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

Autores:

Autores: Andrea Carolina Gamboa Ariza

Paula Andrea Cáceres Ortiz

Trabajo de Grado para Optar el título de Ingeniero Industrial

Director

PhD. Henry Lamos Díaz

Codirector

Ing. David Esteban Puentes Garzón

Universidad Industrial de Santander

Facultad de Ingenierías Físico-Mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2019

**Nota de Proyecto de Grado**

(Va escaneada y con la información completa. Se diligencia al momento de hacer la sustentación)

**Autorización de Uso a Favor de la UIS**

Va escaneada y con la información completa, se puede descargar de la página de la Biblioteca: <http://tangara.uis.edu.co/biblioweb/pags/pub/informacion/recurso_ejemplos_TG.jsp>

***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**

Pág.

[Introducción 14](#_Toc534140604)

[1. Planteamiento del problema 16](#_Toc534140605)

[2. Justificación del proyecto 19](#_Toc534140606)

[3. Objetivos 20](#_Toc534140607)

[3.1. Objetivo General 20](#_Toc534140608)

[3.2. Objetivos Específicos 20](#_Toc534140609)

[4. Revisión de la literatura 21](#_Toc534140610)

[5. Marco teórico 29](#_Toc534140611)

[5.1. Predicción 29](#_Toc534140612)

[5.1.1. Pasos para la predicción 29](#_Toc534140613)

[5.1.2. Predicción del rendimiento de cultivos 30](#_Toc534140614)

[5.2. Correlaciones 31](#_Toc534140615)

[5.3. Multicolinealidad 31](#_Toc534140616)

[5.4. Aprendizaje automático 32](#_Toc534140617)

[5.4.1. Aprendizaje no supervisado 33](#_Toc534140618)

[5.4.2. Aprendizaje supervisado 34](#_Toc534140619)

[5.4.3. Validación de modelos 41](#_Toc534140620)

[5.4.4. Métricas de ajuste 45](#_Toc534140621)

[5.4.5. Gráficas de resultado 48](#_Toc534140622)

[6. Metodología 49](#_Toc534140623)

[6.1. Conjunto de datos 49](#_Toc534140624)

[6.1.1. Limpieza del conjunto de datos 50](#_Toc534140625)

[6.2. Análisis de correlaciones 52](#_Toc534140626)

[6.2.1. Análisis de correlación para el conjunto de variables fotosintéticas 52](#_Toc534140627)

[6.2.2. Análisis de correlación para el conjunto de variables morfológicas 53](#_Toc534140628)

[6.2.3. Análisis de correlación para el conjunto de variables características químicas del suelo 54](#_Toc534140629)

[6.2.4. Análisis de correlación para el conjunto de variables características físicas del suelo 55](#_Toc534140630)

[6.2.5. Análisis de correlación para el conjunto de variables climáticas 56](#_Toc534140631)

[6.3. Prueba de KMO y esfericidad de Bartlett 57](#_Toc534140632)

[6.3.1. Prueba de KMO y esfericidad de Bartlett para el conjunto de variables fotosintéticas 58](#_Toc534140633)

[6.3.2. Prueba de KMO y esfericidad de Bartlett para el conjunto de variables morfológicas 59](#_Toc534140634)

[6.3.3. Prueba de KMO y esfericidad de Bartlett para el conjunto de características químicas del suelo 59](#_Toc534140635)

[6.3.4. Prueba de KMO y esfericidad de Bartlett para el conjunto de características físicas del suelo 60](#_Toc534140636)

[6.3.5. Prueba de KMO y esfericidad de Bartlett para el conjunto de variables climáticas 61](#_Toc534140637)

[6.4. Selección de variables explicativas 61](#_Toc534140638)

[6.5. Distribución de variable respuesta 62](#_Toc534140639)

[6.5.1. Prueba de hipótesis para la distribución de la variable dependiente 62](#_Toc534140640)

[6.5.2. Histograma para la distribución de la variable dependiente 63](#_Toc534140641)

[6.6. Construcción de modelos con técnicas de Aprendizaje Automático 64](#_Toc534140642)

[6.6.1. Modelo Lineal Generalizado 64](#_Toc534140643)

[6.6.2. Máquinas de Soporte Vectorial 68](#_Toc534140644)

[6.7. Validación de modelos con métricas de ajuste 70](#_Toc534140645)

[7. Resultados 73](#_Toc534140646)

[7.1. Resultados modelos con Aprendizaje Automático 73](#_Toc534140647)

[7.1.1. Modelo Lineal Generalizado 73](#_Toc534140648)

[7.1.2. Máquinas de Soporte Vectorial 77](#_Toc534140649)

[7.2. Comparación de modelos GLM y SVM 80](#_Toc534140650)

[7.3. Discusión de resultados 85](#_Toc534140651)

[8. Conclusiones 87](#_Toc534140652)

[9. Recomendaciones 90](#_Toc534140653)

[10. Referencias Bibliográficas 92](#_Toc534140654)

[11. Apéndices 99](#_Toc534140655)

**Lista de Tablas**

[Tabla 1. Cumplimiento de objetivos 14](#_Toc534022441)

[*Tabla 2. Enlaces canónicos comúnmente utilizados.* 36](#_Toc534022442)

[Tabla 3. *Funciones integradas de kernel.* 38](#_Toc534022443)

[*Tabla 4. Prueba de KMO para conjunto de variables fotosintéticas* 57](#_Toc534022444)

[*Tabla 5. Prueba de KMO para conjunto de variables morfológicas* 58](#_Toc534022445)

[*Tabla 6. Prueba de KMO para conjunto de características químicas del suelo* 58](#_Toc534022446)

[Tabla 7. *Prueba de KMO para conjunto de características físicas del suelo* 59](#_Toc534022447)

[Tabla 8. *Prueba de KMO para conjunto de variables climáticas* 60](#_Toc534022448)

[Tabla 9. Modelo Lineal Generalizado con todas las variables 64](#_Toc534022449)

[Tabla 10. Modelo Lineal Generalizado con las variables seleccionadas 66](#_Toc534022450)

[Tabla 11. Resultados modelo GLM 72](#_Toc534022451)

[Tabla 12. Métricas de ajuste para modelo GLM 74](#_Toc534022452)

[Tabla 13. Resultados Modelo Máquinas de Soporte Vectorial 77](#_Toc534022453)

[Tabla 14 Métricas de ajuste para el Modelo de Máquinas de Soporte Vectorial 78](#_Toc534022454)

[Tabla 15. RMSE Modelos 79](#_Toc534022455)

[Tabla 16. Comparación RMSE modelos 84](#_Toc534022456)

**Lista de Figuras**

[*Figura. 1*. Modelos lineales generalizados. Adaptado de Modelos lineales generalizados (GLM) Cayuela (2010). 30](#_Toc534022497)

[*Figura. 2.*  Métodos de aprendizaje automático. Adaptado de Introducing Machine Learning Math Works (2016). https://www.mathworks.com/tagteam/89703\_92991v00\_machine\_learning\_section1\_ebook\_v12.pdf 33](#_Toc534022498)

[Figura. 3. Relación entre la media y la varianza de los datos bajo distintos supuestos. Adaptado de Modelos lineales generalizados (GLM) (Cayuela, 2010). 36](#_Toc534022499)

[*Figura. 4.* Pasos en los Modelos Lineales Generalizados. Adaptado de Generalized Linear Models for Insurance Data. Jong & Heller (2008) 38](#_Toc534022500)

[Figura. 5 . Validación cruzada K-fold de 5 veces, 30 muestras. Adaptado de ProClassify User's Guide (2006) 43](#_Toc534022501)

[Figura. 6. División de datos. Adaptado de Técnicas ML en Medicina Cardiovascular 44](#_Toc534022502)

[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 (2018) 51](#_Toc534022503)

[Figura. 8. Matriz de correlación del conjunto de variables fotosintéticas 53](#_Toc534022504)

[Figura. 9. Matriz de correlación del conjunto de variables morfológicas 54](#_Toc534022505)

[Figura. 10. Matriz de correlación del conjunto de características químicas del suelo 55](#_Toc534022506)

[Figura. 11. Matriz de correlación del conjunto de características físicas del suelo 56](#_Toc534022507)

[Figura. 12. Matriz de correlación del conjunto de variables climáticas 57](#_Toc534022508)

[Figura. 13. Variables explicativas seleccionadas 62](#_Toc534022509)

[Figura. 14. Distribución de la variable respuesta. Adaptado del software Minitab. 63](#_Toc534022510)

[Figura. 15. Histograma distribución variable respuesta. 64](#_Toc534022511)

[Figura. 16. Selección de mejores parámetros para modelo SVM 70](#_Toc534022512)

[Figura. 17. Gráfico de dispersión de pronósticos. Adaptado del lenguaje Python. 76](#_Toc534022513)

[Figura. 18. Gráfico de residuales. Adaptado del lenguaje Python. 77](#_Toc534022514)

[Figura. 19. Variables significativas identificadas en el modelo GLM 86](#_Toc534022515)

[Figura. 20. Variables significativas identificadas en el modelo SVM 86](#_Toc534022516)

**Lista de Apéndices**

**Resumen**

**Título del proyecto:** Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado[[1]](#footnote-1)\*.

**Autores:** Andrea Carolina Gamboa Ariza

Paula Andrea Cáceres Ortiz[[2]](#footnote-2)\*\*

**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, 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, usando el Modelo Lineal Generalizado (GLM) y las Máquinas de Soporte Vectorial (SVM) que ayudan a identificar las variables más influyentes en el rendimiento de cultivos de cacao las cuales son: Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas como variables representativas para predecir el rendimiento del cacao.

**Abstract**

**Project tittle:** Predictive model for the yield of cocoa crops in Santander using Supervised Machine Learning [[3]](#footnote-3)\*.

**Authors:** Andrea Carolina Gamboa Ariza

Paula Andrea Cáceres Ortiz[[4]](#footnote-4)\*\*

**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, 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) that help to identify the most influential variables in the performance of cocoa crops which are: Diameter of the trunk, Phosphorus (P), Magnesium ( Mg),% Sand,% Hum / Grav, Radiation, Temperature, Humidity and Rains accumulated as representative variables to predict the yield of cocoa.

# Introducción

La agricultura es una de las actividades de mayor contribución al crecimiento económico de la población en Colombia. Esta actividad económica es una de las que más aporta al crecimiento del PIB, gracias al fuerte ritmo que tomó durante el 2017, fue el sector que lideró el porcentaje con una variación de 4,9% *DANE (2018*). Para 2018, el aporte de la agricultura al PIB no se quedó atrás pues 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-Sociedad de Agricultores de Colombia, n.d.)

Este importante crecimiento en los cultivos de cacao contribuye positivamente al 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 (mencionar algunos), 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 D |

# 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[[5]](#footnote-5) (>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).

De lo mencionado anteriormente se evidencia la necesidad de apoyar mediante métodos, modelos enfocados al crecimiento en estas áreas.

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[[6]](#footnote-6) 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 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 que utilizan datos dentro de sus modelos podría ser explicada a partir del análisis que realizó el DNP[[7]](#footnote-7) 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).

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.

# 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.

# Objetivos

## Objetivo General

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

## 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.

# Revisión de la literatura

En la búsqueda 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).

“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” (Hammer et al., 2001) y que 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, Rochberg & Wada 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. 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[[8]](#footnote-8) 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”.

El análisis demuestra que el 84%de la varianza en el área sembrada en septiembre-diciembre y el 81% de la varianza 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 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[[9]](#footnote-9) (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 a 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).

Tack et al. (2015) También utilizaron modelos de regresión lineal para analizar el efecto del tiempo en el rendimiento del trigo mediante datos 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 su estudio en 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.

En la más reciente investigación, (Cunha et al., 2016) implementaron 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 (climáticas, no climáticas, morfología de la planta) y modelos para el pronóstico de rendimientos de los diferentes cultivos, aunque la mayoría de los modelos utilizados por los investigadores son de regresión lineal, se percibe de las últimas investigaciones la implementación otro tipo de modelos (Máquinas de soporte vectorial, Redes neuronales…) los cuales permiten comparar sus resultados con el ánimo de elegir el modelo más acertado en el pronóstico de rendimientos y de esta manera aumentar su precisión.

# Marco teórico

## 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.).

### 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 lineales generalizados. Adaptado de Modelos lineales generalizados (GLM) Cayuela (2010).

### Predicción del rendimiento de cultivos

El rendimiento, 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 en los planificadores agrícolas estimar una predicción para el rendimiento de los cultivos, los cuales ayuden 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. “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” (Gonzalez-Sanchez, 2014).

## Correlaciones

Una correlación es una técnica para analizar la 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.

## 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.

## Aprendizaje automático

El aprendizaje automático o “Machine Learning es una rama de la inteligencia artificial que proporciona métodos con la capacidad de aprender o hacer predicciones sobre datos. Estos métodos crean un modelo a partir de entradas de ejemplo para hacer predicciones o tomar decisiones” (Mitchell, 1997).

“ML no hace suposiciones sobre la estructura correcta del modelo de datos, lo que permite la construcción de modelos complejos” (Díaz, Mazza, Combarro, Gimenez & Gaid, 2017).

Los algoritmos de aprendizaje automático encuentran patrones naturales en los datos que generan información y ayudan a tomar mejores decisiones y predicciones (Baratta, 2016). 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.

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). https://www.mathworks.com/tagteam/89703\_92991v00\_machine\_learning\_section1\_ebook\_v12.pdf

### 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 preferencia del diseñador. 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)

### Aprendizaje supervisado

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).

Se le llama al aprendizaje supervisado debido a la presencia de una variable resultado para guiar el proceso de aprendizaje. A diferencia del no supervisado, donde solo se observan las características y no se tiene en cuenta las mediciones del resultado.

Cada objeto en el conjunto de datos viene con una etiqueta de clase preasignada. La labor en este tipo de aprendizaje es entrenar a un clasificador para que haga el etiquetado. Con mucha frecuencia, el proceso de etiquetado no se puede describir en una forma algorítmica. Entonces se le proporciona a la máquina las habilidades de aprendizaje y a su vez se le presentan los datos etiquetados (Kuncheva, 2004).

El aprendizaje supervisado usa técnicas de clasificación y regresión para desarrollar modelos predictivos (Introducing Machine Learning, n.d.) (Baratta, 2016).

• Clasificación: Esta técnica clasifica los datos de entrada en categorías. Las técnicas de clasificación predicen respuestas discretas.

• Regresión: Las técnicas de regresión predicen respuestas continuas. Es decir, para predecir valores numéricos desconocidos.

#### 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. 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)

(1)

Ciertos tipos de variables respuesta sufren invariablemente la violación de estos dos supuestos de los modelos normales y los GLM ofrecen una buena alternativa para tratarlos (Cayuela, 2010)

El supuesto central que se ha hecho hasta el momento con los modelos lineales es que la varianza es constante (Figura 1a). En el caso de los conteos, sin embargo, donde la variable respuesta está expresada en números enteros y en dónde hay a menudo muchos ceros en los datos, la varianza podría incrementar linealmente con la media (Figura 1b). 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 a la media (Figura 1c). Cuando la variable respuesta siga una distribución Gamma, entonces la varianza incrementa de una manera no lineal con la media (Figura 1d).

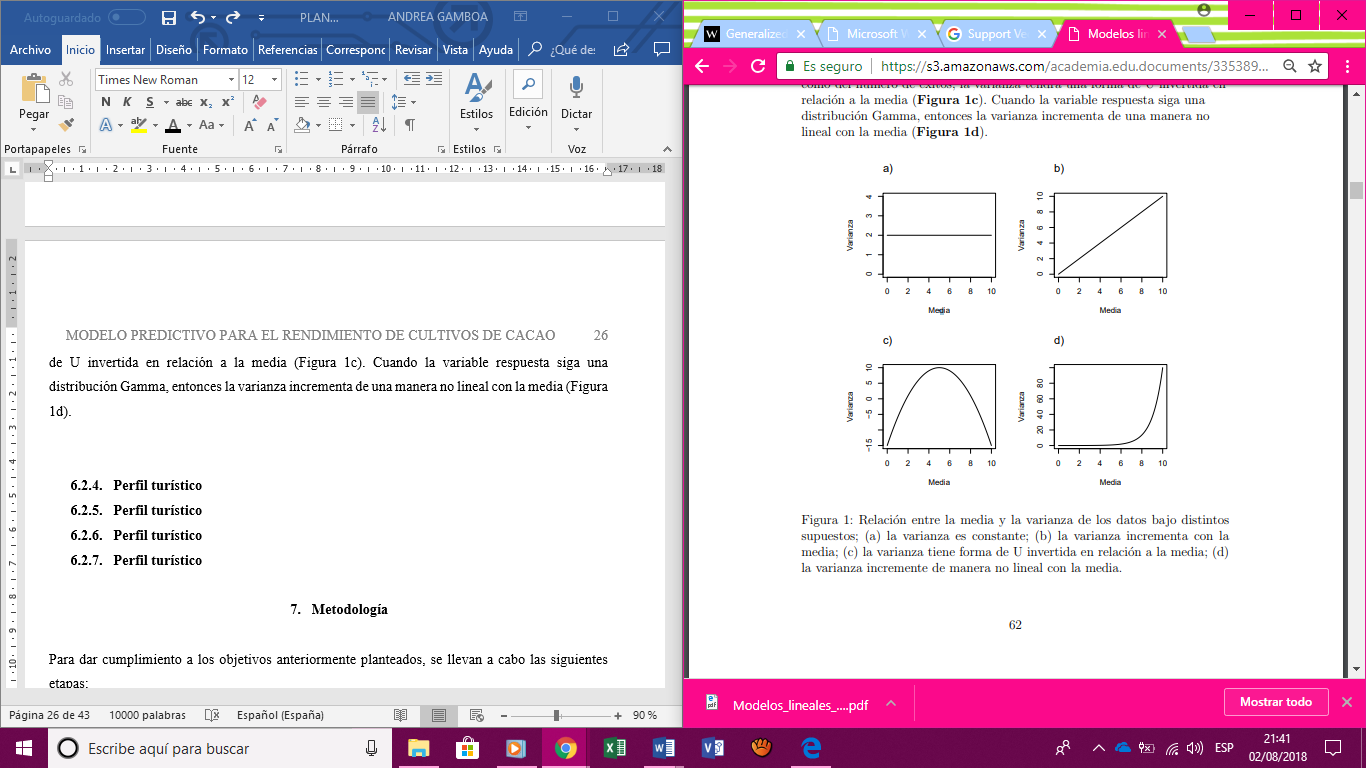


Figura. 3. Relación entre la media y la varianza de los datos bajo distintos supuestos. Adaptado de Modelos lineales generalizados (GLM) (Cayuela, 2010).

##### Función enlace.

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

(2)

Donde g es la función diferenciable; x, las variables independientes y 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 1, se muestran los enlaces canónicos más comunes para las distribuciones de la variable respuesta.

*Tabla 2. Enlaces canónicos comúnmente utilizados.*

|  |  |  |
| --- | --- | --- |
| Función enlace |  | Enlace canónico |
| Identidad |  | Normal |
| Log | ln | Poisson |
| Inverso |  | Gamma(p=1) |
| Raíz cuadrada |  |  |
| Logit |  | Binomial |

*Nota: 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

**→** (3)

##### Pasos para el GLM.

Dada una variable de respuesta , la construcción de un GLM consiste en los siguiente pasos (Jong & Heller, 2008):

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

#### 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 (Kevin, Hoz, Martínez, & Mendoza, n.d.). 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 aprendizaje 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 |  |
| Lineal |  |
| Gaussiano |  |
| Polinomial | donde q se encuentra en el conjunto {2, 3,...} |

*Nota: Adaptado de Pattern Recognition and Machine Learning (2012)*

La base biológica de las máquinas de soporte (SVM) es una clasificación que permite maximizar el espacio en blanco en ambos lados del hiperplano para garantizar la precisión de la clasificación. 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):

(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). Con la condición de permitir la existencia de un error de ajuste, se pueden obtener minimizando la fusión de riesgo 𝑀𝜔,∗ con las variables positivas de holgura (ζ, ζ\*) introducidas en ella como se muestra en la ecuación (5).

(5)

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 (𝛼𝑗−𝛼𝑗∗) se introducen para resolver el problema dual. Se puede obtener el mejor hiperplano de regresión representado en la ecuación (6).

(6)

Donde 𝛼𝑗 𝑦 𝛼𝑗∗ satisfacen la igualdad de 𝛼𝑗×𝛼𝑗∗=0, 𝛼𝑗 ≥0, 𝛼𝑗∗ ≥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.

(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.).

### 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).

#### 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, se 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.

#### 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 aproximador de funciones se ajusta a una función utilizando solo el conjunto de entrenamiento. Luego, se le pide a la función de aproximación que prediga los valores de salida para los datos en el conjunto de prueba. Los errores que comete se acumulan como antes para dar el error de conjunto de prueba absoluto medio, que se utiliza para evaluar el modelo. ( 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).

### Métricas de ajuste

Para la misma tarea, puede haber varios algoritmos y podemos estar interesados 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.

#### Coeficiente de determinación (R2)

Los R2 como medidas de bondad de ajuste son muy populares. Son números que se pueden obtener a partir de un modelo ajustado. 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 el valor estimado de la variable dependiente , se llama i-ésimo residual. Representa el error que se comete al usar para estimar (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).

(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 , llamado suma total de cuadrados (SST) mostrado en la ecuación (9):

(9)

Para saber cuánto se desvían los valores de medidos en la línea de regresión, de los valores de , 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

(10)

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

(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 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 R2 tal y como se muestra en la ecuación (12)

(12)

Expresando este valor como un porcentaje, se puede interpretar a R2 como el porcentaje de la variación de los valores de la variable dependiente que se puede explicar con la ecuación de regresión. (Levin & Rubin, 2004)

#### Error Cuadrático Medio (RMSE)

El Error Cuadrático Medio realiza 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)

(13)

Donde:

* + - * : Es el valor real
      * : Es el valor estimado
      * N: Es el tamaño de la muestra

#### 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)

(14)

Donde:

* + - * : Es el valor real
      * : Es el valor estimado
      * N: Es el tamaño de la muestra

### Gráficas de resultado

En el caso de los modelos, cuando se utiliza una regresión para predecir el valor a partir de un valor en x de un determinado sujeto xi, 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.

#### 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.

#### Gráfico de residuales

Los residuos se utilizan para evaluar la diferencia entre los valores observados y los valores que predice el modelo. Los residuales (Ei) quedan definidos como la diferencia entre un verdadero valor en la variable y ( y su valor predicho como se observa en la expresión (15) Jong & Heller (2008):

(15)

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

# 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.

## Conjunto de datos

El conjunto de datos es el insumo fundamental de todo análisis predictivo, el cual es necesario para encontrar 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. El primero de ellos cuenta con 3 niveles: fertilización al 50%, 100%, 150%. En segundo lugar, se tienen los 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). Y por último, la exposición del cultivo es a sol o a sombra. Considerando estos factores y niveles se cuenta con un total de 60 tratamientos. A cada uno de los tratamientos mencionados anteriormente, le corresponden 15 plantas, por lo que se tiene un total de 900 de ellas a las cuales se les miden sus características fotosintéticas, morfológicas, físicas y químicas del suelo. De igual manera, el cacao producido por estas plantas se evalúa en términos de altura de la planta, número de ramas, diámetro del tronco, peso de la almendra, peso de la mazorca completa, entre otros. En primer lugar, de las características fotosintéticas fueron tomadas 3 muestras, 1 vez al año durante el periodo de 2015-2017 para los 60 tratamientos, para un total de 540 observaciones. 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

### 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 6.

En primera instancia, para el tratamiento de datos perdidos, se realiza una imputación de datos mediante la media. Posteriormente, para reducir la dimensionalidad de las variables, se realiza una normalización de los datos utilizando las variables independientes; que son las que conforman los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas, y teniendo como variable a predecir el rendimiento del cultivo de cacao, obteniendo así finalmente el Dataset.



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 (2018)

## Análisis de correlaciones

Según el autor Faraway (2004), a la hora de especificar el modelo hay que evitar los problemas de multicolinealidad entre todas las variables. Una forma de evitar estos problemas es no incluir variables relacionadas, ya que pueden incrementar la varianza y los coeficientes de regresión dentro de los modelos. Es por esto que con ayuda de la función *corcoef*, del lenguaje de programación Python, se realizó un análisis de correlación como técnica para evaluar la asociación lineal entre variables para 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.

### Análisis de correlación para el conjunto de variables fotosintéticas

En la figura 7, 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 entre sí determinada por su alto coeficiente de correlación.

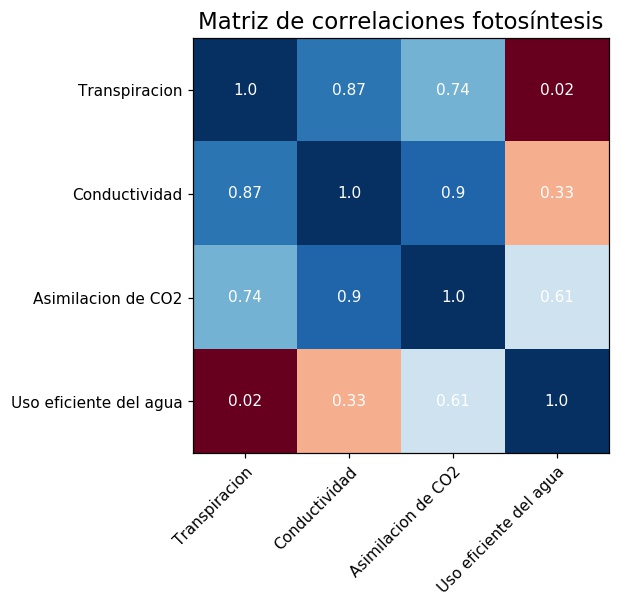


Figura. 8. Matriz de correlación del conjunto de variables fotosintéticas

### Análisis de correlación para el conjunto de variables morfológicas

En la figura 8, 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.

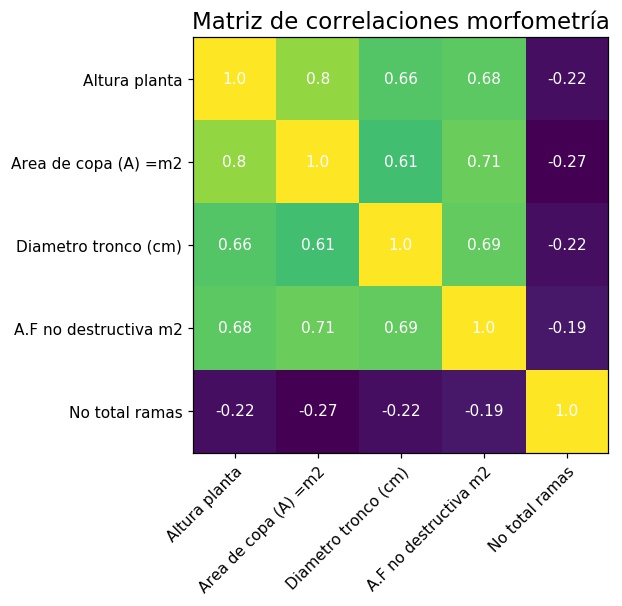


Figura. 9. Matriz de correlación del conjunto de variables morfológicas

### Análisis de correlación para el conjunto de variables características químicas del suelo

En la figura 9, 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.

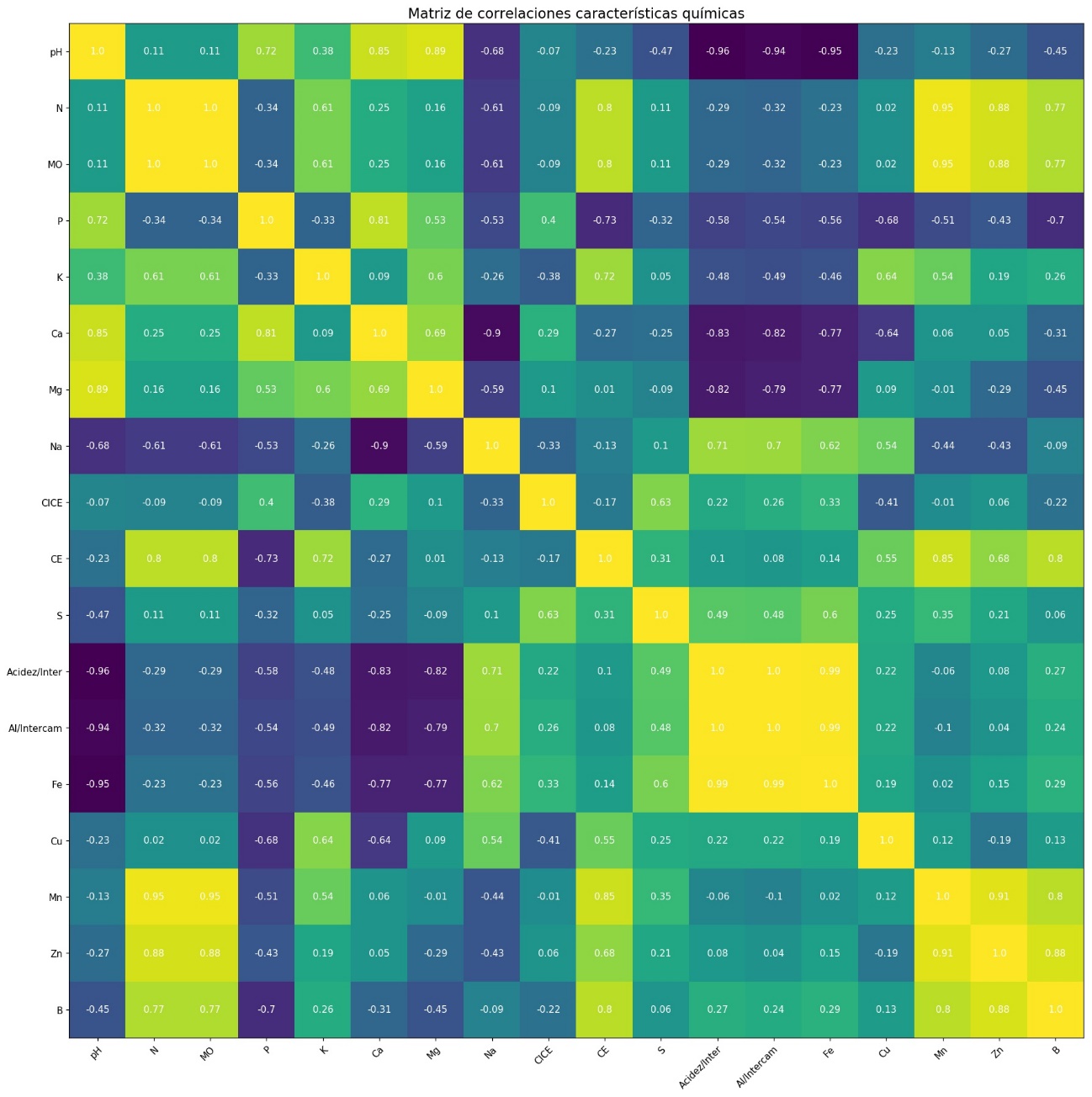


Figura. 10. Matriz de correlación del conjunto de características químicas del suelo

### Análisis de correlación para el conjunto de variables características físicas del suelo

De la figura 10, se observa la matriz de correlación para el conjunto de datos de características físicas del suelo. Allí, se observa que las variables “%Arcilla (%A)” 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)”, “%Arena(%A), “DA”, “DR” y “%Poros”.

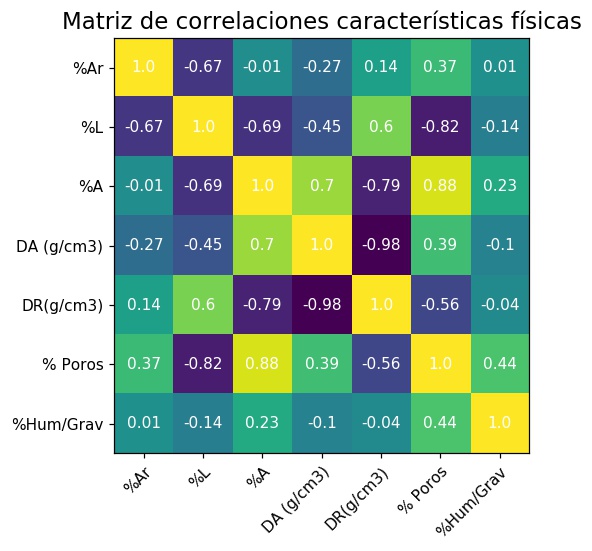


Figura. 11. Matriz de correlación del conjunto de características físicas del suelo

### Análisis de correlación para el conjunto de variables climáticas

En la figura 11, 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.

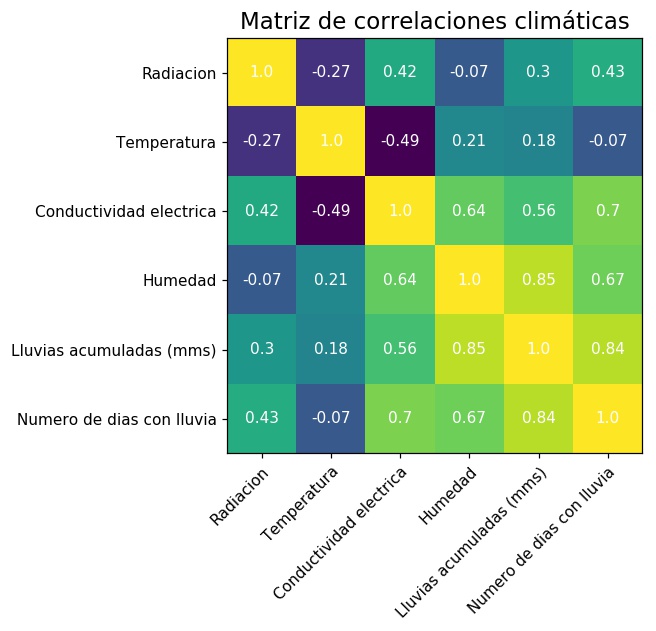


Figura. 12. Matriz de correlación del conjunto de variables climáticas

## 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 las siguientes subsecciones.

### Prueba de KMO y esfericidad de Bartlett para el conjunto de variables fotosintéticas

*Tabla 4. Prueba de KMO para conjunto de variables fotosintéticas*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| KMO |  | ,500 |
| Prueba de esfericidad de Bartlett | Aprox. Chi-cuadrado  Gl  Significancia | ,603  1  ,437 |

*Nota: Adaptado de Software SPSS Statistics.*

En la tabla 4 se muestran los valores obtenidos a partir de la prueba, donde se observa que el KMO es de 0,5; lo cual indica que existe una baja correlación entre las variables explicativas del conjunto de variables fotosintéticas.

### Prueba de KMO y esfericidad de Bartlett para el conjunto de variables morfológicas

*Tabla 5. Prueba de KMO para conjunto de variables morfológicas*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| KMO |  | ,500 |
| Prueba de esfericidad de Bartlett | Aprox. Chi-cuadrado  Gl  Significancia | 255  1  ,100 |

*Nota: Adaptado de Software SPSS Statistics.*

En la tabla 5 se muestran los valores obtenidos a partir de la prueba, donde se observa que el KMO es de 0,5; lo cual indica que existe una baja correlación entre las variables explicativas del conjunto de variables morfológicas.

### Prueba de KMO y esfericidad de Bartlett para el conjunto de características químicas del suelo

*Tabla 6. Prueba de KMO para conjunto de características químicas del suelo*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Medida | Valor |
| KMO |  |  | ,325 |
| Prueba de esfericidad de Bartlett |  | Aprox. Chi-cuadrado  Gl  Significancia | 218  6  ,100 |

*Nota: Adaptado de Software SPSS Statistics.*

En la tabla 6 se muestran los valores obtenidos a partir de la prueba, donde se observa que el KMO es de 0,325; lo cual indica que existe una baja correlación entre las variables explicativas del conjunto de variables seleccionadas para las características químicas del suelo.

### Prueba de KMO y esfericidad de Bartlett para el conjunto de características físicas del suelo

Tabla 7. *Prueba de KMO para conjunto de características físicas del suelo*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Medida | Valor |
| KMO |  |  | ,500 |
| Prueba de esfericidad de Bartlett |  | Aprox. Chi-cuadrado  Gl  Significancia | 21  1  ,100 |

*Nota: Adaptado de Software SPSS Statistics.*

En la tabla 7 se muestran los valores obtenidos a partir de la prueba, donde se observa que el KMO es de 0,5; lo cual indica que existe una baja correlación entre las variables explicativas del conjunto de características físicas del suelo.

### Prueba de KMO y esfericidad de Bartlett para el conjunto de variables climáticas

Tabla 8. *Prueba de KMO para conjunto de variables climáticas*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Medida | Valor |
| KMO |  |  | ,325 |
| Prueba de esfericidad de Bartlett |  | Aprox. Chi-cuadrado  Gl  Significancia | 81  6  ,100 |

*Nota: Adaptado de Software SPSS Statistics.*

En la tabla 8 se muestran los valores obtenidos a partir de la prueba, donde se observa que el KMO es de 0,325; lo cual indica que existe una baja correlación entre las variables explicativas del conjunto de variables seleccionadas para el conjunto de variables climáticas.

## 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 13

Figura. 13. Variables explicativas seleccionadas

## 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 esta sigue.

### 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.

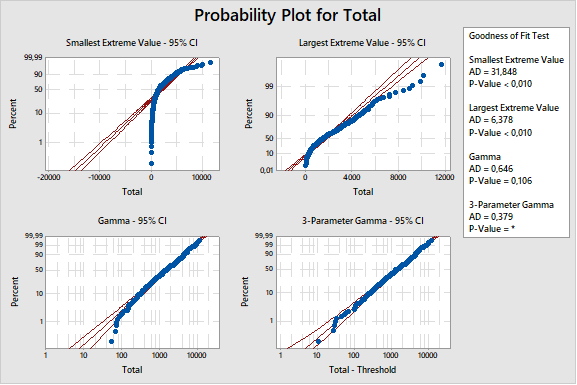


Figura. 14. Distribución de la variable respuesta. Adaptado del software Minitab.

En la figura 14 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, rendimiento de cultivo de cacao, es te tipo Gamma.

### Histograma para la distribución de la variable dependiente

En la figura 15, se muestra el histograma para la distribución de la variable dependiente realizado en el lenguaje de programación Python, donde se observa y una distribución con variables no negativas, sesgadas a la derecha y con grandes valores en la cola superior; característico de una distribución tipo Gamma.

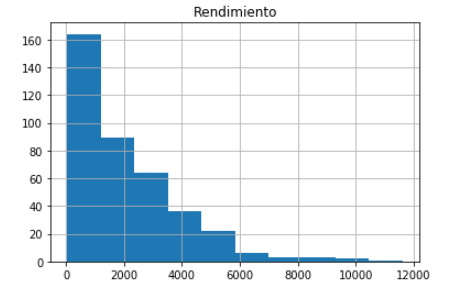


Figura. 15. Histograma distribución variable respuesta.

## 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 la construcción de modelos en este campo de investigación.

### 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.

Dentro de la construcción para los resultados del Modelo Lineal Generalizado, se define dentro del código de programación las variables explicativas (x), la variable a predecir (y). A su vez, se le indica el comportamiento tipo Gamma para la variable respuesta y la función enlace tipo log *sm.genmod.families.links.log.* El cual arroja los valores de los coeficientes para las variables explicativas, el error estándar, el valor p y su respectivo intervalo tal y como se muestra en la tabla 9.

Tabla 9. 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 |
| *%A* | -0.2185 | 0.046 | -4.777 | 0.000 | -0.308 | -0.129 |
| *%Hum/Grav* | 0.4094 | 0.084 | 4.847 | 0.000 | 0.244 | 0.575 |
| *Radiación* | 6.091e-07 | 1.09e-07 | 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 |

*Nota: Adaptado del lenguaje de programación Python.*

Según los autores Jong & Heller (2008) los valores estimados de β se utilizan para definir las variables explicativas 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 para cada variable explicativa. Teniendo en cuenta que la función que define al GLM es la que se muestra en la ecuación (16)

(16)

La cual está constituida por un término de error: ε, variables explicativas: , y coeficientes de regresión: β. Se plantea la siguiente prueba de hipótesis dentro del código de programación:

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

H1: | β | ≠ 0, el coeficiente de la variable respuesta es diferente a cero (es significativa dentro del modelo)

Donde | β | es el coeficiente de regresión de cada variable explicativa.

Considerando que si el P>|Z| o también llamado p-valor mostrado en la tabla 9 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.

Dado lo anterior y a partir de la tabla 9, se identifica que las variables que tengan un p-valor por debajo de 0,05 son aquellas que cuentan con un coeficiente de regresión diferente de cero y en consecuencia, aquellas que tienen mayor incidencia en la variable respuesta. Estas variables se almacenan en otra lista diferente. Para esto, se le indicó al Python que almacenara las variables en la lista llamada *varSelec*. Las variables identificadas como más significativas para el modelo fueron: Diámetro tronco (cm), Fósforo (P), Magnesio (Mg), %Arcilla, %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 y almacenadas previamente en *varSelec.*

Los resultados obtenidos después de entrenar nuevamente el modelo se muestran en la tabla 10.

Tabla 10. Modelo Lineal Generalizado con las variables seleccionadas

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Coef | Std err | z | P>|z| | [0.025 | 0.975] |
| *Diámetro del tronco* | 0.0610 | 0.017 | 3.583 | 0.000 | 0.028 | 0.094 |
| *P* | 0.9988 | 0.174 | 5.757 | 0.000 | 0.659 | 1.339 |
| *Mg* | 2.6765 | 0.394 | 6.798 | 0.000 | 1.905 | 3.448 |
| *%A* | -0.2225 | 0.047 | -4.707 | 0.000 | -0.315 | -0.130 |
| *%Hum/Grav* | 0.3952 | 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* | -0.2980 | 0.059 | -5.024 | 0.000 | -0.414 | -0.182 |
| *Humedad* | -35.4404 | 4.997 | -7.092 | 0.000 | -45.235 | -25.646 |
| *Lluvias Acumuladas* | 0.0088 | 0.001 | 7.019 | 0.000 | 0.006 | 0.011 |

*Nota: Adaptado del lenguaje de programación Python.*

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

### 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 especificando con el código *import SVR.* En primera instancia, se importó el conjunto de datos previamente procesado y consolidado. Es importante tener en cuenta que, para este modelo, al no ser de tipo lineal, 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 a su vez 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: prueba y 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. En primer lugar, se realizó la búsqueda utilizando un *kernel lineal*, obteniendo los mejores parámetros con este tipo de kernel. Posteriormente, se realizó el mismo procedimiento, pero esta vez con un *kernel rbf* donde se obtuvieron otros valores para los parámetros C y Gamma.

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 serán utilizados al momento de entrenar y validar el modelo para general los resultados finales. Los parámetros seleccionados dentro del modelo se presentan en la figura 16.

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



Figura. 16. Selección de mejores parámetros para modelo SVM

## 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, la cual 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. En las siguientes subsecciones, se describe la validación que se realizó para cada uno de los modelos.

Como se mencionó previamente, el conjunto de datos se divide 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.

#### Métricas de ajuste para los modelos

* + - * Coeficiente de determinación (R2)

Para determinar el Coeficiente de Determinación R2 de los modelos GLM y SVM, se utiliza la librería *sklearn.metrics* del lenguaje de programación Python, donde se importa el coeficiente de determinación r2\_score. Este se determina con cada uno de los valores del conjunto de prueba predichos para el rendimiento de cultivos de cacao (y) de cada uno de los valores de las variables independientes (x).

* + - * Error Cuadrático Medio (RMSE)

Para determinar el Error Cuadrático Medio de los modelos GLM y SVM, se utiliza la librería sklearn.metrics del lenguaje de programación Python, donde se importa el código *RMSE.append***.** Para determinar este error, se realiza una diferencia entre cada uno de los valores estimados y los valores reales del conjunto de prueba como se muestra en la ecuación (17)

(17)

* + - * Error Absoluto Medio (MAE)

Para determinar el Error *Absoluto* Medio *de los modelos GLM y SVM,* se utiliza la librería sklearn.metrics del lenguaje de programación Python, donde se importa *el código* MAE.append. Para determinar este error, se realiza una diferencia entre los valores predichos del rendimiento de cultivos de cacao y los valores reales del mismo para el conjunto de prueba y luego se le calcula el promedio a estos valores como se muestra en la ecuación 18

(18)

#### 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: | RMSEGLM | = | RMSESVM |, 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: | RMSEGLM | ≠ | RMSESVM |, 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 (19)

(19)

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.

# Resultados

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

## Resultados modelos con Aprendizaje Automático

### Modelo Lineal Generalizado

Posterior a la programación realizada anteriormente y descrita en la subsección 6.6.1, se procede a generar e interpretar los resultados que arrojó el Modelo Lineal Generalizado, los cuales se presentan en la tabla 11.

Tabla 11. Resultados modelo GLM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Coef | Std err | z | P>|z| | [0.025 | 0.975] |
| *1. Diámetro del tronco* | 0.0610 | 0.017 | 3.583 | 0.000 | 0.028 | 0.094 |
| *2. P* | 0.9988 | 0.174 | 5.757 | 0.000 | 0.659 | 1.339 |
| *3. Mg* | 2.6765 | 0.394 | 6.798 | 0.000 | 1.905 | 3.448 |
| *4. %A* | -0.2225 | 0.047 | -4.707 | 0.000 | -0.315 | -0.130 |
| *5. %Hum/Grav* | 0.3952 | 0.086 | 4.586 | 0.000 | 0.226 | 0.564 |
| *6. Radiación* | 5.073e-07 | 8.64e-08 | 5.869 | 0.000 | 3.38e-07 | 6.77e-07 |
| *7. Temperatura* | -0.2980 | 0.059 | -5.024 | 0.000 | -0.414 | -0.182 |
| *8. Humedad* | -35.4404 | 4.997 | -7.092 | 0.000 | -45.235 | -25.646 |
| *9. Lluvias Acumuladas* | 0.0088 | 0.001 | 7.019 | 0.000 | 0.006 | 0.011 |

*Nota: Adaptado del lenguaje de programación Python.*

Tal y como se mencionó anteriormente, el entrenamiento del modelo para los resultados se realizó con el 80% del total del conjunto de datos, y a su vez, con las variables explicativas previamente identificadas con un p-valor por debajo de 0,05; es decir, aquellas que cuentan con un coeficiente de regresión diferente de cero y, en consecuencia, aquellas que tienen mayor incidencia en la variable respuesta.

En la expresión (20), se presenta la ecuación que corresponde al modelo GLM.

(20)

Por sus coeficientes positivos , 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 (%A), Temperatura y Humedad; las cuales cuentan con un coeficiente negativo, pueden estar impactando de manera negativa al 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, en la cual 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.

#### 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. La validación para este modelo se realizó con el 20% de los datos, el cual había sido previamente dividido del conjunto total de los datos y predestinado para este fin.

La técnica de validación genera un valor por cada repetición para cada una de las métricas de ajuste(R2, RMSE y MAE), en donde se realizan 100 repeticiones de validación para cada una de las métricas de ajuste, para un total de 100 valores de cada una de ellas. Finalmente, se genera un valor promedio derivado de estas 100 repeticiones, el cual se presenta en la tabla 12.

Tabla 12. Métricas de ajuste para modelo GLM

|  |  |
| --- | --- |
| Métrica de ajuste | Valor |
| R2 | 0,1319 |
| RMSE | 1708,2904 |
| MAE | 1028,7805 |

*Nota: Adaptado del lenguaje de programación Python.*

#### 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 17 se observa el gráfico de dispersión de los pronósticos.

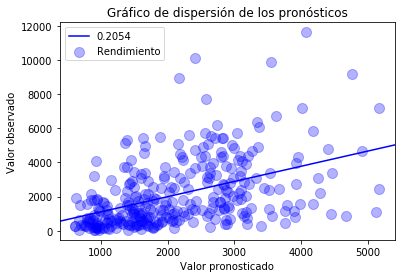


Figura. 17. Gráfico de dispersión de pronósticos. Adaptado del lenguaje Python.

Del gráfico de dispersión de pronósticos 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.

Por otro lado, en la figura 18, se presenta el gráfico de residuales

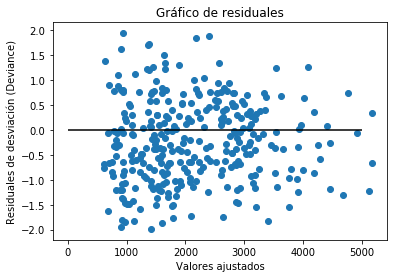


Figura. 18. Gráfico de residuales. Adaptado del lenguaje Python.

Del gráfico de residuales se observa que se tienen una dispersión de residuos al azar, es decir, no tiene una tendencia definida, lo que indica que los valores ajustados no se desvían del comportamiento de los valores observados.

### 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 (21)

(21)

Después de realizar la programación utilizada para la construcción del modelo de Máquinas de Soporte Vectorial descrito en la subsección 6.6.2, en la tabla 13 se presentan los resultados de este modelo para su correspondiente interpretación.

Tabla 13. Resultados Modelo Máquinas de Soporte Vectorial

|  |  |
| --- | --- |
| *Variable* | *Coeficiente* |
| *1. Diámetro tronco* | 439.87306816 |
| *2. P* | 2267.7076112 |
| *3. Mg* | 904.00337998 |
| *4. %A* | -1494.35760507 |
| *5. %Hum/Grav* | 1124.02448836 |
| *6. Radiación* | 621.29006298 |
| *7. Temperatura* | -555.84730059 |
| *8. Humedad* | -1823.33064068 |
| *9. Lluvias acumuladas* | 1760.9283604 |

*Nota: Adaptado del lenguaje de programación Python*

De igual manera que para el modelo GLM, el entrenamiento de este modelo también se realizó con el 80% del conjunto de datos, así mismo las variables de entrada para el desarrollo del modelo SVM siguen siendo las mismas para la posterior comparación de los dos modelos. En el desarrollo del Modelo Lineal Generalizado se hizo el correspondiente análisis para la selección de las variables más influyentes en la variable respuesta.

Como se observa en la tabla 13, 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 puede que sea desalentador debido a que estas estarían impactando de manera negativa sobre él.

#### 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 14.

Tabla 14 Métricas de ajuste para el Modelo de Máquinas de Soporte Vectorial

|  |  |
| --- | --- |
| Métrica de ajuste | Valor |
| R2 | 0,1058 |
| RMSE | 1758,7421 |
| MAE | 926,2623 |

*Nota: 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 de cada modelo, presentados en la tabla 13, con el fin de comparar cuál de los dos modelos contaba con menor error, es decir, con mayor precisión.

Tabla 15. RMSE Modelos

|  |  |
| --- | --- |
| GLM | SVM |
| 1716,91 | 2092,47 |
| 1587,74 | 1757,73 |
| 1785,81 | 1784,31 |
| 1579,53 | 1453,99 |
| 1521,36 | 1521,29 |
| 1688,93 | 1569,92 |
| 1453,91 | 1720,92 |
| 1593,85 | 1883,6 |
| 1754,21 | 1815,32 |
| 1824,01 | 1465,12 |
| 1840,94 | 1729,55 |
| 1628,69 | 2107,77 |
| 1774,41 | 1639,58 |
| 1506,06 | 1747,21 |
| 1955,94 | 1667,57 |
| 1957,2 | 1642,06 |
| 1380,98 | 2227,79 |
| 1539,7 | 1584,76 |
| 1616,7 | 1685,22 |
| 1767,38 | 1405,43 |
| 1412,54 | 1909,94 |
| 1630,8 | 1824,96 |
| 1661,04 | 2063,97 |
| 1764,11 | 1469,51 |
| 1799,64 | 1711,29 |
| 1484,54 | 1482,72 |
| 1560,51 | 1708,96 |
| 1447,79 | 1525,05 |
| 1753,89 | 1965,48 |
| 1393 | 1795,58 |
| 1546,46 | 1292,15 |
| 1543,9 | 1883,17 |
| 1655,32 | 1867,05 |
| 1657,65 | 1485,51 |
| 2296,61 | 1644,01 |
| 1543,3 | 1689,03 |
| 1694,87 | 1866,87 |
| 1887,59 | 2054,3 |
| 2086,49 | 1491,72 |
| 1546,01 | 1651,19 |
| 1994,41 | 1968,58 |
| 1457,85 | 1824,26 |
| 1904,75 | 1624,35 |
| 1822,59 | 1950,71 |
| 1637,38 | 2205,8 |
| 1546,18 | 1836,49 |
| 1645,18 | 1774,85 |
| 1455,46 | 1330,4 |
| 1548,67 | 1745,09 |
| 1653,54 | 1875,55 |
| 1572,81 | 1928,29 |
| 1357,11 | 1574,69 |
| 1742,16 | 1595,8 |
| 1626,6 | 1915,66 |
| 1795,75 | 1666,09 |
| 1734,9 | 1557,72 |
| 1958,03 | 1621,23 |
| 1731,74 | 1922,62 |
| 1417,71 | 1806,29 |
| 1618,23 | 1955,09 |
| 1718,83 | 2006,44 |
| 1773,58 | 1316,61 |
| 1505,42 | 1884,61 |
| 1825,92 | 1557,83 |
| 1572,84 | 1422,01 |
| 1648,72 | 1854,22 |
| 2208,38 | 1366,76 |
| 1672,5 | 1382,94 |
| 1985,81 | 1564,61 |
| 1546,17 | 1560,03 |
| 1792,57 | 1830,85 |
| 1724,06 | 2129,08 |
| 1649,88 | 2138,23 |
| 1846,87 | 1849,87 |
| 1747,37 | 1958,71 |
| 1656,7 | 1913,21 |
| 1825,62 | 1799,67 |
| 1780,45 | 1765,55 |
| 1936,98 | 2004,82 |
| 1463,2 | 2031,83 |
| 1427,3 | 1883,59 |
| 1350,04 | 1542,98 |
| 1780,61 | 1465,97 |
| 1882,08 | 1522,53 |
| 1648,11 | 1837,09 |
| 1450,24 | 2258,81 |
| 2157,95 | 1542,33 |
| 1671,55 | 1898,26 |
| 1555,8 | 1549,68 |
| 1518,38 | 2158,45 |
| 1852,99 | 1474,5 |
| 1810,16 | 1718,59 |
| 1547,07 | 1408,98 |
| 1680,92 | 2374,58 |
| 1662,67 | 1912,3 |
| 1712,76 | 1846,9 |
| 1796,77 | 1961,33 |
| 1619,96 | 2080,07 |
| 1521,71 | 1969,14 |
| 1970,17 | 1566,58 |

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 14

Tabla 16. 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 |

*Nota: 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 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 la misma influencia, tanto de manera negativa como de manera positiva, de las variables explicativas que influyen sobre los rendimientos de cultivos de cacao, las cuales se presentan en la figura 19 y 20 respectivamente.

Figura. 19. Variables significativas identificadas en el modelo GLM

Figura. 20. Variables significativas identificadas en el modelo SVM

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. Pero 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.

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 como las 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.

# Conclusiones

* + - * A partir de la revisión de literatura se observa que las herramientas de Aprendizaje Automático están siendo utilizadas cada vez más en el ámbito de la investigación como herramientas acertadas para la predicción de rendimientos de cultivos tales como maíz, trigo, uva, cebada, nuez, soja, entre otros.
      * Las herramientas de Aprendizaje Automático han venido siendo utilizadas en aumento con el paso de los años por los diferentes investigadores, ya que pueden llegar a tener una alta precisión en la predicción siempre y cuando se seleccionen los modelos más apropiados de acuerdo con las características propias de cada investigación.
      * La información que brindan los pronósticos derivados de las investigaciones que utilizan herramientas de Aprendizaje Automático, pueden llegar a ser bastante útiles para que los diferentes actores del sector agrícola puedan reducir la incertidumbre y a su vez puedan comprender las condiciones más influyentes con el fin de tomar las decisiones más acertadas.
      * Los modelos teóricos encontrados en la revisión de literatura podrían estar sujetos a modificaciones y alteraciones según las características propias de cada investigación, de ahí 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.
      * Es de suma importancia contar con un adecuado procesamiento y limpieza de datos para este tipo de investigaciones, debido a que de esto depende la precisión y confiabilidad del instrumento.
      * Previo a la construcción de un modelo, el investigador debe conocer la distribución de los datos observados para una selección adecuada del modelo a construir con herramientas de Aprendizaje Automático.
      * A partir de una distribución de probabilidad de la variable respuesta (rendimiento de cultivo de cacao) diferente a la normal (tipo Gamma), se puede concluir que las dos herramientas usadas: 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.
      * La validación cruzada en los modelos realizados con herramientas de Aprendizaje Automático puede llegar a ser utilizada no solo para evaluar su ajuste y precisión sino también 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.
      * A partir de los resultados que arrojó tanto el Modelo Lineal Generalizado como el de Máquinas de Soporte Vectorial, se identifica 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.
      * Derivado de los resultados de la presente investigación se observa que, si la humedad, temperatura y porcentaje de arena del cultivo de cacao no están presentes en las cantidades adecuadas y moderadas, el rendimiento del cultivo de cacao se va a ver afectado de una manera negativa.
      * De los resultados de la presente investigación se observa que las variables climáticas (radiación, temperatura, humedad y lluvias acumuladas) fueron las variables que explicaron 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; 4 de las 9 variables identificadas por los modelos como significativas.
      * La construcción y comparación de los resultados de los dos modelos: Modelo Lineal Generalizado y Máquinas de Soporte Vectorial, fue útil para ratificar la influencia de las variables explicativas en los rendimientos de cultivos de cacao, puesto que para los dos modelos se identificaron influencias similares dadas por los valores y signos de sus coeficientes.
      * Las validaciones para los modelos realizados con herramientas de Aprendizaje Automático son de vital importancia para evaluar la precisión obtenida para cada uno de los modelos.
      * Con valores de coeficiente de regresión (R2) para el Modelo Lineal Generalizado y el modelo de Máquinas de Soporte Vectorial de 0.1319 y 0.1058 respectivamente, se puede concluir que al ser mayores a 0.10, estos tienen un buen ajuste y representan una buena alternativa para la predicción del rendimiento de cultivos de cacao en Santander.
      * A partir de los valores derivados de las métricas de ajuste utilizadas para cada uno de los modelos, en especial para el valor de Error Cuadrático Medio (RMSE), 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 los cuales se obtuvo valores de 1708.29 y 1758.74 respectivamente.

# 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.
      * Se sugiere 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, con el fin de aumentar la variabilidad en los datos y ampliar el campo de investigación.
      * A pesar de la importancia que tiene involucrar nuevas tecnologías para el mejoramiento de los cultivos en Colombia, la búsqueda de trabajos dentro de la Universidad Industrial de Santander que utilicen herramientas de Aprendizaje Automático para la predicción agrícola no arrojó ningún resultado, por lo que se recomienda a la comunidad en general, en especial a la EEIE, continuar trabajando en la construcción de modelos haciendo uso de las diferentes metodologías de Aprendizaje Automático, con el fin de mejorar la precisión en los modelos.
      * Al identificar en su mayoría una influencia en la variable respuesta de las variables climáticas, se recomienda analizar y trabajar a profundidad con modelos enfocados en las variables climáticas para los diferentes cultivos.
      * Se recomienda continuar trabajando con diferentes lenguajes de programación que sirvan como herramienta para continuar trabajando con herramientas de Aprendizaje Automático Supervisado para diferentes trabajos de investigación.
      * Se resalta la importancia de continuar trabajando en conjunto con diferentes grupos de interés (gobierno, federaciones, agricultores) dentro de la investigación para lograr cada vez mejores resultados.
      * Se invita a los diferentes investigadores a continuar trabajando en la construcción de modelos con diferentes cultivos existentes a nivel de región y país para brindar información útil y oportuna que ayude a agricultores con la toma de decisiones acertadas.

# 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

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

# Apéndices

1. \* Trabajo de grado. [↑](#footnote-ref-1)
2. \*\* Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: PhD. Henry Lamos Díaz [↑](#footnote-ref-2)
3. \* Degree project. [↑](#footnote-ref-3)
4. \*\* Faculty of Physicomechanical Engineering. Industrial and Business School. Director: PhD. Henry Lamos Díaz. [↑](#footnote-ref-4)
5. Metros sobre el nivel del mar [↑](#footnote-ref-5)
6. Centro Internacional de Agricultura Tropical. [↑](#footnote-ref-6)
7. Departamento Nacional de Planeación. [↑](#footnote-ref-7)
8. El Niño/Southern Oscillation (El fenómeno de El Niño). [↑](#footnote-ref-8)
9. Root-mean-square error: Error cuadrático medio [↑](#footnote-ref-9)