Prototipo de modelo Comercial del cacao y sus derivados para la identificación de oportunidades de negocio para cacaocultores santandereanos por medio del uso de machine learning

Lucas Esteban Quintana Rondón

Asesor

Jhoan Sebastián Báez

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Dedicatoria

A mis padres, hermano y amigos, por todo el apoyo y motivación que me han dado en mi formación como profesional y en la elaboración de mi proyecto de grado.

Agradecimientos

Se le desea agradecer al docente Jhoan Sebastián Báez por todas sus enseñanzas, consejos y conocimientos que me transfirió durante toda la formación tanto profesional como personal, como en la elaboración de este proyecto.

Resumen

En este estudio, empleamos un registro histórico de las exportaciones de cacao desde 2009 hasta 2023 para entrenar una red neuronal, con el fin de determinar qué tan fiable es una exportación de cacao y productos derivados. Durante el entrenamiento se determinó que la mejor opción es el modelo de redes neuronales de perceptrón multicapa (MLP) de clasificación (MLPClassifier) configurado con la función de activación Identify, el solver Adam, alpha = 0.0001, tamaño del lote: *'auto'*, una estructura de tres capas con 106 neuronas en la capa de entrada, 68 neuronas en la capa intermedia y una neurona en la capa de salida. Lo que resultó en buenos desempeños en métricas de evaluación como precisión, recall, f1-score, error cuadrático medio (MSE) y fue corroborado con validación cruzada. Este modelo se realiza con la finalidad de determinar la viabilidad de las exportaciones de cacao y sus productos derivados, principalmente para los cacaocultores santandereanos.

Palabras clave: Cacao, Perceptrón multicapa, Machine learning, Redes neuronales, Exportación

Abstract

In this study, we employed a historical record of cocoa exports from 2009 to 2023 to train a neural network, aiming to determine the reliability of cocoa and cocoa-derived product exports. During the training process, it was determined that the best option is the Multilayer Perceptron (MLP) classification neural network model (MLPClassifier) configured with the Identify activation function, Adam solver, alpha = 0.0001, batch_size: 'auto', and a three-layer structure with 106 neurons in the input layer, 68 neurons in the hidden layer, and one neuron in the output layer. This resulted in good performance metrics such as precision, recall, F1-score, mean squared error (MSE), and was corroborated through cross-validation. This model is designed to determine the viability of cocoa exports and their derived products, mainly for cocoa farmers in Santander.

Keywords: Cocoa, Multilayer Perceptron, Machine Learning, Neural Networks, Exportation

Tabla de contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Glosario

Machine learning: Es un subcampo de la inteligencia artificial en la cual a partir de la información suministrada a un sistema (algoritmo, modelo, etc) este sea capaz de identificar patrones y poder realizar predicciones. (Raschka & Mirjalili, 2017)

Ciencia de datos: La ciencia de datos representa la optimización de procesos y recursos. La ciencia de datos produce conocimientos sobre datos: conclusiones o predicciones procesables basadas en datos que puede utilizar para comprender y mejorar su negocio. (Pierson, 2017)

Data set: Es simplemente un conjunto de datos, es el contenido de una tabla dentro de una base de datos que tiene diferentes columnas, en donde están almacenados los registros en cada una de sus filas. (Solis, 2023)

MLP: El MLP consta de varias capas de neuronas conectadas entre sí, donde las salidas de una capa se convierten en entradas para la siguiente. La primera capa es la de entrada, la última es la de salida, y las intermedias se denominan capas ocultas. (Gamco., 2021)

LSTM: Es una red neuronal recurrente utilizada en el aprendizaje profundo para predecir secuencias de datos. Está diseñada para evitar la pérdida de información crucial que ocurre en las RNN tradicionales durante la retro propagación del error a través de múltiples capas. (Gamco, 2022)

PCA: (Principal Component Analysis o análisis de componente principales), Es una técnica para reducir la dimensionalidad del data set. Aumenta la interpretabilidad, pero, al mismo tiempo, minimiza la pérdida de información. (simplilearn, 2023)

Chi-cuadrado: Test de ajuste de distribuciones que evalúa si los datos de una muestra se ajustan a una distribución teórica determinada. Es una forma de verificar si la distribución teórica que consideramos como válida realmente describe los datos que tenemos. (Saldaña, 2011)

ANOVA: Es un método estadístico que se utiliza para verificar las medias de dos o más grupos que son significativamente diferentes entre sí. (gajawada, 2019)

Estandarización: La estandarización implica transformar las variables para que tengan una media de cero y una desviación estándar de uno. La estandarización es útil cuando las variables tienen diferentes escalas y se quiere que todas tengan un impacto similar en los modelos analíticos, como regresión o clasificación. (Pierson, 2017)

Categorización: La categorización implica agrupar valores de una variable en categorías discretas o clases. Esto se hace cuando los valores de una variable son continuos o numéricos, pero se desea tratarlos como categorías o factores. (Pierson, 2017)

Red neuronal: Una red neuronal es un enfoque de la inteligencia artificial que capacita a las computadoras para procesar datos de manera similar al cerebro humano. (Munar, 2023)

Serie temporal: Una serie temporal es un conjunto de datos medidos en intervalos de tiempo regulares y ordenados de forma cronológica. (Munar, 2023)

Planta del cacao: El cacao es un árbol nativo de la cuenca del río Amazonas en América del Sur. Se encuentra en áreas elevadas de Ecuador, Brasil, Perú y Colombia, así como en

las riberas de los principales afluentes del río Marañón y del Amazonas. (Fuentes & Garcia Jerez, 2021)

LEGIXCOMEX

Plataforma que ofrece herramientas estadísticas, información sobre aranceles, normas, documentos y todos los recursos necesarios para la gestión y análisis del comercio exterior. (legiscomex, 2023)

ICCO

La Organización Internacional del Cacao (ICCO) es una entidad intergubernamental creada en 1973 para implementar el primer Convenio Internacional del Cacao. Este convenio fue negociado en Ginebra durante una Conferencia Internacional del Cacao organizada por las Naciones Unidas. (ICCO, 2023)

Introducción

La industria del cacao desempeña un papel vital en la economía de Santander, Colombia, ofreciendo un potencial significativo para el crecimiento y desarrollo. Como parte de los esfuerzos para mejorar la competitividad de los productores locales de cacao y promover la prosperidad económica en la región, es imperativo identificar y aprovechar oportunidades de exportación lucrativas para el cacao y sus productos derivados. En este contexto, la utilización de tecnologías avanzadas como las redes neuronales presenta un camino prometedor para la predicción del mercado y la identificación de oportunidades comerciales.

Este proyecto tiene como objetivo explorar la aplicación de redes neuronales, específicamente el Perceptrón Multicapa (MLP), en el análisis de datos históricos sobre exportaciones de cacao y productos derivados de Colombia, abarcando el período de 2009 a 2023, tomado de LegisComex. Al aprovechar el poder de esta arquitectura de redes neuronales, buscamos aprovechar los patrones, tendencias y conocimientos valiosos generados por el análisis de datos previos que muy seguramente pueden brindar información a los productores de cacao que puedan informar a los productores de cacao, particularmente para el sector cacaotero en Santander, sobre posibles oportunidades de exportación y ayudar en la planificación estratégica.

Justamente hablando del análisis de datos, este estudio requirió técnicas de procesamiento de datos de tipo estadísticos, como lo son el análisis de varianza (ANOVA) y el análisis de χ-cuadrado, para seleccionar las características más significativas dentro del conjunto de datos. Al identificar las variables clave que impactan significativamente en la dinámica de exportación de cacao, buscamos evitar el sobreajuste de los datos.

A través de un análisis exhaustivo de datos de exportación históricos junto con técnicas avanzadas de aprendizaje automático, este proyecto se esfuerza por capacitar a los productores de cacao en Santander con conocimientos prácticos y visión estratégica. Al aprovechar las redes neuronales y los métodos de procesamiento de datos, buscamos facilitar la toma de decisiones informadas, fomentar la competitividad en el mercado y desbloquear nuevas oportunidades de crecimiento y prosperidad en la industria local del cacao.

Planteamiento del problema

La comercialización del cacao y sus derivados es un aspecto crucial en el comercio colombiano, destacándose como uno de los productos más importantes en las exportaciones del país. Colombia es reconocida por la alta calidad de su cacao, y este sector económico brinda oportunidades significativas para el crecimiento y desarrollo tanto a nivel nacional como en los mercados internacionales. Entre las regiones productoras de cacao en Colombia, el departamento de Santander se destaca como uno de los mayores productores de cacao en grano a nivel nacional, gracias a sus condiciones climáticas y geográficas favorables (Fedecacao, 2023).

Sin embargo, a pesar de la posición prominente de Santander en la producción de cacao en grano, el desarrollo de productos derivados del cacao en la región es insuficiente en comparación con su potencial. La producción de chocolatinas, cacao en polvo, licor de cacao y otros derivados se encuentra en una etapa incipiente. Esto se debe a diversas razones que han obstaculizado el pleno aprovechamiento de estas oportunidades de negocio.

Uno de los principales obstáculos que enfrentan los cacaocultores santandereanos es la falta de conocimiento de las normas y regulaciones internacionales que rigen el comercio de productos cacaoteros. La exportación de cacao y sus derivados implica cumplir con estándares de calidad y requisitos fitosanitarios específicos de cada país de destino. La falta de información y asesoramiento en este aspecto ha llevado a que muchos productos no cumplan con los estándares requeridos, lo que resulta en la pérdida de mercados internacionales y oportunidades de negocio.

Otro desafío crucial es la insuficiencia en la realización de estudios de mercado efectivos. Los cacaocultores santandereanos carecen de información actualizada y precisa sobre las

tendencias del mercado internacional, la demanda de productos derivados del cacao y las preferencias de los consumidores en el extranjero. La falta de datos confiables hace que la toma de decisiones sea poco informada y, en consecuencia, se pierdan oportunidades de negocio.

La calidad de los productos cacaoteros es un tercer obstáculo importante. A menudo, los productos derivados del cacao no cumplen con los estándares de calidad necesarios para la exportación. Esto afecta la reputación de los productos colombianos en el extranjero y limita su acceso a mercados internacionales competitivos.

Además, las pequeñas empresas cacaocultoras en Santander enfrentan dificultades para negociar con compradores internacionales y para establecer relaciones comerciales sólidas. La baja capacidad de negociación y la falta de acceso a información estratégica disminuyen la competitividad de los productores locales en el ámbito internacional.

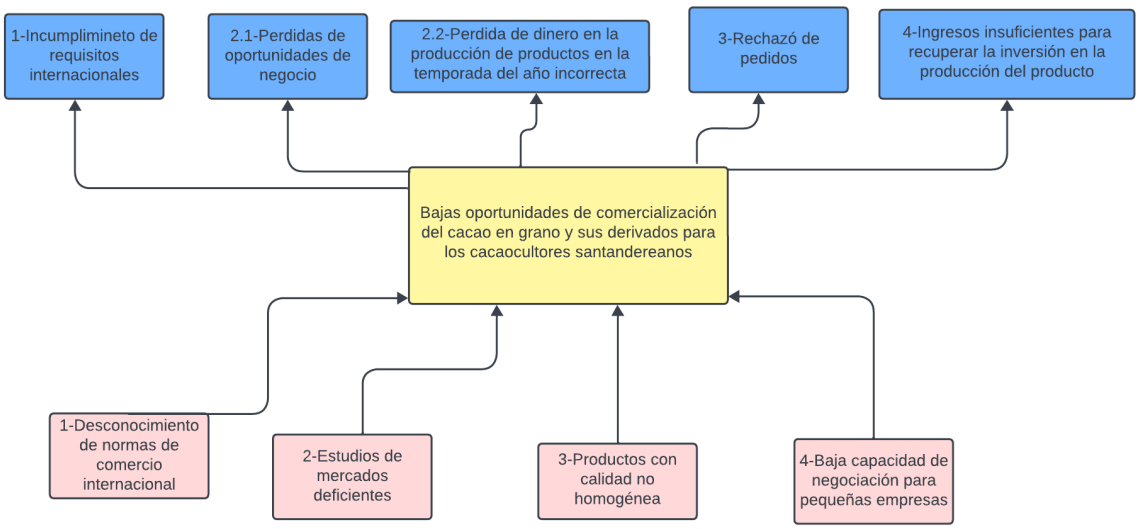
Ante esta problemática, se plantea la necesidad de desarrollar un modelo prototipo de machine learning que analice las exportaciones de cacao colombiano y permita a los cacaocultores santandereanos realizar proyecciones precisas y objetivas. Este modelo sería una herramienta valiosa para comprender las dinámicas del mercado, identificar oportunidades de negocio, y mejorar la toma de decisiones en la producción y comercialización de productos cacaoteros.

La falta de implementación de uso de tecnologías conlleva graves consecuencias. Los cacaocultores santandereanos seguirían perdiendo oportunidades de negocio en el extranjero, manteniendo una dependencia en la exportación de cacao en grano y una producción insuficiente de derivados con mayor valor agregado. Los estudios de mercado deficientes continuarían

resultando en ganancias insuficientes, y la inversión necesaria para producir productos derivados del cacao no se justificaría.

Figura 1

Planteamiento del problema



Fuente. Autoría Propia

En resumen, el desarrollo de un modelo de machine learning es esencial para abordar los desafíos que enfrentan los cacaocultores santandereanos en la comercialización de productos cacaoteros. Este modelo podría marcar la diferencia en la capacidad de la región para aprovechar plenamente las oportunidades de negocio en el mercado internacional, mejorando así la calidad de vida de los productores y contribuyendo al crecimiento económico de Santander y Colombia en su conjunto.

Justificación

En la actualidad el departamento de Santander-Colombia es el mayor productor de cacao en grano en el país pero por desconocimiento de las posibilidades de exportaciones en el país y pocos estudios de mercado, se está perdiendo un gran número de oportunidades comerciales de oferta de productos derivados del cacao como chocolatinas, cacao en polvo, cáscara de cacao, manteca de cacao, entre otros, lo cual causa que el departamento de Santander este perdiendo la oportunidad de ingresos de dinero (Fedecacao, 2023).

El cacao es una de las principales fuentes de ingresos para los agricultores de Santander. Identificar oportunidades comerciales en el extranjero puede aumentar la rentabilidad de los agricultores y contribuir al crecimiento económico de la región. En 2021 Santander represento el 40.6 % de las exportaciones de grano de caco en el país (Fedecacao, 2023).

En un mercado global altamente competitivo, es esencial que los cacaocultores santandereanos ofrezcan productos de alta calidad y sepan identificar oportunidades en el mercado. Un modelo de machine learning puede ayudar a analizar datos relevantes y tendencias comerciales para identificar oportunidades que permitan a los productores mantenerse competitivos.

La finalidad de este proyecto es la construcción de un modelo prototipo de machine learning que le permita al cacaocultor santandereano hacer una proyección objetiva y dinámica para tomar decisiones puedan aprovechar oportunidades de negocio y aumentar sus ganancias

Objetivos

Objetivo general

Crear un Modelo de Machine Learning usando código Python para facilitar la toma de Decisiones de los Cacaocultores Santandereanos en el Mercado Internacional

Objetivos específicos

Recopilar fuentes bibliográficas relevantes relacionadas con el uso del machine learning para la identificación de oportunidades comerciales.

Centralizar la información de las bases de datos de exportaciones de cacao y productos derivados de Colombia, aplicando técnicas de ciencia de datos.

Desarrollar un modelo de Machine Learning utilizando redes neuronales, que incluya la selección de atributos, la ingeniería de características y la elección de algoritmos apropiados.

Entrenar el modelo de Machine Learning con datos históricos y actuales para evaluar su desempeño y ajustarlo según sea necesario, buscando la precisión y la capacidad de brindar información valiosa sobre las oportunidades de mercado para los cacaocultores.

Marco conceptual

Para contextualizar el problema que vamos a desarrollar, es necesario revisar varios detalles cruciales en la exportación de cacao colombiano.

La planta del cacao es un árbol que tiene origen en la región alta de la cuenca del río Amazonas en América del Sur. En las áreas amazónicas de Perú, Ecuador, Colombia y Brasil, así como a lo largo de las orillas de varias redes fluviales importantes que desembocan en los ríos Marañón y Amazonas, se encuentran diversas especies de cacao. Entre ellas, destaca la Theobroma y sus subespecies (Fuentes & Garcia Jerez, 2021).

Con esta planta se pueden crear varios productos derivados de esta como el grano del cacao que es la semilla sana y limpia del fruto del árbol del cacao, con transformaciones bioquímicas al interior, secada, sin mucilago y sin restos de cáscara, está la manteca de cacao que es un producto semisólido, de aspecto graso a temperatura ambiente, de color blanco o ligeramente amarillento, obtenido mediante el procesamiento de los granos de cacao, ya sea por extracción mecánica o por el uso de solventes, luego está el chocolate es un alimento que se obtiene al mezclar azúcar con dos productos derivados del procesamiento de las semillas de cacao, el licor de cacao es el producto resultante de la molienda de las semillas de cacao, ya sean fermentadas o no, tostadas y descascarilladas. Durante el proceso de molienda, se obtiene una masa líquida (Fuentes & Garcia Jerez, 2021).

Como cualquier otro producto en el mercado este tiene sus normas de exportación las cuales se dividen en atributos físicos, atributos sensoriales y atributos químicos (Fuentes & Garcia Jerez, 2021).(Ver Tabla 1)

Tabla 1

Normas exportación del cacao

Atributos físicos	**Atributos sensoriales**	**Atributos químicos**
clasificación de los genotipos: Forastero -Calidad corriente Criollo – calidad alta Trinitario - calidad media	Los niveles más altos de precursores aromáticos en los licores de cacao y en los chocolates provenientes de la variedad criollo se dan como consecuencia de altas concentraciones de aminoácidos, péptidos y azúcares reductores presentes en el grano de cacao, después de terminar la fermentación, el secado y el tostado.	Las metilxantinas. Considerados alcaloides entre los que se destacan la teobromina y la cafeína por los efectos estimulantes sobre el sistema nervioso central. se calcula que la teobromina y la cafeína en grano sin grasa es del 4% y 0,2% del peso seco. Los polifenoles. Son compuestos de gran interés por preservar la salud de la planta de cacao, pero también se han incluido en las sustancias antioxidantes benéficas para la salud humana.

factores que contribuyen en la obtención de un determinado grado de calidad: - El sistema agroforestal en el que se cultiva el cacao. -Las costumbres agroecológicas -La genética del cacao que se manifiesta en la gran cantidad de variedades -Los genotipos, cada uno con sus características propias.	El mercado mundial clasifica el cacao comercializable en dos grandes categorías: -Cacao fino de aroma. -Cacao básico, corriente u ordinario.	Los aldehídos de los azúcares. Se forman durante la fermentación y el secado en el interior del grano y Se encargan de desarrollar el sabor. Se han identificado tres aldehídos importantes en el chocolate debido a su aroma: 2-metilpropanal, 2-metilbutanal y 3-metilbutana.
Buenas prácticas agrícolas -Beneficio del cacao, que contempla a su vez las fases de fermentación y secado. - El almacenamiento, tostado, y sigue el procesado de los		Los ésteres. Son compuestos presentes en granos no fermentados y tostados, pero se pueden formar también durante la fermentación, como el caso del acetato de etilo,

nibs, conchado, refinado y atemperado. -El empacado del producto final o chocolatina, chocolate, bombonería y cosméticos.	que es un producto de esterificación a partir de etanol y ácido acético.

Nota. Adaptado de “Evaluación integral de la calidad sensorial del cacao”, fuente L.Quintana & A.Garcia,2021.

Al conocer ya los conceptos del cacao y sus productos derivados, vamos a tratar los conceptos básicos necesarios durante el desarrollo de este proyecto.

La ciencia de datos representa la optimización de procesos y recursos. La ciencia de datos produce conocimientos sobre datos: conclusiones o predicciones procesables basadas en datos que puede utilizar para comprender y mejorar su negocio. sus inversiones, su salud e incluso su estilo de vida y vida social (Pierson,2017). Dentro de la ciencia de datos existen varias herramientas que ayudan en el ámbito de hacer predicciones para realizar análisis y toma de decisiones en los cuales tenemos a las redes neuronales y el machine learning. El machine learning es una rama de la inteligencia artificial en la cual a partir de la información suministrada a un sistema (algoritmo,

modelo, etc) este sea capaz de identificar patrones y poder realizar predicciones (Raschka & Mirjalili, 2017).

Los tipos de aprendizaje dentro del machine learning los cuales son:

- El aprendizaje supervisado, en este tipo de aprendizaje el modelo aprende a partir de datos de entrenamiento etiquetados es decir se generan predicciones y resultados en base a ejemplos históricos (Raschka & Mirjalili, 2017).
- El aprendizaje no supervisado, en este tipo de aprendizaje el modelo recibe datos sin etiquetar y este debe entender y clasificar esa información por sí mismo (Raschka & Mirjalili, 2017).
- Aprendizaje Reforzado, el modelo puede aprender a tomar decisiones por medio de la interacción con su entorno. El modelo aprende a través de la experiencia y la retroalimentación que recibe (Raschka & Mirjalili, 2017).

El interés particular gira entorno a las redes neuronales, Una red neuronal es un modelo computacional inspirado en el funcionamiento del cerebro humano. Está diseñada para reconocer patrones y aprender de los datos. Las redes neuronales están compuestas por capas de nodos, llamados neuronas, que se conectan entre sí (Munar, 2023).

Hay varios tipos de redes neuronales, pero se explicarán las redes que fueron utilizadas en el desarrollo de este estudio las cuales fueron MLP y LSTM.

La red neuronal MLP (Multilayer Perceptron) , está compuesto por múltiples capas de neuronas interconectadas, en las que las salidas de las neuronas de una capa se convierten en entradas para la siguiente capa. La primera capa se llama capa de entrada,

la última capa se llama capa de salida y las capas intermedias se llaman capas ocultas (Gamco., 2021).

Este modelo de red neuronal cuenta con las funciones de activación que permite a la red aprender y representar relaciones complejas en los datos. A continuación, se describen varias funciones de activación en MLP:

- La función identify de activación lineal simplemente devuelve la entrada tal cual es, sin aplicarle ninguna transformación (Gamco., 2021).
- La función ReLU es una de las funciones de activación más populares debido a su simplicidad y efectividad. Activa solo los valores positivos, dejando los valores negativos en cero (Gamco., 2021).
- La función Tanh , mapea los valores de entrada a un rango entre -1 y 1, lo que puede llevar a un mejor centrado y un aprendizaje más eficiente (Gamco., 2021).
- La función logistic o sigmoid , mapea cualquier valor real a un rango entre 0 y 1, siendo útil en la capa de salida para problemas de clasificación binaria. (Gamco., 2021).

También el modelo MLP cuenta con varios solver los cuales son algoritmos que ajustan los pesos y sesgos para minimizar la función de pérdida. Hay varios tipos de estos los cuales son:

- SGD (Stochastic Gradient Descent), actualiza los pesos usando el gradiente de la función de pérdida evaluado en minibatches de datos (Gamco., 2021).

- Adam (Adaptive Moment Estimation), Ajusta los learning rates de manera adaptativa para cada parámetro (Gamco., 2021) .
- LBFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno), utiliza una aproximación de la matriz Hessiana para encontrar los mínimos de la función de pérdida (Gamco., 2021) .

El modelo LSTM (Long Short-Term Memory) es una variante de las redes neuronales recurrentes (RNN) que se emplea en el aprendizaje profundo para manejar y prever secuencias de datos. La LSTM fue creada para solucionar el problema de la desaparición del gradiente en las RNN convencionales, un problema que surge cuando se retropropaga el error a través de varias capas, lo que lleva a la pérdida de información crucial en el proceso (Gamco, 2022).

Este tipo de modelo cuenta con los epochs que son iteraciones completas sobre el conjunto de datos de entrenamiento, esenciales para ajustar los pesos de la red neuronal durante el proceso de entrenamiento. Un número adecuado de epochs es crucial para lograr un buen equilibrio entre subajuste y sobreajuste, garantizando así que el modelo generalice bien a nuevos datos (Gamco, 2022).

Dentro de la ciencia de datos, existen herramientas estadísticas que permiten tomar las características más significativas de un modelo para evitar el sobreajuste. En los modelos de redes neuronales, hay una gran variedad de estas herramientas, entre ellas se encuentran el análisis univariado, chi-cuadrado, ANOVA y la validación cruzada.

- El análisis univariado implica el análisis de una sola variable. Su objetivo principal es describir la distribución de datos y resumir sus características utilizando estadísticas descriptivas (CEVALLOS TORRES, VALENCIA MARTINEZ, & BARROS MORALES, 2017).
- El chi-cuadrado es una prueba estadística empleada para determinar si hay una asociación significativa entre dos variables categóricas (Saldaña, 2011).
- El ANOVA es una técnica estadística empleada para contrastar las medias de tres o más grupos, con el propósito de determinar si al menos uno de estos grupos presenta diferencias significativas respecto a los demás (gajawada, 2019).
- La validación cruzada es una técnica de evaluación de modelos que se utiliza para evaluar el rendimiento de un modelo de aprendizaje automático de manera más robusta. Implica dividir los datos en subconjuntos y entrenar y evaluar el modelo en diferentes combinaciones de estos subconjuntos (Ochoa, 2019).

Marco teórico

Estado del arte

Las oportunidades de la comercialización del cacao y sus productos derivados siempre han sido muy numerosas siendo el cacao de Latinoamérica uno de los más solicitados se ha realizados estudios de la comercialización del cacao en Perú y en Colombia siendo un ejemplo el departamento de Nariño y Antioquia. En estos estudios se pudo identificar que el cacao colombiano tiene grandes oportunidades comerciales siendo sus mayores compradores América del norte y Europa (Rojas, Melo Mosquera, Agredo Madroñero, & Moncayo Rosero, 2021; Hernández, 2022).

Otros estudios de exportaciones de otros países de América Latina que también exportan cacao como el caso de Ecuador en el cual se hizo un estudio económico del 2014 al 2019 y se llegó a la conclusión de que el cacao es un producto bastante solicitado en Europa, Asia y América del norte, pero es de mencionar que la mayoría de las exportaciones hechas son de grano de cacao y las de sus productos derivados están presentes en menor medida (Quezada, Carvajal Romero, Barrezueta Unda, & Cordova, 2021).

Por otra parte, otro trabajo de estudio de mercado realizado en Colombia para venta de chocolate de alta calidad por medio del ecommerce atreves de este estudio se encontró que este tipo de chocolates se compraban en grandes cantidades en estados unidos (Medina, 2016) .

En la aplicación de machine learning para identificar oportunidades comerciales se encuentra un antecedente en el Perú para poder identificar la demanda del banano por medio del uso de algoritmos de aprendizaje supervisado del tipo Redes Neuronales, en el cual se mostró que la aplicación del machine learning es bastante efectiva al momento de hacer predicciones de tipo comercial (Almeyda, 2022).

Entre otros estudios realizados sobre la aplicación del machine learning en la toma de decisiones de un negocio se encontraron antecedentes en cual por medio del uso de técnicas de aprendizaje supervisado y no supervisado se usaron datos de ventas de una empresa entre el año 2017 y 2018 de ventas, se aplicó para predecir las ventas del 2019 con el cual se tomaron decisiones sobre la cantidad de stock de producto que se necesitaba para evitar pérdidas (Ignacio, 2021).

En otra aplicación del machine learning, se encuentra una investigación realizada para obtener proyecciones de venta de camarón exportado por ecuador por medio del uso del entrenamiento supervisado usando de entrenamiento los datos de ventas de 5 años para predecir el comportamiento del rango de años del 2021 al 2025, el cual mostro un aumento exponencial en las cantidades de producto importado (Cordero-Torres, 2022).

Con los estudios anteriormente mencionados podemos llegar a concluir que la elaboración de este proyecto es factible debido al buen historial que tiene Colombia en temas de exportación del cacao y sus productos derivados y con la aplicación de tecnologías de machine learning podrían aumentar de manera significativa.

Metodología

Investigación mixta

La investigación mixta es una metodología que combina tanto investigación cuantitativa como cualitativa para abordar un problema de investigación. Este enfoque se utiliza cuando se requiere una mejor comprensión del problema de investigación, ya que ninguno de estos métodos por sí solo puede proporcionar toda la información necesaria. (López, 2010)

Fases

Primera fase: Inicio del proyecto

Búsqueda de revisión bibliográfica de proyectos de estudio de mercado en el que se use el machine Learning para identificar oportunidades comerciales.

Actividades:

Revisión bibliográfica

Segunda fase: Recopilación y procesamiento de datos

Para esta etapa, el enfoque principal será la recopilación de datos relacionados con las exportaciones de productos cacaoteros en Colombia. Estos datos representan una pieza clave para comprender las dinámicas comerciales en la industria del cacao y determinar oportunidades estratégicas para los cacaocultores santandereanos.

Los datos que buscaremos incluirán información detallada sobre las exportaciones de cacao y sus productos derivados a nivel nacional y, en particular, a nivel regional en

Santander. Esta desagregación regional es fundamental para comprender cómo se desarrolla el mercado local y cómo los productores en esta región pueden beneficiarse de las tendencias de exportación a nivel nacional.

Una vez que hayamos reunido estos datos, procederemos a realizar un riguroso proceso de limpieza y preprocesamiento. Esto implica la identificación y corrección de posibles errores, la eliminación de valores atípicos y la estandarización de la información. La calidad de los datos es crucial para garantizar la precisión del modelo de machine learning en etapas posteriores de la investigación.

En resumen, en la fase de Recopilación y procesamiento de datos, nos centraremos en la obtención y preparación de información relacionada con las exportaciones de cacao en Colombia, con un enfoque especial en Santander. Este sólido conjunto de datos formará la base sobre la cual construiremos y entrenaremos el modelo de machine learning para identificar oportunidades de negocio para los cacaocultores de la región.

Actividades:

Recolección de información

Proceso de limpieza de la información

Tercera fase: Análisis, selección y diseño de modelo

En la tercera fase llevaremos a cabo un exhaustivo análisis exploratorio de los datos con el propósito de obtener una comprensión más profunda de su distribución, relaciones y posibles correlaciones. Para lograr esto, empleamos diversas técnicas, que incluyen la visualización de datos mediante gráficos y la realización de análisis estadísticos destinados a identificar patrones y tendencias significativas. En particular, consideramos la aplicación de métodos de machine learning tanto supervisados como no supervisados. Para respaldar estas decisiones, nos basamos en técnicas estadísticas como el análisis de correlación, el análisis de componentes principales (PCA) y otras estrategias de selección y extracción de características. Estos enfoques nos permiten reducir la dimensionalidad de los datos y retener únicamente las características más relevantes. La meta final en esta etapa es diseñar un modelo de machine learning que se ajuste de manera óptima a la naturaleza del problema, brindando así una sólida base para identificar oportunidades de negocio en la comercialización de cacao y sus derivados para los cacaocultores santandereanos.

Actividades:

Pruebas y ensayos

Diseño de modelos

Cuarta Fase. Evaluación y validación del modelo

Se llevará a cabo la evaluación del modelo utilizando diversas fuentes de datos, y se emplearán métricas pertinentes al problema en cuestión. En el caso de problemas de regresión, se considerará el error cuadrático medio (MSE), mientras que para problemas de clasificación se evaluarán la precisión y el F1-score. También a su vez se implementarán métricas de evaluación como la matriz de confusión, precisión, exactitud y Recall.

Se usará un porcentaje de entrenamiento 80-20 para los modelos, en el cual 20% de los datos serán de prueba y el otro 80% serán datos de entrenamiento.

La matriz de confusión es un recurso útil para evaluar el rendimiento de un algoritmo de aprendizaje supervisado. En esta matriz, cada columna muestra la cantidad de predicciones para cada clase, mientras que cada fila indica las instancias de la clase real. En términos prácticos, esto nos permite identificar los aciertos y errores del modelo durante el proceso de aprendizaje con los datos. Dentro de la matriz de confusión se utilizan los siguientes términos: (Barrios, 2019).

- Verdaderos Positivos: Son los casos en los que el modelo predice correctamente la clase positiva.
- Verdaderos Negativos: Son los casos en los que el modelo predice correctamente la clase negativa.
- Falsos Positivos: Son los casos en los que el modelo predice incorrectamente la clase positiva cuando en realidad es negativa.

- Falsos Negativos: Son los casos en los que el modelo predice incorrectamente la clase negativa cuando en realidad es positiva.

La precisión Indica el porcentaje de muestras clasificadas correctamente sobre el número total de muestras clasificadas en esa clase. Una baja precisión indica que hay un alto número de falsos positivos. A continuación, se muestra la fórmula de la precisión: (Arias, 2021)

$$Precision = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}$$

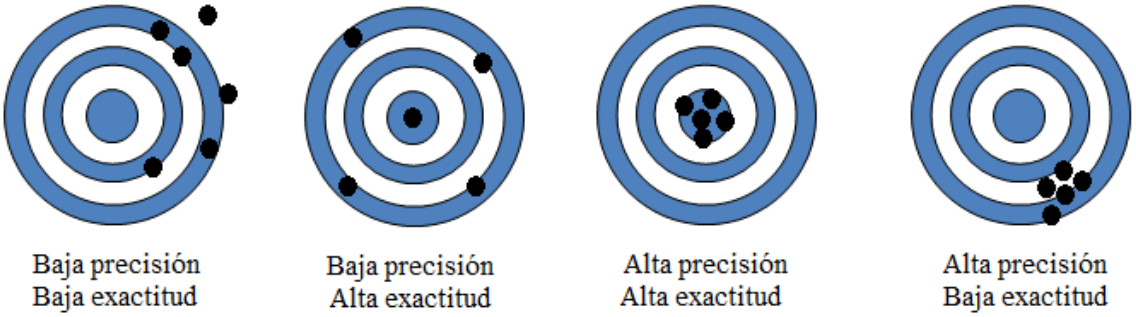
La exactitud el porcentaje de aciertos del modelo, es decir, la cantidad de muestras clasificadas correctamente sobre el total de muestras. A continuación, se muestra la fórmula de la exactitud: (Arias, 2021)

$$Exactitud = \frac{verdaderos\ positivos + verdaderos\ negativos}{verdaderos\ positivos + falsos\ positivos + falsos\ negativos + verdaderos\ negativos}$$

La exactitud y la precisión se complementan entre si como se ve en la siguiente imagen (ver Figura 2):

Figura 2

Precisión y exactitud



Fuente. researchgate ,2012, (https://www.researchgate.net/figure/Figura-2-Precision-y-exactitud_fig2_289674541)

- La baja exactitud y precisión significa que las mediciones son inconsistentes e incorrectas (Arias, 2021) .
- Baja precisión y alta exactitud significa que, en promedio, las mediciones son correctas, pero no son consistentes (Arias, 2021).
- Alta precisión y exactitud significa que las mediciones son consistentes y correctas (Arias, 2021).
- Alta precisión y baja exactitud significa que tus mediciones son consistentes pero incorrectas (Arias, 2021).

El Recall es el porcentaje de muestras de datos que un modelo de aprendizaje automático identifica correctamente como pertenecientes a una clase de interés (la "clase positiva") del total de muestras de esa clase (Arias, 2021).

El error cuadrático medio (MSE) es una medida que evalúa la cantidad de discrepancia entre dos conjuntos de datos. En otras palabras, compara un valor predicho

con un valor observado o conocido. A continuación, se muestra la fórmula de MSE: (Arias, 2021)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$$

F1-score es una métrica muy utilizada en problemas en los que el conjunto de datos a analizar está desbalanceado. Esta métrica combina el precision y el recall, para obtener un valor mucho más objetivo. A continuación, se muestra la fórmula de f1-score: (Arias, 2021)

$$F1 = \frac{2 \times Precision \times recall}{Precision + recall}$$

Una vez que se logre desarrollar un modelo que muestre un desempeño satisfactorio, se procederá a validar su rendimiento utilizando un conjunto de prueba que no haya sido empleado en las etapas previas de entrenamiento y validación. En caso de ser necesario, se realizarán ajustes adicionales con el objetivo de mejorar la precisión y la capacidad de generalización del modelo.

Actividades:

Toma de datos

Pruebas y ensayos

Testeo del modelo

Quinta Fase. Implementación

En la Quinta Fase, llevamos a cabo la implementación del modelo de Machine Learning en un entorno de producción específicamente diseñado para satisfacer las necesidades de los cacaocultores santandereanos. Este proceso implica la integración del modelo en las operaciones diarias de los productores y su infraestructura tecnológica, garantizando su funcionamiento sin problemas.

Una vez que el modelo está en funcionamiento, establecemos un riguroso sistema de monitoreo en tiempo real para evaluar su rendimiento continuamente. Esto implica la recopilación de datos en tiempo real sobre la comercialización de cacao y sus derivados, así como las interacciones con los cacaocultores. El monitoreo constante nos permite detectar posibles desviaciones y evaluar la precisión de las predicciones en un entorno de producción real.

Para mantener la eficacia del modelo, estamos preparados para realizar ajustes y actualizaciones periódicas. Estos ajustes pueden ser necesarios para adaptar el modelo a las cambiantes condiciones del mercado, las preferencias de los consumidores o cualquier otro factor relevante. La prioridad es asegurar que el modelo siga brindando resultados precisos y relevantes a lo largo del tiempo, contribuyendo así a la identificación de oportunidades de negocio óptimas para los cacaocultores santandereanos.

Además, para garantizar la divulgación adecuada de los resultados y el valor generado por el modelo, trabajamos en estrecha colaboración con los cacaocultores y las partes interesadas relevantes. Compartimos información sobre el impacto positivo del modelo en la toma de decisiones, la optimización de la comercialización y la mejora de la

rentabilidad para los cacaocultores de la región. Esto contribuye a su adopción exitosa y al impulso de las oportunidades de negocio en la industria del cacao en Santander.

Actividades:

Perfeccionamiento del modelo

Evaluación del modelo

Despliegue del modelo

Primera fase Inicio del proyecto

Revisión bibliográfica

Basados en la revisión bibliográfica (ver estado del arte) en la que se empleó el uso de machine learning para identificar oportunidades comerciales en la exportación de productos. Los estudios más relevantes fueron los siguientes:

- **Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de Machine Learning**

Datos usados: Se realiza el pronóstico utilizando datos del registro mensual de exportación de banano orgánico (en kilogramos) de Perú, recopilados desde el año 2001 hasta el 2021.

Variables usadas: Peso en kg, VALOR FOB dólar, destino y continente.

Modelos usados: MLP, RNN, LSTM, y GRU

Resultados: Los modelos LSTM, GRU y RNN obtuvieron una mayor precisión en la estimación de la estacionalidad de la serie temporal.

- **Algoritmos de Aprendizaje Supervisado para Proyección de Ventas de Camarón Ecuatoriano con Lenguaje de Programación Python**

Datos usados: Los registros datan a partir de enero del 2011 hasta junio de 2021.

Modelos usados: Regresión múltiple y modelo Ridge

Variables usadas: Los precios del camarón en Estados Unidos se expresan en kilogramos bajo la etiqueta 'EE.UU'. Las exportaciones ecuatorianas se miden en libras y se etiquetan como 'Export'. Las importaciones estadounidenses de camarón indio se cuantifican en millares y se identifican como 'Import'. El precio del petróleo West Texas Intermediate (WTI) se denota en dólares americanos bajo la etiqueta 'Crudo'. Por último, el índice de precios FishPool Index™ (FPI) del salmón se presenta en euros por kilogramo, identificado como 'Salmon'.

Resultados: Los pronósticos de ventas de camarón en este proyecto están condicionados a cumplir con las expectativas de la empresa y reflejar de manera más precisa la gestión de su capital de trabajo, así como para mantener una actitud prudente en las decisiones financieras.

- **Modelo Deep learning para la estimación del potencial exportador de productos no minero-energéticos en Colombia**

Datos usados: Muestra la exportaciones realizadas por Colombia entre 2015 a 2019

Modelo usado: Modelo neuronal Deep Learning por correlación de componentes independientes

Variables usadas: Destino. valor FOB USD, mercado destinos adicionales, acuerdos comerciales, continente, cercanía geográfica y nombre del producto.

Resultados: La construcción de un modelo permite establecer un puente entre la tecnología presente en los softwares utilizados para su desarrollo y la teoría conceptual de inteligencia de mercados. Esto facilita la generación de resultados que pueden servir como fundamentos para tomar decisiones más informadas en el ámbito del comercio internacional.

- **Modelado de estudios de mercado basados en machine learning para empresas dentro del comercio electrónico**

Datos usados: Para la recolección de los datos se estudió el comportamiento del consumidor de servicios médicos de dos áreas de interés específicas a través de los intercambios de texto entre este y la marca del análisis. Al delimitar los competidores y la temporalidad, se obtuvieron los comentarios dando reseña para cada uno de los once hospitales evaluados y se almacenó cada arreglo en un archivo CSV.

Modelo usado: En la aplicación global del análisis sentimental, la propuesta del método ofrece estructuras de procesamiento de lenguaje natural entrenadas con tweets y comentarios de otras redes sociales. Inicialmente, para español las herramientas son limitadas por ello para la fase experimental se aplica los fundamentos de Transformes en el modelo RoBERTuito entrenado con alrededor de cinco mil textos. Se utiliza específicamente pysentimiento, que es la librería variante del análisis sentimental BERT en español.

Variables usadas: Selección de 50 comentarios del total obtenido en redes sociales, los cuales se identificaron dependiendo de la intención presentada por parte del usuario. En el caso del experimento del modelo, se tomaron únicamente los comentarios que son clasificados en la intención de dar reseña con respecto del servicio y atención médica recibida en la visita al establecimiento de estudio.

Resultados: Para un estudio de mercado, la mejor forma de reconocer el tipo de consumidor global y la clasificación dependiendo de la competencia, debe enfocarse en estudiar comentarios de reseñas en las que haya un mensaje que trasmita la intención de retroalimentar acerca del uso de un producto o servicio de una marca especifica.

Con el uso del NLKT y RoBERTuito se automatiza el proceso de leer cada comentario y con la noción de la clasificación inicial, la computadora se encarga de procesarlo para reconocer palabras clave a través de los PoS, estudiar sentimientos dependiendo de un rango positivo, negativo o neutral. También reconoce la emoción más fuerte que trasmite el mensaje a través de puntuaciones.

- **Mapa de oportunidades comerciales de Buenos Aires utilizando modelos de aprendizaje automático**

Datos usados: El conjunto de datos seleccionado, denominado "Rubros", proporciona información del año 2017 sobre los diversos rubros comerciales distribuidos en la geolocalización de la Ciudad Autónoma de Buenos Aires. Después de la selección, se continuó explorando los datos con el objetivo de comprender su composición, lo que llevó a descubrir un total de 2898 muestras, cada una caracterizada por 22 variables distintas.

Modelo usado: El modelo seleccionado para implementar fue un modelo de recomendación utilizando la biblioteca conocida como "LightFM". En esencia, un modelo de recomendación consta de dos elementos principales: los ítems y los usuarios. En el contexto del estudio, los usuarios se representarían mediante los distintos rubros

comerciales presentes en el conjunto de datos, mientras que los ítems serían las diversas formas geométricas distribuidas en la Ciudad Autónoma de Buenos Aires.

Variables usadas: Facturación_prom_actual, Índice crecimiento, Índice estabilidad y Nivel de riesgo.

Resultados: En conclusión, el uso de LightFM demostró resultados destacados en varios indicadores, lo que respalda su recomendación para este modelo de recomendación. Además, el empleo del Análisis de Componentes Principales (PCA) se mostró como una herramienta útil para mejorar los resultados del modelo final. Por otro lado, todos los rubros del conjunto de datos pueden visualizarse en un mapa de la Capital Federal, lo que permite identificar las mejores ubicaciones para cada uno según una escala de colores.

- **Using Econometric Models to Manage the Price Risk of Cocoa Beans: A Case from India**

Datos usados: La información mensual sobre el precio de los granos de cacao se recopila para el periodo comprendido entre abril de 2009 y marzo de 2020 desde la oficina de CAMPCO Limited en Mangalore, y los precios futuros del cacao en el ICE se obtienen del sitio web de investing.com.

Modelo usado: Modelo ARIMA y Modelo VAR

Variables usadas: El riesgo de precio, específicamente para los comerciantes de cacao, fabricantes de chocolate y productores de cacao.

Resultados:

El modelo ARIMA para el precio del Cacao Seco y el precio del Cacao Húmedo, respectivamente. Los correlogramas de los residuos demostraron que el modelo es adecuado para su uso en predicciones.

El modelo VAR determinaron los precios del cacao seco es decir demostraron que las futuras negociaciones del cacao negociados en la plataforma ICE en los Estados Unidos y Londres son los determinantes del precio de los granos de cacao seco en el mercado indio.

Los modelos desarrollados pueden ser utilizados por las entidades comerciales de cacao, los productores, los usuarios industriales y las autoridades gubernamentales para una planificación oportuna y la toma de decisiones gerenciales.

- **Analysis of Multi-Layer Perceptron and Long Short-Term Memory on Predicting Cocoa Futures Price**

Datos usados: El conjunto de datos utilizado está citado de la página Investing.com y abarca desde 2003 hasta 2021 de los precios diarios del cacao.

Modelo usado: MLP y LSTM.

Variables usadas : Date, Open, High, Low, Close y volume

Resultados: Desde la etapa de prueba utilizando el conjunto de datos diarios de los precios futuros del cacao, el rendimiento del modelo de predicción utilizando MLP

para predecir el precio muestra un resultado excelente. El hiperparámetro utilizado para el conjunto de datos diarios consiste en 64 neuronas en la capa de entrada. Para la capa oculta, probamos diferentes números de neuronas, como 16, 32 y 64. La capa de salida consta de 1 neurona, utilizando 500 epochs, y empleamos diferentes tamaños de lote, que son 64 y 128. El optimizador utilizado es el optimizador Adam.

El LSTM también se entrena y prueba para predecir el precio de los futuros del cacao. El LSTM tiene una ventaja sobre el MLP porque este modelo puede memorizar la salida y tiene la probabilidad de usarla como la siguiente entrada, mientras que el MLP no puede memorizarla. Ajustar el tamaño del lote (batch size) en este modelo LSTM es muy importante porque la característica de memoria del LSTM lo hace más lento y requiere mucho tiempo para entrenar. Por lo tanto, cambiar el tamaño del lote a 64 o incluso 128 podría acelerar el tiempo de entrenamiento.

En la construcción del LSTM, utilizamos los mismos hiperparámetros que el modelo MLP. Entrenamos el modelo LSTM con diversas proporciones de datos de prueba y entrenamiento, utilizando 64 neuronas para la capa de entrada. Para la capa oculta, también utilizamos variaciones de 16, 32 y 64, respectivamente. La capa de salida consta de 1 neurona. El número de epochs que utilizamos es de 500, y probamos tamaños de lote de 64 y 128, respectivamente. El optimizador que utilizamos es el optimizador Adam.

Con los descubrimientos encontrados en esta revisión bibliográfica se llega a la conclusión que las redes neuronales se ajustan perfectamente a la problemática en la que

se centra este proyecto y se implementaran los modelos de redes neuronales MLP y LSTM porque son las que mejores resultados tienen en las métricas de evaluación.

Segunda fase Recopilación y procesamiento de datos.

Para el desarrollo de este proyecto se utilizará un data set con la información de las exportaciones del cacao y sus productos derivados desde 2009 a 2023 en Colombia. Estos datos se obtuvieron de la página web LEGISCOMEX. Para ver mas a detalle el proceso de descarga del data set (Ver Apéndice H).

Proceso de limpieza de la información

Para el proceso de limpieza de datos, se hizo enfoque en los siguientes aspectos para evitar problemas de calidad con estos a la hora de hacer el estudio previsto y así determinar cuáles son las columnas útiles del dataset con la información de las exportaciones del cacao y sus productos derivados.

Evitar formatos de datos no procesables

Problema: Es común distribuir información pública en diversos tipos de documentos, siendo el texto el elemento predominante. Estos documentos suelen estar en formato PDF dificultando el procesamiento de estos. (Iniciativa Aporta (datos.gob.es), 2022)

Resultado: En este paso no se encontró ningún problema debido a que los datos que estamos usando se encuentran en formato xlsx lo cual facilitara la lectura de estos.

Utilizar una codificación de caracteres estandarizada

Problema: Es común encontrar conjuntos de datos que incluyen caracteres especiales como acentos, eñes, o signos de puntuación, los cuales pueden ser

interpretados de forma inconsistente por las máquinas. Esto puede dificultar la reutilización de los datos o aumentar la complejidad al tratar de hacerlos comprensibles.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Se revisa a detalle cada columna del dataset y se observa que solo hay uso de caracteres especiales en la columnas Aduana, Aduana De Embarque, Municipio, Agente aduanero(s), Razón social actual Exportador, Dirección agente aduanero, Razón social del importador, Descripción de la partida arancelaria, Descripción de la Mercancía, País de Destino, Departamento Origen, Departamento De Procedencia, Lugar de salida, Vía de transporte, Nacionalidad del medio de transporte, Modalidad de exportación, Moneda de negociación, Continente Destino ,categoría y Dirección del Importador se procede a quitar los caracteres especiales de esas columnas.Principalmente se quitan. , ñ – “ # & de los registros.

Nombrar adecuadamente columnas

Problema: Los nombres de las columnas o variables en un conjunto de datos deben ser claros y fácilmente comprensibles para las personas que los utilizan.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Los nombres de las columnas de la data set son los suficientemente claros y se entiende el tipo de información que estos contienen, a continuación, se muestra una tabla con la descripción de cada columna. (ver Tabla 2)

Tabla 2

Descripción campos data set

Columna	Tipo de datos	características
Año	numérico	Año en el que se realizó la exportación
Mes	numérico	Mes en el que se realizó la exportación
Día	numérico	Día en el que se realizó la exportación
Tipo de declaración	carácter	Tipo de declaración de exportación
Tipo De Datos	carácter	Tipo de datos de la exportación
Aduana	carácter	Lugar donde se realizaron las aduanas para exportar el producto
Aduana De Embarque	carácter	Lugar donde se realizaron las aduanas de embarque para exportar el producto
Agente aduanero(s)	carácter	Entidad encargada de hacer el control de aduanas
NIT del exportador	carácter	Documento de identidad del exportador
Razón social actual Exportador	carácter	La razón social del exportador
Municipio	carácter	Municipio del exportador
Razón social del importador	carácter	Razón social del importador
Dirección del Importador	carácter	Dirección del importador (el lugar donde llegara el cargamento)
Código Partida	carácter	Se compone de una secuencia de 10 números, los cuales se asignan a cualquier artículo que esté destinado a ser importado o exportado. Este código identifica de manera única un producto dentro de un sistema estructurado de descripción y clasificación, fundamentado en el Sistema Armonizado (SA).
Descripción de la partida arancelaria	carácter	Producto que se está exportando
Cantidad(es)	numérico	Peso neto total de la exportación
Peso en kilos netos	numérico	Peso neto total de la exportación
Peso en kilos brutos	numérico	Peso bruto total de la exportación
Número de artículos	carácter	Numero de artículos exportados
País de Destino	carácter	País a donde llegara la exportación
Departamento Origen	carácter	Departamento origen de la exportación
Departamento De Procedencia	carácter	Departamento de origen del producto, cargamento o exportación
Lugar de salida	carácter	Departamento de donde salió la exportación a su destino
Código de embarque	carácter	Tipo de embarque de la exportación
Vía de transporte	carácter	El tipo de transporte usado para importar el producto
Nacionalidad del medio de transporte	carácter	La nacionalidad del transporte es decir de que país proviene el transporte usado
Régimen Exportación	carácter	Régimen de exportación que permite la salida temporal del territorio aduanero nacional de mercancías nacionales o en libre circulación, para ser sometidas a una operación de perfeccionamiento en el exterior, para su posterior reimportación.
Modalidad de exportación	carácter	Son las diferentes formas para la salida de las mercancías de acuerdo con la finalidad que le quiera dar el exportado

Certificado de Origen	carácter	El tipo de documento que certifica o demuestra que los productos cumplen con los requisitos de origen establecidos en un tratado específico.
Sistemas Especiales	carácter	Tipo de sistema especial
Moneda de negociación	carácter	Tipo de moneda con la que hace la negociación
Forma de pago	carácter	Tipo de forma de pago
Valor FOB (USD)	numérico	Valor de la exportación en dólares
Valor FOB (COP)	numérico	Valor de la exportación en pesos colombianos
Valor Agregado Nacional (VAN)	numérico	El valor agregado de las exportaciones es el valor añadido por Colombia a los productos que exporta
Valor Flete	numérico	Valor del flete de la exportación
Valor seguro	numérico	Valor del seguro de la exportación
Valor otros	numérico	Valor de otros gastos para la exportación
Precio Unitario FOB (COP) Peso Neto	numérico	Precio en pesos colombiano del peso neto del producto
Precio Unitario FOB (COP) Peso Bruto	numérico	Precio en pesos colombiano del peso bruto del producto
Precio Unitario FOB (USD) Peso Neto	numérico	Precio en dólares del peso neto del producto
Precio Unitario FOB (USD) Peso Bruto	numérico	Precio en dólares del peso bruto del producto
Precio Unitario FOB (USD) Cantidad	numérico	Precio en dólares de la cantidad
Precio Unitario FOB (COP) Cantidad	numérico	Precio en pesos colombianos de la cantidad
Continente Destino	carácter	Continente a donde se dirige la exportación
categoría	carácter	El tipo de producto que se está exportando

Nota. En esta tabla se muestra los nombres de las columnas, el tipo de datos y la descripción de la información contenida en esa columna.

Publicar datos completos y evitar valores ausentes

Problema: La falta de valores en las tablas es un problema común en muchos conjuntos de datos, que impacta directamente en la calidad de los mismos.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Se procede a buscar en cada columna del dataset la presencia de datos ausentes y solo se encuentran datos faltantes en la columna de Continente Destino, se procede a llenar esos campos vacíos con NA. Adicionalmente se eliminaron algunas columnas a continuación se anexa la tabla con la razón de su eliminación. (Ver Tabla 3)

Tabla 3

Campos eliminadas del data set

Columna	Razón
tipo de datos	No brinda una información relevante para el estudio que se quiere realizar
Aduana De Embarque	La columna de departamento tiene la misma información
NIT del exportador	Se quita porque hay una columna con el nombre de la razón del exportador
municipio	Hay muchos datos nulos en esta columna
Dirección del Importador	Con el nombre del importador y el país del destino es suficiente información
Código de embarque	hay muchos datos nulos en esta columna
Moneda de negociación	Ya hay una columna que dice el valor de la exportación en dólares y otra en pesos colombianos
Dirección del exportador	Con el nombre del exportador y la ciudad de origen es suficiente información

Nota. Columnas que se eliminaron del data set

Evitar la duplicidad de registros

Problema: Esto ocurre cuando un conjunto de datos contiene dos o más registros idénticos. La presencia de datos duplicados aumenta la probabilidad de que los resultados del análisis estén sesgados, disminuyendo así la utilidad de los datos.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: No se encuentra repetida ninguna importación dentro del data set.

Estandarizar valores de datos

Problema: Esta dificultad significativa obstaculiza la correlación entre datos de distintas distribuciones o conjuntos de datos, así como la interoperabilidad y el enlace entre los datos. Incluso complica la comparación de datos tanto dentro como entre organizaciones.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Se revisa la columna de categoría la cual es la que muestra el tipo de producto que se está exportando y se encuentra bien estandarizada según el tipo de producto

Proporcionar una cantidad adecuada de datos para facilitar su análisis

Problema: Es común encontrarse con conjuntos de datos publicados en acceso abierto que están limitados en su alcance o, por el contrario, son excesivamente extensos, llegando incluso a exceder significativamente el propósito original del conjunto de datos. Esto puede hacer que la información sea inutilizable o no aporte un valor claro para los usuarios en su forma actual.(Iniciativa Aporta (datos.gob.es), 2022).

Resultado: Esta problemática no aplica para el dataset utilizado.

Formateo de variables de fecha y hora

Problema: Es recomendable codificar las fechas usando la referencia ISO 8601, que sigue el formato AAAA-MM-DD para las fechas y AAAA-MM-DDThh:mm para las fechas y horas combinadas.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: En el caso del dataset utilizado no aplica debido a que el día {mes y año, se encuentran separados en diferentes columnas.

Formateo de datos numéricos

Problema: En algunas ocasiones, los datos numéricos se presentan con separadores para diferenciar la parte entera de la decimal. Dependiendo de la configuración regional, estos separadores pueden ser un punto o una coma. Además, los separadores de miles pueden ser comas, puntos o incluso espacios en blanco. Estas variaciones pueden ocasionar interpretaciones incorrectas, especialmente al procesar los datos de manera automática.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Se revisa cada columna numérica del dataset y se encuentra que hay datos enteros y decimales en las columnas Cantidad(es), Peso en kilos netos, Peso en kilos brutos,Número de artículos, Valor FOB (USD), Valor FOB (COP), Valor Agregado Nacional (VAN), Valor Flete, Valor seguro, Valor otros, Precio Unitario FOB (COP) Peso Neto, Precio Unitario FOB (COP) Peso Bruto, Precio Unitario FOB (USD) Peso Neto, Precio Unitario FOB (USD) Peso Bruto ,Precio Unitario FOB (USD) Cantidad y Precio Unitario FOB (COP) Cantidad.La corrección de estos valores se procederá a implementar en la fase 3 en Python para automatizar este proceso.

Evitar la mezcla de escalas numéricas

Problema: Una publicación de calidad de un conjunto de datos requiere consistencia en todas las distribuciones que se deriven de él. Las características de cada variable condicionan su análisis y es crucial que sean coherentes para asegurar una interpretación precisa.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: No aplica para el dataset utilizado

Evitar la mezcla de rangos en un mismo conjunto de datos

Problema: La utilización de rangos en un conjunto de datos puede restringir la información disponible para el usuario. Además, en muchas ocasiones, los rangos de datos empleados en diferentes distribuciones son inconsistentes, lo que dificulta la comparación y comprensión adecuada de los datos.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: No aplica para el dataset utilizado.

Incorporar variables con información geográfica

Problema: Las columnas que almacenan información geográfica debe estar estandarizada, que faciliten la interpretación y el análisis de los datos.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: No aplica para el dataset utilizado porque no hay datos de geolocalización geográfica.

Evitar la incorporación de subtotales, totales o agrupamientos

Problema: Es común encontrar filas o columnas de totales o subtotales en tablas de datos, lo que genera agregaciones dentro del conjunto de datos.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: No aplica para el dataset utilizado.

Evitar la fragmentación de datos y de difícil localización

Problema: A menudo, los datos están fragmentados y dispersos en diversas secciones o páginas dentro del portal de un organismo o entidad, e incluso en diferentes sitios web. Esta situación dificulta su descubrimiento y, por consiguiente, su acceso.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Los datos utilizados se encuentran presente en la página web legiscomex y el acceso a ellos no son de difícil acceso y no están fragmentados.

Organizar adecuadamente los datasets disponibles

Problema: Cuando los datos consisten en observaciones numéricas y están asociados con referencias temporales, nos encontramos frente a conjuntos de datos del ámbito estadístico.(Iniciativa Aporta (datos.gob.es), 2022)

Resultado: Para la construcción de este dataset se tuvo que generar individualmente las exportaciones de cada año debido a que legiscomex solo deja sacar máximo los registros de un año, una vez se descargaron todos los años se procedieron a unir en un solo dataset.

Estandarización y categorización de datos

Se estandariza y se categorizan los datos para que estos queden mejor organizados al momento de ser evaluados en el modelo de red neuronal. Para ver más a detalle este proceso (ver Apéndice C).

Tercera fase Análisis, selección y diseño de modelo

Variable objetivo

Antes de hacer la construcción del modelo, se debe tener en claro cuál será la variable objetivo o de salida es decir la predicción del modelo de la red neuronal.

En el modelo estamos buscando poder identificar las oportunidades comerciales del cacao y sus productos derivados en el extranjero, se decidió crear una variable la cual se llama venta fiable. Se procede a realizar un análisis exploratorio de toda la información para identificar que variables se pueden tener en cuenta para determinar si una exportación es fiable o no.

Realizando un análisis exploratorio de los datos, en el que su principal enfoque fue ver la información relacionada al valor en dólares y el peso en kilos netos de las exportaciones de cada uno de los diferentes productos, para poder construir un estándar de evaluación que permita recomendar si una exportación es fiable o no es fiable. (Para ver a detalle porque se escogieron los rangos para la evaluación mirar Apéndice B).

En la ilustración 2 y 3 se muestran la distribución de valor en dólares y peos en kilos netos, a través del uso de deciles, que nos permitieron escoger los rangos de evaluación para crear la variable objetivo.

Figura 3

Deciles valor fob dólar

Categoria	10%	20%	30%	40%	50%	60%	70%	80%	90%
Cacao crudo	87.000000	3160.296000	39622.014000	63742.000000	72978.550000	81075.000000	133222.589000	202895.936000	280771.860000
Cacao en polvo	486.570000	1356.000000	2425.000000	4080.000000	6456.000000	10080.000000	15059.000000	23716.000000	32140.800000
Cacao tostado	393.300000	955.920000	1807.198000	3163.500000	4482.625000	5316.172000	7070.000000	9949.008000	19075.350000
Cascara de cacao	3836.160000	5633.600000	6643.000000	6794.200000	9093.280000	9450.000000	10308.308000	10769.056000	28068.960000
Chocolates	373.195000	905.000000	1574.000000	2530.800000	3842.670000	5829.000000	9195.500000	15866.400000	30272.000000
Manteca de cacao	5563.304000	26780.000000	45000.000000	57272.004000	66983.250000	92000.000000	128840.000000	212000.000000	361200.000000
otras preparaciones	170.500000	1174.880000	3936.900000	9250.920000	16539.400000	39514.512000	61559.961000	74621.424000	87794.692000
pasta de cacao	4678.664000	10331.200000	21164.016000	32760.000000	38703.600000	46867.720000	62207.328000	77154.880000	89991.000000

Fuente. Autoría Propia

Figura 4

Deciles Valor peso kilos netos

Categoria	10%	20%	30%	40%	50%	60%	70%	80%	90%
Cacao crudo	25.000000	496.000000	12496.700000	24978.600000	25000.000000	25127.800000	50000.000000	75128.000000	100119.000000
Cacao en polvo	100.800000	288.000000	625.000000	1104.000000	1800.000000	3000.000000	4838.000000	6000.000000	9750.000000
Cacao tostado	51.200000	100.000000	250.000000	403.640000	600.000000	800.000000	1000.000000	1500.000000	2596.340000
Cascara de cacao	4920.000000	12600.000000	12600.000000	12600.000000	12600.000000	12600.000000	12600.000000	12664.800000	24925.000000
Chocolates	72.000000	190.000000	350.000000	583.200000	957.840000	1517.570000	2419.095000	4095.000000	8640.000000
Manteca de cacao	960.000000	5000.000000	8000.000000	10000.000000	12000.000000	20000.000000	20000.000000	40000.000000	80000.000000
otras preparaciones	42.900000	290.400000	951.735000	2784.000000	4976.000000	9578.400000	14164.000000	16000.000000	18546.000000
pasta de cacao	1000.000000	2479.200000	5000.000000	8000.600000	9988.000000	10017.000000	15208.800000	19998.000000	20020.000000

Fuente. Autoría Propia

A continuación, se muestra como quedó la clasificación de cada tipo de producto:

Tabla 4

Criterios Evaluación variable objetivo

Rango	Chocolates	Cacao en polvo	Cacao crudo	Manteca de cacao	Calificación
Rango de pesos	Menos de 580 kilos	Menos de 1800 kilos	Menos de 25000 kilos	Menos de 8000 kilos	0
Rango valor dólar	Menos de 2500 dólares	Menos de 4000 dólares	Menos de 63000 dólares	Menos de 45000 dólares	0
Rango de pesos	Mayor o igual a 580 kilos	Mayor o igual a 1800 kilos	Más de 25000 kilos	Mayor o igual a 8000 kilos	1
Rango valor dólar	Mayor o igual a 2500 dólares	Mayor o igual a 4000 dólares	Mayor o igual a 63000 dólares	Mayor o igual 45000 dólares	1

Nota. Puntajes de evaluación chocolates, cacao en polvo, cacao crudo y manteca de cacao.

Tabla 5

Criterios evaluación variable objetivo 2

Rango	pasta de cacao	Cacao tostado	otras preparaciones	Cascara de cacao	Calificación
Rango de pesos	Menos de 8000 kilos	Menos de 600 kilos	Menos de 950 kilos	Menos de 12600 kilos	0
Rango valor dólar	Menos de 32000 dólares	Menos de 3100 dólares	Menos de 3900 dólares	Menos de 6600 dólares	0
Rango de pesos	Mayor o igual a 8000 kilos	Mayor o igual a 600 kilos	Mayor o igual a 950 kilos	Mayor o igual a 12600 kilos	1
Rango valor dólar	Mayor o igual a 32000 dólares	Mayor o igual a 3100 dólares	Mayor o igual a 3900 dólares	Mayor o igual a 6600 dólares	1

Nota. Puntajes de evaluación pasta de cacao, cacao tostado, otras preparaciones y cascara de cacao.

Implementación estándares de evaluación en Python

Luego se crea la función para que dependiendo de la categoría y de los rangos estipulados asigne si esas exportaciones fueron fiables o no. (Ver apéndice C).

A continuación, se muestra un fragmento del código:

```
def calcular_calificacion(row):
    categoria = str(row['Categoria'])
    peso_kilos_netos = row['Peso_kilos_netos']
    var_fob_dolar = row['Valor_FOB_USD']
        rangos_calificaciones = {
        'Chocolates': {
            'peso': {
                (0, 579): 0,
                    (580 , float('inf')): 1
            },
            'fob': {
                (0, 2499): 0,

                (2500 , float('inf')): 1
            }
        }
}
if categoria in rangos_calificaciones:
```

```
        calificacion_peso = None
        calificacion_fob = None

        # Verificamos el peso
        for rango, calificacion in
rangos_calificaciones[categoria]['peso'].items():
            if rango[0] <= peso_kilos_netos <= rango[1]:
                calificacion_peso = calificacion
                break

        # Verificamos el valor FOB
        for rango, calificacion in
rangos_calificaciones[categoria]['fob'].items():
            if rango[0] <= var_fob_dolar <= rango[1]:
                calificacion_fob = calificacion
                break

        if calificacion_peso is None or calificacion_fob is None:
            return None

        calificacion_promedio = (calificacion_peso +
calificacion_fob) / 2
        # Determinamos si la venta es confiable o no
```

```
        if calificacion_promedio >=0.5:
            return 1
        #elif calificacion_promedio == 0.5:
            #return 0.5
        else:
            return 0


    else:
        # Si la categoría no está en los rangos predefinidos,
retornamos None
        return 0
df1['venta_fiable'] = df1.apply(calcular_calificacion, axis=1)
```

Análisis componentes principales

Se implementa el PCA para identificar que columnas son más significativas dentro del data set se implementó de la siguiente manera:

Al ver los resultados de la aplicación del PCA en el dataset se llega a la conclusión que no es factible aplicarlo debido a que solo está dando que el número óptimo de componentes es 1 y se necesita más de un componente para que este cumpla correctamente su función de mostrar la importancia de cada columna, se toma la decisión de implementar los métodos chi-cuadrado y ANOVA para poder analizar la relación entre las variables del data set. Para más información del código (ver Apéndice C)

Análisis univariado

Antes de aplicar las técnicas de chi-cuadrado y ANOVA, se lleva a cabo un análisis univariado con el objetivo de simplificar la variedad de categorías en las variables categóricas. Esto se hace para optimizar la eficiencia del proceso de modelado y evitar demoras innecesarias en la presentación de resultados. Al reducir el número de categorías, se facilita la interpretación de los datos y se agiliza el análisis, lo que contribuye a una toma de decisiones más rápida y precisa. Para ver a mas a detalle el proceso del análisis univariado (Ver Apéndice C).

Figura 5

Valores únicos para cada característica categórica reducidos

	index	0
4	Razon_social_importador	161
6	Pais_destino	36
2	Agente_aduanero	14
11	Nacionalidad_medio_transporte	14
3	Razon_social_exportador	9
18	Categoria	8
14	Certificado_origen	8
9	Lugar_salida	6
1	Aduana	6
5	Descripcion_partida_arancelaria	5
7	Departamento_origen	5
8	Departmanento_procedencia	5
17	Continente_destino	5
16	Forma_pago	4
19	trimestre	4
10	Via_transporte	3
20	Cosecha	3
12	Regimen_exportacion	3
15	Sistemas_especiales	2
13	Modalidad_exportacion	2
0	Tipo_de_declaracion	2

Fuente. Autoría Propia

Chi-cuadrado

Se aplica el método de chi-cuadrado para ver el nivel de relación que tiene las columnas categóricas con la variable objetivo-propuesta, estos fueron los resultados, para ver más a detalle el proceso (Ver Apéndice C):

Figura 6

Resultados chi-cuadrado

	Feature	p-value	index	count	unique	top	freq
13	Modalidad_exportacion	0.0	Modalidad_exportacion	57838	2	EXPORTACION DEFINITIVA DE MERCANCIAS DE FABRI...	48108
11	Nacionalidad_medio_transporte	0.0	Nacionalidad_medio_transporte	57838	14	COLOMBIA	13638
2	Agente_aduanero	0.0	Agente_aduanero	57838	14	AGENCIA DE ADUANAS MARIO LONDONO SA NIVEL 1	15071
3	Razon_social_exportador	0.0	Razon_social_exportador	57838	9	COLOMBINA SA	12264
18	Categoria	0.0	Categoria	57838	8	Chocolates	47679
14	Certificado_origen	0.0	Certificado_origen	57838	8	NINGUNO	39831
9	Lugar_salida	0.0	Lugar_salida	57838	6	CARTAGENA	22831
6	Pais_destino	0.0	Pais_destino	57838	36	ECUADOR	10677
1	Aduana	0.0	Aduana	57838	6	CARTAGENA	22489
8	Departmanento_procedencia	0.0	Departmanento_procedencia	57838	5	ANTIOQUIA	18437
7	Departamento_origen	0.0	Departamento_origen	57838	5	ANTIOQUIA	18523
5	Descripcion_partida_arancelaria	0.0	Descripcion_partida_arancelaria	57838	5	Los demas chocolates y demas preparaciones ali...	28538
16	Forma_pago	0.0	Forma_pago	57838	4	CON REINTEGRO	53470
10	Via_transporte	0.0	Via_transporte	57838	3	TRANSPORTE MARITIMO	37087
15	Sistemas_especiales	0.0	Sistemas_especiales	57838	2	NO	36811
17	Continente_destino	0.0	Continente_destino	57838	5	AMERICA	52858
4	Razon_social_importador	0.0	Razon_social_importador	57838	161	OTRO IMPORTADOR	11428

Fuente. Autoría Propia

Las columnas categóricas mostradas en la figura 6 son las que se relacionan más con la variable objetivo.

ANOVA

Se aplica el método ANOVA para saber el nivel de relación de las columnas numéricas con la variable objetivo.

Figura 7

Resultado método anova

	Feature	F-Score	p-value	index	count	mean	std	min	25%
6	Peso_Kilos_brutos	4403.114580	0.000000e+00	Peso_Kilos_brutos	57838.0	0.017541	0.044790	0.0	1.108974e-03
7	Numero_articulos	4215.392450	0.000000e+00	Numero_articulos	57838.0	0.016963	0.027709	0.0	0.000000e+00
5	Peso_kilos_netos	3886.357575	0.000000e+00	Peso_kilos_netos	57838.0	0.016718	0.045284	0.0	9.775622e-04
9	Valor_FOB_COP	3886.357575	0.000000e+00	Valor_FOB_COP	57838.0	0.016718	0.045284	0.0	9.775622e-04
8	Valor_FOB_USD	3056.319379	0.000000e+00	Valor_FOB_USD	57838.0	0.003694	0.011079	0.0	2.647194e-04
11	Valor_flete	2613.305527	0.000000e+00	Valor_flete	57838.0	0.017102	0.050207	0.0	0.000000e+00
0	Año	577.000491	0.000000e+00	Año	57838.0	2016.998565	4.541325	2009.0	2.013000e+03
10	Valor_agregado_nacional	418.740027	0.000000e+00	Valor_agregado_nacional	57838.0	0.000743	0.005857	0.0	0.000000e+00
12	Valor_seguro	168.392275	0.000000e+00	Valor_seguro	57838.0	0.001839	0.022782	0.0	0.000000e+00
4	Cantidades	66.513568	0.000000e+00	Cantidades	57838.0	0.000403	0.016076	0.0	2.880000e-13
13	Valor_otros	48.126762	0.000000e+00	Valor_otros	57838.0	0.000573	0.009508	0.0	0.000000e+00
14	Precio_unitario_FOB_COP_Peso_Neto	39.930331	3.000000e-10	Precio_unitario_FOB_COP_Peso_Neto	57838.0	0.001003	0.004789	0.0	4.465792e-04
15	Precio_unitario_FOB_COP_Peso_Bruto	36.212781	1.800000e-09	Precio_unitario_FOB_COP_Peso_Bruto	57838.0	0.000971	0.004736	0.0	4.388919e-04
16	Precio_unitario_FOB_USD_peso_Neto	23.601287	1.188200e-06	Precio_unitario_FOB_USD_peso_Neto	57838.0	0.001336	0.005040	0.0	6.753800e-04
17	Precio_unitario_FOB_USD_Peso_Bruto	20.109431	7.327600e-06	Precio_unitario_FOB_USD_Peso_Bruto	57838.0	0.001293	0.004969	0.0	6.662772e-04
2	Dia	20.079407	7.443500e-06	Dia	57838.0	15.598741	8.791103	1.0	8.000000e+00

Fuente. Autoría Propia

Las columnas numéricas mostradas en la figura 7 son las que se relacionan más con la variable objetivo.

Construcción del modelo

Una vez establecida la variable objetivo e identificadas las columnas del conjunto de datos que tienen una mayor relación con esta, se puede avanzar en la construcción de los modelos de redes neuronales. En este caso, optamos por implementar los modelos MLP y LSTM. Esta elección se basa en los hallazgos de la revisión bibliográfica realizada durante la fase inicial del proyecto. Estudios similares han demostrado que estos dos tipos de redes neuronales han ofrecido los mejores resultados en términos de predicción y desempeño en problemas similares.

Después de aplicar la codificación one-hot mediante get_dummies, se procedió a la implementación de los modelos MLP y LSTM. Los resultados mostraron que el modelo MLP obtuvo las mejores métricas de evaluación al utilizar las funciones de activación Identidad y ReLu, junto con el solver Adam. Para ver a más a detalle la programación de los modelos (Ver Apéndice C).

Cuarta Fase. Evaluación y validación del modelo

Después de compilar cada uno de los modelos mlp con los diferentes activation y solver que MLP proporciona, fueron los siguientes los que proporcionaron resultados positivos en la aplicación de las métricas de accuracy ,mse ,f1-score,recall y matriz de confusión (se mostraran los mejores resultados de cada modelo) :

Resultado Modelo mlp identify-adam

Se muestra en la Figura 8 los resultados de las métricas en el modelo MLP indentify-adam

Figura 8

Modelo mlp identify-adam

```
Para k=68, La precisión del modelo: 0.9244813278008299
Para k=68, MSE del modelo: 0.07551867219917012
Para k=68, F1-score del modelo: 0.9416292495189224
Para k=68, Recall del modelo: 0.9426369863013698
Para k=68, Matriz de confusión:
[[4560  556]
 [ 536 8808]]
```

Fuente. Autoría Propia

Precisión del modelo (Precision): La precisión mide la proporción de predicciones correctas entre el total de predicciones realizadas. En este caso, el modelo tiene una precisión del 92.45%, lo que indica que el 92.45% de las predicciones del modelo son correctas.

MSE del modelo (Mean Squared Error): El MSE mide el promedio de los cuadrados de los errores, es decir, la diferencia promedio entre las predicciones del modelo y los valores reales. Un MSE de 0.0755 indica que, en promedio, las predicciones del modelo tienen un error cuadrático medio relativamente bajo.

F1-score del modelo: El F1-score es la media armónica de la precisión y el recall. Es una métrica útil cuando deseas tener un equilibrio entre la precisión y el recall. Un F1-score de 0.9416 indica que el modelo tiene un buen equilibrio entre precisión y recall, lo que significa que maneja bien tanto los falsos positivos como los falsos negativos.

Recall del modelo: El recall mide la proporción de verdaderos positivos identificados correctamente entre el total de verdaderos positivos. Un recall de 0.9426 indica que el modelo es capaz de identificar correctamente el 94.26% de los casos positivos reales.

Matriz de confusión: La matriz de confusión muestra la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos generados por el modelo. En este caso, la matriz de confusión muestra que el modelo predijo correctamente 4560 casos negativos y 8808 casos positivos, pero también cometió 536 falsos negativos y 556 falsos positivos. Esto indica que el modelo tiene un buen rendimiento general en la clasificación de ambas clases, con un número relativamente bajo de errores.

Resultado modelo relu-adam

Se muestra en la Figura 9 los resultados de las métricas en el modelo MLP relu-adam

Figura 9

Modelo relu-adam

```
Para k=27, La precisión del modelo: 0.9148686030428769
Para k=27, MSE del modelo: 0.0851313969571231
Para k=27, F1-score del modelo: 0.9331450605550425
Para k=27, Recall del modelo: 0.9194135273972602
Para k=27, Matriz de confusión:
[[4638  478]
 [ 753 8591]]
```

Fuente. Autoría Propia

Precisión del modelo (Precision): La precisión mide la proporción de predicciones correctas entre todas las predicciones realizadas. En este caso, el modelo tiene una precisión del 91.49%, lo que significa que el 91.49% de las predicciones del modelo son correctas.

MSE del modelo (Mean Squared Error): El MSE mide el promedio de los cuadrados de los errores, es decir, la diferencia promedio entre las predicciones del modelo y los valores reales. Un MSE de 0.0851 indica que, en promedio, las predicciones del modelo tienen un error cuadrático medio relativamente bajo.

F1-score del modelo: El F1-score es la media armónica de la precisión y el recall. Es una métrica útil cuando se desea equilibrar la precisión y el recall. Un F1-score de 0.9331 indica que el modelo tiene un buen equilibrio entre precisión y recall, lo que significa que maneja bien tanto los falsos positivos como los falsos negativos.

Recall del modelo: El recall mide la proporción de verdaderos positivos identificados correctamente entre todos los positivos reales. Un recall de 0.9194 indica

que el modelo es capaz de identificar correctamente el 91.94% de los casos positivos reales.

Matriz de confusión: La matriz de confusión muestra la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos generados por el modelo. En este caso, la matriz de confusión muestra que el modelo predijo correctamente 4638 casos negativos y 8591 casos positivos, pero también cometió 753 falsos negativos y 478 falsos positivos. Esto indica que el modelo tiene un buen rendimiento general en la clasificación de ambas clases, con un número relativamente bajo de errores.

Modelo MLP más rápido

Se miden los tiempos de compilación de cada modelo y se llega a la conclusión que el modelo MLP con mejor tiempo de compilación es relu-adam. (Ver figura 10).

MLP identify-adam

Figura 10

Tiempo de compilación mlp identify-adam

```
Para k=68, La precisión del modelo: 0.9244813278008299
Para k=68, MSE del modelo: 0.07551867219917012
Para k=68, F1-score del modelo: 0.9416292495189224
Para k=68, Recall del modelo: 0.9426369863013698
Para k=68, Matriz de confusión:
[[4560  556]
 [ 536 8808]]
Tiempo de ejecución: 132.44579458236694 segundos
```

Fuente. Autoría Propia

MLP relu-adam

Figura 11

Tiempo de compilación mlp relu-adam

```
Para k=27, La precisión del modelo: 0.9148686030428769
Para k=27, MSE del modelo: 0.0851313969571231
Para k=27, F1-score del modelo: 0.9331450605550425
Para k=27, Recall del modelo: 0.9194135273972602
Para k=27, Matriz de confusión:
[[4638  478]
 [ 753 8591]]
Tiempo de ejecución: 119.2399754524231 segundos
```

Fuente. Autoría Propia

Resultado modelo LSTM

Los resultados obtenidos en el modelo lstm indican un comportamiento muy deficiente por lo cual no se considera apropiado usar este modelo para la situación planteada en este proyecto. Debido a que con el número de neuronas presentes no da resultados óptimos y para incluir más se necesita de una herramienta computacional más potente de la que se tiene disponible actualmente.

Validación cruzada y elección del mejor modelo

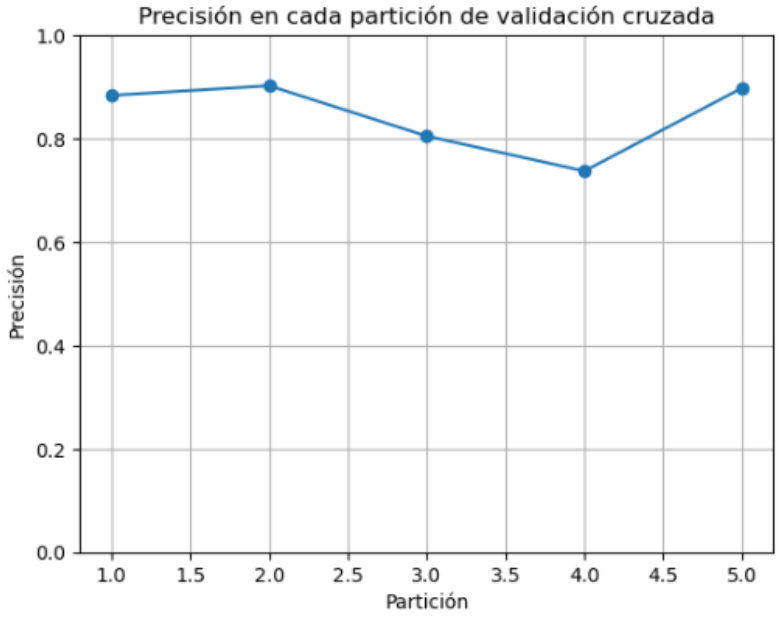
Debido a que los resultados obtenidos fueron muy similares en MLP se escogerá el mejor modelo por medio de validación cruzada.

MLP identify-adam

Se muestra en la figura 33 los resultados obtenidos de la validación cruzada del modelo MLP identify-adam.

Figura 12

Validación cruzada idenify-adam



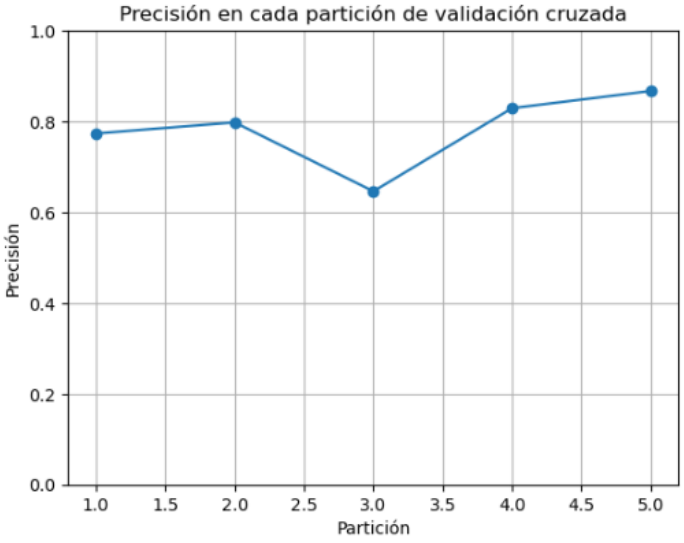
Fuente. Autoría Propia

MLP relu-adam

Se muestra en la figura 34 los resultados obtenidos de la validación cruzada del modelo MLP identify-adam.

Figura 13

Validación cruzada relu-adam

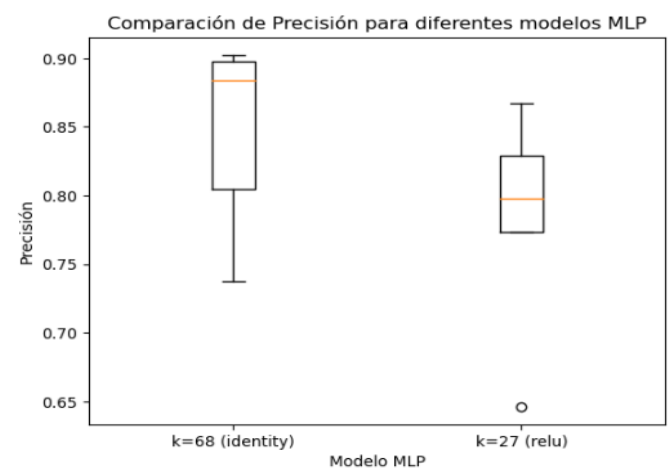


Fuente. Autoría Propia

Modelo relu-adam vs identify-adam

Figura 14

Comparación modelos



Fuente. Autoría Propia

Con los resultados obtenidos de la validación cruzada se llega a la conclusión que el modelo MLP identify-adam es el que tiene mejor rendimiento porque tiene un comportamiento más progresivo.

Quinta Fase. Implementación

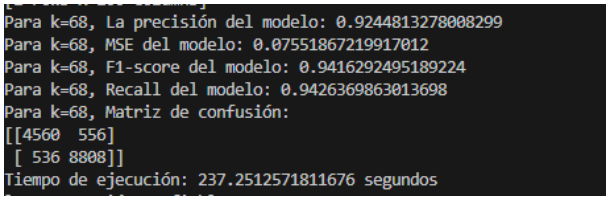
Una vez seleccionado el modelo óptimo, se procede a crear el archivo ejecutable para utilizar la red neuronal. Para ello, se configura un entorno virtual de Python que incluye únicamente los códigos necesarios para la limpieza de datos, el preprocesamiento y el modelo MLP con la función de activación "Identity" y el solver "Adam". Finalmente, se realiza una prueba con un input individual para verificar las predicciones del modelo.

Resultados de la red neuronal en el entorno virtual

Se aplican técnicas de programación orientada a objetos y se utiliza el modelo seleccionado, en este caso MLP con la función de activación "Identity" y el solver "Adam".

Figura 15

Resultado mlp Idenify-adam poo

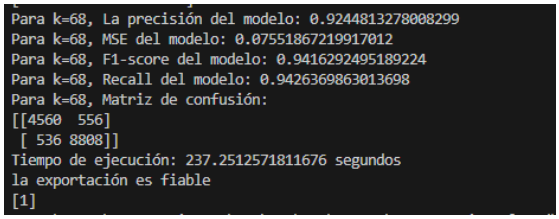


Fuente. Autoría Propia

Se crea una entrada (ver apéndice D) para mostrar si la exportación es fiable. En este caso, el resultado fue 1, lo que indica que la exportación es fiable.

Figura 16

Resultado input mlp Identify -adam



Fuente. Autoría Propia

Conclusiones

Según las revisiones bibliográficas realizadas se llega a la conclusión que las redes neuronales son el método ideal para el desarrollo de estudios para identificar oportunidades comerciales.

Se logra centralizar la información de las bases de datos de exportaciones de cacao y sus derivados utilizando la herramientas de la página legiscomex y técnicas de preprocesamientos de datos.

Se llevo a cabo el desarrollo de un modelo de red neuronal, en el cual se ha prestado especial atención a la selección de atributos y la ingeniería de características. Este proceso ha permitido identificar y retener únicamente aquellos campos que mejor se adaptan a las necesidades planteadas. La meticulosa selección y refinamiento de atributos no solo ha optimizado la precisión del modelo, sino que también ha contribuido significativamente a evitar el sobreajuste.

Con los resultados obtenidos en las métricas de evaluación el modelo que tuvo mejores resultados y que se ajusta más a la necesidad fue el modelo MLP, en el cual los mejores resultados se obtuvieron con el activation Identify y con el solver adam.

Recomendaciones

Automatizar el proceso de limpieza de datos por medio del uso de Python, evitando hacerlo de manera manual desde Excel.

Utilizar un equipo con mejores prestaciones tanto en CPU como en GPU para evitar demoras en los procesos de compilación.

Construir un aplicativo en cual se muestren los resultados de manera gráfica para mejor visualización por parte de un cliente potencial.

Referencias bibliográficas

Almeyda, E. M. (2022). Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de Machine Learning. Universidad de Piura.

Amazon. (2023). *aws*. Obtenido de ¿Qué es una red neuronal?: https://aws.amazon.com/es/what-is/neural-network

Aporta, I. (2022). Guía práctica para la mejora de la calidad de datos abiertos. España.

Arias, L. A. (2021). Evaluación de modelos de machine learning para sistemas de detección de intrusos en dedes IoT. universidad de los andes.

Barrios, J. I. (2019). *juanbarrios*. Obtenido de La matriz de confusión y sus métricas: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

CEVALLOS TORRES, L., VALENCIA MARTINEZ, N., & BARROS MORALES, R. (2017). Análisis Estadístico Univariado. grupo Compás - Universidad de Guayaquil.

Cordero-Torres, B. P. (2022). Algoritmos de Aprendizaje Supervisado para Proyección de Ventas de Camarón Ecuatoriano con Lenguaje de Programación Python. UTE.

Cutler, A., Richard Cutler, D., & Stevens, J. (2011). *Random Forests.* Springer.

ECBTI. (2016). *UNAD*. Obtenido de UNAD: https://academia.unad.edu.co/investigacion-ecbti/cadenas-de-formacion

Fedecacao. (2023). *Fedecacao*. Obtenido de Fedecacao: https://www.fedecacao.com.co/economianacional

Fuentes, L. F., & Garcia Jerez, A. (2021). *EVALUACIÓN INTEGRAL DE LA CALIDAD SENSORIAL DEL CACAO.* Bucaramanga: UNAD.

gajawada, s. k. (2019). *towardsdatascience*. Obtenido de ANOVA for Feature Selection in Machine Learning: https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476

Gamco. (2022). *Gamco*. Obtenido de ¿Qué es LSTM: Long short-term memory? : https://gamco.es/glosario/lstm-long-short-term-memory/#:~:text=La%20LSTM%20fue%20diseñada%20para,información%20importante%20en%20el%20proceso.

Gamco. (2021). *Gamco*. Obtenido de ¿Qué es Perceptrón Multicapa - MLP?: https://gamco.es/glosario/percepatron-multicapa-mlp/

Hernández, M. F. (2022). MODELADO DE ESTUDIOS DE MERCADO BASADOS EN MACHINE LEARNING PARA EMPRESAS DENTRO DEL COMERCIO ELECTRÓNICO. UNIVERSIDAD DEL ISTMO FACULTAD DE INGENIERÍA.

ibm. (2021). *ibm*. Obtenido de ARIMA: https://www.ibm.com/docs/es/spss-modeler/saas?topic=series-arima

ICCO. (2023). *ICCO*. Obtenido de ICCO: https://www.icco.org/

Ignacio, J. (2021). Aplicación de tecnologías de aprendizaje automático para predecir negocios y tomar decisiones empresariales. Universidad Nacional de La Plata.

Kane, F. (2017). *Hands-on data science and Python machine learning.* Packt Publishing.

legiscomex. (2023). *legiscomex*. Obtenido de legiscomex: https://legiscomex.com/es

López, C. P. (2010). *Técnicas de muestreo estadístico.* Ibergarceta.

Medina, S. A. (2016). Estrategias de marketing digital para la comercialización de chocolate premium en el Marketplace de Amazon Estados Unidos por la empresa Girones S.A. de Floridablanca Santander. UNAB.

Molina, M. G. (2017). Proceso de exportación del Grupo de Cacaoteros de Briceño Antioquia.

Munar, P. (2023). *cyberclick*. Obtenido de Data Science: predicciones de series temporales con machine learning: https://www.cyberclick.es/numerical-blog/data-science-predicciones-de-series-temporales-con-machine-learning

Ochoa, L. L. (2019). Evaluación de Algoritmos de Clasificación utilizando Validación Cruzada. National University of St Agustin.

Pierson, L. (2017). *Data science for dummies.* New Jersey: Wiley .

Prabhakaran, S. (2024). *machinelearningplus*. Obtenido de Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python: https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/

Quezada, J., Carvajal Romero, H., Barrezueta Unda, S., & Cordova, K. (2021). Economic analysis of the export of cocoa in Ecuador during the period 2014 – 2019. Universidad Técnica de Machala.

Quiroz Torres, A., & Vásquez Novoa, M. (2019). Oportunidades comerciales en el mercado de Suiza para las exportaciones peruanas de cacao en grano tostado - 2015.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Second Edition: Unlock modern machine learning and deep learning techniques with Python by using the latest cutting-edge open source Python libraries.* Birmingham: Packt Publishing.

Rojas, Y. E., Melo Mosquera, G., Agredo Madroñero, D., & Moncayo Rosero, J. (2021). Oferta exportable del cacao del Departamento de Nariño, (2010-2018). *Oferta exportable del cacao del Departamento de Nariño, (2010-2018)*. Colombia.

Saldaña, M. R. (2011). *dialnet*. Obtenido de La prueba chi-cuadrado o ji-cuadrado: https://dialnet.unirioja.es/servlet/articulo?codigo=3995561

simplilearn. (2023). *simplilearn*. Obtenido de What is Principal Component Analysis?: https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis

Solis, D. C. (2023). *openwebinars*. Obtenido de Datasets: Qué son y cómo acceder a ellos: https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/

Apéndices

Apéndice A

DATA SET Cacao

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link

Apéndice B

Jupyter Notebook Análisis exploratorio categorías

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link

Apéndice C

Jupyter Notebook Modelo Machine learning cacao

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link

Apéndice D

input

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link

Apéndice E

Entorno_virtual_red

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?us|p=drive_link

Apéndice F

Charla divulgación proyecto de grado Lucas Quintana-20240605_134909-Grabación de la reunión

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link

Apéndice G

Repositorio git

https://github.com/lucasquintana1604/Proyecto_aplicado_unad_cacao

Apéndice H

Instrucciones para descargar Data set

https://drive.google.com/drive/folders/1gNp29MRau_miqWCA1wBgSdGTpC_n0j3p?usp=drive_link