Modelo predictivo para el rendimiento de cultivos de cacao y café en el departamento de Santander basado en herramientas de aprendizaje automático profundo y variables climáticas.

Elian Camilo Ricardo Durán Blanco y Dilson Orlando Castro Hernández

Trabajo de grado para optar por el título de Ingeniero Industrial

Director

Henry Lamos Díaz

Ph. D en Física - Matemática

Codirector

David Esteban Puentes Garzón

Msc. en Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2022

AGRADECIMIENTOS

Al docente Henry Lamos, nuestro querido director, por la confianza depositada, la guía y el tiempo dedicado para hacer realidad el presente proyecto.

A nuestro codirector, David Puentes, por todo el apoyo, la entrega y la disponibilidad para brindarnos sus conocimientos y palabras de aliento en cada etapa del proyecto.

Para ellos toda nuestra gratitud, respeto y admiración sincera, sin ellos este proyecto no sería una realidad.

A la Escuela de Estudios Industriales y Empresariales, por ser nuestra casa durante varios años y brindarnos las herramientas para crecer como personas integrales y profesionales competentes.

Al grupo de investigación OPALO, por los espacios y la dirección brindada en este bello camino de la investigación.

A Agronet, la UPRA, el IDEAM y el Ministerio de Agricultura y Desarrollo Rural, por poner a disposición la información utilizada en la presente investigación.

DEDICATORIA

A Dios, por su infinito amor y guía, quien me otorga día a día las herramientas para servir a los demás.

A mi mamá Lucía, mi papá Orlando, y mi hermana Mariana, por su esfuerzo día a día por verme convertido en una mejor persona y un gran profesional.

Al profe David, por sembrar en mí el amor por la enseñanza, a ti amigo, gracias totales.

A Luchis, por su amistad incondicional y sincera, sin duda ha sido un apoyo fundamental en todo este camino.

A Lucas Gómez y Santiago Ceballos quienes con sus consejos y guía han marcado significativamente mi futuro profesional.

A los campistas, por su amistad durante todos estos años de universidad, sin duda me enseñaron el significado de la palabra lealtad.

A mis profesores de la EEIE, por la paciencia, el amor y la entrega en el ejercicio de la docencia.

Finalmente, a Nicolás Gaona y Juan Avella, por su amistad incondicional y sincera.

Dilson Orlando Castro Hernández

DEDICATORIA

A Dios, por ser quién guía mi camino.

A mi mamá Edid, por todo el amor y esfuerzo que ha puesto para convertirme cada día en una persona más humana.

A mi papá Ricardo, por ser mi inspiración de hombre trabajador e inteligente, para mí eres el primer ingeniero de nuestra familia.

A mi hermano Ckristian, quien ha sido mi referente, mi apoyo y mi modelo a seguir.

Al profe David, por creer en mí y mis capacidades, por enseñarme a crecer dimensiones de mi vida más allá de la profesional, gracias a ti amigo.

A Galea, por ser mi casa por un tiempo en mi vida estudiantil, gracias a todos de quienes aprendí

A la familia Posgrados, por brindarme la oportunidad de crecer y madurar profesionalmente.

A la Universidad Industrial de Santander, la Escuela de Estudios, Industriales y Empresariales, a cada uno de sus docentes quienes ponen de su esfuerzo y conocimiento para servir a la sociedad y hacer de Colombia un mejor lugar.

Finalmente, a mi compañero Dilson, por su compañía en este camino, sé que la vida tiene planes gigantes para él.

Elian Camilo Ricardo Durán Blanco

Tabla de Contenido

Introducción 15

1. Planteamiento del problema 19

2. Justificación del proyecto 22

3. Objetivos 24

3.1. Objetivo general 24

3.2. Objetivos específicos 24

4. Revisión de Literatura 25

5. Marco Teórico 33

5.1. Knowledge Discovery in Databases (KDD) 33

5.2. Predicción del rendimiento de cultivos 34

5.3. Aprendizaje Automático 34

5.3.1. Aprendizaje Reforzado 35

5.3.2. Aprendizaje No Supervisado 35

5.3.3. Aprendizaje Supervisado 35

5.4. Aprendizaje Profundo 41

5.4.1. Redes Neuronales 42

5.4.2. Redes Neuronales Recurrentes 46

5.5. Métricas de desempeño y gráficos de resultados 50

5.5.1. Coeficiente de determinación $\boldsymbol{R2}$ 50

5.5.2. Error cuadrático medio (MSE)..... 51

5.5.3. Error cuadrático medio relativo (RMSE)..... 51

5.5.4. Error absoluto medio (MAE)..... 51

5.5.5. Error porcentual absoluto medio (MAPE)..... 52

5.5.6. Gráficos de dispersión de pronósticos..... 52

5.5.6. Gráficos de residuales..... 52

5.5.7. Shap Values..... 53

5.6. Validación de modelos..... 53

5.6.1. Train, Test, Split..... 54

5.6.2. Validación cruzada K-folds..... 54

5.6.3. Underfitting y Overfitting..... 55

5.6.4. Test de normalidad Shapiro Wilk..... 56

5.6.5. ANOVA de un factor..... 56

5.7. Optimización de modelos..... 57

5.7.1. Análisis de correlaciones..... 57

5.7.2. Normalización de datos..... 57

5.7.3. Reducción de la dimensionalidad..... 58

5.7.4 Ajuste de hiperparámetros (GridSearchCV)..... 60

5.7.5. Función de error..... 60

5.7.6. Optimizadores..... 61

5.7.7. Early Stopping 61

5.7.8. Tamaño de lote y épocas 61

6. Metodología 62

6.1. Descripción del conjunto de datos 62

6.2. Limpieza y Preprocesamiento de datos 65

6.3. Análisis exploratorio de Datos 65

6.3.1. Boxplots 66

6.3.2. Detección de outliers 68

6.3.3. Distribución de frecuencia del dataset 69

6.3.4. Análisis de correlación del dataset 71

6.4. Conjunto de datos de entrada 73

6.4.1. Multicolinealidad 74

6.5. Implementación de modelos 76

6.5.1. Implementación de modelos de Aprendizaje Automático y búsqueda de hiperparámetros. 76

6.5.2. Implementación de modelos de Aprendizaje Profundo y búsqueda de hiperparámetros ... 77

6.6. Comparación de Modelos 81

7. Resultados 82

7.1. Comparación de métricas de desempeño entre modelos 82

7.2. Gráficos de resultados 91

7.3. Shap Values 91

7.4. Intervalos de confianza para el error y el ajuste 94

7.5. Discusión de resultados..... 94

8. Conclusiones..... 98

9. Recomendaciones 101

Referencias..... 102

Lista de Tablas

Tabla 1. *Cumplimiento de objetivos* 19

Tabla 2. *Funciones integradas de kernel* 41

Tabla 3. *Estaciones Meteorológicas por tipo en Colombia* 63

Tabla 4. *Porcentaje de outliers por variable* 70

Tabla 5. *Factor de inflación de la varianza por variable* 76

Tabla 6. *Hiperparámetros seleccionados modelos ML* 78

Tabla 7. *Tabla valores promedio de las repeticiones de las métricas modelo café* 88

Tabla 8. *Tabla valores promedio de las métricas de las repeticiones modelo cacao* 88

Tabla 9. *Tabla Desviación Estándar de las repeticiones por modelo café* 89

Tabla 10. *Tabla Desviación Estándar de las repeticiones por modelo cacao* 89

Tabla 11. *Resultados prueba de Shapiro Wilks para las repeticiones de los modelos* 91

Tabla 12. *Anova comparación de modelos café* 91

Tabla 13. *Anova comparación de modelos cacao* 92

Lista de Figuras

Figura 1. *Proceso KDD* 34

Figura 2. *División de Machine Learning* 37

Figura 3. *La neurona* 44

Figura 4. *Comparación de arquitectura de una red feedforward y una red recurrente.* 48

Figura 5. *El módulo que se repite en un RNN contiene solo una capa* 49

Figura 6. *El módulo que se repite en un LSTM contiene cuatro capas que interactúan* 49

Figura 7. *El estado de la celda, representado por una línea horizontal en la parte superir del diagrama.* 0

Figura 8. *Ejemplo de una puerta con una capa sigmoidea* 1

Figura 9. *División del dataset en entrenamiento, prueba y validación* 56

Figura 10. *Ejemplo validación cruzada* 56

Figura 11. *Explicación gráfica underfitting y overfitting* 57

Figura 12. *Boxplots dataset café* 68

Figura 13. *Boxplots dataset cacao* 69

Figura 14. *Distribuciones de probabilidad dataset café* 72

Figura 15. *Distribuciones de probabilidad dataset cacao* 73

Figura 16. *Tabla de correlaciones entre variables dataset café* 74

Figura 17. *Tabla de correlaciones entre variables dataset cacao* 75

Figura 18. *Sequential Forward Selection* 77

Figura 19. *Gráfico de función de perdida y desempeño redes neuronales profundas café* 80

Figura 20. *Gráfico de función de perdida y desempeño redes neuronales profundas cacao* 80

Figura 21. *Gráfico de función de perdida y desempeño red neuronal LSTM café* 81

Figura 22. *Gráfico de función de perdida y desempeño red neuronal LSTM cacao* 82

Figura 23. *Comparativa R2 Train entre modelos* 83

Figura 24. *Comparativa Cross Validation Score entre modelos* 84

Figura 25. *Comparativa R2 Test entre modelos* 85

Figura 26. *Comparativa RMSE y MSE entre modelos* 86

Figura 27. *Comparativa MAE y MAPE entre modelos* 87

Figura 28. *Gráfico de caja y bigotes del MAPE de las repeticiones de cada modelo* 90

Figura 29. *Valores Shap cultivo café, izquierda Perceptrón Multicapa, derecha Bosques Aleatorios* 93

Figura 30. *Valores Shap cultivo cacao, izquierda Perceptrón Multicapa, derecha Bosques Aleatorios* 94

Lista de Apéndices

(Los apéndices se encuentran dentro de una carpeta adjunta)

Apéndice A. Análisis Bibliométrico.

Apéndice B. Visión general de los métodos y variables utilizada para la predicción del rendimiento de cultivos.

Apéndice C. Descripción de variables.

Apéndice D. Gráficos de resultados.

Apéndice E. Intervalos de confianza para el error y el ajuste.

Apéndice F. Metodología para la selección de modelos.

Apéndice G. Artículo Científico.

Resumen

Título: Modelo predictivo para el rendimiento de cultivos de cacao y café en el departamento de Santander basado en herramientas de Aprendizaje Automático Profundo*.

Autores: Dilson Orlando Castro Hernández

Elian Camilo Ricardo Durán Blanco**

Palabras clave: predicción, regresión, Aprendizaje Automático, Aprendizaje Profundo, cultivos, condiciones climáticas, condiciones geográficas.

Descripción: el Aprendizaje Automático y el Aprendizaje Profundo son técnicas utilizadas para identificar patrones y predecir comportamientos futuros, en el sector agrícola permite generar análisis y proyecciones optimizando así la producción y comercialización de los cultivos, apoyando la toma de decisiones a agricultores, gobiernos e intermediarios. Esta investigación tiene como objeto de estudio la producción agrícola anual de los cultivos de cacao y café (años 2007-2021) en los municipios de Colombia, en términos de rendimiento del cultivo (t/Ha). En este estudio se contrastan modelos de Aprendizaje Automático como Regresión Lineal Múltiple, Árbol de Decisión, Bosques Aleatorios, XGBoost y Regresión de Vectores de Soporte; frente a modelos de Aprendizaje Profundo como el Perceptrón Multicapa y la Red Neuronal Recurrente LSTM. Los predictores incluyen siete variables de condiciones ambientales proporcionadas por el IDEAM y condiciones geográficas: área sembrada y la altura sobre el nivel del mar de las cabeceras municipales. Los métodos se compararon utilizando R^2, MSE, RMSE, MAE, MAPE; encontrándose un desempeño superior en los métodos basados en árboles y en redes neuronales**.**

*Trabajo de Grado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: ph.D. Henry Lamos Díaz. Codirector: MSc.David Esteban Puentes Garzón.

Abstract

Project tittle: Predictive model for the yield of cocoa and coffee crops in the department of Santander using Deep Learning techniques.

Authors: Dilson Orlando Castro Hernández

Elian Camilo Ricardo Durán Blanco**

Keywords: forecast, regression, Machine Learning, Deep Learning, crops, weather conditions, geographic conditions.

Description: Machine Learning and Deep Learning are techniques used to identify patterns and predict future behavior. The agricultural sector allows generating analyzes and projections to optimize the production and commercialization of crops, supporting the decision-making of farmers, governments, and intermediaries. This research has as its object of study the annual agricultural production of cocoa and coffee crops (years 2007-2021) in the municipalities of Colombia in terms of crop yield (t/Ha). In this study, machine learning models such as Multiple Linear Regression, Decision Tree, Random Forests, XGBoost, and Support Vector Regression are contrasted with Deep Learning models such as Multilayer Perceptron and Recurrent Neural Network LSTM. The predictors include seven environmental conditions variables provided by the IDEAM and the geographical conditions: sown area and height above sea level of the head-municipalities. Methods were compared using R^2, MSE, RMSE, MAE, MAPE finding superior performance in three based methods and neural networks.

*Bachelor Thesis

** Faculty of Physicomechanical Engineering. Industrial and Business School. Director: ph.D. Henry Lamos Díaz. Codirector: MSc.David Esteban Puentes Garzón.

Introducción

La combinación y rotación de cultivos, la identificación de variedades de cultivos con mejor respuesta a cambios climáticos y las técnicas de irrigación, son algunos ejemplos de adaptaciones en la agricultura ante el riesgo asociado a la incierta variabilidad climática presente históricamente asociada al desarrollo de la agricultura (CEPAL, 2011). Históricamente áreas de estudio como la Meteorología Agrícola han pretendido entender y explicar la acción mutua entre los factores hidro-meteorológicos y la agricultura, incluyendo la ganadería, la silvicultura y la horticultura; con el objetivo de identificar los efectos del clima natural, las modificaciones medioambientales, las condiciones de almacenamiento y transporte, sobre la práctica de la agricultura (IDEAM - Instituto de Hidrología, 2022)

En el año 2020 se evidenció en el país una de las contracciones económicas más fuertes de los últimos años debido a la situación coyuntural de la pandemia del COVID-19, generando un decrecimiento del 15,7% del PIB nacional para el segundo trimestre, a pesar de que, para la fecha algunos sectores económicos hacían nuevamente presencia en el mercado para el proceso de reactivación. Durante este periodo, el sector agropecuario fue el de mejor desempeño, resultando con un crecimiento del 28% para el final del año, así lo evidenciaron las cifras del PIB reveladas por el Departamento Nacional de Estadística (DANE). Concretamente en dinero, dentro de la categoría de agricultura, los cultivos generaron 47 billones de pesos colombianos.

Teniendo en cuenta el potencial que tiene el sector agrícola en Colombia, en el contexto Santandereano gran parte de la producción departamental es el cacao y el café. El cacao cultivado en el departamento de Santander representa alrededor del 42% de la producción nacional con 22

mil 800 toneladas de la producción anual, contando con la participación de más de 17 mil familias cacaoteras en los 40 municipios en los que se cultiva el grano (Gobernación de Santander, 2020). Por otra parte, el café es un cultivo que se encuentra presente en todas las provincias del departamento y son aproximadamente 51.840 hectáreas en 75 municipios, contando con la participación de alrededor 32.602 familias cafeteras en 37.971 fincas representando el 23% de la producción agrícola departamental (Federación Nacional de Cafeteros de Colombia, 2020).

La producción agrícola se ve afectada por diversidad de factores, uno de ellos es el clima, que influye directamente en el crecimiento de los cultivos, por ejemplo, el cambio en las temperaturas en muchos casos termina reduciendo los cultivos deseados y proliferando las malas hierbas, por otra parte, los cambios en los regímenes de lluvias aumentan las probabilidades de fracaso de las cosechas a corto plazo y de la reducción de la producción a largo plazo (Gerald C. Nelson, 2009)

El éxito de cultivos permanentes tales como cacao y café depende de diversos factores cómo el establecimiento de los cultivos en condiciones adecuadas y recomendadas de clima, la elección de semillas que sean ejemplares de buena calidad, la aplicación oportuna y adecuada de prácticas que permitan predecir la rentabilidad y eficiencia de los cultivos y las cosechas para temas relacionados a la planeación y control. En función de esta necesidad, es posible aplicar modelos matemáticos que permitan anticipar la producción de un cultivo, ya sea en función de su comportamiento histórico o en función de las variables climáticas que tengan un efecto en el rendimiento agrícola, y evaluar el mejor método a través de métricas de desempeño.

El Aprendizaje Automático se ha hecho popular debido a que tiene un gran potencial en la predicción de rendimientos agrícolas y la identificación de factores que explican su variabilidad;

esto debido a que la predicción basada en máquinas de aprendizaje junto con la minería de datos y el análisis bayesiano, tiene la capacidad de llevar a cabo análisis de datos más grandes y complejos que los métodos tradicionales, otorgando resultados más rápidos y precisos, en otras palabras, procesamiento computacional más económico y eficiente.

En los últimos años es cada vez más común aplicar estos modelos en el campo de la agricultura basados en variables climáticas. En 2017, Corrales et al. muestran el enfoque del Machine Learning para pronosticar la producción de cacao en la región de Santander haciendo uso de la temperatura promedio diaria, humedad relativa diaria y la tasa de precipitaciones diaria. Por otra parte, Mahdi et al. pronostican el rendimiento agrícola de la papa, uno de los alimentos más consumidos en el mundo haciendo un análisis de precipitaciones con redes LSTM con el fin de ayudar a comprender mejor los cambios en el clima y el rendimiento de los cultivos.

En este proyecto se pretende comparar métodos clásicos de regresión frente a modelos avanzados, en este orden, se aplicarán modelos de Aprendizaje Automático y Aprendizaje Profundo para pronosticar los rendimientos de cacao y café en el departamento de Santander y se identificarán qué variables climáticas influyen en ellos. Las variables climáticas de estudio se obtendrán mediante sensores en las estaciones meteorológicas del Instituto de Hidrología, Meteorología y Estudios Ambientales – IDEAM.

Tabla 1. *Cumplimiento de objetivos*

Objetivos específicos	Cumplimiento
Realizar una revisión de literatura acerca de modelos de Aprendizaje Automático para la predicción de rendimientos agrícolas.	Capítulo 4
Aplicar modelos de Aprendizaje Profundo para la predicción del rendimiento de cultivos de cacao y café en Santander haciendo uso de variables climáticas.	Numeral 6.5.2.
Comparar modelos mediante el uso de métricas para determinar la mejor alternativa para la predicción del rendimiento de cultivos de cacao y café en el departamento de Santander.	Numeral 7.1.
Elaborar un artículo de carácter publicable resumiendo los hallazgos en el proyecto.	Apéndice G

Nota: Fuente propia.

1. Planteamiento del problema

El sector agrícola es diverso y está lleno de contradicciones, a pesar de representar una pequeña proporción de la economía mundial, sigue siendo la actividad central para millones de personas (Howden et al., 2007). En el año 2013, La Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) emitió un informe donde se afirmaba que aproximadamente 2500 millones de personas dependen de la agricultura. Por otra parte, alrededor del 40% de la superficie terrestre del planeta está siendo ocupada para agricultura y ganadería; aproximadamente 1500 millones de hectáreas de tierra son utilizadas para plantar cultivos mientras que 3500 millones se utilizan para pastoreo (Food and Agriculture Organization of the United Nations, 2020).

El incremento en la concentración de gases de efecto invernadero es tal que parece inevitable que se presenten cambios en el clima, los cuales forzarán al sector agrícola a tomar medidas de adaptación. Las capacidades de adaptación son limitadas y por lo tanto es muy probable que el cambio climático afecte la disponibilidad y acceso a alimentos e incremente la volatilidad de los precios (Feldman & Cortés, 2016). A lo largo del siglo XXI, los efectos del cambio climático reducirán el crecimiento económico, complicarán los esfuerzos por reducir la pobreza y afectarán la seguridad alimentaria (Field et al., 2014).

Colombia es un país privilegiado en recursos hídricos, cuenta con seis nevados y más de 48.000 humedales (entre ríos, lagos, lagunas, ciénagas, arrecifes y estuarios, entre otros) habitados por el 87% de la población, no obstante, es un país muy vulnerable a los efectos del cambio climático, en especial a los efectos derivados del fenómeno del Niño y la Niña. En consecuencia, la región Pacífica, los departamentos de la región Andina y los departamentos de la región Caribe sufren

cambios en los patrones de precipitación y temperatura. Según el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (IDEAM), en los primeros meses del año 2021 se presentó una reducción significativa del volumen de precipitaciones. Esta época de sequía y heladas ocasionó la pérdida de más de 10.000 hectáreas de cultivos en el departamento de Cundinamarca, generando preocupación a la seguridad alimentaria nacional.

En el contexto departamental, Santander ha sufrido pérdidas significativas en cultivos al largo del territorio, según Radio Nacional de Colombia, la sequía afectó el 80% de los cultivos presentes en los 12 municipios del departamento. En la provincia de García Rovira, después de 3 meses sin lluvias, cultivos que generalmente producirían toneladas de limones, mandarinas, frijol, maíz y café no pasaron el periodo crítico de crecimiento.

De lo mencionado anteriormente se puede afirmar que la seguridad alimentaria tiene incidencia en una jerarquía de sistemas que operan en componentes económicos, ecológicos, sociales y políticos del planeta Tierra, por lo tanto, pronosticar los rendimientos de los cultivos, o proporcionar una expectativa de cantidades antes de la cosecha es muy importante para toda la cadena de producción agrícola; los agricultores pueden adaptar su gestión, los comerciantes y aseguradores sus esquemas de precios, los proveedores sus existencias, las empresas de logística sus rutas, las autoridades nacionales sus balances alimentarios para orientar la importación o exportación y, finalmente, las organizaciones internacionales de ayuda pueden movilizar socorros (Schauberger et al., 2020).

Diversos países han adoptado el uso de modelos de Aprendizaje Automático en los últimos años, estos métodos permiten resultados relativamente eficientes en relación con su costo. El crecimiento de los volúmenes de datos producidos por sensores remotos ha hecho posible abordar

dichos análisis desde diferentes perspectivas; asimismo, es valioso explorar cómo se han realizado estas investigaciones fuera del contexto colombiano, debido a la naturaleza de los datos, estos modelos pueden ser aplicables a cualquier tipo de cultivo y zona geográfica.

Dada la necesidad de aplicar modelos eficaces para la predicción de cultivos en el país, se propone realizar dichos modelos enfocados en cultivos de café y cacao, cultivos altamente representativos que aportan en mayor forma a la productividad agrícola del país y también en el departamento de Santander.

2. Justificación del proyecto

En 2021 se registró una producción histórica de 69.040 toneladas de cacao, representando un crecimiento del 8,9% respecto al año 2020, comparado con la producción de hace 15 años (2006) el crecimiento fue de 127%, evidenciando el dinamismo de este subsector. A cierre del 2021 Colombia exportó 11.689 toneladas de grano seco, un 4,9% más que el año anterior, y 14.647 toneladas de productos derivados, un 7,1 % más respecto al 2020. Por su parte seis municipios son los principales productores de grano de cacao, con su respectiva participación: Santander (40,6%), Arauca (9,6%), Antioquia (9,6%), Tolima (5,8%), Huila (5,1%) y Nariño (5%) (FEDECACAO, 2022).

Colombia es el mayor productor de café arábigo suave a nivel mundial, en 2021 registró una producción de 12,6 millones de sacos, un 9% menos que el año 2020, este resultado se explica por el efecto del clima en algunas zonas cafeteras del país que pudo afectar la producción. A cierre de 2021 las exportaciones fueron de 12,4 millones de sacos, un 1% menos que el año anterior. Por su parte seis municipios son los principales productores de grano de café, con su respectiva participación: Huila (17,16%), Antioquia (13,84%), Tolima (12,74%), Cauca (11,07%), Caldas (7,06%) y Santander (6,31%). (FEDECACAO, 2021).

Colombia, por su ubicación cercana a la línea del Ecuador, es particularmente vulnerable a los efectos derivados del cambio climático; entre 2008 y 2018 se presentaron dos fenómenos del Niño (periodos de escasez de lluvias) y cuatro fenómenos de La Niña (periodos de exceso de lluvias). En los meses más intensos, El Niño, afectó alrededor de 500 municipios, 44,64% del total de municipios del país; por su parte en los meses más intensos, La Niña, afectó entre 200 y 400

municipios, entre el 17,85% y 35,71% del total de municipios del país (Banco de la República Colombia, 2021).

El efecto del cambio climático es variado en la agricultura, por ende, las condiciones climáticas que benefician a un cultivo pueden afectar a otro. Según Eduard Baquero, presidente ejecutivo de Fedecacao, la producción de cacao disminuyó 30,8% en el primer trimestre del 2022 respecto al mismo periodo del año anterior debido al mal clima, puesto que la flor del cacao se afecta por las variaciones en los niveles de lluvia (La República, 2022). Por otra parte, según un estudio realizado por investigadores de la Universidad de Zúrich, las regiones aptas para sembrar café a nivel global se podrían reducir hasta un 50% a nivel global a 2050 como consecuencia de la temperatura media en países como Brasil, Vietnam, Colombia e Indonesia (El PAÍS, 2022)

Considerando este escenario, surge la necesidad de aplicar herramientas que permitan anticiparse a los efectos del clima en la agricultura. Los modelos predictivos comprenden técnicas que, mediante la recolección de datos históricos, el Big Data, el Aprendizaje Automático y el reconocimiento de patrones, puedan predecir resultados futuros que faciliten la toma de decisiones precisas, ágiles y oportunas (Harvard edX, 2022).

Se pretende obtener pronósticos de rendimientos agrícolas a partir de la aplicación de herramientas matemáticas, estadísticas y computacionales, como los algoritmos de Machine Learning Supervisado y algoritmos de Aprendizaje Profundo. Los modelos desarrollados habilitan la toma de decisiones basadas en datos, en agricultores, gobierno, intermediarios y demás actores, incrementando la productividad y competitividad del sector consecuentemente.

3. Objetivos

3.1. Objetivo general

Construir un modelo predictivo para el rendimiento agrícola de los cultivos de cacao y café en el departamento de Santander a partir de variables climáticas y Aprendizaje Profundo.

3.2. Objetivos específicos

- Realizar una revisión de literatura acerca de modelos de Aprendizaje Automático para la predicción de rendimientos agrícolas.

- Aplicar modelos de Aprendizaje Profundo para la predicción del rendimiento de cultivos de cacao y café en Santander haciendo uso de variables climáticas.

- Comparar modelos mediante el uso de métricas para determinar la mejor alternativa para la predicción del rendimiento de cultivos de cacao y café en el departamento de Santander.

- Elaborar un artículo de carácter publicable resumiendo los hallazgos en el proyecto.

4. Revisión de Literatura

En la revisión de literatura se observan diferentes tipos de cultivos objeto de estudio, tales como arroz, trigo, maíz, cebada, uva, caña de azúcar y naranja, para los cuales se realizan predicciones del rendimiento haciendo uso de modelos de Aprendizaje Automático y Aprendizaje Profundo, intervienen variables explicativas de origen climático, con datos que se obtienen mediante el uso de sensores. También está muy presente en investigaciones recientes el uso de datos de origen satelital, donde se cita constantemente el sensor MODIS.

Autores como Jones (2000), Gaveta (2017), Zhang (Cao et al., 2021a), coinciden en que la variabilidad climática genera riesgos económicos y riesgos para la seguridad alimentaria en todo el mundo debido a sus principales influencias en la agricultura. Es por esto que diferentes países han optado por desarrollar métodos para pronosticar el rendimiento de los cultivos, haciendo uso del gran volumen de datos climáticos obtenidos de instituciones de carácter oficial y gubernamental a través de la instalación de sensores para el estudio del tiempo y también el aprovechamiento del sensor remoto MODIS, este último es el sensor principal de los satélites EOS AM-1 y AQUA cuya función principal según Barker et al. (Anderson, 2016) es tomar datos de cobertura espacial y espectral.

"Rechazar los efectos potenciales del cambio climático en el rendimiento de los cultivos genera modelos cuyas predicciones difieren de los resultados a menudo, comprender esta divergencia es fundamental generar modelos de cómo los cultivos responden al clima para así llegar a proyecciones más confiables y por lo tanto posicionar a los modelos estadísticos, para que sigan desempeñando un papel importante en la anticipación de los impactos futuros del cambio climático en la agricultura" (Lobell & Burke, 2010).

Del primer artículo que se analizó exhaustivamente, se puede observar que Li et al. (Li et al., 2021a) desarrollaron un modelo de Random Forest para estimar el rendimiento para tres cultivos principales de cereales (trigo, maíz y arroz) en toda China, dentro del cual se utilizaron 3 tipos de datos principales que incluyen el clima, los índices de vegetación y las propiedades del suelo. El trigo (Triticum aestivum L.), el maíz (Zea mays L.) y el arroz (Oryza sativa L.) son los tres principales cultivos alimentarios del mundo (Gao et al., 2019; Grundy et al., 2016), ac. - contando aproximadamente el 42,5% del suministro de calorías alimentarias del mundo (FAO, 2018). El clima puede tener efectos significativos en la producción de estos cultivos, por ejemplo, los eventos de alta temperatura extrema, definidos por períodos cortos de temperatura máxima diaria superior a 33 ◦C, pueden afectar en gran medida el número de granos de trigo y maíz en la etapa temprana de llenado de granos (Barlow et al., 2015; Dawson y Wardlaw, 1989). Los eventos de frío extremo con una temperatura mínima diaria inferior a 0 ◦C están estrechamente relacionados con la esterilidad del cultivo y el aborto de los granos formados durante la etapa de floración (Barlow et al., 2015). La sequía y las inundaciones también pueden afectar significativamente el rendimiento de los cultivos. Por ejemplo, la sequía extrema afecta el crecimiento y la arquitectura de las raíces y puede resultar en grandes pérdidas de rendimiento (Schauberger et al., 2020); las inundaciones pueden destruir directamente las tierras agrícolas y también pueden causar anegamiento que es perjudicial para la salud del suelo y que dará como resultado reducciones significativas del rendimiento (Li et al., 2021b).

Los autores usan el modelo de Random Forest, para regresión, se basa en la construcción de una multitud de árboles de decisión (Breiman, 2001), hicieron uso de datos de rendimiento de pruebas de cultivos incluyendo arroz, maíz y trigo, con una ventana temporal del año 2013 al año 2019 y se resalta que el uso de este algoritmo es muy importante ya que proporciona información

confiable sobre la importancia de la característica de cada variable y de esta forma se puede estimar de manera efectiva el error de prueba, bajo un costo computacional del modelo de formación (Stefan, 2018). En general, el rendimiento del modelo fue comparable con estudios anteriores en China, el pronóstico se obtuvo de forma satisfactoria alrededor de uno a tres meses antes de la cosecha del trigo de invierno (r = 0,81–0,85, RMSE = 10,5–11,4%); uno o dos meses antes de la cosecha del maíz de primavera (r = 0,79–0,81, RMSE = 17,1–17,9%), maíz de verano (r = 0,77–0,79, RMSE = 10,2–10,4%), arroz temprano (r = 0,71–0,72, RMSE = 7,4–7,5%), arroz medio (r = 0,78–0,82, RMSE = 7,6–8,3%) y arroz tardío (r = 0,76–0,78, RMSE = 8,6–8,9%).

Dentro de la búsqueda se observa un gran número de algoritmos utilizados en cada investigación, por ejemplo, los autores Jurecka et al. (Jurečka et al., 2021) realizaron un modelo para la predicción del rendimiento de cultivos de cebada de primavera y trigo de invierno para 33 distritos de la República Checa; en este estudio se hizo uso de indicadores basados en evapotranspiración a partir de datos detectados de forma remota, para esto acudieron al uso de Redes Neuronales Artificiales (ANN). Las predicciones basadas en ANN se calcularon para ambos cultivos para todos los distritos juntos y para cada distrito por separado. La raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2) entre los rendimiento observados y pronosticados, variaron con la fecha dentro de la temporada de crecimiento y con el número de entradas ANN utilizadas.

El período con la mayor capacidad predictiva es desde principios de junio hasta mediados de junio. En este período óptimo para la predicción del rendimiento usa un menor número de entradas de ANN, sin embargo, la precisión de la predicción mejoró a medida que se incluyeron más insumos dentro de las ANN. Los valores de RMSE para distritos individuales variaron entre

0,4 y 0,7 toneladas por hectárea mientras que R^2 alcanzó valores de 0,5-0,8 durante el período óptimo.

El uso de datos climáticos para la predicción de rendimiento de cultivos no está limitado a modelos de Aprendizaje Automático, en el año (2019), los autores Cai, Guan & Lobell realizaron el pronóstico del rendimiento del trigo en Australia haciendo uso del modelo de regresión conocido como LASSO y se compararon con tres métodos principales de Aprendizaje Automático: las Máquinas de Soporte Vectorial, los Bosques Aleatorios y las Redes Neuronales Artificiales. Los autores llegaron a la conclusión de que la integración de datos climáticos y datos satelitales pueden mejorar el rendimiento para la predicción del rendimiento. Al descomponer las contribuciones de los datos climáticos y los datos satelitales, se puede encontrar información única y superpuesta para la predicción del rendimiento de cultivos.

Los autores encontraron que, para su contexto, los métodos basados en Aprendizaje Automático superan al método de regresión en el modelado del rendimiento de cultivos. Los resultados del estudio mostraron que la combinación de datos climáticos y datos satelitales se puede lograr explicar la variabilidad del modelo en un 75% en las condiciones óptimas. Los resultados demostraron la capacidad de dictar el rendimiento del trigo con un tiempo de espera de hasta dos meses antes de la cosecha. El mejor modelo desarrollado fueron las Máquinas de Soporte Vectorial (SVM), logró la predicción del rendimiento con un R^2 de 0.73 para el mes de octubre, y un $R^2 = 0.75$ para el mes de diciembre. Dentro de la comparación de modelos, los autores usaron por separado el índice de vegetación mejorado (EVI) y datos climáticos, los autores encontraron que los métodos no lineales obtienen mejoras de rendimiento sobre el uso del método de regresión lineal avanzado al usar sólo datos climáticos que solo usar EVI, lo que indica los impactos de las

variables climáticas en el rendimiento del trigo, es decir, que los datos climáticos proporcionan más respuestas no lineales que EVI. El hallazgo es consistente con la literatura ya que el uso de datos de temperatura y datos de precipitación ejercen respuestas no lineales sobre el rendimiento según estudios de Álvarez (2009); Stöckle et al. (2003); Zheng et al. (2014).

En los últimos años algunos autores han usado la combinación de teledetección y técnicas de Aprendizaje Profundo. Mahdi, Mrittika, Shams, Chowdhury & Siddique en (Mahdi et al., 2020) utilizaron redes LSTM y un modelo proceso gaussiano para la predicción del rendimiento de la papa y un análisis de series de tiempo para la precipitación en distrito de Munshiganj de Bangladesh, aprovechando que el comportamiento de la lluvia tiene un comportamiento cíclico. El uso de redes LSTM es emergente cuando se trata de hacer uso de datos climáticos cómo variables explicativas en los modelos, esto se debe a que una característica de las redes LSTM es que los datos de entrada por lo general poseen características temporales, el uso de este tipo de redes permite el análisis de secuencias de datos y no únicamente de datos individuales. Esta tecnología abre brechas en la investigación debido a la gran cantidad de variables climáticas tienen propiedades temporales y cuyos datos son obtenidos a través de sensores que tienen cierta frecuencia de captación.

Cao et al. (Cao et al., 2021b) también compararon enfoques de Aprendizaje Automático y Aprendizaje Profundo, donde el uso de redes LSTM resaltó, ya que logra explicar entre el 77% y el 87% de la variabilidad de la predicción del rendimiento del cultivo del arroz en China, superando al modelo de Random Forest con un R^2 de entre 0,76 a 0,82; ambos modelos fueron bastante superiores al modelo de regresión LASSO con un R^2 de entre 0,33 a 0,42 en la predicción del rendimiento. Para este estudio se integraron datos disponibles públicamente que incluyen el

rendimiento del arroz y el área de siembra, datos satelitales, que incluyen fluorescencia de clorofila contigua inducida por el sol (SIF), los datos ambientales que contienen el índice de vegetación mejorado (EVI), propiedades del suelo y datos climáticos cómo temperatura mínima, máxima, precipitación, evapotranspiración, etc.

La principal razón por la cual los modelos de Machine Learning y Deep Learning tuvieron mejor desempeño que la Regresión Lineal (LASSO) en las principales áreas de producción de arroz es que estos modelos pueden capturar relaciones complejas y no lineales entre las variables de entrada y el rendimiento. Las redes LSTM, debido a su estructura de red neuronal recurrente que puede incorporar hacen posible, en este tipo de modelos, el establecimiento de relaciones acumulativas y no lineales entre el rendimiento de arroz y factores ambientales.

Continuando con el Aprendizaje Profundo, en el año 2019, el algoritmo propuesto por Ma et al consiste en “the Stacked Sparse Auto-encoder (SSAE)” para la estimación del rendimiento de arroz, haciendo uso de datos climáticos y de espectro radiómetro de imágenes de resolución moderada (MODIS) con el objetivo de elegir escenarios que muestren el mejor desempeño combinado en términos de duración de la temporada de cultivo y períodos de agregación de datos climáticos. El artículo aborda análisis desde la perspectiva temporal como también la espacial. Los autores además de alinear el modelo SSAE con los objetivos del estudio, compararon su desempeño con un modelo de red neuronal artificial (ANN). Según la combinación de datos escogidos, el modelo SSAE superó a ANN, mostrando el error cuadrático medio (RMSE) y el% RMSE de 33,09 kg (10a) −1 (5,21 kg (10a)−1 menor que ANN) y 6,89% (1,14% menor que ANN).

En este estudio, se hizo uso de datos meteorológicos recopilados diariamente que incluyen temperatura máxima, temperatura mínima, temperatura promedio, precipitación y radiación solar

entre el año 2000 y 2013 de 89 estaciones de observación meteorológica y 599 estaciones meteorológicas automáticas operadas por la Administración Meteorológica de Corea (KMA2018). Los datos se interpolaron originalmente para toda Corea del Sur utilizando técnicas de Cressman con datos de cuadrícula de 3 km (Cressman 1959; Hong et al.2012). Por lo tanto, había básicamente 365 capas de cada tipo de datos climáticos para cada año. Con respecto al rendimiento, los datos se recopilaron del año 2000 al 2013 desde el Servicio de Información Estadística de Corea (KSIS 2018). Los datos contienen información sobre la producción total de arroz, el área total de arroz y el rendimiento de arroz para un total de 36 regiones de Jeolla-do.

El Deep Learning se ha convertido en un foco de investigación popular entre la comunidad científica, conteniendo una gran cantidad de arquitecturas como SSAE, Redes Neuronales Convolucionales (CNN) y Redes Neuronales Recurrentes (RNN). Centralmente el SSAE ha mostrado fortalezas en los problemas de regresión y el pronóstico de series de tiempo (Ma et al., 2019a), por eso es tan usado en modelos que incluyen variables climáticas. En estudios recientes, Liu et al obtuvo un resultado superior al algoritmo clásico de regresión de vectores de soporte para procesar datos meteorológicos masivos.

A partir de la revisión de literatura se puede observar que la predicción de rendimiento de cultivos es una tendencia que muchos países a lo largo del mundo han optado por seguir debido a la preocupación mundial por la seguridad alimentaria. Existen métodos que abordan perspectivas desde la simulación hasta los algoritmos de Aprendizaje Automático, estos últimos han destacado debido a que su uso es eficiente y más económico que otros métodos. En la gran mayoría de los artículos las variables explicativas son meteorológicas, esto evidencia que existe un gran interés en entender el comportamiento de las relaciones que hay entre el clima y los rendimientos

agrícolas. Los algoritmos utilizados han evolucionado a lo largo de los años debido al desarrollo de modelos más robustos dentro del Aprendizaje No Supervisado y el Aprendizaje Profundo. La combinación de datos climáticos y otro tipo de variables también ha sido un factor determinante en el tipo de modelo a utilizar para la predicción del rendimiento de cultivos. Finalmente, la comparación de dichos modelos ha facilitado en diversas investigaciones elegir el algoritmo más preciso y acertado para el pronóstico de rendimiento.

En el apéndice A se detalla el análisis bibliométrico derivado de la revisión de literatura realizada.

En el apéndice B se resumen los principales cultivos, algoritmos y variables climáticas utilizadas por los autores que se tomaron en cuenta en la revisión preliminar de la literatura.

5. Marco Teórico

5.1. Knowledge Discovery in Databases (KDD)

El Descubrimiento de Conocimiento en Bases de Datos consiste en una serie de pasos para extraer patrones que se esconden en grandes volúmenes de datos, sus etapas incluyen el preprocesamiento de la información, la selección de datos, la minería de datos que incluye la construcción de modelos y la presentación de resultados. (Fayyad et al., 1996). Las etapas del proceso KDD comprenden:

Figura 1. *Proceso KDD*

Nota: Adaptado de “De la minería de datos al descubrimiento del conocimiento en bases de datos” (Fayyad et al., 1996).

5.2. Predicción del rendimiento de cultivos

El rendimiento agrícola es la cantidad de un cultivo cosechado por unidad de superficie de tierra, su valor puede variar debido a factores interrelacionados como riegos, fertilizantes, rotación de cultivos, morfología de la planta, condiciones del terreno, clima, entre otros. Para cada temporada de crecimiento, los agricultores requieren estimar el rendimiento de todos los cultivos involucrados, lo cual es difícil debido a que el resultado depende de factores interrelacionados (Gonzalez-Sanchez et al., 2014). Un factor relevante para escoger el método predictivo es la relación entre la precisión y rapidez, Gonzalez-Sanchez et al menciona que el Aprendizaje Automático es un enfoque idóneo para lograr soluciones prácticas y efectivas a estos retos ya que las investigaciones afines están orientadas a proponer y comparar métodos con el fin de escoger el que más se ajuste a los datos.

5.3. Aprendizaje Automático

El Aprendizaje Automático (ML) es el proceso mediante el cual se le ayuda a un equipo informático a aprender, utilizando modelos matemáticos de datos. El Aprendizaje Automático usa algoritmos para encontrar patrones en los datos y posteriormente crear modelos de datos capaces de hacer predicciones. Con más datos, repeticiones y experiencia el Aprendizaje Automático mejora su precisión, de manera muy similar a como aprenden los humanos con la práctica. Las aplicaciones del Aprendizaje Automático son diversas: algoritmos de regresión, algoritmos de detección de anomalías, algoritmos de clusterización y sistemas de recomendación. Existen tres

técnicas principales dentro del Aprendizaje Automático: el Aprendizaje Supervisado, el Aprendizaje No Supervisado y el Aprendizaje Reforzado (El Bouchefry & de Souza, 2020)

5.3.1. Aprendizaje Reforzado

Este tipo de aprendizaje utiliza algoritmos que reciben comentarios después de tomar una acción, son capaces de determinar si esa elección fue correcta, incorrecta o neutra. Esta técnica se utiliza en sistemas automatizados que requieren la toma de decisiones pequeñas sin la directriz de un humano, por ejemplo, los autos autónomos aprenden a través de la experiencia y el refuerzo a respetar el límite de velocidad, a permanecer en el carril y detenerse cuando hay peatones (El Bouchefry & de Souza, 2020)

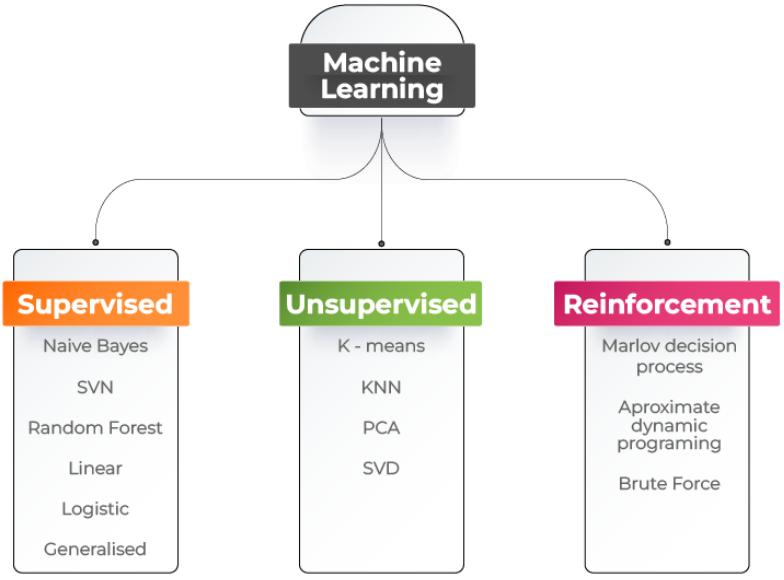
5.3.2. Aprendizaje No Supervisado

Este tipo de aprendizaje etiqueta automáticamente los datos, organizándolos o describiendo su estructura, esta técnica es útil cuando no se sabe el resultado, por ejemplo, cuando se requiere clasificar los clientes en segmentos más pequeños según similitudes en su comportamiento de compra (El Bouchefry & de Souza, 2020)

5.3.3. Aprendizaje Supervisado

Este tipo de aprendizaje genera predicciones a partir de datos etiquetados proporcionados como entradas, esta técnica es útil cuando se sabe el resultado, por ejemplo, para predecir la población de una determinada ciudad a partir de registros históricos. Dentro de los algoritmos de Aprendizaje Supervisado se encuentran la Regresión Lineal, el Árbol de Decisión, los Bosques Aleatorios, las Máquinas de Soporte Vectorial (El Bouchefry & de Souza, 2020).

Figura 2. *División de Machine Learning*



Nota: Adaptado de "Learning in Big Data: Introduction to Machine Learning" (el Bouchefry & de Souza, 2020)

5.3.3.1. Regresión Lineal. El análisis de regresión utilizando más de una variable independiente se le llama análisis de regresión multivariante, en este algoritmo se busca contar la variación de las variables independientes en la variable dependiente en simultáneo. Los supuestos de la regresión lineal múltiple son distribución normal de los datos, ausencia de valores extremos, linealidad y no existencia de múltiples relaciones entre las variables independientes (Buyukozturk, 2002). La Regresión Lineal Múltiple se expresa en la ecuación 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon \quad (1)$$

Donde β_0 es el intercepto con el eje y, β_i son los pesos asociados a las variables independientes, X_i son los valores de las variables independientes y ε es el error (Uyanık & Güler, 2013).

5.3.3.2. Árbol de Decisión (Decision Tree). Para empezar a crear el árbol se compara el valor de la primera característica contra el valor de la variable respuesta, esto se hace para conocer cuál característica puede predecir mejor la variable respuesta, esto puede presentar problemas de impurezas por falta de datos, para evitar esto se utiliza el coeficiente de impureza de Gini, la característica con el menor coeficiente de Gini se convertirá en el nodo raíz, posteriormente se recalcula el índice Gini para las demás características y se construye el árbol de decisión. Las métricas utilizadas están dadas por las ecuaciones 2,3 y 4

$$Entropía\colon -\sum_{i=0}^{c-1} p(t) log_2(p(t)) \tag{2}$$

$$Gini(t)\colon 1 - \sum_{i=0}^{c-1} [p(t)]^2 \tag{3}$$

$$Error\ de\ clasificación(t)\colon 1 - max_i[p(t)] \tag{4}$$

Donde C es el número de clases y $p(t)$ es la proporción de observaciones i en el subconjunto t.

5.3.3.3. Bosques Aleatorios (Random Forest). Este algoritmo surge como la agrupación de varios árboles de clasificación y fue desarrollado para hacerle frente al problema de que los árboles suelen tener buenos resultados en el set de datos de entrenamiento, más cuando se le presentan datos que no ha visto el algoritmo, este no predice las variables de manera óptima. Random Forest es una combinación de estos predictores débiles, donde se tiene que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque (Medina-Merino & Ñique-Chacón, 2017). La combinación de predicciones entregadas por cada árbol pretende reducir la alta varianza que tiene la respuesta de cada árbol individual ante datos de prueba y por tanto mejorar el desempeño del método.

"Estos métodos son atractivos son atractivos principalmente porque son capaces de impulsar métodos débiles y convertirlos en métodos fuertes, con los cuales pueden hacerse predicciones muy precisas" (Zhou, 2012)

5.3.3.4. Refuerzo de Gradientes Extremo (XGBoost). El XGBoost fue el primero de los tres grandes potenciadores de gradiente en árboles de decisión. Los otros dos son LigthGBM de Microsoft (2016) y CatBoots de Yandex, lanzado en 2017. Todos estos desarrollados para abordar problemas de clasificación o regresión. Teniendo presente de los Bosques Aleatorios, este algoritmo en lugar de plantar simultáneamente una carga de árboles independientes y al azar, cada árbol sucesivo plantado se pondera de tal manera que compensa el impacto de los residuales en el árbol anterior (Kaggle, 2021).

5.3.3.5. Máquinas de Soporte Vectorial (SVM). Las máquinas de soporte vectorial fueron inicialmente aplicadas para problemas de clasificación binaria, sin embargo, su uso está extendido a problemas de clasificación, regresión, selección de variables, identificación de outliers y clustering. El caso más simple, es el caso de un conjunto de datos linealmente separable, el modelo determina el hiperplano que maximiza la distancia de cada grupo a este (Vargas et al., 2012). Cuando los datos no son separables linealmente, SVM recurre a una transformación no lineal de espacio conocida como kernel para llevar los datos a una dimensión mayor y poder separar los datos a través de un a hiperplano como en el caso inicial. La transformación inversa deforma dicho hiperplano en una frontera no lineal que separa o ajusta los datos en el espacio original.

La separación de margen duro no resulta muy flexible para cuando hay ruido en los datos o valores atípicos. Una forma de lograr que el modelo se ajuste sólo por la información relevante es establecer una tolerancia. Para esto se establecen variables de holgura que cuantifican la distancia de puntos alejados a la ecuación de regresión o de la frontera de separación (Pisner & Schnyer, 2020).

En la tabla 2 se muestran las funciones integradas del kernel semidefinidas

Tabla 2. *Funciones integradas de kernel*

Nombre del kernel	g(μ)
Lineal	$G(x_j, x_k) = x_j{}' x_k$
Gaussiano	$G(x_j, x_k) = exp\ (-\left\|\left\|x_j - x_k\right\|\right\|^2)$
Base Radial	$G(x_j, x_k) = exp\ (-\frac{\left\|\left\|x_j - x_k\right\|\right\|^2}{2\sigma^2})$

Nota: Adaptado de "Support vector machine" (Pisner & Schnyer, 2020)

La ecuación generalizada para el problema de regresión usando Máquinas de Soporte Vectorial está expresada en la ecuación 5:

$$f(x) = \langle w, x \rangle + b \tag{5}$$

Donde w es un vector de pesos variables, x las observaciones y b es un valor de variación (bias). Con la finalidad de generalizar el problema y aceptar que pueden existir errores entre los valores pronosticados y la función de regresión, se propuso minimizar la ecuación de riesgo (Scholkopf et al., 1997). Dicho problema de minimización lo podemos observar en la ecuación 6.

$$M_{w,\zeta,\zeta*} = \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{j=1}^{1}(\zeta_j, \zeta_j\ *)$$

$$\{y_j - \langle w, x \rangle - h \leq \varepsilon + \zeta_j\ \langle w, x \rangle + h - y_j \leq \varepsilon + \zeta_j\ *\ \zeta_j, \zeta_j\ * \geq 0 \qquad j = 1, \dots, nj \qquad (6)$$

$$= 1, \dots, n$$

Donde C es un parámetro de penalización que balancea la complejidad del modelo y los errores de entrenamiento $\zeta_j, \zeta_j *$ son variables de holgura que establecen el límite de decisión por arriba o por abajo en el ejemplo de entrenamiento. Para resolver el problema de optimización se introducen multiplicadores de Lagrange $(\alpha_j - \alpha_j *)$. El objetivo es obtener el mejor hiperplano de regresión representado en la ecuación número 7.

$$f(x) = \sum_{j=1}^{1} (\alpha_j - \alpha_j *) \langle x, x_j \rangle + b \tag{7}$$

Los valores de los multiplicadores son calculados a través de métodos de optimización para luego estimar los parámetros w y b. Las observaciones cuyo valor de $a_j, a_j *$, son diferentes a 0 son quieres definen el plano de decisión y se conocen como vectores de soporte.

5.4. Aprendizaje Profundo

El Aprendizaje Profundo permite que los modelos computacionales que se componen de múltiples capas de procesamiento aprendan representación con datos con múltiples niveles de abstracción. el Aprendizaje Profundo descubre estructuras intrincadas en grandes conjuntos de datos mediante el uso del algoritmo de retro propagación para indicar cómo una máquina debe cambiar sus parámetros internos que se utilizan para calcular la representación en cada capa a partir de la representación en la capa anterior. (Bengio et al., 2021)

El Aprendizaje Profundo se refiere a una clase de técnicas de aprendizaje de máquina, esta zona puede considerarse como una intersección entre las zonas de redes neuronales, modelado gráfico, optimización, reconocimiento de patrones y procesamiento de datos. Algunas de las

razones importantes de la popularidad del Aprendizaje Profundo actualmente son el aumento drástico en el procesamiento informático con el uso de tarjetas gráficas (Unidades de Procesamiento Gráfico – GPU), un menor costo de hardware y los avances recientes en la investigación del procesamiento de la información y el Aprendizaje Automático.

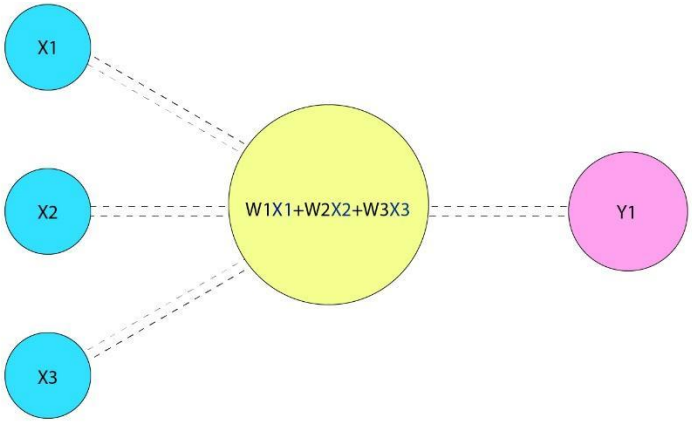
5.4.1. Redes Neuronales

Debido que el campo de la ciencia de datos ha tenido un crecimiento bastante notorio durante la última década, es fácil creer que las primeras investigaciones y trabajos alrededor de las redes neuronales artificiales (ANN) son modernas, sin embargo, las primeras publicaciones rondan entre los años 1943 hasta 1958 por autores como McCulloch, Pitts, Hebb y Rosenblatt.

Las ANN son un método de resolver problemas, de forma individual o combinadas con otros métodos, para aquellas tareas de clasificación, identificación, diagnóstico, optimización o predicción en las que el balance datos/conocimiento se inclina hacia los datos (Izaurieta & Saavedra, 1999), este algoritmo está motivado en modelar la forma de procesamiento de la información en sistemas nerviosos biológicos.

La neurona es la unidad básica de procesamiento que se encuentra en una red neuronal, al igual que las neuronas reales, estas tienen conexiones de entrada a través de los cuales reciben estímulos externos, con estos valores la neurona realizará un cálculo interno y generará un valor de salida, tal cual cómo se observa en la siguiente figura.

Figura 3. *La neurona*



Nota: Adaptado de "Redes Neuronales" (Larrañaga et al., 1997)

Internamente la neurona usa todos los valores de entrada X_1, X_2, X_3 para realizar una suma ponderada, La ponderación de cada una de la entrada viene dada por el peso que se le asigna a cada una de las conexiones de entrada (W_1, W_2, W_3). Es decir, cada conexión que llega a la neurona tendrá asociado un valor que servirá con que intensidad cada variable de entrada afecta el cálculo.

Una red neuronal simple es inherente a limitaciones para la resolución de problemas de separabilidad no lineal, en función de esta necesidad surgió el tipo de red neuronal conocido como Perceptrón Multicapa, el cual es un aproximador universal, en el sentido de que cualquier función continua sobre un espacio de n dimensiones puede aproximarse a través del Perceptrón Multicapa, como un nuevo tipo de función para aproximar o interpolar relaciones no lineales entre datos de entrada y salida (Baldi & Hornik, 1989). La habilidad del perceptrón multicapa para aprender con base en un conjunto de ejemplos, aproximar relaciones no lineales y filtrar ruido de los datos hace que sea un modelo adecuado para abordar problemas reales, sin que esto interfiera para ser uno de los mejores aproximadores universales (Isasi Viñuela & P., 2004).

Hay dos formas diferentes de distribuir las neuronas, la primera manera es colocar dos neuronas en la misma columna o en la misma capa, dos neuronas que se encuentran en la misma capa recibirán la misma información de la capa anterior. La segunda manera es colocar una neurona en frente de otra, de forma secuencial; esto hace que la última neurona procese la información de la capa anterior, haciendo que la red aprenda de forma jerarquizada, entre más capas se añaden, más complejo puede ser el conocimiento que se elabore, dando origen al Aprendizaje Profundo.

Si lo planteamos matemáticamente, cada neurona puede equivalerse a una ecuación de regresión lineal, la conexión de muchas capas de forma secuencial resulta en la concatenación de diferentes ecuaciones de regresión lineal; tanto en las redes neuronales artificiales como en las biológicas, la neurona no sólo transmite la entrada que recibe. Para ajustar esta información de entrada al conocimiento existe un paso adicional, la función de activación, que es análoga a la tasa de potencial de acción disparando en el cerebro. El potencial de acción neuronal se refiere al cambio repentino, rápido, transitorio y que se programa en el potencial de membrana en reposo, solo las neuronas biológicas y los celulares musculares son capaces de generar potencial de acción.

Tal como en las redes biológicas es importante evitar que la red colapse, es necesario que la suma de las capas de como resultado algo diferente a una línea recta. A raíz de esto entran en escena las funciones de activación que permiten cambiar la salida de cada neurona haciendo cambios no lineales, algunos ejemplos de esto son.

5.4.1.1. La Función Escalonada. La salida de la función de activación nos permite realizar una clasificación binaria simple de los valores de entrada de la unidad. La representación matemática está en la Ecuación 8.

$$f(x) = \begin{Bmatrix} 0\ para\ x\ <\ 0 \\ 1\ para\ x\ \geq\ 0 \end{Bmatrix}\ Escalonada \tag{8}$$

5.4.1.2. La Función Sigmoide. Este tipo de funciones permiten mitigar el efecto de outliers (observaciones anormales y extremas en una serie) en el entrenamiento de nuestro modelo. La imagen de este tipo de funciones suele estar contenida en los intervalos o, por lo que valores muy extremos siempre estarán cerca de los límites del intervalo de esa imagen. La representación matemática está en la Ecuación 9.

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}\ Sigmoide \tag{9}$$

5.4.1.3. La Función Tangente Hiperbólica. Este tipo de funciones transforma los valores introducidos a una escala (-1,1), donde los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a -1. Es bastante similar a la función sigmoide, tiende a saturar y matar el gradiente y es bastante usada por su buen desempeño en redes recurrentes. La representación matemática está en la Ecuación 10.

$$f(x) = tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\ TANH \tag{10}$$

5.4.1.4. La Función Relu. También se conocer como "rectificador", esta función transforma los valores introducidos anulando los valores negativos y dejando los positivos tal y como entran. Generalmente se comportan bien con datos de imágenes y tiene un buen desempeño en el uso de redes convolucionales. La representación matemática está en la Ecuación 11.

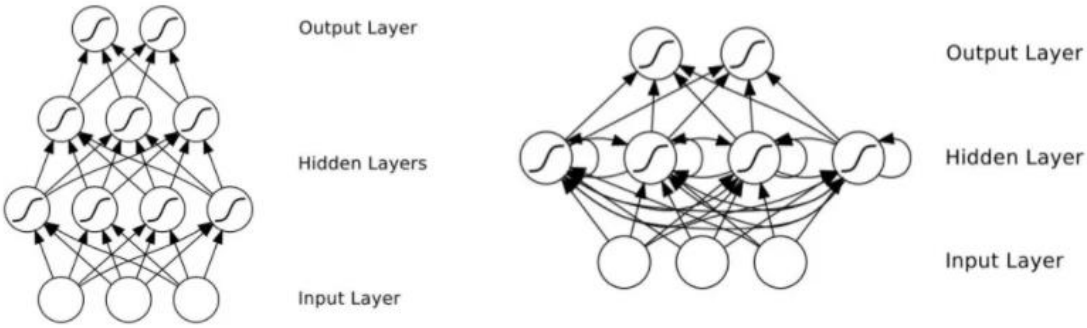
$$f(x) = \begin{Bmatrix} 0\ para\ x\ <\ 0 \\ x\ para\ x\ \geq\ 0 \end{Bmatrix} RELU \tag{11}$$

5.4.2. Redes Neuronales Recurrentes

Las redes neuronales recurrentes son una clase de redes neuronales artificiales donde las conexiones entre las unidades forman un ciclo direccionado. Son modelos de redes neuronales bastante populares que vienen mostrando bastante potencial en diversas áreas de aprendizaje de máquina como procesamiento de lenguaje natural, subtitulado automático de imágenes, traducción y reconocimiento de voz.

La principal idea detrás de las RNN es hacer el uso de informaciones secuenciales, en una red neuronal artificial se asume que todas las entradas y salidas con independientes una de las otras. Sin embargo, para muchas tareas no son la mejor forma de abordar los problemas. Por ejemplo, para predecir la próxima palabra de una oración es necesario las palabras que anteceden la nueva palabra. Se dice que las RNN son recurrentes porque realizan la misma tarea para cada elemento de una oración, siendo la salida dependiente de los resultados anteriores, la siguiente figura ilustra una comparación entra una red feedforward y una red recurrente. (MATTOS PEREIRA, 2017).

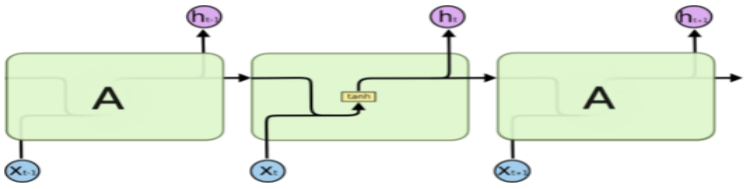
Figura 4. *Comparación de arquitectura de una red feedforward y una red recurrente.*



Nota: Tomado de "Aprendizado profundo: redes lstm" (MATTOS PEREIRA, 2017)

5.4.2.1. Redes LSTM. Las redes LSTM son un tipo especial de RNN, capaces de aprender dependiendo del largo plazo, fueron introducidas inicialmente por el investigador Hochreiter y Schmidhuber (1997) y funcionan muy bien en una grande variedad de problemas, siendo ampliamente utilizadas actualmente. Todas las RNN tienen una forma de cadena de módulos que se repiten en una red neutral. En la RNN, el módulo tiene una estructura bastante simple, por ejemplo, una capa con una función TANH tal como la siguiente figura.

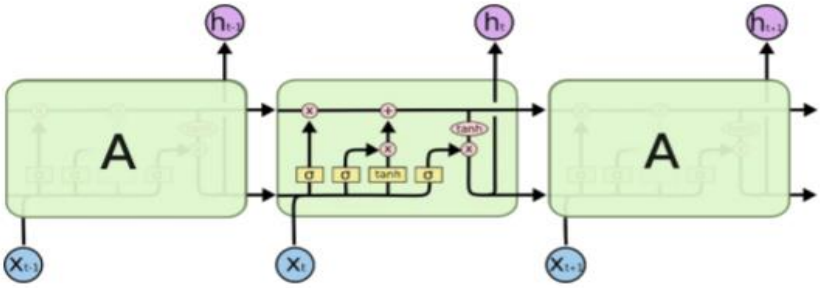
Figura 5. *El módulo que se repite en un RNN contiene solo una capa*



Nota: Tomado de “Aprendizado profundo: redes lstm” (MATTOS PEREIRA, 2017)

Las redes LSTM también tienen esta estructura de cadena, pero el módulo de repetición tiene una estructura diferente. En lugar de tener una sola capa de una red neuronal, hay cuatro, que interactúan de una manera muy específica.

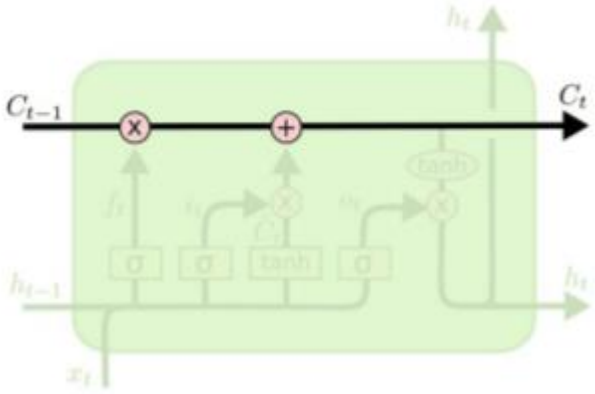
Figura 6. *El módulo que se repite en un LSTM contiene cuatro capas que interactúan*



Nota: Tomado de “Aprendizado profundo: redes lstm” (MATTOS PEREIRA, 2017)

La idea principal de las redes LSTM es crear una representación del estado de la celda, que corresponde a la línea horizontal en la parte superior del diagrama (Figura 7). Este estado de la célula viaja a través de toda la cadena celular, experimentando solo unas pocas interacciones lineales, lo que hace que con los que la información puede fluir sin demasiados cambios.

Figura 7. *El estado de la celda, representado por una línea horizontal en la parte superior del diagrama.*

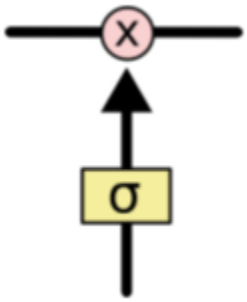


Nota: Tomado de "Aprendizado profundo: redes lstm" (MATTOS PEREIRA, 2017)

Las redes LSTM también tienen la capacidad de eliminar o agregar información al estado de la celda, siendo regulado por estructuras conocidas como puertas (gates).

Las puertas son una forma de permitir que la información fluya en la red neuronal. Ellas están compuestas por una capa sigmoidea de una red neuronal y una multiplicación puntual. En la siguiente figura se ilustra este concepto.

Figura 8. *Ejemplo de una puerta con una capa sigmoidea*



Nota: Tomado de "Aprendizado profundo: redes lstm" (MATTOS PEREIRA, 2017)

Una capa sigmoidea genera números que van de cero a uno, que describen cuánto de cada componente debe pasar por la puerta. Cuanto mayor el valor, mayor la información que transmite esta capa. Una red LSTM tiene tres puertos para proteger y controlar el estado de la celda. (MATTOS PEREIRA, 2017).

5.5. Métricas de desempeño y gráficos de resultados

La optimización y evaluación de modelo se realiza por medio de medidas de error, entre ellas el R^2,MSE, RMSE, MAE y MAPE; así como gráficos de resultados que facilitan visualmente la interpretación de los resultados.

5.5.1. Coeficiente de determinación (R^2)

El coeficiente de determinación es una medida de bondad de ajuste, se interpreta como la proporción de la varianza de la variable dependiente explicada por la variación de las variables independientes; su valor está entre 0 y 1, es adimensional y toma mayor valor a medida que el modelo presenta mejor ajuste (Figuereido, 2011). La presentación matemática está en la ecuación 12.

$$R^2 = \frac{\sum_{i=1}^{N}(A_i - A)^2(P_i - P)^2}{\sum_{i=1}^{N}(A_i - A)^2 \ x \ \sum_{i=1}^{N}(P_i - P)^2} \qquad (12)$$

Donde A_i y P_i son valores medios y pronosticados de los i-ésimo, respectivamente. N es el número de datos de la validación.

5.5.2. Error cuadrático medio (MSE)

El MSE es la función de error más básica y usada en la evaluación de modelos de aprendizaje, calcula el valor medio de los cuadrados de la diferencia entre un valor predicho y un valor observado (Fürnkranz et al., 2011). La presentación matemática está en la ecuación 13.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2 \tag{13}$$

Donde A_i y P_i son valores medios observado y pronosticado respectivamente. N es el número de datos de la validación.

5.5.3. Error cuadrático medio relativo (RMSE).

El RMSE mide la cantidad de error que hay entre dos conjuntos de datos, compara un valor predicho y un valor observado o conocido. El RMSE suele utilizarse junto al MAE para representar el rendimiento del modelo cuando se espera que la distribución de error sea normal (Chai & Draxler, 2014). La presentación matemática está en la ecuación 14.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2} \tag{14}$$

Donde A_i y P_i son valores medios observado y pronosticado respectivamente. N es el número de datos de la validación.

5.5.4. Error absoluto medio (MAE)

Considerando dos series de datos, unos calculados y otros observados, relativos a un mismo fenómeno, el MAE sirve para cuantificar la precisión de una técnica de predicción. A menudo se

requiere una combinación de métricas, por ejemplo, RMSE y MAE, para una mejor evaluación del modelo (Chai & Draxler, 2014). La presentación matemática está en la ecuación 15.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - P_i| \quad (15)$$

Donde A_i y P_i son valores medios observado y pronosticado respectivamente. N es el número de datos de la validación.

5.5.5. Error porcentual absoluto medio (MAPE)

El MAPE es un indicador de desempeño que mide el tamaño del error (absoluto) en términos porcentuales. El error porcentual absoluto medio se usa a menudo en la práctica debido a su interpretación muy intuitiva en términos de error relativo (de Myttenaere et al., 2016).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - P_i|}{A_i} * 100$$

Donde A_i y P_i son valores medios observado y pronosticado respectivamente. N es el número de datos de la validación.

5.5.6. Gráficos de dispersión de pronósticos

Los gráficos de dispersión de pronósticos representan los valores pronosticados frente a los observados, alrededor de la recta de regresión; representa una manera sencilla de observar qué tanto se aleja o se acerca nuestro pronóstico respecto al valor real.

5.5.6. Gráficos de residuales

Los residuales evalúan la diferencia entre los valores observados y pronosticados, el gráfico de residuales muestra los residuos frente a los valores ajustados (pronósticos), con el objetivo de

determinar que los residuos son insesgados y con varianza constante; los residuos deben ubicarse alrededor del cero y sin mostrar algún patrón específico (Elsayir, 2019). El residual se representa por la ecuación 16.

$$e_{ij} = y_{ij} - \hat{y}_{ij} \tag{16}$$

Donde y_{ij} es el valor observado, $\hat{y}_{ij}$ es el valor pronosticado y e_{ij} es el valor del residuo.

5.5.7. Shap Values

Los Shap Values o Shapley Additive Explanations es un método basado teoría de juegos utilizado para la interpretabilidad de los modelos de Aprendizaje Automático, permitiendo el análisis local y global del conjunto de datos. Este método permite cuantificar la contribución de cada característica en la predicción del modelo, y su cambio con valores más altos o bajos (Rodríguez-Pérez & Bajorath, 2020). El cálculo de los Shap Values se representa por la ecuación 17

$$\phi(v) = \frac{1}{|N|!}\sum_{R}[v(P_i^R \cup \{i\}) - v(P_i^R)] \tag{17}$$

Donde ϕ es el valor Shap, N es el número de características, P_i^R es el set del jugador con orden, $v(P_i^R)$ es la contribución del set del jugador con orden y $v(P_i^R \cup \{i\})$ es el aporte del set de jugador con orden y jugador i.

5.6. Validación de modelos

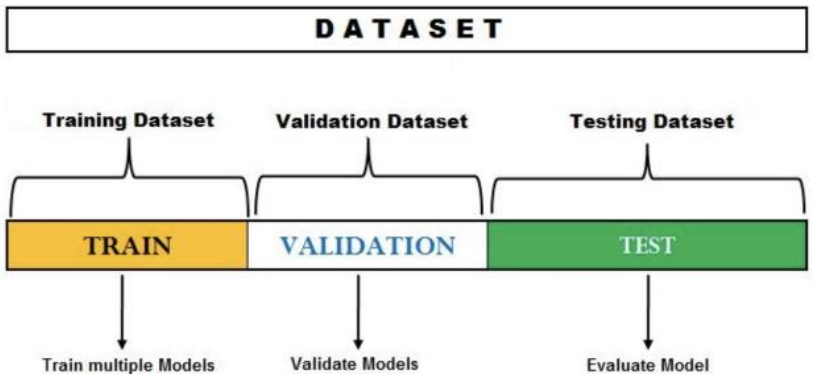
En Aprendizaje Automático, la validación de modelos consiste en el proceso de evaluar el desempeño del modelo entrenado en un conjunto de datos de prueba. El conjunto de datos de prueba o validación es un subconjunto de datos derivado del conjunto de datos de entrenamiento.

La validación del modelo ayuda a encontrar la mejor combinación de hiperparámetros que ayuden al aprendizaje del modelo.

5.6.1. Train, Test, Split

La división del conjunto de datos usando la técnica Train, Test, Split es una variante de la técnica Hold Out, en la cual se toma el conjunto de datos completo y se subdivide en dos subconjuntos, el conjunto de datos de entrenamiento y el de prueba, usualmente esta división suele hacerse 80% de los datos para entrenamiento y 20% para prueba (Vabalas et al., 2019). El modelo se entrena con parte de los datos de entrenamiento y se valida con un subconjunto de datos dentro del set de entrenamiento (Wang & Zheng, 2013).

Figura 9. *División del dataset en entrenamiento, prueba y validación*



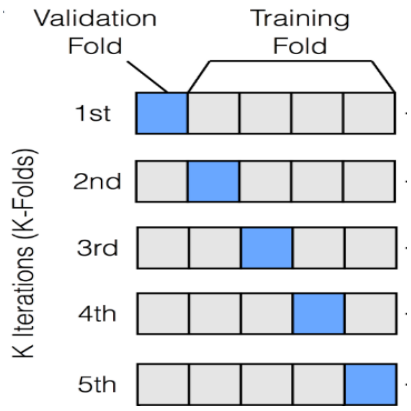
***Nota**:* Tomado de "Model Validation, Machine Learning" (Wang & Zheng, 2013)

5.6.2. Validación cruzada K-folds

La validación cruzada K-folds se usa principalmente para entrenar y validar modelos de Aprendizaje Automático con muestras limitadas y conseguir la mejor selección de hiperparámetros. Esta técnica consiste en dividir el conjunto de observaciones en k grupos o pliegues del mismo tamaño, el primer pliegue corresponde al conjunto de datos de validación y el

modelo se entrena con los k-1 pliegues restantes. El proceso se realiza tantos folds existan, siendo el desempeño del modelo el promedio en las k iteraciones (James et al., 2013).

Figura 10. *Ejemplo validación cruzada*



***Nota**:* Tomado de "Model Validation, Machine Learning" (Wang & Zheng, 2013)

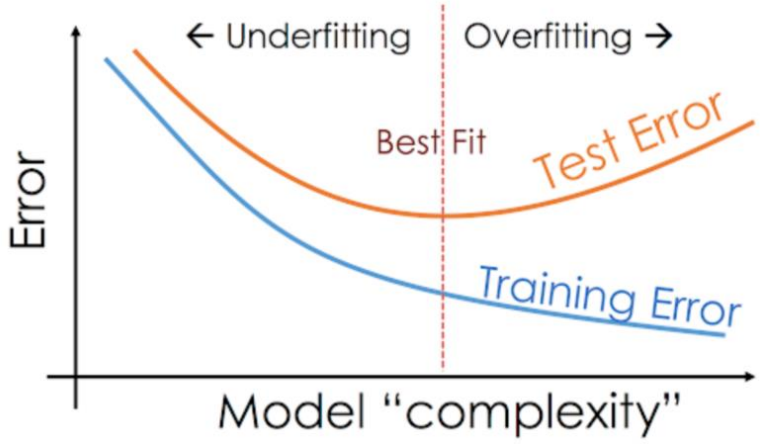
5.6.3. Underfitting y Overfitting

El rendimiento de los modelos de Aprendizaje Automático depende de la capacidad de generalización durante la fase de entrenamiento, donde la máquina puede presentar sobreaprendizaje (overfitting) o subaprendizaje (underfitting) (Badillo et al., 2020).

Overfitting. El sobreajuste se presenta cuando el error de entrenamiento es mucho menor que el error de validación, esto se puede deber a utilizar demasiadas características en el entrenamiento o a un bajo número de muestras.

Underfitting. El subajuste se presente cuando el modelo no puede reducir el error de entrenamiento, esto se puede deber a que el conjunto de entrenamiento contiene menos observaciones que variables, un modelo poco complejo u observaciones limitadas.

Figura 11. *Explicación gráfica underfitting y overfitting*



***Nota**:* Adaptado de "An Introduction to Machine Learning (Badillo et al., 2020)

5.6.4. Test de normalidad Shapiro Wilk

La prueba de Shapiro Wilk evalúa la normalidad en los datos cuando n<50, busca probar la hipótesis nula (Tapia et al., 2021).

$$H_0: los\ datos\ siguen\ una\ distribución\ normal$$

$$H_1: los\ datos\ no\ siguen\ una\ distribución\ normal$$

5.6.5. ANOVA de un factor

El Análisis de la Varianza de un factor es una técnica estadística que compara la media de un factor para múltiples muestras, analizando si existen diferencias significativas entre las medias de los grupos. El resultado del ANOVA es el estadístico F que muestra la diferencia entre la varianza dentro del grupo, de existir diferencias significativas el ratio F será mayor y el valor p será más bajo (Ostertagova et al., 2013). La hipótesis nula que busca probar el ANOVA es

$$H_0: \mu_1 = \mu_2 = \mu_k$$

$$H_a: al\ menos\ una\ media\ es\ diferente$$

5.7. Optimización de modelos

La optimización de modelos es el proceso de encontrar los valores que generen la mejor solución posible para un problema dado bajo un conjunto definido de restricciones. En Aprendizaje Automático se aplican diversas técnicas, métricas y funciones con el objetivo de encontrar la mejor combinación de parámetros que mejoren el aprendizaje del modelo con el mínimo error posible (Touretzky, 2006).

5.7.1. Análisis de correlaciones

La matriz de correlación de Pearson permite visualizar gráficamente el coeficiente de correlación de Pearson, el cuál mide el grado de asociación entre dos variables continuas cuando la asociación es lineal. Esta métrica toma valores en el rango -1 a +1, valores negativos indican una asociación inversa y valores positivos sugieren una asociación directa (Samuels & Gilchrist, 2014). Se representa por la ecuación número 17.

$$r = \frac{n * \sum Xi * Yi - \sum Xi * Yi}{\sqrt{[n * \sum Xi^2 - (\sum Xi)^2] * [n * (\sum Yi)^2 - (\sum Yi)^2]}} \quad (17)$$

Donde Xi son las observaciones de cada característica, Yi son las observaciones de la variable respuesta y n es el número de registros.

5.7.2. Normalización de datos

La normalización de datos es una técnica que permite reducir la varianza al hacer que las variables sean adimensionales, este método consiste en restarle a cada valor del conjunto de datos su media y dividirla en su desviación estándar, según la ecuación 17.

$$z = \frac{x - \mu}{\sigma} \quad (17)$$

Donde x es el valor observado, μ es la media de la distribución y σ es la desviación estándar de la distribución (Jamal et al., 2014)

5.7.3. Reducción de la dimensionalidad

Las técnicas de reducción de la dimensionalidad se refieren al proceso de disminuir el número de dimensiones o características de un conjunto de datos, dentro de sus técnicas se incluyen la extracción y combinación de características, destacando métodos como el Backward Elimination y el Análisis de Componentes Principales – PCA (el Bouchefry & de Souza, 2020).

5.7.3.1. Multicolinealidad. La multicolinealidad consiste en una fuerte dependencia lineal entre las variables independientes, generando una no estimación única de los parámetros y falsas relaciones entre la variable dependiente y los regresores, dando como resultado inferencias estadísticas poco precisas. Para hacer frente a este problema se suelen usar técnicas de selección de características como el Factor de Inflación de la Varianza y el Backward Elimination (Guerrero & Melo, 2017).

5.7.3.2. Factor de Inflación de la Varianza (VIF). El Factor de Inflación de la Varianza (VIF) permite detectar problemas severos de multicolinealidad, cuantifica la influencia de la varianza de una variable independiente por su interacción con las demás. En términos generales, valores de VIF mayores a 5 sugieren que las variables independientes involucradas están altamente correlacionadas (Cuthbert, 1981). El VIF se representa por la ecuación 18

$$VIF_i = \frac{1}{1 - {R_i}^2} \tag{18}$$

Donde ${R_i}^2$ representa el coeficiente de determinación no ajustado para la regresión de la i-ésima variable independiente sobre las demás.

5.7.3.3. Backward Elimination. El Backward Elimination es un método utilizado en el análisis de regresión para seleccionar un subconjunto de características explicativas significativas para el modelo. La técnica recibe un modelo de regresión y calcula a un nivel de significancia del 5% la significancia estadística de las variables independientes, removiendo en cada iteración el predictor menos significativo (valor p más alto) hasta que todos los predictores del modelo sean significativos (Eom et al., 2020).

5.7.3.4. Forward Selection. El Forward Selection es un método utilizado para evaluar el desempeño de un modelo conforme se agregan variables. Consiste en un tipo De regresión que inicia con un modelo vacío y agrega en cada paso la única variable que genera la mejor mejora individual al modelo (Marcano-Cedeno et al., 2010).

5.7.4 Ajuste de hiperparámetros (GridSearchCV)

Los hiperparámetros son los valores de las configuraciones ajustables de un modelo para controlar el proceso de entrenamiento (Microsoft, 2022). Una de las técnicas utilizadas es la búsqueda en rejilla o “GridSearch”, la cuál es una búsqueda exhaustiva que prueba todas las combinaciones de valores de hiperparámetros especificadas en el conjunto de datos de entrenamiento y validación; los mejores hiperparámetros serán aquellos que generen la mejor métrica de desempeño evaluada, por ejemplo, el coeficiente de determinación (Wu et al., 2019).

5.7.5. Función de error

La función de error o función de pérdida es un método utilizado en Aprendizaje Automático para evaluar la efectividad de un algoritmo para modelar los datos, una alta desviación de los valores reales arrojaría un número grande en la función de pérdida (Parmar, 2018). Las funciones de pérdida más usadas son el error cuadrático medio y el error absoluto medio. La función de pérdida

aprende gradualmente a reducir el error en la predicción utilizando algún optimizador (Seifert & Rasp, 2020).

5.7.6. Optimizadores

Los optimizadores son métodos utilizados en Aprendizaje Automático para modificar los atributos del modelo, como los pesos y la tasa de aprendizaje, con el objetivo de minimizar la función de pérdida general y mejorar la precisión del modelo; algunos optimizadores son el Gradiente Descendente (Stochastic, Minibatch), Adam, Momentum, AdaGrad y RMSProp (Choi et al., 2019).

5.7.7. Early Stopping

El Early Stopping o parada temprana es una forma de regularización utilizada para evitar el sobreaprendizaje (overfitting), esta técnica consiste en detener el entrenamiento del modelo en el instante en que el rendimiento del conjunto de datos de entrenamiento aumenta y el del conjunto de validación tiende a empeorar (Orr & Müller, 1998).

5.7.8. Tamaño de lote y épocas

En redes neuronales, el tamaño de lote o batch size hace referencia al número de submuestras de datos de entrenamiento para la entrada, un tamaño de lote pequeño acelera el proceso de aprendizaje y uno grande aumenta la precisión del modelo. Por su parte, las épocas son el número de veces que el conjunto de datos completo pasa, hacia adelante y hacia atrás, a través de un modelo de red neuronal (Veeramsetty et al., 2020).

6. Metodología

Los algoritmos y el código utilizado para el desarrollo de esta fase se encuentran en: https://github.com/Dilson1502/Modelo-Predictivo-Para-Rendimiento-De-Cultivos

6.1. Descripción del conjunto de datos

Los datos son la materia prima que alimenta los algoritmos de Aprendizaje Automático e Inteligencia Artificial. Para efectos del presente proyecto, el conjunto de datos se construyó a través de la unión de diversas fuentes de datos de carácter oficial de instituciones del país.

El Instituto de Hidrología, Meteorología y Estudios Ambientales cuenta a su disposición con la Red de Estaciones Meteorológicas a lo largo del territorio nacional. Su propósito desde su creación ha sido incentivar proyectos que vayan en pro de la importancia de administrar e inventariar los recursos naturales de origen meteorológico para apoyar la toma de decisiones (Bernal, 1978).

Actualmente Colombia cuenta con 8.973 estaciones a lo largo del territorio, del cual el 61.19% se encuentran activas y en funcionamiento, 38.32% se encuentran suspendidas y el 0.47% en mantenimiento, por otra parte, los departamentos con mayor proporción de estaciones activas son Cundinamarca (14.64%), Antioquia (12.98%), Valle del Cauca (10.50%), Boyacá (5.93%), Caldas (5.88%), Santander (4.97%). En la figura 15 a continuación podemos observar la cantidad de estaciones existentes por tipo de medida.

Tabla 3. *Estaciones Meteorológicas por tipo en Colombia*

Tipo de estación	Cantidad	Descripción Estación
Pluviométrica	2074	Precipitación
Limnimétrica	898	Altura del agua

Continuación Tabla 3

Estaciones Meteorológicas por tipo en Colombia

Limnigráfica	**853**	**Nivel de corriente hídrica**
Climática Principal	576	Precipitación, temperatura del aire, temperaturas máxima y mínima a 2 metros, humedad, viento, radiación, brillo solar, evaporación, cantidad de nubes y fenómenos especiales
Climática Ordinaria	401	Precipitación, temperatura del aire, temperaturas máxima y mínima a 2 metros y humedad.
Meteorológica Especial	94	Seguimiento fenómenos especiales Ejemplo: Heladas
Agrometeorológica	57	Temperatura a diferentes profundidades y alturas y las mismas que una estación climática principal
Mareográfica	31	Nivel, temperatura y salinidad de las aguas marinas.
Sinóptica Principal	31	Nubosidad, dirección y velocidad de los vientos, presión atmosférica, temperatura del aire, tipo y altura de las nubes, visibilidad, fenómenos especiales, características de humedad, precipitación, temperaturas extremas, capas significativas de nubes, recorrido del viento y secuencia de los fenómenos atmosféricos

Continuación Tabla 3

Estaciones Meteorológicas por tipo en Colombia

Radio Sonda	11	Temperatura del aire, presión atmosférica, humedad relativa y dirección y velocidad del viento en las capas altas de la atmósfera (tropósfera y baja estratósfera)
Sinóptica Secundaria	5	Visibilidad, fenómenos especiales, tiempo atmosférico, nubosidad, estado del suelo, precipitación, temperatura del aire, humedad del aire, presión y viento.

Para el desarrollo de la presente investigación se utilizaron los registros climatológicos para el periodo comprendido entre los años 2007 al 2021, información proporcionada por el IDEAM descargada a través de una API que permitió realizar la descarga por promedio anual para cada una de las variables con excepción de la precipitación que se filtró por la sumatoria de la lluvia total al final del año; se consideraron humedad relativa, precipitación, presión atmosférica, temperatura media, temperatura máxima, temperatura mínima, velocidad del viento y área sembrada. Como variable predictora se consideraron los rendimientos agrícolas para los cultivos de cacao y café en Colombia proporcionados por las Evaluaciones Agropecuarias Municipales, dependencia del Ministerio de Agricultura y Desarrollo Rural. Asimismo, se consideraron condiciones geográficas como el área sembrada y altura sobre el nivel del mar, de las cabeceras municipales presentes en el estudio. Los conjuntos de datos resultantes contienen 750 registros para el cultivo de cacao y 747 registros para el cultivo de café.

En el apéndice C se encuentra la descripción de las variables consideradas en esta investigación.

6.2. Limpieza y Preprocesamiento de datos

La eficacia de los algoritmos de extracción de conocimiento depende de gran medida de la calidad de los datos, la cual puede ser garantizada por algoritmos de preprocesamiento y limpieza (Herrera, 2016). La preparación de datos según el autor García, S (2015), está formada por una serie de técnicas que tienen el objetivo de inicializar correctamente los datos que servirán de entrada para los algoritmos de minería de datos. La descripción de los datos climáticos de entrada muestra que no en todos los municipios se miden las mismas variables climatológicas, por lo cual se hace necesario realizar una agrupación de datos y la imputación de valores nulos. Se utilizó una imputación por la mediana de cada municipio de cada departamento para todas las variables climáticas, excepto para la temperatura máxima y mínima, que se imputaron con la medición máxima y mínima de cada municipio de cada departamento, esto se debe a que al utilizar otro estimador se pueden presentar datos incoherentes, por ejemplo, que la temperatura máxima sea menor que la temperatura mínima, y viceversa. Posteriormente se eliminaron los registros donde luego de la imputación continuaron los valores nulos, dado que esto indica que no existe ninguna medición de la variable climática observada para dicha combinación de departamento-municipio. Finalmente se descargaron los sets de datos de rendimientos de cultivos de cacao y café para los municipios que contaran con registros de clima y se indexaron con la altura sobre el nivel del mar por separado.

6.3. Análisis exploratorio de Datos

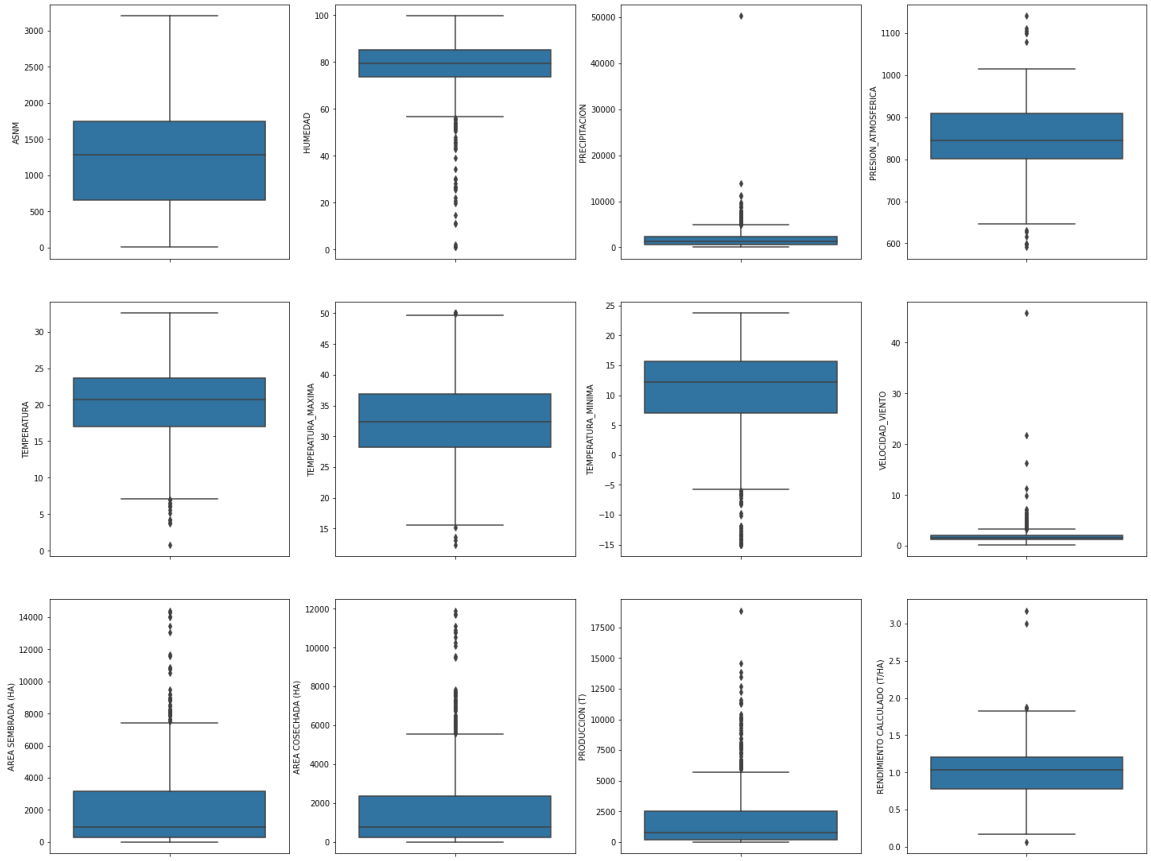
El análisis exploratorio de Datos (EDA) es un proceso utilizado para descubrir patrones, detectar anomalías y validar hipótesis con ayuda de estadísticas resumen y gráficos (Chatfield, 1986). Con el objetivo de realizar inferencias previas a la implementación de modelos se realizó un EDA que

comprende estadísticas descriptivas, análisis de correlaciones, gráfico de distribución de probabilidad y matriz de correlación del dataset.

6.3.1. Boxplots

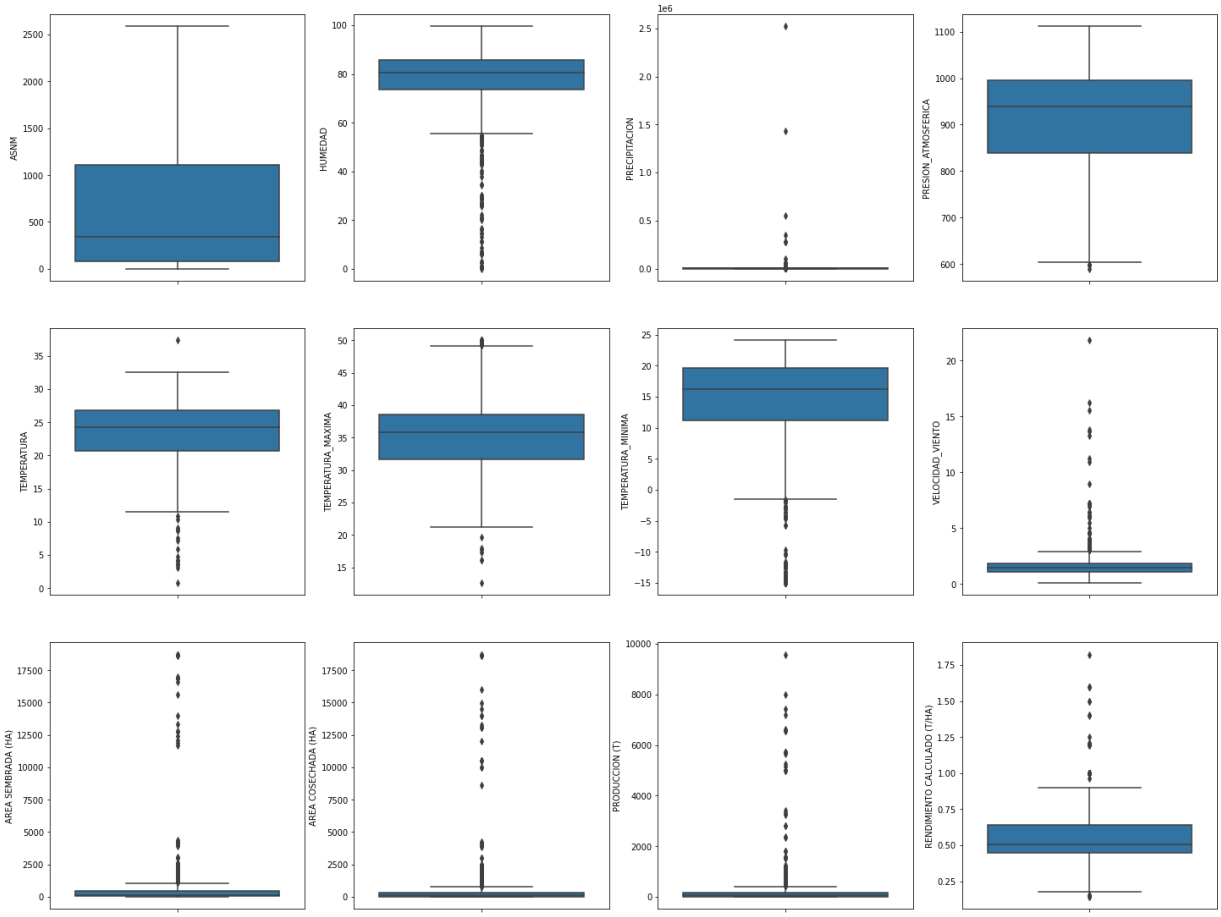
De forma gráfica, los boxplots o diagramas de caja y bigotes permiten visualizar los valores atípicos, comparar las distribuciones e intuir la simetría de cada variable del conjunto de datos. De forma gráfica, se observa mediante los cuartiles la concentración de valores en el 25%, 50% y 75% del total de datos (Edwards et al., 2017). Para ambos conjuntos de datos se observan valores atípicos para la Humedad Relativa, la Precipitación Acumulada, Presión Atmosférica, la Velocidad Promedio del Viento, Temperatura y el Área Sembrada; asimismo, una alta similitud en la distribución de los datos de las variables Área Sembrada, Área Cosechada y Producción, lo cual sugiere a priori una alta correlación entre ellas.

Figura 12. *Boxplots dataset café*



Nota: Tomado de salida del algoritmo en Python

Figura 13. *Boxplots dataset cacao*



Nota: Tomado de salida del algoritmo en Python

6.3.2. Detección de outliers

Los valores atípicos u outliers son puntos de datos diferentes (más grandes, más pequeños) a la distribución del conjunto de datos total, estas observaciones anormales pueden distorsionar la distribución de probabilidad; su existencia se puede deber a mediciones inconsistentes u observaciones equivocadas. Para la detección de outliers se utilizó la regla de Tukey's, para ello se calculó el porcentaje de datos que está por debajo y por encima de 1.5 veces el Rango Intercuartílico IQR (Cuartil 3-Cuartil 1) (Seo, 2006), para los efectos de esta investigación, se

centra la atención en valores atípicos mayores al 10%. Para el café los resultados arrojan valores atípicos en la variable Precipitación. Para el cacao se presentan valores atípicos para las variables: Área Sembrada, Área cosechada y producción; es importante indagar si estos atípicos se deben a errores en la medición o verdaderamente condiciones extremas de los puntos de datos en el momento de la toma. A continuación, se observa el porcentaje de valores atípicos para cada uno de los conjuntos de datos.

Tabla 4. *Porcentaje de outliers por variable*

Variable	% de atípicos dataset café	% de atípicos dataset cacao
ASNM	0.00 %	0.00%
Humedad	5.35 %	8.93%
Precipitación	10.84%	8.00%
Presión Atmosférica	1.87%	0.67%
Temperatura	1.87%	2.27%
Temperatura Máxima	1.74%	4.40%
Temperatura Mínima	5.89%	8.93%
Velocidad Viento	9.10%	9.20%
Área Sembrada	5.76%	11.73%
Área Cosechada	7.63%	12.67%
Producción	7.50%	12.67%
Rendimiento Calculado	0.80%	6.53%

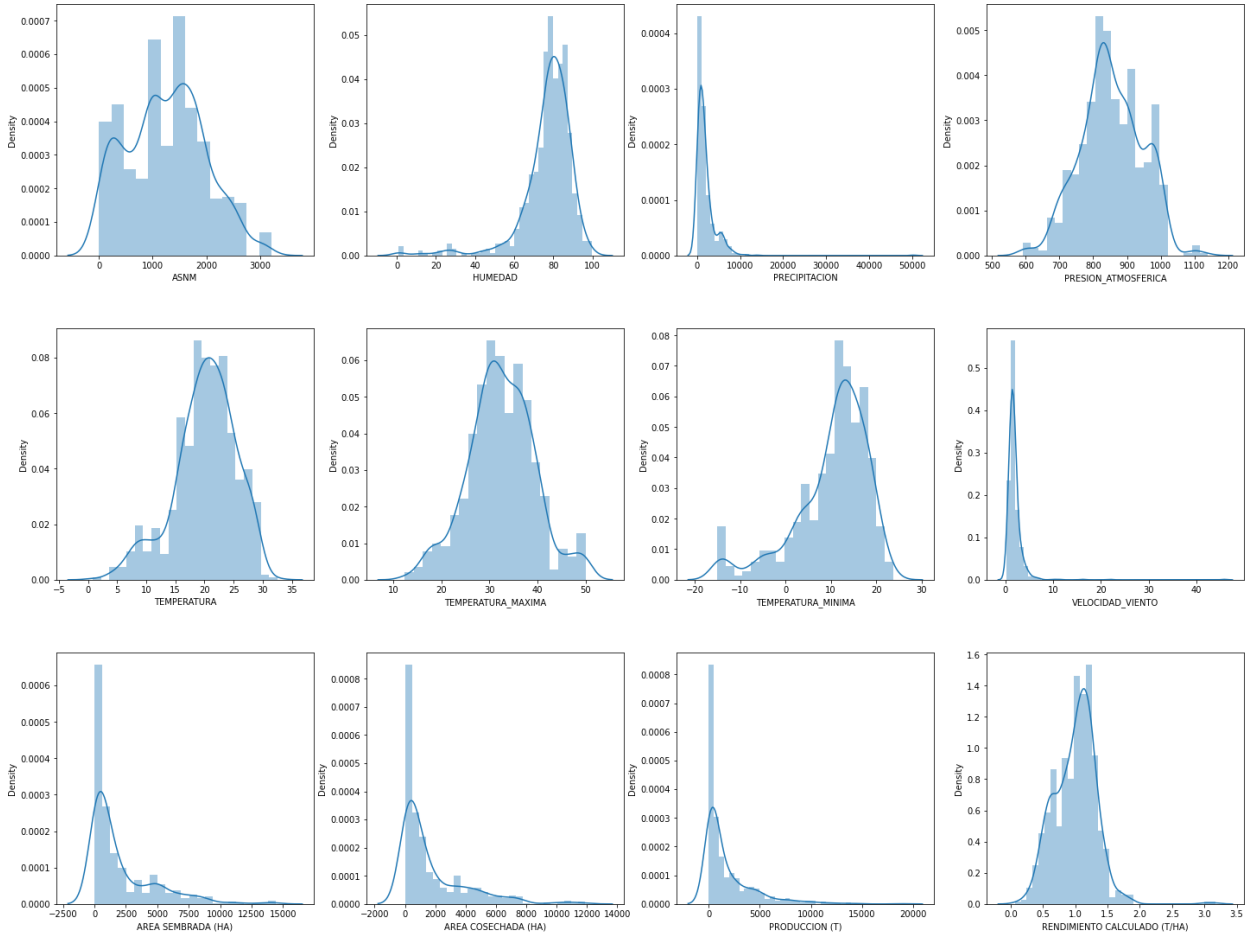
Nota: Adaptado por los autores a partir de Python

6.3.3. Distribución de frecuencia del dataset

La distribución de frecuencias se realizó mediante histogramas, los cuáles son gráficos que permiten visualizar la concentración de los datos en intervalos, el sesgo, la simetría e intuir la

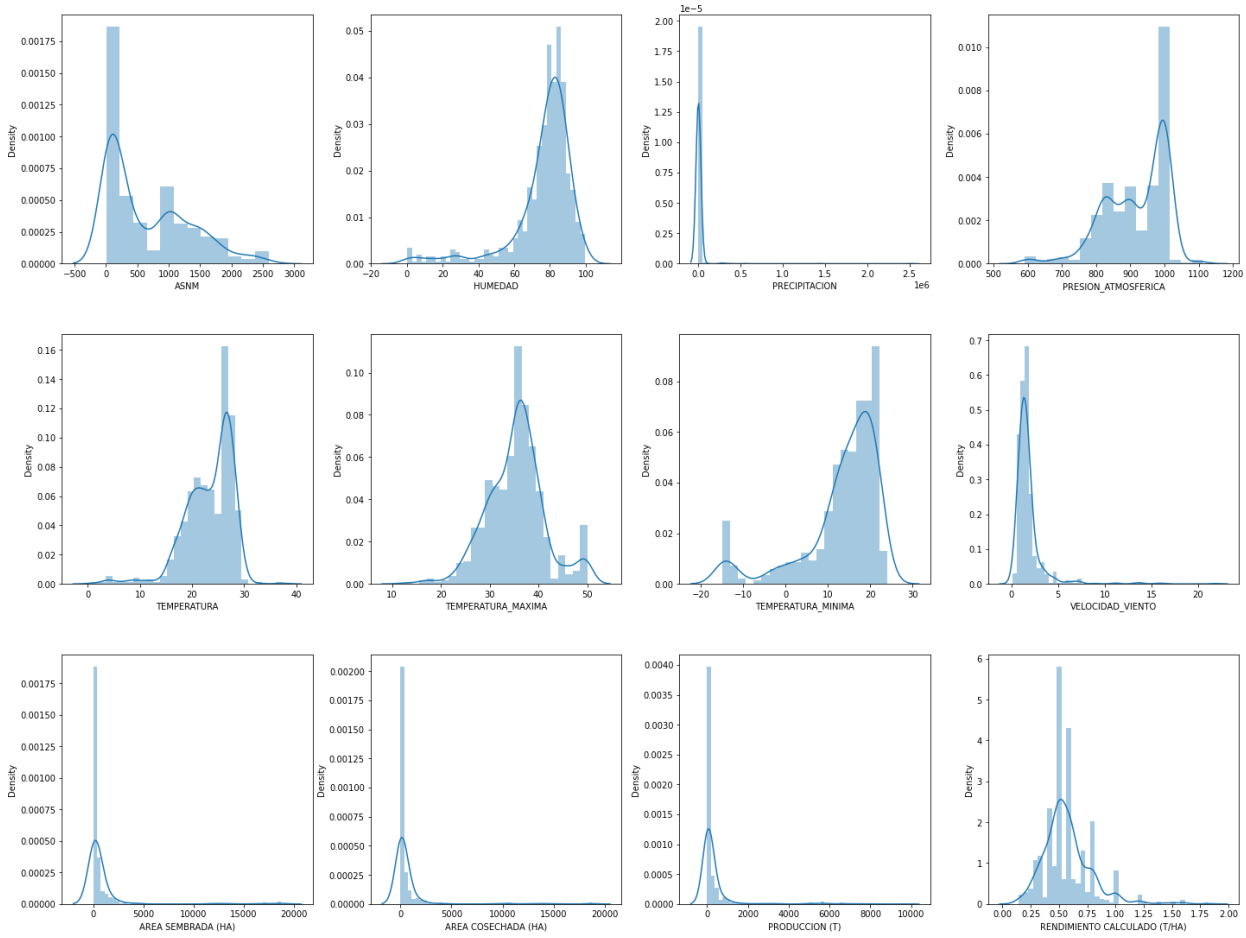
cantidad de modas en la muestra (Weitz, 2020). Para los dos cultivos los histogramas permiten observar que las columnas Humedad Relativa, Precipitación Acumulada y Velocidad Promedio del Viento tienen distribuciones altamente sesgadas; centralmente para el cacao la variable presión atmosférica promedio posee un sesgo a la izquierda. Para ambos gráficos las variables Área Sembrada, Área cosechada y Producción siguen aparentemente la misma distribución de probabilidad, un análisis de correlaciones será necesario para tratar estas variables; el Rendimiento parece seguir una distribución normal.

Figura 14. *Distribuciones de probabilidad dataset café*



Nota: Tomado de salida del algoritmo en Python

Figura 15. *Distribuciones de probabilidad dataset cacao*



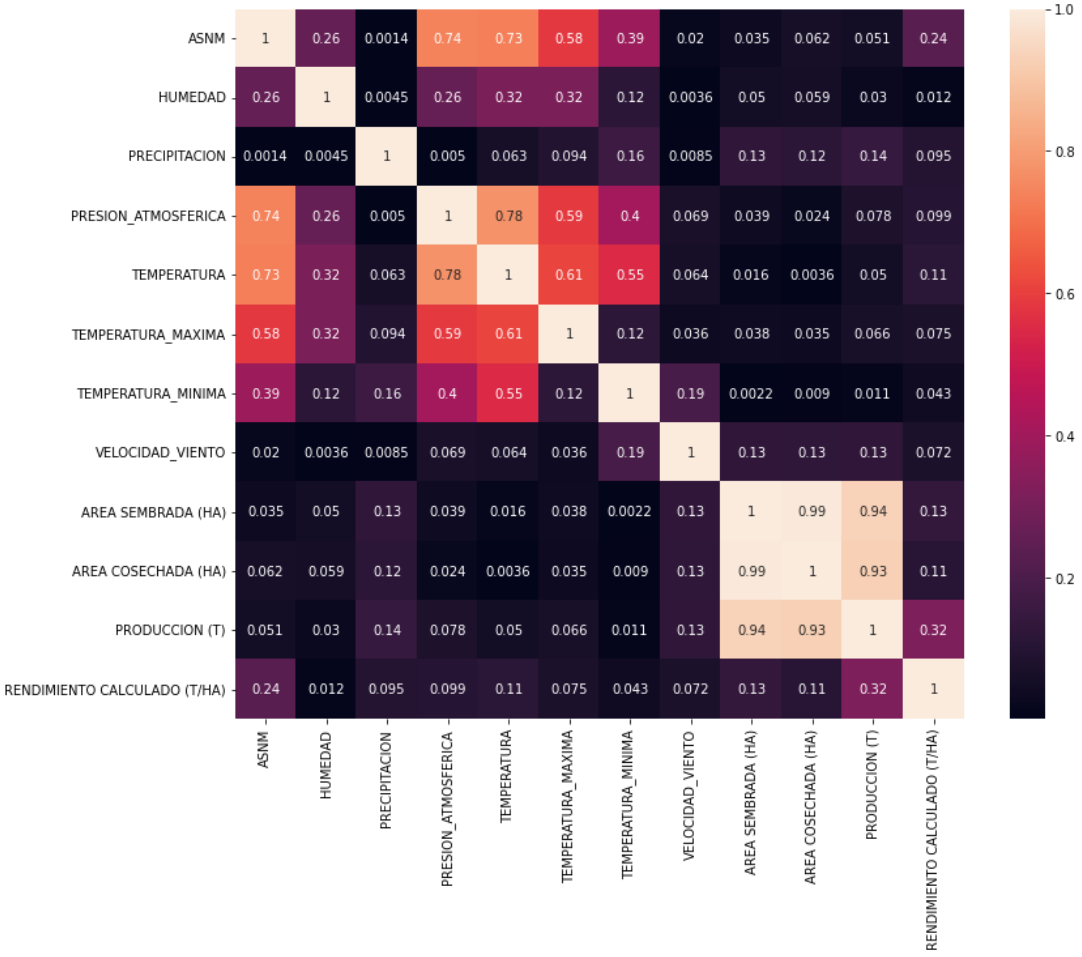
Nota: Tomado de salida del algoritmo en Python

6.3.4. *Análisis de correlación del dataset*

Se construye la matriz de correlación de Pearson con el objetivo de identificar relaciones de asociación entre las variables del conjunto de datos. En general se identifica una alta correlación entre los pares de variables Altura Sobre el Nivel del mar – Presión Atmosférica, Altura Sobre el Nivel del mar – Temperatura Media, Altura Sobre el Nivel del mar – Temperatura Máxima, Temperatura Media – Rendimiento. Se considerará además como predictor la variable Área Sembrada, la cual posee una fuerte correlación con el Área Cosechada y la Producción (0.99 y

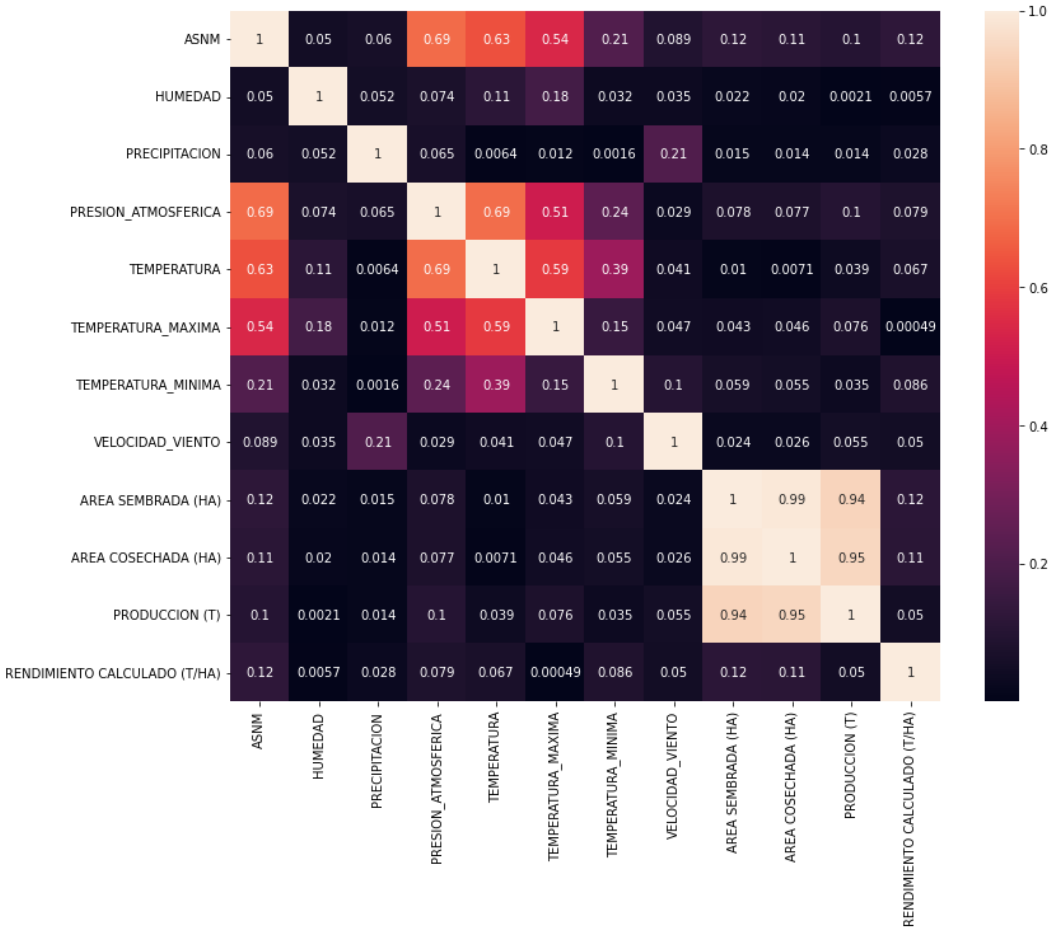
0.94 respectivamente) para los dos conjuntos de datos. En este estudio se opta por predecir la variable dependiente Rendimiento, medido en toneladas de producto cosechada por hectárea sembrada, cuyo coeficiente de correlación con la variable Área Sembrada es de 0.13 para café y 0.12 para cacao.

Figura 16. *Tabla de correlaciones entre variables dataset café*



Nota: Tomado de salida del algoritmo en Python

Figura 17. *Tabla de correlaciones entre variables dataset cacao*



Nota: Tomado de salida del algoritmo en Python

6.4. Conjunto de datos de entrada

Investigaciones similares realizaron predicciones del Área Cosechada y Producción, junto al Rendimiento (García-Arteaga et al., 2020), el análisis de correlaciones sugiere que predecir sobre el Área Cosechada y la Producción podría opacar el efecto de las demás variables sobre la respuesta, por lo anterior, se decide utilizar como variable dependiente únicamente el Rendimiento y como predictores el conjunto completo de variables independientes. Se dividió el conjunto de datos en 80% para entrenamiento y 20% para prueba utilizando la función train, test, split de la

librería Sklearn de Python (Vrigazova, 2021). Posteriormente, con el objetivo de disminuir la varianza, se normalizaron los datos de entrada con ayuda de la función StandardScaler de Sklearn.

6.4.1. Multicolinealidad

Se realizó un análisis profundo de las relaciones de multicolinealidad para evaluar las interacciones entre las variables independientes y su efecto en el aprendizaje de los modelos.

6.4.3.1. Factor de Inflación de la Varianza. Se calculó el Factor de Inflación de la Varianza (VIF) para el conjunto de datos de entrada, como criterio de medida se consideraron valores de VIF mayores a 5 indican relaciones fuertes de multicolinealidad entre pares de variables. Para los conjuntos de datos, se observa que no existen relaciones de colinealidad significativas.

Tabla 5. *Factor de inflación de la varianza por variable*

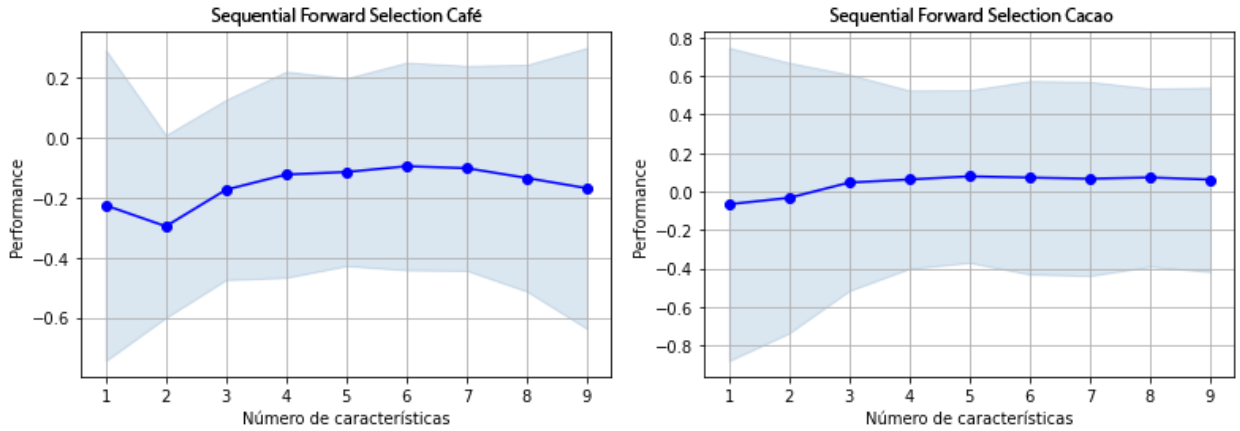
Variable	VIF café	VIF cacao
ASNM	2.638472	2.358117
Humedad	1.168319	1.038452
Precipitación	1.058136	1.083540
Presión Atmosférica	2.797112	2.323597
Temperatura	3.886554	2.872974
Temperatura Máxima	2.091528	2.176489
Temperatura Mínima	1.605048	1.572662
Velocidad Viento	1.113500	1.120613
Área Sembrada	1.080888	1.080000

Nota: Adaptado por los autores a partir de Python

6.4.4.2. Backward Elimination. Se utilizó la técnica Backward Elimination por medio de Bosques Aleatorios, la técnica sugiere que para el café las características significativas son la Altura Sobre el Nivel del Mar, Precipitación acumulada, Presión Atmosférica, Temperatura Promedio y Velocidad del Viento. Para el cacao las características significativas son la Altura Sobre el Nivel del Mar, Humedad, Temperatura Mínima, Velocidad del Viento y Área Sembrada.

6.4.4.3. Sequential Forward Selection. Se utilizó la técnica Forward Selection con el objetivo de evaluar el impacto de agregar o quitar variables del conjunto de datos de entrada. Cómo se observa en la figura 17 con 6 y 5 variables significativas se alcanza el mejor ajuste en el entrenamiento para el café y cacao respectivamente, no obstante, el incluir la totalidad de las variables no impacta drásticamente el desempeño, pero si otorga mayor interpretabilidad del conjunto de datos completo. Según estos análisis se decidió considerar las 9 variables independientes como parte del conjunto de datos de entrada para los modelos.

Figura 18. *Sequential Forward Selection*



Nota: Tomado de salida del algoritmo en Python

6.5. Implementación de modelos

Para la implementación de modelos se realizó la búsqueda de los mejores hiperparámetros en el entrenamiento y posteriormente se calcularon las métricas de desempeño para cada uno de los modelos de Aprendizaje Automático y Aprendizaje Profundo propuestos.

6.5.1. Implementación de modelos de Aprendizaje Automático y búsqueda de hiperparámetros.

Dentro de los modelos de Aprendizaje Automático se utilizaron la Regresión Lineal Múltiple, Árbol de Decisión, XGBoost, Bosques Aleatorios y Máquinas de Soporte Vectorial. Se realizó la búsqueda exhaustiva de hiperparámetros utilizando la Búsqueda en rejilla (cuadrícula) con ayuda de la librería GridSearchCV de Sklearn en Python, la mejor combinación de parámetros fue aquella que mejor ajuste generara en el conjunto de datos de entrenamiento, haciendo uso de la validación cruzada con k=5 folds. Los hiperparámetros seleccionados para la búsqueda en cada modelo se especifican a continuación.

Tabla 6. *Hiperparámetros seleccionados modelos ML*

Modelo	Modelo café	Modelo cacao
Linear Regression	No se hizo búsqueda de hiperparámetros	
Decision Tree	min_samples_split: [6] max_depth: [1] max_features: ['auto'] max_leaf_nodes: [10]	min_samples_split: [6] max_depth: [1] max_features: ['s'] max_leaf_nodes: [10]
Support Vector Regression	kernel: ['rbf'] gamma: ['auto'] C: [0.1]	kernel: ['rbf'] gamma: ['scale'] C: [1]
XGBoost	min_samples_split: [4] max_depth: [100] max_features: ['log2'] max_leaf_nodes: [100] n_estimators: [20]	min_samples_split: [2] max_depth: [100] max_features: ['log2'] max_leaf_nodes: [100] n_estimators: [20]
Random Forest	min_samples_split: [8] max_depth: [10] max_features: ['log2'] max_leaf_nodes: [100] n_estimators: [300]	min_samples_split: [2] max_depth: [100] max_features: ['log2'] max_leaf_nodes: [500] n_estimators: [500]

6.5.2. Implementación de modelos de Aprendizaje Profundo y búsqueda de hiperparámetros

Para la implementación de modelos de Aprendizaje Profundo se implementó el Perceptrón Multicapa y la red LSTM, la búsqueda de hiperparámetros se hizo exhaustivamente variando el

número de capas, neuronas, funciones de activación y optimizadores, de los cuales se seleccionó la mejor combinación de hiperparámetros como se muestra a continuación.

La arquitectura del Perceptrón Multicapa se diseñó de la siguiente manera: capa de entrada [128 neuronas, activación = sigmoide], capas intermedias [64,30,20,5,5,10,15,5] neuronas, con activación relu en cada capa, capa de salida [1 neurona, activación = lineal]. La función de pérdida utilizada fue el error cuadrático medio, el optimizador Adam y la métrica de desempeño el error cuadrático medio. El modelo compilado se entrenó con el 80% y se validó con el 20% del total de datos de entrenamiento. Se utilizaron 150 épocas, un batch size = 220 y verbose =1. Los hiperparámetros seleccionados se tomaron realizando pruebas de ensayo y error y observando el comportamiento del ajuste y las funciones de pérdida.

Figura 19. *Gráfico de función de perdida y desempeño redes neuronales profundas café*

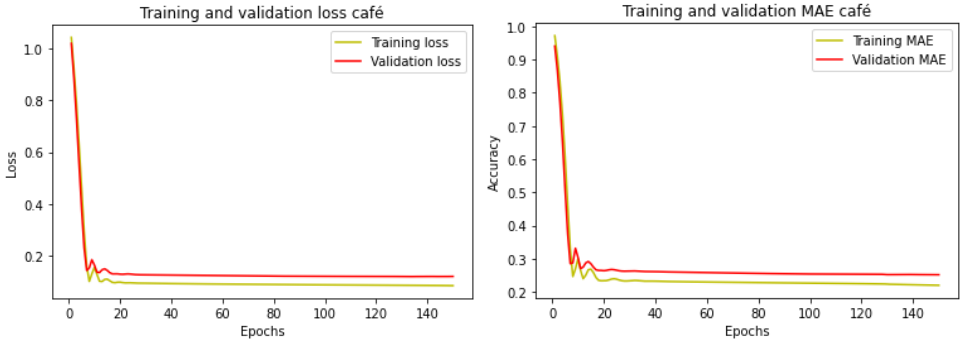
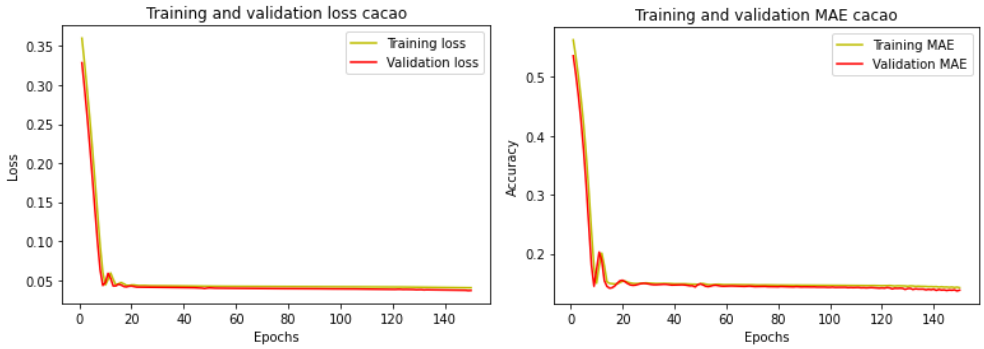


Figura 20. *Gráfico de función de perdida y desempeño redes neuronales profundas cacao*



Nota: Tomado de salida del algoritmo en Python

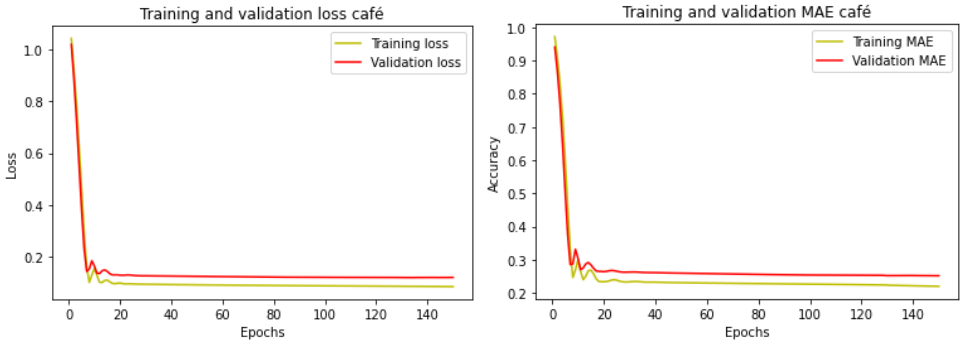
Los gráficos sugieren que después de 20 épocas la función de pérdida del error y el desempeño del ajuste se estabilizan.

Red Neuronal LSTM

La arquitectura de la red LSTM se diseñó de la siguiente manera: capa de entrada [tipo: LSTM, 10 neuronas, activación = relu], capa intermedia 1 [tipo: densa, 5 neuronas, activación = relu], capa intermedia 2 [tipo: LSTM, 10 neuronas, activación = relu], capa intermedia 3 [tipo: densa, 5 neuronas, activación = relu], capa intermedia 4 [tipo: densa, 10 neuronas, activación = relu], capa de salida [1 neurona, sin activación]. La función de pérdida utilizada fue el error cuadrático medio, el optimizador Adam y la métrica de desempeño el error cuadrático medio. El modelo compilado se entrenó con el 80% y se validó con el 20% del total de datos de entrenamiento. Se utilizaron 150 épocas, un batch size = 220 y verbose =1. Los hiperparámetros seleccionados se tomaron realizando pruebas de ensayo y error y observando el comportamiento del ajuste y las funciones de pérdida.

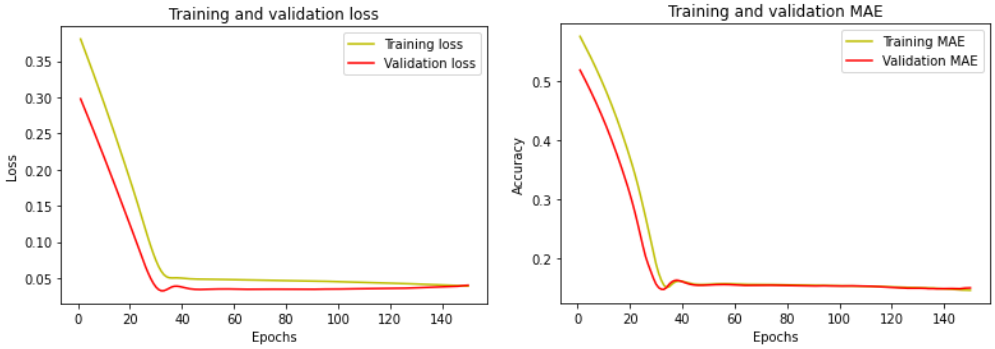
Los parámetros seleccionados se determinaron por los gráficos de función de pérdida y desempeño del mse y mae, conforme aumentó el número de épocas, tanto para el conjunto de datos de entrenamiento como el de validación.

Figura 21. *Gráfico de función de perdida y desempeño red neuronal LSTM café*



Nota: Tomado de salida del algoritmo en Python

Figura 22. *Gráfico de función de perdida y desempeño red neuronal LSTM cacao*



Nota: Tomado de salida del algoritmo en Python

Los gráficos sugieren que después de 20 y 40 épocas la función de pérdida del error y el desempeño del ajuste se estabilizan, respectivamente.

6.6. Comparación de Modelos

Para el cálculo de las métricas de desempeño se entrenaron los modelos con los mejores hiperparámetros que arrojó la búsqueda en rejilla, posteriormente se calculó y validó el ajuste en el conjunto de datos de entrenamiento utilizando k=5 folds, el ajuste en la totalidad del conjunto de datos de entrenamiento, el ajuste en el conjunto de datos de prueba y las métricas MSE, RMSE, MAE y MAPE para el cada uno de los modelos seleccionados, este proceso se ejecutó con 10 repeticiones para cada una de las métricas de desempeño. Posteriormente se grafican los pronósticos, los residuales y se realizan pruebas estadísticas para verificar la existencia de normalidad y facilitar la interpretabilidad de los modelos.

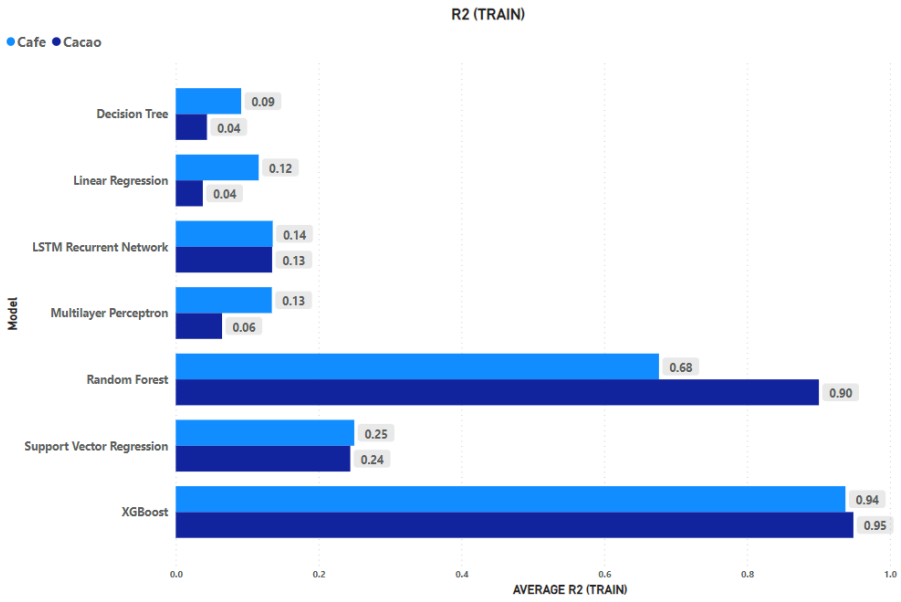
7. Resultados

En esta sección se contrastan los resultados obtenidos en la implementación de cada uno de los modelos, con la mejor selección de hiperparámetros y los resultados de las métricas de desempeño en las repeticiones ejecutadas.

7.1. Comparación de métricas de desempeño entre modelos

Se incluyen los gráficos correspondientes a las métricas de desempeño promedio de las repeticiones por modelo y por cultivo: R^2, RMSE, MSE, MAE y MAPE.

Figura 23. *Comparativa R^2 Train entre modelos*



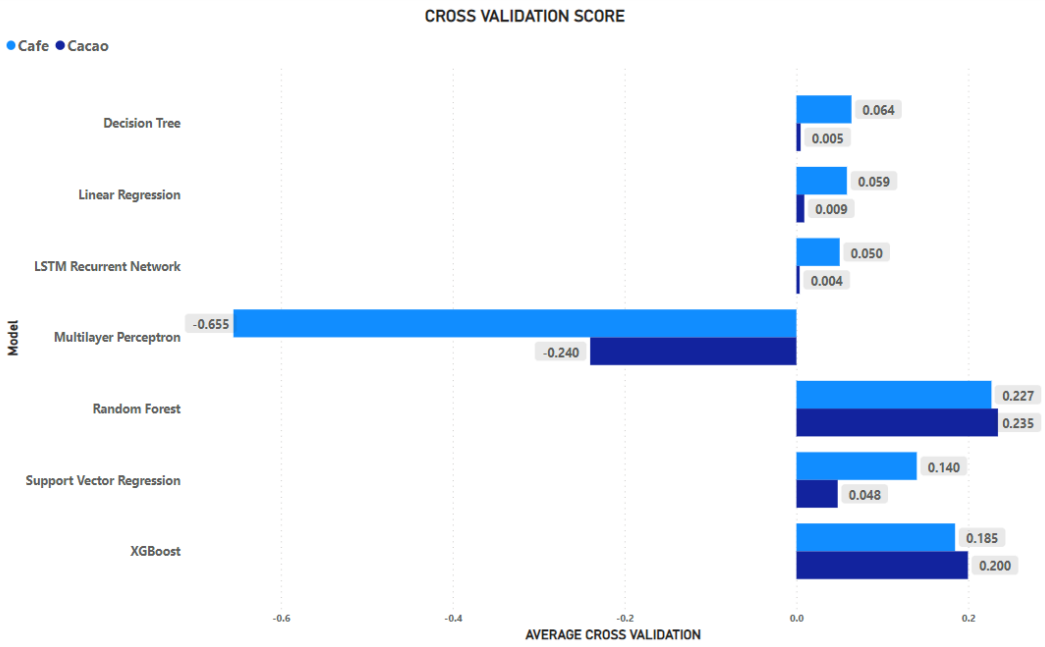
Nota: Tomado de salida del algoritmo en Python

El coeficiente de determinación R^2 evalúa la fuerza de relación lineal entre los valores observados y las predicciones, en otras palabras, cuantifica la variabilidad explicada por el modelo; el R^2 es libre de escala y está estrechamente relacionado con el MSE. El R^2 no indica si un modelo

de regresión es adecuado, se puede tener un valor bajo para un modelo adecuado o un valor alto para un modelo que no se ajusta a los datos.

Los valores más altos de R^2 (train) para café y cacao se obtuvieron en XGBoost, Random Forest y Support Vector Regression.

Figura 24. *Comparativa Cross Validation Score entre modelos*



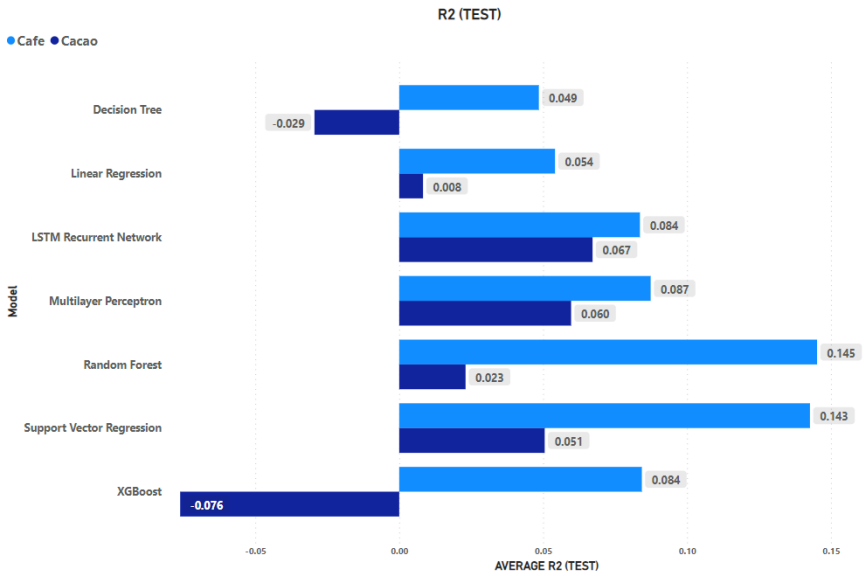
Nota: Tomado de Microsoft Power BI, con datos de la salida del algoritmo en Python

La validación cruzada es una técnica para medir la capacidad de generalización de los modelos, en nuestro caso se observa que para todos los modelos en ambos cultivos hay una disminución del ajuste en el caso de los datos de entrenamiento frente a la validación cruzada.

Los valores más altos de Cross Validation Score para café se obtuvieron en Random Forest, XGBoost y Support Vector Regression. Los valores más altos de Cross Validation Score para cacao se obtuvieron en Random Forest, XGBoost y Support Vector Regression. El modelo de

Perceptrón Multicapa se ajustó negativamente para ambos cultivos. Un R^2 negativo indica que las predicciones son peores que predecir por medio de la media, por lo que el modelo planteado no se ajusta a los datos, esto ocurre, por ejemplo, cuando el modelo predice un valor altamente negativo siendo el valor observado positivo.

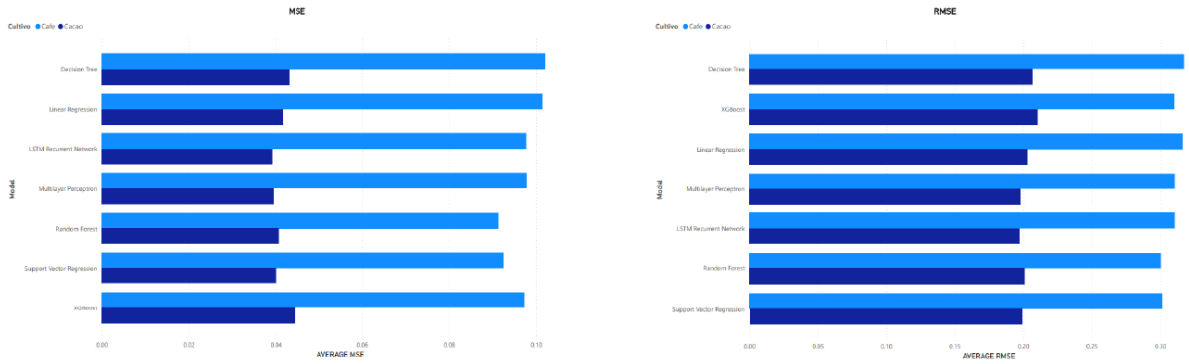
Figura 25. *Comparativa R^2 Test entre modelos*



Nota: Tomado de Microsoft Power BI, con datos de la salida del algoritmo en Python

En el caso de ajuste en el conjunto de datos de prueba (R^2 test) se evidencia el desempeño real de los modelos en datos que nunca ha visto. Los valores más altos de R^2 test para café se obtuvieron en Random Forest, Support Vector Regression y Perceptrón Multicapa. Los valores más altos de R^2 test para cacao se obtuvieron en red LSTM, Perceptrón Multicapa y Support Vector Regression. Los modelos de XGBoost y Decision Tree se ajustaron negativamente para el cultivo de cacao.

Figura 26. *Comparativa RMSE y MSE entre modelos*

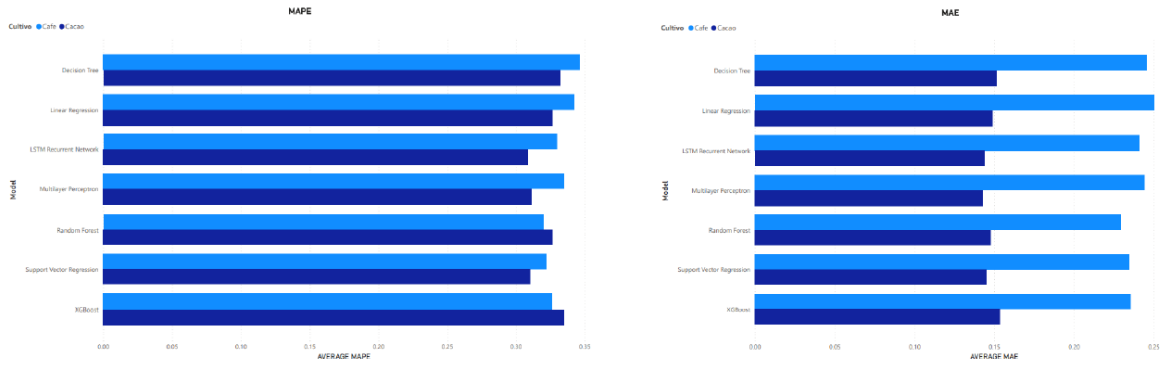


Nota: Tomado de Microsoft Power BI, con datos de la salida del algoritmo en Python

El MSE mide el error cuadrático promedio del valor observado frente a las predicciones, su valor siempre será positivo por lo que entre más cercano sea al cero mejor es el desempeño del modelo. Si se hace una predicción mala, la cuadratura empeorará aún más el error sobrestimando el desempeño del modelo. Por otro lado, si todos los errores son pequeños (menores que 1) se puede subestimar el desempeño del modelo. El RMSE introduce la raíz cuadrada al MSE para que la escala de los errores sea igual a la escala de los valores observados.

Los valores más bajos de MSE y RMSE para café se obtuvieron en Random Forest, Support Vector Regression, XGBoost. Los valores más bajos de MSE Y RMSE para cacao se obtuvieron en Support Vector Regression, Red LSTM y Perceptrón Multicapa.

Figura 27. *Comparativa MAE y MAPE entre modelos*



Nota: Tomado de Microsoft Power BI, con datos de la salida del algoritmo en Python

El MAE es el error promedio de las diferencias absolutas entre los valores observados y los pronósticos, es una puntuación lineal, por lo que cada diferencia se pondera por igual en el promedio. Esta métrica penaliza errores grandes mejor que lo hace el MSE, por lo que no es tan sensible a los valores atípicos. Cuando se tiene certeza de la existencia outliers en los datos es conveniente utilizar el MAE, de otro modo si solo hay valores inesperados que requieren especial cuidado es conveniente utilizar MSE.

El MAPE suele ser la expresión de error más fácil de entender, debido a que expresa la exactitud como porcentaje, el valor de MAPE es entonces el porcentaje de error del valor pronosticado frente al valor observado, un valor más pequeño indica un mejor ajuste.

Los valores más bajos de MAE y MAPE para café se obtuvieron en Random Forest, Support Vector Regression y XGBoost. Los valores más bajos de MAE y MAPE para cacao se obtuvieron en Perceptrón Multicapa, red LSTM y Support Vector Regression.

Se incluyen las tablas de los valores promedio de las repeticiones por modelo y por cultivo.

Tabla 7. *Tabla valores promedio de las repeticiones de las métricas modelo café*

Modelo	MAPE	MAE	MSE	RMSE	R2 (Train)	R2 (Test)	CV
LR	0.3425	0.2502	0.1014	0.3161	0.1161	0.0541	0.0587
DT	0.3465	0.2455	0.1021	0.3173	0.0915	0.0485	0.0640
SVR	0.3222	0.2346	0.0925	0.3015	0.2499	0.1427	0.1401
XGBoost	0.3263	0.2355	0.0973	0.3101	0.9377	0.0842	0.2270
RF	0.3201	0.2291	0.0912	0.3002	0.6766	0.1451	-0.2413
ANN	0.3301	0.2425	0.0974	0.3097	0.1299	0.0913	0.0226
LSTM	0.3164	0.2331	0.0924	0.3020	0.1794	0.1327	0.0226

Tabla 8. *Tabla valores promedio de las métricas de las repeticiones modelo cacao*

Modelo	MAPE	MAE	MSE	RMSE	R2 (Train)	R2 (Test)	CV
LR	0.3267	0.1488	0.0415	0.2031	0.0378	0.0082	0.0091
DT	0.3442	0.1545	0.0428	0.2065	0.0340	-0.0285	0.0040
SVR	0.3103	0.1450	0.0399	0.1990	0.2443	0.0505	0.0474
XGBoost	0.3351	0.1536	0.0443	0.2103	0.9488	-0.0760	0.1996
RF	0.3266	0.1476	0.0406	0.2010	0.9005	0.0230	0.2345
ANN	0.3332	0.1510	0.0423	0.2050	-0.0003	-0.0096	-0.1633
LSTM	0.3071	0.1399	0.0374	0.1929	0.1178	0.1044	-0.032

Nota: Adaptado por los autores a partir de Python

Tabla 9. *Tabla Desviación Estándar de las repeticiones por modelo café*

Modelo	MAPE	MAE	MSE	RMSE	R2 (Train)	R2 (Test)	CV
LR	0.0531	0.0220	0.0269	0.0408	0.0187	0.0682	0.0369
DT	0.0601	0.0247	0.0268	0.0405	0.0117	0.0367	0.0252
SVR	0.0535	0.0226	0.0270	0.0423	0.0215	0.0434	0.0203
XGBoost	0.0502	0.0186	0.0226	0.0350	0.0022	0.0950	0.0642

RF	0.0479	0.0175	0.0225	0.0354	0.0148	0.0554	0.0308
ANN	0.0550	0.0220	0.0259	0.0404	0.0203	0.0819	0.4776
LSTM	0.0491	0.0205	0.0226	0.0366	0.0237	0.0943	0.1026

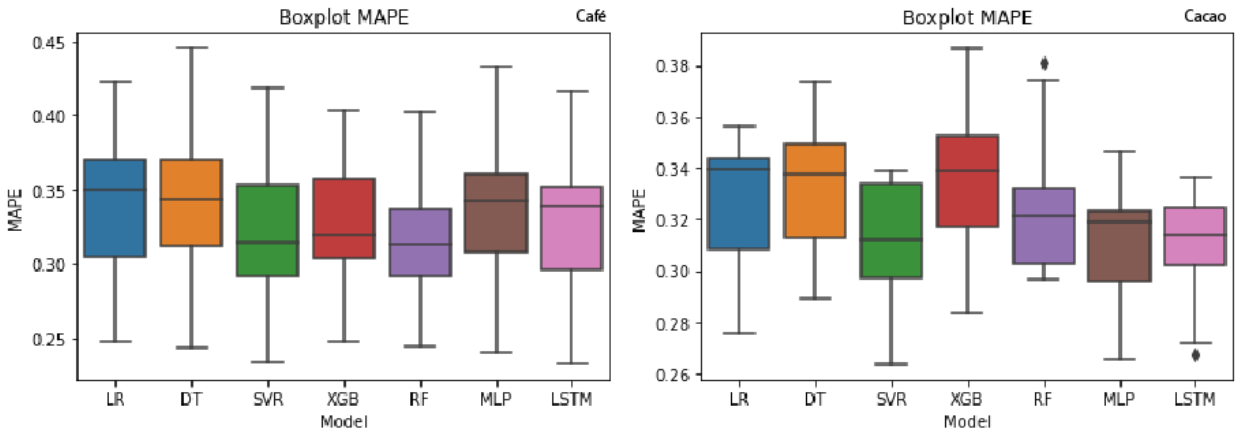
Tabla 10. *Tabla Desviación Estándar de las repeticiones por modelo cacao*

Modelo	MAPE	MAE	MSE	RMSE	R2 (Train)	R2 (Test)	CV
LR	0.0291	0.0096	0.0073	0.0180	0.0055	0.0295	0.0096
DT	0.0248	0.0072	0.0062	0.0150	0.0091	0.0555	0.0219
SVR	0.0256	0.0111	0.0079	0.0200	0.0107	0.0247	0.0233
XGBoost	0.0296	0.0056	0.0048	0.0116	0.0017	0.1458	0.0361
RF	0.0299	0.0074	0.0057	0.0143	0.0034	0.0923	0.2334
ANN	0.0273	0.0095	0.0074	0.0188	0.0036	0.0120	0.2124
LSTM	0.0270	0.0092	0.0063	0.0163	0.0078	0.0368	0.1193

Nota: Adaptado por los autores a partir de Python

Se incluyen las desviaciones estándar para cada una de las métricas para cada cultivo, menores valores de desviación estándar indicarán menor variación entre mediciones. Para el cultivo de café los modelos con menores desviaciones en error y ajuste son Random Forest, XGBoost y Support Vector Regression. Para el cultivo de cacao los modelos con menores desviaciones en error y ajuste son Perceptrón Multicapa, red LSTM y Random Forest.

Figura 28. *Gráfico de caja y bigotes del MAPE de las repeticiones de cada modelo*



Nota: Tomado de salida del algoritmo en Python

Se construyeron los boxplots para ilustrar la dispersión del MAPE para las corridas en cada algoritmo, así los valores de MAPE oscilan entre 0.2 y 0.4 para los cultivos de café y cacao en los distintos modelos. La prueba Shapiro-Wilk se realizó para probar la existencia de normalidad en las mediciones y posteriormente implementar un Análisis de la Varianza, verificando que no existen diferencias significativas entre modelos en ambos cultivos para esta métrica de error.

Tabla 11. *Resultados prueba de Shapiro Wilk para las repeticiones de los modelos*

Variable	Estadístico Café	Valor-P Café	Estadístico Cacao	Valor-P Cacao
Linear Regression	0.979646	0.963249	0.846022	0.052083
Decision Tree	0.978671	0.957663	0.945908	0.620414
Support Vector Regression	0.990118	0.996955	0.913647	0.306999
XGBoost	0.952723	0.700777	0.975776	0.938637
Random Forest	0.958334	0.766726	0.859109	0.074485
Multilayer Perceptron	0.972581	0.913695	0.883413	0.142779
LSTM	0.980539	0.967989	0.875483	0.115754

Nota: Adaptado por los autores a partir de Python

Tabla 12. *Anova comparación de modelos café*

	df	sum_sq	mean_sq	F	PR(>F)
C(Modelo)	6.0	0.006063	0.001011	0.3550	0.904323
Residual	63.0	0.179294	0.002846	-	-

Nota: Adaptado por los autores a partir de Python

Tabla 13. *Anova comparación de modelos cacao*

	df	sum_sq	mean_sq	F	PR(>F)
C(Modelo)	6.0	0.007413	0.001235	1.6297	0.153771
Residual	63.0	0.047758	0.000758	-	-

Nota: Adaptado por los autores a partir de Python

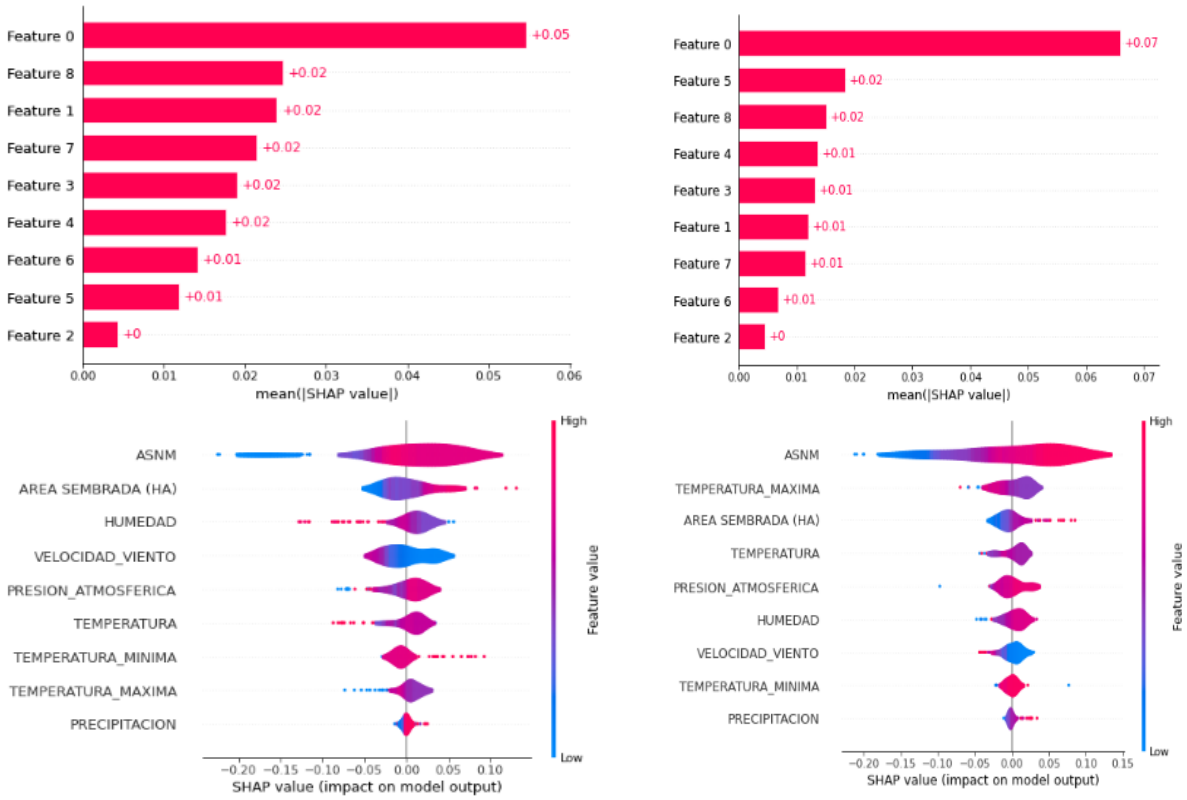
7.2. Gráficos de resultados

En el Apéndice D se encuentran los gráficos de resultados para cada modelo, los gráficos de dispersión de pronósticos muestran una distribución de los pronósticos alrededor de la recta de regresión, con sus respectivos coeficientes de ajuste en el conjunto de datos de prueba.

7.3. Shap Values

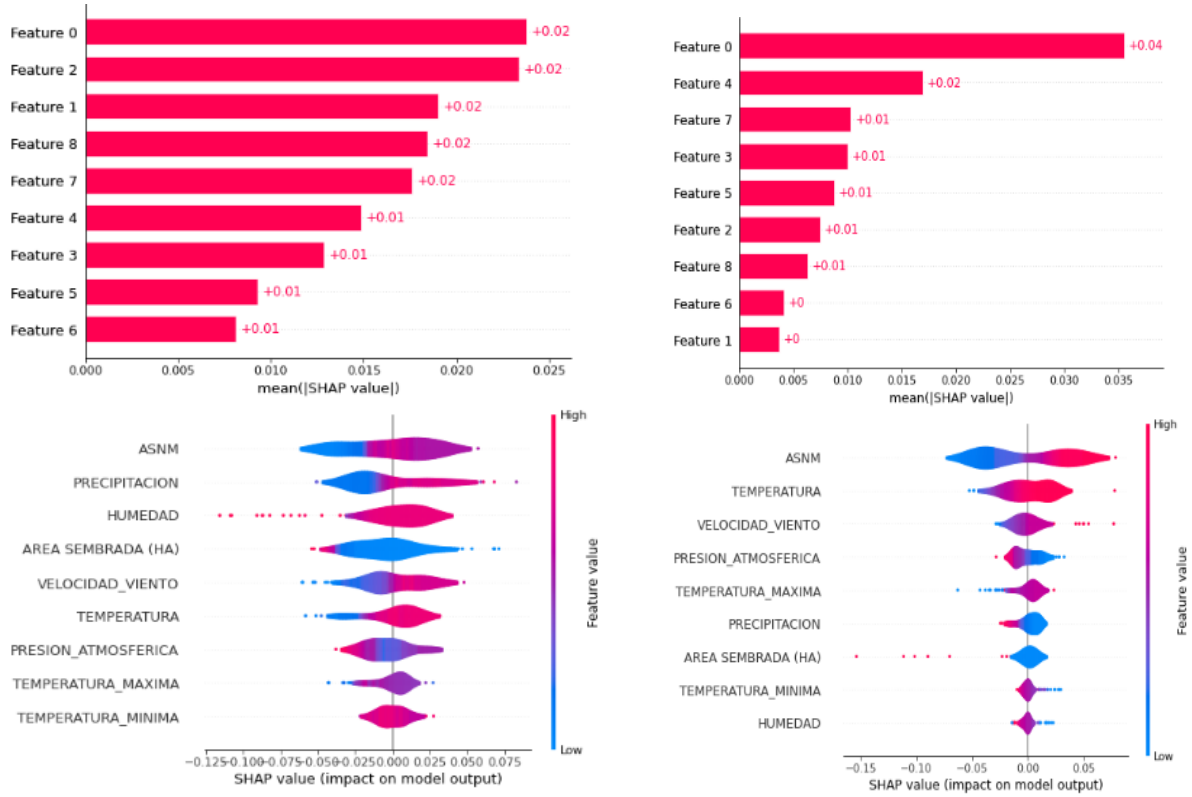
Un algoritmo de Aprendizaje Automático puede generar predicciones precisas, sin embargo, su naturaleza de "caja negra" puede dificultar la adopción de este en la práctica al carecer de interpretabilidad. El valor de Shapley calcula el promedio de las contribuciones marginales en todas las permutaciones posibles, los Shap son una extensión del valor de Shapley utilizado para explicar el resultado de cualquier modelo de Aprendizaje Automático. Estos valores permiten cuantificar la contribución (positiva o negativa) de cada predictor a la variable respuesta de forma global y local, como el gráfico de importancia de una variable. Es importante aclarar que los valores Shap no identifican causalidad. Para los mejores modelos de Aprendizaje Profundo y Automático se construyeron las gráficas de importancia global y local de cada variable, así como las relaciones de los predictores con la variable respuesta, cuando estos toman valores positivos y negativos.

Figura 29. *Valores Shap cultivo café, izquierda Perceptrón Multicapa, derecha Random Forest*



Nota: Tomado de salida del algoritmo en Python

Figura 30. *Valores Shap cultivo cacao, izquierda Perceptrón Multicapa, derecha Random Forest*



Nota: Tomado de salida del algoritmo en Python

Para el cultivo del café las condiciones que mayormente contribuyen al rendimiento son la altura sobre el nivel del mar, el área de siembra, la temperatura media, la humedad, la presión atmosférica y finalmente la precipitación acumulada.

Para el cultivo del cacao las condiciones que mayormente contribuyen al rendimiento son la altura sobre el nivel del mar, la precipitación acumulada, la humedad, la temperatura media, el área de siembra y finalmente la presión atmosférica.

Ambos modelos convergen en que geografías con mayor altitud, una mayor precipitación acumulada, mayor humedad promedio, mayor área de siembra, temperaturas medias no extremas, presión atmosférica medianamente alta y una mayor precipitación acumulada contribuyen positivamente a la predicción del rendimiento agrícola.

7.4. Intervalos de confianza para el error y el ajuste

En el Apéndice E se encuentran los intervalos de confianza del 95% para el R^2 (test), MSE y MAPE, calculados luego de probar normalidad en las repeticiones con la prueba de Shapiro- Wilk.

7.5. Discusión de resultados

Para la selección de los mejores modelos por cultivo se ordenaron de 1 a 7 los valores de las métricas de desempeño promedio resultante de las repeticiones, para el caso de los errores se buscan valores pequeños y para los ajustes (R^2) valores cercanos al 1. Posteriormente se calcularon los promedios de las posiciones de cada modelo, de este modo una menor posición indicará un mejor desempeño promedio de los modelos tanto para el ajuste como para el error. En el apéndice F se encuentra la metodología propuesta.

Para el café los modelos que arrojan los menores errores y mejores ajustes son, Random Forest, Support Vector Regression y XGBoost. Para el cacao los modelos que arrojan los menores errores son red LSTM, Perceptrón Multicapa y Support Vector Regression; los modelos que arrojan los mejores ajustes son, Random Forest, Support Vector Regression y XGBoost.

La variabilidad explicada por los mejores modelos seleccionados es en promedio, del 14.517% para el café y del 0.672 % para el cacao. Para el cultivo de cacao los modelos XGBoost y Decision Tree se ajustaron negativamente en el conjunto de datos de prueba, esto indica que el modelo planteado no se ajusta a los datos. El R^2 por sí sólo no indica si un modelo de regresión es

adecuado, un bajo R^2 puede deberse a la naturaleza aleatoria propia del fenómeno, además, aún con un bajo valor de esta métrica se pueden tener predictores estadísticamente significativos. Un análisis complementario consideró examinar las gráficas de residuos en búsqueda de sesgos y existencia de aleatoriedad, validando la superioridad de los modelos basados en árboles y las redes neuronales para capturar no linealidades en los datos.

La validación cruzada en el entrenamiento permitió un panorama más acertado acerca de la capacidad de generalización de los modelos. De esta forma, el ajuste para el café en el conjunto de datos de entrenamiento alcanzó valores entre 25% y 95%, en tanto, la validación cruzada alcanzó valores entre 14% y 22.7%. Para el cacao en el conjunto de datos de entrenamiento alcanzó valores entre 24% y 95%, en tanto, la validación cruzada alcanzó valores entre 14% y 23.5%. La validación cruzada permitió tener un valor más acertado de la capacidad de generalización de los modelos.

Una comparación entre la validación cruzada y el desempeño en el conjunto de datos de prueba permitió evidenciar posibles problemas de overfitting y underfitting. Para el café los modelos Random Forest, XGBoost y Support Vector Regression tienen menores diferencias en el desempeño del entrenamiento y la prueba. Para el cacao los modelos Random Forest, XGBoost y la red LSTM tienen menores diferencias en el desempeño del entrenamiento y la prueba. Buenas alternativas para disminuir las variaciones del aprendizaje de los modelos comprenden utilizar modelos medianamente complejos, recolectar más datos, aplicar técnicas de ingeniería de características o incluir regularizaciones.

Asimismo, las métricas de error permiten cuantificar los errores entre los valores observados y las predicciones, como se mencionó anteriormente. El análisis de la varianza

corroboró la uniformidad en los resultados del MAPE, métrica de error expresada en porcentaje, lo que en indica que el porcentaje de coincidencia (1-MAPE) entre las predicciones y el valor real se encuentra entre el 64% y 72% para el café, y entre el entre 66% y 71% para el cacao.

La búsqueda de hiperparámetros y el costo computacional es un factor determinante en el momento de implementar un modelo, así la búsqueda exhaustiva (GridSearch) para los modelos basados en árboles y la selección de capas, neuronas, optimizadores e hiperparámetros para el caso de las redes neuronales generan un alto costo computacional medido en Hardware, Software y tiempo.

Este tipo de investigaciones deben ir acompañadas del juicio de expertos que conozcan del comportamiento en la práctica del fenómeno en estudio. Para el café las condiciones que permiten una condición óptima son: temperaturas entre los 19 y 21,5 grados centígrados, el cultivo debe estar localizado en altitudes entre 1200 y 1800 metros sobre el nivel del mar, las lluvias acumuladas deben estar entre 1800 y 2800 milímetros anuales, el brillo solar debe encontrarse entre 135 y 165 horas de sol al mes y humedad del 40% al 80% (Bancolombia, 2018b). Para el cacao las condiciones que permiten una condición óptima son: temperaturas entre los 23 y 25 grados centígrados (con variaciones de temperatura máxima y mínima menores a los 9 grados centígrados), el cultivo debe estar localizado en altitudes entre 1000 y 1400 metros sobre el nivel del mar, las lluvias acumuladas deben estar entre 1500 y 2500 milímetros anuales, y humedad del 70% al 78% (Bancolombia, 2018a).

El juicio de expertos y los resultados obtenidos derivados de la interpretabilidad de los modelos coinciden en que cultivos localizados en mayores altitudes, temperaturas cálidas, con ligeras variaciones entre temperaturas máximas y mínimas, mayor humedad promedio y una mayor

cantidad de lluvias distribuidas uniformemente a lo largo del año, son condiciones que favorecen el rendimiento de los cultivos de café y cacao. Un hallazgo adicional indica que una mayor velocidad del viento y una presión atmosférica mayormente alta contribuyen positivamente al rendimiento de ambos cultivos.

8. Conclusiones

Dada la necesidad de buscar estrategias orientadas a mejores prácticas que maximicen el rendimiento de cultivos en el sector cacaotero y cafetero, surge la presente investigación desarrollada con herramientas de Aprendizaje Profundo y Aprendizaje Automático. La precisión de los modelos desarrollados depende directamente de la disponibilidad y calidad de los datos, un adecuado preprocesamiento de estos y la correcta selección de los algoritmos que mejor describan los datos en cada investigación.

Los modelos basados en árboles, las máquinas de soporte y las redes neuronales demostraron ser una buena opción para solucionar problemas de regresión no lineales, gestionando la deficiencia de datos y la existencia de sesgos de una buena manera. El uso de métricas de desempeño debe ir acompañado de gráficos de dispersión y cálculos de error que permitan una mejor interpretación del comportamiento de los modelos, guiarse únicamente del coeficiente de determinación puede ser engañoso y debe ir acompañado de métricas como el MSE y el MAPE. El uso de la técnica Shap Values facilitó la interpretabilidad de los modelos utilizados, centrándose en la importancia de las variables más allá de sólo los resultados de las métricas de desempeño.

Los mejores modelos se seleccionaron teniendo en cuenta aspectos como la capacidad de generalización, la gestión del error y la variabilidad explicada, así para el café los mejores modelos fueron Random Forest (R^2test = 0.1452, MSE = 0.0913, MAPE = 0.3201), Support Vector Regression (R^2test = 0.1427, MSE = 0.2346, MAPE = 0.3223), y XGBoost (R^2test = 0.0843, MSE = 0.0973, MAPE = 0.3263).

Para el cacao los mejores modelos fueron red LSTM (R^2test = 0.0672, MSE =0.0392, MAPE = 0.3087), Perceptrón Multicapa (R^2test = 0.0597, MSE = 0.0394, MAPE = 0.3115), y

Random Forest (R^2test = 0.023, MSE = 0.0406, MAPE = 0.3267). Las redes neuronales y los modelos basados en árboles demostraron un buen desempeño en el manejo del error y la variabilidad explicada para ambos cultivos.

Los modelos y la literatura coinciden en que la altura sobre el nivel del mar, la temperatura, la humedad, la precipitación y la presión atmosférica son variables climáticas que inciden altamente en el rendimiento de los cultivos en estudio. Mayores altitudes, temperaturas cálidas, con ligeras variaciones entre temperaturas máximas y mínimas, mayor humedad promedio y una mayor cantidad de lluvias distribuidas uniformemente a lo largo del año, son condiciones que favorecen el rendimiento de los cultivos de café y cacao. Adicional a la literatura, los modelos indican que a mayor velocidad del viento y una presión atmosférica mayormente alta contribuyen positivamente al rendimiento de ambos cultivos.

El presente estudio es relevante en la medida que los hallazgos sirven de insumo para la toma de decisiones estratégicas acerca de la ubicación y selección de los cultivos en estudio en el departamento de Santander y a nivel nacional. Los agricultores y actores involucrados que deseen maximizar el rendimiento de sus cultivos de café y cacao deberán ubicar sus cultivos en geografías con mayor altura sobre el nivel del mar, con temperaturas cálidas (no extremas), zonas altamente húmedas, con abundante precipitación distribuida uniformemente a lo largo del año, con alta velocidad del viento, presión atmosférica moderadamente alta y utilizando mayor área de siembra posible.

Es importante considerar la capacidad de cómputo y la eficiencia de los algoritmos al momento de tomar una decisión respecto al modelo a implementar, este factor es clave dado que, en la investigación se generaron 10 repeticiones de cada una de las métricas para cada algoritmo,

consumiendo aproximadamente 6 horas de procesamiento continuo en la nube (utilizando recursos de CPU y GPU). Futuras investigaciones sugieren la implementación de más repeticiones, muestras y algoritmos, buscando obtener la mejor relación costo computacional-desempeño del modelo-generación de valor en el ejercicio práctico.

9. Recomendaciones

Para fortalecer este tipo de investigaciones es imprescindible la existencia de datos precisos, constantes y abiertos, en la actualidad el 61% del total de estaciones agrometeorológicas se encuentran activas, de las cuáles el 4,97% se encuentran en el departamento de Santander (IDEAM, 2022). Según lo anterior se recomienda a las entidades correspondientes velar por una cobertura integral y ampliada de los registros agrometeorológicos en los municipios de Colombia, así como la toma de datos granulares de rendimientos de cultivos, por ejemplo, mensual o trimestralmente, con distinción georreferenciada de cada una de las mediciones. Una menor granularidad de los datos junto a sensores precisos permitirá construir modelos más robustos incluyendo, por ejemplo, análisis de series de tiempo junto a modelos de aprendizaje de máquina.

Se recomienda a gobierno, federaciones, agricultores y demás actores del sector, la toma de decisiones basadas en datos, este tipo de investigaciones robustecen las decisiones orientadas a maximizar el rendimiento de cultivos y mitigar las pérdidas que afectan, entre otros, la seguridad alimentaria y la economía local. Se sugiere replicar el presente estudio para otro tipo de cultivos, sectorizando según las condiciones geográficas, condiciones climáticas y condiciones propias del cultivo (tipo de clon, resistencia a plagas, entre otros) y el área de siembra (profundidad, acidez del suelo, presencia de fertilizantes, cantidad de materia orgánica, entre otros).

Por su parte, se invita a la academia, en especial a los estudiantes de la Universidad Industrial de Santander, a trabajar en conjunto para fortalecer el desarrollo del campo colombiano mediante la aplicación de nuevas tecnologías, como el Aprendizaje Automático y el Aprendizaje Profundo.

Referencias

Anderson, L. O. (2016). SENSOR MODIS: UMA ABORDAGEM GERAL SENSOR MODIS : UMA ABORDAGEM GERAL Liana Oighenstein Anderson Marcelo Lopes Latorre Yosio Edemir Shimabukuro Egídio Arai Osmar Abílio de Carvalho Júnior São José dos Campos. *Incluye Información de La SE Función Pública*, *January 2003*, 58.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics*, *107*(4), 871–885. https://doi.org/10.1002/cpt.1796

Baldi, P., & Hornik, K. (1989). Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. In *Neural Networks* (Vol. 2).

Banco de la República | Colombia. (2021, October 15). *Choques climáticos y sus efectos sobre el sector agrícola en Colombia*.

Bancolombia. (2018a, August 23). *Guía completa: lo que debes saber para Cultivar Cacao*.

Bancolombia. (2018b, August 30). *Guía completa: Cómo cultivar Café*.

Barlow, K. M., Christy, B. P., O'Leary, G. J., Riffkin, P. A., & Nuttall, J. G. (2015). Simulating the impact of extreme heat and frost events on wheat crop production: A review. *Field Crops Research*, *171*, 109–119. https://doi.org/10.1016/j.fcr.2014.11.010

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58–65. https://doi.org/10.1145/3448250

Bernal, E. (1978). *RED METEOROLÓGICA DE COLOMBIA* (Vol. 1). Instituto Colombiano de Hidrología, Meteorología y Adecuación de Tierras.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, *274*(November 2020), 144–159. https://doi.org/10.1016/j.agrformet.2020.108275

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2021a). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agricultural and Forest Meteorology*, *297*(April 2020), 108275. https://doi.org/10.1016/j.agrformet.2020.108275

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2021b). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agricultural and Forest Meteorology*, *297*(November 2020). https://doi.org/10.1016/j.agrformet.2020.108275

CEPAL. (2011). *Agricultura y cambio climático: instituciones, políticas e innovación*.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chatfield, C. (1986). Exploratory data analysis. *European Journal of Operational Research*, *23*(1), 5–13. https://doi.org/10.1016/0377-2217(86)90209-2

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). *On Empirical Comparisons of Optimizers for Deep Learning*. http://arxiv.org/abs/1910.05446

Cuthbert, D. (1981). *Origins of the variance inflation factor as recalled*.

de Myttenaere, A., Golden, B., le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing*, *192*, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Edwards, T. G., Özgün-Koca, A., & Barr, J. (2017). Interpretations of Boxplots: Helping Middle School Students to Think Outside the Box. *Journal of Statistics Education*, *25*(1), 21–28. https://doi.org/10.1080/10691898.2017.1288556

el Bouchefry, K., & de Souza, R. S. (2020a). Learning in Big Data: Introduction to Machine Learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (pp. 225–249). Elsevier. https://doi.org/10.1016/B978-0-12-819154-5.00023-0

el Bouchefry, K., & de Souza, R. S. (2020b). Learning in Big Data: Introduction to Machine Learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (pp. 225–249). Elsevier. https://doi.org/10.1016/B978-0-12-819154-5.00023-0

el PAÍS, & NICHOLAS DALE. (2022, March 13). *Los nuevos horizontes del café ante el cambio climático*.

Elsayir, H. A. (2019). Residual Analysis for Auto-Correlated Econometric Model. *Journal of Mathematics and Statistics*, *15*(1), 99–111. https://doi.org/10.3844/jmssp.2019.99.111

Eom, H., Son, Y., & Choi, S. (2020). Feature-Selective Ensemble Learning-Based Long-Term Regional PV Generation Forecasting. *IEEE Access*, *8*, 54620–54630. https://doi.org/10.1109/ACCESS.2020.2981819

Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowlegde Discovery in Databases. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9078*(3), 637–648. https://doi.org/10.1007/978-3-319-18032-8_50

FEDECACAO. (2021). *Producción de café de Colombia cierra 2021 en 12,6 millones de sacos*. Por El Impacto Del Paro Nacional y El Efecto Del Clima.

FEDECACAO. (2022, July 11). *La producción cacaotera nacional sigue creciendo: en 2021 logra un nuevo récord histórico*.

Feldman, A. J. L., & Cortés, D. H. (2016). *Climate Change and Agriculture: A Review of the Literature with Emphasis on Latin America: Vol. LXXXIII* (Issue 4).

Food and Agriculture Organization of the United Nations. (2020). *Impact of COVID-19 on agriculture, food systems and rural livelihoods in Eastern Africa: Policy and programmatic options*.

Fürnkranz, J., Chan, P. K., Craw, S., Sammut, C., Uther, W., Ratnaparkhi, A., Jin, X., Han, J., Yang, Y., Morik, K., Dorigo, M., Birattari, M., Stützle, T., Brazdil, P., Vilalta, R., Giraud-Carrier, C., Soares, C., Rissanen, J., Baxter, R. A., … De Raedt, L. (2011). Mean Squared Error. In *Encyclopedia of Machine Learning* (pp. 653–653). Springer US. https://doi.org/10.1007/978-0-387-30164-8_528

García, S., Luengo, J., & Herrera Francisco. (2015). *Data Preprocessing in Data Mining* (1st ed., Vol. 72). Springer International Publishing.

García-Arteaga, J. J., Zambrano-Zambrano, J. J., Alcivar-Cevallos, R., & Zambrano-Romero, W. D. (2020). Predicción del rendimiento de cultivos agrícolas usando Aprendizaje Automático. *Revista Arbitrada Interdisciplinaria Koinonía*, *5*(2), 144. https://doi.org/10.35381/r.k.v5i2.1013

Gerald C. Nelson, M. (2009). Cambio Climático: El impacto en la agricultura y los costos de adaptación. In *Cambio Climático: El impacto en la agricultura y los costos de adaptación*. International Food Policy Research Institute. https://doi.org/10.2499/0896295370

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (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

Guerrero, S. C., & Melo, O. O. (2017). Una metodología para el tratamiento de la multicolinealidad a través del escalamiento multidimensional. *Ciencia En Desarrollo*, *8*, 9–24.

Harvard edX. (2022). *Cursos de Modelo predictivo*. Https://Www.Edx.Org/Es/Aprende/Modelo-Predictivo#:~:Text=Los%20Modelos%20Predictivos%20son%20un,Mediante%20t%C3%A9cnicas%20de%20an%C3%A1lisis%20de.

Herrera, F. (2016). *Big Data: Preprocesamiento y calidad de datos*. www.highlycited.com

Hochreiter, S., & Urgen Schmidhuber, J. ¨. (1997). *Long Short-Term Memory*.

Howden, S. M., Soussana, J.-F., Tubiello, F. N., Chhetri, N., Dunlop, M., & Meinke, H. (2007). *Adapting agriculture to climate change*. www.pnas.orgcgidoi10.1073pnas.0701890104

IDEAM - Instituto de Hidrología, M. y E. A. (2022). *METEOROLOGÍA AGRÍCOLA*. Aplicaciones Metereologicas.

Isasi Viñuela, & P., & G. L. (2004). Redes de neuronas artificiales. Un Enfoque Práctico. *Editorial Pearson Educación SA Madrid España.*, 46–51.

Izaurieta, F., & Saavedra, C. (1999). Redes Neuronales Artificiales. *Charlas de Fisica*, 1–15. https://doi.org/10.1016/S0210-5691(05)74198-X

Jamal, P., Ali, M., Faraj, R. H., Ali, P. J. M., & Faraj, R. H. (2014). 1-6 Data Normalization and Standardization: A Technical Report. In *Machine Learning Technical Reports* (Vol. 1, Issue 1). https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#

James, G., Witten, D., Hastie, T., & Robert Tibshirani. (2013). *An Introduction to Statistical Learning* (2nd ed., Vol. 112). Springer.

Jurečka, F., Fischer, M., Hlavinka, P., Balek, J., Semerádová, D., Bláhová, M., Anderson, M. C., Hain, C., Žalud, Z., & Trnka, M. (2021). Potential of water balance and remote sensing-based evapotranspiration models to predict yields of spring barley and winter wheat in the Czech Republic. *Agricultural Water Management*, *256*(July). https://doi.org/10.1016/j.agwat.2021.107064

Kaggle. (2021). *An introduction to XGBoost regression*. https://www.kaggle.com/code/carlmcbrideellis/an-introduction-to-xgboost-regression

Larrañaga, P., Inza, I., & Moujahid, A. (1997). *Redes Neuronales*. http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t8neuronales.pdf

Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Liu, D. L., Li, Y., He, J., Feng, H., Yang, G., & Yu, Q. (2021a). Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agricultural and Forest Meteorology*, *308–309*(October 2020). https://doi.org/10.1016/j.agrformet.2021.108558

Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Liu, D. L., Li, Y., He, J., Feng, H., Yang, G., & Yu, Q. (2021b). Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agricultural and Forest Meteorology*, *308–309*. https://doi.org/10.1016/j.agrformet.2021.108558

Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, *150*(11), 1443–1452. https://doi.org/10.1016/j.agrformet.2010.07.008

Ma, J. W., Nguyen, C. H., Lee, K., & Heo, J. (2019a). Regional-scale rice-yield estimation using stacked auto-encoder with climatic and MODIS data: a case study of South Korea. *International Journal of Remote Sensing*, *40*(1), 51–71. https://doi.org/10.1080/01431161.2018.1488291

Ma, J. W., Nguyen, C. H., Lee, K., & Heo, J. (2019b). Regional-scale rice-yield estimation using stacked auto-encoder with climatic and MODIS data: a case study of South Korea. *International Journal of Remote Sensing*, *40*(1), 51–71. https://doi.org/10.1080/01431161.2018.1488291

Mahdi, M. D., Mrittika, N. J., Shams, M., Chowdhury, L., & Siddique, S. (2020). A Deep Gaussian Process for Forecasting Crop Yield and Time Series Analysis of Precipitation Based in Munshiganj, Bangladesh. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 1331–1334. https://doi.org/10.1109/IGARSS39084.2020.9323423

Marcano-Cedeno, A., Quintanilla-Dominguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, 2845–2850. https://doi.org/10.1109/IECON.2010.5675075

MATTOS PEREIRA, M. (2017). *APRENDIZADO PROFUNDO: REDES LSTM*. 210093.

Medina-Merino, R. F., & Ñique-Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*, *0*(010), 165. https://doi.org/10.26439/interfases2017.n10.1775

Microsoft. (2022, September 26). *Ajuste de hiperparámetros de un modelo (v2)*. Azure. https://learn.microsoft.com/es-es/azure/machine-learning/how-to-tune-hyperparameters

Orr, Genevieve., & Müller, K.-Robert. (1998). *Neural networks : tricks of the trade*. Springer.

Ostertagova, E., Ostertag, O., & Ostertagová, E. (2013). Methodology and Application of One-way ANOVA. *American Journal of Mechanical Engineering*, *1*(7), 256–261. https://doi.org/10.12691/ajme-1-7-21

Parmar, R. (2018, September 2). *Common Loss functions in machine learning*. Towards Data Science. https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of*

Computer-Aided Molecular Design, *34*(10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0

Samuels, P., & Gilchrist, M. (2014). *Pearson Correlation*. https://www.researchgate.net/publication/274635640

Schauberger, B., Jägermeyr, J., & Gornott, C. (2020). A systematic review of local to regional yield forecasting approaches and frequently used data resources. *European Journal of Agronomy*, *120*(April), 126153. https://doi.org/10.1016/j.eja.2020.126153

Scholkopf, B., Kah-Kay Sung, Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, *45*(11), 2758–2765. https://doi.org/10.1109/78.650102

Seifert, A., & Rasp, S. (2020). Potential and Limitations of Machine Learning for Modeling Warm-Rain Cloud Microphysical Processes. *Journal of Advances in Modeling Earth Systems*, *12*(12). https://doi.org/10.1029/2020MS002301

Seo, S. (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. http://d-scholarship.pitt.edu/7948/1/Seo.pdf

Tapia, F., Ernesto, C., Cevallos, F., Carlos, K. L., Flores Tapia, E., & Lissette, K. (2021). PRUEBAS PARA COMPROBAR LA NORMALIDAD DE DATOS EN PROCESOS PRODUCTIVOS: ANDERSON-DARLING, RYAN-JOINER, SHAPIRO-WILK Y KOLMOGÓROV-SMIRNOV. *Periodicidad: Semestral*, *23*(2), 2021.

Touretzky, D. (2006). Optimization Techniques. In *Artificial Neural Networks* (15th ed., Vol. 7, pp. 486–782). Spring.

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences, 106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, *14*(11). https://doi.org/10.1371/journal.pone.0224365

Vargas, J., Conde, B., Paccapelo, V., & Zingaretti, L. (2012, August 8). MÁQUINAS DE SOPORTE VECTORIAL: METODOLOGÍA Y APLICACIÓN EN R. *MÁQUINAS DE SOPORTE VECTORIAL: METODOLOGÍA Y APLICACIÓN EN R*.

Veeramsetty, V., Singal, G., & Badal, T. (2020). Coinnet: platform independent application to recognize Indian currency notes using deep learning techniques. *Multimedia Tools and Applications*, *79*(31–32), 22569–22594. https://doi.org/10.1007/s11042-020-09031-0

Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research Journal*, *12*(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015

Wang, H., & Zheng, H. (2013). Model Validation, Machine Learning. In *Encyclopedia of Systems Biology* (pp. 1406–1407). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_233

Weitz, D. (2020, April 15). *Histograms, Why & How, Storytelling, Tips & Extensions*. Towards Data Science. https://towardsdatascience.com/histograms-why-how-431a5cfbfcd5#:~:text=A%20histogram%20provides%20a%20visual,or%20gaps%20in%20the%20data.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Zhou Zhi-Hua. (2012). *Emsemble Methods: Foundations and Algorithms*.