Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado.

Autores:
Andrea Carolina Gamboa Ariza, Paula Andrea Cáceres Ortiz

Trabajo de Grado para Optar el título de Ingeniero Industrial

Director
Henry Lamos Díaz
Ph.D en Física-Matemática

Codirector
David Esteban Puentes Garzón
Ingeniero Industrial

Universidad Industrial de Santander
Facultad de Ingenierías Físico-Mecánicas
Escuela de Estudios Industriales y Empresariales
Bucaramanga
2019

AGRADECIMIENTOS

Al profesor Henry Lamos, nuestro apreciado director, por la confianza, orientación y tiempo dedicado en la realización del presente proyecto.

A nuestro querido codirector, David Puentes, por todo el apoyo, empeño y conocimientos brindados en cada etapa del proyecto. Para ellos toda nuestra gratitud; sin ellos nada de esto hubiera sido posible.

A la escuela de Estudios Industriales y Empresariales, por ser el espacio del saber que nos brindó todas las herramientas para nuestra formación personal y profesional a lo largo de estos años.

Al grupo de investigación ÓPALO, por los espacios y el conocimiento brindado en nuestros primeros pasos en el camino de la investigación.

A la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), por brindarnos la información necesaria; la cual fue vital para la realización de la presente investigación.

DEDICATORIA

A Dios, por ser el que guía mis pasos, ilumina mi vida y me llena de tantas bendiciones para ser feliz.

A mi mamá Doris y a mi abuelita Flor que gracias a su amor, compañía y cuidado el camino se hace más fácil.
A mi abuelito, que desde el cielo ha sido mi ejemplo y mi aliciente para seguir adelante.

A mi tía Nínive por su apoyo incondicional en mi formación personal y profesional.
A mis tíos Néstor y Adriana por quererme como una hija más y orientar mi camino.

A toda mi familia Ariza Ávila, mi ejemplo a seguir, porque gracias a su apoyo y cariño he salido adelante.
A Jorge por su paciencia, apoyo y cariño incondicional.

Y por último, pero no menos importante, a Paula por ser mi compañía y equipo en este maravilloso proyecto.

Andrea Carolina Gamboa Ariza

DEDICATORIA

A Dios, por darme la oportunidad de vivir y por estar conmigo en cada paso que doy, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo el periodo de estudio.

A mis padres Alfonso Cáceres Mancilla y Yolanda Ortiz Ramos por darme la vida, amarme tanto, creer en mí y porque siempre me apoyaron cuando el camino se ponía difícil. Papitos gracias por darme una carrera para mi futuro, todo esto es para ustedes.

A mis hermanos, Silvia Juliana Cáceres, Camila Cáceres y Carlos Cáceres, por estar conmigo y apoyarme siempre, los quiero mucho.

A mi sobrina, Sophie Alejandra Ibáñez, para que vea en mí un ejemplo a seguir.

A Julián por ser mi compañero y cómplice incondicional

.

Por último y de suma importancia, a mi compañera en este viaje, Andrea Carolina Gamboa, infinitas gracias por confiar en mí, sin ella nada de esto hubiese sido posible.

Paula Andrea Cáceres Ortiz

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

(Apéndices en CD adjunto)

Resumen

Título del proyecto: Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado[*].

Autores: Andrea Carolina Gamboa Ariza
Paula Andrea Cáceres Ortiz[**]

Palabras clave: Aprendizaje Automático, predicción, cultivos, Santander.

Descripción: Las herramientas de Aprendizaje Automático representan una buena alternativa para el sector agrícola, dado que permiten apoyar a los agricultores, gobierno y demás actores del sector en la toma de decisiones a partir de pronósticos en los rendimientos de los cultivos, los cuales se definen como el volumen de producto cosechado por unidad de área (Kg/Ha), proporcionando así información para mejorar su productividad. El presente trabajo tiene como objeto de estudio un cultivo experimental de cacao en Santander, ubicado en el centro de investigación La Suiza, y su propósito es predecir el rendimiento del cultivo a través de un conjunto de variables fotosintéticas, morfológicas, climáticas, químicas y físicas del suelo. Haciendo uso del Modelo Lineal Generalizado (GLM) y las Máquinas de Soporte Vectorial (SVM), se identificaron las variables explicativas de mayor impacto tanto de forma negativa como de forma positiva sobre la variable rendimiento de cultivo de cacao, con las cuales el decisor puede determinar la mejor combinación para obtener un mayor rendimiento dentro de los recorridos de las variables. La construcción y comparación de los resultados de los dos modelos, fue útil para ratificar que las variables explicativas: Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas son las variables que explican en mayor medida el rendimiento del cultivo de cacao.

[*] Trabajo de grado.

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz, Ph.D en Física-Matemática. Codirector: David Esteban Puentes Garzón, Ingeniero Industrial.

Abstract

Project tittle: Predictive model for the yield of cocoa crops in Santander using Supervised Machine Learning[*].

Authors: Andrea Carolina Gamboa Ariza

Paula Andrea Cáceres Ortiz[**]

Keywords: Machine Learning, prediction, crop, Santander

Description: Supervised Machine Learning represent a good alternative for the agriculture, in the way that it allows to support farmers, government and other stakeholders in the decision-making process based on crop yield forecast, which are defined as the volume of product harvested per unit area (Kg / Ha), thus providing information to improve their productivity. This investigation has as object of study an experimental culture of cocoa in Santander, located in the research center La Suiza, and its purpose is to predict the yield of the crop through a set of photosynthetic, morphological, climatic, chemical and physical variables. Using the Generalized Linear Model (GLM) and the Vector Support Machines (SVM), the explanatory variables with the greatest impact were identified both negatively and positively on the cocoa crop yield variable, with which the decision maker can determine the best combination to obtain a better performance within the paths of the variables. The construction and comparison of the results of the two models, was useful to ratify that the explanatory variables: Diameter of the trunk, Phosphorus (P), Magnesium (Mg),% Sand, % Hum/Grav, Radiation, Temperature, Humidity and Rains accumulated are the variables that explain to a greater extent the performance of the cocoa crop.

[*] Degree project.

[**] Faculty of Physicomechanical Engineering. Industrial and Business School. Director: Henry Lamos Díaz, Ph.D in Physics-Mathematics. Co-director: David Esteban Puentes Garzón, Industrial Engineer.

Introducción

La agricultura es una de las actividades de mayor contribución al crecimiento económico de la población en Colombia de acuerdo al aumento del 4,9% para el 2017, siendo así uno de los sectores que lideró el incremento del PIB para este año (DANE, 2018). Para 2018, el aporte de la agricultura al PIB continuó al alza según las últimas cifras presentadas "En el primer trimestre de 2018, el valor agregado de agricultura tuvo una serie positiva de 2,0%" (DANE, 2018).

El cultivo de cacao contribuye en gran medida a dicho crecimiento, debido a que fue el cultivo que más creció porcentualmente en producción, pues pasó de producir 56.785 toneladas de cacao en el 2016 a producir 60.535 en el 2017, traduciéndose en un incremento del 6,6% y creándose un récord para el país en este sector (FEDECACAO, 2018). Del mismo modo este cultivo ha venido sustituyendo gradualmente hectáreas ilícitas brindando tranquilidad e ingresos dignos a los agricultores (Semana- Comercio, 2018), por lo cual se ratificó como "el cultivo de la paz" según lo confirma el gobierno nacional ante agricultores y diferentes agremiaciones del sector cacaotero (SAC, n.d.)

Este importante crecimiento en los cultivos de cacao afianza el estudio realizado en la Universidad Nacional sobre los proyectos productivos, en donde se expone que "el 60% de los excombatientes quieren especializarse en actividades agropecuarias" (Semana-Agricultura, 2017), esto se puede llevar a cabo gracias a las más de 40 millones de hectáreas aptas para la siembra con las que cuenta Colombia delimitadas a partir del mapa básico realizado por la Unidad de Planificación Rural Agropecuaria (UPRA, n.d.).

La situación presentada anteriormente, favorece al departamento de Santander el cual tiene una alta participación en el sector agrícola "Teniendo presente que de los 87 municipios del Departamento 78 de ellos basan sus actividades en el Sector Agropecuario como principal renglón económico y cerca del 50% de la población vincula sus ingresos con actividades originadas en el Sector Rural" (Secretaría de agricultura, 2017).

En aras de contribuir con uno de los objetivos del Plan de Desarrollo Departamental, el cual busca "Fortalecer la agricultura familiar de tal forma que se garantice la Seguridad Alimentaria en el Departamento de Santander, integrando redes de intercambio de bienes y servicios rurales

locales, regionales y/o nacionales" (Plan de Desarrollo Departamental, 2016-2018) y gracias a los avances recopilados en los diferentes trabajos como el de los investigadores Chen, Wu, & Liu (2016), Zheng, Chen, Han, Zhao, & Ma (2009), Logan, McLeod, & Guikema (2016) y Pagani et al. (2017), se origina el presente trabajo con el propósito de predecir los rendimientos de los cultivos a partir de la construcción de modelos de Machine Learning (Modelos de Aprendizaje Automático) que ayuden a agricultores, entidades gubernamentales y demás actores en la toma de decisiones importantes a fin de lograr un incremento en la productividad, calidad, sostenibilidad y competitividad de este sector.

Tabla 1.

Cumplimiento de objetivos.

Objetivos específicos	**Cumplimiento**
Realizar una revisión de literatura sobre la aplicación de los modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimientos agrícolas.	Capítulo 4
Aplicar modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimiento agrícola en cultivos de cacao	Numeral 6.6
Validar los modelos con métricas de ajuste para determinar la alternativa que mejor represente los rendimientos agrícolas en cultivos de cacao	Numeral 7.2
Elaborar un artículo publicable resumiendo los hallazgos encontrados en el proyecto.	Apéndice E

1. Planteamiento del problema

Colombia, gracias a su ubicación geográfica; ubicada en la zona ecuatorial, con diversidad de pisos térmicos que van desde los 0 m.s.n.m[1] (>24ºC) hasta los 4.000 (<6ºC) (Earthtrends, 2011) siendo el primer país latinoamericano con mayores tasas de precipitación anuales, el décimo a nivel mundial y el cuarto país en América Latina con disponibilidad de tierras para la producción agrícola (FAO, n.d.), lo convierten en un país con amplias alternativas para la producción agroindustrial. Así lo ratifican las cifras del DANE para el año 2017, las cuales indican que la agricultura fue la actividad económica que más le aportó al PIB con 4,9 puntos porcentuales (DANE, n.d).

Sin embargo, para el primer trimestre de 2018 "De las siete actividades que presentaron crecimiento por encima del promedio de la economía, la de primer lugar con 6,1% fueron las actividades financieras y de seguros. La agricultura se posicionó en octavo lugar, quedando por debajo del promedio del PIB (el cual fue de 2,2% para el primer trimestre de 2018) con una tasa de crecimiento de 2,0 puntos porcentuales (DANE, 2010).

Además, debido a la situación de posconflicto por el que atraviesa el país un 60% de excombatientes de las FARC desean formarse en agricultura, según un estudio de la Universidad Nacional (Semana-Agricultura, 2017), lo cual podría ser posible gracias a las más de 40 millones de hectáreas aptas para la siembra (UPRA, n.d.) más específicamente a los cultivos de cacao, los cuales pueden jugar un papel clave en el desarrollo del posconflicto dado a que "el mapa del país donde hay más cacao y el mapa de las zonas más conflictivas es más o menos el mismo" tal y como lo afirmó el embajador de Estados Unidos en Colombia, Kevin Whitaker, en su visita al CIAT[2] en 2017 (CIAT, 2017).

El cultivo de cacao en Colombia, el cual se consideró como el cultivo de la paz, tiene un alto potencial especialmente por ser reconocido mundialmente como "Fino y de aroma en el mundo" según lo indica la International Cocoa Organization (PROCOLOMBIA, n.d.). Situación que

[1] Metros sobre el nivel del mar
[2] Centro Internacional de Agricultura Tropical.

beneficia considerablemente al departamento de Santander, debido a que es la región con más áreas en hectáreas productoras del cultivo. (Ministerio de Agricultura, 2014).

Para 2017, este cultivo alcanzó una producción de 60.535 toneladas dentro de las cuales 11.688 fueron para exportación. A pesar de las cifras alentadoras, "el sector sigue desarrollándose por debajo de su potencial" (CIAT & Universidad de Perdue, 2016). La falta de herramientas e investigaciones en el campo que ayuden a los agricultores del sector cacaotero a tomar decisiones acertadas y que en consecuencia puedan aumentar la productividad y competitividad, podrían ser algunas de las casusas que expliquen este fenómeno.

Aun cuando "Se evidencia la necesidad de aplicar un modelo de desarrollo económico sostenible soportado en la innovación y la tecnología" (Secretaría de agricultura, 2017), de la revisión de literatura y trabajos previos no se evidencian investigaciones que promuevan el mejoramiento de la agricultura que involucren nuevas tecnologías en el país. A diferencia de otros países como Indonesia, donde se encontraron investigaciones que datan de años atrás en los cuales usan "Las predicciones cuantitativas a partir de datos estadísticos previos de los efectos de factores climáticos en las cosechas para proporcionar herramientas en aras de gestionar la seguridad alimentaria" (Naylor, Falcon, Rochberg, & Wada, 2001).

La falta de uso de herramientas basadas en datos podría ser explicada a partir del análisis que realizó el DNP[3] del gobierno de Colombia en la "Política de explotación de datos: Big Data" (Mejía, 2018), el cual presenta un panorama de 4 factores: datos digitales, cultura de datos, capital humano para la explotación de datos y valor social y económico de los datos. Entre otros, uno de los grandes problemas es que tan solo el 37% de las entidades, de una muestra de 150 entidades colombianas, utilizan datos para la predicción (DNP, 2017)(Mejía, 2018) .

Otra de las situaciones encontradas es que solo el 9% de las 150 entidades tiene al menos 1 proyecto de explotación de datos en los que involucra el uso de algoritmos (DNP, 2017).

[3] Departamento Nacional de Planeación.

Uno de los indicadores que espera conseguir el DNP para el 2020 es que el 90% de las entidades públicas tengan al menos 1 proyecto de aprovechamiento de datos que involucre el uso de algoritmos.

En aras de contribuir con los escenarios presentados anteriormente, se tiene el objetivo de desarrollar, comparar y aplicar modelos que más se ajusten a los rendimientos observados para la predicción del rendimiento de cultivos de cacao, mediante herramientas del Aprendizaje Automático, con el fin de beneficiar el crecimiento y mejoramiento de dichos cultivos.

2. Justificación del proyecto

A lo largo de los años se ha venido evidenciado un crecimiento significativo en la producción de cacao en Colombia, lo cual seguirá cobrando fuerza, ya que durante el desarrollo del posconflicto el cacao se ratificó como "el cultivo de la paz" según lo confirma el gobierno nacional ante agricultores y diferentes agremiaciones del sector cacaotero (SAC n.d.) explicando que gracias a la siembra de cacao es posible sustituir los cultivos ilícitos y a su vez generar un aporte significativo al crecimiento de la economía del país.

Según el presidente ejecutivo de la federación nacional de cacaoteros (Fedecacao) Colombia pasó de producir 56.785 toneladas de cacao en el 2016 a producir 60.535 en el 2017, traduciéndose en un incremento del 6,6% y creándose un nuevo récord para el país en este sector (FEDECACAO, 2018). Este aumento en la producción llevó a que las exportaciones presentaran un incremento significativo respecto a años anteriores de acuerdo con las cifras presentadas por Fedecacao, las cuales indican que en el 2017 se vendieron al exterior 11.688 toneladas cuando en 2016 fueron 10.550 lo que significa un incremento del 11% (SAC, n.d.).

A pesar de las cifras alentadoras, un estudio del CIAT y la Universidad de Purdue, establecen que "el sector cacaotero sigue desarrollándose por debajo de su potencial" (Gutiérrez, 2017), por lo que se hace necesario desarrollar estrategias que promuevan la coordinación entre los diferentes

actores y la inclusión de tecnologías para mejorar las prácticas agrícolas en vez de enfocarse en aumentar el área cultivada.

Considerando la búsqueda de estrategias para mejorar las prácticas agrícolas en el sector cacaotero, surge la alternativa de trabajar con herramientas de aprendizaje automático o Machine Learning ML supervisado el cual consiste en hacer predicciones a partir de datos históricos de algunas características que influyen. El aprendizaje supervisado permite buscar patrones en datos históricos relacionando todos los campos hacia un objetivo en común (Gonzalez, 2014).

Por lo tanto, se pretende lograr resultados confiables comparando diferentes modelos de aprendizaje supervisado en cultivos cacaoteros para la predicción y pronóstico de rendimientos agrícolas y además determinar las variables influyentes. Los modelos elaborados permiten apoyar a los agricultores y demás actores en la toma de decisiones y así incrementar la productividad, calidad, sostenibilidad y competitividad de este sector.

3. Objetivos

3.1. Objetivo General

Plantear un modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado

3.2. Objetivos Específicos

- Realizar una revisión de literatura sobre la aplicación de los modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimientos agrícolas.
- Aplicar modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimiento agrícola en cultivos de cacao.
- Validar los modelos con métricas de ajuste para determinar la alternativa que mejor represente los rendimientos agrícolas en cultivos de cacao.
- Elaborar un artículo publicable resumiendo los hallazgos encontrados en el proyecto.

4. Revisión de la literatura

En la revisión de literatura se observan diferentes tipos de cultivos objetos de estudio tales como maíz, trigo, uva, cebada, nuez, soja, arroz, aceituna, manzana, jilo, coco a partir de los cuales se pronostican tendencias en los rendimientos haciendo uso de modelos de aprendizaje supervisado.

Los autores Motha y Heddinghaus (1986), Stephens (1988), Walker (1989), Genovese y Terres (1999), ABARE (2004), coinciden en que "Las grandes fluctuaciones anuales en los rendimientos y la producción son motivo de gran preocupación para agricultores, políticos, entidades transportadoras. Es por esto que, para abordarlas, varios países han desarrollado métodos operativos para pronosticar los rendimientos de cultivos" (Hansen, Potgieter, & Tippett, 2004).

Así como el investigador Hammer et al. (2001) el cual afirma que:

> "La información anticipada sobre la producción probable y su distribución geográfica es útil para las agencias de manejo y comercialización que gestionan la logística de almacenamiento y transporte y las ventas de exportación en el entorno de comercialización recientemente desregulado, y al gobierno en relación con las intervenciones de política".

Debido a los bajos valores de error "la confiabilidad y la capacidad de estos modelos de pronóstico justifican su uso para apoyar el proceso de toma de decisiones y mejorar la eficiencia agronómica y económica" (Cunha, Ribeiro, & Abreu, 2016).

De la primera investigación que arrojó la búsqueda, se observa que los investigadores Naylor, Falcon &Wada, (2001) hicieron uso de modelos de regresión lineal para predecir la producción y las variables que más influyen en los rendimientos de cultivo de arroz, al igual que los autores Kar & Kumar (2014).

Ambas investigaciones se realizaron con el objetivo de "evaluar la productividad de la cosecha de arroz por adelantado utilizando datos de atributos meteorológicos y fisiológicos de la planta" (Kar & Kumar, 2014) y así "proporcionar una herramienta adicional para gestionar la seguridad alimentaria" (Naylor et al., 2001).

La creciente variabilidad climática ha venido afectando los cultivos de la región de Indonesia, especialmente para el de arroz el cual es el principal alimento básico para los habitantes de la región y una de las actividades que más genera ingresos y empleo. El estudio pretende mostrar el impacto del fenómeno de El Niño en los cultivos utilizando datos estacionales recolectados de 1971-1998 sobre cultivos plantados, arroz cosechado, producción y rendimientos de arroz.

Para medir el impacto del ENSO[4] en la producción de arroz, primero examinamos dos vínculos intermedios: las asociaciones entre el ENSO y la lluvia y entre la lluvia y la producción de arroz, y luego examinamos la asociación directa entre el ENSO y la producción de arroz. Los enlaces entre variables se cuantifican utilizando análisis de correlación y regresión (Naylor et al., 2001).

El análisis demuestra que el 84% de la variabilidad en el área sembrada en septiembre-diciembre y el 81% de la variabilidad en el área sembrada en enero-abril se explica por las precipitaciones presentadas en esos meses, lo cual explica que la disminución en las precipitaciones de dicho año retrasa las plantaciones de arroz hasta que la lluvia sea adecuada para los cultivos. A su vez, las fechas de siembra tienen un impacto directo en la cosecha. Luego de la información suministrada, los agricultores y formuladores de políticas podrán seleccionar los cultivos y la fecha de siembra para estabilizar cuestiones de demanda en los productos.

"Para abordar las preocupaciones de los mercados de productos básicos y los programas de alivio de la sequía, varios países exportadores de granos, incluida Australia, han desarrollado métodos operativos para pronosticar los rendimientos de cultivos regionales y la producción agregada" (ABARE, 2004). A partir de lo anterior, otra de las investigaciones derivadas de la preocupación por las temporadas de sequía producidas por el fenómeno de El Niño, fue la de los investigadores Hansen et al., (2004) quienes al mismo tiempo utilizaron un modelo de regresión lineal para predecir los rendimientos de trigo en función de los predictores climáticos estacionales con datos de precipitaciones tomados del distrito y estableciendo una distribución de probabilidad alrededor de cada pronóstico, generando como resultado una influencia significativa de las variables climáticas en los cultivos de trigo; cultivo que también fue objeto de investigación para

[4] El Niño/Southern Oscillation (El fenómeno de El Niño).

Ceglar, Toreti, Lecerf, Van der Velde, & Dentener (2016), en el que mediante regresión de mínimos cuadrados parciales se identificaron las variables meteorológicas y el periodo en el cual estas tienen máxima influencia en los rendimientos de trigo durante su crecimiento en diferentes zonas de Francia. Para el desarrollo de este proyecto se utilizaron datos reportados en 92 regiones de Francia, en donde se presentaban cuatro tipos de clima principales (marítima, mediterránea, continental y montañosa), todo esto para lograr una comparación de las relaciones climáticas con el rendimiento y la variabilidad de los cultivos de trigo a partir de los diferentes tipos de clima.

Dentro de la búsqueda se observa un gran número de investigaciones en las que se utilizan los modelos de regresión lineal. Autores como Mkhabela et al. (2005) y Nadler & Bullock (2011) enfocaron sus esfuerzos en pronosticar los rendimientos de cultivos de maíz con ayuda de estos modelos.

Los cultivos de maíz del sur de África, más específicamente de Swaziland, fueron objeto de estudio para los investigadores Mkhabela, Mkhabela, & Mashinini (2005) donde el clima varía de templado con veranos calurosos a inviernos secos, lo que tiene como consecuencia inseguridad alimentaria y pérdida de ingresos debido a la reducción de los rendimientos, motivo por el cual se buscó “proporcionar estimaciones precisas y oportunas de la producción de maíz a los gobiernos y otras partes interesadas en la seguridad alimentaria para una intervención oportuna en caso de déficit” (Mkhabela et al., 2005) utilizando elementos como NDVI (Índice de vegetación de diferencia normalizada) derivados del NOAA Advanced Very Resolution Radiometer (AVHRR). De la relación lineal positiva entre el rendimiento de maíz y el índice de vegetación de diferencia normalizada para 3 regiones (Middleveld, Lowveld y Lubombo), se concluye que las precipitaciones afectan directamente al cultivo. A su vez se puede observar que los factores climáticos no son los únicos que tiene incidencia en los cultivos, sino que además existen otros factores que determinan su rendimiento los cuales son de vital importancia involucrar para futuras investigaciones.

El trabajo de Nadler & Bullock (2011) se concentró en estudiar los cambios en las características climáticas para las praderas canadienses. Los datos de entrada para la investigación fueron registros entre 1921 y 2000 de temperatura máxima diaria, temperatura mínima diaria y

cantidad diaria de precipitación, los cuales se adquirieron del Centro de Investigación de Cerealera y Oleaginosas del Este de Agricultura y Agroalimentación de Canadá.

Cinco de las siete regiones de estudio mostraron tendencias significativas en la temporada de crecimiento del cultivo, lo que indica que el calentamiento durante esas temporadas fue más pronunciado y determinó la variabilidad del cultivo.

Un descubrimiento importante derivado de las tendencias a largo plazo y resultados del modelo fue que, a lo largo de los años, debido a la variabilidad climática (Precipitaciones más variables y aumentos en la temperatura) hicieron que los cultivos se adaptaran a las condiciones y se generara una mayor variedad en los cultivos con mayor potencial de rendimiento.

Por otro lado, los cultivos de nogal (nuez) fueron objeto de estudio para los autores Lobell, Cahill, & Field, (2007) que pretenden, en primer lugar, proporcionar una compresión cuantitativa de las relaciones entre el rendimiento de los cultivos y la incidencia de los cambios climáticos en estos últimos y en segundo lugar evaluar el impacto neto del clima en las tendencias de rendimiento observadas durante el periodo de estudio (1980-2003) analizando tres variables climáticas: temperatura mínima, temperatura máxima y precipitación las cuales se obtuvieron del Servicio Nacional de Estadísticas Agrícolas (NASS) del Departamento de Agricultura de los Estados Unidos (USDA). Esta vez para la región de California, en la cual la agricultura representa una actividad económica importante, motivo por el cual despierta el interés de estudiar las relaciones de rendimiento en el clima las cuales pueden proporcionar una base para pronosticar la producción de cultivos dentro de un año para proyectar el impacto de los futuros cambios climáticos.

Para efectos de esta investigación, se ajusta una tendencia lineal para producir series de tiempo utilizando una regresión múltiple utilizando las 2 variables más importantes seleccionadas para cada cultivo como predictoras. Teniendo como resultado un alto valor de R2 (Coeficiente de determinación), lo cual indica una estrecha relación entre las variables y los rendimientos de los cultivos.

Por otra parte, según Tack, Barkley, & Nalley, (2015) también coinciden en que existe una fuerte relación de los rendimientos en los cultivos y el clima. Para este caso en concreto, se

cuantifica la relación entre las variables climáticas y los rendimientos de trigo los cuales evidencian un crecimiento rápido. Los avances más recientes han utilizado gran variedad de especificaciones que han presentado cambios en la forma de incluir los datos de temperatura en la ecuación de rendimiento estadístico.

Se utilizó análisis de regresión para analizar el efecto del clima en el rendimiento del trigo con la recopilación de un conjunto de datos que combina resultados de ensayos de campo de variedad de trigo de Kansas para en los periodos de 1985 al 2013.

Trabajos como los de Alvarez (2009), Mavromatis (2014) y Kouadio, Djaby, Duveiller, El Jarroudi, & Tychon (2012) coinciden en la realización de un modelo adecuado para estimar el rendimiento de trigo. Para el primero de ellos las variables analizadas fueron las características del suelo y los factores climáticos; utilizados como variables de entrada y tomados de registros meteorológicos para la provincia de Pampa, Argentina en el intervalo de tiempo de 1995-2004. Se utilizaron Redes Neuronales Artificiales y modelos lineales de regresión como metodologías para estimar el rendimiento del cultivo, el cual se correlacionó con el agua disponible en el suelo y el contenido de carbono orgánico. Comparando el modelo pronosticado con el observado y a partir de un RMSE[5] (0,05) se concluye que, en comparación con los modelos de regresión lineal, las Redes neuronales artificiales pudieron realizar una predicción más ajustada la cual explica el 64%. de la varianza en el rendimiento de los cultivos obteniendo como resultado que el factor climático con mayor efecto sobre el rendimiento fue la precipitación. Mientras que para el segundo trabajo se utilizaron modelos de regresión lineal en relación con los procesos bióticos y abióticos en la situación de producción para los cultivos de trigo ubicados en Bélgica.

Los modelos lineales generalizados (GLM) fueron utilizados dentro de su metodología por los autores Park, Hwang & Vlek, los cuales coinciden dentro de su investigación en que “Los procesos de toma de decisiones en la agricultura a menudo requieren modelos fiables de respuesta de cultivos para evaluar el impacto de la gestión específica de la tierra” (Park, Hwang, & Vlek, 2005).

En su trabajo, los investigadores Tack, Barkley & Nalley (2015) también utilizaron modelos de regresión lineal para analizar el efecto del tiempo en el rendimiento del trigo mediante datos

[5] Root-mean-square error: Error cuadrático medio

específicos de ubicación. Para este caso se encuentran con limitaciones provenientes del conjunto de datos meteorológicos obtenidos dado que el clima en cada lugar de interés es muy variante. El proyecto presentado proporcionó importantes ideas para el mejoramiento del trigo y la toma de decisiones agrícolas relacionadas con el clima cambiante, también se ofrecen oportunidades para intensificar los esfuerzos de investigación para aumentar la resistencia al estrés por calor durante el desarrollo de cada una de las etapas de crecimiento.

La calidad del grano de trigo se ve directamente afectada por varios factores agroambientales y ambientales, es por esto que (Toscano et al., 2014) tienen como objetivo determinar los principios generales que indican como en ambientes mediterráneos el Contenido de Proteína de Grano (GPC) se ve afectado por estos factores planteando un modelo de sistema con alta capacidad de predicción.

Inicialmente evaluaron la capacidad del sistema Delphi; modelo utilizado en el desarrollo del proyecto, para simular el GPC, en las principales cuencas de suministro italianas, también se analizaron las relaciones entre los errores Delphi y las variables durante las etapas de floración y llenado de grano.

Los resultados encontrados en dicho proyecto se evaluaron mediante regresión con GPC observada, mientras que los errores se calcularon realizando un análisis de correlación lineal con variables ambientales.

Los autores Oteros et al. (2014) Aguilera & Ruiz-Valenzuela (2014) y García-Mozo, Yaezel, Oteros, & Galán (2014) coinciden en el estudio de los cultivos de aceituna utilizando diferentes modelos de regresión lineal como el de tipificación y modelado de regresión de mínimos cuadrados para el pronóstico de la producción de dicho cultivo. Algunos de los resultados que se obtuvieron en estos proyectos fueron que los índices más altos de polen y la disponibilidad de agua durante la primavera están relacionadas con un aumento en la producción, además, una disminución de la producción de la aceituna se relaciona con el aumento de la temperatura del aire durante el invierno y el verano.

De la búsqueda se encontraron trabajos en los que se realizan modelos para la predicción de cultivos, además de los anteriormente mencionados, de coco, soja, jilo, caña de azúcar, maní,

manzanas abordados por los investigadores Toggweiler & Key (2001), Fishman et al. (2010), Rolim, Novo, Pantano, & Trani (2011), Pagani et al. (2017), Moreto & Rolim (2015) y Logan, McLeod, & Guikema (2016) para Sri Lanka, Estados Unidos, Brasil, España y Nueva Zelanda respectivamente. Los investigadores coinciden en utilizar modelos de regresión lineal para las predicciones de dichos cultivos encontrando que para el primer trabajo las precipitaciones son las condiciones climáticas que más influyen en el cultivo, en el segundo las concentraciones elevadas de ozono en el suelo y en el tercero las características de la planta (altura, numero de hojas) están relacionadas directamente con la producción del cultivo.

Los investigadores Cunha et al. (2016) implementaron en su investigación modelos AHP y regresión paso a paso para modelar y ajustar pronósticos en la producción de los cultivos de uva en la región de Alentejo, Portugal utilizando datos del periodo de 1998 a 2014 de variables agronómicas y climáticas de la floración para estimar el potencial de producción y las variables que afectan al cultivo.

A partir de la revisión de literatura se puede observar que en la mayoría de los artículos las variables utilizadas son meteorológicas, es decir variables que describen el clima, por lo tanto, se evidencia que existe un fuerte interés en determinar qué tipo de relaciones hay entre el clima y los rendimientos del sector agrícola. Sin embargo, se observa que con el paso del tiempo las investigaciones involucran cada vez más variables (morfología de la planta, características de suelos, manejos agrícolas, etc.) y modelos para el pronóstico de rendimientos de los diferentes cultivos, donde la mayoría son modelos de regresión lineal y en menos medida otro tipo de modelos (Maquinas de Soporte Vectorial, Redes Neuronales, etc.) los cuales permiten comparar sus resultados con el ánimo de elegir la mejor alternativa.

5. Marco teórico

5.1. Predicción

> “La predicción numérica es una estimación del valor de una variable continua y ordenada a partir de un modelo utilizando un conjunto de entrenamiento” (Berzal, n.d.).

5.1.1. Pasos para la predicción. Para la predicción es necesario otorgar datos históricos de entrada a un modelo diseñado para transformar dichos datos en variables de salida que den idea de un evento futuro. Se debe tener en cuenta que para la evaluación y construcción de un modelo son necesarios una serie de pasos (Cayuela, 2010) los cuales se detallan a continuación:

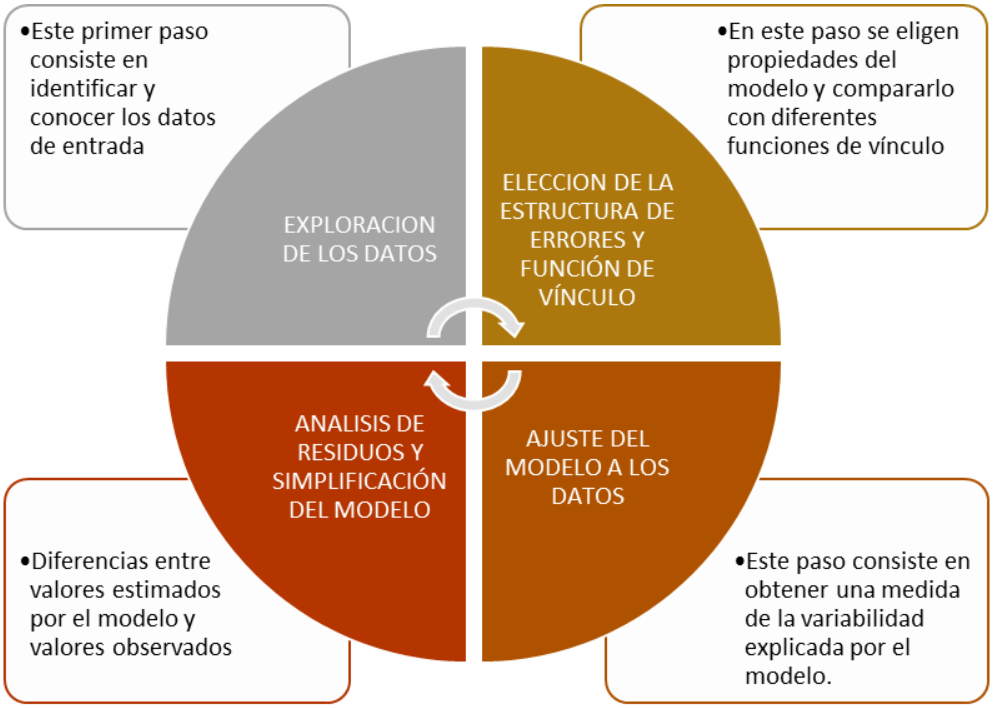

Figura. 1. Modelos de predicción. Adaptado de Modelos lineales generalizados (GLM) Cayuela (2010).

5.1.2. Predicción del rendimiento de cultivos. El rendimiento, definido como el volumen de producto cosechado por unidad de área (Kg/Ha), al verse afectado por muchos factores interrelacionados: riegos, fertilizantes, rotación de cultivos, morfología de la planta, condiciones de la tierra, factores climáticos entre otros, es de gran interés para los planificadores agrícolas estimar el rendimiento de los cultivos, de tal manera que apoyen a las diferentes entidades (agricultores, empresarios, gobierno) a tomar decisiones acertadas a partir de las tendencias y variables que más afecten el cultivo.

A lo largo de los años se han usado estimadores simples, como el promedio de rendimientos o el ultimo rendimiento obtenido. Sin embargo, estos últimos no han sido de alta precisión. Debido a esto, los investigadores Gonzalez-Sanchez (2014) afirman que:

> Por lo tanto, se han desarrollado métodos más eficientes que se pueden clasificar como modelos basados en datos (…). Las técnicas Aprendizaje automático se basan en estructuras no paramétricas y semiparamétricas, y la validación se basa en la precisión de predicción (Breiman, 2001). Trabajos previos sugieren que los modelos basados en datos tienen una mejor adaptabilidad para la planificación de cultivos debido a su implementación y rendimiento.

5.2. Correlaciones

El coeficiente de correlación es una medida que se usa para determinar el grado de asociación lineal entre dos variables continuas. Esta relación se mide a través de coeficientes que van desde +1,0 a -1,0; cuanto más cerca estén a ellos, mayor será la relación entre las variables. Valores entre |-0,1| a |-0,3| tendrán relación baja, entre |-0,3| a |-0,5| las variables tendrán relación moderada y cuando el coeficiente está entre |-0,5| y |-1| las variables tienen una fuerte relación entre ellas. (Faraway, 2004)

5.3. Multicolinealidad

La multicolinealidad en regresión es una condición que ocurre cuando algunas variables explicativas incluidas en el modelo están correlacionadas con otras variables explicativas. De acuerdo con los investigadores Del Valle Moreno & Guerra Bustillo (2012), la multicolinealidad severa es problemática, ya que puede incrementar la varianza de los coeficientes de regresión haciéndolos inestables, y al estar inestables se puede incurrir en consecuencias indeseables dentro de los modelos, algunas de ellos son:

1. Coeficientes insignificantes, aun cuando exista una relación significativa entre las variables explicativas y la variable respuesta.

2. Los coeficientes de los términos muy correlacionados pueden tener el signo equivocado.

3. Los estimadores en presencia de multicolinealidad se vuelven inestables haciendo que sus varianzas sean grandes.

5.4. Aprendizaje automático

El Aprendizaje Automático es una rama de la inteligencia artificial que proporciona métodos con capacidad de aprender o hacer predicciones sobre datos. Gracias a los avances en la tecnología informática, se tiene la capacidad de almacenar y procesar grandes cantidades de datos lo cual es posible gracias al aprendizaje automático (Baratta, 2016). El también llamado Machine Learning según Mitchell (1997):

> Se dice que un programa de computadora aprende de la experiencia con respecto a algunas clases de tareas y la medida de rendimiento, si su desempeño en las tareas mejora con la experiencia (…) Estos métodos crean un modelo a partir de entradas de ejemplo para hacer predicciones o tomar decisiones.

Los algoritmos de Aprendizaje Automático encuentran patrones naturales en los datos que generan información y ayudan a tomar mejores decisiones y predicciones. En la literatura, a menudo se llaman predictores a las variables independientes o bien a las entradas y a las salidas o variables dependientes se les llama variables respuesta. El ML no hace suposiciones sobre la estructura correcta del modelo de datos, lo que permite la construcción de modelos complejos Baratta (2016).

A continuación, se visualiza un esquema de clasificación para el aprendizaje automático.

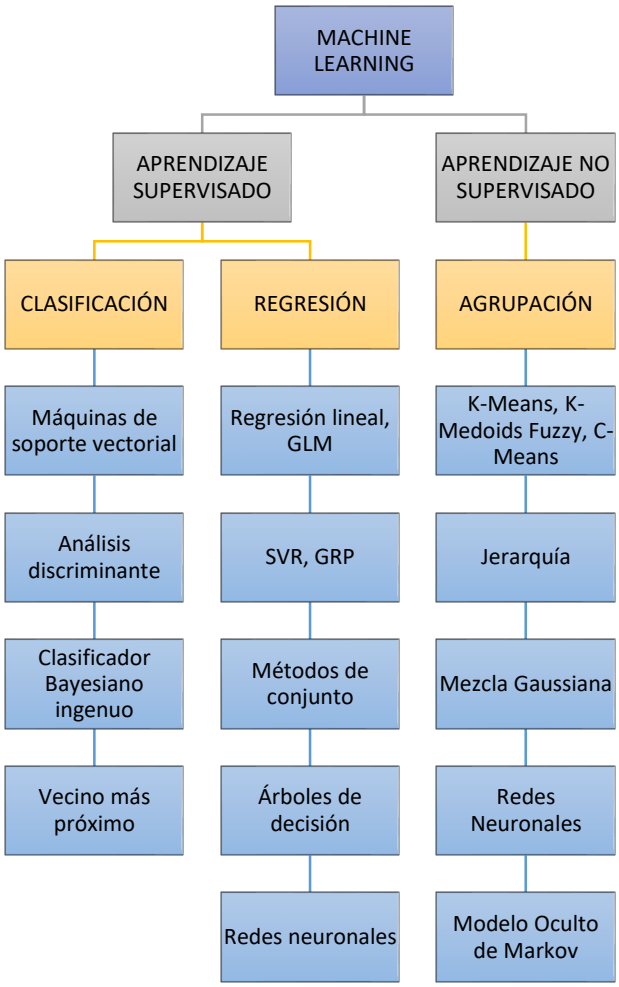

Figura. 2. Métodos de aprendizaje automático. Adaptado de Introducing Machine Learning. Math Works (2016).

5.4.1. Aprendizaje no supervisado. En la categoría de reconocimiento de patrones, existe el tipo de problema no supervisado (también llamado aprendizaje no supervisado), el problema es descubrir la estructura del conjunto de datos si los hay. Esto generalmente significa que el usuario desea saber si hay grupos en los datos, y qué características hacen que los objetos sean similares dentro de un grupo. La elección de un algoritmo es una cuestión de características del problema. Diferentes algoritmos pueden presentar diferentes estructuras para el mismo conjunto de datos.

Una característica de este tipo de aprendizaje es que no hay ninguna verdad fundamental contra la cual comparar los resultados.

La única indicación de qué tan bueno es el resultado es la estimación subjetiva del usuario (Kuncheva, 2004).

5.4.2. Aprendizaje supervisado. Se le llama al aprendizaje supervisado debido a la presencia de una variable resultado (etiqueta de salida o clase) para guiar el proceso de aprendizaje, donde se le proporciona a una máquina la habilidad de clasificar y pronosticar las etiquetas de nuevas observaciones (Kuncheva, 2004). A diferencia del no supervisado, donde solo se observan las características y no se tiene en cuenta las mediciones del resultado.

Dentro de los diferentes métodos de aprendizaje automático supervisado para desarrollar modelos predictivos, se distingue entre métodos de calificación y métodos de regresión (Introducing Machine Learning, n.d.) (Baratta, 2016). El método de clasificación es una técnica que clasifica los datos de entrada en categorías y predice respuestas discretas. Por otro lado, las técnicas de regresión predicen respuestas continuas. Es decir, valores numéricos desconocidos; por lo que a partir de su definición se encuentra que las técnicas de regresión son las más apropiadas para la presente investigación.

En el siguiente apartado se definen los métodos de regresión utilizados en el presente trabajo.

5.4.2.1. Modelo Lineal Generalizado. J. A. Nelder and R. W. M. Wedderbur (2000) propusieron por primera vez el Modelo Lineal Generalizado (GLM de las siglas en inglés de Generalized Linear Models); una extensión de los modelos lineales que considera distribuciones no normales del término de error aleatorio en la variable dependiente. Los GLM son una alternativa cuando la variable respuesta pertenece a la familia de distribuciones exponenciales (Binomiales, Poisson, gamma, etc.) con varianzas no constantes, donde la media está relacionada con las variables explicativas. Al igual que en el modelo de regresión clásico, un modelo GLM está constituido por un componente aleatorio (término error: ε), variables explicativas: x y coeficientes de regresión: β como se muestra en la ecuación (1).

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \varepsilon \quad (1)$$

El supuesto central que se ha hecho hasta el momento con los modelos lineales es que la varianza es constante (Figura 3a). En el caso de los conteos, sin embargo, donde la variable respuesta está expresada en números enteros y en dónde hay a menudo un gran número de

observaciones con valores de cero, la varianza podría incrementar linealmente con la media (Figura 3b). Con proporciones, donde hay un conteo del número de fallos de un evento, así como del número de éxitos, la varianza tendrá una forma de U invertida en relación con la media (Figura 3c). Cuando la variable respuesta sigue una distribución Gamma, entonces la varianza incrementa de una manera no lineal con la media (Figura 3d).

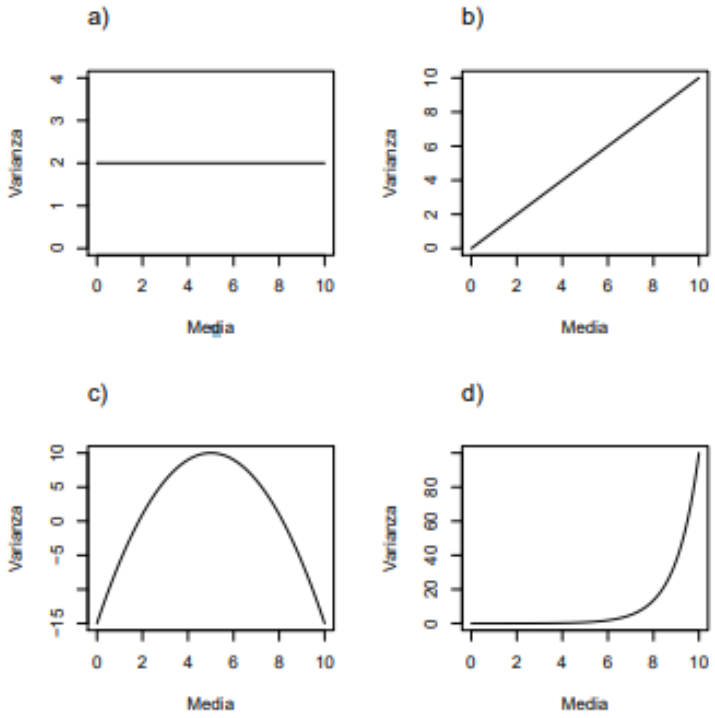

Figura. 3. Relación entre la media y la varianza de los datos bajo distintos supuestos. Adaptado de Modelos lineales generalizados (GLM) (Cayuela, 2010).

5.4.2.1.1. Función enlace: La ecuación (2) muestra la función enlace para los Modelos Lineales Generalizados, la cual es la ecuación que determina cómo se relaciona la media con las variables explicativas contenidas en x.

$$g(\mu) = x'\beta \quad (2)$$

Donde g es la función diferenciable; x, las variables independientes y β los coeficientes de regresión relacionados con las variables explicativas. Es necesario elegir el enlace canónico g correspondiente a la distribución de la variable respuesta. En la tabla 2, se muestran los enlaces canónicos más comunes para las distribuciones de la variable respuesta.

Tabla 2.
Enlaces canónicos comúnmente utilizados.

Función enlace	$g(\mu)$	Enlace canónico
Identidad	μ	Normal
Log	$\ln \mu$	Poisson
Inverso	μ^p	Gamma(p=1)
Raíz cuadrada	$\sqrt{\mu}$	Normal
Logit	$ln\frac{\mu}{1-\mu}$	Binomial

Adaptado de Generalized Linear Models for Insurance Data. Jong & Heller (2008).

De acuerdo con los autores Jong & Heller (2008), a pesar de que la función inversa es la función canónica más común para la distribución Gamma, la función logit es la función que determina de forma más adecuada cómo se relaciona la media con las variables explicativas y facilita la interpretación de los coeficientes. La ecuación (3) muestra la función enlace cuando g adapta una función canónica logit.

$$ln(\tfrac{\mu}{\mathrm{n}}) = x'\beta \quad \rightarrow \quad ln(\mu) = ln(\mathrm{n}) + x'\beta \ (3)$$

5.4.2.1.2. Pasos para el GLM. Con referencia a lo presentado en la subsección 5.1.1., se presenta en la figura 4 los pasos para la construcción de un Modelo Lineal Generalizado según los investigadores Jong & Heller (2008).

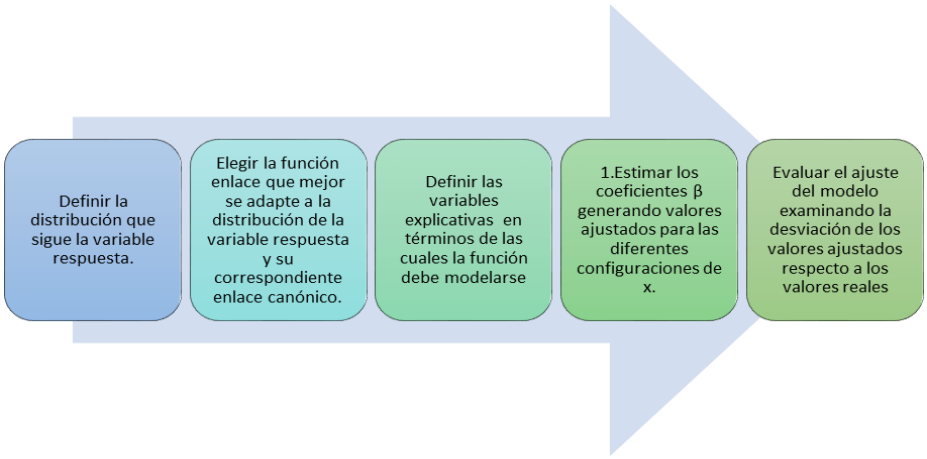

Figura. 4. Pasos en los Modelos Lineales Generalizados. Adaptado de Generalized Linear Models for Insurance Data. Jong & Heller (2008).

5.4.2.2. Máquinas de soporte vectorial. Las Máquinas de Soporte Vectorial o también llamadas SVMs por sus siglas en inglés "Support Vector Machines", fueron desarrolladas originalmente por Vapnik. Inicialmente eran utilizadas para problemas de clasificación; los cuales tienen la tarea de asignar una clase a la que pertenece un conjunto de datos. Sin embargo, hoy en día se usan en problemas de regresión, los cuales cuentan con un panorama más amplio para la investigación, debido a que estos permiten predecir valores continuos, con lo cual es el método más apropiado para la presente investigación. El objetivo de las SVMs es construir un hiperplano como superficie de decisión tal que el margen de separación entre los ejemplos positivos y negativos sea máximo (Net- & Pino, 2012), esto se conoce como hiperplano óptimo de separación. La descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de entrenamiento con muy poco o ningún conocimiento de los datos fuera de esta frontera.

La constante C, el gamma y el kernel son los parámetros con los que cuenta las SVM. Los datos son mapeados por medio de un kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento (Betancour, 2005).

En la tabla 3 se muestran las funciones integradas de kernel semidefinidas.

Tabla 3.
Funciones integradas de kernel.

Nombre del kernel	$\boldsymbol{g(\mu)}$
Lineal	$G(x_j, x_k) = x_j{}' x_k$
Gaussiano	$G(x_j, x_k) = exp(-\|\|x_j - x_k\|\|^2)$
Polinomial	$G(x_j, x_k) = (1 + x_j{}' x_k)^q$ donde q se encuentra en el conjunto {2, 3, ...}

Adaptado de Pattern Recognition and Machine Learning (2012).

La regresión de vectores de soporte (SVR) es la extensión de SVM cuando se aplican para tratar problemas de regresión (Vapnik et al., 1997; Basak et al., 2007). La función lineal f (x) se puede expresar como se muestra en la ecuación (4):

$$f(x) = \langle \omega, x \rangle + h \quad (4)$$

Donde ⟨,⟩ indica el producto escalar, ω es un vector de peso ajustable, x es la información de entrada y h es el umbral escalar. La función de pérdida (yj, fxj)), introducida en SVR, describe un modelo que indica que no existe diferencia entre los valores actuales y predictivos si el valor de diferencia entre ellos es menos, que es la principal diferencia entre las funciones de regresión lineal (Yoon et al. 2011). Para permitir la existencia de un error de ajuste, se puede minimizar la siguiente función de riesgo: $M\omega,*$ con las variables positivas de holgura (ζ, ζ*) introducidas en ella como se muestra en la ecuación (5).

$$M_{\omega,\zeta,\zeta*} = \frac{1}{2} \| \omega \|^2 + C \sum_{j=1}^{1} (\zeta_j, \zeta_j *) \quad (5)$$

$$\begin{cases} y_j - \langle \omega, x \rangle - h \leq \varepsilon + \zeta_j \\ \langle \omega, x \rangle + h - y_j \leq \varepsilon + \zeta_j * \quad j = 1, \ldots, nj = 1, \ldots, n \\ \zeta_j, \zeta_j * \geq 0 \end{cases}$$

La constante C en el modelo SVM se usa como un parámetro de penalización en términos del error, es decir controla la compensación entre los errores de entrenamiento y los márgenes rígidos, creando así un margen que permita algunos errores en la clasificación a la vez que los penaliza conciliando el riesgo empírico y el riesgo de confianza ζ, ζ* son variables de holgura que calculan el error de los lados hacia arriba y hacia abajo, respectivamente.

Los multiplicadores de Lagrange ($\alpha j - \alpha j*$) se introducen para resolver el problema dual. Se puede obtener el mejor hiperplano de regresión representado en la ecuación (6).

$$f(x) = \sum_{j=1}^{1} (\alpha_j - \alpha_j *) \langle x, x_j \rangle + h \quad (6)$$

Donde αj y $\alpha j*$ satisfacen la igualdad de $\alpha j \times \alpha j* = 0$, $\alpha j \geq 0$, $\alpha j* \geq 0$, y j=1...n y se pueden determinar maximizando la forma dual. La clave para mejorar la precisión de predicción es la función SVM kernel y la optimización de parámetros. Las funciones que satisfacen el teorema de Mercer pueden ser usadas como productos punto y por ende como kernels. A continuación, en la ecuación (7), un kernel polinomial de grado d para construir un clasificador SVM.

$$K(x_i x_j) = (1 + x_i \bullet x_j)^d \quad (7)$$

Por otra parte, el parámetro gamma define hasta qué punto llega la influencia de un solo ejemplo de entrenamiento, con valores bajos que significan "lejos" y valores altos que significan "cerca". Para mayor comprensión, un Gamma bajo, significa que los puntos están alejados de la línea de separación, mientras que un gamma con valor alto presenta cercanía de los puntos a la línea de separación (Gavidia Bovadilla Dirigida por & Oñate Ibañez de Navarra Ing Eduardo Soudah Prieto, n.d.).

5.4.3. Validación de modelos. La validación de los modelos en conjuntos de prueba es una de las acciones más importantes a la hora de construir un modelo predictivo (Logan et al., 2016). Este importante paso consiste en tomar el conjunto completo de datos disponibles y dividirlo aleatoriamente en dos subconjuntos; el primero de ellos son los datos aplicados a la fase de entrenamiento y el segundo para la fase de prueba. Generalmente, se consideran dos terceras partes para el conjunto de entrenamiento y la otra tercera parte para el conjunto de prueba. (Mendoza, 2008).

5.4.3.1. Cross-Validation. La técnica Cross-Validation, o bien llamada validación cruzada, es usada comúnmente para estimar diferentes configuraciones de parámetros (Kohavi, n.d.). A su vez, se utiliza para evaluar la precisión de los modelos derivados de la predicción en aprendizaje automático supervisado. Esta técnica tiene dos clasificaciones para la estimación de precisión de los modelos (FH, 2006) las cuales se discuten en las siguientes secciones.

- Validación cruzada K-fold

Una iteración de la validación cruzada de k, en primera instancia, genera una permutación aleatoria del conjunto de muestra y se reparte en K subconjuntos de aproximadamente el mismo tamaño. De los K subconjuntos, un solo subconjunto se retiene como los datos de validación para probar el modelo (testset) y los subconjuntos K - 1 restantes se usan juntos como datos de entrenamiento (trainset). Posteriormente, el modelo se entrena en la composición y su precisión se evalúa en el conjunto de prueba. La capacitación y evaluación del modelo se repite K veces, y cada uno de los K subconjuntos se usa exactamente una vez como conjunto de prueba como se muestra en la figura 5.

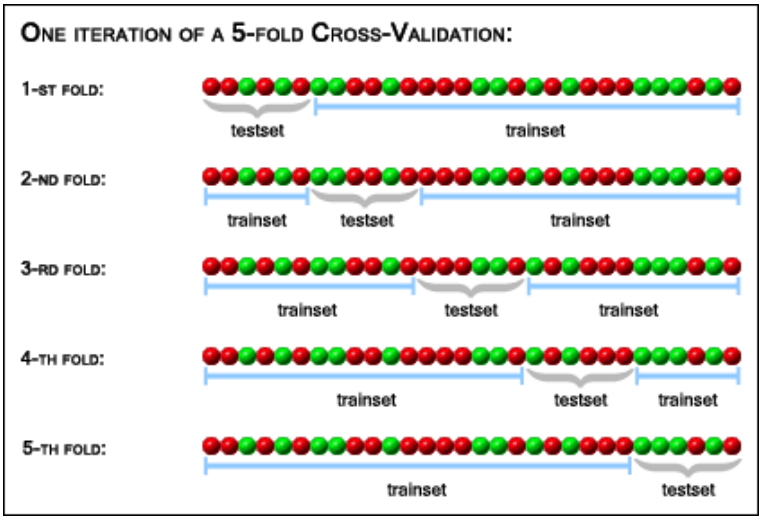

Figura. 5 . Validación cruzada K-fold de 5 veces, 30 muestras. Adaptado de ProClassify User's Guide (2006).

- Validation cruzada Leave-one-out

Como su nombre lo indica, esta técnica de la validación cruzada consiste en el uso de una única muestra del conjunto de muestras original como datos de validación, y las muestras restantes como datos de entrenamiento. Esto se repite de modo que cada muestra del conjunto de muestras se use exactamente una vez como datos de validación. Cabe resaltar que para en este caso, no hay necesidad de generar permutaciones aleatorias para la validación cruzada de dejar uno fuera y repetirla, porque los conjuntos de datos de entrenamiento y validación para cada uno de los pliegues son siempre los mismos, y por lo tanto se determina el resultado de la estimación de la precisión.

5.4.3.2. Hold Out. La técnica Hold-Out consiste en tomar el conjunto completo de datos disponibles y dividirlo aleatoriamente en dos subconjuntos; el primero de ellos son los datos aplicados a la fase de entrenamiento y el segundo para la fase de prueba. El algoritmo de aprendizaje de funciones se ajusta a una función utilizando solo el conjunto de entrenamiento.

Luego, se le pide al modelo de aprendizaje que pronostique los valores de salida para los datos en el conjunto de prueba (Mendoza, 2008).

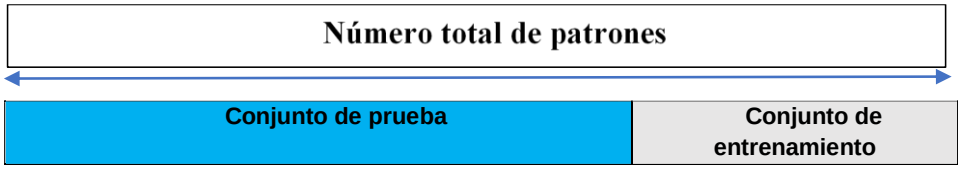

Figura. 6. División de datos. Adaptado de Técnicas ML en Medicina Cardiovascular.

De esta manera, se crea un modelo únicamente con los datos de entrenamiento. Con el modelo creado se generan datos de salida que se comparan con el conjunto de datos reservados para realizar la validación (que no han sido utilizados en el entrenamiento), por lo que no han sido utilizados para generar el modelo Los estadísticos obtenidos con los datos del subconjunto de validación son los que nos dan la validez del método empleado en términos de error. Una aplicación alternativa de este método consiste en repetir el proceso hold-out, tomando distintos conjuntos de datos de entrenamiento (aleatorios) un determinado número de veces, de manera que se calculan los estadísticos de la regresión a partir de la media de los valores en cada una de las repeticiones (Pérez-Planells, Delegido, Rivera-Caicedo, & Verrelst, 2015).

5.4.4. Métricas de ajuste. Para la misma tarea, puede haber varios algoritmos y se puede tener interés en encontrar el más eficiente" (Cayuela, 2010), motivo por el cual se hacen necesarias las métricas de ajuste aplicadas al modelo con el fin de validar y escoger el que más se ajuste a los datos.

5.4.4.1. Coeficiente de determinación (R2). Los R2 como medidas de bondad de ajuste son muy populares. Tienen una interpretación natural como la proporción de la varianza explicada de la variable respuesta por el modelo, está entre 0 y 1, es adimensional y es mayor a medida que el modelo presenta mejor ajuste (Guzmán, 2012).

Para la i-ésima observación de la muestra, la desviación entre el valor observado de la variable dependiente y_i y el valor estimado de la variable dependiente $\hat{y}_i$, se llama i-ésimo residual y representa el error que se comete al usar para estimar y_i (Cardona Madariaga, González Rodríguez, Rivera Lozano, & Cárdenas Vallejo, 2013).

La suma de los cuadrados de esos residuales es lo que se minimiza en el método de mínimos cuadrados. También se le conoce como la suma de los cuadrados debidos al error (SSE) como se muestra en la ecuación (8).

$$SSE = \left(y_{i-}\hat{y}_i\right)^2 \quad (8)$$

El valor de SSE es una medida del error que se comete al usar la ecuación de regresión para calcular los valores de la variable dependiente en la muestra.

Otro valor de importancia es la medida del error incurrido al usar para estimar y_i, llamado suma total de cuadrados (SST) mostrado en la ecuación (9):

$$SST = \sum(y_i - \bar{y})^2 \quad (9)$$

Para saber cuánto se desvían los valores de $\hat{y}_i$ medidos en la línea de regresión, de los valores de $\bar{y}$, se calcula otra suma de cuadrados, la cual se muestra en la ecuación (10). A esa suma se le llama suma de cuadrados debida a la regresión, y se representa por SSR

$$SSR = \sum(\hat{y}_i - \bar{y})^2 \quad (10)$$

Existe una relación entre las tres sumas, representada a continuación en la ecuación (11):

$$SST = SSR + SSE \quad (11)$$

Ahora bien, es posible entender cómo se pueden emplear las tres sumas de cuadrados para suministrar una medida de la bondad de ajuste para la ecuación de regresión. Esa ecuación tendría un ajuste perfecto si cada valor observado de la variable independiente estuviera sobre la línea de regresión. En este caso cada diferencia $y_i - \hat{y}_i$ sería cero, por tanto, SSE=0.

De la ecuación (11) se tendría que SST=SSR y por consiguiente la relación SSR/SST sería igual a 1 como el máximo ajuste. De manera análoga, los ajustes menos perfectos darán como resultado mayores valores de SSE. En consecuencia, de (11) se deduce que el máximo valor de SSE se tiene cuando SSR es cero (Cardona Madariaga et al., 2013)

La relación SSR/SST, se denomina coeficiente de determinación y se representa por R^2 tal y como se muestra en la ecuación (12)

$$R^2 = \frac{SSR}{SST} \quad (12)$$

5.4.4.2. Error Cuadrático Medio (RMSE). El Error Cuadrático Medio considera una diferencia entre los valores estimados y los valores reales, estas diferencias se elevan al cuadrado y se calcula el promedio de todas ellas. Como su nombre lo indica, a este promedio se le debe calcular su raíz cuadrada. RMSE mide la magnitud de error (Negrón Baez, 2014). Su ecuación está representada en (13).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (13)$$

$$RMSE = \sqrt{MSE}$$

Donde:

- y_i: Es el valor real
- $\hat{y}_i$: Es el valor estimado
- N: Es el tamaño de la muestra

5.4.4.3. Error Absoluto Medio. El resultado del Error Absoluto Medio demuestra que tan cercano es la predicción hecha al resultado real, se parte haciendo una diferencia entre el valor obtenido y el valor real en valor absoluto y luego se le calcula el promedio (Negrón Baez, 2014), cuya representación se muestra en la ecuación (14).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (14)$$

Donde:

- y_i: Es el valor real
- $\hat{y}_i$: Es el valor estimado
- N: Es el tamaño de la muestra

5.4.5. Gráficas de resultado. En el caso de los modelos, cuando se utiliza una regresión para predecir el valor $\hat{y}_i$ a partir de un valor en x de un determinado sujeto x_i, es probable que se cometan errores en la predicción realizada. Las gráficas de resultado ayudan a tener una interpretación visual de dichos errores, de manera que se observa el ajuste entre los datos observados y los predichos por el modelo. En las siguientes subsecciones se presentan las gráficas para la evaluación de resultado más utilizadas en los diferentes modelos lineales.

5.4.5.1. Gráfico de dispersión de pronósticos. Como su nombre lo indica, en las gráficas de dispersión de pronósticos en los modelos lineales, se puede observar la relación entre los valores pronosticados y los valores observados dentro de un modelo; indicando de una manera visual cómo se dio el ajuste de cada uno de los valores pronosticados respecto a los valores observados dentro del modelo.

5.4.5.2. Gráfico de residuales. Los residuos se utilizan para evaluar la diferencia entre los valores observados y los valores que predice el modelo. Los residuales (E_i) quedan definidos como la diferencia entre un verdadero valor en la variable y ($\boldsymbol{y_i}$) y su valor predicho $\hat{y}_i$ como se observa en la expresión (15) (Jong & Heller, 2008):

$$E_i = y_i - \hat{y}_i \quad (15)$$

Gráficamente, un residual corresponde a un punto del diagrama de dispersión representado por su distancia vertical a la recta de regresión.

6. Metodología

En el presente capitulo se describen de forma detallada las actividades realizadas para la construcción, implementación y validación de los modelos con herramientas de Aprendizaje Automático para la predicción de rendimientos de cultivos de cacao en Santander. Cabe resaltar que para ello se hizo uso del lenguaje de programación Python en su versión 3, el cual se implementó para el desarrollo y programación de los modelos de Máquinas de Soporte Vectorial y Modelo Lineal Generalizado.

6.1. Conjunto de datos

El conjunto de datos es el insumo fundamental de todo análisis predictivo, el cual es necesario para pronosticar la variable respuesta, puesto que es la fuente que contiene la información de variables y su relación entre ellas. Para efectos de la presente investigación, el conjunto de datos de entrada empleados en el desarrollo y validación del modelo fueron suministrados por AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) los cuales fueron

tomados a partir de un cultivo experimental ubicado en el centro de investigación La Suiza en Rio Negro, Santander, para los años 2015, 2016 y 2017. El cultivo experimental de cacao está compuesto por 3 factores: fertilización, clon y exposición. Para el factor fertilización se tiene en cuenta 3 niveles: fertilización al 50%, 100%, 150%. En segundo lugar, para el factor clon se eligen 10 tipos de clones más representativos de Santander clasificados en 5 regionales (SCC19, SCC-52, SCC-61, SCC-64 Y SCC-83) y 5 universales (ICS-95, CNN-51, ETT-8, TSH565 y ICS-1). Por último, la exposición del cultivo es a sol o a sombra. Considerando estos factores y sus correspondientes niveles se tiene un total de 60 tratamientos (3 × 10× 2). A cada uno de los tratamientos le corresponden 15 plantas, por lo que se tiene un total de 900 de ellas a las cuales se les miden características fotosintéticas, morfológicas, físicas y químicas del suelo. De igual manera, el cacao producido por estas plantas se evalúa en términos de características morfológicas de la planta (altura de la planta, número de ramas, diámetro del tronco, peso de la almendra, peso de la mazorca completa, número de mazorcas, entre otros). Además, para las características fotosintéticas fueron tomadas 3 muestras, en tres momentos, una vez al año durante el periodo de 2015-2017 en cada uno de los 60 tratamientos, para un total de 540 observaciones (3 muestras × 60 tratamientos × 3 años). Por otro lado, se cuenta con 2160 observaciones para las características morfológicas de la planta tomadas una vez por semestre, realizando 3 repeticiones para 2 árboles por repetición para los 60 tratamientos. A su vez, se cuenta con 7239 observaciones de condiciones ambientales del cultivo experimental, las cuales fueron medidas con ayuda de sensores que realizaron recuentos para el periodo comprendido entre 2015 y 2017. Adicionalmente, se obtiene información de variables climáticas y se acude a fuentes secundarias como el IDEAM para obtener información sobre niveles de lluvia en la región.

En el apéndice A, se presenta la descripción de las variables independientes clasificadas dentro de las características mencionadas anteriormente.

6.1.1. Limpieza del conjunto de datos. La calidad de los datos según el autor Morbey, G (2013), está definida como el grado de cumplimiento de todos los requisitos definidos para los datos que se necesitan para un propósito específico. La calidad de los datos puede verse afectada debido a diversos problemas que se pueden presentar durante el proceso de la toma de estos (Maydanchik, A., 2007). Según los autores Rossin & D. (1999), al tratarse de la precisión predictiva de los modelos, esta puede verse afectada en gran magnitud debido a los errores en los datos utilizados para la predicción. Con el fin de garantizar la calidad de los datos, se adaptaron los pasos descritos en el trabajo realizado por los autores Corrales, Corrales, & Ledezma (2018), donde identifican tareas de limpieza de datos para la construcción de modelos predictivos, los cuales se muestran a continuación en la figura 7.

En primera instancia, para el tratamiento de datos perdidos se realiza una imputación de datos mediante la media, método que fue utilizado por los autores Corrales, Corrales, & Ledezma (2018) en su investigación, demostrando ser una buena técnica para la imputación de los datos perdidos. Posteriormente, para asegurar que todas las variables observadas, que son las que conforman los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas sean adimensionales, se realiza una normalización de las variables independientes, la cual consiste en quitarle a cada valor del conjunto su media y dividirla en su desviación estándar tal y como se muestra en la ecuación (16):

$$z = \frac{x - \mu}{\sigma} \quad (16)$$

Donde:

- z: Es el valor normalizado
- x: Es el valor a normalizar
- μ: Es la media aritmética de la distribución
- σ: Desviación estándar de la distribución

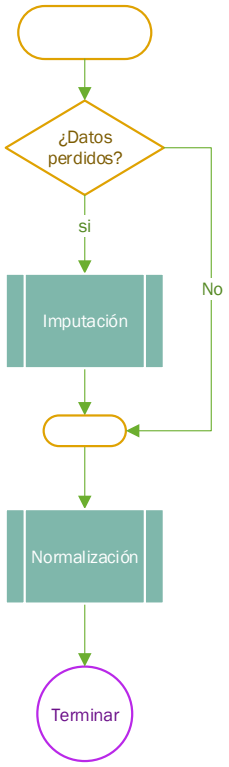

Figura. 7. Proceso de limpieza de datos. Adaptado de How to address the data quality issues in regression models: A guided process for data cleaning (Corrales, D. C., Corrales, J. C., & Ledezma, A., 2018).

6.2. Análisis de correlaciones

Según el autor Faraway (2004), a la hora de especificar el modelo hay que verificar si no existe un problema de multicolinealidad entre todas las variables. En el caso de que el supuesto de multicolinealidad no se cumpla, una forma de evitar el problema es no incluir las variables que se relacionen, y de esta manera no incrementaría la varianza. Es por esto, que con ayuda de la función *corcoef*, del lenguaje de programación Python, se realizó un análisis de correlación para evaluar la asociación lineal entre las variables explicativas que conforman cada uno de los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas, los cuales se presentan en las siguientes subsecciones.

6.2.1. Análisis de correlación para el conjunto de variables fotosintéticas. En la figura 8, se presenta la matriz de correlación para el conjunto de datos de variables fotosintéticas. En base a esta, se observa que las variables “Transpiración” y “Uso eficiente de agua” son las variables que presentan menor correlación entre sí. Por el contrario, las demás variables; “Conductividad” y “Asimilación de CO2”, presentan una gran similitud determinada por su alto coeficiente de correlación.

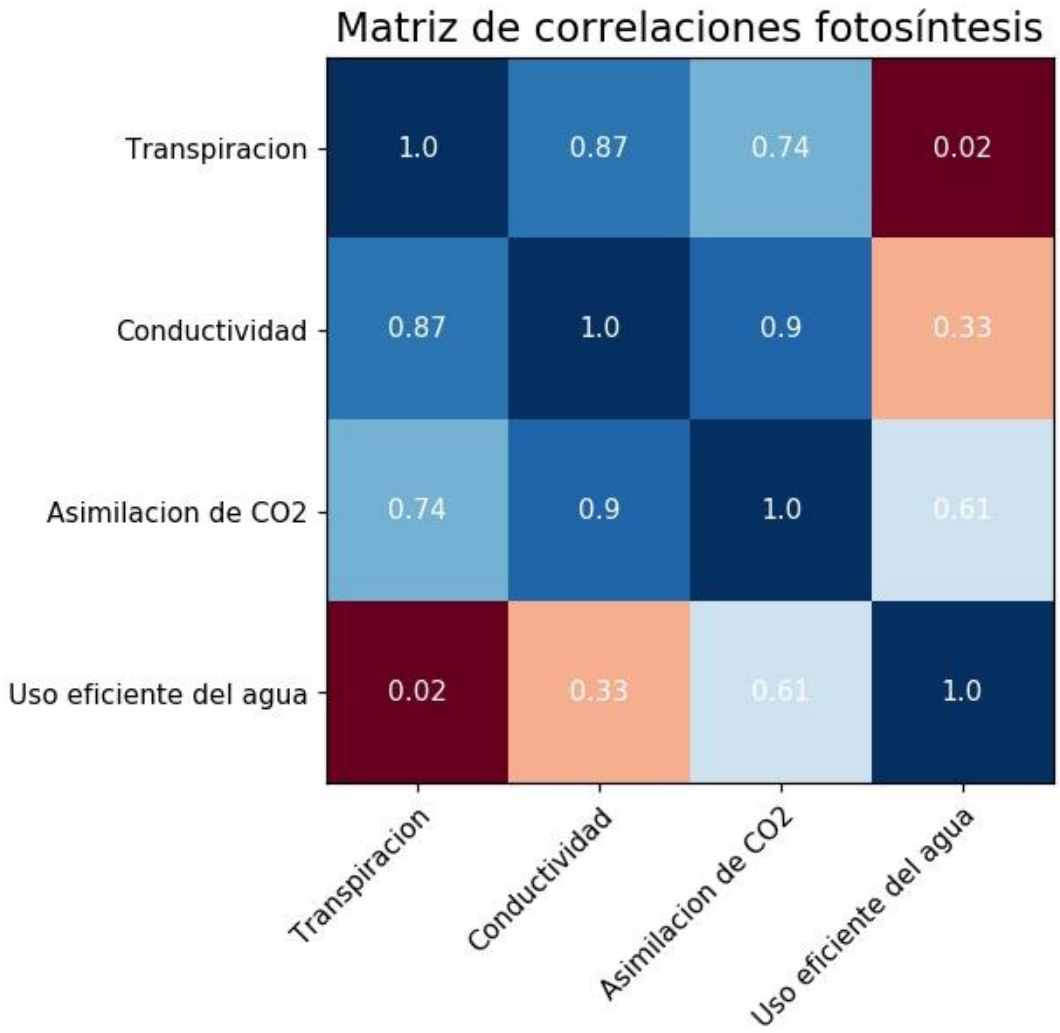

Figura. 8. Matriz de correlación del conjunto de variables fotosintéticas.

6.2.2. Análisis de correlación para el conjunto de variables morfológicas. En la figura 9, se observa la matriz de correlación para el conjunto de datos de variables morfológicas. De esta, se observa que las variables “Diámetro del tronco” y “Número total de ramas” son las variables que presentan una baja correlación entre sí y que, por el contrario, las demás variables tienen similitud entre sí.

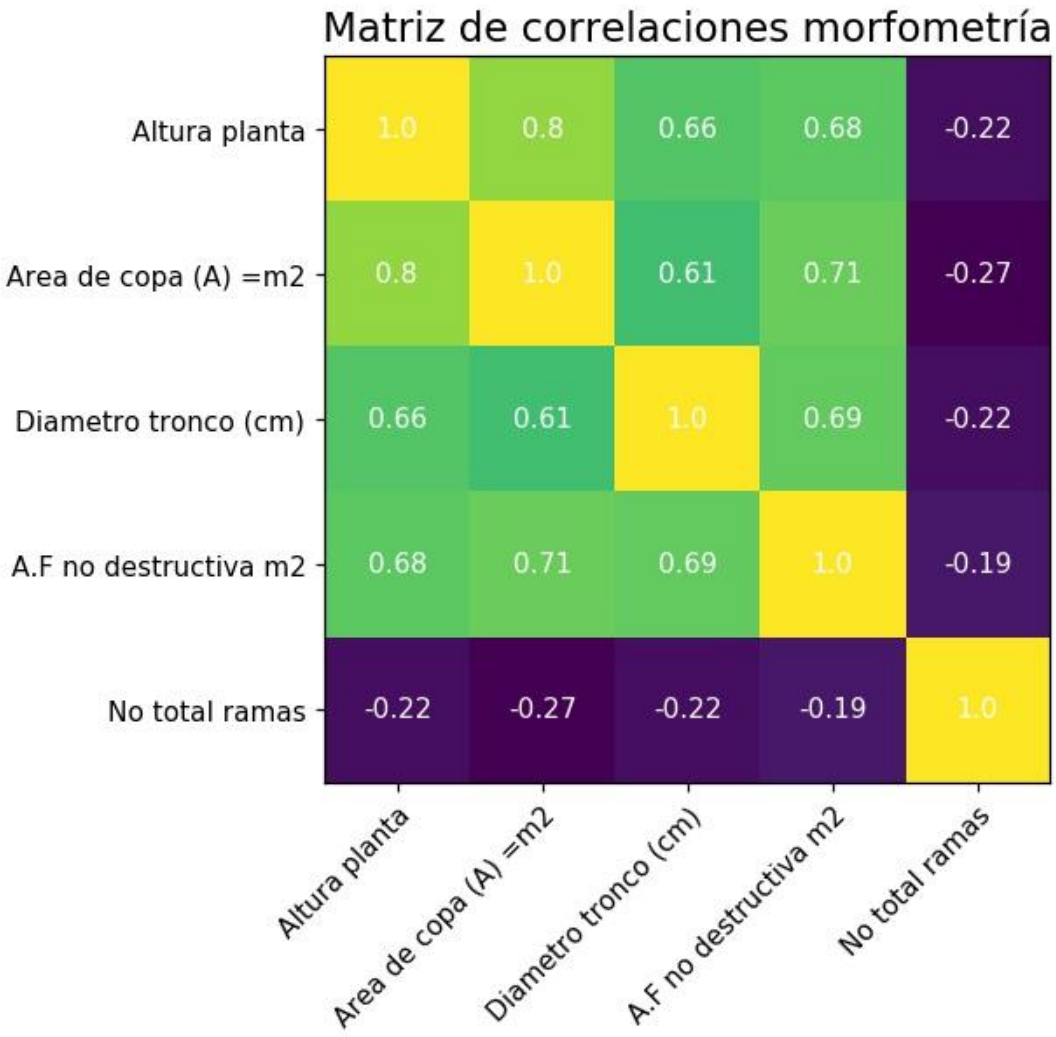

Figura. 9. Matriz de correlación del conjunto de variables morfológicas.

6.2.3. Análisis de correlación para el conjunto de variables características químicas del suelo. En la figura 10, se muestra la matriz de correlación para el conjunto de datos de características químicas del suelo. De ella, se observa que las variables “Fosforo (P)”, “Materia orgánica (MO), Magnesio (Mg) y Sodio (Na)” son las variables que presentan una baja correlación entre sí, a diferencia de las demás variables que presentan un alto coeficiente de correlación.

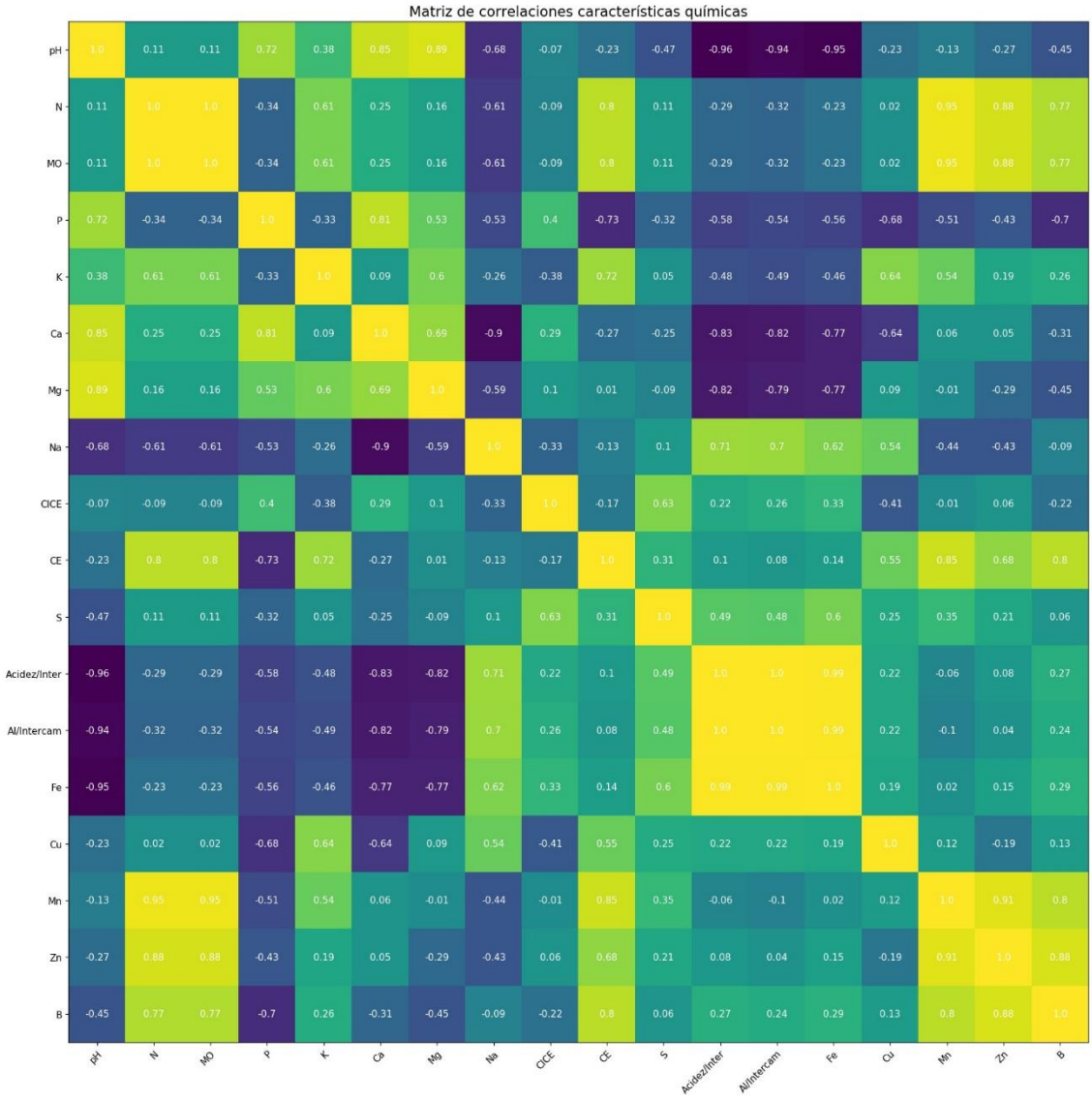

Figura. 10. Matriz de correlación del conjunto de características químicas del suelo

6.2.4. Análisis de correlación para el conjunto de variables características físicas del suelo. De la figura 11, se observa la matriz de correlación para el conjunto de datos de características físicas del suelo. Allí, se observa que las variables "%Arena (%Ar)" y "%Humedad gravimétrica (%Hum/Grav)" son aquellas que tienen menor similitud entre sí, deducido por su bajo coeficiente de correlación. Por el contrario, las variables que presentan mayor coeficiente de correlación son "%Limo(%L)", "%Arcilla(%A), "DA", "DR" y "%Poros".

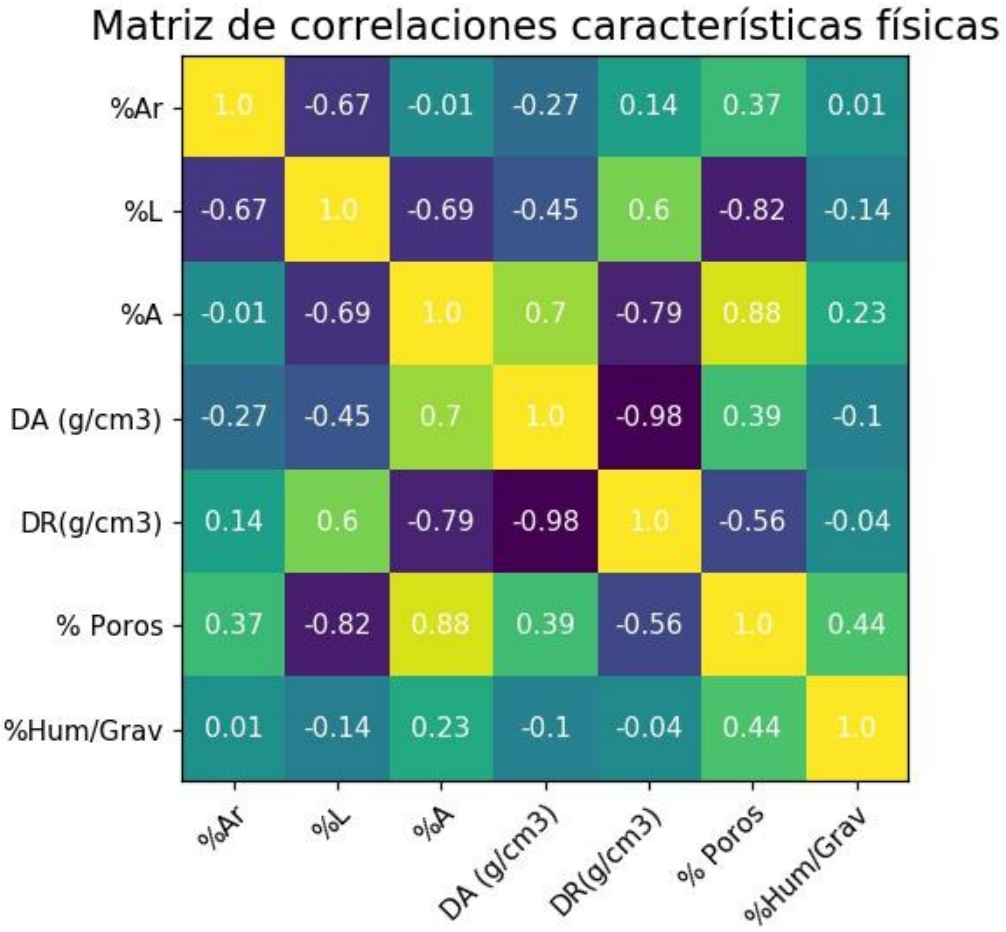

Figura. 11. Matriz de correlación del conjunto de características físicas del suelo.

6.2.5. Análisis de correlación para el conjunto de variables climáticas. En la figura 12, se visualiza la matriz de correlación para el conjunto de datos de variables climáticas. De ella, se observa que las variables "Radiación", "Temperatura y "Humedad" son las variables que

presentan menor correlación entre sí. Por el contrario, las demás variables; “Conductividad eléctrica”, “Lluvias acumuladas” y “Número de días con lluvia”, presentan una gran similitud entre sí determinada por su alto coeficiente de correlación.

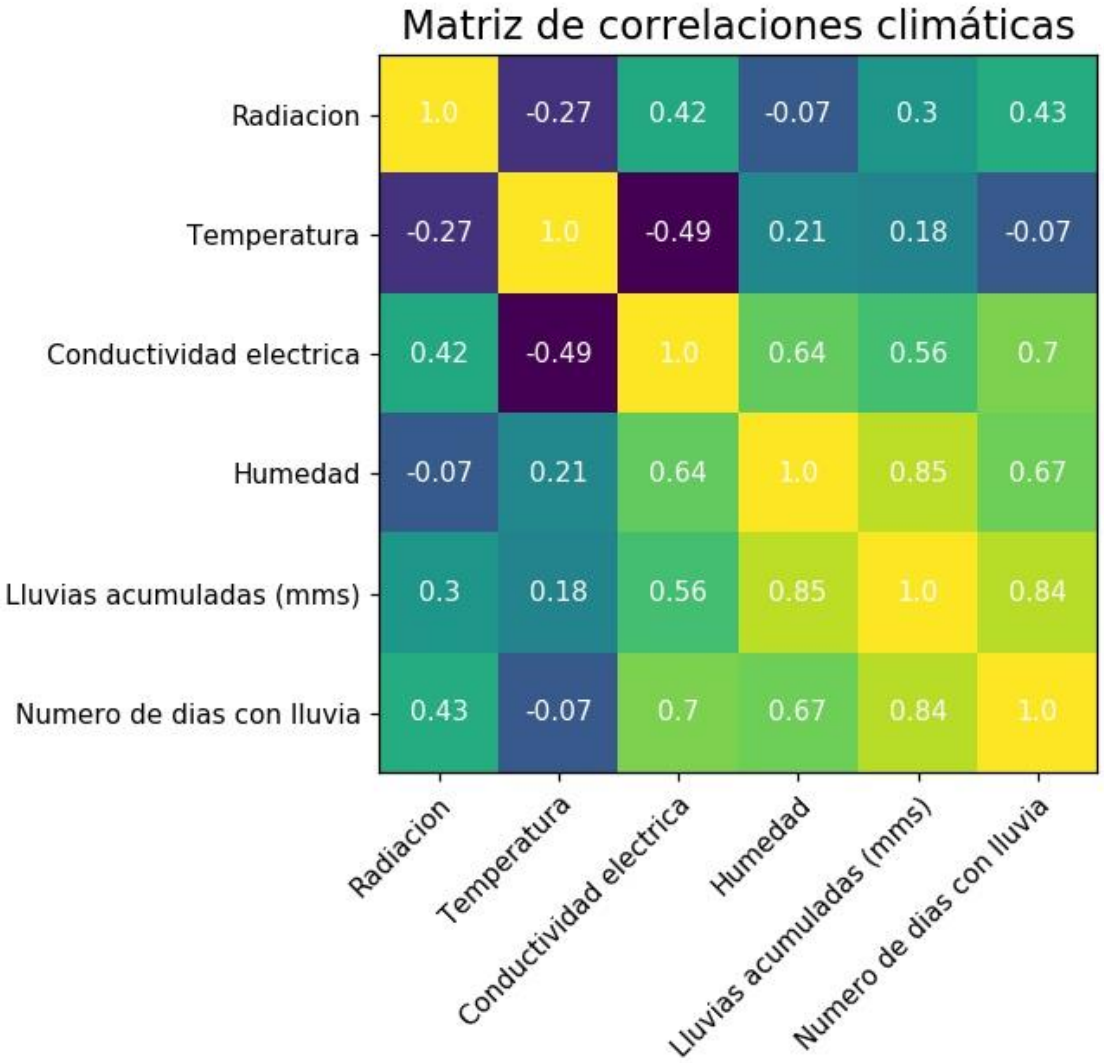

Figura. 12. Matriz de correlación del conjunto de variables climáticas.

6.3. Prueba de KMO y esfericidad de Bartlett

Adicional al análisis de correlación, se realizó la prueba de KMO y esfericidad de Bartlett con ayuda del software SPSS Statistics para las variables explicativas con más baja correlación descritas anteriormente que conforman cada uno de los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas. La prueba KMO consiste en evaluar las correlaciones parciales entre las variables, en donde los valores pequeños

(menores a 0,5) del valor KMO, indican una baja correlación entre las variables. Los valores para cada uno de los conjuntos de datos se presentan en la siguiente subsección.

6.3.1. Prueba de KMO y esfericidad de Bartlett para los conjuntos de variables.

		VALOR				
	MEDIA	Variables Fotosintéticas	Variables Morfológicas	Características Químicas del suelo	Características Físicas del suelo	Variables Climáticas
KMO		0,5	0,5	0,325	0,5	0,325
PRUEBA DE ESFERICIDAD DE BARTLET	Aprox chi-cuadrado	0,603	255	218	21	81
	Gl	1	1	6	1	6
	Significancia	0,437	0,1	0,1	0,1	0,1

Figura. 13. Prueba KMO y esfericidad de Bartlett**.** Adaptado de software SPSS.

En la figura 13 se muestran los valores obtenidos a partir de la prueba KMO y esfericidad de Bartlett, donde se observa que el KMO para las variables fotosintéticas, variables morfológicas y para las características físicas del suelo es de 0,5 y en cuanto a las características químicas del suelo y las variables climáticas se obtiene un valor de 0.325; lo cual indica que existe una baja correlación entre las variables explicativas de todos los conjuntos de variables.

6.4. Selección de variables explicativas

A partir del análisis de correlaciones, soportado en la prueba de KMO y esfericidad de Bartlett descrita anteriormente, se procede a seleccionar las variables correlacionadas en menor proporción para cada uno de los conjuntos de datos como variables explicativas, o bien llamadas variables de entrada para cada uno de los modelos, las cuales se presentan en la figura 14.

Figura. 14. Variables explicativas seleccionadas**.**

6.5. Distribución de variable respuesta

Dada la importancia de conocer la distribución de la variable dependiente: Rendimiento del cultivo de cacao, se tomaron los valores observados de la variable dependiente del conjunto de datos para conocer su distribución. A continuación, se describe la prueba de hipótesis realizada para determinar la distribución que ésta sigue.

6.5.1. Prueba de hipótesis para la distribución de la variable dependiente. Para determinar la distribución que sigue la variable dependiente, se utilizó el software Minitab en el cual se contrasta la siguiente hipótesis:

Ho: Los datos siguen una distribución tipo Gamma.

H1: Los datos siguen una distribución diferente a Gamma.

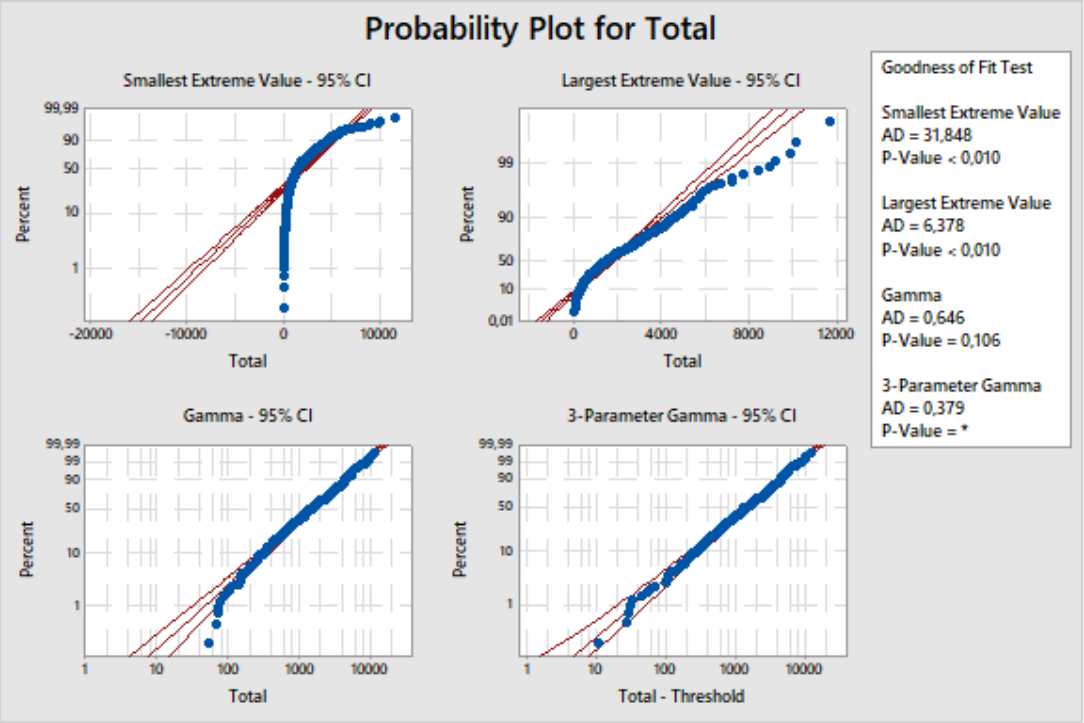

Figura. 15. Distribución de la variable respuesta. Adaptado del software Minitab.

En la figura 15 se puede observar la respuesta obtenida a partir del software Minitab, en donde se visualiza un P-Valor de 0,106 > 0,05; lo cual indica que no hay evidencia para rechazar la hipótesis nula, determinando que la distribución que sigue la variable dependiente es tipo Gamma, como se puede evidenciar en su histograma con variables no negativas, sesgadas a la derecha y con grandes valores en la cola superior (figura 16).

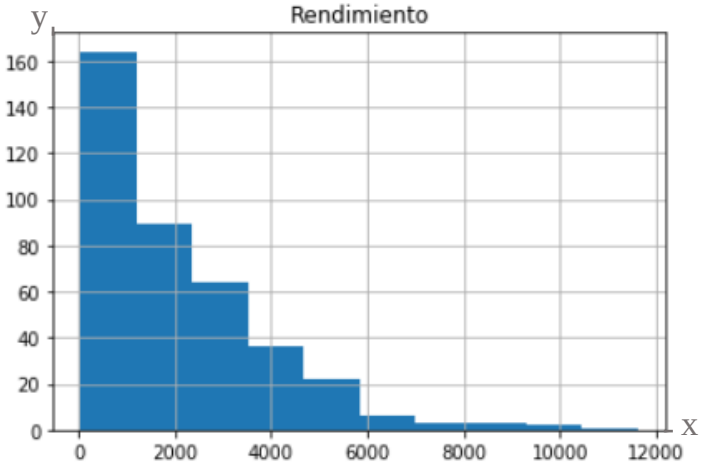

Figura. 16. Histograma distribución variable respuesta.

6.6. Construcción de modelos con técnicas de Aprendizaje Automático

Para la construcción de los modelos se utilizó el lenguaje de programación Python. Esta herramienta cuenta con una extensa variedad de librerías para desarrollar modelos con herramientas de Aprendizaje Automático, lo que representa un lenguaje apropiado para esta investigación.

6.6.1. Modelo Lineal Generalizado. Para la programación del Modelo Lineal Generalizado, se utilizó la librería *statsmodels.api* del lenguaje de programación Python. En primera instancia, se importó el conjunto de datos previamente procesado y consolidado.

A partir del análisis de correlaciones, prueba KMO y esfericidad de Bartlett, previamente explicadas en las subsecciones 6.3 y 6.4, se procede a definir dentro del modelo las variables exógenas o explicativas (x), previamente seleccionadas y mencionadas en la subsección 6.4 (Figura 13), y la variable endógena a predecir: rendimiento del cultivo (y).

De la sección 6.5, a partir de la prueba de hipótesis descrita para la distribución de la variable respuesta, se comprueba su comportamiento tipo Gamma; lo cual ratifica el uso apropiado del Modelo Lineal Generalizado al tener una distribución de la variable respuesta perteneciente a la familia de las distribuciones exponenciales, por lo que se emplea la función de enlace log. En la tabla 4 se muestran los resultados para los coeficientes de regresión, desviación estándar y valor p.

Tabla 4.
Modelo Lineal Generalizado con todas las variables.

Variables	**Coef**	**Std err**	**Z**	**P>|z|**	**[0.025**	**0.975]**
Uso eficiente del agua	0.019	0.049	0.384	0.701	-0.078	0.116
Transpiración	0.1403	0.072	1.954	0.051	0.000	0.281
Diámetro del tronco (cm)	0.0607	0.015	3.998	0.000	0.031	0.091
P	0.9526	0.151	6.295	0.000	6.656	1.249
MO	-0.5472	0.473	-1-157	0.247	-1.474	0.379
Na	-2.1315	7.464	-0.286	0.775	-16.761	12.498
Mg	2.9475	0.407	7.245	0.000	2.150	3.745
%Arena	-0.2185	0.046	-4.777	0.000	-0.308	-0.129

[Continuación Tabla 4]

Variables	Coef	Std err	Z	P>|z|	[0.025	0.975]
%Hum/Grav	0.4094	0.084	4.847	0.000	0.244	0.575
Radiación	6.091e-0	1.09e-0	5.587	0.000	3.95e-07	8.23e-07
Temperatura	-0.2819	0.075	-3.774	0.000	-0.428	-0.135
Humedad	-35.3668	4.657	-7.595	0.000	-44.494	-26.240
Lluvias acumuladas	0.0090	0.001	6.983	0.000	0.006	0.012

Según los autores Jong & Heller (2008) los valores estimados de β se utilizan para definir las variables independientes que mejor explican la variable respuesta, con lo cual se contrasta una hipótesis para encontrar estas variables de acuerdo con los valores estimados de β. La prueba de hipótesis planteada es:

. Se plantea la siguiente prueba de hipótesis dentro del código de programación:

Ho: $|\beta| = 0$, el coeficiente de la variable respuesta es igual a cero (no es significativa para el modelo)

H1: $|\beta| \neq 0$, el coeficiente de la variable respuesta es diferente a cero (es significativa dentro del modelo)

Donde $|\beta|$ es el coeficiente de regresión de cada variable independiente.

Considerando que si el p-valor mostrado en la tabla 5 toma un valor inferior a 0,05, se rechaza la hipótesis nula y, por el contrario, si este valor es superior a 0,05 no se tiene evidencia significativa para rechazar.

Dado lo anterior y a partir de la tabla 5, se identifica que las variables más significativas para el modelo son: Diámetro tronco (cm), Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas (mms), lo que quiere decir que se excluyeron para este caso las variables: Sodio (Na), Materia Orgánica (Mo), transpiración y Uso eficiente del agua.

Finalmente, el modelo fue entrenado nuevamente con las variables identificadas.

Los resultados obtenidos después de entrenar nuevamente el modelo se muestran en la tabla 5.

Tabla 5.
Modelo Lineal Generalizado con las variables seleccionadas.

Variables	Coef	Std err	z	P>|z|	[0.025	0.975]
Diámetro del tronco	0.061	0.017	3.583	0.000	0.028	0.094
P	0.999	0.174	5.757	0.000	0.659	1.339
Mg	2.677	0.394	6.798	0.000	1.905	3.448
%Arena	-0.223	0.047	-4.707	0.000	-0.315	-0.130
%Hum/Grav	0.395	0.086	4.586	0.000	0.226	0.564
Radiación	5.073e-07	8.64e-08	5.869	0.000	3.38e-07	6.77e-07
Temperatura	-0.298	0.059	-5.024	0.000	-0.414	-0.182
Humedad	-35.440	4.997	-7.092	0.000	-45.235	-25.646
Lluvias Acumuladas	0.009	0.001	7.019	0.000	0.006	0.011

Adaptado del lenguaje de programación Python.

El código de la construcción del Modelo Lineal Generalizado se muestra en al apéndice B.

6.6.2. Máquinas de Soporte Vectorial. Para la programación del modelo de Máquinas de Soporte Vectorial se utilizó la librería *sklearn.svm* importando las Máquinas de Soporte para Regresión. En primera instancia, se importó el conjunto de datos previamente procesado y consolidado. Es importante tener en cuenta que, para este modelo, no es necesario tener en cuenta supuestos para su desarrollo.

El modelo SVM se entrenó con las mismas variables definidas como exógenas o explicativas (x) previamente seleccionadas en el modelo GLM; donde a partir de su prueba de hipótesis almacenó las variables significativas. Por otro lado, la variable endógena a predecir (y), sigue siendo el rendimiento del cultivo.

Para el presente modelo, desde un primer momento se realiza una división 80/20 para el conjunto de datos obteniendo dos subconjuntos: 20% para la prueba y 80% para el entrenamiento, con el fin de poder hacer posteriormente la búsqueda y selección de los mejores parámetros. Acto seguido, se procede a realizar esta búsqueda en el subconjunto de puntos de datos de entrenamiento con ayuda de la validación cruzada, teniendo en cuenta que el modelo de Máquinas de Soporte Vectorial depende de una función kernel y una serie de hiperparámetros adecuados para su entrenamiento (Bishop, 2012). Para determinar los parámetros C, Gamma y kernel del

modelo SVM, se utilizó la función *GridSearch* del lenguaje Python, el cual trabaja con el método de validación cruzada.

Luego de encontrar los mejores parámetros para el *kernel tipo lineal* y el *kernel tipo rbf*, se realiza la predicción del modelo para cada uno de los dos tipos de kernel, esta vez indicando en cada función los mejores parámetros para cada uno de ellos hallados previamente.

Para obtener una predicción acertada de la variable respuesta en el presente modelo, se realiza nuevamente la búsqueda de parámetros con la función *GridSearch*, pero esta vez incluyendo dentro de ellos los dos tipos de *kernel*; lineal y rbf, con el fin de encontrar los valores más apropiados para los parámetros. Valores que son utilizados al momento de entrenar y validar el modelo. Los parámetros seleccionados dentro del modelo se presentan en la figura 17.

El código de la construcción del modelo de Máquinas de Soporte Vectorial se muestra en al apéndice C.

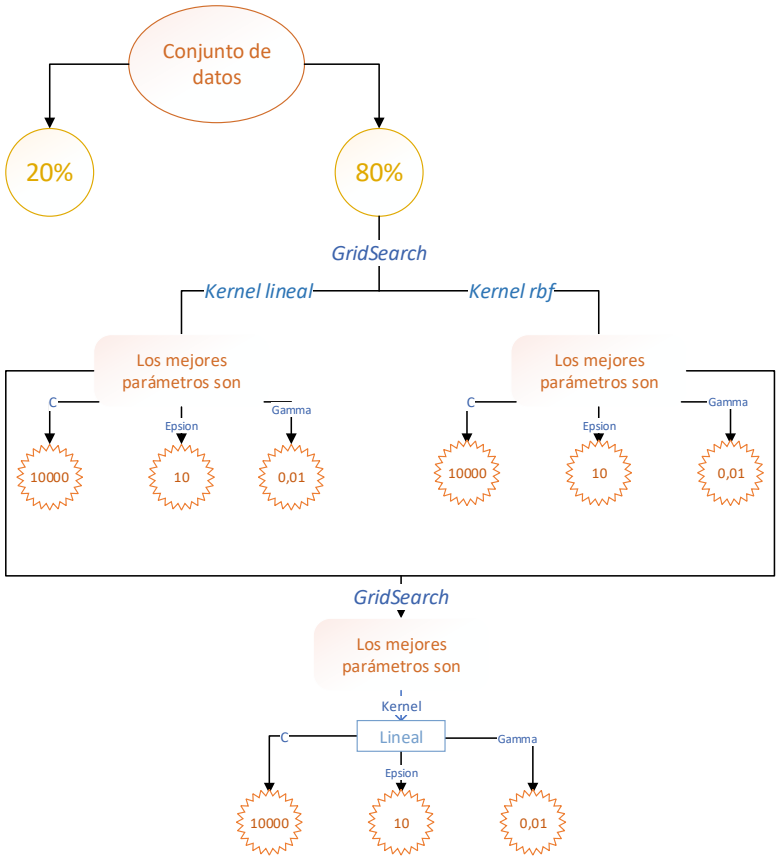

Figura. 17. Selección de mejores parámetros para modelo SVM.

6.7. Validación de modelos con métricas de ajuste

Para la validación de los modelos de Máquinas de Soporte Vectorial y Modelo Lineal Generalizado, se utilizó la técnica Hold Out, en la que se divide el conjunto de datos con una relación 80/20; utilizando el 80% de los datos para el entrenamiento y el otro 20% restante para la prueba de los modelos. En primera instancia, se crea una repetición de prueba en la cual se especifica dentro del código que realice una repetición de 100 veces para los errores de métricas de ajuste. Finalmente, el código genera un valor promedio de las 100 repeticiones para cada una de las métricas de ajuste.

6.7.1. Prueba de hipótesis para igualdad de medias. A partir de la técnica Hold Out y de cada uno de los valores de error RMSE, se planteó una prueba de hipótesis sobre la igualdad de medias para verificar cuál de los dos modelos contaba con menor error, es decir, con mayor precisión. La prueba de hipótesis se realizó con ayuda del software Minitab y se muestra a continuación:

Ho: $| RMSE_{GLM} | = | RMSE_{SVM} |$, el Error Cuadrático Medio del modelo GLM es igual al Error Cuadrático Medio del modelo SVM (Los dos modelos tienen la misma precisión para la predicción)

H1: $| RMSE_{GLM} | \neq | RMSE_{SVM} |$, el Error Cuadrático Medio del modelo GLM es diferente al Error Cuadrático Medio del modelo SVM (El modelo GLM tiene mayor precisión para la predicción)

Considerando que, si el Valor P toma un valor inferior a 0,05, se rechaza la hipótesis nula y, por el contrario, si este valor es superior a 0,05 se acepta la hipótesis nula. Por otro lado, el software realiza la diferencia de los valores promedio del Error Cuadrático Medio como se muestra en la ecuación (17)

$$RMSE_{GLM} - RMSE_{SVM} \quad (17)$$

Teniendo en cuenta que, si la diferencia es negativa, el Error Cuadrático Medio del modelo GLM es menor, es decir, tiene mayor precisión para la predicción en comparación con el modelo SVM. Y, por el contrario, si la diferencia es positiva, el Error Cuadrático Medio del modelo SVM es menor, es decir, tiene mayor precisión para la predicción en comparación con el modelo GLM,

estas afirmaciones se cumplen siempre y cuando no haya un valor de 0 en el intervalo de confianza para las diferencias.

7. Resultados

Luego de las actividades, descritas anteriormente, se procede en el presente capitulo a presentar los resultados que arrojó el estudio.

7.1. Resultados modelos con Aprendizaje Automático

7.1.1. Modelo Lineal Generalizado. Posterior a la programación y descrita en la subsección 6.6.1, se procede a generar e interpretar los resultados para el Modelo Lineal Generalizado, los cuales se presentan en la tabla 6.

Tabla 6.
Resultados modelo GLM.

Variables	Coef	Std err	z	P>|z|	[0.025	0.975]
1. Diámetro del tronco	0.061	0.017	3.583	0.000	0.028	0.094
2. P	0.999	0.174	5.757	0.000	0.659	1.339
3. Mg	2.677	0.394	6.798	0.000	1.905	3.448
4. %Arena	-0.223	0.047	-4.707	0.000	-0.315	-0.130
5. %Hum/Grav	0.395	0.086	4.586	0.000	0.226	0.564
6. Radiación	5.073e-0	8.64e-08	5.869	0.000	3.38e-07	6.77e-07
7. Temperatura	-0.298	0.059	-5.024	0.000	-0.414	-0.182
8. Humedad	-35.440	4.997	-7.092	0.000	-45.235	-25.646
9. Lluvias Acumuladas	0.008	0.001	7.019	0.000	0.006	0.011

Adaptado del lenguaje de programación Python.

En la expresión (18), se presenta la ecuación que corresponde al modelo GLM.

$$ln\frac{\beta 0+\beta 1x1+\beta 2x2+\beta 3x3+..\beta nxn+\varepsilon}{1-(\beta 0+\beta 1x1+\beta 2x2+\beta 3x3+..\beta nxn+\varepsilon)} \quad (18)$$

Por sus coeficientes positivos β, se puede inferir que las variables: Diámetro del tronco, Fosforo (P), Magnesio (Mg), %Hum/Grav, Radiación y Lluvias acumuladas tienen un efecto positivo sobre el rendimiento del cultivo de cacao y permiten explicar la variabilidad de la variable respuesta. Por otro lado, las variables %Arena, Temperatura y Humedad; las cuales cuentan con un coeficiente β negativo, pueden estar teniendo relación inversa con el rendimiento de cultivos de cacao. Las variables más influyentes para la variable respuesta encontradas en la presente investigación para el modelo GLM, fueron contrastadas con investigaciones como las de Zheng, Chen, Han, Zhao, & Ma (2009). En esta investigación, el Aprendizaje Automático Supervisado fue el protagonista para identificar las variables explicativas con mayor influencia en el rendimiento de cultivos de soja en el noreste de China y a pesar de no ser el mismo cultivo estudiado, se encontró que las lluvias acumuladas en el terreno de siembra, al igual que para la presente investigación, tienen una gran incidencia en la variabilidad del cultivo y que a su vez, el fosforo, en una proporción adecuada suministrada al cultivo, podría llegar a tener una incidencia positiva dentro del mismo.

7.1.1.1. Validación con métricas de ajuste para el modelo GLM. Como se mencionó en la subsección 6.7, el Modelo Lineal Generalizado se validó con la técnica Hold Out donde se generó un valor promedio derivado de estas 100 repeticiones, el cual se presenta en la tabla 7.

Tabla 7.
Métricas de ajuste para modelo GLM

Métrica de ajuste	Valor
R^2	0,1319
RMSE	1708,2904
MAE	1028,7805

Adaptado del lenguaje de programación Python.

7.1.1.2. Gráficas de resultado. En la presente subsección, se presentan las gráficas de resultado del modelo GLM, la cuales permiten tener una interpretación visual del ajuste del modelo. En la figura 18 se observa el gráfico de dispersión de los pronósticos.

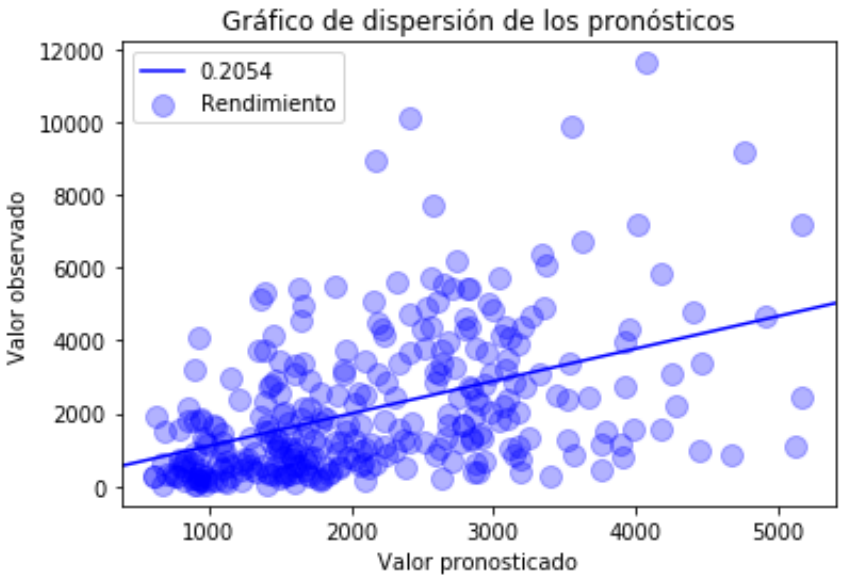

Figura. 18. Gráfico de dispersión de pronósticos. Adaptado del lenguaje Python.

En la figura se muestra la dispersión entre el valor pronosticado y el valor observado, el cual ilustra que cada uno de los puntos pronosticados están un tanto dispersos del valor observado. De lo anterior se concluye que los valores que pronosticó el modelo están alejados del valores reales.

Por otro lado, en la figura 19, se presenta el gráfico de residuales donde se observa que no se sigue algún patrón de comportamiento

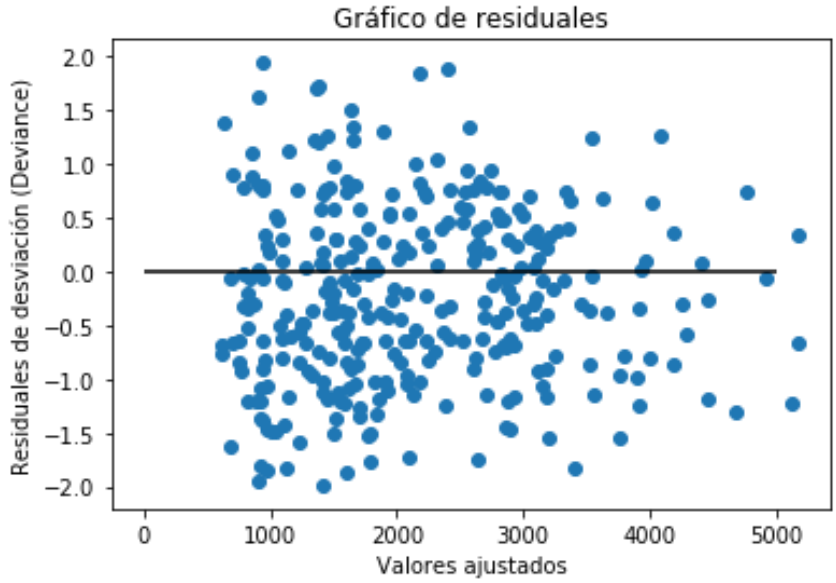

Figura. 19. Gráfico de residuales. Adaptado del lenguaje Python.

7.1.2. Máquinas de Soporte Vectorial. Recordando que para el modelo de Máquinas de soporte Vectorial se hizo la selección de los mejores parámetros con el fin de que el modelo se entrenara y arrojara resultados más acertados, se utilizó el kernel lineal cuya función se muestra a continuación en la ecuación (19).

$$G(x_j, x_k) = x_j{}' x_k \quad (19)$$

Después de realizar la programación utilizada para la construcción del modelo de Máquinas de Soporte Vectorial descrito en la subsección 6.6.2 y teniendo en cuenta que estos valores se obtuvieron utilizando un kernel lineal, en la tabla 8 se presentan los resultados de este modelo, obtenidos utilizando un kernel lineal para su correspondiente interpretación.

Tabla 8.
Resultados Modelo Máquinas de Soporte Vectorial.

Variable	***Coeficiente***
1. Diámetro tronco	439.873
2. P	2267.708
3. Mg	904.003
4. %Arena	-1494.358
5. %Hum/Grav	1124.024
6. Radiación	621.290
7. Temperatura	-555.847
8. Humedad	-1823.330
9. Lluvias acumuladas	1760.928

Adaptado del lenguaje de programación Python.

Como se observa en la tabla 8, las variables Diámetro del tronco, Fosforo (P), Magnesio (Mg), %Hum/Grav, Radiación y Lluvias acumuladas presentan coeficientes positivos, con lo que se concluye que, al igual que con el Modelo Lineal Generalizado, estas variables influyen de manera positiva sobre los rendimientos del cultivo de cacao y permiten explicar la variabilidad de la variable respuesta. Por otro lado, con la presencia del resto de las variables con coeficiente negativo, el rendimiento de cultivos de cacao podría disminuir a medida que estas aumenten.

7.1.2.1. Validación con métricas de ajuste para el modelo SVM. La validación del Modelo de Máquinas de Soporte vectorial se realizó con la técnica Hold Out y con el resto de los datos

correspondiente al 20% del total de ellos ya que, como se ha mencionado anteriormente, el otro 80% fue utilizado para el entrenamiento de este mismo.

Para esta técnica los datos se validaron a partir de 100 repeticiones para cada una de las métricas de ajuste (R2, RMSE y MAE), escogidas para la validación de cada uno de los modelos del presente proyecto. A partir de esto a cada métrica de ajuste le corresponden 100 valores para finalmente presentarse un valor promedio de cada conjunto de 100 datos de las 3 métricas de ajuste presentado en la tabla 9.

Tabla 9.
Métricas de ajuste para el Modelo de Máquinas de Soporte Vectorial.

Métrica de ajuste	Valor
R^2	0,1058
RMSE	1758,7421
MAE	926,2623

Adaptado del lenguaje de programación Python.

7.2. Comparación de modelos GLM y SVM

Como se presentó en subsecciones anteriores, cada uno de los dos modelos construidos (Modelo Lineal Generalizado y Máquinas de Soporte Vectorial) se validó con métricas de ajuste que permitieron evaluar el desempeño de cada uno de ellos. Con lo cual se extrajeron los valores de cada repetición del Error Cuadrático Medio, en la tabla 10 se presenta el promedio, la desviación estándar, el valor mínimo y máximo de este error para cada uno de los modelos, con el fin de comparar cuál contaba con menor error, es decir, con mayor precisión.

En el apéndice D se presentan los valores correspondientes a las 100 repeticiones para cada Error Cuadrático Medio.

Tabla 10.
RMSE Modelos.

	GLM	SVM
PROMEDIO	1685,604	1758,742
DESVIACIÓN ESTÁNDAR	186,136	235,177
VALOR MÍNIMO	1350,040	1292,150
VALOR MÁXIMO	2296,610	2374,580

A partir de la prueba de hipótesis descrita en la subsección 6.7.1.2, se obtienen los resultados de la comparación de los valores de la métrica de ajuste RMSE para cada uno de los modelos en la tabla 11.

Tabla 11.
Comparación RMSE modelos.

	RMSE	**Intervalo de confianza**	**Valor P**	**Diferencia μ (GLM) - μ (SVM)**
SVM	1758,7421	[-132,30; -14,0]	0,016	-50,6806
GLM	1708,0615			

Adaptado del software Minitab.

De la tabla anterior, se observa que el valor p es de 0,016, es decir, toma un valor inferior a 0,05; lo que indica que valor del Error Cuadrático Medio del Modelo GLM es diferente al valor del Error Cuadrático Medio del Modelo SVM. Por otro lado, teniendo en cuenta que la diferencia $RMSE_{GLM} - RMSE_{SVM}$ es negativa, el modelo GLM tiene menor Error Cuadrático Medio en comparación al modelo SVM, lo cual indica que el modelo GLM tiene mayor precisión para la predicción en comparación con el modelo SVM.

7.3. Discusión de resultados

A partir de los dos modelos predictivos: Modelo Lineal Generalizado y Máquinas de Soporte Vectorial, y de cada uno de los coeficientes de cada modelo, se evidenció que cada uno de ellos identificó las mismas variables, tanto aquellas que influyen de manera negativa como de manera positiva sobre la variable respuesta, las cuales se presentan en la figura 20 y 21 respectivamente.

En ese orden de ideas, a medida que: el diámetro del tronco de la planta, el fosforo y magnesio presente en el terreno de cultivo, la radiación, lluvias acumuladas y la humedad gravimétrica son mayores, va a aumentar el rendimiento del cultivo de cacao. Por otro lado, si la humedad, temperatura y porcentaje de arena presente en el cultivo no están presentes en las cantidades adecuadas y moderadas, el rendimiento del cultivo de cacao se va a ver afectado.

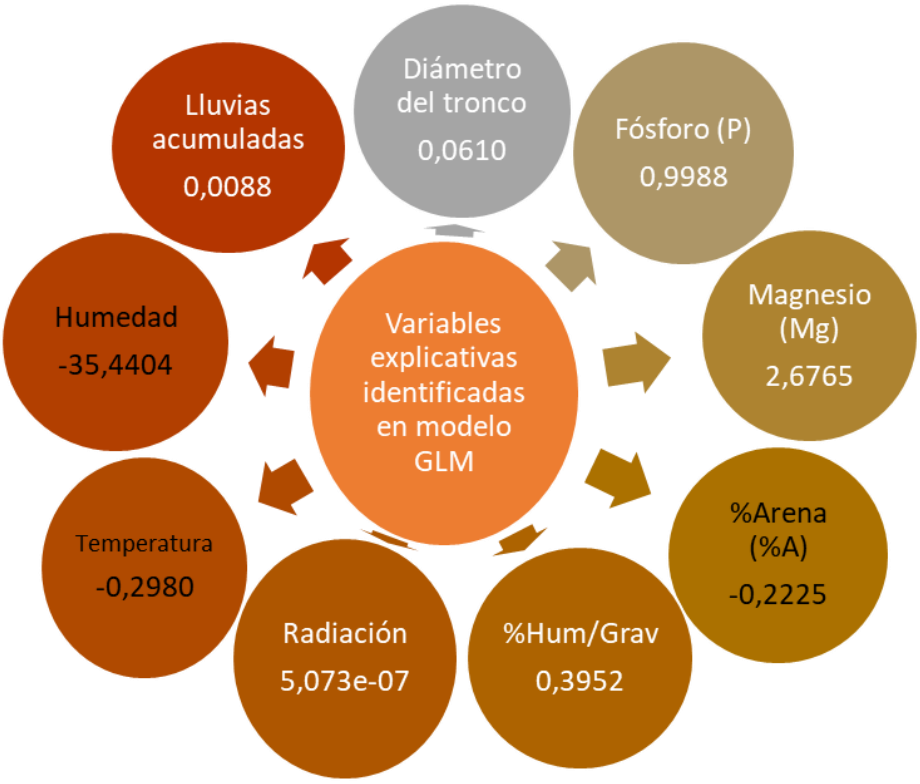

Figura. 20. Variables significativas identificadas en el modelo GLM.

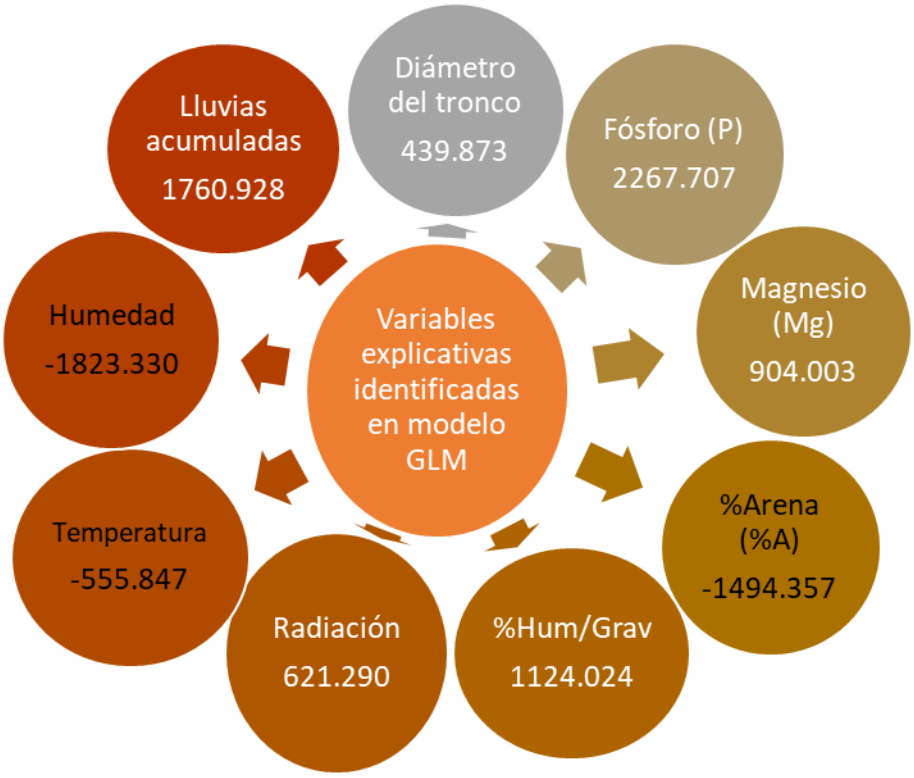

Figura. 21. Variables significativas identificadas en el modelo SVM

De los dos modelos se observa que, las variables climáticas son las variables que explicaron en mayor medida a la variable respuesta (rendimiento del cultivo de cacao) así como en investigaciones de Chattopadhyay & Mitra (2018) y Chen, Wu, & Liu (2016), en donde encontraron que factores climáticos como: horas de sol, rango de temperatura, lluvias, humedad, temperatura media afectan en gran medida a los rendimientos de cultivos de cultivos de soja y arroz, respectivamente.

8. Conclusiones

Considerando la búsqueda de estrategias para aumentar los rendimientos y mejorar las prácticas agrícolas en el sector cacaotero, surge el presente trabajo desarrollado con herramientas de Aprendizaje Automático supervisado.

En primera instancia, se observa que éstas herramientas han venido siendo utilizadas en aumento con el paso de los años por diferentes autores en el ámbito de la investigación, siendo herramientas precisas y prácticas, motivo por el cual se utilizan en la presente investigación.

Es importante tener en cuenta que la alta precisión en los modelos desarrollados se puede llegar a obtener siempre y cuando se cuente con un adecuado procesamiento y limpieza de datos y que, a su vez, se seleccionen los modelos más apropiados de acuerdo con las características propias de cada investigación. De ahí se deriva la importancia de conocer a profundidad estas características, especialmente de los datos de entrada, con el fin de que los investigadores construyan los modelos de la manera más apropiada según su objeto de investigación.

Además del conocimiento de las características propias de los datos de entrada, previo a la construcción del modelo, también se debe tener en cuenta la aplicación de diferentes pruebas que ayuden a determinar la distribución que siguen los datos observados. Para la presente investigación, la distribución de probabilidad de la variable respuesta (rendimiento del cultivo de cacao) es diferente a la normal, debido a que sigue una distribución de probabilidad tipo Gamma; por lo que se puede concluir que las dos herramientas de Aprendizaje Automático utilizadas para

la construcción de los modelos: Modelo Lineal Generalizado y Máquinas de Soporte Vectorial, son adecuadas para la predicción de la variable respuesta, ya que estos modelos se ajustan a los datos con características de distribuciones no normales de la variable respuesta.

Dado que cada investigación cuenta con sus propias características, se evidencia que los modelos encontrados en la revisión de literatura podrían estar sujetos a modificaciones y alteraciones. No obstante, sin importar las diferencias de las características de cada modelo, se evidencia que los modelos desarrollados con herramientas de Aprendizaje Automático pueden llegar a ser bastante útiles para los diferentes actores del sector agrícola, por cuanto estos pueden reducir la incertidumbre con la información brindada y, a su vez, comprender las condiciones más influyentes con el fin de tomar las decisiones más acertadas.

Las validaciones para los dos modelos realizados con herramientas de Aprendizaje Automático son de vital importancia para evaluar la precisión obtenida tanto para el Modelo Lineal Generalizado como para el de Máquinas de Soporte Vectorial. De esta validación, se obtuvo un coeficiente de regresión (R2) de 0.1319 y 0.1058 respectivamente. A pesar de no tener un alto valor, sí se pudo identificar las variables que influyen, dando prioridad a éstas por encima de la precisión del instrumento. Por otro lado, a partir de los valores de Error Cuadrático Medio (RMSE) como otra métrica de ajuste para cada uno de los modelos, se concluye que el Modelo Lineal Generalizado tiene mayor precisión en comparación al modelo de Máquinas de Soporte Vectorial debido a su más bajo valor de error RMSE; para el cual se obtuvo valores de 1708.29 y 1758.74 respectivamente.

Además del uso de la validación cruzada para evaluar el ajuste y precisión del instrumento, se pudo encontrar que esta también puede llegar a ser utilizada para encontrar los parámetros más adecuados con los que debe contar cada modelo de acuerdo con las características propias de cada investigación.

Para los dos modelos, GLM y SVM, se identificaron las variables explicativas de mayor impacto tanto de forma negativa como de forma positiva sobre la variable rendimiento de cultivo de cacao, con las cuales el decisor puede determinar la mejor combinación para obtener un mayor rendimiento dentro de los recorridos de las variables. La construcción y comparación de los

resultados de los dos modelos, fue útil para ratificar la influencia de las variables explicativas en los rendimientos de cultivos de cacao.

Los modelos presentaron información de los aspectos fundamentales que se deben tener en cuenta en los cultivos de cacao. Tanto el Modelo Lineal Generalizado como las Máquinas de Soporte Vectorial, identifican que a medida que el diámetro del tronco de la planta de cacao, el fosforo y magnesio presente en el terreno de cultivo, la radiación, las lluvias acumuladas y la humedad gravimétrica son mayores, va a aumentar el rendimiento del cultivo de cacao.

En primera instancia, el diámetro del tronco influye positivamente debido a que al tener la planta un mayor diámetro del tronco se puede almacenar en él un mayor número de mazorcas de cacao, aumentando así el rendimiento del cultivo.

Por otro lado, las características químicas del suelo: el fosforo y magnesio, son dos de los nutrientes requeridos por las plantas para un crecimiento optimo, por lo que son requeridos en grandes cantidades. Esto explica la influencia positiva determinada por los modelos en el rendimiento de los cultivos de cacao.

Por último, las condiciones climáticas óptimas: lluvias y humedad en el cultivo, ayudan a aumentar su rendimiento. De los resultados de la presente investigación, se observa que estas variables son las que explican en mayor medida a la variable respuesta rendimiento del cultivo de cacao, siendo el mayor número de variables identificadas dentro de un mismo conjunto de datos identificadas por los modelos como significativas.

Otras de las variables consideradas como significativas dentro de los modelos son: humedad, temperatura y porcentaje de arena del cultivo. Sin embargo, se observa que, si estas no están presentes en cantidades moderadas, el rendimiento del cultivo de cacao se va a ver afectado de una manera negativa.

Teniendo en cuenta los a aspectos fundamentales determinados por los modelos que influyen al rendimiento de cultivo de cacao mencionados anteriormente, es de suma importancia que los agricultores pongan en marcha el uso de diferentes fertilizantes que contribuyan a buenas cantidades de fosforo y magnesio en los cultivos de cacao, así como la selección de zonas donde haya precipitaciones constantes con terrenos templados para garantizar el buen rendimiento de los

cultivos de cacao. De esta manera, al estimular la práctica de labores que mejoren la calidad de este grano, se satisfacen las necesidades de las industrias procesadoras que demandan un grano que proporcione las características de sabor y aroma, garantizando con una buena calidad, el aumento en la exportación y precio del cacao siendo de gran ayuda para los actores del sector agrícola en Santander.

9. Recomendaciones

Se recomienda a los actores del sector tomar en consideración los resultados de este tipo de investigaciones para seguir mejorando los rendimientos de los diferentes cultivos y optar por la toma de nuevas observaciones en otras regiones del departamento de Santander para otros cultivos, y a su vez, para diferentes tipos de clones de cultivos de cacao implementando la construcción de diferentes modelos con diversos cultivos existentes a nivel de región y país para brindar información útil y oportuna que ayude a agricultores con la toma de decisiones acertadas , con el fin de aumentar la variabilidad en los datos y ampliar el campo de investigación, trabajando en conjunto con diferentes grupos de interés (gobierno, federaciones, agricultores) dentro de la investigación para lograr cada vez mejores resultados.

Por otra parte, se sugiere analizar y trabajar a profundidad con modelos enfocados en las variables climáticas para los cultivos debido a la influencia que estas presentan en la variable respuesta de los modelos, con ayuda de diferentes lenguajes de programación que sirvan como herramienta para continuar trabajando con la técnica de Aprendizaje Automático Supervisado de diferentes trabajos de investigación. Es por esto por lo que se invita a la comunidad y en especial a los estudiantes de la Universidad Industrial de Santander que han realizados proyectos involucrando nuevas tecnologías para el mejoramiento de los cultivos en Colombia mediante estas técnicas, a trabajar en conjunto con el fin de tener resultados más acertados y de mejorar la precisión de los modelos.

Referencias Bibliográficas

Aguilera, F., & Ruiz-Valenzuela, L. (2014). Forecasting olive crop yields based on long-term aerobiological data series and bioclimatic conditions for the southern Iberian Peninsula. *Spanish Journal of Agricultural Research*, *12*(1), 215–224. https://doi.org/10.5424/sjar/2014121-4532

Alvarez, R. (2009). Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy*, *30*(2), 70–77. https://doi.org/10.1016/j.eja.2008.07.005

Baratta, A. (2016). Introducción a Machine Learning. *SUNQU*, 197–224.

Berzal, F. (n.d.). *Clasificación y predicción*. Retrieved from http://elvex.ugr.es/idbis/dm/slides/3 Classification.pdf

Betancour, G. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia Et Technica*, (27), 67–72. https://doi.org/10.22517/23447214.6895

Bishop, C. M. (2012). *Pattern Recognition and Machine Learning*. https://doi.org/10.1021/jo01026a014

Cardona Madariaga, D. F., González Rodríguez, J. L., Rivera Lozano, M., & Cárdenas Vallejo, E. H. (2013). Aplicación de la regresión lineal en un problema de pobreza. *Revista Interacción*, *12*, 73–84. https://doi.org/ISSN 1665-9627

Cayuela, L. (2010). *Modelos lineales generalizados (GLM)*. Retrieved from https://s3.amazonaws.com/academia.edu.documents/33538949/3-Modelos_lineales_generalizados.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1534283770&Signature=SwPRBT18Y6cj24fcheD0xBQUxvE%3D&response-content-disposition=inline%3B filename%3DModelos_lineale

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., & Dentener, F. (2016). Impact of

meteorological drivers on regional inter-annual crop yield variability in France. *Agricultural and Forest Meteorology*, *216*, 58–67. https://doi.org/10.1016/j.agrformet.2015.10.004

Chattopadhyay, M., & Mitra, S. K. (2018). Assessing the predictability of different kinds of models in estimating impacts of climatic factors on food grain availability in India. *Opsearch*, *55*(1), 50–64. https://doi.org/10.1007/s12597-017-0314-9

Chen, H., Wu, W., & Liu, H.-B. (2016). Assessing the relative importance of climate variables to rice yield variation using support vector machines. *Theoretical and Applied Climatology*, *126*(1–2), 105–111. https://doi.org/10.1007/s00704-015-1559-y

CIAT. (2017). Retrieved August 13, 2018, from https://blog.ciat.cgiar.org/es/que-papel-puede-jugar-el-cacao-para-la-paz-en-colombia/

Corrales, D. C., Corrales, J. C., & Ledezma, A. (2018). How to address the data quality issues in regression models: A guided process for data cleaning. *Symmetry*, *10*(4), 1–20. https://doi.org/10.3390/sym10040099

Cunha, M., Ribeiro, H., & Abreu, I. (2016). Pollen-based predictive modelling of wine production: application to an arid region. *European Journal of Agronomy*, *73*, 42–54. https://doi.org/10.1016/j.eja.2015.10.008

DANE - PIB 2018. (2010). Retrieved from http://www.dane.gov.co/files/investigaciones/boletines/pib/bol_PIB_Itrim18_producion_y_gasto.pdf

Del Valle Moreno, J., & Guerra Bustillo, W. (2012). La Multicolinealidad en modelos de Regresión Lineal Múltiple. *Computación Y Matemática Aplicada*, *21*(4), 80–83.

FAO Marco programático Colombia (2015-2019). (n.d.). Retrieved from http://www.fao.org/3/a-bp556s.pdf

Faraway, J. J. (2004). Extending Linear Model With R. https://doi.org/10.1161/CIRCRESAHA.110.231803

FEDECACAO. (2018). Retrieved August 9, 2018, from http://www.fedecacao.com.co/

portal/index.php/es/2015-04-23-20-00-33/551-en-2017-colombia-alcanzo-nuevo-record-en-produccion-de-cacao

FH, J. (2006). ProClassify User's Guide - Cross-Validation Explained. Retrieved December 20, 2018, from http://genome.tugraz.at/proclassify/help/pages/XV.html

Fishman, J., Creilson, J. K., Parker, P. A., Ainsworth, E. A., Vining, G. G., Szarka, J., … Xu, X. (2010). An investigation of widespread ozone damage to the soybean crop in the upper Midwest determined from ground-based and satellite measurements. *Atmospheric Environment*, *44*(18), 2248–2256. https://doi.org/10.1016/j.atmosenv.2010.01.015

García-Mozo, H., Yaezel, L., Oteros, J., & Galán, C. (2014). Statistical approach to the analysis of olive long-term pollen season trends in southern Spain. *Science of the Total Environment*, *473–474*, 103–109. https://doi.org/10.1016/j.scitotenv.2013.11.142

Gavidia Bovadilla Dirigida por, G. E., & Oñate Ibañez de Navarra Ing Eduardo Soudah Prieto, E. (n.d.). *Clasificadores Basados en Máquinas de Soporte Vectorial para el Diagnóstico y Predicción de la Enfermedad de Alzheimer*. Retrieved from https://web.cimne.upc.edu/users/esoudah/publications/Tesis_Gio_2012_SVM_Alzheimer.pdf

Gonzalez-Sanchez, a. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439

Guzmán, D. (2012). Comparación de algunos R2 como medidas de bondad de ajuste en modelos lineales mixtos, 58.

Hansen, J. W., Potgieter, A., & Tippett, M. K. (2004). Using a general circulation model to forecast regional wheat yields in northeast Australia. *Agricultural and Forest Meteorology*, *127*(1–2), 77–92. https://doi.org/10.1016/j.agrformet.2004.07.005

Introducing Machine Learning. (n.d.). Retrieved from https://www.mathworks.com/tagteam/89703_92991v00_machine_learning_section1_ebook_v12.pdf

J. A. Nelder and R. W. M. Wedderbur. (2000). Generalized Linear Models By. *Jisuanji*

Xuebao/Chinese Journal of Computers, *23*(9), 370–384. https://doi.org/10.1002/jia2.25041

Jong, P. de, & Heller, G. (2008). *Generalized Linear Models for Insurance Data*.

Kar, G., & Kumar, A. (2014). Forecasting rainfed rice yield with biomass of early phenophases, peak intercepted PAR and ground based remotely sensed vegetation indices. *Journal of Agrometeorology*, *16*(1), 94–103.

Kevin, A., Hoz, D., Martínez-palacio, U. J., & Mendoza-palechor, F. E. (n.d.). ML en Medicina Cardiovascular.

Kohavi, R. (n.d.). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, (0), 0–6.

Kouadio, A. L., Djaby, B., Duveiller, G., El Jarroudi, M., & Tychon, B. (2012). Cinétique de décroissance de la surface verte et estimation du rendement du blé d'hiver. *Biotechnology, Agronomy and Society and Environment*, *16*(2), 179–191.

Kuncheva, L. I. (Ludmila I. (2004). *Combining pattern classifiers : methods and algorithms*. J. Wiley.

Lobell, D. B., Cahill, K. N., & Field, C. B. (2007). Historical effects of temperature and precipitation on California crop yields. *Climatic Change*, *81*(2), 187–203. https://doi.org/10.1007/s10584-006-9141-3

Logan, T. M., McLeod, S., & Guikema, S. (2016). Predictive models in horticulture: A case study with Royal Gala apples. *Scientia Horticulturae*, *209*, 201–213. https://doi.org/10.1016/j.scienta.2016.06.033

Mavromatis, T. (2014). Pre-season prediction of regional rainfed wheat yield in Northern Greece with CERES-Wheat. *Theoretical and Applied Climatology*, *117*(3–4), 653–665. https://doi.org/10.1007/s00704-013-1031-9

Mejía, L. F. (2018). *Departamento Nacional de Planeación*. Retrieved from https://colaboracion.dnp.gov.co/CDT/Prensa/Presentación Big Data Política explotación

datos.pdf

Mitchell, T. M. (1997). *(Mcgraw-Hill International Edit) Thomas Mitchell-Machine learning-McGraw Hill Higher Education (1997)*. https://doi.org/10.1016/j.cub.2007.11.035

Mkhabela, M. S., Mkhabela, M. S., & Mashinini, N. N. (2005). Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. *Agricultural and Forest Meteorology*, *129*(1–2), 1–9. https://doi.org/10.1016/j.agrformet.2004.12.006

Moreto, V. B., & Rolim, G. D. S. (2015). Agrometeorological models for groundnut crop yield forecasting in the Jaboticabal, São Paulo State region, Brazil. *Acta Scientiarum. Agronomy*, *37*(4), 403. https://doi.org/10.4025/actasciagron.v37i4.19766

Nadler, A. J., & Bullock, P. R. (2011). Long-term changes in heat and moisture related to corn production on the Canadian Prairies. *Climatic Change*, *104*(2), 339–352. https://doi.org/10.1007/s10584-010-9881-y

Naylor, R., Falcon, W., Rochberg, D., & Wada, N. (2001). Using El Nino/Southern Oscillation climate data to predict rice production in Indonesia. *Climatic Change*, 255–265. https://doi.org/10.1023/A:1010662115348

Negrón Baez, P. A. (2014). Redes neuronales sigmoidal con algoritmo LM para pronóstico de tendencia del precio de las acciones del IPSA. Retrieved from http://opac.pucv.cl/pucv_txt/txt-5500/UCE5728_01.pdf

Net-, N., & Pino, R. (2012). Predicción del índice IBEX-35 aplicando Máquinas de Soporte Vectorial y Redes, 1353–1360.

Oteros, J., Orlandi, F., García-Mozo, H., Aguilera, F., Dhiab, A. Ben, Bonofiglio, T., … Galán, C. (2014). Better prediction of Mediterranean olive production using pollen-based models. *Agronomy for Sustainable Development*, *34*(3), 685–694. https://doi.org/10.1007/s13593-013-0198-x

Pagani, V., Stella, T., Guarneri, T., Finotto, G., van den Berg, M., Marin, F. R., … Confalonieri,

R. (2017). Forecasting sugarcane yields using agro-climatic indicators and Canegro model: A case study in the main production region in Brazil. *Agricultural Systems*, *154*(March), 45–52. https://doi.org/10.1016/j.agsy.2017.03.002

Park, S. J., Hwang, C. S., & Vlek, P. L. G. (2005). Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. *Agricultural Systems*, *85*(1), 59–81. https://doi.org/10.1016/j.agsy.2004.06.021

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de Teledeteccion*, *2015*(44), 55–65. https://doi.org/10.4995/raet.2015.4153

PROCOLOMBIA. (n.d.). Retrieved August 13, 2018, from http://www.inviertaencolombia.com.co/sectores/agroindustria.html

Referencia, M. De, & Mendoza, L. L. P. O. (2008). Índice General.

Rolim, G. D. S., Novo, M. D. C. D. S. S., Pantano, A. P., & Trani, P. E. (2011). Modelagem agrometeorolÃ³gica para estimaÃ§Ã£o do desenvolvimento e da produÃ§Ã£o de jilÃ³\rAgrometeorological model to estimate development and production of "jilÃ³," (19), 832–837.

Rossin, D. F., & D., K. B. (1999). Data Quality in Linear Regression Models: Effect of Errors in Test Data and Errors in Training Data on Predictive Accuracy, *0*(November), 33–43.

SAC-Sociedad de Agricultores de Colombia. (n.d.). Retrieved August 10, 2018, from http://sac.org.co/es/noticias/536-el-cacao-sera-el-cultivo-de-la-paz.html

Secretaría de agricultura. (2017). Retrieved August 10, 2018, from http://www.santander.gov.co/index.php/secretaria-agricultura

Semana-Agricultura. (2017). Retrieved August 10, 2018, from https://www.semana.com/educacion/articulo/farc-en-que-se-quieren-formar-los-exguerrilleros-de-las-farc/531826

Tack, J., Barkley, A., & Nalley, L. L. (2015). Effect of warming temperatures on US wheat yields.

Proceedings of the National Academy of Sciences, *112*(22), 6931–6936. https://doi.org/10.1073/pnas.1415181112

Toggweiler, J., & Key, R. (2001). Ocean circulation: Thermohaline circulation. *Encyclopedia of Atmospheric Sciences*, *4*(December 2007), 1549–1555. https://doi.org/10.1002/joc

Toscano, P., Gioli, B., Genesio, L., Vaccari, F. P., Miglietta, F., Zaldei, A., … Porter, J. R. (2014). Durum wheat quality prediction in Mediterranean environments: From local to regional scale. *European Journal of Agronomy*, *61*, 1–9. https://doi.org/10.1016/j.eja.2014.08.003

UPRA. (n.d.). Retrieved August 10, 2018, from http://www.upra.gov.co/uso-y-adecuacion-de-tierras/evaluacion-de-tierras/zonificacion

Zheng, H., Chen, L., Han, X., Zhao, X., & Ma, Y. (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. *Agriculture, Ecosystems and Environment*, *132*(1–2), 98–105. https://doi.org/10.1016/j.agee.2009.03.004