Un modelo de pronóstico para la volatilidad de los futuros del café colombiano basado en noticias antes y durante la pandemia de la COVID-19 y variables económicas.

Daniel Alejandro Corredor León y Miguel Ángel Sanabria Núñez

Trabajo de grado para optar por el título de Ingeniero Industrial

Director

Henry Lamos Díaz

Ph.D. en Matemáticas – Física.

Codirector

David Esteban Puentes Garzón

M.Sc. en Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingenierías Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2023

Dedicatorias

A Dios quien con sus bendiciones me ha dado salud, fortaleza y entendimiento.

A mis padres, quienes siempre fueron mi apoyo en los momentos difíciles, con sus palabras y consejos siempre me dieron el impulso y fuerza necesaria para seguir adelante.

A Alejandra, quien con su amor incondicional me acompañó durante todo este proceso, con el cariño, paciencia y palabras correctas.

A mi hermana, porque siempre estuvo pendiente de mí, escuchándome y dándome sus mejores deseos.

Al profe David, quien con amabilidad siempre ofreció su tiempo para ayudarnos, guiarnos y aconsejarnos.

Daniel Alejandro Corredor León

A mis padres quienes me impulsan a ser mejor cada día.

A mi compañero de tesis, quien me hacía reaccionar cuando pensaba que no podía continuar.

Y, finalmente, a los que no creyeron en mí, con su actitud lograron que tomará más impulso.

Miguel Ángel Sanabria Núñez

Agradecimientos

En primer lugar, a Dios, quien fue el guía. A nuestros padres y familia, que nos acompañaron en este proceso universitario, gracias por creer en nosotros y apoyarnos.

A la Universidad Industrial de Santander, la cual nos abrió sus puertas para forjarnos como profesionales, con sentido crítico y humano.

A la Escuela de Estudios Industriales y Empresariales y a todos los docentes que con sus conocimientos aportaron en nuestra formación como ingenieros.

Al Grupo de Investigación Ópalo, quienes nos abrieron sus puertas, dándonos herramientas fundamentales para la investigación, ofreciéndonos su ayuda siempre con la mejor disposición.

Al profesor Henry Lamos, por admitir dirigir nuestro proyecto, gracias por el tiempo y brindarnos su experiencia en el ámbito científico, fue fundamental para llevar a feliz término este trabajo.

A nuestro codirector, David Esteban Puentes Garzón, quien siempre fue guía y luz en cada etapa del proyecto, siendo un maestro y contribuyendo tanto en nuestra formación profesional como personal. Para él toda nuestra admiración. Siempre lo llevaremos en nuestros corazones.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Apéndices

Los apéndices están adjuntos y pueden visualizarse en la base de datos de la biblioteca UIS

Apéndice A. Herramienta computacional para visualización.

Apéndice B. Artículo científico.

Apéndice C. Conjuntos de datos.

Apéndice D. *Scripts* con el código en el lenguaje de programación Python.

Resumen

Título: Un modelo de pronóstico para la volatilidad de los futuros del café colombiano basado en noticias antes y durante la pandemia de la COVID-19 y variables económicas[*]

Autores:

Daniel Alejandro Corredor León, Miguel Ángel Sanabria Núñez[**]

Palabras clave: Aprendizaje Automático, Pronóstico, Precio Del Café Colombiano, COVID-19, Análisis De Sentimientos, Volatilidad.

Descripción:

La crisis económica producida por la pandemia de la COVID-19 ha tenido efectos negativos a nivel mundial, provocando incertidumbre y volatilidad en los mercados, desde las negociaciones de materias primas hasta en los precios de productos agrícolas. De igual manera, la rápida difusión de la información relacionada con la pandemia, sin comprobar su veracidad, ha aumentado el pánico de los inversores. En este estudio se examina si las Variables Económicas (VE) y las noticias aportan valor predictivo a la volatilidad de los futuros del café colombiano antes y durante la pandemia. Por lo que se plantean cuatro modelos de aprendizaje automático (Random Forest, Extreme Gradient Boosting, Extreme Learning Machine y Long Short-Term Memory), junto a variables exógenas con observaciones diarias divididas en cuatro y ocho conjuntos de datos en el evento antes y durante la pandemia respectivamente. Lo anterior, con el fin de evaluar qué modelo tiene mejor capacidad de generalización y que conjunto de datos le aporta predicciones más precisas. Los resultados experimentales muestran que antes de la pandemia las variables económicas en promedio produjeron mejores previsiones, mientras que durante la pandemia del nuevo coronavirus las noticias cobraron mayor relevancia, proporcionando mejoras en la predicción. El aporte de esta investigación, gira en torno a contribuir a la extensa literatura que existe alrededor de predicción de series temporales y para que entes gubernamentales, productores, administradores e inversionistas puedan comprender mejor el comportamiento de la volatilidad del

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Diaz, Ph.D. en Física, Matemáticas. Codirector: David esteban puentes Garzón, M. Sc. Ingeniería Industrial.

café colombiano antes y durante la pandemia de la COVID-19 , anticipándose a periodos de alta incertidumbre.

Abstract

Project tittle: A forecasting model for the volatility of Colombian coffee futures based on news before and during the COVID-19 pandemic and economic variables [*]

Authors:

Daniel Alejandro Corredor León, Miguel Ángel Sanabria Núñez[**]

Keywords: Machine Learning, Forecasting, Colombian Coffee Price, COVID-19, Sentiment Analysis, Volatility.

Description:

The economic crisis caused by the COVID-19 pandemic has had negative effects worldwide, causing uncertainty and volatility in the markets, from the negotiation of raw materials to the prices of agricultural products. Similarly, the rapid dissemination of information related to the pandemic, without verifying its veracity, has increased investor panic. This study examines whether Economic Variables (EV) and news provide predictive value to the volatility of Colombian coffee futures before and during the pandemic. Therefore, four machine learning models (Random Forest, Extreme Gradient Boosting, Extreme Learning Machine and Long Short-Term Memory) are proposed, together with exogenous variables with daily observations divided into four and eight data sets in the node before and during the pandemic, respectively. The above, to evaluate which model has better generalization capacity and which data set provides more accurate predictions. The experimental results show that before the pandemic the economic variables on average produced better forecasts, while during the pandemic of the new coronavirus the news became more relevant, providing improvements in the prediction. The contribution of this research revolves around contributing to the extensive literature that exists around time series prediction and so that governmental entities, producers, administrators, and investors can better understand the behavior

[*] Bachelor's degree

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Diaz, Ph.D. en Física, Matemáticas. Codirector: David esteban puentes Garzón, M. Sc. Ingeniería Industrial.

of the volatility of Colombian coffee before and during the COVID-19 pandemic, anticipating periods of high uncertainty.

Introducción.

El sector primario de la economía colombiana aporta una gran participación del PIB nacional, con una contribución del 14.1%, donde el café es uno de los productos con mayor participación en este rubro, en cuanto a exportaciones para 2021 este commodity ocupó el tercer lugar con un 7.7%, viéndose superado solamente por el petróleo y la hulla con 27.1% y 10.6% respectivamente (Ministerio de Comercio, Industria y Turismo, 2022). Asimismo, en la actualidad el sector cafetero afronta periodos de incertidumbre, según el último informe de la Federación Nacional de Cafeteros las variables que rigen el precio del café han tenido gran volatilidad, principalmente el precio del contrato C[1] que tuvo un incremento del 28%, es decir, unas 142.2 ¢/lb con máximos no vistos desde 2014 de 207.8 ¢/lb para julio de 2021 y las tasas de cambio que para 2021 tuvieron fluctuaciones del 1,6% respecto 2020 llegando a máximos históricos del grano de $1.905.000 por carga de 125 kg (Federación Nacional de Cafeteros, 2021).

En consecuencia, es evidente la fortaleza de la moneda estadounidense frente al peso colombiano y las repercusiones que tiene en la economía nacional, esta dinámica es explicada por políticas monetarias y tasas de interés por parte de la Reserva Federal de Estados Unidos para hacerle frente al impacto de la COVID-19 sobre la economía de dicho país (Federación Nacional de Cafeteros, 2021). En este orden de ideas, la pandemia por la COVID-19 está teniendo un papel fundamental en la toma de decisiones de gobiernos, inversores y la economía mundial, debido al riesgo e incertidumbre que esta genera, además del rol preponderante que han tenido las noticias

[1] El contrato C es un contrato de futuros donde se fija el precio de referencia mundial de los cafés arábigos lavados siendo un contrato donde el valor es tomado como referencia para el precio internacional del café colombiano y otras variedades de arábiga que se negocian en la New York Board of Trade (NYBOT). Estos precios se caracterizan por ser volátiles y responden no sólo a los factores de oferta y demanda, sino también a la actividad de agentes especuladores en los mercados financieros siendo el principal determinante del precio del café a nivel global.

alrededor de esta enfermedad provocando aumento de la inseguridad pública, compras de pánico, rumores infundados, noticias falsas que acrecentaron el pánico y grandes fluctuaciones en los precios de los productos agrícolas (Liu et al., 2022). Además, las noticias asociadas con la COVID-19 están relacionadas a la volatilidad de los mercados bursátiles (Weng et al., 2021).

Por lo anterior, en la actualidad las problemáticas mundiales causadas por la pandemia de la COVID-19 han generado gran interés en estudiar las diferentes capaces de hacer previsiones a partir de factores y variables que puedan influir en un resultado, así mismo la aplicación de métodos adecuados, modelando fenómenos y de este modo facilitar la toma de decisiones. En cuanto a modelos de pronóstico a partir de técnicas de aprendizaje automático diversos autores han propuesto modelos no lineales ideales para la predicción del precio y/o volatilidad de diferentes productos agrícolas. Estos estudios se enfocan en su mayoría en el uso de modelos estadísticos para el análisis de series temporales, donde se emplean técnicas de aprendizaje automático para la previsión de commodities donde se incluyen productos agrícolas como café, tubérculos, soya, aceite de soya, maíz, aceite de palma y cacao.

Por lo tanto, el presente proyecto busca pronosticar la volatilidad de los futuros del café colombiano, por medio de técnicas de aprendizaje automático con la herramienta de programación Python, permitiendo el desarrollo, análisis y selección de los modelos apropiados para la previsión de los futuros del café colombiano, asimismo, teniendo en cuenta variables económicas que afectan el mercado de valores, la influencia que ha tenido la pandemia de la COVID-19 y las noticias relacionadas con esta enfermedad en la volatilidad del precio del café colombiano.

Finalmente, este trabajo busca servir como referencia para próximos proyectos que quieran abordar y explorar pronósticos de volatilidad con la influencia de variables económicas y de la pandemia de la COVID-19 sobre diferentes materias primas y productos agrícolas de exportación

como por ejemplo petróleo, gas, oro, flores, banano, aguacate y cacao entre otros, basados en técnicas de aprendizaje automático que permitan tener un impacto positivo sobre la economía nacional y asimismo que sirva como insumo para productores como mecanismo de previsión a la hora de comercializar sus productos.

Tabla de cumplimiento de objetivos

Tabla 1

Tabla de cumplimiento de objetivos.

Objetivo	Apartado Relacionado
1. Realizar una revisión de literatura sobre modelos de aprendizaje automático y análisis de minería de texto para el pronóstico de la volatilidad de productos agrícolas.	Capítulo 4
2. Comparar modelos de aprendizaje automático para el pronóstico de la volatilidad de productos agrícolas	Capítulo 4 y 7
3. Seleccionar el modelo de aprendizaje automático para el pronóstico de la volatilidad del café colombiano basado en métricas de bondad de ajuste.	Capítulo 8
4. Desarrollar una herramienta computacional para la visualización de los resultados del modelo de pronóstico desarrollado.	Apéndice A
5. Elaborar un artículo investigativo de carácter publicable sobre los resultados obtenidos en el proyecto.	Apéndice B

Resultados Esperados.

- Modelo predictivo para la volatilidad de los futuros del café colombiano basado en noticias antes y durante la pandemia de la COVID-19 y variables económicas.
- Artículo científico con los hallazgos obtenidos en la investigación realizada.
- Libro de trabajo de grado con las evidencias de la investigación

1. Planteamiento del Problema

El café es uno de los alimentos más apetecidos en el mundo, con un consumo mundial en el periodo 2020 - 2021 cercano a 166.3 millones de sacos de 60 kg, donde los mercados más importantes fueron el europeo, asiático y norteamericano con 54.1, 36.6 y 30.9 millones de sacos respectivamente, destacando a Colombia como el tercer país a nivel de exportación de este grano (International Coffee Organization, 2021). Por tal motivo, es uno de los commodities más importantes en la economía nacional, representado alrededor del 15% del PIB agrícola, convirtiéndolo así en el primer producto de exportación no minero ni petrolero del país. Además, actualmente es el sustento de más de medio millón de familias que se benefician de manera directa, ubicados en cerca de 23 departamentos y generando más de 700.000 empleos directos en todo el territorio nacional. (Federación Nacional de Cafeteros de Colombia, 2017).

Es importante anotar que el café tuvo una reducción de la producción al cierre del 2021 de 13.4 millones frente a los 14.1 millones de sacos de 60 Kg recolectados en 2020, por el impacto de las protestas en el país durante mayo y junio, asimismo, el consumo nacional disminuyó 1.5% en relación con el de 2020, esto a raíz del levantamiento del confinamiento causado por la pandemia de la COVID-19. Por otro lado, el precio del café colombiano tuvo máximos históricos en junio de 2021 con un valor $1.905.000 COP por carga de 125 kg y un precio total de la cosecha por 9.9 billones de COP, superando en un 14% lo recaudado en 2020, gracias a la fortaleza del dólar estadounidense frente al peso colombiano a causa de las decisiones de la Reserva Federal de Estados Unidos (FED) en cuanto a tasas de interés y política monetaria para disminuir el efecto de la pandemia de la COVID-19 sobre la economía estadounidense (Federación Nacional de Cafeteros , 2021). Por lo tanto, la pandemia de la COVID-19 ha supuesto retos importantes no solo en la producción, sino también en el comportamiento del precio, ya que el consumo interno y la

exportación del grano generan cambios en la balanza comercial, estando sujetas a las fluctuaciones del precio, debido a la volatilidad que se ha presentado en los últimos años en las tendencias del mercado mundial, suponiendo riesgos en cuanto a la producción y comercialización del grano para los agricultores, importadores y la economía nacional (Deina, y otros, 2021).

Sumado a los problemas descritos, la pandemia de la COVID -19 ha causado problemas económicos y financieros en todo el mundo, ya que las medidas para reducir la propagación del virus han ralentizado la economía mundial, causadas por las cuarentenas y las restricciones de movilidad, provocando que los inversores sean reacios al riesgo, lo que lleva a una menor confianza de las empresas y los consumidores. (OCDE, 2020). Asimismo, la pandemia COVID-19 ha sido el centro de atención de los medios de comunicación alrededor del mundo, quienes enfatizaron en la gravedad y consecuencias de esta enfermedad. Por tal motivo, las noticias derivadas de la pandemia COVID-19 juegan un papel importante en los sentimientos de los inversores (Weng et al., 2021). En este orden de ideas, las noticias tienen poder predictivo y esto no debe pasarse por alto al momento de pronosticar las dinámicas en el comercio de productos básicos, los inversores individuales prefieren centrarse en mercados atractivos, por lo que, de hecho, las noticias pueden atraer o repeler a los inversores y, por lo tanto, generar retornos inesperados (Narayan, 2019).

Ahora bien, las metodologías utilizadas para predecir los precios de las materias primas y la volatilidad de los commodities son cada vez más relevantes, no solamente contribuyen a la toma de decisiones, sino también a toda la cadena de productividad. Por tal motivo, existen estrategias y herramientas que ayudan a resolver problemas de predicción, donde es posible encontrar tanto modelos clásicos apoyados en el análisis de series temporales como basados en aprendizaje automático (ML). Sin embargo, la precisión de los resultados puede mejorar usando enfoques de última generación, como los es una ELM (Extreme Learn Machine), la cual es una herramienta

eficiente para resolver tareas de predicción, siendo adecuados para enfrentar problemas desafiantes de pronóstico debido a que presentan altas capacidades de aproximación y generalización (Deina, y otros, 2021). En este caso, se aplicarán algunos modelos de aprendizaje automático para el pronóstico de la volatilidad de los futuros del café colombiano basado en variables económicas y las noticias durante la pandemia de la COVID-19, dichas variables se obtendrán mediante conjuntos de datos disponibles en la página de la Federación Colombiana De Cafeteros, Departamento Administrativo Nacional de Estadística y RavenPack.

2. Objetivos

2.1 Objetivo General

Elaborar un modelo de aprendizaje automático para el pronóstico de la volatilidad del café colombiano basado en variables económicas y la influencia de las noticias antes y durante la pandemia de la COVID-19.

2.2 Objetivos Específicos

Realizar una revisión de literatura sobre modelos de aprendizaje automático y análisis de minería de texto para el pronóstico de la volatilidad de productos agrícolas.

Comparar modelos de aprendizaje automático para el pronóstico de la volatilidad de productos agrícolas

Seleccionar el modelo de aprendizaje automático para el pronóstico de la volatilidad del café colombiano basado en métricas de bondad de ajuste.

Desarrollar una herramienta computacional para la visualización de los resultados del modelo de pronóstico desarrollado.

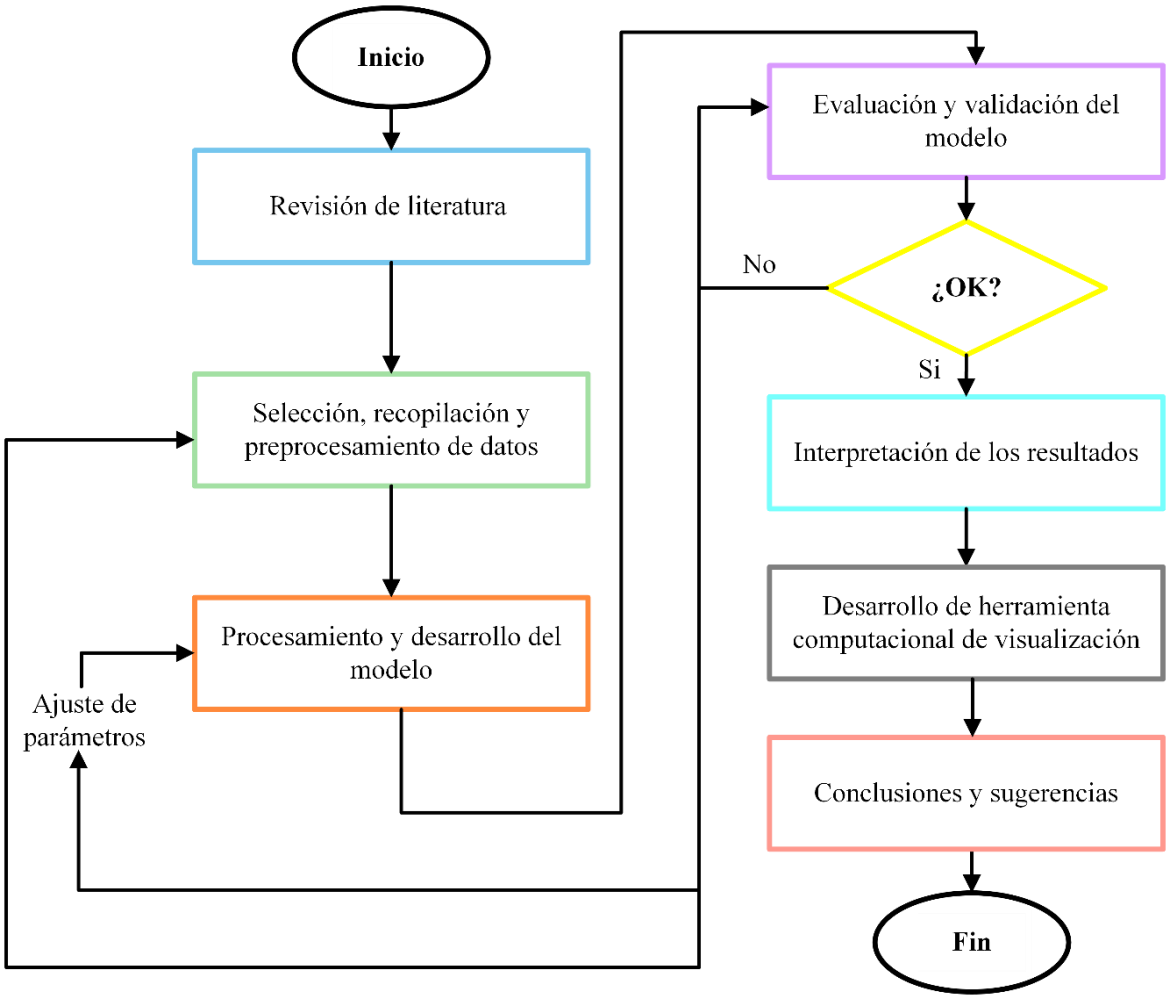
Elaborar un artículo investigativo de carácter publicable sobre los resultados obtenidos en el proyecto.

3. Metodología.

Para el desarrollo del presente proyecto se aplicará la metodología KDD (*Knowledge Discovery in databases)* o mejor conocida en español como "Descubrimiento de conocimiento en bases de datos" propuesta por (Fayyad et al., 1996) quienes la definen como "El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y esencialmente entendibles a partir de los datos". La Figura 1 muestra la metodología KDD que consta de 9 actividades, que, para la presente investigación se dividió en 7 fases.

Figura 1

Flujo de la metodología del proyecto.



3.1.1 Fase 1: Revisión de literatura

1. Crear una ecuación de búsqueda en la base de datos Scopus, con su respectivo análisis bibliométrico con base en los resultados extraídos por medio de programas como Bibliometrix y VOSviewer.

2. Analizar y describir el comportamiento de la temática en el tiempo, es decir, la manera en que diversos autores abordan la temática de pronóstico de volatilidad de productos agrícolas antes y durante la pandemia de COVID-19 y las variables económicas que inciden en este.

3. Identificar los diferentes modelos de aprendizaje automático y los diversos enfoques que han propuesto en la literatura.

3.1.2 Fase 2: Selección, recopilación y preprocesamiento de datos (datos objetivo)

En esta etapa se identifica el conocimiento relevante y prioritario, identificando un grupo de datos objetivo, donde se puede seleccionar todo el conjunto de datos o una muestra representativa de este, sobre el cual se desarrollará la investigación. Asimismo, analizar la calidad de los datos, donde se sustraen los datos atípicos, se emplean estrategias para el manejo de datos desconocidos, nulos, duplicados y técnicas estadísticas como media, moda, máximo, mínimo y regresiones lineales para su reemplazo.

1. Seleccionar la fuente de datos estructurados del precio diario del café colombiano y los índices producidos por los medios durante la pandemia de la COVID-19.

2. Establecer los conjuntos de datos necesarios en la base de datos de la FNC (precio diario del café colombiano), DANE (variables económicas colombianas) y RavenPack

(índices de medios antes y durante la COVID-19) sobre los que se va a realizar el estudio.

3. Elegir el método idóneo para la extracción de dichos datos.
4. Reducción de la dimensionalidad en los conjuntos de datos seleccionados, con el fin de reducir el número de variables, utilizando aquellas que sirvan como representación de los datos.
5. Tratamiento o eliminación de datos atípicos.
6. Limpiar y homogeneizar los datos desconocidos, nulos y duplicados.
7. Crear la serie de tiempo tanto de las variables de los precios diarios del café colombiano como de los índices de los medios colombianos.
8. Adaptar los datos para el alistamiento y el desarrollo del modelo.

3.1.3 Fase 3: Procesamiento y desarrollo de los modelos

1. Seleccionar el modelo según el comportamiento de las series temporales y la manera en que mejor se adapta para la previsión de volatilidad basado en noticias relacionadas a la COVID-19.
2. Desarrollar el modelo bajo el cual se desarrollará el estudio.
3. Calibrar los hiperparámetros del modelo seleccionado.
4. Entrenar el modelo a partir de datos de entrenamiento de los cuales aprenda el modelo de ML.
5. Ejecutar el modelo.

3.1.4 Fase 5: Evaluación y validación de los modelos

En esta fase se evaluará y dará validez al modelo, donde se solventarán inconvenientes y/o limitaciones potenciales durante el desarrollo de fases previas a través de diversas herramientas como matrices de comparación, matrices de desempeño o KPI's.

1. Evaluar el desempeño del modelo y como se adapta este, donde se corregirán posibles errores, y si fuere el caso, retornar a etapas previas para realizar nuevas iteraciones.
2. Validar el modelo a través de métricas de desempeño y comparación.
3. Remover patrones redundantes o irrelevantes.

3.1.5 Fase 6: Interpretación de los resultados

Esta etapa se consolidará e interpretarán los hallazgos según los criterios de éxito de la investigación, los cuales permitan un profundo análisis y descripción de los fenómenos descubiertos, asimismo se documentará el conocimiento y patrones encontrados.

1. Interpretar y analizar los patrones descubiertos a través de una profunda reflexión acerca del nuevo conocimiento encontrado o alcanzado.
2. Documentar hallazgos en términos útiles que faciliten la comprensión e interpretación de los resultados con el fin de permitir un acceso más sencillo para un uso posterior.

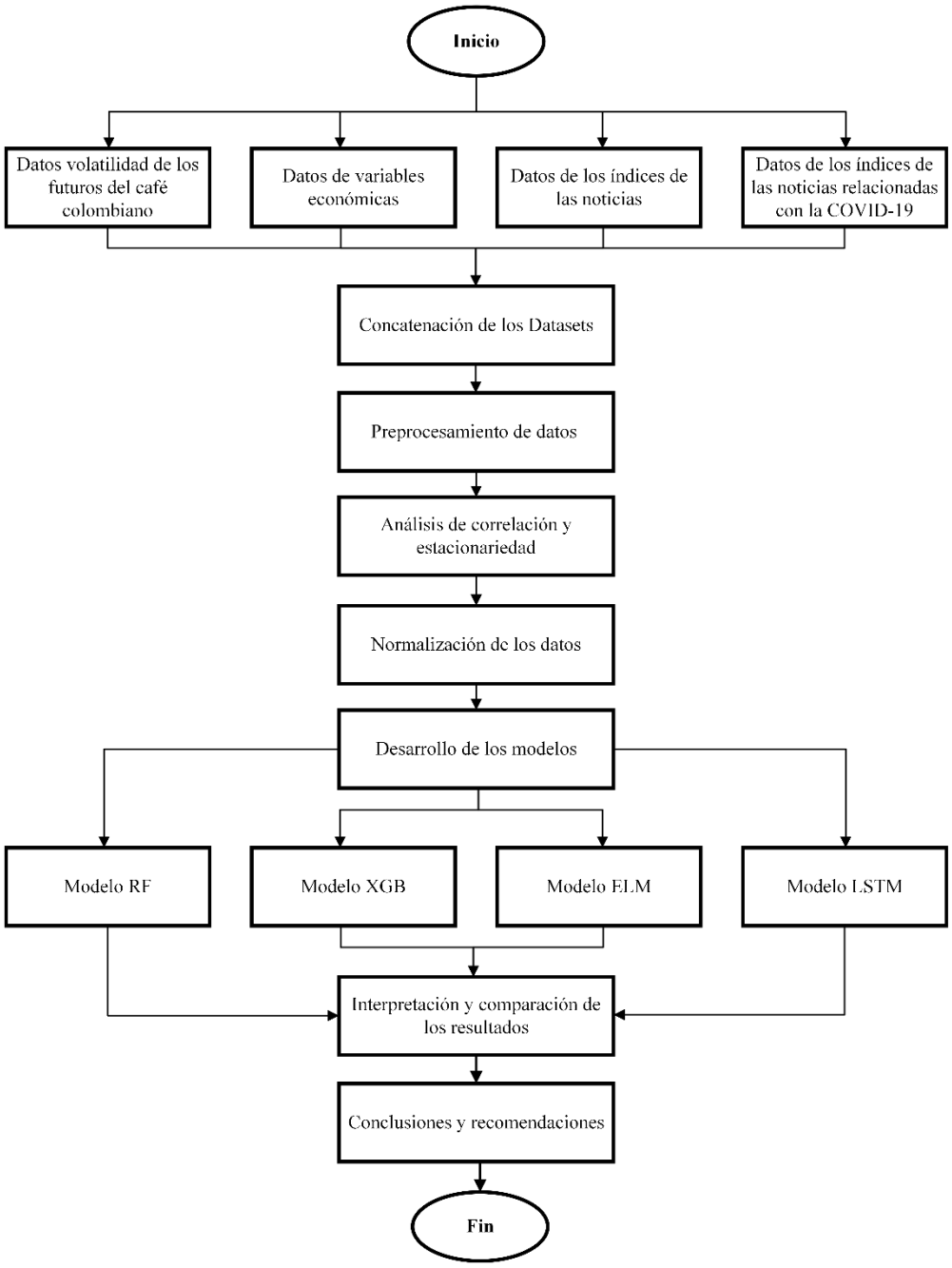
3.1.6 Fase 7: Conclusiones y recomendaciones.

1. Concluir acerca del marco metodológico, descubrimientos y resultados obtenidos durante el desarrollo de la investigación.
2. Presentar una serie de recomendaciones para trabajos e investigaciones futuras.

En la Figura 2, se presenta las fases de trabajo 2,3,4,5,6 y 7 de forma detallada.

Figura 2

Flujo metodológico del alistamiento, creación del modelo y conclusiones.



4. Revisión de la Literatura.

4.1 Análisis Bibliométrico

El análisis bibliométrico tiene como finalidad proporcionar información sobre los resultados de procesos investigativos, la evolución, visibilidad y estructuras de las diversas temáticas que son objeto de estudio en el tiempo, asimismo, se explora el impacto de estos procesos en el entorno y las fuentes de información relacionadas a este. Para la ejecución de la revisión de literatura se realiza una ecuación de búsqueda construida de manera iterativa, yendo de lo general a lo particular, donde se realizan búsquedas en diversas bases de datos disponibles como lo son: ScienceDirect (Elsevier), Web of Science (ISI Web of Knowledge), SpringerLink (Springer Science+Business Media) y Scopus (Elsevier). Recalcando esta última por la robustez en su repositorio bibliográfico y su disposición al proceso de filtrar y limitar la búsqueda.

Figura 3

Ecuación de búsqueda.

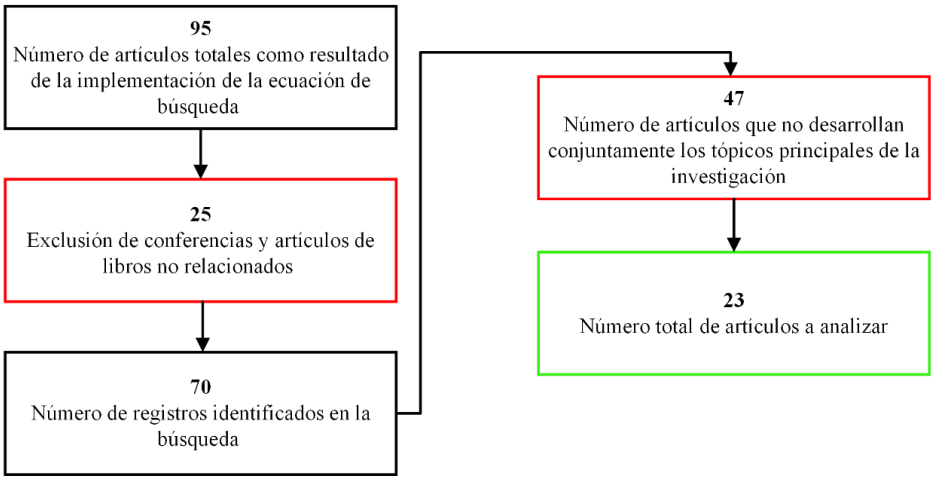
(Forecast) AND (volatility OR price OR fluctuation OR uncertainty OR Price) AND ("agricultural product" OR "agricultural commodities" OR "agricultural produce") AND ("Machine Learning" OR Machine-Learning OR "artificial intelligence") AND (News OR "fake news")

La ecuación es implementada en la base de datos Scopus, donde se encuentra un volumen total de 95 artículos mediante cuatro tópicos de búsqueda (predicción de precios, productos agrícolas, modelos basados en técnicas de aprendizaje automático y difusión de noticias falsas), a dicho número de artículos se efectúan dos depuraciones, para la primera se utiliza como criterio limitar la búsqueda a literatura relacionada con publicaciones científicas y revisiones de literatura, excluyendo conferencias y capítulos de libros no relacionados, lo cual genera un conjunto 70 artículos. Finalmente, en la segunda y última depuración se tiene en cuenta como criterio principal

que se desarrollen conjuntamente los cuatro tópicos mencionados anteriormente, dando como resultado 23 artículos relacionados con la investigación expuestos en la Figura 4.

Figura 4

Artículos seleccionados para la revisión de literatura.

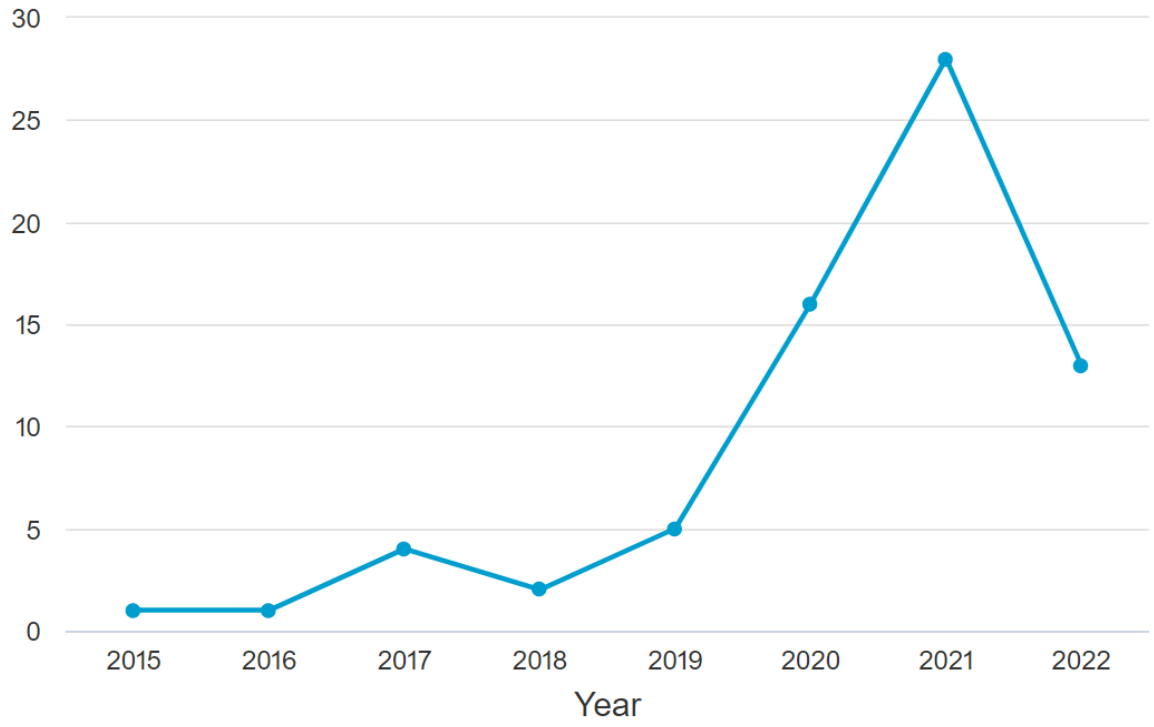


El análisis bibliométrico se realiza mediante el lenguaje de programación R con la herramienta Bibliometrix cuyo paquete permite importar datos bibliográficos de Scopus, también se cuenta con la herramienta para análisis de estudios bibliométricos de Scopus y VOSviewer.

Mediante el análisis, es posible el rastreo de los comienzos del trabajo intelectual del tema de interés en la Figura 5, mediante la publicación del autor Zhang, &., Choudhry, T.(2015) con el artículo *"Forecasting the daily dynamic hedge ratios by GARCH models: evidence from the agricultural futures markets"*, desde al año 2015 se puede observar que hubo un crecimiento moderado hasta el año 2019, sin embargo a partir del 2020 y 2021 hay un aumento de artículos publicados, esto debido a la crisis mundial de la COVID-19 lo que convirtió en gran tema de interés de investigación con un pico de 28 artículos y obteniendo una cantidad en el 2022 vigente de 13 artículos en total.

Figura 5

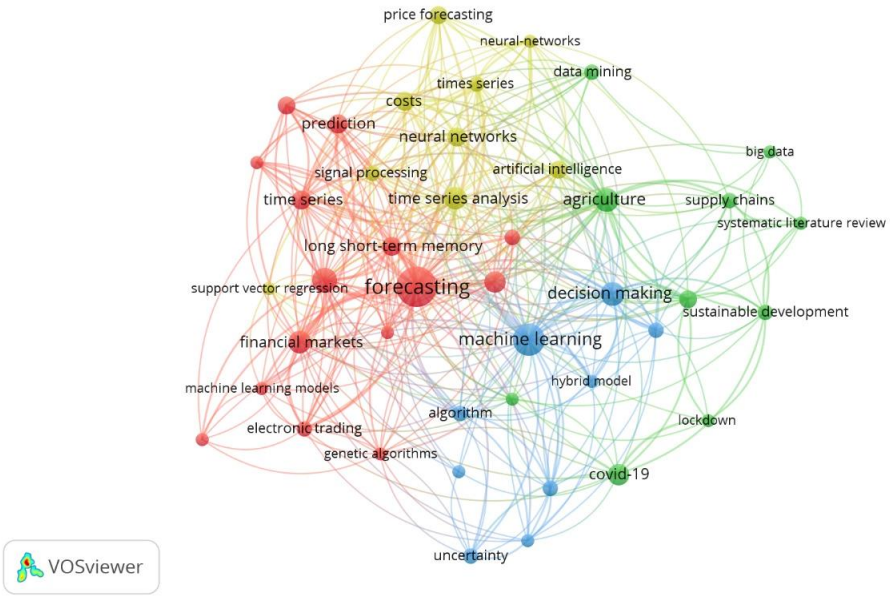
Publicaciones elaboradas por año.



Haciendo una exploración global, en la Figura 6 es posible evidenciar que se generan tres clústeres. El primero (color rojo), expone la relevancia del pronóstico o previsión en los mercados financieros a través de modelos de aprendizaje automático como LSTM (Long Short-Term Memory) y SVR (Support Vector Regression). En el segundo (verde y azul), se evidencia la relación entre el aprendizaje automático y la agricultura, los cuales se apoyan en la minería de datos para la toma de decisiones muy probablemente debido a la incertidumbre generada en torno a la COVID-19. El tercero y último grupo (color amarillo) exhibe la manera en que la inteligencia artificial permite por medio del análisis de series de tiempo y el estudio de modelos basados en redes neuronales el pronóstico de precios de productos agrícolas.

Figura 6

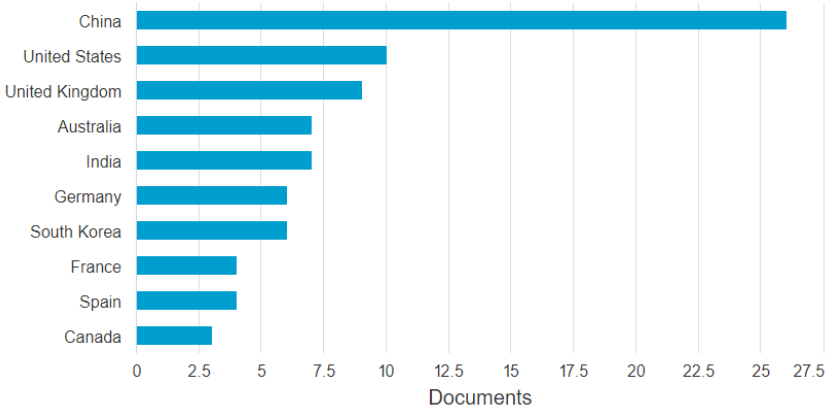
Mapa de red de palabras clave.



Finalmente, en la Figura 7 se muestra que el continente asiático, en particular China (26) es la principal zona de investigación de dichas temáticas, seguido por Estados Unidos (10), Reino Unido (9) y Australia (7), cabe resaltar que la colaboración entre países con mayor contribución intelectual es: China, Reino Unido, Estados Unidos, Australia e India.

Figura 7

Aportes investigativos por país.



4.2 Revisión de literatura

Como se menciona en el análisis bibliométrico, las temáticas que se desarrollarán en esta investigación han tomado relevancia desde el 2020 por el interés de estudiar los efectos de la pandemia en productos básicos como oro, petróleo y productos agrícolas, por lo que la mayoría de los estudios se han enmarcado alrededor del análisis del petróleo y en algunos casos a productos agrícolas. Se encontraron modelos enfocados en la predicción de precio y/o volatilidad de commodities con el componente de análisis de sentimientos en el contexto de la pandemia por la COVID-19 en productos, por lo tanto, se decide dar prelación a aquellos artículos que enmarcan estos dos ítems.

4.2.1 Modelos predictivos de volatilidad y/o precio

Los modelos predictivos para el análisis de los precios se presentan en varios campos, como estudios financieros, trabajos académicos y proyectos de investigación, donde a partir de un pronóstico con bajo margen de error es posible tomar decisiones, siendo de esta forma, la predicción de precios fundamental para generar ganancias, influencias en la toma de decisiones y, por lo tanto, en la asignación de recursos y el bienestar económico (Xu, 2019)

De este modo, el análisis del comportamiento de los precios es un tema que ha sido estudiado y realizado varios años a través del uso de modelos matemáticos como el análisis de regresión lineal, donde “se usa para la identificación de relaciones potencialmente causales o bien, cuando no existen dudas sobre su relación causal, para predecir una variable a partir de la otra” (Dagnino, 2014), siendo utilizado por Rotem Zelingher D. M., (2021) como punto de referencia para estimar la fluctuación del precio en función de los cambios en la producción o el rendimiento regionales, usando un modelo lineal generalizado (GML) donde calculaba la probabilidad de un aumento de precio dado los valores de producción o rendimiento.

Uno de los modelos más utilizados para la predicción de precios son las redes neuronales, en el artículo de (Xu & Zhang, 2022) proponen pronosticar problemas en conjuntos de datos de precios diarios durante períodos de más de cincuenta años para la soja y el aceite de soja mediante redes neuronales autorregresivas no lineales (NARNN) de dos neuronas ocultas y dos retardos y de tres neuronas ocultas y tres retardos y NARNN con entradas exógenas (NARNN-X) de seis neuronas ocultas y cinco retardos y de cuatro neuronas ocultas, una gran ventaja de la red neuronal en comparación con otros enfoques no lineales para series de tiempo es que una clase de redes neuronales multicapa pueden aproximarse bien a una gran clase de funciones (TaoWang, 2010).

Según (Indranil Ghosha, 2021) "la red neuronal (DNN) y la red de memoria a largo plazo (LSTM) se utilizan luego en el conjunto de funciones procesadas para evaluar escrupulosamente la cantidad de previsibilidad de dichos activos", donde mide la Introspección de la previsibilidad del miedo del mercado en el contexto indio en términos de volatilidad de los mercados de valores, eligiendo el índice de volatilidad implícita (VIX) que es un indicador para dar cuenta de la volatilidad de los precios de las opciones en los futuros y la desviación estándar móvil de 20 días de los rendimientos NIFTY para tener en cuenta la volatilidad, respectivamente, durante la línea de tiempo en curso de COVID-19. Es más, "los hallazgos sugieren que, a pesar de exhibir un alto grado de características volátiles, tanto el VIX de India como la volatilidad histórica se pueden predecir utilizando las arquitecturas propuestas de manera efectiva y brindan conocimientos prácticos procesables" (Indranil Ghosha, 2021).

Según Luo et al., (2019) aprovecharon modelos HAR integrados con estructuras ocultas de cambio de régimen de Márkov para tener en cuenta los efectos de saltos, apalancamiento y especulación para realizar modelos de volatilidad realizados de cinco futuros de productos agrícolas, a saber, maíz, algodón, arroz Indica, Aceite de palma y soja. También, Los autores indican el enfoque de aprendizaje profundo conocido como red de memoria a corto plazo (LSTM)

como método de predicción del precio del petróleo crudo, este método mitiga la dificultad que posee la red neuronal recurrente (RNN) y de este modo permitir dependencias de largo alcance entre instancias de datos lejanos entre sí.

En Ko et al., (2023) realiza una comparación entre el modelo clásico de series de tiempo conocido como el modelo de promedio diferencial de series de tiempo (ARIMA) contra el modelo de red neuronal de IA conocido como modelo de memoria a corto plazo (LSTM), también se enfoca en predecir los precios de la carne de cerdo con base en artículos de noticias mediante la incorporación de métodos de modelado de temas y aprendizaje profundo, utilizando distribuciones de probabilidad de palabras en datos de texto para recuperar un conjunto de palabras clave principales llamado tema y evaluando el rendimiento del enfoque propuesto a través de extensos experimentos con métodos estadísticos de última generación (ARIMAX y Ridge).

En Li, y otros, (2021) realizan un pronóstico del precio de productos ganaderos basado en redes neuronales GRU donde exhibe una capacidad de ajuste no lineal eficiente. Para evaluar el desempeño de pronóstico del método de pronóstico, este estudio mide el efecto de pronóstico del método desde tres dimensiones. La primera dimensión utiliza el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de correlación (R^2) para cuantificar el rendimiento del método. Se propuso el método AE-VMD y el método MA-LZ para aumentar la precisión de la descomposición de la señal y el reconocimiento de frecuencia, se planteó la red neuronal AH-GRU, preservando así por completo la información de las series temporales de datos globales y multimodales, Combinado con información estática e información dinámica (cantidad de precipitación, temperatura, precios de otros productos y otras series temporales) inicialmente. El efecto de la información estática sobre la fluctuación de precios se introdujo en el proceso de pronóstico para proporcionar pronósticos específicos para el precio de los productos pecuarios.

Para pronosticar la dinámica a corto plazo del precio futuro del petróleo crudo, Romero-Meza (2022) propuso un algoritmo basado en la Feedforward Hidden Layers, donde hibrida el optimizador de lobo gris mejorado (IGWO) con la máquina de aprendizaje extrema (ELM) para desarrollar un modelo de predicción para predecir el precio futuro del petróleo crudo conjuntos de datos. La ventaja de usar este algoritmo es que no requiere ningún ajuste fino de los parámetros en cada iteración, lo que hace que la convergencia del error sea más rápida, y también el ELM tiene una tasa de aprendizaje más rápida y una ejecución de especulación más alta que las técnicas de aprendizaje basadas en gradientes.

4.2.2 Análisis de sentimientos en el contexto de la pandemia COVID-19

Con la velocidad con la que se desarrolla internet y las avanzadas tecnologías de procesamiento de datos, el análisis de sentimientos ofrece la capacidad de explorar noticias de internet sobre productos básicos, la influencia de este sobre el precio de productos y el comportamiento de venta (Tseng et al,. 2017). En la actualidad, herramientas computacionales como el análisis de sentimientos han cambiado la dinámica de hacer negocios, haciendo uso de técnicas como minería de texto, procesamiento de lenguaje natural (NLP), con la intención de analizar y evaluar los sentimientos y actitudes mediante análisis de opinión, por tal motivo, la finalidad del análisis de sentimientos es un conjunto de técnicas que tienen la capacidad de extraer información basadas en opiniones a partir de datos sin un procesamiento previo (Aditya Bhardwaj, 2015).

El impacto de las noticias en los mercados financieros ha sido de interés para académicos, por ejemplo, en Narayan (2019) afirman que estas pueden llegar a tener un poder predictivo, por lo tanto, no se puede ignorar a la hora de predecir la dinámica futura de los mercados. Asimismo, no solamente las noticias financieras resultan útiles para predecir el rendimiento de las acciones,

sino que las noticias en general son una fuente de información para el mercado de valores. Del mismo modo, las materias primas sensibles como el petróleo, la energía, productos agrícolas y acciones poseen alta volatilidad y están influenciadas por las noticias (Tseng et al., 2018).

Sumado a esto, la pandemia COVID-19 ha llevado a los investigadores a estudiar el papel que juega la incertidumbre al momento de pronosticar la volatilidad de los movimientos de los productos básicos, debido a que la volatilidad es un insumo importante para la toma de decisiones de inversión y las opciones de cartera (Rangan Gupta, 2022). Además, en (Haroon & Rizvi, 2020) afirman que pandemias como la de COVID-19 causan frenesí en los medios, donde los participantes de los mercados financieros no tienen la capacidad de evaluar dicha cantidad de noticias y su efecto económico, encontrando que el pánico generado por los medios de comunicación está relacionado con una mayor volatilidad en los mercados financieros alrededor del mundo y esta asociación es más fuerte para las sectores que tuvieron mayor afectación durante la pandemia.

Dentro de la literatura fue posible encontrar diversas técnicas para la clasificación y análisis de sentimientos, se centran en enfoques basado en léxico y aprendizaje automático, para efectos de este trabajo el estudio y la orientación será en su mayoría en modelos apoyados en Machine Learning, los cuales se presentan a continuación:

En el artículo propuesto por Rangan Gupta (2022) titulado *"Forecasting the realized variance of oil-price returns: a disaggregated analysis of the role of uncertainty and geopolitical risk"* plantean como referencia un modelo autorregresivo (AR) con un término rezagado, donde añaden rastreadores de volatilidad del mercado de valores (EMV), incertidumbre de la política económica (EPU) y amenaza y riesgos geopolíticos globales (GPR) siendo útiles para predecir la

varianza realizada o *volatilidad realizada* de los rendimientos del precio del petróleo cuando el horizonte de pronóstico es lo suficientemente largo. Sumado a esto, muestran como la adición de la incertidumbre asociada al COVID-19 mejora significativamente la precisión de las previsiones, de esta forma afirman que la incertidumbre y los riesgos geopolíticos sirven para predecir las varianzas realizadas de los movimientos de los precios de otros recursos fósiles y quizás de las materias primas agrícolas.

En (Liu et al., 2022), con su artículo *"Dynamic impact of negative public sentiment on agricultural product prices during COVID-19"* presentan un modelo de autorregresión de parámetros variables en el tiempo (TVP-VAR) el cual pudo capturar las respuestas de impulso de los precios de los productos agrícolas a los sentimientos negativos del público en línea en cada punto de la serie temporal y demostrar el impacto dinámico del sentimiento negativo en línea del COVID-19 en diferentes precios de productos agrícolas. Además, muestra como el pánico y el sentimiento público logra generar desequilibrios entre la oferta y demanda, asimismo fluctuaciones de precios en los productos agrícolas, capturando los comportamientos de los compradores bajo pánico.

(Lei Chai, 2020) con *"A multi-source heterogeneous data analytic method for future price fluctuation prediction"* proponen un método de análisis de datos heterogéneos de múltiples fuentes (MHDA) para la predicción del precio de los futuros, un mapa de relaciones de dominio específico para analizar la influencia de todas las entidades relacionadas y un diccionario para capturar el sentimiento en eventos noticiosos y un modelo gaussiano de Márkov de mezcla oculta (HMGMM) para capturar las dependencias temporales subyacentes para la predicción de precios en los datos de series temporales de múltiples fuentes. Lo modelos empleados los colocan a prueba con datos de los futuros del aceite de palma de China, demostrando la efectividad del modelo y que las

estrategias de inversión se ven afectadas por sesgos psicológicos como el exceso de confianza o el pesimismo.

Por su parte, (Ye et al., 2021) *"A heterogeneous graph enhanced LSTM network for hog price prediction using online discussion"* proponen LSTM mejorado con gráficos heterogéneos (HGLTSM), explorando la influencia de las discusiones de foros de inversores del precio del cerdo, proponiendo un modelo que predice los precios semanales del cerdo. Además, afirman que se han efectuado estudios anteriormente para productos básicos a través de modelos métodos de regresión como el promedio móvil integrado autorregresivo (ARIMA), siendo el ARIMA generalizado el cual mostró mejor rendimiento en el pronóstico de precios del cacao, donde este desempeño se obtiene en series temporales lineales, cuando las series no son lineales o no estacionarias la eficacia de este disminuye significativamente. Por su parte en (Weng et al., 2021) con "*Volatility forecasting of crude oil futures based on a genetic algorithm regularization online extreme learning machine with a forgetting factor: The role of news during the COVID-19 pandemic"* propusieron máquina de aprendizaje extremo en línea de regularización de algoritmos genéticos con factor de olvido (GA-RFOS-ELM), para estimar los efectos de las noticias durante la pandemia de COVID-19 en la volatilidad de los futuros de petróleo crudo, determinando que dicho modelo supera estadísticamente a todos los modelos de referencia en cuanto a la precisión del pronóstico. Además, logran determinar que las noticias relacionadas con COVID-19 están asociadas con la volatilidad en los mercados de valores.

En conclusión, en torno a las investigaciones que se han realizado en diferentes países y la relevancia que ha tenido tanto para gobiernos como académicos en la toma de decisiones apoyadas en el uso de herramientas computacionales a partir de modelos de aprendizaje automático en la predicción de precios y/o volatilidad de commodities como petróleo o productos agrícolas, se hace

necesario explorar acerca del comportamiento y la predicción de los futuros del café colombiano basado en variables económicas y las noticias antes y durante la pandemia de la COVID-19.

5. Marco de Referencia.

Colombia es uno de los productores de café más reconocidos de arábigo, siendo el segundo productor de arábica suave a nivel mundial y manteniendo altos estándares de calidad, gracias al arduo trabajo de miles de familias campesinas que desempeñan esta labor (National Coffee Association, 2017). En la actualidad, el mercado cafetero colombiano está atravesando periodos de incertidumbre, donde factores como la inflación de 9.23% para abril de 2022, la pandemia de la COVID-19, la devaluación del peso frente al dólar estadounidense y precios del grano a nivel mundial no vistos desde 2012 con un aumento del 80%, han generado preocupación tanto en el mercado financiero nacional como internacional (Portafolio, 2022).

Sumado a los problemas descritos, la pandemia a causa de la COVID-19 ha tenido incidencia importante en la economía colombiana, donde se ha alterado la oferta y la demanda del sector cafetero. Según un informe de la Organización Internacional del Café (OIC) la pandemia ha generado fuerte volatilidad en los precios del café, estableciendo un enorme reto para este sector que para inicios de la pandemia se encontraba en un periodo extenso de precios bajos, siendo los países exportadores como Colombia quienes más han sufrido las consecuencias derivadas de la pandemia, como resultado existe incertidumbre en cuanto a la producción, empleo, ingresos, exportaciones y consumo interno en el mercado cafetero nacional (Organización Internacional del Café, 2020).

En consecuencia, la producción nacional de café ha tenido una disminución para el primer trimestre de 2022 de un 16% respecto a el mismo periodo en 2021, lo que se traduce en una reducción de 500.000 sacos de 60 kg respecto al año inmediatamente anterior, por ende, las

exportaciones presentaron una reducción cercana al 10% frente a los casi 3,5 millones sacos de 60 kg exportados en 2021 (Federación Nacional de Cafeteros, 2022). Sin embargo, la carga de café de 125 kilos ha llegado a máximos históricos de $2.307.000 COP, lo que se ha traducido en una posición favorable para las familias cafeteras y para los cerca de 2.5 millones de empleos entre directos e indirectos que esta actividad genera (Ministerio de Agricultura, 2022).

Entorno a estas problemáticas se han realizado diversos estudios que pretenden analizar el impacto de la pandemia de la COVID-19 en los mercados internacionales, la influencia que tienen la noticias entorno a esta enfermedad y el efecto de estos factores en los precios de los productos agrícolas. Por ejemplo, (Zhang, Hu, & Ji, 2020) expone que los riesgos del mercado financiero mundial han crecido en respuesta a la pandemia, generando incertidumbre y pérdidas económicas, la cuales han derivado en que los mercados sean muy volátiles e impredecibles. Además, (Haroon & Rizvi, 2020) exploran la relación entre el sentimiento generado por las noticias derivadas de la COVID-19 y la volatilidad de los mercados, donde logran afirmar que el pánico generado por los medios de comunicación está relacionado con un incremento en la volatilidad de los mercados financieros mundiales, siendo más fuerte esta relación para los sectores productivos que tuvieron una mayor afectación durante la pandemia, donde el sentimiento y la cobertura de los medios no fueron relevantes en cuanto a la volatilidad de los precios. Sin embargo, (Liu et al., 2022) afirman que el sentimiento publico negativo afectó los precios de los productos agrícolas durante la COVID-19 y que se acentuó en los periodos de propagación y recesión del virus.

En este orden de ideas, resulta de interés el estudio de técnicas que permitan la previsión volatilidad o del precio de commodities en el contexto de la pandemia y asimismo el efecto de las noticias sobre los productos básicos. En particular, (Weng et al., 2021) a través de técnicas de aprendizaje automático afirman que las noticias tienen valor predictivo parar la volatilidad del

petróleo durante la pandemia del nuevo coronavirus, asimismo, manifiestan que modelos de Machine Learning en el contexto de la pandemia de COVID-19, resultan eficaces y eficientes en el pronóstico de volatilidad de los futuros del petróleo crudo. Además, (Deina, y otros, 2021) realizaron un estudio de previsión de precio del café basada en el uso de Extreme Learn Machines (ELM), encontrado que las redes neuronales, en particular el ELM, llega a niveles de desempeño por encima de otros modelos, manifestando que en investigaciones posteriores es posible desarrollar diferentes arquitecturas de redes neuronales, conjuntos y series de otros países.

En definitiva, existen diversos estudios que han explorado la influencia de las noticias durante la pandemia de la COVID-19 y como se relacionan a la hora de hacer previsiones de volatilidad y precio basados en técnicas de aprendizaje automático, dichas investigaciones se han realizado en su mayoría para productos como petróleo, oro y algunos productos agrícolas, principalmente en países como China, Estados Unidos, Reino Unido, Gran Bretaña, Australia e India. Por lo tanto, existe la necesidad de indagar las implicaciones que se han tenido antes y durante la pandemia a causa del nuevo coronavirus, las noticias relacionadas a este en cuanto a la volatilidad de los futuros del café colombiano y las incidencias que se derivan al momento de pronosticar los futuros de este.

5.1 Marco de Antecedentes.

El estudio del comportamiento de la volatilidad ha sido un tema de gran interés en los últimos años, lo que ha generado un variado número de investigaciones y publicaciones respecto al tema, algunos de estos estudios se enfocan en la previsión de la volatilidad mediante el uso de diferentes técnicas de predicción, también teniendo en cuenta la influencia de factores económicos y las noticias frente al comportamiento de la volatilidad, con el uso de técnicas de aprendizaje automático y el análisis de sentimientos. A continuación, se exhiben tres proyectos de autores que

han incursionado en el uso de estas técnicas, buscando obtener un mejor desarrollo del mismo proyecto, aplicándolo a diferentes campos *científicos*, académicos, económicos, etc.

En su trabajo (Weng et al., 2021), *"Volatility forecasting of crude oil futures based on a genetic algorithm regularization online extreme learning machine with a forgetting factor: The role of news during the COVID-19 pandemic"* proponen una máquina de aprendizaje extremo en línea de regularización de algoritmos genéticos con factor de olvido (GA-RFOS-ELM), para estimar los efectos de las noticias durante la pandemia de COVID-19 en la volatilidad de los futuros de petróleo crudo. Las noticias durante la pandemia de COVID-19 exhiben actualidad, donde los nuevos datos atraen más énfasis, mientras que los datos más antiguos se olvidan gradualmente. Teniendo en cuenta la actualidad de las noticias relacionadas con COVID 19, el modelo GA-RFOS-ELM fortalece la capacidad de búsqueda óptima del algoritmo genético. Puede ser eficaz y eficiente utilizar un mecanismo de aprendizaje fragmento por fragmento con un tamaño de fragmento fijo o variable, lo que ilustra que se necesita la capacidad de aprendizaje en línea. Además, el precio del petróleo crudo es no lineal y no estacionario. También el método GA-RFOS-ELM se compara con varios modelos econométricos y algoritmos de aprendizaje automático, incluidos autorregresivos (AR) donde suele emplearse para resolver problemas de series temporales, árboles de regresión (RT), regresión bayesiana (Bayes), regresión de vector de soporte (SVR) que son modelos clásicos de aprendizaje automático basados en diferentes teorías, que se utilizan comúnmente como métodos de referencia de aprendizaje automático. Además, se consideran ELM y OS-ELM, como dos algoritmos importantes en el desarrollo de la familia de algoritmos ELM, y al mismo tiempo, GA-OS-ELM se presenta como un enfoque de referencia para comparar el modelo que proponen. Según los autores, en la literatura no es seguro que criterios son más apropiados para evaluar las predicciones de modelos de volatilidad. Por lo tanto, eligieron cuatro funciones diferentes, error

cuadrático medio (RSME), error absoluto medio (MAE), error porcentual absoluto medio (MAPE) y error medio (MDE), muestra que el modelo GA-RFOS-ELM propuesto supera a otros modelos en la predicción de la futura volatilidad de los precios del petróleo crudo. Estas experiencias muestran la superioridad del modelo propuesto. Mientras tanto, prueban el papel de las noticias durante la pandemia de COVID-19. Los resultados empíricos muestran que las noticias durante la pandemia de COVID-19 afectan las decisiones de los inversores individuales, pero las noticias exhiben oportunidad debido a la explosión de información, lo que también es consistente con la cognición básica de los seres humanos. Los autores recomiendan utilizar el modelo de GA-RFOS-ELM para aplicarse a predicciones de otras variables temporales u otros mercados, como maíz, cobre y oro, proporcionando así otros modelos más precisos para el pronóstico de volatilidad en futuras investigaciones, especialmente considerando otras variables, construyendo otros índices mediante minería de texto de noticias relacionadas con COVID-19.

El proyecto *"Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI"* (Indranil Ghosha, 2021), justifica en comprender el patrón inherente del miedo del mercado en el contexto indio durante el caos de los eventos sin precedentes de la pandemia de COVID -19 que encarnan la primera y la segunda ola de infecciones. En la investigación se tiene en cuenta el índice de volatilidad implícita (VIX) que es un indicador para dar cuenta de la volatilidad de los precios de las opciones en los futuros. VIX está marcado como un activo importante para capturar el miedo del mercado en las principales economías desarrolladas y en desarrollo. Se eligió India VIX (IV) como instrumento de este en el contexto indio. Por otro lado, es una práctica común estimar la volatilidad realizada de los activos financieros y la plétora de materias primas heterogéneas para evaluar la cantidad de volatilidad histórica (HV) y los rasgos caóticos en la atmósfera comercial.

IV y HV son indicadores principales del miedo del mercado e intenta reconocer el patrón alborotador del mismo a través del despliegue sistemático de marcos de investigación integrados. Los descubrimientos sugieren tanto IV como HV de exponer un alto grado de no linealidad y rasgos no paramétricos, sin embargo, se pueden predecir con un nivel de precisión viable en situaciones extremadamente desafiantes. Observaron que IV manifiesta la volatilidad de los mercados futuros, resultaron ser comparativamente más predecible que su homólogo histórico, HV. También es importante señalar que las variables macroeconómicas poseen un mejor control sobre IV que sobre HV. Por lo tanto, negociar en el mercado de futuros mirando cifras de características explicativas destacadas sería más rentable y menos riesgoso desde la perspectiva de los inversores. El índice de volumen de búsqueda de Google (GSVI) de palabras clave para extraer el sentimiento afín de manera efectiva han prevalecido para tener muy poco impacto en IV o HV en el contexto indio. Por lo tanto, dedujeron que el miedo inherente en el mercado indio no es muy sensible a la información que flota en los portales web. Los indicadores macroeconómicos y técnicos explican en gran medida la variación de las series temporales subyacentes holísticamente, pero las construcciones de sentimientos de GSVI juegan su papel en la predicción a nivel local. Entonces, el estudio se puede aprovechar de manera eficiente para decodificar el miedo del mercado en los mercados financieros indios con un alto grado de precisión. La selección cuidadosa de características explicativas y el filtrado riguroso de características a través del algoritmo Boruta que es una versión extendida de bosque aleatorio (RF) que simplemente agrega un alto nivel de aleatoriedad en su marco metodológico para evaluar características significativas e insignificantes, donde contribuyen igualmente al éxito final en la generación de predicciones de calidad por los cuatro modelos, Aumento de gradiente extremo (XGB), Árboles extremadamente aleatorios (ERT), Red neuronal profunda (DNN) y Red neuronal de memoria a corto y largo plazo (LSTM), mientras que la IA explicable ha sido enormemente competente en descubrir el patrón oculto de

influencia de las características independientes elegidas. La cantidad de precisión obtenida clasifica los marcos para una consideración seria como instrumentos comerciales. Los marcos se pueden ampliar fácilmente para medir la volatilidad en otros activos influyentes.

El proyecto de " (Martín, 2003)", donde busca diseñar un modelo no lineal para el análisis y predicción de la serie de tiempo del precio externo del café colombiano por medio de redes neuronales artificiales, con el objetivo de obtener alcances y limitaciones de las redes neuronales artificiales frente al modelo ARIMA siendo este un modelo clásico de predicción lineal. Gracias a las características de flexibilidad y adaptabilidad de las RNA presenta una alternativa, no lineal, para el análisis y predecir el precio externo del café colombiano, teniendo como ventaja a diferencia de los modelos tradicionales que intentan ajustar los datos a un modelo, las RNA fabrica un modelo que se ajusta a los datos. Un criterio que utilizaron para la comparación de los dos modelos, se basa en la relación entre la varianza de los errores de modelo no lineal y la de los errores del modelo lineal, siendo esta relación de 0.78069, lo que significa que el modelo de RNA reduce la varianza del error del modelo de ARIMA en un 22%, lo que el autor deduce que el modelo RNA es superior al ARIMA, por lo que se puede inferir que el precio externo del café Colombiano, se puede analizar y predecir con un modelo econométrico no lineal.

Los tres proyectos que se analizaron tienen relación directa con el proyecto en desarrollo, debido a que presentan objetivos de desarrollo en común, donde comparan modelos para la obtención del más optimo, asimismo hacen uso de máquinas de aprendizaje automático y utilización de técnicas de análisis de sentimientos, cabe resaltar que el último artículo es de gran importancia pues es necesario saber el comportamiento de los datos del café, en base a esto, resultando importante, ya que sirven como base para el desarrollo de las fases del proyecto, la terminación y/o finalización del mismo.

6. Marco Teórico

6.1 Pronóstico

El pronóstico es una estimación de lo que se espera que llegue a suceder en relación con una variable, es decir, es un método para predecir lo que sucederá en el futuro, se puede utilizar en varios niveles, facilitando la toma de decisiones y asimismo para mitigar los efectos adversos de alguna variable en particular, centrándose en la reducción de la incertidumbre teniendo como base cualquier tipo de información que sea relevante (Guillermo Westreicher, 2020).

6.2 Series de tiempo

Una serie temporal es un conjunto de datos, observaciones o valores que se miden en un momento determinado y se sitúan en un orden cronológico. Los datos pueden estar espaciados uniforme o desigualmente. Cuando se recopilan series de tiempo, a menudo se realiza un análisis para identificar patrones en los datos. Básicamente, se trata de comprender lo que sucede en el tiempo. La capacidad de procesar datos de series temporales es una habilidad esencial en el mundo moderno. Son dependientes del tiempo, por lo tanto, en este caso, la suposición básica del modelo de regresión de que las observaciones son independientes no se cumple, donde por lo general suelen tener tendencia y mayoría de las series de tiempo tienden a mostrar algún tipo de tendencia estacional, es decir, la variabilidad típica durante un período de tiempo (Briega, 2016).

Según (Villavicencio, 2011), en el análisis de series temporales, existen tres componentes principales cuyo trabajo conjunto da como resultado los valores calculados, los cuales son:

La *Componente de tendencia,* que está definida como un cambio a largo plazo que se produce por un cambio a largo plazo de la media. La tendencia es posible identificarla a través de un movimiento suave de la serie a largo plazo.

La *Componente estacional*, se refiere cuando una serie temporal presenta cierta periodicidad o variación en cierto periodo (semanal, mensual, etc.). Estos periodos son fáciles de observar e interpretar, además, existe facilidad de medir de manera explícita o incluso de eliminar de la serie de datos, a este proceso se le denomina *desestacionalización de una serie de tiempo*.

La *Componente aleatoria,* la particularidad de esta componente es que no responde a ningún patrón de comportamiento, ya que es el resultado de factores aleatorios que tienen injerencia en una serie de tiempo.

6.3 Proceso estocástico estacionario

En primer lugar, es necesario definir que es una *serie de tiempo estacionaria y no estacionaria.* Según Villavicencio (2011), una serie de tiempo es estacionaria cuando sus características de medias y varianza no cambian a través del tiempo y cuya covarianza sólo se encuentra en función del rezago. Mientras que, una serie de tiempo no es estacionaria, cuando su tendencia o su variabilidad fluctúa en el tiempo. Los cambios en la media señalan que existe una tendencia a crecer o decrecer a largo plazo, por lo tanto, la serie no se encuentra en torno a un valor constante.

(Villavicencio, 2011) expone que, un proceso estocástico se describe a través de una serie de datos que evolucionan en el tiempo, donde un proceso estocástico es estacionario si la media y varianza son constantes en el tiempo y si el valor de la covarianza entre dos periodos depende exclusivamente del rezago entre el par de periodos de tiempo y no del tiempo en el cual se ha calculado la covarianza.

Sea X_t una serie de tiempo con las siguientes propiedades:

$$Media \qquad \mathrm{E}\,(X_t) = \mathrm{E}(X_{t+k}) = \mu \tag{1}$$

$$Varianza \quad \mathrm{V}(X_t) = \mathrm{V}(X_{t+k}) = \sigma^2 \tag{2}$$

$$Covarianza \quad \gamma_k = \mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)] \tag{3}$$

Donde γ_k, la covarianza al rezago *k*, es la covarianza entre los valores de X_t y X_{t+k}, que están separados *k* periodos. Por lo tanto, si una serie temporal es estacionaria, su media, su varianza y covarianza (en diferentes rezagos) se mantienen iguales sin importar el momento en sean calculadas, es decir, que son invariantes en el tiempo.

6.4 Autocorrelación

(Villavicencio, 2011) afirma que en las series de tiempo suelen presentarse escenarios donde los valores que toma una variable en el tiempo no son independientes entre ellas. A continuación, se presentan dos maneras de cuantificar dicha dependencia.

6.4.1 Función de autocorrelación simple.

Villavicencio expone que la autocorrelación mide la correlación entre dos variables que están divididas en k periodos para series estacionarias.

$$\rho_j = corr\,(X_j, X_{j-k}) = \frac{cov(X_j, X_{j-k})}{\sqrt{V(X_j)}\sqrt{V(X_{j-k})}} \tag{4}$$

Donde se cumplen las siguientes propiedades:

- $\rho_0 = 1$
- $-1 \leq \rho_j \leq 1$
- Simetría $\rho_j = \rho_{-j}$

6.4.2 Función de autocorrelación parcial (PACF)

Según (Villavicencio, 2011), la función autocorrelación parcial cuantifica la correlación entre dos variables separadas por *k* periodos.

$$\pi_j = corr\ (X_j, X_{j-k}/X_{j-1}X_{j-2} \dots X_{j-k+1}) \quad (5)$$

$$\pi_j = \frac{cov(X_j - \widehat{X_j}, X_{j-k} - \widehat{X}_{j-k})}{\sqrt{V(X_j - \widehat{X_j})}\sqrt{V(X_{j-k} - \widehat{X}_{j-k})}} \quad (6)$$

6.5 Pruebas de correlación

Las pruebas de correlación resultan de suma importancia tanto en el estudio de series temporales como en el aprendizaje automático, ya que permiten identificar patrones y relaciones entre variables, seleccionar las variables más relevantes para predecir una variable de interés y evaluar la calidad de los modelos.

Para (Lahura, 2003), una de las intenciones fundamentales del estudio de series temporales es encontrar contenido empírico capaz de describir relaciones teóricas entre diferentes variables. En este orden de ideas, existen diversos métodos que logran encontrar relaciones económicas de tipo lineal y que involucren dos variables (*relaciones multivariadas)* capaces de medir la fuerza con que estas se relacionan, esto es denominado como *coeficiente de correlación*. Este coeficiente permite establecer la dirección o sentido y la cercanía o fuerza entre dos variables o conjuntos de datos.

6.5.1 Coeficiente de correlación de Pearson.

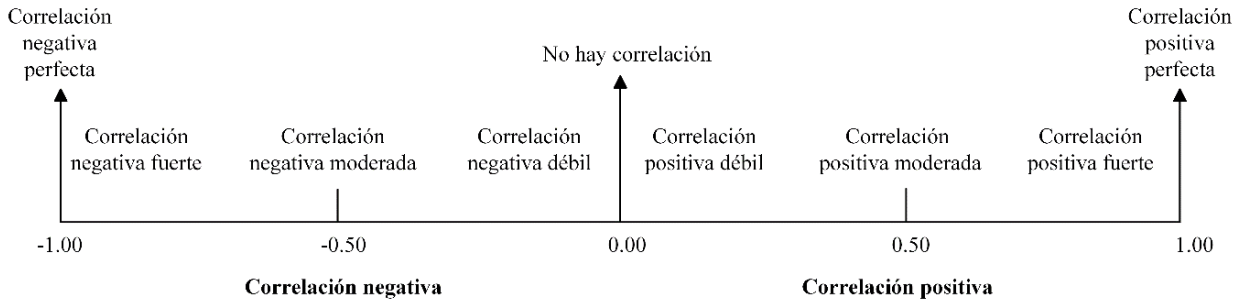
El coeficiente de correlación de Pearson mide la dependencia lineal entre dos variables x e y. También se conoce como prueba de correlación paramétrica porque depende de la distribución de los datos y está definido por:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \quad (7)$$

Donde *x* e *y* son dos vectores de tamaño *n*. Además, m_x y m_y son las medias de las variables x e y. La manera de interpretar los resultados del coeficiente de correlación de Pearson se presenta en la Figura 8.

Figura 8

Interpretación de la correlación lineal



6.6 Prueba de raíz unitaria

Según (Mahadeva & Robinson, 2009), un serio problema a la hora de pronosticar series temporales es que a menudo estas tienen cierta tendencia y es difícil emplear métodos que permitan identificarla. Para solventar este inconveniente, es necesario probar si las series de tiempo presentan estacionariedad, estas pruebas se denominan *pruebas de raíz unitaria*, además, la verificación de la estacionariedad es factor fundamental para pronosticar, ya que, puede explicar la clase de procesos que son necesarios para construir modelos predictivos, esto con el fin de realizar previsiones con mayor precisión. Por su parte (Wolters & Hassler, 2006), afirman que la prueba Dickey Fuller es ampliamente utilizada para determinar la estacionariedad de series de tiempo que serán ingresadas a modelos de aprendizaje automático.

6.6.1 Caminata aleatoria

$$y_t = \delta(y_{t-1}) + \varepsilon_t \quad (8)$$

Donde y_t es la variable predictora, y_{t-1} es el valor de rezago de la serie, ε_t es el error de ruido blanco y δ la raíz del proceso.

6.6.2 Prueba Dickey Fuller (DF).

El test de Dickey Fuller (Dickey & Fuller, 1979), prueba la hipótesis de una raíz unitaria en presencia de una tendencia lineal y se calcula de la siguiente manera:

$$\Delta y_t = \alpha + \delta(y_{t-1}) + \varepsilon_t \quad (9)$$

Donde α es el intercepto o derivada del proceso.

6.6.3 Prueba Dickey Fuller aumentada (ADF).

La diferencia entre ADF y DF es que la prueba Dickey-Fuller aumentada es la complejidad agregada al modelo de regresión, donde, al incluir retrasos de orden *p*.

$$\Delta y_t = \alpha + \rho(y_{t-1}) + \delta_1(\Delta y_{t-1}) + \cdots + \delta_p(\Delta y_{t-p}) + \varepsilon_t \quad (10)$$

Donde α es una constante, $\delta_1$es el coeficiente de tendencia en el tiempo.

Para las tres pruebas anteriores se establece la siguiente prueba de hipótesis:

$$H_0: \delta = 0$$

$$H_a: \delta < 0$$

Donde:

H_0: es la hipótesis nula (existencia de raíz unitaria o no estacionario)

H_a: es la hipótesis alternativa (no tiene raíz unitaria o estacionario)

Para las tres pruebas anteriormente expuestas, existen dos posibles resultados, si no se rechaza la hipótesis nula, la serie de tiempo tiene raíz unitaria, por lo tanto, la serie tendrá tendencia estocástica y será *no estacionaria*. Por otra parte, si el valor critico es menor que el estadístico de prueba existirá evidencia significativa para rechazar la hipótesis nula, concluyendo que no está presente una raíz unitaria, que la serie no tiene tendencia estocástica y, por lo tanto, la serie de tiempo es estacionaria.

6.7 Modelos para calcular volatilidad en series de tiempo

La volatilidad hace referencia a los cambios observados en una serie temporal a través del tiempo. En el campo de la economía, según (Andersen et al., 2006) este término es ampliamente usado para describir la variabilidad de una serie de tiempo. Además, como lo menciona (Engle & Patton, 2007) esta es una de las áreas más activas en la investigación de modelos predictivos, donde prácticamente todos los usos financieros de los modelos de volatilidad implican aspectos de pronóstico de rendimientos futuros, por tal motivo, existen diferentes métodos que permiten cuantificar dicha variabilidad.

6.7.1 Rentabilidad continua o equivalente

Según (Novales, 2017), existen dos tipos de rentabilidades útiles como factor de volatilidad, ya que han sido ampliamente usadas en el ámbito académico y financiero por permitir describir la variación de una serie temporal, estas son: *rentabilidades porcentuales y rentabilidades logarítmicas*.

Rentabilidad porcentual:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{11}$$

Rentabilidad logarítmica

$$r_t = \left[\ln\left(\frac{P_t}{P_{t-1}}\right)\right] * 100 \quad (12)$$

Donde P_t indica el precio de cierre de en el día t. Además, es posible afirmar que cuando R_t es cercano a cero, ambos tipos de rentabilidades son aproximadamente iguales.

6.7.2 Volatilidad intradía.

Así mismo, Novales (2017) expone que si existe la necesidad de ver el comportamiento de un activo en intervalos regulares de tiempo es posible usar:

$$R_{t+j/m} = \ln(\frac{S_{t+j/m}}{P_{t+(j-1)/m}}) \quad (13)$$

Donde se suponen *m* observaciones diarias, para estimar la varianza diaria.

$$\sigma^2_{m,t+1} = \sum_{j=1}^{m} R^2_{t+j/m} \quad (14)$$

Expresión que podría empelarse para la validación de modelos de previsión de volatilidad o utilizarse directamente en la predicción de volatilidad. Donde a medida que crecen el número de observaciones intradía *m*, la medida de la varianza realizada confluye a la propia varianza diaria.

6.8 Métodos de ML para pronósticos

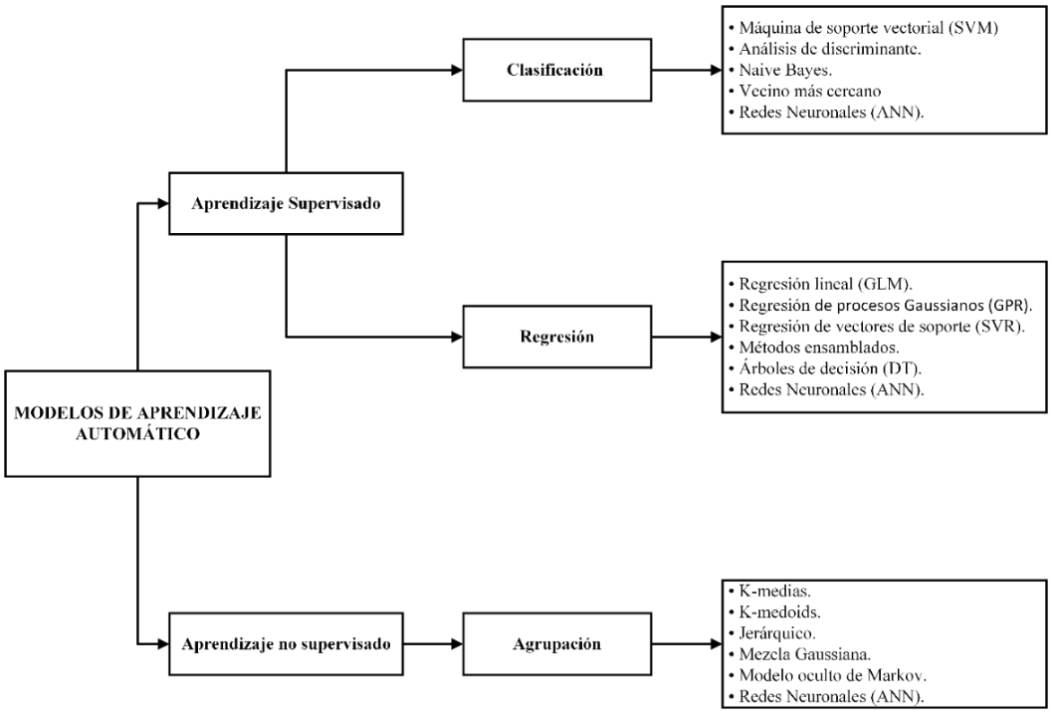
6.8.1 Machine Learning.

Es una rama de la inteligencia artificial permite que las máquinas aprendan sin estar programadas explícitamente para hacerlo. Es conocido como el reconocimiento de patrones, siendo capaz de transformar un patrón de datos en un programa de computadora idóneo para extraer conclusiones de nuevos conjuntos de datos en los que no se había entrenado previamente

(MathWorks, 2022). La Figura 9, presenta un resumen de los tipos de modelos de aprendizaje automático.

Figura 9

Métodos de aprendizaje automático.



Nota. Adaptado de (MathWorks, 2022).

Aprendizaje supervisado.

Estos algoritmos cuentan con una fase de preaprendizaje a partir de un sistema de etiquetas relacionadas con una serie de datos que permiten tomar decisiones o hacer previsiones. Utiliza un conjunto conocido de datos (llamado conjunto de datos de entrenamiento) para preparar un algoritmo y hacer predicciones usando un conjunto conocido de datos de entrada (llamado característica) y una respuesta conocida. El conjunto de datos de entrenamiento contiene datos de entrada etiquetados que coinciden con la respuesta deseada o el valor de salida. El algoritmo de

aprendizaje supervisado intentará construir un modelo a partir de ahora estableciendo la relación entre las características y los datos de salida y prediciendo los valores de respuesta del nuevo conjunto de datos (MathWorks, 2022).

Aprendizaje no supervisado

El aprendizaje supervisado es una técnica de aprendizaje automático que se utiliza para entrenar modelos que puedan predecir la salida correspondiente para nuevas entradas, y se basa en la utilización de un conjunto de datos etiquetados para entrenar el modelo y validar su capacidad de generalización. En diversos ámbitos se emplean estas técnicas, por ejemplo, el reconocimiento de voz, la detección de spam en correos electrónicos, el diagnóstico médico y la predicción del precio de las acciones. Entre los algoritmos de aprendizaje supervisado más populares se encuentran los árboles de decisión, las redes neuronales, la regresión lineal y la regresión logística. (MathWorks, Aprendizaje automatico, 2020).

6.9 Modelos clásicos

6.9.1 Modelo ARIMA.

Son modelos paramétricos que buscan mantener la representación de la serie en función de las interrelaciones temporales de sus elementos. Este tipo de modelos caracterizan las series como variables aleatorias o sumas o diferencias ponderadas o no ponderadas en la serie resultante, propuestos por Yule y Slutzky en 1920, siendo los cimientos de los procesos de media móviles y autorregresivos que han obtenido resultados significativamente notables gracias al libro de Box Jenkins sobre modelos ARIMA en 1970.

El modelo *ARIMA (p, d, q)* se define como:

$$\phi_p(L)\,\Delta^d\,Y_t = \delta + \theta_q\,(L)\,a_t \tag{15}$$

Donde el polinomio autorregresivo estacionario $\Phi_p(L)$ y el invertible de medias móviles $\Theta_q(L)$ no tienen raíces comunes.

Por predictor óptimo (o predicción óptima) es el mejor predictor en el sentido de que minimiza una función de pérdida particular. Por lo general se minimiza el Error Cuadrático Medio de Predicción (ECMP). Por lo tanto, YT () es el mejor predictor para minimizar ECMP cuando se cumplen las siguientes condiciones:

$$E\,[Y_{T+\ell} - Y_T(\ell)]^2 \;\leq\; E[Y_{T+\ell} - Y_T^*(\ell)]^2 \quad \forall Y_T^*(\ell) \tag{16}$$

Se puede demostrar que, el predictor por punto óptimo viene dado por la esperanza condicionada al conjunto de información:

$$Y_T(\ell) \;=\; E[Y_{T+\ell}|\,I_T] \;=\; E\big[Y_{T+\ell}|Y_{T,}\,Y_{T-1}, Y_{T-2}, Y_{T-3}, \ldots\big] \;=\; E_T[Y_{T+\ell}] \tag{17}$$

No hay garantía de que este valor esperado condicional sea una función lineal del historial de la serie. Sin embargo, si el proceso sigue una distribución normal, es posible demostrar que el valor esperado condicional se puede representar como una función lineal del conjunto de información de TI. Por lo tanto, bajo el supuesto de normalidad, el predictor óptimo en términos de minimizar ECMP es lineal. Si no se cumple este supuesto, la proyección lineal de $Y_{T+\ell}$ en su pasado proporcionaría el predictor óptimo dentro de la clase de predictores lineales. La predicción óptima por intervalo se construirá a partir de la distribución del error de predicción que, bajo el supuesto de que RBN (0, σ^2), es la siguiente:

$$e_T(\ell) \;=\; Y_{T+\ell} \;-\; Y_T(\ell) \;\sim\; N(0, V(e_T(\ell))) \tag{18}$$

Caracterizando se obtiene:

$$\frac{Y_{T+\ell} - Y_T(\ell) - 0}{\sqrt{V(e_T(\ell))}} \sim N(0,1) \tag{19}$$

De forma que el intervalo de predicción de probabilidad (1 − α) % es:

$$\left[Y_T(\ell) - N_{\alpha/2}\sqrt{V(e_T(\ell)),} \qquad Y_T(\ell) + N_{\alpha/2}\sqrt{V(e_T(\ell)),} \right] \tag{20}$$

6.10 Modelos ensamblados

Los modelos ensamblados o el *Ensemble Learning* parte de la premisa de combinar múltiples modelos de aprendizaje automático y combinar sus resultados, es decir, a partir de una serie de predicaciones con la combinación adecuada es posible lograr una mayor precisión en la generalización de un modelo de previsión, siendo superior a cualquier resultado que se pueda obtener individualmente (Brown, 2011). Además, según (Džeroski et al., 2009) en su libro "*Encyclopedia of Complexity and Systems Science"* afirman que este tipo de modelos presentan una mayor robustez y precisión cuando existen conjuntos de datos de diferentes fuentes donde el mismo tipo de objetos se describen en términos diversos atributos, por lo tanto, este tipo de modelos pueden ser de utilidad para la presente investigación donde se emplearán datos de diversos orígenes como variables exógenas.

Para construir un modelo de ensamble se debe definir primero el algoritmo base (e.g árboles regresivos, regresión lineal, redes neuronales, máquinas de soporte vectorial), el cual será el que generará varias predicciones que a medida que se obtengan se irán agregando. Finalmente, se añade el algoritmo de agregación quienes operan los *inputs* de los modelos base para que estos sean independientes. En este estudio se empelaron los siguientes modelos:

6.10.1 Random Forest (RF)

Formulado por (Breiman, 2001), los bosques aleatorios (RF) son un algoritmo de aprendizaje automático supervisado basado en árboles de decisión que se usa para problemas de regresión, donde incluso sin ajustar los hiperparámetros produce un gran resultado. En este algoritmo las puntuaciones Z de cada predictor de estrada sobre el atributo duplicado se determinan con el fin de conocer los factores cruciales de los predictores a través de las métricas de puntuación Z.

Para un vector de entrada particular x_t, una variable duplicada (o sombra) x'_t de orden aleatorio se produce. Luego, se eliminan las correlaciones y la aleatoriedad entre predictores de sombras y salidas. A continuación, se presenta el modelo:

$$MDA = \frac{1}{m_{tree}} \sum_{m=1}^{m_{tree}} \frac{\sum_{t \in OOB} I(y_t = f(x_t)) - \sum_{t \in OOB} I(y_t = f(x_t^n))}{|OOB|} \quad (21)$$

OOB es el error de predicción de cada uno de los ensayos de entrenamiento mediante agregación Bootstrap, $y_t = f(x_t^n)$ los valores previstos antes y después de la permutación e *I* representa la función indicadora. Las puntuaciones Z se calculan mediante:

$$Z - score = \frac{MDA}{std} \quad (22)$$

Donde *std* es la desviación estándar de las pérdidas de precisión y luego se calcula la puntuación Z máxima entre atributos duplicados (MZSA). En segundo lugar, las puntuaciones Z de los predictores (entradas) se igualan con los duplicados correspondientes y se valoran mediante distribución variable.

Finalmente, las entradas de $Z - score < MZSA$ se marcan como "sin importancia" y se eliminan permanentemente hasta que las entradas hayan $Z - score > MZSA$ estén marcados como "Confirmado", el algoritmo se detiene cuando se confirman todos los parámetros de entrada o se alcanza el umbral de iteración.

6.10.2 Extreme Gradient Boosting (XGB)

Propuesto por (Friedman, 2001), es un método de aprendizaje automático supervisado basado en la aproximación de funciones por medio de la optimización de perdidas específicas y la aplicación de técnicas de regularización, además, emplea su propio método de creación de árboles en los que la puntuación de similitud y la ganancia determinan los mejores de nodos. El modelo se representa matemáticamente de la siguiente manera:

$$S_C = \frac{(\sum_{i=1}^{n} Residual_i)^2}{\sum_{i=1}^{n}[Probabilidad\ Previa_i * (1 - Probabilidad\ Previa_i)] + \lambda} \quad (23)$$

Donde $Residual_i$ es el $valor\ observado - valor\ predecido$, $Probabilidad\ Previa_i$ es la probabilidad de un evento calculada y λ denota el parámetro de regularización. Luego de conocer el puntaje de similitud para cada hoja, se calcula la ganancia a través de la siguiente formula:

$$G = Left\ leaf_{similarity} + Right\ leaf_{similarity} - Root_{similarity} \quad (24)$$

Por lo tanto, la división de nodos con la mayor ganancia se elige la mejor división para el árbol. Finalmente, la manera en que se realiza el cálculo de la salida para el algoritmo es la siguiente:

$$Value = \frac{\sum_{i=1}^{n} Residual_i)}{\sum_{i=1}^{n}[Probabilidad\ Previa_i * (1 - Probabilidad\ Previa_i)] + \lambda} \quad (25)$$

6.11 Modelos basados en redes neuronales (RNA)

6.11.1 Modelo Long Short-Term Memory (LSTM)

Se desarrolla a partir de la red neuronal recurrente (RNN). En comparación con RNN, Long Short-Term Memory LSTM agrega un mecanismo de olvido, que también puede resolver el problema de explosión de gradiente. La estructura de LSTM, una unidad particular llamada celda de memoria es similar a un acumulador y una neurona cerrada. El siguiente paso secuencial tiene un peso paralelo y copia el valor real y la acumulación de su estado. El LSTM tiene un mecanismo de auto conexión controlado por una puerta de multiplicación que aprende y decide cuándo borrar el contenido de la memoria por otra unidad.

LSTM consta de tres puertas: puerta de entrada, puerta de salida y puerta de olvido. Su proceso de propagación hacia adelante se puede expresar mediante las ecuaciones:

Puerta de entrada:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (26)$$

Puerta de salida:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (27)$$

Puerta de olvido:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (28)$$

La información de la celda C_T y la información oculta h_t se actualizan mediante la ecuación:

$$\widehat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (29)$$

$$C_t = f_t \times C_{t-1} + i_t \times \widehat{C}_t \quad (30)$$

$$h_t = o_t \times \tanh(C_t) \quad (31)$$

Donde W es la matriz de pesos y b es el sesgo, que se actualizan y optimizan en el proceso de formación.

LSTM no pasa toda la información histórica hacia atrás, sino que olvida selectivamente parte del contenido histórico, memoriza parte del contexto histórico y agrega nueva información de entrada a la transferencia hacia atrás. Luego usa el algoritmo de retro propagación para actualizar los parámetros y optimizar el modelo (Man et al., 2022).

6.11.2 Máquinas de aprendizaje Extremas (ELM)

Se pueden emplear para entrenar redes neuronales Feedforward (SLFN) de una sola capa oculta. En ELM, la iniciación de los nodos ocultos se realiza aleatoriamente y antes de que se corrija sin ajuste iterativo. Además, los nodos ocultos de ELM ni siquiera necesitan ser neuronas por igual. El parámetro libre que tiene que aprender son las conexiones (o pesos) entre la capa de salida y la capa oculta. Como tal, ELM se desarrolla como un modelo lineal en el parámetro que, en última instancia, se ocupa de resolver un sistema lineal. A diferencia de los métodos tradicionales de aprendizaje, ELM es significativamente más eficiente y tiene una mayor tendencia a lograr un óptimo global. (Musatafa Abbas Abdul Albad, 2017)

La salida de un SLFN con nodos ocultos (aditivo o RBF nodos) puede ser representado por:

$$f_{\tilde{N}}(x) = \sum_{i=1}^{\tilde{N}} \beta_i G(a_i, b_i, x), \qquad x \in R^n, \quad a_i \in R^n \quad (32)$$

Donde a_i y b_i son los parámetros de aprendizaje de los nodos ocultos y βi el peso que conecta "i" el nodo oculto al nodo de salida. G (ai, bi, x) es la salida del nodo oculto con respecto a la entrada x. Para el nodo oculto aditivo con la función de activación g(x): R →R, G (a_i, b_i, x) es dado por:

$$G(a_i, b_i, X) = g(a_i * x + b_i), \qquad b_i \in R \tag{33}$$

Donde a_i es el vector de pesos que conecta la capa de entrada con la del nodo oculto y b_i es el sesgo del nodo oculto, a_i. x denota el producto interno de los vectores a_i y x en Rn. Para el nodo oculto RBF con función de activación g(x): R →R, G (a_i, b_i, x) es dado por:

$$G(a_i, b_i, x) = g(b_i \|x - a_i\|), \qquad b_i \in R^+ \tag{34}$$

Donde a_i y b_i son el centro y el factor de impacto del RBF. R+ indica el conjunto de todos los valores reales positivos. La red es un caso especial de SLFN con nodos RBF en su capa oculta. Cada nodo RBF tiene su propio centroide y factor de impacto, y su salida viene dada por una función radialmente simétrica de la distancia entre la entrada y el centro (N. Y. Liang et al., 2006).

6.11.3 Máquina de aprendizaje extremo secuencial en línea (OS-ELM)

El algoritmo OS-ELM tiene como objetivo hacer frente a las necesidades emergentes en varias aplicaciones de aprendizaje en línea. OS-ELM se considera un algoritmo rápido y se prefiere a otros algoritmos porque OS-ELM elimina el paso de reentrenamiento al recibir nuevos datos. OS-ELM puede aprender de los datos de entrenamiento a través de un mecanismo de fragmento por fragmento con longitud constante o variable (N. Y. Liang et al., 2006), en el algoritmo OS-ELM, hay tres capas o nodos que son la capa de entrada, la capa oculta y la capa de salida. La capa de entrada tiene las características extraídas, la capa oculta tiene sesgos y la capa de salida tiene las clases finales del algoritmo. La matriz de salida (H) de la capa oculta se calcula como la siguiente ecuación:

$$H = W_1 * X_1 + B_1 \tag{35}$$

Donde W indica los pesos de entrada que vinculan la capa de entrada a la capa oculta, X se refiere a las características en la capa de entrada, y B indica sesgos de la capa oculta. Los pesos de entrada (W) y sesgos ocultos (B) se generan aleatoriamente con un rango entre −1 y 1. Para N

muestras distintas arbitrarias (X_j, t_j), donde $X_j \in Rd$, y $t_j \in R^m$, redes neuronales Feedforward de una sola capa (SLFN) con n nodos ocultos y la función de activación g(x) se puede modelar matemáticamente como la siguiente ecuación:

$$f(X) = \sum_{i=1}^{n} \beta_i g(\omega_i * x_j + b_i) = t_j, \quad j = 1, 2, \dots, N \tag{36}$$

Además, se puede compactar y reescribir de la siguiente manera:

$$H\beta = T \tag{37}$$

Donde:

$$H = \begin{pmatrix} g(\omega_1 * x_1 + b_1) & \cdots & g(\omega_n * x_1 + b_n) \\ \vdots & \ddots & \vdots \\ g(\omega_1 * x_N + b_1) & \cdots & g(\omega_n * x_N + b_n) \end{pmatrix}_{N \times n}, \quad \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_n^T \end{bmatrix}_{n \times m}, T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{38}$$

Los pesos de salida ($\widehat{\beta}$) se estiman, entonces de acuerdo con la siguiente ecuación:

$$\widehat{\beta} = H^{\dagger} T \tag{39}$$

Donde $H^{\dagger}$ es la inversa generalizada de Moore-Penrose (pseudo inversa) de la matriz de salida de la capa oculta H, y se calcula de la siguiente manera:

$$H^{+} = (H^T H)^{-1} H^T \tag{40}$$

OS-ELM se ejecuta para aprender las muestras de entrenamiento de forma sucesiva e incremental. El proceso de aprendizaje de OS-ELM consta de dos pasos, el paso de inicialización y el paso de aprendizaje secuencial. En el paso de inicialización, la matriz de salida de la capa oculta H_0 y los pesos de salida de la inicial β_0 se calculan como las siguientes ecuaciones:

$$H_{k+1} = g(W * X_{k+1} + B) \tag{41}$$

$$P_0 = (H_0^T H_0)^{-1} \tag{42}$$

$$\beta_0 = P_0 H_0^T T_0 \tag{43}$$

En el paso de aprendizaje secuencial, la matriz de salida de la capa oculta H_{k+1} se actualizará para la nueva muestra. Además, la matriz de pesos de salida β_{k+1} se actualizará de acuerdo con las siguientes ecuaciones:

$$P_{K+1} = P_k - P_k H_{k+1}^T (I + H_{k+1} + P_k H_{k+1}^T)^{-1} H_{k+1} P_k \tag{44}$$

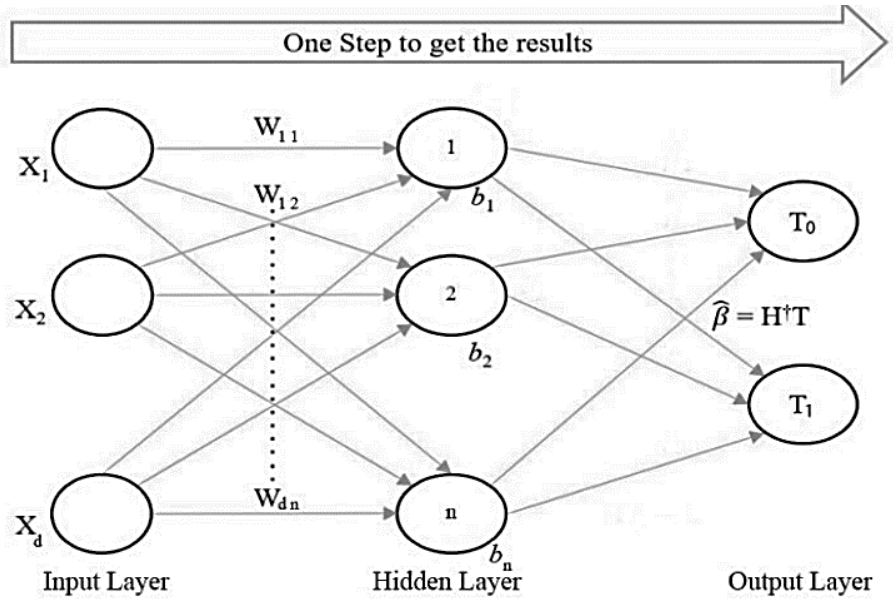
$$\beta_{k+1} = \beta_k + P_{k+1} H_{k+1}^T (T_{k+1} - H_{k+1} \beta^k \tag{45}$$

Cuando se entrenan todas las muestras, el OS-ELM se puede utilizar para la predicción de un vector de entrada desconocido. En el algoritmo OS-ELM, la capa de entrada se implementa aleatoriamente antes de realizar más cálculos para obtener la capa de salida y los resultados finales. En la Figura 10 se muestra la arquitectura del algoritmo OS-ELM, donde las clases finales se etiquetan comoT_0 y T_1.

La OS-ELM tiene dos grandes fases. La fase de refuerzo entrena a los SLFN utilizando la técnica ELM con algunos datos de entrenamiento en la fase de inicialización, y luego descarta estos datos de entrenamiento de refuerzo cuando finaliza la etapa de refuerzo. Después de la etapa de impulso, el OS-ELM aprende los datos de entrenamiento fragmento por fragmento y descarta todos los datos de entrenamiento cuando finaliza el proceso de aprendizaje de datos. La Figura 10 presenta la arquitectura del modelo OS-ELM.

Figura 10

Arquitectura del modelo OS-ELM.



Nota. Adaptado de (Taha Al-Dhief et al., 2021)

6.12 Análisis de sentimientos

El análisis de sentimiento o *sentiment analysis* consiste en evaluar las emociones, actitudes y opiniones. Las herramientas de análisis de sentimiento emplean tecnologías avanzadas de aprendizaje automático, como el procesamiento del lenguaje natural (NLP), el análisis de texto y la ciencia de datos, para identificar, extraer y estudiar información subjetiva, clasificando textos como positivo, negativo o neutral (QuestionPro, 2022).

6.12.1 Índices de noticias como variables exógenas

Según (Sha et al., 2019) modelar y pronosticar volatilidad en los mercados de valores es un ejercicio importante tanto en el ámbito académico como en la econofísica. Por lo tanto, en la literatura existen estudios que avalan el uso de variables exógenas como predictores, esto con el fin de mejorar la precisión de modelos de previsión. Para efectos de este estudio, se utilizarán tres

predictores exógenos basados en noticias antes y durante la pandemia de la COVID-19 y cuatro predictores adicionales que están estrechamente ligados con noticias sobre la pandemia del nuevo coronavirus.

Para efectos de la investigación se utilizarán datos sintéticos de las noticias relacionadas antes y durante la COVID-19 en Colombia obtenidos de Fred.org y RavenPack sitios web gratuitos que permiten encontrar datos estadísticos sobre noticias y noticias relacionadas la nueva pandemia de coronavirus respectivamente, incluidas las últimas noticias y temas de actualidad. RavenPack.com es un sitio web que rastrea la información más reciente sobre el nuevo coronavirus mediante el análisis de miles de fuentes en todo el mundo para identificar tendencias y patrones clave que surgen de las noticias. Una serie de índices rastrean los niveles de cobertura de los medios por país y resaltan algunos de los temas más importantes, incluidos los niveles de pánico, exageración de los medios y noticias falsas. A continuación, se presentan en primer lugar, los tres índices basados en noticias antes y durante la pandemia de la COVID-19. Finalmente, se describen los cuatro índices de las noticias relacionadas con el nuevo coronavirus.

6.12.1.1. Índice de volatilidad implícita (VIX). El VIX en es un índice creado por la *Junta de Opciones de Chicago (CBOE)*, siendo una medida de volatilidad del mercado de valores basado en el índice de opciones S&P500. Este índice se calcula en tiempo real y a menudo es conocido como el *índice del miedo*. Según (Demeterfi et al., 1999), el VIX busca representar la raíz cuadrada del precio de ejercicio en un intercambio de varianza de valor cero, esto se calcula de la siguiente manera:

$$P = (\sigma_R^2 - K) \tag{46}$$

Donde σ^2_R es la varianza de los rendimientos realizados durante el próximo período de 30 días y K es el precio de ejercicio igual a:

$$K = \sqrt{\sigma^2_{R\ t*(t+30)}} \quad (47)$$

Por lo tanto, este índice se ha convertido en referencia de la volatilidad esperada en el mercado de valores, ya que es el principal estimador a nivel mundial del sentimiento de los inversores, debido a su correlación con el precio de las acciones futuras tranzadas en el mercado de valores (Zhichao et al,. 2022).

6.12.1.2. Rastreador de volatilidad del mercado de valores respecto a enfermedades infecciosas (EMV-ID). Es un rastreador basado en periódicos que se basa a partir del índice de volatilidad (VIX) y con la volatilidad realizada de los rendimientos del S&P500(Baker et al,. 2019). Para construir el EMV-ID se especifican los términos en cuatro conjuntos de palabras de la siguiente manera:

- E: {económico, economía, financiero}
- M: {"mercado de valores", acciones, "Standard and Poors"}
- V: {volatilidad, volátil, incierto, incertidumbre, riesgo, riesgoso}
- ID: {epidemia, pandemia, virus, gripe, enfermedad, coronavirus, MERS, SARS, ébola, H5N1, H1N1}

En segundo lugar, se obtienen recuentos diarios de artículos periodísticos a través de tecinas de NLP que contengan al menos un término en cada E, M, V e ID en aproximadamente 3000 periódicos estadounidenses. En tercer lugar, se escalan los recuentos de EMV-ID sin procesar según el recuento de todos los artículos en el mismo día. En un paso final, se modifica la escala de forma multiplicativa de la serie, es decir, se iguala el nivel del VIX usando el índice EMV general y luego se escala este índice ID-EMV para reflejar la proporción de los artículos EMV-ID con

respecto al total de artículos EMV. Así mismo, (Li et al,. 2020) afirman que el EMV-ID tiene poder predictivo como variable exógena en volatilidades de mercados bursátiles.

6.12.1.3. Índice de incertidumbre económica relacionada con las políticas (EPU). Propuesto por (Baker et al., 2016), es un indice que se basa en noticias diaras tomadas de los archivos de periodicos del servicio *NewsBank*, esta base de datos contiene los archivos de miles de periódicos y otras fuentes de noticias de todo el mundo. A través de NLP se cuantifica la cantidad de artículos que contienen al menos un término de cada uno de los 3 conjuntos de términos. El primer conjunto es *económico o economía*. El segundo es *incierto o incertidumbre*. El tercer conjunto es *legislación o déficit o regulación o congreso o reserva federal o casa blanca*.

Al mismo tiempo, se adicionan datos completos de actualizaciones diarias que proporcionan un conjunto continuo de 60 observaciones por día, lo que brinda el conjunto completo de resultados utilizados para actualizar el Índice basado en noticias diarias. Además, en (Dutta et al,. 2021) exponen que el EPU tiene relación con el modelado de la prevision de voaltilidad de commodities como el petroleo y monedas digitales como el Bitcoin.

6.12.1.4. Índice de pánico (PI). Mide el nivel de conversaciones en las noticias que hacen referencia al pánico o la histeria relacionados con el coronavirus. Los valores oscilan entre 0 y 100, donde un valor de 7.00 indica que el 7 % de todas las noticias a nivel mundial hablan de términos relacionados con el pánico y el COVID-19. Cuanto más alto es el valor del índice, más referencias al pánico se encuentran en los medios. El índice se calcula como el recuento diario de historias distintas que mencionan palabras clave de pánico y coronavirus, dividido por el recuento diario total de historias distintas.

6.12.1.5. Índice de exageración de los medios (HY). Evalúa el porcentaje de noticias que hablan sobre el nuevo Coronavirus. Los valores oscilan entre 0 y 100, donde un valor de 75.00 indica que el 75 % de todas las noticias a nivel mundial hablan de la COVID-19. El índice se calcula como el recuento diario de historias distintas que mencionan el coronavirus, dividido por el recuento diario total de historias distintas.

6.12.1.6. Índice Noticias Falsas (FNI). Estima el nivel de comentarios de los medios sobre el nuevo virus que hace referencia a información errónea o noticias falsas relacionados con la COVID-19. Los valores oscilan entre 0 y 100, donde un valor de 2.00 indica que el 2 por ciento de todas las noticias a nivel mundial hablan de noticias falsas y COVID-19. Cuanto mayor sea el valor del índice, más referencias a noticias falsas se encuentran en los medios. El índice se calcula como el recuento diario de historias distintas que mencionan palabras clave relacionadas con noticias falsas y el coronavirus, dividido por el recuento diario total de historias distintas.

6.12.1.7. El índice de sentimiento del país (CSI). Mide el nivel de sentimiento de todas las entidades en un país determinado relacionado con las noticias junto y el coronavirus. El índice oscila entre -100 y 100, donde un valor de 100 es el sentimiento más positivo, -100 es el más negativo y 0 es neutral.

6.13 Formas de normalización de datos.

De acuerdo con (Asesh, 2022), los métodos de normalización tienen un impacto significativo en la precisión de los modelos de aprendizaje automático, ya que permiten expresar las variables como cantidades adimensionales. Además, cuando existen datos de distribuciones multivariados con medias y varianzas diferentes, la normalización se convierte en una etapa fundamental para asegurar que no se vea afectada la capacidad de previsión de un modelo. A

continuación, se presentan dos tipos de normalización comúnmente empleadas para modelos de aprendizaje automático.

6.13.1 Normalización min-max.

La normalización *min-max* incluye eliminar el mínimo y dividir por la diferencia entre los valores mínimo y máximo del punto de datos de la serie temporal para transformar los datos en valores entre 0 y 1. Estos enfoques de normalización son sencillos y se usan ampliamente para eliminar el sesgo en las variables con valores más grandes cuando se comparan con datos de valores más bajos.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{48}$$

6.13.2 Normalización escalado estándar

Es una técnica de normalización en donde los datos transformados no poseen escala al convertir la distribución estadística de los datos en $\mu = 0$ y $\sigma = -1$ y está descrita por:

$$z = \frac{x - \mu}{\sigma} \tag{49}$$

Donde μ y σ representan la media y desviación poblacional.

6.14 Métricas de ajuste

La evaluación y validación de los modelos de aprendizaje automático es una fase fundamental, ya que permite determinar el rendimiento y la capacidad de generalización respecto a la previsión de valores futuros de cada modelo. En la actualidad, existen una serie de medidas de error que permiten cuantificar el desempeño de los algoritmos de aprendizaje automático. Para efectos de esta investigación se emplearán los siguientes:

Raíz cuadrada del error medio (RMSE), permite probar el desempeño de manera cuantitativa a través del cálculo de la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado, indicando el ajuste absoluto del modelo de datos y cuan cerca están los puntos observados de los valores predichos de un modelo. Un valor bajo del RMSE indica un mejor ajuste, así mismo, es una buena medida para informar la precisión con que un modelo predice. Sin embargo, el RMSE amplía y penaliza con mayor fuerza los errores mayor magnitud, por lo que su interpretación por sí sola no es suficiente, por lo que debe ir acompañada de otras medidas de error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{50}$$

Error medio absoluto (MAE), es una métrica de evaluación de modelos de aprendizaje automático. El error absoluto medio de un modelo con respecto a un conjunto de prueba es la media de los valores absolutos de los errores de predicción individuales en todas las instancias del conjunto de prueba . Cada error de predicción es la diferencia entre el valor real y el valor previsto para la instancia.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{51}$$

Error porcentual medio absoluto (MAPE)¸ es la medida más común utilizada para cuantificar el error en modelos de aprendizaje automático, ya que las unidades de la variable se escalan a unidades porcentuales, su cálculo se hace a través del error absoluto promedio para cada periodo de tiempo menos los valores reales divididos por los reales. Además, debido a que se utilizan errores porcentuales absolutos, se evita el problema de errores positivos y negativos que se anulan entre sí.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{O_i - P_i}{O_i}\right| \quad (52)$$

Donde n es el número de observaciones. O_i y P_i son los valores medidos y pronosticados para la *i*-ésima observación respectivamente.

Con el fin de realizar una comparación y determinar qué modelo de aprendizaje automático se ajusta mejor respecto a los datos observados se emplea el *índice de mejora del RMSE*.

$$RI_{RMSE=} = \frac{RMSE_R - RMSE_E}{RMSE_R} \quad (53)$$

Donde $RMSE_R$ y $RMSE_E$ son los valores de desempeño del modelo que se tome de referencia y del que se desea comparar.

7. Caso de estudio

7.1 Selección, recopilación y preprocesamiento de datos.

En esta fase se identifica el conocimiento relevante y prioritario, identificando un grupo de datos objetivo, donde se puede seleccionar todo el conjunto de datos o una muestra representativa de este, sobre el cual se desarrollará la investigación. Asimismo, se analiza la calidad de los datos, se sustraen los datos atípicos, se emplean estrategias para el manejo de datos desconocidos, nulos, duplicados y técnicas estadísticas como media, moda, máximo, mínimo y regresiones para su reemplazo.

7.1.1 Selección y Recopilación de datos.

Para efectos de esta investigación se estudiarán dos conjuntos de datos principales, donde el primero estará comprendido en el nodo antes de la pandemia del nuevo coronavirus (01/01/2010 – 31/12/2019) y el segundo durante la pandemia de la COVID-19 (01/01/2020-06/07/2022), los cuales, a su vez, se dividen en un total de 10 variables que serán ingresadas como *inputs* en el

modelo de aprendizaje automático que posteriormente se desarrollará. A continuación, se presenta el primer conjunto de datos donde estarán las *variables económicas* (VE), la cuales se conforman de la siguiente manera:

(Brockwell & Davis, 2016) exponen en su libro *"Introduction to Time Series and Forecasting"* la importancia de que todas las series de tiempo que vayan a ser empleadas para modelos de previsión deben compartir la misma temporalidad, es decir, estas deben tener los mismos recuentos (diarios, mensuales, anuales, etc.), por tal motivo, se selecciona la *tasa representativa del mercado (TRM)* y *el precio interno del café colombiano* debido a que presentan recuentos diarios en su medición.

Precio interno del café colombiano, este lo proporciona la Federación Nacional de Cafeteros de Colombia, donde ofrece a todos los cafeteros la garantía de compra, a partir de la publicación de un precio base de mercado, sus unidades se encuentran en COP (pesos colombianos), su cálculo se realiza de acuerdo con la cotización de cierre en la Bolsa de Nueva York del día, la tasa de cambio del día y el diferencial o prima de referencia para el café colombiano (Federación Nacional de Cafeteros de Colombia, 2022). Para efectos de este estudio, se toman los precios históricos de cierre diario del precio interno por carga de 125 kg desde el *01/01/2010* hasta *06/07/2022*, dicho conjunto de datos se encuentra alojado en la página de la FNC como datos de uso abierto.

Tasa representativa del mercado (TRM), es la cantidad de pesos colombianos (COP) por cada dólar estadounidense (USD). Su cálculo se realiza en base a las operaciones de compra y venta de divisas que se transan en el mercado financiero que se negocian en el mercado cambiario de Colombia con cumplimiento del mismo día que se pacta la negociación de las divisas (Banco de la República de Colombia, 2022). Para la presente investigación se toma la serie histórica de cierre

diaria de la *TRM* desde el *01/01/2010* hasta *06/07/2022,* el conjunto de datos se encuentra alojado en la página web del Banco de la República de Colombia.

En segundo lugar, se encuentran los índices de noticias relacionadas con la pandemia de la COVID-19, los cuales se presentan a continuación:

Índices antes y durante de la pandemia de la COVID-19. Como se mencionó en el capítulo 6 donde se explica que existen ciertos índices que se relacionan con las fluctuaciones e incertidumbre del mercado de valores. Por lo tanto, se usarán tres índices que logran captar el sentimiento de los inversores y personas en general por medio de análisis de texto de los periódicos más importantes del mundo, así mismo, a través del comportamiento y los cambios que se presentan en el mercado de valores. Los *índices de noticias* (IN) que se usarán para el evento antes y durante la pandemia de la COVID-19 tendrán la misma temporalidad, comprendida desde *(01/01/2010)* - *(06/07/2022)* y serán extraídos de Fred.com quienes proporcionan el acceso y uso de manera libre. Los índices se presentan a continuación:

El *índice de volatilidad implícita (VIX),* refleja las fluctuaciones macroeconómicas afectadas por los cambios en las redes sociales o en la percepción de los inversores.

El *índice de volatilidad del mercado de valores de enfermedades infecciosas* basado en el análisis de texto de los periódicos más reputados del mundo que contengan palabras relacionadas a *economía, mercado, volatilidad* y *enfermedades*.

El *índice de incertidumbre política* cuantifica la cobertura periodística de la incertidumbre económica relacionada con los resultados de búsqueda de 10 periódicos importantes de Estados Unidos.

En cuanto al evento durante la pandemia, se emplearán las observaciones tanto de los tres índices anteriormente mencionados como de los *índices durante la pandemia de la COVID-19* (IC). Estos índices serán tomados de RavenPack.com, empresa que brinda datos para uso de manera abierta, además, son quienes rastrean la información más reciente sobre el nuevo coronavirus mediante análisis y procesamiento de lenguaje natural (NLP) de distintas fuentes de información alrededor del mundo para identificar patrones y tendencias clave que surgen de las noticias. En este sitio web, se encuentran una serie de índices que rastrean los niveles de cobertura de los medios por país y resaltan algunos de los temas más importantes, los índices que se usarán como variables en el modelo son: *niveles de pánico (PI)*, *exageración de los medios (MHI)*, *noticias falsas (FNI)* y *el sentimiento nacional (SI)*. Los índices que se usarán para el desarrollo de la investigación estarán representados como variables cuantitativas desde inicio de la pandemia, es decir los valores de cierre diarios desde *01/01/2020* hasta *06/07/2022.*

Finalmente, se encuentra el ultimo conjunto de datos, el cual hace referencia al *índice de volatilidad de los futuros del café colombiano* (Vol). Este índice se calcula a partir de la serie de datos históricos de los futuros del café colombiano que se recopilan en el sitio web Investing.com quienes proporcionan los datos históricos de cierre diario de manera abierta y de libre uso, en el periodo comprendido entre *01/01/2010 y 06/07/2022*. Para calcular el factor de volatilidad según (Wang et al., 2018) en este tipo de estudios, es posible tomar los rendimientos absolutos del precio de los futuros como indicador de volatilidad, el cual se calcula de la siguiente manera:

$$r_t = \left[\ln\left(\frac{P_t}{P_{t-1}}\right)\right] * 100 \tag{54}$$

Donde P_t indica el precio de cierre de los futuros del café colombiano en el día t.

En la Tabla 2 y 3 se presenta un resumen de los conjuntos de datos seleccionados que serán objeto de estudio en esta investigación, con su respectiva fuente, temporalidad y unidades.

Tabla 2

Datos extraídos antes la pandemia

Variable	Fuente	Temporalidad	Unidades
Índice de Volatilidad	Investing.com	*(01/01/2010)* - *(31/12/2019)*	USD
Precio interno café colombiano (carga 125 Kg)	Federación Nacional de Cafeteros (FNC)	*(01/01/2010)* - *(31/12/2019)*	COP/ 125 kg
Tasa Representativa del Mercado (TRM)	Banco de la República de Colombia	*(01/01/2010)* - *(31/12/2019)*	COP/USD
Índice de volatilidad del mercado de valores de USA (VIX)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2010)* - *(31/12/2019)*	Adimensional
Volatilidad del Mercado de Valores de Enfermedades Infecciosas (EMV-ID)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2010)* - *(31/12/2019)*	Adimensional
Índice de Incertidumbre de Política-Económica (EPU)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2010)* - *(31/12/2019)*	Adimensional

Tabla 3

Datos extraídos durante la pandemia.

Variable	Fuente	Temporalidad	Unidades
Índice de Volatilidad	Investing.com	*(01/01/2020)* - *(06/07/2022)*	USD
Precio interno café colombiano (carga 125 Kg)	Federación Nacional de Cafeteros (FNC)	*(01/01/2020)* - *(06/07/2022)*	COP/ 125 kg
Tasa Representativa del Mercado (TRM)	Banco de la República de Colombia	*(01/01/2020)* - *(06/07/2022)*	COP/USD

Variable	Fuente	Temporalidad	Unidades
Índice de volatilidad del mercado de valores de USA (VIX)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2020) - (06/07/2022)*	Adimensional
Volatilidad del Mercado de Valores de Enfermedades Infecciosas (EMV-ID)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2020) - (06/07/2022*	Adimensional
Índice de Incertidumbre de Política-Económica (EPU)	Banco de la Reserva Federal de St. Louis (FRED)	*(01/01/2020) - (06/07/202)*	Adimensional
Índice de Pánico (PI)	RavenPack.com	*(01/01/2020) - (06/07/2022)*	Porcentaje
Índice de Exageración de los medios (MHI)	RavenPack.com	*(01/01/2020) - (06/07/2022)*	Porcentaje
Índice de Noticias Falsas (FNI)	RavenPack.com	*(01/01/2020) - (06/07/2022)*	Porcentaje
Índice del Sentimiento Nacional (SI)	RavenPack.com	*(01/01/2022) - (06/07/2022)*	Porcentaje

7.1.2 Fase 3: Preprocesamiento de datos

Las fases de preprocesamiento y limpieza de datos son tareas cruciales para que un conjunto de datos se pueda emplear para modelos de aprendizaje automático. Además, estas fases son las que más tiempo requieren, ya que son un factor fundamental en cuanto a la calidad de los hallazgos y conclusiones a las que se puedan llegar.

En esta investigación se siguió la metodología propuesta por (Han et al., 2012) para el preprocesamiento de datos en modelos de aprendizaje automático, donde la primera actividad es la detección de valores faltantes.

7.1.2.1. Primera actividad: tratamiento de datos faltantes. Para dicha actividad, se utilizó interpolación spline cúbica para las variables: TRM, VIX e Vol, donde según (Xu & Zhang, 2022) es un excelente método para aproximar datos faltantes en recuentos diarios y no afecta los resultados en modelos de previsión. Por otra parte, para los índices durante la pandemia de la

COVID-19 se tuvo otro tratamiento, debido a que RavenPack.com en su base de datos presenta los valores nulos como recuentos vacíos y el lenguaje de programación utilizado en este trabajo los cuantificaba como valores faltantes o *NaN*. Por tal motivo, para las variables PI, MHI y FNI se realiza la imputación con la función *.fillna()* la cual imputa los valores según el valor especificado que para este caso es *cero*. Las demás variables no cuentan con valores faltantes, las que si fueron imputadas se encuentran en la Tabla 4.

Tabla 4

Variables imputadas.

Variable	Datos faltantes	Porcentaje	Método
Índice de Volatilidad	9	0,20%	*Spline cúbico*
Precio interno del café	10	0,22%	*Spline cúbico*
VIX	9	0,20%	*Spline cúbico*
PI	26	2,91%	*.fillna(0)*
MHI	18	2,00%	*.fillna(0)*
FNI	146	18,91%	*.fillna(0)*

7.1.2.2. Segunda actividad: concatenación de los conjuntos de datos. Como se mencionó anteriormente, este estudio se divide en dos partes, esto con el fin de evaluar el comportamiento de la volatilidad del café colombiano antes y durante la pandemia del nuevo coronavirus basado en las noticias y variables económicas. Por tal motivo, la segunda actividad fue la concatenación de los Datasets necesarios para evaluar la capacidad de generalización de los modelos que se formularán en la presente investigación según qué variables se incluyan, ya que (Weng et al., 2021) afirman que es importante evaluar la capacidad de los modelos a través de la inclusión de las variables exógenas gradualmente debido a que ciertos eventos podrán tener injerencia en las fluctuaciones de la volatilidad.

De acuerdo a lo anterior, para el evento antes de la pandemia resultaron 4 conjutos de datos, mientras que, para las variables empleadas durante la pandemia resultaron 8 Datasets. A continuación, se presenta la división de los conjuntos de datos:

Conjuntos de datos concatenados evento antes de la pandemia

1. Vol. Futuros.
2. Vol. Futuros y Variables Económicas.
3. Vol. Futuros ~~y~~ e Índices de Noticias.
4. Vol. Futuros, Variables Económicas e Índices de Noticias.

Conjuntos de datos concatenados evento durante la pandemia

1. Vol. Futuros.
2. Vol. Futuros y Variables Económicas.
3. Vol. Futuros y Índices de Noticias.
4. Vol. Futuros y Índices de la COVID-19.
5. Vol. Futuros, Variables Económicas e Índices de Noticias.
6. Vol. Futuros, Variables Económicas e Índices de la COVID-19.
7. Vol. Futuros, Índices de Noticias e Índices de la COVID-19.
8. Vol. Futuros, Variables Económicas, Índices de Noticias e Índices de la COVID-19.

7.1.2.3. Tercera actividad: tratamiento de datos atípicos. Para el tratamiento de datos atípicos, de acuerdo con *(Beltrán Martínez, 2009)* existen algoritmos robustos como aquellos basados en árboles de decisión y redes neuronales con la capacidad de ser indiferentes ante los datos atípicos por lo que es posible ignorarlos. Por tal motivo, para esta investigación no se eliminaron datos atípicos puesto que por la naturaleza de los modelos que se van a emplear y por

la información relevante que estos pueden representar se decide ignorarlos y seguir con la siguiente actividad.

7.1.2.4. Cuarta actividad: caracterización de los datos. Para esta actividad se realiza la observación de diferentes medidas de tendencia central, tanto de las variables del evento antes como durante la pandemia de la COVID-19, esto con el fin de entender el comportamiento de Vol, Ve, IN e IC.

Tabla 5

Características de los datos antes de la pandemia.

Variable	Recuentos	Media	Desviación estándar	Mediana	Mínimo	Máximo
Vol	3.653	1,21	1,14	1,17	0,00	11,79
TRM	3.653	2.439,80	587,56	2.407,26	1.748,41	3.522,48
P. Interno	3.653	7,45E+05	1,45E+05	7,13E+05	3,70E+05	1,16E+06
VIX	3.653	16,85	5,74	16,14	8,75	49,68
EMV-ID	3.653	0,43	0,90	0,50	0,00	15,91
EPU	3.653	109,71	63,82	102,24	3,32	586,55

Tabla 6

Características de los datos durante la pandemia.

Variable	Recuentos	Media	Desviación estándar	Mediana	Mínimo	Máximo
Vol	918	1,36	1,25	1,04	0,00	9,56
TRM	918	3.760,04	192,56	3.777,17	3.253,89	4.259,86
P. Interno	918	1,45E+06	4,57E+05	1,22E+06	8,13E+05	2,31E+06
VIX	918	24,70	9,32	22,95	12,10	82,69
EMV-ID	918	17,42	13,33	14,17	0,00	112,93
EPU	918	205,90	138,04	157,76	20,63	861,10
PI	918	3,83	6,46	1,79	0,00	73,53
MHI	918	36,96	18,61	36,26	0,00	93,77

Variable	Recuentos	Media	Desviación estándar	Mediana	Mínimo	Máximo
FNI	918	0,92	2,86	0,35	0,00	46,49
SI	918	-4,95	8,01	-3,76	-32,82	22,52

Según las Tablas 5 y 6 es posible afirmar que la media de la volatilidad del café colombiano, las variables económicas y los índices de las noticias fueron mayores durante la pandemia COVID-19, lo que quiere decir que los recuentos de cada una de estas variables incrementaron. Así mismo, sucede con la desviación lo que nos indica una mayor dispersión de los datos durante la pandemia del nuevo coronavirus.

Además, se evidenció que, aunque en el periodo de la pandemia hubo mayor variabilidad y dispersión en la volatilidad del café colombiano, el pico más fuerte se observó antes de la misma, con un máximo de 11.79 lo que representa un 18.91% más alto. Por el contrario, para la TRM, precio interno, VIX, EMV-ID y EPU se presentó un aumento durante el evento de la COVID-19 de 18.91%, 20.93%, 98.13%, 66.45%, 609.81% y 46.81% respectivamente, siendo el índice de la volatilidad del mercado de valores respecto a enfermedades infecciosas (EMV-ID) el que tuvo el mayor incremento, debido a las connotaciones y repercusiones que tiene la pandemia por el SARS-COV-2 sobre la sensibilidad de este índice.

7.1.2.5. Quinta actividad: análisis de estacionariedad. Para este estudio se utilizó la función *adfuller()* creada por (Seabold & Perktold, 2010) donde se realiza un análisis de estacionariedad por medio de la prueba Dickey Fuller Aumentada (ADF), la cual es una prueba estadística comúnmente usada en el análisis de series temporales con el fin de probar si una serie de tiempo dada es estacionaria o no. Esta prueba consiste en que la H_0 asume la no estacionariedad, por lo que si llegase a existir evidencia significativa para rechazar H_0 estaríamos afirmando que la serie de tiempo es estacionaria.

Por lo tanto, surge la necesidad de hacer este análisis debido a que modelos como los autorregresivos o modelos clásicos como por ejemplo el modelo ARIMA requieren este análisis para determinar si una serie temporal es estacionaria y, por lo tanto, adecuada para ser modelada por este método. Si se encuentra evidencia de una raíz unitaria, se requerirá un proceso de diferenciación para transformar la serie temporal en una serie estacionaria antes de modelarla. En la Tabla 7 se muestran los resultados de la prueba ADF para las variables que serán objeto de estudio.

Tabla 7

Resultados de la prueba ADF.

Variable	**Test Estadístico**	**V.C. 1%**	**V.C. 5%**	**V.C. 10%**	**Valor *p<5%***	**Resultado**	**Estacionaria**
Vol. Futuros	-8,159	-3,432	-2,862	-2,567	0,001	*Rechazar*	*Si*
TRM	-1,179	-3,432	-2,862	-2,567	0,388	*No Rechazar*	*No*
P. Interno	-3,534	-3,432	-2,862	-2,567	0,007	*Rechazar*	*Si*
VIX	-5,710	-3,432	-2,862	-2,567	0,001	*Rechazar*	*Si*
EMV-ID	-3,819	-3,432	-2,862	-2,567	0,002	*Rechazar*	*Si*
EPU	-4,136	-3,432	-2,862	-2,567	0,001	*Rechazar*	*Si*
PI	-22,156	-3,438	-2,865	-2,568	0,001	*Rechazar*	*Si*
MHI	-2,113	-3,438	-2,865	-2,568	0,239	*No Rechazar*	*No*
FNI	-29,269	-3,438	-2,865	-2,568	0,001	*Rechazar*	*Si*
SI	-4,660	-3,438	-2,865	-2,568	0,001	*Rechazar*	*Si*

Según la Tabla 7 es posible afirmar que para las variables TRM y MHI al tener un valor p>5% poseen un comportamiento NO estacionario, es decir que estas no varían dependiendo el tiempo. Mientras que para las demás variables de este estudio al tener que p<5% existe evidencia significativa para rechazar H_0, por lo que es posible afirmar que su comportamiento es estacionario.

7.1.2.6. Sexta actividad: análisis de correlación. A través del coeficiente de correlación de Pearson por medio de la librería *pandas* con la función *.corr()*, la cual cuantifica la asociación

lineal entre dos variables numéricas, donde existe un rango [-1+1], donde +1 indica una correlación positiva perfecta y -1 una negativa perfecta. Para esta actividad se realiza el análisis respecto a las observaciones de la volatilidad de los futuros del café colombiano tanto para el evento antes como durante la pandemia de la Covid-19. En las Figuras 11 y 12 se presentan los gráficos de correlación.

Figura 11

Análisis de correlación volatilidad evento antes de la pandemia.

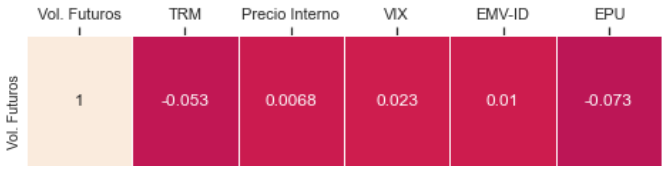


Figura 12

Análisis de correlación volatilidad evento durante la pandemia.

Según las Figuras 11 y 12, es posible determinar que las variables económicas, los índices de noticias y de al COVID-19 poseen una correlación mayor durante la pandemia, que, aunque no sean correlaciones fuertes si existió un crecimiento, especialmente en el Índice de Volatilidad del mercado de valores (VIX). Así mismo, cabe resaltar que la TRM presentó un cambio en su comportamiento respecto al evento antes de la pandemia, ya que paso de tener una correlación negativa débil, a tener un positiva débil, es decir, en una pequeña medida está relacionada directamente.

En las Tablas 8 y 9, se presenta un resumen general de la correlación de las demás variables que serán objeto de estudio en esta investigación.

Tabla 8

Análisis de correlación variables empleadas evento antes de la pandemia

Variable 1	Variable 2	Correlación	Tipo
TRM	Precio Interno	26,00%	*Débil +*
TRM	VIX	-25,00%	*Débil -*
TRM	EMV-ID	4,60%	*Ninguna*
TRM	EPU	-29,00%	*Débil -*
Precio Interno	VIX	23,00%	*Débil +*
Precio Interno	EMV-ID	2,20%	*Ninguna*
Precio Interno	EPU	5,40%	*Ninguna*
VIX	EMV-ID	2,50%	*Ninguna*
VIX	EPU	34,00%	*Débil +*
EMV-ID	EPU	-2,70%	*Ninguna*

Tabla 9

Análisis de correlación variables empleadas evento durante la pandemia

Variable 1	Variable 2	Correlación	Tipo	Variable 1	Variable 2	Correlación	Tipo
TRM	P. Interno	66,00%	*Fuerte +*	VIX	MHI	37,00%	*Débil +*
TRM	VIX	37,00%	*Débil +*	VIX	FNI	2,70%	*Ninguna*
TRM	EMV-ID	21,00%	*Débil +*	VIX	SI	-14,00%	*Débil -*
TRM	EPU	17,00%	*Débil +*	EMV-ID	EPU	66,00%	*Fuerte*
TRM	PI	4,60%	*Ninguna*	EMV-ID	PI	22,00%	*Débil +*
TRM	MHI	14,00%	*Débil +*	EMV-ID	MHI	43,00%	*Débil +*
TRM	FNI	0,14%	*Ninguna*	EMV-ID	FNI	5,70%	*Ninguna*
TRM	SI	-5,10%	*Ninguna*	EMV-ID	SI	-19,00%	*Débil -*
P. Interno	VIX	-10,00%	*Débil -*	EPU	PI	19,00%	*Débil +*
P. Interno	EMV-ID	-23,00%	*Débil -*	EPU	MHI	53,00%	*Fuerte +*
P. Interno	EPU	-39,00%	*Débil -*	EPU	FNI	5,20%	*Ninguna*
P. Interno	PI	-11,00%	*Débil -*	EPU	SI	-33,00%	*Débil -*
P. Interno	MHI	-34,00%	*Débil -*	PI	MHI	43,00%	*Débil +*
P. Interno	FNI	-3,80%	*Ninguna*	PI	FNI	5,20%	*Ninguna*
P. Interno	SI	27,00%	*Débil +*	PI	SI	-6,90%	*Ninguna*
VIX	EMV-ID	65,00%	*Fuerte +*	MHI	FNI	17,00%	*Débil +*
VIX	EPU	56,00%	*Fuerte +*	MHI	SI	-36,00%	*Débil -*
VIX	PI	17,00%	*Débil +*	FNI	SI	-11,00%	*Débil -*

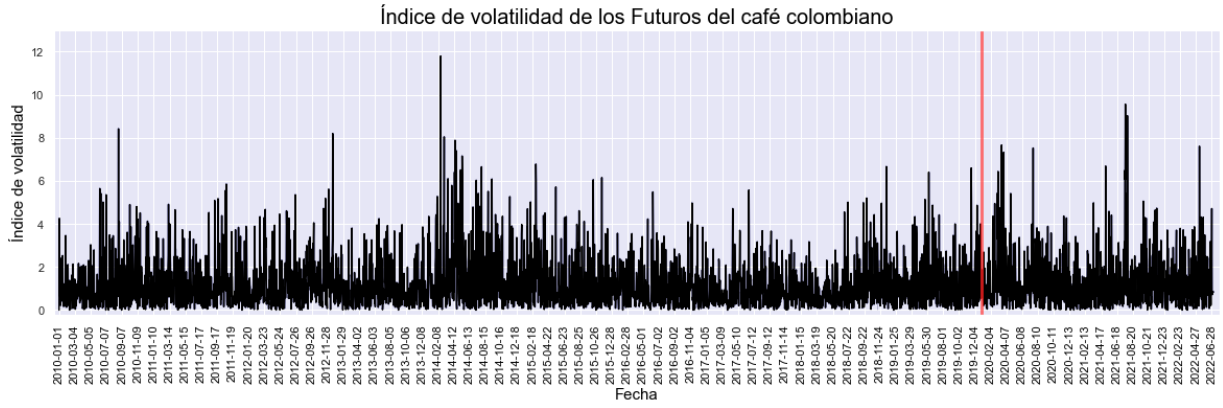
Según las Tablas 8 y 9, existió un aumento durante la pandemia de las correlaciones entre VE e IN, pasando de tener correlaciones débiles a fuertes directas como, por ejemplo: la TRM y

Precio interno y tanto VIX como EPU con EMV-ID. Por otra parte, el comportamiento de los índices de las COVID-19 presentó una correlación muy pequeña respecto a las demás variables.

7.1.2.7. Séptima actividad: creación de la serie de tiempo. Para esta actividad se crean las series de tiempo para todas las variables, sin embargo, para efectos de este estudio en el presenten escrito, se mostrará únicamente la visualización de la serie de tiempo de la volatilidad de los futuros del café colombiano. Donde la línea vertical roja señala el inicio de la pandemia de la COVID-19 el 01/01/2020.

Figura 13

Serie de tiempo de volatilidad de los futuros del café colombiano.



A partir de la serie de tiempo de la volatilidad de los futuros del café colombiano en la Figura 12, es posible apreciar que durante el periodo 08/02/2014 a 13/06/2014 existió una alta volatilidad sostenida, la cual coincide con el mismo coportamiento de la TRM y el Precio interno del café colombiano, así mismo, sucede con los picos sostenidos durante la pandemia del 04/02/2020 – 07/04/2020 y del 30/06/2021 – 20/08/2021 donde a las variables económicas se les suman los índices de las noticias y en menor medida los de la COVID-19, ya que como se vio en el análisis de correlación hubo una mayor influencia de dichas variables en el periodo que se desarrolló el nuevo coronavirus.

7.1.2.8. **Octava actividad: Normalizacion de los datos.** Para esta actividad se empleó el escalado estándar o future scaling, también conocido en el campo de la estadística como normalización con z score, el cual consiste en transformar las observaciones para que cada variable este en el mismo rango, garantizando que ninguna serie de tiempo o variable tenga un mayor peso, además, por este método se obtiene una distribución de los datos con media cero y varianza unitaria, logrando así, que el alistamiento y el entrenamiento en modelos de aprendizaje automático sean más rápidos y efectivos (Scikit-learn: Machine Learning in Python, 2011). La ecuación 62 normaliza los datos.

$$z = \frac{x - \mu}{\sigma} \tag{62}$$

7.2 Procesamiento y desarrollo de los modelos.

Para esta fase, en primer lugar, se tuvieron en cuenta dos modelos ensamblados RF y XGB, seguidamente de dos modelos basados en redes neuronales LSTM y ELM. Se toma como punto de partida la metodología propuesta por (Weng et al., 2021), donde los modelos RF, XGB y ELM se toman como referencia para comparar el rendimiento del modelo principal LSTM.

El modelo LSTM se toma como modelo principal ya que ha sido ampliamente usado tanto para el análisis de sentimiento de los inversores antes y durante la pandemia de la COVID-19, como la incidencia de las noticias en el pronóstico de volatilidad de futuros de commodities y petróleo junto a variables exógenas (G. Li et al., 2020), (Y. Li et al., 2021) y (Ghosh & Sanyal, 2021). Por tal motivo, LSTM es un modelo robusto e idóneo para el propósito de esta investigación.

Así mismo, se elige el modelo ELM por su excelente capacidad de generalización de la volatilidad junto con variables exógenas de noticias ligadas a la COVID-19. Además, en las

investigaciones efectuadas por (Weng et al., 2021), (J. Wang et al., 2018) y (X. Wang & Han, 2015) emplean este modelo para problemas de predicción de la volatilidad de commodities asociados al pánico y el sentimiento de los inversores a través del uso de índices de noticias, donde el ELM a través de los resultados obtenidos en la validación, indica que sus rendimientos y previsiones son superiores a los de las técnicas de pronóstico tradicionales.

Finalmente, según (Adhikari & Agrawal, 2012), (Allende & Valle, 2017) y (Laurinec et al., 2019) los *modelos ensamblados* son métodos predilectos a la hora de hacer análisis y previsión de series temporales, ya que estos consideran tanto a modelos univariados como modelos acompañados de variables exógenas, sin reducir su eficiencia, simplicidad, robustez y flexibilidad. Además, según (Ghosh & Sanyal, 2021) RF y XGB son altamente exitosos en la estimación precisa de los movimientos futuros de productos agrícolas.

Todos los modelos mencionados se desarrollaron en un computador Asus Windows 11 Home con un procesador Intel® Core™ i5-1240P CPU 1.70 GHz, memoria RAM de 8.00 GB y un sistema operativo de 64 bits, en el lenguaje de programación Python a través de la interfaz gráfica de usuario Anaconda Navigator 2.1.4. ejecutándose localmente en Jupyter Notebook 6.4.8.

7.2.1 Bosque aleatorio (RF)

El algoritmo de bosque aleatorio (RF) es una técnica que puede emplearse en tareas tanto de clasificación como de regresión, adaptándose bien a valores categóricos y continuos, además, debido a su naturaleza no lineal lo convierten en un modelo versátil en el campo del aprendizaje automático y la ciencia de datos. Este método utiliza "conjuntos paralelos" que ajustan varios clasificadores de árboles de decisión en paralelo, en diferentes submuestras de conjuntos de datos y utiliza votos por mayoría o promedios para el resultado final (Sarker, 2021). Además, minimiza el problema de ajuste excesivo, aumentando la precisión y el control de la predicción. Por lo tanto,

el modelo RF con múltiples árboles de decisión suele ser más preciso que un modelo basado en un solo árbol de decisión.

En este estudio se empleó la función .*RandomForestRegressor* (Pedregosa et al., 2011), donde se crea una serie de árboles de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la predicción y controlar el sobreajuste. En este caso, se buscó la mejor combinación de hiperparámetros para cada uno de los cuatro conjuntos de datos antes y los ocho conjuntos de datos durante la pandemia de la COVID-19. Por tal motivo, inicialmente se estableció la partición de datos de entrenamiento *(train)* y prueba *(test)* con el fin de obtener resultados más fiables, siendo común dividirlos en una relación de 70% *train* y 30% *test* (Alpaydin, 2020). Seguidamente, se iteró el modelo con métricas de validación del *test* a partir del *RMSE* para el número de retardos del modelo *lags* que es una variable importante en modelos de aprendizaje automático y puede afectar significativamente la precisión de las predicciones y *n_estimators* que describe la cantidad de árboles que representan el número de árboles dentro del modelo.

Finalmente, los parámetros *bootstrap*, *criterion*, *max_features* y *min_samples_leaf*, los cuales representan: el criterio para medir la calidad de la división de cada árbol, el número de características a buscar y el número mínimo de muestras requeridas para dividir un evento interno se mantuvo por defecto. En la Tabla 10 se muestran los mejores hiperparámetros de cada conjunto de datos antes y durante la pandemia del nuevo coronavirus.

Tabla 10

Selección de hiperparámetros RF antes de la pandemia de la COVID-19

RF	lags	n_esitmators	criterion	Max_features	Min_samples_leaf
Vol	15	200	MSE	1.0	1.0
Vol y VE	3	300	MSE	1.0	1.0

RF	lags	n_esitmators	criterion	Max_features	Min_samples_leaf
Vol y IN	15	300	MSE	1.0	1.0
Vol, VE y IN	3	200	MSE	1.0	1.0

Según la Tabla 10 los conjuntos de datos que necesitaron la menor cantidad de *lags* fueron Vol-VE y Vol-VE-IN, lo que indica que poseen una dependencia temporal menos compleja, facilitando que el modelo para estos casos generalice mucho más rápido y logre establecer con mayor velocidad la dependencia temporal entre las variables exógenas. Además, Vol-VE y VOL-IN necesitaron un número mayor de árboles lo que puede generar un mejor desempeño en el conjunto de datos de entrenamiento, ya que el modelo puede ajustarse mejor a los patrones y relaciones en los datos.

Tabla 11

Selección de hiperparámetros de RF durante de la pandemia de la COVID-19

RF	lags	n_esitmators	criterion	Max_features	Min_samples_leaf
Vol	7	50	MSE	1.0	1.0
Vol y VE	15	50	MSE	1.0	1.0
Vol e IN	7	200	MSE	1.0	1.0
Vol e IC	3	50	MSE	1.0	1.0
Vol, VE e IN	15	50	MSE	1.0	1.0
Vol, VE e IC	15	300	MSE	1.0	1.0
Vol, IN e IC	8	50	MSE	1.0	1.0
Vol, VE, IN e IC	15	200	MSE	1.0	1.0

Según la Tabla 11 el conjunto de datos durante la pandemia de la COVID-19 que requirió menos *lags* fue Vol-IC, presentando una dependencia temporal de menor complejidad y que probablemente con IC, se logró generalizar mejor el modelo durante la pandemia a diferencia de las demás variables. Por otra parte, los modelos Vol, Vol-VE, Vol-IC, Vol-VE-IN y Vol-IN-IC necesitaron una menor cantidad de árboles lo que se traduce en que fue más rápido su entrenamiento y se usaron menos recursos computacionales. Sumado a esto, puede ser menos

propenso al sobreajuste, lo que puede ayudar a mejorar su capacidad de generalización a nuevos datos.

7.2.2 Aumento de gradiente extremo (XGBoost)

El aumento de gradiente extremo (XGB), al igual que RF, es un algoritmo que genera un modelo final basado en una serie de modelos individuales, que por lo general son árboles de decisión (Sarker, 2021). El gradiente es empelado con el fin de minimizar la función pérdida, de manera análoga a como las redes neuronales usan el descenso de gradiente para optimizar los pesos. De este modo, el XGB es un algoritmo de aumento de gradiente que tiene en cuenta aproximaciones más detalladas al determinar eficientemente el mejor modelo, calculando gradientes de orden dos de la función de pérdida, reduciendo así, el ajuste excesivo y mejorando la generalización y rendimiento del modelo. Las ventajas principales son la rapidez al momento de generalizar y su gran rendimiento en grandes conjuntos de datos.

Para este modelo se empleó la función *xgb.XGBRegressor* Implementación de la API scikit-learn para regresión XGBoost (Pedregosa et al., 2011), la cual es una biblioteca de aumento de gradiente distribuida, optimizada y diseñada para ser altamente eficiente, flexible y portátil, proporcionando un impulso de árbol paralelo que resuelve muchos problemas de ciencia de datos de una manera rápida y precisa. Para este estudio se buscaron los mejores hiperparámetros para cada uno de los datasets, donde el método de selección de los mejores hiperparámetros fue a partir de las iteraciones que presentaron menor *RMSE* en el *test* variando del número de *lags* o retardos del modelo y *n_estimators* que indica el número de árboles potenciados por gradiente (equivalente al número de impulsos) y *early_stopping_rounds* que activa la parada anticipada, donde la métrica de validación necesita mejorar al menos una vez cada *early_stopping_rounds* para seguir entrenando.

Finalmente, *learning_rate* y *max_depth*, que son el aumento de la tasa de aprendizaje y profundidad máxima del árbol para aprendizaje de base se mantuvo por defecto. Además, la partición de *train* y *test* fue de 70% y 30% respectivamente. Los mejores hiperparámetros de cada conjunto de datos antes y durante la pandemia de la COVID-19 se presentan en las Tablas 12 y 13.

Tabla 12

Selección de hiperparámetros de XGB antes de la pandemia de la COVID-19

XGB	lags	n_esitmators	early_stopping_rounds	learning_rate	max_depth
Vol	15	200	80	0.3	6
Vol y VE	15	300	80	0.3	6
Vol e IN	8	300	80	0.3	6
Vol, VE e IN	15	200	80	0.3	6

Según la Tabla 12 el conjunto de datos de Vol-IN necesitó un menor número de *lags*, debido a que los datos de la volatilidad y los índices de noticias como variables de entrada presentan estructuras temporales menos complejas, lo que le permite al modelo lograr un buen rendimiento con menor cantidad de características retrasadas. Por otra parte, que *n_estimators* sea menor para Vol y Vol-VE-IN indica que para los datos antes de la pandemia de la COVID-19 se encontró un numero óptimo donde las características seleccionadas son lo suficientemente representativas en la relación entre las variables de entrada y la variable de salida, lo que reduce la necesidad de tener un gran número de árboles en el modelo.

Tabla 13

Selección de hiperparámetros de XGB durante la pandemia de la COVID-19

XGB	lags	n_esitmators	early_stopping_rounds	learning_rate	max_depth
Vol	8	50	80	0.3	6
Vol y VE	3	50	80	0.3	6
Vol e IN	15	200	80	0.3	6
Vol e IC	7	50	80	0.3	6

XGB	lags	n_esitmators	early_stopping_rounds	learning_rate	max_depth
Vol, VE e IN	15	50	80	0.3	6
Vol, VE e IC	10	300	80	0.3	6
Vol, IN e IC	5	50	80	0.3	6
Vol, VE, IN e IC	3	200	80	0.3	6

Según la Tabla 13 que para los conjuntos de datos Vol-IC y Vol-IN-IC durante la pandemia de la COVID tengan tanto un número de *lags* como de *n_estimators* bajo proporciona indicios de que el conjunto de características es suficientemente representativo de la relación entre las variables de entrada y la variable de salida, por lo tanto, las variables de las noticias e índices de la COVID-19 proporciona un buen rendimiento predictivo y ayudan a mejorar la generalización del modelo.

7.2.3 Extreme Learning Machine (ELM)

La máquina de aprendizaje extremo (ELM) a diferencia de los métodos basados en gradiente, esta asigna valores aleatorios a los pesos entre la capa de entrada la oculta y a los sesgos en la capa oculta. Las funciones de activación no lineales de la capa oculta proporcionan no linealidad al sistema. Por tanto, puede considerarse un sistema lineal, donde el único parámetro que requiere es el peso entre la capa oculta y la capa de salida. Por tanto, el ELM converge mucho más rápido que los algoritmos tradicionales porque aprende sin iteración. El análisis teórico demostró que el modelo ELM tiene más probabilidades de alcanzar la solución óptima global con parámetros aleatorios por encima de las redes neuronales tradicionales con todos los parámetros a entrenar (J. Wang et al. 2021).

Para este modelo se utilizó la función *.ELMRegressor* implementado en la biblioteca de aprendizaje automático de Python, scikit-learn (Pedregosa et al., 2011). ELMRegressor es un tipo de red neuronal de una sola capa oculta, donde los pesos de la capa oculta se inicializan de manera aleatoria y se calcula analíticamente la capa de salida. Se empleó la metodología de (Weng et al.,

2021) donde para los modelos ELM se hace una partición de 70% *train* y 30% *test* y emplea la función de activación *sigmoid* en el parámetro *activation*. Además, se hacen iteraciones del modelo con métrica de validación *RMSE* variando el número de *lags* o retardos y *hidden_layer_sizes* que es el número de neuronas de capa oculta en el modelo ELM, controla el tamaño del modelo y la capacidad de aprendizaje. Por lo general, el número de neuronas debe ser menor que el número de muestras de datos de entrenamiento, ya que, de lo contrario, el modelo aprenderá perfectamente el conjunto de entrenamiento, lo que resultará en un sobreajuste.

Finalmente, los demás parámetros del modelo se mantienen por defecto, destacando *solver='adam'* que funciona bastante bien en términos de tiempo de entrenamiento y puntuación de validación. Los mejores hiperparámetros del modelo ELM de cada conjunto de datos antes y durante la pandemia de la COVID-19 se presentan en las Tablas 14 y 15.

Tabla 14

Selección de hiperparámetros de ELM antes de la pandemia de la COVID-19

ELM	lags	hidden_layer_sizes	activation	solver
Vol	15	300	sigmoid	adam
Vol y VE	15	500	sigmoid	adam
Vol e IN	15	300	sigmoid	adam
Vol, VE e IN	10	300	sigmoid	adam

Según la Tabla 14 el conjunto de datos antes de la pandemia de la COVID-10 de Vol-VE-IN necesitó una menor cantidad de lags, posiblemente tanto las variables económicas como los índices de las noticias proporcionan información adicional que pueden ayudar a predecir la variable objetivo sin la necesidad de utilizar tantos retardos, ya que estas variables proporcionan información adicional que puede ayudar a explicar la variabilidad en la variable objetivo. Por otra parte, que tres conjuntos de datos obtuvieran las mejores métricas de validación con 300 neuronas

de la capa oculta significa que los modelos pueden ser capaces de aprender relaciones no lineales complejas en los datos de manera similar. Sin embargo, esto no significa necesariamente que los modelos tengan un rendimiento similar en todos los casos, debido a que es importante evaluar cada modelo en función de su rendimiento en los datos de los errores del *test* para determinar cuál es el mejor modelo y, por ende, la mejor capacidad de generalización.

Tabla 15

Selección de hiperparámetros de ELM durante la pandemia de la COVID-19

ELM	lags	hidden_layer_sizes	activation	solver
Vol	10	300	sigmoid	adam
Vol y VE	10	90	sigmoid	adam
Vol e IN	7	50	sigmoid	adam
Vol e IC	8	200	sigmoid	adam
Vol, VE e IN	15	80	sigmoid	adam
Vol, VE e IC	10	50	sigmoid	adam
Vol, IN e IC	7	300	sigmoid	adam
Vol, VE, IN e IC	10	50	sigmoid	adam

A partir de la Tabla 15 se evidencia que los conjuntos de datos que necesitaron una menor cantidad de *lags* fueron Vol-IN, Vol-IC y Vol-IN-IC, lo que puede indicar que las variables exógenas de índices de noticias e índices relacionados con la pandemia de la COVID-19 proporcionaron información adicional que puede ayudar a explicar la variabilidad en la variable objetivo. No obstante, será necesario validar las métricas de ajuste del *test* para determinar que variables le permitieron al modelo generalizar mejor.

7.2.4 Modelo de memoria a corto plazo (LSTM)

Es un modelo con una arquitectura basada en redes neuronales recurrentes (RNN). El modelo LSTM posee una serie de enlaces de retroalimentación. Además, es adecuado para analizar y generalizar datos que están distribuidos de manera secuencial, es ideal para clasificar, procesar y

predecir datos basados en series temporales, diferenciándolo de otro tipo de algoritmos basados en redes neuronales (Sarker. 2021).

Para el desarrollo de este modelo se utilizó la función *LSTM* implementada en la biblioteca de Python de aprendizaje automático *TensorFlow* (Abadi et al., 2016). Este modelo se utiliza para modelar relaciones de dependencia a largo plazo en los datos de entrada. En el modelo desarrollado se realiza una segmentación aproximada de 80% *train* y 20% *test* propuesto por (Ghosh & Sanyal, 2021). Para este modelo se hicieron 100 iteraciones de cada conjunto de datos *"epochs"* con métrica de validación *RMSE* para controlar la cantidad de veces que se recorre todo el conjunto de datos durante el proceso de entrenamiento del modelo. Además, dentro de los *epochs* se variaron tanto el número de *lags* como las *neuronas* o *units* que son el número de unidades en la capa oculta el cual controla la complejidad del modelo y la capacidad para aprender patrones.

Los demás hiperparámetros se mantuvieron por defecto, destacando que *activation*, *optimizer* y *loss*, que es la función de activación que controla la salida de las unidades, algoritmo de optimización utilizado durante el entrenamiento y la función de pérdida utilizada para evaluar el rendimiento del modelo durante el entrenamiento respectivamente. Los mejores hiperparámetros del modelo LSTM de cada conjunto de datos antes y durante la pandemia de la COVID-19 se presentan en las Tablas 16 y 17.

Tabla 16

Selección de hiperparámetros de LSTM antes de la pandemia de la COVID-19

LSTM	Lags	units/Neuronas	activation	Optimizer	loss
Vol	15	32	tanh	adamax	mean_squared_error
Vol y VE	15	128	tanh	adamax	mean_squared_error
Vol e IN	10	16	tanh	adamax	mean_squared_error
Vol, VE e IN	15	64	tanh	adamax	mean_squared_error

A partir de la Tabla 16 se evidencia que el conjunto de datos antes de la pandemia de la COVID-19 que necesitó un menor número de *units* fue donde se combinaron los datos de la volatilidad, las variables económicas y los índices de noticias, lo cual implica que las entradas exógenas del modelo le permitieron generalizar más rápido y con un menor uso de recursos computacionales. Sin embargo, es necesario hacer las validaciones del modelo a partir de las métricas de ajuste para poder determinar cuál presentó mejor capacidad predictiva.

Tabla 17

Selección de hiperparámetros de LSTM durante la pandemia de la COVID-19

LSTM	lags	units/Neuronas	activation	Optimizer	loss
Vol	15	128	tanh	adamax	mean_squared_error
Vol y VE	15	512	tanh	adamax	mean_squared_error
Vol e IN	8	32	tanh	adamax	mean_squared_error
Vol e IC	10	16	tanh	adamax	mean_squared_error
Vol, VE e IN	15	32	tanh	adamax	mean_squared_error
Vol, VE e IC	10	64	tanh	adamax	mean_squared_error
Vol, IN e IC	10	32	tanh	adamax	mean_squared_error
Vol, VE, IN e IC	10	64	tanh	adamax	mean_squared_error

Según la Tabla 17 es notorio que existió un comportamiento similar al de la Tabla 16. Con la particularidad que se sumaron los conjuntos de datos donde están involucrados los índices relacionados con la pandemia, presentando indicios de que dichas variables le permiten generar predicciones precisas con pocos recursos computacionales. No obstante, es necesario analizar las métricas de ajuste para afirmar qué variables permitieron una mayor capacidad predictiva y cuáles presentaron una mayor influencia.

7.3 Evaluación y validación de los modelos.

Según (Weng et al., 2021a) y (J. Wang et al., 2018a) la evaluación y validación de modelos de aprendizaje automático es posible realizarla a través de las siguientes métricas de ajuste: error

cuadrático medio (RMSE), error medio absoluto (MAE) y error porcentual medio (MAPE). Además, se presentan únicamente los valores del *test* de las mejores métricas encontradas en el modelo, tanto antes como durante la pandemia de la COVID-19, ya que, la intención principal de este estudio es presentar un modelo de aprendizaje automático que logre pronosticar la volatilidad del café colombiano junto a variables exógenas y poder determinar cuáles ayudaron a generalizar el modelo, teniendo en cuenta ambos contextos.

7.3.1 Resultados de RF

Con el fin de evaluar el rendimiento del algoritmo RF antes y durante la pandemia, se presentan los errores del test, ya que estos representan datos no observados por el modelo durante el *train*. Si el modelo tiene un buen desempeño en el conjunto de prueba, esto indica que puede generalizar bien a nuevos datos y que no ha aprendido únicamente los patrones específicos del conjunto de entrenamiento. En las Tabla 18 se presentan los rendimientos del modelo para los cuatro conjuntos de datos antes de la pandemia, mientras que en la Tabla 19 se muestran los resultados para los ocho Datasets analizados durante el contexto de la pandemia del nuevo coronavirus.

Tabla 18

Resultados del rendimiento de RF datos antes de la pandemia de la COVID-19.

Test RF	RMSE	MAE	MAPE
Vol	0.972	0.732	40.040
Vol y VE	1.047	0.821	46.664
Vol e IN	0.974	0.738	40.587
Vol, VE e IN	1.024	0.793	44.964

Según la Tabla 18 para los conjuntos de datos antes de la pandemia de Vol y Vol-IN, se evidencia que en promedio el algoritmo se comportó de manera similar, siendo superior el

rendimiento en el conjunto de datos donde se encuentra solo la volatilidad, esto se debe según la literatura a que el algoritmo encontró relaciones más sencillas de interpretar, además, a RF le afecta el rendimiento cuando las variables exógenas que lo acompañan están altamente correlacionadas entre ellas, en este caso el conjunto de datos de VE donde se encuentran los datos de la *TRM* y *Precio interno del café colombiano* tiene una correlación fuerte. Así mismo, RF es sensible cuando alguna de las variables de entrada tiene un componente no estacionario, como es el caso de la TRM, provocando la disminución en la precisión del modelo, ya que la tendencia no es capturada por los árboles de decisión individuales. Por lo tanto, donde se ven involucradas estas variables el rendimiento del modelo cae alrededor de un 15% en sus métricas. Por otra parte, con IN el modelo generaliza prácticamente igual que al conjunto de datos de volatilidad individual, lo que indica que estas variables, aunque no le generan más información, no le producen ruido.

Tabla 19

Resultados del rendimiento de RF datos durante la pandemia de la COVID-19.

Test RF	RMSE	MAE	MAPE
Vol	1.184	0.893	44.052
Vol y VE	2.489	2.262	133.903
Vol e IN	1.142	0.855	41.636
Vol e IC	1.178	0.907	45.985
Vol, VE e IN	2.426	2.215	131.076
Vol, VE e IC	2.378	2.173	129.106
Vol, IN e IC	1.142	0.874	44.094
Vol, VE, IN e IC	2.696	2.479	145.931

Según la Tabla 19, se evidencia el mismo comportamiento de los resultados antes de la pandemia en cuanto a el aumento del error cuando el conjunto de VE está presente, siendo más notorio debido a que en este caso a la cantidad de observaciones es menor. Por otra parte, para este modelo se evidenció que IN le permitió al modelo una mejor generalización y por ende presentar las mejores métricas de validación, así mismo, donde estuvo involucrada IC hubo buen rendimiento, aunque no proporcionó información adicional al modelo, tampoco le generó ruido.

En resumen, aunque se obtuvieron excelentes métricas de validación de este modelo tanto antes como durante la pandemia, se evidencia una gran sensibilidad al presentar una menor cantidad de datos y limitaciones de generalización con el conjunto de datos de variables económicas (VE).

7.3.2 Resultados de XGB

Para obtener los resultados de este modelo se ajustó el parámetro *early_stopping_rounds* a 80, el cual monitorea el rendimiento en un conjunto de validación diferente al conjunto de entrenamiento después de cada iteración del algoritmo. La intención principal de este parámetro es que, si el rendimiento del modelo deja de mejorar, el entrenamiento se detiene y regresa al modelo con el mejor rendimiento. En las Tablas 20 y 21 se presenta el mejor modelo por cada conjunto de datos antes y durante la pandemia de la COVID-19.

Tabla 20

Resultados del rendimiento de XGB datos antes de la pandemia de la COVID-19.

Test XGB	RMSE	MAE	MAPE
Vol	0.969	0.726	39.439
Vol y VE	0.956	0.716	38.488
Vol e IN	1.013	0.755	40.363
Vol, VE e IN	1.039	0.780	42.242

A partir de la Tabla 20 el modelo con los conjuntos de datos antes de la pandemia de la COVID-19 que en promedio tuvo los mejores resultados fue Vol-VE, el cual presenta las mejores métricas de RMSE, MAE y MAPE, con 0.969, 0.716 y 40.363. Siendo superior incluso al conjunto de datos de referencia para el modelo donde solamente se encuentran las observaciones de Vol, lo cual indica que las variables económicas antes de la pandemia por el nuevo coronavirus

proporcionaron información relevante y tienen una mayor influencia en la volatilidad del café colombiano.

Tabla 21

Resultados del rendimiento de XGB datos durante la pandemia de la COVID-19.

Test XGB	RMSE	MAE	MAPE
Vol	1.183	0.916	47.427
Vol y VE	1.411	1.184	68.813
Vol e IN	1.162	0.896	43.982
Vol e IC	1.178	0.916	48.352
Vol, VE e IN	1.675	1.436	84.121
Vol, VE e IC	1.760	1.532	91.851
Vol, IN e IC	1.167	0.904	48.091
Vol, VE, IN e IC	1.683	1.443	84.742

Para los conjuntos de datos durante la pandemia de la COVID-19 se evidencia que para este modelo los conjuntos de datos Vol, Vol-IN, Vol-IC y Vol-IN-IC se presentan en promedio las mejores métricas, resaltando que las observaciones de Vol-IN fueron superiores, lo cual indica que las noticias tienen influencia en la volatilidad de los futuros de café colombiano y que, por lo tanto, le proporciona al modelo información adicional para mejorar la capacidad de generalización. Por otra parte, los conjuntos de datos donde están presentes las observaciones de VE los resultados presentan gran error, generando ruido al modelo y reduciendo su capacidad predictiva, indicando que su información no es relevante durante este periodo de tiempo. XGB presentan la misma limitación de RF, ya que su arquitectura converge en la construcción de múltiples arboles de decisión afectando la capacidad de capturar la tendencia de la volatilidad de los futuros del café colombiano.

7.3.3 Resultados del test ELM

Con el fin de realizar la validación y evaluación del rendimiento del algoritmo ELM antes y durante la pandemia, se presentan los errores del test, ya que estos representan datos no observados por el modelo durante el *train* y permiten determinar la capacidad de generalización para cada conjunto de datos antes y durante la pandemia de la COVID-19, además de interpretar cuales variables fueron influyentes para este en cada contexto. En las Tablas 22 y 23 se presentan los rendimientos del modelo.

Tabla 22

Resultados del rendimiento de ELM datos antes de la pandemia de la COVID-19.

Dataset	RMSE	MAE	MAPE
Vol	0.968	0.712	36.474
Vol y VE	0.956	0.696	35.641
Vol e IN	0.973	0.728	38.462
Vol, VE e IN	0.964	0.710	36.836

Según la Tabla 22 se evidencia que el modelo ELM tiene un mejor rendimiento en promedio cuando VE está involucrada en los conjuntos de datos, ya que la capacidad de generalización del modelo mejora respecto a la del modelo de referencia donde se encuentra solo Vol, lo que significa que VE aporta en este modelo información relevante para mejorar la capacidad predictiva de la volatilidad de los futuros del café colombiano antes de la pandemia de la COVID-19. Por otra parte, los índices de las noticias antes de la pandemia no presentan información adicional al modelo, por lo que en el conjunto de observaciones de Vol-IN el RMSE, MAE y MAPE aumentan un 1.83%, 4.22% y 6.91% respectivamente.

Tabla 23

Resultados del rendimiento de ELM datos durante la pandemia de la COVID-19.

Test ELM	RMSE	MAE	MAPE
Vol	1.144	0.883	43.866
Vol y VE	1.147	0.886	44.477
Vol e IN	1.136	0.875	42.378
Vol e IC	1.163	0.889	43.365
Vol, VE e IN	1.164	0.895	43.061
Vol, VE e IC	1.151	0.873	42.527
Vol, IN e IC	1.158	0.892	42.471
Vol, VE, IN e IC	1.163	0.944	50.277

A partir de la Tabla 23 se puede evidenciar que los conjuntos de observaciones donde se incluyó IN e IC se obtuvieron las mejores métricas de validación, lo cual indica que las noticias durante la pandemia de la COVID-19 proporcionan información más amplia y completa sobre la volatilidad del mercado y la incertidumbre global, lo que permitió predecir la volatilidad del café colombiano de manera más precisa. Sin embargo, el modelo donde se encuentran todos los conjuntos de datos fue el que peor rendimiento tuvo con un aumento del MAPE del 18%, esto puede significar según la literatura consultada que las características de baja variabilidad en VE respecto a IN e IC perjudiquen la capacidad de generalización del modelo y por ende le supongan mayor ruido en sus predicciones.

7.3.4 Resultados del test LSTM

Para este modelo se realizaron 100 iteraciones del modelo llamadas *epochs*, las cuales permiten que el modelo alcance su capacidad de ajuste y aprenda la relación entre las entradas y las salidas del problema que se está resolviendo, sin que llegue a sobre ajustarse. Este modelo según la literatura, por su arquitectura basada en redes neuronales es robusto para tareas donde las variables presentan multicolinealidad y no estacionariedad, dificultades presentes en los modelos

RF y XGB. En las tablas 24 y 25 se presentan los cuatro conjuntos de datos empleados antes de la pandemia y los ocho conjuntos usados durante la COVID-19.

Tabla 24

Resultados del rendimiento de LSTM datos antes de la pandemia de la COVID-19.

Test LSTM	RMSE	MAE	MAPE
Vol	0.963	0.726	39.198
Vol y VE	1.027	0.736	36.066
Vol e IN	0.984	0.746	40.010
Vol, VE e IN	1.016	0.767	40.609

Según la tabla 24 el conjunto de datos de Vol y Vol-VE tienen las mejores métricas de validación, especialmente este último, el cual muestra un MAPE de 36.066% indicando que este modelo bajo esos *inputs* tuvo un bajo porcentaje de error absoluto promedio en relación con los valores reales, lo que se traduce en que la información que le suministran las observaciones en el periodo de tiempo estudiado le permite generalizar mejor y obtener predicciones más precisas. Por otra parte, aunque IN no presente información que contribuya a mejorar la predicción, tampoco le genera mayor ruido al modelo al compararlo con el de referencia (Vol) teniendo diferencias en el RMSE, MAE y MAPE apenas de 2.164%, 2.692 y 2.071 respectivamente.

Tabla 25

Resultados del rendimiento de LSTM datos durante la pandemia de la COVID-19.

Test LSTM	RMSE	MAE	MAPE
Vol	1.112	0.843	40.436
Vol y VE	1.511	1.155	47.839
Vol e IN	1.110	0.844	39.743
Vol e IC	1.150	0.878	43.030
Vol, VE e IN	1.115	0.820	38.845
Vol, VE e IC	1.135	0.855	40.651
Vol, IN e IC	1.149	0.846	37.843
Vol, VE, IN e IC	1.112	0.816	37.863

A partir de la tabla 25 se evidencia que los conjuntos de datos donde el modelo tuvo un mejor desempeño fueron donde estuvieron tanto IN como IC, especialmente en Vol-IN-IC y Vol-VE-IN-IC quienes en promedio tuvieron el mejor desempeño general, con un RMSE, MAE y MAPE de 1.149, 0.846, 37.843%, 1.112, 0.816 y 37.863 respectivamente. Lo que permite inferir que la volatilidad del café colombiano se ve afectada en mayor medida por los eventos y noticias actuales, que por los factores económicos a largo plazo. Por lo tanto, IN e IC proporcionan información más relevante y oportuna, aumentando la capacidad predictiva del modelo. Sin embargo, cabe resaltar que, aunque VE no mejora las métricas del modelo, cuando esta junto a IN las métricas mejoran, esto se debe a que, al estar estas observaciones agrupadas, permiten una mayor correlación con la variable objetivo y, por lo tanto, el modelo puede capturar patrones en los datos contribuyendo a mejorar su capacidad de generalización.

8. Interpretación de los resultados

En esta sección, se expone un resumen de los resultados de los modelos de pronóstico tanto para el evento antes como durante la pandemia de la COVID-19, ya que, como se ha mencionado en apartados anteriores de este documento, es importante analizar qué o cuáles variables tuvieron mayor influencia en la volatilidad de los futuros del café colombiano en cada contexto. Por lo tanto, se presentará un análisis por separado de los periodos antes mencionados, ya que tanto los conjuntos de datos como la cantidad de observaciones difieren, de lo contrario se podría estar sesgando la información y el conocimiento encontrado. Los resultados de los modelos RF, XGB, ELM y LSTM para cada evento antes y durante la pandemia se presentan en las Tablas 26 y 27 respectivamente.

8.1 Resultados generales de validación antes de la pandemia de al COVID-19

A partir de la Tabla 26 es posible evidenciar que, en promedio, según las métricas de validación RMSE, MAE y MAPE, el conjunto de datos que permitió una mejor generalización de los diferentes modelos fue VE, quien ofreció información relevante y oportuna, permitiendo describir mejor la predicción de la volatilidad de los futuros del café colombiano. Por otra parte, en general, el mejor modelo con mejores métricas de error fue el modelo ELM, esto debido a que según (J. Wang et al., 2018) al incorporar variables exógenas que están correlacionadas entre sí, el modelo es significativamente más preciso. Por lo tanto, los resultados generales del conjunto de datos antes de la pandemia de la COVID-19 de Vol-VE, especialmente con el modelo ELM, en cuanto a RMSE, MAE y MAPE con un 6.906%, 5.524%, 1.178% es superior a LSTM, sugiriendo que el comportamiento de las variables económicas como la TRM y Precio interno del café colombiano atraen la atención y expectativas de los inversores hacia el mercado de futuros del café colombiano.

Tabla 26

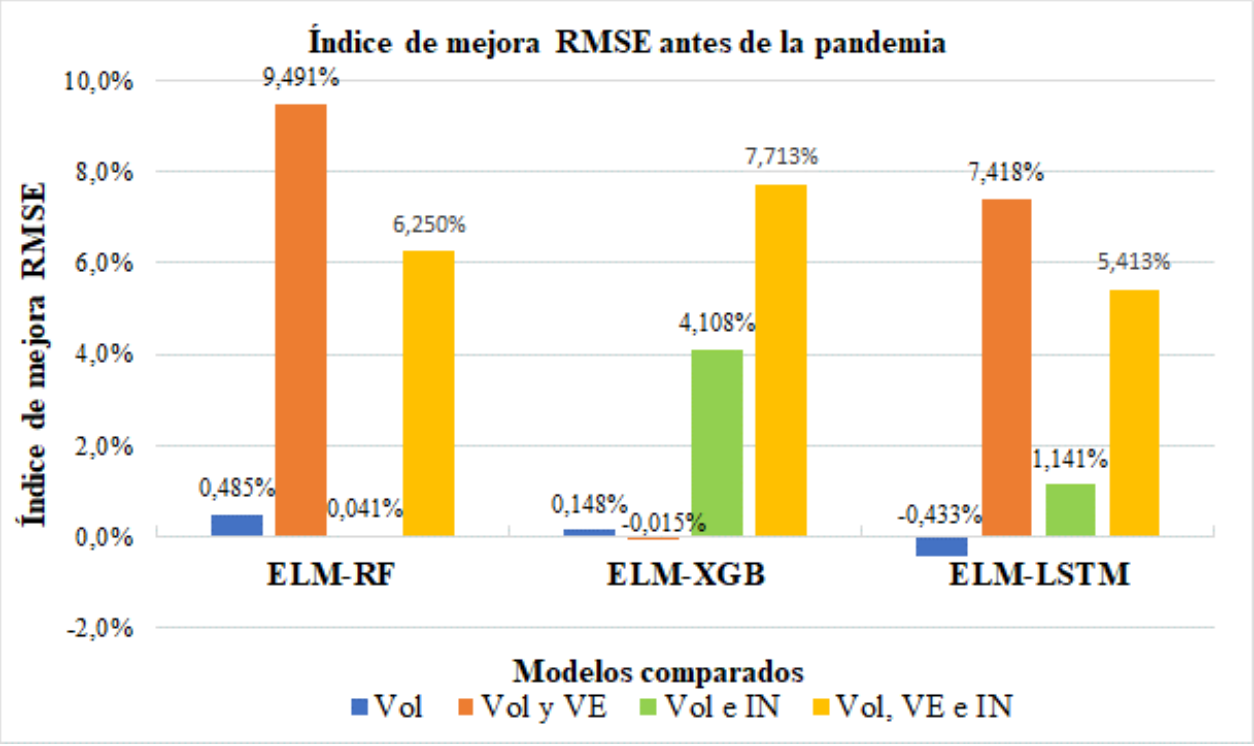
Resumen de errores de predicción de los modelos antes de la pandemia.

Modelo	Error	Vol	Vol y VE	Vol e IN	Vol, VE e IN
RF	RMSE	0.972	1.047	0.974	1.024
	MAE	0.732	0.821	0.738	0.793
	MAPE	40.040	46.664	40.587	44.964
XGB	RMSE	0.969	0.956	1.013	1.039
	MAE	0.726	0.716	0.755	0.780
	MAPE	39.439	38.488	40.363	42.242
ELM	RMSE	0.968	*0.956*	0.973	0.964
	MAE	0.712	*0.696*	0.728	0.710
	MAPE	36.474	*35.641*	38.462	36.836
LSTM	RMSE	0.963	1.027	0.984	1.016
	MAE	0.726	0.736	0.746	0.767
	MAPE	39.198	36.066	40.010	40.609

Sumado a lo anterior, en la Figura 14 se presenta el índice de mejora del RMSE, el cual reafirma que, ELM junto a el conjunto de datos Vol-VE es el mejor método de predicción de volatilidad de los futuros del café colombiano, aunque se vea superado en este índice por XGB, en MAE y MAPE, ELM lo supera en 2.915%, 7.986% respectivamente. Por lo tanto, es posible afirmar que para el evento antes de la pandemia este modelo tuvo el mejor rendimiento y capacidad predictiva.

Figura 14

Comparación de modelos índice de mejora del RMSE con ELM de referencia antes de la pandemia

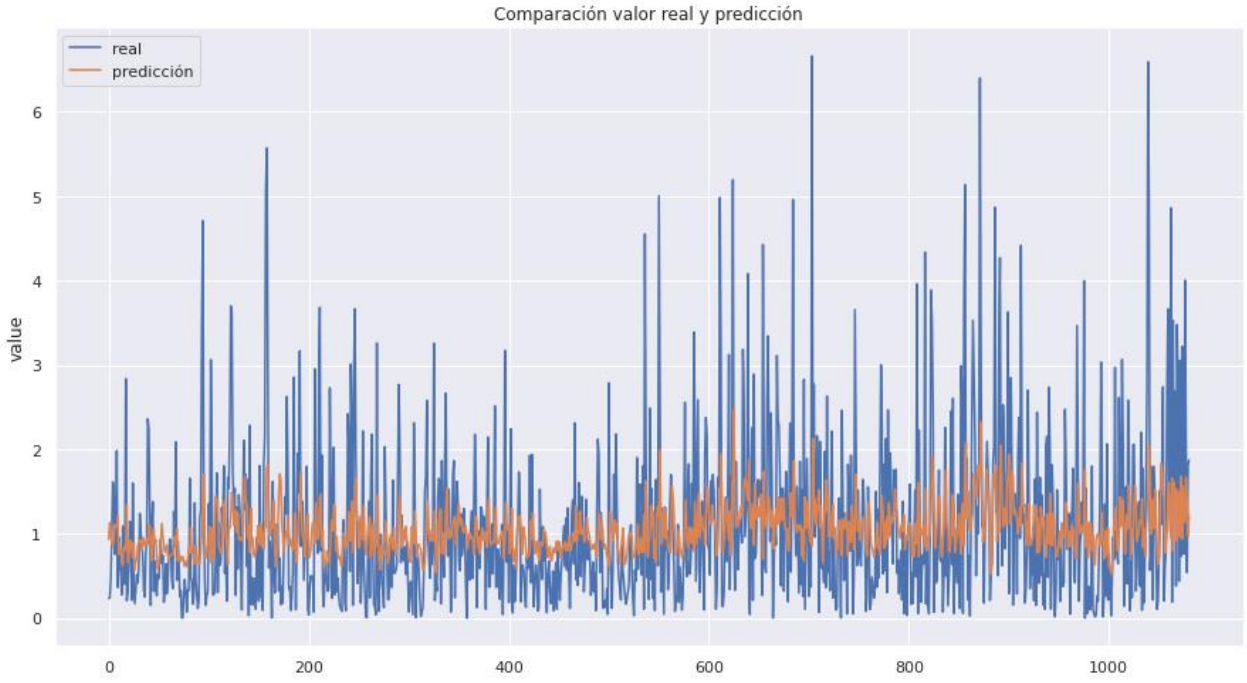


Finalmente, en la Figura 15 se muestra el mejor modelo con el mejor conjunto de datos basado en el desempeño de las métricas de validación, es decir el modelo ELM junto con Vol-VE

como *inputs*, evidenciando un ajuste adecuado, donde los puntos que la línea naranja sobrepasan la azul, representan sobre estimación del rendimiento.

Figura 15

Desempeño ELM con Vol-VE antes de la pandemia.



8.2 Resultados generales de validación evento durante la pandemia de la COVID-19

En la Tabla 27, se presentan los resultados de los cuatro modelos y los ocho conjuntos de datos que fueron objeto de estudio durante la pandemia de la COVID-19. En términos generales, los modelos tienen valores similares para las tres métricas de validación. Sin embargo, se observan algunas diferencias en la precisión de los modelos, dicho esto, el modelo en promedio con mejor desempeño es el LSTM con las combinaciones de Vol-IN-IC y Vol-VE-IN-IC permitiendo que estos datasets presentaran información relevante para describir y predecir con precisión la volatilidad de los futuros del café colombiano, especialmente donde se encuentran las nueve variables exógenas, ya que presentan los valores más bajos para RMSE, MAE y MAPE, superando

en un 4.603, 15.765% y 32.787% respectivamente al modelo ELM que según la literatura era el más adecuado para estas tareas. Lo anterior, se presenta ya que, según (Y. Liang et al., 2022) LSTM es adecuado para predecir series temporales, aun cuando la cantidad de observaciones es baja, mejorando la precisión del pronóstico de volatilidad de commodities en periodos de incertidumbre.

Tabla 27

Resumen de errores de predicción de los modelos antes de la pandemia.

Modelo	Error	Vol	Vol y VE	Vol e IN	Vol e IC	Vol, VE e IN	Vol, VE e IC	Vol, IN e IC	Vol, VE, IN e IC
RF	RMSE	1.184	2.489	1.142	1.178	2.426	2.378	1.142	2.696
	MAE	0.893	2.262	0.855	0.907	2.215	2.173	0.874	2.479
	MAPE	44.052	133.903	41.636	45.985	131.076	129.106	44.094	145.931
XGB	RMSE	1.183	1.411	1.162	1.178	1.675	1.760	1.167	1.683
	MAE	0.916	1.184	0.896	0.916	1.436	1.532	0.904	1.443
	MAPE	47.427	68.813	43.982	48.352	84.121	91.851	48.091	84.742
ELM	RMSE	1.144	1.147	1.136	1.163	1.164	1.151	1.158	1.163
	MAE	0.883	0.886	0.875	0.889	0.895	0.873	0.892	0.944
	MAPE	43.866	44.477	42.378	43.365	43.061	42.527	42.471	50.277
LSTM	RMSE	1.112	1.511	1.110	1.150	1.115	1.135	1.149	1.112
	MAE	0.843	1.155	0.844	0.878	0.820	0.855	0.846	0.816
	MAPE	40.436	47.839	39.743	43.030	38.845	40.651	37.843	37.863

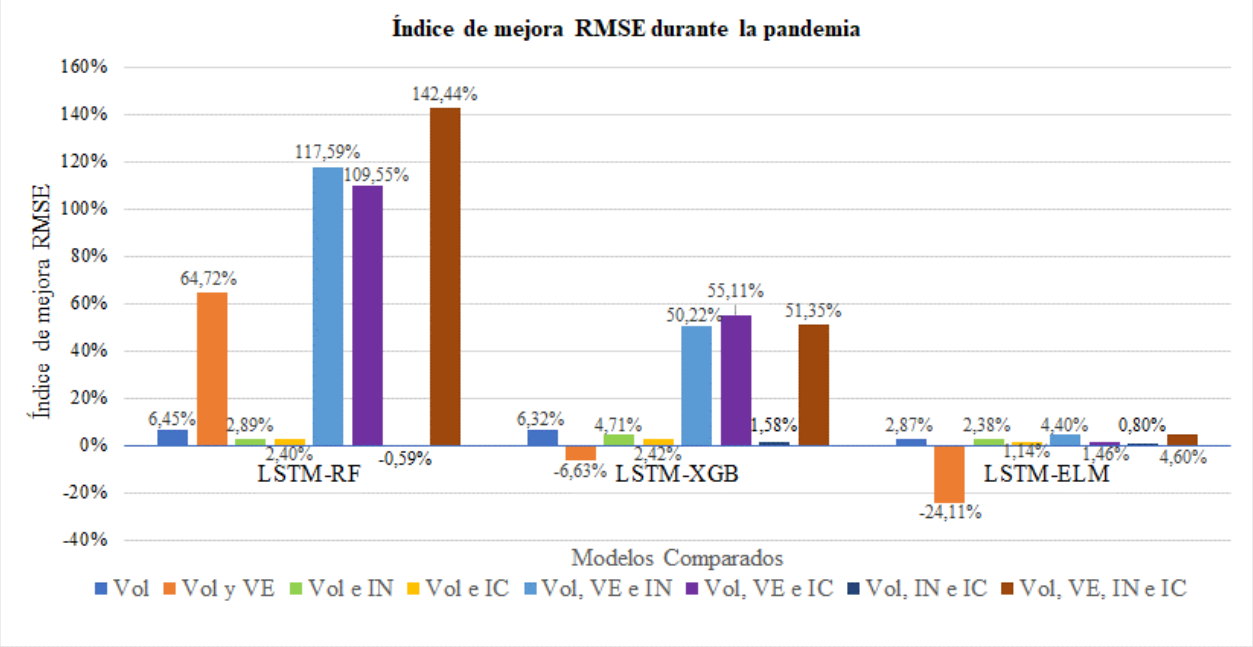
En la Figura 16, se presenta el índice de mejora del RMSE, donde se comparan los modelos empleados en el evento antes de la pandemia de la COVID-19, confirmando que en promedio LSTM tuvo mejores métricas. Sin embargo, hubo una excepción, ELM logró captar información relevante del conjunto de datos Vol-VE, aunque no es significativo su rendimiento frente a LSTM con Vol-VE-IN-IC. Además, RF y XGB tienen una baja considerable en sus métricas de desempeño, especialmente cuando están presentes las observaciones asociadas a VE, esto se debe a tres razones principales, la primera, es que la cantidad de datos de entrada no son lo suficientemente grade (918 observaciones) presentando dificultades para identificar patrones complejos y evitar el sobreajuste. La segunda, es que en modelos basados en árboles de decisión

el rendimiento disminuye considerablemente cuando las variables exógenas que lo acompañan están altamente correlacionadas como es el caso de la TRM y el precio interno del café colombiano. Finalmente, el componente no estacionario de TRM hace que estos modelos tengan dificultades

Por lo tanto, se reafirma que LSTM con Vol-VE-IN-IC fue el mejor modelo, logrando una mejor generalización y ajuste, sin dejar de lado el excelente rendimiento de Vol-IN-IC, lo cual a su vez indica que las noticias cobraron un papel relevante en el contexto del nuevo coronavirus, teniendo repercusiones en la volatilidad del café colombiano y brindando información más relevante y oportuna, aumentando la capacidad predictiva del algoritmo, lo cual le permitió modelar la volatilidad de los futuros del café colombiano.

Figura 16

Comparación de modelos índice de mejora del RMSE con LSTM de referencia durante la pandemia

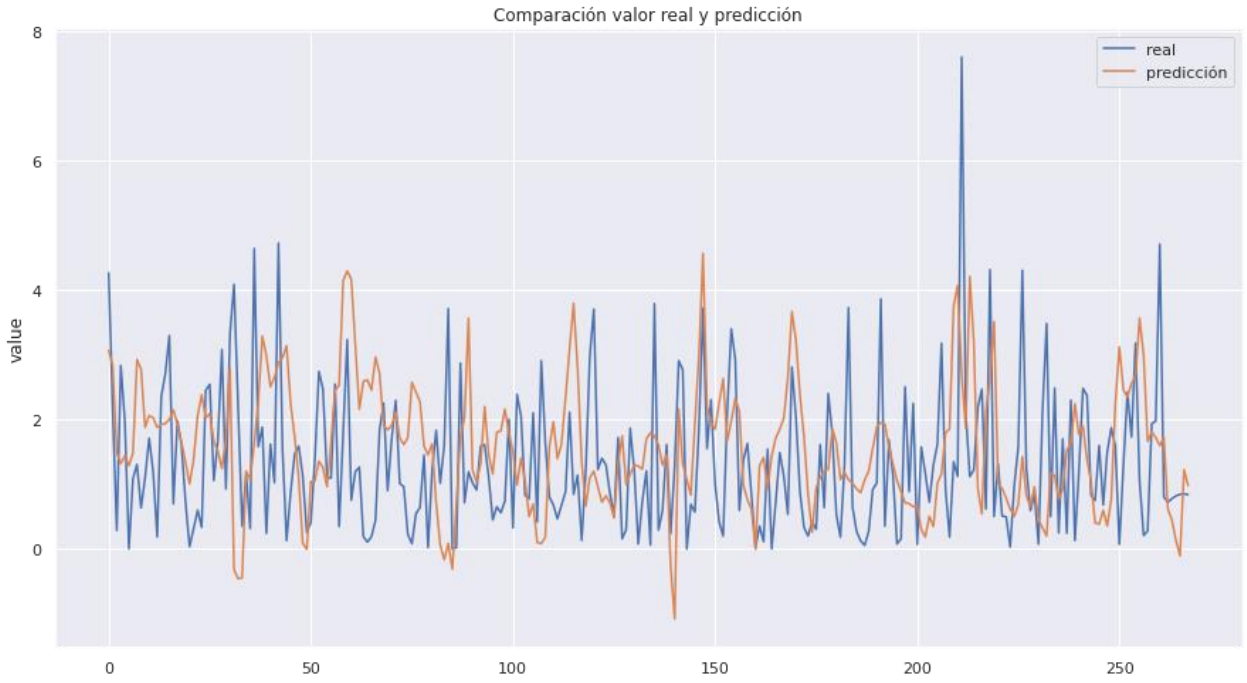


Finalmente, en la Figura 15 se muestra la serie de tiempo del modelo más adecuado, con el mejor conjunto de datos basado en el desempeño de las métricas de validación, es decir el modelo

LSTM junto con VE-IN-IC con entradas exógenas, evidenciando un ajuste adecuado, donde en las instancias que la línea naranja sobrepasan la azul, representan una sobre estimación de la volatilidad de los futuros del café colombiano, así mismo, se evidencia que aunque el modelo presenta un buen ajuste hay periodos que este disminuye, viéndose afectado por la cantidad limitada de observaciones, con apenas 275 en el *test* para el evento durante la pandemia.

Figura 17

Desempeño LSTM con Vol-VE-IN-IC durante la pandemia



9. Herramienta computacional

Para la fase final, según la metodología KDD, que fue la base del proceso metodológico de este estudio, se plantea decidir la ruta para exponer los patrones, información y nuevo conocimiento encontrado. Por lo tanto, se determina que la divulgación y el conocimiento hallado se realiza a través de una herramienta computacional, donde se vinculan las series de tiempo modeladas por los algoritmos Random Forest, Extreme Gradient Boosting, Extreme Learning Machine y Long Short-Term Memory para los diferentes conjuntos de datos junto a las variables exógenas agregadas con sus respectivas combinaciones, de modo que quienes interactúen con la herramienta tengan la posibilidad de observar el comportamiento de los modelos y su capacidad de generalización tanto en el contexto antes como durante la pandemia COVID-19. Además, se presentará un despliegue de los resultados de error de los mejores modelos encontrados, de esta manera se podrá visualizar el comportamiento de las métricas de validación.

El diseño y creación del código del prototipo del aplicativo web se desarrolló a través del framework *Dash*, el cual es un marco de código abierto que permite relacionar el código realizado en Python proporcionando una interfaz sencilla para vincular controles de usuario. El código de dicho aplicativo y el de los notebooks de Python se encuentran en la página web de hosting de repositorios GitHub (Corredor León, 2023).

10. Conclusiones

Esta investigación aporta en la extensa literatura que hay en torno a la previsión de volatilidad de commodities y productos agrícolas a través de técnicas de aprendizaje automático. Además, a partir de la revisión de literatura, se encuentra que la predicción de series temporales es una actividad ampliamente estudiada en la comunidad científica, especialmente en productos como el petróleo, gas, y productos agrícolas como la soya, cacao, maíz, café entre otros. Por otra parte, la búsqueda de artículos relacionados con la pandemia del nuevo coronavirus y predicciones de volatilidad tuvo un aumento significativo desde el año 2020. Por lo tanto, la búsqueda de modelos predictivos que permitan realizar previsiones en el contexto de la COVID-19 fue liderada por países como China, Estados Unidos, Reino Unido, Australia e India, debido a la gran influencia que tuvo la pandemia en dichos países en materia económica.

Existen diversos métodos utilizados para llevar a cabo tareas de previsión, siendo los modelos de *machine learning* los más populares. Entre ellos, destacan los modelos ensamblados (EM), específicamente el Random Forest y Extreme Gradient Boosting, así como los basados en redes neuronales artificiales (ANNs), como el Extreme Learning Machine (ELM) y el Long Short-Term Memory (LSTM). Estos modelos son altamente efectivos para manejar grandes volúmenes de datos, relaciones no lineales y múltiples variables de entrada, además de ser menos sensibles a observaciones ruidosas, lo que se traduce en pronósticos más precisos.

En este estudio, inicialmente, se presenta la necesidad de conocer el comportamiento de la serie de tiempo de la volatilidad del café colombiano. Por medio de la prueba Dickey Fuller Aumentada, que con un valor p de 0.001>5%, hubo evidencia significativa de que su comportamiento es estacionario y no lineal. En segundo lugar, se evidencia que existen variables que están correlacionadas con la volatilidad de los futuros del café colombiano, principalmente la

TRM y VIX, quienes sufrieron un incremento en esta métrica durante la pandemia en un 156% y 83.57% respectivamente.

En este documento se desarrollaron cuatro modelos de aprendizaje automático, dos modelos ensamblados RF y XGB y dos modelos basados en redes neuronales ELM y LSTM, siendo estos últimos quienes obtuvieron las mejores métricas de validación. Además, se estudia la relación y los cambios que pueden producir las variables exógenas en las fluctuaciones de los futuros del café colombiano, entre las cuales se encuentran: las variables económicas (TRM y Precio Interno del café colombiano), las noticias por medio de índices cuantitativos (VIX, EPU, EMV-ID) e índices basados en las noticas de la COVID-19 (PI, FNI, MHI y SI), separando los datos en dos periodos de estudio diferentes, el primero, antes de la pandemia de la COVID-19 y el segundo, durante esta. Con el fin de evaluar cuales variables de entrada tuvieron efectos en la predicción, se agregaron cuatro conjuntos de datos para el evento antes de la pandemia y ocho durante esta.

Para el contexto antes de la pandemia del nuevo coronavirus se evidenció que el mejor modelo, en general, que logra captar información relevante y generar predicciones más precisas fue el ELM, junto a las variables económicas (VE), es decir el Vol-VE, con valores de RMSE, MAE y MAPE de 0.956, 0.696 y 35.641% respectivamente, sugiriendo que las tanto la *TRM* y *el precio interno del café colombiano* permiten describir mejor la predicción de la volatilidad de los futuros del café colombiano presentando influencia en sus fluctuaciones durante dicho periodo.

Por otra parte, durante la pandemia del nuevo coronavirus el mejor modelo según los resultados de validación para los diferentes conjuntos de datos fue LSTM, esto debido a su excelente capacidad de predecir series temporales, aun cuando la cantidad de observaciones es baja, mejorando la precisión del pronóstico. Además, se encontró que el modelo LSTM, en combinación con los datasets Vol-IN-IC y Vol-VE-IN-IC, proporciona información valiosa para

describir y predecir con precisión la volatilidad de los futuros del café colombiano. Este modelo destacó en particular en presencia de las nueve variables exógenas, demostrando los valores más bajos para RMSE, MAE y MAPE, con 1.112 , 0.816 y 37.863%, superando al modelo ELM en un 4.603%, 15.765% y 32.787% respectivamente.

En conclusión, se puede afirmar que el modelo LSTM en combinación con el Dataset Vol-VE-IN-IC, demostró ser el más adecuado para estas tareas, gracias a su mejor capacidad de generalización y ajuste. Sin embargo, no se puede ignorar el excelente rendimiento del modelo con los *inputs* de Vol-IN-IC, lo cual sugiere que las noticias relacionadas con la COVID-19 tuvieron un papel relevante en el comportamiento y fluctuaciones de la volatilidad del café colombiano durante la pandemia del nuevo coronavirus.

Finalmente, en la presente investigación, se desataca la efectividad de los modelos Extreme Learning Machine (ELM) y de Memoria a Largo Plazo (LSTM) en el análisis y predicción de la volatilidad del café colombiano, antes y durante la pandemia de la COVID-19 respectivamente. Los resultados obtenidos han resaltado la utilidad de ambos modelos, en especial durante situaciones de alta incertidumbre como la actual pandemia. Asimismo, se ha comprobado que la inclusión de variables exógenas, tales como las variables económicas (VE), índices de noticias (IN) e índices de noticias basadas en la COVID-19 (IC), proporcionan información relevante para la obtención de predicciones más precisas. No obstante, estos modelos pueden requerir actualizaciones frecuentes a medida que cambian las condiciones económicas y sociales que pueden afectar la volatilidad de los futuros del café colombiano.

11. Recomendaciones

En este estudio, aunque se presentan buenas métricas de validación, no se establece de manera concluyente que pueda ser aplicable en inversiones o participación en el mercado de valores. Por lo tanto, se sugiere que se empleen técnicas de aprendizaje automático *o machine learning* en tareas de previsión de series temporales, como el *algorithm trading*, únicamente como apoyo en la toma de decisiones de inversión por parte de personas con experiencia y conocimiento en el tema.

Como se pude evidenciar, los modelos ELM y LSTM resultaron robustos para predecir series temporales en contextos de incertidumbre, volatilidad y observaciones limitadas. Por lo tanto, se recomienda que para trabajos futuros se utilicen variaciones de estos modelos como por ejemplo utilizar algoritmos genéticos para optimizar los parámetros de entrada, factores de olvido o variaciones de tipo: ELM-K, ELM-HPO, ELM-Reg, ELM-Ensemble. S-LSTM, LSTM-ATTN, B-LSTM, C-LSTM, entre otras.

Finalmente, se recomienda utilizar este estudio como base para realizar pronósticos de volatilidad de otros commodities importantes para el mercado colombiano, como el petróleo, gas, oro, TRM, banano, cacao, entre otros. Es recomendable incluir diferentes variables exógenas que puedan contribuir al modelado y predicción, como el S&P500, las acciones de Ecopetrol o COLCAP. De esta manera, se puede obtener un análisis más completo y preciso de la volatilidad en el mercado colombiano y tomar decisiones informadas en cuanto a inversiones y participación en el mercado de valores.

Referencias bibliográficas

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). *TensorFlow: A System for Large-Scale Machine Learning*. 1–21. https://tensorflow.org.

Adhikari, R., & Agrawal, R. (2012). A Novel Weighted Ensemble Technique for Time Series Forecasting. *https://link-springer-com.bibliotecavirtual.uis.edu.co/book/10.1007/978-3-642-30217-6*, 38-49.

Aditya Bhardwaj, Y. N. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *ScienceDirect*, 1-5.

Alpaydin, E. (2020). *Introduction to Machine Learning Second Edition*.

Al-Dhief, F. T., Latiff, N. M., Malik, N. N., Abbas, M., & Mohammed, M. A. (2021). Voice Pathology detection and classification by adopting online sequential extreme learning machine. *Universiti Teknologi Malaysi*.

Allende , H., & Valle, C. (2017). Ensemble Methods for Time Series Forecasting. *SPRINGER*, 217-232.

Asesh, A. (2022). Normalization and Bias in Time Series Data. *SPRINGER*, 88-97.

Banco de la República de Colombia. (2022). *Banco de la República de Colombia*. Obtenido de https://www.banrep.gov.co/es/estadisticas/trm

Beltrán Martínez, B. (2009). Minería de datos. *Primavera*, 13-25.

Breiman, L. (2001). Random Forest. *Machine Learning*, 6-31.

Briega, R. L. (2016). Series de tiempo con Python. *Matemáticas, análisis de datos y python*.

Brockwell, P., & Davis, R. (2016). Introduction to Time Series and Forecasting . *SPRINGER*, pp 227–257.

Brown, G. (2011). Ensemble Learning. *Encyclopedia of Machine Learning*, 312-320.

Corredor León, D. A. (2023). *danielcorredor99/Proyecto-de-Grado*. https://github.com/danielcorredor99/Proyecto-de-Grado

Chuluunsaikhan, T., Ryu, G.-A., Yoo, K.-H., Rah, H., & Nasridinov, A. (2020). Incorporating Deep Learning and News Topic Modeling for Forecasting Pork Prices: The Case of South Korea. *Chungbuk National University*.

Dagnino, J. (2014). Regresión lineal. *Revista chilena de anestesia*, 7.

Deina, C., do Amaral Prates, M. H., Rodrigues, C. H., Ribeiro Martins, M. S., Trojan, F., Stevan Jr, S. L., & Valadares Siqueira, H. (2021). A methodology for coffee price forecasting based on extreme learning machines. *ELSEVIER*, 2.

Dickey, D., & Fuller, W. (1979). Distributions of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 427-431.

Džeroski, S., Panov, P., & Ženko, B. (2009). Encyclopedia of Complexity and Systems Science. *SPRINGER*, 5317-5325.

Engle, R., & Patton, A. (2007). What good is a volatility model? *ELSEVIER*, 47-63.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. www.aaai.org

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The kdd Process for Extracting Useful Knowledge from Volumes of Data. *Comunications of the ACM* , 27-34.

Federación Nacional de Cafeteros de Colombia. (2017). *FNC en Cifras.* Obtenido de https://federaciondecafeteros.org/static/files/FNCCIFRAS2017.pdf

Federación Nacional de Cafeteros. (2021). *Informe del gerente 89 congreso de cafeteros.*

Federación Nacional de Cafeteros. (6 de 4 de 2022). *federaciondecafeteros*. Obtenido de Producción de café de Colombia cae 13% en marzo: Producción de café de Colombia cae 13% en marzo

Federación Nacional de Cafeteros de Colombia. (2022). *Tebla precio interno de referencia para la compra de café en colombia.* Bogotá.

Friedman, J. (2001). Greddy Function Approximation: A Gradient Boosting Machine. *The annals of statistics*, 1189-1232.

Ghosh, I., & Sanyal, M. K. (2021). Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights*, *1*(2), 100039. https://doi.org/10.1016/J.JJIMEI.2021.100039

Ghosh, I., & Sanyal, M. (2021). Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights*.

Guillermo Westreicher. (2020). Predicción (estadística). *Economipedia*.

Haroon, O., & Rizvi, S. A. (2020). COVID-19: Media coverage and financial markets behavior—A sectoral inquiry. *Elsevier*.

Huang, G.-B., Liang, N.-Y., Rong, H.-J., Saratchandran, P., & Sundararajan, N. (2005). On-Line Sequential Extreme Learning Machine. *the IASTED International Conference on Computational Intelligence*.

Indranil Ghosha, M. K. (2021). Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights*.

International Coffee Organization. (2021). *World coffee consumption.* Obtenido de https://www.ico.org/prices/new-consumption-table.pdf

KumarNarayan, P. (2019). Can stale oil price news predict stock returns? *Elsevier*.

Lahura, E. (2003). EL COEFICIENTE DE CORRELACIÓN Y CORRELACIONES ESPÚREAS.

Lei Chai, H. X. (2020). A multi-source heterogeneous data analytic method for future price fluctuation prediction. *Elsevier*.

Li, G., Chen, W., Li, D., Wang, D., & Xu, S. (2020). *Comparative Study of Short-Term Forecasting Methods for Soybean Oil Futures Based on LSTM, SVR, ES and Wavelet Transformation*. 12007. https://doi.org/10.1088/1742-6596/1682/1/012007

Li, K., Shen, N., Kang, Y., Chen, H., Wang, Y., & He, S. (2021). Livestock Product Price Forecasting Method Based on Heterogeneous GRU Neural Network and Energy Decomposition. *IEEE Access*, 158322-158330.

Li, Y., Jiang, S., Li, X., & Wang, S. (2021). The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach. *Energy Economics*, *95*, 105140. https://doi.org/10.1016/J.ENECO.2021.105140

Liang, Y., Lin, Y., & Lu, Q. (2022). Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM. *Expert Systems with Applications*, *206*, 117847. https://doi.org/10.1016/J.ESWA.2022.117847

Liang, Y., Lin, Y., & Lu, Q. (2022). Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM. Expert Systems with Applications, 206, 117847. https://doi.org/10.1016/J.ESWA.2022.117847

Liu, Y., Liu, S., Ye, D., Tang, H., & Wang, F. (2022). Dynamic impact of negative public sentiment on agricultural product prices during COVID-19. Journal of Retailing and Consumer Services, 64. https://doi.org/10.1016/J.JRETCONSER.2021.102790

Luo, K. J. (2019). Forecasting realized volatility of agricultural commodity futures with infinite Hidden Markov HAR models. *International Journal of Forecasting*.

Mahadeva, L., & Robinson, P. (2009). Prueba de raíz unitaria para ayudar a la construcción de un modelo. *CENTRO DE ESTUDIOS MONETARIOS LATINOAMERICANOS* , 1-10.

Man, Y., Yang, Q., Shao, J., Wang, G., Bai, L., & Xue, Y. (2022). Enhanced LSTM Model for Daily Runoff Prediction in the Upper Huai River Basin, China. Engineering. https://doi.org/10.1016/J.ENG.2021.12.022

Martín, I. G. (2003). análisis y predicción de la serie de tiempo del precio externo del café colombiano utilizando redes neuronales artificiales. *Pontificia Universidad Javeriana*.

MathWorks. (2020). *Aprendizaje automatico*. Obtenido de mathwork: https://la.mathworks.com/discovery/supervised-learning.html?s_tid=srchtitle_aprendizaje%20supervisado_1

MathWorks. (2022). *Cree modelos predictivos a partir de datos de entrada y respuesta conocidos con técnicas de Machine Learning*. Obtenido de https://la.mathworks.com/discovery/supervised-learning.html?s_tid=srchtitle_aprendizaje%20supervisado_1

Ministerio de Agricultura. (15 de 2 de 2022). *Minagricultura*. Obtenido de Proyectamos que la producción de café estará en 13,2 millones de sacos este 2022, lo que representará un crecimiento de 5% frente al año anterior": ministro Rodolfo Zea Navarro: https://www.minagricultura.gov.co/noticias/Paginas/Proyectamos-que-la-producci%C3%B3n-de-caf%C3%A9-estar%C3%A1-en-13,2-millones-de-sacos-este-2022,-lo-que-representar%C3%A1-un-crecimiento-de-5-f.aspx

Ministerio de Comercio, Industria y Turismo. (06 de 04 de 2022). *Perfil de Colombia.* Obtenido de Ministerio de Comercio, Industria y Turismo: https://www.mincit.gov.co/getattachment/1c8db89b-efed-46ec-b2a1-56513399bd09/Colombia.aspx#:~:text=PIB%20per%20c%C3%A1pita%20(PPP%202021,terap%C3%A9uticos%20(2%2C8%25).

Musatafa Abbas Abdul Albad, S. T. (2017). Extreme Learning Machine: A Review. *International Journal of Applied Engineering Research*, 4610-4623.

Narayan, P. K. (2019). Can stale oil price news predict stock returns? *Energy Economics*, 430-444.

National Coffee Association. (10 de 11 de 2017). *https://www.ncausa.org/*. Obtenido de Coffee Around the World: https://www.ncausa.org/About-Coffee/Coffee-Around-the-World

Novales, A. (2017). Midiendo el riesgo en mercados Önancieros. 27-80.

OCDE. (2020). *Coronavirus: The world economy at risk.*

Organización Internacional del Café. (2020). *Efectos de la COVID-19 en el sector mundial del café.*

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. Journal of Machine Learning Research, 12, 2825–2830. http://scikit-learn.sourceforge.net.

Portafolio. (26 de 4 de 2022). *Por qué tomar café es cada vez más caro en Colombia*. Obtenido de https://www.portafolio.co/economia/finanzas/cafe-por-que-tomarlo-es-cada-vez-mas-caro-en-colombia-y-america-latina-564569

QuestionPro. (2022). *QuestionPro*. Obtenido de QuestionPro : https://www.questionpro.com/blog/es/herramienta-de-analisis-de-sentimientos/

Rangan Gupta, C. P. (2022). Forecasting the realized variance of oil-price returns: a disaggregated analysis of the role of uncertainty and geopolitical risk. *Elsevier*.

Romero-Meza, R. ,.-V. (2022). Prophesying the Short-Term Dynamics of the Crude Oil Future Price by Adopting the Survival of the Fittest Principle of Improved Grey Optimization and Extreme Learning Machine. *Mathematics*.

Rotem Zelingher, D. M. (2021). Assessing the Sensitivity of Global Maize Price to Regional Productions Using Statistical and Machine Learning Methods. *ELSEVIER*.

Rotem Zelingher, D. M. (2021). Assessing the Sensitivity of Global Maize Price to Regional Productions Using Statistical and Machine Learning Methods. *Frontiers in Sustainable Food Systems*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 1–21. https://doi.org/10.1007/S42979-021-00592-X/FIGURES/11

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference.*, 1-5.

TaoWang, J. (2010). Nonlinearity and intraday efficiency tests on energy futures markets. *energy economics*.

Tseng, K. K., Lin, R. F. Y., Zhou, H., Kurniajaya, K. J., & Li, Q. (2018). Price prediction of e-commerce products through Internet sentiment analysis. Electronic Commerce Research, 18(1), 65–88. https://doi.org/10.1007/S10660-017-9272-9/FIGURES/8

Villavicencio, J. (2011). Introducción a Series de Tiempo. 1-2.

Villavicencio, J. (2011). Introducción a Series de Tiempo. 6-11.

Wang, J., Athanasopoulos, G., Hyndman, R. J., & Wang, S. (2018a). Crude oil price forecasting based on internet concern using an extreme learning machine. *International Journal of Forecasting*, *34*(4), 665–677. https://doi.org/10.1016/J.IJFORECAST.2018.03.009

Wang, J., Lu, S., Wang, S. H., & Zhang, Y. D. (2021). A review on extreme learning machine. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-11007-7

Wang, X., & Han, M. (2015). Improved extreme learning machine for multivariate time series online sequential prediction. *Engineering Applications of Artificial Intelligence*, *40*, 28–36. https://doi.org/10.1016/J.ENGAPPAI.2014.12.013

Weng, F., Zhang, H., & Yang, C. (2021). Volatility forecasting of crude oil futures based on a genetic algorithm regularization online extreme learning machine with a forgetting factor: The role of news during the COVID-19 pandemic. *Resources Policy*, *73*, 102148. https://doi.org/10.1016/J.RESOURPOL.2021.102148

Wolters, J., & Hassler, U. (2006). Modern Econometric Analysis . 41-55.

Xu, X. (2019). Contemporaneous and Granger causality among US corn cash and futures prices. *European Review of Agricultural Economics*, 663–695.

Xu, X., & Zhang, Y. (2022). Soybean and Soybean Oil Price Forecasting through the Nonlinear Autoregressive Neural Network (NARNN) and NARNN with Exogenous Inputs (NARNN–X). *Intelligent Systems with Applications*.

Ye, K., Piao, Y., Zhao, K., & Cui, X. (2021). A heterogeneous graph enhanced lstm network for hog price prediction using online discussion. *Agriculture (Switzerland)*, *11*(4). https://doi.org/10.3390/AGRICULTURE11040359

Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters*.