Evaluación de modelos de aprendizaje automático para la predicción de variables asociadas a la producción en cultivos de caña de azúcar usando índices de vegetación y variables edáficas

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires

Área Biometría y Mejoramiento

Camilo Alberto Herrera Rozo

Estadístico

Universidad del Valle

Escuela para Graduados *Ing. Agr. Alberto Soriano*

Facultad de Agronomía – Universidad de Buenos Aires

COMITÉ CONSEJERO

Director de Tesis

Pablo A. Cipriotti

Ingeniero Agrónomo (Universidad de Buenos Aires)

Doctor en Ciencias Agropecuarias (Universidad de Buenos Aires)

JURADO DE TESIS

Director de Tesis

Pablo A. Cipriotti

Ingeniero Agrónomo (Universidad de Buenos Aires)

Doctor en Ciencias Agropecuarias (Universidad de Buenos Aires)

JURADO

Mónica Balzarini

Ingeniera Agrónoma (Universidad Nacional de Córdoba)

Doctora con Especialidad en Estadística Aplicada (Louisiana State University, USA)

JURADO

Santiago Rámon Verón

Ingeniero Agrónomo (Universidad de Buenos Aires)

Doctor en Ciencias Agropecuarias (Universidad de Buenos Aires)

JURADO

Rosa Teresa Boca

Ingeniera Agrónoma (Universidad de Buenos Aires)

Doctora en Ciencias Agropecuarias (Universidad de Buenos Aires)

Fecha de defensa de la tesis: 13 Agosto 2019

A mis padres, Fabio y Nubia

A mis hermanos, Andrés, Jaime y Angélica

A mi compañera de muchas travesías y compañera de vida María Isabel

Quienes son mis mayores pilares.

Agradecimientos:

Esta tesis fue posible gracias a todas aquellas personas e instituciones con las cuales tuve la oportunidad de compartir y aprender. En primer lugar, agradezco a mi familia por el apoyo incondicional que siempre me han brindado. Sus sabias palabras y oportunos consejos han sido de gran ayuda y motivación en todo momento. Llevar a buen término tanto esta investigación como los demás logros que he obtenido no hubiera sido posible sin su compañía.

Agradezco al Centro de Investigación de la Caña de Azúcar de Colombia (Cenicaña), entidad que me brindó el apoyo para poder realizar esta investigación con información disponible y guía constante, al Dr Javier Ali Carbonell que creyó en este proyecto y puso a disposición el centro en todas las fases de esta investigación, a Héctor Chica por su apoyo y consejos, Paulo, César, Andrés y a las distintas dependencias de Cenicaña que me brindaron acompañamiento desde el planteamiento del proyecto hasta su conclusión.

Agradezco a mi director de tesis Dr. Pablo A. Cipriotti muchas gracias por la confianza, valiosísimos aportes y por enseñarme el valor de trabajar de manera independiente y de ver todo con ojos críticos, siempre desde lo constructivo. Estuvo atento desde la distancia como un apoyo incondicional tanto en la fase de proyecto como en la culminación de esta investigación, gracias por su paciencia y todo su tiempo invertido.

A mis Jefes y compañeros de trabajo tanto en Buenos Aires con la empresa RS reliable que siempre estuvieron dispuestos para que avanzara en mis cursos brindando espacios y disponibilidad, como a mis Jefes y compañeros en Cali Colombia en las empresas Indra - Celsia que han impulsado a la culminación de esta etapa y que con sus palabras de aliento me apoyaron en estos años.

A los miembros de la Escuela para Graduados Alberto Soriano, donde me formé durante esta etapa, especialmente a mis profesores y compañeros. Una mención especial al Dr Rodolfo Cantet que fue un apoyo y guía fundamental desde el primer momento en que decidí aventurarme en cursar esta maestría, tanto en el ámbito académico como en lo personal fue clave su apoyo.

Un especial agradecimiento Carolina García que ha sido una conexión con Buenos Aires desde que volví a Colombia y sin su ayuda la interacción hubiera sido casi imposible. A Sebastián Munilla ya que su guía en los cursos y apoyo en Buenos Aires fue clave para integrarme en la facultad.

A todos los amigos y conocidos que fueron parte de nuestra estadía en Buenos Aires en especial a Pilar, Diana, Carolina, Oscar, Juan David, José Luis, Natalia quienes en algún momento fueron como nuestra familia, gracias por los agradables momentos vividos durante este tiempo.

Finalmente pero no por esto menos importante a mi compañera de travesía, hoy mi compañera de vida Maria Isabel que con su apoyo constante e incondicional ha hecho que sea posible que avancemos juntos tanto personalmente como profesionalmente.

Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.

Camilo Alberto Herrera Rozo.

Índice

Índice de tablas

Índice de figuras

Resumen

La revolución de la informática ha impulsado nuevos desarrollos que mejoran la capacidad de tomar decisiones en agricultura. La caña de azúcar es considerada como el cultivo agrícola más importante del planeta según la Unesco. Conociendo la importancia del cultivo y las nuevas posibilidades computacionales, se generó una metodología para el procesamiento de la información disponible, la cual incluía imágenes multiespectrales, información de características edáficas e información muestreada de las variables a estimar. Con los datos estandarizados se evaluaron y ajustaron cuatro algoritmos de aprendizaje automático (BaggedCART, PLS, Random Forest y Cubist) para la predicción de índices de área foliar (IAF) y rendimiento de cultivo en toneladas de caña por hectárea (TCH). En el proceso de evaluación de los algoritmos se usaron estrategias de big data para lograr procesar y evaluar múltiples parámetros en búsqueda del mejor modelo tanto a nivel de cada algoritmo evaluado como a nivel global buscando el mejor de los algoritmos para las variables estimadas. El mejor ajuste se obtuvo con el algoritmo de random forest, tanto para la predicción de IAF como la predicción del TCH en términos de precisión de la respuesta. Igualmente, este algoritmo presentó los mejores tiempos de ajuste y entrenamiento de los modelos. Los resultados en cuanto a la precisión y modelado tanto de IAF como de TCH a partir de algoritmos de aprendizaje automático indican que los modelos propuestos pueden usarse para predecir estas variables asociadas a las producción de un cultivo y ayudar en el análisis de cultivos con datos fáciles de recolectar y con las nuevas tecnologías disponibles para ayudar en la toma de decisiones en la agroindustria de la caña de azúcar.

Palabras clave:

BaggedCART, Random Forest, Regresión, Cubist, IAF, Metodología, Caña de Azúcar, Predicción, Producción, Aprendizaje Automático.

Abstract

The computer science revolution has driven new developments that improve the ability to make decisions in agriculture. Sugarcane is considered the most important agricultural crop in the world according to Unesco. Knowing the importance of the crop and the new computational possibilities, a methodology for the processing of the available information was generated, which included multispectral images, information of edaphic characteristics and sampled information of the variables to be estimated. With the standardized data, four machine learning algorithms (BaggedCART, PLS, Random Forest and Cubist) were evaluated and adjusted for the prediction of leaf area indexes (LAI) and crop yield in tons of cane per hectare (TCH). In the process of evaluating the algorithms, big data strategies were used to process and evaluate multiple parameters in search of the best model both at the level of each evaluated algorithm and at the global level, looking for the best algorithms for the estimated variables. The best fit was obtained with the random forest algorithm, both for the prediction of IAF and the prediction of the TCH in terms of accuracy of the response. Likewise, this algorithm presented the best adjustment and training times for the models. The results regarding the accuracy and modeling of both IAF and TCH from machine learning algorithms indicate that the proposed models can be used to predict these variables associated with the production of a crop and help in the analysis of crops with easy data. to collect and with the new technologies available to help in decision-making in the sugarcane agroindustry.

Key words:

BaggedCART, Random Forest, Regression, Cubist, IAF, Methodology, Sugarcane, Prediction, Production, Machine Learning.

1. Introducción

1.1. La Revolución Informática en el Agro

En los últimos años se han venido dando una cantidad importante de adelantos en el agro a nivel mundial. Estos adelantos han sido impulsados por nuevos desarrollos en distintas tecnologías empleadas para la captura de información. Conjuntamente, desde la informática se han mejorado las capacidades disponibles para procesar grandes volúmenes de datos. El foco de este desarrollo en el agro se ha centrado en los sistemas de agricultura de precisión llevando a tecnologías robotizadas, automatizadas y con inteligencia artificial.

En agricultura como en otras ciencias aplicadas, generar datos puede mejorar la toma de decisiones. Debido a esto, de manera constante se está buscando innovar en las fuentes de información desde el software, dispositivos de captura de información, maquinaria agrícola, prácticas de manejo de cultivo, análisis de suelos, imágenes satelitales, modelos digitales de elevación, cartografía digital de suelos, estaciones meteorológicas, entre otras (**?**).

Actualmente existe una disponibilidad de datos de gran magnitud y variedad sin precedentes que están generando incertidumbre sobre su valor entre los productores agropecuarios. Lo que preocupa, no solo es que la agricultura esté sumergida en "océanos" de datos, sino que estos están creciendo cada vez más rápido. La cantidad de datos que existe es tan grande que los métodos tradicionales de almacenamiento, procesamiento y análisis parecen ser insuficientes para aprovecharlos y potencializar su valor para el productor.

En 2015 el INTA en Argentina mostró cómo se pasó de tener un dato por hectárea, a mapas de rendimientos que muestran entre 400, 600 y hasta 800 datos por hectárea (**?**). "El campo se puede ver en resolución de medio metro por medio metro, incluso menos. Para todo esto, ya hay máquinas preparadas en pequeña escala. Lo que se viene, es llevar estos procesos a grandes superficies o cultivos extensivos, bajo sistemas que tienen inteligencia artificial mediante una fijación de criterios de trabajo por productores o técnicos capacitados" (**?**).

En la era de los datos analógicos, la cual finalizó en la década del 2000, el análisis y procesamiento de los datos en la agricultura se pensaba dentro de un panorama de constante escasez de información, ya que normalmente la recopilación de datos era costosa y consumía mucho tiempo. En contraste, la era digital ha permitido avanzar de manera eficiente en los procedimientos para generar, almacenar y analizar datos (**?**). La recopilación de datos, que décadas atrás podía demorar días o meses, ahora se puede obtener en minutos, de diversas fuentes y en formato digital. Estos datos digitales pueden

ser leídos, almacenados y procesados automáticamente con ayuda de potentes computadoras.

Los nuevos desarrollos disponibles actualmente para aplicaciones agrícolas abren las puertas al procesamiento de la información en tiempo real. Estas aplicaciones permiten crear bases de datos agronómicas que se pueden acceder incluso a través de teléfonos inteligentes por parte de los usuarios. Esto ocasiona el nacimiento de una nueva agricultura basada en el procesamiento y transformación del producto primario (los datos), junto a las tecnologías de procesos industriales.

"No existe un término adecuado para describir lo que está sucediendo, pero si puede servir de ayuda, es posible que estemos ante la era de la "datificación digital de la agricultura". Esta datificación puede alterar significativamente la manera en que generamos conocimiento en la agricultura y puede ser una fuente de nuevas oportunidades. Aprovechar los agro datos digitales masivos, o "Agro Big Data" como se les conoce, ha sido considerada la revolución tecnológica más importante de la agricultura para el siglo XXI " (**?**). Esta revolución informática es transversal para todo el sector del agro, a nivel mundial uno de los sectores con gran interés en hacer uso de estas capacidades es el sector agro-industrial del cultivo de caña de azúcar.

1.2. El Cultivo de la Caña de Azúcar

La Organización de las Naciones Unidas para la educación, la ciencia y la cultura (Unesco), considera la caña de azúcar el cultivo agrícola más importante del planeta, esto debido a que su principal producto, el azúcar, proporciona un alto porcentaje de calorías para la población mundial, además que los subproductos de este cultivo son altamente utilizables. El cultivo de la caña de azúcar también cumple un rol fundamental en la economía mundial debido a su valor económico e importante dinámica en el comercio internacional, siendo una mercancía clave en las exportaciones de países en desarrollo, por lo cual muchas economías alrededor del mundo interactúan socioeconómicamente alrededor de este.

La caña de azúcar (*Saccharum spp)* descrita en **?** es un cultivo de zonas tropicales y subtropicales que se propaga mediante la plantación de esquejes de caña, donde de cada nudo sale una planta nueva e idéntica a la original. Este cultivo es considerado un recurso natural renovable, ya que se puede restaurar por procesos naturales a una velocidad superior al consumo por parte de los seres humanos, este cultivo es fuente de azúcar, biocombustible, fibra, fertilizante y muchos otros productos y subproductos con sustentabilidad ecológica.

Los principales subproductos de la industria azucarera son el bagazo y las melazas. Las melazas, el principal subproducto, es la materia prima para las industrias del alcohol y sus derivadas, actualmente el exceso de bagazo es usado como materia prima para la industria del papel. Además,

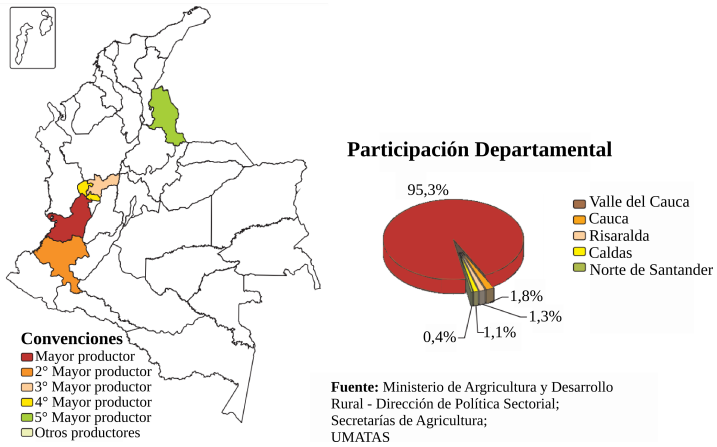


Figura 1.1: Mapa de la distribución del cultivo de caña de azúcar por departamentos en Colombia y su participación porcentual

en la mayoría de los molinos azucareros es factible cogenerar energía usando el bagazo de caña como combustible (?).

A nivel mundial, la producción anual de caña de azúcar es de casi 1.700 millones de toneladas y abarca un área de 24 millones de ha. El mayor productor es Brasil, que con 720 millones de toneladas genera más del 40 % de la producción mundial. Esto sumado a las cifras de India y China da como resultado que los tres países son responsables de dos tercios de la producción mundial de caña de azúcar en un área de casi 15 millones de hectáreas.

Los factores climáticos, en particular el abastecimiento de agua, son los que más influyen en la producción del cultivo. Mientras que el rendimiento promedio de la caña de azúcar en el mundo es cercano a las 60t/ha, algunos países tienen una producción promedio de 100t/ha o más. Dentro de los grandes productores que generan más de 20 millones de toneladas cada año, Colombia, Argentina, Australia, Filipinas y Brasil suelen tener rendimientos promedio de 80t/ha o más.

El sector azucarero colombiano se encuentra ubicado principalmente en el valle geográfico del río Cauca, que abarca 47 municipios desde el norte del departamento del Cauca, la franja central del Valle del Cauca, hasta el sur del departamento de Risaralda. En esta región hay 225.560 hectáreas sembradas con caña para azúcar, de las cuales, el 25 % corresponde a tierras propias de los ingenios y el restante 75 % a más de 2.750 cultivadores de caña. Dichos cultivadores abastecen a 13 ingenios de la región (Cabaña, Carmelita, Manuelita, María Luisa, Mayagüez, Pichichí, Risaralda, Sancarlos, Tumaco, Ríopaila-Castilla, Incauca y Providencia). Desde 2005, cinco de los trece ingenios tienen destilerías anexas para la producción de alcohol carburante (Incauca, Manuelita, Providencia, Mayagüez y Risaralda) (?), en los últimos años el cultivo se ha extendido también en el departamento de Norte de Santander aun con una baja participación como se observa en la Figura ??.

Gracias al clima privilegiado de la región, en Colombia se cuenta con las condiciones ideales para el

crecimiento de la caña de azúcar, radiación solar permanente e intensa a lo largo del año, amplitud adecuada de temperatura entre el día y la noche, disponibilidad de agua, lluvias adecuadas y fertilidad en los suelos, y al contrario de lo que sucede en el resto del mundo (con excepción de Hawaii y el norte de Perú), se puede plantar y cosechar caña durante todos los meses del año. Esta condición agroclimática, sumada al avance tecnológico impulsado por el Centro de Investigación de la Caña de Azúcar (Cenicaña), que funciona con el aporte de todos los cultivadores e ingenios, ha llevado a que la región se especialice en el cultivo y ostente el liderazgo en productividad a nivel mundial: más de 14 toneladas de azúcar procesada por hectárea al año (TAH). Actualmente se producen en Colombia alrededor de dos millones cuatrocientas mil toneladas anuales de azúcar, de las cuales se exporta cerca del 41 %, lo que convierte a Colombia en un oferente de tamaño medio dentro de los estándares del mercado internacional del azúcar (**?**).

Existen macro-nutrientes y micro-nutrientes fundamentales para el desarrollo del cultivo de caña de azúcar, los suelos aportan un gran porcentaje de estos nutrientes, ya que la caña de azúcar crece bien en diferentes tipos de suelos, pero prefiere los suelos francos o franco-arcillosos, bien drenados y profundos. El pH óptimo para su desarrollo es de 6,5 (ligeramente ácido), aunque tolera desde suelos ácidos hasta suelos alcalinos (**?**), con pH próximo o menor de 4,5, la acidez del suelo limita la producción del cultivo, principalmente por la presencia de aluminio intercambiable y de algunos micro-nutrimentos en exceso como hierro y manganeso que pueden ocasionar toxicidad y muerte de la planta (**?**), es importante conocer la importancia de estos nutrientes debido a que se deseamos evaluar la capacidad de relacionar éstos con los rendimientos finales de los cultivos para mejorar el conocimiento de los procesos naturales vinculados al rendimiento de este cultivo.

Existen 16 elementos nutritivos esenciales para la caña de azúcar: el carbono, el hidrógeno y el oxígeno no son minerales y la planta los toma del dióxido de carbono y del agua. El nitrógeno es necesario en grandes cantidades y ayuda a maximizar la producción de materia seca y el rendimiento. La demanda de nitrógeno es máxima durante el macollaje y en el período de gran crecimiento. El fósforo es necesario temprano en el desarrollo de la planta para asegurar un buen crecimiento de las raíces y estimular el macollaje. El potasio es necesario en grandes cantidades, en niveles mayores a los del nitrógeno. La mayor parte de este potasio se usa en el tallo de la caña y la máxima demanda de potasio ocurre durante el período de gran crecimiento en el cual la absorción es mayor que para cualquier otro nutriente. En la etapa temprana del ciclo de vida del cultivo se absorbe una cantidad significativa de calcio, lo que juega un papel importante en el enraizamiento y en la integridad de las células de la planta. En comparación, el azufre y el magnesio se absorben más lentamente y en estadios de crecimiento posteriores, ambos son importantes para la calidad del azúcar y las proporciones son mayores en el tallo de la planta que en otras partes.

Finalmente, los micronutrientes clave que se absorben en mayores cantidades son el hierro y el

manganeso, estos aseguran un crecimiento libre de estrés, mejoran el desempeño fotosintético y el rendimiento. El boro y el zinc, aunque son necesarios en cantidades menores juegan un papel específico en fomentar el crecimiento de tejidos nuevos en la planta, otros micronutrientes necesarios son el cloro, cobre y molibdeno; estos últimos, aunque son necesarios para el normal desarrollo de la planta, se requieren en cantidades muy pequeñas (**?**), lo anterior muestra la importancia de incluir información respecto estos componentes presentes en el suelo donde se realizó el experimento. Para más detalles de los aportes de los componentes presentes en el suelo para el cultivo de caña de azúcar ver **??**.

Un cultivo experimental posee unas características únicas que permite observar la variación del comportamiento de un cultivo según un grupo de características que se pueden controlar, en particular poder controlar el tamaño del lote, el manejo que se realiza del suelo, la densidad de siembra, sobretodo el seguimiento que se le puede dar al cultivo durante su crecimiento entre otros factores permite mejorar las conclusiones respecto a algún fenómenos que se esté estudiando, por lo que es común hacer uso de estos experimentos para la validación minimizar el posible error asociado a estos factores.

1.3. Datos Masivos - «Big Data»

Los datos se han convertido en un torrente que fluye en todas las áreas de la economía global, las empresas producen un creciente volumen de datos transaccionales, capturando miles de millones de bytes de información sobre sus clientes, proveedores, operadores y productos. Millones de sensores en red están siendo incorporados en el mundo físico en dispositivos tales como teléfonos móviles, medidores de energía inteligentes, automóviles y máquinas industriales que crean y comunican datos de manera constante actualmente en la era del internet de las cosas.

Entonces como las empresas y organizaciones interactúan con individuos y productos, están generando una enorme cantidad de datos digitales, muchos de estos datos son creados por ejemplo como subproductos de otras actividades (**?**), de la misma manera ocurre hoy día en el agro donde cada día se posee más cantidad de datos disponibles para analizar y ser aprovechados por parte de la industria.

En el caso del cultivo de caña de azúcar se puede ver que tanto a nivel local como a nivel mundial existen extensiones importantes del cultivo como nuevas tecnologías que permiten tener un mayor detalle en la información que puede recolectarse, esto hace que la cantidad de información que es posible obtener para el mejoramiento del cultivo sea un reto importante tanto para su uso como para el manejo de dicha información. Esta información puede ser aprovechada por las nuevas

tendencias aplicadas en el agro, una de estas tendencias es lo que se denomina como "big data" o datos masivos, donde se requiere de herramientas que permitan obtener beneficios de los datos digitales de la agricultura de una manera eficiente.

En el manejo de estos datos se pueden presentar problemas técnicos relacionados con **volumen** de los datos, **variedad** de las fuentes de información, **velocidad** de procesamiento o **veracidad** de los resultados, considerados como las 4 V´s ("cuatro V") del big data (**?**). En muchas ocasiones estos problemas tienen solución desde la informática, por medio de infraestructura adecuada para el procesamiento de la información y sobre todo por la aplicación de métodos adecuados para lograr extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas (estructurados o no estructurados) y en tiempos adecuados.

Tecnología y técnicas, probablemente son la manera para capturar valor de estos datos masivos, las organizaciones deben desplegar nuevas tecnologías, por ejemplo (almacenamiento, informática y software analítico) y técnicas (es decir generar nuevos tipos de análisis mediante estas herramientas). La gama de retos tecnológicos y las prioridades establecidas para abordarlos varían dependiendo de la madurez de los datos en las instituciones. Los nuevos problemas y el creciente poder informático impulsarán el desarrollo de nuevas técnicas analíticas, también hay una necesidad de innovación continua en tecnologías y técnicas que ayuden a individuos y organizaciones a integrar, analizar y visualizar consecuentemente el creciente torrente de datos.

Aunque técnicas y tecnología se refieren a términos similares estos tienen un significado distinto. La técnica es un concepto intangible relacionado al «know-how» o conocimiento para sacar el máximo provecho de una tecnología. Por otro lado la tecnología se refiere al conjunto de equipos que permiten realizar alguna actividad aplicando dichas técnicas o conocimiento, esto es un concepto más tangible y permite aplicar las técnicas, ya sean estas soluciones tecnológicas o equipos requeridos para aplicar las técnicas disponibles.

Se han desarrollado y adaptado una amplia variedad de técnicas y tecnologías para agrupar, manipular, analizar y visualizar grandes volúmenes de datos. Estas técnicas y tecnologías se basan en varias disciplinas científicas, entre ellos la estadística, la informática, las matemáticas aplicadas y hasta la economía. Esto significa que una organización que tiene la intención de obtener valor del "big data" tiene que adoptar un enfoque flexible y multidisciplinario. Algunas técnicas y tecnologías se desarrollaron en un mundo con acceso a volúmenes mucho más pequeños y variedad de datos escasa, pero se han adaptado con éxito para que sean aplicables a conjuntos de datos diversos, procesamiento o simulación de información que suele generar grandes cantidades de datos.

El "big data" no está ligado exclusivamente a grandes volúmenes de datos involucrados al principio de un análisis, este es un error de concepto muy común. Esta no es una características excluyente para que un problema se catalogue como que requiere el uso de técnicas de big data, algunos pro-

blemas pueden requerir abordarse mediante estas técnicas debido a características de los procesos a realizar con los datos originales que involucra un crecimiento iterativo de información, esto es muy común en procesos de simulación, inclusive en proceso de validación donde se deben realizar constantes iteraciones para la búsqueda y optimización de los algoritmos (?), es muy común y requerido hacer uso de estos procesos que permitan partir el problema, almacenar datos y optimizar procesos que son computacionalmente exhaustivos, de manera que logren menores tiempos y procesamientos computacionalmente más eficientes.

1.4. Datos Derivados de Imágenes

Captar datos no es algo nuevo, lo que realmente ha avanzado es la cantidad de datos que se captura en diversos formatos, lo que posibilita mezclar y procesar distintas fuentes de datos para obtener información valiosa de los cultivos, las imágenes son una de estas fuentes de información potencial que hoy se puede aprovechar. Ya hace más de 60 años que se toman fotografías a color e infrarrojo para seguir el crecimiento de plantas (?). En la actualidad, estos métodos están siendo revaluados para realizar análisis de la variabilidad espacial en la agricultura de precisión, ya que las imágenes aéreas se pueden adquirir rápidamente durante los períodos críticos del crecimiento de las plantas (?).

Existen una infinidad de métodos para la evaluación del crecimiento y desarrollo de los cultivos, entre estos métodos encontramos nuevas aplicaciones como la percepción remota o teledetección, la cual se define como el grupo de técnicas para la obtención de información sobre las propiedades físicas de ciertas superficies u objetos y su entorno, desde distancias relativamente grandes, sin contacto físico con ellos (?). Las imágenes adquiridas por sensores desde plataformas aéreas o satelitales en el mundo agropecuario tienen un potencial que se ha venido explorando con mayor énfasis en las últimas décadas (?).

La agricultura de precisión y el manejo sitio-específico (AEPS), se define como el arte de realizar las prácticas agronómicas requeridas por una especie vegetal, de acuerdo con las condiciones espaciales y temporales del sitio donde se cultiva, para obtener de ellas el máximo rendimiento que puede alcanzarse de un cultivo (?).

La medición de variables biofísicas como área foliar (AF) y el índice de área foliar (IAF) en cultivos de caña de azúcar en Colombia se realiza en la actualidad por métodos destructivos directos sobre superficies de muestreo pequeñas. El resultado de producción expresado por la cantidad de toneladas de caña por hectárea (TCH) es otro resultado para medir la producción en un cultivo y este solo se puede medir hasta la etapa final del cultivo (cosecha), sin poder tener predicciones precisas

con anterioridad. Estas características plantean un panorama interesante de investigación con la finalidad de utilizar nuevos métodos de predicción, sobre áreas de mayor tamaño (exhaustivos) y con mayor precisión, para los cuales la estadística y los modelos de aprendizaje automático como herramienta de investigación y análisis puede ser una ayuda invaluable.

Usar métodos para adquirir, procesar y analizar imágenes del mundo real con el fin de producir información es una disciplina denominada visión artificial (**?**), al igual que los humanos usamos nuestros ojos y cerebro para comprender el mundo que nos rodea, la visión artificial intenta producir el mismo efecto para que la información pueda ser entendida y asimilada de manera correcta para sacar conclusiones de una escena o imagen disponible como abstracción de los datos.

El índice de área foliar (IAF) es el cociente entre el área foliar (AF) y la unidad de superficie de suelo (**?**). La información precisa y oportuna sobre el IAF tiene gran importancia y aplicaciones en la agricultura para la predicción del rendimiento y la evaluación de estrés en distintos cultivos, y en la ecología para el estudio de la producción primaria y el cambio ambiental (**?**).

El valor máximo del IAF encontrado en caña de azúcar es de 8 (**?**). En Colombia las mediciones preliminares indican que los valores de IAF varían entre 4 y 7, en cultivos de 8 a 9 meses de edad. Las principales aplicaciones de las técnicas de teledetección están dentro de los campos de la inteligencia agrícola, la gestión agrícola y la investigación ecológica (**?**).

Los índices de vegetación son combinaciones de distintas bandas espectrales mediante operaciones algebraicas básicas, la función de estos índices de vegetación es realzar la cubierta vegetal en función de su respuesta espectral y atenuar los detalles de otros componentes como el suelo o la iluminación, etc. (**?**)

Desde los inicios de la percepción remota los índices espectrales de vegetación han sido útiles y fáciles de calcular para relacionarlos con diversas variables agronómicas, aunque los índices espectrales de vegetación en muchos casos muestran excelentes relaciones con estas variables, es necesario calibrar o comprender la equivalencia de sus valores en la predicción de contenidos de clorofila, IAF o biomasa (**?**).

A nivel mundial países como Francia y Brasil han realizado trabajos para estimar algunos parámetros biofísicos e inclusive para pronosticar la producción de caña de azúcar. En Francia **?** haciendo uso del índice de vegetación normalizado (NDVI) en cultivos con variabilidad espacial (independiente a las etapas de crecimiento de los cultivos), demostraron que a una escala estacional, el patrón de crecimiento dentro de un campo depende de la etapa fenológica del cultivo mientras que a escala anual los mapas NDVI revelaron patrones estables. Lo anterior permite concluir que es necesario conocer el ciclo de crecimiento del cultivo para interpretar correctamente los patrones espaciales y usar imágenes de una fecha única podría ser insuficiente para el diagnóstico de la situación de los cultivos o para aplicaciones en predicción.

En ? se evaluaron las calificaciones visuales subjetivas de crecimiento del cultivo de la caña de azúcar para la predicción de parámetros de rendimiento del cultivo y paralelamente se realizaron mediciones de los IAF. En este caso se encontró que existen relaciones importantes entre estos sistemas de evaluación visual, los índices IAF y la predicción de la población, pero estas relaciones son válidas sólo en algunos estados del crecimiento del cultivo. Por lo tanto, las predicciones no eran buenas fuera de algunos periodos del desarrollo del cultivo, sobre todo en las primeras etapas del crecimiento, lo que impide tomar decisiones tempranas respecto al manejo del cultivo.

En Brasil ? propusieron un método para realizar predicción del rendimiento sobre cultivos de caña de azúcar usando índices de vegetación espectral, mediante análisis de componentes principales e información histórica de los cultivos. Para dicho estudio se utilizaron imágenes (ETM +) / Landsat-7 e imágenes ASTER/Terra. Este método comprende varias etapas y permite una síntesis de la información tanto de la imagen como de la información histórica, normalizando todas las variables en conjunto, y haciendo posible expresar todos los datos en imágenes síntesis.

Sudáfrica es el líder en producción de caña de azúcar en África y uno de los más grandes productores del mundo, ahí se realizó el monitoreo de factores de estrés del cultivo de caña y determinaron que es de vital importancia para tomar acciones preventivas y de mitigación sobre el cultivo. ? exploró el potencial de usar sensores remotos en cultivos de caña de azúcar, mediante el uso de imágenes Landsat TM y ETM+ con las cuales generó modelos de predicción de rendimiento de caña aplicando algoritmos de random forest optimizados, logrando una predicción buena para algunas de las variedades estudiadas.

Posteriormente (?) exploro el uso de algoritmos de random forest en caña de azúcar haciendo uso de imagenes hyperespectrales para estimar concentración de nitrógeno en las hojas concluyendo un uso potencial de este algoritmo para dicha predicción.

En ? se realizaron análisis de correlación del IAF forestal con 12 índices de vegetación extraídos de las reflectancias en imágenes Hyperion, luego correlacionaron cada uno de los índices con mediciones de IAF en el campo. Los resultados indican que muchas bandas multiespectrales en la región SWIR y algunas en la región NIR tienen el mayor potencial en la formación de índices para la predicción IAF. Las longitudes de onda de banda más eficaces se centraron cerca de 820, 1040, 1200, 1250, 1650, 2100 y 2260 nm. Índices de vegetación derivados de las bandas R e NIR no produjeron correlaciones tan altas con IAF como aquellas con bandas en las regiones SWIR y NIR. Basándose en su alta correlación con las mediciones de IAF, SR y NDVI se recomienda para la predicción de IAF utilizando imágenes basadas en datos multiespectrales.

Recientemente ? explora la posibilidad de predicción del IAF mediante imágenes multiespectrales obtenidas con radioespectrómetro de campo (GER 3700). Para la búsqueda del mejor modelo hizo uso de PLS y RandomForest con resultados interesantes donde se observaron resultados más

prometedores mediante el modelo de random-forest donde se mejora el R^2 y se minimizaba el RMSE.

En **?** se evaluó el rendimiento de cuatro máquinas de aprendizaje automático para la predicción del IAF a través de series de tiempo producto de imágenes MODIS con el objetivo de completar información incompleta en las series. El estudio mostró que al crecer el tamaño de la muestra para todos los métodos evaluados, lo que se lograba era la estabilización del R^2 y del RMSE sin variaciones muy altas después de los 3000 datos ingresados para el entrenamiento de los modelos. Adicionalmente, se mostró que el tiempo de entrenamiento y la rapidez de convergencia de los modelos dependía de su estructura interna.

En (**?**) se presenta como ha sido el interés creciente en el uso de vehículos aéreos no tripulados, drones o UAV , con aplicaciones en cultivos agrícolas en Colombia, lo que en gran medida se debe a los grandes beneficios que se obtienen mediante la implementación de estos sistemas.

(**?**) haciendo uso de estos vehículos se generaron mosaicos de imágenes infrarrojas con los cuales fue posible calcular algunos índices de vegetación como el NDVI, estos índices se usaron para estimar el contenido de biomasa en edades tempranas de cultivos de caña de azúcar, donde se encontraron correlaciones de hasta un $R^2 = 0.70$ es importante aclarar que los autores recomiendan para futuras aplicaciones se hace necesario calibración radiométrica para convertir los valores de respuesta de las imágenes a valores de reluctancia de una manera adecuada.

Finalmente en (**?**) se explora la predicción de la productividad en caña de azúcar (TCH) desde la visión de la percepción remota, en esta publicación se presentan desarrollos asociados en dos distintas áreas de estudio con distintos tipos de sensores, el segundo de estos estudio hace uso de la misma imagen multiespectral base utilizada en el presente estudio al igual que la respuesta de la productividad, en éste estudio se realizó el uso las bandas espectrales e índices de vegetación de manera directa, donde se observa que haciendo uso de regresiones lineales múltiples los resultados para modelar el TCH son pobres o al menos no generan buenos resultados en términos de coeficientes de determinación, por otro lado plantea una primera aproximación al modelado de datos obtenidos mediante percepción remota.

1.5. Técnicas de Análisis

Existen muchas técnicas estadísticas que el agricultor puede aprovechar para la predicción de variables del cultivo, entre estas está el aprendizaje automático que es una rama de la Inteligencia Artificial (IA) cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender (**?**). De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Una máquina es un sistema

organizado capaz de transformar cierto mensaje de entrada en otro de salida, de acuerdo con algún principio de transformación (?). Si tal principio está sujeto a cierto criterio de validez de funcionamiento, y si el método de transformación se ajusta a fin de que se tienda a mejorar el funcionamiento del sistema de acuerdo con este criterio, se dice que el sistema aprende, estas máquinas pueden proporcionar soluciones a distintos problemas en sistemas agrícolas complejos de manera eficaz (?).

En ? se define a las máquinas de aprendizaje o el aprendizaje automático como el campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas siguiendo ciertos criterios asignados con la finalidad de predecir hechos futuros, realizar recomendaciones, clasificaciones de elementos, eventos, tags, etc. Esto sería un gran avance en las técnicas de predicción en cultivos de caña de azúcar, sobre todo cuando la información proviene de imágenes digitales complementada por información de campo, proceso que genera un flujo grande de información, el cual requiere métodos estadísticos capaces de tratar con estructuras de datos multivariados y no lineales.

Las máquinas de aprendizaje están relacionadas directamente a la estadística computacional solo que con un enfoque a la realización de predicciones mediante del uso de computadoras, las máquinas de aprendizaje están fuertemente ligadas a la optimización matemática. Se pueden separar en dos áreas principales denominadas aprendizaje supervisado y aprendizaje no supervisado, el primero está más enfocado la predicción de resultados futuros cuando se tiene conocimiento previo de las relaciones que se desean modelar, a partir de ejemplos previamente etiquetados, mientras en el aprendizaje no supervisado este está más alineado al análisis exploratorio de datos, donde los que se busca es reconocer patrones en datos que no han sido etiquetados previamente (?).

Las herramientas matemáticas basadas en la minería de datos y máquinas de aprendizaje proporcionan un marco adecuado para la extracción de información útil a partir de grandes bases de datos, así como también pueden conducir al descubrimiento de conocimiento (?).

La regresión por mínimos cuadrados parciales o PLS, árboles de clasificación y regresión usando BaggedCART, el algoritmo de random forest o bosques aleatorios y los árboles de aprendizaje empírico (Cubist) son metodologías de mucho interés para el desarrollo de esta investigación pues aunque son métodos cada uno desarrollado para aplicaciones originales diferentes, a priori son altamente escalables y previamente no se han probado sobre cultivos de caña para los objetivos desarrollados en esta investigación.

En la actualidad existe una infinidad de métodos de aprendizaje automático que se podrían evaluar, publicados y disponibles existen más de 450 distintos tipos de modelos, los cuales podríamos probar en este proceso, desde el modelado clásico, fundamentados en supuestos de distribución, pasando por un modelado basado en kernels hasta llegar a planteamientos mucho más cercanos a la tendencia

actual en procesamiento de imágenes como lo sería modelado mediante redes neuronales profundas o "Deep Learning".

Sin embargo desde la fase de planteamiento de esta investigación se pretendía abordar dos enfoques con los datos disponibles, el primero un poco más clásico, a partir de PLS algo que es más cercano al modelado multivariado asociado a proyecciones de matrices y el segundo lado es abordar un enfoque a los arboles de regresión y distintas variaciones de estos, donde se comenzaba desde lo más general hasta lo particular, iniciando con un modelo básico de árbol que incluye ajustes automáticos en su aprendizaje, pasando por modelos de árbol más complejos enfocados en conjuntos de modelos en búsqueda del consenso, hasta llegar a modelos de árboles que vienen más desde la algoritmia que desde la estadística y han sido poco explorados en agricultura con lo es Cubist.

Es de destacar que las metodologías de modelamiento mediante árboles y conjuntos de modelos para predicción son es un enfoque ampliamente utilizado a nivel mundial al punto de que gran cantidad de las competencias del prestigioso sitio `www.kaggle.com` tiene como factor común el uso de este tipo de modelado con rendimientos que pocos modelos logran mejorar.

A continuación se presenta brevemente cada uno de los modelos a evaluar, para posteriormente desarrollarlos con más detalle en la sección metodológica.

1.5.1. Regresión por Mínimos Cuadrados Parciales (PLS)

La regresión por mínimos cuadrados parciales se introdujo hace casi treinta años y ha tenido un gran desarrollo en áreas como la quimiometría, donde se analizan datos que se caracterizan por muchas variables predictoras, con problemas de multicolinealidad, y pocas unidades experimentales en estudio (**??**). En **?** se propone el método PLS, la idea de Wold era dotar a la práctica estadística de una alternativa analítica para aquellas situaciones en que no se tenían las hipótesis básicas de la modelación estadística.

Es una forma particular de análisis multivariante, relacionado con la regresión de componentes principales (PCR) posee valiosas ventajas teóricas y computacionales que han llevado a innumerables aplicaciones. PLS se utiliza para encontrar las relaciones fundamentales entre dos matrices (X e Y), es decir, un enfoque de variable latente para el modelado de las estructuras de covarianza en estos dos espacios.

Un modelo de PLS trata de encontrar el sentido multidimensional en el espacio X que explica la dirección de la máxima varianza multidimensional en el espacio Y (**?**). Estos métodos tienen ventajas intrínsecas cuando se los compara con métodos univariados. Todas las variables relevantes son incluidas en el modelo PLS. La suposición básica de todos estos modelos es que el sistema o proceso estudiado depende de un número pequeño de variables latentes (V.L.). Este concepto es

similar al de componentes principales. Las variables latentes son estimadas como combinaciones lineales de las variables observadas.

1.5.2. BaggedCART

BaggedCART es la conjunción entre dos metodologías los árboles de clasificación y regresión (CART) (?) y máquinas de aprendizaje Bagging, la cual se refiere a un conjunto de meta-algoritmos diseñado para mejorar la estabilidad y precisión de los algoritmos de aprendizaje automático. También reduce la varianza y ayuda a prevenir el sobreajuste de los modelos, esta metodología puede ser usada junto a cualquier tipo de método, el bagging es un caso especial de los modelos de enfoque de promedios.

En el **algoritmo CART** el resultado es en general, un árbol de decisión, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición), formando así grupos homogéneos respecto a la variable que se desea discriminar (?). Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos históricos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos.

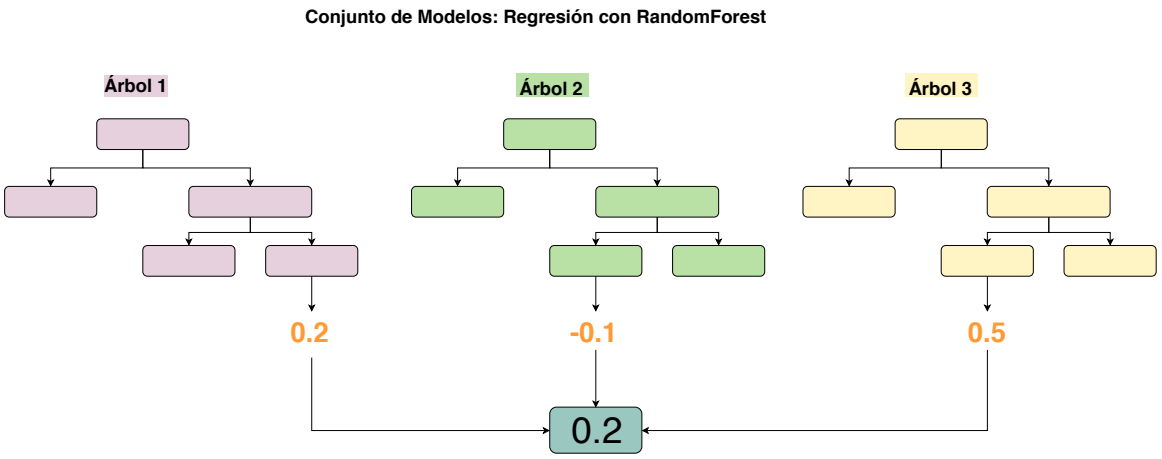
1.5.3. Bosques Aleatorios (Random Forest)

En Breiman (????) se demostró que hay una ganancia sustancial en la precisión en los métodos de clasificación y regresión mediante el uso de un conjunto de bosques donde cada árbol en el conjunto se cultiva de acuerdo a un parámetro aleatorio. Las predicciones finales se obtienen de las agregaciones sobre el conjunto de datos. Como los componentes de base del conjunto son predictores con estructura de árbol, y donde cada uno de estos bosques se construye utilizando una introducción de aleatoriedad, se conoce a estos procedimientos como bosques aleatorios.

El modelo de random forest se considera como un refinamiento de los modelos de BaggedCART, es una combinación de bosques predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de éstos. Esta es una modificación sustancial del bagging [Bootstrap aggregating] que construye una larga colección de árboles no correlacionados y luego los promedia (ver figura ??), en esto se diferencia del método de BaggedCART.

1.5.4. Árboles de Aprendizaje Empírico Cubist

Los árboles de aprendizaje empírico Cubist son una mejora sobre el modelo M5 de Quinlan (?) que se basa en generar modelos predictivos basados en reglas, se expresa como un árbol de decisión en el cual sus hojas terminales contienen modelos de regresión lineal. Este modelo se basa en



Se observan nodos principales y sus particiones, hasta que se llega el nivel más bajo o nodos terminales, a este nivel, lo que se tiene es un valor, el cual es utilizado junto al resto de nodos terminales para ser promediados y entregar el valor final correspondiente a la respuesta de un bosque aleatorio. Usualmente estos bosques contienen cientos de árboles con los cuales se realiza este mismo procedimiento.

Figura 1.2: Promediado de ramas finales usando árboles aleatorios (random forest). Ejemplo para 3 árboles

los predictores utilizados en las decisiones anteriores del árbol. También hay modelos lineales intermedios en cada árbol. Una predicción se realiza mediante el modelo de una regresión lineal en el nodo terminal del árbol, el árbol se reduce a un conjunto de reglas, que inicialmente son caminos de la parte superior del árbol a la parte inferior del árbol, hay reglas que se eliminan a través de la poda o mediante combinación para simplificar el árbol. Detalles sobre este método se pueden encontrar en **??** donde se trató de recrear este modelo usando una reconstrucción racional que es la base para el modelo M5P o Cubist.

En este contexto, para analizar estructuras de datos complejas derivadas de la agricultura de precisión, estos cuatro métodos, PLS o regresión por mínimos cuadrados parciales (**?**), árboles de clasificación y regresión (CART) (**?**), random forest o bosques aleatorios (**?**) y árboles de aprendizaje empírico Cubist (**?**), fueron de particular interés, dado que son herramientas potencialmente importantes de aplicar en casos de aprendizaje en regresión o clasificación estadística (**??**). Estos métodos fueron aplicados a los datos derivados de imágenes multiespectrales (**?**) e información tomada en campo.

1.6. Objetivos

Existe la necesidad cada vez mayor de automatizar y mejorar procesos que son costosos o de características destructivas relacionados a la toma de datos y análisis de éstos, se busca facilitar la toma de decisiones apuntando a sistemas de agricultura de precisión y manejo sitio-específico en cultivos de caña de azúcar con el uso de modelos de aprendizaje automático para predicción.

1.6.1. Objetivo General:

Evaluar modelos de aprendizaje automático integrando datos de imágenes multiespectrales e información tomada en campo para la predicción de variables asociadas a la producción en parcelas experimentales de caña de azúcar.

1.6.2. Objetivos Específicos:

- Establecer una metodología para el procesamiento de datos que provienen de diversas fuentes de información, tanto datos tomados en campo mediante muestreo (suelo y variables asociadas a la producción del cultivo) como imágenes multiespectrales, llevándolos a sistemas de información geográfica en escalas comparables.
- Ajustar modelos de aprendizaje automático para encontrar los modelo que mejores resultados entreguen en términos de predicción y precisión en el modelado, relacionando información de imágenes multiespectrales y datos tomados en campo con variables asociadas a la producción del cultivo de caña de azúcar tales como el índice de área foliar (IAF) y la producción medida en toneladas de caña por hectárea (TCH).
- Evaluar y estimar los parámetros que mejoren la precisión de cuatro modelos de aprendizaje automático, regresión por mínimos cuadrados parciales [PLS], árboles de regresión [BaggedCART], bosques aleatorios [Random Forest] y árboles de aprendizaje empírico [Cubist], aplicados a datos tomados en parcelas experimentales de caña de azúcar.
- Determinar el nivel de importancia de las variables incluidas en la evaluación de los algoritmos de aprendizaje automático para la predicción del índice de área foliar (IAF) en parcelas experimentales de caña de azúcar y la producción medida en toneladas de caña de azúcar por hectárea (TCH).

2. Metodología

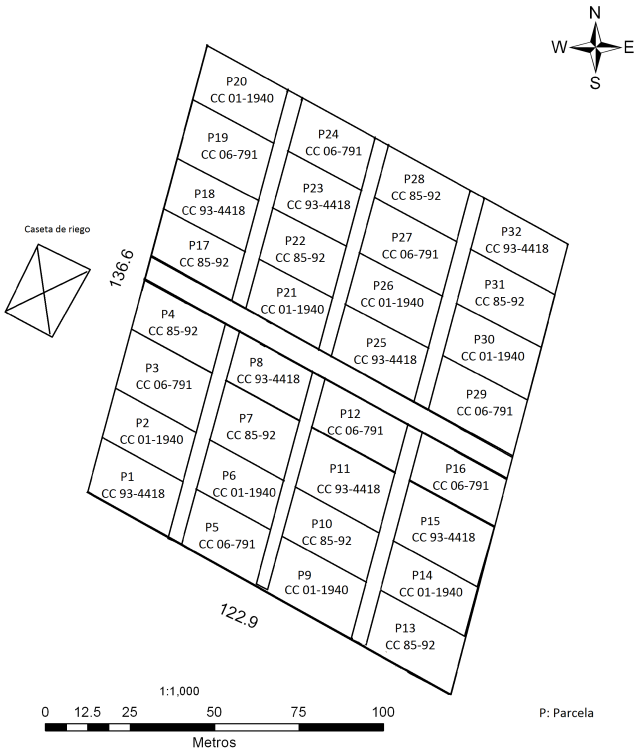
2.1. Descripción del Área de Estudio

Se cuenta con parcelas experimentales de caña de azúcar ubicado en el valle geográfico del río Cauca, Colombia. Este cultivo experimental ubicado en las coordenadas geográficas Latitud 3°21'36.6"N y Longitud: 76°17'48.9"W, se utilizó dada la necesidad de realizar tomas de información sobre cultivos donde se pudieron controlar las distintas fuentes de variación, como lo son los sistemas de riego, el tamaño del lote productivo, la fecha y densidad de siembra, el manejo del suelo, el seguimiento del cultivo durante el crecimiento, y su fertilización, entre otros factores.

De esta manera se limitó la variabilidad para realizar una aproximación de las variables que se desea estimar, esto a lo largo de 32 parcelas distintas. Cada una de estas parcelas posee dimensiones aproximadas 30x15 metros con las distintas variedades de caña (CC 01-1940, CC 06-791, CC 85-92, CC 93-4418) ubicadas de forma aleatoria, y separado en dos bloques de cultivo de acuerdo al tipo de riego, delimitado por un surco principal. En el diseño presentado en la Figura: **??** se tienen las parcelas P1-P16 con fertiriego (fertilización usando como medio de transporte y localización el riego, riego por goteo) y desde la P17-P32 con un sistema de riego convencional (riego por tubería de ventanas), este experimento fue sembrado en el mes de octubre de 2014.

La información disponible del área de estudio, se compone de datos recolectados en distintas instancias en el área de ubicación del cultivo, hasta el momento posterior a la cosecha, manteniendo la identificación de cada parcela y ubicación del dato recolectado en cada caso.

1. Imagen Multiespectral: Esta imagen fue tomada el 7 de abril de 2015, es decir cuando la caña tenía una edad de 7 meses. Se empleó un Drone tipo UAV equipado con una cámara Gamaya Oxi nir-25 de resolución espacial de 2MP (2048 x 1088 px), con un tamaño de píxel equivalente a 35 x 35cm (figura **??**). Esta cámara entrega 25 bandas espectrales (**?**) entre los rangos de 660 y 900 nm. Dada la sobreexposición de algunas bandas, fueron seleccionadas 17 bandas espectrales, las cuales van desde los 666nm hasta los 882nm, estas bandas son estrechas es decir que captan la radiación en un rango de 10 nm, como se describe en la tabla **??**. Este espectro corresponde al área de transición del rojo visible al infrarrojo. La imagen final tiene un tamaño de píxel de 35cm lo que permite tener una resolución superior a la mayoría de los sensores remotos.



Al interior del lote experimental se encuentran las parcelas enumeradas de 1 a 32, adicionalmente se incluye información de la variedad de caña sembrada en cada parcela

Figura 2.1: Esquema de diseño experimental del área de cultivo involucrado denominado (Lote 14 - Estación Experimental Cenicaña).

La cámara Oxi nir-25 esta catalogada como una cámara hiperespectral aunque no posee cientos de bandas su definición como hiperespectral está dado por que los rangos de sus bandas son contiguos https://gamaya.com/smallest-hyperspectral-camera/, aun así como no se consideraron la totalidad de sus bandas y se puede perder en algunos momentos esta continuidad, se hablara de imágenes multiespectrales a lo largo de este documento.

Figura 2.2: Cámara multiespectral utilizada para la toma de información de referencia: Gamaya Oxi nir-25

Banda	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Marca de Clase (nm)	671	680	707	720	733	744	758	770	783	795	815	825	836	847	857	867	877
Longitud de Onda (nm)	666-676	675-685	702-712	715-725	718-738	739-749	753-763	765-775	778-788	790-800	810-820	820-830	831-841	842-852	852-862	862-872	872-882

Tabla 2.1: Detalle de Bandas espectrales disponibles en la imagen multiespectral utilizada

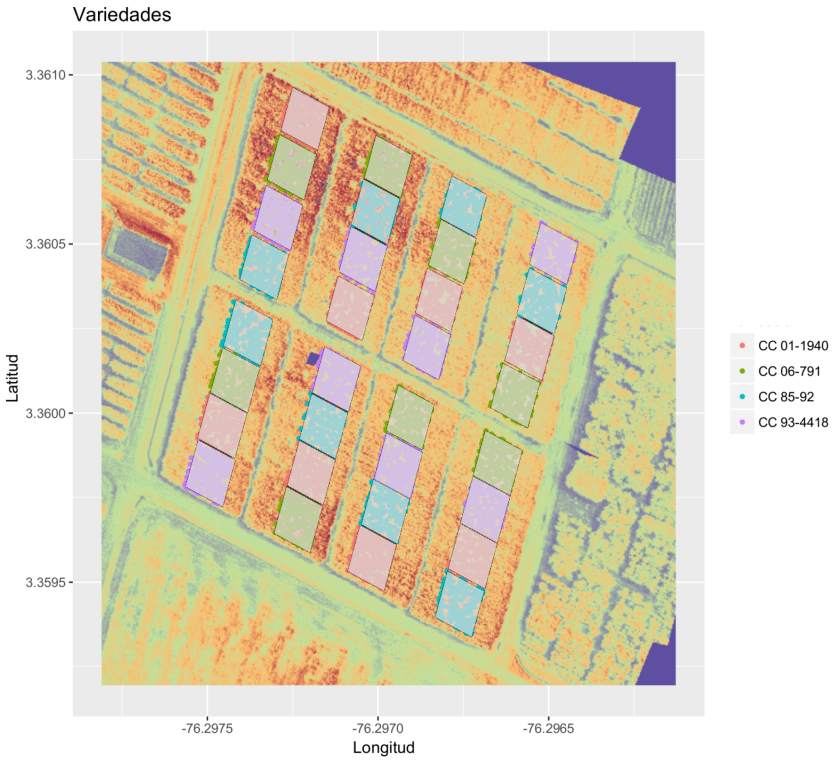
2. Información de variables físico químicas del suelo (Edáficas): Se tomaron 96 muestras del suelo a través de los surcos del experimento, previo a la fecha de siembra del mismo aproximadamente en el mes de Julio de 2014 georeferenciando cada punto de muestreo. Posteriormente fueron procesadas en los laboratorios del centro de investigación, de estas muestras se midieron variables como pH, Mo (Materia Orgánica), P (Fósforo), K (Potasio), Fe (Hierro), Zn (Zinc), B (Boro), Ca (Calcio), Mg (Magnesio), Ca/Mg, Na (Sodio), Mn (Manganeso), Cu (Cobre), Arena, Arcilla, Limo y textura asociada a cada punto muestreado.

3. Información de respuesta del cultivo: Se posee información tomada como respuesta del experimento entre las que se encuentran IAF (índice de área foliar) y el TCH (toneladas de caña por hectárea). En el caso del IAF esta información fue recolectada por el área de fisiología de Cenicaña el día 20 de abril de 2015 y fue consolidada por parcela, por lo tanto solo se posee un dato (valor) por parcela, ya que previamente se recolectó la información por medio de muestras aleatorias sobre cada parcela (10 individuos) sobre los cuales de midió respuesta y se promedió creando un perfil de respuesta común por parcela, en el caso del TCH este se obtuvo en el momento de la cosecha del cultivo en el mes de agosto de 2015.

2.2. Procesamiento de Información

La información tenía distintas escalas, resolución y detalles relativos al área de estudio, por esta razón fue primordial organizar esta información, el proceso principal consistió en llevar todos los datos disponibles a un sistema de información geográfica.

- Paso 1: El proceso de unificación de información se realiza partiendo de la imágen multiespectral en formato raster para tener una referencia del campo de trabajo, esta información básicamente eran 17 bandas espectrales, mediante las cuales se calcularon índices de vegetación. Los índices de vegetación calculados fueron (CIred [Red Chlorophyll], DD[Double Difference Index ?], NDVI[Normalized Difference Vegetation (Normalized Difference 750/705 Chl ND)], SR[Simple Ratio], ZTM[? Index], Lnbr [Low NarrowBand Ratio], MTCI [Meris Terrestrial Chlorophyll Index] y MSR [Modified Simple Ratio]). Estos índices son una fuente de información como entrada para el cálculo de los modelos evaluados. En el caso de la imagen multiespectral base, se realizaron varias adaptaciones de cada uno de estos índices para hacer uso de las bandas espectrales disponibles, se adjunta en apéndice (??) dicha equivalencia. Se descartó el uso directo de las bandas espectrales originales, ya que éstas no entregaron aporte significativo en los modelos previamente evaluados e incrementaron de manera importante los tiempos de procesamiento, en (?) donde se hizo uso de estas bandas de manera directa para modelar el TCH se observaron resultados pobres en términos de precisión.



Se observa la delimitación mediante polígonos internos de las parcelas y la variedad de caña sembrada en cada una, graficados sobre la imagen multiespectral base reconstruida en pseudocolor. El color en el que se resalta cada parcela corresponde a la variedad de caña cultivada en cada parcela.

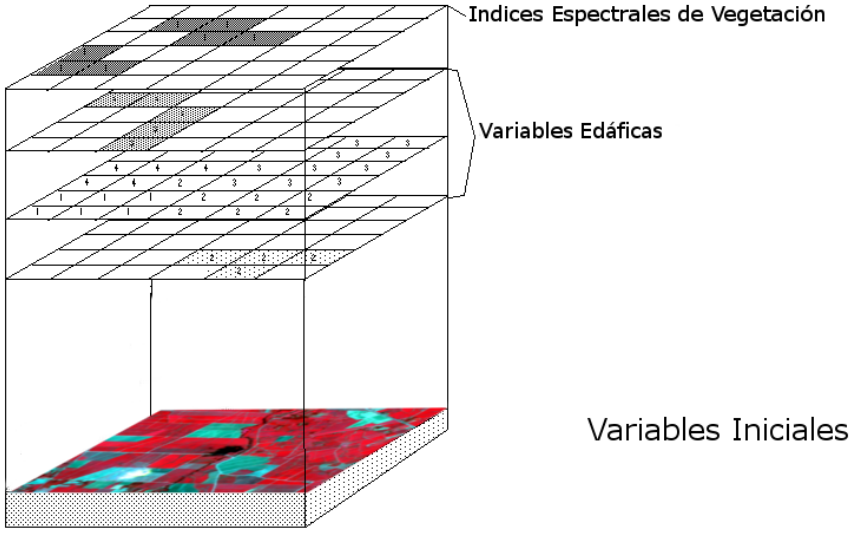
Figura 2.3: Área de muestreo en parcelas incluidas en cultivo experimental, limitada por poligonos

- Paso 2: Lectura de información de suelos, las variables (edáficas) contienen componentes químicos y físicos de los que se tomaron muestras previo a la siembra. Esta información procesada conjuntamente permite generar mapas intermedios con la información de cada variable en formato ráster, dado que la información inicial era información georeferenciada a cada punto muestreado, lo que se realizó fue una interpolación ajustada de manera automáticamente mediante el algoritmo de krige (autokrige), el cual busca el ajuste del mejor modelo para cada una de las variables incluyendo un proceso de validación cruzada para asegurarse que así sea (véase **??**). Así esta información junto con la información puntual de las imágenes se unieron en una única fuente de información donde la unidad de muestreo es el píxel tomando como referencia el tamaño de pixel entregado por la imagen multiespectral, es decir 35 x 35 cm.

- Paso 3: Toda esta información se llevó a una única base de datos, dicha base de datos tiene en su estructura la identificación de la parcela, la variedad de caña sembrada, todas las variables físico químicas del suelo y los índices de vegetación calculados. De esta forma se generó un archivo final, con el cual se comenzó a hacer el procesamiento de información para

el entrenamiento y validación de los algoritmos propuestos.

- Paso 4: Para poder tener una referencia precisa de los datos de las parcelas, se generaron polígonos que cubren áreas específicas dentro de las parcelas a estudiar (figura **??**). Estos polígonos de dimensiones 15mX15m sirven como una máscara para extraer información, lo que genera un límite de muestreo para realizar un etiquetado claro de la información de imágenes y campo versus la información etiquetada de cada parcela y los resultados, ademas esta clara delimitacion minimiza el problema de tener algun tipo de efecto de borde en el analisis de los datos de cada parcela.



Representación gráfica de cómo se apila la información en cada capa (variable) alineándola de manera vertical mediante la ubicación geográfica.

Figura 2.4: Esquema de apilamiento de información para todas las variables inicialmente incluidas en el estudio

Dado que esta información se procesó de manera secuencial, se tuvo especial cuidado en considerar correcciones y ajustes numéricos a lo largo del procesamiento de los datos para asegurar la coherencia de la información relacionada, haciendo de este conjunto de datos un sistema de información geográfica (Figura: **??**).

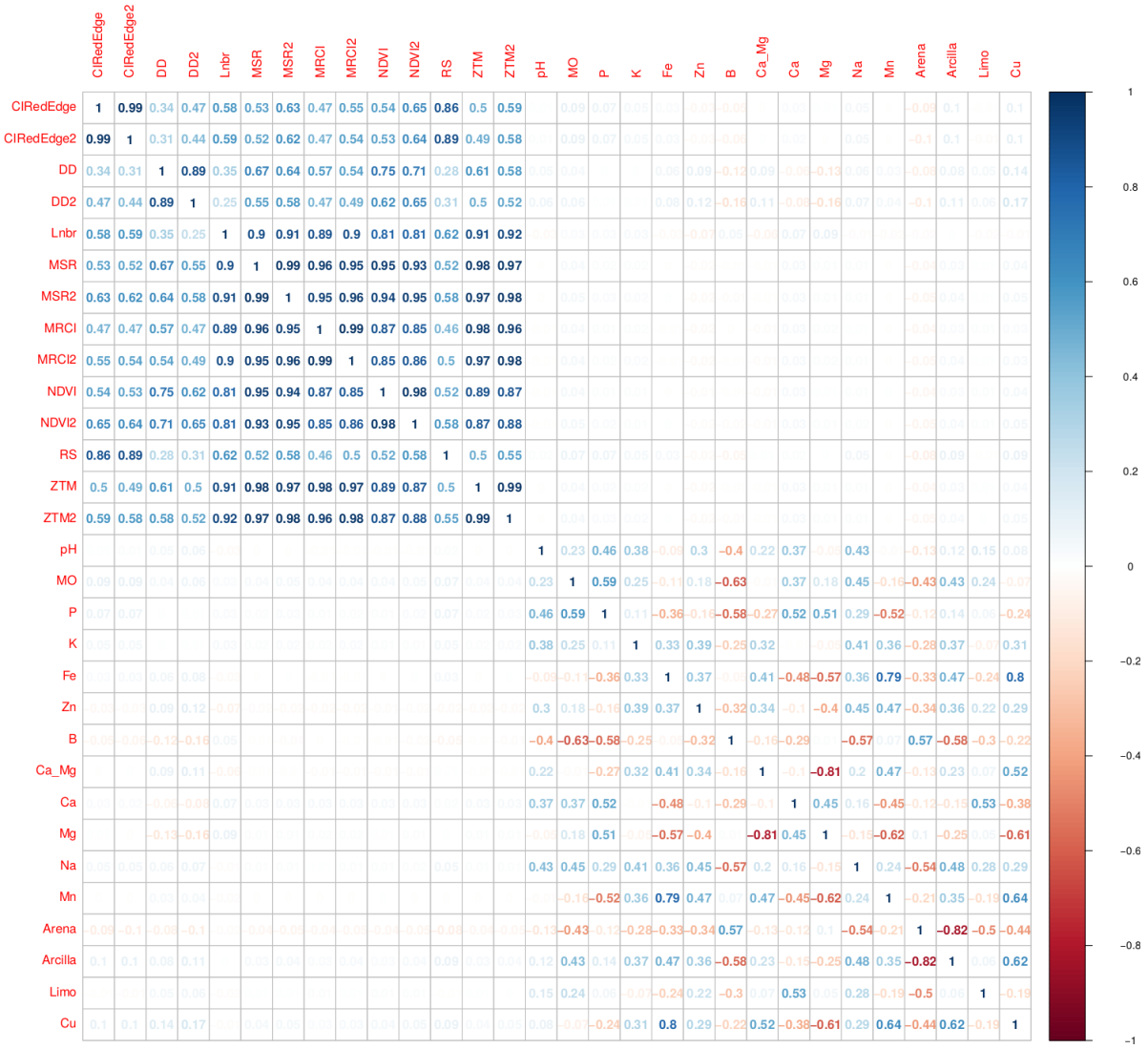
Las variables que hacen parte del conjunto de datos base para el análisis a realizar son las siguientes:

Índices Calculados con la Imagen Multiespectral						
CIRedEdge	CIRedEdge2	DD	DD2	Lnbr	MSR	MSR2
MTCI	MTCI2	NDVI	NDVI2	SR	ZTM	ZTM2

Ver apéndice **??** para conocer información detallada de estos índices.

Variables Edáficas							
pH	MO	P	K	Fe	Zn	B	Cu
Ca_Mg	Ca	Mg	Na	Mn	Arena	Arcilla	Limo

Se observa la correlación existente entre las variables originales, donde podemos ver que existen variables con correlaciones altas como se ve a continuación:

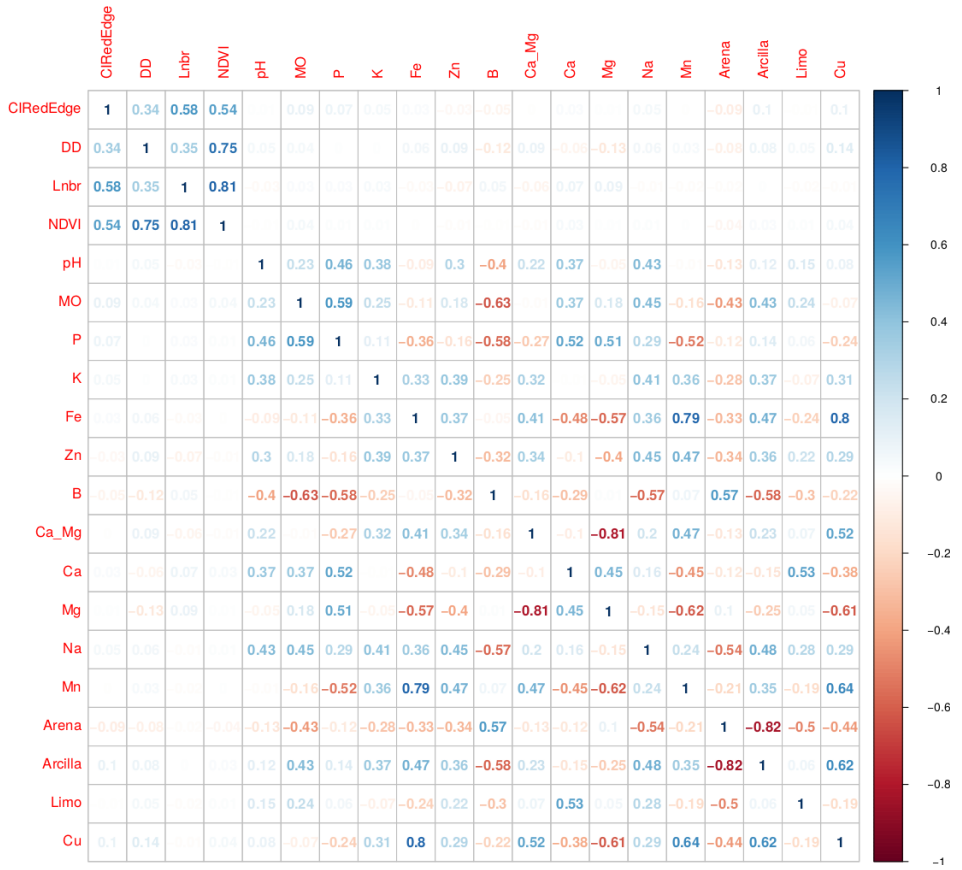


En cada eje se encuentran las variables y en su intercepto se representa el nivel de correlación de pearson entre pares de variables, correlaciones positivas en azul y correlaciones negativas en rojo, en la intercepción se presenta la magnitud de la correlación, claramente esta matriz es simétrica, por lo cual solo es necesario analizar la diagonal superior o inferior para comprender las correlaciones.

Figura 2.5: Correlación de Pearson entre variables originalmente incluidas en el estudio

Uno de los problemas encontrados al entrenar los modelos planteados fue que a medida que se incluían más variables en un modelo los tiempos de ejecución y entrenamiento se hacían más largos, pasando de minutos a horas o incluso días, es decir se multiplicaban de manera exponencial a medida que se incluían más variables, por lo tanto se hacía más compleja la evaluación de estos,

esto en gran medida por la cantidad de índices planteados en una fase inicial, por esta razón se decidió escoger solo una de cada par de variables con correlaciones superiores al 85 % y todas las demás variables con correlaciones bajas entre sí, esto para evitar incluir y alimentar los modelos con variables que entreguen información redundante.



La interpretación de este grafico es idéntica a la que se realizó en el grafico??

Figura 2.6: Correlación de Pearson entre variables seleccionadas para seguir siendo incluidas en el estudio.

Índices Calculados con la Imagen Multiespectral			
CIRedEdge	DD	Lnbr	NDVI

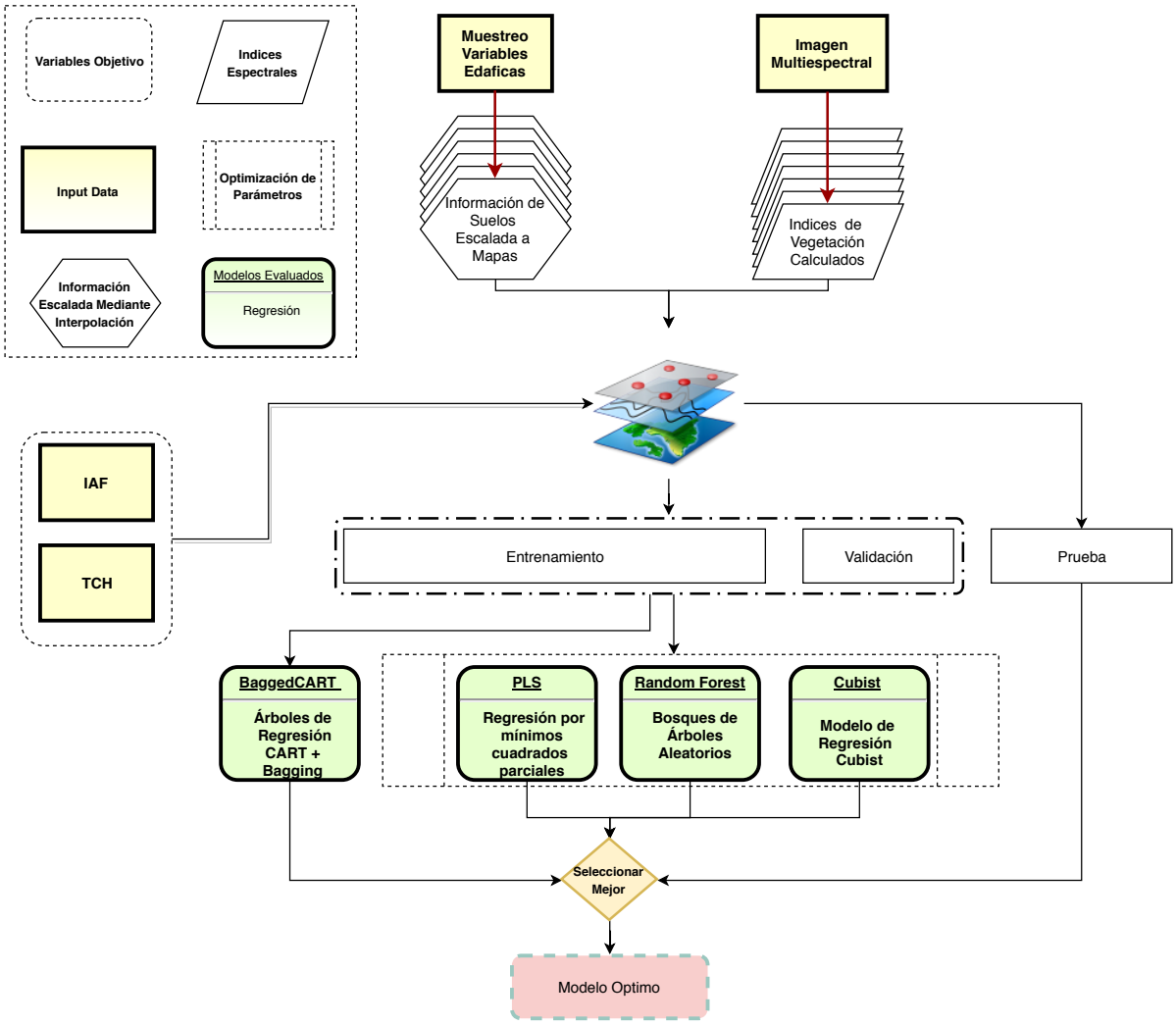
Variables Edáficas							
pH	MO	P	K	Fe	Zn	B	Cu
Ca_Mg	Ca	Mg	Na	Mn	Arena	Arcilla	Limo

Con las variables seleccionadas se procedió a la siguiente fase de entrenamiento, validación y pruebas de modelos para la predicción de IAF y TCH. El conjunto de datos total con información completa cuenta con 50.673 puntos o píxeles, los cuales corresponden a datos georeferenciados correspondientes a las áreas identificadas para cada parcela en estudio, de estos datos se hizo uso del 75 % para el proceso de entrenamiento del modelo y el 25 % restante de la información se dejó

disponible para realizar las pruebas y evaluación de la precisión de los algoritmos. Todo el procesamiento de la información se realizó mediante el software estadístico *R* (**?**). En la evaluación de los modelos se hizo uso de la librería caret (**?**), esta librería junto con los paquetes de procesamiento paralelo y los relacionados a cada modelo, complementan como herramientas el análisis de los datos.

2.3. Flujo de Trabajo de la Metodología

La figura **??** muestra el flujo de trabajo desarrollado.



Cuatro modelos (BaggedCART, PLS, Random Forest y Cubist) fueron desarrollados y evaluados hasta la optimización de sus parámetros usando información de índices de vegetación, calculados de una imagen multiespectral e información relativa a variables edáficas muestreadas en campo para predecir IAF (Índice de área foliar) y producción TCH (toneladas de caña por hectárea), se genera un sistema de información geografica, el cual se separa en (datos de entrenamiento, datos de validación) y datos de prueba, con los primeros dos se entrenan y validan los modelos, con los datos de prueba se evalua la precisión de los modelos sobre datos que el modelo no conoce previamente.

Figura 2.7: Flujo de trabajo general de la metodología

Se incluye en el apéndice de este documento el algoritmo global que presenta todas las fases del procesamiento de datos y evaluación de los modelos a modo de guía (Ver **??**).

2.4. Detalles y Características Asociadas a los Modelos Evaluados

Cada uno de los modelos presentados en la sección **??**, tienen vinculados características que hacen que estos modelos difieran entre ellos y por lo tanto unos puedan tener mejores o peores características para modelar y predecir las variables de interés y su relación con los datos de índices de vegetación y variables edificas. A continuación se recorrerán dichos modelos y se describirán algunas de estas características.

2.4.1. Regresión por Mínimos Cuadrados Parciales (PLS)

De acuerdo a lo presentado en la seccion **??** aunque PLS está relacionado con la regresión de componentes principales (PCR) se diferencia de éste, al no usar hiperplanos de máxima varianza entre la variable respuesta y las variables independientes, lo que PLS busca es encontrar una regresión lineal mediante proyecciones de la variable respuesta (Y) y las variables independientes (X) en un nuevo espacio.

Dado que tanto X como Y se proyectan en espacios nuevos al PLS se le denomina como factor de modelos bilineales (**?**).

PLS es un método efectivo para construir modelos predictivos cuando los factores son muchos y altamente colineales. Se debe tener en cuenta que el énfasis está en predecir las respuestas y no necesariamente en tratar de entender la relación subyacente entre las variables (**?**). PLS no asume la normalidad y estima por mínimos cuadrados de forma recursiva.

El modelo subyacente de PLS multivariante es:

$X = TP^T + E$

$Y = UQ^T + F$

En nuestro caso

- X esta compuesto de las variables [Índices de vegetación y Variables Edáficas].
- Y representaría la respuesta es decir: IAF o TCH.

Por lo tanto de acuerdo al modelo subyacente de PLS presentado X es una matriz de $n \times m$ predictores, Y es una matriz de $n \times p$ respuestas; T y U son matrices $n \times l$ que son, respectivamente, proyecciones de X y proyecciones de Y ; P y Q son, respectivamente de dimensión $m \times l$ y $p \times l$ matrices ortogonales; y las matrices E y F son los términos de error donde se supone que estos son independientes e idénticamente distribuidos de manera aleatoria con distribución normal. Las descomposiciones de X e Y se hace para maximizar la covarianza de T y U.

2.4.2. BaggedCART

BaggedCART es un método compuesto por dos métodos que se desarrollaron de manera independiente cada uno para aplicaciones particulares. El algoritmo CART es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos nodos o hijos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado (**?**). Las divisiones se seleccionan de modo que la "impureza" de los hijos sea menor que la del grupo madre y éstas están definidas por un valor de una variable explicativa (**?**). El objetivo es particionar la respuesta en grupos homogéneos y a la vez mantener el árbol razonablemente pequeño. Para dividir los datos se requiere un criterio de particionamiento el cual determina la medida de impureza, esta última establecerá el grado de homogeneidad entre los grupos.

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos (**?**):

1. Construcción del árbol máximo:

   Se comienza creando un árbol que describe el conjunto de entrenamiento, el que generalmente está sobreajustado, es decir que gran parte de los niveles y nodos no producen una mejor clasificación o regresión y esto puede ser demasiado complejo, cada grupo o respuesta es caracterizado por la media como respuesta numérica, gráficamente un árbol se representa con el nodo raíz (los datos sin ninguna división) al iniciar y las ramas con hojas debajo (cada hoja es el final de un grupo).

2. Escoger el tamaño correcto del árbol

   a) Calidad del Nodo:

   La función de impureza es una medida que permite determinar la calidad de un nodo. Existen varias medidas de impureza (criterios de particionamiento) que permiten analizar varios tipos de respuesta, las tres funciones de impureza principales son presentadas por **?**.

   b) Poda del árbol:

   El árbol obtenido es generalmente sobreajustado, por tanto debe ser podado, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño "adecuado" del árbol. **?** introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol. Computacionalmente el procedimiento descrito es complejo, pero se puede consultar en **???**.

3. Selección del árbol óptimo mediante un procedimiento de validación cruzada ("cross-validation"): De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad (**?**), por tanto se requiere estimar con precisión el error de predicción y en general esta predicción se hace utilizando un procedimiento de validación cruzada.

El otro algoritmo involucrado en el algoritmo de BaggedCART es bagging o agregación bootstrap:

Este es un procedimiento aplicado en términos generales, el procedimiento reduce la varianza de un método de aprendizaje estadístico, este es particular y frecuentemente usado en el contexto de árboles de decisión.

Si tenemos un conjunto de n observaciones $Z_1, Z_2, ..., Z_n$ independientes e idénticamente distribuidas cada una con varianza igual a σ^2, la varianza de la media de las observaciones $\overline{Z}$ está dado por σ^2/n en otras palabras, el promedio del conjunto de observaciones reduce la varianza. Por supuesto, esto no es práctico porque generalmente tenemos acceso a múltiples conjuntos de entrenamiento. En lugar, podemos iniciar el remuestreo "bootstrap" tomando muestras repetidas del conjunto de datos de entrenamiento.

Esta característica genera B diferentes datos de entrenamiento remuestreados o "bootstraped" nosotros entonces entrenaremos en el B_ceavo conjunto remuestreado de entrenamiento en el orden a obtener $\hat{f}^{*b}(x)$, la predicción para el punto en X, tendríamos entonces un promedio de todas las predicciones posibles:

$$\hat{f}^{*b}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

A este procedimiento es lo que llamamos bagging, en el proceso de bagging, el número de árboles creados no es un parámetro crítico en cuanto a que por mucho que se incremente el número, no se aumenta el riesgo de sobreajuste. Alcanzado un determinado número de árboles, la reducción del error de validación se estabiliza.

La idea principal de BaggedCART es utilizar la inestabilidad de los árboles individuales generando varios árboles mediante un proceso de generación de varias muestras con reemplazo y una votación de los árboles para la predicción de un consenso. Los árboles utilizados para la regresión y los árboles utilizados para la clasificación tienen algunas similitudes, pero también algunas diferencias como el procedimiento utilizado para determinar dónde se divide cada árbol,

Principalmente esto tiene que ver con definir una medida de impureza que puede ser por medio de la suma de mininos cuadrados o las desviaciones mínimas absolutas con los cuales la elección de

particionamiento se vincula a la minimización del error de acuerdo a la métrica escogida. En **?** se describe con detalle cada parte del procedimiento inmerso detrás de los BaggedCART.

2.4.3. Bosques Aleatorios (Random Forest)

Random forest o 'Bosques Aleatorios' es una combinación de árboles predictores donde cada árbol depende de los valores de un vector aleatorio probado independientemente y cada uno de estos valores posee la misma distribución. En muchos problemas el rendimiento del algoritmo random forest es muy similar a la del boosting (Ver **??**), pero este es más simple de entrenar y ajustar. Como una consecuencia el random forest es popular y es ampliamente utilizado.

El algoritmo de random forest o bosques aleatorios (**??**) se describe de la siguiente manera:

Algoritmo 1 Bosques aleatorios para Regresión

Donde B corresponde al número de árboles a considerar en un bosque aleatorio.

1. Para b=1 a B:
   - *a*) Sacar una muestra (bootstrap) **Z*** de tamaño N del conjunto de datos de entrenamiento. aproximadamente el 70 % del conjunto de entrenamiento.
   - *b*) Cultivar el árbol aleatorio T_b con los datos remuestreados, por repetición recursiva se realizan los siguientes pasos para cada nodo terminal del árbol, hasta que haya alcanzado el mínimo nodo de tamaño n_{min} .
     1) Se selecciona m variables aleatoriamente de las P variables.
     2) Se escogen los mejores variables/puntos de división entre las m variables.
     3) Se divide el nodo en dos nodos hijos.
2. Se extrae el conjunto de árboles T_{b1}^{B}

Para hacer una predicción de un nuevo punto de X: En regresión: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Este algoritmo proporciona una mejora con respecto a los árboles generados en bagging a modo de incluir pequeñas cantidades que descorrelacionan los árboles, esto reduce la varianza del modelo cuando se promedian los árboles. Al igual que en el bagging, se construye una serie de árboles de decisión sobre muestras de entrenamiento (remuetreadas), pero en la construcción de estos árboles de decisión, cada vez que se da una "separación" en un árbol, esta es considerada, así se realiza una selección aleatoria de m predictores donde son elegidos los candidatos de división de todo el conjunto de p predictores. Se le permite que la división use sólo uno de esos predictores m. En cada selección de las muestras remuestreadas, se elige una selección fresca de m predictores que se toma para cada división, y por lo general se elige $m \approx \sqrt{p}$ es decir, el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores.

Muestras Fuera de la Bolsa (Out of Bag Samples):

Una de las características importantes de los bosques aleatorios es el uso de muestras fuera de la bolsa en la literatura (OOB) para cada zi o muestra del conjunto de entrenamiento. Una predicción de error OOB es casi idéntica a la obtenida por validación cruzada N veces. Por lo tanto, a diferencia de muchos otros estimadores no lineales, los bosques aleatorios pueden ser entrenados en forma secuencial, y la validación cruzada se puede realizar a lo largo de la secuencia sin que esto implique un problema. Una vez se estabiliza el error fuera de la bolsa (esto significa que el error OOB no varía de manera significativa entre iteraciones) , la formación del árbol definitivo se puede dar por terminada.

Importancia de las Variables en Random Forest

Una medida muy útil en Random Forest es la entregada por el error OOB (Out of Bag Error) o importancia de variables basada en permutaciones. En cada árbol construido mediante muestreo «bootstrap» sobre el 70 % del conjunto de entrenamiento, (el 30 % restante de las otras observaciones no se usan para crecer el árbol de regresión). Para calcular la importancia de la variable j (los valores de la variable j se permutan aleatoriamente mientras se mantienen todas las otras variables predictoras fijas). Estos datos modificados OOB se pasan nuevamente por el árbol y se calculan los valores predichos. Calculado en base a la diferencia en el error OOB entre el conjunto de datos (real/no permutado) y el conjunto de datos permutados en todos los árboles. En otras palabras, la importancia de la variable j se evalúa por cuanto es peor la exactitud de predicción de los árboles aleatorios cuando la variable j es permutada al azar (**??**). Dicha importancia se puede visualizar mediante el uso de gráficos que muestran la incidencia de una variable sobre la predicción final de un nodo (IncNodePurity), lo que da una visión clara de la importancia de ciertas variables en la consecución de un modelo en evaluación, para la clasificación, la impureza del nodo se mide con el índice de Gini y para la regresión, se mide por la suma residual de cuadrados.

2.4.4. Árboles de Aprendizaje Empírico - Cubist

Como observamos en la sección **??** el algoritmo de cubist es una mejora sobre el modelo M5 de Quinlan el cual consiste en modelos predictivos basados en reglas, las cuales se expresan como un árbol de decisión en el que sus hojas terminales contienen modelos de regresión lineal. Asociado al modelo de cubist se encuentran dos conceptos de importancia los cuales son boosting y los denominados commites.

Boosting

El Boosting está basado en el cuestionamiento planteado por **?**, ¿Puede un conjunto de clasifica-

dores débiles crear un clasificador robusto? (**?**). El boosting consiste en combinar los resultados de varios clasificadores débiles para obtener un clasificador robusto. Cuando se añaden estos clasificadores débiles, se hace de modo que estos tengan diferente peso en función de la exactitud de sus predicciones. Luego de que se añade un clasificador débil, los datos cambian su estructura de pesos: los casos que son mal clasificados ganan peso y los que son clasificados correctamente pierden peso. Así, los clasificadores débiles se centran de mayor manera en los casos que fueron mal clasificados por los clasificadores débiles.

Committees

El concepto de "committees" es similar al de "boosting" mediante el desarrollo de una serie de árboles de forma secuencial con los pesos ajustados. Sin embargo, la predicción final es el promedio simple de las predicciones de todos los miembros de los "committees", una idea más cerca de "bagging".

El primer árbol sigue el procedimiento descrito en la sección anterior, árboles subsiguientes se crean utilizando versiones ajustadas al resultado del conjunto de entrenamiento: si el modelo predice en exceso un valor, la respuesta se ajusta para bajar al próximo modelo (y así sucesivamente). A diferencia del boosting tradicional los pesos en cada etapa en cada committee no se utilizan para promediar las predicciones de cada árbol del modelo; la predicción final es un promedio simple de las predicciones de cada árbol del modelo.

Cubist posee una innovación llamada correcciones por instancia, mediante el uso de vecinos cercanos para ajustar las predicciones de las reglas, esto lo que genera es un ajuste de las reglas de predicción haciendo uso de los vecinos más cercanos, así el algoritmo usa argumentos de sus vecinos para ajustar sus propias reglas. Ver (**?**) para detalles del ajuste.

2.5. Evaluación y Ajuste de Modelos

2.5.1. Teoría para la Decisión Estadística

En esta sección se introduce brevemente la teoría que sirve como marco para el desarrollo de los modelos. Primero consideramos el caso de una salida cuantitativa, y nos situamos en el mundo de las variables aleatorias y espacios de probabilidad. Sea $X \in \mathbb{R}^p$ un vector aleatorio de entrada de valores reales y $Y \in \mathbb{R}$ una variable aleatoria de salida con distribución conjunta $Pr(X, Y)$, buscamos una función $f(X)$ para predecir valores Y dados por la entrada X. Esta teoría requiere una función de pérdida $L\left(Y, f\left(X\right)\right)$ para penalizar errores en la predicción, y de lejos la más común y conveniente es la pérdida de error cuadrático: $L\left(Y, f\left(X\right)\right) = \left(Y, f\left(X\right)\right)^2$. Esto nos lleva a un criterio para elegir $f(X)$.

$$EPE(f) = E(Y-f(X))^2$$

$$= \int [y-f(x)]^2 Pr(dx, dy).$$

la predicción del error esperado al cuadrado, condicionado en X, podemos escribir EPE como:

$$EPE(f) = E_X E_{Y|X}([Y-f(X)]^2|X)$$

y vemos que esto es suficiente para minimizar EPE puntual:

$$f(x) = argmin_c E_{Y|X}([Y-c]^2|X = x)$$

La solución es:

$$f(x) = E(Y|X = x),$$

La esperanza condicional, es conocida como la función de regresión. Entonces la mejor predicción de Y en cualquier punto de $X = x$ es el promedio condicional, entendiendo por mejor a la medida del promedio de los errores al cuadrado (**?**).

2.5.2. Validación Cruzada

Una técnica ampliamente utilizada para evaluar los resultados de un análisis estadístico o de modelos de aprendizaje automático es el proceso de validación cruzada, el cual se realiza generando una partición de datos separando datos de entrenamiento y datos de prueba en ciertas proporciones, donde se garantiza que estas participaciones son independientes entre sí, con la primera parte de los datos se entrena el modelo o los modelos posteriormente con los datos de prueba, los cuales son datos que el modelo no ha tenido conocimiento previamente, con los valores estimados por el modelo y los valores observados es posible calcular medidas de evaluación de la precisión o calidad en la predicción. (**?**) Esta técnica se suele utilizar en estudios donde el objetivo es la predicción y se requiere estimar que tan preciso es un modelo cuando este es llevado a la práctica, es decir en un entorno donde los datos son totalmente nuevos para el modelo y por lo tanto podemos probar la capacidad del modelo de encontrar las relaciones intrínsecas en los datos de entrenamiento para lograr extrapolar los resultados a un entorno general.

La validación cruzada probablemente es el método más simple y ampliamente usado para estimar el error en predicción, este método estima directamente el error esperado de una muestra-extra

$Err = E\left[L\left(Y, \hat{f}(X)\right)\right]$, el error medio generalizado cuando el método $\hat{f}(X)$, se aplica a una muestra de prueba independiente de la distribución conjunta de X e Y. Como se mencionó anteriormente, se podría esperar que la validación cruzada estime el error condicional, con el conjunto de entrenamiento τ el cual se mantiene fijo (**?**).

Validación Cruzada K-veces ("K-Fold")

Sería ideal que, en el caso que se tuvieran datos suficientes, se dejará de lado un conjunto de datos para la validación y los datos restantes se usarán para evaluar el desempeño de nuestro modelo de predicción. En general no es posible dado que los datos son a menudo escasos, y la precisión de la validación es perjudicada. Alternativamente, si se repite el procedimiento de validación cruzada en K ocasiones, cada una con datos distintos para entrenamiento y validación, se gana en precisión. Para ello se dividen los datos en K partes más o menos del mismo tamaño; por ejemplo, cuando $K = 4$, el escenario podría ser como en la figura: **??**.

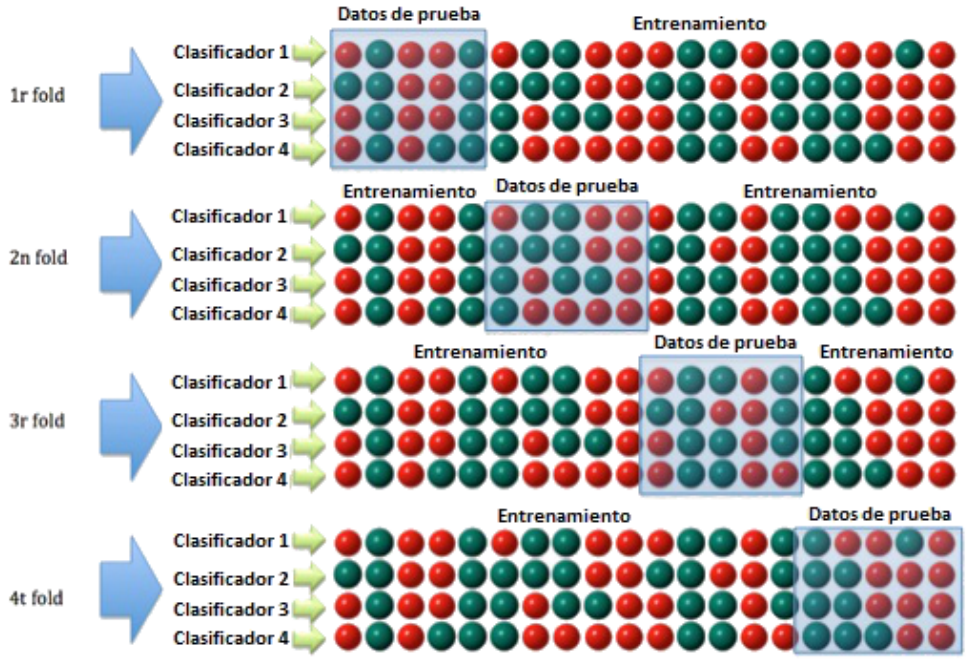


Figura 2.8: Esquema de validación cruzada usando el método de k-fold, aplicada a muestras de datos para realizar entrenamiento de modelos de aprendizaje automático, By Joan.domenech [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)]

Para la figura **??** en $k = 3$, se ajusta el modelo con las otras k-1 partes de los datos, es decir con $(k = 1, k = 2, k = 4)$ y se calcula el error de predicción del modelo para la $k - ésima$ parte, esto para predecir la parte de orden k (en este caso la parte 3) de los datos, se realiza este proceso para cada $k = 1, 2, 3, .., K$ y se combinan las estimaciones del error de predicción.

Ahora sea $\kappa : 1, ..., N \rightarrow 1, ..., K$ una función de indexación la cual indica a qué partición es

asignada la observación i por la aleatorización. Denotemos por $\hat{f}^{-\kappa}(X)$ la función de ajuste, computada con la $k-ésima$ parte de los datos eliminada, entonces la validación cruzada estimada del error de predicción es:

$$CV\left(\hat{f}\right) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i\right)\right)$$

Es común realizar una elección de $k = 5\,ó\,10$, pero incluso se podría tener una elección de $K = N$ lo que es conocido como una validación cruzada dejando solo uno fuera "leave-one-out " y donde el estimador de validación cruzada es aproximadamente in-sesgado, pero no se utiliza mucho este caso ya que cuando tenemos tantos conjuntos de validación, existen muchos conjuntos casi iguales. Por esta razón en la mayoría de las veces se opta por selección de un K un poco más conservador como 5 o 10, lo que es suficiente para obtener buenos resultados (**?**).

2.5.3. Entrenamiento de Modelos y Afinación de Parámetros

Los algoritmos de aprendizaje automático evaluados fueron parametrizados para obtener el mejor funcionamiento en cada uno de los problemas abordados. Una dificultad aquí es que configurar un algoritmo para cierto problema, en sí, puede ser una investigación en sí misma, esto se da por la cantidad posible de parámetros que puede contener un modelo y las posibles combinaciones de estos parámetros, lo que se convierte en un problema importante de optimización matemática.

Al igual que la selección de "el mejor" modelo para un problema, donde no se puede saber de antemano cual parámetro para un modelo será mejor para un problema en particular. Lo mejor es investigar empíricamente con experimentos controlados.

Una manera de encontrar estos parámetros óptimos es crear una cuadricula (Grid) con los parámetros que se desean evaluar en cada modelo y seguir el algoritmo (**??**) que permitirá ajustar un modelo para cada parámetro o combinación de parámetros, (dependiendo del modelo) **?**, y así mediante un proceso iterativo se realiza la búsqueda de los mejores parámetros que maximizan el desempeño del algoritmo ajustado a los datos, dicho desempeño se mide de acuerdo a una métrica que determina el investigador como parámetro de calidad, de acuerdo al objetivo que este determine que mejora el rendimiento del modelo, por ejemplo minimización del RMSE o maximización del R^2 en problemas de regresión.

Algoritmo 2 Búsqueda del modelo optimo, mediante la afinación de parámetros

1. Define el conjunto de parámetros a evaluar para cada modelo
   - **Para** cada parámetro definido **realicé**
     - **Para** cada iteración por remuestreo **realicé**
       - Retenga una muestra específica
       - Ajuste el modelo con la muestra de entrenamiento
       - Realice la predicción con las muestras asignadas para prueba
     - **Fin**
     - Calcule el rendimiento promedio de las predicciones obtenidas a través de la muestra seleccionada
   - **Fin**
   - Determine el conjunto de parámetros óptimos
   - Ajuste el modelo final a todo el conjunto de entrenamiento usando los parámetros óptimos.

Es importante saber que este proceso puede ser tan extenso y complejo, como la cantidad de parámetros y combinaciones que se puedan probar para un único modelo, y la necesidad computacional se asociará con la cantidad de datos vinculados a estas pruebas y a la cantidad de procesos de remuestreo asignados a cada evaluación, en la figura **??** este proceso se muestra como un par de flechas azules de iteración en la sección de uso de datos de Entrenamiento.

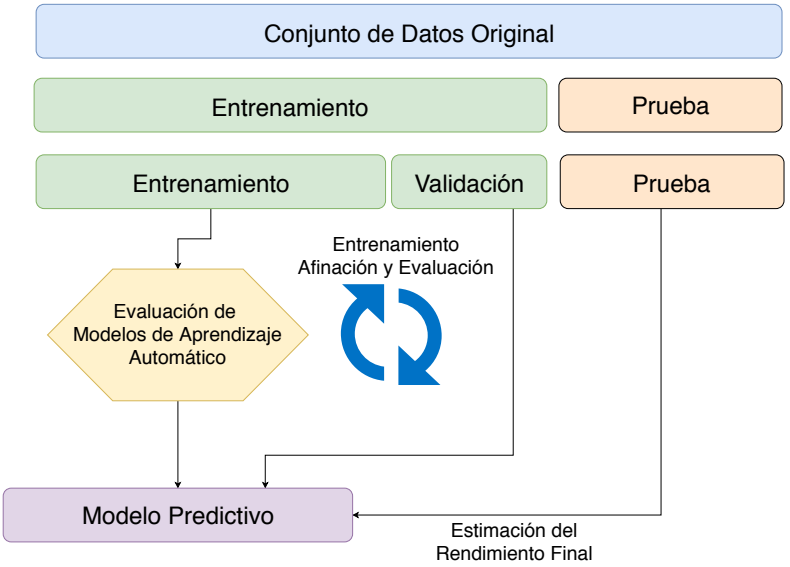


Figura 2.9: Esquema de selección y partición de datos para la evaluación de modelos de aprendizaje automático supervisado

2.5.4. Medidas de Desempeño en Modelos de Aprendizaje Automático Supervisado

A continuación se presentan algunas características relacionadas a los modelos de aprendizaje automático estas asociadas a la medición del error de predicción (el término "error" representa aquí la diferencia entre el valor predicho y el valor verdadero), es importante elegir un estimador

apropiado del error (**?**), para el caso de abordar problemas de regresión las principales métricas son el error absoluto medio, el error cuadrático medio, la raíz de este error como el coeficiente de determinación o R^2:

1. El error absoluto medio (MAE, por sus siglas en inglés) es la suma de las diferencias absolutas entre las predicciones y los valores reales. Da una idea de cuán equivocadas estaban las predicciones es decir cuantifica la magnitud del error, pero no tiene idea de la dirección (por ejemplo, por encima o por debajo de la predicción). Esta es una métrica que es sensible a valores atípicos por lo cual no es usado como un estimador robusto del error.
   Su formulación es MAE: $\frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y_i} \mid$ donde n es la cantidad de individuos , y_i son los valores reales u observados y $\hat{y_i}$ es el valor predicho.

2. El error cuadrático medio o MSE (Mean squared error) es un estimador que mide el promedio de la diferencia entre el estimador y lo que se estima al cuadrado, se puede evaluar el rendimiento de un modelo con una comparación de cuánto se desvían en promedio las predicciones de los datos reales. El MSE es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia entre estos se puede producir ya sea debido a la aleatoriedad o porque el estimador no tiene en cuenta información que podría producir una predicción más precisa (**?**).
   Error cuadrático medio (MSE): $\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y_i})^2$

3. El error cuadrático medio (o MSE, por sus siglas en inglés) es muy parecido al error absoluto medio, ya que proporciona una idea general de la magnitud del error. Al tomar la raíz cuadrada del error cuadrático medio este se convierte en las unidades originales de la variable de salida y puede ser significativa para la descripción y presentación de resultados. A este ajuste se le conoce como raíz del error cuadrático medio (RMSE) este es usado con mayor frecuencia para análisis del error en regresión.
   Este es una analogía con la desviación estándar, al tomar la raíz cuadrada del MSE se produce el (RMSE o RMSD), que tiene las mismas unidades que la cantidad que se estima, esto es importante, por que se puede medir el error en las mismas unidades de la variable sobre la cual se intenta ajustar el modelo y por lo tanto predecir; el RMSE es un estimador in-sesgado, el RMSE es la raíz cuadrada de la varianza.
   Raíz del error cuadrático medio (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y_i})^2}$

4. La métrica R^2 o R cuadrado, proporciona un indicador de la bondad del ajuste de un conjunto de predicciones respecto a los valores reales. En la literatura estadística, esta medida se llama coeficiente de determinación. Este es un valor entre 0 y 1, donde cero indica sin ajuste y 1 ajuste perfecto.

El R^2 o coeficiente de determinación es también una manera estándar de medir cuánto se adapta el modelo a los datos, esté mide la proporción de variabilidad total de la variable dependiente (Y) respecto a su media. el R^2 proporciona una medida de qué tan bien son replicados los datos observados por el modelo.

Es usual expresar esta medida en porcentaje (multiplicándola por cien).

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

En nuestro caso particular se hará uso de las medidas de ajuste relacionadas con problemas de regresión que poseen mejores resultados debido a su robustes, los cuales son:

- El RMSE
- El R^2 o coeficiente de determinación.

De acuerdo a esto un modelo será más preciso que otro si el RMSE es más pequeño y el R^2 tiende a 1 o en términos porcentuales a 100 %, lo que significa que el modelo es capaz de modelar la respuesta cometiendo un bajo nivel de error.

2.5.5. Errores Asociados a los Modelos de Aprendizaje Automático

Los tipos de errores asociados a un problema de modelamiento de datos haciendo uso de aprendizaje automático son los denominados 'sesgo', 'varianza' y 'la tasa de error optima o error irreducible'.

Cuando se tiene un conjunto de datos, generalmente en aprendizaje automático lo que se hace es partir este conjunto en 2 partes, (conjunto de entrenamiento y conjunto de pruebas), debido a que es necesario ajustar parámetros de un modelo evaluado este conjunto realmente se consideraría como 3 particiones el conjunto de entrenamiento general compuesto por [conjunto de entrenamiento y conjunto de validación] y el conjunto de pruebas, este esquema se puede observar claramente en la figura **??**.

- El conjunto de entrenamiento: Es típicamente el 75 % de los datos y se puede partir en dos, conjunto de entrenamiento y conjunto de validación.
  - El conjunto de entrenamiento: Es típicamente el 70 % de los datos. Como su nombre indica, este se usa para entrenar un modelo de aprendizaje automático.
  - El conjunto de validación: Este suele ser el 30 % de los datos. Este conjunto no se utiliza durante el entrenamiento. Se utiliza para probar la calidad del modelo entrenado.
- El conjunto de pruebas: Este conjunto suele ser el 25 % de los datos. Su único propósito es informar la precisión del modelo final.

Los errores en el conjunto de validación se usan para guiar la elección del modelo. Aunque este conjunto no se usa para entrenamiento, el hecho de que se usó para la selección del modelo hace que sea una mala elección para informar la precisión final del modelo, por esta razón se realiza este esquema de validaciónes y pruebas para la evaluación de los modelos de aprendizaje automático (**?**).

En aprendizaje automático los errores asociados a un modelo son la suma de tres tipos de errores:

- error debido a sesgo del modelo ajustado a los datos (Sesgo).
- el error debido a la varianza del modelo ajustado a los datos (Varianza).
- el error irreductible (equivalente a la varianza real de los datos, no asociada al modelado).

La siguiente ecuación resume las fuentes de errores:

$TotalError = ErrorIrreducible + Varianza + Sesgo$

Demostración (**?**):

$$
\begin{aligned}
E\left[\left(y-\hat{f}\right)^2\right] &= E\left[y^2+\hat{f}^2-2y\hat{f}\right] \\
&= E\left[y^2\right]+E\left[\hat{f}^2\right]-E\left[2y\hat{f}\right] \\
&= Var[y]+E\left[y\right]^2+Var[\hat{f}]+E\left[\hat{f}\right]^2-2fE\left[\hat{f}\right] \\
&= Var[y]+Var[\hat{f}]+\left(f^2-2fE\left[\hat{f}\right]+E\left[\hat{f}\right]^2\right) \\
&= Var[y]+Var[\hat{f}]+\left(f-E\left[\hat{f}\right]\right)^2 \\
&= Var[y]+Var[\hat{f}]+E\left[f-\hat{f}\right]^2 \\
&= \sigma^2+Var[\hat{f}]+Sesgo\left[\hat{f}\right]^2
\end{aligned}
$$

Estos errores variaran de acuerdo a la complejidad asociada al modelo entrenado, ver figura **??** , si un modelo es muy complejo su error en la validación tiende a ser mayor comparado con su error en el entrenamiento, esto significa que el modelo se ajusta mucho a los datos de entrenamiento perdiendo la capacidad de generalización a esto se le denomina **sobreajuste (overfitting)** en este caso el sesgo es bajo pero la varianza es alta, ahora si por lo contrario, el error tanto en el entrenamiento como en la validación son altos lo que puede ocurrir es que el modelo no tiene un grado de complejidad tal que permita generalizar el conocimiento a partir de los datos que conoce, a esto se le denomina **subajuste (underfitting)** y en este caso el error de sesgo es alto y la varianza es baja.

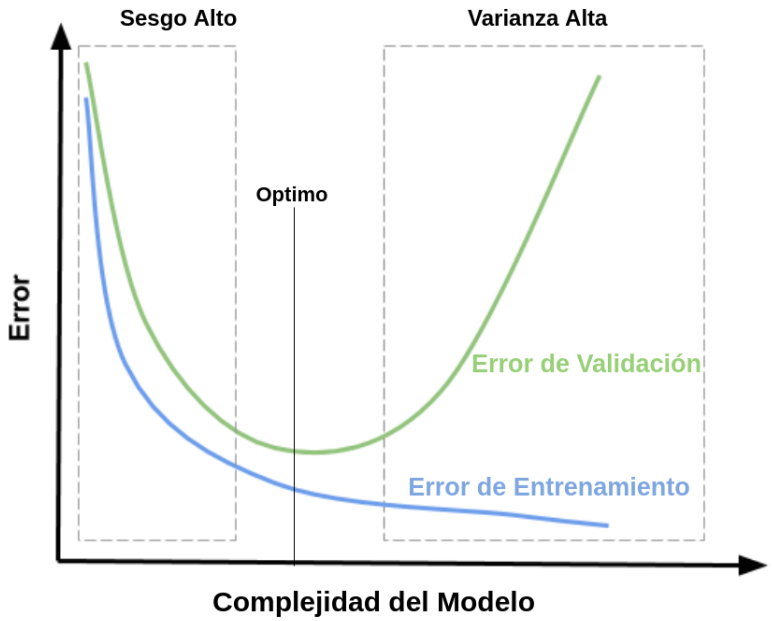


Figura 2.10: Errores asociados a la complejidad de modelo

Lo que se busca en el modelado mediante aprendizaje automático es encontrar un punto óptimo, donde se minimice el error total, es decir que no se entrene un modelo al punto de sobreajustarlo a los datos de entrenamiento pero que tampoco se genere un modelo tan simple que no sea capaz de aprender y generalizar el conocimiento obtenido.

2.5.6. Evaluaciones Inferenciales sobre el Rendimiento de los Modelos

Basándonos en los métodos planteados en (?) y (?) para hacer inferencias sobre diferencias entre modelos. Al momento de comparar modelos hacemos uso de las métricas evaluadas en nuestro caso el RMSE y el R^2, con los cuales se calculan todas las diferencias por pares y se prueban para evaluar si dichas diferencia son igual a cero. Para la comparación de múltiples modelos (más de 2 modelos) se usa una corrección de Bonferroni, donde el nivel de confianza es 1 - ((1-0.95) / p) donde p es el número de comparaciones por pares que se están realizando.

2.5.7. Consideraciones en las Variables de Respuesta TCH e IAF

El IAF es un índice que relaciona el área foliar de las plantas a unidades de superficie del suelo, por esta razón es válido realizar predicciones a nivel de pixel, ya que este índice representa un valor independiente a la escala espacial que se desee modelar, esto mismo no ocurre con el TCH, ya que como su nombre lo dice, el TCH es (Cantidad de Toneladas de Caña de Azúcar por Hectarea), por esta razón no debe considerarse como una predicción puntual a nivel de pixel, dado que cada predicción en cada punto (píxel) supondría tener pixeles de una hectarea de área algo que no es cierto en ningún caso, por esta razón la predicción del TCH se debe hacer de manera puntual y

posteriormente dichas predicciones se deben promediar para estimar el TCH general del área de estudio, de esta manera se podrá captar la variabilidad a lo largo del cultivo y a si mismo tener predicciones más cercanas a la realidad.

2.5.8. Consideraciones Asociadas al Diseño Experimental de las Parcelas

Un aspecto importante a tener en cuenta es el diseño experimental planteado en la (**??**), dicho diseño se planteó en el área de estudio con la finalidad de realizar otras investigaciones adjuntas al campo experimental, particularmente existen dos variaciones, la primero relacionada con el tipo de riego aplicado en las parcelas de cultivo y la segunda asociada a la variedad de caña sembrada, la primera variación lo realizaron para validar el impacto económico en el uso de distintos tipos de riego y distribución de fertilizantes en el cultivo, esta variación no genera ningún impacto en la respuesta del cultivo ya que dicha evaluación estaba enfocada en una evaluación económica más que en una evaluación asociada al rendimiento del cultivo, adicionalmente se adjunta en anexos el detalle de una prueba anova donde se presentan estas variaciones y su interacción (**??**), donde se observa claramente que la variación del tipo de riego no genera diferencias en la respuesta de la producción, ni en su interacción con las variedades de caña de azúcar, por otro lado las variedades de caña si generan diferencias significativas, mas estas están relacionadas al nivel de producción propio de cada variedad de caña de azúcar, por esta razón estas variedades de caña fueron aleatorizadas en el área de estudio y no existen parcelas contiguas, con la finalidad de separar los efectos asociados a esta variabilidad de la evaluación de los algoritmos, además esta variación sirvió para darle más conocimiento a los modelos evaluados en términos de variabilidad.

2.6. Aplicando Big Data para Entrenar y Validar Modelos

El uso de herramientas de big data se realizó como una estrategia para minimizar tiempos de procesamiento de los datos, dado que este procesamiento incluye usar información proveniente de múltiples fuentes que fue escalada a la resolución asociada a la imagen multiespectral base con resolución de 35X35 cm.

Esto genera una cantidad de observaciones que al momento de evaluar los modelos requerían grandes capacidades computacionales, por lo que se realizó el procesamiento en equipos con posibilidades de procesamiento paralelo, de no ser así en una computadora de escritorio (probado en un equipo desktop procesador core i7 4 nucleos y 12GB Ram) los tiempos de procesamiento se hacían muy largos, llegando hasta las semanas para la evaluación de los modelos planteados y en muchas ocasiones los procesos no se completaban por falta de capacidad de almacenamiento en memoria de las iteraciones, ya fuera para predecir el IAF o TCH.

La razón fundamental por la que estos algoritmos requieren altas características de procesamiento y funcionan bien en esquemas de procesamiento paralelo, es debido a que los algoritmos se basan en procesos iterativos que pueden ejecutarse independientemente para mejorar la respuesta a medida de que se hayan los parámetros óptimos. Estos procesos iterativos demandan altos requisitos computacionales para guardar en memoria la mayor cantidad de iteraciones sin causar fallas en los sistemas informáticos por saturación o lentitud, a lo que las herramientas de big data son una ayuda fundamental.

En este ámbito se exploró el procesamiento de los datos directamente en infraestructura en la nube con las siguientes características:

36 núcleos físicos de alto desempeño, lo que se traduce en el equivalente a 132 núcleos únicos de 2.9GHz en un equipo de escritorio y 60 GB de Memoria Ram.

Una máquina con estas características completa todo el procesamiento de la información y evaluación de modelos en menos de 2 horas de cómputo, la importancia de este procedimiento es presentar bases para hacer uso de estas herramientas para futuros modelamientos y sobre todo para conocer las limitantes a las que se enfrentarían en investigaciones futuras donde se desee ampliar el uso de estos algoritmos en estudios que puedan desprenderse de la investigación aquí desarrollada.

2.7. Evaluación y Parametrización de Modelos Elegidos

En esta sección se realizará un recorrido por algunas características relacionadas a los algoritmos y su aplicación para la predicción de variables relacionadas con la producción en caña de azúcar. Para cada modelo (exceptuando BaggedCART) se hizo una variación de parámetros de manera guiada, configuración de parámetros en la tabla **??**, dichos parámetros ajustables fueron variados en cada modelo. Adicionalmente a esta variación se generaron entrenamientos y pruebas mediante el algoritmo de k-fold, lo que permite realizar un re-muestreo interno sobre la información para evitar sobreajuste y sesgo que pueda presentarse en el entrenamiento y calibración de los modelos.

Parámetros Evaluados:

Modelo	**Parámetro**	**Valores**
PLS	ncomp	1,2,3,4,...,20
Cubist	committees	1,10,50,100
Cubist	neighbors	0,1,3,5,7,9
Random Forest	mtry	1,2,3,4,...,20
BaggedCART	-	-

Tabla 2.2: Parámetros de configuración evaluados en cada modelo probado para la búsqueda del modelo con mejor rendimiento

Cada uno de los parámetros evaluados tiene una interpretación según el contexto propio del algoritmo evaluado como se presenta a continuación:

- PLS:
  - **ncomp** es el número de componentes que se desea ajustar, se evaluaron todos los posibles componentes disponibles para las variables predictoras ingresadas al modelo.
- Bosques Aleatorios - Random Forest:
  - **mtry** es el número de variables seleccionadas al azar como candidatos en cada división para la creación de los árboles por parte del modelo.
- Cubist:
  - **committees:** En el modelo cubist se utiliza un esquema similar al boosting, pero en este caso se le llama committees y determina la cantidad de iteraciones realizadas para crear los modelos de árbol, a diferencia del boosting en los committees los pesos en cada nivel del árbol no se utiliza para promediar las predicciones de cada árbol.
  - **neighbors:** los vecinos cercanos pueden ser utilizados para ajustar las predicciones del modelo basado en reglas asociadas a la cercanía.

En los algoritmos evaluados tanto para el caso de IAF como para el TCH, se configuró cada modelo de manera que permitiera minimizar el sesgo y el sobreajuste a los datos disponibles para entrenamiento de los mismos. Aun teniendo en cuenta esto, es difícil en este caso en particular minimizar el grado de sobreajuste, ya que la cantidad de parcelas no es muy grande teniendo en cuenta que la escala con la que se tenían los datos para las variables de respuesta está limitada al resultado promedio por parcela más que por unidad píxel, aunque es técnicamente imposible obtener esta información a tal detalle, en este caso esto se da como una limitante propia del conjunto de datos disponible.

Por esta razón fue importante encontrar formas para minimizar ese posible sobreajuste a los datos. En (**?**) se propone añadir ruido blanco a los datos de respuesta del modelo, este ruido blanco se genera de manera aleatoria con una distribución normal centrada en cero y en nuestro caso con una desviación de 1/(10*SD(variable respuesta)) un décimo de la desviación estándar perceptible en la variable respuesta. Al agregar este ruido a la respuesta en el conjunto de entrenamiento, lo que hace es minimizar la posibilidad de que el modelo se sobreajuste a los datos de entrenamiento, es decir ya no se tendrían 32 valores respuesta los cuales en este caso son resúmenes (promedios) de la variable respuesta, lo que se tendría sería un espectro de variación alrededor de los resultados originales, esto le permitiría a los modelos encontrar detalles a mayor nivel en su búsqueda de adquirir conocimiento

de los datos, así finalmente se pueden tener modelos con una mayor posibilidad de generalización (para datos con rangos similares a los datos disponibles), de esta manera se espera que los modelos resultantes sean modelos más robustos.

En términos generales se siguió una configuración donde se realizó un re-muestreo sobre el conjunto de entrenamiento llamado K-fold donde K=10 (ver la sección (**??**)), adicionalmente para cada parámetro se repitió el procedimiento tres (3) veces, a esto se le llama (repeated K-Fold) con el objetivo principal de realizar el cálculo del RMSE y el R^2 para cada modelo y variación del parámetro propuesto.

Por otro lado se evaluó con el conjunto de datos disponible la importancia que tenía el porcentaje de muestreo asignado a entrenamiento y prueba, ya que en la bibliografía disponible se hace referencia a particiones de 75 % para entrenamiento y 25 % para prueba, pero el buen desempeño de esta elección esta supeditada a los datos. Usando uno de los modelos con mejores resultados y con mayor requerimiento de recursos se evaluó la posible ganancia en la variación de la proporción de entrenamiento de un mismo modelo usando el algoritmo de random forest (con 200 árboles, y usando mtry=13) obteniendo los siguientes resultados:

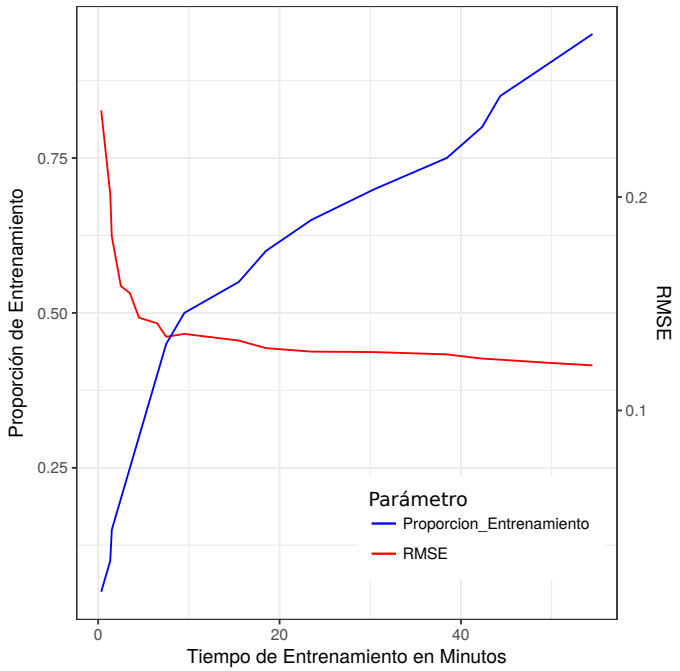


Figura 2.11: Evaluación de modelos de bosques aleatorios - Random Forest respecto a la ganancia en términos de minimización de RMSE y tiempo de procesamiento vs porcentaje de datos asignados al conjunto de entrenamiento.

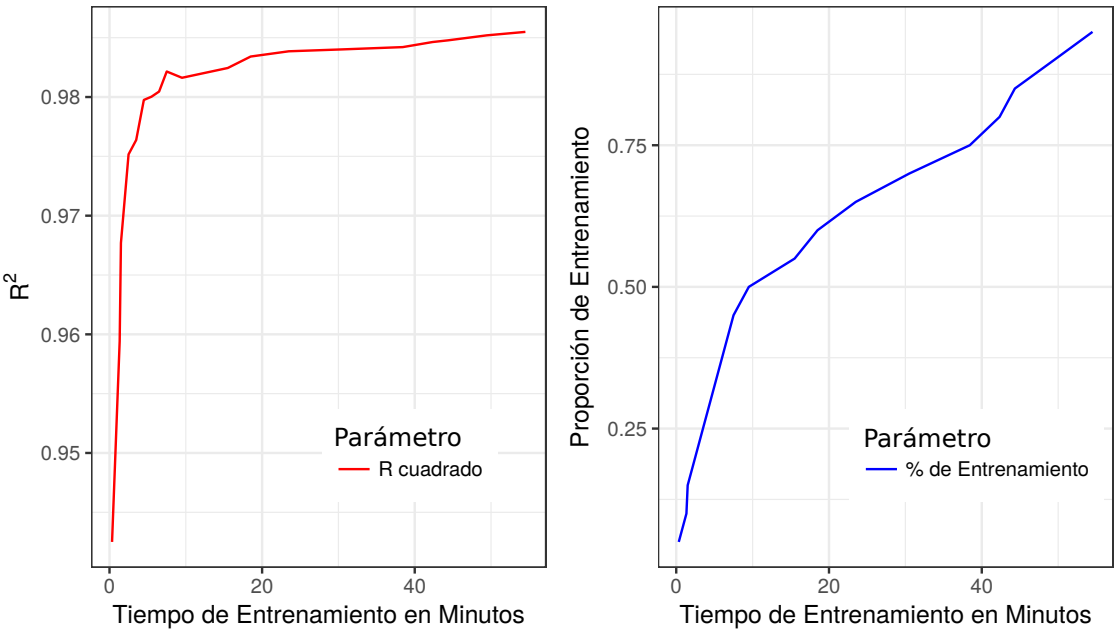


Figura 2.12: Evaluación de modelos de bosques aleatorios - random forest respecto a la ganancia en términos de R^2 y tiempos de procesamiento, respecto al porcentaje de datos asignados al conjunto de entrenamiento

Como se observa la ganancia en R^2 para el conjunto de datos a medida que crece la proporción de muestreo es de poco más de 1 % entre proporciones aproximadas a 20 % de los datos hasta proporciones mayores a 75 % que es lo recomendado por la literatura. Por lo contrario el incluir más datos incrementa los tiempos de procesamiento en el entrenamiento de los modelos, esto lo que significa es que por lo menos en los datos disponibles para nuestra investigación, la ganancia en R^2 y minimización de RMSE no es tan grande dentro de cada método, pero si hay un requerimiento importante de recursos y tiempo al momento de usar proporciones mayores para la selección del conjunto de entrenamiento de los datos, al menos en el proceso de optimización de los modelos propuestos es importante tener esta consideración.

Este resultado es un dato adicional relacionado al funcionamiento de los modelos, ya que su desempeño va a depender de la proporción de datos usados para el entrenamiento y prueba de los mismos, por esta razón en todos los casos para que los resultados fueran comparables se usaron selecciones idénticas para poder realizar comparaciones para los resultados presentados.

3. Resultados

Esta sección se presentan los resultados obtenidos en esta investigación en la siguiente secuencia: En primer lugar, se presentan aspectos importantes que se deben considerar para el análisis e interpretación de los resultados, en segundo lugar se realizará el recorrido por los resultados más importantes para la predicción del IAF; en tercer lugar se realizará el recorrido por los resultados más importantes para la predicción del TCH; y finalmente se presentarán otros resultados complementarios que ayudan a comprender aspectos asociados al uso de modelo de aprendizaje automático y finalmente se hará una síntesis respecto a los resultados obtenidos en este trabajo.

El objetivo general de esta investigación consistía en la evaluar modelos de aprendizaje automático integrando datos de imágenes multiespectrales e información tomada en campo para la predicción de variables asociadas a la producción (IAF y TCH) en parcelas experimentales de caña de azúcar, uno de los resultados destacables en esta investigación fue el desarrollo de una metodología que permite lograr bases de datos coherentes y capaces de unificar información de las distintas fuentes disponibles, volcando los datos en un sistema de información geográfica, esto permite posteriormente realizar la evaluación y modelado de los datos, apuntando a sistemas de agricultura de precisión en cultivos de caña de azúcar y facilitando la toma de decisiones.

Un aspecto central en la comparación de los modelos es que cada uno de los modelos evaluados a continuación presentan características particulares, pero en todos los casos el conjunto de entrenamiento es idéntico al igual que los conjuntos de datos seleccionados para realizar el proceso de validación y prueba como se presentó en **??** y en **??**, así no hay ventajas para ninguno modelo en particular, por lo tanto los resultados dependen totalmente de la capacidad de cada modelo para encontrar la mejor estrategia hasta lograr un criterio de parada y ajuste del modelo.

Existen varios aspectos sobre los cuales se puede realizar la medición de la calidad y resultados de un modelo enfocado a la predicción, alineados a esto a continuación se presentan dos aspectos que serán considerados para determinar qué combinación de factores nos lleva a concluir que un modelo o una combinación interna de parámetros dentro de un modelo, hace que este sea mejor que otro modelo u combinación en evaluación, el primer aspecto a presentar será respecto a los tiempos de entrenamiento y generación de dichos modelos, el cual es un aspecto de gran interés cuando se desea generar metodologías escalables que puedan ser utilizadas a futuro por el sector de la agroindustria como es nuestro caso, por otro lado se realizará la evaluación a partir de las métricas presentadas en la sección **??**.

3.1. Predicción IAF (Índice Área Foliar)

Respecto a los tiempos de ejecución relacionados a la evaluación de los modelos, tenemos que estos tiempos se dividen en el tiempo de evaluación de todas las variaciones en los parámetros propios de cada modelo que podemos observar en la tabla (**??**) y otro es el tiempo de ejecución tomado para la creación del modelo final o mejor modelo encontrado en cada algoritmo, estos tiempos corresponden a los tiempos finales de ejecución en una maquina con capacidad de procesamiento en paralelo con las características presentadas en **??**, Este mismo proceso en computadoras de escritorio tendría tiempos de ejecución que pueden multiplicarse hasta por 600 veces según pruebas realizadas en una primer instancia de esta investigación, razón por la cual se usaron algoritmos paralelizables para solucionar el problema desde un enfoque de big data.

	Modelo	**Usuario***	**Sistema****	**Transcurrido**	Iterados
Todos	**PLS**	1,212	0,048	1,860	20
	BaggedCART	2,784	0,00	5,979	-
	RandomForest	18,172	1,688	1698,719	20
	Cubist	155,608	0,224	569,255	24
Final	**PLS**	0,100	0,000	0,098	1
	BaggedCART	1,6	0,00	1,6	1
	RandomForest	8,228	1,032	13,189	1
	Cubist	152,804	0,000	152,790	1

* El tiempo de usuario es el tiempo de cargado en CPU para la ejecución de instrucciones del usuario en la llamada del proceso.
**El tiempo del sistema es el tiempo para la ejecución en CPU por parte del sistema en respuesta a la llamada del proceso.
▽ Se presenta la información de tiempo en segundos.
Todos: Tiempo tomado para evaluar todos los modelos asociados a la parametrización del modelo entrenado.
Final: Tiempo Tomado para el entrenamiento del modelo final o mejor modelo segun la parametrización realizada en cada modelo.
Iterados: cantidad de modelos entrenados en búsqueda de parámetros optimos para cada modelo.

Tabla 3.1: Tiempos de ejecución de los proceso de optimización y búsqueda de parámetros aplicados a la búsqueda del mejor modelo para predicción de IAF

En cuanto a tiempos de ejecución en la tabla (**??**) observamos dos características que ayudarían a medir la eficiencia de estos modelos, esta eficiencia se puede considerar en dos aspectos, primero la eficiencia de optimización de los modelos propuestos medidos por los tiempos totales tomados para todas las pruebas realizadas en cada modelo (todos los modelos); segundo la eficiencia en el cálculo del mejor modelo (modelo final). Dado que nuestro objetivo es medir la eficiencia respecto al mejor modelo encontrado podríamos decir que en términos de tiempo el ranking de velocidad en el ajuste y entrenamiento de los modelos estaría dado por:

1. PLS
2. BaggedCART
3. RandomForest

4. Cubist

Los tiempos de entrenamiento de los modelos, juegan un papel crucial en el momento en que se desea realizar el ajuste de los parámetros disponibles para dichos modelos, debido a que puede existir infinidad de parámetros a ajustar dependiendo del modelo evaluado, por esta razón algoritmos que se demoren tiempos altos, tienden a generar un incremento exponencial en los tiempos de ejecución y validación de los resultados, esto significa que la evaluación final de un modelo puede pasar de 3 minutos a varias horas por vez, dependiendo de los parámetros a ser evaluados y las capacidades computacionales disponibles, así, si otro algoritmo requiere ajuste de menos parámetros y tiene un rendimiento similar en términos de capacidad predictiva y minimización del error cometido en la predicción, la decisión será más fácil de tomar en términos del esfuerzo computacional requerido.

Para las siguientes evaluaciones los modelos fueron entrenados con el 75 % de la información disponible y se evaluó mediante k-fold sobre estos datos, el otro 25 % de la información se dejó como conjunto de prueba, siguiendo el esquema presentado en la figura **??**.

3.1.1. BaggedCART

En el modelo de BaggedCART no se realiza un proceso dirigido para la estimación de parámetros que mejoren el funcionamiento de este, internamente el modelo realiza únicamente un proceso de búsqueda del número óptimo de replicaciones bootstrap, donde se encontró que 25 replicaciones hacen óptimo el ajuste del modelo y lo hace estable para la predicción.

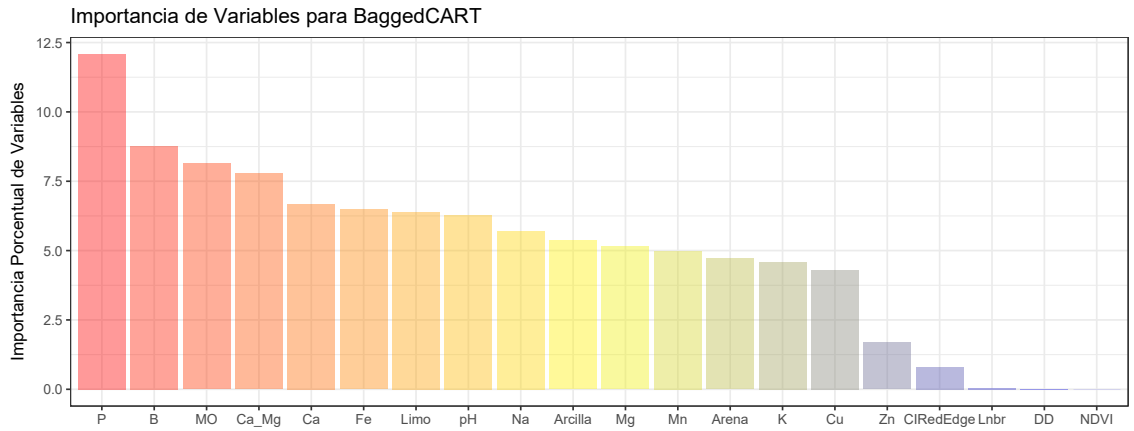


Figura 3.1: Importancia de Variables según modelo final de BaggedCART aplicado a predicción de IAF. Ordenados de mayor a menor nivel de importancia según sus nivel de importancia

Se evaluó la importancia de las variables predictoras sobre el resultado (ver la figura (**??**)), se observa que las variables más importantes para el modelo son: P, B, Mo, Ca/Mg, Ca los cuales son considerados nutrimentos del cultivo y cuya abundancia o deficiencia tienen un impacto importante en el crecimiento de la planta, y por ende hay una estrecha relación con el índice de área foliar que

esta pueda tener. Es importante resaltar que las variables con información de índices de vegetación, aparecen al final del nivel de importancia y son las variables que menos aportan en el modelo.

Este modelo aplicado a los datos seleccionados para la fase de prueba obtuvo un R^2 de 81,93 % lo que habla de la capacidad que tiene el modelo de predecir sobre nuevos conjuntos de datos, mientras el RMSE fue de 0,4288 dado que el RMSE posee la misma escala que la variable respuesta IAF, esto significaría que le modelo se puede llegar a desviar en menos de 1 unidad en la predicción realizada, respecto a la realidad.

3.1.2. Regresión por Mínimos Cuadrados Parciales (PLS)

Se realizó la estimación del modelo con PLS variando el parámetro #**Componentes**, se determinó que de las 20 variables incluidas en el modelo inicialmente con 19 Componentes se minimizaba el RMSE, de acuerdo a lo presentado en la sección **??** y alineado con lo presentado en (**?**) para PLS, este criterio fue el que se consideró para determinar que éste parametro maximiza el desempeño del modelo, minimizando el error del mismo. En la figura (**??**) se observa cómo a medida en que crece la cantidad de componentes incluidos en el modelo, el RMSE disminuye hasta ser mínimo, se puede ver el detalle en la tabla Anexa **??** que incluir más variables hace que el RMSE aumente y por lo tanto esto indicaría generación de ruido al incluir más componentes en el modelo.

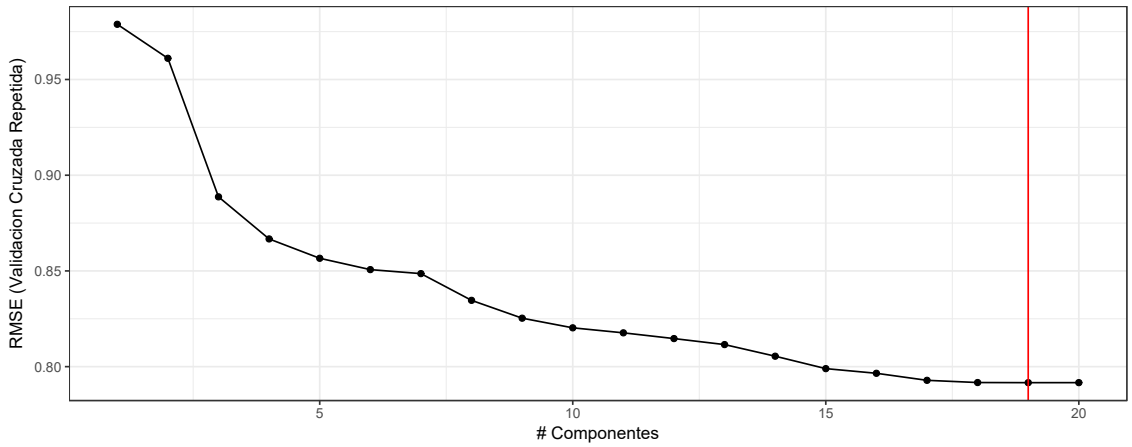


Figura 3.2: Validación cruzada para la búsqueda de la cantidad optima de componentes para la predicción de IAF mediante modelo PLS para la minimización del RMSE

Se evaluó la importancia de las variables, ver la figura (**??**), donde se observa que las variables más importantes para el modelo son: B, Na, Cu, Mo, Lnbr en este modelo las variables relacionadas con información de índices de vegetación a diferencia que el modelos de baggedCART no son las variables que menos aportan en el modelo, por el contrario la variable P, que era la que más aportaba en dicho modelo en PLS considera que esta da un aporte mínimo, y las variables relativas a índices de vegetación tienen un aporte medio.

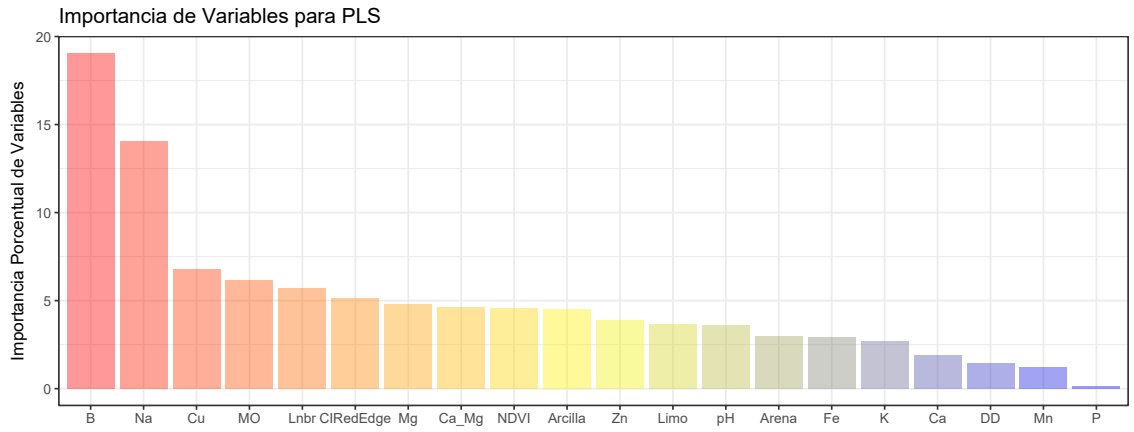


Figura 3.3: Importancia de las variables según modelo final de PLS en predicción de IAF.

Mediante este modelo entrenado previamente se obtuvieron las métricas asociadas a la fase de entrenamiento que se pueden ver en tabla anexa **??**, mientras sobre el conjunto de datos de prueba se obtuvo un R^2 de 37,76 % y RMSE de 0,7879, en términos generales este modelo está muy cercano a los resultados obtenidos con el modelo anterior de BaggedCART.

3.1.3. Cubist

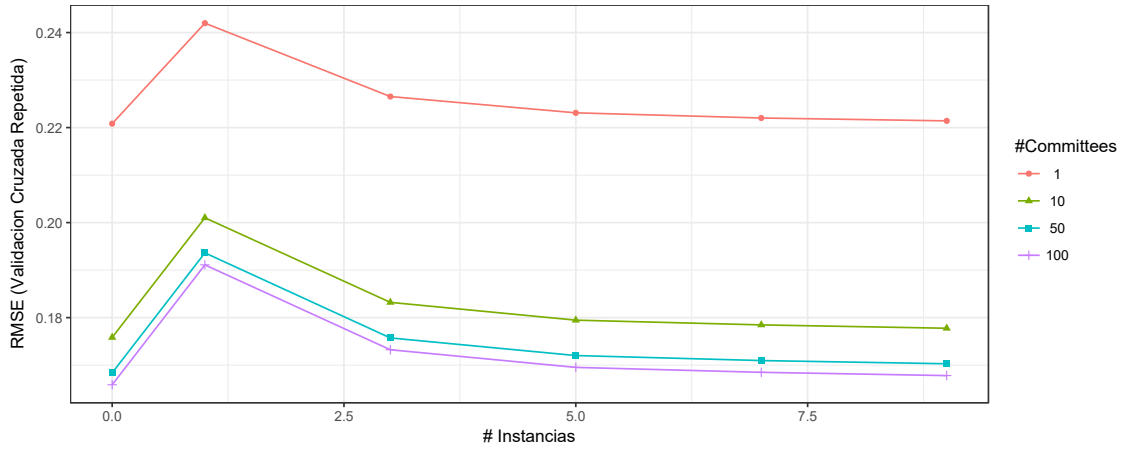


Figura 3.4: Representación gráfica del RMSE obtenido en los modelos resultantes para la conjugación de los parámetros a considerar en el modelo Cubist para la predicción de IAF

El cubist es un modelado de regresión utilizando reglas con correcciones basadas en añadir instancias, dichas instancias están compuestas por la combinación de dos parámetros, la variación conjunta de estos parámetros en la figura (**??**) llevo al modelo que incluía 100 committees y 0 vecinos cercanos logrando la minimización del RMSE el cual es el objetivo a lograr de acuerdo a lo presentado en **??**, ésta configuración de commites y vecinos cercanos se vería como árboles independientes para cada una de las reglas de predicción generadas, el detalle de esta búsqueda se puede consultar en **??**.

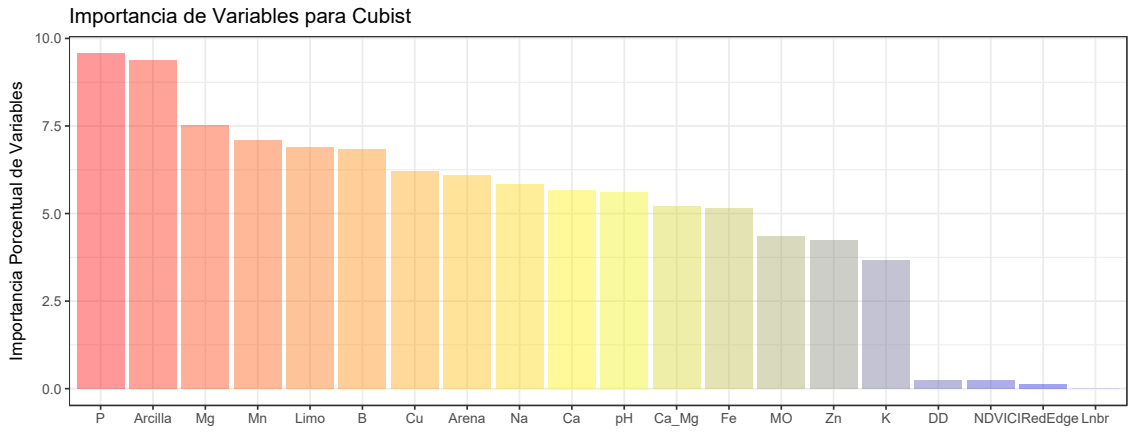


Figura 3.5: Importancia de las variables para modelo final de Cubist aplicado en la predicción del IAF.

Posteriormente se evaluó la importancia de las variables como se observa en la figura (**??**), donde las variables más importantes para el modelo cubist son: P, Arcilla, Mg, Mn, Limo y en último lugar se encuentran las variables relativas a información de índices de vegetación como las variables que menos aportan en el modelo.

En la figura (**??**) se observa cómo a medida que se varian los parámetros el RMSE disminuye hasta ser mínimo, se puede ver el detalle en la tabla anexa **??** .

El mejor modelo entrenado se puso a prueba entregando un R^2 que aumenta considerablemente, llegando a un 98,18 % y un RMSE de 0,1348 valor que es menor con respecto a los dos modelos presentados previamente.

3.1.4. Bosques Aleatorios (Random Forest)

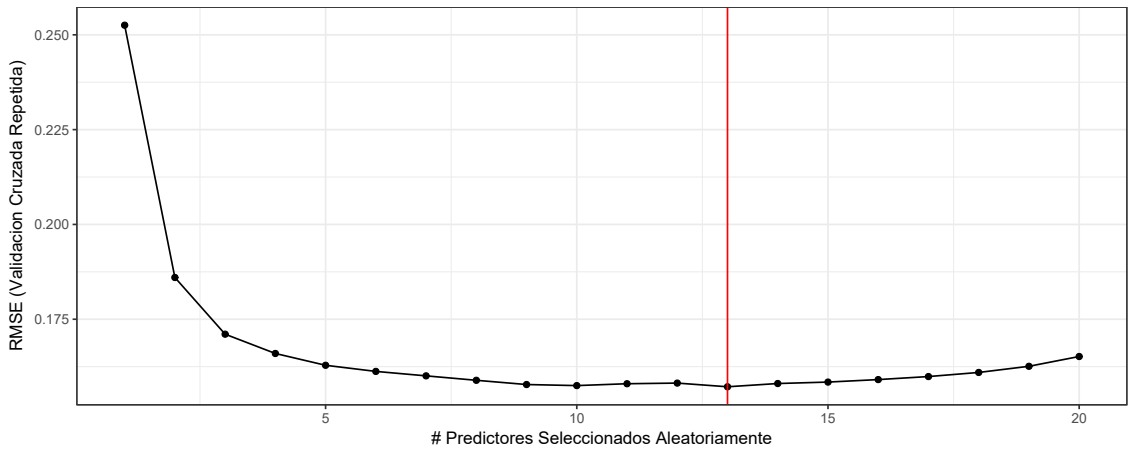


Figura 3.6: Proceso de validación cruzada para la búsqueda de la cantidad optima de variables para la predicción de IAF mediante el modelo de Random Forest medido por la minimización del RMSE y estabilización del mismo

El último modelo evaluado fue el Random Forest en el cual se realiza un ajuste variando el pará-

metro #**mtry**, el cual determinó que de las 20 variables incluidas en el modelo inicialmente con 13 predictores muestreados para la división de las ramas se minimizaba el RMSE. En la figura (**??**) se observa cómo a medida en que crece la cantidad de componentes incluidos en el modelo, el RMSE disminuye hasta ser mínimo en 13 predictores, incluir más variables lo único que generaría sería ruido y se perdería precisión, ver detalle en tabla adjunta **??**.

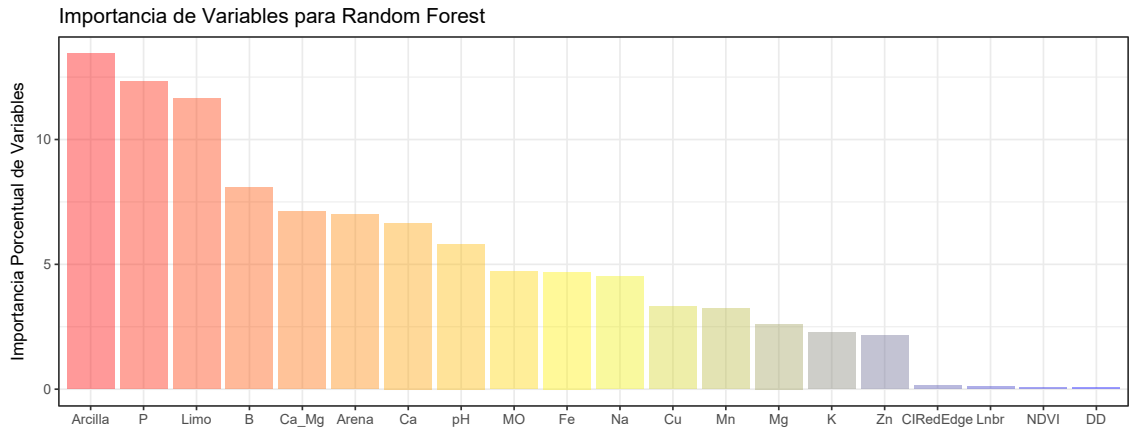


Figura 3.7: Importancia de las variables según el modelo final de RandomForest para predicción de IAF

Se evaluó la importancia de las variables (figura (**??**)), donde se observa que las variables más importantes para el modelo son: Arcilla, P, Limo, B, Ca/Mg y en último lugar las variables relativas a información de índices de vegetación como las variables que menos aportan en el modelo.

Aplicando el mejor modelo a el conjunto de datos de prueba se obtuvo un R^2 que aumenta, llegando a un 98,41 % y un RMSE de 0,12698 este es el mínimo con respecto a los tres modelos evaluados previamente, pero se debe revisar cuidadosamente entre el modelo Cubist y el Random Forest en busca de determinar cuál de estos es mejor en términos de costo eficiencia computacional debido que su rendimiento en términos de precisión fue muy similar.

3.1.5. Comparación Múltiple de Modelos Utilizados para Estimar IAF

Siguiendo la metodología planteada en (**??**), se consideró cada uno de los resultados previamente presentados, lo que nos muestra cómo fue el rendimiento de cada modelo evaluado de una manera independiente. Usando la información resultante de cada uno de estos modelos y teniendo en cuenta la cantidad de información disponible de cada re-muestreo realizado internamente para la evaluación de los parámetros de los modelos, se hizo uso de esta información para entregar una visión más clara de los resultados hallados, lo que nos permite hacer una comparación:

El objetivo final en problemas de regresión es minimizar la raíz cuadrada del error cuadrático medio (RMSE) y maximizar el ajuste del modelo medido por el R^2, estas métricas se miden con la

finalidad de obtener un criterio para poder comparar y calificar los modelos en búsqueda del mejor de estos.

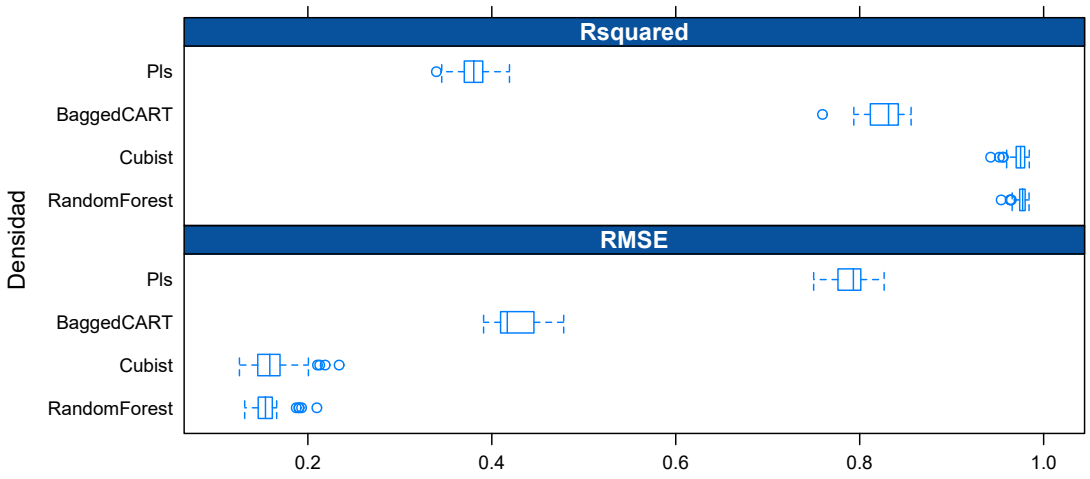


Figura 3.8: Comparación de Métricas (RMSE y R^2) en predicción de IAF.

En la figura (??) se observa que los modelos que minimizan de una manera más eficiente el RMSE y adicionalmente generaron un mejor ajuste del modelo medido por el R^2 son los modelos Cubist y Random Forest.

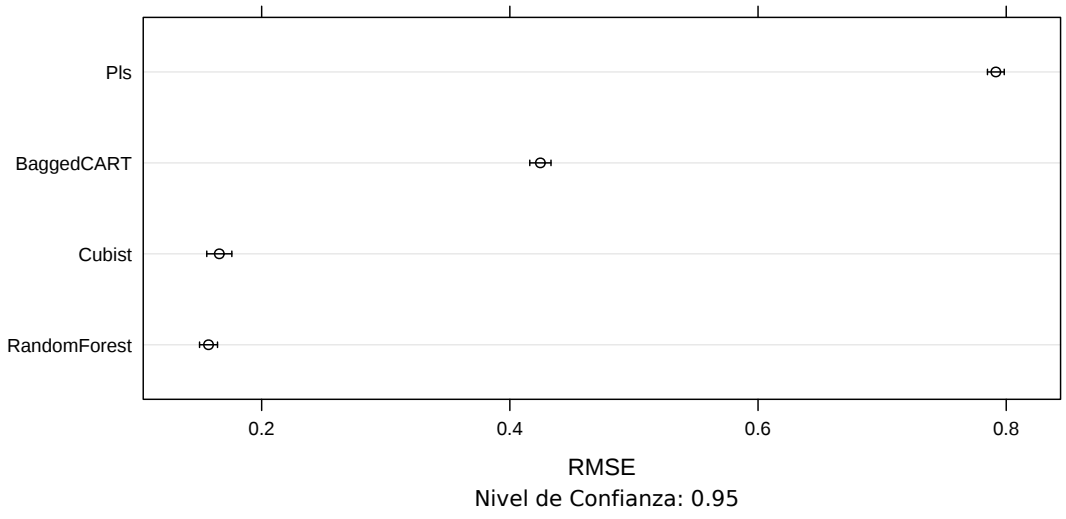


Figura 3.9: Comparación de RMSE estimado entre los distintos modelos y re-muestreos en predicción de IAF.

Se ajustaron intervalos de confianza de 95 % a partir de los datos de remuestreo resultante del proceso de validación para el RMSE, como se observa en la figura (??) el modelo con peor desempeño es el PLS, seguido de BaggedCART, y el mejor desempeño se disputa entre los modelos de

Random Forest y Cubist, no hay una definición clara entre estos, ya que los intervalos de confianza se solapan para este par de modelos.

Haciendo una comparación por pares de modelos mediante la diferencia del RMSE, según el remuestreo realizado con los datos de cada modelo podemos determinar qué modelo es mejor que otro, según el error de cada uno de los modelos, obteniendo así que el modelo Cubist y el Random Forest son casi idénticos en desempeño como se observa en la figura **??**. Sin embargo, se podría considerar mejor el modelo de Random Forest debido a que su variabilidad fue inferior, esto se puede ver tanto en la figura de comparación **??** como en el rango intercuartílico del boxplot de la figura **??**. No obstante, estos intervalos se solapan por lo cual no es totalmente concluyente la diferencia entre este par de modelos, esto se prueba estadísticamente usando los métodos planteados en (**??**) haciendo uso de la prueba de Bonferroni para comparaciones múltiples, ver **??** donde observamos que en el único par de comparaciones donde no se pueden concluir diferencias es en la comparación de Cubist vs Random Forest.

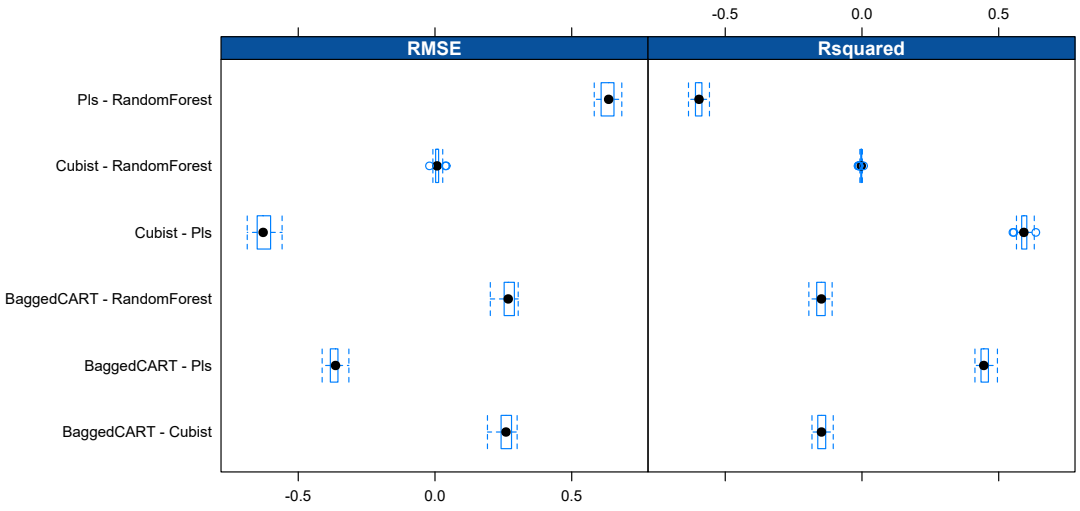


Figura 3.10: Diferencias entre modelos comparados para la predicción de IAF.

Haciendo uso del modelo entregado por el modelo de Random Forest en la figura **??** se presenta gráficamente la predicción realizada sobre la totalidad del área sembrada de las parcelas experimentales y se compara con la información de los valores originales (en el área determinada para muestreo, entrenamiento y validación), este grafico se presenta para apreciar el comportamiento del modelo entrenado y su precisión en la predicción, resaltando que se incluye predicción para un área mayor al área de entrenamiento del modelo, la idea de presentar esta imagen es observar la capacidad predictiva del modelo inclusive en área no observadas previamente.

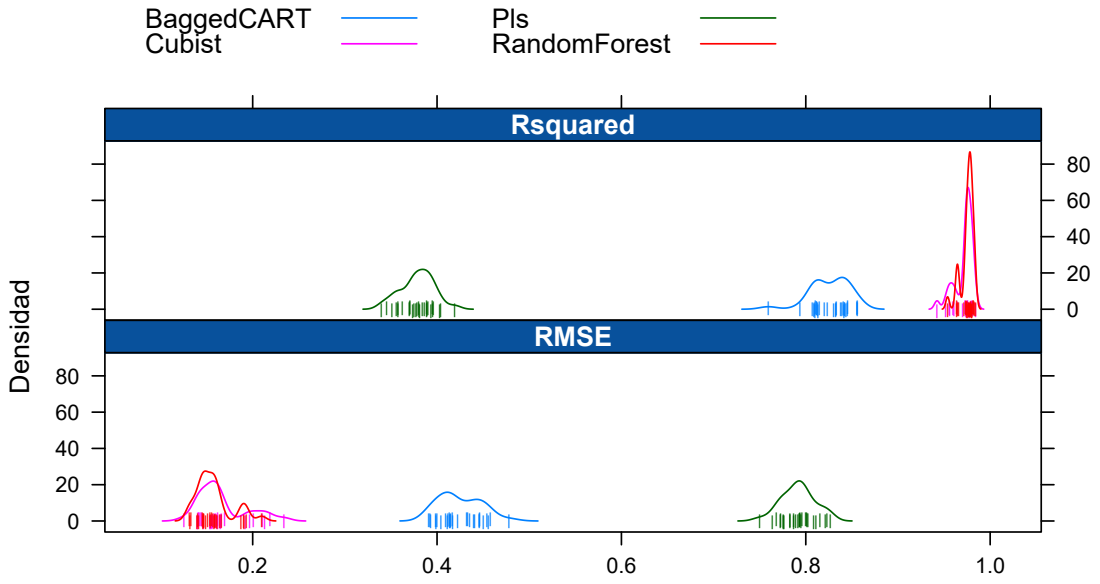
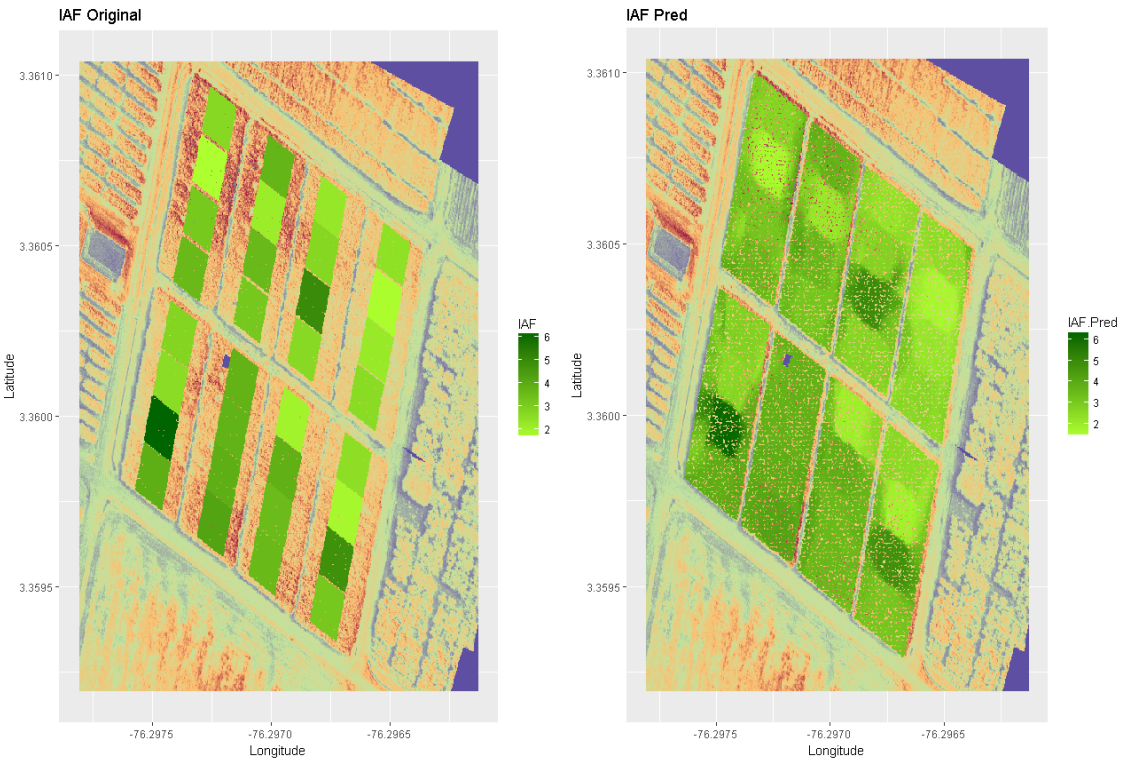


Figura 3.11: Densidad de respuesta en R^2 y RMSE para los modelos evaluados en la predicción de IAF.



A la izquierda se grafican los polígonos de muestreo originales con los valores de respuesta para IAF, a la derecha se presenta la predicción realizada usando el algoritmo de Random Forest sobre el área total sembrado de caña de azúcar relativa al experimento, en verde se presenta la correspondencia del IAF, ambas imágenes poseen de fondo una representación visual en pseudocolor referencia del área de estudio.

Figura 3.12: Comparación entre el valor de IAF registrado en área de muestreo y la predicción realizada usando el modelo generado en el modelo Random Forest sobre las parcelas completas del experimento.

3.2. Resultados Predicción de TCH (Toneladas de Caña por Hectárea)

En términos generales se siguió una configuración al igual que para la predicción del IAF donde se agregó ruido blanco a la respuesta en el conjunto de datos de entrenamiento y se realizó un re-muestreo sobre el conjunto de entrenamiento llamado K-fold donde K=10 de acuerdo a lo presentado en **??**. Adicionalmente, para cada parámetro se repitió el procedimiento tres (3) veces, (repeated K-Fold) con el objetivo de realizar el cálculo del RMSE y medir la bondad de ajuste del modelo a los datos R^2 para cada modelo y variación del parámetro propuesto, otro aspecto a considerar es la interpretación de los resultados del TCH a nivel de pixel para su correcta interpretación, se debe tener en cuenta las consideraciones presentadas en (**??**).

Respecto a los tiempos de ejecución relacionados a la evaluación de los modelos, tenemos que estos tiempos se dividen en el tiempo de evaluación de todas las variaciones en los parámetros propios de cada modelo que podemos observar en la tabla (**??**) y otro es el tiempo de ejecución tomado para la creación del modelo final o mejor modelo encontrado en cada algoritmo,

	Modelo	**Usuario***	**Sistema****	**Transcurrido**	Iterados
Todos	**PLS**	1,028	0,032	1,608	20
	BaggedCART	2,92	0,00	6,74	-
	RandomForest	18,068	1908	1729,745	20
	Cubist	184,988	0,340	693,368	24
Final	**PLS**	0,100	0,000	0,098	1
	BaggedCART	1,712	0,000	1,711	1
	RandomForest	7,940	1,320	13,382	1
	Cubist	181,424	0,000	181,408	1

* El tiempo de usuario es el tiempo de cargado en CPU para la ejecución de instrucciones del usuario en la llamada del proceso.

**El tiempo del sistema es el tiempo para la ejecución en CPU por parte del sistema en respuesta a la llamada del proceso.

▽ Se presenta la información de tiempo en segundos.

Todos: Tiempo tomado para evaluar todos los modelos asociados a la parametrización del modelo entrenado.

Final: Tiempo Tomado para el entrenamiento del modelo final o mejor modelo segun la parametrización realizada en cada modelo.

Iterados: cantidad de modelos entrenados en búsqueda de parámetros optimos para cada modelo.

Tabla 3.2: Tiempos de ejecución para el proceso de optimización y búsqueda de parámetros aplicados a la búsqueda del mejor modelo para predicción de TCH

En cuanto a tiempos de ejecución en el cuadro (**??**) observamos dos características que ayudarían a medir la eficiencia de estos modelos, esta eficiencia se puede considerar en dos aspectos, primero la eficiencia de optimización de los modelos propuestos medidos por los tiempos totales tomados para todas las pruebas realizadas en cada modelo (todos los modelos); segundo la eficiencia en el cálculo del mejor modelo (modelo final). Dado que nuestro objetivo es medir la eficiencia respecto al mejor modelo encontrado podríamos decir que en términos de tiempo el ranking de velocidad en el ajuste y entrenamiento de los modelos estaría dado por:

1. PLS
2. BaggedCART
3. RandomForest
4. Cubist

Esta información es importante al momento de pensar en aplicar estos modelos a nuevos conjuntos de datos, adicionalmente son clave para determinar una evaluación efectiva de los modelos evaluados.

Para las siguientes evaluaciones los modelos fueron entrenados con el 75 % de la información disponible y se evaluó mediante k-fold sobre estos datos, el otro 25 % de la información se dejó como conjunto de prueba, siguiendo el esquema presentado en la figura **??**.

3.2.1. BaggedCART

En el modelo de BaggedCART no se realiza un proceso dirigido para la estimación de parámetros que mejoren el funcionamiento de este, internamente el modelo realiza únicamente un proceso de búsqueda del número óptimo de replicaciones bootstrap, donde se encontró que 25 replicaciones hacen óptimo el ajuste del modelo y lo hace estable para la predicción.

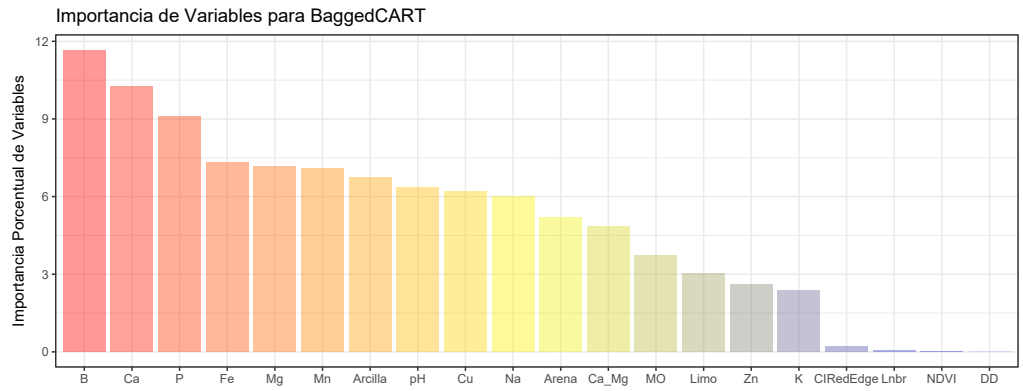


Figura 3.13: Importancia de Variables según modelo final de BaggedCART aplicado a predicción de TCH

Se evaluó la importancia de las variables predictoras sobre el resultado (ver la figura **??**), se observa que las variables más importantes para el modelo son: B, Ca, P, Fe, Mg los cuales son considerados nutrimentos del cultivo y cuya abundancia o deficiencia tiene un impacto importante en el crecimiento de la planta y por lo tanto en su rendimiento. Al igual que en los resultados vistos para IAF resalta que las variables relativas a información de índices de vegetación, aparece al final en términos de importancia, aportando menos al modelo.

Mediante este modelo se obtuvo un R^2 de 85,81 % lo que habla de la capacidad que tiene el modelo de predecir sobre nuevos conjuntos de datos. Mientras el RMSE de 11,209 mide la distancia entre los valores predichos y los valores observados, entre más pequeño es este valor más preciso es el modelo en términos de variabilidad o desviaciones respecto a los valores observados.

dado que el RMSE posee la misma escala que la variable respuesta TCH, esto significaria que le modelo se puede llegar a desviar en menos de 12 toneladas por hectárea en la predicción realizada, respecto a la realidad, lo que es muy bueno teniendo en cuenta que la variable respuesta está alrededor de 100 a 250 toneladas por hectárea .

3.2.2. Regresión por Mínimos Cuadrados Parciales (PLS)

Se realizó la estimación del modelo PLS variando el parámetro #**Componentes**, se determinó que de las 20 variables incluidas en el modelo inicialmente con todas las 20 Componentes de minimizaba el RMSE, por lo tanto no hay una ganancia real al convertir las variables originales en componentes. En la figura **??** se observa cómo a medida en que crece la cantidad de componentes incluidos en el modelo, el RMSE disminuye hasta ser mínimo, se puede ver el detalle en la tabla Anexa **??**.

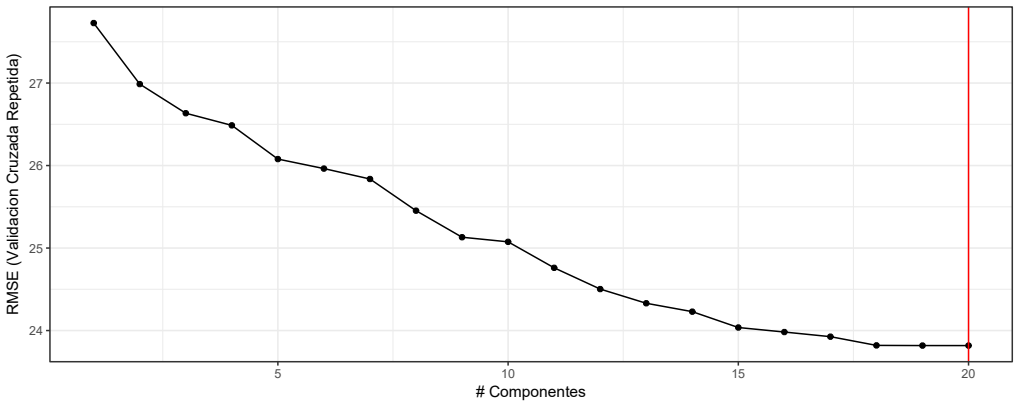


Figura 3.14: Proceso de validación cruzada para la búsqueda de la cantidad optima de componentes para la predicción de TCH mediante PLS para la minimización del RMSE y estabilización del mismo.

Se evaluó la importancia de las variables, ver la figura (**??**), donde se observa que las variables más importantes para el modelo son: B, K, pH, Na, Ca/Mg en este modelo las variables relacionadas con información de índices de vegetación se encuentran entre las variables de importancia media para el modelo, de manera similar a lo que ocurrió en el caso de la predicción de IAF, pero de la misma forma este modelo es el que peor desempeño tiene con un R^2 muy bajo y un RMSE alto.

Mediante este modelo entrenado previamente se obtuvieron las metricas asociadas a la fase de entrenamiento que se pueden ver en tabla anexa **??**, mientras sobre el conjunto de datos de prueba se obtuvo un R^2 de 34,73 % y RMSE de 23,64 en terminos generales este modelo tiene un comportamiento peor que el modelo presentado anteriormente el de BaggedCART.

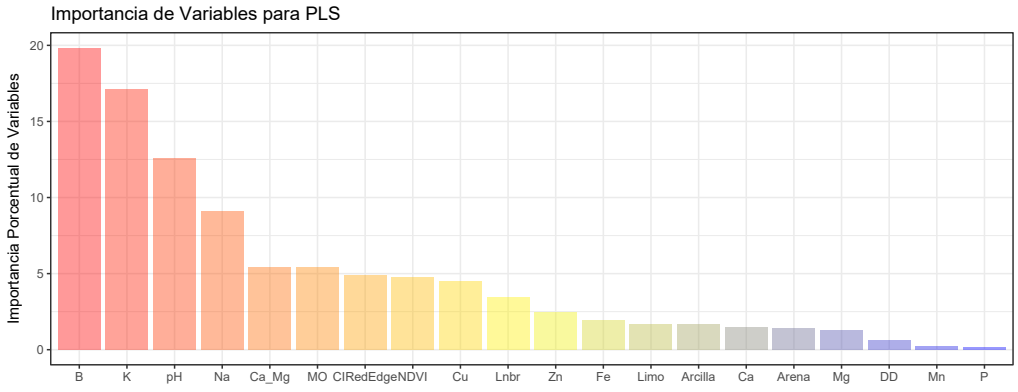


Figura 3.15: Importancia de las variables según modelo final de PLS en predicción de TCH

3.2.3. Cubist

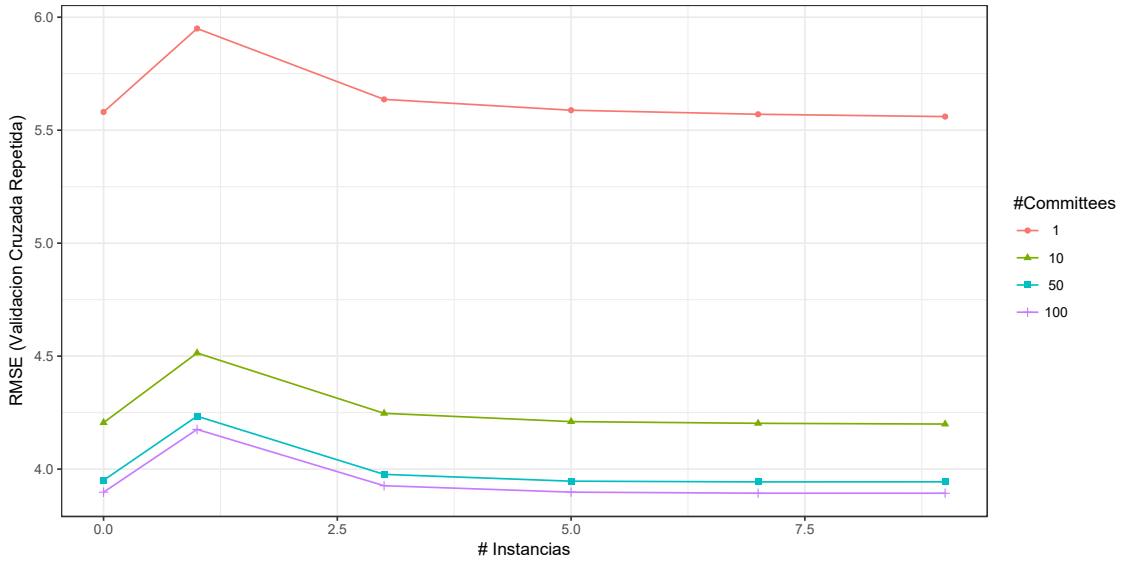


Figura 3.16: Evaluación de resultados obtenidos mediante aplicación del modelo Cubist para TCH. Representación gráfica del RMSE obtenido en los modelos resultantes para la conjugación de los distintos parámetros a considerar en el modelo Cubist para la predicción de TCH

El cubist es un modelado de regresión utilizando reglas con correcciones basadas en añadir instancias, dichas instancias están compuestas por la combinación de dos parámetros, la variación conjunta de estos parámetros en la figura?? muestra cómo el RMSE se minimiza según la configuración que se realice, para este caso el modelo que minimizó el RMSE fue el modelo que incluía 100 committees y considerando 9 vecinos cercanos. Esto significa que en este caso el TCH se ajustó por medio de árboles entrenados conjuntamente considerando hasta 9 vecinos cercanos o árboles con características de similaridad para la consecución del modelo de predicción, así esta instancia minimiza el error del modelo, el detalle de esta búsqueda se puede consultar en ??.

Posteriormente se evaluó la importancia de las variables como se observa en la figura ??, donde se observa que las variables más importantes para el modelo son: P, Arena, B, Mg, Mn y en último

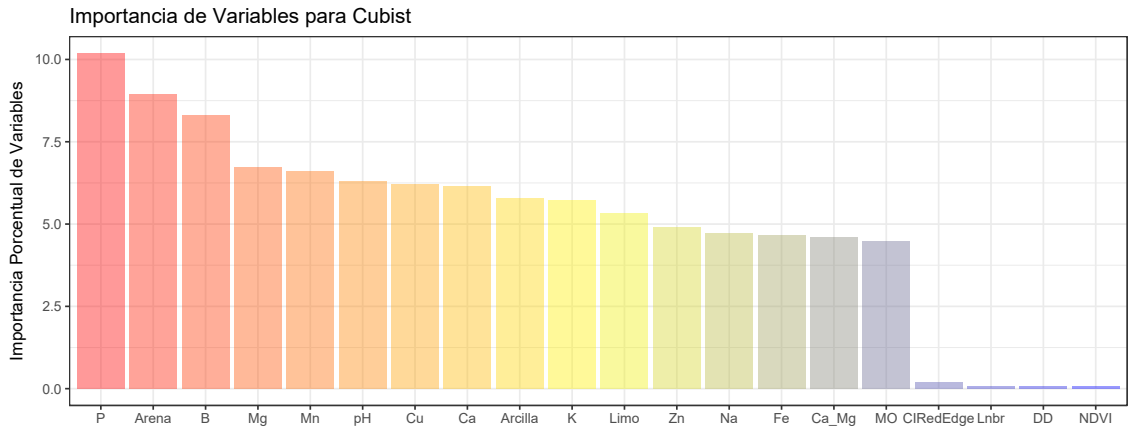


Figura 3.17: Importancia de las variables para modelo final Cubist aplicado en la predicción del TCH

lugar las variables relativas a información de índices de vegetación como las variables que menos aportan en el modelo.

El mejor modelo entrenado se puso a prueba entregando un R^2 que aumenta considerablemente, llegando a un 98,08 % y un RMSE de 4,054 valor que es menor con respecto a los dos modelos presentados previamente.

3.2.4. Bosques Aleatorios (Random Forest)

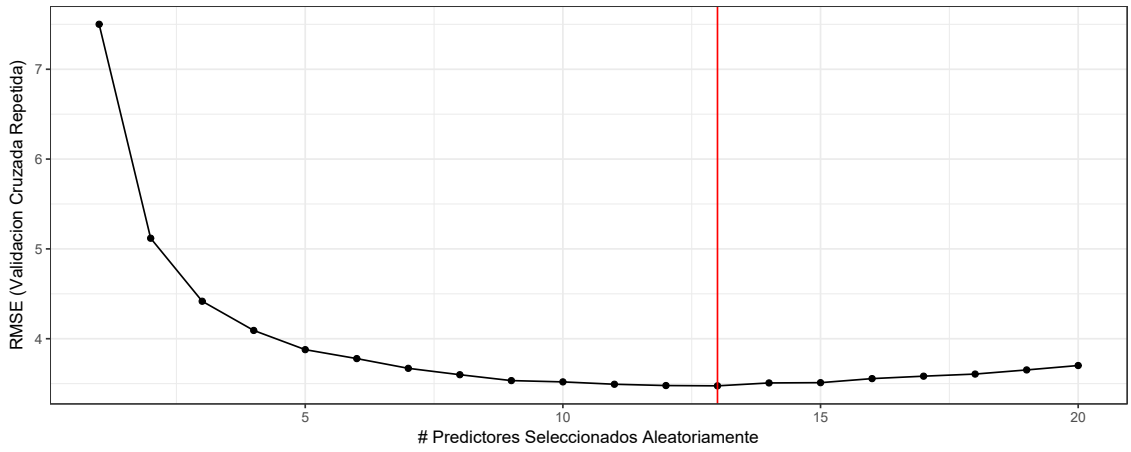


Figura 3.18: Proceso de validación cruzada para la búsqueda de la cantidad optima de variables para la predicción de TCH, mediante el modelo de Random Forest, en rojo se denota la selección de cantidad optima de variables medido por la minimización del RMSE y estabilización del mismo

Para el modelo de Random Forest se realiza un ajuste variando el parámetro #**mtry** , el cual determinó que de las 20 variables incluidas en el modelo inicialmente con 13 predictores muestreados para la división de las ramas se minimizaba el RMSE. En la figura **??** se observa cómo a medida en que crece la cantidad de componentes incluidos en el modelo, el RMSE disminuye hasta ser

mínimo en 13 predictores, incluir más variables lo único que generaría sería ruido y se perdería precisión, ver detalle en tabla adjunta **??**.

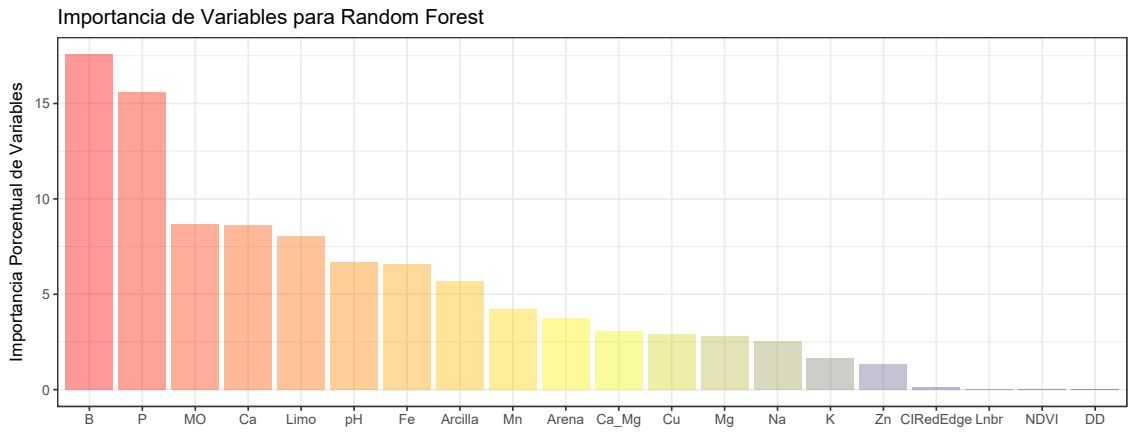


Figura 3.19: Importancia de las variables según el modelo final de Random Forest para predicción de TCH

Se evaluó la importancia de las variables (ver la figura **??**), donde se observa que las variables más importantes para el modelo son: B, P, Mo, Ca, Limo y en último lugar las variables relativas a información de índices de vegetación como las variables que menos aportan en el modelo.

Aplicando el mejor modelo a el conjunto de datos de prueba se obtuvo un R^2 que aumenta, llegando a un 98,50 % y un RMSE de 3,603 este es el mínimo con respecto a los tres modelos evaluados previamente, pero se debe revisar cuidadosamente entre el modelo Cubist y el Random Forest en busca de determinar cual de estos es mejor en terminos de costo eficiencia computacional debido que su rendimiento en terminos de precisión fue muy similar, al igual que ocurrio en la comparación de modelos de los modelos evaluados para el IAF.

3.2.5. Comparación Múltiple de Modelos Utilizados para Estimar TCH

Siguiendo la metodología planteada en (**??**), cada uno de los resultados previamente presentados nos muestra cómo fue el rendimiento de cada modelo evaluado de una manera independiente, usando la información resultante de cada uno de los modelos evaluados y teniendo en cuenta la cantidad de información disponible de cada re-muestreo realizado internamente para la evaluación de los parámetros de los modelos. Se hizo uso de esta información para entregar una visión más clara de los resultados hallados, lo que nos permite hacer una clara comparación:

El objetivo final en problemas de regresión es minimizar la raíz cuadrada del error cuadrático medio (RMSE) y maximizar el ajuste del modelo medido por el R^2, estas métricas se miden con la finalidad de obtener un criterio para poder comparar y calificar los modelos en búsqueda del mejor de estos.

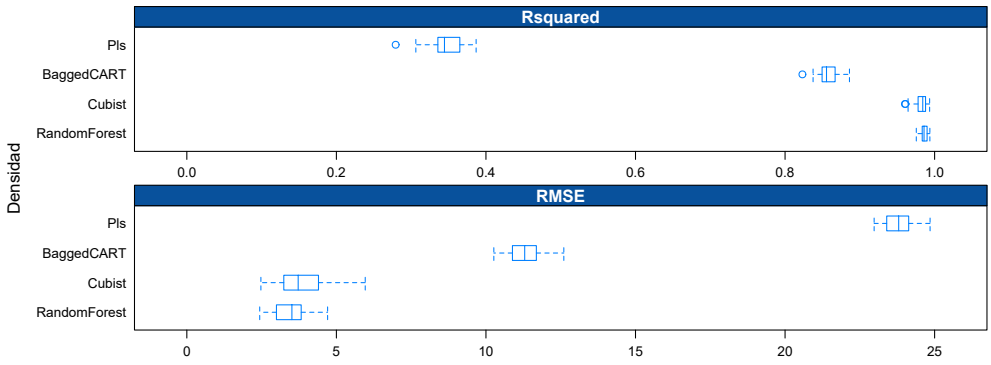


Figura 3.20: Comparación de Métricas (RMSE y R^2) en predicción de TCH

En la figura ?? se observa que los modelos que minimizan de una manera más eficiente el RMSE y adicionalmente que generan un mejor ajuste del modelo medido por el R^2 son los modelos Cubist y Random Forest.

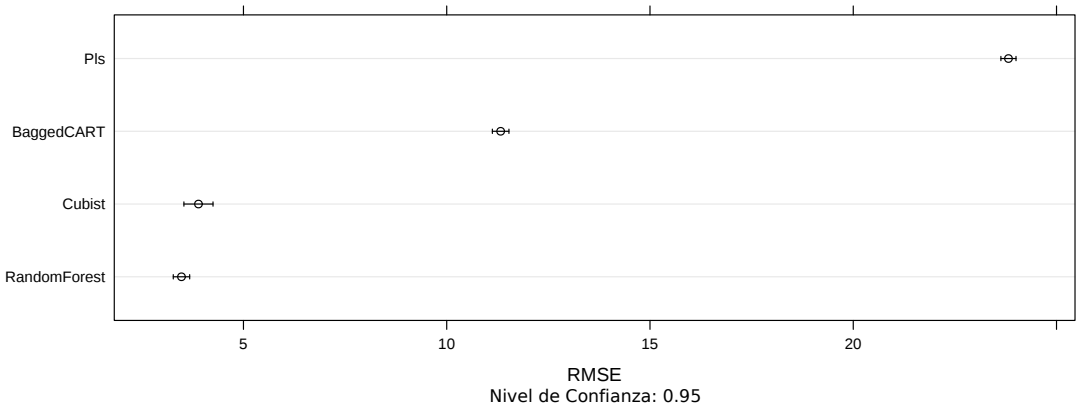


Figura 3.21: Comparación de RMSE estimado entre los distintos modelos y re-muestreos en predicción de TCH.

Se ajustaron intervalos de confianza de 95 % a partir de los datos de remuestreo resultante del proceso de validación para el RMSE, como se observa en la figura ?? el modelo con peor desempeño es el PLS, seguido de BaggedCART, y el mejor desempeño se disputa entre los modelos de Random Forest y Cubist, no hay una definición clara entre estos, ya que los intervalos de confianza se solapan para este par de modelos.

Haciendo una comparación por pares de modelos mediante la diferencia del RMSE, según el re-muestreo realizado con los datos de cada modelo podemos determinar que modelo es mejor que otro, según el error de cada uno de los modelos, obteniendo así que el modelo Cubist y el Random Forest son casi idénticos como se observa en la figura ??. Sin embargo, se podría considerar mejor el modelo de Random Forest debido a que su variabilidad fue inferior, esto se puede ver tanto en la figura de comparación ?? como en el rango intercuartílico del boxplot de la figura ??. No obstante,

estos intervalos se solapan por lo cual no es totalmente concluyente la diferencia entre este par de modelos, esto se prueba estadísticamente usando los métodos planteados en (??) haciendo uso de la prueba de Bonferroni para comparaciones múltiples, ver ?? donde observamos que en el único par de comparaciones que donde no se pueden concluir diferencias es en la comparación de Cubist vs Random Forest.

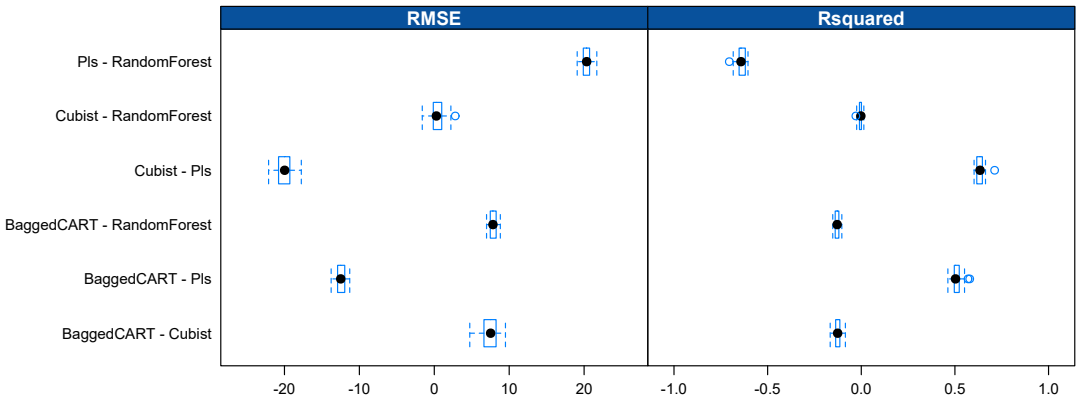


Figura 3.22: Diferencias entre modelos comparados para la predicción de TCH

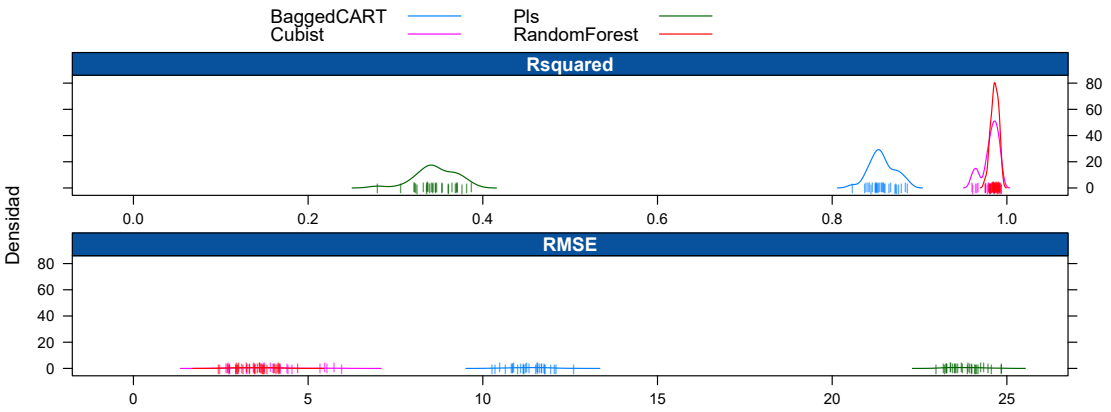
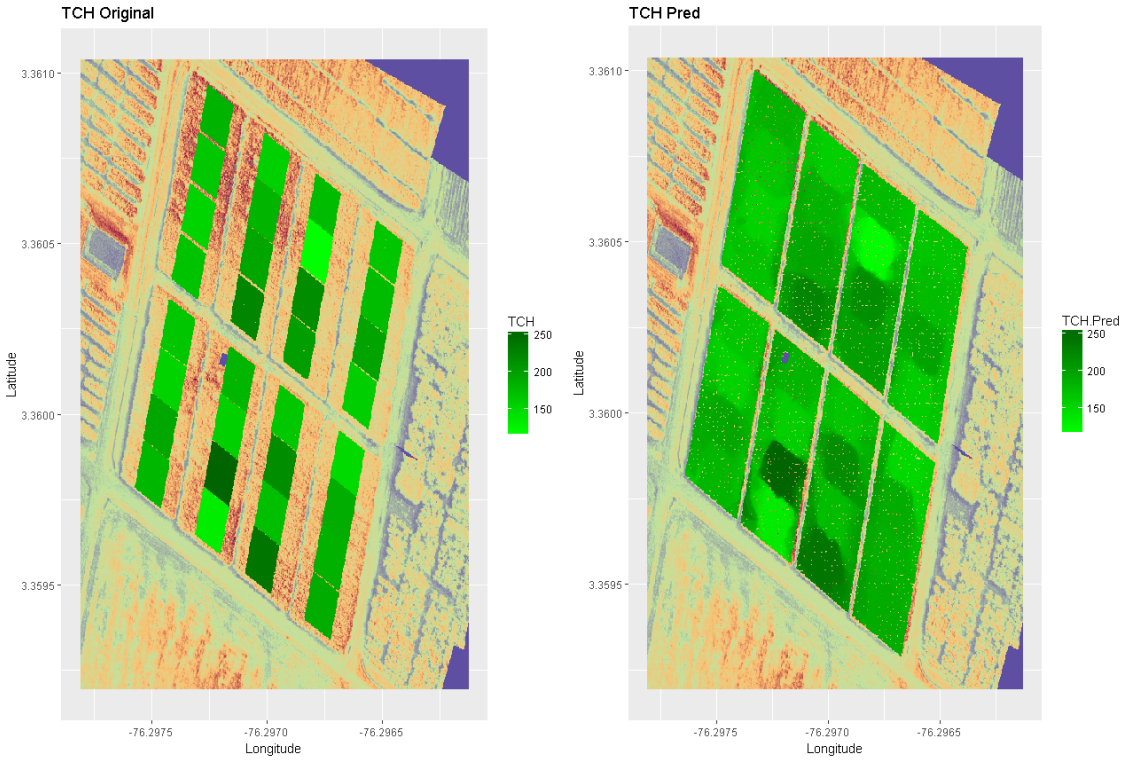


Figura 3.23: Densidad de respuesta en R^2 y RMSE para los modelos evaluados en la predicción de TCH.

Haciendo uso del modelo entregado por el modelo de Random Forest en la figura ?? se presenta gráficamente la predicción realizada sobre la totalidad del área sembrada de las parcelas experimentales y se compara con la información de los valores originales (en el área determinada para muestreo, entrenamiento y validación), este grafico se presenta para apreciar el comportamiento del modelo entrenado y su precisión en la predicción, resaltando que se incluye predicción para un área mayor al área de entrenamiento del modelo, la idea de presentar esta imagen es observar la capacidad predictiva del modelo inclusive en área no observadas previamente.



A la izquierda se grafican los polígonos de muestreo originales con los valores de respuesta para TCH, a la derecha se presenta la predicción realizada usando el algoritmo de Random Forest sobre el área total sembrado de caña de azúcar relativa al experimento, en verde se presenta la correspondencia del TCH, ambas imágenes poseen de fondo una representación visual en pseudocolor referencia del área de estudio.

Figura 3.24: Comparación entre el valor de TCH registrado en área de muestreo y la predicción realizada usando el modelo generado en el modelo Random Forest sobre las parcelas completas del experimento.

3.3. Resultados Complementarios

El mejor modelo encontrado para cada método planteado posee una configuración de parámetros que hace más adecuado el proceso de ajuste de los modelos, a continuación en la tabla (??) se encuentra un resumen de los parámetros estimados:

IAF	**BaggedCART**	replicaciones de bootstrap = 25
	PLS	componentes = 19
	Cubist	committees = 100 : vecinos cercanos = 0
	Random Forest	mtry = 13
TCH	**BaggedCART**	replicaciones de bootstrap = 25
	PLS	componentes = 20
	Cubist	committees = 100 : vecinos cercanos = 9
	Random Forest	mtry = 20

Tabla 3.3: Parámetros que optimizan cada modelo para las distintas variables respuesta estimadas IAF y TCH

Finalmente se presenta un cuadro resumen que incluye el listado Top 5 de variables más importantes en cada modelo evaluado. Estos datos sirven de referencia para futuras pruebas donde se desee re-entrenar los modelos en búsqueda de mejorar el conocimiento para la prediccion de las variables de interes.

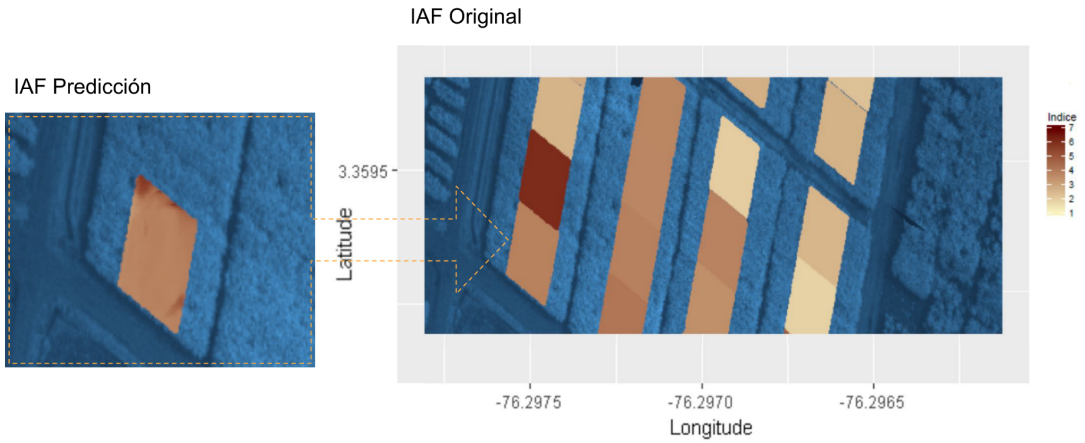
		1ro	**2do**	**3er**	**4to**	**5to**
IAF	**BaggedCART**	P	B	MO	Ca_Mg	Ca
	PLS	B	Na	Cu	MO	LnBr
	Cubist	P	Arcilla	Mg	Limo	B
	Random Forest	Arcilla	P	Limo	B	Ca/Mg
TCH	**BaggedCART**	B	Ca	P	Fe	Mg
	PLS	B	K	pH	Na	Ca/Mg
	Cubist	P	Arena	B	Mg	Mn
	Random Forest	B	P	MO	Ca	Limo

Tabla 3.4: Top 5 de variables importantes en cada modelo evaluado para la predicción de IAF y TCH

3.3.1. Validación para la Predicción (dejando una parcela por fuera del conjunto de entrenamiento a la vez)

Uno de los conceptos importantes a tener en cuenta en el uso de aprendizaje automático para modelar datos, tiene que ver con el conocimiento que adquieren los algoritmos, dicho conocimiento está basado en los datos que estos algoritmos han observado o en las relaciones que el modelo fue capaz de encontrar en los datos, lo que le da la capacidad de interpolar sobre información a la que no ha tenido la posibilidad de observar previamente.

Por esta razón se realizó la evaluación de los modelos encontrados como mejores anteriormente, se usó el método de Random Forest para la predicción de parcelas enteras, que fueron dejadas fuera de entrenamiento. De este ejercicio se obtuvieron resultados similares a los presentados en la gráfica (**??**), donde podemos ver que la predicción es muy cercana a la realidad:



A la izquierda se presenta una pequeña sección, asociada a una parcela experimental en su área de muestreo, donde el modelo realiza la predicción de IAF, a la derecha se presenta una sección más amplia con los valores de respuesta del IAF reales para una sección más amplia de las parcelas experimentales, en azul se presenta el fondo, esta imagen sirve únicamente de referencia del área de estudio, para poder apreciar mejor el espacio en el cual se da la referencia.

Figura 3.25: Comportamiento de predicción haciendo validación cruzada, dejando una parcela completa fuera de los datos (Caso de IAF)

En general el comportamiento en las pruebas realizadas fue bueno, a excepción de las parcelas que poseen los valores mayores y los valores menores entre todas las parcelas, en ese caso la predicción fue menos precisa, debido a que los modelos como random forest estiman el resultado mediante la información aprendida, pero en caso de no tener previamente información de valores mayores y menores al rango conocido, el modelo no puede estimar estos nuevos valores y entregará predicciones lo más parecidas en el rango observado según las características ingresadas en el entrenamiento con grados altos de similaridad en los predictores, más no es capaz de extrapolar los resultados para modelos como lo es el random forest que trabaja sobre los promedios de los bosques de árboles entrenados.

Otra particularidad observada es que en algunas áreas dentro de la parcela de validación es posible pensar que el valor estimado tiene un grado de abstracción mayor al presente en los datos originales, debido a que el modelo ha aprendido particularidades que pueden repetirse en ciertas zonas internas a una parcela de cultivo donde estos datos originalmente se perdieron, por las características implícitas en los datos al realizar la agregación de los datos a una escala de parcela. Esta es una característica a resaltar de estos modelos que puede servir en un futuro para detectar otras características que puedan tener efecto sobre áreas particulares de un cultivo.

3.4. Resumen de Resultados

Se hizo uso de los datos disponibles una vez puestos estos en un sistema de información geográfica que permitiera tener los datos en escalas comparables, se realizó la predicción de IAF (índice de

área foliar) y del TCH (toneladas de caña de azúcar por hectárea), con el objetivo de evaluar los modelos planteados obteniendo de manera general los siguientes resultados:

El modelo de BaggedCART o árboles de regresión con re muestreo: Fue el único modelo que no permitía un proceso de ajuste de parámetros de manera manual para búsqueda del mejor ajuste (Ver **??**), pero sí entregó las métricas asociadas RMSE y R^2 (Ver **??** y **??**) , lo que permite hacer comparación con los otros modelos evaluados.

Se evaluó el modelado mediante PLS (Mínimos cuadrados parciales): Se tuvo un resultado aceptable en algunos de los parámetros evaluados, éste es el modelo más rápido en aprender y validar resultados (Ver Tablas: **??** y **??**) esto debido a su presentación matemática y sencillez computacional, razón por la que brinda una facilidad en términos de tiempo, pero entregó los resultados menos precisos (Ver: **??** y **??**).

El modelo Cubist tuvo resultados muy buenos en términos del rendimiento en RMSE y R^2 (Ver: **??** y **??**) por esta razón se comparo su rendimiento directamente con el último modelo evaluado (Random Forest), el cual entregó resultados bastante precisos en las evaluaciones realizadas (Ver: **??** y **??**). En términos de desempeño estos dos últimos son los modelos que requieren mayor esfuerzo computacional para hacer la búsqueda de los mejores parámetros de ajuste, pero en términos de precisión son los mejores modelos entre los evaluados.

Por otro lado los tiempos computacionales requeridos para entrenamiento y validación de estos modelos serán tomados en cuenta para poder tomar la determinación de cuál puede ser el mejor modelo para realizar la predicción de las variables en estudio (IAF y TCH), esto debido a que en las comparaciones múltiples no fue posible concluir diferencias significativas entre los modelos que entregaron mejores resultados en las métricas RMSE y R^2 (Ver: **??** y **??**).

4. Discusión

Los nuevos flujos de información han llevado a que nuevas metodologías se implementen en torno al análisis de datos, muchas veces estas metodologías tienden a salirse de la concepción clásica de diseño experimental con la finalidad de maximizar el aprovechamiento de la información, de igual manera a lo largo de la historia de la agroindustria se han dado muchos adelantos en estas nuevas metodologías, pero estos avances son asíncronos debido a que se cubren necesidades distintas según el sector del agro y sus propios intereses.

La premisa en la actualidad es mejorar la toma de decisiones en agricultura, donde la innovación es la manera de materializar este objetivo, razón por la cual los análisis deben ajustarse a las nuevas tendencias en términos de capacidades tanto técnicas como computacionales, para hacer frente a este nuevo paradigma, con cantidades importantes de datos que pueden ser extraídos de un campo agrícola.

Apuntando hacia la agricultura de precisión se evaluaron cuatro métodos enfocados en el concepto de máquinas de aprendizaje, para la predicción del IAF y el TCH. Lo anterior se realizó con la finalidad de validar el funcionamiento y alcance de estos métodos aplicados a información proveniente de múltiples fuentes donde se incluyó información indirecta a partir de una imagen multiespectral e información directa tomada en campo mediante muestreos in situ. Para desarrollar estos modelos se requiere realizar procesos de ajuste y búsqueda de parámetros previos para optimizar los resultados de cada algoritmo evaluado. Para esta parte se hizo uso de herramientas enmarcadas en el concepto de bigdata con la finalidad de mejorar los tiempos que en una fase inicial fueron privativos para la evaluación de los modelos y el desempeño al momento de evaluar las metodologías utilizadas.

Los algoritmos evaluados en esta investigación hacen parte de las diversas metodologías disponibles para abordar el modelado de datos, tres de estos modelos se basan en el enfoque de árboles de regresión (BaggedCART, Random Forest y Cubist) y uno se plantea con un enfoque de métodos multivariados, el análisis de mínimos cuadrados parciales o PLS que busca relaciones fundamentales entre matrices para modelar la estructura de covarianzas implícita en los datos.

Existen algunas relaciones entre los modelos de aprendizaje automático y el modelado estadístico clásico que pueden llevarnos a pensar que son lo mismo, sin embargo existen diferencias en su funcionamiento, ya que en el modelado estadístico clásico, los datos nos guían a la selección de un modelo estocástico que sirve como la abstracción para hacer afirmaciones probabilísticas sobre cuestiones de interés, como hipótesis, predicciones y pronósticos. En este es necesaria la intervención de un experto humano que determine las fórmulas a aplicar y las ponga a prueba con base en su relación causa - efecto. Mientras que el aprendizaje automático no funciona de esta forma, este

trabaja en sentido inverso, y comienza con el resultado, que enseña a un ordenador que se ocupará de descubrir automáticamente los factores que lo impulsan. Por lo tanto la misma información es la que entrena los algoritmos (?). Además las relaciones entre datos que el aprendizaje automático maneja pueden ser increíblemente complejas, incluyendo cientos de posibles causas, interacciones y respuestas no lineales. Si se realiza correctamente, el resultado en un modelo de aprendizaje automático puede ser mucho más preciso que el que se obtiene con el modelado clásico. Además, tiene la capacidad de ajustarse automáticamente para mejorar con el tiempo, esta es una característica clave para la selección de estos modelos en un entorno aplicado, donde se pueden mejorar los resultados a partir de nuevos datos.

Una de las características de los modelos de aprendizaje automático está dada por la complejidad que puede tener interpretar las relaciones encontradas por los modelos sobre los datos, es difícil en la mayoría de casos poder detallar cuales son las relaciones involucradas y en cuales se basa el modelo entrenado. En el caso de PLS podemos interpretar los aportes a los componentes, mientras en el caso de los árboles de decisión visualmente es posible considerarlo, pero dependiendo de la complejidad del fenómeno modelado estos árboles pueden ser de dimensiones muy grandes, lo que hace casi imposible lograr apreciar en detalle todo el conocimiento condensado en el árbol final.

Uno de los procesos incluidos en el procesamiento y evaluación de los algoritmos es la evaluación de múltiples parámetros para determinar la mejor parametrización de los modelos. Este proceso en algunos algoritmos es configurable y permite explorar todas las combinaciones posibles. Sin embargo, esto requiere tiempo y criterios para definir cual es la mejor configuración para la predicción de nuestra variable respuesta. Este procedimiento fue fundamental en la fase de evaluación de los algoritmos, ya que si la predicción se realizaba sin esta calibración se podían extraer conclusiones equivocadas.

La eficiencia se describe como la relación entre el resultado alcanzado y los recursos utilizados. En este caso como todos los métodos propuestos se evaluaron con la misma cantidad de información y se alcanzaron los resultados esperados (entrenamiento y generación de modelos), nuestro recurso a evaluar es el tiempo que cada método se tomó en generar el mejor modelo.

Para el caso de la predicción del IAF los tiempos de ejecución del modelo final en cada algoritmo entregaron que el método más rápido fue el PLS, seguido por el método de BaggedCART. Estos dos algoritmos se toman tiempos inferiores a 2 segundos en obtener un modelo listo para ser usado. En tercer lugar se encuentra el método de random forest con más de 10 segundos y por último se encuentra el método de Cubist con más de 2,5 minutos para la generación de un modelo (ver detalle en ??). Es importante aclarar cómo se ha hecho a lo largo de este documento que estos tiempos corresponden a los algoritmos evaluados en máquinas pensadas para ejecutar algoritmos mediante procesamiento en paralelo y dotadas con recursos que permiten mejorar la respuesta en

tiempos y rendimiento para el entrenamiento y prueba de los modelos evaluados, como se presento en **??**.

Por otro lado para el caso de la predicción de TCH, los tiempos de ejecución del modelo final en cada algoritmo entregaron resultados casi idénticos que en el caso de IAF. En la tabla **??** se observa que de igual manera PLS y BaggedCART son los algoritmos más rápidos y tanto Random Forest como Cubist son los más lentos. No obstante, existen algunas variaciones que se atribuiría a la complejidad implícita para la predicción del TCH lo cual eleva los tiempos de entrenamiento del método Cubist hasta los 3 minutos.

Sin embargo, al momento de evaluar la precisión de los mejores modelos encontrados el panorama es bien distinto. Para la predicción de IAF los métodos de PLS y BaggedCART entregaron resultados menos precisos debido a que el RMSE era alto. Lo anterior significa que el error al momento de predecir es mayor y el R^2 era bajo en ambos casos. Mientras los algoritmos de Random Forest y Cubist entregaron resultados más precisos debido a que el RMSE fue mucho menor que el producido por los dos modelos previos y el R^2 alcanzó valores altos mayores al 0,98 en ajuste.

Para la predicción del TCH se observó el mismo panorama que para el IAF donde el PLS y el BaggedCART se destacaron como los modelos menos precisos, mientras que en el caso del Random Forest y Cubist entregaron resultados mucho más precisos.

Adicionalmente, se compararon conjuntamente los modelos mediante una prueba múltiple de Bonferroni (Ver **??** y **??**) de esta prueba se observó que tanto para la predicción del TCH como del IAF, no había diferencia entre el RMSE para los modelos Random Forest y Cubist, los cuales tuvieron mejor precisión. Este resultado es valioso para nuestro objetivo de comparar los modelos, ya que podemos unir todas las características incluida la eficiencia, donde podemos concluir que el mejor modelo en términos generales seria el método de Random Forest. Lo anterior debido a que no hay una diferencia clara con el método Cubist en términos de precisión (Ver **??** y **??**), pero en términos de eficiencia computacional el modelo de Random Forest es más rápido en la generación del modelo (Ver: **??** y **??**).

En la tabla **??** se incluyen los parámetros escogidos para cada modelo, los cuales varían según el objetivo planteado. Esto se da debido a que los métodos evaluados reaccionan de acuerdo a la complejidad intrínseca que tenga la variable respuesta y dependiendo de esa complejidad en cada algoritmo se necesitarán ajustes en términos de información necesaria para encontrar relaciones existentes en las variables predictoras (**??**).

Finalmente, cada método entrega la importancia que tiene cada variable en la consecución de la predicción, como se observa en la tabla (**??**) no hay concordancias en la importancia de cada variable entre la predicción de IAF y la predicción de TCH, como tampoco existe esta concordancia entre los diferentes métodos evaluados.

Hay algunas variables que se relacionan mejor para la predicción del IAF y otras que se relacionan mejor con el TCH, pero esto depende de gran manera de la manera de selección interna que realiza cada algoritmo en la busqueda de su objetivo de predicción. Se podría llegar a esperar correspondencia entre las variables que se intentan modelar, ya que finalmente estas representan procesos biológicos subyacentes, pero según los resultados queda demostrado que dichos procesos no se ven representados de la misma manera en cada uno de los modelos evaluados.

Para nuestro caso hablaremos de las variables que mejor se relacionan con cada predicción centrándonos en aquellas involucradas en los modelos Random Forest y Cubist. Para la predicción del IAF tenemos la concentración de P (Fósforo), B(Boro), Mg (Magnesio), Arcilla, Limo y la relación Ca/Mg , cada uno de estos componentes poseen funciones propias en el cultivo de la caña de azúcar, en **?** se encuentra de manera detallada la descripción de las funciones fisiológicas estudiadas de los componentes del suelo en el cultivo de caña de azúcar, para conocer en detalle ver el **??**.

El fósforo, el boro y el magnesio tienen impacto en el desarrollo de la planta, considerando las hojas como los órganos donde se puede observar de manera más directa el impacto. Precisamente es ahí donde el índice de área foliar se vería afectado por variaciones de estos nutrientes. Adicionalmente, la textura del suelo juega un papel importante en la retención y transporte de estos nutrientes, por lo cual hay una implicancia que vale la pena estudiar posteriormente en búsqueda del nivel de estas relaciones, a niveles más generales en cultivos de caña de azúcar. En **?** se encontró una relación importante entre el fósforo (P) y el índice de área foliar en cultivos de maíz, para caña de azúcar no se encontraron estudios tan específicos relacionando IAF y el fósforo, aunque en **?** se evaluaron distintos niveles de fósforo en cultivos de caña donde se pudo concluir que una cantidad excesiva del mismo tiene efectos adversos en el crecimiento del cultivo, lo que se ve expresado tanto en la altura del cultivo como en el IAF del mismo.

Para el caso del Boro y el Magnesio no se encontraron estudios directos que relacionen estos nutrientes y la respuesta de IAF, por otro lado en **?** se muestra que ambos nutrientes tienen una relación importante con el IAF, ya que el boro se encarga del transporte del azúcar a través de la planta y por lo tanto contribuye al desarrollo apical del cultivo, mientras que el magnesio tiene un papel importante en el desarrollo y funcionamiento adecuado de los ápices de la raíz de la planta. Una deficiencia de magnesio hace que la planta se debilite y su desarrollo se retarde por lo tanto hay una relación directa entre estos nutrientes y el IAF alcanzado por la planta.

Para la predicción del TCH tenemos la concentración de P (Fósforo), B(Boro), Arena, MO (Materia Orgánica), Ca (Calcio), Mn (Manganeso) y Limo. Para este caso algunas de las variables que se encontraron como de importancia para los modelos generados también resultaron ser de importancia en el caso de la predicción del IAF como por ejemplo P, B y Limo. No obstante, otras variables adicionales de importancia resaltaron en estos modelos como el Calcio, la Materia Orgánica y el

Manganeso, estos últimos nutrientes y características del suelo posibilitan el mejor crecimiento de la planta de caña de azúcar, incluidas características como mejores diámetros de tallo en el caso del calcio, o en el caso de la materia orgánica (MO) está relacionada con la probabilidad de obtener mayor respuesta a la aplicación de Nitrógeno (N) que es el componente más importante para altos rendimientos de caña de azúcar de manera general. Por otro lado el manganeso se encarga de actividades fisiológicas como la actividad enzimática y fotosíntesis, todas estas finalmente son estrategias para incrementar el rendimiento del cultivo de caña de azúcar (ver **??**).

En **?** se realizó un estudio para determinar la aptitud de tierra en México para la siembra de caña de azúcar donde se presenta la importancia que tienen las características edáficas en la determinación de que tan aptos son los suelos en términos del rendimiento potencial del cultivo, este estudio se basó en información previa presentada por **???** y muestra la relación que hay entre las variables edáficas en el potencial de estas para determinar el rendimiento del cultivo.

En el procesamiento de los datos disponibles para esta investigación se pudo observar la ganancia en R^2 y la disminución del RMSE evaluado con el método de Random Forest en términos de la proporción del conjunto de entrenamiento y prueba seleccionado. Esto en perspectiva al tiempo de entrenamiento, que se toma una computadora a medida que se incluye más información. Este aspecto es importante, debido a que metodologías como BaggedCART, RandomForest y Cubist realizan una cantidad grande de iteraciones y submuestreos para el entrenamiento y búsqueda de los modelos. Se debe considerar siempre los recursos cuando hablemos de estos modelos. En perspectiva modelos como PLS funcionan de una manera más rápida y sustenta los resultados en transformaciones y procesamiento entre matrices, lo que hace que el tiempo y capacidades computacionales necesarias sean menores.

? logró obtener un modelo con un $R^2 = 0,84$ y un $RMSE = 0,55$ mediante la calibración de imágenes del cultivo de caña de azúcar haciendo uso del índice NDVI. En este fue requerido un proceso de corrección atmosférica de la reflactancia y poder captar la variación del cultivo en varias de las etapas de su crecimiento para poder lograr un ajuste correcto. Gran parte del valor que tiene la metodología presentada por el autor se da en términos de poder hacer seguimiento al índice de vegetación para estimar a partir de comportamientos anteriores los siguientes valores de NDVI y por lo tanto su relación en términos del IAF. En perspectiva este estudio puede dar en gran parte un indicio de las razones por las cuales los índices de vegetación utilizados en nuestra investigación no tuvieron un aporte importante en ninguno de los modelos generados, debido a que solo poseíamos una imagen multiespectral a lo largo de todo el ciclo crecimiento del cultivo, aunque esta imagen se obtuvo a los 7 meses donde el cultivo se encuentra en una etapa fenológica clave para su estudio, este hecho no fue suficiente para hacer una correcta calibración en la búsqueda de relaciones funcionales entre las variables productivas y los índices espectrales. Esto va en concordancia con lo encontrado

por ?, ya que conocer el patrón de crecimiento del cultivo es descrito como una característica fundamental para lograr estimar el IAF y realizar predicciones de este.

? realizó un recorrido para la predicción de IAF en campos de pino mediante la búsqueda de correlaciones entre el cálculo de índices de vegetación derivados de datos multiespectrales hasta 168 bandas espectrales. En este estudio el R^2 más alto encontrado se situó entre 0,73 y 0,75 para unas combinaciones muy puntuales de las bandas multiespectrales. Lo anterior, da un panorama de incertidumbre respecto a los índices de vegetación usados en nuestra investigación, ya que estos fueron adaptados para las bandas espectrales que se poseían, como se observa en la imagen del ??. Para mejorar el conocimiento acerca del aporte de los índices de vegetación para la predicción del IAF, se requiere de una revisión más amplia para la busqueda de su utilidad y relación de los índices calculados con los rangos espectrales disponibles.

? realizó un extenso recorrido validando metodologías para la predicción de IAF sobre pastizales en Europa mediante el uso de datos multiespectrales logrando un $R^2 = 0,719$ y $RMSE = 0,289$, este RMSE se expresaba en terminos porcentuales del rango de IAF es decir una variación de 28,9 %, En esta investigación el modelo más preciso fue el de random forest para la predicción del IAF donde las bandas espectrales más útiles se encontraron en los rangos de NIR. Para estos rangos, la dispersión estructural está asociada con la absorción de ligninas y taninos, mientras el rango de NIR lejano y SWIR cobraban importancia cuando la dispersión estructural estaba asociada a absorción de agua. Lo anterior nos lleva a repensar de una manera más objetiva, la cantidad de factores que se deben considerar en el estudio del IAF a partir de imágenes multiespectrales. Esto debido a que se requiere de un grado alto de calibración del método, lo cual implicaría contar con una gran cantidad de recursos para la toma de información y así lograr tener resultados tan buenos como los presentados en estos estudios. Aun así, el enfoque presentado en estos estudios donde se incluye el método de random forest ofrece una ruta a considerar para investigaciones futuras.

En ? se pretendía hacer uso de sensores remotos landsat7 para la predicción de la producción mediante clasificación de los niveles de rendimiento de caña de azúcar en Tucumán Argentina. Se encontraron una serie de problemas muy comunes con estas imágenes relacionados a la disponibilidad de las imágenes en periodos críticos de la zafra por excesiva nubosidad o necesidad excesiva de correcciones en las imágenes para su uso, esto no permitió realizar un ajuste adecuado para la predicción de la producción. Posteriormente en (?) se usaron imágenes Aster donde se observó el comportamiento promedio de tres niveles distintos de producción en las bandas 2 y 3 que van de 630 a 860 nanómetros. Sin embargo no se ajustaron modelos a los datos ya que los resultados fueron meramente descriptivos. En perspectiva podemos ver que el uso de imágenes para predicción de la producción del cultivo de caña se ha limitado a hacer uso de información histórica, series temporales o imágenes en campo para predicción de rangos de producción más no en términos de

regresión por culpa de la incertidumbre relacionada a la costosa calibración y ajuste de los sensores.

En los modelos de arboles de regresión, para el caso de BaggedCART y Random Forest estos poseen un alto grado de similaridad (?), ya que el algoritmo de random forest se considera una mejora al algoritmo de BaggedCART. En ? se abordaron los conceptos que relacionan estos algoritmos de donde se expone que estos modelos BaggedCART aunque tienen ventajas de interpretabilidad importantes, son inferiores en capacidad de modelar datos más complejos por lo que random forest genera una mejora importante en el rendimiento de la predicción.

En ? se exploraron varios métodos de aprendizaje automático para la predicción de cobertura del suelo, entre estos métodos se encontraban random forest y Cubist, estos tuvieron resultados buenos en la predicción, los cuales podían ser mejores con implementaciones y parametrizacion optimizadas de los algoritmos. Además se observó que en términos de tiempo de ejecución, cubist fue muy superior a los otros métodos. Este resultado es importante ya que en comparación con nuestra investigación, este resultado es distinto, lo que lleva a pensar que dichos tiempos de convergencia o ajuste de los modelos dependen de la complejidad en la estructura de los datos estudiados.

En ? hicieron uso de pruebas de laboratorio de las propiedades físicas y químicas de suelos, para comparar las condiciones edafoclimáticas, las variables biofísicas y su relación con la ganancia de biomasa. Los resultados muestran que de la información geográfica (SIG) y los datos edáficos y climáticos registrados en campo permiten anticipar las respuestas fisiológicas de la plantación y por lo tanto el rendimiento de la misma, este resultado es muy importante ya que es acorde con los resultados presentados en nuestra investigación. AquaCrop es el modelo de crecimiento de cultivos desarrollado por la FAO, este simula la respuesta de rendimiento de cultivos herbáceos. En caña de azúcar se ha avanzado en el desarrollo y prueba de este modelo (?) encontrando la importancia en variables de tipo climáticas, textura del suelo y el potencial de fertilidad logrando un R^2 de 0,91 a 0,92 en algunas estaciones donde se evaluó. Esta información valida la importancia de variables como la textura del suelo y propiedades del mismo para la predicción del rendimiento del cultivo.

En ? se buscaba predecir el (Carbono orgánico del suelo), el cual desempeña un papel fundamental en la función física, química y biológica de los suelos, mediante muestras de suelo donde se analizó los espectros de reflectancia de estos suelos, se hizo uso de los modelos de Cubist y PLS para estimar el carbono orgánico del suelo, donde con los mismos datos y variaciones en sus parámetros se obtuvo los mejores resultados con el algoritmo de Cubist, un resultado importante que se alinea con los resultados obtenidos en nuestra investigación es la capacidad superior de predicción por parte del algoritmo Cubist con respecto al rendimiento del PLS, adicional a esto, este estudio presenta en su metodología una gran similaridad en la ejecución del procesos de evaluación y prueba de los algoritmos evaluados, características que validan el proceso realizado en la presente investigación.

Los datos disponibles del área de estudio poseían unas características particulares que podemos

considerarlos como limitantes del estudio por lo que pueden tener un impacto en los resultados, este impacto principalmente estaría relacionado a la posibilidad de generalización de los modelos desarrollados para predicción de variables biofisicas y variables de producción en cultivos de caña de azúcar, básicamente por la cantidad limitada de información con la que se contó y con la cual los modelos fueron alimentados. Estas características se pueden nombrar como posibles factores a mejorar en trabajos futuros realizando réplicas del mismo con mayor cantidad de información que cubra distintas latitudes a lo largo del Valle del Cauca u otras regiones.

La primera limitante se da debido a que las parcelas incluidas en el estudio se encontraban en un área de extensión limitada, un área pequeña donde posiblemente no existe la variabilidad suficiente para poder encontrar representatividad a los distintos tipos de suelo presentes en todo el Valle del Cauca. Esto reduce la posibilidad de extrapolar resultados a áreas donde las características de suelos cambian en un porcentaje alto respecto a lo medido en los datos disponibles para esta investigación.

La segunda limitante está dada por la imagen tomada mediante la cámara multiespectral la cual es una única imagen, esto limita la posibilidad de hacer un seguimiento a lo largo del crecimiento del cultivo, lo que puede resultar en un bajo nivel de importancia de los índices de vegetación calculados a partir de la imagen. Aunque se fue cuidadoso en la selección del periodo de toma de la imagen, estas variables (índices de vegetación calculados) no tuvieron un impacto significativo en casi ninguno de los modelos evaluados. Inclusive en estudios recientes como el realizado por **?** se encontró que existe una dificultad muy importante para poder hacer uso de imágenes aéreas multiespectrales en cultivos de caña por el hecho que se requiere un alto nivel de calibración. En esta investigación al tener una única toma y adicionalmente no poseer más imágenes tomadas con esta cámara no permite realizar una correcta calibración por medio de contrastes y validación cruzada de la imagen final, lo que significa un nivel de ruido adicional presente en las imágenes que no puede ser estandarizado de forma correcta.

Otra posible limitante asociada a los datos se da como el posible error asociado al proceso de interpolación espacial mediante el método de krige, ya que se llevó información de 96 muestras tomadas a lo largo de toda el área de estudio a una resolución de mayor detalle ($35 \times 35cm$), lo que en general se controló con ajustes y validación cruzada pero que a pequeña escala el error puede ser mucho mayor.

Finalmente la última limitante posible relacionada a los resultados es el nivel de sobreajuste de los modelos evaluados a los datos de entrenamiento, este se puede materializar por las características implícitas en los datos disponibles. Sin embargo a lo largo de toda la metodología se buscaron estrategias que permitieran minimizar dicho sobreajuste, incluyendo una sección donde se realizó predicción sobre parcelas dejando parcelas completas fuera de las muestras de entrenamiento, lo

que mostró que los métodos evaluados tienen un buen rendimiento para estimar estas, con una limitante relacionada a los bordes o parcelas con IAF o TCH de valores más altos o más bajos, donde los modelos pierden dominio en la respuesta y fallan generando un error mayor.

Aunque estas limitantes tienen un impacto sobre los resultados en términos de generalización, no resta validez a los resultados obtenidos respecto a la comparación de los modelos propuestos para la predicción de variables asociadas a la producción en cultivos de caña de azúcar, esto debido a que los cuatro modelos evaluados, hicieron uso de los mismos conjuntos de datos para cada fase de entrenamiento, validación y prueba, por esta razón los resultados obtenidos en términos de comparar la capacidad de modelar los datos son independientes a las consecuencias respecto a la generalización de los mismos.

Algunas de las características del presente estudio se dan ya que la información utilizada en esta investigación fue información disponible o recolectada en medio de otros estudios paralelos llevados a cabo por el centro de investigación en caña de azúcar (Cenicaña) sobre el área de estudio. Esto conllevo a que la metodología se ajustara para adapta el procesamiento de la información. Esta metodología trata de ir en un entorno ad-hoc en búsqueda de algunos resultados particulares asociados a la evaluacion de los algoritmos propuestos para evaluación.

Establecer una metodología para el procesamiento de datos que provienen de diversas fuentes de información, tanto datos tomados en campo mediante muestreo (suelo y variables asociadas a la producción del cultivo) como imágenes multiespectrales, llevándolos a sistemas de información geográfica en escalas comparables fue un factor clave que permitió explorar de una manera objetiva los datos para la aplicación de los modelos y consecución de los resultados obtenidos. Adicionalmente, obtener información con una buena resolución permite tener más información disponible para los procesos de entrenamiento y prueba de este tipo de modelos que ayuda a la búsqueda de patrones implícitos en los datos.

Una de las desventajas más importantes encontradas en los datos disponibles fue que los datos de las variables respuesta, en el caso del IAF, se tomaron mediante un muestreo sobre cada parcela y se agregaron (promedios) como una medida única que representa cada parcela. Para el caso del TCH se sumó y se escaló a rendimiento por hectárea. Esta agregación presente en los datos base puede considerarse como un problema en los datos debido a que al momento de hacer predicción solo se van a tener 32 respuestas únicas cada una relacionada a una parcela. Además en cada parcela existe una variabilidad que no quedó representada en los datos. En la práctica estas variables se encontrarán de manera similar, debido a que medir el IAF o el TCH por ejemplo se hace de manera destructiva en colombia y sería muy ineficiente medirlo a una menor resolución. Por lo tanto, los valores agregados son los más cercano que se podrá encontrar en un contexto productivo real. Este problema y cómo minimizar el impacto se abordó a lo largo del desarrollo de la metodología

planteada en esta investigación (Ver **??**).

Finalmente, los modelos evaluados en esta investigación funcionaron adecuadamente para la predicción de IAF y TCH, unos modelos con mejor rendimiento que otros como se exploro a lo largo de la sección de **??**. Cabe aclarar que estos resultados tienen una validez limitada a áreas con características de suelos similares a las presentes en las variables incluidas en este estudio como se presenta en la tabla (**??**) y en la sección (**??**) podemos ver que estos modelos funcionan muy bien en rangos de la variable respuesta conocidos y en estas condiciones es capaz de predecir y encontrara relaciones intrínsecas en los datos disponibles con resultados precisos. Esta aclaración se realiza conociendo que las parcelas con las que se entrenaron los modelos no se encuentran dispersas por distintas áreas de la geografía del Valle del Cauca, por el contrario éstas se encuentran ubicadas de manera contigua. Este hecho limita la variabilidad espacial explorada e impide una correcta generalización de estos modelos para su aplicación directa en otras áreas más amplias. Aun así esta investigación brinda un marco de referencia importante donde se evalúan modelos de aprendizaje automático y abre la posibilidad de aplicar este tipo de modelos en la agroindustria de la caña de azúcar en mayores extensiones de tierra donde se incluya una variedad más amplia de información.

5. Conclusiones Generales

En esta investigación se desarrolló una metodología que permite realizar el procesamiento de información de manera conjunta de índices de vegetación calculados a partir de imágenes multiespectrales e información de variables edáficas tomadas en campo y modelar la relación con las variables índice de área foliar y toneladas de caña de azúcar por hectárea. Uno de los principales retos en esta investigación se dio por las características de la información disponible, debido a la variabilidad y diferencia de escalas relacionada a cada fuente de datos. Se establecio una metodología para el procesamiento de estos datos llevandolos a sistemas de informacion geografica en escalas comparables, para estos se uso escala entregada por las imágenes multiespectrales con píxeles de 35 x 35 cm. A esta resolución se llevaron todas las variables tanto los índices de vegetación calculados como los datos de variables edáficas que provenían de muestreos, los cuales fueron escalados mediante interpolaciones espaciales. En el caso de las variables respuesta (IAF y TCH) se conservaron los valores disponibles por parcela. Posteriormente, estas variables se usaron para el entrenamiento del modelo con una adición de ruido blanco mientras para la fase de validación y prueba se utilizaron los valores registrados a nivel de parcela.

Con los datos procesados se evaluaron y ajustaron cuatro algoritmos de aprendizaje automático (BaggedCART, Pls, Random Forest y Cubist) para la predicción del índice de área foliar (IAF) y rendimiento de cultivo en toneladas de caña de azúcar por hectárea (TCH). Este proceso demostró que dichos modelos encuentran relaciones intrincecas en los datos que permiten realizar la predicción de las variables IAF y TCH.

Se realizó una ponderación de los modelos evaluados tanto para la predicción de IAF y TCH, donde se usó un criterio de valor que vincula la eficiencia de los modelos en términos de tiempo requerido para su entrenamiento y la precisión de la predicción obtenida por el mejor modelo encontrado en el proceso de parametrización. El primer lugar lo obtuvo el modelo del Random Forest debido a que este modelo minimizó el RMSE y maximizó el R^2. Aunque en términos de precisión tanto este modelo como el modelo Cubist tuvieron resultados muy buenos y no se puede concluir diferencia entre estos. Sin embargo, en términos de eficiencia computacional Random Forest fue más rápido en la obtención de modelos lo que lo lleva a ser considerado el mejor modelo entre todos los evaluados. Posteriormente se tendría al modelo Cubist con excelente precisión pero con tiempos de entrenamientos mayores, seguido por el algoritmo de BaggedCART y finalmente el modelo de PLS. Estos últimos dos algoritmos son bastante rápidos en el proceso de generación del modelo, pero la precisión de los mismos es menor, razón por la cual se encuentran en las últimas posiciones del ranking entre los modelos evaluados.

Se observa que existe una diferencia en la selección de las variables más importantes para cada modelo, queda demostrado que esta importancia depende del tipo de modelo que se les aplique a los datos, si bien se pensaría que al ser el proceso biológico subyacente se esperaría diferencias menores, esto no ocurre ni para la predicción del IAF ni para el TCH, estas diferencias estarían dadas por la diferencia entre las metodologías utilizadas, ya sea trabajar con proyecciones de matrices como

ocurre con PLS o mediante modelos de árboles, los criterios de selección de las variables y su importancia varias en cada uno de estos, encontrando así distintas relaciones entre las variables para modelar la respuesta cada uno de los modelos desde su especificación y criterios.

El mejor modelo entre los evaluados logró un R^2 de $98,4\,\%$ y un RMSE de $0,126$, esto que se traduce en un error relativo de $3,23\,\%$ en la predicción del índice de área foliar (IAF). Mientras para la predicción del TCH se logró un R^2 de $98,5\,\%$ y un RMSE de $3,60$ lo que se traduce en un error relativo de $1,92\,\%$ en la predicción de toneladas de caña por hectárea (TCH).

Cada modelo evaluado entregó una ponderación de las variables más importantes para la predicción según criterios internos de cada modelo. Estos criterios están relacionados con la capacidad de estas variables para aportar información valiosa para la obtención de la predicción. Se observó que estas variables tienen una relación importante con las características del cultivo de caña de azúcar implicadas en el crecimiento de las hojas, que se ve expresado en el IAF, y la capacidad de obtener mejor producción y plantas con mejores capacidades de rendimiento que se expresa en el TCH. Esto hace consistente las relaciones implícitas encontradas en los modelos evaluados.

Uno de los mayores retos encontrados en esta investigación se dio principalmente por la poca cantidad de datos disponibles para las variables respuesta, ya sea IAF o TCH. Los modelos evaluados junto a estrategias como adición de ruido blanco en la variable de respuesta pueden minimizar la posibilidad de sobre-ajuste del modelo. Sin embargo las características del suelo presentes en todas las parcelas estudiadas poseen una variación limitada debido a que se encontraban en un área acotada. Debido a lo anterior no se podría lograr una generalización de los modelos para su aplicación en áreas más amplias. Adicionalmente, otros factores relevantes para el cultivo, como información climática o información histórica, relacionados con la producción de las parcelas no fueron incluidos en los modelos, pero darían un panorama muy amplio de estudio a desarrollar en investigaciones posteriores.

Tanto la metodología desarrollada para el procesamiento de la información, como los modelos evaluados tienen la capacidad de ser adaptados para la predicción de las variables estudiadas (IAF y TCH) en áreas de cultivos extensivos, mediante la provisión de mayor información en la fase de entrenamiento y validación de los modelos. A nivel de esta investigación los resultados son prometedores, ya que permite una comparación clara entre los modelos evaluados y plantea un panorama amplio de investigación con modelos de aprendizaje automático en agricultura, por otro lado estos resultados son en esta instancia limitados al área de estudio, razón por la cual se deben evaluar más estas metodologías en contextos donde se posea mayor cantidad de información, si se desea llevar estas metodologías a la agroindustria de la caña de azúcar en un futuro cercano.

A. Tablas de Equivalencias - Índices de Vegetación Calculados

Índice	Fórmula
Red Edge Chlorophyll Index	$CI_{red\ edge} = \frac{R_{NIR}}{R_{red\ edge}} - 1$
Double Difference Index	$DD = (R749 - R720) - (R701 - R672)$
Low Narrowband Ratio	$Labr = \frac{R820 - R701}{R820 + R701}$
Modified Simple Ratio	$MSR = \frac{\left(\frac{R750}{R705}\right) - 1}{\sqrt{\left(\frac{R750}{R705}\right) + 1}}$
Meris Terrestrial Chlorophyll Index	$MTCI = \frac{R750 - R710}{R710 + R680}$
Normalized Difference Vegetation Index	$NDVI = \frac{R750 - R705}{R750 + R705}$
Simple Ratio	$SR = \frac{R_{ref}}{R_{ind}}$
Zarco-Tejada & Miller Index	$ZTM = \frac{R750}{R710}$

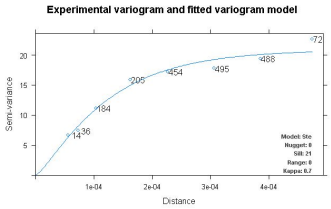
http://www.indexdatabase.de/

Figura A.1: Índices de vegetación según literatura

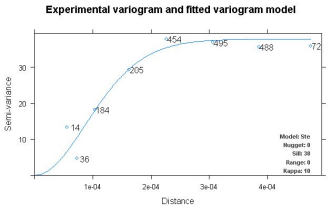
Índice	Fórmula
Red Edge Chlorophyll Index	$CI_{red\ edge} = \frac{(R770 + R783 + R795)/3}{(R720 + R733)/2} - 1$
Red Edge Chlorophyll Index 2	$CI_{red\ edge}2 = \frac{(R770 + R783 + R795 + R815)/4}{(R720 + R733)/2} - 1$
Double Difference Index	$DD = (R744 - R720) - (R707 - R671)$
Double Difference Index 2	$DD2 = (R758 - R720) - (R707 - R671)$
Low Narrowband Ratio	$Lnbr = \frac{R825 - R707}{R825 + R707}$
Modified Simple Ratio	$MSR = \frac{\left(\frac{R744}{R707}\right) - 1}{\sqrt{\left(\frac{R744}{R707}\right) + 1}}$
Modified Simple Ratio	$MSR2 = \frac{\left(\frac{R758}{R707}\right) - 1}{\sqrt{\left(\frac{R758}{R707}\right) + 1}}$
Meris Terrestrial Chlorophyll Index	$MTCI = \frac{R744 - R707}{R707 + R680}$
Meris Terrestrial Chlorophyll Index 2	$MTCI2 = \frac{R758 - R707}{R707 + R680}$
Normalized Difference Vegetation Index	$NDVI = \frac{R744 - R707}{R744 + R707}$
Normalized Difference Vegetation Index 2	$NDVI2 = \frac{R758 - R707}{R58 + R707}$
Simple Ratio	$SR = \frac{R877}{R720}$
Zarco-Tejada & Miller Index	$ZTM = \frac{R744}{R707}$
Zarco-Tejada & Miller Index 2	$ZTM2 = \frac{R758}{R707}$

Figura A.2: Índices de vegetación calculados adaptación realizada para bandas espectrales disponibles

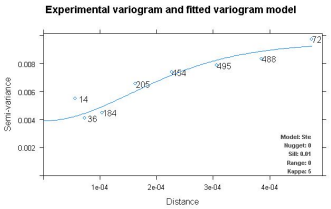
B. Modelos Krige Ajustados a Variables Edáficas



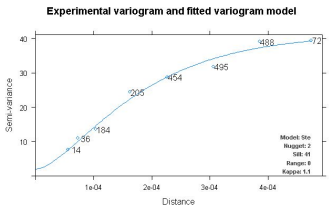
(a) Arcilla



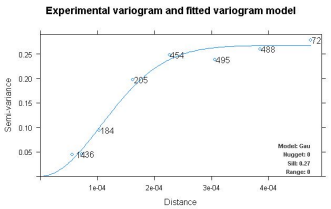
(b) Arena



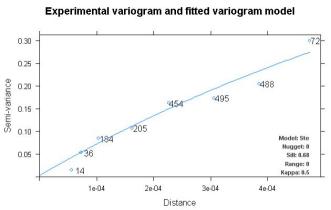
(c) Boro



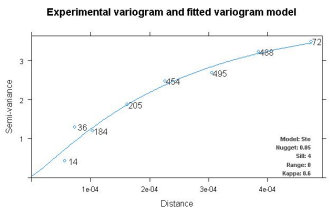
(d) Calcio



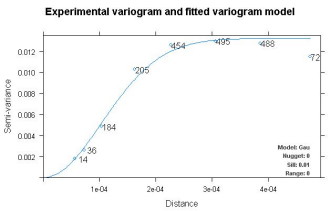
(e) Ca_Mg



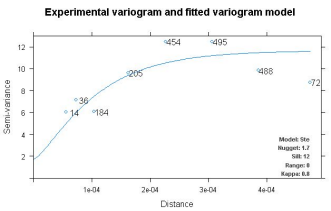
(f) Cobre



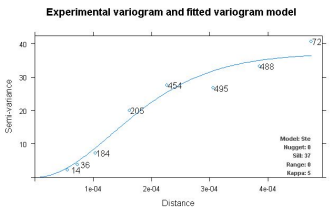
(g) Hierro



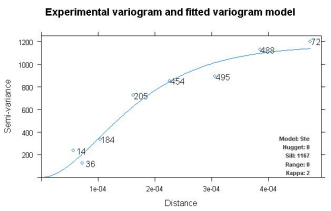
(h) Potasio



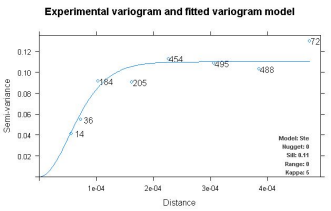
(i) Limo



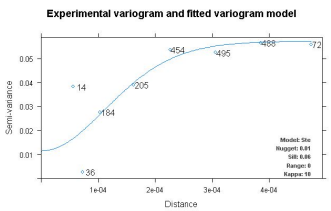
(j) Magnesio



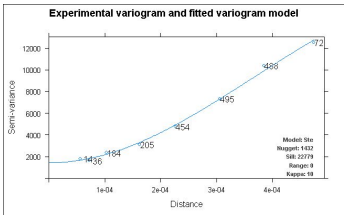
(k) Manganeso



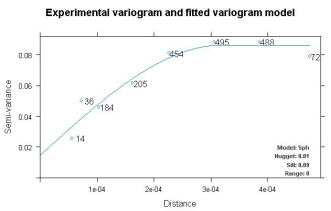
(l) Materia Organica



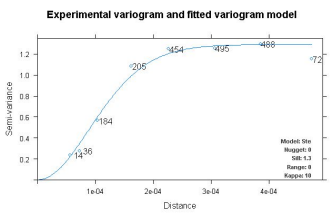
(m) Sodio



(n) Fósforo



(ñ) PH



(o) Zinc

Figura B.1: Semivariogramas generados en proceso intermedio de Interpolación de variables edáficas

C. Comparaciones Múltiples

C.1. IAF

```
Models: BaggedCART, Cubist, Pls, RandomForest
Metrics: RMSE, Rsquared
Number Of Differences: 6
P-value Adjustment: Bonferroni

Upper Diagonal: Estimates Of The Difference
Lower Diagonal: P-value For H0: Difference = 0
RMSE         BaggedCART Cubist    Pls       RandomForest
BaggedCART               0,258737 -0,367027  0,267420
Cubist       < 2,2e-16            -0,625764  0,008683
Pls          < 2,2e-16  < 2,2e-16            0,634447
RandomForest < 2,2e-16  0,006729  < 2,2e-16

Rsquared     BaggedCART Cubist    Pls       RandomForest
BaggedCART              -0,145930  0,447370 -0,149391
Cubist       < 2,2e-16             0,593300 -0,003461
Pls          < 2,2e-16  < 2,2e-16           -0,596761
RandomForest < 2,2e-16  0,002273  < 2,2e-16
```

C.2. TCH

```
Models: Baggedcart, Cubist, Pls, Randomforest
Metrics: Rmse, Rsquared
Number Of Differences: 6
P-value Adjustment: Bonferroni

Upper Diagonal: Estimates Of The Difference
Lower Diagonal: P-value For H0: Difference = 0
Rmse         Baggedcart Cubist   Pls      Randomforest
Baggedcart                7,4333 -12,4912   7,8511
Cubist       <2e-16               -19,9245   0,4178
Pls          <2e-16     <2e-16             20,3424
Randomforest <2e-16     0,1233   <2e-16

Rsquared     Baggedcart Cubist    Pls       Randomforest
```

```
Baggedcart              -0,124379  0,511093 -0,129005
Cubist       < 2e-16                0,635471 -0,004627
Pls          < 2e-16    < 2e-16              -0,640098
Randomforest < 2e-16    0,05945    < 2e-16
```

D. Características asociadas al diseño experimental de las parcelas experimentales

```
ANOVA
Df Sum Sq Mean Sq F value   Pr(>F)
Variedad        3  16727    5576  13.023 2.99e-05 ***
Riego           1    271     271   0.634    0.434
Variedad:Riego  3    453     151   0.353    0.788
Residuals      24  10275     428
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = TCH ~ Variedad + Riego + Variedad:Riego, data = base)

$Variedad                 diff        lwr         upr     p adj
CC 06-791-CC 01-1940  -63.48037 -92.019826 -34.9409241 0.0000138
CC 85-92-CC 01-1940   -39.63300 -68.172451 -11.0935491 0.0041928
CC 93-4418-CC 01-1940 -27.59512 -56.134576   0.9443259 0.0606307
CC 85-92-CC 06-791     23.84737  -4.692076  52.3868259 0.1249509
CC 93-4418-CC 06-791   35.88525   7.345799  64.4247009 0.0100234
CC 93-4418-CC 85-92    12.03788 -16.501576  40.5773259 0.6548418

$Riego                      diff       lwr      upr    p adj
Fertirriego-Convencional 5.82525 -9.273077 20.92358 0.433667
```

E. Tablas Resultados de Modelos Evaluados PLS, Cubist , Random Forest

ncomp	RMSE	Rsquared	RMSESD	RsquaredSD
1	0,9788328	0,05062780	0,02673372	0,01536021
2	0,9610735	0,08418739	0,02581673	0,01443689
3	0,8886993	0,21698373	0,02261733	0,01273004
4	0,8667137	0,25529698	0,01929348	0,01538872
5	0,8566091	0,27263634	0,01812214	0,01734500
6	0,8506877	0,28263101	0,01838083	0,01786145
7	0,8485931	0,28614602	0,01870711	0,01756886
8	0,8346347	0,30940768	0,01979161	0,01674272
9	0,8253175	0,32466369	0,01931291	0,01622394
10	0,8202781	0,33286549	0,01852725	0,01686934
11	0,8176636	0,33705777	0,01774343	0,01749650
12	0,8146666	0,34191712	0,01874164	0,01787099
13	0,8115275	0,34696443	0,01894222	0,01854979
14	0,8054568	0,35670666	0,01869125	0,01713453
15	0,7989645	0,36703467	0,01858542	0,01668065
16	0,7965320	0,37088965	0,01820061	0,01788385
17	0,7928571	0,37673640	0,01817948	0,01830253
18	0,7917006	0,37857189	0,01804305	0,01827436
19	0,7916351	0,37867436	0,01809052	0,01833126
20	0,7916466	0,37865979	0,01809750	0,01833395

Tabla E.1: Resultados numéricos de la variación del RMSE y R^2 según el número de componentes seleccionados para el algoritmo PLS en estimación de IAF

committees	neighbors	RMSE	Rsquared	RMSESD	RsquaredSD
1	0	0,2208254	0,9503914	0,04235160	0,01915121
1	1	0,2419795	0,9416241	0,04067204	0,01965788
1	3	0,2265327	0,9482849	0,04079824	0,01878943
1	5	0,2231063	0,9496856	0,04115745	0,01870737
1	7	0,2220182	0,9500749	0,04180201	0,01890239
1	9	0,2214144	0,9503103	0,04195190	0,01892585
10	0	0,1758400	0,9682484	0,03345879	0,01302892
10	1	0,2010175	0,9592671	0,03011109	0,01300124
10	3	0,1832136	0,9657141	0,03293423	0,01324690
10	5	0,1794677	0,9670380	0,03304376	0,01304822
10	7	0,1784714	0,9673630	0,03336025	0,01310722
10	9	0,1777651	0,9675997	0,03346146	0,01310639
50	0	0,1683784	0,9710131	0,02954270	0,01121192
50	1	0,1936289	0,9622518	0,02644643	0,01120521
50	3	0,1757273	0,9685460	0,02910815	0,01143931
50	5	0,1720173	0,9698182	0,02912561	0,01122900
50	7	0,1709647	0,9701491	0,02950997	0,01132425
50	9	0,1702906	0,9703678	0,02962303	0,01133388
100	0	0,1658710	0,9719744	0,02704597	0,01005578
100	1	0,1911033	0,9632935	0,02425330	0,01008569
100	3	0,1732380	0,9695162	0,02678077	0,01030882
100	5	0,1695251	0,9707744	0,02676727	0,01010378
100	7	0,1684896	0,9710974	0,02715316	0,01020467
100	9	0,1677990	0,9713219	0,02722513	0,01020329

Tabla E.2: Resultados numéricos de la variación del RMSE y R^2 según el número de commites y vecinos cercanos seleccionados para el algoritmo Cubist en estimación de IAF

mtry	RMSE	Rsquared	RMSESD	RsquaredSD
1	0,2525391	0,9484704	0,01324545	0,006904650
2	0,1860092	0,9678765	0,01537575	0,006484299
3	0,1710597	0,9718924	0,01683790	0,006482813
4	0,1659695	0,9732179	0,01682775	0,006377352
5	0,1628562	0,9740249	0,01792198	0,006653474
6	0,1612404	0,9744631	0,01778995	0,006475261
7	0,1600675	0,9747356	0,01866199	0,006717912
8	0,1588985	0,9750582	0,01881065	0,006715525
9	0,1577808	0,9753940	0,01804211	0,006400607
10	0,1574998	0,9754151	0,01929637	0,006851252
11	0,1579895	0,9751925	0,02005874	0,007206278
12	0,1581606	0,9751587	0,01902014	0,006787063
13	0,1571885	0,9754352	0,01936690	0,006849508
14	0,1580455	0,9751578	0,01918427	0,006799633
15	0,1584291	0,9750208	0,01877046	0,006678712
16	0,1590841	0,9747439	0,02013547	0,007142414
17	0,1598837	0,9744682	0,02017479	0,007165787
18	0,1609660	0,9740993	0,02020547	0,007263794
19	0,1625909	0,9735363	0,02053735	0,007394208
20	0,1651713	0,9726602	0,02090666	0,007532916

Tabla E.3: Resultados numéricos de la variación del RMSE y R^2 según el número de mtry seleccionado para el algoritmo Random Forest en estimación de IAF

ncomp	RMSE	Rsquared	RMSESD	RsquaredSD
1	27,72718	0,1147970	0,5060253	0,02345254
2	26,98737	0,1611499	0,6844556	0,03046338
3	26,63418	0,1825688	0,4845372	0,02619873
4	26,48700	0,1915800	0,4662637	0,02661508
5	26,07847	0,2162667	0,4237207	0,02556427
6	25,96327	0,2232259	0,4374749	0,02716938
7	25,83694	0,2307012	0,4463378	0,02637894
8	25,45329	0,2534248	0,5069996	0,02777987
9	25,13115	0,2721365	0,4832659	0,02836559
10	25,07530	0,2753691	0,4713089	0,02736624
11	24,76046	0,2933265	0,4370756	0,02713255
12	24,50269	0,3078429	0,4618389	0,02731288
13	24,33064	0,3175000	0,4410449	0,02662788
14	24,22905	0,3233522	0,5118291	0,02753743
15	24,03677	0,3339400	0,4795328	0,02469072
16	23,98189	0,3369662	0,4843703	0,02409237
17	23,92586	0,3400715	0,4754884	0,02438276
18	23,81995	0,3458800	0,4945913	0,02350172
19	23,81794	0,3459881	0,4954162	0,02335574
20	23,81775	0,3460013	0,4977405	0,02342956

Tabla E.4: Resultados numéricos de la variación del RMSE y R^2 según el número de componentes seleccionados para el algoritmo PLS en estimación de TCH

committees	neighbors	RMSE	Rsquared	RMSESD	RsquaredSD
1	0	5,580518	0,9626459	1,2851419	0,016585084
1	1	5,949951	0,9582583	1,1920832	0,016314408
1	3	5,636398	0,9621286	1,2438638	0,016142084
1	5	5,588384	0,9627082	1,2514252	0,016191533
1	7	5,570421	0,9628965	1,2637692	0,016307118
1	9	5,560163	0,9629952	1,2721054	0,016382841
10	0	4,205557	0,9784722	0,9980091	0,010266902
10	1	4,513809	0,9756281	0,9037623	0,010017939
10	3	4,246915	0,9781349	0,9769778	0,010136254
10	5	4,210218	0,9784696	0,9857009	0,010158907
10	7	4,202624	0,9785255	0,9914543	0,010201486
10	9	4,199386	0,9785444	0,9959566	0,010236731
50	0	3,950522	0,9809846	0,9495134	0,009460525
50	1	4,234172	0,9784745	0,8823902	0,009342049
50	3	3,976660	0,9807644	0,9426675	0,009407545
50	5	3,946580	0,9810253	0,9467949	0,009413568
50	7	3,943348	0,9810492	0,9486659	0,009439383
50	9	3,943734	0,9810406	0,9507631	0,009462931
100	0	3,897807	0,9814363	0,9590029	0,009397043
100	1	4,175685	0,9790188	0,8872053	0,009228579
100	3	3,926425	0,9811916	0,9519070	0,009358194
100	5	3,898002	0,9814373	0,9554914	0,009353895
100	7	3,893295	0,9814752	0,9573895	0,009376768
100	9	3,893218	0,9814728	0,9585992	0,009388878

Tabla E.5: Resultados numéricos de la variación del RMSE y R^2 según el número de committees y vecinos cercanos seleccionados para el algoritmo Cubist en estimación de TCH

mtry	RMSE	Rsquared	RMSESD	RsquaredSD
1	7,501008	0,9511670	0,4906692	0,008307207
2	5,118851	0,9735257	0,4700066	0,005421772
3	4,417340	0,9793488	0,4493741	0,004447287
4	4,092182	0,9818806	0,4556789	0,004246348
5	3,878360	0,9835625	0,4022551	0,003543220
6	3,778777	0,9841858	0,4481625	0,003892534
7	3,670459	0,9849914	0,4551768	0,003767470
8	3,599319	0,9854641	0,4603138	0,003794448
9	3,533734	0,9859167	0,4640743	0,003746237
10	3,519338	0,9859860	0,4502100	0,003587300
11	3,493144	0,9860759	0,5045853	0,004066053
12	3,478754	0,9861463	0,5099880	0,004103489
13	3,475380	0,9860997	0,5408318	0,004343551
14	3,507679	0,9857819	0,5635474	0,004515502
15	3,510710	0,9857157	0,5768428	0,004662636
16	3,555860	0,9852476	0,6261318	0,005126376
17	3,582950	0,9849950	0,6397601	0,005193459
18	3,606548	0,9847269	0,6726750	0,005596481
19	3,652525	0,9842998	0,6811806	0,005764729
20	3,701614	0,9838404	0,6987193	0,006013650

Tabla E.6: Resultados numéricos de la variación del RMSE y R^2 según el número de mtry seleccionado para el algoritmo random forest en estimación de TCH

F. Descripción de Variables de Importancia para los Modelos Evaluados

En **?** se encuentra de manera detallada la descripción de las funciones fisiológicas estudiadas de los componentes del suelo en el cultivo de caña de azúcar.

- El fósforo (P) al igual que el nitrógeno y el potasio se considera un nutrimento primario en el

cultivo de caña de azúcar, además es esencial para la síntesis de la clorofila y está íntimamente relacionado con la formación de la sacarosa, la deficiencia de fósforo reduce el macollamiento y el desarrollo de la planta.

- El boro (B) es un micro-nutrimento encargado del transporte de azúcar a través de las membranas celulares, es una sustancia mitótica de la planta de caña de azúcar. La deficiencia de boro en la planta se manifiesta por el escaso desarrollo apical, debido a su inmovilidad dentro de la planta. Los entrenudos se tornan cortos, las hojas detienen su desarrollo.

- La disponibilidad del magnesio (Mg) en el suelo, al igual que la de calcio, depende de la fracción intercambiable y de su balance en relación con este último nutrimento y con el potasio. La deficiencia del magnesio en la caña de azúcar se manifiestan por la aparición, en las hojas más viejas, de manchas cloróticas pequeñas con la parte central necrosada que se tornan de color rojizo-oscuro.

- La relación Ca/Mg tiene que ver con la absorción del calcio por la planta, la cual está estrechamente relacionada con el contenido en la fracción intercambiable y con la proporción en que se encuentre en el suelo en relación con el magnesio y el potasio.

- El Limo, la Arcilla y la Arena hacen parte de la descripción de la textura de el suelo, lo que indica el contenido relativo de partículas de diferente tamaño como la arena, el limo y la arcilla, la textura tiene que ver con la facilidad con que se puede trabajar el suelo, la cantidad de agua y aire que retiene y la velocidad con que el agua penetra en el suelo y lo atraviesa. Estas características son determinantes al momento de retener agua, gases como los nutrimentos que la planta toma, así el suelo se puede considerar como un reservorio donde las plantas toman el agua necesaria para los proceso de transpiración y para el transporte de nutrimentos del suelo a los tejidos de la planta.

- Calcio (Ca): El calcio es esencial para el crecimiento de los meristemos y, particularmente, para el desarrollo y funcionamiento adecuados de los ápices de las raíces, los síntomas de deficiencia de calcio en la caña de azúcar se manifiestan por la aparición, en las hojas más viejas, de manchas cloróticas pequeñas con la parte central necrosada que se tornan de color rojizo-oscuro. La intensidad de las manchas aumenta con la edad de las hojas y pueden unirse hasta formar áreas necróticas. Las hojas jóvenes deficientes en calcio se vuelven cloróticas y extremadamente débiles. La planta se debilita y su desarrollo se retarda; en consecuencia, los tallos presentan un diámetro reducido, son más delgados hacia el punto de crecimiento y su corteza es suave.

- MO (Materia Orgánica): La materia organiza es un componente del suelo que sirve para

caracterizar este según el contenido de M.O se da una categoría de suelo: Contenido de M.O. (%) Baja Menor de 2 Mediana Entre 2 y 4 Alta Mayor de 4 Estas categorías están relacionadas con la probabilidad de obtener respuesta a las aplicaciones de nitrógeno (N), así, a menor contenido de M.O. mayor será la respuesta a la aplicación de N.

- Mn (Manganeso): se encarga de actividades fisiológicas como la actividad enzimática y fotosíntesis.

- Mg (Magnesio): La disponibilidad de magnesio en el suelo, al igual que la de calcio, depende de la fracción intercambiable y de su balance en relación con este último nutrimento y con el potasio. La deficiencia del magnesio en la caña de azúcar son parecidos a los del calcio.

- K (Potasio): Las plantas absorben potasio en la forma elemental (K+). Es un elemento muy móvil dentro de la planta e importante en la formación de aminoácidos y proteínas, Las plantas que crecen en suelos deficientes en potasio presentan baja actividad fotosintética y son susceptibles a enfermedades y a estrés por sequía. Los síntomas de deficiencia de potasio en caña de azúcar se manifiestan como un marcado amarillamiento de las hojas, especialmente en el ápice y los márgenes, que termina con el necrosamiento de las áreas afectadas.

- Na (Sodio): El Sodio es un elemento relacionado con la apertura y cierre de estomas, con el balance hidrico y con la actividad del potasio; actuá también como activador enzimático de algunas reacciones.

- Fe (Hierro): la actividad fisiólogica en la planta: Actividad de las enzimas, transporte de electrones, metabolismo de ácidos nucleicos, síntesis de clorofila y fotosíntesis.

- Cu (Cobre): El Cobre es un micro-nutrimento encargado de la fotosíntesis y la resistencia a plagas y enfermedades, como responsable actividades fisiológicas de las enzimas. Las deficiencias de cobre son frecuentes en suelos que han recibido altas aplicaciones de abonos orgánicos. Estas deficiencias se manifiestan por una aparente marchitez de las hojas, debido al debilitamiento de las paredes celulares, que no debe relacionarse con el estrés por falta de agua.

- pH: Es una escala usada para medir la alcalinidad o acidez de cualquier cosa, los cultivos de caña de azúcar necesitan niveles óptimos de pH para funcionar correctamente, pH óptimo para su desarrollo es de 6,5 (ligeramente ácido), aunque tolera suelos ácidos hasta alcalinos, con un pH próximo o menor de 4,5 , la acidez del suelo limita la producción, principalmente por la presencia de aluminio intercambiable y de algunos micronutrimentos como hierro y manganeso que pueden ocasionar toxicidad y muerte de la planta.

G. Algoritmo en Lenguaje R - Procesamiento y Evaluación Realizada

```
# Procesa datos

# Cargar Librerias
library("dplyr")
library("readr")
library("randomForest")
library("rgdal")
library("maptools")
library("raster")
library("ggplot2")
library("reshape2")
library("doParallel")
library("caret")
library("gridExtra")
library("gstat")
library("sp")
library("automap")

#Configurar multiples nucleos para procesamiento
cl <- makePSOCKcluster(32)
clusterEvalQ(cl, library(foreach))
registerDoParallel(cl)

#Obtener imagen Base
x <- stack('./Datos/201504-CO_0408_lote_14_Refl.tif')
#Obtener mascara de recorte
load("./Datos/extend2.RData")
#Aplica Reporte a Imagen Original
x<-crop(x,newext)
##mapea como puntos el raster
map.p <- rasterToPoints(x)
# convierte a data frame
df <- data.frame(map.p)
# genera una funcion que realiza el calculo de los indices en acuerdo con la imagen
    hiperespectral disponible.
indices<-function(R671,R680,R707,R720,R733,R744,R758,R770,R783,R795,R815,R825,R836,R847,
    R857,R867,R877){
  CIRedEdge<-(((R770+R783+R795)/3)/((R720+R733)/2))-1
  CIRedEdge2<-(((R770+R783+R795+R815)/4)/((R720+R733)/2))-1
  DD<-(R744-R720)-(R707-R671)
  DD2<-(R758-R720)-(R707-R671)
  Lnbr<- (R825/R707)/(R825+R707)
  MSR<-((R744/R707)-1)/sqrt((R744/R707)+1)
  MSR2<-((R758/R707)-1)/sqrt((R758/R707)+1)
  MTCI<-(R744-R707)/(R707+R680)
  MTCI2<-(R758-R707)/(R707+R680)
```

```
  NDVI<-(R744-R707)/(R744+R707)
  NDVI2<-(R758-R707)/(R758+R707)
  SR<-(R877)/(R720)
  ZTM<-(R744)/(R707)
  ZTM2<-(R758)/(R707)
  #genera una lista con los calculos generados y se devuelve el resultado
  resultado<-list(CIRedEdge,CIRedEdge2,DD,DD2,Lnbr,MSR,MSR2,MTCI,MTCI2,NDVI,NDVI2,SR,ZTM,
      ZTM2)
  return(resultado)
}

##Calculo todos los indices sobre la imagen
calc<-indices(df[,3],df[,4],df[,5],df[,6],df[,7],df[,8],df[,9],df[,10],df[,11],df[,12],df
    [,13],df[,14],df[,15],df[,16],df[,17],df[,18],df[,19])

##Adjunto a la base de datos de todos los indices calculados sobre la imagen
Bas<-do.call(cbind.data.frame,calc)
#se asignan nombres a las columnas
names(Bas)<-c("CIRedEdge","CIRedEdge2","DD","DD2","Lnbr","MSR","MSR2","MTCI","MTCI2","
    NDVI","NDVI2","SR","ZTM","ZTM2")

#se genera un data frame con los indices y se le adjuntan las coordenadas espaciales
bas2<-cbind(df[,1:2],Bas)

#### se genera un grafico como linea base del area de estudio
#es un promedio simple de todas las bandas visibles para la camara para generar un pseudo
    color de vizualizacion
df<-cbind(df[,c(1,2)],rowSums(df[,3:19])/17)
colnames(df) <- c("Longitude", "Latitude", "Pseudo_Col")
#se grafica la imagen con pseudo color asignado
w<-ggplot(data=df, aes(y=Latitude, x=Longitude)) +  geom_raster(aes(fill=Pseudo_Col))
w$data$Pseudo_Col[w$data$Pseudo_Col==0]
w #esta variable se graba para ser usada a futuro
rm(indices,calc,Bas,df)

#se lee la informacion de suelos (muestreado en campo)
suelos <- rgdal::readOGR("./Datos/shp_lote14_quimicas.shp")
# se leen poligonos que posteriormente se usaran para usar unicamente informacion de las
    parcelas previamente identificadas.
polyg <- rgdal::readOGR("./Datos/polig2.shp")
#se asigna el sistemas de coordenadas para cada shapefile leido
proj4string(polyg)<-"+proj=tmerc +lat_0=4.596200416666666 +lon_0=-77.07750791666666 +k=1
    +x_0=1000000 +y_0=1000000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs"
proj4string(suelos)<-"+proj=tmerc +lat_0=4.596200416666666 +lon_0=-77.07750791666666 +k=1
    +x_0=1000000 +y_0=1000000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs"

#se pasa la infomacion de los shapefiles a dataframes
suelo<-data.frame(suelos)[,1:49]
#se seleccionan las variables que contienen las coordenadas espaciales
names(suelo)[c(15,16)]<-c("y","x")
coordinates(suelo) = ~x+y
x.range <- c(newext@xmin,newext@xmax)
```

```
y.range <- c(newext@ymin,newext@ymax)
#se genera una grilla que permite generar un area de trabajo en la cual se puede
    extrapolar datos posteriormente
grd <- expand.grid(x=seq(from=x.range[1], to=x.range[2], by=0.000001), y=seq(from=y.range
    [1], to=y.range[2], by=0.000001))
coordinates(grd) <- ~ x+y
gridded(grd) <- TRUE
suelo.grid<-grd
plot(suelo.grid, cex=1.5)
points(suelo, pch=1, col='red', cex=1)
title("Interpolation Grid and Sample Points")
#en el grafico se puede ver los puntos muetreados en campo y el area en el cual se debria
     extrapolar estos datos a futuro.

rm(x.range,y.range,newext,grd)

##############################

#se genera una funcion para intrepolar y extrapolar los datos al tamano de la grilla
    generada.
fitmod<-function(v1,suelo.grid,model.variog){
  w<<-formula(paste(v1,"~ 1"))
   krig<-krige(formula=w, locations=suelo, newdata=suelo.grid, model=model.variog)
   return(krig)
}



#se genera un vector de las variables a ajustar y aplicar modelo krige.

vars<-c("ALTITUDE","pH","MO","P","K","Fe","Zn","B","Ca_Mg","Ca","Mg","Na","Mn","Arena","
    Arcilla","Limo","Cu")
vario<-sue<-list()

# se genera un ciclo que va aplicando una busqueda del mejor modelo para cada variable de
     suelos
for(i in 1:length(vars)){
  v1<-vars[i]
   variogram <- autofitVariogram(formula(paste(v1,"~ 1")),suelo)
   vario[[i]]<-variogram
  # variogram
   plot(variogram)
   model.variog<-variogram$var_model
   ajustado<-fitmod(v1,suelo.grid,model.variog)
   sue[[i]]<- raster(ajustado["var1.pred"])
   print(i)
}

##el objeto vario guarda el variograma de cada modelo ajustado.
##el objeto sue guarda una imagen raster de el resultado de cada modelo.
save.image("Krige_Save.RData") #se guarda imagen de lo procesado hasta el momento.

##############################
```

```
salida <- brick(stack(sue[[1]],sue[[2]],sue[[3]],sue[[4]],sue[[5]],sue[[6]],sue[[7]],sue
    [[8]],sue[[9]],sue[[10]],sue[[11]],sue[[12]],sue[[13]],sue[[14]],sue[[15]],sue[[16]],
    sue[[17]]))
names(salida)<-vars

map.salida <- rasterToPoints(salida)
# converte a data frame
ajustado2 <- data.frame(map.salida)

image<-salida
# seleccion de coordenadas (puntos muestrales)
s_points<-data.frame(bas2$x,bas2$y)
names(s_points)<-c("latitude","longitude")

#extraccion de valores exactos o por interpolacion simple en el caso que sea necesario
    para cada dato.
rp <- rasterize(polyg, salida, 'parcela')
rp <- extract(rp, s_points[,1:2], method='simple')
rp2 <- rasterize(polyg, salida, 'variedad')
rp2 <- extract(rp2, s_points[,1:2], method='simple')

xdat<- extract(x, s_points[,1:2], method='simple')
data <- extract(image, s_points[,1:2], method='simple')
rp[is.na(rp)]<-0
rp2[is.na(rp2)]<-0
parcela<-as.numeric(rp)
variedad<-as.numeric(rp2)
variedad[variedad>0]<-levels(polyg$variedad)[variedad[variedad>0]]
parcela[parcela>0]<-levels(polyg$parcela)[parcela[parcela>0]]
df2 <- cbind(bas2,xdat,data,parcela=rp,variedad=rp2)
suppressWarnings(salida2<-rasterFromXYZ(df2))

proj4string(salida2)<-"+proj=longlat +datum=WGS84 +no_defs"
bf <- writeRaster(salida2, filename="./Capas_Final/All_Data.tif", options="INTERLEAVE=
    BAND",overwrite=TRUE)
bf2 <- writeRaster(salida2, filename="./Capas_Final/Capa_.tif", options="INTERLEAVE=BAND"
    ,suffix="names", bylayer=TRUE,overwrite=TRUE)

parcela<-as.factor(rp)
variedad<-as.factor(rp2)
levels(variedad)<-c("None",levels(polyg$variedad))
levels(parcela)<-c("None",levels(polyg$parcela))
df_Proc <- cbind(bas2,data,parcela,variedad)
save(df_Proc,file="salvadoall.RData")
df_Proc2<-df_Proc[df_Proc$variedad!="None",]
rm(list=ls()[!ls() %in% c("df_Proc2")])

all<-df_Proc2
####info de pixels en estudio
pixels<-c(table(all$parcela,all$variedad))[c(table(all$parcela,all$variedad))>1]
###pixes promedio por parcela
```

```
mean(pixels)
# desviacion de cantidad pixels
sd(pixels)
##areas de muestreo 13.92 M2
rm(list=ls()[!ls() %in% c("df_Proc2","all")])

# write.csv2(df_Proc2,"Salida_Final.csv",row.names = F,na="") #Guardado temporal

##############################


Resultado <- data.frame(read_csv("./Datos/base.csv")) %% dplyr::select(1:3,40:54)
Resultado$AF<-Resultado$IAF*(2*1.5) #AREA FOLEAR m2
df_Proc2$parcela<-as.numeric(as.character(df_Proc2$parcela))
df_all<-inner_join(df_Proc2,Resultado,by=c("parcela"="Parcela"))
all<-df_all[complete.cases(df_all),]
save(all,file="./Salida/salvado_Datos.RData")
rm(list = ls())
############################## Entrenamiento y Evaluacion de  Modelos


# IAF ------------------------------------------------------------------

load("./Salida/salvado_Datos.RData")

Data<-all[,c(3:16,18:33)]

#buscamos correlacion de variables
descrCor <-  cor(Data)
#se hace un corte para descartar las variables con mas de 85% de correlacion dejando una
    sola
highlyCorDescr <- findCorrelation(descrCor, cutoff = .85)
#se guarda registro de las variables a descartar
variablesHC<-names(Data[,highlyCorDescr])
#se continua trabajando con los datos despues de eliminar las altamente correlacionadas
Data<-filteredDescr <- Data[,-highlyCorDescr]
# agregamos la variable respuesta
Data$y<-all$IAF

#generamos una semilla para que el proceso sea reproducible
set.seed(325)
#Generamos un set de datos para entrenamiento y otro para test
inTraining <- createDataPartition(Data$y, p = .75, list = FALSE)
#set de datos de entrenamiento
DataTR<-Data[ inTraining,]
#set de datos de prueba
DataTE<-Data[ -inTraining,]

#generamos un data set de entrenamiento
BaseTR<-DataTR
noise<-BaseTR$y+rnorm(dim(BaseTR)[1],0,(1/10*sd(BaseTR$y)))
BaseTR$y<-noise
```

```
#generamos un data set de pruebas
BaseTE<-DataTE
BaseTE_IAF<-BaseTE

#se eliminan los objetos que no se van a usar
rm(filteredDescr,inTraining,descrCor,highlyCorDescr,noise)

set.seed(325)
#se configura el control (aplica kfold) con n=10 y el experimento se repite 3 veces para
    calcular las metricas.
control <- trainControl(method="repeatedcv", number=10, repeats=3, allowParallel = TRUE)

####PLS
set.seed(325)
plsGrid <-  expand.grid(ncomp=(1:20))
modelpls <- train(y~., data=BaseTR, method="pls", trControl=control,tuneGrid = plsGrid)
save(modelpls,file="pls_IAF.RData")

####bagCART
set.seed(325)
modelbag <- train(y~., data=BaseTR, method="treebag", trControl=control, tuneLength=5)
save(modelbag,file="treebag_IAF.RData")

###Randomforest
#Randomly Selected Predictors (MTRY)
set.seed(325)
RFGrid <-  expand.grid(mtry=(10:20))
modelparRF <- train(y~., data=BaseTR, method="parRF", trControl=control, tuneGrid =
    RFGrid)
save(modelparRF,file="parRF_IAF10.RData")

####Cubist
set.seed(325)
cubistGrid <-  expand.grid(committees= c(1, 10, 50, 100),neighbors= c(0, 1,3, 5,7,9))
modelCubist <- train(y~., data=BaseTR, method="cubist", trControl=control, tuneGrid=
    cubistGrid)
save(modelCubist,file="cubist_IAF.RData")

rm(list = ls())


# TCH ----------------------------------------------------------------------

load("~/Salida/salvado_Datos.RData")

Data<-all[,c(3:16,18:33)]

#buscamos correlacion de variables
descrCor <-  cor(Data)
#se hace un corte para descartar las variables con mas de 85% de correlacion dejando una
    sola
```

```
highlyCorDescr <- findCorrelation(descrCor, cutoff = .85)
#se guarda registro de las variables a descartar
variablesHC<-names(Data[,highlyCorDescr])
#se continua trabajando con los datos despues de eliminar las altamente correlacionadas
Data<-filteredDescr <- Data[,-highlyCorDescr]
# agregamos la variable respuesta
Data$y<-all$TCH

#generamos una semilla para que el proceso sea reproducible
set.seed(325)
#Generamos un set de datos para entrenamiento y otro para test se entrena
inTraining <- createDataPartition(Data$y, p = .75, list = FALSE)
#set de datos de entrenamiento
DataTR<-Data[ inTraining,]
#set de datos de prueba
DataTE<-Data[ -inTraining,]

#generamos un data set de entrenamiento
BaseTR<-DataTR
noise<-BaseTR$y+rnorm(dim(BaseTR)[1],0,(1/10*sd(BaseTR$y)))
BaseTR$y<-noise

#generamos un data set de pruebas
BaseTE<-DataTE
BaseTE_TCH<-BaseTE

#se eliminan los objetos que no se van a usar
rm(filteredDescr,inTraining,descrCor,highlyCorDescr,noise)

set.seed(325)
#se configura el control (aplica kfold) con n=10 y el experimento se repite 3 veces para
    calcular las metricas.
control <- trainControl(method="repeatedcv", number=10, repeats=3, allowParallel = TRUE)

####PLS
set.seed(325)
plsGrid <-  expand.grid(ncomp=(1:20))
modelpls <- train(y~., data=BaseTR, method="pls", trControl=control,tuneGrid = plsGrid)
save(modelpls,file="pls_TCH.RData")

####bagCART
set.seed(325)
modelbag <- train(y~., data=BaseTR, method="treebag", trControl=control, tuneLength=5)
save(modelbag,file="treebag_TCH.RData")

###Randomforest
#Randomly Selected Predictors (MTRY)
set.seed(325)
RFGrid <-  expand.grid(mtry=(1:20))
modelparRF <- train(y~., data=BaseTR, method="parRF", trControl=control, tuneGrid =
    RFGrid)
save(modelparRF,file="parRF_TCH.RData")
```

```
####Cubist
set.seed(325)
cubistGrid <-  expand.grid(committees= c(1, 10, 50, 100),neighbors= c(0, 1,3, 5,7,9))
modelCubist <- train(y~., data=BaseTR, method="cubist", trControl=control, tuneGrid=
    cubistGrid)
save(modelCubist,file="cubist_TCH.RData")


# Comparacion de Modelos ------------------------------------------------------

# Esta seccion del codigo - Aplica tanto para IAF como para TCH

####PLS
load("./pls_IAF.RData")
Imp <- varImp(modelpls, scale = TRUE)
plot(Imp, top = 20)
pred <- predict(modelpls, BaseTE)
postResample(pred = pred, obs = BaseTE$y)
plot(modelpls$finalModel, plottype = "scores", comps = 1:3)

####bagCART
load("./treebag_IAF.RData")
Imp <- varImp(modelbag, scale = TRUE)
plot(Imp, top = 20)
pred <- predict(modelbag, BaseTE)
postResample(pred = pred, obs = BaseTE$y)

###Randomforest
load("./IAF/parRF_IAF.RData")
pred <- predict(modelparRF, BaseTE)
postResample(pred = pred, obs = BaseTE$y)
varImpPlot(modelparRF$finalModel, top = 20,main="")

####Cubist
load("./cubist_IAF.RData")
Imp <- varImp(modelCubist, scale = TRUE)
plot(Imp, top = 20)
pred <- predict(modelCubist, BaseTE)
postResample(pred = pred, obs = BaseTE$y)


# Comparaciones Multiples -----------------------------------------------------

# se ponen todos los modelos a comparar en una lista y se usa la informacion de
    remuestreo incluida en cada modelo

resamps2 <- resamples(list(BaggedCART = modelbag,
                           Cubist = modelCubist,
                           Pls = modelpls,
                           RandomForest= modelparRF)) ## funcion resamples de libreria
                               Caret
```

```
resamps2
summary(resamps2)

trellis.par.set()
bwplot(resamps2, layout = c(2, 1),scales = list(x = list(relation = "free")),xlim = list(
    c(0, 25), c(0, 1)))

trellis.par.set(caretTheme())
dotplot(resamps2, metric = "RMSE")

difValues <- diff(resamps2)
difValues
summary(difValues)

bwplot(difValues, layout = c(2, 1),scales = list(x = list(relation = "free")),xlim = list
    (c(-25, 25), c(-1, 1)))
dotplot(difValues)

library(ggplot2)
##bagg no  posee grafico, dado que no se realizo tunning con este.
ggplot(modelparRF) + theme(legend.position = "top")
ggplot(modelpls) + theme(legend.position = "top")
ggplot(modelCubist) + theme(legend.position = "top")
```