Análisis estadístico de la fluctuación de precios pagados por una compañía agroindustrial en la compra de lotes de café de acuerdo con la zonificación geográfica entre 2021 – 2023

Edwin Ernesto Useche Mahecha

Asesor

Mireya García

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Ciencia de Datos y Analítica

2024

Resumen

El café, un producto esencial para la economía global, ha enfrentado una notable volatilidad en el mercado entre 2021 y 2023. Factores como el cambio climático, las políticas comerciales internacionales, las fluctuaciones cambiarias y las alteraciones en las cadenas de suministro generadas por la pandemia de COVID-19 han afectado directamente en los precios del mercado C y, por ende, en los precios pagados a los productores locales. Ante esta realidad, este estudio desarrolló una estrategia de analítica de datos orientada a comprender las dinámicas de mercado y optimizar la toma de decisiones para una compañía agroindustrial colombiana.

A partir de un análisis exhaustivo, se identificaron variables clave como el precio del mercado C, la TRM, y atributos específicos del café, incluyendo calidad, variedad, certificación y factores geográficos. Se emplearon técnicas estadísticas y de machine learning, como la regresión lineal múltiple y modelos de series temporales (SARIMAX), para analizar el comportamiento del precio del café según distintas calidades. Los resultados mostraron que existe una fuerte correlación entre el precio del mercado C y los precios pagados por calidades como RTB, mientras que las calidades A y AA presentaron problemas de heterocedasticidad, sugiriendo la implementación de modelos GARCH para mejorar la precisión del análisis.

Los hallazgos permitieron formular recomendaciones estratégicas, como la adopción de técnicas de codificación categórica más avanzadas para datos cualitativos y la inclusión de factores exógenos que reflejen las condiciones del mercado global. Este enfoque integral no solo proporciona una base sólida para decisiones informadas, sino que también posiciona a la compañía para enfrentar desafíos futuros en un entorno volátil y competitivo, fortaleciendo su sostenibilidad y rentabilidad a largo plazo.

***Palabras clave*:** precio del mercado C, análisis estadístico, regresión, series temporales, SARIMAX.

Abstract

Coffee, an essential product for the global economy, has experienced significant market volatility between 2021 and 2023. Factors such as climate change, international trade policies, exchange rate fluctuations, and supply chain disruptions caused by the COVID-19 pandemic have directly impacted C market prices and, consequently, the prices paid to local producers. In response to this reality, this study developed a data analytics strategy aimed at understanding market dynamics and optimizing decision-making for a Colombian agribusiness company.

Through a comprehensive analysis, key variables were identified, including the C market price, the exchange rate (TRM), and specific coffee attributes such as quality, variety, certification, and geographic factors. Statistical and machine learning techniques, including multiple linear regression and time series models (SARIMAX), were employed to analyze the behavior of coffee prices across different quality categories. The results showed a strong correlation between the C market price and the prices paid for qualities such as RTB, while qualities A and AA exhibited heteroskedasticity issues, suggesting the implementation of GARCH models to improve analysis accuracy.

The findings enabled the formulation of strategic recommendations, such as adopting more advanced categorical encoding techniques for qualitative data and including exogenous factors that reflect global market conditions. This comprehensive approach not only provides a solid foundation for informed decision-making but also positions the company to address future challenges in a volatile and competitive environment, strengthening its long-term sustainability and profitability.

***Keywords*:** coffee price, statistical analysis, regression, time series, SARIMAX

.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El café ha sido, durante siglos, uno de los pilares fundamentales de la economía global. Más allá de ser una bebida apreciada por millones de personas en todo el mundo, su cultivo y comercialización representan una fuente clave de ingresos para numerosos países productores. Entre ellos, Colombia ocupa un lugar destacado gracias a la calidad y prestigio de su café, características que la han posicionado como líder en el mercado internacional. Sin embargo, el sector cafetero enfrenta desafíos cada vez más complejos, como la volatilidad de los precios internacionales, exacerbada por fenómenos globales recientes como el cambio climático, la pandemia de COVID-19 y las fluctuaciones en las tasas de cambio.

Entre 2021 y 2023, estos factores han generado incertidumbre significativa, afectando no solo el precio del café en el mercado C, sino también el precio pagado a los productores locales. Para compañías agroindustriales colombianas que compran café directamente a los productores, esta inestabilidad ha creado la necesidad de adoptar estrategias innovadoras que optimicen la toma de decisiones. En este contexto, la analítica de datos emerge como una herramienta esencial para comprender las dinámicas de precios y desarrollar respuestas adaptativas que garanticen la sostenibilidad y rentabilidad del negocio.

El presente proyecto tiene como objetivo principal diseñar una estrategia de análisis de datos que permita identificar las variables más influyentes en la variabilidad del precio del café y aplicar métodos estadísticos avanzados, como análisis de series temporales y técnicas de machine learning, para optimizar la interpretación de esta información. A través de esta estrategia, se busca no solo entender mejor el comportamiento del mercado, sino también generar recomendaciones prácticas que permitan a las compañías agroindustriales mitigar riesgos, mejorar la rentabilidad y fortalecer su competitividad en un mercado dinámico y retador. Con

ello, se pretende contribuir tanto al desarrollo de la industria como al bienestar de los pequeños productores que dependen de esta actividad.

Justificación

La importancia del café en la economía global es innegable. Como uno de los productos agrícolas más comercializados en el mundo, su precio afecta directamente a las economías de muchos países productores, así como a la vida de millones de pequeños agricultores que dependen de este cultivo y de las compañías que les compran directamente sus productos. Dada esta realidad, una estrategia de analítica de datos robusta no solo es deseable, sino esencial para fortalecer la sostenibilidad económica de las compañías agroindustriales.

Identificar las variables más significativas que inciden en la variabilidad del precio del café permitirá tener una comprensión más clara de las dinámicas que lo afectan. Esto, a su vez, facilitará la implementación de estrategias de mitigación de riesgos y optimización de beneficios. Los métodos estadísticos adecuados proporcionarán una base científica para el análisis, evitando conclusiones erróneas basadas en información sin sustento o tomada por actores de forma subjetiva.

Además, la evaluación continua de los resultados obtenidos a través de esta estrategia permitirá ajustes y mejoras constantes, asegurando su relevancia y efectividad a lo largo del tiempo. En última instancia, una estrategia de analítica de datos bien implementada fortalecerá la adaptabilidad de la compañía agroindustrial participe en el desarrollo del proyecto, permitiendo adaptarse mejor a las condiciones cambiantes del mercado y contribuyendo a una mayor estabilidad económica.

Objetivos

Objetivo General

Mejorar la rentabilidad y calidad del café a distribuir mediante una estrategia de análisis de datos que considere al precio y volumen de compra de café de especialidad en el periodo comprendido entre 2021 – 2023.

Objetivos Específicos

Identificar las variables más significativas que inciden en la variabilidad del precio del café.

Determinar las mediciones estadísticas tanto descriptivas como inferenciales y/o de data science que mejor se adapten para el análisis de las variables seleccionadas.

Evaluar los resultados obtenidos del desarrollo de la estrategia de analítica planteada.

Identificar las oportunidades de mejora de acuerdo a los resultados obtenidos.

Presentar las recomendaciones que surjan del desarrollo de la estrategia de analítica planteada

Planteamiento del Problema

El mercado del café, especialmente entre los años 2021 y 2023, ha experimentado una notable volatilidad en los precios de los lotes. Esta volatilidad puede ser atribuida a una variedad de factores tanto internos como externos. Entre ellos, las condiciones climáticas extremas como lo fueron las sequias en Brasil afectaron la producción, mientras que la pandemia de COVID-19 ha perturbado las cadenas de suministro globales, impactando los costos de transporte y logística. Además, las políticas comerciales internacionales y las fluctuaciones en las tasas de cambio han añadido capas adicionales a la volatilidad de los precios.

Colombia no ha sido ajena a los efectos que esto ha generado, y compañías agroindustriales que compran café directamente a los productores, han tenido que afrontar distintos retos que van desde: conseguir un mayor apalancamiento financiero para hacer frente al incremento de los precios internacionales de negociación, hasta implementar estrategias para competir con otros actores del mercado que buscan acaparar a los productores con los que se tiene una relación comercial de larga data, lo cual suele traducirse en un mayor efectivo a desembolsar.

A pesar de la abundancia de datos disponibles, la falta de una estrategia analítica sistemática para su interpretación impide una comprensión profunda de los factores que más influyen en el precio del café. Esta carencia se traduce en decisiones basadas en información incompleta o en suposiciones, lo que puede llevar a pérdidas económicas significativas para la organización. En un mercado tan competitivo, la capacidad de anticipar y responder a las fluctuaciones del precio es crucial para mantener la rentabilidad y la estabilidad.

Marco de Referencia

Factores Economicos y Politicos

El precio del café está influenciado por factores económicos y políticos. La volatilidad de las tasas de cambio es uno de los elementos críticos que afectan los precios del café. (Bittencourt & Agudelo, 2021) analizaron el impacto de la volatilidad de la tasa de cambio en el comercio colombiano, encontrando que las fluctuaciones en las tasas de cambio pueden causar inestabilidad en los precios de los productos exportados, al igual que variaciones significativas en los precios de los productos que son negociados en una divisa unificada como lo es el dólar.

Además, políticas de internacionalización económica, como las descritas por (Romero, 2020), han tenido efectos directos en la competitividad del mercado colombiano a nivel internacional.

Impacto de Factores Climáticos en el Precio del Café

Uno de los factores más estudiados en relación con las fluctuaciones de precios en el mercado del café es el clima. Estudios como el de (Bastianin et al., 2018) han demostrado mediante el desarrollo de modelos de series temporales que eventos climáticos extremos, como el fenómeno de El Niño, tienen un impacto significativo en la producción de café en Colombia. Estos eventos pueden reducir la oferta de café, lo que a su vez incrementa los precios en el mercado internacional.

De manera similar (Avelino et al, 2015) analizaron la crisis de la roya del café en Colombia y Centroamérica entre 2008 y 2013, concluyendo que las enfermedades causadas por cambios climáticos también afectan negativamente la producción y elevan los precios.

Precio C del café

Se conoce como mercado C en la industria del café, al mercado de materias primas en donde los agentes de la bolsa de valores de New York, establecen día a día el precio futuro de los contratos de café a nivel global (Kilimanjaro Specialty Coffees, s.f.). Como se consigna en un artículo de (Boydell, 2020) fundamentalmente, el Precio C se define según la oferta y la demanda, esto quiere decir que el precio se define en el punto en el cual la oferta equivale a la demanda.

Competitividad y Sostenibilidad en la Producción de Café

La competitividad de las exportaciones de café es otro tema crucial para el análisis del mercado. (Cerquera et al., 2020) analizan a través de modelos de regresión la competitividad de las exportaciones de café producido en el departamento de Huila, de este estudio se identifica que factores como la calidad del producto y las estrategias de marketing son esenciales para mantener una posición favorable en el mercado internacional.

Además, la sostenibilidad en la producción de café ha sido un tema de interés creciente. (Ordoñez-Jurado & Castillo-Marín, 2021) evaluaron la sostenibilidad de los sistemas de producción de café en algunos municipios del departamento de Nariño, destacando la importancia de prácticas agrícolas sostenibles para garantizar la estabilidad a largo plazo del sector.

Análisis Estadístico

El análisis estadístico es una rama de la estadística que implica la recopilación, exploración y presentación de grandes volúmenes de datos para descubrir patrones subyacentes y tendencias significativas. Como se evidencia en el texto de (Urdan, 2017), este proceso incluye técnicas descriptivas, inferenciales y predictivas, que son fundamentales para la toma de

decisiones basadas en datos. El análisis estadístico es esencial en diversas áreas de investigación y negocios, ya que permite interpretar datos y convertirlos en información útil. Los métodos descriptivos incluyen medidas de tendencia central y dispersión, mientras que los métodos inferenciales permiten hacer predicciones y generalizaciones sobre una población basada en una muestra.

Regresiones

La regresión es una técnica estadística utilizada para modelar y analizar las relaciones entre una variable dependiente y una o más variables independientes. Los modelos de regresión permiten predecir el valor de la variable dependiente basándose en los valores de las variables independientes. (Montgomery et al., 2012) Existen varios tipos de regresiones, siendo las más comunes la regresión lineal y la regresión múltiple. Estas técnicas son ampliamente utilizadas en economía, ciencias sociales y ciencias naturales para identificar y cuantificar relaciones causales.

Considerando los objetivos establecidos en el estudio, será la regresión múltiple la que se implementará, es una herramienta fundamental en el análisis de datos, ya que permite modelar la relación entre una variable dependiente y múltiples variables independientes. Este enfoque es especialmente útil cuando se desea comprender el impacto relativo de diferentes factores sobre la variable de interés. A través de la regresión lineal múltiple, se pueden identificar patrones complejos que no pueden ser capturados por modelos univariantes, y ajustar los efectos de múltiples variables al mismo tiempo. (Montgomery et al., 2012).

La regresión lineal múltiple facilita la interpretación de los coeficientes, que indican el cambio promedio en la variable dependiente asociado a un cambio en una de las variables independientes, mientras las demás se mantienen constantes. Tal como indica (Fox, 2015) esto proporciona una visión más profunda y detallada de las relaciones entre las variables, al incluir

más predictores relevantes en el modelo, se puede mejorar la precisión de las predicciones y obtener resultados más sólidos para la toma de decisiones.

Análisis de Series Temporales

El análisis de series temporales es una técnica estadística utilizada para analizar datos que se han recopilado a lo largo del tiempo. Este tipo de análisis es esencial para entender las dinámicas temporales de los datos y hacer predicciones futuras. (Shumway & Stoffer, 2017) definen que las series temporales pueden ser univariadas (una sola variable) o multivariadas (varias variables), y se utilizan en diversas disciplinas, incluyendo economía y control de calidad.

Las características principales de un análisis de series de tiempo suelen estar enfocadas en la tendencia y la estacionalidad, donde la primera hace referencia a la dirección general en la que las observaciones se desarrollan, que permite identificar un aumento o disminución a lo largo de la secuencia analizada. Mientras que la estacionalidad presenta las variaciones que ocurren en intervalos regulares específicos, los cuales pueden ser diarios, semanales mensuales o anuales (Auffarth, 2021).

Este estudio estará enfocado en series temporales multivariadas, donde se hará uso del modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors), como evidencian (Hyndman & Athanasopoulos, 2018) el modelo SARIMAX es una extensión del modelo ARIMA que permite incorporar tanto componentes estacionales como variables exógenas en el análisis de series temporales. Este enfoque es útil cuando se desea modelar series temporales que no solo exhiben patrones de tendencia y estacionalidad, sino que también se ven influenciadas por factores externos que pueden afectar la dinámica de la serie.

La incorporación de variables exógenas mejora la precisión del modelo al capturar la relación entre la serie temporal y otros factores externos que tienen un impacto en las

observaciones, permitiendo modelar de manera más completa y precisa la evolución de una serie temporal, proporcionando pronósticos más robustos y explicaciones más detalladas de las dinámicas subyacentes (Hyndman & Athanasopoulos, 2018).

Técnicas de Machine Learning

Machine learning (aprendizaje automático) es un subcampo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos que permiten a las computadoras aprender y hacer predicciones basadas en datos. Estas técnicas son particularmente útiles para analizar grandes conjuntos de datos y encontrar patrones complejos que serían difíciles de detectar mediante métodos estadísticos tradicionales. Las técnicas de machine learning se dividen generalmente en dos categorías: supervisado y no supervisado. De acuerdo con (Murphy, 2012) el aprendizaje supervisado utiliza datos etiquetados para entrenar modelos que pueden predecir resultados futuros, mientras que el aprendizaje no supervisado busca patrones ocultos en datos no etiquetados

Modelos Estadísticos Aplicados al Análisis del Precio del Café

Diversos estudios han aplicado métodos estadísticos para analizar y predecir los precios del café. (Monsalve, 2016) desarrolló un aplicativo en Excel para la enseñanza y modelación de series de tiempo, utilizando el precio externo del café colombiano como caso de estudio. Este enfoque permite identificar patrones y tendencias en los datos históricos, facilitando la predicción de precios futuros.

Por otro lado, (Velásquez & Aldana , 2007) emplearon redes neuronales artificiales para modelar el precio del café colombiano en la bolsa de Nueva York, demostrando que este método puede capturar relaciones no lineales en los datos y mejorar la precisión de las predicciones.

Asimismo, (Ordoñez-Jurado & Castillo-Marín, 2021) evaluaron la sostenibilidad en los sistemas de producción de café en Nariño, Colombia, aplicando algoritmos de machine learning para identificar las prácticas más sostenibles y rentables.

Metodología

Este proyecto se desarrollará bajo un enfoque mixto, donde el enfoque cuantitativo estará orientado en el planteamiento y desarrollo de la estrategia de analítica de datos sobre la compra de lotes de café en el periodo comprendido entre 2021 – 2023.

El enfoque cualitativo se encuentra encaminado al análisis e interpretación de los resultados obtenidos de la estrategia de analítica de datos. En la Tabla 1 la metodología a seguir con base a los objetivos a alacnzar en este proyecto.

Objetivos Específicos y Metodología

Tabla 1

Metodología a Partir de los Objetivos Específicos

Objetivo específico	Metodología	Herramientas
Identificar las variables más significativas que inciden en la variabilidad del precio del café.	Análisis descriptivo del perfil de las variables (media, varianza, etc.) Análisis de correlación de variables. Búsqueda de datos atípicos, normalización y limpieza de datos.	Python R
Determinar las mediciones estadísticas tanto descriptivas como inferenciales y/o de data science que mejor se adapten para el análisis de las variables seleccionadas.	Desarrollar y ejecutar la estrategia de analítica de datos establecida, donde se debe tener delimitado mediciones estadísticas y/o de data science a trabajar.	Modelos de machine learning Estadística descriptiva e inferencial Modelos de regresión Python R
Evaluar los resultados obtenidos del desarrollo de la estrategia de analítica planteada.	Analizar los resultados obtenidos del desarrollo de la estrategia de analítica de datos.	Python R

Identificar las oportunidades de mejora de acuerdo a los resultados obtenidos.	Establecer las áreas que son susceptibles de mejora, tras considerar los resultados obtenidos en el desarrollo del proyecto.	Python R
Presentar las recomendaciones que surjan del desarrollo de la estrategia de analítica planteada.	Presentar los resultados, las recomendaciones y las oportunidades de mejora.	Informe final

Nota. Definición de la metodología a trabajar para alcanzar el resultado definido en cada uno de los objetivos específicos.

Comprensión de los Datos

Se hace uso de tres bases de datos, cada una de ellas comprenden el periodo entre 2021 y 2023.

Transacciones de Compras de la Compañía

Este conjunto de datos se obtiene de la información almacenadas en los sistemas de información de la compañía, en los cuales se almacena un registro detallado de cada una de las compras efectuadas. El conjunto de datos contiene tanto datos cualitativos como datos cuantitativos, con un total de 48 variables y 32.348 filas. No todas las variables son incluidas en el análisis, en las actividades de preprocesamiento se delimita el conjunto de datos, como se puede ver en la Tabla 2.

Tabla 2

Definición de las Variables Conjunto de Datos de Compras

Variables	Observación	Incluida
Lot	Identificador de lotes comprados.	No
Lot/Set	Identificador de si el lote esta individual o integrado a un set.	No
State	Estado de compra.	No
Group State	Estado agrupado de compra.	Si
Pricing Date	Fecha en la que se realiza la compra.	Si
Local Cost/Kg	Precio pagado por cada Kg comprado, en pesos colombianos.	No
Total Local Cost	Precio pagado en total por cada lote, en pesos colombianos.	No
Created at	Fecha en la que se registra por primera vez el lote en el ERP.	No
Updated at	Última fecha en la que se registra alguna transacción con el lote en el ERP.	No
Producer	Nombre del productor.	Si
Association	Nombre de la asociación, si el productor hace parte de alguna	Si
National ID	Identificación del productor.	No
Gender	Genero del productor.	Si
Age	Edad del productor.	Si
PECA State	Define si se le brindo asesoramiento en temas de optimización y sostenibilidad de su cosecha.	Si
Phone Number	Numero celular del productor.	No
Parchment Kg	Cantidad de Kg disponibles compradas después de extraída la muestra para análisis de calidad.	No
Parchment Initial Kg	Cantidad de Kg comprados originalmente, sin extraer la muestra para análisis de calidad.	Si
Cost Parchment USD/Kg	Precio pagado por cada Kg comprado en dólares americanos.	Si

Estimated Green Kg	Cantidad de café excelso que se espera obtener de cada lote, después de realizar el proceso de trilla.	No
Green 60Kg Bags	Cantidad de sacos de 60 Kg de excelso que se espera obtener. Dato calculado al dividir Estimated Green Kg entre 60.	No
Total Cost USD	Precio pagado en total por cada lote en dólares americanos.	Si
#Lots	Número de lotes comprados, por defecto es 1.	No
Pricing	Define si ya fue establecido el precio con el productor	No
Local Currency	Divisa en la que se negocia con el productor	No
Origin	País donde se efectúa la compra, en este caso todas las compras se realizan en Colombia	No
Farm	Nombre de la finca de donde se obtiene el café.	No
Farm2	Nombre secundario de la finca de donde se obtiene el café, aplica si esta parcelada.	No
WH Location	Id de la bodega que será el destino final de los lotes comprados.	No
WH Origin	Id de la bodega en la que se realiza la compra.	No
Variety	Variedad de café comprada.	Si
Process Type	Tipo de proceso de secado aplicado.	Si
Coffee Type	Tipo de café comprado, se agrupa en 3 tipos principales: seco, natural y honey.	Si
Grade	Grado de calidad asignado de acuerdo con los resultados de los análisis de calidad.	No
Quality	Categorización del grado de calidad.	Si
Yield 14	Registro del análisis de calidad, donde se pasa el café a través de una malla y se evalúa la proporción de café que queda en esta (evalúa el tamaño de los granos).	No

Yield 15	Registro del análisis de calidad, donde se pasa el café a través de una malla y se evalúa la proporción de café que queda en esta (evalúa el tamaño de los granos).	No
Parchment 60Kg Bags	Cantidad de sacos de 60 Kg comprados. Dato calculado al dividir Parchment Kg entre 60.	No
Factor 14	Factor de rendimiento sobre malla 14, a menor valor mejor rendimiento.	Si
Factor 15	Factor de rendimiento sobre malla 15, a menor valor mejor rendimiento.	Si
Area	Área en metros cuadrados de la finca.	No
Cost Green USD/Lb	Costo estimado de la libra de excelso que se espera obtener.	No
Certificate	Certificaciones internacionales que pueda tener los lotes comprados, al tener una certificación se realiza un pago de prima por este aspecto.	Si
Country	País en el que se realiza la compra.	No
Department	Departamento en el que se realiza la compra.	Si
Municipality	Municipio en el que se realiza la compra.	Si
Humidity Green	Grado de humedad obtenido de los análisis realizados a la muestra, recomendablemente debe estar entre 9.5-11%.	Si
Latitude	Latitud geográfica de la finca, se refiere a coordenadas.	No
Longitude	Longitud geográfica de la finca, se refiere a coordenadas.	No

Nota. Descripción de las variables del conjunto de datos de las compras historicas de café por parte de la compañía.

Tasa Representativa del Mercado (TRM)

El conjunto de datos se obtuvo de los datos abiertos del Banco de la Republica de Colombia. Estos datos muestran la tasa de cambio en la que se negocia un dólar americano en pesos colombianos para cada uno de los días del periodo a analizar.

Mercado C

Este conjunto de datos contiene el registro histórico de los precios negociados en la bolsa de valores de New York para la libra de café (lb/USD), los datos son obtenidos del portal financiero Investing.com, que es una plataforma en la cual se hace seguimiento a los precios negociados para distintos commodities, entre ellos el café.

Preprocesamiento de los Datos

Considerando los objetivos planteados en el estudio y los modelos que se van a desarrollar (regresión lineal múltiple y un análisis de series de tiempo), es necesario evaluar la estructura inicial de los datos, con el fin de preparar los datos; ya sea eliminando características que no son de interes, imputando datos vacíos, identificando y dando tratamiento a los outliers, o haciendo otras transformaciones necesarias, que permitan tener los datos en la estructura mas adecuada posible para incluirlos en los modelos indicados.

Las variables originales de cada uno de los conjuntos de datos, se redujeron en con base al conocimiento que se tiene del negocio y que permite establecer las variables que son relevantes para el análisis, adicionalmente se aplican análisis estadisticos con los cuales se identifican categorías claves en las variables sobre las cuales se centran los análisis.

Transacciones de Compras de la Compañía

De las 48 variables que tenia el conjunto de datos original, se reduce la dimensionalidad a 20 variables. Con base a algunas de las variables que se incluyen en el conjunto de datos final, se

analiza las categorías que las componen, esto para identificar si hay una concentración de datos una única o pocas categorías y centrar el análisis en estas.

Con base a los análisis realizados a algunas de las categorías, se redujeron las categorías en algunas carcateristicas, lo que conlleva a la eliminación de algunos registros, reduciendo el conjunto de datos a 31.216 filas, lo que corresponde aproximadamente al 96% del total de los datos originales. Como se evidencia en la Tabla 3 y en el Apéndice A, del análisis realizado a algunas de las características se identifica información clave para el desarrollo del proyecto.

Tabla 3

Preprocesamiento del Conjunto de Datos de Compras

Variables	Análisis/Cambios efectuados
Group State	El 99% de las oportunidades de compras se efectúan, por lo tanto se eliminan los registros de las compras no efectuadas.
Pricing Date	Se identifica las tendencias en las compras, donde se evidencia que aproximadamente en julio de cada año se tiene un pico en las compras efectuadas.
Producer	Al ser información sensible se seudonimiza el nombre del productor.
Association	Al ser información sensible se seudonimiza el nombre de la asociación.
Gender	Se determina que aproximadamente el 85% de los productores son hombres.
Age	Se identifica valores atípicos, que se presentan por errores en el ingreso de la información de los productores en el sistema de información de la compañía.
Parchment Initial Kg	Se identifica los rangos mas comunes entre los que se distribuyen las compras y los valores atípicos.
Cost Parchment USD/Kg	Se encontraban registradas compras a un costo de 0 (café producido en fincas de la compañía), para evitar que esto genere ruido en los resultados se elimina esos registros.

Total Cost USD	Analisis realizado a nivel de la cantidad y el costo unitario.
PECA State	Define si se le brindo asesoramiento en temas de optimización y sostenibilidad de su cosecha.
Certificate	Se imputa los valores vacíos que corresponden a lotes que no tienen asociada ninguna certificación.
Variety	Se identifica las variedades que se compran con mayor frecuencia.
Process Type	Aproximadamente el 98% de los lotes comprados fueron sometidos a un proceso de secado mediante la técnica de lavado, por lo que se elimina los registros que correspondan a métodos de secado diferentes.
Coffee Type	Aproximadamente el 99% de los lotes comprados se adquieren es estado seco, por lo que se elimina los registros que correspondan a otros tipos.
Quality	Se identifica las calidades que se compran con mayor frecuencia, considerando que es una variable que tiene relación directa con el precio no se elimina ningún registro.
Factor 14	Se identifica que los valores atípicos en el factor de rendimiento se deben en mayor medida al tipo de secado al que sea sometido el café.
Factor 15	Se identifica que los valores atípicos en el factor de rendimiento se deben en mayor medida al tipo de secado al que sea sometido el café.
Department	Se identifica los departamentos en los que se realiza una mayor cantidad de compras.
Municipality	Se identifica los municipios en los que se realiza una mayor cantidad de compras.
Humidity Green	Se valida que la humedad de los lotes se encuentra dentro de los estándares establecidos por la compañía.

Nota. Analisis y cambios realizados a cada una de las variables del conjunto de datos de compras.

Considerando que el estudio se encamina a determinar cuales son las variables que tienen un mayor impacto en el precio pagado por la compañía en la compra de los lotes de café y como ha evoluciado el precio pagado a lo largo del periodo evaluada respecto a estas caracteristicas, se determina que en base a la variable Cost Parchment USD/Kg se identificaran los valores atípicos, lo que conlleva a que el conjunto de datos se divida en dos; uno que contendrá los registros que no se consideren valores atípicos en este análisis inicial y otro conjunto que contendrá únicamente los registros con valores atípicos.

Se emplea el método de rango intercuartílico para identificar los outliers, esto teniendo en cuenta que este método suele tener mejor resultado en la identificación de los datos atípicos para datos que no están distribuidos normalmente. Una vez implementada la segmentación del conjunto de datos, se obtiene un dataset sin outliers con 31.086 registros con un que se encuentra entre aproximadamente los 2 USD hasta los 7 USD por Kg de café comprado, mientras que el dataset que contiene los outliers contiene solo 130 registros. En el Apéndice A se detalla las transformaciones efectuadas al conjunto de datos.

Tasa Representativa del Mercado (TRM)

Para este conjunto de datos, se eliminan algunos campos que son redundantes en la información suministrada, como lo son los campos de año, día del mes, mes e Id mes, como se indica en la Tabla 4. Esta información es posible analizarla partiendo del campo de fecha. Adicionalmente, se cambia el tipo de datos de los campos Fecha (dd/mm/aaaa) y TRM, para que sean consecuentes con la información que almacenan.

Tabla 4

Estructura Original Datos TRM

Año	Fecha (dd/mm/aaaa)	TRM	Día del mes	Mes	Id Mes
2023	2023-12-28	3844,81	28	Diciembre	12
2023	2023-12-29	3822,05	29	Diciembre	12
2023	2023-12-30	3822,05	30	Diciembre	12
2023	2023-12-31	3822,05	31	Diciembre	12

Nota. Estructura en la que se obtienen los registros sobre la tasa representativa del mercado del Banco de la República.

Mercado C

Dado que la información que se obtiene del portal financiero Investing.com solo registra datos para fechas en las que la bolsa de valores de New York opere, solo se tiene registro de precios para los días hábiles, como se puede ver en la Tabla 5. Considerando lo anterior se hace necesario crear un nuevo dataframe que comprenda todas las fechas entre el 1 de enero de 2021 y el 31 de diciembre de 2023, al cual se trasladan los valores del dataframe original y para las fechas que no tiene un valor asignado se le imputa el valor de la fecha inmediatamente anterior.

Tabla 5

Estructura Original Datos Mercado C

Fecha	Último	Apertura	Máximo	Mínimo	Fecha
22/12/2023	192,8	192,2	194,25	189,2	22/12/2023
26/12/2023	194,35	192,8	195,4	191,95	26/12/2023
27/12/2023	197,75	194,9	199,95	194,05	27/12/2023
28/12/2023	198	196,75	201,35	195,15	28/12/2023

29/12/2023	188,3	197,2	198,15	185,65	29/12/2023

Nota. Estructura en la que se obtienen los registros sobre el precio del mercado C.

Una vez se efectua este cambio, se hace la conversión de los valores de la variable Apertura que son los que muestran el precio con el que se negocia el café, llevandolos a USD/Kg, para esto se divide este valor entre 100 y después se multiplica por 2,2046, que es la cantidad de libras que hay un kilogramo, tal como se muestra en la Tabla 6.

Tabla 6

Estructura Final Datos Mercado C

Fecha	Apertura
22/12/2023	4,2372
23/12/2023	4,2372
24/12/2023	4,2372
25/12/2023	4,2372
26/12/2023	4,2504
27/12/2023	4,2967

Nota. Estructura en la que se obtienen los registros sobre el precio del mercado C.

Modelos

Considerando los objetivos planteados y la naturaleza de los datos que se analizan, la presente metodología se enfoca en la implementación de técnicas de análisis de datos para identificar las características más relevantes que explican el costo pagado por kilogramo de café por parte de la compañía, así como para analizar el comportamiento del precio del café a lo largo del tiempo. Para alcanzar estos objetivos, se emplearán dos enfoques principales: un modelo de

regresión lineal múltiple y un modelo de series de tiempo, diseñados para abordar distintas dimensiones del problema.

Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple es el enfoque más adecuado para el análisis de los factores que determinan el costo por kilogramo de café, ya que permite modelar de manera eficiente la relación entre una variable dependiente y múltiples variables independientes, tanto cualitativas como cuantitativas (Montgomery et al., 2012). A diferencia de otros métodos más complejos, como las redes neuronales o los modelos no lineales, que requieren grandes volúmenes de datos y presentan dificultades interpretativas, la regresión lineal múltiple facilita la identificación de variables significativas de manera clara y cuantificable.

Este modelo es ampliamente utilizado en la investigación económica y empresarial debido a su capacidad para proporcionar resultados interpretables, lo que facilita la toma de decisiones estratégicas, como la mejora en la rentabilidad y calidad de los productos (Kutner et al., 2004). Además, su simplicidad y robustez permiten una evaluación precisa de los factores que impactan el comportamiento de la variable objetivo, ofreciendo ventajas en términos de confiabilidad y facilidad de comunicación de los resultados (Field, 2009). Por estas razones, la regresión lineal múltiple es la metodología más adecuada para los objetivos de este análisis

Este enfoque permite identificar y cuantificar la relación entre la variable objetivo, el costo por kilogramo de café, y diversas variables independientes, de tipo cualitativo, por ejemplo, la certificación, el tipo de proceso de secado y la calidad; y cuantitativas como lo son los factores de rendimiento, la humedad de los granos, la tasa de cambio y el precio establecido mediante el mercado C. El objetivo es determinar qué variables tienen un impacto significativo

en el costo, lo que facilitará la toma de decisiones informadas para mejorar la rentabilidad y la calidad del café.

Inicialmente, se procede a unir los tres conjuntos de datos con el fin de tener en el mismo dataset la información referente a las compras, junto con la TRM y el precio establecido por el mercado C para cada día. Se toma como campo clave la fecha en cada uno de los conjuntos de datos y se crea un nuevo dataset que contiene esta información.

Con el objetivo de evitar que se aumente la dimensionalidad del conjunto de datos y teniendo en cuenta que dentro del análisis a realizar mediante la regresión lineal múltiple se incluirán variables que originalmente son categóricas, se emplea la clase LabelEncoder de la librería scikit-learn para codificar las categorías de las variables categóricas de forma tal que puedan ser incluidas en el modelo, este método asigna un número a cada una de las categorías de cada campo categorico incluido. Es útil es el contexto de este ejercicio, ya que no se cuenta con variables que tengan un número extenso de categorías y adicionalmente se evita aumentar la dimensionalidad del conjunto de datos.

Con el conjunto de datos definido y con el objetivo de identificar si exista una dependencia significativa entre variables, se procede a establecer la correlación entre las variables. Como se muestra en el Apéndice B, en este análisis se evidencia que el precio de apertura del mercado C tiene una correlación positiva aproximada del 80% con el precio pagado por Kg de café por la compañía, mientras que con las demás variables no se evidencia una correlación considerable.

El modelo de regresión lineal tiene como variable objetivo Cost parchment USD/Kg (precio pagado por cada Kg de café comprado por la compañía), mientras que tiene 12 variables predictoras. Aunque el resultado del modelo muestra que con estas variables es posible explicar

aproximadamente el 74% del comportamiento del precio pagado por Kg como se evidencia en la Figura 1, se evidencia serios problemas en el cumplimiento de los supuestos para una regresión lineal.

Partiendo de un análisis visual del comportamiento de los residuos generados en la regresión, que se encuentra en el Apéndice B, se identifica que el modelo no cumple con el supuesto de normalidad, donde se presentan una desviación significativa de los valores en los extremos. De igual manera se observa que los residuos tienen una ligera tendencia a variar en amplitud dependiendo de los valores predichos, esto indica que posiblemente la varianza de los residuos no es constante, es decir que hay indicios de heterocedasticidad.

Figura 1

Resultados Regresión Lineal Múltiple Original

OLS Regression Results

Dep. Variable:	Cost Parchment USD/Kg	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.742
Method:	Least Squares	F-statistic:	5950.
Date:	Mon, 18 Nov 2024	Prob (F-statistic):	0.00
Time:	14:56:13	Log-Likelihood:	-11764.
No. Observations:	24868	AIC:	2.355e+04
Df Residuals:	24855	BIC:	2.366e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.8271	0.175	21.878	0.000	3.484	4.170
Parchment Initial Kg	-2.431e-05	6.63e-06	-3.665	0.000	-3.73e-05	-1.13e-05
PECA State	0.0024	0.003	0.867	0.386	-0.003	0.008
Certificate	-0.0574	0.002	-31.443	0.000	-0.061	-0.054
Variety	-0.0042	0.001	-7.043	0.000	-0.005	-0.003
Quality	-0.0368	0.001	-25.259	0.000	-0.040	-0.034
Factor 14	-0.0742	0.003	-28.445	0.000	-0.079	-0.069
Factor 15	0.0290	0.002	18.918	0.000	0.026	0.032
Humidity Green	0.0038	0.007	0.524	0.601	-0.010	0.018
Department	-0.0081	0.002	-4.045	0.000	-0.012	-0.004
Municipality	0.0007	0.000	3.671	0.000	0.000	0.001
TRM	0.0002	7.51e-06	22.131	0.000	0.000	0.000
Apertura	0.9516	0.004	251.775	0.000	0.944	0.959

Por lo indicado anteriormente, se realiza la limpieza de los outliers identificados mediante la regresión lineal, en este punto se identifica 1.205 registros como outliers, lo que corresponde a aproximadamente el 4% de los 31.086 registros con los que se trabaja originalmente la regresión.

La eliminación de los identificados, demostró ser una decisión adecuada en comparación con otras posibles estrategias de tratamiento de valores atípicos. Este enfoque permitió mitigar de manera efectiva los efectos negativos de estos puntos extremos en la estabilidad del modelo, favoreciendo un ajuste más robusto sin introducir distorsiones que podrían surgir de alternativas como la transformación o el recorte de los datos.

Adicional a limpieza de los outliers, mediante la clase SequentialFeatureSelector de la librería sklearn se realizan iteraciones de los posibles modelos que pueden formar a partir de las variables indicadas, con el fin de identificar la combinación de variables que representen el comportamiento del precio pagado por Kg de café por parte de la compañía, obteniendo como resultados que las variables: Certificate, Variety, Quality, Factor 14, Factor 15, TRM y Apertura.

Una vez que se tienen definidas las variables con las que se planteará el modelo de regresión lineal final, se divide el conjunto de datos en un segmento de entrenamiento y un segmento de prueba, donde el primero equivale al 80% de la data y el segundo equivale al 20% restante. En los resultados obtenidos, se evidencia que todas las variables incluidas en el modelo son estadísticamente significativas para explicar el comportamiento del precio pagado por Kg de café, donde se explica aproximadamente el 88% del comportamiento, ver la Figura 2.

El coeficiente mas significativo es el que corrresponde a Apertura, que indica que por cad unidad de incremento en Apertura, el costo del pergamino aumenta en aproximadamente 1,009 USD/Kg. Esto sugiere que esta variable tiene un impacto directo y fuerte en el costo, siendo consistente con el análisis de correlación realizado inicialmente.

Figura 2

Resultados Regresión Lineal Múltiple Final

OLS Regression Results

Dep. Variable:	Cost Parchment USD/Kg	R-squared:	0.877
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	2.438e+04
Date:	Mon, 18 Nov 2024	Prob (F-statistic):	0.00
Time:	15:28:59	Log-Likelihood:	-1764.7
No. Observations:	23904	AIC:	3545.
Df Residuals:	23896	BIC:	3610.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.5563	0.104	43.810	0.000	4.352	4.760
Certificate	-0.0607	0.001	-53.789	0.000	-0.063	-0.058
Variety	-0.0012	0.000	-2.936	0.003	-0.002	-0.000
Quality	-0.0271	0.001	-27.733	0.000	-0.029	-0.025
Factor 14	-0.0662	0.002	-35.184	0.000	-0.070	-0.063
Factor 15	0.0107	0.001	9.155	0.000	0.008	0.013
TRM	0.0002	5.04e-06	32.307	0.000	0.000	0.000
Apertura	1.0088	0.003	394.250	0.000	1.004	1.014

El comportamiento de los residuos de este modelo de regresión, muestra que los supuestos de normalidad y homocedasticidad que anteriormente parecían incumplirse, presentan un comportamiento adecuado, como se muestra en el Apéndice B. De igual manera aplica el test de Shapiro-Wilk para comprobar la normalidad de los residuos, donde se tiene un resultado de p-value de 1,36e-13, por lo tanto se acepta la hipótesis de normalidad de los residuos.

Como paso final se computan las predicciones del modelo sobre el conjunto de datos de pruebas, sobre estas predicciones se calculan las métricas de evaluación que se consolidan en la Tabla 7. Se evidencia que en la base de pruebas el modelo es capaz de explicar aproximadamente el 88% de la variabilidad del precio pagado por Kg de café, siendo este valor ligeramente mayor al obtenido sobre el conjunto de entrenamiento, de igual manera el resultado del R^2 Ajustado es prácticamente igual al R^2, con lo que se puede inferir que las variables incluidas son relevantes y que no hay sobreajuste significativo en el modelo, reforzando la calidad de la selección de las variables para el modelo.

Tabla 7

Métricas de Evaluación de las Predicciones de la Regresión Lineal Múltiple

Métrica	Resultado
RMSE	0,2579
RSE	0,2589
MAE	0,2077
R^2	0,884
R^2 Ajustado	0,884

Nota. Métricas que explican el comportamiento de las predicciones sobre el conjunto de datos de pruebas

En relación a los errores, el modelo muestra un buen desempeño predictivo, considerando que en promedio, el modelo comete RMSE de aproximadamente USD 0,25. Este valor esta en torno al 4-7% del rango de precios pagados por la compañía por Kg de café comprado (entre 3,7 – 5,0 USD), por lo que esta cerca al rango de aceptación establecido del 5%. Adicionalmente, el RSE también es de aproximadamente USD 0,25 indicando que las predicciones están, en promedio, cerca de los valores reales, mostrando un bajo nivel de error residual.

Modelo de Series de Tiempo

El análisis de series de tiempo multivariado es fundamental para estudiar la relación entre el precio pagado por Kg café y otras variables económicas, como el precio internacional definido por el mercado C. En este contexto, se utilizará el modelo SARIMAX, que es adecuado para manejar datos de series temporales con componentes estacionales y variables externas que influyen en la dinámica del precio del café. Este enfoque es especialmente útil cuando se busca modelar el comportamiento de una variable dependiente mientras se incorpora la influencia de otras variables explicativas, como los precios internacionales (Hyndman & Athanasopoulos,

2018). A diferencia de otros métodos de series de tiempo multivariadas, como el VAR (Vector AutoRegressive), el SARIMAX permite integrar de manera más efectiva los componentes estacionales y las variables exógenas, evitando hacer un análisis de dos vías.

El uso de SARIMAX para el análisis de series de tiempo multivariadas tiene ventajas significativas frente a otros métodos. Mientras que el modelo VAR también permite manejar múltiples variables, no tiene la capacidad de modelar explícitamente las estacionalidades o incluir variables exógenas que puedan influir en las series dependientes. Por el contrario, el modelo SARIMAX puede incorporar tanto componentes estacionales como exógenos, lo que es crucial para el análisis del precio del café, ya que este se ve afectado no solo por su propia historia, sino también por factores externos como el precio del café en el mercado internacional. (Brockwell & Davis, 2013).

De igual forma, el modelo SARIMAX ofrece una flexibilidad adicional al incorporar tanto términos autorregresivos como de promedios móviles, lo que lo hace más adecuado para capturar la compleja dinámica temporal de las series de precios. Al integrar las influencias externas y estacionales de manera eficaz, se obtiene un modelo que no solo es capaz de realizar pronósticos más precisos, sino que también proporciona una interpretación clara de cómo las diferentes variables externas afectan la serie de tiempo principal (Hamilton, 2020).

Para este análisis temporal se realiza la transformación del conjunto de datos de compras, dado que en este conjunto de datos se pueden encontrar registros de más de una compra efectuada el mismo día, por lo tanto preparar los datos para el análisis de series temporales se calcula el promedio del precio pagado en cada una de las fechas por calidad de grano comprado, donde en caso de que en alguna fecha particular no se evidencien compras de una calidad en especifico, se asignará un valor de 0.

A este conjunto de datos se le incluye el valor del mercado C para cada fecha, definido en el campo apertura, tal como se observa en la Tabla 8, esto con base a lo evidenciado en el análisis realizado mediante la regresión lineal múltiple, donde se concluyó que esta es la característica que tiene una mayor con relación al precio pagado por Kg.

Tabla 8

Estructura Final Conjunto de Datos Para Series de Tiempo

	A	AA	AAA	ML	RTB	Apertura
6/04/2021	0,0000	3,3798	0,0000	0,0000	3,1267	2,7348
7/04/2021	3,0235	3,3444	0,0000	0,0000	3,1614	2,8329
8/04/2021	3,0235	3,3584	3,6041	0,0000	3,1741	2,8384
9/04/2021	2,9207	3,1048	3,6041	0,0000	3,1741	2,8638
10/04/2021	2,8773	3,2361	3,6041	0,0000	3,1790	2,8638

Nota. Estructura que se plantea para la aplicación del modelo de series temporales.

Una vez se tiene definido el conjunto de datos a trabajar, se calcula la correlación entre las variables con el fin de determinar la relación del precio del mercado C respecto al precio pagado para cada una de la calidades, donde se evidencia que mientras menor sea la calidad comprada (las calidades van de menor a mayor a menor de la siguiente forma: RTB, A, AA, AAA y ML), existe una mayor relación con el precio de bolsa definido internacionalmente, como se puede evidenciar en el Apéndice C.

En el Apéndice C se compila el comportamientos de los patrones de las series A, AA y RTB respecto a la variable Apertura.

Con base a la dinámica de compras del café con calidad AAA y ML, que suele darse con la oportunidad de compra ocasional de estas calidades y no con una dinámica constante como sucede con las demás calidades, sumado a la correlación que no es significativa para estas dos

calidades, se excluye del análisis de series temporales a las calidades AAA y ML, y se procede a analizar en parejas la relación existente entre el precio establecido mediante el mercado C (Apertura) y las calidades restantes que fueron compradas.

En el Apéndice C se compila los gráficos de comportamiento de los patrones de las series A, AA y RTB respecto a la variable Apertura.

Los patrones de las series RTB y Apertura están altamente alineados en todas las dimensiones, donde en ambos casos se evidencia una tendencia creciente desde 2021 hasta mediados de 2022, seguida de una tendencia decreciente que persiste hasta 2023, la cual es mas fuerte para el precio pagado por el café calidad RTB. De igual manera la estacionalidad, sugiere que ambas series tienen fluctuaciones cíclicas regulares con periodicidad aparente de un año, indicando estacionalidad anual.

Un comportamiento similar se presenta con los patrones de las series A y Apertura, y AA y Apertura , donde las series comparten estacionalidades similares a los presentadas anteriormente, pero se evidencia que las tendencias se van incrementando en las observaciones de compra de las calidades A y AA, lo que sugiere a mayor calidad el precio pagado por Kg por parte de la compañía el precio tiende a subir en mayor proporción a la que presenta el precio internacional, definido por el mercado C.

Posterior al análisis realizado con base a las gráficas del comportamiento de las observaciones, tendencias y estacionalidad, mediante un modelo SARIMAX se evaluará como Apertura afecta a cada una de las calidades de café compradas de manera independiente.

Con el fin de reducir la complejidad del modelo y optimizar el computo del mismo, se reduce la granularidad del conjunto de datos, esto mediante el promedio de un único dato por semana para cada una de las variables, considerando esto se establecerá como estacionalidad 52,

que hace referencia a la cantidad de semanas en un año. Adicionalmente se hará uso de la clave auto_arima, esta se encarga de identificar automáticamente los mejores parámetros (p, d y q para order y P, D, Q para seasonal_order), donde order define la estructura del modelo ARIMA, mientras que seasonal_order especifica la estructura estacional del modelo SARIMAX. (Brockwell & Davis, 2013).

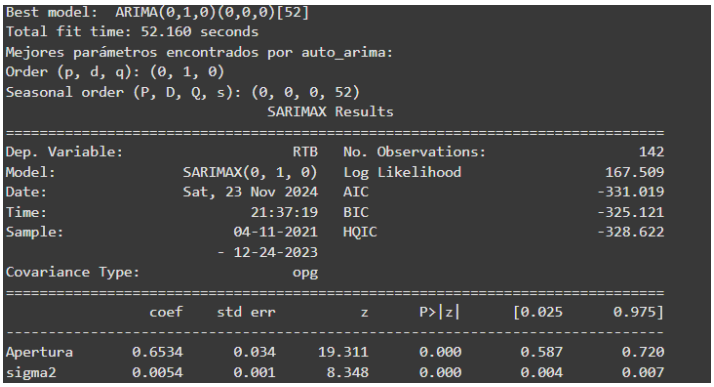
Los resultados del SARIMAX donde RTB es la variable dependiente y Apertura es la variable exógena, ver Figura 3, muestran que en los parámetros el modelo no encuentra patrones complejos o ciclos autorregresivos significativos ni a nivel de serie completa ni a nivel estacional. En esta serie se ha evidenciado una tendencia, la cual es eliminada por diferenciación, motivo por el que el parámetro d es igual a 1.

El coeficiente de Apertura indica que en promedio, un aumento de una unidad en el valor de Apertura está asociado con un aumento de 0,6534 en el precio de compra de la calidad RTB. El estadístico muestra que esta variable es altamente significativa, por hay evidencia fuerte de que Apertura tiene un impacto real en RTB.

Los resultados del modelo donde A es la variable dependiente y Apertura es la variable exógena, ver Figura 4, presentan resultados muy similares a los obtenidos en el modelo de RTB. Los parámetros generados por el método auto_arima siguen indicando que Apertura captura suficientemente bien las variaciones que afectan a A (al igual que a RTB), eliminando la necesidad de que el modelo dependa de patrones estacionales internos en la serie de tiempo.

Figura 3

Resultados SARIMAX Para RTB con Apertura Como Regresor



```
Best model:  ARIMA(0,1,0)(0,0,0)[52]
Total fit time: 52.160 seconds
Mejores parámetros encontrados por auto_arima:
Order (p, d, q): (0, 1, 0)
Seasonal order (P, D, Q, s): (0, 0, 0, 52)
                               SARIMAX Results
==============================================================================
Dep. Variable:                    RTB   No. Observations:                  142
Model:               SARIMAX(0, 1, 0)   Log Likelihood                 167.509
Date:                Sat, 23 Nov 2024   AIC                           -331.019
Time:                        21:37:19   BIC                           -325.121
Sample:                    04-11-2021   HQIC                          -328.622
                         - 12-24-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Apertura       0.6534      0.034     19.311      0.000       0.587       0.720
sigma2         0.0054      0.001      8.348      0.000       0.004       0.007
```

En promedio un aumento de una unidad en el valor de Apertura, incide en un aumento de 0,6777 en el precio de compra de la calidad A, de igual manera el estadístico establece que que el valor de Apertura es estadísticamente significativo para capturar el comportamiento del valor pagado por cada Kg de café de calidad A.

Finalmente los resultados del modelo donde AA es la variable dependiente -ver Figura 5-, presentan una diferencia sustancial respecto a los anteriores modelos, ya que se incluye un parámetro autorregresivo de orden, lo que indica que el precio pagado por el café calidad AA en una semana está influenciado por su propio valor rezagado de la semana inmediatamente anterior. El resto de parámetros generados son iguales a los sugeridos por el método auto_arima para los demás modelos.

Figura 4

Resultados SARIMAX Para A con Apertura Como Regresor

```
Best model:  ARIMA(0,1,0)(0,0,0)[52]
Total fit time: 41.505 seconds
Mejores parámetros encontrados por auto_arima:
Order (p, d, q): (0, 1, 0)
Seasonal order (P, D, Q, s): (0, 0, 0, 52)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      A   No. Observations:                  142
Model:               SARIMAX(0, 1, 0)   Log Likelihood                62.202
Date:                Sat, 23 Nov 2024   AIC                           -120.405
Time:                        21:38:44   BIC                           -114.507
Sample:                    04-11-2021   HQIC                          -118.008
                         - 12-24-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Apertura       0.6777      0.092      7.366      0.000       0.497       0.858
sigma2         0.0242      0.002     14.879      0.000       0.021       0.027
```

Los coeficientes muestran que respecto a Apertura se tiene una variación de aproximadamente 0,63 con cada dólar que incrementa el precio internacional del mercado C, mientras que se evidencia una disminución respecto a su propio valor de la semana inmediatamente anterior, que es prácticamente de 0,63 dólares. Esto sugiere un proceso de ajuste, donde cambios en una semana tienden a ser compensados en la siguiente semana, ver Figura 5.

Figura 5

Resultados SARIMAX Para AA con Apertura Como Regresor

```
Best model:  ARIMA(1,1,0)(0,0,0)[52]
Total fit time: 18.206 seconds
Mejores parámetros encontrados por auto_arima:
Order (p, d, q): (1, 1, 0)
Seasonal order (P, D, Q, s): (0, 0, 0, 52)
                               SARIMAX Results
==============================================================================
Dep. Variable:                     AA   No. Observations:                  142
Model:               SARIMAX(1, 1, 0)   Log Likelihood                64.823
Date:                Sat, 23 Nov 2024   AIC                           -123.645
Time:                        22:22:08   BIC                           -114.799
Sample:                    04-11-2021   HQIC                          -120.050
                         - 12-24-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Apertura       0.6309      0.060     10.442      0.000       0.513       0.749
ar.L1         -0.6283      0.052    -12.174      0.000      -0.729      -0.527
sigma2         0.0233      0.002     13.069      0.000       0.020       0.027
```

Respecto a las predicciones generadas para cada uno de los modelos de series de tiempo, se generán los graficos que comparan el comportamiento de las observaciones del conjunto de datos respecto a la predicción generada a través de SARIMAX.

Los gráficos del forecast que se encuentran en el Apéndice C, evidencian que las predicciones generadas para la calidad RTB son las que se ajustan con una menor desviación respecto a las observaciones del conjunto de datos, mientras que las predicciones de la calidad A y AA aunque mantienen una estructura similar a las observaciones, presentan un desfase hacia la derecha, lo que implica que el valor que se predice suele corresponder a un valor muy cercano a las observaciones inmediatamente anteriores, lo que puede llevar a una toma de decisiones erróneas..

Como se puede ver en la Tabla 9, los coeficientes de los tres modelos de series de tiempo son estadísticamente significativos, aun con este resultado es necesario validar el comportamiento de los residuos

Tabla 9

Resultado de los Residuos de los Modelos de Series de Tiempo

Métrica	RTB	A	AA
Sigma² (varianza del error)	0,0054	0,0242	0,0233
P Sigma²	0,0000	0,0000	0,0000
Ljung-Box (Q)	6,2300	2,9300	6,9900
Prob (Q)	0,0100	0,0900	0,0100
Jarque-Bera (JB)	0,1600	117,9600	61,1800
Prob (JB)	0,9200	0,0000	0,0000
Heteroskedasticity (H)	1,1200	1,9800	3,6300
Prob (H)	0,7000	0,0200	0,0000

Nota. Métricas que explican el comportamiento de los errores en las series de tiempo.

Se evidencia que para el modelo donde la variable dependiente es RTB, Aunque el modelo presenta baja varianza residual, residuos normales y homocedasticidad, presenta una autocorrelación residual significativa definida por el test de Ljung-Box. Esto indica que los errores no son independientes y podrían reflejar patrones sistemáticos no capturados por el modelo actual. Esto puede deberse a que el modelo actual no no captura adecuadamente la estructura temporal de los datos. Este problema puede abordarse incrementando el valor del parámetro p del modelo ARIMA, que es el elemento que captura el comportamiento autorregresivo de la variable dependiente.

Mientras que para los modelos planteados para la calidad A y AA se presentan problemas de no normalidad de los residuos (basado en el test de Jarque-Bera) y problemas de heterocedasticidad, lo que significa que la varianza de los residuos no es constante a lo largo del tiempo. Considerando esta situación, existe la posibilidad de reevaluar los modelos en este caso empleando modelos GARCH.

Mientras GARCH modela la varianza condicional de una serie temporal, siendo útil para capturar la heterocedasticidad en datos financieros como retornos de inversión o precios de activos (Engle, 1982), SARIMAX se centra en modelar la media condicional, incorporando componentes estacionales y variables exógenas, lo que lo hace adecuado para series con patrones periódicos o dependientes de factores externos (Hyndman & Athanasopoulos, 2018).

Se evidencia que para el modelo donde la variable dependiente es RTB, Aunque el modelo presenta baja varianza residual, residuos normales y homocedasticidad, presenta una autocorrelación residual significativa definida por el test de Ljung-Box. Esto indica que los errores no son independientes y podrían reflejar patrones sistemáticos no capturados por el modelo actual. Esto puede deberse a que el modelo actual no no captura adecuadamente la

estructura temporal de los datos. Este problema puede abordarse incrementando el valor del parámetro *p* del modelo ARIMA, que es el elemento que captura el comportamiento autorregresivo de la variable dependiente.

Mientras que para los modelos planteados para la calidad A y AA se presentan problemas de no normalidad de los residuos (basado en el test de Jarque-Bera) y problemas de heterocedasticidad, lo que significa que la varianza de los residuos no es constante a lo largo del tiempo. Considerando esta situación, existe la posibilidad de reevaluar los modelos en este caso empleando modelos GARCH.

Mientras GARCH modela la varianza condicional de una serie temporal, siendo útil para capturar la heterocedasticidad en datos financieros como retornos de inversión o precios de activos (Engle, 1982), SARIMAX se centra en modelar la media condicional, incorporando componentes estacionales y variables exógenas, lo que lo hace adecuado para series con patrones periódicos o dependientes de factores externos, tal como indican (Hyndman & Athanasopoulos, 2018).

Conclusiones

En esta investigación se logró segmentar de manera efectiva las compras de lotes de café realizadas por la compañía entre 2021 y 2023, identificando patrones y las variables más influyentes en la determinación del precio pagado. Utilizando técnicas de estadística descriptiva, se analizó la composición del conjunto de datos y se determinó que los lotes secados mediante un proceso de lavado y convertidos en pergamino seco representaban aproximadamente el 98% del total de las compras, por lo que se decidió enfocar el análisis en esta categoría predominante.

Con el objetivo de estudiar el comportamiento del precio pagado, se realizó una depuración de datos para eliminar valores atípicos. Posteriormente, a través de la correlación de Pearson, se identificó que el precio de apertura del mercado C es la variable más fuertemente relacionada con el precio pagado por la compañía. Para profundizar este hallazgo, se desarrolló un modelo de regresión lineal múltiple, confirmando que el precio de apertura del mercado C tiene un impacto estadísticamente significativo en el precio pagado por kilogramo. Adicionalmente, este análisis reveló que las ubicaciones geográficas de las bodegas donde se adquieren los lotes no presentan una influencia significativa, sugiriendo que el precio pagado es consistente entre diferentes puntos de compra.

Las variables más influyentes identificadas fueron: el precio del mercado C, la TRM, los factores de rendimiento, la calidad, la variedad y el certificado del café comprado. Dado el carácter histórico de los datos, se planteó un análisis de series temporales, evaluando el precio promedio semanal pagado por cada calidad en relación con el precio del mercado C en el mismo periodo, incluyendo las variables con mayor significancia según la regresión lineal múltiple.

La aplicación del modelo SARIMAX para series temporales indicó que, para la calidad RTB (el café de menor calidad adquirido por la compañía), el precio de apertura del mercado C

tiene una incidencia altamente significativa. Sin embargo, los resultados sugieren que el modelo puede mejorarse incorporando una evaluación de los ciclos autorregresivos del precio de la calidad RTB. En contraste, los modelos para las calidades A y AA mostraron problemas con los supuestos de homoscedasticidad, independencia y normalidad de los residuos, lo que impide realizar inferencias y predicciones válidas. Por ello, se recomienda explorar el uso de modelos GARCH, que son más adecuados para capturar la heterocedasticidad inherente en estos datos.

En conclusión, esta investigación permitió identificar las variables con mayor significancia estadística para explicar el precio pagado por cada kilogramo de café, así como comprender la variabilidad de estos precios a lo largo del periodo evaluado. Estos hallazgos ofrecen una base para optimizar las decisiones de compra de la compañía y sugieren oportunidades de mejora mediante el uso de modelos más avanzados en futuros análisis.

Recomendaciones

Con base en los resultados de esta investigación, se recomienda analizar de manera independiente los datos que no fueron considerados en este estudio, como los lotes de café procesados con métodos de secado diferentes al lavado y aquellos registros identificados como atípicos en el precio pagado por kilogramo. Para analizar el conjunto de datos sin eliminar outliers, se podrían emplear técnicas como la transformación de variables a través de logaritmos, raíces cuadradas, o potencias inversas, o la asignación de pesos para reducir su impacto en los resultados. Este análisis permitiría ampliar la cobertura al 100% de los datos disponibles, lo cual proporcionaría una visión más completa y fundamentada para la toma de decisiones en todos los posibles escenarios.

Además, se sugiere mejorar la precisión y eficacia de los modelos implementados mediante el escalado de las variables categóricas utilizando métodos como OneHotEncoder, en lugar del LabelEncoder aplicado en este estudio. Este cambio transformaría las variables de forma equitativa, evitando asignar una importancia indebida a determinadas categorías, como puede suceder con LabelEncoder. Sin embargo, es importante tener en cuenta que este método puede incrementar la carga computacional del modelo y añadir complejidad a la interpretación de los resultados, aspectos que deben ser evaluados cuidadosamente.

Para profundizar en el análisis temporal del precio pagado por la compañía, se recomienda incluir en los modelos parámetros que permitan considerar los ciclos autorregresivos tanto en la variable dependiente (el precio pagado por cada calidad de café) como en la variable exógena (el precio del mercado C). Esto contribuiría a reducir el impacto de la autocorrelación residual y mejorar la precisión de las predicciones.

Finalmente, para abordar los problemas de heterocedasticidad observados en los modelos de las calidades A y AA, se sugiere la implementación de modelos basados en GARCH**.** Estos modelos son especialmente útiles para capturar y analizar patrones en la variabilidad de los precios, ofreciendo una herramienta robusta para comprender mejor su comportamiento y apoyar estrategias de optimización.

En resumen, estas recomendaciones buscan fortalecer el análisis realizado, mejorar la capacidad predictiva de los modelos y ampliar el alcance del estudio para beneficiar la toma de decisiones estratégicas de la compañía.

Referencias

Auffarth, B. (2021). *Machine Learning for Time-Series with Python: Forecast, predict, and detect anomalies with state-of-the-art machine learning methods.* Packt Publishing Ltd.

Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., . . . Morales, C. (2015). The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions. *Food Security, 7*(2), 303-321. doi:https://doi.org/10.1007/s12571-015-0446-9

Bastianin, A., Lanza, A., & Manera, M. (2018). Economic impacts of El Niño Southern Oscillation: evidence from the Colombian coffee market. *Munich Personal RePEc Archive, 49*(5), 623-633. https://mpra.ub.uni-muenchen.de/89984/

Bittencourt, M., & Agudelo, P. (2021). Impactos de la volatilidad cambial del comercio colombiano con sus principales socios comerciales. *EconoQuantum*, 57-81. doi:https://doi.org/10.18381/eq.v18i2.7209

Boydell, H. (2020). *Precio C: ¿Deberíamos fijar de otra manera el precio del café?* . Perfect Daily Grind Español: https://perfectdailygrind.com/es/2018/11/06/precio-c-deberiamos-fijar-de-otra-manera-el-precio-del-cafe/

Brockwell, P., & Davis, R. (2013). *Introduction to Time Series and Forecasting.* Springer Science & Business Media.

Cerquera, O., Pérez, V., & Sierra, J. (2020). Análisis de la competitividad de las exportaciones del café del Huila. *Tendencias, 21*(2), 19-44. doi:https://doi.org/10.22267/rtend.202102.139

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica, 50*(4), 987. doi:https://doi.org/10.2307/1912773

Field, A. (2009). *Discovering statistics using SPSS.* SAGE Publications.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications.

Hamilton, J. (2020). *Time Series analysis.* Princeton University Press.

Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Kilimanjaro Specialty Coffees. (s.f.). *¿Qué significa un precio «C» del café por sobre los 200 cts/lb?* https://www.coffeekilimanjaro.com/educacion/que-significa-un-precio-c-del-cafe-por-sobre-los-200-ctslb#:~:text=Qu%C3%A9%20es%20el%20precio%20%22C,de%20caf%C3%A9%20a%20nivel%20global

Monsalve, O. (2016). Un aplicativo en Excel para la enseñanza y modelación de una serie de tiempo: Precio externo del café colombiano. Universidad Tecnológica de Pereira.

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis.* John Wiley & Sons.

Murphy, K. (2012). *Machine learning: A Probabilistic Perspective.* MIT Press.

Ordoñez-Jurado, H., & Castillo-Marín, J. (2021). Evaluation of the sustainability in coffee production systems (Coffea Arabica L) in La Unión, Nariño, Colombia. *Revista de Ciencias Agrícolas, 1,* 110-122. doi:https://doi.org/10.22267/rcia.223901.177

Romero, R. (2020). Impact of economic internationalization Policies in Colombia, Peru and Chile. *Cuadernos de Administración, 36*(66), 78-91. doi:https://doi.org/10.25100/cdea.v36i66.8516

Shumway, R., & Stoffer, D. (2017). *Time Series Analysis and Its Applications: With R Examples.* Springer.

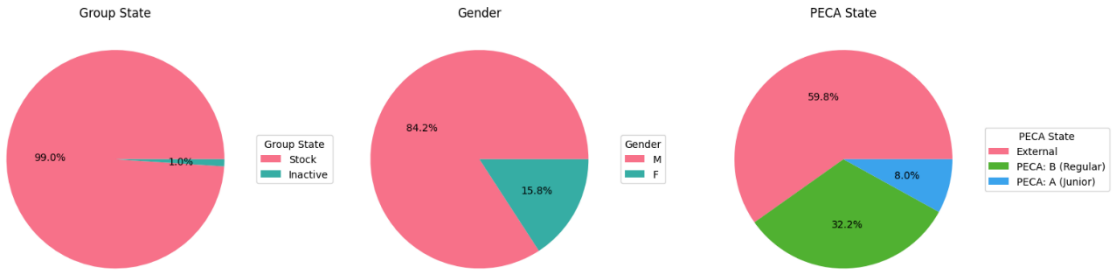
Urdan, T. (2017). *Statistics in Plain English.* Routledge.

Velásquez , J., & Aldana , D. (2007). Modelado del precio del café colombiano en la bolsa de nueva york usando redes neuronales artificiales. Universidad Nacional de Colombia - Sede Medellín.

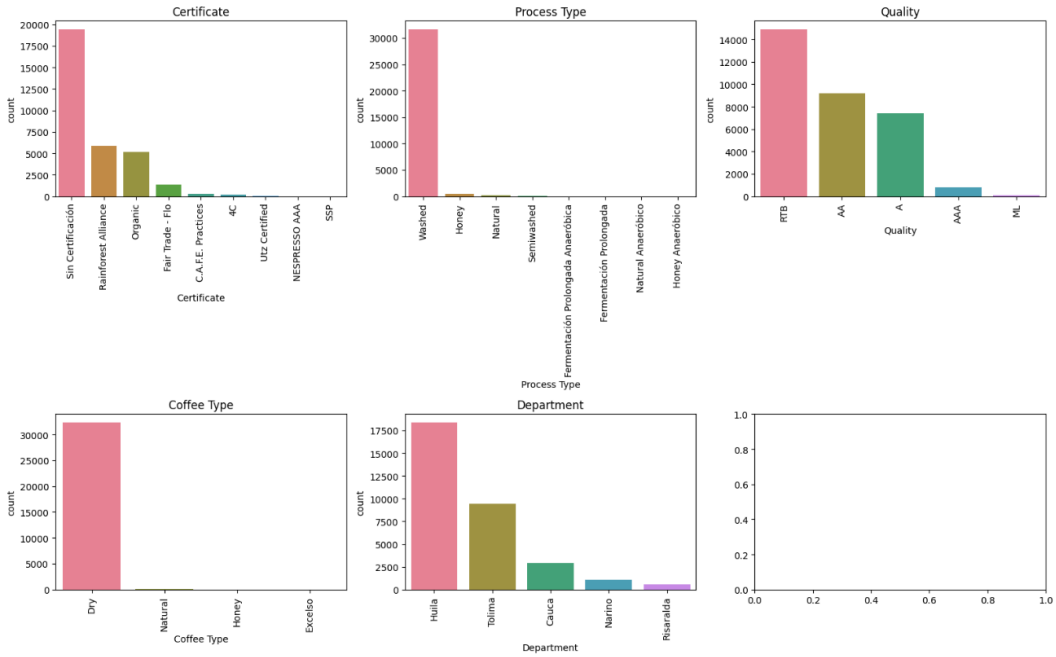
Apéndices

Apéndice A

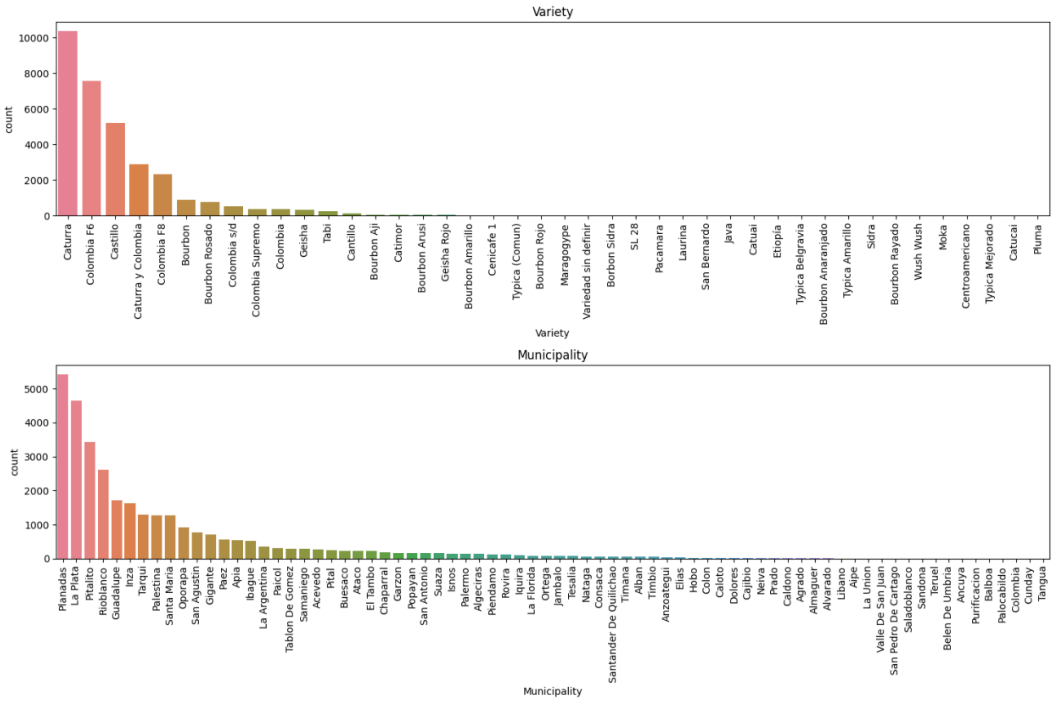
Preprocesamiento del Conjunto de Datos de Compras



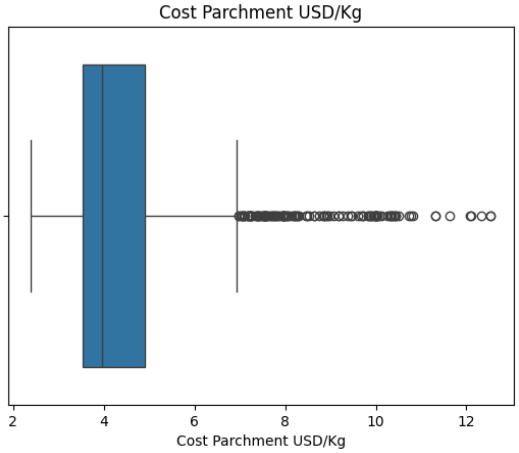
Nota. El gráfico representa la distribución de variables con máximo 3 categorías



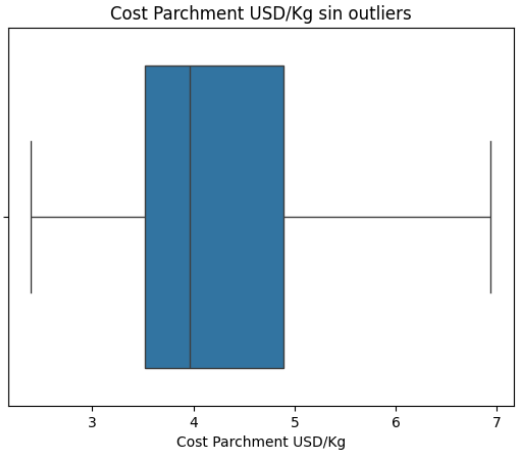
Nota. El gráfico representa la distribución de variables con máximo 10 categorías



Nota. El gráfico representa la distribución de variables con más de 10 categorías



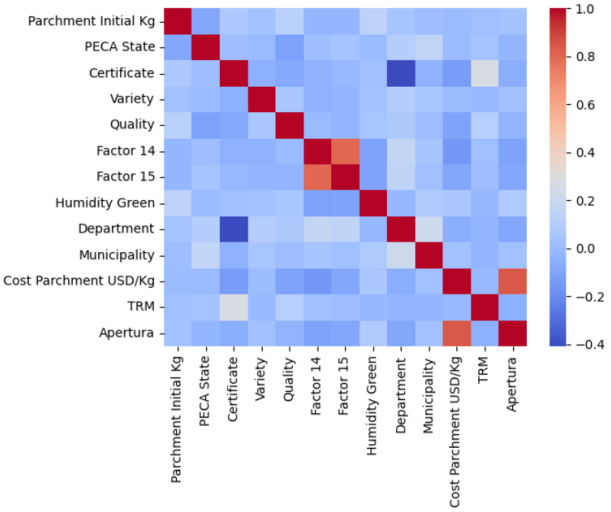
Nota. El gráfico representa la distribución de de la variable Cost-Parchmetn USD/Kg sin realizar ningún preproceamiento.



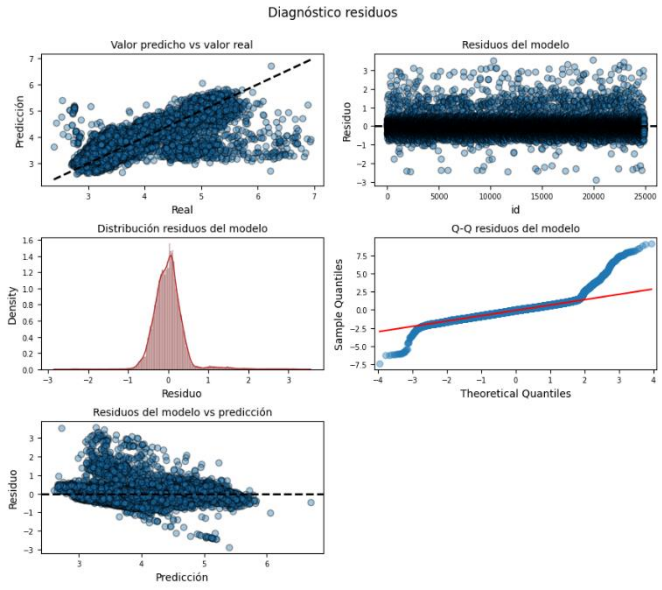
Nota. El gráfico representa la distribución de de la variable Cost-Parchmetn USD/Kg una vez se aplica la eliminación de los outliers mediante el método intercuartílico.

Apéndice B

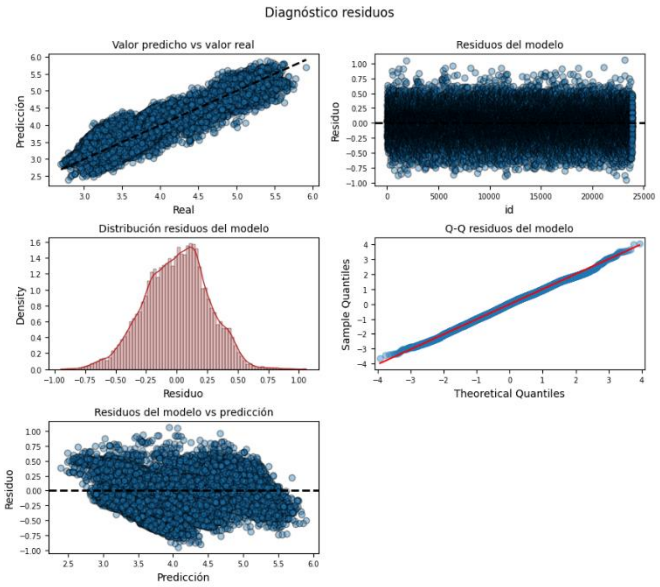
Gráficos de la Aplicación del Modelo de Regresión Lineal Múltiple



Nota. Correlación de las variables previo a realizar el análisis de regresión lineal.



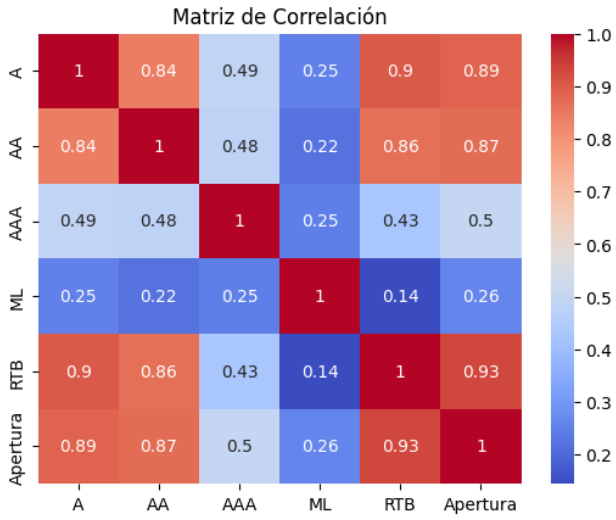
Nota. El gráfico muestra el comportamiento de los residuos obtenidos en la regresión original, evidenciando que hay serios problemas con el cumplimiento de los supuestos de regresión.



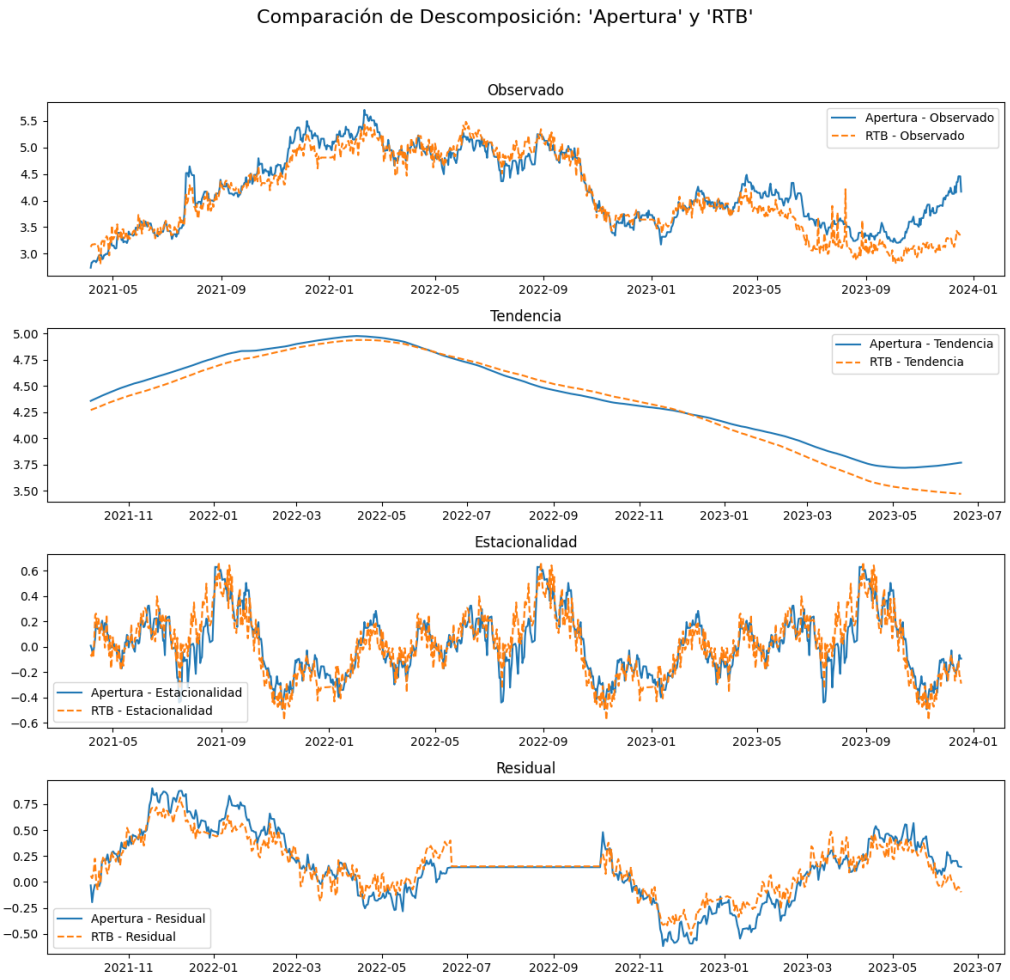
Nota. El gráfico muestra el comportamiento de los residuos obtenidos en la regresión final, donde a diferencia de la regresión original, los residuos tienen un comportamiento que se ciñe a los supuestos que debe cumplir una regresión.

Apéndice C

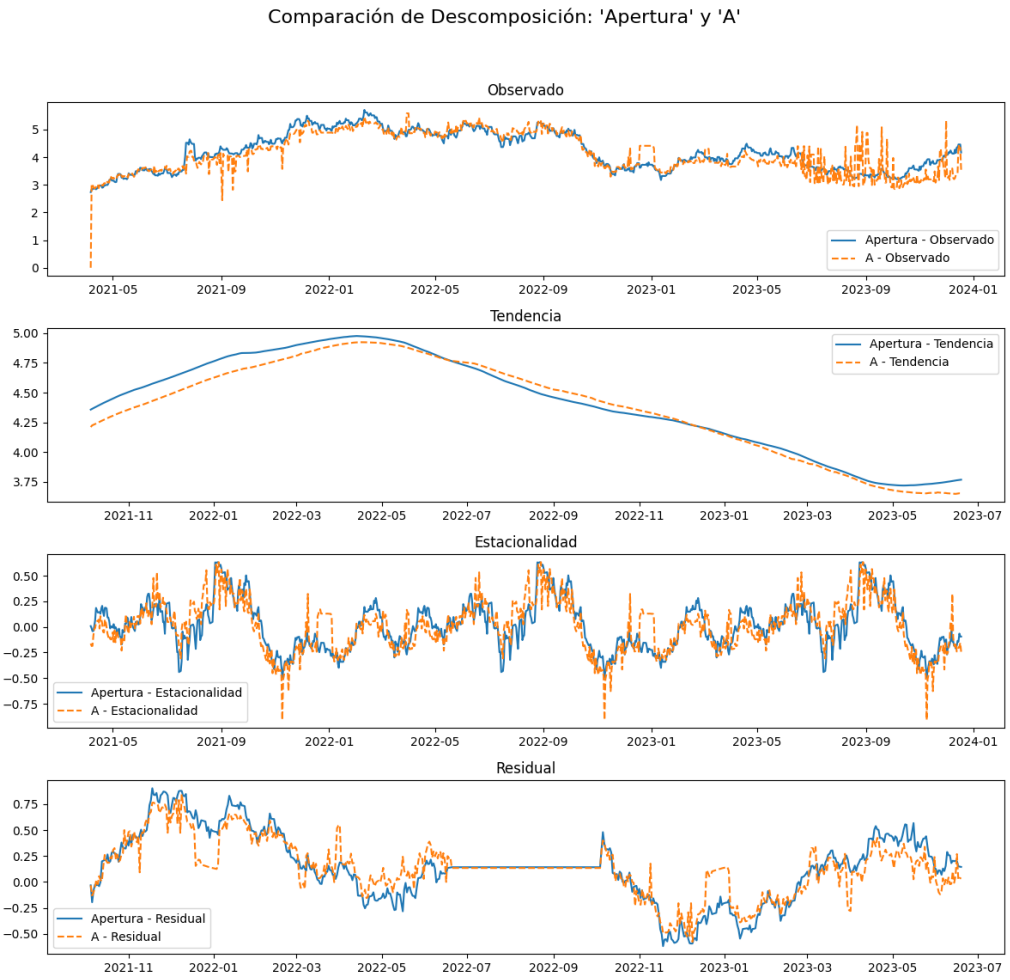
Gráficos de la Aplicación del Modelo de Series Temporales



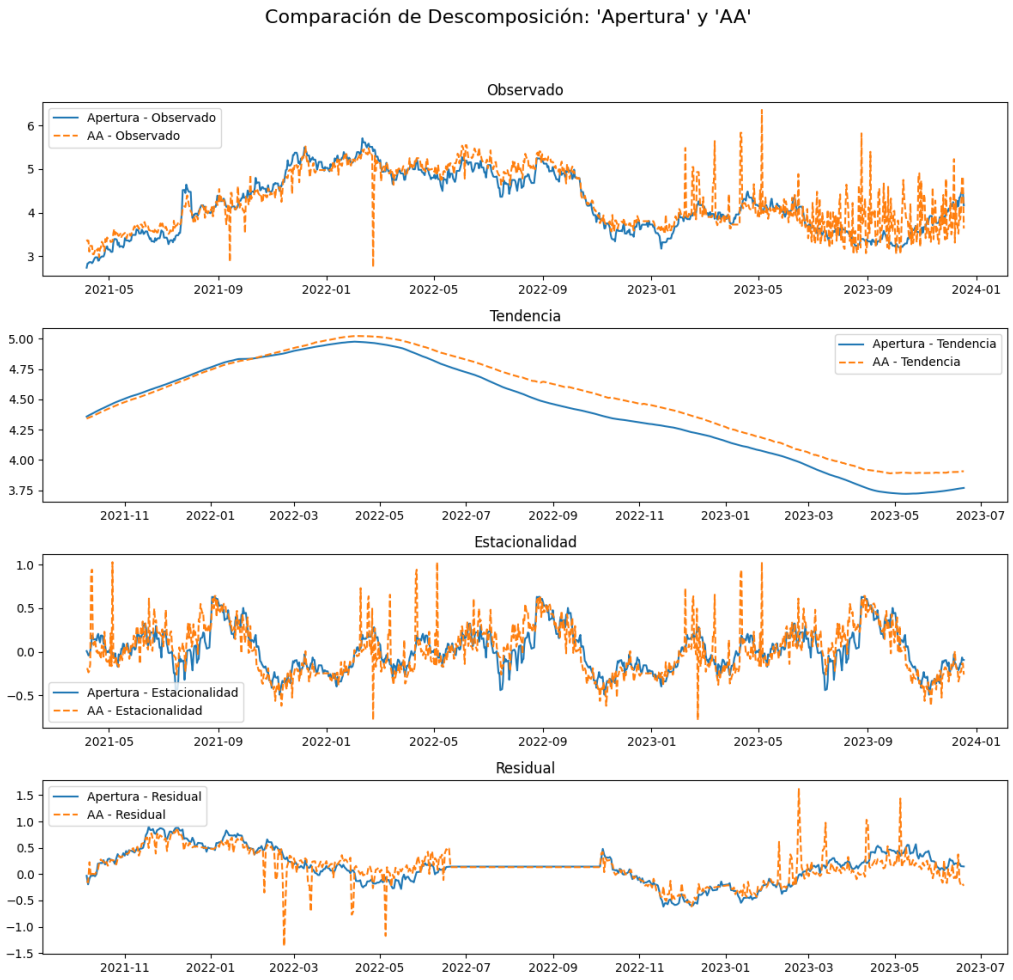
Nota. Correlación de las variables que se emplean para el desarrollo del modelo de series temporales.



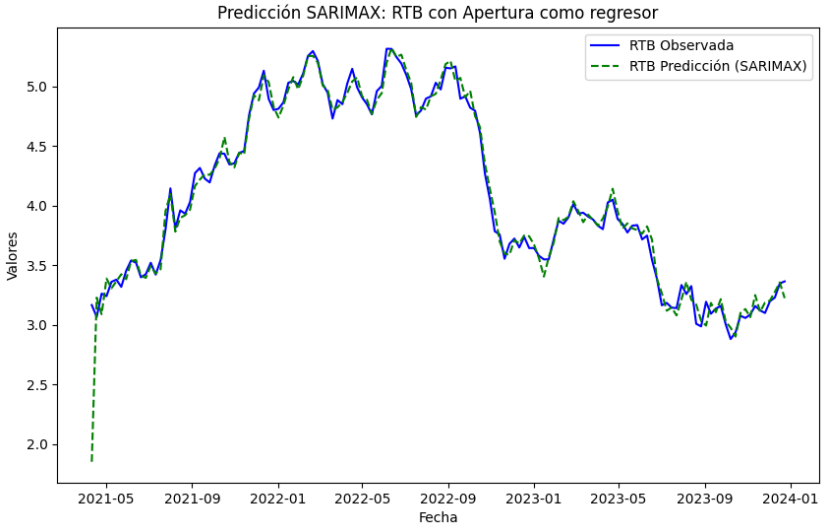
Nota. El gráfico muestra el la descomposición de Apertura y RTB, representando las observaciones, la tendencia existente, la estacionalidad y el comportamiento de los residuos.



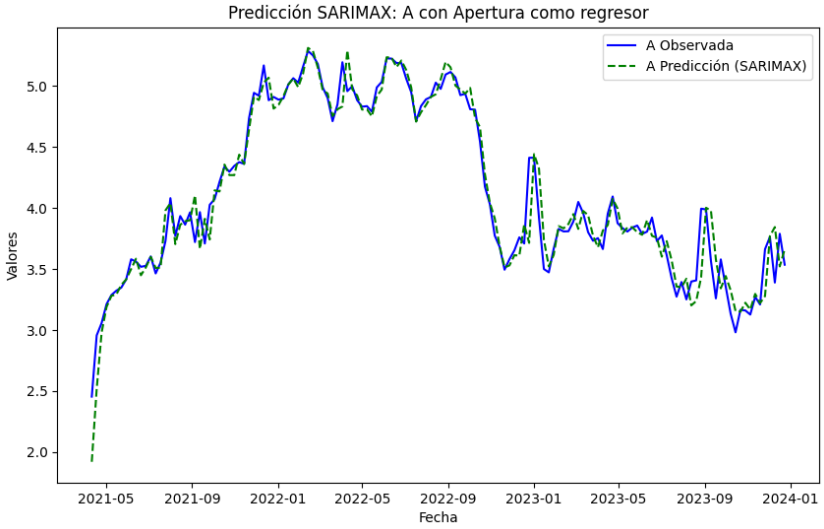
Nota. El gráfico muestra el la descomposición de Apertura y A, representando las observaciones, la tendencia existente, la estacionalidad y el comportamiento de los residuos.



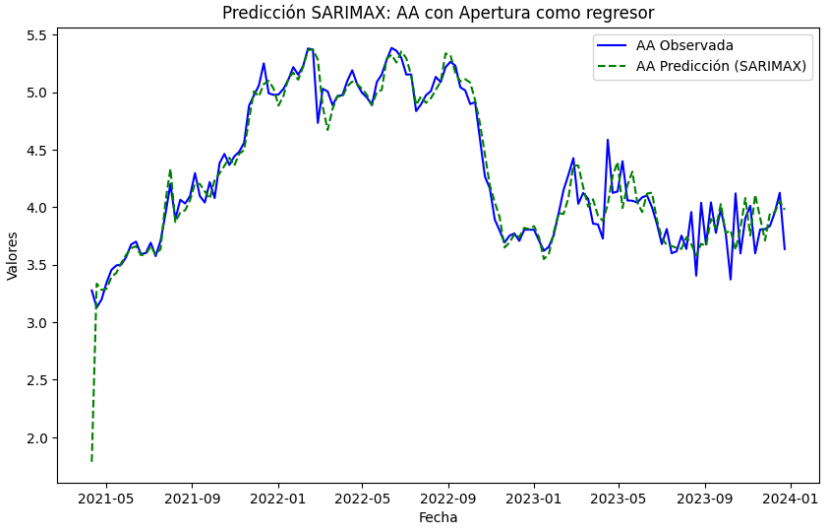
Nota. El gráfico muestra el la descomposición de Apertura y AA, representando las observaciones, la tendencia existente, la estacionalidad y el comportamiento de los residuos.



Nota. El gráfico muestra el ajuste que tiene las predicciones generadas de la aplicación del modelo de series temporales con SARIMAX para RTB con Apertura como regresor, con respecto a las observaciones reales



Nota. El gráfico muestra el ajuste que tiene las predicciones generadas de la aplicación del modelo de series temporales con SARIMAX para A con Apertura como regresor, con respecto a las observaciones reales



Nota. El gráfico muestra el ajuste que tiene las predicciones generadas de la aplicación del modelo de series temporales con SARIMAX para AA con Apertura como regresor, con respecto a las observaciones reales

Apéndice D

Script de Python con el Desarrollo de los Modelos

https://drive.google.com/file/d/1z6hJ9c4gubzzFSxe4WKqOwpADPMYUEN4/view?usp=sharing