Todos los datos vs Floración	-0,83	0,412	
Todos los datos vs Cosecha	-2,03	0,051	

Tabla 12. Resultados de significancia estadística para las distintas etapas del cultivo de la variable de temperatura mínima en la región Naranjal utilizando la prueba t-test pareado

A continuación, en la Tabla 13 se muestran los resultados finales de los experimentos para todas las variables de clima de las diferentes regiones cafeteras estudiadas. La Tabla 13 muestra el tipo de dataset elegido (Todos los datos TD, etapas del cultivo EC) y los métodos seleccionados dependiendo del tipo de dataset seleccionado.

Región	Atributo	Dataset elegido	Métodos seleccionados			
			Todos los datos	Formación de hojas	Floración	Cosecha
Naranjal	MinTem	EC		ANN-3-5	ANN-2-10	Holt-winters
Naranjal	MaxTem	TD	ANN-3-10			
Naranjal	AvgTem	TD	ANN-1-15			
Naranjal	PR	EC		ANN-3-5	ANN-2-5	Holt-winters
Naranjal	Hum	EC		ANN-2-10	Holt-winters	ANN-1-10
Naranjal	Sol	TD	ANN-2-15			
Naranjal	Pd	TD	ANN-3-5			
Naranjal	Pn	EC	ANN-2-5			
Jazmín	MinTem	TD	ANN-3-5			
Jazmín	MaxTem	TD	ANN-1-10			
Jazmín	AvgTem	TD	ANN-1-15			
Jazmín	PR	TD	ANN-3-10			
Jazmín	Hum	TD	Holt-winters			
Jazmín	Sol	TD	Holt-winters			
Jazmín	Pd	TD	ANN-1-10			
Jazmín	Pn	TD	ANN-1-10			
Santagueda	MinTem	TD	ANN-3-10			
Santagueda	MaxTem	TD	Holt-winters			
Santagueda	AvgTem	TD	Box-jenkins			
Santagueda	PR	TD	ANN-1-10			
Santagueda	Hum	TD	Holt-winters			
Santagueda	Sol	TD	ANN-3-5			
Santagueda	Pd	TD	Holt-winters			
Santagueda	Pn	EC		ANN-1-5	Holt-winters	ANN-2-15

Tabla 13. Métodos de pronósticos de series de tiempo seleccionados para las diferentes variables de clima de las regiones Jazmín, Naranjal y Santagueda

4.2 Método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos

En esta sección, se describe la construcción del método de conjunto de clasificadores teniendo en cuenta los resultados de la selección de características del capítulo 3 y las características mencionadas en capítulos anteriores sobre el comportamiento de la roya en el tiempo.

Para la construcción del método de conjunto, fueron considerados dos segmentos de análisis: configuración de las entradas al método de conjunto, y configuración de los parámetros de los algoritmos RF y BT.

4.2.1 Configuración de atributos de entrada al método de conjunto

La literatura científica [25] [3] afirma que, la medición de la incidencia de roya IR_1 (donde el subíndice 1 corresponde al día de medición 1) depende de las condiciones ambientales de los días anteriores a la medición observada, no solamente a las condiciones observadas el mismo día de la captura del dato, por ejemplo, la medición de la IR del día 28 del mes (IR_{28}) es la consecuencia de las condiciones ambientales del periodo de tiempo dado entre el 01 y el 28 día del mes observado. De esta forma, el formato de entrada de una variable dentro del conjunto de datos de entrenamiento y de prueba debe ser una serie de tiempo tal como se muestra en la Figura 14, donde *t* corresponde a un día de medición de un atributo de clima y su correspondiente valor IR_1, *t-1* al valor del atributo de clima el día anterior a la medición IR_1, *t-2* al valor del atributo de clima dos días anteriores a la medición IR_1, etc.

Dado que convertir en una columna del conjunto de datos de entrada cada día de una variable de clima (t, t-1,t-2,…etc) hasta llegar a los 28 días anteriores a la medición de roya observada implica un aumento considerable en la dimensionalidad del dataset, teniendo en cuenta que este proceso debe hacerse para cada variable de clima que haga parte de los atributos de entrada al método de conjunto, se opta por representar las series de tiempo en cuatro muestras que generalicen el comportamiento de los 15 días anteriores a la fecha de medición de la IR , tal como se muestra en la Figura 14 (recuadros verdes). De esta manera, por ejemplo, si el conjunto de datos de entrada tiene cuatro atributos de entrada, el total de columnas de dicho dataset sería de 4 atributos x 28 días = 112 columnas. Al tomar únicamente 4 muestras de cada atributo que representen los 15 días previos a la toma de la muestra de IR, tendremos un dataset de 4 atributos x 4 días representativos = 16 columnas.

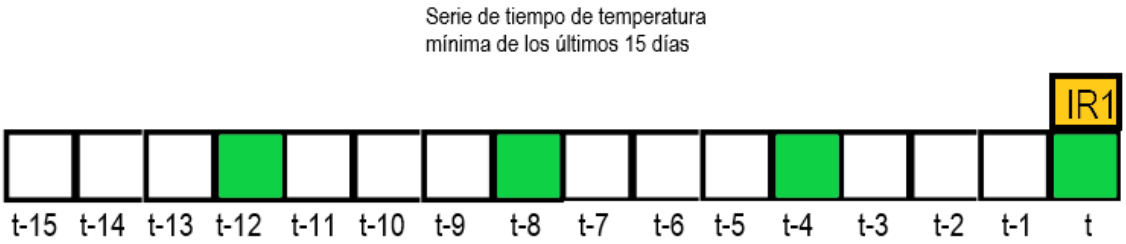


Figura 14. Serie de tiempo de la temperatura mínima en los últimos 15 días a la fecha de la medición de incidencia de roya IR1

Bajo las consideraciones de la Figura 14, el formato del conjunto de datos de entrenamiento del método de conjunto debe construirse de la siguiente manera:

1. Seleccionar una muestra de la IR
2. Seleccionar de cada atributo los valores de los días *t, t-4, t-8 y t-12* a la fecha de la muestra de IR seleccionada en el paso 1(estos son los representantes del comportamiento ambiental en los últimos 15 días a la fecha de medición de la IR)
3. Utilizar cada valor seleccionado como un atributo de entrada del conjunto de datos

La tabla 14 muestra el conjunto de datos para la etapa de formación de hojas durante el segundo semestre del año con la configuración mencionada anteriormente.

MinTem				**AvgTem**				**ΔT**				**IR**
t-12	**t-8**	**t-4**	**t**	**t-12**	**t-8**	**t-4**	**T**	**t-12**	**t-8**	**t-4**	**t**	**IR_1**
18,5	17	16,8	16,6	22	21,7	21,4	20,9	6	8	5	5	0,14

Tabla 14. Estructura del conjunto de datos de entrenamiento para una instancia de la etapa de formación de hojas durante el segundo semestre del año

4.2.2 Configuración de los métodos de conjunto Random Forest y Bagging Tree

Como segunda parte del proceso de modelado, en esta sección son mostrados los algoritmos de machine learning seleccionados en el proceso de selección de

características, así como su configuración en la arquitectura del método y la configuración final del método de conjunto.

Los métodos de ensamble Bagging Tree [61] y Random Forest [62] cuentan con las siguientes características comunes de implementación:

- Combinación de diferentes árboles de decisión
- Divide el número de muestras del dataset de entrada en partes iguales
- Cada nuevo subconjunto de datos creado es la entrada a un árbol de decisión individual. Por ejemplo, al tener una dataset de 100 instancias, se divide el dataset en 4 partes iguales de 25 instancias cada uno, de esta manera los métodos de conjunto construirán 4 árboles de decisión independientes
- El resultado final del método de ensamble es el promedio de los resultados parciales otorgados por cada árbol (Figura 15)

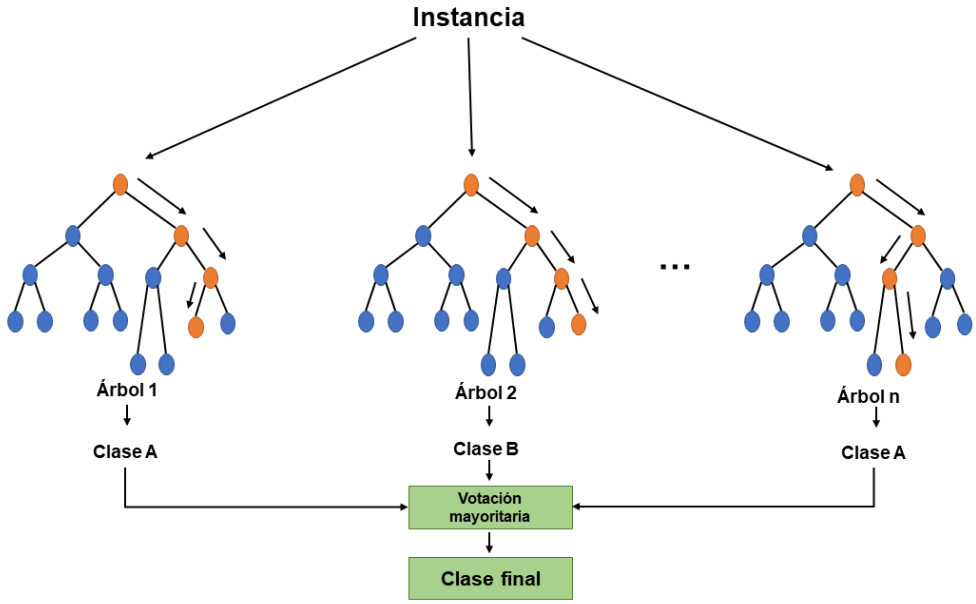


Figura 15. Estructura general simplificada de los métodos Random Forest y Bagging Tree para n árboles

La diferencia entre los métodos de ensamble radica en la forma de dividir cada nodo para crear una nueva ramificación. Para el caso del método BT, el algoritmo considera

todas las características de entrada para cada nodo a dividir. Por otro lado, RF considera algunas características seleccionadas aleatoriamente (Figura 16).

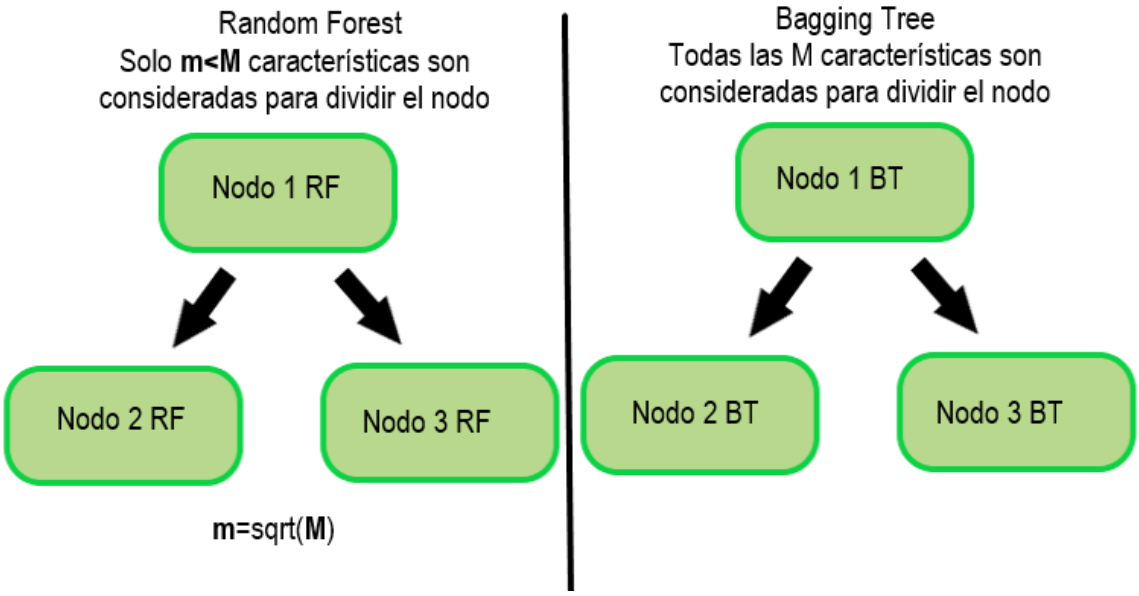


Figura 16. Diferencias entre RF y BT para la división de un nodo dentro de los árboles de decisión

La implementación de los métodos de ensamble en el método de conjunto se realizó con el paquete SCIKIT-LEARN del lenguaje de programación Python. Los parámetros técnicos configurados para ambos métodos de ensamble fueron:

- N_estimators: 10
- Max_depth: 4
- Random_state: 1

Donde, *n_estimators* nos indica el número de árboles que deseamos construir dentro de los métodos de ensamble, *max_depth* limita la profundidad de los árboles, es decir que especifica el número máximo de nodos a construir en cada árbol, y *random_state* permite que los experimentos se puedan replicar, ya que durante cada ejecución la estructura de los árboles puede cambiar durante el entrenamiento. La Figura 17 muestra la estructura final del método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos.

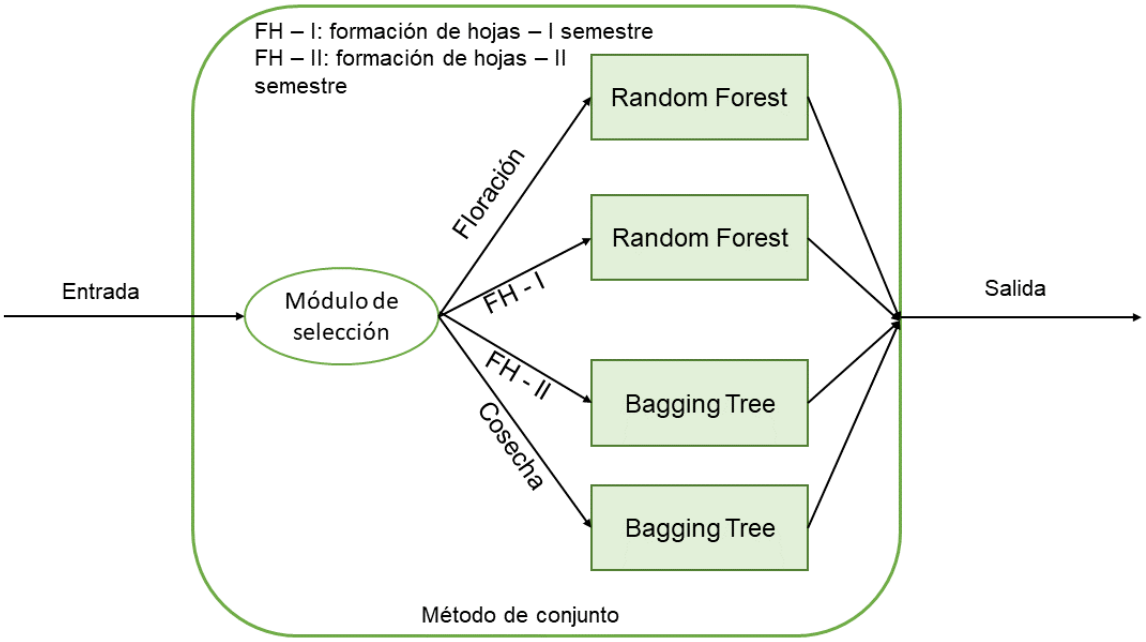


Figura 17. Método de conjunto de clasificadores para la detección de la incidencia de roya en cultivos de café colombianos

Dado que los resultados del capítulo 3 en la sección de selección de características muestran que, dependiendo la etapa del cultivo, ciertos atributos deben ser considerados en cada etapa, el método de conjunto contiene un módulo de selección de atributos, el cual se encarga de tomar únicamente los atributos necesarios para cada etapa del cultivo y así enviarlos posteriormente a los modelos de machine learning correspondientes que fueron elegidos en el análisis de selección de características, puesto que estos fueron los que mejor se adecuaron a los datos disponibles.

4.3 Resumen

En este capítulo fueron descritos las actividades realizadas entorno a la construcción del método de conjunto de clasificadores para la detección de roya. En primer lugar, se realizó un estudio sobre los atributos de los diferentes conjuntos de datos con el objetivo de poder realizar tareas de pronóstico de las condiciones ambientales para

otorgar la posibilidad de realizar pronósticos de la roya 5 días en el futuro. En este punto fueron seleccionados los algoritmos de pronóstico para cada una de las variables de clima analizadas (Tabla 13). Por otra parte, retomando los resultados de capítulos anteriores, se construyó y describió el método de conjunto de clasificadores para la detección de la incidencia de roya en las regiones de Colombia analizadas.

Capítulo 5: experimentación y evaluación

En este capítulo se presentan las pruebas realizadas para medir las capacidades del método de conjunto para la detección de la Incidencia de Roya en cultivos de café colombianos. Primero se realiza una prueba orientada a cada región cafetera analizada, y segundo, se realiza una prueba general donde se utiliza una de las tres regiones como conjunto de datos de entrenamiento, y las regiones restantes como conjuntos de datos de prueba, esto con el objetivo de evaluar las capacidades de generalización del método de conjunto en entornos de prueba distintos al entorno de entrenamiento.

5.1 Método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos

Con el objetivo de evaluar las capacidades del método de conjunto de clasificadores, el cual reúne todos los procesos mencionados anteriormente en esta propuesta, se realizaron tres pruebas de evaluación. La primera prueba se enfoca en cada una de las regiones analizadas utilizando las variables resultantes del proceso de selección de características. La segunda prueba considera las mismas variables y la precipitación total. De esta manera es posible evaluar el impacto de dicha variable en las tareas de detección de la enfermedad.

Para cada región se construyeron cuatro subconjuntos de datos, que corresponden a los resultados de las etapas del cultivo dados por la selección de características: floración, formación de hojas en el primer semestre del año (formación de hojas I), formación de hojas en el segundo semestre del año (formación de hojas II), y cosecha. El entrenamiento del método de conjunto se realizó con el 70% de los datos, el 30% restante se utilizó para validar las salidas del método de conjunto.

Finalmente, en la tercera prueba se realiza el mismo proceso de la prueba 1, con la diferencia que se utiliza el dataset de la región “Naranjal” como conjunto de datos de entrenamiento, y los datasets de las regiones “Santagueda” y “Jazmín” como conjuntos de datos de prueba.

Las métricas utilizadas para la evaluación del método de conjunto son: el Error Absoluto Medio (MAE), el coeficiente de correlación de Pearson (CCP), y el error cuadrático medio (MSE; Mean Squared Error), el cual corresponde a la suma de los errores cuadráticos de las muestras [63]. La expresión matemática que representa al MSE es el siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{pred\,i}\right)^2 \quad (5.1)$$

Donde, n es el número de residuos, y $\left(y_i - y_{pred\,i}\right)$ es el error en la predicción.

Los resultados de la primera prueba, la cual utiliza un conjunto de datos de entrenamiento y de prueba para cada región utilizando las variables resultantes de la selección de características se resumen en la tabla 15.

Región cafetera	Subconjunto de datos	MAE	MSE	CCP
Jazmín	Floración	16%	4.5%	-0.03
	Formación de hojas I	13%	2.3%	0.25
	Formación de hojas II	2%	0.1%	0.71
	Cosecha	9%	1.1%	0.27
Santagueda	Floración	8%	0.9%	0.54
	Formación de hojas I	3%	0.1%	0.28
	Formación de hojas II	0.5%	4.24 x10^{-5}%	0.69
	Cosecha	11%	2.1%	0.73
Naranjal	Floración	10%	1.6%	0.14
	Formación de hojas I	5%	0.4%	0.11
	Formación de hojas II	4%	0.2%	0.31
	Cosecha	11%	2%	0.48

Tabla 15. Resultados de MAE, MSE y CCP utilizando el método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos y los atributos resultantes de la selección de características

Los resultados de la tabla 15 muestran que, el menor error obtenido se da en la región “Santagueda” en el subconjunto de datos de formación de hojas en el segundo semestre del año (MAE=0.5% y MSE=4.24 x10^{-5}%), y el error más alto del método de conjunto se dio con la región Jazmín en la etapa de floración, obteniendo una MAE del

16% y un MSE del 4.5%. Los resultados muestran la facilidad de detección de incidencia de roya por parte del método de conjunto para cada subconjunto de datos, puesto que, para las tres regiones analizadas, en los periodos de formación de hojas en el segundo semestre del año y la cosecha, el error en la detección es en promedio del 2.1% y 10.3% respectivamente, siendo estos los mejores resultados, ya que además de tener bajos porcentajes de error, los resultados muestran niveles altos de correlación entre los datos observados y los predichos por el método de conjunto en estos periodos. Estos resultados demuestran que las variables de clima utilizadas para la detección de roya en estos periodos contienen suficiente información para detectar el comportamiento de la roya durante estas etapas. Las figuras 18 y 19 muestran los puntos originales analizados y las detecciones realizadas por el método de conjunto en los periodos correspondientes a la formación de hojas en el segundo semestre del año y la cosecha para la región “Santagueda”.

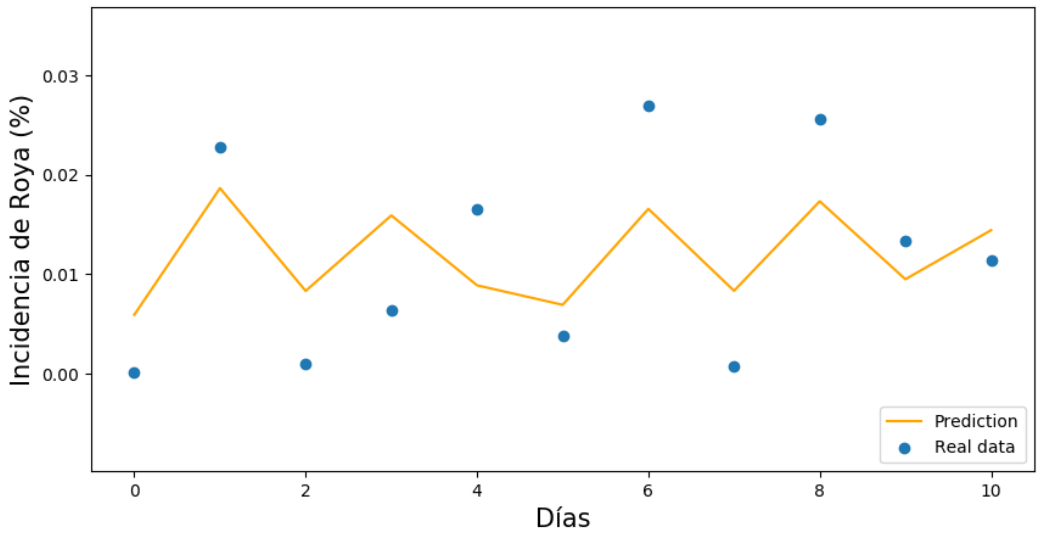


Figura 18. Detección de la IR durante el periodo de formación de hojas en el segundo semestre del año para la región “Santagueda” con la primera prueba de evaluación

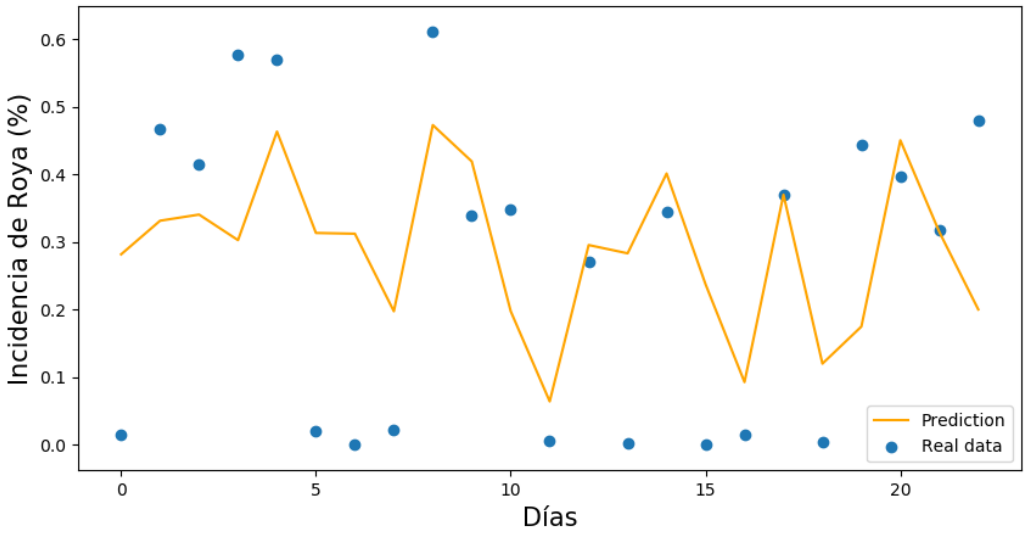


Figura 19. Detección de la IR durante el periodo de cosecha en el segundo semestre del año para la región “Santagueda” con la primera prueba de evaluación

Las figuras 18 y 19 muestran que, las detecciones realizadas por el método de conjunto se aproximan a los puntos originales observados, lo cual indica la facilidad de generalización que tiene el método de conjunto en estos periodos del cultivo para la región descrita. En segundo lugar, los periodos de formación de hojas para el primer semestre del año y la floración obtuvieron en promedio errores de 7% y 11,3% respectivamente. Aunque estos errores son cercanos al error promedio presentado por el periodo de cosecha, los coeficientes de correlación nos indican que aún se guarda cierto nivel de incertidumbre en cuanto a la consistencia del método de conjunto para realizar predicciones precisas en estos periodos del cultivo. Estos resultados nos indican lo siguiente: para que el método de conjunto tenga un mayor grado de exactitud en la detección de la enfermedad deben considerarse variables adicionales a las propuestas para estas etapas.

Las figuras 20 y 21 muestran los puntos originales analizados y las detecciones realizadas por el método de conjunto en los periodos correspondientes a la formación de hojas en el primer semestre del año y la floración para la región “Santagueda”. Los resultados completos de las pruebas se encuentran en el Anexo C.

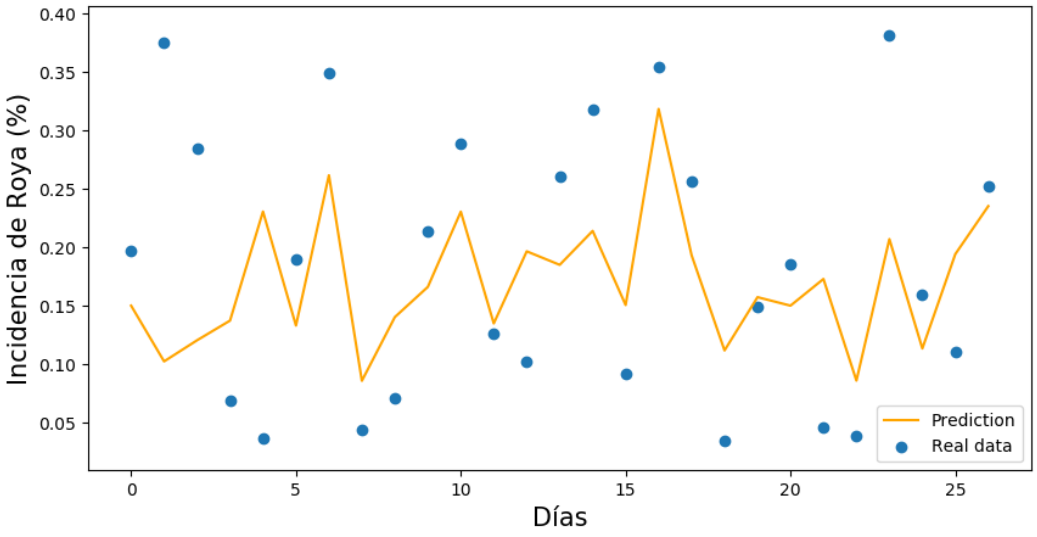


Figura 20. Detección de la IR durante el periodo de floración para la región “Santagueda” con la primera prueba de evaluación

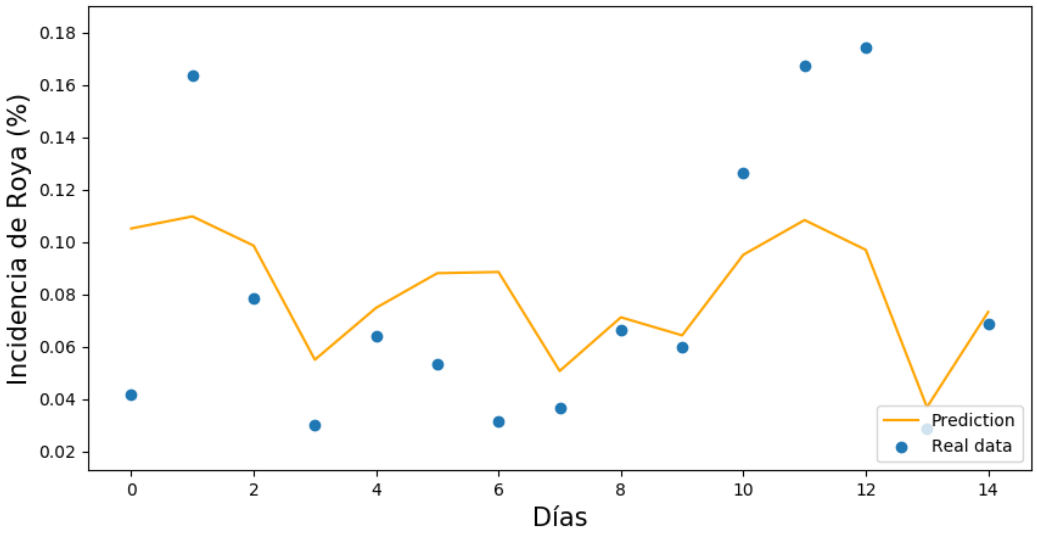


Figura 21. Detección de la IR durante el periodo de formación de hojas en el primer semestre del año para la región “Santagueda” con la primera prueba de evaluación

Los resultados de las figuras 20 y 21 muestran que para el método de conjunto es difícil realizar detecciones de valores altos de la IR en las etapas del cultivo analizadas.

Para el caso del periodo de formación de hojas en el primer semestre del año significa un nivel incertidumbre con valores por encima del 25%, mientras que para la floración significa incertidumbre por encima del 11%.

La segunda prueba utiliza adicionalmente información de la precipitación en las regiones para la detección de roya en los cultivos de café. La tabla 16 muestra los resultados obtenidos en la detección para cada región analizada.

Región cafetera	Subconjunto de datos	MAE	MSE	CCP
Jazmín	Floración	18.71%	5.5%	-0.02
	Formación de hojas I	12%	2%	0.34
	Formación de hojas II	3.2%	0.2%	0.62
	Cosecha	10%	1.4%	-0.11
Santagueda	Floración	7.4%	0.9%	0.53
	Formación de hojas I	2.5%	$9.1x10^{-4}$%	0.69
	Formación de hojas II	0.7%	$7.38x10^{-5}$%	0.51
	Cosecha	12.5%	2.1%	0.78
Naranjal	Floración	10.6%	1.8%	0.19
	Formación de hojas I	4%	0.3%	0.16
	Formación de hojas II	3.1%	0.1%	0.58
	Cosecha	9.7%	1.3%	0.54

Tabla 16. Resultados de MAE, MSE y CCP utilizando el método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos y los atributos resultantes de la selección de características

En general, los resultados de la Tabla 16 muestran el mismo comportamiento encontrado en la primera evaluación, donde los periodos de formación de hojas en el segundo semestre del año y la cosecha presentan los mejores resultados en la detección de la enfermedad, por otra parte, en los periodos de formación de hojas en el primer semestre del año y en la floración, los resultados presentan un mayor grado de incertidumbre. Sin embargo, se observa que la inclusión de la precipitación favoreció y empeoró ciertos resultados. En el caso de la floración y la cosecha, los cambios en los resultados no fueron muy alejados, sin embargo, en la mayoría de los casos la inclusión de la precipitación mejoró los resultados de error y correlación. Por otra parte, en los periodos de formación de hojas durante el primer semestre del año de todas las regiones, la inclusión de la precipitación mejoró considerablemente el error y la correlación de los resultados, mientras que en la formación de hojas del

segundo semestre afecto el rendimiento en la detección de roya en este periodo del cultivo.

Los resultados de las pruebas 1 y 2 muestran que la precipitación es una variable climática fundamental en el análisis de la roya en los cultivos de café, puesto que la inclusión de dicha variable en los experimentos mejoró el rendimiento del método de conjunto en general, sin embargo, se concluye que hace falta realizar un proceso de transformación adicional de la variable de precipitación, con el objetivo de identificar una representación de la precipitación que permita relacionar con mayor claridad el impacto que tiene la lluvia en el desarrollo de la roya.

La tercera prueba de evaluación, la cual consta de utilizar los cuatro subconjuntos de datos de la región “Naranjal” como conjuntos de datos de entrenamiento y los conjuntos de datos de las regiones “Santagueda” y “Jazmín” como conjuntos de prueba, muestran los siguientes resultados en la tabla 17:

Región cafetera	Subconjunto de datos	MAE	MSE	CCP
Jazmín	Floración	16.8%	4.1%	0.132
	Formación de hojas I	14.7%	3%	0.155
	Formación de hojas II	3.9%	0.2%	0.383
	Cosecha	17.1%	3.8%	0.021
Santagueda	Floración	9.4%	1.5%	0.122
	Formación de hojas I	3.8%	0.2%	0.018
	Formación de hojas II	3.3%	0.1%	0.167
	Cosecha	25.7%	9.3%	0.245

Tabla 17. Resultados de MAE, MSE y CCP utilizando el método de conjunto de clasificadores para la detección de roya en cultivos de café colombianos con el conjunto de datos “Naranjal” como dataset de entrenamiento

Los resultados de la tabla 17 muestran que, para la región “Jazmín” los errores analizados y los coeficientes de correlación son semejantes a los resultados dados en la primera prueba realizada (Tabla 15). Sin embargo, para los periodos de formación de hojas durante el segundo semestre del año y la cosecha, los resultados desmejoran considerablemente.

Realizando un contraste entre las variables de clima de la región “Jazmín” y la región “Naranjal” es posible observar diferencias en la clase objetivo (IR). La Incidencia de Roya en la región “Jazmín” para el periodo de formación de hojas durante el segundo semestre del año que presenta las características mostradas en la tabla 18, y la IR en la región “Naranjal” durante el mismo periodo del cultivo (Tabla 19) no comparten distribuciones de datos semejantes, ya que, para esta prueba el conjunto de datos de entrenamiento (datos de la región “Naranjal”) no posee ejemplos superiores al 1% de incidencia de roya (Tabla 19), mientras que el conjunto de datos de prueba correspondiente a la región “Jazmín” si los posee. Por lo tanto, es difícil para el método de conjunto en este caso detectar porcentajes de la IR que no conoce. En otras palabras, el método de conjunto está recibiendo instancias de las cuales no tiene ninguna información, lo cual puede explicar el desmejoramiento en la detección de la IR por parte del método de conjunto para el periodo de formación de hojas durante el segundo semestre del año para la prueba realizada.

Variable	Número de muestras	Valor Promedio	Desviación estándar	Valor Mínimo	Valor Máximo
IR	90	6%	6%	0%	17%

Tabla 18. Estadística descriptiva para la Incidencia de Roya durante el periodo de formación de hojas en el segundo semestre del año para la región “Jazmín”

Variable	Número de muestras	Valor Promedio	Desviación estándar	Valor Mínimo	Valor Máximo
IR	90	0%	0.006%	0%	1%

Tabla 19. Estadística descriptiva para la Incidencia de Roya durante el periodo de formación de hojas en el segundo semestre del año para la región “Naranjal”

Casos similares se presentan en la comparación entre los subconjuntos de datos de la región “Santagueda” y la región “Naranjal” (Anexo C), demostrando que las condiciones ambientales tienen un impacto distinto en el progreso de la IR para cada región analizada. Por lo tanto, el método de conjunto propuesto es adaptable a las distintas regiones analizadas de Colombia para la detección de la incidencia de roya

en cultivos de café, siempre y cuando los datos de entrenamiento utilizados pertenezcan a las cercanías de la región cafetera de interés.

5.2 Resumen

En este capítulo se realizaron tres pruebas para evaluar las capacidades del método de conjunto propuesto en esta Maestría. La primera prueba consiste en utilizar un conjunto de datos de entrenamiento para cada región analizada, y sus respectivos conjuntos de datos de prueba, utilizando las variables de clima resultantes del proceso de selección de características. Por otra parte, fue llevada a cabo una segunda prueba que adiciona la precipitación en las variables de clima. Finalmente, se realiza una tercera prueba de evaluación donde se utilizó el conjunto de datos de la región “Naranjal” para entrenar el método de conjunto, mientras que los conjuntos de datos de las regiones “Jazmín” y “Santagueda” fueron usados como conjuntos de prueba. Las métricas utilizadas en las distintas evaluaciones son: error absoluto medio (MAE), error medio cuadrático (MSE) y el coeficiente de correlación de Pearson (CCP).

Los resultados de la primera prueba muestran que el método de conjunto tiene gran facilidad para detectar el comportamiento de la IR en los periodos de cultivo correspondientes al segundo semestre del año (formación de hojas en el segundo semestre del año y cosecha), mientras que para los periodos del primer semestre del año existe un poco más de incertidumbre en la detección. Po otro lado, la segunda prueba muestra que la precipitación es una variable fundamental que debe incluirse en el estudio del progreso de la enfermedad, sin embargo, es necesario realizar un estudio de mayor profundidad sobre dicha variable con el objetivo de encontrar la representación más adecuada del atributo en los datos, que permita a los algoritmos de aprendizaje automático relacionar de mejor manera el impacto de la precipitación en el desarrollo de la enfermedad. Finalmente, la tercera prueba muestra que utilizando el conjunto de datos de una región para detectar las dos restantes, las capacidades del método de conjunto para una adecuada detección de la IR se disminuyen, debido a que el comportamiento de la incidencia de roya es distinto para cada región.

Capítulo 6: prototipo

En este capítulo se presenta un prototipo que implementa el método de conjunto para la detección de la incidencia de roya en cultivos de café colombianos. Para la construcción del prototipo se utilizó como base la metodología ágil de desarrollo SCRUM [19][20].

6.1 Metodología SCRUM

6.1.1 Pila de producto

Siguiendo los pasos de la metodología, en primer lugar, se construye la pila de producto con la cual se trabajará en las distintas iteraciones del proyecto. A continuación, se define la pila de producto en la tabla 20.

Id	Historia de usuario	Importancia	Estimación	Como probarlo	Notas
1	Organizar datos de entrada	300	4	1. El usuario ingresa su dataset en formato csv (test)	
				2. El sistema despliega por consola la dimensión del dataset modificado	Se requiere caso de uso del sistema
				3. El número de columnas del nuevo dataset corresponde al número de atributos multiplicado por 4	

2	Elegir región cafetera	80	2	1. El usuario elige la región cafetera con la cual desea trabajar de una lista desplegable	
				2. El sistema le permite al usuario continuar con el siguiente paso dentro de la interfaz de usuario	Se requiere caso de uso del sistema
3	Ver porcentaje de incidencia de roya	150	10	1. El usuario ingresa su dataset en formato csv (test)	
				2. El usuario ingresa la región con la cual quiere trabajar	Se requiere caso de uso del sistema
				3. El sistema despliega en un mensaje el porcentaje de incidencia de roya en un valor de 0% a 100%	
4	Pronóstico de incidencia de roya	100	8	1. El usuario ingresa su dataset en formato csv (test)	
				2. El usuario ingresa la región con la cual quiere trabajar	
				3. El sistema realiza los pronósticos de los siguientes 5 días de cada variable de clima	Se requiere diagrama de secuencia del sistema

				4. El sistema hace la detección de la incidencia de roya para el día quinto y lo despliega en un mensaje al usuario	

Tabla 20. Pila de producto para el prototipo del sistema para la detección de la incidencia de roya en Colombia

Una vez identificadas las actividades necesarias para el buen funcionamiento de la primera versión del sistema, el siguiente paso es establecer las iteraciones o sprints, con el objetivo dividir el prototipo en una serie de entregables funcionales.

6.1.2 Iteraciones (Sprints)

Para la elección de las iteraciones de trabajo, se tienen en cuenta únicamente las actividades que se encuentran en la pila de producto, se analizan las importancias de cada actividad y la estimación inicial para realizar cada una de ellas. Teniendo en cuenta esto, la tabla 21 presenta las 3 iteraciones construidas para el desarrollo del prototipo.

SPRINT	Meta del Sprint	Historias de usuario
1	Preparar los datos del usuario para el método de conjunto	1. Organizar datos de entrada 2. Elegir región cafetera
2	Detección de la IR	3. Ver porcentaje de incidencia de roya
3	Pronóstico de la IR	4. Pronóstico de la incidencia de roya

Tabla 21. Sprints del prototipo de acuerdo a las actividades de la pila de producto

6.1.2.1 SPRINT 1: “Preparar los datos del usuario para el método de conjunto”

Para la determinación de las tareas puntuales que cada historia de usuario contiene en este sprint, se realizó el diagrama de casos de uso correspondiente (Figura 22).

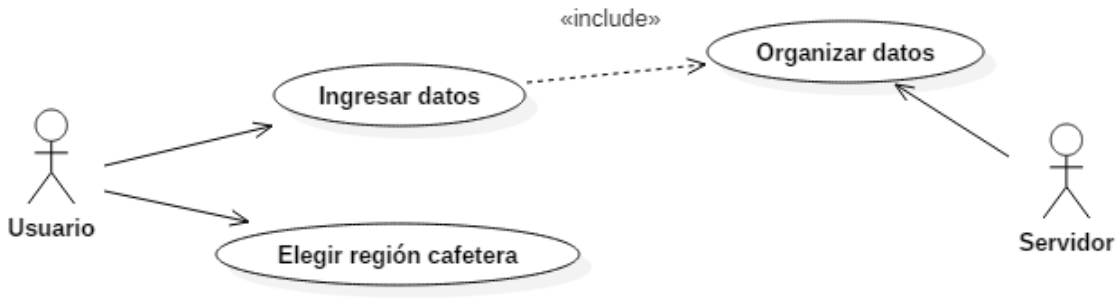


Figura 22. Caso de uso de las historias de usuario 1 y 2

Considerando el diagrama de la Figura 22, se muestran a continuación las tareas de las historias de usuario y los productos resultantes de cada tarea (Tabla 22 y 23).

Historia de Usuario - id:1			
Tareas de la historia	**Tipo de producto**	**Producto**	**Notas**
1. Escribir test	Python test	Test.read_datasets()	Este método contiene todas las condiciones a cumplir de la historia 1 en la sección “Como probarlo”
2. Leer archivos csv	Python function	Dataset_Tools. getDataset()	Se detalla en el diagrama de secuencia
3. Elegir de cada variable 4 muestras que representen el comportamiento de 15 días de la variable	Python function	Dataset.Tools. configurateDataset()	Se detalla en el diagrama de secuencia
4. Diseñar GUI	Python GUI	GUI.initInterface	Figura 24
5. Crear módulo final de entrega	Python script	Dataset_tools.py	

Tabla 22. Tareas y productos para la historia de usuario: “Organizar datos de entrada”

Historia de Usuario - id:2			
Tareas de la historia	**Tipo de producto**	**Producto**	**Notas**
1. Construir GUI para la lista desplegable de las regiones cafeteras y botón para ir al siguiente paso	Python GUI	GUI.initInterface	Figura 24
2. Construir método de selección de región cafetera	Python function	Dataset_Tools. loadModel()	Se detalla en el diagrama de secuencia

Tabla 23. Tareas y productos para la historia de usuario: “Elegir región cafetera”

A continuación, el siguiente diagrama de secuencia de la Figura 23 describe el comportamiento del sistema cuando el usuario ingresa un conjunto de datos para ser analizado mediante el método de conjunto. Allí se pueden apreciar cómo se relacionan las distintas funciones creadas en este sprint, así como sus entradas y salidas. Por su parte la Figura 24 muestra la interfaz de usuario creada en esta iteración.

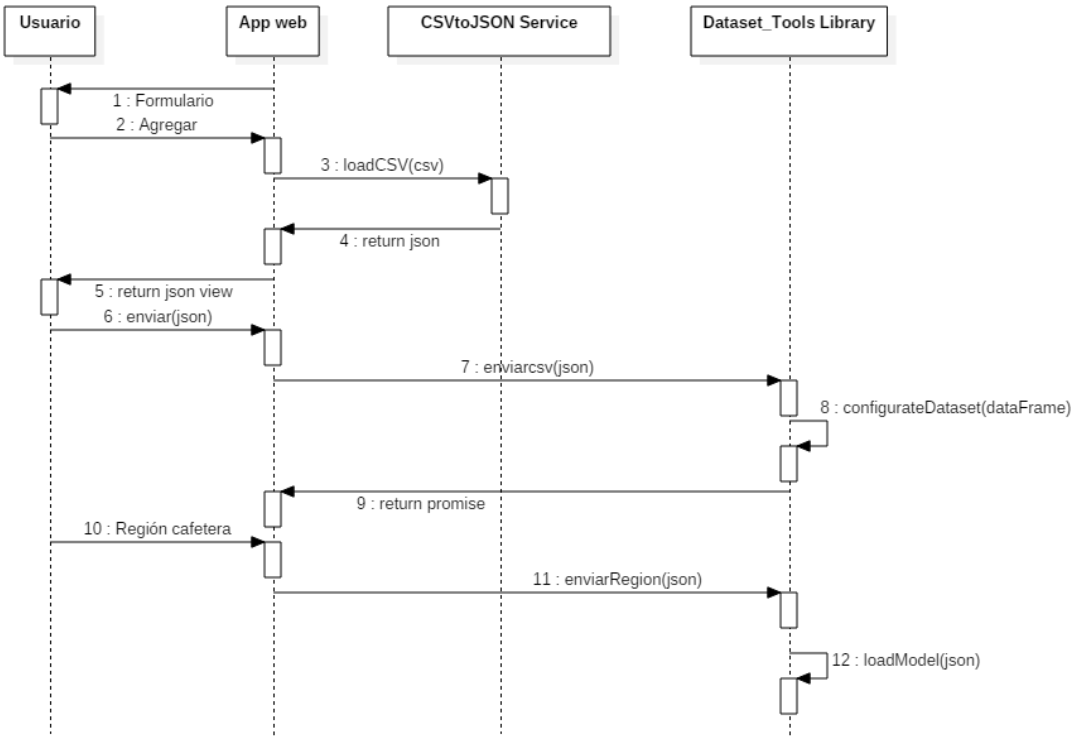
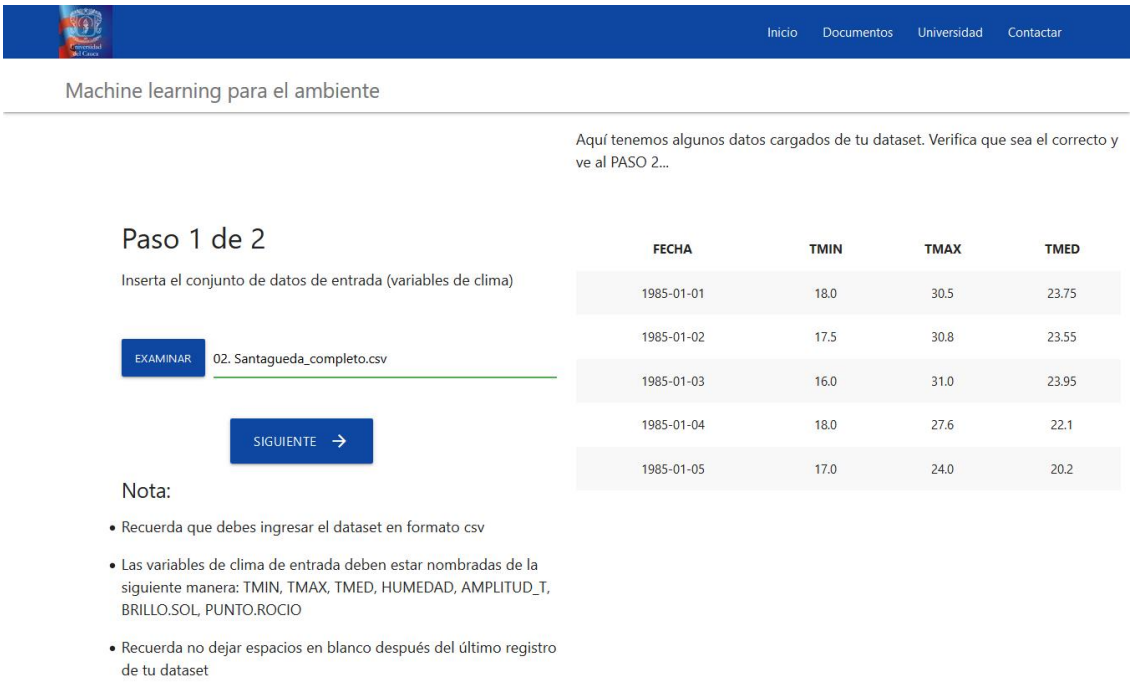
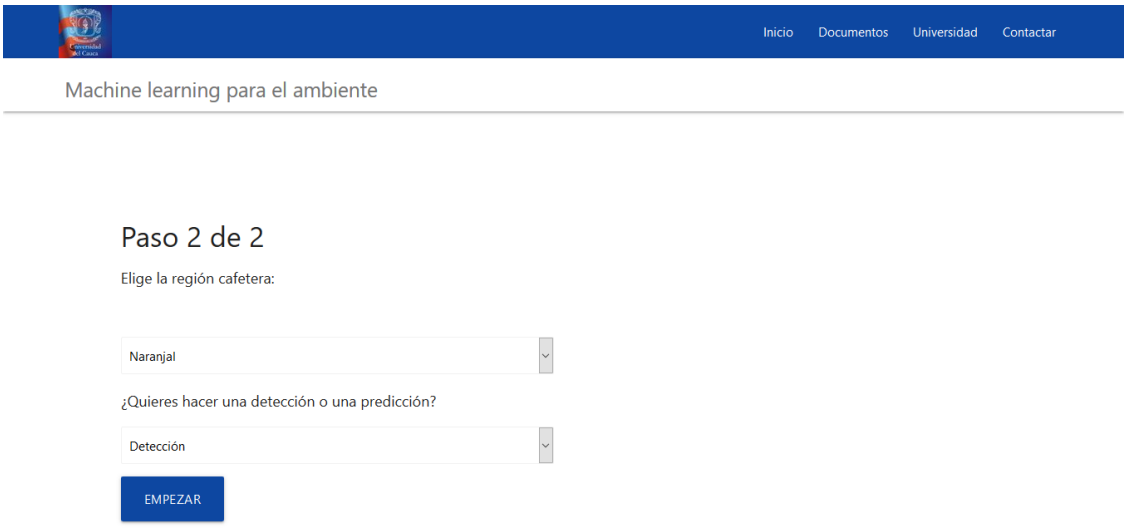


Figura 23. Diagrama de secuencia del sistema para el SPRINT 1 del prototipo



(a)



(b)

Figura 24. Interfaz de usuario para el ingreso del conjunto de datos y la elección de la región con la cual se quieren analizar los datos

Con el cumplimiento de las tareas de este sprint, se tiene un primer sistema capaz de preparar los datos del usuario al formato requerido por el método de conjunto para la correcta aplicación de sus algoritmos internos.

6.1.2.2 SPRINT 2: “Detección de la incidencia de roya”

De la misma forma que en la primera iteración, partimos del diagrama de casos de uso del sistema para especificar las tareas correspondientes a cada historia de usuario relacionada a este sprint. La Figura 25 muestra el diagrama de casos de uso para esta iteración.

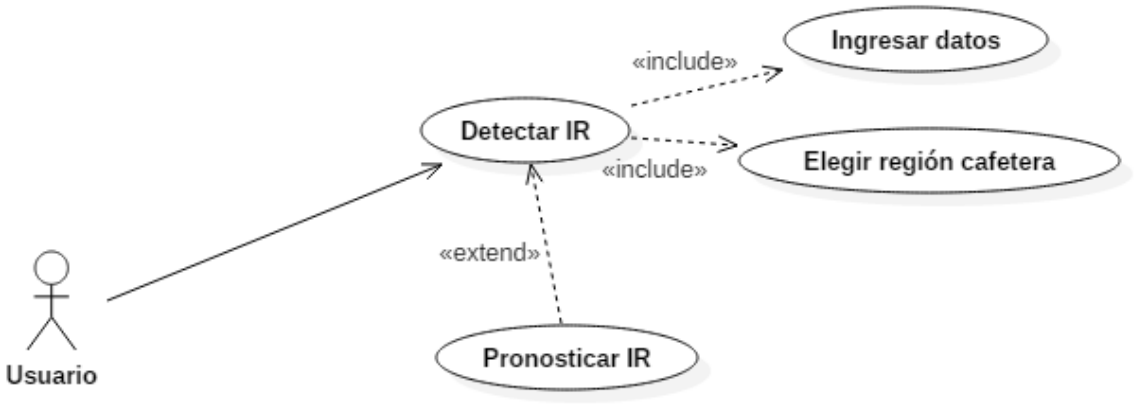


Figura 25. Caso de uso de la historia de usuario “Ver porcentaje de incidencia de roya”

Las tareas específicas y los productos resultantes de ellos se describen en la tabla 24.

Historia de Usuario - id:3			
Tareas de la historia	**Tipo de producto**	**Producto**	**Notas**
1. Importar módulo de configuración de datos de entrada	Integración de módulos	Dataset_configuration library	Esta librería permite utilizar todas las funciones del script Dataset_tools.py
2. Importar región cafetera que el usuario desea utilizar	Modelo de machine learning	Modelos entrenados para las regiones “Jazmín”, “Naranjal” y “Santagueda”	

3. Construir método de conjunto	Python function	EnsembleMethods. predictEnsembaleMethod()	
4. Construir GUI de visualización de incidencia de roya	Python GUI	GUI.prediction	
5. Entregar valor de incidencia de roya a la GUI	Test	GUI.message	Esta tarea comprueba que todos los requerimientos para dar por finalizada la historia se cumplen

Tabla 24. Tareas y productos para la historia de usuario: “Ver porcentaje de incidencia de roya”

6.1.2.3 SPRINT 3: “Pronóstico de la incidencia de roya”

Considerando que el pronóstico es un caso de uso extendido de la detección de la IR, a continuación, se muestran las tareas específicas y los productos resultantes de ellos en la tabla 25.

Historia de Usuario - id:3			
Tareas de la historia	**Tipo de producto**	**Producto**	**Notas**
1. Leer datos csv	Python function	Dataset_Tools. getDataset()	
2. Construir módulo de pronóstico de series de tiempo de acuerdo a la región	Python function	Time_Series_Tools. loadModels()	
3. Entregar pronósticos como datos de entrada al método de conjunto	Integración	TimeSeries_configuration library EnsembleMethods. TimeSeriesPrediction()	
4. Configurar datos de entrada al		Dataset.Tools. configurateDataset()	

formato requerido por el método de conjunto			
5. Realizar pronóstico de la incidencia de roya	Test	GUI.message	Esta tarea comprueba que todos los requerimientos para dar por finalizada la historia se cumplen

Tabla 25. Tareas y productos para la historia de usuario: “Pronóstico de incidencia de roya”

6.2 Resumen

En este capítulo se expusieron los elementos requeridos por la metodología SCRUM para el desarrollo del prototipo que implementa el método de conjunto de clasificadores construido en este proyecto de Maestría. En primer lugar, se muestra la pila de producto que describe las historias de usuario que fueron construidos para la primera versión del prototipo, todo esto apoyado de diagramas de casos de uso y diagramas de secuencias. Finalmente se muestran las interfaces principales del prototipo para este proyecto de Maestría.

Capítulo 7: conclusiones y trabajos futuros

7.1 Conclusiones

El establecimiento de un método de conjunto para la detección de roya en cultivos de café colombianos, se obtienen las siguientes conclusiones:

1. El mecanismo propuesto para la generación de muestras sintéticas de roya en Colombia mejora considerablemente el número de valores atípicos generados en los datos sintéticos. Esto se debe a la inclusión de conocimiento experto en el proceso de construcción de datos, lo cual permite representar de mejor manera el comportamiento de la curva de progreso de la roya en las regiones cafeteras de Colombia

2. Los experimentos de selección de características demostraron que las variables de clima afectan de manera distinta el progreso de la roya dependiendo de la etapa del ciclo de vida en la que se encuentre el cultivo de café. Esta propuesta realiza un proceso de selección de características para cada región cafetera analizada, definiendo las variables de clima más adecuadas para el estudio de la roya en dichas zonas. Sin embargo, es necesario realizar un estudio de mayor profundidad que considere además de los resultados del algoritmo RFE, conocimiento experto que permita identificar la manera más adecuada de procesar ciertas variables de clima como la precipitación en los cultivos, para relacionar apropiadamente dichas variables con el comportamiento de la enfermedad, y de esta manera poder usarlas en modelos de detección de roya en el café

3. El método de conjunto propuesto para la detección de roya en el café permite simular escenarios para estudiar el comportamiento de la roya ante la combinación de ciertas condiciones ambientales. Los experimentos demostraron que el método de conjunto propuesto tiene una buena fiabilidad en la detección de roya durante los periodos de formación de hojas del segundo semestre del año y la cosecha, mientras que, en los periodos de floración y

formación de hojas del primer semestre del año, los resultados muestran la existencia de un cierto nivel de incertidumbre sobre la respuesta entregada.

4. Para que el método de conjunto propuesto pueda ser implementado como un modelo predictivo, es necesario aumentar el conocimiento del sistema con información de otros factores que son igualmente importantes en el estudio de la roya en cultivos de café, tales como, calendarios de aspersiones de funguicidas, propiedades de la planta (especie, edad, etc), información sobre la distribución de la cosecha principal y el periodo principal de floración, entre otras que permiten ampliar la comprensión del impacto que tiene cada uno de estos factores en el desarrollo de la enfermedad

7.2 Trabajos futuros

1. Explorar sistemas mixtos que permitan incluir adecuadamente información sobre las propiedades físicas del cultivo y sobre controles de fertilización de los cultivos. Este conocimiento puede contribuir al ajuste de los porcentajes de incidencia de roya entregados por el método de conjunto

2. Realizar un análisis de mayor profundidad de series de tiempo en las variables de clima, con el objetivo de ampliar las capacidades de predicción de los distintos modelos de machine learning para la detección de roya, brindando la oportunidad a los modelos de ser utilizados fuera de entornos de simulación

3. Utilizar enfoques de visión por computadora como una alternativa para la recolección de muestras de incidencia de roya, con el objetivo de aumentar el número de datos registrados sobre la enfermedad

REFERENCIAS

[1] "Evolución anual de la producción mundial de café."

[2] J. Avelino, M. Cristancho, S. Georgiou, P. Imbach, L. Aguilar, G. Bornemann, P. Läderach, F. Anzueto, A. J. Hruska, and C. Morales, "The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions," *Food Secur.*, vol. 7, no. 2, pp. 303–321, Apr. 2015.

[3] C. Rivillas Osorio, C. Serna Giraldo, M. Cristancho Ardila, and Á. Gaitán Bustamante, *La Roya del cafeto en Colombia*. 2009.

[4] D. C. Corrales, A. Ledezma, A. J. P. Q., J. Hoyos, A. Figueroa, and J. C. Corrales, "A new dataset for coffee rust detection in Colombian crops base on classifiers," *Sist. Telemática*, vol. 12, no. 29, pp. 9–23, 2014.

[5] C. A. A. DI GIROLAMO, NETO RODRIGUES, L. H. A.; MEIRA, "Modelos de predição da ferrugem do cafeeiro (Hemileia vastatrix Berkeley & Broome) por técnicas de mineração de dados. - Portal Embrapa." [Online]. Available: https://www.embrapa.br/informatica-agropecuaria/busca-de-publicacoes/-/publicacao/991078/modelos-de-predicao-da-ferrugem-do-cafeeiro-hemileia-vastatrix-berkeley--broome-por-tecnicas-de-mineracao-de-dados.

[6] D. C. Corrales, A. Figueroa, A. Ledezma, and J. C. Corrales, "An Empirical Multi-classifier for Coffee Rust Detection in Colombian Crops," Springer International Publishing, 2015, pp. 60–74.

[7] O. Arbelaitz and B. Sierra Araujo, *Aprendizaje automático : conceptos básicos y avanzados : aspectos prácticos utilizando el software WEKA*. Pearson, 2006.

[8] T. G. Dietterich, "Ensemble Methods in Machine Learning."

[9] Z.-H. (Computer scientist) Zhou, *Ensemble methods : foundations and algorithms*. Taylor & Francis, 2012.

[10] D. C. Corrales, G. German, J. P. Rodriguez, L. Agapito, and J. C. Corrales, "Lack of Data: Is It Enough Estimating the Coffee Rust with Meteorological Time Series?," *Comput. Sci. Its Appl. - ICCSA 2017*, vol. 10405, pp. 3–16, 2017.

[11] J. A. Pérez Toro, *Economía cafetera y desarrollo económico en Colombia*. Bogotá, 2013.

[12] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. R. Daimlerchrysler, C. Shearer, and R. W. Daimlerchrysler, "Step-by-step data mining guide," *SPSS inc*, vol. 78, pp. 1–78, 2000.

[13] C. Rivillas Osorio, C. Serna Giraldo, M. Cristancho Ardila, and Á. Gaitán Bustamante, *La roya del cafeto en Colombia*. Cenicafé, 2011.

[14] S. Sierra, C. Rivillas Osorio, G. Gómez, and C. Leguizamón, "Recomendaciones para el control químico de la roya del cafeto para 1993 (zonas con cosecha principal en el primer semestre del año). Avances técnicos Cenicafé," 1993.

[15] S. Sierra, O. Rivillas, G. Gómez, and C. Leguizamón, "Recomendaciones para el control químico de la roya del cafeto para 1993 (zonas con cosecha principal en el segundo semestre del año). Avances técnicos Cenicafé," 1993.

[16] G. A. Alvarado, *El café y la roya. Caldas*. Cenicafé, 2011.

[17] M. Abreu and L. Villas, *Minería de datos para Series Temporales*, no. August. 2015.

[18] S. S. Francia, *Multiclasificadores : Métodos y Arquitecturas*. 2006.
[19] McGraw-Hil, *Diccionario de términos científicos y técnicos*. 2009.
[20] I. D. López, F. Campo, S. A. Ordoñez, J. C. Corrales, A. Figueroa, C. León, and D. C. Corrales, "Plataforma para el seguimiento de variables meteorológicas y ambientales para el sector agropecuario," *VII Congr. ibérico agroIngeniería y ciencias hortícolas*, p. ISBN – 10: 84-695-8844-3, 2013.
[21] D. C. Corrales, "Toward detecting crop diseases and pest by supervised learning," *Ing. y Univ.*, vol. 19, no. 1, p. 207, Jul. 2015.
[22] D. C. Corrales, A. Ledezma, A. J. Peña Q., J. Hoyos, A. Figueroa, and J. C. Corrales, "A new dataset for coffee rust detection in Colombian crops base on classifiers," *Sist. y Telemática*, vol. 12, no. 29, p. 9, Jun. 2014.
[23] D. C. Corrales, A. J. Peña Q, C. León, A. Figueroa, and J. C. Corrales, "Early warning system for coffee rust disease based on error correcting output codes: a proposal," *Rev. Ing. Univ. Medellín*, vol. 13, no. 25, pp. 57–64, 2014.
[24] E. Lasso, T. T. Thamada, C. A. A. Meira, and J. C. Corrales, "Graph Patterns as RepLasso, E., Thamada, T. T., Meira, C. A. A., & Corrales, J. C. (2015). Graph Patterns as Representation of Rules Extracted from Decision Trees for Coffee Rust Detection. Communications in Computer and Information Science, 405–414.resen," *Commun. Comput. Inf. Sci.*, pp. 405–414, 2015.
[25] E. J. Girón Buitrón, D. C. Corrales, J. Avelino, J. A. Iglesias, and J. C. Corrales, "Rule-based expert system for detection of coffee rust warnings in colombian crops," *J. Intell. Fuzzy Syst.*, pp. 1–11, 2019.
[26] M. E. Cintra, C. A. A. Meira, M. C. Monard, H. A. Camargo, and L. H. A. Rodrigues, "The use of fuzzy decision trees for coffee rust warning in Brazilian crops," in *2011 11th International Conference on Intelligent Systems Design and Applications*, 2011, pp. 1347–1352.
[27] C. B. Pérez-Ariza, A. E. Nicholson, and M. J. Flores, "Prediction of Coffee Rust Disease Using Bayesian Networks," *Proc. 6th Eur. Work. Probabilistic Graph. Model. PGM*, 2012.
[28] O. Luaces, L. H. Rodrigues, C. A. Alves Meira, and A. Bahamonde, "Using nondeterministic learners to alert on coffee rust disease," *Expert Syst. Appl.*, vol. 38, no. 11, pp. 14276–14283, 2011.
[29] O. Luaces, L. H. A. Rodrigues, C. A. A. Meira, J. R. Quevedo, and A. Bahamonde, "Viability of an Alarm Predictor for Coffee Rust Disease Using Interval Regression.," pp. 337–346, 2010.
[30] I. Cano and V. Torra, "Generation of synthetic data by means of fuzzy c-Regression," in *2009 IEEE International Conference on Fuzzy Systems*, 2009, pp. 1145–1150.
[31] Y. Park, J. Ghosh, and M. Shankar, "Perturbed Gibbs Samplers for Generating Large-Scale Privacy-Safe Synthetic Health Data," in *2013 IEEE International Conference on Healthcare Informatics*, 2013, pp. 493–498.
[32] A. Eshmawi and S. Nair, "Semi-Synthetic Data for Enhanced SMS Spam Detection," in *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems - MEDES '14*, 2014, pp. 206–212.
[33] L.-S. Lin, D.-C. Li, W.-H. Yu, and Y.-M. Hsueh, "Generating Multi-modality Virtual Samples with Soft DBSCAN for Small Data Set Learning," in *2015 3rd*

International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence, 2015, pp. 363–368.

[34] Y.-S. Lin and T.-I. Tsai, “Using virtual data effects to stabilize pilot run neural network modeling,” in *Proceedings of 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS)*, 2013, pp. 463–468.

[35] C. Rusu, R. Morisi, D. Boschetto, R. Dharmakumar, and S. A. Tsaftaris, “Synthetic Generation of Myocardial Blood–Oxygen-Level-Dependent MRI Time Series Via Structural Sparse Decomposition Modeling,” *IEEE Trans. Med. Imaging*, vol. 33, no. 7, pp. 1422–1433, Jul. 2014.

[36] A. Mohammadian, H. Aghaeinia, and F. Towhidkhah, “Incorporating prior knowledge from the new person into recognition of facial expression,” *Signal, Image Video Process.*, 2014.

[37] U. Morbiducci, R. Ponzini, G. Rizzo, M. E. Biancolini, F. Iannaccone, D. Gallo, and A. Redaelli, “Synthetic dataset generation for the analysis and the evaluation of image-based hemodynamics of the human aorta,” *Med. Biol. Eng. Comput.*, vol. 50, no. 2, pp. 145–154, Feb. 2012.

[38] J. A. Malpica, “Splines Interpolation in High Resolution Satellite Imagery,” Springer Berlin Heidelberg, 2005, pp. 562–570.

[39] L. Tu, M. Styner, J. Vicory, S. Elhabian, R. Wang, J. Hong, B. Paniagua, J. C. Prieto, D. Yang, R. Whitaker, and S. M. Pizer, “Skeletal Shape Correspondence Through Entropy,” *IEEE Trans. Med. Imaging*, vol. 37, no. 1, pp. 1–11, Jan. 2018.

[40] J. WANG, Y. DANG, N. XU, and S. DING, “Grey interpolation approach for small time-lag samples based on grey dynamic relation analysis,” *J. Syst. Eng. Electron.*, vol. 29, no. 1, pp. 105–115, 2018.

[41] C. Cheng, W. Xu, and J. Wang, “A Comparison of Ensemble Methods in Financial Market Prediction,” in *2012 Fifth International Joint Conference on Computational Sciences and Optimization*, 2012, pp. 755–759.

[42] D. K. Barrow, S. F. Crone, and N. Kourentzes, “An evaluation of neural network ensembles and model selection for time series prediction,” in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8.

[43] Yu-Feng Deng, Xing Jin, and Yi-Xin Zhong, “Ensemble SVR for prediction of time series,” in *2005 International Conference on Machine Learning and Cybernetics*, 2005, p. 3528–3534 Vol. 6.

[44] R. Soelaiman, A. Martoyo, Y. Purwananto, and M. H. Purnomo, “Implementation of recurrent neural network and boosting method for time-series forecasting,” in *International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering 2009*, 2009, pp. 1–8.

[45] J. D. Wichard and M. Ogorzalek, “Time series prediction with ensemble models,” in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, vol. 2, pp. 1625–1630.

[46] 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, Nov. 2015.

[47] R. H. Bartels, J. C. Beatty, and B. A. Barsky, *An Introduction to Splines for Use in Computer Graphics and Geometric Modelling*. San Francisco: Morgan

Kaufmann, 1998.
[48] “Algebra_Angel_Cap3.Pdf,” in *cimat*, 2016, pp. 148–150.
[49] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006.
[50] K. Yan and D. Zhang, “Feature selection and analysis on correlated gas sensor data with recursive feature elimination,” *Sensors Actuators B Chem.*, vol. 212, pp. 353–363, Jun. 2015.
[51] “rfe function | R Documentation,” 2017. [Online]. Available: https://www.rdocumentation.org/packages/caret/versions/6.0-79/topics/rfe. [Accessed: 05-Apr-2018].
[52] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.
[53] J. P. Rodríguez Muñoz, “SERIES DE TIEMPO PARA EL PRONÓSTICO DE ROYA EN CULTIVOS DE CAFÉ COLOMBIANOS A PARTIR DE DATOS SINTÉTICOS,” Universidad del Cauca, 2017.
[54] P. S. Kalekar and P. Bernard, “Time series Forecasting using Holt-Winters Exponential Smoothing Under the guidance of,” no. 4329008, pp. 1–13, 2004.
[55] Z. Deljac, M. Kunstic, and B. Spahija, “A comparison of traditional forecasting methods for short-term and long-term prediction of faults in the broadband networks,” *2011 Proc. 34th Int. Conv. MIPRO*, no. 1, 2011.
[56] A. Tealab, H. Hefny, and A. Badr, “Forecasting of nonlinear time series using ANN,” *Futur. Comput. Informatics J.*, 2018.
[57] M. Abromowitz and I. A. Stegun, “Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical tables, 9th printing,” 1972.
[58] C. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005.
[59] H. E. SOPER, A. W. YOUNG, B. M. CAVE, A. LEE, and K. PEARSON, “ON THE DISTRIBUTION OF THE CORRELATION COEFFICIENT IN SMALL SAMPLES. APPENDIX II TO THE PAPERS OF "STUDENT" AND R. A. FISHER. A COOPERATIVE STUDY,” *Biometrika*, vol. 11, no. 4, pp. 328–413, May 1917.
[60] J. Demsar, “Statistical Comparisions of Classifiers over Multiple Data Sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[61] C. Strobl, J. Malley, and G. Tutz, “An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests.,” *Psychol. Methods*, vol. 14, no. 4, pp. 323–348, 2009.
[62] M. R. Segal, “Machine Learning Benchmarks and Random Forest Regression,” 2003.
[63] R. J. Urbanowicz and J. H. Moore, “ExSTraCS 2.0: description and evaluation of a scalable learning classifier system,” *Evol. Intell.*, vol. 8, no. 2–3, pp. 89–116, Sep. 2015.
[64] K. Schwaber and J. Sutherland, “The Scrum Guide,” *Scrum.Org and ScrumInc*, no. July, p. 17, 2013.

[65] H. Kniberg, P. De, J. Sutherland, and M. Cohn, *Una historia de guerra Ágil SCRUM Y XP DESDE LAS TRINCHERAS*. 2007.