Modelo de redes neuronales para la predicción del precio del cacao en polvo con y sin adición de azúcar en Colombia

Maria Kamila Muñoz Acosta

Freddy Isaías Reyes Martínez

Asesora

Edith Johana Morales Liberato

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Nota de Aceptación

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

2025

Dedicatoria

Dedicamos este proyecto a nuestras familias, cuyo apoyo incondicional y paciencia nos han impulsado a culminar esta etapa académica con éxito. A nuestros docentes y tutores de la Universidad Nacional Abierta y a Distancia (UNAD), por su guía, conocimientos y enseñanzas, que han sido fundamentales en la construcción de este trabajo. También dedicamos este esfuerzo a todos los productores y actores de la industria cacaotera en Colombia, cuya labor incansable nos inspiró a desarrollar este modelo de redes neuronales, con la esperanza de que pueda contribuir al análisis y predicción de precios en el sector. Finalmente, a cada persona que cree en el poder de la ciencia de datos y la analítica para transformar sectores productivos, generando conocimiento y oportunidades para el futuro.

Agradecimientos

Expresamos nuestro más sincero agradecimiento a la Universidad Nacional Abierta y a Distancia (UNAD) por brindarnos las herramientas y el conocimiento necesario para desarrollar este proyecto. A nuestros docentes y asesores, por su orientación y paciencia, que fueron clave en cada etapa del trabajo y a todas las personas que, directa o indirectamente, contribuyeron con sus conocimientos y experiencias a la realización de este proyecto.

Resumen

Este estudio se centra en la alta volatilidad de los precios del cacao en polvo, un problema que afecta la planificación estratégica y la competitividad de los productores colombianos. Para abordar esta situación, se propone un modelo de predicción basado en redes neuronales artificiales, una herramienta que, según estudios previos es ideal para analizar los patrones no lineales en datos económicos.

El objetivo principal es desarrollar un modelo que permita comprender las tendencias históricas de los precios y predecir futuras fluctuaciones, considerando tanto las características intrínsecas del cacao en polvo como factores externos que influyen en su comercialización. Se espera que este modelo contribuya a una mejor toma de decisiones en el sector cacaotero colombiano y a la creación eficiente de estrategias de crecimiento para el sector.

Además, el estudio se enfoca en el cacao en polvo con y sin adición de azúcar debido a la oportunidad que la creciente demanda de productos orgánicos y saludables representa para los exportadores colombianos. Al analizar este mercado específico, se busca optimizar las estrategias de exportación y aprovechar las tendencias del consumidor.

Palabras claves: Redes neuronales, predicción de precios, modelado predictivo, inteligencia artificial, aprendizaje automático, cacao en polvo.

Abstract

This study focuses on the high volatility of cocoa powder prices, a problem that affects the strategic planning and competitiveness of Colombian producers. To address this situation, a prediction model based on artificial neural networks is proposed, a tool that, according to previous studies is ideal for analyzing non-linear patterns in economic data.

The main objective is to develop a model that allows us to understand historical price trends and predict future fluctuations, considering both the intrinsic characteristics of cocoa powder and external factors that influence its marketing. This model is expected to contribute to better decision-making in the Colombian cocoa sector and to the efficient creation of growth strategies for the sector.

In addition, the study focuses on the European market, where the growing demand for organic and healthy products represents an opportunity for Colombian exporters. By analyzing this specific market, it seeks to optimize export strategies and take advantage of consumer trends.

Keywords: Neural networks, price prediction, predictive modeling, artificial intelligence, machine learning, cocoa powder.

Tabla de Contenido

Lista de Tabla

Lista de Figuras

Lista de Apéndices

Introducción

La volatilidad de los precios del cacao representa un desafío significativo para los productores colombianos, afectando directamente su planificación estratégica y su capacidad para competir a nivel nacional e internacional. Al ser un producto cuyo precio se determina por la dinámica de la oferta y la demanda, los productores están expuestos a las fluctuaciones del mercado, sin tener un control directo sobre ellas (Fedecacao, 2021).

De acuerdo con Montoya-Restrepo et al. (2015) la situación se ve agravada por factores como la baja productividad de los cultivos, atribuida a la vejez de las plantas, plagas y enfermedades, así como a las dificultades que enfrentan los agricultores para implementar prácticas de manejo integral de cultivos. Estos problemas generan altos costos de producción y un elevado riesgo de pérdidas para los productores.

La falta de herramientas precisas para predecir la evolución de los precios del cacao exacerba la vulnerabilidad de los productores. Esta incertidumbre limita su capacidad para tomar decisiones informadas sobre la comercialización de su producto, dificultando la implementación de estrategias de mitigación de riesgos y la optimización de sus operaciones. Como resultado, existe una alta inestabilidad económica en la industria cacaotera colombiana.

Por esta razón este proyecto propone construir un modelo de redes neuronales multicapa que permita determinar el comportamiento futuro del precio del cacao más específicamente del cacao en polvo con y sin adición de azúcar teniendo en cuenta las exportaciones de este producto desde Colombia al resto del mundo. Esto se hace con el fin de aprovechar la oportunidad que se está generando con el cambio de los hábitos de consumo que se está desplazando hacia consumidores con mayor conciencia y enfocados en los productos orgánicos y saludables.

Definición del Problema

El mercado internacional del cacao ha venido sufriendo grandes cambios debido a las dificultades que presentan los productores al momento de suplir la demanda, los países de África Occidental enfrentan bajos rendimientos en los cultivos de cacao como consecuencia del envejecimiento de los cultivos, las plagas y los problemas de distribución logística. Según Montoya-Restrepo et al. (2015) la baja productividad en Colombia se debe a los factores antes mencionados y a los problemas que presenta el agricultor al momento de darle un manejo integral al cultivo. Esto genera sobre costos y un gran riesgo de pérdidas para el agricultor que ya de por si puede considerase pobre.

A todo esto, se suman los problemas socioculturales como las condiciones laborales que siempre han estado marcadas por temas relacionados con la explotación laboral e infantil sobre todo en los países africanos, y los efectos adversos del cambio climático se convierten en factores importantes configurando el mercado como se conoce y cambiando las perspectivas a futuro (Foodcom Experts, 2024).

Otro factor importante para considerar, son los cambios en las preferencias de los consumidores. En la actualidad, las personas son mucho más conscientes de los productos que consumen día a día y todo está enfocado a una cultura más saludable al momento de comer, por esta razón las condiciones de cultivo y producción, el proceso de transformación y la disponibilidad de la información sobre el producto, son altamente valorados al momento de consumir productos derivados el cacao. El cacao el polvo con y sin adición de azúcar ha sido el impulsor de la demanda de los derivados de cacao en los mercados desarrollados y emergentes porque es un ingrediente fundamental al momento de conseguir otros productos de valor agregado como las chocolatinas, bombones y otros postres.

Existe un marcado contraste cuando se observa la situación del mercado del cacao en cada uno de los continentes. El sitio Foodcom Experts (2024) señala que a diferencia de África Occidental que suministra cerca del 70% de la producción de cacao en el mundo pero que atraviesa por una compleja situación debido a lo anteriormente mencionado, los países del sudeste asiático se han convertido en una alternativa para cubrir la demanda porque en estos países se han mejorado las prácticas agrícolas, Malasia y Filipinas son un ejemplo de países que se han dedicado a cambiar las prácticas de cultivo y ahora se centran en las prácticas sostenibles y de rentabilidad de largo plazo.

En el caso de Estados Unidos y Canadá, estos países son los principales consumidores de los productos derivados del cacao y están muy enfocados en demandar productos de la más alta calidad, de características ecológicas e invierten de mejorar las prácticas de los países productores para suplir las necesidades de los consumidores que ahora son más exigentes.

Adicionalmente, América Latina y particularmente Ecuador se empiezan a convertir en un lugar clave para suplir la demanda de cacao en el mundo porque cuenta con un clima estable y favorable para aumentar la competitividad internacional. Además, este país ha establecido políticas gubernamentales enfocadas en convertir el cacao en un producto importante al momento de hablar de exportación.

Sin embargo, la alta volatilidad en los precios del cacao y sus derivados también representa un gran problema para los productores porque dificulta la planificación estratégica y la toma de decisiones, ya que los productores se enfrentan a la incertidumbre sobre los precios futuros dificultando que sean competitivos a nivel nacional e internacional (Mora & Fontalvo, 2010).

Las dinámicas de los mercados internos de cacao en el mundo funcionan de maneras distintas, por ejemplo, según explican Abbott et al. (2019) en los países de África es muy común

que los compradores de cacao viajen hasta las fincas y negocien directamente con el productor porque las fincas se encuentran bastante retiradas y la información de cómo se encuentra el mercado del cacao es escasa, es decir, el comprador tiene la ventaja y en la mayoría por no decir todas las veces abusa de este poder.

Contrario a lo que ocurre en Colombia, donde la dinámica del mercado es distinta, ya que el comprador no viaja a la finca y es el productor es el que lleva su producto a los lugares donde se encuentran los compradores centrales o a las estaciones de compra en los que por lo general sólo encontrará dos opciones de grandes compradores que son Casa Luker o Nutresa, y el productor tiene en menor o mayor medida una relación con los asociados de dichas empresas por lo que tendrá un mayor conocimiento del comportamiento del mercado (Abbott et al., 2019).

En los lugares del país donde el cultivo del cacao es pequeño o relativamente nuevo, no están estos grandes compradores y existen mayores dificultades para la venta del producto y más para obtener un precio justo. De ahí que la falta de herramientas precisas de predicción de precios en el sector cacaotero limite la capacidad de los productores de negociar precios justos y puedan optimizar sus estrategias comerciales.

Como resultado, existe el riesgo de pérdidas financieras significativas y de inestabilidad económica en la industria cacaotera colombiana. Por lo anterior se ve la necesidad de utilizar técnicas de ciencia de datos en la búsqueda de soluciones que permitan una mejor comprensión y anticipación de las tendencias de precios se vuelve evidente para garantizar la viabilidad a largo plazo de la industria del cacao en Colombia.

Justificación

De acuerdo con Abbott et al. (2019) el gobierno colombiano y los gobiernos locales son los encargados de brindarles a los agricultores y productores las ayudas técnicas que necesiten a las fincas pequeñas y medianas. Sin embargo, estos programas no son realmente de ayuda porque suelen convertirse en cursos de corta duración en los que no se consigue la suficiente evidencia para evaluar los resultados y el impacto de dichos programas, y son aún más ineficientes cuando se trata de solucionar problemas persistentes.

La ley le exige a Fedecacao retribuir el 3% de la Cuota de Fomento Cacaotero a los lugares de donde se obtuvo, aquí empiezan las desventajas porque esto implica que las zonas que más producen serán las que más dinero obtengan y las zonas más pequeñas o emergentes no puedan recibir los recursos suficientes para mejor su productividad. En consecuencia, aunque Fedecacao tiene una red de asistencia técnica integral, en algunos sectores donde el cultivo del cacao es nuevo o muy pequeño se percibe que la presencia es escasa y el impacto que puede tener es casi inexistente (Ríos et al., 2017).

Los diferentes actores entrevistados por Ríos et al. (2017) perciben que las políticas implementadas hasta ahora no han producido los resultados esperados, porque no se han incrementado significativamente los volúmenes de producción y no se ha podido adoptar nuevas tecnologías por la cobertura limitada de la asistencia técnica. Adicionalmente, debido a que la variabilidad en los precios del cacao asociada a los ciclos de producción, la oferta y la demanda, dificulta la toma de decisiones de los productores y se ve una clara necesidad de buscar soluciones que permitan una mejor comprensión y anticipación de las tendencias de precios para garantizar la viabilidad a largo plazo de la industria del cacao en Colombia.

Los modelos de predicción para bienes básicos han evolucionado significativamente y un estudio de 2012 resaltó la superioridad de las redes neuronales en este ámbito. Gracias a su

capacidad para captar patrones no lineales, estas herramientas superan a otros métodos tradicionales, ofreciendo una mayor precisión en la predicción de precios, especialmente en sectores como el energético.

Por esta razón este proyecto propone construir un modelo de redes neuronales multicapa que permita determinar el comportamiento futuro del precio del cacao más específicamente del cacao en polvo con y sin adición de azúcar teniendo en cuenta las exportaciones. Esto se hace con el fin aprovechar la oportunidad que se está generando con el cambio de los hábitos de consumo que se está desplazando hacia consumidores con mayor conciencia y enfocados en los productos orgánicos y saludables.

Esta iniciativa no solo tiene implicaciones prácticas para la industria del cacao, sino que contribuirá al avance del conocimiento en el campo de la predicción de precios de productos básicos utilizando técnicas de aprendizaje automático.

Objetivos

Objetivo General

Evaluar el comportamiento del precio del cacao en polvo con y sin adición de azúcar mediante un modelo de redes neuronales multicapa.

Objetivos Específicos

Describir el comportamiento histórico del precio del cacao en polvo con y sin adición de azúcar con base en las exportaciones.

Desarrollar un modelo de redes neuronales multicapa con base en el comportamiento histórico de las exportaciones.

Evaluar la capacidad predictiva del modelo de redes neuronales multicapa sobre el comportamiento futuro del precio del cacao en polvo con y sin adición de azúcar.

Marco Teórico

En esta sección se elaborará una revisión bibliográfica de los conceptos generales a partir de los cuales se puede sustentar la importancia de evaluar un modelo de red neuronal multicapa para la predicción del precio del cacao en polvo con y sin adición de azúcar. Algunos de los conceptos a considerar son el aprendizaje automático (*machine learning*) y modelos de aprendizaje supervisado porque ayudará a entender porque el modelo escogido es el que mejor puede predecir el comportamiento de los precios, una breve explicación de qué es y cómo funciona el modelo de redes neuronales multicapa y entender cómo funciona la producción de cacao en el país.

Aprendizaje Automático

El aprendizaje automático o mejor conocido como machine learning hace referencia al conjunto de algoritmos que se crean y se entrenan para que las maquinas puedan aprender con base a los comportamientos del pasado del entorno que se quiere analizar, dichos algoritmos tienen la capacidad de encontrar relaciones y características únicas en los datos de manera que se pueda encontrar información nueva y de valor (Casas et al, 2019). Lo más importante del machine learning es tener claro el tipo de datos con los que se está trabajando y los algoritmos adecuados para analizar esos datos porque no todos los datos pueden ser tratados con todos los modelos.

Según Zhou (2021) el aprendizaje automático es lo que se necesita porque mejora el rendimiento al momento de aprender y los sistemas informáticos aprenden a través de la experiencia que está representada en datos y su misión es construir modelos a partir de dichos datos. En la actualidad, el aprendizaje automático ha ayudado a las industrias a crecer de manera exponencial y ha demostrado que usarlo de la manera correcta y eficiente puede ayudar a que las empresas perduren en el tiempo y como ejemplo de esto tenemos el *e-commerce* que crece día a

día gracias a la identificación de patrones de consumo o las ofertas de vuelos que se obtienen de analizar las búsquedas en los navegadores.

Métricas de Evaluación de Modelos Predictivos

Las métricas de evaluación son fundamentales para medir el desempeño y la calidad de un modelo predictivo, dichas métricas permiten analizar qué tan bien el modelo representa la relación entre las variables de entrada y los valores reales, asegurando que las predicciones sean precisas y útiles en el contexto del problema (Casas et al, 2019). A continuación, se describen las principales métricas utilizadas para evaluar la efectividad y la capacidad explicativa del modelo en este proyecto.

- MAE (Mean Absolute Error): definida como el error promedio absoluto entre las predicciones y los valores reales, indica qué tan cerca están en promedio las predicciones del modelo respecto a los datos reales. Un MAE más bajo implica una mayor precisión en las predicciones diarias y es útil para interpretar los errores de una manera directa y sin sesgo hacia valores extremos.

- MSE (Mean Squared Error): El MSE elevar al cuadrado los errores individuales, de esta forma penaliza de forma más severa los errores grandes, haciendo que el modelo sea más sensible a los valores atípicos. Esto hace que se identifiquen configuraciones que mantienen un bajo error general y que también manejan correcta los casos de predicciones fuera del rango. Un MSE bajo indica que el modelo es preciso, y además maneja bien los valores extremos o los datos menos comunes.

- R^2 (Coeficiente de Determinación): Indica la capacidad explicativa del modelo, es decir, calcula la proporción de la variabilidad en los datos que es explicada por el modelo. Un R^2 cercano a 1 significa que el modelo es capaz de capturar la mayor parte de la variación en los

datos. Ya que el objetivo del proyecto es crear un modelo explicativo y preciso, el coeficiente de determinación es muy importante para evaluar qué tanto el modelo capta la variabilidad en los datos.

- Tasa de Efectividad: La tasa de efectividad indica el porcentaje de predicciones que caen dentro de un rango aceptable de error. Esta métrica se alinea directamente con los objetivos del proyecto, ya que ayuda a evaluar la practicidad del modelo en un contexto de aplicación real.

Modelos de Aprendizaje Supervisado

Los modelos de aprendizaje supervisado se usan cuando se tiene claro cuáles son los resultados que se quieren obtener porque como explica Kane (2017) los modelos tienen claro desde el principio las agrupaciones o patrones que se buscan analizar y el ejemplo más común de este modelo son las predicciones de los precios de los autos, donde se entrena el modelo con la información histórica de las ventas de los autos y sus respectivos precios de forma que pueda aprender su comportamiento y pueda replicar a comportamientos futuros.

Por lo general, una forma sencilla de entender los modelos de aprendizaje supervisado es que compararán los resultados esperados con los datos proporcionados y para ello hay que brindarle un conjunto de entrenamiento con los datos de entrada y de salida del modelo (Casas et al, 2019). El problema consiste en poder predecir la variable dependiente Y respecto a los valores conocidos de las variables independientes X, cuando Y es una variable categórica se quiere resolver un problema de clasificación, si Y continúa el problema es de regresión (Ríos Insua & Gómez-Ullate, 2019).

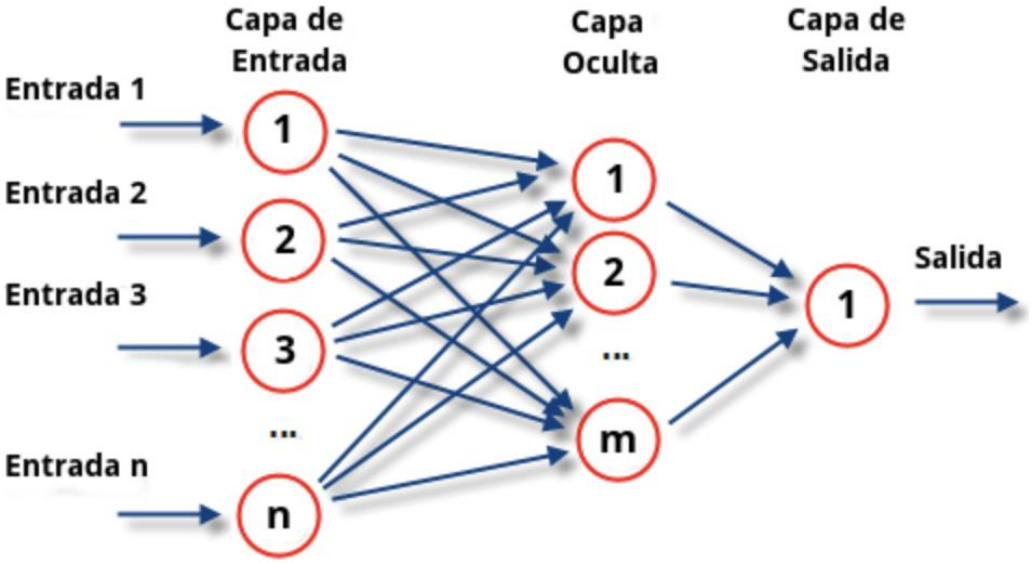
Modelos de Redes Neuronales

La forma más sencilla entender cómo funciona un modelo de redes neuronales es entendiéndolo desde el funcionamiento de las redes neuronales biológicas, entonces, la

información pasa a través de varias capas donde las variables son transformadas hasta que se obtiene un resultado final (Gil, 2016). Este modelo nace en 1943 cuando McCulloch y Pitts entendieron el cerebro humano como una unidad de cómputo donde cada neurona actúa como procesador digital y es útil porque la realidad no siempre se explica linealmente. El modelo se creó para imitar o simular el aprendizaje del cerebro humano de forma que cada nodo forma capas de entrada, ocultas y de salida, y cada nodo produce una señal de salida y recibe varias señales de entrada (Véliz, 2020).

Figura 1

Estructura del Modelo de Redes Neuronales



Nota. Tomado de *Qué son las redes neuronales y sus funciones,* Martínez E., 2024, ATRIA Innovation.

Según Véliz (2020) la forma de entender cada una de las capas presentes en el modelo es la siguiente: Una capa de entrada, también llamada sensorial, está compuesta por nodos o neuronas que reciben la información del entorno y proviene de las variables predictoras, la capa

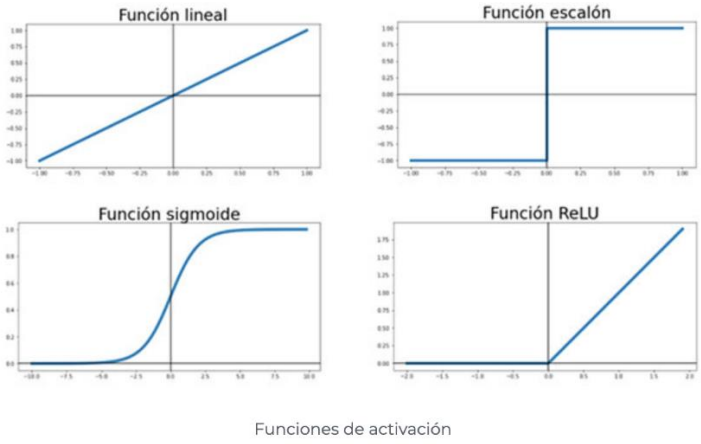
de salida está compuesta por nodos que representan la respuesta del sistema a través de la variable dependiente (p. 232).

Las capas de entrada pueden entenderse como las variables predictoras que transmitirán señales a cada capa escondida para producir una sola señal de salida, aunque no se relacionan directamente con las capas de entrada si influyen en la modelación de la capa de salida. Otra forma de entender las capas del modelo es la descrita por Villada (2016), donde las capas de entrada son la información disponible para desarrollarlo, las capas ocultas se entienden como activación del modelo y las capas de salida, la variable de respuesta al medio exterior.

Al igual que las neuronas biológicas disparan impulsos eléctricos, las funciones de activación en las redes neuronales determinan la intensidad con que se transmite la información entre las neuronas artificiales. Son el equivalente a la tasa de disparo de una neurona real. Entre las funciones de activación más conocidas de se encuentran: función sigmoidal, función tangente hiperbólica, función rectificadora y función *softmax* (Sosa y Zamora, 2022).

Figura 2

Funciones de Activación del Modelo de Redes Neuronales



Nota. Tomado de *Redes Neuronales y Deep Learning. Capítulo 2: La neurona*, Alonso F., 2021, Space S.A

Aquí también entra el concepto de función de pérdida o la función de coste, Sosa y Zamora (2022) la explican como las funciones de coste guían el aprendizaje de una red neuronal. Al calcular la diferencia entre las salidas predichas y las esperadas, estas funciones indican a la red cómo ajustar sus parámetros para mejorar su desempeño.

Modelos de Predicción en el Mercado de Materias Primas

La predicción de precios en el mercado de materias primas se ha abordado con modelos estadísticos y análisis fundamentales. Sin embargo, estos métodos a menudo tienen limitaciones para capturar la complejidad de los datos y las interacciones entre variables. En respuesta a estos desafíos, se ha interesado en el uso de modelos de inteligencia artificial, particularmente redes neuronales multicapa, para el pronóstico de los precios. En el contexto del mercado de materias primas, las redes neuronales han demostrado su capacidad para capturar tendencias y patrones no lineales que pueden pasar desapercibidos para otros enfoques analíticos (Villada, 2016).

Estos modelos aprenden patrones y relaciones complejas a partir de conjuntos de datos, lo que los hace adecuados para predicción de series temporales y análisis de datos financieros. El uso de redes neuronales en la predicción de precios del cacao ofrece la promesa de modelos más precisos y robustos, capaces de adaptarse a la dinámica cambiante del mercado y de proporcionar información valiosa para la toma de decisiones en la industria cacaotera.

Optimización Bayesiana

De acuerdo con Cátedra Santalucía de Analytics for Education (s.f.), la optimización bayesiana es una técnica de optimización que se usa para encontrar los mejores valores de los hiperparámetros en un modelo de aprendizaje automático. La optimización bayesiana es una técnica muy poderosa que puede ser aplicada para solucionar problemas de optimización no convexa de cajas negras.

La optimización bayesiana está fundamentada en el uso de modelos probabilísticos con el propósito de modelar la función objetivo de forma que a medida que cambia la combinación de hiperparámetros, también lo hace el error en el aprendizaje automático. Este modelo probabilístico se actualiza a medida que se evalúan nuevos puntos en el espacio de búsqueda.

El Mercado del Cacao

El cacao es un cultivo crucial para las economías de varios países en América Latina, África y Asia, y su comercio internacional genera miles de millones de dólares cada año. El mercado del cacao es altamente influenciado por factores tanto internos como externos. Entre los factores internos se incluyen la producción agrícola, los costos de mano de obra, las políticas gubernamentales y la infraestructura de transporte. Por otro lado, los factores externos incluyen las condiciones climáticas, las tendencias del mercado global, las fluctuaciones en los precios de otras materias primas y los cambios en las políticas comerciales internacionales (Amaya, 2020).

Figura 3

Plata de Cacao

Nota. Tomada de *Estrategias país para la oferta de cacaos especiales -Políticas e iniciativas privadas exitosas en el Perú, Ecuador, Colombia. Colombia más competitiva,* Ríos et al., (2017).

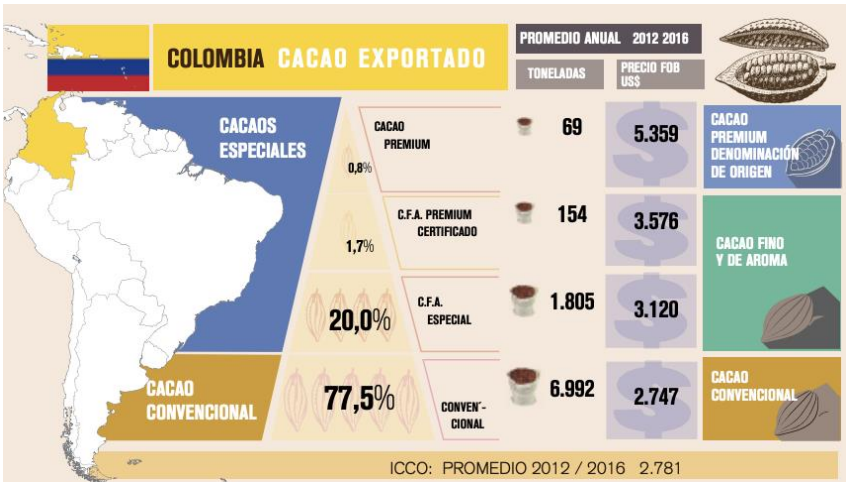
La volatilidad en el precio del cacao puede tener un impacto significativo en los productores, procesadores y consumidores finales. La capacidad de predecir con precisión las tendencias futuras del precio del cacao es esencial para tomar decisiones informadas en todos los niveles de la cadena de suministro.

Exportaciones de Cacao

La producción del grano de cacao ha ganado relevancia a nivel mundial, convirtiéndose en el cuarto producto alimenticio más exportado. Su impacto es significativo en el PIB de los países productores, especialmente en Costa de Marfil, que aporta cerca del 36% del cacao comercializado globalmente. Asia y América siguen en importancia, destacando América por su clima favorable que da un sabor y aroma distintivos al cacao. El cacao colombiano, con características especiales muy demandadas en el mercado europeo, ha encontrado una ventaja competitiva con el TLC entre Colombia y la UE, beneficiándose de diversas desgravaciones arancelarias (Amaya, 2020).

Figura 4

Exportaciones de Cacao en el Periodo 2012 a 2016



Nota. Tomada de *Estrategias país para la oferta de cacaos especiales -Políticas e iniciativas privadas exitosas en el Perú, Ecuador, Colombia. Colombia más competitiva,* Ríos et al., (2017).

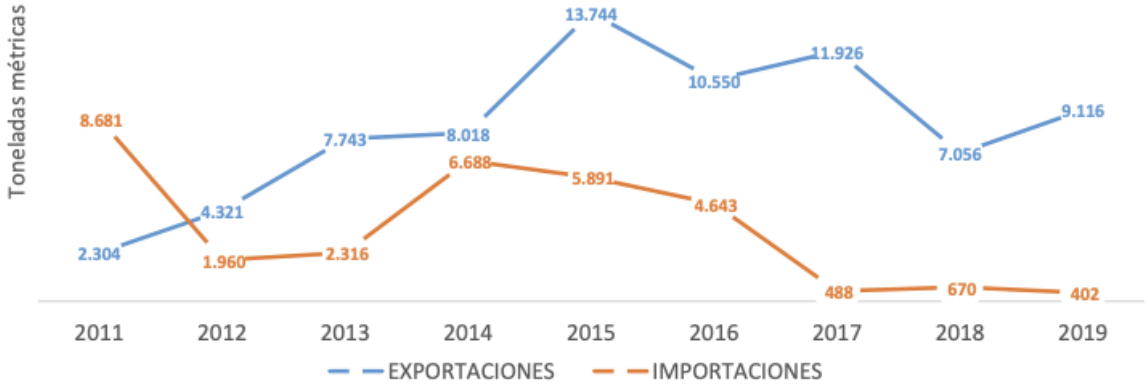
De acuerdo con Cobaleda (2022) a diferencia de países como Ecuador que destinan la mayor parte de su producción de cacao a la exportación, Colombia tiene un enfoque más diversificado. Una porción significativa de su cacao se transforma en productos como chocolate dentro del país. No obstante, las exportaciones colombianas de cacao y sus derivados han experimentado un crecimiento notable.

Sin embargo, el sector cacaotero colombiano se enfrenta a múltiples obstáculos que impiden su crecimiento. Entre los principales desafíos se encuentran la baja productividad, caracterizada por rendimientos por hectárea insuficientes, y una infraestructura deficiente que limita la comercialización y encarece la producción. Estos problemas se agudizan por la falta de acceso a tecnologías modernas, la utilización de variedades tradicionales de bajo rendimiento y la presencia de enfermedades y plagas (Montoya-Restrepo et al., 2015).

El cultivo de cacao se ha posicionado como una herramienta clave para transformar regiones afectadas por conflictos, generando empleo y desarrollo económico a través de este sistema productivo. Los datos muestran un aumento del 29% en el área sembrada y del 41% en la producción de cacao en el período 2010-2019. No obstante, los rendimientos promedio nacionales se han mantenido relativamente estables en torno a los 450 kg/ha, lo que indica una brecha significativa con respecto a los rendimientos potenciales (Ministerio de Agricultura y Desarrollo Rural, 2020).

Figura 5

Importaciones y Exportaciones de Cacao en el Periodo 2011 a 2019



Nota. Tomado de *Cadena de cacao*, Direccion de cadenas agrícolas y forestales, 2020, Minagricultura.

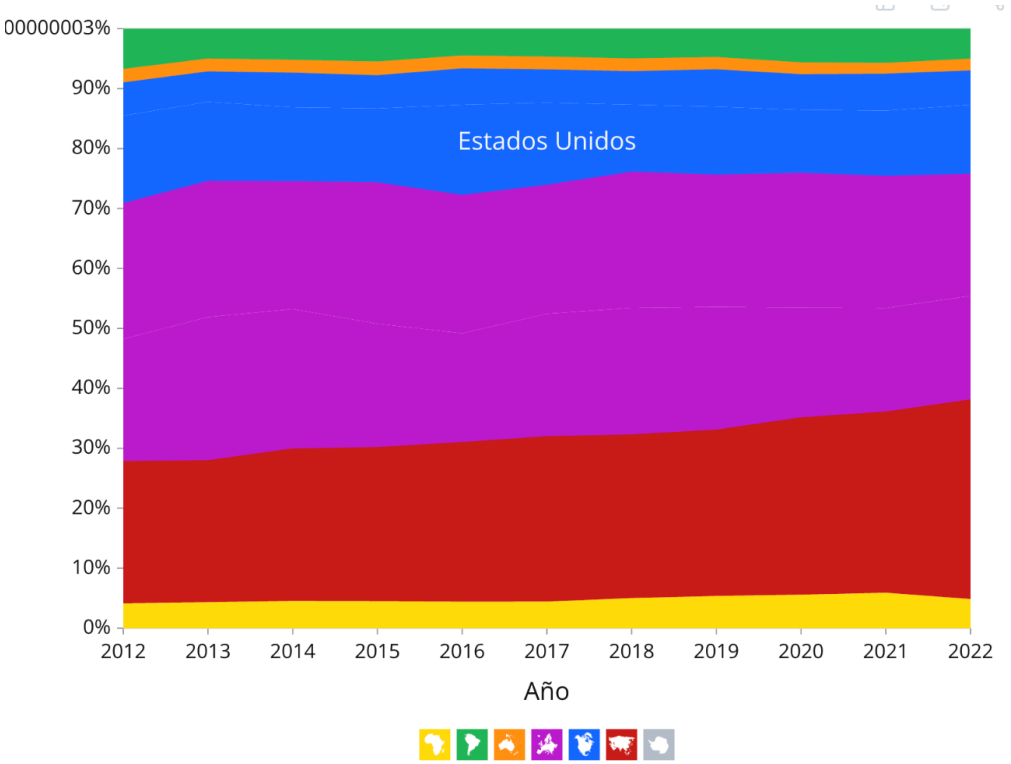
El Ministerio de Agricultura y Desarrollo Rural (2020) señala que las exportaciones de cacao en grano en Colombia mostraron un crecimiento sostenido entre 2011 y 2015, pero experimentaron una caída en 2016 debido a diversos factores. A partir de 2017, las exportaciones volvieron a aumentar, aunque a un ritmo más lento, y se vieron afectadas por la disminución del valor del cacao en el mercado internacional.

No obstante, ya no solamente se exporta el cacao en grano, aunque sigue siendo la variedad que más se exporta en el país. FORBES Colombia (s.f.) destaca que en el año 2023 los productos semielaborados y elaborados representaron alrededor de US$ 131 millones más de lo que se obtuvo en 2022. En los últimos años las empresas entendieron que darle un valor agregado al cacao en grano podría darles la posibilidad de competir en el mercado internacional, ya que actualmente se pueden encontrar más de 15 marcas con calidad excepcional y unos 400 transformadores del cacao en grano.

Cuando se analiza las exportaciones de cacao en polvo con y sin adición de azúcar en el periodo de tiempo comprendido entre el 2014 y el 2023 se puede evidenciar que en el continente europeo España se destaca como el principal destino de las exportaciones, recibiendo una cantidad significativamente mayor de envíos en comparación con los demás países. A pesar de la cantidad de países de la Unión Europea representados (color magenta en la figura 6), la mayor parte de las exportaciones se concentra en un número reducido de ellos, principalmente España, Reino Unido, Bélgica y Alemania.

Figura 6

Participación de las Exportaciones de Cacao en Polvo sin Azúcar.

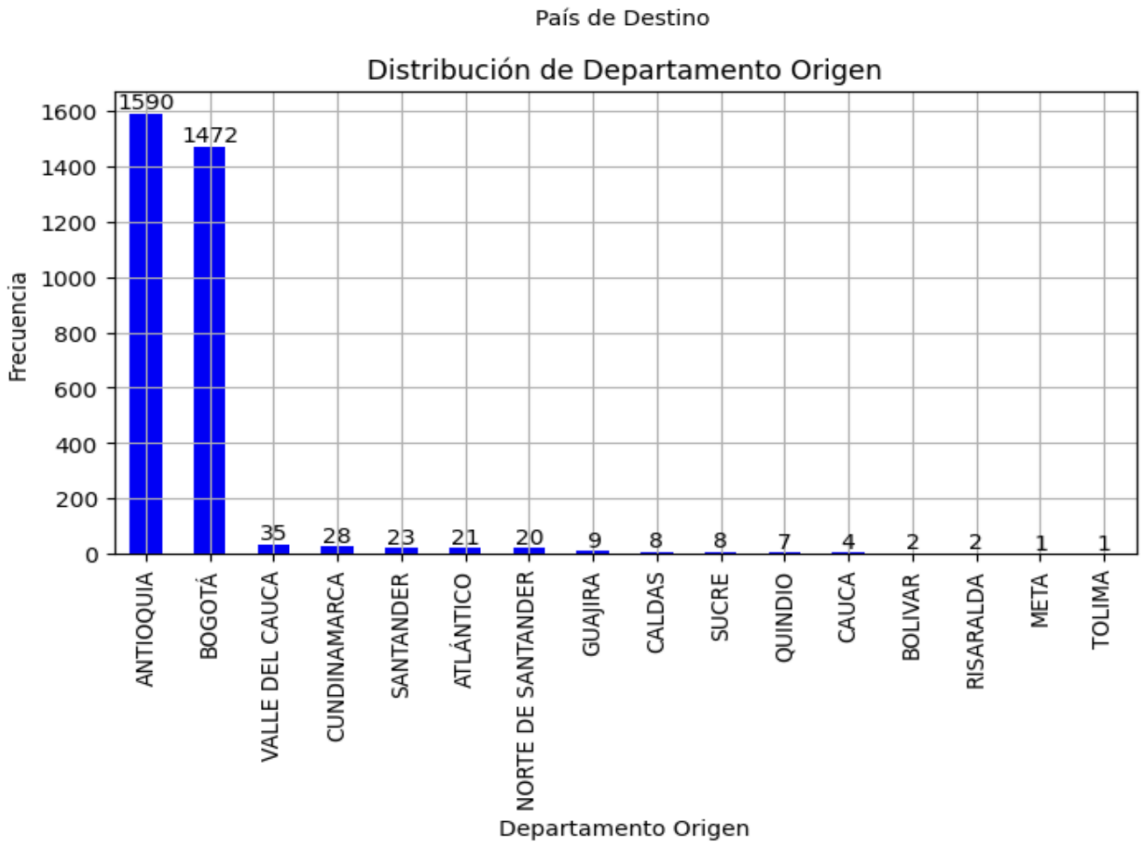


Nota. Tomado de *Cacao en polvo, sin adición de azúcar ni otro edulcorante en Colombia* (s/f), Observatorio de Complejidad Económica.

Otro patrón distintivo de estas exportaciones es el departamento de origen de las empresas encargadas de realizar estas exportaciones. La figura 7 señala que Bogotá se destaca como el principal departamento de origen, concentrando una proporción significativamente mayor de la exportación en comparación con los demás departamentos, esto habla del nivel de intermediación tan grande existente en el sector porque las principales regiones cacaoteras tienen muy poca participación en las exportaciones.

Figura 7

Frecuencias de las Exportaciones 2014 - 2023 por Departamento de Origen



Nota. Elaboración propia con información de Exportaciones de cacao y sus preparaciones, fuente de los datos Dirección De Impuestos Y Aduana De Colombia DIAN a través de la base de datos Legis Comex.

Según FEDECACAO (2023) los principales departamentos productores de cacao son Santander, Arauca, Antioquia, Tolima, Huila y Nariño. Pero solo el departamento de Antioquia cuenta con empresas que transforman los granos de cacao en polvo para la exportación, esto habla de un factor de oportunidad para que los demás departamentos puedan hacer lo mismo y reducir la cantidad de intermediarios que existen hasta llegar a la exportación y de esta forma conseguir mejores precios para los involucrados.

A pesar de que se ha evidenciado una reducción en la producción cuando se compara cómo ha evolucionado la producción en el periodo comprendido entre 2012 y 2022 se puede observar que se ha mantenido y un crecimiento en la producción. Además, se ha observado que, aunque las exportaciones del cacao en grano disminuyeron en 2022 en comparación con las toneladas exportadas en 2021, las exportaciones del cacao en polvo con y sin adición de azúcar tuvieron un comportamiento positivo (FEDECACAO, 2023).

Figura 8

Exportaciones Anuales de Cacao en Polvo con y sin Azúcar

Nota. Elaboración propia con información de Exportaciones de cacao y sus preparaciones, fuente de los datos Dirección De Impuestos Y Aduana De Colombia DIAN a través de la base de datos Legis Comex.

La figura 8 muestra que la gran mayoría de las declaraciones de exportación definitivas se concentran en el año 2022. La forma del violín es más ancha en esta zona, lo que indica una mayor densidad de datos y a medida que retrocedemos en el tiempo (años 2020, 2018, etc.), la densidad de datos disminuye significativamente. Fedecacao ha venido implementado medidas como la renovación de cacaotales envejecidos para combatir las pérdidas de cultivos debido a cambios climáticos o las plagas (FEDECACAO, 2023).

Características del Cacao en Polvo

El cacao en polvo, obtenido tras extraer la grasa y moler las semillas de cacao, es esencial en la industria alimentaria. Se utiliza en la elaboración de galletas, tortas y otros tipos de productos de panadería, así como en la fabricación de bebidas de chocolate, chocolates, coberturas y bombones, y en la aromatización de helados, glaseados y bebidas. Además de sus propiedades sensoriales, el cacao en polvo natural es rico en polifenoles antioxidantes, asociados con una buena salud cardiovascular.

Figura 9

Cacao en Polvo

Nota. Tomado de *Cacao: ¿cómo es su cultivo y producción,* Experts F., 2024, Foodcom S.A

Los polifenoles del cacao, como los flavonoles, son similares a los presentes en el vino y el té, y se asocian con la capacidad antioxidante del cacao, que protege contra daños celulares causados por radicales libres, reduciendo el riesgo de enfermedades como el cáncer y enfermedades cardiovasculares. Además, se ha investigado su relación con la disminución de la presión arterial y otros efectos positivos, como analgesia, anti-inflamación y actividad antimicrobiana (Esteve, 2016).

El cacao en polvo natural es también una fuente importante de fibra dietética, cuyo consumo se vincula con la prevención de trastornos y algunas enfermedades que llegan a ser comunes en los países desarrollados, como obesidad, diabetes y cáncer. Debido a estos beneficios, el cacao en polvo se considera un ingrediente funcional prometedor para la industria alimentaria. La alcalinización es un proceso que mejora las propiedades sensoriales y tecnológicas del cacao en polvo.

Este proceso, que implica el tratamiento con álcalis como carbonato sódico o potásico, reduce la acidez, la astringencia y aumenta la solubilidad del cacao. Además, desarrolla aromas, sabores y colores característicos. Sin embargo, la alcalinización puede alterar el perfil nutricional del cacao al reducir su contenido de polifenoles y capacidad antioxidante, aunque estos efectos dependen de las condiciones específicas del proceso y la composición del cacao (Esteve, 2016).

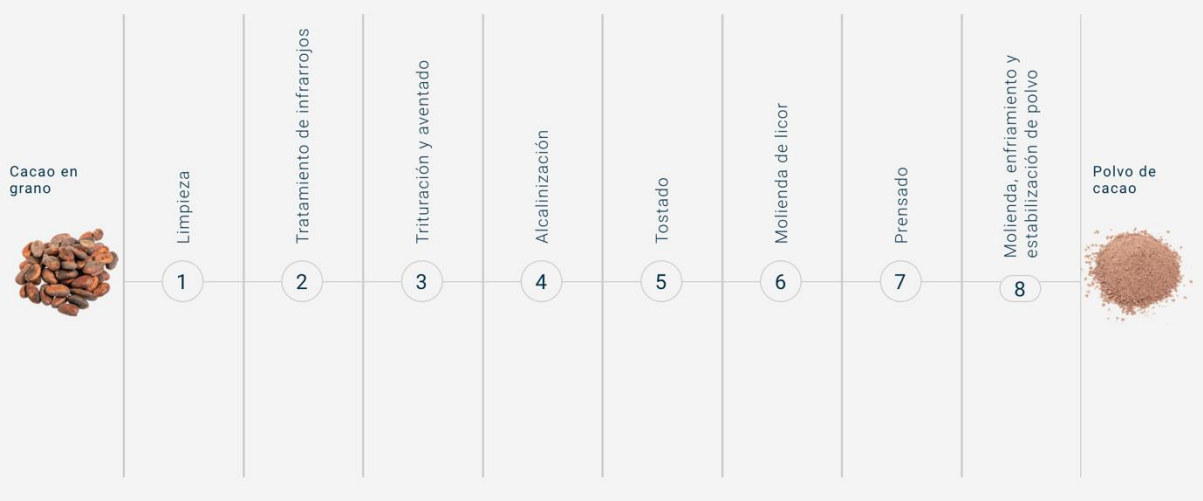
Procesamiento del Cacao en Polvo

El proceso comienza con la entrada del cacao a la máquina tostadora, donde se cocina a 85 grados durante 30 minutos. Este paso ablanda la cáscara que lo recubre. Al salir de la tostadora, el cacao tiene la cáscara blanda y se transfiere a la máquina descascarilladora, que separa los nibs de la pepita del cacao. Utilizando una malla, la máquina clasifica las pepitas por tamaño (medianas y grandes). Luego, las pepitas se muelen en un molino, obteniendo el licor de

cacao. Este licor se lleva a una prensa, donde se convierte en polvo y manteca de cacao, productos listos para el empaquetado. De este proceso se obtienen tres productos: nibs, polvo de cacao y manteca de cacao (Preciado et al., 2021).

Figura 10

Procesamiento del Cacao



Nota. Tomado de *Procesamiento del cacao (s.f.)*, Buhlergroup.com

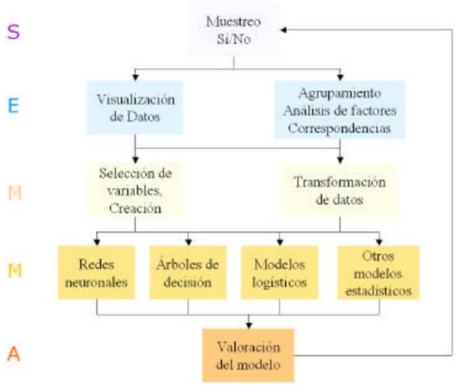
Metodología

El aumento en la capacidad de almacenamiento, la generación masiva de datos y el desarrollo de algoritmos más eficientes han impulsado la creación de diversas metodologías para estructurar estudios relacionados con el análisis predictivo y la minería de datos. Estas metodologías, desarrolladas por profesionales del campo, permiten organizar y ejecutar de manera clara y eficiente los diferentes tipos de análisis.

La metodología que se aplica en este proyecto es SEMMA, llamada así por sus siglas en inglés (*Sample*, *Explore*, *Modify*, *Model*, *Assess*), desarrollada por SAS Institute como parte de su software de minería de datos, SAS Enterprise Miner, se convirtió en una de las metodologías más reconocidas en el campo de machine learning por su efectividad para procesar grandes volúmenes de datos para encontrar patrones y generar información útil para la toma de decisiones (Rodríguez et al., s.f.). SEMMA está centrada en los aspectos técnicos y estadísticos del análisis de datos, optimizando el flujo de trabajo desde la selección de los datos hasta la construcción y evaluación de modelos.

Figura 11

Metodología SEMMA



Nota. Tomado de *Metodologías para la realización de proyectos de data mining,* Rodríguez et al. (s.f.)

En primer lugar, se recolectan y preparan los datos, eliminando inconsistencias y transformándolos en un formato compatible con el modelo. A continuación, se diseña la arquitectura del modelo, seleccionando el tipo de modelo y sus componentes. El entrenamiento del modelo consiste en ajustar sus parámetros para aprender las relaciones entre los datos de entrada y salida. Una vez entrenado, se evalúa su desempeño en un conjunto de datos no vistos previamente. Si los resultados no son óptimos, se itera sobre el proceso, ajustando hiperparámetros, modificando la arquitectura o recolectando más datos.

Al momento de trabajar en la predicción de precios de un producto como el cacao en polvo que tiene un comportamiento no lineal y en el que intervienen tantos factores externos, los modelos de regresión no son tan efectivos ya que uno de sus principios fundamentales es la relación lineal que existe entre las variables de forma que no puede capturar de forma clara las interacciones entre diferentes variables independientes, lo cual es crucial para explicar el comportamiento de los precios.

Un modelo que, si es capaz de capturar dichas interacciones no lineales entre las variables en la predicción de precios, es el modelo de redes neuronales artificiales. Este modelo de redes neuronales se ha convertido en una herramienta popular al momento de modelar los fenómenos económicos y financieros por su poder de cálculo y la implementación ya que se puede llegar a los resultados sin necesidad un análisis previo de las variables e identifica patrones de comportamiento difíciles de obtener con otro tipo de modelos (Arango, 2017).

Para obtener datos comerciales de confianza y donde se maneje los datos históricos, se realizó la descarga de las exportaciones anuales de cacao en polvo con y sin adición de azúcar que se encuentran disponibles en la base de datos de Legis Comex, se descargaron datos que

comprenden el periodo de 2010 a 2023, además, se incluyó el salario mínimo colombiano de cada uno de estos años. Esta data está conformada por 68 columnas y 6343 registros.

Muestreo (Sample)

Al observar el conjunto de datos obtenidos, los cuales van desde el año 2010 hasta el 2023, se identificó que los datos anteriores a 2014 presentan un comportamiento diferente en comparación con los demás datos. Esta variabilidad en los datos históricos puede estar relacionada con cambios en el mercado, la introducción de nuevas tecnologías o modificaciones en las políticas de exportación, factores que alteran los patrones y afectan la creación de los modelos predictivos.

Por lo tanto, para asegurar la relevancia y la precisión del análisis, se opta por tomar en cuenta únicamente los datos de los últimos 10 años, es decir, desde 2014 hasta 2023. Esta decisión se fundamenta en que los datos más recientes reflejan de manera más precisa las condiciones actuales y el comportamiento reciente del mercado, siendo así más adecuado para el modelado y mejorando la capacidad de generalización. De esta forma el conjunto de datos se reduce a 3.390 registros.

Exploración (Explore)

En esta fase del proyecto se realizó un análisis exploratorio de datos, iniciando con calcular una tabla resumen con las estadísticas descriptivas, en el caso de las variables numéricas se observaron para cada variable el promedio, la desviación estándar, sus valores mínimos y máximos y los rangos intercuartiles (25% - 50% - 75%); Para las variables categóricas se calculó el número de valores o categorías únicas, su valor más repetido, y la frecuencia de dicho valor. De igual manera se identificaron las variables con valores nulos y el porcentaje de estos, se identificaron valores atípicos en la variable objetivo “Precio Unitario FOB (USD) Peso Neto”.

Mediante la observación e interpretación del conjunto de datos, junto con un análisis de estadísticas descriptivas para las variables tanto numéricas como categóricas se identificó la existencia de variables que no son pertinentes para los objetivos de este trabajo, ya que tienen las siguientes características: hacen referencia a identificadores del número de registros, presentan información repetida en otras, presentan más del 80% de sus registros como valores nulos; Además de lo anterior y basados en el marco teórico y al estudio del estado del arte, se realizó un análisis de características Redundantes o Irrelevantes, con lo que se encontró un número importante de variables que no se consideran relevantes para el objetivo de este proyecto, ya que hacen referencia a códigos del proceso aduanero e identificadores de identidad de las partes involucradas en el proceso de exportación.

Al analizar la variable objetivo "Precio Unitario FOB (USD) Peso" mediante histogramas y gráfica *boxplot* se observa que esta tiene unos valores atípicos que pueden llegar a generar conflictos con el modelo.

Por lo tanto, por medio del cálculo intercuartil se identifican los limites superior e inferior:

$$IQR = Q_3 - Q_1 \quad (1)$$

$$limite\ inferior = Q_1 - 1.5 * IQR \quad (2)$$

$$limite\ superior = Q_3 - 1.5 * IQR \quad (3)$$

Con los límites establecidos se identifican 159 registros que al estar por fuera de los limites son considerados valores extremos, por lo cual se procede a eliminar dichos valores, se vuelve a graficar las distribuciones para poder observar los cambios en la variable. La figura 14 evidencia una mejoría en la distribución de la variable objetivo, cobrando un mayor sentido los valores atribuidos para trabajar con aquellos que se ajustaron a los límites establecido.

Figura 12

Histograma de la Variable Objetivo Después de Quitar los Valores Atípicos

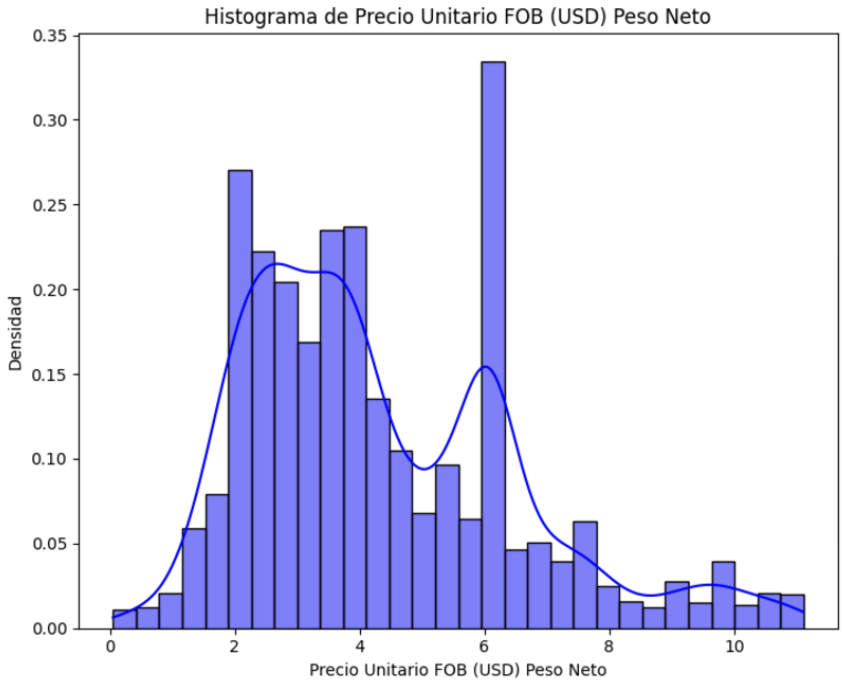
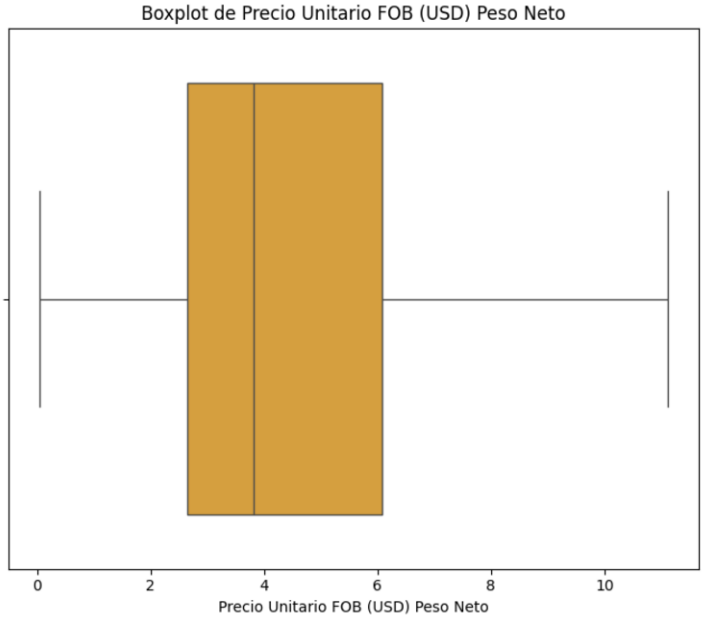


Figura 13

Boxplot de la Variable Objetivo Después de Quitar los Valores Atípicos



Una vez aplicada la eliminación de los valores atípicos se obtiene un conjunto de datos con un tamaño de 18 columnas con 3.231 registros para poder crear el modelo. Sin embargo, hay que seguir reduciendo la cantidad de variables con las que se va a trabajar con el fin de modelar solo con variables que sean realmente relevantes para el ejercicio.

Luego, con un conjunto de datos más limpio y pertinente se realiza un análisis bidimensional que pueda explicar la relación de las diferentes variables respecto a la variable objetivo. El análisis de correlación de Pearson y Kendall entre las variables numéricas y el "Precio Unitario FOB (USD) Peso Neto" reveló patrones consistentes de relaciones mayoritariamente débiles.

En Pearson, la variable "Peso en kilos netos" mostró una correlación media negativa ($r = -0.349$, $p < 0.001$), mientras que en Kendall presentó una correlación débil negativa ($r = -0.259$, $p < 0.001$). Otras variables como "Número de artículos" y "Valor Agregado Nacional (VAN)" mostraron correlaciones débiles positivas en ambos métodos, aunque con variaciones en magnitud. de la misma forma variables como "Valor Flete" y "Valor seguro" tuvieron correlaciones débiles negativas en ambas métricas.

En general, las correlaciones, aunque significativas para la mayoría de las variables, fueron principalmente débiles en fuerza. Esto sugiere que el "Precio Unitario FOB (USD) Peso Neto" no está fuertemente asociado con estas variables. Estas observaciones destacan la necesidad de explorar variables adicionales o enfoques más complejos para comprender mejor las relaciones subyacentes en el conjunto de datos.

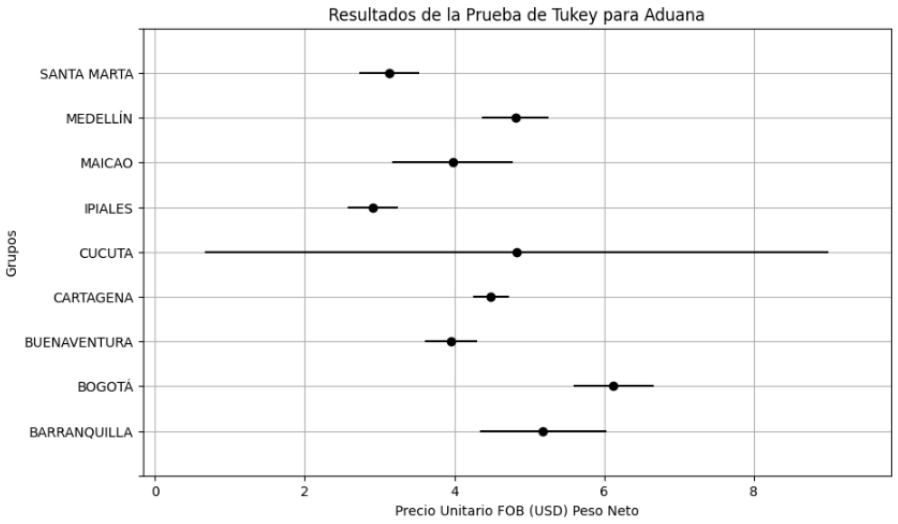
Debido a que los análisis de correlaciones no mostraron una relación fuerte entre ninguna de las variables independientes con nuestra variable dependiente, no es posible descartar ninguna de estas variables en este punto del trabajo.

Para las variables categóricas, se ejecuta una prueba ANOVA para comparar las medias de cada una de las variables categóricas e identificar si tienen un efecto sobre la variable objetivo. Con los resultados de la prueba ANOVA se realiza una prueba Post Hoc de Tukey, esta prueba se utiliza para identificar específicamente cuáles pares de grupos tienen diferencias significativas en sus medias. Con los resultados de la Post Hoc de Tukey se observó que en todas las variables existen grupos que se diferencian en sus medias con otros grupos, pero de igual manera en todas también existen grupos que sus medias se solapan.

A continuación, se observa el resultado grafico de la prueba de Turkey para la variable Aduana, en donde se puede identificar diferencias significativas entre los pares de aduanas de Santa Marta – Medellín, Santa Marta – Bogotá, Santa Marta – Barranquilla, Medellín – Ipiales, Ipiales – Cartagena, Buenaventura – Bogotá; Sin embargo, también se identifican aduanas que se solapan entre sí, como es el caso de Santa Marta – Ipiales, Medellín – Bogotá, y Cúcuta con todas las aduanas.

Figura 14

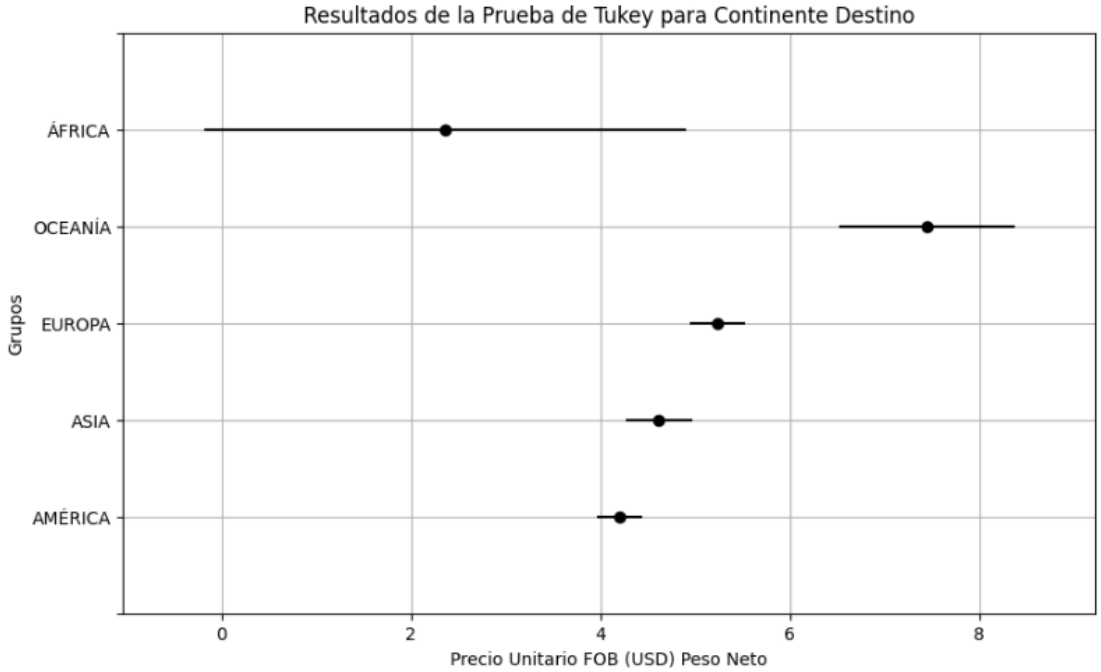
Resultados de la Prueba Post Hoc de Tukey para Aduana



En el caso de la variable continente de destino, el resultado grafico de la prueba de Turkey muestra diferencias significativas entre los destinos África – Oceanía, Oceanía – Europa, Oceanía – África, Oceanía – América, Europa – América, pero también muestra pares de destinos solapados como son África – Asia, África –América, Asia – América.

Figura 15

Resultados de la Prueba Post Hoc de Tukey para Continente Destino



Mediante la aplicación de esta prueba a cada una de las variables categóricas se lograron identificar variables como "Aduana" ($F = 30.253$, $p < 0.001$), "Municipio" ($F = 30.269$, $p < 0.001$), y "Descripción de la partida arancelaria" ($F = 88.186$, $p < 0.001$) las cuales destacaron por su relevancia estadística, indicando que los valores del precio unitario varían significativamente entre los diferentes niveles de estas categorías. De manera similar, se encontraron diferencias significativas para "País de destino", "Departamento de origen", "Vía de transporte", "Modalidad de exportación", "Forma de pago" y "Continente destino", con valores p inferiores a 0.001.

En contraste, la variable "Moneda de negociación" (F = 1.305, p = 0.259) no presentó diferencias significativas entre grupos, lo que sugiere que el "Precio Unitario FOB (USD) Peso Neto" no depende del tipo de moneda utilizada en las transacciones, es por ellos que en este punto se descarta la utilización de dicha variable.

Modificar (Modify)

Con el fin de escoger las variables más representativas para la correcta creación del modelo, se usó el método de *Random Forest*, el cual es un modelo de ensamble basado en múltiples árboles de decisión, al crear estos árboles y combinarlos, *Random Forest* puede modelar relaciones no lineales y de alta interacción entre las variables calculando la importancia de cada variable en función de su capacidad para reducir el error de predicción.

Para aplicar este proceso primero fue necesario la codificación de las variables categóricas mediante la técnica de One-Hot Encoding, la cual transforma cada variable categórica en un conjunto de variables binarias (0 o 1), donde cada categoría única de la variable original se convierte en una nueva columna. Además de esto es necesario realizar un escalado de las variables con el objetivo de que las variables tengan una media de 0 y una desviación estándar de 1, lo cual ayuda a evitar que las variables con mayores magnitudes dominen a las variables con menores magnitudes.

El modelo de Random Forest para la selección de las variables más adecuadas se configuro de la siguiente manera:

Figura 16

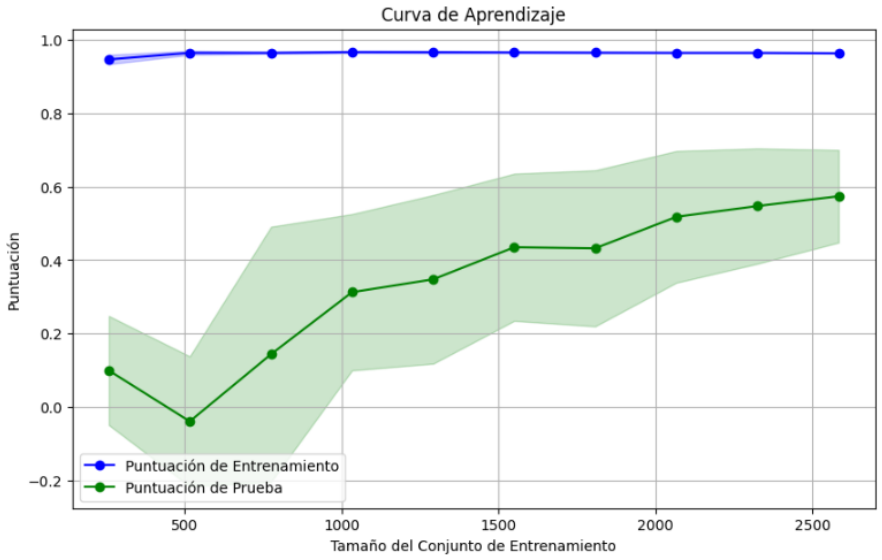
Configuración del Modelo de Random Forest

```
# Crear el modelo de Random Forest
rf = RandomForestRegressor(
    n_estimators=200,               # número de árboles
    max_depth=None,                 # profundidad máxima de los árboles
    min_samples_split=2,            # Número mínimo de muestras para dividir un nodo
    min_samples_leaf=1,             # Número mínimo de muestras por hoja
    max_features=None,              # número de características
    criterion="squared_error",      # Criterio de división mse "squared_error" o mae  "absolute_error"
    random_state=16                 # Semilla
)
```

Una vez realizado el método de Random Forest se obtiene la siguiente curva de aprendizaje, en donde la puntuación del conjunto de entrenamiento se mantiene cercado a 1 en los diferentes tamaños de datos, mientras que la puntuación del conjunto de prueba va en aumento a medida que aumenta el tamaño del conjunto de entrenamiento, llegando hasta una puntuación de 0.6 con un conjunto de entrenamiento de 2500 registros.

Figura 17

Curva de Aprendizaje del Modelo



En la siguiente imagen se puede ver los primeros 3 niveles del primer árbol de decisión del modelo de Ramdon Forest, en él se puede identificar como la variable "Peso en kilos netos" es la primera en dividir los datos en la raíz, con un umbral de `<= 0.288`, separando los valores en dos ramas: izquierda (menores o iguales) y derecha (mayores), con un error cuadrático medio de `4.675` y una predicción promedio (value) de `4.417`. En el segundo nivel, la rama izquierda utiliza la variable "País de Destino_ESTADOS UNIDOS" para dividir en dos subramas según si el destino es Estados Unidos (`<= 0.5`), mientras que, en la rama derecha, la división vuelve a realizarse con la misma variable. En el tercer nivel, la rama izquierda del segundo nivel continúa dividiéndose con "Peso en kilos netos" <= -0.615 y "Número de artículos" <= 0.783, mientras que en la rama derecha aparece la variable "País de Destino_COREA DEL SUR", seguida de una nueva división por "Peso en kilos netos". Las divisiones sucesivas buscan minimizar el error cuadrático en cada uno de los nodos, mostrando el número de muestras (samples), la predicción promedio (value) y el squared_error, siendo "Peso en kilos netos" y "País de Destino" las variables con mayor poder predictivo ya que aparecer en los niveles iniciales del árbol:

Figura 18

Primeros 3 Niveles del Primer Árbol de Decisión.

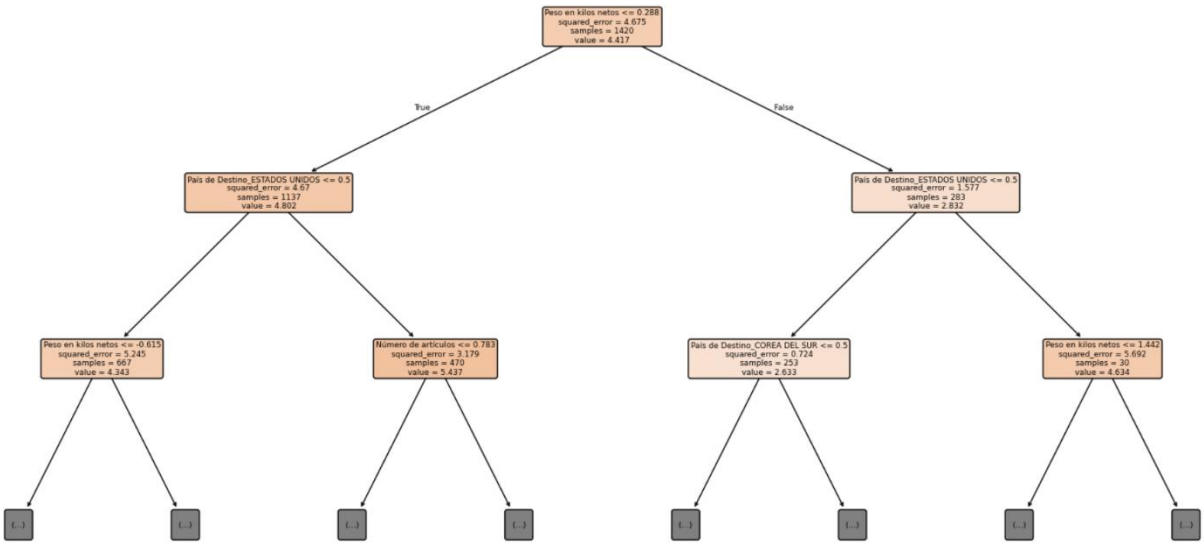


Tabla 1

Resultados de la Prueba de Importancia

Variable	Importancia
Peso en kilos	0.5873
País de Destino	0.5100
Fecha de Declaración de Exportación Definitiva	0.1368
Departamento Origen	0.0671
Número de artículos	0.0619
Descripción de la partida arancelaria	0.0541
Forma de pago	0.0482
Valor seguro	0.0384
Valor Flete	0.0187
Continente Destino	0.0176
Municipio	0.0173
Aduana	0.0127
Moneda de negociación	0.0061
Vía de transporte	0.0059
Salario mínimo COP	0.0021
Modalidad de exportación	0.0010
Valor Agregado Nacional (VAN)	-0.0000
Valor otros	-0.0002

El análisis de importancia de variables utilizando *Random Forest* destacó los factores clave que influyen en el "Precio Unitario FOB (USD) Peso Neto". La variable "Peso en kilos netos" obtuvo la mayor puntuación (0.5873), indicando que es el principal determinante del

precio. "País de Destino" (0.5100) también mostró una contribución significativa, lo que refleja la influencia de la localización geográfica en los costos asociados.

Variables como "Fecha de Declaración de Exportación Definitiva" (0.1368), "Departamento Origen" (0.0671), "Número de artículos" (0.0619) y "Descripción de la partida arancelaria" (0.0541) tuvieron impactos moderados, lo que sugiere que características relacionadas con el tiempo, la ubicación y el tipo de producto afectan el precio unitario. Otras variables como "Forma de pago" (0.0482), "Valor seguro" (0.0384) y "Valor Flete" (0.0187) aportaron en menor medida, pero pueden llegar a ser relevante en algunos casos.

Por otro lado, variables como "Moneda de negociación" (0.0061) y "Vía de transporte" (0.0059), muestran una influencia demasiado baja en el precio unitario, mientras que las variables "Valor Agregado Nacional (VAN)" (-0.0000) y "Valor otros" (-0.0002) no presentaron ninguna relevancia significativa.

Finalmente, las variables seleccionadas para la realización del modelo fueron las que obtuvieron un nivel de importancia mayor a 0.01, a continuación, se relacionan:

Tabla 2

Variables Definitivas que se Usaran en el Modelo

Variable	Descripción	Clasificación
Fecha de Declaración de Exportación Definitiva	Fecha en que se declaró la exportación.	Numérica
Aduana	Aduana desde se registró o legalizo la exportación.	Categórica
Municipio	Municipio de origen de la exportación.	Categórica
Peso en kilos netos	Peso en kilos del producto exportado.	Numérica
Número de artículos	Numero de artículos relacionados	Numérica

Descripción de la partida arancelaria	Indica el tipo de cacao en polvo (con o sin adición de azucares)	Categórica
País de Destino	País de destino de la exportación.	Categórica
Departamento de origen	Departamento de origen.	Categórica
Forma de pago	Forma de pago.	Categórica
Valor Flete	Costo asociado al transporte de la mercancía.	Numérica
Valor seguro	Cantidad de dinero que se establece para asegurar la mercancía durante su transporte internacional.	Numérica
Continente Destino	Continente hacia donde se dirige el producto.	Categórica

Modelado (Model)

Una vez se han definido las variables independientes el siguiente paso es encontrar los hiperparametros para la construcción del modelo de red neuronal multicapa, para esto se utiliza la optimización bayesiana, la cual es un enfoque para optimizar funciones costosas o de alta dimensión utilizando el teorema de Bayes para crear un modelo probabilístico de la función objetivo.

Una vez que se tienen los mejores hiperparametros a evaluar para cada configuración de la red neuronal, se inicia con la construcción de los modelos, cabe resaltar que para todos los modelos se utilizaron las variables independientes codificadas y escaladas. Ya que este es un problema de regresión, se establece un porcentaje de error permisible entre las predicciones y los valores reales de la variable objetivo, este valor será del **10%.**

Se emplea la técnica de validación cruzada (*Cross validation*) para evaluar los rendimientos de manera más confiable y robusta, ya que de esta forma se divide el conjunto de

datos en varias partes para entrenar y evaluar el modelo en diferentes subconjuntos, lo que ayuda a obtener una estimación más precisa de su capacidad de generalización en datos no vistos, se utiliza una configuración de 5 *Fold*. Las métricas seleccionadas para evaluar los modelos fueron: *Mean Absolute Error* (Error Absoluto Medio) MAE, *Mean Squared Error* (Error Cuadrático Medio) MSE, coeficiente de determinación R^2, y la tasa de efectividad.

De igual manera, se decide iterar el modelo de redes neuronales con una optimización bayesiana enfocada en maximizar la tasa de efectividad como criterio principal, así mismo utilizando el MAE como función de pérdida en la red neuronal, pues esta métrica es menos sensible a los valores atípicos que el MSE y ha mostrado mayor consistencia en las configuraciones previas. Como resultado, se optimiza la precisión del modelo respecto a la tasa de efectividad, mientras se preserva un error absoluto bajo.

Modelo de Regresión Lineal Múltiple

Inicialmente se desarrolló un modelo de regresión lineal múltiple para tener un punto de partida, se utilizó el 30% del conjunto de datos como datos de prueba para realizar la evaluación del modelo y se obtuvieron los siguientes resultados:

Tabla 3

Métricas de Evaluación para el Modelo de Regresión Lineal Múltiple

Fold	MSE	MAE	R^2	Tasa de efectividad (%)
1	2.7317	1.1483	0.4069	24.88
2	2.6114	1.1551	0.4438	25.69
3	2.9204	1.1987	0.4095	27.09
4	2.6447	1.1795	0.4589	22.75
5	2.5859	1.1464	0.4117	26.16
Promedio	2.6988	1.1656	0.4262	25.32

Nota. Métricas de evaluación del modelo.

Como se observa en la tabla 3 el modelo de regresión tiene una consistencia aceptable en las métricas MSE y MAE en los diferentes Fold, esto indica que tiene una capacidad razonable para generalizar en las diferentes particiones de los datos. Al observar el coeficiente de determinación R^2 está en promedio en 0.42, se infiere que el modelo logra capturar el 42% de la varianza de los datos, además la tasa de efectividad promedio está en 25.31%, es decir solo el 25.31% de las predicciones cumplen con el criterio del error aceptable menor al 10%.

Red Neuronal de 3 Capas con Función de Pérdida MSE

Se empieza probando un modelo de redes neuronales de 3 capas con una función de pérdida MSE para empezar a iterar y encontrar el modelo más adecuado para predecir el precio del cacao en polvo con y sin adición de azúcar. Configuración de búsqueda:

- Número de neuronas en las capas 1, 2 y 3 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).

- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').
- Con dicha configuración se obtuvieron los siguientes resultados

Tabla 4

Modelo de Redes Neuronales con 3 Capas y Función de Pérdida MSE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.6087	0.6903	0.6507	55.33
1	2	1.3405	0.6557	0.7145	56.50
2	3	1.4136	0.6508	0.7141	58.66
3	4	1.6716	0.7738	0.6580	47.36
4	5	1.7389	0.6776	0.6044	58.04
Promedio	3	1.5547	0.6897	0.6683	55.18

Nota. Métricas de evaluación del modelo.

La tabla 4 muestra que el modelo logra un error medio absoluto (MAE) promedio de 0.689, un error cuadrático medio (MSE) promedio de 1.554, y un coeficiente de determinación (R^2) promedio de 0.668, con una tasa de efectividad promedio del 55.18%. Aunque los resultados son consistentes entre los folds, se observan ligeras variaciones en las métricas, siendo el Fold 2 el que presenta la mejor efectividad (58.66%) y el Fold 3 el menor rendimiento con 47.36%.

Red Neuronal de 3 Capas con Función de Pérdida MAE

En la segunda búsqueda de hiperparametros se cambia la función de pérdida de la red neuronal por MAE. Configuración de búsqueda:

- Número de neuronas en las capas 1, 2 y 3 (valores enteros entre 50 y 2200).

- Número de épocas de entrenamiento (valores enteros entre 50 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

La evaluación se realiza mediante validación cruzada configurado para 5 fold, los resultados con los datos de validación de cada interacción fueron los siguientes:

Tabla 5

Modelo de Redes Neuronales con 3 Capas y Función de Pérdida MAE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.7467	0.6755	0.6207	59.50
1	2	1.4872	0.6274	0.6832	65.01
2	3	1.5265	0.6317	0.6913	62.22
3	4	1.6534	0.6798	0.6617	62.53
4	5	1.7325	0.6517	0.6058	60.99
Promedio	3	1.6293	0.6532	0.6526	62.05

Nota. Métricas de evaluación del modelo.

El rendimiento de este modelo por validación cruzada muestra un MAE promedio 0.653, donde su fold más bajo fue en 0.6274 y el más alto en 0.6798. El MSE promedio está en 1.6293 mientras su valor mínimo está en el fold 1 con 1.487221 y el mayor en el fold 1 con 1.746740, con una R^2 promedio de 0.6526, lo que significa que el modelo explica aproximadamente el 65% de la variabilidad de los datos. En cuanto a la tasa de efectividad promedio de 62.06%, se observa un rendimiento estable ya que la diferencia entre la tasa más baja y la más alta entre los fold es de

5.51. En general, los resultados de esto modelo muestran consistencia entre sus diferentes fold, además de esto los resultados promedios son aceptable para el desarrollo del proyecto.

Red Neuronal de 3 Capas, con Función de Pérdida MAE y Optimización Bayesiana Maximizando Tasa de Efectividad

Después, se prueba un modelo de red neuronal que se encuentra con la optimización bayesiana es un modelo de red neuronal de 3 capas con el que se está buscando directamente la mejora de la tasa de efectividad usando función de pérdida MAE. Configuración de búsqueda:

- Número de neuronas en las capas 1, 2 y 3 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

Se realizaron 100 interacciones en la que cada una evaluó una configuración diferente de la red neuronal, al finalizar obtenemos una tabla ordenada de los 10 mejores resultados (MAE más bajos). Al probar esta configuración de parámetros en una red neuronal usando el método de validación cruzada se obtienen los siguientes resultados:

Tabla 6

Modelo de Redes Neuronales con 3 Capas y Función de Pérdida MAE y Optimización Bayesiana.

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.6919	0.6852	0.6326	59.19
1	2	1.3385	0.6247	0.7149	64.08
2	3	1.4926	0.6332	0.6982	63.00
3	4	1.5121	0.6587	0.6906	63.15
4	5	1.7437	0.6537	0.6033	60.99
Promedio	3	1.5558	0.6511	0.6679	62.08

Nota. Métricas de evaluación del modelo.

La tabla 6 muestra que el modelo logra un error medio absoluto (MAE) promedio de 0.651, un error cuadrático medio (MSE) promedio de 1.556, y un coeficiente de determinación (R^2) promedio de 0.668, con una tasa de efectividad promedio del 62.09%. Aunque los resultados son consistentes entre los folds, se observan ligeras variaciones en las métricas, siendo el Fold 2 el que presenta la mejor efectividad (64.09%) y el Fold 1 el menor rendimiento con 59.19%.

Red Neuronal de 5 Capas, con Función de Pérdida MSE

Para la tercera búsqueda se evaluó por medio de optimización bayesiana un modelo de 5 capas, optimizando el MSE de una red neuronal la búsqueda de estos hiperparametros se realizó de la siguiente manera:

- Número de neuronas en las capas 1, 2, 3, 4 y 5 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).

- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

La implementación de esta red con el mejor conjunto de hiperparametros dio como resultado la siguiente tabla:

Tabla 7

Modelo de Redes Neuronales con 5 Capas y Función de Pérdida MSE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.5150	0.7035	0.6710	51.62
1	2	1.2495	0.6652	0.7339	50.92
2	3	1.3788	0.6637	0.7212	55.88
3	4	1.4831	0.6684	0.6965	56.50
4	5	1.8910	0.8485	0.5698	37.92
Promedio	3	1.5035	0.7099	0.6785	50.57

Nota. Métricas de evaluación del modelo.

Los resultados para esta red neuronal no son del todo satisfactorios, en ella encontramos un MAE promedio de 0.709921 y el MSE promedio de 1.5035, un el R^2 promedio de 0.678523 sugiere que el modelo es capaz de explicar más del 67% de la variabilidad de los datos, aunque estos resultados son aceptables, al ver cada fold se observa una variabilidad considerable entre ellos, especialmente en el Fold 5, donde tanto el MAE como el MSE son más altos y el R^2 es más bajo.

En cuanto a la tasa de efectividad promedio de 50.57% es la más baja de los modelos de redes neuronales evaluados en este trabajo. En esta medida también se ve una gran variabilidad

sobre todo en el fold 5, donde la tasa de efectividad bajo a un 37.92%. En general este modelo no se adapta de manera efectiva al conjunto de datos.

Red Neuronal de 5 Capas, con Función de Pérdida MAE

Para la cuarta búsqueda se realizó optimizando el MAE de una red neuronal la búsqueda de estos hiperparametros se realizó de la siguiente manera:

- Número de neuronas en las capas 1, 2, 3, 4 y 5 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

Este modelo de red neuronal es configurado utilizando los mejores hiperparametros obtenidos para la configuración de 5 capas con función MAE, los cuales fueron los siguientes:

Tabla 8

Modelo de Redes Neuronales con 5 Capas y Función de Pérdida MAE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.5495	0.6475	0.6635	61.51
1	2	1.3509	0.5820	0.7123	68.26
2	3	1.3775	0.5971	0.7214	65.32
3	4	1.4838	0.6267	0.6965	64.86
4	5	1.6933	0.6275	0.6148	65.32
Promedio	3	1.4909	0.6162	0.6817	65.05

Nota. Métricas de evaluación del modelo.

Los resultados para este modelo de red son un MAE promedio de 0.616203 y MSE promedio de 1.490996 indican que el modelo tiene una capacidad aceptable para predecir los valores. El R^2 promedio de 0.681732 indica que el modelo explica el 68% de la variabilidad en los datos. La tasa de efectividad promedio de 65.06% es un indicador de que el modelo está realizando predicciones dentro del error aceptable en 2/3 del conjunto de datos.

Al observar los diferentes fold, se observa que existe una variabilidad en la tasa de efectividad, donde la menor está en 61.51% mientras que la mayor en 68.26%, mostrando una diferencia de 6.75 puntos porcentuales.

Red Neuronal de 5 Capas, con Función de Pérdida MAE y Optimización Bayesiana Maximizando Tasa de Efectividad

Se prueba un segundo modelo de redes neuronales que busca la mejora de la tasa de efectividad usando la función de pérdida MAE, pero aumentando el número de capas en el modelo. Configuración de búsqueda:

- Número de neuronas en las capas 1, 2, 3, 4 y 5 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

La implementación de esta red con el mejor conjunto de hiperparametros dio como resultado la siguiente tabla:

Tabla 9

Modelo de Redes Neuronales con 5 Capas y Función de Pérdida MAE y Optimización Bayesiana

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.6199	0.6629	0.6482	59.19
1	2	1.4551	0.5814	0.6901	67.49
2	3	1.5296	0.5713	0.6907	67.80
3	4	1.6266	0.6505	0.6672	64.24
4	5	1.9188	0.6698	0.5635	63.46
Promedio	3	1.6300	0.6272	0.6519	64.43

Nota. Métricas de evaluación del modelo.

Los resultados muestran que el modelo tiene un error medio absoluto (MAE) promedio de 0.627, un error cuadrático medio (MSE) promedio de 1.630, y un coeficiente de determinación (R^2) promedio de 0.652, con una tasa de efectividad promedio del 64.44%. Al ver cada uno de los fold, el numero 3 destaca con la menor MAE (0.571) y la mayor tasa de efectividad (67.80%), mientras que el fold 5 presenta el menor rendimiento.

Red Neuronal de 7 Capas con Función de Pérdida MSE

Para esta búsqueda se realizó la iteración del modelo optimizando el MSE de una red neuronal, la búsqueda de los hiperparametros se realizó de la siguiente manera:

- Número de neuronas en las capas 1, 2, 3, 4, 5, 6 y 7 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).

- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

En la tabla de resultados observamos un MAE promedio de 0.7009 y MSE promedio de 1.532, junto a un R^2 promedio de 0.6726 indica que el modelo explica el 67% de la variabilidad en los datos. La tasa de efectividad promedio de 53.54%.

Tabla 10

Modelo de Redes Neuronales con 7 Capas y Función de Pérdida MSE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	2.0958	0.8671	0.5449	40.34
1	2	1.2334	0.6171	0.7373	59.44
2	3	1.4114	0.6801	0.7146	57.43
3	4	1.3630	0.6739	0.7211	54.79
4	5	1.5587	0.6663	0.6454	55.72
Promedio	3	1.5325	0.7009	0.6726	53.54

Nota. Métricas de evaluación del modelo.

Si se analiza cada uno de los fold, se puede observar que el modelo tiene un alto grado de variabilidad en sus diferentes fold en todas las métricas, la más notable es en el caso de la tasa de efectividad, donde en el fold 1 tiene una tasa del 40.34% y en el fold 2 tiene 59.44%, esto señala que el modelo no es completamente consistente en su desempeño.

Red Neuronal de 7 Capas, con Función de Pérdida MAE

En la sexta y última búsqueda se realizó optimizando el MAE de una red neuronal la búsqueda de estos hiperparametros se realizó de la siguiente manera:

- Número de neuronas en las capas 1, 2, 3, 4, 5, 6 y 7 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

Este modelo presento un MAE promedio de 0.6333, mostrando una buena estabilidad en los folds. El MSE promedio es de 1.6406, con fluctuaciones menores, lo que sugiere una pequeña dispersión de errores. El coeficiente R^2 promedio de 0.6499 indica que el modelo explica el 65% de la variabilidad en los datos, manteniendo estabilidad en los diferentes folds, aunque en el fold 5 bajo a 0.5805. La tasa de efectividad promedio es 65.98%, teniendo un mínimo en el fold 1 de 63.06% y alcanzando un máximo de 68.27% en el fold 2.

Tabla 11

Modelo de Redes Neuronales con 7 Capas y Función de Pérdida MAE

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.7360	0.6507	0.6230	63.06
1	2	1.4403	0.6014	0.6932	68.26
2	3	1.5781	0.6139	0.6809	67.02
3	4	1.6043	0.6517	0.6718	65.17
4	5	1.8443	0.6485	0.5804	66.40
Promedio	3	1.6406	0.6332	0.6499	65.98

Nota. Métricas de evaluación del modelo.

Red Neuronal de 7 Capas, con Función de Pérdida MAE y Optimización Bayesiana Maximizando Tasa de Efectividad

Finalmente, se ejecuta un modelo de redes neuronales con 7 capas buscando directamente la mejora de la tasa de efectividad usando una función de pérdida MAE. Configuración de búsqueda:

- Número de neuronas en las capas 1, 2, 3, 4, 5, 6 y 7 (valores enteros entre 200 y 1500).
- Número de épocas de entrenamiento (valores enteros entre 80 y 150).
- Tamaño del lote (valores enteros entre 32 y 120).
- Optimizador (algoritmo de optimización para entrenar la red, se elige entre varias opciones como 'adam', 'sgd', 'rmsprop', 'adamax', 'nadam'.
- Función de activación de las capas ocultas (se elige entre 'relu' o 'tanh').

La implementación de esta red con el mejor conjunto de hiperparametros dio como resultado la siguiente tabla:

Tabla 12

Modelo de Redes Neuronales con 7 Capas y Función de Pérdida MAE y Optimización Bayesiana

	Fold	MSE	MAE	R^2	Tasa de efectividad (%)
0	1	1.9277	0.6723	0.5814	62.90
1	2	1.4609	0.6347	0.6888	64.39
2	3	1.5359	0.6165	0.6894	65.01
3	4	1.5899	0.6485	0.6747	64.08
4	5	1.8105	0.6457	0.5881	65.63

Promedio	3	1.6650	0.6435	0.6445	64.40

Nota. Métricas de evaluación del modelo.

En esta configuración de red neuronal la dio como resultados error medio absoluto (MAE) promedio de 0.644, un error cuadrático medio (MSE) promedio de 1.665, y un coeficiente de determinación (R^2) promedio de 0.645, con una tasa de efectividad promedio del 64.41%. El Fold 3 obtiene el menor MAE (0.6165) y el mayor R^2 (0.6894), mientras que el Fold 1 presenta el mayor MSE (1.9277) y el menor R^2 (0.5814).

Evaluación (Assess)

A continuación, se presenta una tabla con las métricas promedio obtenidas en el proceso de validación, las cuales reflejan el comportamiento de los diferentes modelos desarrollados en términos de precisión y efectividad.

Tabla 13

Resultados de los Diferentes Modelos

Configuración del modelo	MAE	MSE	R^2	Tasa de efectividad %
Red neuronal de 3 capas, con función de pérdida MSE	0.689712	1.554701	0.668378	55.184108
Red neuronal de 3 capas, con función de pérdida MAE	0.653269	1.629310	0.652606	62.055881
Red neuronal de 3 capas, con función de pérdida MAE, optimizada para mayor tasa de efectividad.	0.651166	1.555807	0.667961	62.086936
Red neuronal de 5 capas, con función de pérdida MSE	0.709921	1.503514	0.678523	50.572253

Red neuronal de 5 capas, con función de pérdida MAE	0.616203	1.490996	0.681732	65.058355
Red neuronal de 5 capas, con función de pérdida MAE, optimizada para mayor tasa de efectividad.	0.627227	1.630065	0.651971	64.439877
Red neuronal de 7 capas, con función de pérdida MSE	0.700961	1.532527	0.672696	53.547882
Red neuronal de 7 capas, con función de pérdida MAE	0.633295	1.640624	0.649909	65.986669
Red neuronal de 7 capas, con función de pérdida MAE, optimizada para mayor tasa de efectividad.	0.643599	1.665029	0.644536	64.407769

Nota. Métricas de evaluación del modelo.

Tomando en cuenta la tabla resumen anterior y los análisis de cada configuración realizados en la fase de modificación se llega a la conclusión que la red neuronal de 5 capas con función de pérdida MAE es la mejor configuración para realizar predicciones en el conjunto de datos. Esta configuración obtiene el menor MAE y MSE, el mayor R^2 y la segunda mejor tasa de efectividad con el 65.06%. Esto indica que esta configuración ofrece la mayor precisión en la predicción, la mejor capacidad explicativa y la efectividad más alta, haciéndola ideal para los objetivos del proyecto.

Cabe mencionar que la red neuronal de 7 capas con función de pérdida MAE es la configuración con mejor tasa de efectividad con un 65.99%, sin embargo, sus métricas MAE, MSE y R^2 son ligeramente inferiores a las de la red de 5 capas con función de pérdida MAE, lo que indica una menor precisión y capacidad explicativa.

Es por esto por lo que la configuración de 5 capas con función de pérdida MAE es la seleccionada como mejor opción, ya que proporciona un equilibrio óptimo entre todas las métricas clave, garantizando no solo una alta efectividad, sino también un desempeño predictivo consistente y confiable en el conjunto de datos analizado.

El modelo seleccionado cuenta con los siguientes hiperparámetros:

- Capa de entrada recibe 12 variables, después de la codificación de las variables categóricas el número total de variables que recibe esta capa es de 107.
- Número de neuronas en la capa oculta 1: 679 con función de activación relu.
- Número de neuronas en la capa oculta 2: 441 con función de activación relu.
- Número de neuronas en la capa oculta 3: 1452 con función de activación relu.
- Número de neuronas en la capa oculta 4: 451con función de activación relu.
- Número de neuronas en la capa oculta 5: 1344 con función de activación relu.
- Capa de salida con función de activación lineal.
- Optimizador de la red “adamax”.
- Función de pérdida de la red “MAE”.
- Numero de épocas: 137.
- Numero de batch size: 54

A continuación, un resumen de las capas:

- Capa de entrada: recibe 107 variables, salida de capa 679, tiene 73.332 parámetros (incluyen tanto los pesos como los sesgos de la capa)
- Capa oculta 1: recibe 679, salida de capa 441, tiene 299.880 parámetros (incluyen tanto los pesos como los sesgos de la capa)

- Capa oculta 1: recibe 441, salida de capa 1452, tiene 641.784 parámetros (incluyen tanto los pesos como los sesgos de la capa)
- Capa oculta 1: recibe 1452, salida de capa 451, tiene 655.303 parámetros (incluyen tanto los pesos como los sesgos de la capa)
- Capa oculta 1: recibe 451, salida de capa 1344, tiene 607.488 parámetros (incluyen tanto los pesos como los sesgos de la capa)
- Capa oculta 1: recibe 1344, salida de capa 1, tiene 1.345 parámetros (incluyen tanto los pesos como los sesgos de la capa)

Una vez entrenada la red se obtienen las siguientes gráficas tanto de la función de pérdida (MAE) como del MSE durante el entrenamiento, en azul con los datos de entrenamiento y en naranja con los datos de validación. En ellas se puede observar como a medida que avanzan las épocas el error disminuye, es decir que el modelo está aprendiendo.

Figura 19

Pérdida Durante el Entrenamiento

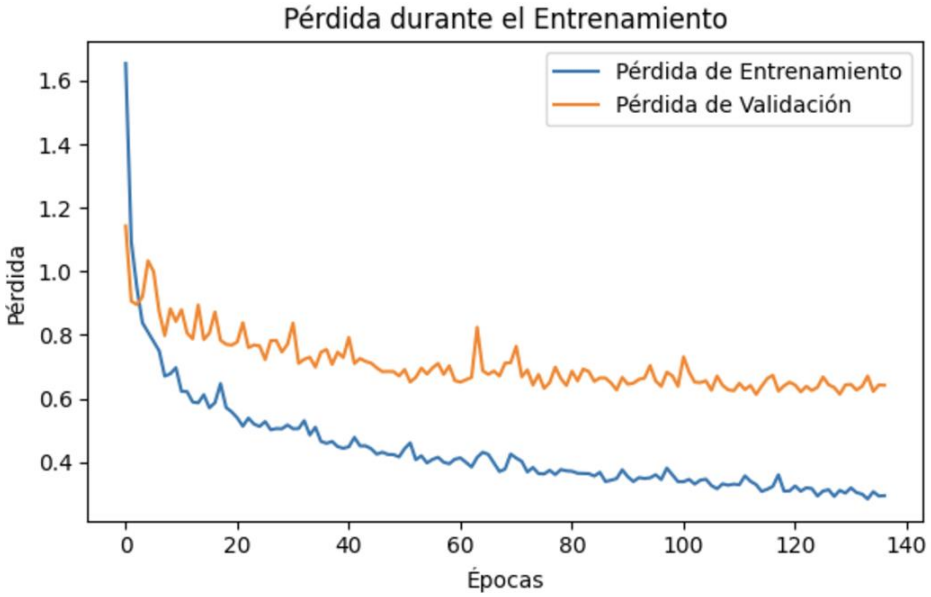
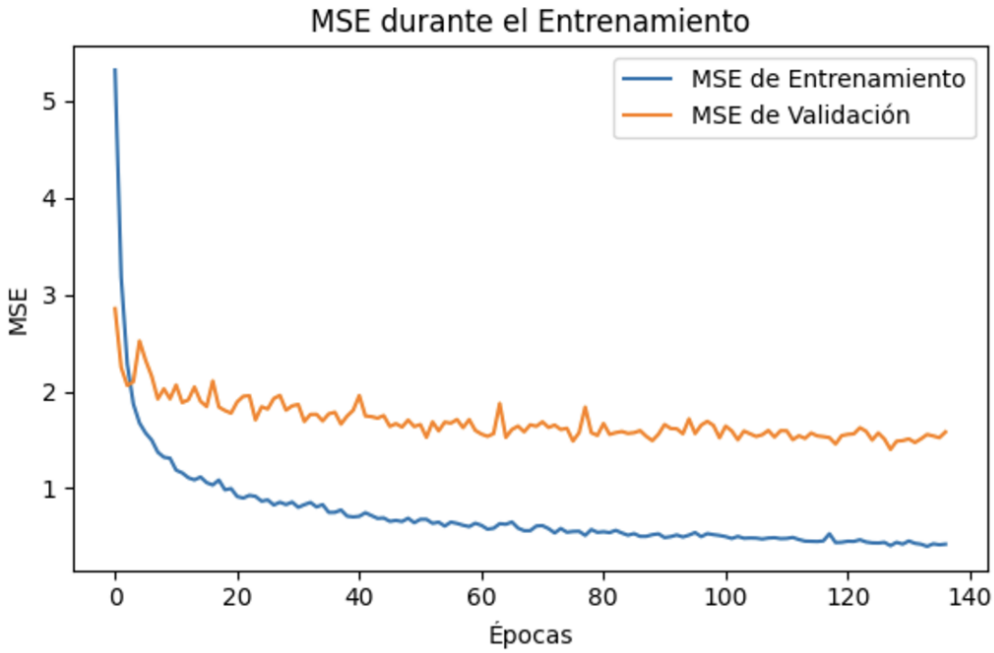


Figura 20

MSE Durante el Entrenamiento



Al evaluar el rendimiento del modelo, se establece un umbral de error aceptable del 10% al igual a como se realizó en la fase de Modelado (Model) y usando el conjunto de datos de prueba, el modelo predice de forma correcta el 61.86% de las observaciones dentro de un rango de error aceptable del 10%. El Error Cuadrático Medio (MSE) es de 1.66, lo que refleja el promedio de los cuadrados de los errores entre las predicciones y los valores reales, indicando que, en promedio, los errores del modelo no son demasiado grandes. El Error Absoluto Medio (MAE), con un valor de 0.65, muestra que el modelo tiene un error promedio de aproximadamente 0.65 unidades entre las predicciones y los valores reales, lo que sugiere una precisión razonable en las predicciones.

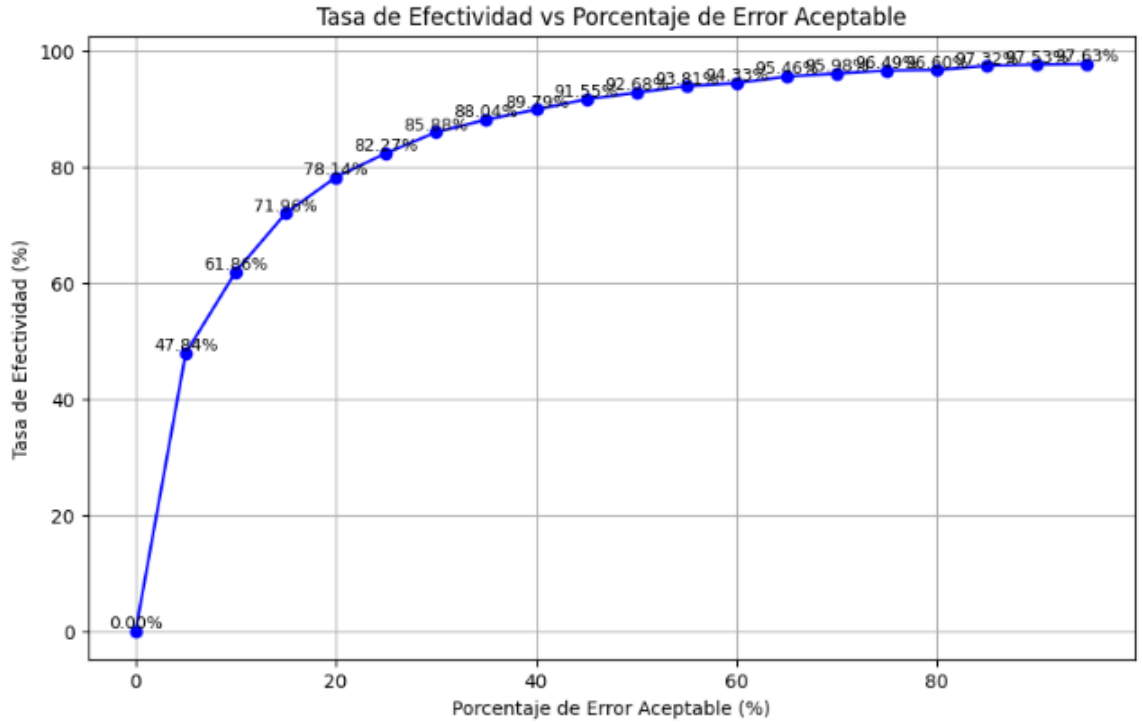
Los resultados comparativos entre los valores reales y las predicciones señalan una variabilidad en los porcentajes de error. Por ejemplo, en algunas predicciones como el registro

428 el valor real es de 6.28 mientras que la predicción es de 7.71, teniendo así un error considerablemente alto de –22.81%; Pero por otro lado en los registros 2332, 3100 y 1718 la predicción es casi exacta, con errores entre 0.22% y 0.28%.

Usando solo los datos de prueba, se realizaron comparaciones para calcular la tasa de efectividad del modelo con diferentes porcentajes de error aceptable. En la figura 23 se puede apreciar como el modelo predice correctamente el 47.84% de los datos de prueba con un error aceptable del 5%, a medida que este error aceptable se hace mayor también lo hace su tasa de efectividad.

Figura 21

Tasa de Efectividad con Diferentes Errores Aceptables



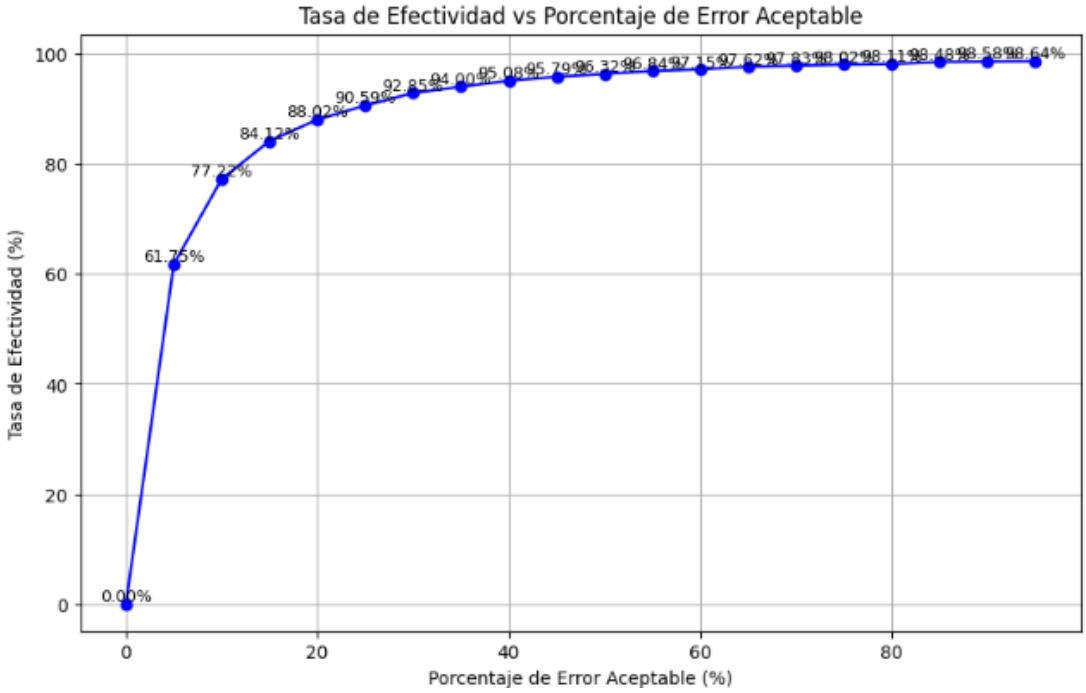
Al ordenar la tabla de resultados de las predicciones usando los datos de prueba de forma ordenada descendiente, mostrando los 25 valores de porcentaje de error absoluto más altos, se

obtiene que 15 de estos 25 errores altos corresponden a registros donde el valor real es inferior a 1.5 dólares, es decir donde el valor real es bajo, indicando que una de las limitantes o dificultades del modelo es la predicción cuando el valor de la variable objetivo está cerca de su límite inferior.

Para finalizar con la evaluación del modelo, se realizan las predicciones con la totalidad del conjunto de datos (entrenamiento y prueba), en este caso los resultados que se obtuvieron señalan una tasa de efectividad de 77.22% con un error permisible del 10%. De igual manera se graficó la tasa de efectividad con diferentes errores aceptables, en ella se puede ver como el modelo predice correctamente el 61.75% de los datos de prueba con un error aceptable del 5%, a medida que este error aceptable se hace mayor también lo hace su tasa de efectividad.

Figura 22

Tasa de Efectividad con Diferentes Errores Aceptables con Todos los Datos



Conclusiones

Entre los hallazgos del proyecto se tiene que Colombia como país tiene un gran potencial de exportación del cacao en polvo con y sin adición de azúcar porque se pudo evidenciar un crecimiento en las exportaciones de dichos productos, incluso cuando las exportaciones de cacao en grano disminuyeron en 2022.

Esto debido a que Colombia ha sido uno de los países que ha empezado a trabajar en poder diversificar sus productos y además darle un valor agregado al producto, sumado a los cambios en las tendencias de consumo, productos como el cacao en polvo se valoran en gran medida y ayuda abrir las puertas de la industria internacional del cacao a los productores nacionales.

Fedecacao debe jugar un papel fundamental en el desarrollo del sector cacaotero colombiano mejorando su presencia en los lugares donde aún es muy deficiente. Ya que, al brindar atención integral a los productores, promoviendo prácticas sostenibles como el rejuvenecimiento de las plantas y el control de plagas, y fomentando la adopción de tecnología, contribuye a aumentar la productividad y posicionar al país como un líder en el mercado internacional del cacao.

El análisis también revelo una marcada concentración geográfica de empresas exportadoras en la ciudad de Bogotá. A pesar de que las principales regiones productoras de cacao se encuentran en otras zonas del país, son las empresas bogotanas las que dominan el comercio exterior de este producto. Esta situación evidencia un alto grado de intermediación en la cadena de valor del cacao, con las regiones productoras teniendo una participación relativamente baja en los beneficios de la exportación.

Al momento de desarrollar un modelo de redes neuronales para entender el comportamiento del precio del cacao a partir de las exportaciones se puede identificar que la

variable objetivo “Precio Unitario FOB (USD) Peso Neto”, presenta valores atípicamente altos los cuales son producto de altas demandas y eventos poco comunes en los mercados, de igual forma se identificaron valores bajos, que si bien no llegan a ser considerados atípicos si se destacan por su valor por debajo de 1.5 dólares, los cuales dificultan la interpretación directa de las tendencias.

Se utilizo la metodología SEMMA para crear de manera estructurada el modelo y ayudo a observar correlaciones débiles con el precio unitario, lo que sugiere no solo la no linealidad del conjunto de datos con la variable objetivo sino también la influencia de factores externos no representados directamente en los datos de las exportaciones, como políticas de comercio internacional, fluctuaciones de la demanda global y condiciones climáticas adversas que afectan los cultivos de cacao. Otros factores no relacionados en la data y ligados directamente a la producción como la disponibilidad de mano de obra, costos de insumos agrícolas también pueden contribuir a la variabilidad observada en los precios.

Al momento de crear el modelo se utilizó la técnica de optimización bayesiana y posterior validación cruzada para identificar la mejor configuración de red y de hiperparametros, encontrando que la red neuronal de 5 capas ocultas con función de pérdida MAE es la mejor configuración para capturar las relaciones no lineales entre las variables, minimizar el error absoluto medio y maximizar la tasa de efectividad, mostrando consistencia en las predicciones y estabilidad durante la validación cruzada.

La capacidad predictiva del modelo final mediante la validación cruzada fue aceptable, con un MAE promedio de 0.616 y un coeficiente de determinación (R^2) de 0.682. Esto demuestra que el modelo es capaz de explicar el 68% de la variabilidad en los datos, siendo adecuado para predecir tendencias generales del precio del cacao. Adicionalmente, el modelo mostró una tasa de efectividad del 65.06% con un error permisible del 10%, destacando que sus mayores porcentajes

de errores se encuentran en casos donde los precios reales estuvieron por debajo de 1.5 dólares, mientras que sus errores más bajos están en los casos donde el precio real estuvo en un rango de 1.6 a 6.1 dólares.

Esto evidencia una limitación en la predicción de valores extremos, especialmente en los valores más bajos, posiblemente debido a la menor representatividad de estos casos en el conjunto de datos o a la sensibilidad del modelo a distribuciones no balanceadas, esta limitación podría ser trabajada con la integración de datos externos (factores climáticos, condiciones de mercado o producción) o mediante técnicas adicionales como entrenar submodelos especializados para diferentes rangos de precios (bajos, intermedios y altos) y combinar sus predicciones mediante un ensemble de tipo *Stacking ensemble* o *Weighted ensemble*.

Los resultados obtenidos en este estudio proporcionan una base sólida para futuras investigaciones en el área de la predicción de precios de los insumos básicos. Aunque el modelo de redes neuronales propuesto ha mostrado un rendimiento prometedor, es necesario realizar estudios adicionales para explorar otras técnicas de aprendizaje automático y evaluar su aplicabilidad en diferentes contextos. Asimismo, se recomienda un análisis más detallado de los factores que influyen en el rendimiento del modelo, con el objetivo de identificar nuevas oportunidades de mejora.

Recomendaciones

Dado la naturaleza compleja y dinámica de los mercados de *commodities* como el cacao, es fundamental explorar una variedad de arquitecturas de redes neuronales para obtener modelos predictivos robustos y precisos. En este sentido, se recomienda investigar arquitecturas recurrentes como LSTM o GRU para capturar las dependencias temporales en los datos históricos de precios. Asimismo, se sugiere explorar modelos de atención para que la red pueda focalizarse en los factores más relevantes al momento de realizar una predicción, como los índices de producción, las condiciones climáticas o las fluctuaciones en la demanda de productos derivados del cacao.

Otra recomendación importante sería ampliar los conjuntos de datos utilizados para entrenar los modelos de predicción de precios del cacao es fundamental para mejorar su generalización y robustez. Incorporar datos históricos más extensos, así como variables adicionales relacionadas con factores económicos, climáticos y geopolíticos, permitiría capturar patrones más complejos y reducir el riesgo de sobreajuste. Además, la inclusión de datos de alta frecuencia, como los precios intradías, podría revelar dinámicas del mercado que no son evidentes en datos de menor frecuencia.

Finalmente, para adaptar los modelos pre-entrenados a la tarea específica de predecir los precios del cacao, podría ser necesario realizar un ajuste fino (fine-tuning) en las capas superiores de la red neuronal. Esto permitirá que el modelo aprenda a capturar las características específicas del mercado del cacao que no están presentes en los datos de entrenamiento originales. Además, la incorporación de capas adicionales especializadas para el análisis de series temporales podría mejorar aún más el rendimiento del modelo.

Referencias

Abbott, P. C., Benjamin, T. J., Burniske, G. R., Croft, M. M., Fenton, M. C., Fernando, R. C., & Wilcox, M. D., Jr. (2019). *Análisis de la cadena productiva del cacao en Colombia*.

Alonso, F. (2021, abril 8). *Redes Neuronales y Deep Learning. Capítulo 2: La neurona*. Future Space S.A. https://www.futurespace.es/redes-neuronales-y-deep-learning-capitulo-2-la-neurona/

Amaya A, S. (2020). *Análisis de las exportaciones de cacao desde Colombia hacia la Unión Europea. 2013 – 2018*. Universitaria Agustiniana.

Arango, F. O. (2017). *Pronóstico de precios de petróleo: una comparación entre modelos GARCH y redes neuronales diferenciales*. Investigación Económica, 76(300), 105–126.

Cacao en polvo, sin adición de azúcar ni otro edulcorante en Colombia. (s/f). Observatorio de Complejidad Económica. https://oec.world/es/profile/bilateral-product/cocoa-powder/reporter/col?marketConcentrationViewSelector=latestTrendsViewOption1&shareMarket=valueMarket

Carlos Véliz. (2020). *Aprendizaje automático. Introducción al aprendizaje profundo.* El Fondo Editorial de la Pontificia Universidad Católica del Perú.

Casas Roma, J. Nin Guerrero, J. & Julbe López, F. (2019). *Big data: análisis de datos en entornos masivos.* Editorial UOC.

Cátedra Santalucía de Analytics for Education [@catedraAfE]. (s/f). *Introducción a la Optimización Bayesiana 3: ¿Qué es la Optimización Bayesiana* https: //www.youtube.com/watch?v=ZI-2jGt4FxU&list=PL5_Uyo65b_AYmFf2u3iTwhLz52TlKX4HM&index=3

Cobaleda Lasso, L. G. (2022). *Análisis Económico de las Exportaciones de Cacao en Colombia durante el periodo 2010-2020.*

Delgado, Y. (2021). *Desarrollo de un modelo predictivo de precio de mora de castilla en Bogotá implementando técnicas de aprendizaje automático.* Tesis de grado Universidad Jorge Tadeo Lozano.

Esteve, J. M. (2016). *Estudio del valor nutricional y funcional de cacao en polvo con diferentes grados de alcalinización.* Tesis de grado Máster en ciencias e ingeniería de alimentos, universidad politécnica de Valencia. https://riunet.upv.es/bitstream/handle/10251/65834/-DUR%C3%81%20-%20Estudio%20del%20valor%20nutricional%20y%20funcional%20de%20cacao%20en%20polvo%20con%20diferentes%20grados%20de%20alcal....pdf?sequence=1

Experts, F. (2024, julio 12). *Cacao: ¿cómo es su cultivo y producción?.* Foodcom S.A. https://foodcom.pl/es/cacao-como-es-su-cultivo-y-produccion/

Fedecacao. (2023, Febrero 7). *Producción cacaotera presentó una reducción del 10% en 2022 por lluvias.* Sitefedecacao. https://www.fedecacao.com.co/post/producci%C3%B3n-cacaotera-present%C3%B3-una-reducci%C3%B3n-del-10-en-2022-por-lluvias

Federación Nacional de Cacaoteros. (2021). *Cacao ¿Qué está pasando con los precios?*. Fedecacao. https://www.fedecacao.com.co/post/cacao-qu%C3%A9-est%C3%A1-pasando-con-los-precios#:~:text=El%20cacao%2C%20as%C3%AD%20como%20el,y%20el%20comportamiento%20del%20d%C3%B3lar.

Foodcom Experts. (2024, Agosto 1*). Mercado del cacao en 2024 - dinámica y análisis del mercado.* Foodcom S.A. https://foodcom.pl/es/panorama-del-mercado-del-cacao-2024-informe-global/

Gil Serna, J. G. (2016). *Estimación de un pronóstico de exportaciones de café suave colombiano: Redes Neuronales Artificiales y ARDL (ene-dic 2012).*

Kane, F. (2017). *Hands-On Data Science and Python Machine Learning.* Packt Publishing.

Marcano, A & Quintanilla, J & Cortina, M. (2010). *Feature Selection Using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network* .https://www.researchgate.net/publication/224207758_Feature_selection_using_Sequential_Forward_Selection_and_classification_applying_Artificial_Metaplasticity_Neural_Network

Martínez, E. (2024, 7 marzo). *Qué son las redes neuronales y sus funciones.* ATRIA Innovation. ATRIA Innovation. https://atriainnovation.com/blog/que-son-las-redes-neuronales-y-sus-funciones/

Martinez, E., & Pedraza L. (2012). *Implementación web de redes neuronales artificiales aplicadas a la predicción de series de tiempo.* Tesis de grado Universidad de la costa. https://repositorio.cuc.edu.co/bitstream/handle/11323/1253/IMPLEMENTACION%20WEB%20DE%20REDES%20NEURONALES%20ARTIFICIALES%20APLICADAS%20A%20LA%20PREDICCION%20DE%20SERIES%20DE%20TIEMPO.pdf?sequence=1&isAllowed=y

Montes, S. 2023. *Conozca a los nuevos 'cacaos' del país y su impacto en el exterior.* FORBES Colombia (s.f.).

Montoya-Restrepo, I. A., Montoya-Restrepo, L. A., & Lowy-Ceron, P. D. (2015). *Oportunidades para la actividad cacaotera en el municipio de Tumaco, Nariño, Colombia.* Revista Entramado, 48-59.

Mora, R., & Fontalvo, Y. (2010). *Análisis del sector cacao y la volatilidad de los precios; aplicando operaciones de cobertura con derivados.* Diplomado Universidad del Magdalena.

Pemice, S.A. (2024*). El problema de la reducción dimensional. Análisis de Componentes Principales (PCA).* Revista Mutis, 14(1), 1-21. https://doi.org/10.21789/22561498.2057

Preciado, K. & Martinez, P. & Casas, O. (2021). *Transformación del cacao orgánico en polvo.* Tesis de grado Especialización en gerencia de proyectos, universidad piloto de Colombia. 37-41. https://repository.unipiloto.edu.co/bitstream/handle/20.500.12277/11406/Transformaci%C3%B3n%20Cacao%20Org%C3%A1nico.pdf?sequence=3

Procesamiento del cacao. (s/f). Buhlergroup.com. https://www.buhlergroup.com/global/es/industries/Cocoa-Chocolate/cocoa.html

Ríos Insua, D. & Gómez-Ullate Oteiza, D. (2019). *Big data: conceptos, tecnologías y aplicaciones: (ed.).* Editorial CSIC Consejo Superior de Investigaciones Científicas. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/122031?page=42

Ríos, F., Ruiz, A., Lecaro, J., & C., R. (2017). *Estrategias país para la oferta de cacaos especiales -Políticas e iniciativas privadas exitosas en el Perú, Ecuador, Colombia. Colombia más competitiva.*

Rodríguez Montequín, M. T., Álvarez Cabal, J. V., Mesa Fernández, J. M., & González Valdés, A. (s/f). *Metodologías para la realización de proyectos de data mining.* 257–265.

Sosa, L., Zamora, L. (2022). Estructura de redes neuronales (MLP) y su aplicación como aproximador universal. Monografías de proyecto de grado universidad distrital Francisco José de Caldas. https://repository.udistrital.edu.co/bitstream/handle/11349/30489/SosaJerezLexlyVanessa2022.pdf?sequence=1&isAllowed=y

Ugalde Binda, N., & Balbastre-Benavent, F. (2022). *Investigación cuantitativa e investigación cualitativa: buscando las ventajas de las diferentes metodologías de investigación*.

Revista De Ciencias Económicas, 31(2), 179–187. https://doi.org/10.15517/rce.v31i2.12730

Vargas, S; García, M; Claros, L. (2023) *El Cacao y sus transformados: Caso Asprobelén y Asoproagro.* AgriLAC Resiliente. https://cgspace.cgiar.org/server/api/core/bitstreams/911a70be-7f0a-4e55-bb17-63a45f185def/content

Vidal, A.G. (2015). *Selección de variables: una revisión de métodos existentes.* Universidad de la Coruña. (pp. 15-22). http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1263.pdf

Villada, F. (2016). *Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro.* Información tecnológica, 143-150.

Zhou, Z. (2021). *Machine learning.* Springer Nature.

Apéndices

Apéndice A

Lista Total de Variables Originales

Característica	Clasificación
fila	int64
Año	int64
Mes	int64
Dia	int64
Año de la Declaración Definitiva	int64
Mes de la declaración definitiva	int64
Día de la Declaración Definitiva	int64
Capítulo Del Arancel	object
Tipo de declaración	object
Fecha de Declaración de Exportación Definitiva	int64
Número de la declaración definitiva	int64
Fecha de Declaración de Exportación Anterior	float64
Número de declaración de exportación anterior	object
Modalidad de importación	object
Fecha De Declaración De Importación Anterior	float64
Número De Declaración De Importación Anterior	object
Tipo De Datos	object
Exportación en Tránsito	object
Aduana	object
Aduana De Embarque	object
Oficina MinComercio	object
Código Agente aduanero	int64

Agente aduanero(s)	object
Usuario	object
Código De Usuario	float64
NIT del exportador	int64
Razón social actual Exportador	object
Municipio	object
Dirección agente aduanero	object
Clase de Exportación	object
Razón social del importador	object
Dirección del Importador	object
Código Partida	int64
Descripción de la partida arancelaria	object
Descripción de la Mercancía	object
Unidad comercial	object
Cantidad(es)	float64
Peso en kilos netos	float64
Peso en kilos brutos	float64
Número de artículos	int64
País de Destino	object
Departamento Origen	object
Departamento De Procedencia	object
Lugar de salida	object
Fecha de Embarque	int64
Número de Autorización de Embarque	object
Código de embarque	object
Vía de transporte	object
Nacionalidad del medio de transporte	object
Régimen Exportación	object
Modalidad de exportación	object
Certificado de Origen	object
Sistemas Especiales	object

Moneda de negociación	object
Forma de pago	object
Valor FOB (USD)	float64
Valor FOB (COP)	float64
Valor Agregado Nacional (VAN)	float64
Valor Flete	float64
Valor seguro	float64
Valor otros	float64
Precio Unitario FOB (COP) Peso Neto	float64
Precio Unitario FOB (COP) Peso Bruto	float64
Precio Unitario FOB (USD) Peso Neto	float64
Precio Unitario FOB (USD) Peso Bruto	float64
Precio Unitario FOB (USD) Cantidad	float64
Precio Unitario FOB (COP) Cantidad	float64
Continente Destino	object
Dólar del día	float64
Salario mínimo COP	int64

Apéndice B

Enlace al Video de Presentación

Video Presentación - Proyecto de Grado II. https://unadvirtualedu-my.sharepoint.com/:v:/g/personal/mkmunoza_unadvirtual_edu_co/EcD51ccoRx5BiGe0BTMt8pAB6yfHMkUbvt_Q2iJMeda5pQ?e=1QNczP