CÁLCULO DE LA DISPONIBILIDAD DE NITRÓGENO EN SUELOS ESTABLECIDOS CON CULTIVO DE CAFÉ UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Universidad del Cauca ®

Proyecto de Trabajo de Grado

Julian Egas Daza

Andrés Felipe Bravo Portilla

Director: Msc. Juan Fernando Casanova Olaya

Codirector: PhD. Juan Carlos Corrales Muñoz

Asesor: PhD. Cristhian Nicolás Figueroa Martínez

Asesora: PhD. María Cristina Ordoñez Díaz

Departamento de Telemática

Facultad de Ingeniería Electrónica y Telecomunicaciones

Universidad del Cauca

Popayán, Cauca, julio de 2022

CÁLCULO DE LA DISPONIBILIDAD DE NITRÓGENO EN SUELOS ESTABLECIDOS CON CULTIVO DE CAFÉ UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Julian Egas Daza

Andrés Felipe Bravo Portilla

Trabajo de grado presentado a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para obtener el título de: Ingeniero en Electrónica y Telecomunicaciones

Director: Msc. Juan Fernando Casanova Olaya
Codirector: PhD. Juan Carlos Corrales Muñoz
Asesor: PhD. Cristhian Nicolás Figueroa Martínez
Asesora: PhD. María Cristina Ordoñez Díaz

Departamento de Telemática

Facultad de Ingeniería Electrónica y Telecomunicaciones

Universidad del Cauca

Popayán, Cauca, julio de 2022

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Ecuaciones

Resumen

La agricultura cumple un rol fundamental en América Latina y especialmente en Colombia genera un impacto significativo en el empleo directo e indirecto generado a raíz de la producción y comercialización de productos agrícolas, mejorando la calidad de vida de millones de personas.

En Colombia, el cultivo de café ha cobrado gran importancia, siendo uno de los cultivos que mayor peso tiene dentro de la canasta de productos agrícolas, responsable de generar 2,5 millones de empleos directos e indirectos y de cuya actividad dependen el 25% de la población rural colombiana. , a pesar de la su importancia, el cultivo de café ha venido presentando una disminución de la producción en los últimos años, en el 2021 disminuyó un 13% comparado con el año anterior, posiblemente por consecuencia de factores como la variabilidad y el cambio climático, incremento del costo de los insumos y el uso de prácticas inadecuadas en el manejo nutricional del cultivo, lo que genera un incremento de costos de fertilizantes, que se traduce en bajas rentabilidades para el productor, afectando su calidad de vida. En este sentido, es necesario el desarrollo de nuevas herramientas tecnológicas, que facilite el uso de prácticas adecuadas de fertilización, permitiendo la reducción de costos de fertilizantes y así mejorar la rentabilidad de los productores.

En ese orden de ideas, la presente investigación se centra en estimar la disponibilidad de N en suelos establecidos con cultivo de café a partir de un enfoque híbrido que plantea la integración de modelos basados en procesos y datos. Dicha estimación puede apoyar al caficultor a ejecutar una fertilización eficiente de N, reduciendo costos en su aplicación y a su vez manteniendo los niveles óptimos para el cultivo.

Abstract

Agriculture plays a fundamental role in Latin America and especially in Colombia, generating a significant impact on direct and indirect employment generated as a result of the production and marketing of agricultural products, improving the quality of life of millions of people.

In Colombia, the cultivation of coffee has gained great importance, being one of the crops that has the greatest weight within the basket of agricultural products, responsible for generating 2.5 million direct and indirect jobs and on whose activity 25% of the population depend. The entire Colombian rural population. However, despite its importance, coffee cultivation has been showing a decrease in production in recent years, in 2021 decreased by 13% compared to the previous year, possibly due to factors such as variability and climate change, increased cost of inputs and the use of inadequate practices in the nutritional management of the crop, which generates an increase in fertilizer costs, which translates into low returns for the producer, affecting their quality of life. In this sense, it is necessary to develop new technological tools that facilitate the use of adequate fertilization practices, facilitating the reduction of fertilizer costs and, in this way, improve the profitability of producers.

In that order of ideas, the present investigation focuses on estimating the availability of N in soils established with coffee cultivation from a hybrid approach that proposes the integration of models based on processes and data. Said estimation could support the coffee grower to execute an efficient N fertilization, reducing costs in its application and in turn maintaining the optimal levels for the crop.

1. Introducción

1.1. Planteamiento del problema

Los principales objetivos para el mejoramiento del sector agrícola en Colombia son aumentar la producción del cultivo, mejorar su calidad, reducir los costos operativos y la contaminación ambiental [1]. La potencial producción de un cultivo depende de atributos relacionados con los componentes de clima, suelo, planta, manejo del cultivo y su interrelación. Una de las claves para poder conseguir un incremento en la productividad de un cultivo es la determinación del estado nutricional del mismo, dado que el exceso o la deficiencia de nutrientes pueden causar daños severos que ocasionan la disminución del rendimiento en el cultivo [2].

El diagnóstico confiable del estado nutricional es determinante para aumentar la producción y parte esencial del manejo de un cultivo, ya que sirve como base para el uso racional de los suplementos nutricionales, evitando el desperdicio de recursos financieros y reduciendo el impacto ambiental [2]. En cuanto al café, el nutriente más crítico para su productividad es el nitrógeno (N), dado que determina el establecimiento y desarrollo de la planta. Además, su aplicación adecuada mejora la productividad en el año de suministro y aporta al crecimiento vegetativo de la planta para garantizar producciones futuras [3]. Por otro lado, el N excesivo o mal programado puede reducir el rendimiento del cafetal, aumentar el riesgo de enfermedades y plagas en el cultivo, entre otros efectos adversos [4] [5].

En este sentido, para optimizar la producción es necesario hacer coincidir el suministro de N con los requisitos del cultivo, mediante la determinación precisa y oportuna de la disponibilidad de N en el cultivo [5]. A pesar de que es conocida la importancia del N en los cultivos, muchos caficultores no pueden acceder a una medición o estimación adecuada, ya sea por desconocimiento de herramientas que faciliten dicha estimación, falta de apoyo económico u otros factores [6]. Algunos productores acuden a aplicaciones basadas en gestión de calendario o al muestreo ocasional de suelo. En casos extremos, toman decisiones reactivas ante los síntomas de deficiencia de N, como el color amarillento de las hojas. Estos métodos son comúnmente adaptados una vez que el daño fisiológico ya se ha infligido en el cultivo. A largo plazo, los métodos basados en muestreo no son ideales para equilibrar las necesidades de nutrientes en el cultivo, la absorción de nutrientes provenientes de fuentes naturales y el destino a corto y largo plazo del fertilizante aplicado, lo que resulta en una producción insostenible y un déficit de rendimiento [5] [3].

Por otro lado, existen métodos científicos para la determinación del N en cultivos agrícolas. Según [7] la estimación del N en la planta se puede realizar con dos métodos: destructivo y no destructivo. El método destructivo consiste en prácticas que resultan dañinas para el cultivo, el más común es el análisis químico de laboratorio que requiere muestras de la planta, que son llevadas al laboratorio, donde se realiza el proceso de secado y posterior determinación de N mediante análisis químico [5]. Un método no destructivo por el contrario consiste en

estimaciones a partir de datos y/o imágenes, entre los más populares está la teledetección óptica del estado de N en la planta por medio de índices espectrales, tema que ha sido ampliamente estudiado durante la última década [2] y la estimación de N por modelos de predicción creados con algoritmos de aprendizaje automático, los cuales utilizan datos relacionados al cultivo (pH, humedad, salinidad, entre otros) y datos meteorológicos como entradas para realizar la predicción del N [8].

El uso de técnicas de aprendizaje automático para generar modelos predictivos requiere de datos históricos del cultivo, sin embargo, Colombia no cuenta con conjuntos de datos que cumplan dichas necesidades, la información es escasa y en muchos casos incompleta [9], por lo tanto, resulta insuficiente. De esta manera, se ha propuesto la integración de modelos basados en datos y modelos basados en procesos, como los modelos agronómicos, que facilitan la generación de datos considerando métodos numéricos y analíticos, y permiten simular los procesos del sistema agrícola con un alto nivel de confiabilidad [10], [11].

En consecuencia del grave impacto fisiológico que implica una deficiencia de N en los suelos establecidos con cultivo de café [3], para los caficultores colombianos es clave hacer coincidir el suministro de N con los requisitos del cultivo con el fin de mitigar este problema [5]. Sin embargo, aun cuándo es conocida la importancia del N, ha sido un desafío para los caficultores colombianos aplicar la cantidad adecuada, esto debido a la desinformación sobre las condiciones iniciales en las que se encuentra el suelo al iniciar el proceso de siembra, lo anterior puede conllevar a pérdidas económicas originadas por el mal uso del fertilizante, o disminución en el rendimiento del cultivo. En consecuencia, la presente investigación estará enfocada en el cálculo de la disponibilidad de N en suelos establecidos con cultivo de café por medio de técnicas de aprendizaje automático y modelos agronómicos, con el fin de apoyar al caficultor en el manejo del cultivo, ahorrando recursos agrícolas y económicos.

De acuerdo con lo expuesto anteriormente, en este estudio se busca dar respuesta a la siguiente pregunta de investigación:

¿Es posible estimar la disponibilidad de nitrógeno en suelos establecidos con cultivo de café integrando técnicas de aprendizaje automático y modelos agronómicos?

1.2. Escenario de motivación

En Colombia, el café sigue siendo uno de los cultivos que mayor peso tiene dentro de la canasta de productos agrícolas, a pesar de la disminución circunstancial de la producción en los últimos años, explicada por la caída en la producción del grano por diversos factores, cómo puede ser, problemas asociados al manejo del cultivo, el aumento en los costos de los insumos necesarios, entre otros. El café, más que un bien agrícola, es el pilar de un tejido social e institucional , responsable de generar 2,5 millones de empleos directos e indirectos y de cuya actividad dependen 545 mil familias colombianas, estas familias componen aproximadamente el 25%

de toda la población rural colombiana [12]. De la misma forma, el café aporta la tercera parte del empleo rural del país y es la actividad que más contribuye con la redistribución del ingreso en el campo. En 2020 el valor de la cosecha llegó a $9 billones, alcanzando el máximo de los últimos 20 años, un flujo importante de recursos que aportan a la economía regional. De ahí la importancia de esta actividad como eje para la disminución de la pobreza y potencial generador de condiciones de paz en Colombia [7].

En cuanto a las personas involucradas en la producción agrícola, los pequeños productores juegan un papel fundamental. En América Latina y el Caribe, el 81% de la actividad agrícola es llevada a cabo por pequeños agricultores, agrupando a una población de alrededor de 60 millones de personas, creando entre el 57% y el 77% del empleo agrícola en la región, a nivel país, fueron registradas un total de 540 mil familias caficultoras, sólo en el departamento del Cauca se encuentran ubicadas 90 mil, de las cuales en un 99% están compuestas por pequeños productores [9]. Los pequeños agricultores no sólo producen la mayor parte de los alimentos que consumen los países de la región, sino que también desarrollan actividades agrícolas diversificadas, con las que juegan un papel fundamental a la hora de aportar a la sostenibilidad del medio ambiente y la conservación de la biodiversidad [8]. Sin embargo, en la actualidad, los pequeños productores enfrentan diversos retos asociados a la sostenibilidad de su actividad y el diseño de estrategias que permitan mantener o aumentar su productividad.

El plan de fertilización es un factor determinante en la producción agrícola ya que puede determinar un incremento o caída en el rendimiento. Sin embargo, para que un plan de fertilización sea eficaz, se debe contar con un previo análisis de nutrientes en el suelo, esto con el fin de determinar el nivel de nutrientes con el que la planta cuenta inicialmente para evitar posibles desperdicios. Lo anterior conlleva a un manejo óptimo de los recursos ambientales y económicos, ya que el alza de los fertilizantes ha aumentado los costos de producción un 10% en los últimos años, lo que afecta principalmente al pequeño productor [13].

En este orden de ideas, en la presente investigación, se estudia la disponibilidad de nitrógeno en cultivos de café, y con ello generar una herramienta de apoyo a los caficultores de la región caucana, que les permita estimar la cantidad de nitrógeno en sus terrenos y aplicar buenas prácticas de manejo encaminadas al incremento de las tasas de rendimiento de sus cultivos, que podría conllevar a una mejora económica.

1.3. Objetivos

1.3.1. Objetivo general

Calcular la disponibilidad de nitrógeno en suelos establecidos con cultivo de café mediante la integración de técnicas de aprendizaje automático y modelos agronómicos

1.3.2. Objetivos específicos

- ❖ Conformar un conjunto de datos sobre disponibilidad de nitrógeno en el suelo haciendo uso de modelos agronómicos
- ❖ Desarrollar un modelo basado en técnicas de aprendizaje automático a partir de los datos generados para la estimación de la disponibilidad de nitrógeno
- ❖ Evaluar el modelo desarrollado para la estimación de la disponibilidad de nitrógeno en suelos establecidos con cultivo de café

1.4. Partes de la memoria

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

- ***Capítulo 1***: presenta la introducción, el planteamiento del problema de investigación y la estructura general del trabajo realizado.

- ***Capítulo 2***: denominado “Estado actual del conocimiento”, hace referencia a las tecnologías y conceptos en los que está fundamentada la presente investigación, además de las experiencias previas llevadas a cabo en otras investigaciones de aprendizaje automático y agricultura que se relacionan con la expuesta en el presente trabajo.

- ***Capítulo 3***: denominado “Calibración del modelo agronómico QUEFTS a la región de estudio”, presenta el proceso de construcción de un conjunto de datos del rendimiento del cultivo de café y disponibilidad de nitrógeno en Colombia mediante el modelo agronómico QUEFTS. El proceso descrito en el capítulo comprende la selección, adquisición, estructuración y calibración del modelo agronómico.

- ***Capítulo 4***: denominado “Modelo de datos para estimar la disponibilidad de nitrógeno en suelos establecidos con cultivo de café”, presenta el proceso de construcción de un conjunto de datos del rendimiento del cultivo de café en Colombia haciendo uso del modelo agronómico calibrado en el capítulo 3. Así como también expone el uso de dichos datos para llevar a cabo un proceso de entrenamiento utilizando distintas técnicas de aprendizaje automático para la generación de un modelo que permite la estimación de la

disponibilidad de Nitrógeno en suelos establecidos con café en Colombia.

- ***Capítulo 5***: contiene la síntesis de los resultados de la presente investigación, así como las principales contribuciones y elementos a tener en cuenta en el desarrollo de trabajos futuro

2. Estado actual del conocimiento

Este capítulo expone la generación de la base conceptual. recopila los conceptos y tecnologías fundamentales para la presente investigación. De igual manera, describe las investigaciones recientes y/o representativas que han sido desarrolladas en torno a la aplicación de algoritmos de aprendizaje automático en el campo de la agricultura.

El capítulo está dividido en los siguientes apartados:

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

2.1. Conceptos y definiciones fundamentales

2.1.1. Nitrógeno en la agricultura

El nitrógeno (N) es un nutriente vegetal clave en todo el crecimiento y desarrollo de la planta ya que afecta tanto el rendimiento como la calidad de la cosecha. Todas las plantas pueden absorber y utilizar formas inorgánicas y orgánicas de N del suelo. Estas formas de N son tomadas del suelo directamente por las raíces y se asimilan en las hojas para sintetizar proteínas [14].

La baja disponibilidad de N en el suelo es a menudo el principal factor nutritivo que limita el crecimiento y el rendimiento de las plantas de un cultivo, es por ello que la aplicación de fertilizantes inorgánicos de N se ha convertido en una estrategia importante y rentable utilizada en sistemas agrícolas de todo el mundo. Sin embargo, la forma en que el N es aprovechado por las plantas depende de variables ambientales fuera del suministro de N [15].

Para lograr y mantener un equilibrio de N dentro de la planta durante toda la temporada (desde la siembra hasta la cosecha) es indispensable realizar mediciones constantes al cultivo. La estimación del estado de N de la planta puede dividirse en dos tipos principales: destructivos y no destructivos, el método más común de medición destructiva es un análisis químico del suelo que está asociado con la técnica de Kjeldahl y es largo y costoso. Por otro lado están los no destructivos, estos métodos no infringen ningún daño a la

plantación, entre estos destacan la teledetección óptica del estado de N en la planta por medio de índices espectrales y la estimación de N por medio de modelos de predicción haciendo uso de algoritmos de inteligencia artificial, está demostrado que los métodos no destructivos se equiparan en precisión al método tradicional [16], con la ventaja de ser más económicos, ecológicos y manteniendo un monitoreo constante del N [3].

2.1.2. Aprendizaje automático

El aprendizaje automático o Machine learning (ML) es un segmento de la inteligencia artificial encargado de identificar patrones a partir de datos y usarlos para hacer pronósticos o tomar decisiones en nuevos registros. Se dice que un programa de computadora aprende de la experiencia (E) con respecto a alguna tarea (T) y alguna medida de rendimiento (P), si su rendimiento en T, medido por P, mejora con la experiencia E. En otras palabras, el aprendizaje automático está relacionado con la cuestión de cómo construir programas informáticos que mejoren automáticamente con la experiencia. El aprendizaje automático puede clasificarse en tres clases principales: supervisado, no supervisado y por refuerzo [17] [18].

El aprendizaje supervisado consiste en aprender de un conjunto de entrenamiento de datos etiquetados que son proporcionados por un supervisor externo conocedor. Cada ejemplo (instancia) es una descripción de una situación junto con una especificación, la etiqueta de la acción correcta que el sistema debe tomar para esa situación, que puede ser identificar una categoría a la que pertenece la situación (clasificación) ó hacer una predicción del valor de una clase (regresión). Por otro lado, el aprendizaje no supervisado generalmente trata de encontrar estructuras ocultas en colecciones de datos sin etiquetar. Por último en el aprendizaje por refuerzo, se busca determinar las acciones que debe escoger el sistema de aprendizaje en un entorno dado, con el fin de maximizar una señal de recompensa numérica, no se le dice al sistema qué acciones tomar, sino que debe descubrir qué acciones producen la mayor recompensa al probarlas [19].

2.1.2.1. Aprendizaje automático en la agricultura

Uno de los múltiples campos de enfoque del aprendizaje automático es la agricultura, ya que el avance en la ciencia y la tecnología ha llevado a generar una gran cantidad de datos agrícolas. Por lo tanto, surge una corriente de investigación de los datos disponibles y su integración con procesos tales

como la predicción del rendimiento [20], detección de enfermedades del cultivo [21], la identificación del estrés hídrico [22], mapeo de cultivos y maleza [23], predicción del estado nutricional del cultivo [3], entre otras. El aprendizaje automático posee la capacidad de procesar una gran cantidad de entradas y manejar tareas no lineales [16], por ende, su uso para optimizar múltiples tareas agronómicas va en aumento.

2.1.3. Modelos agronómicos

Un modelo agronómico es un modelo de simulación en la agricultura, es una representación simplificada de un sistema real, su utilidad radica no sólo en reproducir la realidad, sino en simplificar y permitir que los procesos más importantes sean identificados, estudiados y permitan predecir los resultados [10]. Por lo tanto, los modelos se pueden utilizar para organizar y reunir el conocimiento de un tema específico, con el fin de mostrar interacciones entre muchos factores [24].

Debemos tener en cuenta que entre más procesos internos y/o externos sean incluidos, la complejidad del modelo será mucho mayor y el proceso de calibración, simulación y validación tendrá un mayor grado de dificultad, consumirá muchos más recursos en hardware y serán requeridos más datos. De esta manera un modelo agronómico, debe ser al mismo tiempo lo bastante complejo y exhaustivo en su concepción para representar el sistema integralmente, y lo bastante simple y comprensible en sus estructuras cuantitativa y dinámica para poder ser aplicado con facilidad. No existe un único modelo agronómico universal, por el contrario, se han desarrollado numerosos modelos en diferentes cultivos que pueden ser simples y describir solamente un proceso, o pueden ser complejos y representar varios procesos y sus interacciones, como la distribución de biomasa y el carbono, la disponibilidad de agua en el suelo, la producción, entre otras. Los modelos agronómicos pueden clasificarse principalmente en empíricos y mecanicistas [10].

2.1.3.1. Modelos agronómicos empíricos

Los modelos empíricos, también llamados modelos estadísticos, describen relaciones entre variables sin referirse a los procesos correlacionados, es decir, describen lo que sucede, sin decir cómo sucede, lo que resulta en un enfoque de caja negra. Las relaciones matemáticas del modelo no

corresponden necesariamente a un proceso biológico, químico o físico, por tanto, no explican el mecanismo de la relación. Los modelos empíricos examinan o representan datos y, por lo tanto, no se adquiere nueva información [24], [25].

2.1.3.2. Modelos agronómicos mecanicistas

Un modelo agronómico mecanicista es generado a partir de procesos o simulaciones, intentan representar relaciones causa-efecto entre las variables, esto ocurre cuando el comportamiento de un sistema se describe matemáticamente, mediante ecuaciones, esa representación del sistema es un modelo matemático. El modelo matemático representa hipótesis asumidas cuantitativamente sobre el sistema real, lo que permite deducir sus consecuencias [25]. Los modelos mecanicistas pueden clasificarse como deterministas y estocásticos, mientras que los modelos deterministas presentan un resultado único o una solución única, los modelos estocásticos dan la probabilidad de un resultado. La principal ventaja de los modelos mecanicistas es que se pueden transferir a otro conjunto de condiciones y, por lo tanto, ofrecen más posibilidades para manipular y mejorar el sistema, lo que los hace ideales para la construcción de escenarios. Los dos propósitos principales de los modelos mecanicistas son mejorar la comprensión científica de los procesos y predecir las consecuencias de la manipulación del sistema de cultivo, con esto pueden ayudar a ahorrar recursos ambientales y costos [24].

2.2. Trabajos relacionados

2.2.1. Estudios sobre modelos agronómicos

Dada la importancia de los modelos agronómicos en este proyecto particular, puesto que son necesarios para la generación de datos confiables en el sector agrícola. Esta sección presenta una investigación sobre los modelos agronómicos más importantes y cuál podría ser utilizado para el cultivo de café, la información se encuentra resumida en la tabla 1.

Tabla 1: Resumen sobre modelos agronómicos

Modelos	Variables a predecir	Disponibilidad
The DNDC Model	Crecimiento del cultivo, la temperatura del suelo, los regímenes de humedad, la dinámica del carbono del suelo, la lixiviación de nitrógeno y las emisiones de gases traza [26], [27].	Abierto, con documentación gratuita
ROTH-CNP	Reservas de carbono, nitrógeno y fósforo, los cambios de reserva, su equilibrio y los flujos de nutrientes [28].	Cerrado, probado únicamente en Reino Unido con bases de datos privadas
Century	Flujos de carbono, nitrógeno, fósforo y azufre [29].	Cerrado, documentación paga
DayCent	Flujos de C y N entre la atmósfera, la vegetación y el suelo [30]	Abierto, con documentación gratuita
APSIM	Balance hídrico, transformaciones de N y P, pH del suelo y erosión [31]	Abierto para uso no comercial, con documentación gratuita
Stics	Equilibrio térmico, radiativo, de agua, carbono y nitrógeno en la escala de tiempo del ciclo del cultivo [32]	Abierto para uso no comercial, con documentación gratuita
EPIC	Rendimiento, irrigación, absorción de nutrientes (carbono, nitrógeno y fósforo), el ciclo del nitrógeno y el carbono, lixiviación ,impactos del cambio climático y los efectos del dióxido de carbono atmosférico [33], [34]	Abierto con documentación gratuita
DSSAT	Crecimiento del cultivo, rendimiento según cambios en	Abierto con documentación gratuita

	las variables de entrada, manejo del agua, fertilidad del suelo (Carbono, Nitrógeno) [35]	
WOFOST	Rendimiento, la biomasa, irrigación, crecimiento, absorción ,desarrollo fenológico, biomasa, el índice de área foliar, el uso del agua y la humedad del suelo [36], [37]	Abierto con documentación gratuita
QUEFTS	Rendimiento, biomasa, Requerimientos de N,P, K [38]	Abierto con documentación gratuita
SAFERNAC	Rendimiento, biomasa, Requerimientos de N,P, K [39]	Abierto con documentación gratuita
DynACof	Productividad primaria neta, crecimiento, rendimiento, balance energético e hídrico [40]	Abierto con documentación gratuita

La tabla 1 muestra que existe un vacío en cuanto al desarrollo de un modelo agronómico para el café en nuestra región, dado que los más avanzados y completos modelos usados a nivel mundial, no cuentan con adaptación al café. Sin embargo, se hallaron dos modelos desarrollados para el cultivo de café, adaptados a las condiciones de otras regiones; uno de ellos es el DynACof [40] que fue diseñado para realizar una predicción sobre el rendimiento del cultivo, teniendo en cuenta particularmente variables de manejo y el modelo SAFERNAC [39] que es una adaptación del modelo QUEFTS [38] el cual fue desarrollado para obtener la predicción de rendimiento y la cantidad de nutrientes clave (N, P y K) necesarios para alcanzar la máxima producción.

2.2.2. Estudios sobre integración entre modelos agronómicos y algoritmos de aprendizaje automático

Se realizó una revisión bibliográfica siguiendo la metodología de Kitchenham para generar revisiones sistemáticas de literatura (SLR, Systematic Literature Reviews) en ingeniería de software. Se emplearon las cadenas de búsqueda ("crop model" OR "agronomic model" OR "crop modeling" OR "crop simulation" OR "biophysical model") AND ("machine learning" OR "deep learning" OR "artificial intelligence") en las bases de datos Web Of Science (WOS), Scopus y Science Direct, todas las búsquedas se realizaron entre los años 2010-2021 y se encontraron 846 documentos, de los cuales fueron descartados aquellos que no se relacionaban con

el tema de investigación o estaban duplicados, obteniendo así, solamente 11 estudios que realizaron una integración entre aprendizaje automático o machine learning y modelos agronómicos. A continuación, se presenta la descripción de los trabajos más destacados en este campo:

- En el estudio [41], se desarrolló un enfoque de modelado híbrido que combina las características de un modelo de crecimiento de cultivos (CROPGRO Soybean) y un modelo de regresión de vectores de soporte (SVR) para la evaluación integral de la variabilidad del agua subterránea bajo diferentes umbrales de riego de soja en toda la temporada de crecimiento. Se calibró CROPGRO para simular los requisitos de riego diarios de la soya. Estos resultados se utilizaron como datos de entrada en la SVR para evaluar la respuesta prevista de los niveles diarios de agua subterránea a diferentes demandas de riego. Los resultados muestran que el modelo híbrido es capaz de evaluar la respuesta del agua subterránea a múltiples escenarios de riego.

- El articulo [42], evaluó el potencial de cuatro algoritmos de aprendizaje automático (LASSO Regression, Ridge Regression, random forest(RF), Extreme Gradient Boosting (XGBoost) y sus combinaciones) y un simulador de sistemas de cultivo (APSIM) para la predicción de rendimiento y pérdidas de N. El conjunto de datos simulado incluyó más de tres millones de datos. XGBoost fue el algoritmo más preciso en la predicción del rendimiento con un error cuadrático medio relativo (RRMSE) del 13,5%. Se encontró que en todos los modelos ML, el error de predicción de rendimiento disminuyó entre un 10% y un 40% a medida que el conjunto de datos de entrenamiento aumentó de 0,5 a 1,8 millones de puntos de datos.

- Zhang. et. al. calibraron el modelo agronómico AquaCrop aprovechando los algoritmos avanzados de inferencia bayesiana y las imágenes multiespectrales UAV a escalas de campo. En particular, las imágenes se aplican primero para obtener el valor de área del dosel(CC) mediante el uso de aprendizaje automático, esta área se usa para el cálculo del índice de área foliar y luego es suministrada al modelo AquaCrop. Los valores de CC predichos por el enfoque bayesiano son consistentes, tienen un RMSE de 0.0271, menor que el enfoque de optimización convencional (0.0514) [43].

- En la investigación [44], fue presentado un método para superar limitaciones actuales de la estimación del rendimiento de los cultivos, este consiste en combinar los datos de teledetección, el modelado de cultivos y algoritmos de ML. Los resultados del modelo de cultivo SARRA-O y los datos medidos se usaron para la calibración. El modelo de rendimiento final se construyó sobre la interacción entre la biomasa aérea en la floración y el estrés hídrico del cultivo durante las fases reproductiva y de maduración. Los resultados

mostraron que el modelo RF fue el de mejor rendimiento concluyendo así, que el método propuesto es un enfoque eficaz para pronosticar los rendimientos de los cultivos de maíz en entornos donde los datos de campo son escasos

- Feng et. al. desarrolló un modelo híbrido incorporando los resultados del modelo APSIM y los indicadores para impactos de los eventos climáticos extremos (ECE) específicos de la etapa de crecimiento (es decir, heladas, sequía y estrés por calor) en el modelo de RF, y el modelo de regresión lineal múltiple (MLR). Los resultados mostraron que el modelo híbrido APSIM + RF podría explicar el 81% de las variaciones de rendimiento observadas en el sureste de Australia, que tuvo una mejora del 33% en la precisión del modelado en comparación con el modelo APSIM solo y el 19% mejora en comparación con el modelo híbrido APSIM + MLR [11].

- Yamamoto et. al. utilizó las redes neuronales convolucionales (CNN) con el fin de evaluar la capacidad del aprendizaje profundo identificando el conocimiento de la fisiología de las plantas detrás de un modelo de cultivo simplemente mediante el aprendizaje. Fue utilziado el modelo de cultivo SIMRIW para generar conjuntos de datos. Los mismos se introdujeron en CNN. Los modelos entrenados fueron evaluados mediante la visualización de mapas de prominencia (técnica usada en redes convolucionales para evaluar el modelo). Los resultados indicaron que CNN determinó el índice de desarrollo del arroz con cáscara, que se implementó en el modelo de cultivo, además el modelo pudo darse cuenta de que el arroz con cáscara se volvía sensible a las temperaturas medias y máximas diarias durante períodos específicos [45].

- Folberth et. al implementó dos enfoques de aprendizaje automático (aumento extremo del gradiente y RF) para desarrollar metamodelos para la predicción de resultados de modelos de cultivos cuadriculados globales(GGCMs) en resoluciones espaciales precisas, los algoritmos de aprendizaje automático se entrenan con las simulaciones de maíz de forma global provenientes del modelo EPIC-IIASA. Los resultados muestran una precisión alta con $R^2 > 0.96$ para las predicciones de los rendimientos de maíz, así como las externalidades hidrológicas, la evapotranspiración y el agua disponible del cultivo [46].

- En esta investigación [47], primero se desarrolló un enfoque híbrido de pronóstico de rendimiento para el trigo mediante la combinación de APSIM (un modelo agronómico basado en procesos), biomasa simulada, extremos

climáticos, NDVI y SPEI (índice estandarizado de precipitación y evapotranspiración), posteriormente estos datos se usaron para entrenar los modelos RF y regresión lineal múltiple. El sistema basado en RF superó al de regresión lineal múltiple. Los pronósticos de rendimiento satisfactorios se produjeron un mes antes de la cosecha (RMSE = 0,70) y dos meses antes de la cosecha (RMSE = 1,01). Además, los eventos de sequía a lo largo de la temporada de crecimiento se identificaron como el principal factor que causó pérdidas de rendimiento durante la última década.

- Saravi et. al hizo uso de un modelo de sistemas agrícolas(DSSAT) para evaluar los impactos de la cantidad de riego y el tiempo de aplicación en el rendimiento del cultivo. Se utiliza y se capacita una red de aprendizaje profundo mediante la incorporación de grandes cantidades de entradas de los modelos DSSAT, es decir, fecha de precipitación, cantidad de precipitación, cantidad de riego de fecha de riego y rendimiento de maíz al final de la temporada de crecimiento. DSSAT fue utilizado para generar un dataset con 100.000 instancias sobre aplicaciones de riego aleatorio durante períodos de 200 días, con estos datos se entrenó la red neuronal Multilayer Perceptron (MLP) variando el número de capas y neuronas. Tres posibles combinaciones de este modelo alcanzaron una precisión superior al 98% [48].

- En el estudio [49], se investigó si la combinación del modelado de cultivos y el ML mejora las predicciones de rendimiento de maíz en EEUU. Se utilizaron cinco modelos ML para abordar la investigación. Los resultados sugieren que agregar variables del modelo de cultivo de simulación (APSIM) como características de entrada a los modelos ML puede marcar una diferencia significativa en el rendimiento de los modelos ML, y puede aumentar el rendimiento ML hasta en un 29%. El análisis sugirió que las variables APSIM más importantes eran las relacionadas con el agua del suelo y, en particular, el estrés por sequía promedio de la temporada de crecimiento

.

- Bai et. al. realizó un experimento de campo de tres años (2014-2016) en la provincia de Jilin. Los datos de 2014 se utilizaron para calibrar el modelo DSSAT y los datos de 2015 se utilizaron para la validación. Después de la calibración y validación, se utilizaron el modelo DSSAT y un algoritmo genético (GA) para optimizar el programa de fertilización N del maíz en menos de 20 años (1973-1992) de datos meteorológicos para Changchun. En los períodos de calibración y validación, el error cuadrático medio normalizado (nRMSE) para el rendimiento de grano fue de 1,45% y 1,61%. Se obtiene un nuevo programa de fertilizantes nitrogenados que exhibe una

cantidad de 198 kg / ha, con el cual se puede aumentar los beneficios económicos en un 8,4% y un 12,4% [50].

De acuerdo con el análisis de los temas investigados mediante el estado del arte, podemos concluir que las investigaciones sobre la integración entre modelos agronómicos y algoritmos de aprendizaje automático han podido solventar problemas como la escasez de datos [44], e incluso se han logrado mejoras en la precisión al momento de medir el rendimiento [11], [49]. En el presente proyecto se resalta el cultivo y país de estudio, ya que ninguna investigación se realizó para el cultivo de café, ni en Colombia, haciendo énfasis en el uso de un modelo agronómico desarrollado para café [39], [40], el cual no ha sido empleado con el enfoque híbrido propuesto. Además, para este caso en particular, se manifiestan restricciones en relación al conjunto de datos iniciales para entrenar un modelo de predicción, por esto la integración de estas dos técnicas se convierte en una alternativa para lograr los resultados esperados en la investigación.

2.2.3. Aportes y brechas de los trabajos relacionados

En los numerales anteriores se pudo establecer que existen diversas investigaciones realizadas aplicando exitosamente la integración entre algoritmos de aprendizaje automático y modelos agronómicos. Algunos trabajos han solventado problemas como la escasez de los datos, mientras otros han utilizado dicha integración para lograr mejoras en la precisión del modelo. Con esto, fue posible identificar los posibles aportes de los trabajos encontrados a la presente investigación y las brechas existentes entre estos, tal como se muestra en la Tabla 2

Tabla 2: Aportes y brechas de los trabajos relacionados

Sección - trabajos	Aportes	Brechas
.Estudios sobre modelos agronómicos	Proponen un modelo matemático para el cálculo de rendimiento de un cultivo, cantidad de nutrientes, balance hídrico, entre otros. Presentan ejemplos de calibración de modelos matemáticos para cultivos como el trigo, la soja. Realizan un modelamiento de las relaciones existentes entre diferentes variables incidentes en el óptimo	Los modelos agronómicos más avanzados y completos no cuentan con adaptación para el café. La calibración de los modelos se ha realizado principalmente en

	desarrollo de un cultivo, tales como variables climáticas, edáficas y de manejo del cultivo.	países europeos, asiáticos o norteamericanos.
Estudios sobre integración entre modelos agronómicos y algoritmos de aprendizaje automático	Solucionan problemas graves para el modelado cómo la escasez de datos. Estudian la precisión de diferentes técnicas de aprendizaje automático en integración con modelos agronómicos Presentan una mejora en la precisión del modelo después de la integración. Analizan el uso de algoritmos de aprendizaje automático complejos, como el aprendizaje profundo Exploran el uso de la integración para predecir los impactos de los eventos climáticos extremos en los cultivos	Los estudios son realizados a nivel global o en países de otros continentes. Los estudios son realizados para diversos cultivos, especialmente cultivos transitorios como la soya, trigo, maíz, entre otros. Los estudios se enfocan principalmente en el rendimiento del cultivo.

2.3. Conclusiones acerca del estado actual del conocimiento

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

- Existe gran variedad de investigaciones dedicadas al estudio y calibración de los modelos agronómicos. Se encontraron distintos enfoques como lo son la predicción del rendimiento, cantidad de nutrientes clave necesarios, fertilización, irrigación, entre otros. Los cultivos estudiados comprenden el trigo, avena, caña de azúcar, arroz, soya, maíz, café, papa, entre otros.

- Mediante el uso de la integración entre modelos agronómicos y algoritmos de aprendizaje automático es posible construir soluciones para distintos problemas como lo es la escasez de datos, además, con la integración se logró mejorar la precisión de los modelos predictivos que únicamente

utilizaron algoritmos de aprendizaje automático.

- Máquinas de vectores de soporte, regresión lineal múltiple y RF, son algunas de las técnicas utilizadas en la integración con modelos agronómicos. Otros algoritmos empleados en menor medida, pero con interesantes resultados, son los de aprendizaje profundo.

- Después de analizar las investigaciones acerca de la integración entre modelos agronómicos y algoritmos de aprendizaje automático, se concluye que la presente investigación propone un nuevo enfoque que puede aportar no solamente a la agricultura colombiana, sino también al campo ingenieril. Utilizando un modelo agronómico novedoso en el café, calibrando el mismo para la región caucana y analizando diferentes algoritmos de aprendizaje automático, logrando estimar la disponibilidad de N en suelos con cultivo de café.

- Existen diversos modelos agronómicos desarrollados y calibrados para diferentes cultivos, sin embargo, se lograron identificar únicamente dos modelos que cuentan con enfoque en el cultivo de café, de los cuales, solo uno considera las variables edáficas, de manejo de cultivo, y por ende, la estimación de la disponibilidad de nitrógeno dentro del modelado.

3. Calibración del modelo agronómico QUEFTS a la región de estudio

Este capítulo presenta el proceso llevado a cabo para la calibración del modelo agronómico QUEFTS a la región de estudio, es decir, el departamento del Cauca Dicho modelo fue utilizado para la construcción de un conjunto de datos sobre producción de café, condiciones foliares iniciales y la fertilización aplicada hasta la cosecha. Se toma como referencia la metodología para calibración y validación de un modelo agronómico utilizada en [51]. La Figura 1 muestra las fases propuestas por dicha metodología que fueron implementadas en la presente investigación.

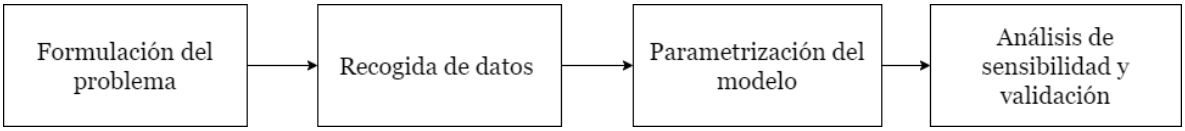


Figura 1. Fases que conforman la calibración y validación. [51]

La fase de formulación del problema está abarcada en el primer capítulo de la presente investigación. Teniendo esto en cuenta se contemplaron las siguientes actividades en las cuales va a estar dividido el capítulo actual:

- Recogida de datos: en esta primera etapa es importante que se definan con claridad y exactitud los datos que el modelo va a requerir para la calibración y poder producir los resultados deseados.

- Parametrización del modelo QUEFTS: en esta fase se explica detalladamente el funcionamiento del modelo QUEFTS, así mismo se describen los cambios realizados durante el proceso de calibración

- Análisis de sensibilidad: se establece un plan de pruebas para realizar un análisis de sensibilidad que permita ajustar valores de algunos parámetros del modelo con fin de mejorar su efectividad para la región caucana

- Validación del modelo QUEFTS: se construye un conjunto de datos sintéticos a partir del modelo y se evalúa el rendimiento

3.1. Recopilación de datos

Para esta primera fase de recolección de datos utilizamos las cadenas de búsqueda ("coffee" OR “coffe*”) AND ("nitrogen" OR "nitrate" OR "ammonium"), así como también (“café” OR “coffe*” OR “cafe*”) AND (“nitrógeno” OR “nitrato” OR “amonio”) en las bases de datos Google Scholar y Sciencedirect. Además de realizar una búsqueda en las publicaciones de Cenicafé de la federación nacional de cafeteros,

para encontrar un total de 12 registros que cumplieron con los criterios de búsqueda, es decir, que fueran plantaciones de café arábica variedad Caturra, y estuvieran ubicados en el departamento del Cauca, o en suelos con características similares, provenientes de los estudios [52]–[56], adicionalmente fue necesario el asesoramiento con 2 expertos en caficultura caucana, para establecer los lineamientos necesarios para filtrar adecuadamente las investigaciones y seleccionar la información relevante para calibrar el modelo agronómico a la región de estudio.

3.2. Parametrización del modelo QUEFTS

El modelo agronómico QUEFTS originalmente no tiene entre sus opciones trabajar con el cultivo de café, ya que este no hace parte de los cultivos que vienen previamente parametrizado por el modelo, por esta razón debe ser creado en el archivo quefts_crops_pairs.csv, para revisar este archivo por favor dirigirse a la sección A de los anexos. Las variables requeridas se obtuvieron de la revisión de la literatura en el paso 3.1 y son las siguientes:

- NminStore
- NminVeg
- NmaxStore
- NmaxVeg
- PminStore
- PminVeg
- PmaxStore
- PmaxVeg
- KminStore
- KminVeg
- KmaxStore
- KmaxVeg
- New_Yzero
- Nfix

_minVeg, _maxVeg, _minStore, _maxStore corresponden a la concentración mínima y máxima de "_" (N, P o K) en órganos vegetativos y en órganos de almacenamiento, los significados de New_Yzero y Nfix se pueden encontrar en la tabla 3:

Tabla 3. Lista de acrónimos del modelo QUEFTS

Acrónimo	Significado
SOC	Carbono orgánico en el suelo
Kex	Potasio Intercambiable
PBray	Fósforo extraído con el método de Bray
$\boldsymbol{\beta}$N	Beta N
$\boldsymbol{\beta}$P	Beta P
$\boldsymbol{\alpha}$P	Alpha P
$\boldsymbol{\alpha}$K	Alpha K
NrTrees	Número de árboles por hectárea
New_Yzero	Biomasa máxima de órganos vegetativos teniendo el rendimiento en cero de órganos de almacenamiento
Nfix	La fracción de la absorción de nitrógeno de un cultivo, proporcionada por la fijación biológica.
iYratA	Eficiencia fisiológica (PhE) en la máxima acumulación del nutriente i
iYratD	Eficiencia fisiológica (PhE) en la máxima dilución del nutriente i
Si	Suministro disponible del nutriente i-ésimo
Ui(j)	Se refiere a la absorción del i-ésimo nutriente en relación con j
Ui	Absorción real de i-ésimo nutriente
$Y(i)^a$	Rendimiento con la máxima acumulación del nutriente i-ésimo
$Y(i)^d$	Rendimiento con la máxima dilución del nutriente i-ésimo
Y_{ij}	Rendimiento para el par de nutrientes i y j
Y0	Máximo rendimiento potencial

Una vez el cultivo de café ha sido creado, empezamos a describir a profundidad el modelo, el proceso que se realiza en el modelo internamente será explicado en 4 pasos, desde que son ingresados los datos iniciales hasta el cálculo del rendimiento

y demás salidas del modelo.

Paso 1 (evaluación del suministro de nutrientes disponibles): el primer paso es calcular el suministro de nutrientes autóctonos del suelo, estos serán los valores iniciales de nutrientes antes del proceso de siembra, es decir, sin fertilización. Esto es logrado llamando a la función NutSupply, modificada de acuerdo con [39], también agregamos los parámetros de entrada NrTrees y New_Yzero para mejorar la practicidad del modelo.
Los parámetros de entrada son pH, SOC, Kex, PBray, NT y newYzero. La función esta descrita en el algoritmo nutSupply mostrado a continuación, para consultar el código completo, por favor dirigirse a la sección A de los anexos:

$$nutSupply1 <- function(pH, SOC, Kex, Pbray, Nt, newYzero)\{$$

$$NrTrees = NT$$

$$New\_Yzero = newYzero$$

$$N\_base\_supply = (0.25 * ([\mathrm{pH}] - 3) * \beta N * [SOC]$$

$$P\_base\_supply = (1 - 0.5 * ([\mathrm{pH}] - 6)^2) * \alpha P * [SOC] + (\beta P * [Pbray])$$

$$K\_base\_supply = \frac{((2 - 0.2 * ([\mathrm{pH}]) * \alpha K * [Kex])}{[SOC]} \}$$

Las salidas del algoritmo son N_base_supply, P_base_supply, K_base_supply, NrTrees y New_Yzero. Las variables base_supply corresponden al suministro de nutrientes del suelo autóctono. Las variables α y β tienen como objetivo ajustar dichas variables según el suelo de cada región, tema que será tratado en la sección 3.3 del presente capítulo.

Agregando la variable fD para manejar el número de árboles por hectárea en el modelo, se adiciona la fórmula encontrada en [39], pero fue ajustada a 5000 árboles, considerando la máxima densidad de siembra establecida en el departamento del Cauca [13], tal como se muestra en la ecuación 1:

$$fD = \left(0.4 * \left(\frac{NrTrees}{1000}\right)\right) - \left(0.4 * \left(\frac{NrTrees}{1000}\right)^2\right) \quad (1)$$

Esta variable es utilizada después para calcular los valores totales de suministro de nutrientes del suelo, como se observa en la ecuación 2.

$$i\_supply = \left(i\_base\_supply + (i\_fertilizer * i\_recovery) + (i\_o\_fertilizer * i\_recovery\_o)\right) * fD \quad (2)$$

Donde i = N, P, K; También se eliminó la variable temporada de rebrote y se agregó la posibilidad de abono orgánico y su tasa de absorción orgánica para calcular los

aportes totales de nutrientes, considerando que un alto porcentaje de sistemas productivos de café realizan un manejo nutricional mixto, donde utilizan los residuos del beneficio de café como abono orgánico.

Paso 2 (relación entre el suministro de nutrientes disponibles y la absorción real): las relaciones entre el suministro de nutrientes y la absorción real se calcularon utilizando la función de absorción, también se eliminaron los parámetros 'Zero', tal como se muestra en el algoritmo de la función uptake:

$$uptake(Si, iYratA, iYratD, Sj, jYratA, jYratD)\ \{$$

$$if\left(Si\ <\left(jYratA\ *\ \frac{(Sj\ )}{iYratD}\right)\right)\ \{$$

$$Ui(j)\ =\ Si;$$

$$\}$$

$$else\ if\left(Si\ <\left((Sj) * \frac{2\ *\ jYratD}{iYratA - \frac{jYratA}{iYratD}}\right)\right)\ \{$$

$$Ui(j)\ =\ Si\ -\left(0.25\ * \frac{\left(Si - \frac{jYratA\ *\ Sj}{iYratD}\right)^2}{\left(\frac{jYratD}{iYratA} - \frac{jYratA}{iYratD}\right) * Sj}\right)$$

$$\}\ else\ \{$$

$$Ui(j)\ =\ jYratD\ *\ (Sj)\ /\ iYratA;$$

$$\}$$

$$\}$$

Donde **i, j = N, P, K, i ≠ j;** Ui(j) se refiere a la absorción de i-ésimo nutriente en relación con j, si i = N, j podría ser P o K; **Si** = es el suministro de nutriente i disponible, obtenido en el paso 1; **iYratA** = eficiencia fisiológica (PHE) a la máxima acumulación de nutriente i (kg grano kg−1 nutriente i); **iYratD** = eficiencia fisiológica (PhE) a la dilución máxima del nutriente i (kg grano kg−1 nutriente i).

La absorción real Ui será el valor mínimo de **Ui(j)** y **Ui(k)**, donde **i, j, k = N, P, K, i ≠ j ≠ k**; por ejemplo: UN será el valor mínimo de UN(P) y UN(K)

Paso 3 (relación entre la absorción real y los rangos de rendimiento): los principios utilizados en QUEFTS [38] en esta etapa, establecen que los rangos de rendimiento calculados entre el rendimiento **Y(i)^{a}** en la acumulación máxima (a) y el rendimiento **Y(i)^{d}** en la dilución máxima (d), como funciones de la Ui de captación real, tal como se muestra en la ecuación 3:

$$Y\ (i)^a\ =\ iYratA\ *\ Ui, i=\ \ N, P, K$$
$$Y\ (i)^d\ =\ iYratD\ *\ Ui, i=\ \ N, P, K \tag{3}$$

Paso 4 (combinación de rangos de rendimiento con estimaciones de rendimiento final): en este paso final, los rangos de rendimiento son combinados para cada par de nutrientes, y luego promediados los rendimientos estimados para obtener una estimación de rendimiento final, fue utilizada la función yield para calcular el rendimiento $\mathbf{Y_{ij}}$ para el par de nutrientes i y j. Esta función es modificada según [39] y podemos observarla en el siguiente algoritmo:

$$yield(Ui, iYratD, iYratA, Y(j)^a, Y(j)^d, Y(k)^d, Y0)\ \{$$
$$YxD\ =\ min\ (\{\ Y(j)^d, Y(k)^d, Y0\});$$
$$if\ (Ui\ ==\ 0\ ||\ YxD\ ==\ 0)\{$$
$$Y_{ij}\ =\ 0;$$
$$\}$$
$$else\{$$
$$Y_{ij}\ =\ Y(j)^a\ +(2\ *\ (YxD\ -\ Y(j)^a)\ *\ (Ui\ -\ Y(j)^a\ /\ iYratD))\ /\ (YxD\ /\ iYratA\ -\ Y(j)^a\ /\ iYratD)\ -\ (YxD\ -\ Y(j)^a)\ *\left(\frac{Ui\ -\ Y(j)^a}{iYratD}\right)^2\ \frac{\left(\frac{Ui\ -\ Y(j)^a}{iYratD}\right)^2}{\left(\frac{YxD}{iYratA\ -\ \frac{Y(j)^a}{iYratD}}\right)^2}$$
$$\}$$
$$\}$$

Donde **i, j, k = N, P, K, i ≠ j ≠ k;** Y0 corresponde al rendimiento potencial máximo o biomasa alcanzable (en ausencia de limitación de nutrientes) (kg ha^{-1}) para órganos de almacenamiento.

La estimación de rendimiento final Y(U) se calcula siguiendo la ecuación 4.

$$limYield\ =\ min(\{Y(N)^d, Y(P)^d, Y(K)^d, Yo\})$$
$$Y(U)\ =\ min\left(limYield, \frac{(YNP\ +\ YPN\ +\ YNK\ +\ YKN\ +\ YPK\ +\ YKP)}{6}\right) \tag{4}$$

Con esto, es obtenida la salida principal del modelo, es decir, el rendimiento esperado junto con las brechas de cada nutriente necesarias para alcanzar el

máximo de productividad y las estimaciones de dichos nutrientes en el suelo durante todo el periodo de siembra y cosecha.

3.3. Análisis de sensibilidad

Continuando con la calibración del modelo, presentamos el plan de pruebas realizado al modelo QUEFTS modificado según el paso 3.2. Este se ha diseñado con el propósito de lograr ajustar la sensibilidad de los parámetros usados a la región de estudio. El plan de pruebas consistió en ir variando cada una de las variables a ajustar, dejando las demás constantes. Luego comparar los resultados con los estudios realizados en la región de estudio [52], [53]. En este orden de ideas fue calculado el error, el cual es la diferencia entre el rendimiento calculado por el modelo y el real presentado en los estudios. El valor final será el que muestre una mejor adaptación al rendimiento presentado en la región, es decir, el valor que conlleve a la menor diferencia entre el rendimiento real y el entregado por el modelo.

A continuación, son visualizadas las gráficas que muestran el error calculado (Kg/ha) durante el análisis de sensibilidad. Los archivos que dan soporte a esta información estan en la sección A de los anexos. El primer paso fue realizar un análisis de sensibilidad a la eficiencia fisiológica (PHE) a la máxima acumulación de nutrientes: nitrógeno (N), fósforo (P) y potasio (K), estas variables se encargan de la adaptación del modelo frente a los nutrientes acumulados en el suelo, es decir, son las encargadas de definir en qué medida deben ser procesados los nutrientes acumulados internamente en los cálculos del modelo agronómico. Esto es indispensable para el correcto funcionamiento del modelo ya que cada suelo de una región es considerablemente distinto y asimila en una manera diferente la absorción de nutrientes. En ese orden de ideas, las figuras 2, 3 y 4 muestran el análisis de sensibilidad realizado para las variables PHE a la máxima acumulación de nutrientes N, P y K.

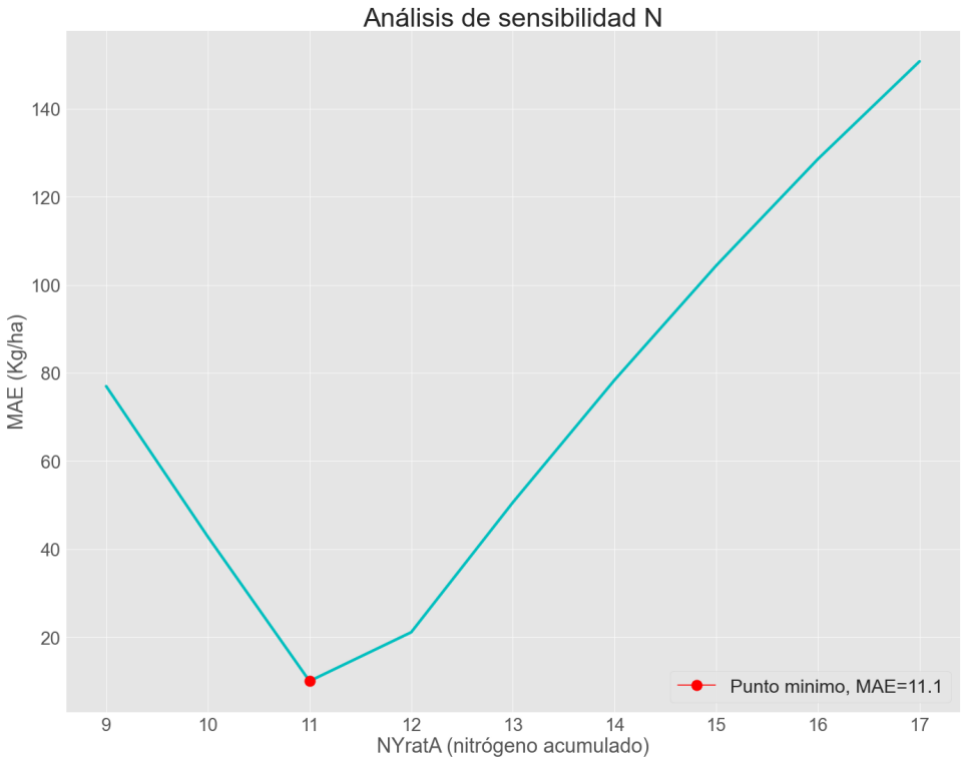


Figura 2. Análisis de sensibilidad para PHE a máximo nitrógeno acumulado.

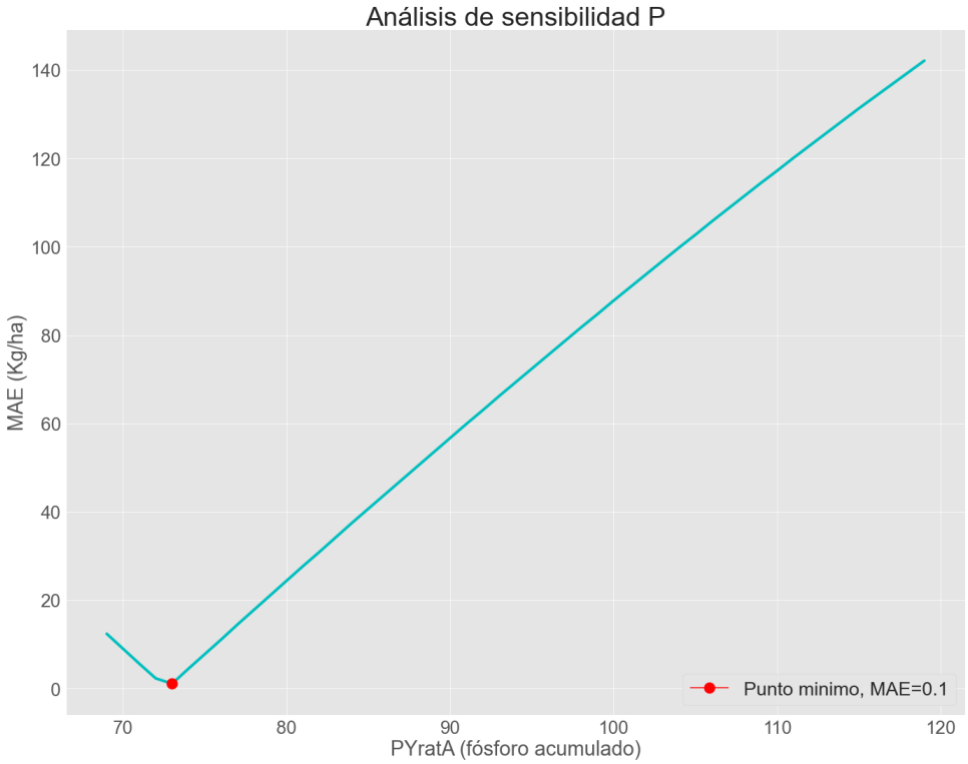


Figura 3. Análisis de sensibilidad para PHE a máximo fósforo acumulado.

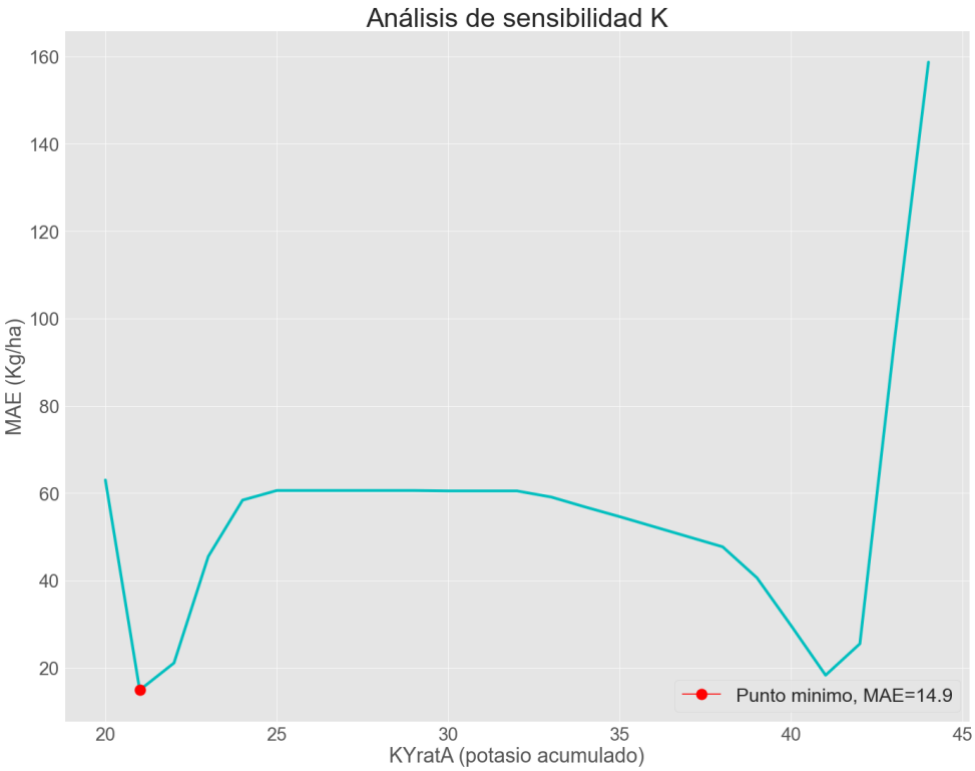


Figura 4. Análisis de sensibilidad para PHE a máximo potasio acumulado.

Una vez se hallan los valores adecuados para el PHE acumulado de todos los nutrientes, el siguiente paso consistió en ajustar los parámetros intrínsecos del modelo, es decir, los parámetros que se encuentran en las ecuaciones usadas para calcular el suministro de nutrientes (N, P y K) autóctonos del suelo. Estos nutrientes son calculados por medio de las ecuaciones descritas en el paso 1 de la sección 3.2 del presente capitulo.

Las figuras 5-8 muestran el análisis de sensibilidad realizado para todas las variables involucradas en el cálculo del suministro mencionado previamente.

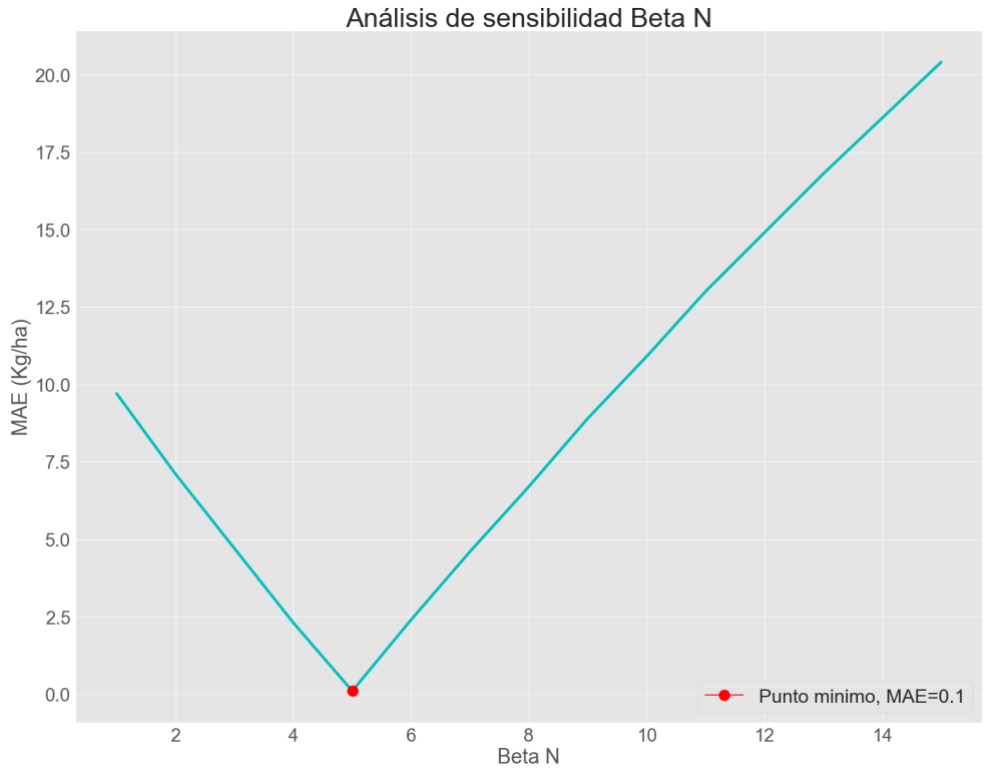


Figura 5. Análisis de sensibilidad para Beta N.

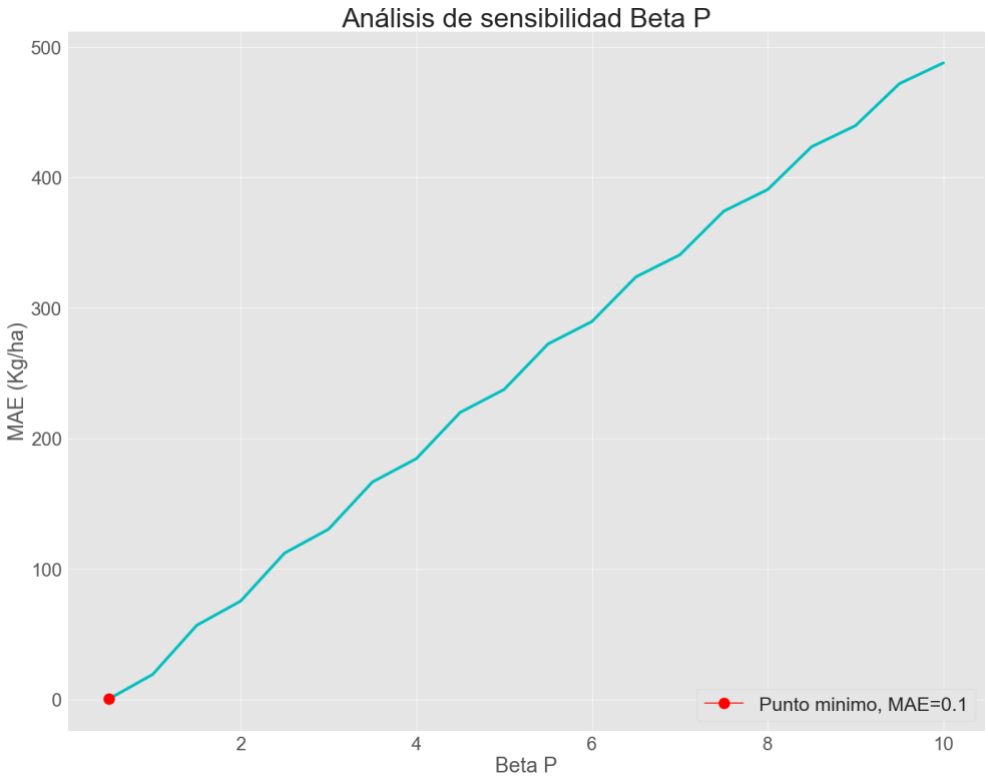


Figura 6. Análisis de sensibilidad para Beta P.

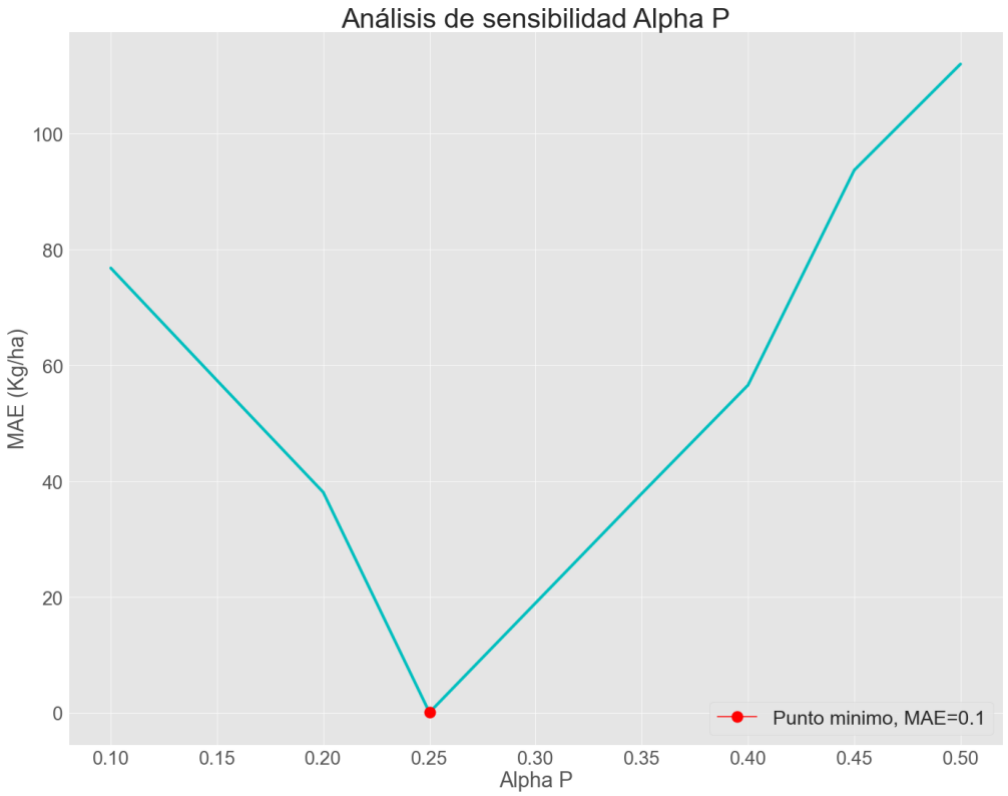


Figura 7. Análisis de sensibilidad para Alpha P.

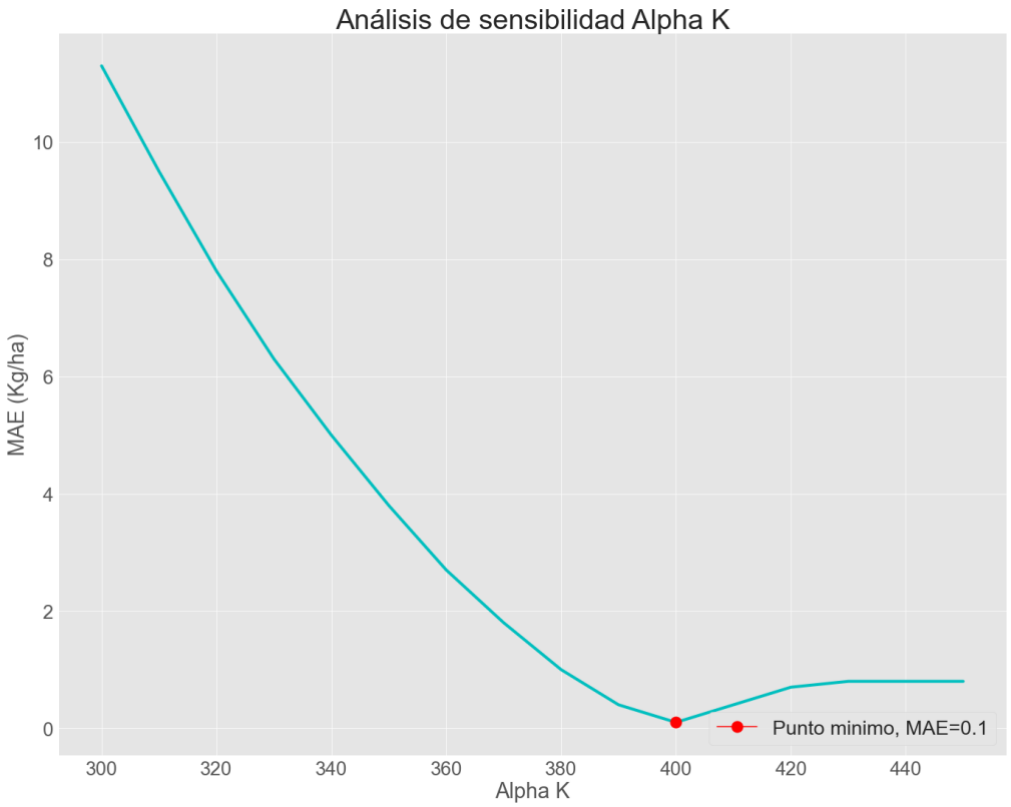


Figura 8. Análisis de sensibilidad para Alpha K.

La información del análisis de sensibilidad realizado a lo largo de la presente sección se encuentra resumida en la siguiente tabla, junto con los valores finales, que serán los valores que tomarán dichas variables en el modelo para continuar con el desarrollo de la investigación:

Tabla 4. Resultados del análisis de sensibilidad

Variable	Rango de pruebas	Tasa de cambio	Valor final
NYratA	7-16	1	11
PYratA	70-120	1	77
KYratA	20-45	1	22
BetaN	1-15	1	5
BetaP	0.5-10	0.5	0.5
AlphaP	0.1-0.5	0.05	0.25
AlphaK	300-450	10	400

Con los datos calculados y resumidos en la tabla 4 se da por terminado el proceso de calibración del modelo agronómico QUEFTS a la región del departamento del Cauca, se puede decir que el modelo resultante de la presente sección está adaptado al funcionamiento del suelo en el departamento y por consiguiente se encuentra calibrado, esto gracias a los análisis de sensibilidad realizados tanto a los parámetros PHE acumulados para todos los nutrientes, como a los realizados para las variables que intervienen en el cálculo de la disponibilidad de nutrientes autóctonos del suelo. Con esto, se le da paso a la siguiente sección en dónde se debe validar los resultados obtenidos.

3.4. Validación del modelo QUEFTS

Para la validación del modelo agronómico QUEFTS en el departamento del Cauca, se presenta la construcción de un conjunto de datos a partir del paso 3.1, se recolectaron un total de 12 experimentos a partir de los estudios [52]–[56], es importante resaltar que los experimentos cuentan con una densidad de siembra que va desde 2500 hasta las 7843 plantas por hectárea, de la misma manera, tienen diferentes planes de fertilización, esto con el fin de compararlos con los resultados generados por el modelo agronómico QUEFTS resultante de los pasos 3.1, 3.2 y 3.3. Así mismo, en esta sección se describen las diferentes métricas base de la evaluación del modelo generado. Mediante las cuales, se calcula el error, que será

la diferencia entre el rendimiento calculado por el modelo y el real presentado en los estudios, así como otras métricas claves para poder evaluar adecuadamente un modelo agronómico

3.4.1. Métricas de evaluación del modelo

Para determinar la precisión del modelo agronómico calibrado con el objetivo de calcular la disponibilidad de N en suelos con cultivo de café, evaluamos los datos estimados frente a los reales mediante el cálculo de diferentes métricas estadísticas. A continuación, describimos las métricas utilizadas para la evaluación del desempeño de modelos agronómicos:

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

$$MAEE = \frac{1}{n}\sum_{i=0}^{n}|ri - ei| \quad (5)$$

  Donde ri indica el valor real en la posición i, ei es el i-ésimo valor estimado y n es el número de datos estudiados.

- ➔ *Coeficiente de determinación (R^2):* estima la dispersión combinada contra la dispersión simple de la serie observada y pronosticada. Su valor oscila entre 0 y 1, donde un valor de 0 significa que no hay ninguna correlación y un valor de 1 significa que la dispersión de la predicción es igual a la de la observación. es calculado con la siguiente ecuación [51]:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(ri - \bar{r})(ei - \bar{e})}{\sqrt{\sum_{i=1}^{n}(ri - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(ei - \bar{e})^2}}\right)^2 \quad (6)$$

  En dónde ri indica el valor real en la posición i, ei es el i-ésimo valor estimado, $\bar{r}$ corresponde a la media de los valores reales, $\bar{e}$ corresponde a la media de los valores estimados por el modelo y n es el número de datos estudiados.

- ➔ *Porcentaje de sesgo (PBIAS):* mide la tendencia promedio de los valores simulados a ser mayores o menores que los observados. El valor óptimo es 0.00, los valores de baja magnitud que indican una simulación precisa del modelo. Los valores positivos indican una subestimación del modelo y los

valores negativos indican una sobreestimación del modelo, el cálculo de este porcentaje se realiza con la siguiente ecuación [51]:

$$PBIAS = \frac{\sum_{i=1}^{n}(ri - ei) * (100)}{\sum_{i=1}^{n} ri} \quad (7)$$

Dónde ri indica el valor real en la posición i, ei es el i-ésimo valor estimado y n es el número de datos estudiados.

3.4.2. Proceso de validación del modelo QUEFTS

Como primera medida son seleccionados los registros con los que se realizará la calibración, luego es calculado el rendimiento usando el modelo QUEFTS, y posteriormente calculamos el error medio absoluto (MAE). En la figura 9 visualizamos los resultados obtenidos en el proceso de validación, observamos el rendimiento calculado por el modelo agronómico para cada uno de los registros comparado con el real, evidenciando las posibles diferencias entre ellos. Los archivos que dan soporte a esta información y la encontrada a lo largo del presente apartado están en los anexos en la sección A.

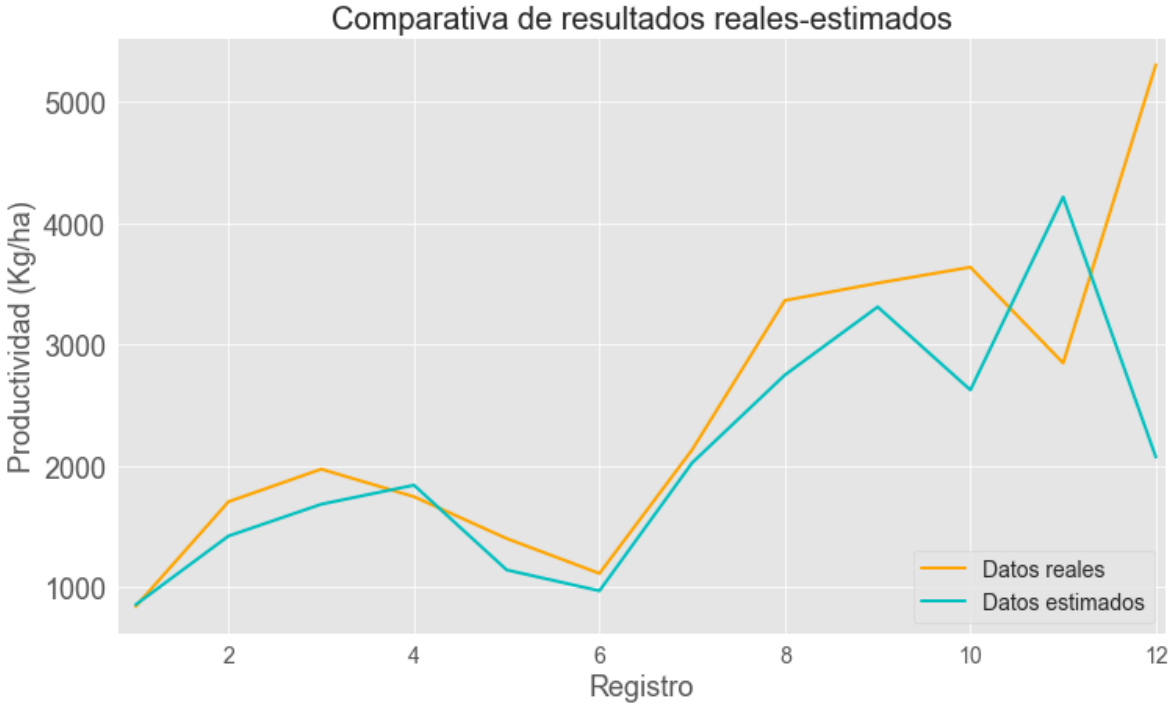


Figura 9. Comparación de resultados reales-estimados.

Teniendo en cuenta los resultados obtenidos en la Figura 9, procedimos a sacar el error, información que esta registrada en la Figura 10**,** donde podemos observar la diferencia entre cada uno de ellos.

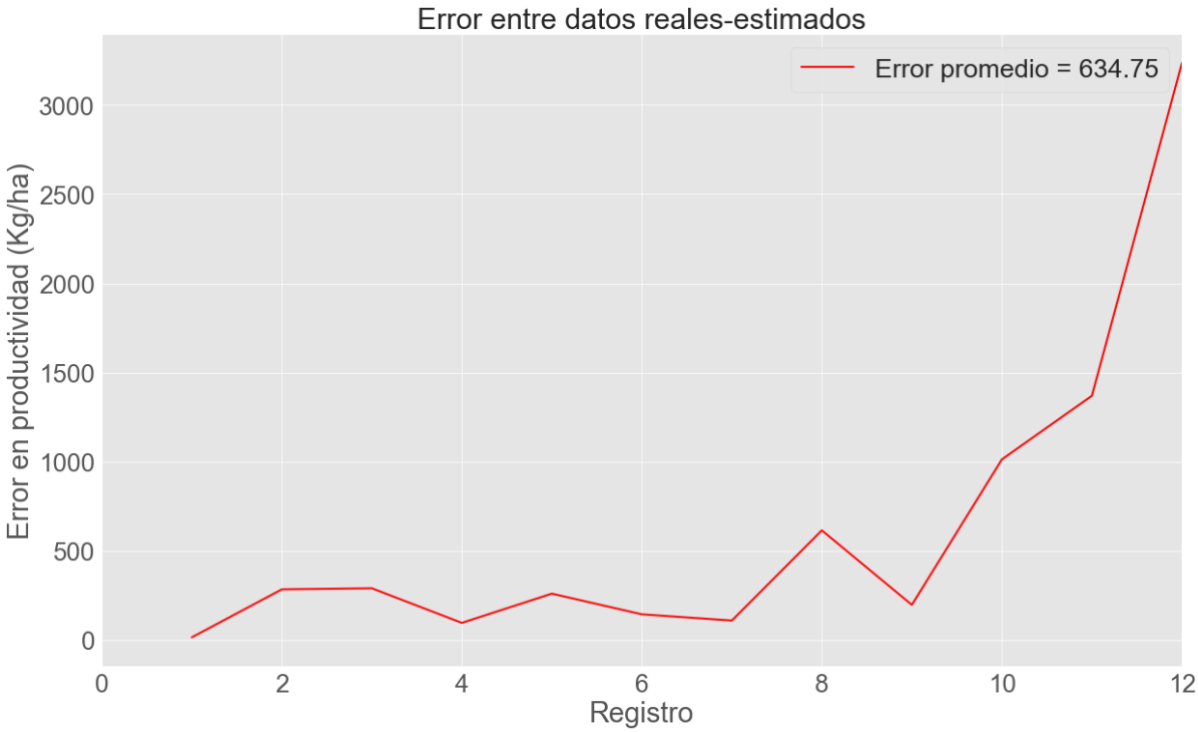


Figura 10. Error entre datos reales-estimados.

Analizando los resultados obtenidos en las figuras 9 y 10, obtuvimos en promedio un error (MAE) de 634.75 Kg/ha, junto con un R^2 de 0.24. Sin embargo, revisando detalladamente observamos que los registros 10,11 y 12 resultaron con un error considerablemente más grande que los demás, dichos registros pertenecen a los experimentos realizados con la densidad de siembra por hectárea mayor, es decir, 7000, 7000 y 7843 plantas por hectárea respectivamente. Teniendo esto en cuenta, decidimos eliminar dichos registros del proceso de validación y repetir el proceso. A su vez, concluimos que el modelo QUEFTS resultante de los pasos 3.1, 3.2 y 3.3 no está calibrado para estudios que contengan 7000 plantas por hectárea o más, esto debido al alto error presentado en el proceso de validación, en promedio 1872 Kg/ha, que puede observarse en la Figura 11.

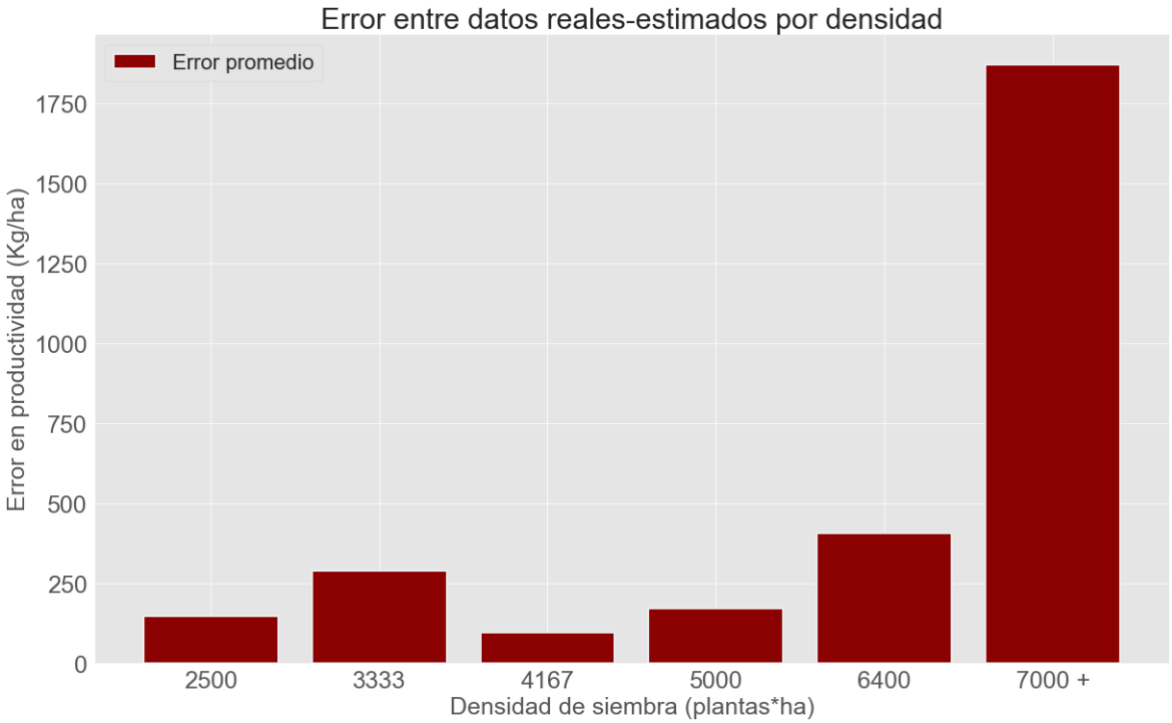


Figura 11. Error entre datos reales-estimados por densidad de siembra.

Una vez eliminados los registros que tienen densidad de siembra de 7000 o más plantas/ha, repetimos el procedimiento de validación y se obtuvo un MAE de 222.3 Kg/ha, tal como se muestra en las figuras 12 y 13.

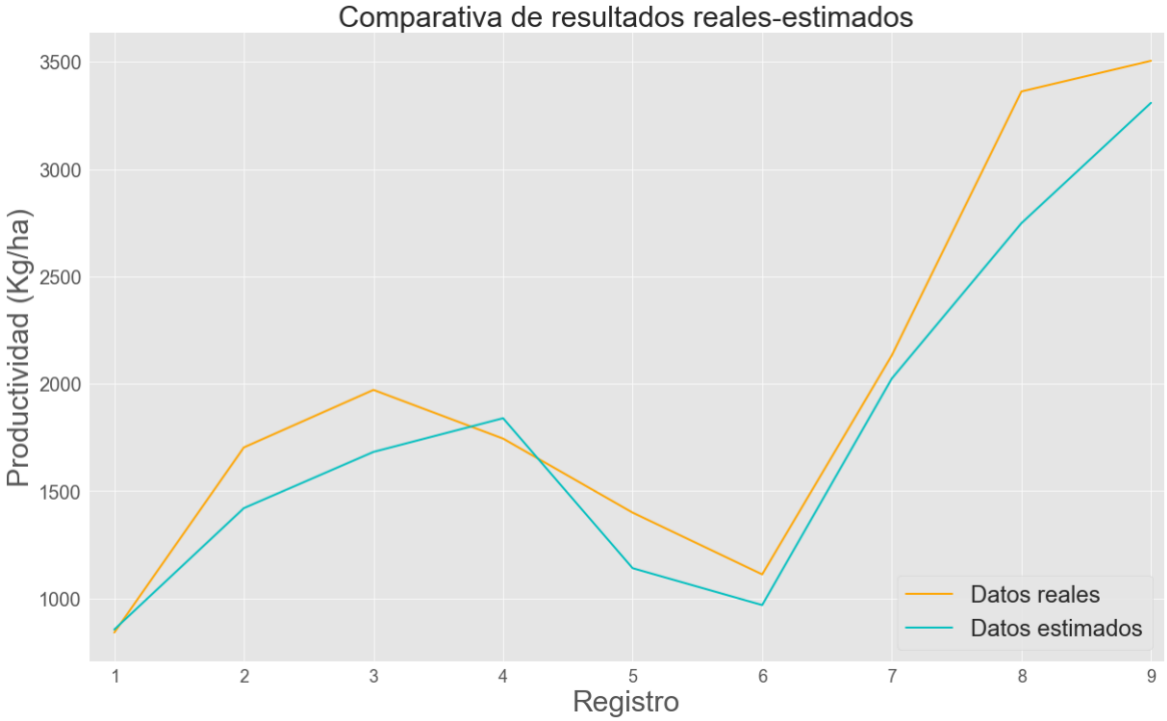


Figura 12. Comparación entre datos reales-estimados.

Figura 13. Error entre datos reales-estimados por densidad de siembra.

Con el modelo agronómico propuesto alcanzamos un error MAE de 222.3 Kg/ha, R^2 de 0.898 y un PBIAS de 10.03%, éste último indica que el modelo está subestimando ligeramente los valores reales, por ende, realizamos una serie de pruebas, con el fin de encontrar el mejor modelo agronómico posible. En ese orden de ideas, fueron probados distintos enfoques y encontramos que el modelo alcanza su mejor rendimiento, es decir su error MAE más bajo cuándo trabaja con densidad de siembra de hasta 5000 plantas/ha, tal como queda evidenciado en la Figura 14.

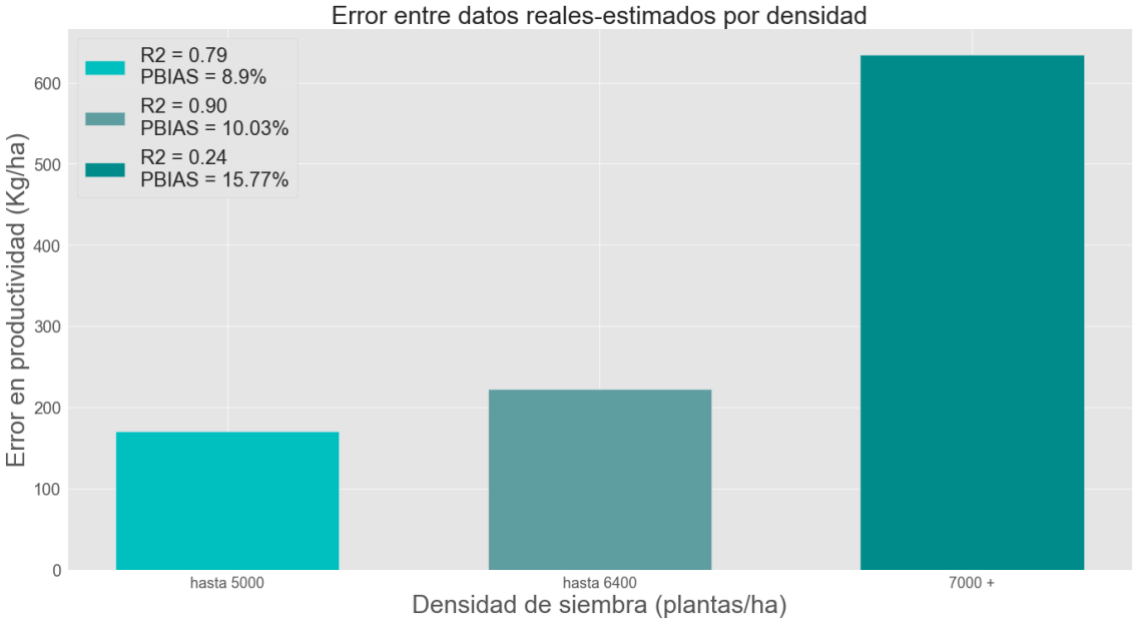


Figura 14. Error entre datos reales-estimados según densidad de siembra.

Como podemos observar en la Figura 14, logramos un modelo agronómico con un error MAE de 170.1 Kg/ha, R^2 de 0.79 y PBIAS de 8.9% para fincas con plantaciones de hasta 5000 plantas/ha, resultado con el cual es dado por finalizado el proceso de validación del modelo agronómico para la región caucana. Por consiguiente, concluimos que el modelo esta validado y representa correctamente el comportamiento de la caficultura en el departamento del cauca en fincas cafeteras de hasta 5000 plantas/ha. Por tal motivo, para el desarrollo del presente proyecto es necesario trabajar con ese máximo rango de densidad.

En esta sección es importante tener en cuenta, que el modelo ya fue calibrado, pero ha sido validado únicamente para el rendimiento del café, ya que esta es la principal salida del modelo agronómico y las demás son dependientes de la misma. Sin embargo, la presente investigación está centrada en la disponibilidad de Nitrógeno (N) en suelos establecidos con cultivo de café, por tal motivo fue necesario hacer una validación para dicho rubro que es entregado por el modelo agronómico. En ese orden de ideas, en la Figura 15 muestra el error entre los datos reales y estimados para la disponibilidad de N en los registros usados para la calibración del modelo.

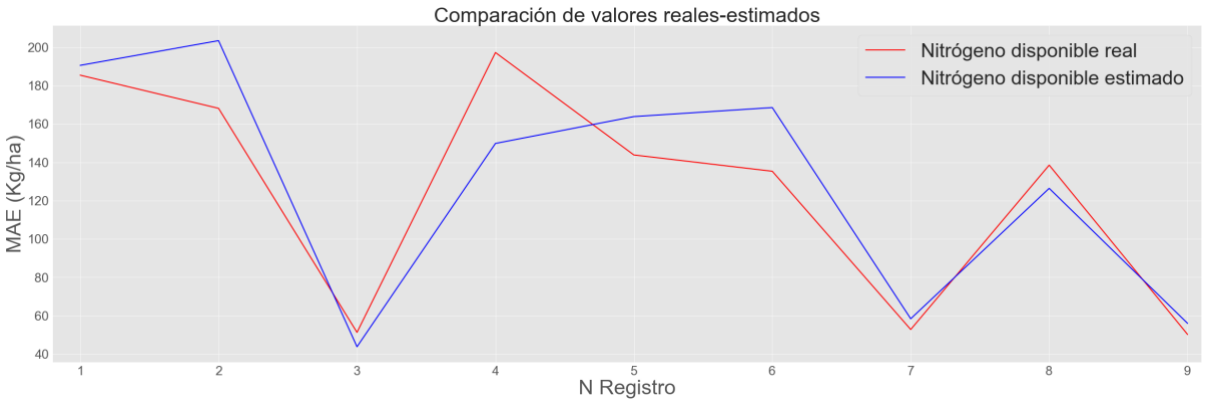


Figura 15. Error entre datos reales-estimados para disponibilidad de N.

Tal como podemos observar en la Figura 15, el modelo agronómico QUEFTS calibrado y validado a lo largo del presente capítulo muestra adaptación a la disponibilidad de N en suelos establecidos con cultivo de café, logrando un error MAE promedio de 19.11 Kg/ha y un R^2 de 0.80. Teniendo esto en cuenta, queda validado el modelo agronómico tanto para el rendimiento como para la disponibilidad de N en suelos establecidos con café.

Es importante tener en cuenta que existen una serie de suposiciones y pre requisitos, propios del modelo QUEFTS inicial [39] y por ende aplican al modelo resultante del presente capítulo, se deben cumplir para el correcto funcionamiento del modelo, y estas son:

- ➔ La fertilidad del suelo se concibe como la capacidad de un suelo para proporcionar a las plantas nitrógeno, fósforo y potasio como macronutrientes primarios. Por lo tanto, el sistema asume que otros nutrientes son mucho menos limitantes que esos tres, y por ende no los toma en cuenta.

- ➔ La irradiancia solar y la disponibilidad de humedad deben estar en óptimas condiciones
- ➔ El suelo debe estar bien drenado, es decir tener un mínimo de clase 3 de drenaje [58]
- ➔ El suelo es lo suficientemente profundo, es decir que tiene más de 90 cm de profundidad.

Por otro lado, el modelo también hace algunas suposiciones sobre los rangos que deben manejar algunas de las variables utilizadas, esto para garantizar el correcto funcionamiento del mismo, los rangos que deben tenerse en cuenta para que el modelo presentado en este capítulo funcione adecuadamente están detallados en la Tabla 5

Tabla 5: Rangos de variables para el modelo resultante del capítulo 3, Fuente propia

Variable	Rango
pH	4.5 - 7.0
SOC en la capa superior del suelo (0-20 cm)	menos de 70 g/Kg
P Bray en la capa superior del suelo (0-20 cm)	menos de 30 mg/Kg
KEX en la capa superior del suelo (0-20 cm)	menos de 30 mmol/Kg
Densidad de árboles	hasta 5000 plantas/ha

En consecuencia, podemos decir que el modelo agronómico trabajado a lo largo del presente capítulo esta calibrado y validado correctamente únicamente cuando se satisfacen todas y cada una de las suposiciones mencionadas previamente, así como también, que los rangos de las variables utilizadas por el modelo están dentro de lo establecido en la Tabla 5. Una vez se cumplen dichas condiciones, procedemos a la construcción de un conjunto de datos que sirva de entrenamiento para un modelo de aprendizaje automático, tema que será abordado en el próximo capítulo.

3.5. Conclusiones acerca de la calibración del modelo agronómico QUEFTS

Este capítulo expone el proceso de calibración del modelo agronómico QUEFTS a la región de estudio. Como primera medida, se llevamos a cabo una revisión de la literatura con el fin de obtener los datos necesarios para el desarrollo del capítulo.

Por otra parte, se hizo la parametrización y posterior análisis de sensibilidad a la región estudiada, por último, fue realizada la validación del modelo y establecieron unas suposiciones para el correcto funcionamiento del mismo. En este sentido, tomando como base el análisis y los resultados presentados en el presente capítulo, se concluye:

- Aun cuándo contamos con acceso a una cantidad limitada de registros para la calibración, el modelo agronómico QUEFTS demostró ser capaz de estimar con un error MAE relativamente bajo (170.1 Kg/ha) el rendimiento de los cultivos de café en suelos caucanos

- Fue corroborada la importancia de la densidad de siembra en el rendimiento de un cultivo de café, ya que el modelo se probó para distintas densidades desde 2500 hasta 7800 plantas por hectárea, pero obtuvo los mejores resultados con estudios de hasta 5000 plantas por hectárea, por tal motivo, el modelo estudiado fue calibrado y validado únicamente para cultivos en el departamento del Cauca que cuenten con hasta 5000 plantas/ha.

- El modelo agronómico QUEFTS fue calibrado y validado correctamente a la región del departamento del Cauca, sin embargo, esto se logra únicamente cuándo se cumple con una serie de suposiciones y se manejan ciertos rangos para algunas variables de entrada del modelo, descritos en la sección 3.4.2.

- El modelo agronómico fue validado correctamente en la región de estudio, es decir, la región caucana a través de una serie de procedimientos presentados en este capítulo. Logrando en su mejor versión un error MAE de 170.1 Kh/ha, R^2 de 0.79 y PBIAS de 8.9% en la estimación del rendimiento y un error MAE de 9.11 Kg/ha, con un R^2 de 0.80 en la estimación de la disponibilidad de N

4. Modelo de datos para estimar la disponibilidad de nitrógeno en suelos establecidos con cultivo de café

Este capítulo presenta el proceso llevado a cabo para la obtención del conjunto de datos sobre disponibilidad de nitrógeno en suelos establecidos con cultivo de café, que luego fue utilizado como base del modelo de predicción del rendimiento del mencionado cultivo. Tanto la construcción del conjunto de datos, como el entrenamiento del modelo, fueron realizados siguiendo el marco de trabajo de CRISP-DM (Cross Industry Standard Process for Data Mining) [59]. La Figura 16 muestra las fases propuestas por dicha metodología que fueron implementadas en la presente investigación.

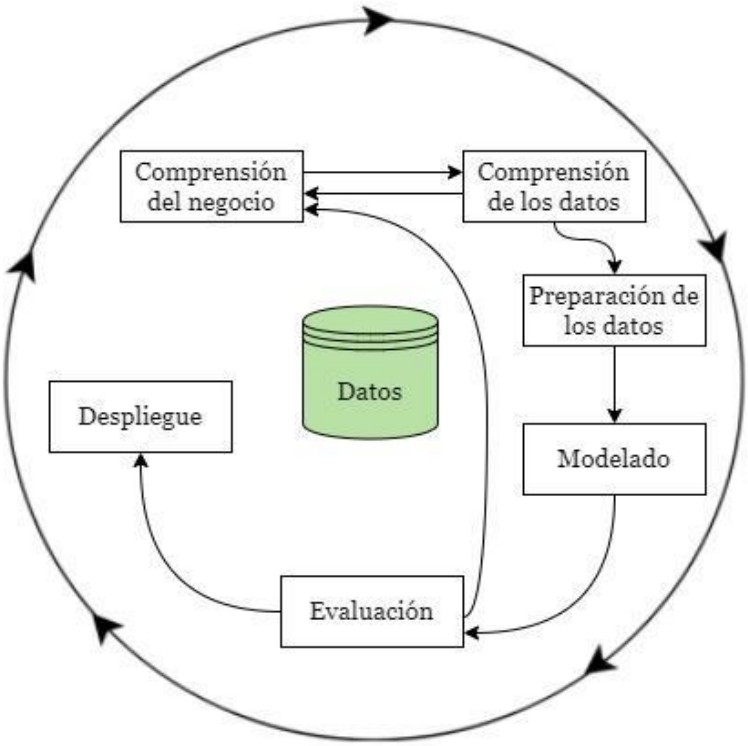


Figura 16: Fases propuestas por el modelo CRISP-DM. Adaptado de [60].

El modelo CRISP-DM ofrece un resumen del ciclo de vida de un proyecto de minería de datos. Este contiene las fases del proyecto, así como sus respectivas tareas y resultados. El ciclo de vida es desglosado en seis fases interconectadas entre sí por flechas que expresan las relaciones más relevantes y frecuentes: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación

y despliegue. Las tres primeras fases fueron la base para la consecución del conjunto de datos sobre disponibilidad de Nitrógeno en suelos establecidos con cultivo de café. Teniendo esto en cuenta, el capítulo está dividido en los siguientes apartados:

- ➔ *Comprensión del negocio:* expone los objetivos que se buscan con la construcción del conjunto de datos y el entrenamiento del modelo, para tener claridad de los registros esperados. Como parte de esto, fue estudiada la importancia del Nitrógeno en un cultivo agrícola en general y el del café, así como las variables que inciden en su disponibilidad.

- ➔ *Comprensión de los datos:* se presenta el proceso de recolección de datos iniciales y la descripción de estos.

- ➔ *Preparación de los datos:* expone las fases de selección, estructuración, limpieza e integración de los registros recolectados mediante el proceso mencionado en la comprensión de los datos.

4.1. Comprensión del negocio

Buscamos construir un conjunto de datos que describa la disponibilidad de N en un cultivo de café, con el objetivo de ser utilizado luego como insumo principal para el entrenamiento de un modelo que permita estimar dicha variable. Es por esto que, en la presente investigación, es necesario entender cómo afecta la disponibilidad de N al cultivo de café, así como las variables que pueden incidir en sus reservas, ya sea de una forma positiva o negativa.

El N es el nutriente más limitante para la producción de café en Colombia. Un suministro adecuado es especialmente importante durante el período vegetativo de crecimiento, ya que es conocido que cualquier deficiencia durante este período tiene efectos duraderos en la producción posterior de granos de café. Su importancia es tan grande que puede producir pérdidas de hasta 60% en el rendimiento del cultivo cuándo no es aplicado ningún fertilizante que contenga N [61].

Por otra parte, el N es un nutriente esencial, que le da color verde a las plantas porque es parte de la clorofila y regula el crecimiento vegetativo de la planta. Cuando el nitrógeno es deficiente o hay escasez temporal, las hojas contendrán relativamente poca clorofila, se volverán de color grisáceo o amarillento en lugar de verde oscuro saludable y pueden tener un crecimiento anormal, con posibles consecuencias negativas en el rendimiento del cultivo [13].

La disponibilidad de N suele medirse en kilogramo por hectárea [kg x ha^{-1}]. En Colombia, las recomendaciones de la FNC piden implementar las dosis de N a los 2, 6, 10, 14 y 18 meses después del trasplante, dependiendo del contenido de materia orgánica del suelo y la disponibilidad de agua. La urea es la fuente más común por su alto contenido de N (46%) y bajo precio por unidad de N. La cantidad total de N aplicada durante esta etapa varía de 100 a 125 gramos de urea por planta, lo que equivale a tasas de hasta 600 kg x ha^{-1} para plantaciones con densidades de hasta 10,000 plantas ha^{-1} [61]**.**

La disponibilidad de N en el suelo depende de diversos factores, que van más allá de la fertilización. Por tal motivo, es necesario entender el ciclo del N y su fijación biológica. La disponibilidad de elementos en los suelos, debemos considerar desde los diferentes compartimientos orgánicos y minerales y su interacción con la biomasa microbiana. Puesto que, la atmósfera es el reservorio más grande de nitrógeno con un 79% en su forma elemental (N_2), sin embargo, metabólicamente no está disponible para su asimilación por las plantas superiores que no poseen mecanismos para procesarlo. El nitrógeno en forma gaseosa es la principal fuente primaria de entrada para los ecosistemas, existen organismos simbiontes y de vida libre con la capacidad de fijarlo de la atmósfera a formas que sean aprovechables por las plantas, es decir, en amonio (NH_4^+) y en nitrato (NO_3^-). Esto mediante procesos como la nitrificación y la fijación de nitrógeno, que son complejos y envuelven una variedad de actividad microbiana, plantas y animales, lo que representa una entrada al ciclo terrestre del nitrógeno, de gran importancia para los ecosistemas [62]. Las comunidades microbianas poseen un papel principal, ya que de ellas dependen funciones como hacer disponibles los nutrientes para ellas mismas y para otras formas de vida como las plantas, dinámica esencial para el mantenimiento de los ciclos biogeoquímicos [63].

Existen varios factores inherentes que tienen influencia en la actividad microbiana y por ende en la disponibilidad de N en los suelos: los factores naturales: cambios climáticos (lluvia y temperatura), las condiciones del sitio, como la humedad, la aireación del suelo (niveles de oxígeno), el contenido de sal, la profundidad, la inclinación de la pendiente, niveles de materia orgánica, entre otras, así mismo, las pérdidas normales del ciclo del nitrógeno por lixiviación, escorrentía o desnitrificación y los factores antropogénicos, es decir, contaminación y manejo agrícola. Han sido empleadas diversas técnicas para estimar la actividad microbiana. No obstante, su implementación es dispendiosa, sólo estima una parte de la comunidad (<1%) produciendo información limitada y que no da cuenta de la potencial diversidad funcional (número y distribución de funciones) y actividad de las comunidades [63], [64]**.**

Los fertilizantes juegan un rol importante en la preservación del N, ya que cuándo son aplicados a tiempo, es decir cuándo la planta está en crecimiento y en floración para el caso del café, pueden ayudar al mantenimiento y/o crecimiento de la flora

microbiana, y por ende en la recuperación de N por parte de los cultivos, evitar la pérdida de N por volatilización y adaptarse al método de fertilización para reducir las pérdidas y maximizar la recuperación por parte de los cultivos, lo que puede resultar en un posible aumento en la eficiencia del uso de N por parte del cultivo y en consecuencia evitar posibles sobrecostos, daños ambientales u otros problemas asociados al mal uso del N [63], [64].

4.2. Comprensión de los datos

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

4.2.1. Análisis de fuentes de datos

Como primera medida fue realizado un análisis de la principal fuente de datos para esta investigación, es decir, el modelo agronómico QUEFTS previamente calibrado a la región de estudio en el capítulo 3. Procedimos a entender a profundidad los datos de entrada y salida del modelo. En las tablas 6 y 7 podemos observar los datos involucrados durante el proceso de modelamiento y las salidas obtenidas durante el mencionado proceso.

Tabla 6. Variables de entrada para el modelo QUEFTS, Fuente propia

Atributo	Descripción	Tipo	Unidad
PH	PH medido en el suelo	Numérico	pH
SOC	Carbono orgánico en el suelo	Numérico	g/Kg
Kex	Potasio intercambiable	Numérico	mmol/ Kg
PBray	Fósforo medido con el método de Bray	Numérico	mg/Kg
NrTrees	Número de plantas por hectárea	Numérico	plantas/ha^{-1}
Crop	Tipo de cultivo	Nominal	-
NminStore	Concentración mínima de N en órganos de almacenamiento	Numérico	Kg/Kg

NmaxStore	Concentración máxima de N en órganos de almacenamiento	Numérico	Kg/Kg
NminVeg	Concentración mínima de N en órganos vegetativos	Numérico	Kg/Kg
NmaxVeg	Concentración máxima de N en órganos vegetativos	Numérico	Kg/Kg
PminStore	Concentración mínima de P en órganos de almacenamiento	Numérico	Kg/Kg
PmaxStore	Concentración máxima de P en órganos de almacenamiento	Numérico	Kg/Kg
PminVeg	Concentración mínima de P en órganos vegetativos	Numérico	Kg/Kg
PmaxVeg	Concentración máxima de P en órganos vegetativos	Numérico	Kg/Kg
KminStore	Concentración mínima de K en órganos de almacenamiento	Numérico	Kg/Kg
KmaxStore	Concentración máxima de K en órganos de almacenamiento	Numérico	Kg/Kg
KminVeg	Concentración mínima de K en órganos vegetativos	Numérico	Kg/Kg
KmaxVeg	Concentración máxima de K en órganos vegetativos	Numérico	Kg/Kg
SeasonLength	Duración de la temporada	Numérico	días
Leaf_att	Biomasa de cultivo alcanzable(en ausencia de limitación de nutrientes) para hojas	Numérico	Kg/ha^{-1}
Stem_att	Biomasa de cultivo alcanzable(en ausencia de limitación de nutrientes) para tallos	Numérico	Kg/ha^{-1}
Store_att	Biomasa de cultivo alcanzable(en ausencia de limitación de nutrientes) para órganos de almacenamiento (Rendimiento ideal)	Numérico	Kg/ha^{-1}
New_Yzero	Biomasa máxima de órganos	Numérico	Kg/ha^{-1}

	vegetativos teniendo el rendimiento en cero de órganos de almacenamiento		
N	Fertilización con N aplicada al cultivo durante toda la temporada (siembra-cosecha)	Numérico	Kg/ha^{-1}
P	Fertilización con P aplicada al cultivo durante toda la temporada (siembra-cosecha)	Numérico	Kg/ha^{-1}
K	Fertilización con P aplicada al cultivo durante toda la temporada (siembra-cosecha)	Numérico	Kg/ha^{-1}

Tabla 7. Variables de salida del modelo QUEFTS, Fuente propia

Atributo	Descripción	Tipo	Unidad
N_uptake	Absorción de N por parte del cultivo	Numérico	Kg/ha^{-1}
P_uptake	Absorción de P por parte del cultivo	Numérico	Kg/ha^{-1}
K_uptake	Absorción de K por parte del cultivo	Numérico	Kg/ha^{-1}
N_supply	Disponibilidad de N del suelo (con fertilización)	Numérico	Kg/ha^{-1}
P_supply	Disponibilidad de P del suelo (con fertilización)	Numérico	Kg/ha^{-1}
K_supply	Disponibilidad de K del suelo (con fertilización)	Numérico	Kg/ha^{-1}
leaf_lim	Biomasa limitada de nutrientes en las hojas	Numérico	Kg/ha^{-1}
stem_lim	Biomasa limitada de nutrientes en los tallos	Numérico	Kg/ha^{-1}
store_lim	Biomasa limitada de nutrientes en los órganos de almacenamiento (Rendimiento)	Numérico	Kg/ha^{-1}
N_gap	N necesario para alcanzar la biomasa especificada	Numérico	Kg/ha^{-1}

P_gap	P necesario para alcanzar la biomasa especificada	Numérico	Kg/ha^{-1}
K_gap	K necesario para alcanzar la biomasa especificada	Numérico	Kg/ha^{-1}

El atributo **N_supply** que se encuentra en la tabla 7, es la clase objetivo, que buscamos predecir en la presente investigación, ya que corresponde a la disponibilidad de N en el suelo, calculada después de hallar el suministro del suelo autóctono y añadirle el N resultante de la fertilización durante la temporada de cosecha. Para ese objetivo, contamos con un conjunto de datos explicado en las tablas 6 y 7, que es la base para el entrenamiento del modelo que permita predecir la disponibilidad de N en suelos con cultivo de café en el departamento del Cauca en Colombia, que es presentado en el siguiente capítulo.

4.2.2. Recolección de datos

La recolección de los datos fue llevada a cabo en dos etapas. La primera etapa fue el entendimiento de la región caucana con la ayuda de expertos en el tema, es necesario entender los distintos tipos de suelos que pueden ser encontrados a lo largo de la región, y como varían en sus cantidades de pH, carbono orgánico, potasio intercambiable, entre otras propiedades del suelo. En consecuencia, en la segunda etapa procedemos a programar el modelo agronómico proveniente del capítulo 3 para generar los datos simulados con los cuáles el modelo de inteligencia artificial se va a entrenar. Para el desarrollo de este apartado se llevará a cabo el proceso presentado en la Figura 17.

Figura 17: Fases propuestas para la recolección de datos. Fuente propia

Como está expuesto en el proceso de la Figura 16, en primera instancia es realizada la fase de entendimiento de propiedades del suelo en la región de estudio. Proceso en el cuál determinamos en qué proporción serán variadas las propiedades del suelo (pH, SOC, entre otras) según lo que usualmente encontramos en suelos ubicados en el departamento del Cauca. Este paso es clave, ya que es necesario simular los posibles escenarios que se encuentran en el departamento del Cauca, así como también las posibles fertilizaciones que comúnmente son aplicadas con

los nutrientes N, P y K. De esta manera se busca lograr un modelado más robusto, para dicho procedimiento contamos con el asesoramiento de expertos en el tema que cuentan con el conocimiento de campo necesario para determinar qué rangos de estos atributos son ideales para el desarrollo del proyecto, y, en consecuencia, dichos rangos representen de la forma más acertada posible toda la variedad de suelos encontrados en el departamento. Dicha información esta resumida en la Tabla 8

Tabla 8. Propiedades del suelo para el modelado con QUEFTS, Fuente propia

Propiedad	Rango de variabilidad	Tasa de cambio
pH	4.7 - 5.9	0.1
SOC	9 - 35	1
Kex	0.1 - 1.8	0.1
P Bray	2 - 1.5	0.1
Nr Trees	2000 - 5000	10
N	150 - 500	10
P	10 - 150	10
K	50 - 350	10

La siguiente fase consistió en la construcción de un algoritmo en R, el cual varia los atributos de entrada del modelo agronómico QUEFTS según la información presentada en la Tabla 8. El mencionado algoritmo consistió en escoger aleatoriamente un valor para cada una de las propiedades dentro del rango preestablecido, para después ejecutar el modelo agronómico previamente calibrado y, por último, registrar todas las variables de salida (véase en la Tabla 6) dentro de un archivo csv. El algoritmo es el encargado de repetir una y otra vez el proceso descrito anteriormente hasta completar un total de 15000 registros almacenados en el csv, dicho archivo servirá como base para el resto de la investigación, es decir, servirá como datos de entrenamiento para el modelo de aprendizaje automático para predecir la disponibilidad de N en suelos cafeteros, abordado en el siguiente capítulo. Además del conjunto de datos de entrenamiento mencionado anteriormente, generamos un archivo csv de la misma forma y cumpliendo con las mismas condiciones, pero esta vez con 1500 registros, el 10% del total de registros anteriores. Esto con el fin de generar un conjunto de datos de prueba (test), tema que será abordado y explicado a profundidad en el próximo capítulo.

4.3. Preparación de los datos

La fase de preparación de los datos está compuesta por varias tareas, entre las que encontramos: la selección de los datos que serán utilizados para el modelamiento en el siguiente capítulo, así como la verificación de su calidad y limpieza.

El primer paso en este apartado es la selección de los datos que serán utilizados para el modelamiento con aprendizaje automático, para esto filtramos los atributos que serán seleccionados del conjunto de datos resultante del capítulo 3, por consiguiente, procedimos a eliminar todas las variables de salida del modelo agronómico, expuestas en la Tabla 7, con la excepción de la clase que se busca predecir, es decir **N_Supply**, esto fue realizado porque las salidas del modelo agronómico podrían estar relacionadas linealmente con la clase objetivo, conformando una relación que buscamos evitar, ya que podría traer problemas al rendimiento, además porque el objetivo es utilizar el modelo de aprendizaje automático directamente, prescindiendo del uso del modelo agronómico, por tal motivo las salidas del mismo no deben ser tomadas en cuenta para la fase de entrenamiento.

En consecuencia, los datos que conforman el conjunto de datos de entrenamiento son los que conforman la entrada del modelo agronómico, descritos en la Tabla 6, además de la clase objetivo **N_Supply** (véase Tabla 7). Conformando un conjunto de datos con 32 atributos, incluida la clase que se busca predecir y los otros son los atributos que permiten que la predicción sea posible.

En un segundo paso, llevamos a cabo un diagnóstico de la calidad de los datos [65], además de realizar una limpieza de los mismos. En la Figura 18 presentamos las tareas llevadas a cabo para este paso.

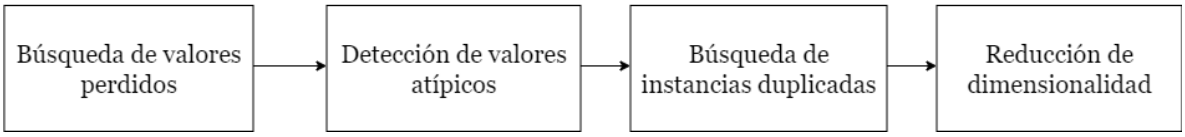


Figura 18: Proceso de limpieza de datos. Adaptado de [66]

- ➔ *Búsqueda de valores perdidos:* espacios en blanco, valores como “NaN” o “null” y caracteres especiales como “*” o “?” fueron buscados para verificar que no existiesen valores perdidos en los conjuntos de datos. Al realizar la búsqueda, no se encontraron valores perdidos en el conjunto de datos generado

- ➔ *Detección de valores atípicos:* para detectar si existían valores que se encontraran desviados significativamente de los demás, para esto se utilizaron dos algoritmos, DBSCAN (Density-based spatial clustering of applications with noise) y LOF (Local Outlier Factor). Los valores atípicos y extremos encontrados con los dos algoritmos mencionados previamente fueron eliminados del conjunto de datos

- ➔ *Búsqueda de instancias duplicadas:* se utilizó un filtro para detectar si existían casos con esta anomalía. El filtro no encontró instancias con dicho problema, dado que el conjunto de datos fue construido aleatoriamente, reduciendo así este problema

- ➔ *Reducción de dimensionalidad:* para cumplir este objetivo se analizaron los atributos que mayor grado de correlación tenían con la clase objetivo. Una mayor información acerca de este proceso de selección de atributos, se expone en la sección 5.2.3.

Luego de los procesos mencionados anteriormente, se obtuvo como resultado un conjunto de datos procesados que agrupa 32 atributos entre los que se encuentran propiedades del suelo tales como pH, SOC, biomasas alcanzables en condiciones ideales (ausencia de limitación de nutrientes) para hojas, tallos y órgano de almacenamiento, número de árboles, concentraciones mínimas y máximas para acumulación y dilución de cada uno de los nutrientes, así como variables de manejo del cultivo, tales como la fertilización utilizada con los nutrientes N,P y K. Dicho conjunto de datos es la base para el entrenamiento de un modelo de aprendizaje automático que permita predecir la disponibilidad de N en suelos establecidos con cultivo de café en la región caucana, desarrollado en la siguiente fase de la metodología propuesta.

4.4. Modelado

Posteriormente, procedemos a ejecutar la fase de modelado donde se realiza la construcción de un modelo de aprendizaje automático que permita la estimación de la disponibilidad de N en suelos establecidos con cultivo de café. A continuación, en la Figura 19 se exponen las tres actividades realizadas como parte del proceso de modelamiento:

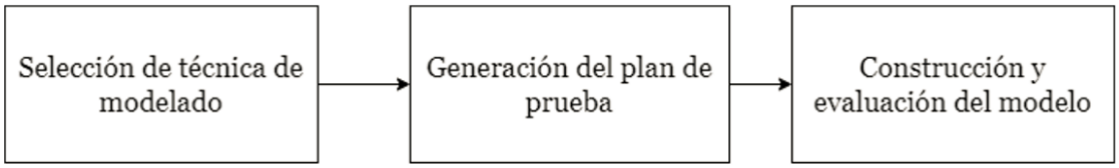


Figura 19: Etapas que conforman la fase de modelado.

- **Selección de la técnica de modelado:** en esta primera etapa se analizaron diversos algoritmos de aprendizaje automático para encontrar el que presente el mejor comportamiento con base en el conjunto de datos expuesto en el capítulo 1. Tomando como punto de partida diferentes investigaciones en las que se realizó una evaluación de técnicas de aprendizaje automático, en el presente proyecto fueron analizados distintos algoritmos entre los que encuentramos: linear regression, decision tree, support vector machine for regression, perceptrón multicapa y random forest.

- **Generación del plan de pruebas:** son descritos los parámetros de desempeño que permiten validar y entender los resultados obtenidos a través del entrenamiento del modelo de aprendizaje automático.

- **Construcción y evaluación del modelo:** son expuestas todas las pruebas realizadas durante el proceso de entrenamiento del modelo para la estimación de la disponibilidad de N en suelos establecidos con cultivo de café.

La selección de la técnica de modelado y la selección de atributos fueron realizadas utilizando la versión 3.8.4 del software Weka (Waikato Environment for Knowledge Analysis), programa que provee un conjunto de algoritmos implementados y herramientas para la consecución de proyectos de aprendizaje automático, incluyendo tanto la preparación de los datos de entrada como el entrenamiento y la evaluación de los modelos. Dicho software incluye diversos algoritmos para tareas de regresión, clasificación, agrupamiento, asociación y selección de atributos que, combinados con técnicas de análisis estadístico y visualización de resultados, permiten construir modelos basados en aprendizaje automático y el seguimiento de metodologías como CRISP-DM [67]. Sin embargo, la construcción y evaluación del modelo fueron realizadas en el lenguaje Python sobre el ambiente Anaconda, que ofrece una mayor robustez y parametrización del modelo, es decir, la posibilidad de probar con diferentes valores los hiperparámetros del modelo y en consecuencia, obtener los mejores resultados posibles, así como también ofrece la posibilidad de desplegar el modelo en un ambiente empresarial productivo, razón que podría contemplarse para realizar trabajos futuros de la presente investigación. Además, Python es el lenguaje de programación más utilizado en el campo de la ciencia de datos y uno de los de mayor crecimiento en los últimos años [68], por ende, es de resaltar la importancia del uso de dicho lenguaje en la investigación.

4.5. Selección de la técnica de modelado

Con el objetivo de seleccionar la técnica adecuada para generar un modelo que permita la estimación de la disponibilidad de N en suelos establecidos con cultivo de café, se tomaron como punto de partida los resultados obtenidos a partir de la generación del estado actual del conocimiento expuesta en el capítulo 2, en dónde se analizaron las integraciones entre algoritmos de aprendizaje automático y datos generados por modelos agronómicos.

Por una parte, un primer grupo de investigaciones, presentadas en [11], [41], [42], [46], [47], lograron comprobar la eficacia del uso de un enfoque híbrido conformado por modelos agronómicos y algoritmos de inteligencia artificial en la predicción de rendimiento, pérdidas de N, niveles diarios de agua, entre otros. haciendo uso de distintos algoritmos cómo el modelo de regresión de vectores de soporte(SVR), RF, regresión lineal múltiple (MLR), Extreme Gradient Boosting (XGBoost). Igualmente, un segundo grupo de investigaciones, expuestas en [45], [48], expusieron un enfoque híbrido distinto, conformado por modelos agronómicos y algoritmos de aprendizaje profundo, es decir redes neuronales. Fue utilizado para predecir los impactos de la cantidad de riego y el tiempo de aplicación en el rendimiento del cultivo y el índice de desarrollo del arroz con una alta efectividad, los algoritmos que se utilizaron fueron redes convolucionales (CNN) y multilayer perceptron (MLP).

Con base en las publicaciones mencionadas anteriormente, fueron elegidas las siguientes técnicas de aprendizaje supervisado: multilayer perceptron, support vector machine for regression(SVR) y random forest. Que serán evaluadas en su configuración por defecto mediante el software mencionado previamente WEKA. Sin embargo, debemos tener en cuenta que el comportamiento de una u otra técnica de aprendizaje automático al intentar modelar una variable objetivo depende del dominio de aplicación y los tipos de datos, en ese orden de ideas, evaluamos la precisión de los algoritmos de aprendizaje automático mediante el entrenamiento de un modelo basado en el conjunto de datos presentado en el capítulo 4 mediante dicho software, y así definir la técnica que muestre un mejor rendimiento para el conjunto de datos específico. Para ello, fue analizada la correlación y los errores de las diferentes técnicas al predecir la disponibilidad de N en suelos establecidos con cultivo de café. En ese orden de ideas, la tabla 9 muestra tres parámetros (descritos en la sección 5.2.2.) el coeficiente de correlación, el error medio absoluto y el error cuadrático medio, resultados del entrenamiento de los modelos con las técnicas mencionadas previamente provenientes del estado del arte, así como también de otras que mostraron una mejor adaptación a los datos y por ende unos mejores resultados.

Tabla 9. Métricas de evaluación de técnicas de aprendizaje automático en WEKA obtenidas mediante validación cruzada. Fuente propia

Algoritmo/Métricas	Mean Absolute Error (MAE)	Coeficiente de correlación	Root Mean Squared Error (RMSE)
Random Forest Regressor	20.50	0.9828	26.56
Bagging Regressor	33.94	0.93	43.78
Decision Tree Regressor	10.45	0.99	13.44
SVR	7.6	0.99	12.68
Multi layer Perceptron (MLP)	0.54	0.99	0.62

Los resultados encontrados en la tabla 9, muestran algunas métricas de las 5 técnicas de aprendizaje automático que mostraron un mejor desempeño en el software WEKA, además, se puede evidenciar que el algoritmo con el mejor rendimiento es el Multilayer Perceptron (MLP) con un coeficiente de correlación (CC) de 0.99, un error MAE de 0.54 y un RMSE de 0.62, seguido del SVR que mostró un CC de 0.99 con errores MAE de 7.6 y RMSE de 12.58, otro algoritmo que mostró un buen rendimiento fue el Decision Tree Regressor que alcanzó un cc de 0.99, con errores MAE de 10.45 y RMSE de 13.44, en última instancia se encuentran los algoritmos Bagging Regressor y RF, que muestran unos errores considerablemente más grandes. En consecuencia, tomamos la decisión de elegir el MLP cómo el algoritmo a utilizar en la presente investigación ya que muestra el menor porcentaje de error y mayor coeficiente de correlación, tema que será tratado con más detalle en la sección 5.3. del presente capítulo.

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

4.6. Generación del plan de pruebas

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

4.6.1. Validación cruzada

Existen varios enfoques que permiten entrenar y evaluar un modelo de regresión. El primero de ellos consiste en realizar los dos procesos mencionados utilizando el mismo conjunto de datos. En este, como primera medida, se utilizan todas las instancias para el entrenamiento y construcción del modelo; luego, se elige una pequeña porción de ellas como el conjunto de pruebas. Las instancias de dicho conjunto, sin sus respectivas clases objetivo, son utilizadas para predecir nuevos valores mediante el modelo ya construido. Finalmente, comparamos los valores predichos por el modelo con los valores reales en el conjunto de prueba y así conocer la precisión del mismo. Sin embargo, este enfoque presenta una limitación; debido a que el modelo es evaluado utilizando una porción del mismo conjunto de datos con el que fue construido, es probable que se obtenga una alta precisión con el conjunto de entrenamiento, pero no una alta precisión en conjuntos por fuera de la muestra [69].

La precisión de entrenamiento es el porcentaje de predicciones correctas que hace el modelo cuando utilizamos el conjunto de pruebas para evaluarlo. No obstante, registrar una alta precisión de entrenamiento no es necesariamente algo positivo, ya que puede resultar en un ajuste excesivo de los datos, teniendo así un modelo altamente capacitado para el conjunto de datos, pero no para nuevos registros, es

decir, un modelo no generalizado. Por el contrario, la precisión fuera de la muestra es el porcentaje de predicciones correctas que el modelo hace a partir de datos desconocidos; un buen modelo de aprendizaje automático debe registrar entonces una alta precisión fuera de la muestra [69], [70].

Un segundo enfoque, que permite mejorar la precisión fuera de la muestra, consiste en dividir el conjunto de datos en dos diferentes, uno de entrenamiento y uno de prueba, mutuamente excluyentes. Esto proporciona una evaluación más precisa, ya que el conjunto de datos con el que se realiza la evaluación no hace parte del entrenamiento del modelo, siendo así una solución más realista. Sin embargo, al dividir el conjunto de datos en dos, los valores que pertenecen al conjunto de prueba ahora no harán parte del modelamiento, perdiéndose así datos valiosos que podrían mejorar la precisión del mismo [71]

Buscando utilizar todos los datos tanto en el entrenamiento como la evaluación y obtener la mayor precisión fuera de la muestra, aparece un tercer enfoque, llamado validación cruzada (cross validation). Este enfoque consiste en dividir el conjunto de datos en k subconjuntos para luego tomar uno de ellos como el conjunto de pruebas, mientras que los restantes k-1 son convertidos en el conjunto de entrenamiento. Esta tarea es repetida k veces, durante las cuales cada uno de los subconjuntos es utilizado una única vez como conjunto de prueba. Finalmente, son promediados todos los resultados para generar así el modelo final [72]. En el caso de la presente investigación, es utilizado un k = 10, es decir, dividimos el conjunto de datos en 10 subconjuntos, donde cada uno de ellos es utilizado una vez como conjunto de prueba y el restante 90% de los datos como conjunto de entrenamiento.

4.6.2. Métricas de evaluación del modelo

Para determinar la precisión del modelo de regresión entrenado con el objetivo de predecir el rendimiento del cultivo de café, se deben evaluar los datos estimados frente a los reales mediante el cálculo de diferentes métricas estadísticas. A continuación, describimos las métricas utilizadas para la evaluación del desempeño del modelo expuesto en la sección 5.3.

- ➔ *Coeficiente de Correlación (CC):* es una de las métricas estadísticas más utilizadas para la evaluación de diferentes resultados de investigaciones. El CC permite medir qué tan fuerte es la relación existente entre dos o más variables [73]. En el caso de la inteligencia artificial, el cálculo del CC proporciona el grado de relación entre los datos reales y los generados por el modelo. Esta métrica se define mediante la siguiente ecuación:

$$CC = \frac{cov(r,e)}{\sigma r\ \sigma e} = \frac{\sum_{i=1}^{n}[(ri-\bar{r})(ei-\bar{e})]}{\sqrt{\sum_{i=1}^{n}(ri-\bar{r})^2\ \sum_{i=1}^{n}(ei-\bar{e})^2}} \quad (8)$$

Donde $cov(r,e)$ es la covarianza de los datos reales con los estimados, σr y σe se refieren a la desviación estándar de los mismos, n es el número de datos estudiados, ri indica el valor real en la posición i, $\bar{r}$ es la media de los datos reales, ei es el valor estimado en la posición i-ésima y $\bar{e}$ es la media de los datos estimados. Tanto $\bar{r}$ como $\bar{e}$ se pueden calcular mediante las siguientes ecuaciones:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} ri \qquad \bar{e} = \frac{1}{n}\sum_{i=1}^{n} ei \quad (9)$$

El CC puede tomar valores entre -1 y 1, de la siguiente forma: un resultado igual a 1 indica una correlación positiva perfecta, mientras que al ser igual a -1 indica una correlación negativa perfecta. Un CC igual a 0 significa que no existe ninguna relación entre las variables estudiadas [74].

➔ *Error cuadrático medio (RMSE, por sus siglas en inglés):* es una métrica estadística, que al igual que el MAE (abordado en la sección 3.4.1.), también permite estimar la magnitud promedio del error. Elevar al cuadrado los errores antes de promediarlos, otorga un peso relativamente alto a los mayores de ellos, siendo así mucho más útil el uso del RMSE cuando no se desea que el modelo pueda generar errores grandes. Esta métrica se calcula mediante la siguiente ecuación [75]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ri-ei)^2} \quad (10)$$

Donde ri indica el valor real en la posición i, ei es el i-ésimo valor estimado y n es el número de datos estudiados

4.6.3. Selección de atributos

En el campo de la ciencia de datos, es muy común encontrar conjuntos de datos que contienen una gran cantidad de atributos, donde varios de ellos pueden ser irrelevantes, o en el peor de los casos, desfavorables para el proceso de modelamiento. Por consiguiente, llevar a cabo una selección de atributos puede traer importantes mejoras, entre las cuales se encuentran: aumentar la precisión del modelo, reducir tiempo y costo computacional, entre otras. En este orden de ideas, existen dos opciones principales para llevar a cabo una selección de atributos: conocimiento experto o un método tradicional de selección [76].

Considerando lo mencionado anteriormente y para el caso particular de la presente investigación fue necesario del uso de ambos métodos para la selección de atributos, por una parte fue requerido el asesoramiento de dos expertos en caficultura implementada en el departamento del cauca, teniendo en mente lograr modelar con la menor cantidad posible de atributos, y en consecuencia, lograr que el modelo se ejecute de la manera más fácil y económica posible para los caficultores, en otras palabras, dándole prelación a los atributos que sean de fácil obtención para los mismos. Después del proceso antes mencionado, se obtiene un conjunto de datos filtrado para continuar con la investigación y el siguiente paso fue elegir dos métodos tradicionales de selección de atributos para evaluar cuáles de ellos inciden directamente en la variable objetivo, buscando la optimización del modelo, estos fueron implementados mediante el software Weka y son:

- *CfsSubsetEval:* permite elegir un subconjunto de atributos que tengan la mayor capacidad predictiva individual (correlación alta con la clase) y el menor grado de redundancia entre ellos (correlación baja con las demás características) [77].

- *CorrelationAttributeEval:* evalúa la importancia de cada atributo mediante la obtención de la correlación entre él y la clase [78].

4.7. Construcción y evaluación del modelo

El primer paso para la construcción y evaluación del modelo fue la implementación de los modelos con todos los algoritmos mencionados previamente en la sección 3.1, para comprobar si el algoritmo Multilayer Perceptron MLP es el que mejor se adapta a los datos de la presente investigación, como se explicó previamente la construcción y evaluación del modelo va a ser en el ambiente anaconda bajo el lenguaje de programación Python, mientras que la selección del algoritmo fue realizada en WEKA, por tal motivo se procede a evaluar de nuevo todos los modelos, pero esta vez en lenguaje python.

En primera instancia, se van a probar los modelos con todos los atributos, en otras palabras, se evalúa el rendimiento de los distintos algoritmos de aprendizaje automático utilizando la configuración por defecto en la predicción de la disponibilidad de N haciendo uso de todos los atributos del conjunto de datos de entrenamiento, mismo conjunto de datos con el que se realizaron los cálculos en el software WEKA en la sección 5.1. Durante estos experimentos fueron utilizados un total de 31 atributos y un conjunto de datos con 15000 instancias para entrenar el modelo, utilizando validación cruzada con k = 10, es decir, se divide el conjunto de datos en 10 subconjuntos, donde cada uno de ellos se utiliza una vez como conjunto de prueba y el restante 90% de los datos como conjunto de entrenamiento, tal cómo se explicó en la sección 5.2.2. En ese orden de ideas, las Figuras 20-24 reflejan el comportamiento de cada uno de los algoritmos de aprendizaje automático escogidos para el desarrollo del presente trabajo frente a los datos de entrada provenientes del capítulo 4, por lo tanto, dichos datos serán tratados como datos de entrenamiento en lo que resta del capítulo.

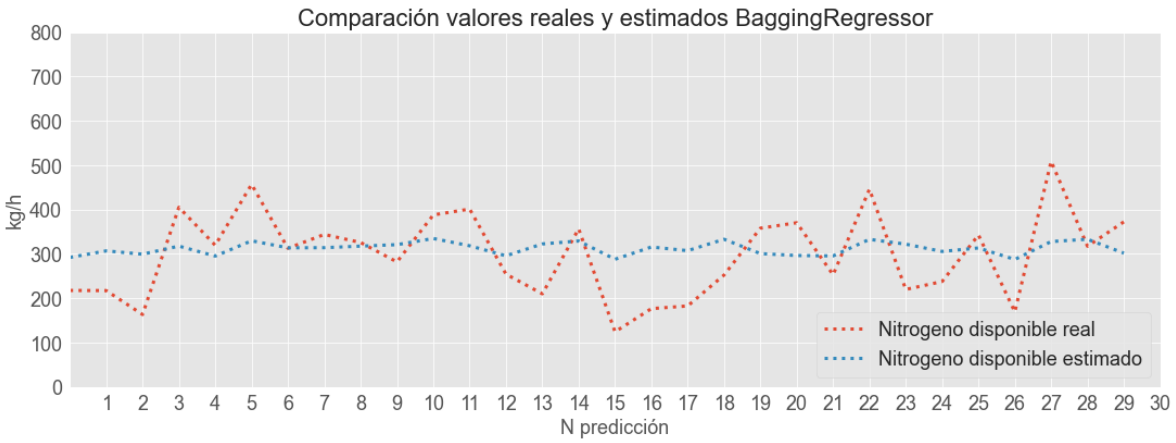


Figura 20. Datos de entrenamiento vs estimados de N con algoritmo BaggingRegressor.

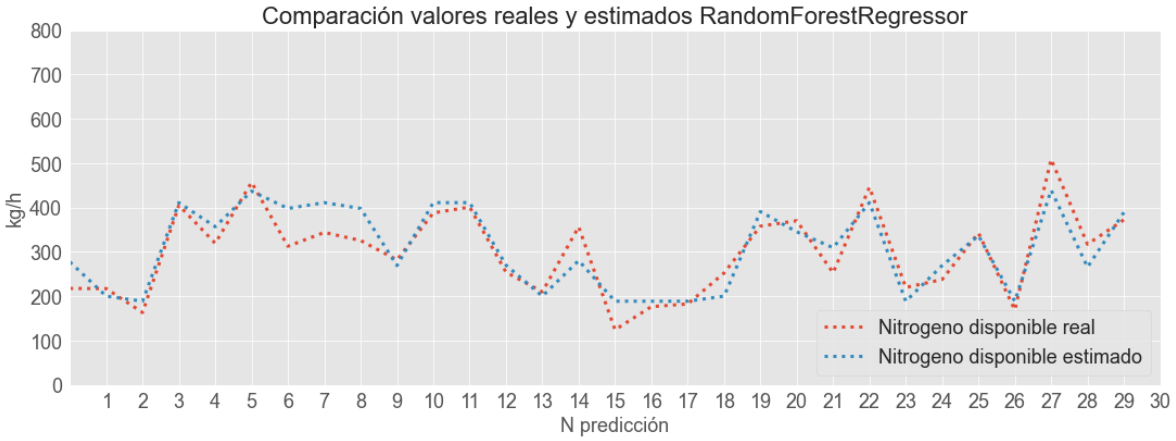


Figura 21. Datos de entrenamiento vs estimados de N con algoritmo RF.

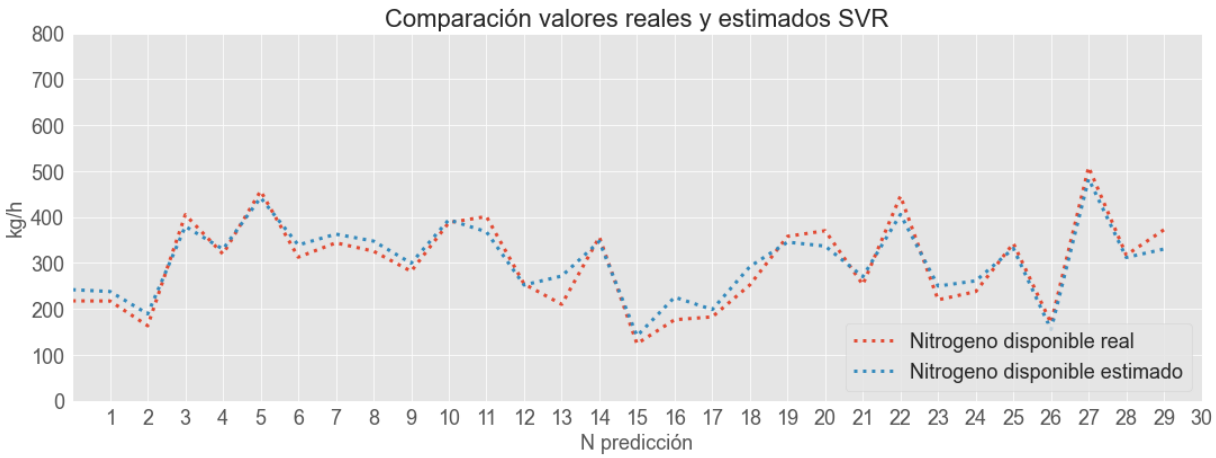


Figura 22. Datos de entrenamiento vs estimados de N con algoritmo SVR

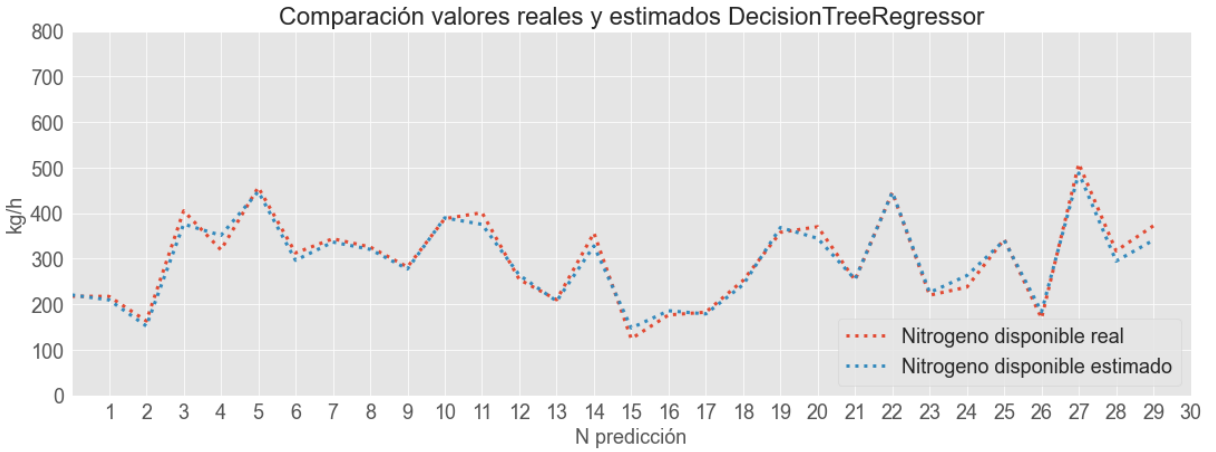


Figura 23. Datos de entrenamiento vs estimados de N con algoritmo DecisionTreeRegressor.

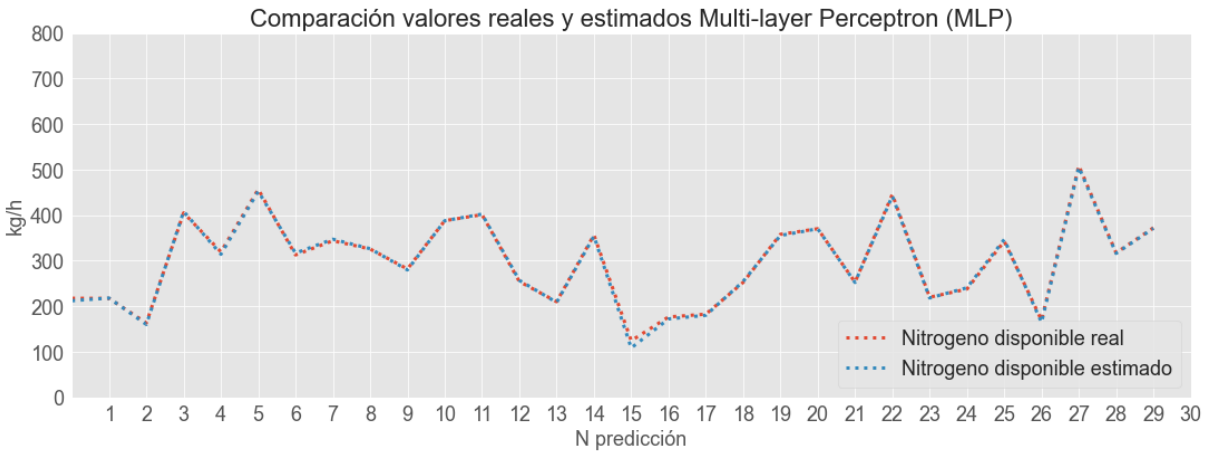


Figura 24. Datos de entrenamiento vs estimados de N con algoritmo MLP.

En las Figuras 20-24 se pueden observar 30 registros de entrenamiento sobre disponibilidad de N en suelos con cultivo de café y las predicciones generadas con los algoritmos de aprendizaje automático: BaggingRegressor, RF, SVR, DecisionTreeRegressor y MLP. En un primer acercamiento se obtuvieron resultados poco esperados que arrancan desde un RMSE de 100.5, 68.54% de precisión y un R^2 de 0.2 por el primer algoritmo BaggingRegressor. Mostrando alguna mejora se encuentra el algoritmo RF que logró un RMSE de 49.53, 86.42% de precisión y un R^2 de 0.8. Por otro lado, el algoritmo SVR, mostró unos resultados relativamente buenos con un 15.32 de RMSE, 96.36 de precisión y un R^2 de 0.98. Sin embargo, con un mejor rendimiento se encuentra el algoritmo DecisionTreeRegressor con un RMSE de 8.7, una precisión del 97.76% y un R^2 de 0.99. Por último, y logrando el mejor resultado, el algoritmo MLP que alcanzó un 2.97 de RMSE, 99.22% de precisión y un R^2 de 0.99. Demostrando que puede replicar los datos de una manera muy acertada, con errores muy bajos. Es importante tener en cuenta que los modelos construidos en Python podrían tener resultados ligeramente diferentes a los encontrados previamente con el software Weka en la sección 5.1, esto podría deberse a que los modelos son construidos por cada herramienta manejando sus propios hiper parámetros, lo que resulta en modelos distintos, aunque manejen el mismo algoritmo. Las métricas de evaluación para los modelos observados en las Figuras ya mencionadas se encuentran en la tabla 10:

Tabla 10: Métricas de evaluación de modelos con 31 atributos. Fuente propia

Algoritmo/Métricas	**Random Forest Regressor**	**Multi layer Perceptron (MLP)**	**Bagging Regressor**	**Decision Tree Regressor**	**SVR**
Mean Absolute Error (MAE)	39.38	2.18	85.37	6.57	4.49
Mean Squared Error (MSE)	100.5	8.83	10098.9	75.73	35.13
Root Mean Squared Error (RMSE)	49.53	2.97	100.49	8.7	5.92
Mean Absolute Percentage Error (MAPE)	13.58	0.78	31.46	2.24	1.51
Accuracy	86.42	99.22	68.54	97.76	98.49
R^2	0.8	0.99	0.20	0.99	0.99

Analizando los resultados consignados en la Tabla 10, queda comprobado lo mencionado en la sección 3.1. mostrando que el algoritmo MLP en su configuración por defecto fue el mejor y obtuvo mejores métricas de evaluación en prácticamente todos los rubros evaluados en comparación con los demás algoritmos. Sin embargo, el buen desempeño de los modelos y el bajo error presentado podría deberse a un sobre ajuste a los datos de entrenamiento, dado que las evaluaciones han sido realizadas por medio de validación cruzada en el mismo conjunto de datos de entrenamiento, por tal motivo, se busca que el modelo sea lo más robusto posible, y en ese orden de ideas se prueban nuevamente todos los modelos, pero esta vez en un conjunto de datos de prueba o test, el cuál fue generado en la sección 2.2. Se busca evaluar el desempeño de los modelos, pero con datos nuevos, es decir datos con los cuales no hayan sido entrenados, de esta manera se podría lograr un mejor desempeño en el campo de acción. En la tabla 11 se muestran los desempeños de los modelos de aprendizaje automático utilizando las configuraciones por defecto en los datos de prueba.

Tabla 11: Métricas de evaluación de modelos con datos de prueba. Fuente propia

Métricas/Algoritmo	**Random Forest Regressor**	**Multi layer Perceptron (MLP)**	**Bagging Regressor**	**Decision Tree Regressor**	**SVR**
Mean Absolute Error (MAE)	41.75	2.29	79.59	5.23	3.57
Mean Squared Error (MSE)	2705.46	9.62	8872.60	61.93	22.09
Root Mean Squared Error (RMSE)	52.01	3.10	94.19	7.87	4.7
Mean Absolute Percentage Error (MAPE)	15.0	0.87	29.65	1.8	1.21
Accuracy	85.0	99.13	70.35	98.2	98.79
R^2	0.76	0.99	0.23	0.99	0.99

En conclusión, el algoritmo MLP demostró ser el que garantiza mejores resultados tanto con los datos de entrenamiento, como en los de prueba. El siguiente paso será seleccionar los atributos buscando un mejor modelo, este es un punto clave de la investigación ya que, como se ha mencionado a lo largo de la investigación, en

Colombia y especialmente en el departamento del Cauca no se encuentran bases de datos confiables para la caficultura, de ahí la necesidad de buscar otras alternativas para solventar dicha problemática, tal como en este caso y el enfoque híbrido planteado con un modelo agronómico. Una de las principales razones del problema es la dificultad de conseguir los datos en la zona tanto de manejo del cultivo como los agrícolas y edáficos, en busca de darle solución a este objetivo y también a futuro una implementación real en el campo, se debe lograr la mayor eficiencia con la menor cantidad de atributos posible, encontrando así un óptimo modelo.

Para esto se contó con el asesoramiento de dos expertos para determinar qué atributos son indispensables para el desarrollo del proyecto y a su vez son fáciles de obtener en el campo, esto porque un modelo agronómico puede llegar a requerir una gran cantidad de datos de entrada que para su obtención requieren de un gran esfuerzo técnico y económico. En consecuencia, con el modelo de aprendizaje automático abordado en el presente capítulo la idea es obtener un óptimo resultado en la predicción y de la misma manera se busca que su implementación se logre de forma sencilla, para esto es necesario que la obtención de los parámetros de entrada del modelo se haga lo más fácil posible para el usuario final, es decir el caficultor caucano. En este sentido, en la tabla 12 se muestra el listado de atributos seleccionados por los expertos, dejando un total de 12, eliminando variables teóricas y otras que resultaban de una difícil extracción.

Tabla 12: Parámetros de entrada para el modelado con aprendizaje automático, Fuente propia

Atributo	Unidad
PH	pH
SOC	g/Kg
Kex	mmol/ Kg
PBray	mg/Kg
NrTrees	plantas/ha^{-1}
Leaf_att	Kg/ha^{-1}
Stem_att	Kg/ha^{-1}
Store_att	Kg/ha^{-1}
New_Yzero	Kg/ha^{-1}

N	Kg/ha^{-1}
P	Kg/ha^{-1}
K	Kg/ha^{-1}
N_supply	Kg/ha^{-1}

En consecuencia, con un total de 12 atributos se obtuvo un modelo que mostró su mejor desempeño con el algoritmo MLP alcanzando un 9.98 de RMSE, 97.22% de precisión y un R^2 de 0.99, manteniendo de esta forma un rendimiento relativamente muy bueno y disminuyendo a menos de la mitad la cantidad de atributos utilizados en el primer modelo, demostrando que el modelo es capaz de adaptarse satisfactoriamente a los datos aun cuando se reducen a unos pocos los atributos para el entrenamiento.

El paso siguiente fue la aplicación de los filtros *CfsSubsetEval* y *CorrelationAttributeEval* a los atributos, esto en búsqueda del objetivo mencionado anteriormente, como resultado se encuentra que algunos atributos guardan especial relación entre ellos. Por lo cual, se identifican un grupo de atributos que podrían ser seleccionados para eliminarse. Además, se identifican posibles agrupaciones de atributos más relevantes que podrían utilizarse. Con esto en mente, y buscando el mejor modelo posible, se ejecutaron pruebas de 16 modelos en los cuales se fue variando y eliminando los atributos de entrada, teniendo en cuenta los filtros ejecutados.

Consiguiendo así un modelo final con 9 atributos, eliminando las variables Stem_att, Store_att y New_Yzero de los atributos presentes en la tabla 12, dicho modelo logró alcanzar un RMSE de 9.65, una precisión del 97.35% y un R^2 de 0.99. Logrando mejorar un poco la precisión de este modelo comparada con la lograda inicialmente con los atributos mencionados en dicha tabla, el RMSE mostró una ligera mejora y el R^2 se mantuvo estable, por este motivo se selecciona este como el modelo con el mejor rendimiento alcanzado, demostrando que el modelo de aprendizaje automático con el algoritmo MLP desarrollado en el presente capítulo se adapta a los datos de entrenamiento y los reproduce en una manera relativamente exacta aun cuándo se le suministra sólo 9 atributos. Por lo tanto, se puede observar en la Figura 25 cómo se comporta el modelo seleccionado con 30 registros de entrenamiento frente a los datos modelados.

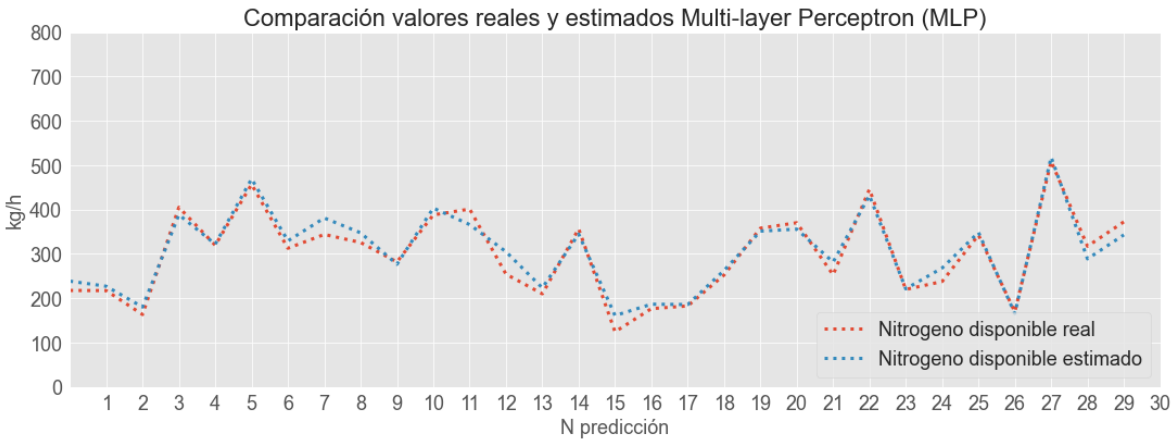


Figura 25. Datos de entrenamiento vs estimados de N con algoritmo MLP-9 atributos. Fuente: propia

Se tiene un modelo con un buen rendimiento que fue entrenado con pocos atributos y mantuvo resultados en el conjunto de datos de entrenamiento y de prueba. Sin embargo, buscando cumplir uno de los objetivos de esta investigación se espera que el modelo se aplique en la mayor cantidad de escenarios posibles en el campo, por tal motivo, se busca que funcione con los menos atributos posibles, sabiendo que esto podría conllevar posibles consecuencias aumentando el error. Teniendo esto en mente, se construyeron diez modelos con distintas combinaciones de atributos, las cuales se encuentran descritas en la tabla 13.

Tabla 13. Parámetros de entrada para el modelo de aprendizaje automático, Fuente propia

Combinación	Conjunto de atributos	RMSE
c1	pH, SOC, Kex, PBray, NrTrees, N,P,K	20.95 Kg/ha
c2	SOC, Kex, PBray, NrTrees, N,P,K	20.37 Kg/ha
c3	pH, SOC, Kex, PBray, NrTrees	105.84 Kg/ha
c4	pH, SOC, NrTrees, N, P	15.12Kg/ha
c5	pH, SOC, Kex	107.65 Kg/ha
c6	pH, SOC, Kex, PBray, NrTrees, N	30.5 Kg/ha
c7	pH, SOC, Kex, NrTrees, N	16.38 Kg/ha

c8	pH , SOC, PBray, N, P, K	42.72 Kg/ha
c9	pH , SOC, NrTrees, N	15.27 Kg/ha
c10	pH ,SOC	106.79 Kg/ha

La información de los modelos planteados con las combinaciones de atributos presentadas en la tabla 13, se encuentra consignada en la Figura 26**,** mediante la cual puede observarse que la combinación 9, compuesta por los atributos pH , SOC, NrTrees, N, presenta el mejor comportamiento por sobre todas las evaluadas, logrando un error de 15.27 Kg/ha con tan solo 4 atributos disponibles en el conjunto de datos de entrenamiento. Es importante resaltar que la combinación 4 logró un error ligeramente más bajo con 15.12 Kg/ha, sin embargo, se decide descartar porque trabaja con un atributo más y la mejora en el error no es significativa. En consecuencia, se da por cumplido el propósito planteado y se mantiene un buen rendimiento en la estimación de N.

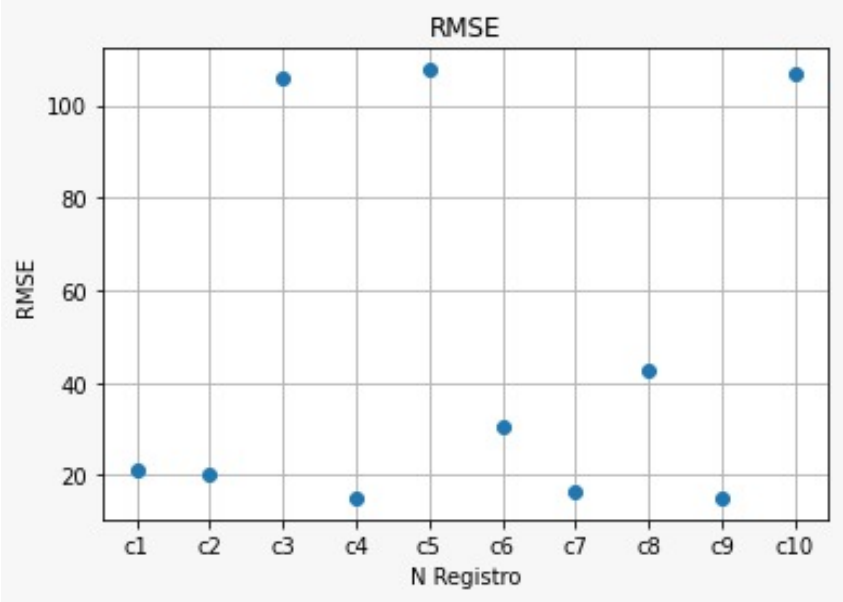


Figura 26. Error RMSE para 8 modelos con algoritmo MLP. Fuente: propia

Por otra parte, el modelo se logró probar con datos reales del municipio de Cajibío del departamento del Cauca, facilitados por la empresa Ecotecma. Los datos otorgados corresponden a un conjunto de datos de 186 registros, que cuenta con la disponibilidad de N en el suelo y 2 atributos: pH y SOC, la combinación 10 fue planteada con el fin de poder medir la efectividad del modelo en estos datos reales. Sin embargo, los resultados no fueron los esperados con los datos de prueba, alcanzando un error RMSE de 106.79 Kg/ha, el cual es el segundo error más grande de todas las combinaciones consignadas en la tabla 13. En ese orden de ideas, en la Figura 27 podemos observar el comportamiento del modelo frente a los datos reales.

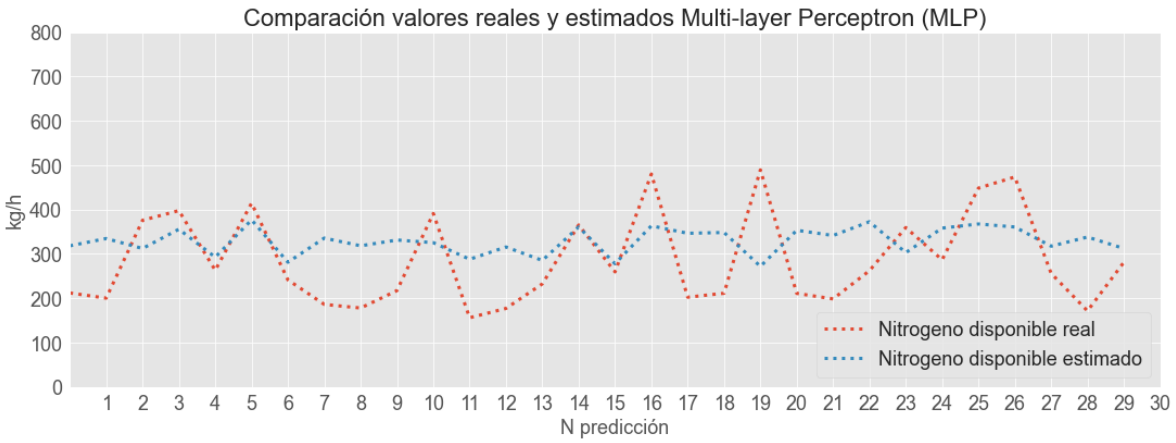


Figura 27. Datos reales vs estimados de N con algoritmo MLP-2 atributos. Fuente: propia

Los resultados obtenidos en la Figura 27, muestran un error grande, alcanzando un RMSE de 110.37 Kg/ha con una precisión del 65.8% para la predicción de N en los datos reales, esto puede ser explicado debido a los pocos atributos sobre los que se tiene registro, que resultan insuficientes para que el modelo haga una correcta estimación de la disponibilidad de N en el suelo. Esto puede observarse también en los datos de prueba, mediante los cuales la combinación c10 alcanza un RMSE de 106.79 Kg/ha, que resulta ser el segundo error más alto de todas las combinaciones evaluadas. Información que quedó consignada en la tabla 13.

Por esta razón se plantea como trabajo a futuro probar el modelo con un conjunto de datos reales que cuente con al menos los atributos planteados en la combinación 9, mostrada en dicha tabla, es decir pH, SOC, NrTrees y N. Ya que, si se tiene este mínimo de atributos de entrenamiento se podría garantizar una estimación de la disponibilidad de N con un error aceptable, y de esta manera poder validar los resultados obtenidos en esta investigación.

4.8. Conclusiones acerca del modelo de datos para la estimación de N en suelos establecidos con cultivo de café

En este capítulo fueron expuestas todas las fases de CRISP-DM necesarias para el modelado de datos en esta investigación, como primer paso se realizó la comprensión del negocio, comprensión de los datos y preparación de los datos. Con esto, fue generado un conjunto de datos generados sintéticamente por el modelo agronómico QUEFTS calibrado en el capítulo 3 para Colombia, que es insumo principal para el entrenamiento del modelo que permite estimar la disponibilidad de N en suelos con café. Durante el proceso de construcción del modelo base para la presente investigación, se llevó a cabo un análisis del desempeño de cinco técnicas de aprendizaje automático frente al conjunto de datos de entrenamiento, a su vez, se eligieron tanto los enfoques de entrenamiento y evaluación como los métodos de

selección de atributos necesarios para realizar con éxito el proceso de modelamiento y cumplir con el objetivo planteado de utilizar la menor cantidad de atributos posibles. Por último, se construyeron y evaluaron dieciséis modelos de aprendizaje automático, de los cuales fue seleccionado uno para la estimación de Nitrógeno en suelos establecidos con cultivo de café en el municipio del Cauca. En este orden de ideas, tomando como base los procesos y resultados expuestos en el presente capítulo, se concluye:

- Siguiendo la metodología CRISP-DM se llevaron a cabo varios procesos iterativos que permitieron el entendimiento del problema, la recolección de datos y la construcción de un conjunto de datos que describe la disponibilidad de N en suelos establecidos con cultivo de café en el departamento del Cauca. Dicho conjunto de datos, al encontrarse organizado y efectivamente descrito, será utilizado como insumo principal para esta investigación

- En Colombia, el acceso a los datos agrícolas es limitado. Específicamente, la construcción de un conjunto de datos de la disponibilidad de N en suelos establecidos con cultivo de café es una tarea compleja, debido a que existen pocos datos disponibles y el acceso a estos presenta una gran limitación. Sin embargo, existen enfoques diferentes para abordar esta problemática, tal como el presentado en la presente investigación que busca a través de modelos agronómicos reproducir el comportamiento agrícola, en este caso de la caficultura en el departamento del Cauca, conformando así un conjunto de datos que sirva de insumo para el modelo de aprendizaje automático.

- Aun cuando existen diferentes factores influyentes en la disponibilidad de N en cultivos de café, se determinó que a partir de pocas variables tanto edáficas como de manejo del cultivo, es posible estimar la disponibilidad de N con un R^2 de hasta 0.99 y un error MAE de 2.18. Comprobando así. que el modelo de aprendizaje automático construido a lo largo de este capítulo se adapta a los datos de entrenamiento y reproduce un error bajo y un R^2 bastante bueno en el conjunto de datos de prueba, lo que nos indica que el modelo se ajusta muy bien a los datos.

- Fue validado el resultado encontrado en primera instancia por WEKA y consolidado después en Python, mostrando que el algoritmo Multilayer Perceptron (MLP) es el que muestra mejores métricas de rendimiento tanto en los conjuntos de datos de validación, como en el de prueba. Alcanzando un RSME de 9.87 y un R^2 de 0.99 en la mejor versión trabajando únicamente con 9 atributos, con lo cual se obtiene una predicción de la disponibilidad de N en suelos con café con un error relativamente bajo, y un modelo bien adaptado a los datos.

- Cuatro de las cinco técnicas de aprendizaje automático analizadas presentaron buenos resultados, en ese orden de ideas los algoritmos DecisionTreeRegressor y MLP obtuvieron el mejor desempeño respecto a la predicción de la disponibilidad de N en cultivos de café. No obstante, esto no excluye la posibilidad de que otros algoritmos o enfoques sean analizados en trabajos futuros. Por ejemplo, el problema de investigación puede ser abordado mediante series de tiempo, con lo cual enfoques como los algoritmos metaheurísticos, aprendizaje profundo o algoritmos genéticos, podrían ser utilizados para estimar el rendimiento del cultivo de café y así generar una comparación con los resultados de la presente investigación.

- Fueron probaron modelos con distintas agrupaciones de atributos para el entrenamiento, y fueron seleccionadas las que mantuvieron buenos resultados, resultados consignados en la tabla 13, se logró alcanzar con la combinación 9, un error de 15.27 Kg/ha con tan solo 4 atributos disponibles en el conjunto de datos de entrenamiento: pH, SOC, NrTrees y N. De esta manera, se cumple con uno de los retos de la presente investigación, manteniendo resultados aceptables en la estimación de la disponibilidad de N en suelos con café, y a su vez, utilizando la menor cantidad de atributos, garantizando así, que pueda ejecutarse el modelo en la mayor cantidad de escenarios posibles en el campo.

5. Conclusiones y trabajo futuro

Este capítulo resalta las conclusiones más importantes obtenidas a través de la consecución de cada uno de los objetivos de la presente investigación. Así mismo, fueron plasmadas las contribuciones logradas con el trabajo desarrollado. Por último, se proponen diferentes trabajos futuros que pueden mejorar los resultados obtenidos en la presente investigación y con ello aportar a la calidad de vida de los caficultores colombianos.

5.1. Conclusiones

En la presente investigación se estimó la disponibilidad de N en suelos establecidos con cultivo del café a partir de la integración de modelos basados en procesos, modelos basados en datos y conocimiento experto, en busca de generar una herramienta que permitiese estimar la disponibilidad de N en el suelo, de esta forma apoyar a una fertilización más eficiente del cultivo y con ello contribuir a mejorar la calidad de vida de todas las personas involucradas en la cadena productiva del café en Colombia. En ese orden de ideas, primero se calibró el modelo agronómico QUEFTS para café a la región de estudio, es decir el departamento del Cauca. Posteriormente, se construyó un conjunto de datos de entrenamiento a partir del modelo calibrado, por último, fue construido un modelo de aprendizaje automático que permite estimar la disponibilidad de N en suelos establecidos con cultivo de café. Acorde con el trabajo realizado y como resultado de los procesos mencionados, en cada capítulo se exponen diversas conclusiones específicas de los mismos. Por su parte, en la presente sección se proponen las siguientes conclusiones generales de la investigación:

- ➔ En la literatura científica si bien existen diversas investigaciones con propuestas para la estimación o predicción de la disponibilidad de nitrógeno en cultivos de café. Dichas propuestas abarcan desde las técnicas tradicionales que implican desde el muestreo del suelo, hasta técnicas de visión por computadora, para estimar el N mediante imágenes satelitales. Así mismo, la mayor parte de los estudios están dirigidos a cultivos extensivos, comprendiendo el trigo, arroz, caña de azúcar, soya, algodón, entre otros. Por su parte, en cuanto a la fertilización del N para el café en Colombia, actualmente se basa en una planeación de acuerdo al calendario cafetero, recomendaciones que hace la FNC, sin embargo, no siempre resultan acertadas dada la heterogeneidad de los suelos, condiciones meteorológicas y de manejo del cultivo. En consecuencia, la presente investigación cobra relevancia en el campo de la caficultura colombiana, generando una propuesta que, basada en un enfoque híbrido con algoritmos de aprendizaje automático, modelos agronómicos y conocimiento experto, permite estimar

la disponibilidad de N en el suelo, y por consecuencia podría conllevar a un mejor manejo de los fertilizantes que suministran el N al cultivo.

- ➔ El modelo agronómico QUEFTS fue calibrado y validado con éxito el a la región de estudio, es decir, el departamento del cauca, aun cuándo se contó con acceso a una cantidad limitada de registros para la calibración, este demostró ser capaz de estimar con un error MAE relativamente bajo el rendimiento de los cultivos de café en suelos caucanos Logrando en su mejor versión un error MAE de 170.1 Kh/ha, un R^2 de 0.79 y PBIAS de 8.9%. Además, se validó la efectividad del modelo en la estimación de la disponibilidad de N en suelos establecidos con cultivo de café, alcanzando para la estimación de este rubro un error MAE de 9.11 Kg/ha, con un R^2 de 0.80 en la estimación de la disponibilidad de N.

- ➔ La disponibilidad de N en suelos establecidos con café, guarda una estrecha relación con diversas condiciones climáticas, físicas y químicas. La lluvia, temperatura y factores cómo la humedad, la aireación del suelo (niveles de oxígeno), el contenido de sal, la profundidad, la inclinación de la pendiente, entre otros, influyen directamente en el ciclo del N y en consecuencia en su disponibilidad, que a su vez puede verse afectada por factores edáficos como el SOC, así como por variabilidades en el manejo que se le da al cultivo. Por lo tanto, se podría plantear en un trabajo futuro un modelo que tome en cuenta esas variables para la estimación de la disponibilidad de N, para de esta forma lograr un modelo más robusto y una posible mejora en sus métricas, cómo el RMSE o el R^2.

- ➔ Fueron construidos modelos con diferentes algoritmos de aprendizaje automático y se evaluó su precisión, teniendo los mejores resultados con las técnicas multilayer perceptrón y DecisionTreeRegressor, con un coeficiente de correlación de 0.99 en el conjunto de datos de entrenamiento. Fueron evaluados algunos algoritmos seleccionados con base en estudios previos, así como también, esto no excluye la posibilidad de analizar otros algoritmos que podrían arrojar resultados exitosos. De la misma manera, se construyeron modelos con distintos atributos seleccionados, alcanzando un error de 15.27 Kg/ha con tan solo 4 atributos disponibles en el conjunto de datos de entrenamiento: pH, SOC, NrTrees y N. Por consiguiente, se alcanza uno de los beneficios que se buscan con el presente proyecto, y es que muestre el mejor desempeño posible con la menor cantidad posible de atributos, en búsqueda del beneficio del usuario final, en este caso el caficultor colombiano, así como también, esperando que la aplicabilidad del modelo sea mayor y poder utilizarlo en casos dónde los datos iniciales sean escasos.

5.2. Trabajos Futuros

En la presente investigación es propuesta una herramienta que busca ayudar a los caficultores colombianos a mejorar la eficiencia de la fertilización de N. Para ello, se estima la disponibilidad de N en suelos con cultivo de café mediante un enfoque híbrido de análisis. En ese sentido, con relación a los resultados obtenidos en la presente investigación, se proponen los siguientes trabajos futuros:

- Utilizar un conjunto de datos con más registros sobre estudios con café que cuenten con distintos manejos de cultivo (fertilización) y distinta densidad de plantas, así como la disponibilidad de N en el suelo, información necesaria para probar el modelo resultante de la presente investigación con el entrenamiento mediante un enfoque híbrido compuesto por el modelo agronómico QUEFTS calibrado a la región de estudio, así como de algoritmos de aprendizaje automático. En consecuencia, poder validar los resultados obtenidos y por consiguiente, la aplicabilidad del modelo. Por otro lado, si se cuenta con más atributos se pueden plantear posibles mejoras al modelo, ya que está demostrado que existe una alta correlación entre las condiciones del cultivo y el ambiente que lo rodea en general y su disponibilidad de N, se propone utilizar estas variables adicionales para obtener una mejora en las métricas de evaluación del modelo estimando la disponibilidad de N. En ese orden de ideas, variables como el porcentaje de sombra en el cultivo, la pendiente del terreno, la cantidad de irrigación, entre otras, deberían ser analizadas. Así mismo, se pueden plantear trabajos distintos haciendo uso de las mismas variables, pero para el análisis de clases objetivo diferentes, como la cantidad de fertilizante a utilizar, o la incidencia de posibles enfermedades, así como la incidencia de la cantidad de agua en el suelo en la disponibilidad de N.

- Analizar distintas técnicas de aprendizaje automático, así como otros enfoques ya sean híbridos o no para la consecución y el tratamiento de los datos, las cinco técnicas de aprendizaje automático analizadas presentaron buenos resultados. Sin embargo, resulta tentadora la posibilidad de analizar otros algoritmos y enfoques en trabajos futuros. Por ejemplo, el problema de investigación puede ser abordado mediante algoritmos genéticos, o técnicas avanzadas de aprendizaje profundo. Por otro lado, se podría utilizar el enfoque híbrido planteado en esta investigación, pero con el uso de otro modelo agronómico, como puede ser el Dynacof, un modelo agronómico para café que tiene en cuenta variables adicionales como la incandescencia del sol, o la irrigación, una variable que podría servir para darle distintos enfoques y predecir distintas clases objetivo, como por ejemplo los niveles

de humedad en el suelo, o analizar la incidencia de las cantidades de agua en el suelo con el rendimiento del cultivo.

- Desarrollar un sistema de recomendaciones que permita mejorar las tasas de eficiencia del uso de N y por ende mantener o mejorar el rendimiento en los cultivos. Esto haciendo uso de la disponibilidad de N estimada en la presente investigación. Se propone un sistema de recomendaciones basado en contenido, que permita generar recomendaciones de distintos fertilizantes, así como de la cantidad que se debe utilizar para mantener niveles óptimos de N en el suelo, garantizando de esta manera una mejor eficiencia del mismo, y en consecuencia, un uso responsable de los fertilizantes, que pueda asegurar las condiciones óptimas de N en el suelo para el desarrollo normal del cultivo, así como también evite el desperdicio de fertilizantes, lo que podría conllevar en pérdidas económicas para el caficultor, o en algunos casos podría conllevar a un cultivo más propenso a enfermedades, de esta manera también se puede mitigar el daño ambiental, ya que el exceso de fertilización es una de las principales causas de contaminación de N en las fuentes hídricas encontradas en el subsuelo del cultivo.

6. Referencias

[1] Federación Nacional de Cafeteros *(consultado el 20 de octubre de 2020)*, “Gobierno Nacional y Federación Nacional de Cafeteros firman agenda 2030 para el sector cafetero”, [Online]. Available: https://federaciondecafeteros.org/wp/listado-noticias/gobierno-nacional-y-federacion-nacional-de-cafeteros-firman-agenda-2030-para-el-sector-cafetero/.

[2] J. G. A. Barbedo, “Detection of nutrition deficiencies in plants using proximal images and machine learning: A review”, *Comput. Electron. Agric.*, vol. 162, pp. 482–492, jul. 2019, doi: 10.1016/j.compag.2019.04.035.

[3] A. Chemura, O. Mutanga, J. Odindi, y D. Kutywayo, “Mapping spatial variability of foliar nitrogen in coffee (Coffea arabica L.) plantations with multispectral Sentinel-2 MSI data”, *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 1–11, abr. 2018, doi: 10.1016/j.isprsjprs.2018.02.004.

[4] lack S.h, naderi A, S. saidat, A. A, G. H. Mohammadi, y M. S.H, “The Effects of Different Levels of Irrigation, Nitrogen and Plant Population on Yield, Yield Components and Dry Matter Remobilization of Corn at Climatical Conditions of Khuzestan”, *J. Sci. Technol. Agric. Nat. Resour.*, ene. 2008.

[5] Y. Fu , G. Yang, Z. Li, H. Li, X. Xu, X. Song, Y. Zhang, D. Duan, C. Zhao y L. Chen , “Progress of hyperspectral data processing and modelling for cereal crop nitrogen monitoring”, *Comput. Electron. Agric.*, vol. 172, p. 105321, may 2020, doi: 10.1016/j.compag.2020.105321.

[6] S. Sadeghian Khalajabadi y H. Duque Orrego, “Formulaciones Generales de Fertilizantes: Alternativas para una nutrición balanceada de los cafetales en Colombia”. https://www.cenicafe.org/es/publications/AVT0483.pdf (consultado el 20 de octubre de 2020).

[7] Y. Li, D. Wang, V. Lasoukanh, X. Yang, W. Li, y Y. Zhao, “Prediction of carbon, nitrogen and phosphorus contents of Leymus Chinensis based on soil chemical properties using artificial neural networks”, *Nongye Gongcheng XuebaoTransactions Chin. Soc. Agric. Eng.*, vol. 30, pp. 104–111, feb. 2014, doi: 10.3969/j.issn.1002-6819.2014.03.014.

[8] Y. Li, S. Liang, Y. Zhao, W. Li, y Y. Wang, “Machine learning for the prediction of L. chinensis carbon, nitrogen and phosphorus contents and understanding of mechanisms underlying grassland degradation”, *J. Environ. Manage.*, vol. 192, pp. 116–123, may 2017, doi: 10.1016/j.jenvman.2017.01.047.

[9] Cenicafe (consultado el 20 de octubre de 2020), “Nuestras Publicaciones, Revista Cenicafe”, [Online]. Available: www.cenicafe.org”. https://www.cenicafe.org/es/index.php/nuestras_publicaciones/revista_cenicafe.

[10] “G. Guarín. “Impacto de la Variabilidad Climática en la Producción de Banano en el Urabá Antioqueño”. Maestría thesis, Universidad Nacional de Colombia, Sede Medellín. 2011.”

[11] P. Feng, B. Wang, D. L. Liu, C. Waters, y Q. Yu, “Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts

on wheat yield in south-eastern Australia", *Agric. For. Meteorol.*, vol. 275, pp. 100–113, sep. 2019, doi: 10.1016/j.agrformet.2019.05.018.
[12] *ANEIA - Universidad de Los Andes* (consultado el 11 de septiembre de 2021), "El valor del café colombiano, ¿en el grano o en la taza?", [Online]. Available:. https://agronegocios.uniandes.edu.co/2021/09/11/el-valor-del-cafe-colombiano-en-el-grano-o-en-la-taza/.
[13] Cenicafe, "Fertilidad del suelo y nutrición del café en Colombia". 2008.
[14] M. Andrews y P. Lea, "Do plants need nitrate? The mechanisms by which nitrogen form affects plants", *Ann. Appl. Biol.*, vol. 163, pp. 174–199, sep. 2013, doi: 10.1111/aab.12045.
[15] N. Vigneau, M. Ecarnot, G. Rabatel, y P. Roumet, "Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat", *Field Crops Res.*, vol. 122, núm. 1, pp. 25–31, abr. 2011, doi: 10.1016/j.fcr.2011.02.003.
[16] A. Chlingaryan, S. Sukkarieh, y B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review", *Comput. Electron. Agric.*, vol. 151, pp. 61–69, ago. 2018, doi: 10.1016/j.compag.2018.05.012.
[17] A. Mechelli y S. Vieira, "Machine Learning - 1st Edition". https://www.elsevier.com/books/machine-learning/mechelli/978-0-12-815739-8 (consultado el 21 de octubre de 2020).
[18] J. Lee, J. Shin, y M. Realff, "Machine Learning: Overview of the Recent Progresses and Implications for the Process Systems Engineering Field", *Comput. Chem. Eng.*, vol. 114, oct. 2017, doi: 10.1016/j.compchemeng.2017.10.008.
[19] Richard S. Sutton, Andrew G, *Reinforcement Learning An Introduction second edition*, 2018.
[20] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, y K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey", *Comput. Electron. Agric.*, vol. 155, pp. 257–282, dic. 2018, doi: 10.1016/j.compag.2018.10.024.
[21] J. Behmann, A.-K. Mahlein, T. Rumpf, C. Römer, y L. Plümer, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection", *Precis. Agric.*, vol. 16, pp. 239–260, jun. 2015, doi: 10.1007/s11119-014-9372-7.
[22] A. Goap, D. Sharma, A. K. Shukla, y C. Rama Krishna, "An IoT based smart irrigation management system using Machine learning and open source technologies", *Comput. Electron. Agric.*, vol. 155, pp. 41–49, dic. 2018, doi: 10.1016/j.compag.2018.09.040.
[23] A. Tellaeche, X. P. Burgos-Artizzu, G. Pajares, y A. Ribeiro, "A vision-based method for weeds identification through the Bayesian decision theory", *Pattern Recognit.*, vol. 41, núm. 2, pp. 521–530, feb. 2008, doi: 10.1016/j.patcog.2007.07.007.
[24] E. J. Link, *Investigation and modeling of the optimization potential of adapted nitrogen fertilization strategies in corn cropping systems with regard to minimize nitrogen losses*, 2005.
[25] D. Dourado-Neto, D. A. Teruel, K. Reichardt, D. R. Nielsen, J. A. Frizzone, y O.

O. S. Bacchi, “Principles of crop modeling and simulation: I. uses of mathematical models in agricultural science”, *Sci. Agric.*, vol. 55, núm. SPE, pp. 46–50, 1998, doi: 10.1590/S0103-90161998000500008.
[26] University of New Hampshire (consultado el 21 de octubre de 2020), “The DNDC Home Page”. [Online]. Available: https://www.dndc.sr.unh.edu/.
[27] J. Qiu, C. Li, L. Wang, H. Tang, H. Li, y E. Van Ranst, “Modeling impacts of carbon sequestration on net greenhouse gas emissions from agricultural soils in China”, *Glob. Biogeochem Cycles*, vol. 23, mar. 2009, doi: 10.1029/2008GB003180.
[28] S. E. Muhammed, K. Coleman, L. Wu, V. Bell, A. Davies, J. Quinton, E. Carnell, S. Tomlinson, A. Dore, U. Dragosits, P. Naden, E. Tipping, A. Whitmore, “Impact of two centuries of intensive agriculture on soil carbon, nitrogen and phosphorus cycling in the UK”, *Sci. Total Environ.*, vol. 634, pp. 1486–1504, sep. 2018, doi: 10.1016/j.scitotenv.2018.03.378.
[29] W. J. Parton, “The CENTURY model”, en *Evaluation of Soil Organic Matter Models*, Berlin, Heidelberg, 1996, pp. 283–291. doi: 10.1007/978-3-642-61094-3_23.
[30] Natural Resource Ecology Laboratory (consultado el 21 de octubre de 2020), “NREL-DayCent: Daily Century Model”. [Online]. Available: https://www2.nrel.colostate.edu/projects/daycent-home.html.
[31] *APSIM* (consultado el 21 de octubre de 2020), “What is APSIM?” . [Online]. Available: https://www.apsim.info/apsim-model/
[32] STICS (consultado el 21 de octubre de 2020), “Agroclim STICS - About us ?”. [Online]. Available: https://www6.paca.inrae.fr/stics_eng/About-us.
[33] Texas university (consultado el 21 de octubre de 2020), “EPIC | EPIC & APEX Models”. [Online]. Available: https://epicapex.tamu.edu/epic/.
[34] International institute for applied system analysis (consultado el 21 de octubre de 2020), “The Environmental Policy Integrated Model (EPIC)”. [Online]. Available: https://iiasa.ac.at/web/home/research/researchPrograms/EcosystemsServicesandManagement/EPIC.en.html
[35] DSSAT (consultado el 21 de octubre de 2020), “DSSAT Overview”, *DSSAT.net*. [Online]. Available: https://dssat.net/about/.
[36] C. dr Ajw. W. C. form, “WOFOST - WOrld FOod STudies”, *WUR*, el 15 de agosto de 2019. [Online]. Available: https://www.wur.nl/en/Research-Results/Research-Institutes/Environmental-Research/Facilities-Tools/Software-models-and-databases/WOFOST.htm (consultado el 21 de octubre de 2020).
[37] A. de Wit, H. Boogaard, D. Fumagali, S. Janssen, R. Knapen, D. van Kraalingen, I. Supit y K. van Diepen , “25 years of the WOFOST cropping systems model”, *Agric. Syst.*, vol. 168, pp. 154–167, ene. 2019, doi: 10.1016/j.agsy.2018.06.018.
[38] Regional Agronomy (consultado el 21 de octubre de 2020), “The QUEFTS model — Regional Agronomy”. [Online]. Available: https://reagro.org/methods/explanatory/quefts.html.
[39] G. Maro, “Developing a Coffee Yield Prediction and Integrated Soil Fertility Management Recommendation Model for Northern Tanzania | Request PDF”,

doi: 10.9734/IJPSS/2014/6883.
[40] R. Vezy, G. le Maire, M. Christina, S. Georgiou, P. Imbach, H. Hidalgo, E. Alfaro, F. Charbonnier, P. Lehner y O. Roupsard , “DynACof: A process-based model to study growth, yield and ecosystem services of coffee agroforestry systems”, *Environ. Model. Softw.*, vol. 124, p. 104609, feb. 2020, doi: 10.1016/j.envsoft.2019.104609.
[41] S. M. Guzmán, J. O. Paz, M. L. M. Tagert, A. E. Mercer, y J. W. Pote, “An integrated SVR and crop model to estimate the impacts of irrigation on daily groundwater levels”, *Agric. Syst.*, vol. 159, pp. 248–259, ene. 2018, doi: 10.1016/j.agsy.2017.01.017.
[42] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, y S. V. Archontoulis, “Maize yield and nitrate loss prediction with machine learning algorithms”, *Environ. Res. Lett.*, vol. 14, núm. 12, p. 124026, dic. 2019, doi: 10.1088/1748-9326/ab5268.
[43] T. Zhang, J. Su, C. Liu, y W.-H. Chen, “Bayesian calibration of AquaCrop model for winter wheat by assimilating UAV multi-spectral images”, *Comput. Electron. Agric.*, vol. 167, p. 105052, dic. 2019, doi: 10.1016/j.compag.2019.105052.
[44] L. Leroux, M. Castets, C. Baron, M.-J. Escorihuela, A. Bégué, y D. Lo Seen, “Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices”, *Eur. J. Agron.*, vol. 108, pp. 11–26, ago. 2019, doi: 10.1016/j.eja.2019.04.007.
[45] K. Yamamoto, “Distillation of crop models to learn plant physiology theories using machine learning”, *PLOS ONE*, vol. 14, núm. 5, p. e0217075, may 2019, doi: 10.1371/journal.pone.0217075.
[46] C. Folberth, A. Baklanov, J. Balkovič, R. Skalský, N. Khabarov, y M. Obersteiner, “Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning”, *Agric. For. Meteorol.*, vol. 264, pp. 1–15, ene. 2019, doi: 10.1016/j.agrformet.2018.09.021.
[47] P. Feng, B. Wang, D. Li, C. Waters, D. Xiao, L. Shi y Q. Yu , “Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique”, *Agric. For. Meteorol.*, vol. 285–286, p. 107922, may 2020, doi: 10.1016/j.agrformet.2020.107922.
[48] B. Saravi y A. Pouyan Nejadhashemi, “Quantitative model of irrigation effect on maize yield by deep neural network | SpringerLink”. https://link.springer.com/article/10.1007/s00521-019-04601-2 (consultado el 7 de marzo de 2021).
[49] M. Shahhosseini, G. Hu, S. V. Archontoulis, y I. Huber, “Coupling Machine Learning and Crop Modeling Improves Crop Yield Prediction in the US Corn Belt”, *ArXiv200804060 Cs Q-Bio*, jul. 2020, Consultado: el 19 de febrero de 2021. [En línea]. Disponible en: http://arxiv.org/abs/2008.04060
[50] Y. Bai y J. Gao, “Optimization of the nitrogen fertilizer schedule of maize under drip irrigation in Jilin, China, based on DSSAT and GA”, *Agric. Water Manag.*, vol. 244, p. 106555, feb. 2021, doi: 10.1016/j.agwat.2020.106555.
[51] B. M. Shehu, B. Lawan, J. Jibrin, A. Kamara, I. Mohammed, J. Rurinda, S. Zingore, P. Craufurd, A. Adam y R. Merckx , “Balanced nutrient requirements for maize in the Northern Nigerian Savanna: Parameterization and validation of QUEFTS model”, *Field Crops Res.*, vol. 241, p. 107585, sep. 2019, doi: 10.1016/j.fcr.2019.107585.

[52] D. Capa, J. Pérez-Esteban, y A. Masaguer, “Unsustainability of recommended fertilization rates for coffee monoculture due to high N2O emissions”, *Agron. Sustain. Dev.*, vol. 35, núm. 4, pp. 1551–1559, oct. 2015, doi: 10.1007/s13593-015-0316-z.
[53] F. F. Valencia, J. R. R. Sáenz, y H. D. M. Franco, “Densidad de siembra de Coffea arabica variedad tabi en sistemas agroforestales, en tres zonas cafeteras de Colombia”, p. 6, 2016.
[54] F. Farfán-Valencia y J. E. Baute-Balcázar, “efecto del arreglo espacial del café y del sombrío sobre la producción de café”, p. 11.
[55] F. Farfán-Valencia y J. E. Baute-Balcázar, “La fertilización mineral como complemento a la fertilización con abono orgánico en el cultivo del café.”, *Rev. Cenicafé*, núm. 71–1, pp. 48–53, jul. 2020, doi: 10.38141/10778/1119.
[56] L. Salazar y S. Sadeghian, “Producción de café (Coffea Arabica L.) en respuesta al manejo específico por sitio de la fertilidad del suelo”, *Rev. Investig. Agrar. Ambient.*, vol. 7, p. 25, Diciembre 2016, doi: 10.22490/21456453.1555.
[57] L. Kouadio, R. C. Deo, V. Byrareddy, J. F. Adamowski, S. Mushtaq, y V. Phuong Nguyen, “Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties”, *Comput. Electron. Agric.*, vol. 155, pp. 324–338, Diciembre 2018, doi: 10.1016/j.compag.2018.10.014.
[58] FAO, “Guidelines for soil description”, en *FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS*, 4a ed.,
[59] IBM Corporation, “Manual CRISP-DM de IBM SPSS Modeler”, 2021.
[60] R. Wirth y J. Hipp, “CRISP-DM: Towards a standard process model for data mining”, *Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, ene. 2000.
[61] A. Salamanca-Jimenez, T. A. Doane, y W. R. Horwath, “Nitrogen Use Efficiency of Coffee at the Vegetative Stage as Influenced by Fertilizer Application Method”, *Front. Plant Sci.*, vol. 8, p. 223, mar. 2017, doi: 10.3389/fpls.2017.00223.
[62] H. Celaya-Michel, A. E. Castellanos-Villegas, H. Celaya-Michel, y A. E. Castellanos-Villegas, “Mineralización de nitrógeno en el suelo de zonas áridas y semiáridas”, *Terra Latinoam.*, vol. 29, núm. 3, pp. 343–356, sep. 2011.
[63] L. E. C. Rincón y F. A. A. Gutiérrez, “Dinámica del ciclo del nitrógeno y fósforo en suelos Nitrogen and phosphorus cycles dynamics in soils”, núm. 1, p. 11, 2012.
[64] United States Department. of Agriculture. USDA. NRCS, “Soil Nitrogen Guide”. USDA, 2012. [Online]. Available: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_053274.pdf
[65] D. C. Corrales, A. Ledezma, y J. C. Corrales, “A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal”, Department of Computer Science and Engineering, Universidad Carlos III de Madrid, *J. Comput.*, vol. 10, núm. 6, pp. 396–405, nov. 2015, doi: 10.17706/jcp.10.6.396-405.
[66] J. Rincon-Patino, E. Lasso, y J. Corrales, “Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data”, *Sustainability*, vol. 10, p. 3498, sep. 2018, doi: 10.3390/su10103498.
[67] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. Witten, “The

WEKA data mining software: An update", *SIGKDD Explor Newsl*, vol. 11, pp. 10–18, nov. 2008.
[68] *Anaconda* (consultado el 26 de marzo de 2022),"Anaconda | State of Data Science 2021", . [Online]. Available: https://www.anaconda.com/state-of-data-science-2021 (consultado el 26 de marzo de 2022).
[69] G. Zhang, M. Y. Hu, B. Eddy Patuwo, y D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *Eur. J. Oper. Res.*, vol. 116, núm. 1, pp. 16–32, jul. 1999, doi: 10.1016/S0377-2217(98)00051-4.
[70] L. Tashman, "Out-of sample tests of forecasting accuracy: an analysis and review", *Int. J. Forecast.*, vol. 16, pp. 437–450, ene. 2000.
[71] C. Bergmeir y J. Benítez, "On the use of cross-validation for time series predictor evaluation", *Inf. Sci.*, vol. 191, pp. 192–213, may 2012, doi: 10.1016/j.ins.2011.12.028.
[72] S. Abu-Nimeh, D. Nappa, X. Wang, y S. Nair, "A comparison of machine learning techniques for phishing detection", ene. 2007, vol. 269, pp. 60–69. doi: 10.1145/1299015.1299021.
[73] R. Taylor, "Interpretation of the Correlation Coefficient: A Basic Review", 1990, doi: 10.1177/875647939000600106.
[74] J. D. Rincon-Patino y J. C. Corrales, "Análisis del rendimiento de café basado en técnicas de aprendizaje automático". Universidad del Cauca, 2019.
[75] E. G. L. Sambony y J. C. Corrales, "Sistema experto basado en emparejamiento de patrones", Universidad del Cauca , 2017 p. 93. http://dx.doi.org/10.22395/rium.v15n29a5
[76] D. Corrales, E. Lasso, A. Ledezma Espino, y J. Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?", *J. Intell. Fuzzy Syst.*, vol. 34, pp. 1–11, may 2018, doi: 10.3233/JIFS-169470.
[77] S. Rathore y A. Gupta, "A Comparative Study of Feature-Ranking and Feature-Subset Selection Techniques for Improved Fault Prediction", presentado en ACM International Conference Proceeding Series, feb. 2014. doi: 10.1145/2590748.2590755.
[78] D. Ferreira, H. Peixoto, J. Machado, y A. Abelha, "Predictive Data Mining in Nutrition Therapy", en *2018 13th APCA International Conference on Automatic Control and Soft Computing (CONTROLO)*, jun. 2018, pp. 137–142. doi: 10.1109/CONTROLO.2018.8516413.

CÁLCULO DE LA DISPONIBILIDAD DE NITRÓGENO EN SUELOS ESTABLECIDOS CON CULTIVO DE CAFÉ UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Anexos

Proyecto de Trabajo de Grado

Julian Egas Daza

Andrés Felipe Bravo Portilla

Director: Msc. Juan Fernando Casanova Olaya

Codirector: PhD. Juan Carlos Corrales Muñoz

Asesor: PhD. Cristhian Nicolás Figueroa Martínez

Asesora: PhD. María Cristina Ordoñez Díaz

Departamento de Telemática

Facultad de Ingeniería Electrónica y Telecomunicaciones

Universidad del Cauca

Popayán, Cauca, abril de 2022

Anexo A

El anexo A presenta las carpetas que contienen el código en lenguaje R del modelo agronómico QUEFTS calibrado durante la investigación, así como los archivos de soporte utilizados durante el proceso de calibración y validación del mismo, toda la información recién mencionada se encuentra en la carpeta Modelo agronómico.

Disponibles en:

https://github.com/andresf5/codigos_modelos_agronomico

Anexo B

El anexo B presenta las carpetas que contienen el código en un jupyter notebook en lenguaje Python, así como los notebooks de soporte utilizados durante el proceso de validación del mismo y además se encuentran los registros para las combinaciones de distintos modelos presentada en el capítulo 4, toda la información recién mencionada se encuentra en la carpeta Modelo aprendizaje automático.

Disponibles en:

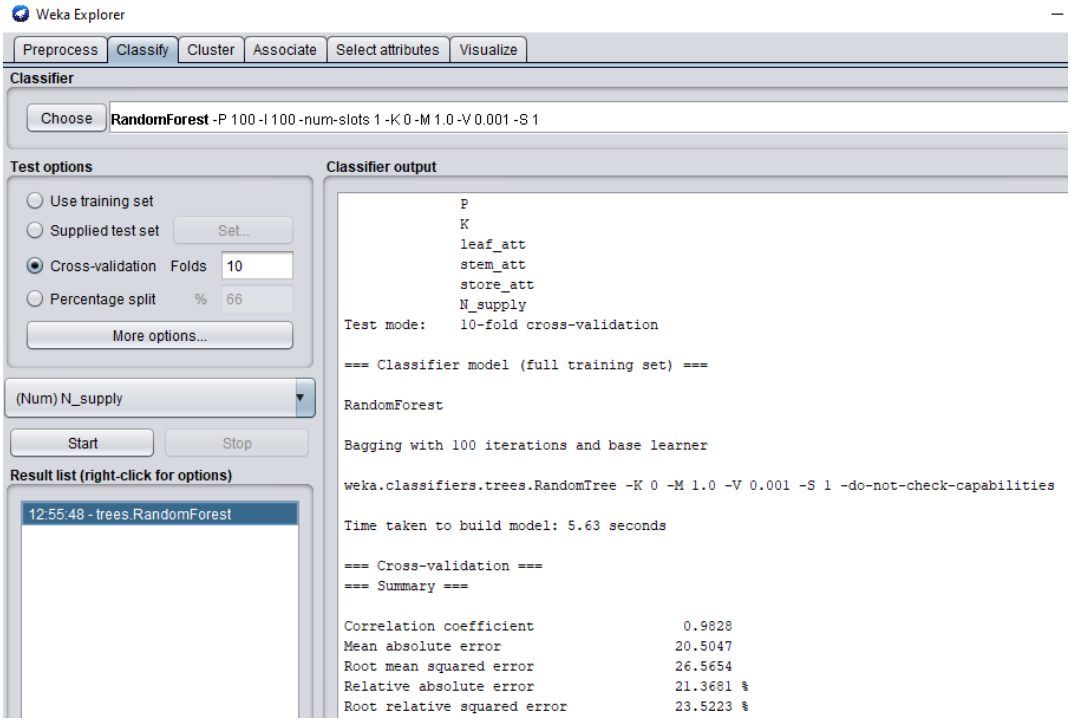
https://github.com/andresf5/codigos_modelos_agronomico

Anexo C

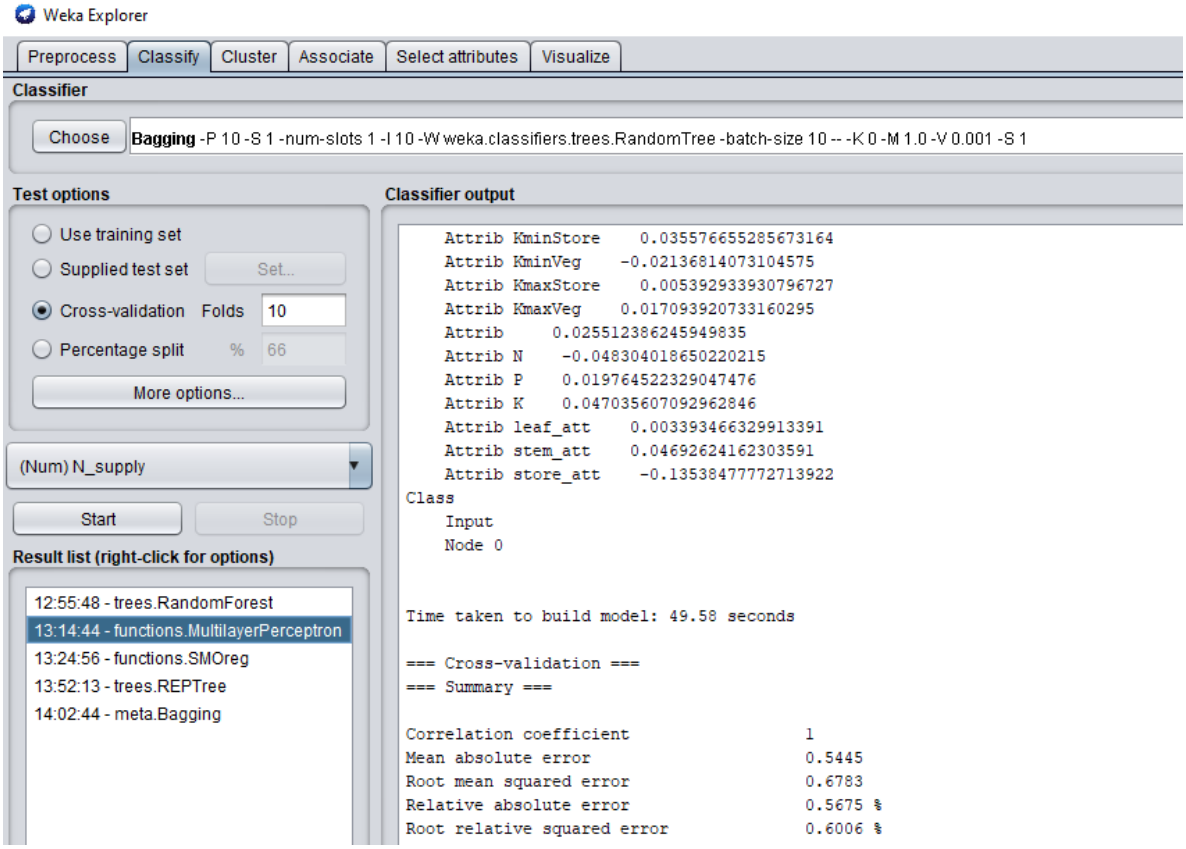
El anexo C presenta los pantallazos del software Weka utilizados durante el desarrollo del presente proyecto, en los procesos de selección de técnica de aprendizaje automático y selección de filtros

Selección de técnica de aprendizaje automático.

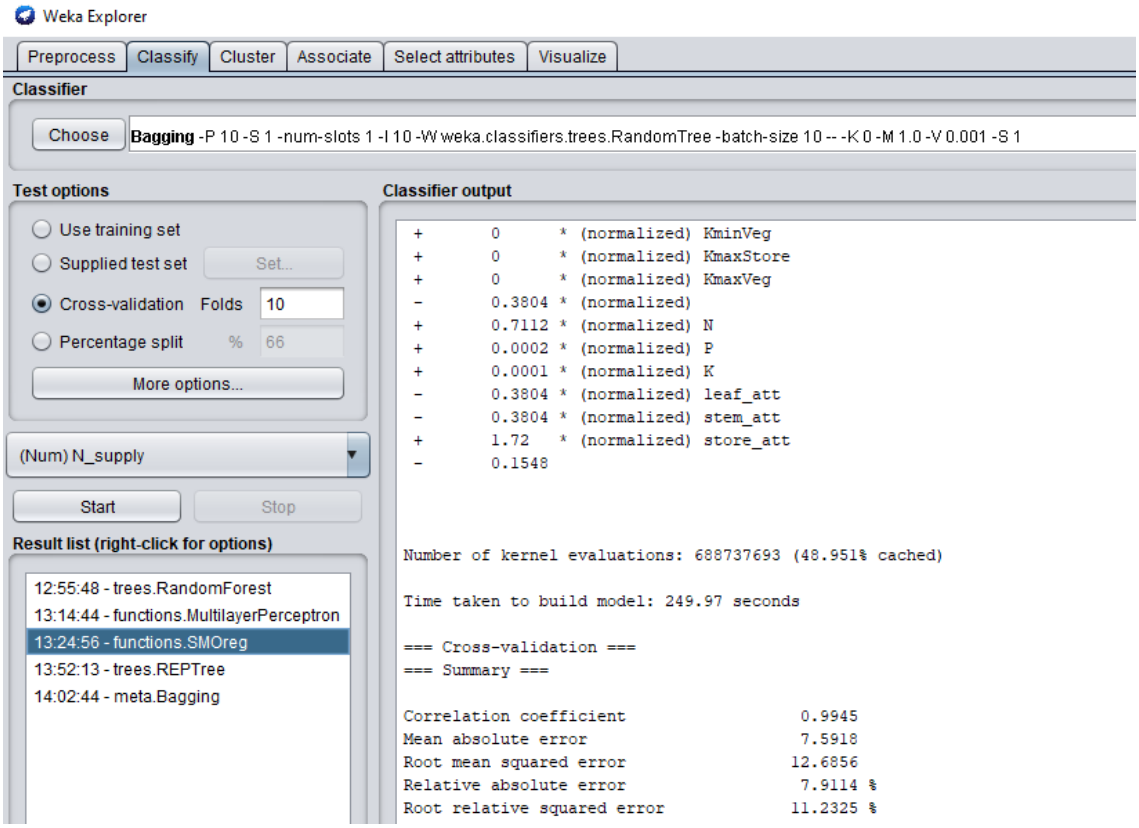
RandomForest



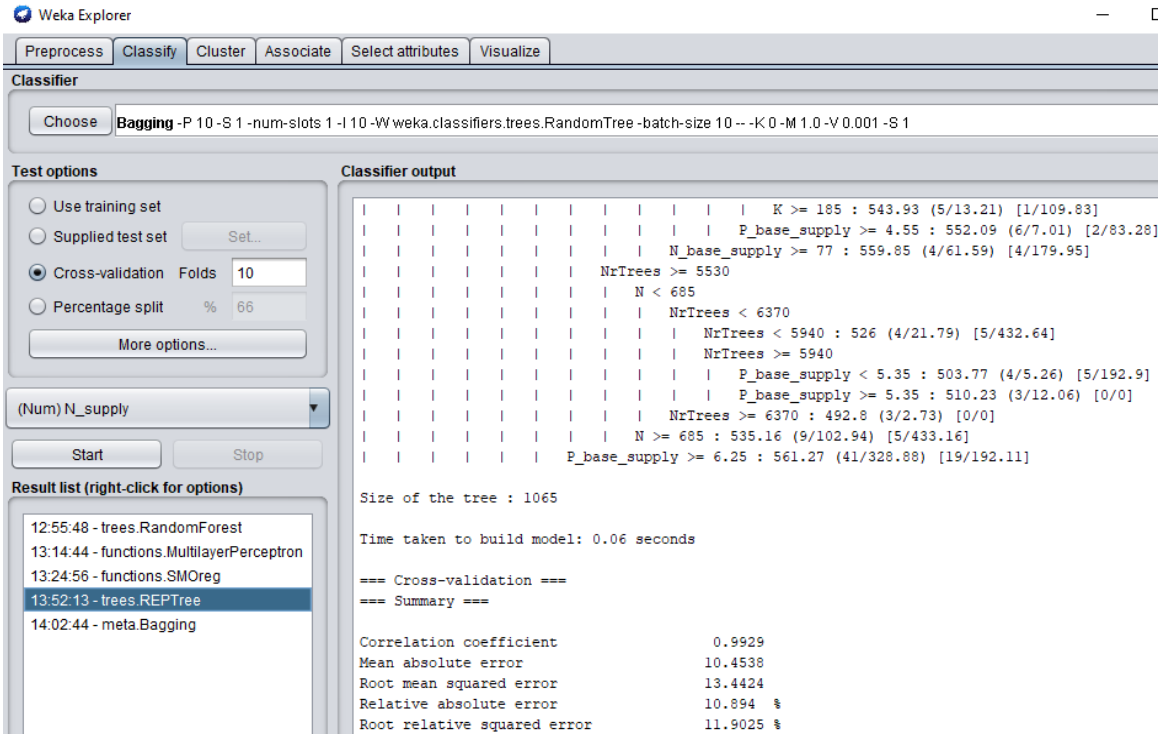
Multilayer Perceptron MLP



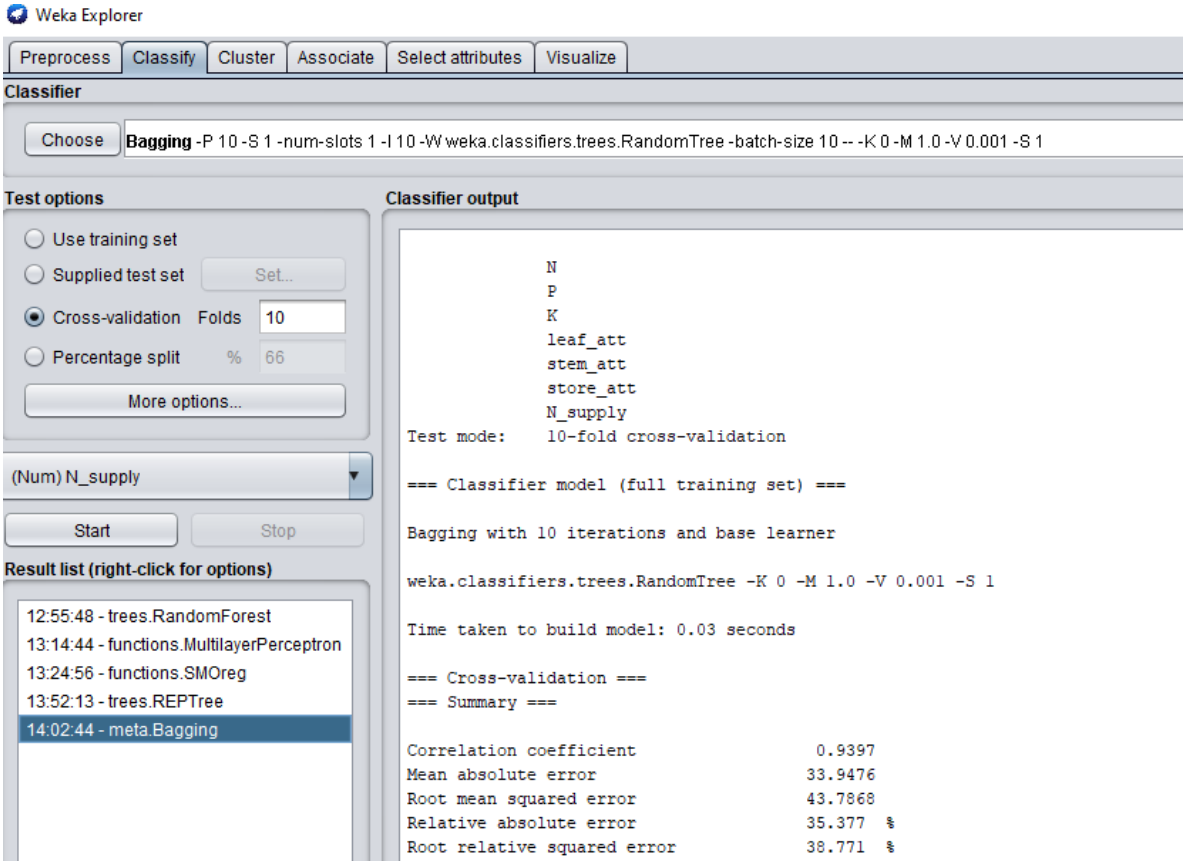
SVR



Decision Tree

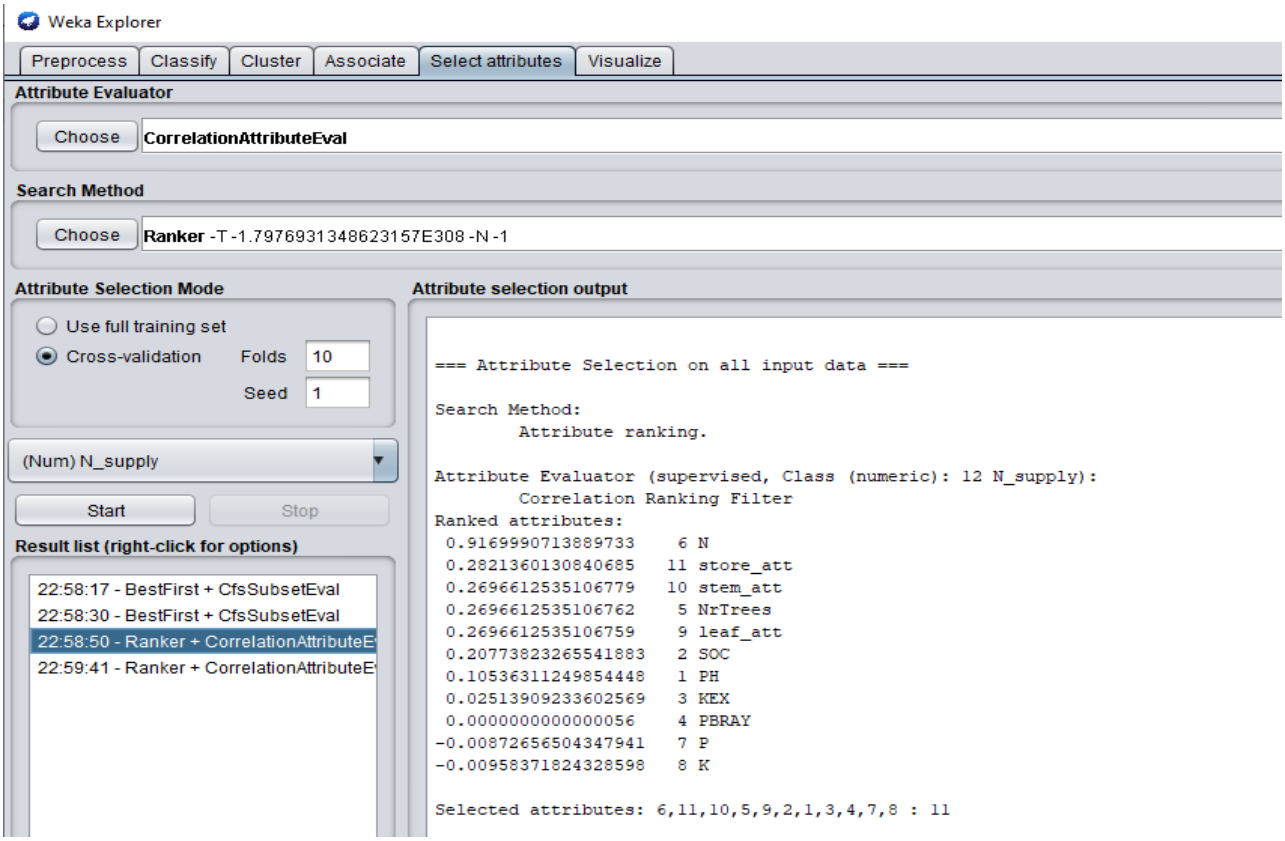


Bagging Regressor



Filtros para selección de Atributos

CorrelationAtrributeEval



CfsSubsetEval