MODELACIÓN DEL PRECIO PARA LA COMPRA Y VENTA DE ACEITE DE SOYA EMPLEANDO METODOLOGÍAS DE SERIES DE TIEMPO BASADAS EN MACHINE LEARNING

NIDIA BEATRIZ MUNEVAR QUIROGA
LEONARDO ANDRES PALACIOS CORDOBA

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

DANIEL ENRIQUE GONZALEZ GOMEZ
Director

Diego Luis Linares
Jurado

Julián Gil
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en CIENCIA DE DATOS.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

BOGOTA, 15 DE FEBRERO DE 2024

Acta de Correcciones al Documento de Trabajo de Grado

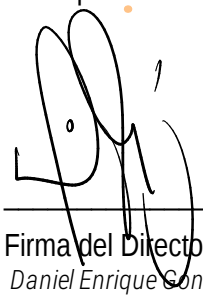
Santiago de Cali, 15 de Febrero de 2024

Autor: NIDIA BEATRIZ MUNEVAR QUIROGA y LEONARDO ANDRES PALACIOS CORDOBA

Título del Trabajo de Grado: MODELACIÓN DEL PRECIO PARA LA COMPRA Y VENTA DE ACEITE DE SOYA EMPLEANDO METODOLOGÍAS DE SERIES DE TIEMPO BASADAS EN MACHINE LEARNING

Director: DANIEL ENRIQUE GONZALEZ GOMEZ

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del Director del Trabajo de Grado
Daniel Enrique González Gómez

Santiago de Cali, 15 de febrero de 2024

Ingeniero
Diego Luis Linares Ospina
Director Maestría en Ciencia de Datos
Facultad de Ingeniería
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Proyecto de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado “Modelación Del Precio Para La Compra Y Venta De Aceite De Soya Empleando Metodologías De Series De Tiempo Basadas En Machine Learning”, el cual fue realizado por la estudiante Nidia Beatriz Munevar Quiroga con código 8972783 y Leonardo Andrés Palacios Córdoba con código 8975315 perteneciente a la Maestría en Ciencia de Datos, bajo la dirección del profesor Daniel Enrique González Gómez.

El suscrito director del Proyecto de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

_____ _____

Nidia Beatriz Munevar Quiroga
C.C. 53.161.082 de Bogotá

Leonardo Andres Palacios Córdoba
C.C. 1152469372 de Medellín

_____

Daniel Enrique González Gómez
C.C. 16.669.372 de Cali

FICHA RESUMEN
PROYECTO DE TRABAJO DE GRADO

TÍTULO: MODELACIÓN DEL PRECIO PARA LA COMPRA Y VENTA DE ACEITE DE SOYA EMPLEANDO METODOLOGÍAS DE SERIES DE TIEMPO BASADAS EN MACHINE LEARNING

1. ÁREA DE TRABAJO: Ciencia de Datos
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Nidia Beatriz Munevar Quiroga y Leonardo Andrés Palacios Córdoba
4. CORREO ELECTRÓNICO: nbmunevarq17@javerianacali.edu.co, leopalacios11@javerianacali.edu.co.
5. DIRECCIÓN Y TELEFONO: Diagonal 45 sur 5H – 41 Bogotá , 3108527750; Carrera 75DA # 2b Sur – 320 Medellín, 3212552350.
6. DIRECTOR: Daniel Enrique González Gómez
7. VINCULACIÓN DEL DIRECTOR: Profesor planta Universidad Javeriana Cali.
8. CORREO ELECTRÓNICO DEL DIRECTOR: dgonzalez@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica):
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5): pronostico, modelo, aceite de soya, ciencia de datos, machine learning, precios.
13. FECHA DE INICIO: Enero 2023
14. DURACIÓN ESTIMADA (En meses): 12 meses
15. RESUMEN:

El proyecto aplicado realizado es la modelación del precio para la compra y venta de aceite de soya empleando metodologías de series de tiempo basadas en técnicas clásicas y en modelos de machine learning que se plantea ante una necesidad de los actores que requieren mejorar sus decisiones y de esta forma su rentabilidad. Los precios de las materias primas afectan directamente al mercado y a los precios de los bienes producidos a partir de estas materias, es decir, los valores terminan impactando al comprador final, por lo que se requiere mejorar los métodos de pronósticos empleados incorporando herramientas de ciencia de datos y de esta manera integrar otros elementos del mercado que afectan los precios y su dinámica. El objetivo fue desarrollar un modelo de series de tiempo basado en técnicas de machine learning capaz de estimar y ajustar el precio del aceite de soya incorporando factores inflacionarios, políticos, de

demanda, histórico de ventas y la cotización del precio del aceite de soya en el mercado de futuros. A partir de la construcción de modelos de ciencias de datos que permitan pronosticar el precio del aceite de soya bajo las restricciones del mercado de futuros, se evalúan los diferentes modelos construidos y selecciona el modelo que ofreció los mejores resultados, con el modelo seleccionado se visualizan los resultados obtenidos mediante una aplicación web que permite a los decisores actuar de manera eficiente y hacer seguimiento al comportamiento de los precios en tiempo real. Como resultado se obtuvo un modelo de pronóstico de precios de compra y venta de materia prima, el dataset resultante del preprocesamiento realizado para obtener el pronóstico, el documento resultante con la descripción de la metodología empleada y el dashboard que permite el monitoreo de precios y visualización del pronóstico de compra y venta de materias primas. Finalmente, la metodología empleada puede ser escalada a otros productos con el propósito de ser utilizada en el pronóstico de otras materias primas agrícolas.

Tras la construcción y evaluación de varios modelos multivariados, el modelo Convolucional fue seleccionado por ofrecer los mejores resultados en términos de precisión, capturando eficazmente la dinámica del mercado. Los resultados obtenidos se visualizan a través de una aplicación web diseñada para facilitar a los decisores una actuación eficiente y un seguimiento en tiempo real del comportamiento de los precios. Como resultado final, se obtuvo un modelo robusto para el pronóstico de precios de compra y venta de aceite de soya, un dataset procesado para el pronóstico, una documentación detallada de la metodología utilizada y un dashboard para el monitoreo y visualización de precios. Esta metodología, tiene el potencial de ser aplicada a la predicción de precios de otras materias primas agrícolas, ampliando su utilidad en diversos sectores del mercado.

MODELACIÓN DEL PRECIO PARA LA COMPRA Y VENTA DE ACEITE DE SOYA EMPLEANDO METODOLOGÍAS DE SERIES DE TIEMPO BASADAS EN MACHINE LEARNING

NIDIA BEATRIZ MUNEVAR QUIROGA
Código 8972783

LEONARDO ANDRES PALACIOS CORDOBA
Código 8975315

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
DANIEL ENRIQUE GONZALEZ GOMEZ

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, FEBRERO 15 DE 2024

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

La soya es uno de los productos agrícolas más populares en todo el mundo debido a su rico valor nutricional, por esta razón el mercado de futuros de la soya es uno de los más desarrollados en cuanto a los productos agrícolas se refiere. Los futuros de la soya son contratos en los que se establece la compra o venta de una cantidad específica de soya en una fecha futura determinada a un precio acordado de antemano.

En el mercado de venta y compra de materias primas agrícolas intervienen diferentes actores, los precios son públicos y son afectados por diferentes variables tales como el precio del petróleo, la tasa de cambio, el clima entre otros elementos. La necesidad de los actores es mejorar sus decisiones y de esta forma su rentabilidad, los precios de las materias primas afectan directamente al mercado y a los precios de los bienes producidos a partir de estas, es decir, los valores terminan impactando al comprador final.

Los métodos de estimación son técnicas en las que se emplean datos históricos con el objetivo de poder hacer análisis de tendencias y patrones sobre el conjunto de datos. Además, para pronosticar se debe suponer que el comportamiento futuro de la variable o las variables de interés son similares a su comportamiento pasado. Los pronósticos se emplean en diversos mercados productivos, en áreas de empresas y en entidades gubernamentales con el fin de tener una guía en la toma de algunas decisiones, donde los modelos estadísticos son utilizados para realizar estimaciones sobre el precio de los futuros agrícolas. Esto ayuda a entender el comportamiento y la volatibilidad en los precios de futuros, así como la correlación que puede existir entre productos agrícolas sustitutos como el maíz y la soya. De igual forma, permite hacer entendimientos desde la de la oferta y la demanda en cuanto a la sustitución de los productos.

Los futuros de soya se negocian en bolsas de valores y en plataformas de trading, estas plataformas se utilizan para proteger a las empresas y los agricultores contra la volatilidad de los precios cambiantes de la soya. Además, para poder operar con éxito los futuros de la soya es importante comprender los conceptos básicos del producto y los factores clave que influyen en su precio como los cambios en la demanda, productos sustitutos y cambios en las políticas gubernamentales.

En lo que se refiere a este trabajo, abordaremos la problemática de pronosticar el precio de los futuros de un producto agrícola como lo es la soya. Para dar solución a esta problemática se realizaron modelos estadísticos y técnicas de aprendizaje automático que permitieron estimar el precio de la soya. Los modelos construidos se evaluaron y se compararon contra los datos históricos para identificar el mejor modelo. Una vez que se definió el modelo multivariado que presenta mejores ajustes se procedió a desarrollar una aplicación web que permite visualizar el pronóstico generado con el algoritmo de mejor rendimiento.

1. DEFINICIÓN DEL PROBLEMA

Actualmente los actores que participan en la compra y venta de materias primas agrícolas han desarrollado una diversidad de estrategias encaminadas a optimizar la rentabilidad utilizando sus experiencias. Debido a que los mercados de futuros cada vez requieren mayores factores y variables para poder ser operados con éxito, los métodos de pronósticos que se emplean habitualmente desde la experiencia quedan cortos, de esta forma el incorporar herramientas de Ciencia de Datos que ayuden a mejorar las estimaciones y a optimizar los procesos de compra y ventas de futuros resultan ganadores, ya que estos pueden integrar y tener en cuenta otros elementos del mercado que afectan los precios y su dinámica.

1.1 PLANTEAMIENTO DEL PROBLEMA

En el mercado de venta y compra de materias primas agrícolas como el aceite de soya intervienen diferentes actores como son cultivadores, intermediarios y compradores que son regulados mediante el registro de sus transacciones en bolsa de valores, precios que son públicos y otros factores como:

Inflación: La inflación está al nivel más alto, en cuatro décadas y se han aumentado los costos de producción de los agricultores. El aumento de los fertilizantes, la energía, los equipos, el transporte, los gastos de mano de obra y el aumento del valor de la tierra significan que los agricultores deben recibir precios más altos de los cultivos para mantener el ritmo de la inflación. Cuando todos los gastos de los insumos aumentan, los precios de la producción tienen que seguir el mismo ritmo, o la producción se vuelve económicamente inviable.

Condiciones meteorológicas: Determinan las cosechas anuales; la sequía o las inundaciones conllevan a una reducción de los suministros, provocando escasez y explosiones de precios.

Crecimiento población mundial: La demanda de productos agrícolas básicos que alimentan al mundo está aumentando con la población mundial que crece a un ritmo de aproximadamente 20 millones de personas cada trimestre.

Cambio en la política energética de EE.UU: Se aleja de los combustibles fósiles en favor de fuentes alternativas y renovables, aumentando las necesidades de etanol a base de maíz y biodiésel a base de soya [1].

Además de las dificultades presentadas en las variaciones de los precios también existen restricciones adicionales que reglamentan las transacciones para evitar los problemas de especulación.

Cada cosecha es negociada en bolsa con un tiempo de caducidad y una ventana de tiempo asociada para realizar las transacciones entre los diversos actores, haciendo posible solo realizar negociaciones hasta una semana antes de la terminación del contrato.

Es importante el desarrollo de un modelo de pronóstico de precios para la compra y venta de aceite de soya empleando metodologías de series de tiempo basadas en machine learning, dado que los actuales pronósticos tradicionales que se apoyan en un análisis univariado al ser evaluados muestran mayores diferencias en las estimaciones de los precios y no tienen en cuenta todos los factores adicionales que afectan los precios.

Se emplean técnicas de ciencia de datos, las cuales tienen un campo interdisciplinar que combina machine learning, estadista, análisis avanzado, minería de datos, big data y programación, con el objetivo de extraer conocimiento oculto y útil a partir de los datos, mediante procesos de descubrimiento o de formulación y verificación de hipótesis [2].

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo se pueden incorporar técnicas de series de tiempo basadas en machine learning a la estimación de precios del aceite de soya (materias primas agrícolas) que permita mejorar la toma de decisiones?

Para abordar esta interrogante de manera exhaustiva, se plantearon las siguientes preguntas secundarias:

- ¿Qué factores clave deben ser modelados con precisión para estimar el precio del aceite de soya, y cómo pueden las técnicas de machine learning optimizar este proceso?
- ¿Cuál es el enfoque más adecuado para incorporar las variables y restricciones del mercado en un algoritmo predictivo que sea capaz de pronosticar con precisión el precio del aceite de soya?
- ¿Cómo se determina cuál modelo proporciona la mejor estimación y ajuste para el precio del aceite de soya, evaluando su fiabilidad y precisión?
- ¿De qué manera se pueden presentar los resultados de la estimación de precios de manera que sean accesibles y útiles para los tomadores de decisiones, permitiendo un seguimiento efectivo y en tiempo real?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo de series de tiempo basado en técnicas de machine learning capaz de estimar y ajustar el precio del aceite de soya incorporando factores inflacionarios, meteorológicos, demanda, histórico de ventas y la cotización del precio del aceite de soya en el mercado de futuros.

2.2 OBJETIVOS ESPECÍFICOS

- Estudiar las características y el comportamiento del precio de la serie de tiempo de los futuros de soya.

- Construir un algoritmo predictivo que permita pronosticar el precio del aceite de soya bajo las incertidumbres y restricciones del mercado de futuros.

- Comparar los diferentes modelos construidos con técnicas estadísticas y machine learning para seleccionar el algoritmo que ofrece el mejor ajuste y resultados.

- Visualizar los resultados obtenidos mediante aplicación web que permita a los decisores actuar de manera eficiente y hacer seguimiento al comportamiento de los precios en tiempo real.

3. MARCO DE REFERENCIA

3.1 MARCO TEÓRICO

A continuación, se presentan los conceptos principales que han servido como base en el desarrollo del proyecto. Estos conceptos están organizados por temas clave como: series de tiempo y machine learning, modelos, herramientas y técnicas analíticas, características de series de tiempo e indicadores de precisión.

- **Series de Tiempo y Machine Learning:**

3.1.1 Series de Tiempo:

Las observaciones de una variable Y que se recaban en el trascurso del tiempo se conocen como datos de una serie de tiempo o simplemente una serie de tiempo. Las observaciones de una serie de tiempo generalmente están relacionadas unas con otras, (autocorrelacionadas) [2]. Esta dependencia genera patrones de variabilidad que pueden utilizarse para pronosticar valores futuros y ayudar en la administración de las operaciones de los negocios.

3.1.2 Machine learning:

Es una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. Sin embargo, machine learning no es un proceso sencillo. Conforme el algoritmo ingiere datos de entrenamiento, es posible producir modelos más precisos basados en datos. Un modelo de machine learning es la salida de información que se genera cuando entrena su algoritmo de machine learning con datos. Después del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Por ejemplo, un algoritmo predictivo creará un modelo predictivo. Cuando se proporciona el modelo predictivo con datos, la salida es un pronóstico basado en los datos que entrenaron al modelo [4].

- **Modelos:**

3.1.3 Metodología Box-Jenkins:

Modelos que generan pronósticos exactos con base en una descripción de patrones históricos de los datos. Los modelos autorregresivos integrados de promedio móvil son una clase de modelos lineales que son capaces de representar tanto series de tiempo estacionarias como no estacionarias [5]. Se basa en un enfoque iterativo para identificar un posible modelo a partir de una clase general de modelos. Luego, el modelo seleccionado se compara con los datos históricos

para ver si describe la serie con exactitud, donde el modelo se encuentran bien ajustado si los residuos son generalmente pequeños y están distribuidos aleatoriamente [2].

3.1.3.1 ARIMA:

Es un modelo que analiza las autocorrelaciones para diversos retrasos temporales. Este análisis permite la comparación entre los patrones de autocorrelaciones muestrales obtenidos de la serie de tiempo y el patrón teórico de autocorrelación asociado a un modelo ARIMA específico. Los pronósticos se derivan directamente de un modelo ajustado [2]. Una serie de tiempo ARIMA tiene la siguiente forma:

$$\Delta^d{}_{y_t} = \delta + \sum_{i=1}^{p} \emptyset_i \, \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \, \varepsilon_{t-i} + \varepsilon_t$$

Donde $\Delta^d y_t$ representa la d-ésima diferencia de la serie, utilizada para convertir series no estacionarias en estacionarias; δ es una constante o término de intercepción; φi son los coeficientes del componente autorregresivo que relacionan una observación con sus p retrasos anteriores; θ_i son los coeficientes del componente de media móvil que relacionan el error actual con q errores pasados; y ϵ_t es el término de error en el tiempo t, reflejando innovaciones o choques aleatorios.

Este modelo integra un enfoque autorregresivo (p) y de media móvil (q) con diferenciación (d) para abordar la estacionariedad de la serie, proporcionando un marco completo para el análisis y pronóstico de series de tiempo mediante la metodología Box-Jenkins. Este proceso incluye la identificación, estimación y verificación del modelo para asegurar la adecuación y precisión de los pronósticos generados [19].

3.1.4 Elman:

La red Elman es una red neuronal recurrente, lo que significa que tiene conexiones que retroceden en el tiempo. Estas conexiones permiten a la red “recordar” entradas anteriores, lo que puede ser útil al trabajar con series temporales [9].

3.1.5 Jordan:

Una red neuronal Jordan es un tipo especial de red neuronal que puede recordar información pasada para ayudar en predicciones futuras. En lugar de simplemente tomar una entrada y producir una salida, esta red toma tanto la entrada actual como su propia salida anterior para

hacer su próxima predicción. Es como si tuviera una pequeña memoria de lo que hizo anteriormente. Si se intenta predecir el clima, en lugar de solo mirar el clima del día actual, también considera la predicción del día anterior. La red Jordan trata de reducir el overfitting (sobreajuste), el cual es un concepto clave en el aprendizaje automático y se refiere a una situación en la que un modelo de machine learning se ajusta demasiado bien a los datos de entrenamiento, capturando incluso el ruido o las pequeñas fluctuaciones en los datos. Como resultado, el modelo tiene un rendimiento deficiente cuando se le presenta nuevos datos que no formaron parte del conjunto de entrenamiento. El overfitting es un problema común que puede llevar a una falta de generalización en los modelos [10].

3.1.6 Prophet:

Prophet es un modelo descomponible de series temporales que tiene en cuenta tres factores importantes: la tendencia, la estacionalidad y los días festivos. Inicio como un software de código abierto que fue desarrollado internamente en Facebook, para afrontar dos de los problemas más comunes en las metodologías de predicción, el primero presente en las herramientas de predicción más automáticas disponibles, las cuales tendían a ser inflexibles e incapaces de ajustarse a suposiciones adicionales; y la segunda la cual hacía que las herramientas de predicción más robustas requirieran un analista especializado en la ciencia de datos [15].

Prophet es especialmente útil para series de tiempo que tienen patrones estacionales fuertes y varios puntos de inflexión o “cambios de tendencia”. Fue diseñado para manejar datos diarios con al menos un año de historia y se espera que funcione bien con datos que tienen patrones estacionales y fechas festivas [6].

3.1.7 Redes Neuronales:

Las redes neuronales imitan la forma como el cerebro del hombre reconoce patrones para identificar información y aprender de sus propios errores. Ellas pueden reconocer objetos, imágenes, audios, textos escritos manualmente, patrones abstractos e ideas. El atractivo de las redes neuronales es su habilidad para descifrar patrones dentro de sistemas complejos. Las redes neuronales han sido aplicadas a las finanzas, en problemas como detección de patrones, asociación y clasificación. Las redes neuronales son más precisas que los modelos algorítmicos tradicionales; ellas reconocen patrones dentro de los datos y ofrecen una alternativa a los métodos convencionales. Las redes neuronales se construyen estructurando en una serie de niveles o capas, compuestas por nodos o capas [3].

La estructura de una neurona cuenta con unos pesos que son constantes y se inicializan aleatoriamente y durante el proceso de aprendizaje serán modificadas. La salida de la neurona se define de la siguiente manera:

$$NET = \sum_{i=1}^{N} X_i w_i + U$$

$$S = f(NET)$$

Donde la salida de la red neuronal está compuesta por la sumatoria de unas entradas (X) que se multiplican por unos pesos (w) que a su vez la multiplicación de los pesos por la entrada, se suman a un umbral (U) para generar la salida (S) de la red neuronal.

3.1.8 Redes Convolucionales:

Las redes neuronales convolucionales (CNNs) aplican la operación de convolución en el dominio discreto, que se fundamenta en la suma ponderada de los valores de píxeles en las imágenes (o de elementos en señales) y un filtro o kernel de convolución. La versión discreta de la convolución, que es central para el funcionamiento de las CNNs, se puede expresar matemáticamente como:

$$s(t)(x * w)(t) = \sum_{a=-\infty}^{\infty} x(a) \cdot w(t-a)$$

En el contexto de las CNNs, esta operación se simplifica aún más debido a la naturaleza finita de las imágenes y señales procesadas, así como de los filtros utilizados. Por lo tanto, para una imagen 2D y un filtro 2D, la convolución se describe como:

$$s(i,j) = (I * K)(i,j) = \sum m \sum n I(m,n) \cdot K(i-m, j-n)$$

donde:

- S(i,j) es el valor de la salida en la posición (i,j),
- I representa la imagen de entrada,
- K es el kernel o filtro de convolución,
- m y n son los índices que recorren el filtro.

Las CNNs utilizan esta operación de convolución para procesar la entrada a través de múltiples capas convolucionales, cada una diseñada para extraer diferentes características de los datos de entrada. Esto permite a las CNNs aprender desde características visuales simples en las primeras capas hasta características más complejas y abstractas en capas más profundas [17].

3.1.9 Redes RNN-LSTM:

Las Redes Neuronales Recurrentes (RNN) son un tipo avanzado de modelo de aprendizaje automático que se especializa en identificar patrones complejos, lo que las hace particularmente eficaces para resolver problemas específicos. Su diseño es ideal para mitigar los problemas asociados con el sobreajuste paramétrico, centrando su capacidad en mejorar la precisión de las predicciones. Dentro de las RNN, una arquitectura destacada es la de las Redes de Larga Memoria de Corto Plazo (LSTM, por sus siglas en inglés Long Short-Term Memory). Las LSTM representan una de las técnicas más avanzadas y exitosas en el campo del aprendizaje profundo, especialmente en aplicaciones como la predicción de series temporales, el reconocimiento de escritura, debido a su habilidad única para capturar dependencias temporales a largo plazo en los datos paramétricos [18].

- **Herramientas y Técnicas Analíticas:**

3.1.10 Minería de texto:

La minería de texto busca encontrar información que sea relevante y útil de una fuente de datos por medio de la identificación, análisis y exploración de patrones donde se pueda extraer información importante para un propósito en particular o para incorporar información cualitativa a un modelo [5].

3.1.11 Promedio Móvil:

Es una técnica que consiste en calcular un valor promedio en base a los datos históricos que se dispongan y utilizar este valor como pronóstico para un periodo en el futuro. Para eliminar la aleatoriedad de los datos obtenidos hay que considerar el promedio de los últimos valores observados y usarlo como pronóstico para un periodo próximo. El número de observaciones que se habrán de utilizar en este promedio se especifica de antemano y permanece constante [16].

Cuando se aplica promedio móvil a una serie de tiempo, cada punto de la serie transformada (promediada) es el promedio de un número determinado de puntos anteriores, actuales y futuros de la serie original. Este número de puntos que decides promediar se llama “ventana” del promedio móvil [7]. Se define de la siguiente manera:

$$\hat{X}_t = \frac{\sum_{t=1}^{n} X_{t-1}}{n}$$

Donde $\widehat{\boldsymbol{X}}_{\boldsymbol{t}}$ promedio de ventanas en unidades del en periodo **t**. Además, $\boldsymbol{X}_{\boldsymbol{t-1}}$ son las ventanas reales en unidades de los periodos anteriores a **t** y **n** es el número de datos.

- **Características de Series de Tiempo:**

3.1.12 Rezago:

El concepto de rezago es fundamental para analizar y modelar series de tiempo porque permite entender cómo los valores pasados pueden influir en los valores presentes o futuros de la serie. Al analizar los rezagos, podemos identificar patrones, hacer predicciones más precisas y entender mejor la dinámica subyacente de los datos [7].

3.1.13 Descomposición de una serie de tiempo:

Se refiere a los patrones o tendencias que se repiten a intervalos regulares, como cada día, mes, trimestre o año, dependiendo de la frecuencia de los datos. En otras palabras, es como un ciclo que se repite en el tiempo.

- Observed: Serie de tiempo original.
- Tendencia (trend): Muestra la dirección general en la que se mueven los datos a largo plazo, sin tener en cuenta las fluctuaciones estacionales o irregulares. Son movimientos a largo plazo en una serie de tiempo que en ocasiones pueden describirse mediante una línea recta o una curva suave. Si la tendencia parece ser aproximadamente lineal, es decir si aumenta o disminuye como una línea recta se representa de la siguiente manera:

$$\hat{T}_t = b_0 + b_1 t$$

  donde $\widehat{\boldsymbol{T}}_{\boldsymbol{t}}$ es el valor pronosticado de la tendencia para el tiempo **t**. El símbolo **t** representa el tiempo, el coeficiente de la pendiente, **b1**, es el incremento o decremento promedio de **T** para cada incremento de un periodo en el tiempo [2].

- Estacionalidad (seasonal): Representa las fluctuaciones que ocurren en intervalos regulares, como los cambios diarios, mensuales o anuales, debido a la estacionalidad.
- Error o Residuo (random): Es la parte de la serie de tiempo que no se puede atribuir ni a la tendencia ni a la estacionalidad. Captura la variabilidad en los datos que no se puede explicar por los otros dos componentes [8]. Es decir que, un residuo es la diferencia entre un valor real observado y su valor de pronóstico, se calcula de la siguiente forma:

$$e_t = Y_t - \hat{Y}_t$$

donde $\boldsymbol{e_t}$ representa el error de pronóstico en el periodo **t**. El símbolo $\boldsymbol{Y_t}$ es el valor real en el periodo **t** y $\boldsymbol{\hat{Y}_t}$ valor del pronóstico en el periodo de tiempo.

3.1.14 Estacionariedad:

Una serie es estacional cuando un patrón relacionado con el calendario se repite a sí mismo durante un intervalo de tiempo específico (generalmente un año). Por lo tanto, las observaciones de la misma posición, en diferentes periodos estacionales, tienden a estar relacionadas [2].

La prueba de Dickey-Fuller, específicamente la prueba ADF (Augmented DickeyFuller), es una prueba estadística utilizada para determinar si una serie temporal tiene una raíz unitaria, es decir, si es no estacionaria y presenta alguna forma de estructura temporal como una tendencia o una estacionalidad [8]. Por tanto, una serie de tiempo estacionaria es aquella cuyas propiedades estadísticas básicas, como la media y la varianza, permanecen constantes en el tiempo, es decir que, varía alrededor de un nivel fijo sin crecimiento ni decrecimiento a medida que pasa del tiempo, por tanto, se dice que es estacionaria [2].

3.1.15 Diferenciación:

Diferenciar una serie temporal es un proceso utilizado para hacer que una serie no estacionaria se vuelva estacionaria. Su objetivo es transformar la serie de datos para estabilizar la media de la serie temporal, eliminando tendencias y efectos estacionales. En otras palabras, se busca que las propiedades de la serie (como la media y la varianza) no cambien con el tiempo [8].

- **Indicadores de precisión:**

3.1.16 MSE (Mean Squared Error)

Esta métrica representa el promedio de los errores cuadrados entre las predicciones y los valores reales. Penaliza fuertemente los errores grandes, es muy útil para saber qué tan cerca es la línea de ajuste de la regresión a las observaciones [2]. Su estimación se hace de la siguiente forma:

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

Donde $\boldsymbol{y_i}$ es el resultado real esperado y $\boldsymbol{\hat{y}_i}$ es la predicción del modelo.

3.1.17 MAE (Mean Absolute Error)

Representa el promedio de las diferencias absolutas entre las predicciones y los valores reales. Proporciona una idea del tamaño del error sin considerar su dirección. El MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio [2]. Su formulación se representa de la siguiente manera:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Donde $\boldsymbol{y_i}$ es el resultado real esperado y $\boldsymbol{\hat{y}_i}$ es la predicción del modelo.

3.1.18 RMSE (Root Mean Squared Error)

Es la raíz cuadrada del MSE y proporciona una idea del tamaño del error en las mismas unidades que la variable de interés, permitiendo que su interpretación sea más fácilmente entendible al tener los valor en las unidades originales [2]. Su representación es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

$$RMSE = \sqrt{MSE}$$

3.2 ANTECEDENTES

Las operaciones con futuros son contratos legalmente establecidos para la compra y venta de activos en una fecha específica, estos iniciaron con los mercados centrales de granos, donde los agricultores pueden vender sus productos para una entrega inmediata o entrega futura, y el precio es la única variable remanente. Inicialmente las operaciones comenzaron con la comercialización de materias primas agrícolas tradicionales, como cereales y ganado en mercados bursátiles. El creciente interés en los mercados globales ha acelerado la atención de los medios y ha atraído el interés de operadores de todo el mundo, donde a partir del estudio de los mercados, los operadores determinan la dirección prevista de los precios de las materias primas [11]. El mercado de derivados y sus negociaciones, tales como las operaciones con futuros, involucran riesgos de perdidas sustanciales donde los factores como lo es el país juegan un papel importante, un ejemplo es China, que se embarca en el aumento de la producción doméstica de commodities importados, para disminuir su dependencia de mercados externos, principalmente en soja, maíz, canola y trigo, lo que conlleva a vender grano doméstico caro y reemplazar con granos baratos importados [12].

La necesidad de pronósticos está en todas las líneas funcionales de las organizaciones, así como en todos tipos de sectores productivos. Los pronósticos son absolutamente necesarios para

avanzar en el ambiente del negocio actual, el cual es cambiante y altamente interactivo; donde intervienen factores importantes en el ajuste de los precios, como lo son las tasas de interés, el crecimiento económico e indicadores inflacionarios. Los métodos de análisis de regresión, descomposición, suavizamiento y promedios móviles autorregresivos son técnicas para realizar pronósticos, basadas en datos, las cuales son altamente eficientes. Con la difusión de los modelos computacionales y la amplia disponibilidad de paquetes de software avanzado, se generan fácilmente pronósticos de valores futuros para las variables de interés, lo que ha permitido obtener resultados más exactos [2]. Según [13], en su estudio indica que las redes neuronales son algoritmos que ayudan capturar características no lineales de los índices de bolsa y han probado que pueden ser entrenadas con información suficiente para identificar dichas relaciones no lineales entre los valores de entrada y salida, donde la superioridad de los modelos basados en redes neuronales con respecto a otras técnicas se debe a que estos capaces de identificar valores atípicos y datos erróneos para mejoran las estimaciones, ya que hay valores que están muy ligados al comportamiento de la economía y sus variables. [14] empleo redes neuronales para pronosticar series temporales con el fin de predecir lluvias en la toma decisiones agrícolas, en este estudio se incorporaron datos masivos, con el uso de nuevas tecnologías como el Big Data orientadas a obtener mayor velocidad, veracidad, precisión y eficacia en los pronósticos.

En ciencias de datos las técnicas empleadas para pronosticar series de tiempo son muchas; se pueden construir pronósticos con técnicas estadísticas, modelos de Machine Learning (aprendizaje de maquina) y algoritmos de Deep Learning (aprendizaje profundo). En el presente trabajo, se desarrollan diferentes modelos empleando las técnicas mencionadas anteriormente, desde modelos más sencillos hasta construir algoritmos más complejos particularmente, aquellos que son útiles en la estimación de pronósticos para series de tiempo.

4. METODOLOGÍA

En el contexto de este trabajo, se utiliza la metodología CRISP-DM, la cual es común en el desarrollo de proyectos y trabajos de ciencias de datos. Esta metodología cuenta con una serie de fases que van entrelazadas en una forma cíclica, esto indica que el proyecto no acaba una vez que se arrojan los resultados, sino que se pueden generar nuevas iteraciones, con el objetivo de mejorar el proyecto. La cual esta descrita y definida en los siguientes pasos:

- Entender el modelo de negocio. En este caso, comprender el mercado de futuros, específicamente para la soya.
- Comprensión y análisis de los datos.
- Preparación y transformación de datos.
- Construcción de los modelos para series de tiempo.
- Evaluación de los modelos de series de tiempo obtenidos.
- Desarrollo de un tablero de control o dashboard.

4.1 ENTENDIMIENTO DEL MERCADO DE FUTUROS DE SOYA

Se identificaron las variables que teóricamente pueden influir en el aumento o disminución del precio del precio del aceite de soya, se hizo una aproximación a los datos e información que puede ser relevante, tomando como referencia estudios que se han realizado sobre el pronóstico de futuros y materias primas. Investigación acerca del mercado de futuro, sus características y tendencias actuales e históricas.

Las variables escogidas para llevar a cabo el modelamiento para determinar el precio del aceite de soya fueron:

- Aceite de soya
- Precio del petróleo
- Precio del dólar
- Bien sustituto (maíz)
- Clima del país productor (Argentina)
- Situación política del país productor (Argentina)
- Situación económica país productor (Argentina)
- Efecto de la guerra de Ucrania
- Situación económica país consumidor (China)

Entre los principales productores de soya a nivel mundial se encuentran países como Brasil (114 millones de toneladas), Estados Unidos (96 millones de toneladas) y Argentina (55 millones de toneladas). Se escogió Argentina dada su situación política, acceso a datos e idioma. Y el país

consumidor predominante es China, consume el 60 % de la producción internacional, el consumo de soya empezó a aumentar desde que se descubrió que satisfacía las necesidades nutricionales del ganado y las aves, que exigen raciones de alta calidad nutricional y sanitaria, así como una elevada densidad energética y contenido proteico, sumado a que cerca de la mitad de los cerdos del mundo están en China, alimentándose estos principalmente de soya, como factor adicional, China también utiliza grandes cantidades de soya en la alimentación de peces de piscifactoría.

Actualmente tenemos acceso a través de una conexión a la plataforma yahoo finance a los datos de las siguientes variables: aceite de soya, petróleo, precio del dólar, bien sustituto (maíz). El tener acceso a dichas variables ayuda a entender de los datos y a la generar y construir las tablas requeridas para desarrollar análisis exploratorios.

Con las variables, clima del país productor (Argentina), situación política del país productor (Argentina), situación económica país productor (Argentina) y situación económica país consumidor (China). Son variables que requieren de un proceso más largo y tedioso para poder acceder a los datos, dado que estas variables tienen una fuente información diferente para cada una:

- Clima del país productor (Argentina): Del país productor Argentina se escogió la ciudad de Buenos Aires la cual es una de las ciudades en las que se produce soya. La temperatura varía en cada ciudad más aun teniendo en cuanta la posición geográfica del país, y por lo tanto no es un dato uniforme.

  La conexión fue posible a través de la plataforma openweather (https://openweathermap.org/) la cual nos permite un periodo de prueba de 15 días y se validó el reemplazo de la misma por una plataforma sin restricción en el acceso.

  Se tuvo acceso a la plataforma web de estadísticas de la ciudad de Buenos Aires https://www.estadisticaciudad.gob.ar/eyc/?p=27702 y se obtuvo la serie de tiempo desde Enero 1991 a mayo 2023, al validar los datos se identifica que se deben hacer transformaciones a los datos, porque la información está organizada por columnas para cada año y no por filas.

- Situación política del país productor (Argentina): Esta es una variable dummy que indica el cambio de periodo presidencial en el cual es posible la reelección, donde el valor de tiene el valor de 1 indican las fechas en las que se dieron cambios de gobiernos.

  Esta variable es relevante porque los cambios de gobierno y de política generan un cambio en el panorama político de un país.

- Situación económica país productor (Argentina): Esta variable se define por el IPC (Índice de Precios al Consumidor Nacional) del país y al cual estamos accediendo a través de la página del gobierno argentino https://www.datos.gob.ar/series/api/series/?ids=148.3_INIVELNAL_DICI_M_26 el cual contiene diferentes dataset.

  La serie de tiempo tiene datos desde diciembre de 2016 para el índice de precios del consumidor, el cual nos permite identificar la variación de la economía Argentina.

- Efecto guerra de Ucrania: Ucrania y Rusia juegan un papel importante en la producción de fertilizantes como UREA, Amoniaco Anhidro, Nitrogenados, Potásicos y Fosfatos y se maneja por lo tanto una variable dummy binaria la cual nos permitirá identificar como fue afectada la agricultura, en especial la producción de soya en el periodo de la guerra.

- Situación económica país consumidor (China): China y Estados Unidos son los mayores países consumidores de soya, lo usan para el consumo de sus animales en gran parte. El Ipc (Índice de Precios al Consumidor Nacional) de China no es una información abierta, solo es publicada de forma esporádica con una conclusión en la que se determina que China realizó una apertura o cierre de su mercado, se emplea esta información para tener una variable dummy binaria que ayuda a identificar el estado de la situación económica de China.

El uso de variables dummy permite incluir en el modelo variable cualitativas para tener en cuenta todos los factores y evaluar cómo se afecta el modelo de acuerdo a la ocurrencia de la misma.

4.1.1 Búsqueda de datos información

Se consultaron las principales variables que se requieren para estudiar el comportamiento del precio del aceite de soya a través de fuentes que monitorean y estudian los mercados financieros como es el caso de yahoo finance, en la que se obtiene el histórico del precio del petróleo, dólar y el precio del maíz como sustituto del aceite de soya. Además, se consultan fuentes de información específicas de los países que producen aceite de soya, como lo son su ipc y las temperaturas medias que se han tenido en dichos países.

4.2 ANÁLISIS EXPLORATORIO DE DATOS

En esta tarea se analizó y se investigaron las características que se encuentran alrededor de los datos que se extrajeron de las diferentes fuentes de información, donde se observa que tipo de datos son, cantidad de datos por cada variable, medias, medianas, frecuencias, máximos y mínimos. De igual forma, se visualizan los datos con el fin de tener un mejor panorama del comportamiento, la tendencia y los patrones que se puedan evidenciar, esto nos permitió

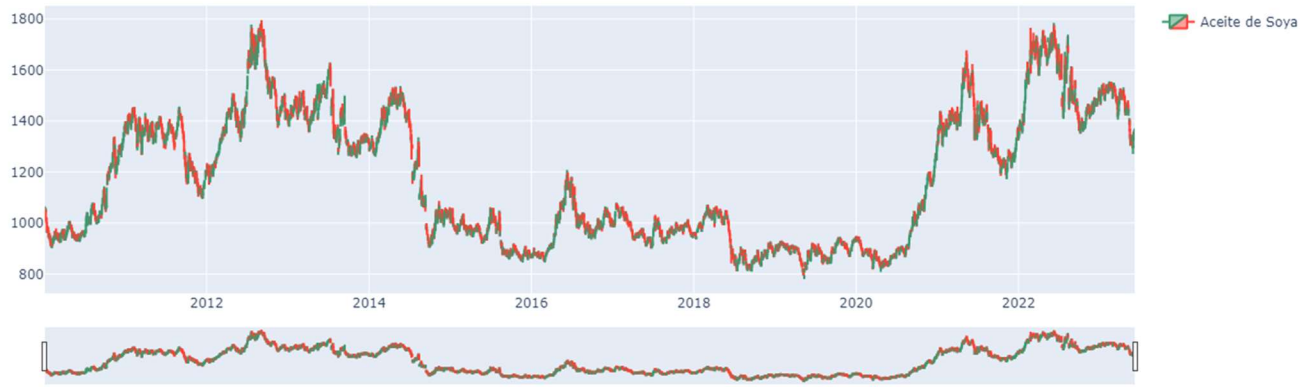
entender la manera con la que se deben manipular los datos al momento de hacer las trasformaciones y modelos.

4.2.1 Precio aceite de soya

Se estudia preliminarmente la tendencia histórica del precio de los futuros de soya, esto permite observar su comportamiento en cuanto a apertura y cierre en una ventana de tiempo determinada. El rango de tiempo de la serie de datos del aceite de soya es del 1 de Enero de 2010 hasta el 30 de Septiembre de 2023.

A continuación, en la Figura 1 se presenta la serie de tiempo precio futuros de soya.

Figura 1. Serie de tiempo precio futuros de soya.



Fuente: Elaboración propia.

Se puede observar que entre enero del 2010 y septiembre del 2023 la serie de tiempo del precio de los futuros de soya presenta picos de aumento y disminución (máximos y mínimos), donde los valores más bajos que tomo el precio de los futuros de soya fueron en el año 2016 y entre el período comprendido entre los años 2018 a 2020, posterior a esta última caída el precio ha vuelto a crecer, sin presentar caídas que lleguen a los valores entre los periodos anteriormente mencionados. Este análisis ayuda a observar el tratamiento de datos que se debe realizar sobre el precio de los futuros de soya, ya que en algunos periodos se observa que no hubo medición en la plataforma donde se obtuvo la información.

Se utilizan los conceptos de promedios móviles, rezago, descomposición, estacionariedad y diferenciación con el fin de comprender y modelar los patrones y las relaciones dentro de los datos temporales de la serie antes de aplicar los modelos que permiten evaluar el impacto de las variables adicionales mencionadas (precio del petróleo, precio del dólar, etc).

4.2.2 Promedio Móvil

Se aplica promedio móvil a la serie de tiempo del precio de los futuros de soya, donde cada punto de la serie transformada se obtiene el promedio de un número determinado de puntos anteriores. Este proceso se realiza para determinar y encontrar tendencias que se puedan apreciar en el comportamiento de la serie tempo.

A continuación, en la Figura 2 se presenta la serie de tiempo del precio de futuros de soya con promedios móviles.

Figura 2. Serie de tiempo del precio de futuros de soya con promedios móviles.

Fuente: Elaboración propia.

Entre los años 2013 y mediados del 2014 se pueden ver cambios en la tendencia de la serie de tiempo de futuros de la soya, tanto para el promedio móvil de 200 días, como para el de 500 días, el cual es más marcado. Entre los años 2021 y mediados del 2023 se puede apreciar cambios en la tendencia de la serie de tiempo de futuros de la soya, tanto para el promedio móvil de 200 días, como para el de 500 días el cual es más marcado. Se podría llegar a validar por medio de un mayor estudio de este tiempo si la afectación fue causada por el desarrollo de la pandemia del covid-19 la cual inicio en marzo de 2020 e inicio a retroceder en Agosto de 2021 cuando se inició el uso de las vacunas. Al suavizar las fluctuaciones menores, a través de los promedios móviles se logró resaltar las tendencias subyacentes en los datos.

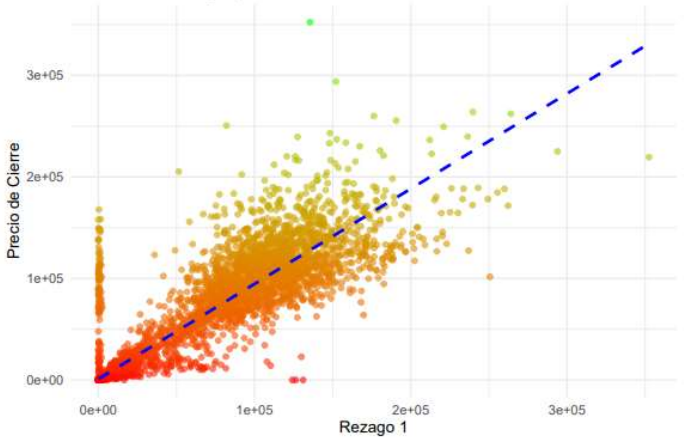
4.2.3 Rezagos

Una parte importante en el estudio de las características y el comportamiento de la serie de tiempo del precio de los futuros de soya, es analizar los rezagos de la serie de tiempo del precio

de los futuros de soya, para comparar dichos rezagos con los valores de la serie original, esto con el fin de poder determinar si existe correlación entre los rezagos y los valores de la serie de tiempo original.

A continuación, en la Figura 3 se presenta el gráfico de rezagos.

Figura 3. Gráfico de rezagos.



Fuente: Elaboración propia.

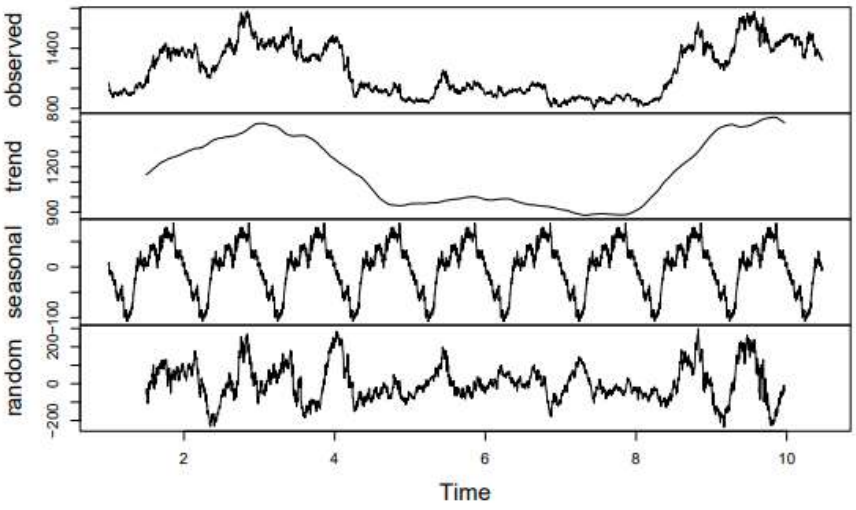
Se puede observar un patrón claro o una agrupación de puntos en el gráfico de rezago 1, por lo tanto, es probable que exista una autocorrelación significativa. Se puede considerar modelos de series de tiempo como ARIMA que toman en cuenta la autocorrelación para hacer predicciones más precisas de ser necesario.

4.2.4 Descomposición de la serie de tiempo

La descomposición de una serie de tiempo es un procedimiento matemático que consiste en dividir la serie de tiempo analizada en varias series temporales diferentes con el objetivo comprender y estudiar el comportamiento de la serie de tiempo para buscar patrones y confirmar tendencias evidenciadas.

A continuación, en la Figura 4 se presenta la descomposición Serie de tiempo precio futuros de soya.

Figura 4. Descomposición Serie de tiempo precio futuros de soya.



Fuente: Elaboración propia.

Es posible deducir que, la serie de tiempo de los precios del aceite de soya muestra patrones claros y consistentes, esto sugiere que la serie temporal tiene ciclos regulares que se repiten a intervalos fijos; se pueden identificar en qué momentos del ciclo tienden a ocurrir los valores altos y bajos de la serie.

4.2.5 Estacionariedad

Se emplea la prueba de Dickey-Fuller para saber si la serie de tiempo es estacionaria o no. Una serie estacionaria es aquella cuyas propiedades, como la media y la varianza, no cambian con el tiempo. En esta prueba, tenemos algo llamado valor p, que es como un termómetro que nos dice qué tan seguros estamos de la estacionariedad de la serie de tiempo estudiada. Un valor p pequeño (menor que 0.05) nos indica que la serie es estacionaria. Mientras que un valor p grande (mayor que 0.05) nos dice que la serie no es estacionaria. En este caso, el valor p es 0.5657, que es bastante grande, es decir que, la serie de tiempo del precio de los futuros de soya no es estacionaria.

4.2.6 Diferenciación

Antes de realizar cualquier diferenciación (d = 0), el valor p de la prueba de Dickey-Fuller aumentada es 0.5657422, lo que es mayor que 0.05. Por lo tanto, no puedes rechazar la hipótesis nula de que existe una raíz unitaria, y se concluye que la serie original no es estacionaria. Después de diferenciar la serie una vez (d = 1), el valor p de la prueba de Dickey-Fuller aumentada es 0.01, lo cual es menor que 0.05.

La serie de tiempo original no es estacionaria, pero después de realizar una diferenciación, la serie

resultante sí es estacionaria. Fue necesario transformarla o diferenciarla para eliminar la tendencia y estabilizar la varianza, antes de aplicar modelos de series temporales como ARIMA.

4.2.7 Precio del petróleo

En el contexto estudiado en el presente trabajo, se estudian los precios del petróleo, ya que estos pueden impactaren directamente los futuros agrícolas desde múltiples formas. Debido a que las subidas en los precios del petróleo pueden estimular la demanda combustible derivados del petróleo, lo que conllevaría a empujar hacia arriba los precios de los aceites vegetales, entre los que se encuentran la soya.

El petróleo es la principal clave de combustible para el transporte de productos agrícolas, un alza de los precios del petróleo impacta en los costos logísticos, limitando la competitividad de las economías más distantes de sus mercados de consumo. La exportación de productos agrícolas se vuelve más complejo con los costos de transporte subiendo.

A continuación, en la Figura 5 se presenta la serie de tiempo precio del petróleo.

Figura 5. Serie de tiempo precio del petróleo.

Fuente: Elaboración propia.

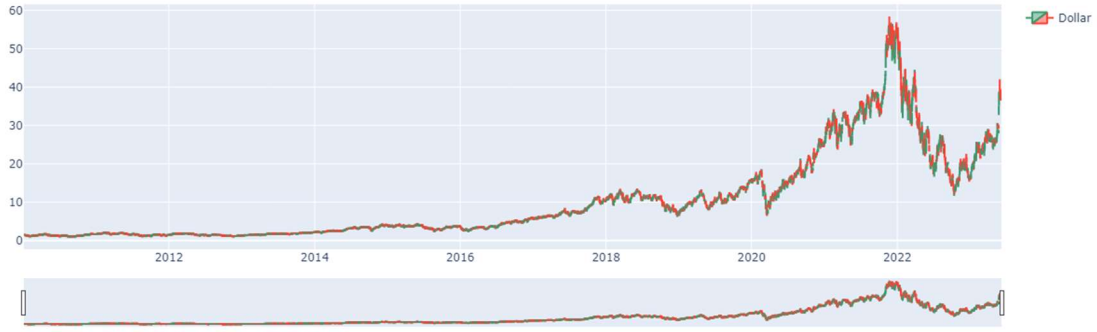
En la Figura 5 se analiza el comportamiento del precio del petróleo crudo porque dentro de los precios de producción del aceite de soya se encuentra el factor del transporte, donde los insumos, materias primas y productos deben ser transportados. Además, en la figura 5 se observa un aumento significativo en el precio del petróleo, se visualiza que en algunos periodos está alcanzando y superando los niveles más altos obtenidos desde 2014. Este tipo de tendencias ejerce una presión al alza en el precio de futuros como lo son la soya y el maíz.

4.2.8 Precio del dólar

La cotización del precio de los futuros agrícolas en los mercados financieros está dada en dólares,

por esta razón se estudia el comportamiento histórico del precio del dólar, ya que la tendencia de esta moneda influye en el valor del precio en el mercado de los futuros.
A continuación, en la Figura 6 se presenta la serie de tiempo precio del dólar.

Figura 6. Serie de tiempo precio del dólar.



Fuente: Elaboración propia.

En la Figura 6 se puede apreciar que a partir del año 2020 se ha presentado un aumento significativo en el dólar en comparación a los años anteriores donde los niveles se ven con una tendencia medida. El alza del dólar a partir del año 2020 coincide con el inicio de la pandemia covid-19.

4.2.9 Bien Sustituto (maíz)

Como se estudió anteriormente la inflación de los países productores empujan alza de la soya, por esta razón y los conflictos geopolíticos es importante analizar el comportamiento de un bien sustituto como lo es el maíz, lo que permite ver si la tendencia es igual a la de la soya y si tienen relación los precios de ambos futuros.

A continuación, en la Figura 7 se presenta la serie de tiempo precio del maíz.

Figura 7. Serie de tiempo precio del maíz.

Fuente: Elaboración propia.

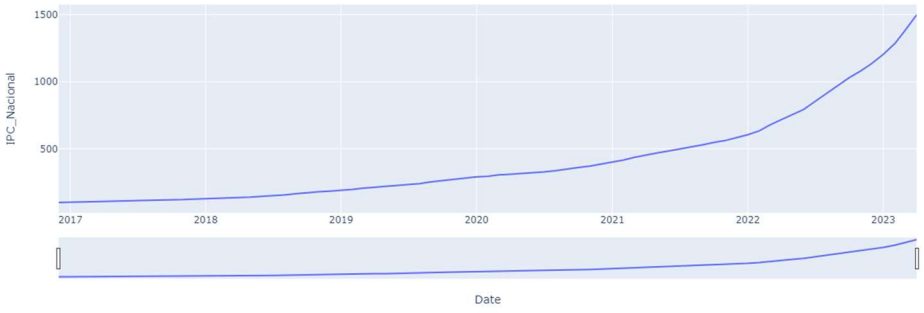
Se puede observar que la serie se tiempo del precio del maíz en ciertos periodos de tiempo su comportamiento es similar al que se produce con la serie del precio de los futuros de soyas, debido a que son bienes que satisfacen un tipo de necesidad parecido se esperaría que un aumento en el precio de uno de los bienes disminuya su consumo y aumente el consumo del otro bien. Analizando el comportamiento de las series de tiempo el comportamiento de ambos bienes en cuando a su precio en el mercado de futuros es similar.

4.2.10 Situación económica país productor (IPC Argentina)

Además de estudiar el comportamiento histórico del aceite de soya, también se estudian variables que nos ayudan entender la situación de los países productores de los futuros de soya, como lo es Argentina, donde analizamos el índice de precio al consumidor, ya que es una variable de influye en la relación positiva o negativa con el precio del aceite de soya.

A continuación, en la Figura 8 se presenta la serie de tiempo de la situación económica país productor (IPC Argentina)

Figura 8. Serie de tiempo de la situación económica país productor (IPC Argentina)



Fuente: Elaboración propia.

Estudiar el ipc de un país productor como lo es Argentina, nos permite evidenciar que el ipc de Argentina presenta un constante crecimiento a lo largo del tiempo. Los temas inflacionarios afectan la economía y es pertinente revisar lo que sucede en los países productores, debido a que el aumento del ipc hace que los gastos de insumos para producir futuros agrícolas (aceite de soya) aumenten, lo que conlleva a los precios de la producción sigan el mismo patrón.

4.2.11 Efecto de la guerra de Ucrania

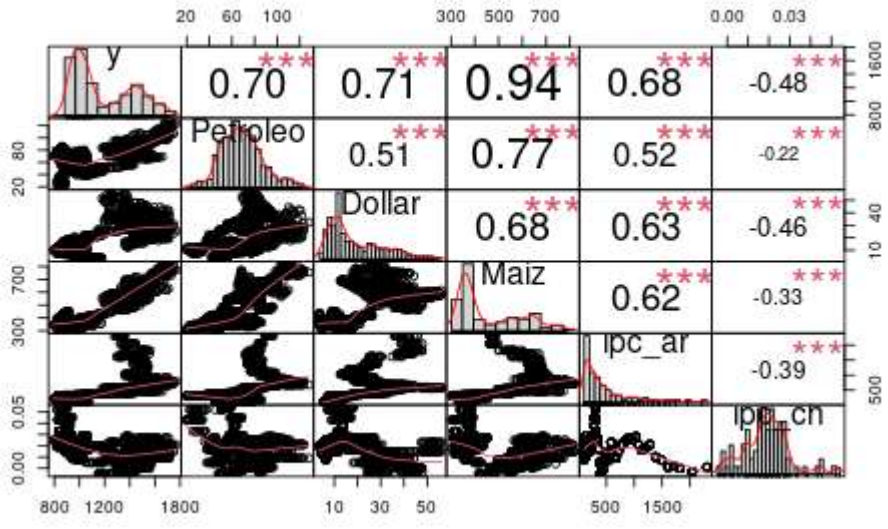
No solo los temas inflacionarios y de transporte influyen en el alza de los precios de las materias primas agrícolas, las tenciones geopolíticas hacen que se provoquen movimientos en los precios de las meterías primas, como lo es el conflicto entre Rusia y Ucrania, que puede generar un detonante en los precios de materias primas agrícolas. Rusia es de las principales exportadoras de este tipo de futuros, en especial del trigo, la tensión que se tiene con Ucrania puede crear escasez con los futuros debido a la producción y el transporte por bloqueos, lo que genera poca oferta aumentando los precios.

4.2.12 Correlación de variables

A una vez que se tienen los datos y el comportamiento de las variables o factores que pueden afectar el precio de los futuros de soya, se hace análisis de correlación para describir el grado de relación lineal que existente entre las variables cuantitativas, para este análisis se dispone de la variable objetivo que hace referencia al precio de los futuros de soya y un conjunto de variables regresoras o exógenas en este caso que permiten calcular el indicador (coeficiente de correlación) que mide la relación asociación lineal entre las variables.

A continuación, en la Figura 9 se presenta la Matriz de correlación.

Figura 9. Matriz de correlación.



Fuente: Elaboración propia.

El coeficiente de correlación de Pearson indica que cuando el indicador toma valores positivos entre (0.5 y 1.0) se puede decir que hay una relación lineal positiva fuerte entre el precio del aceite de soya y variables externas como el petróleo, el dólar, el precio del maíz y el ipc de Argentina, es decir que, cuando se da el crecimiento del precio de los futuros de soya, se asocia con un crecimiento de las variables exógenas. Con respecto al ipc de China que es un país consumidor, se puede observar que existe una relación negativa débil con el precio de los futuros de soya.

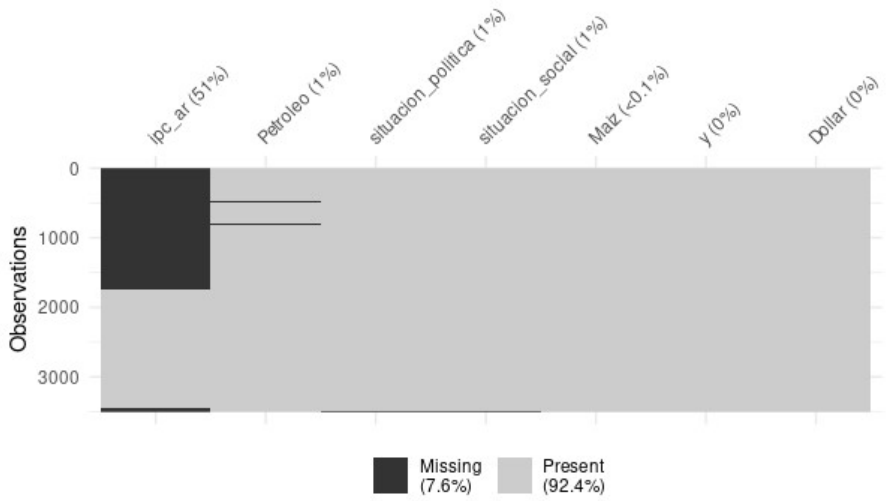
4.3 PREPARACIÓN Y TRANSFORMACIÓN DE DATOS

Una vez se tienen las fuentes de información a los diferentes portales que hacen reportes y miden periódicamente las variables que requerimos para el proyecto y con las fases de entendimiento del negocio y comprensión del comportamiento de los datos, inicia la etapa en la que se organizan los datos en pro de los objetivos trazados, se limpian, se clasifican, se seleccionan o se sustituyen datos.

Para la etapa de la preparación de los datos, se emplea el software R, el cual ayuda hacer el tratamiento de la información, de esta forma se obtienen los dataset o tablas que se van a utilizar a lo largo de la construcción de los modelos de series de tiempo.

En el análisis exploratorio de datos preliminar se encontraron datos faltantes, esto debido a que en las fuentes de información reportadas en algunas fechas no se reportaron los precios e indicadores. Por medio de un mapa de datos faltantes, se identifica el número de datos faltantes para cada una de las variables estudiadas.

Figura 10. Mapa de valores faltantes.



Fuente: Elaboración propia.

En el mapa de valores faltantes, se puede apreciar que del total del conjunto de datos hay un 7.6% de datos nulos, donde la variable del ipc de Argentina es el campo con mayor cantidad de datos faltantes, esta variable tiene un 51% de datos faltantes y la siguen el precio del petróleo, situación política y situación social.

Para dar solución a los datos faltantes que se encuentran en los diferentes campos, se emplea la técnica de vecinos más cercanos que se encarga se rellenar los valores faltantes para un conjunto de datos utilizando información de los vecinos más cercanos. La idea principal detrás de este método es estimar los valores desconocidos basándose en los valores conocidos de observaciones similares, en este caso se hace promediando los valores de los vecinos, asignando un peso basado en la similitud.

4.4 CONSTRUCCIÓN DE LOS MODELOS ESTADÍSTICOS PARA PROYECCIONES

Una vez que se ha realizado el estudio y preprocesamiento de los datos de la serie temporal del precio de futuros de soya, se da inicio a la etapa de creación de modelos predictivos. Donde se implementaron modelos ARIMA, Prophet, redes neuronales Elman y Jordan, empleando exclusivamente la serie temporal del precio de futuros de soya como variable independiente, lo que establece una base univariada para comparaciones futuras. Esta fase inicial del modelado permite establecer un punto de referencia esencial antes de incorporar variables externas en modelos más complejos, como Prophet, redes neuronales convolucionales y redes neuronales recurrentes LSTM, con el fin último de analizar, comparar y evaluar la precisión de estos modelos en la predicción del precio del aceite de soya en el mercado de futuros.

En la fase de modelado de los algoritmos univariados y multivariados del proyecto se realiza una partición del conjunto de datos en entrenamiento y prueba. Para el conjunto de entrenamiento se toman los datos en el periodo de tiempo comprendido entre junio de 2016 a septiembre del 2023, obteniendo de esta forma 1.844 registros para el entrenamiento de los modelos. Por otra parte, para el conjunto de prueba se selecciona una ventana de tiempo de 15 días, entre el 30 de septiembre al 14 de octubre del 2023, la cual sería la ventana de tiempo a pronosticar; esto debido a que en el mercado de futuros agrícolas las negociaciones sobre los contratos de productos agrícolas en especie negociados tienen diversas fechas de vencimiento mensual, denominadas meses contractuales donde se ejecutan los traspasos entre vendedores y compradores, por esta razón no es viable tomar una ventana de tiempo superior a un mes que es donde se dan los intercambios en el mercado de futuros agrícolas.

La principal métrica de error para evaluar el rendimiento de los modelos fue el RMSE (Root Mean Squared Error). Sin embargo, también se consideraron errores como el MSE (Mean Squared Error) y el MAE (Mean Absolute Error) para una evaluación más completa.

4.4.1 Modelos de series de tiempo Univariados

Los modelos de series de tiempo univariados seleccionados para evaluar la predicción de la serie del aceite de soya fueron: modelo ARIMA, Prophet, Redes neuronales Elman y redes neuronales Jordan.

Para interpretar y comprobar la efectividad de las predicciones, se usaron las métricas de rendimiento como el Error Cuadrático Medio (MSE), el Error Medio Absoluto (MAE) y la Raíz del Error Cuadrático Medio (RMSE), para las cuales tuvimos acceso a los datos reales de los futuros de soya y a las predicciones arrojadas en cada modelo.

4.4.1.1. ARIMA

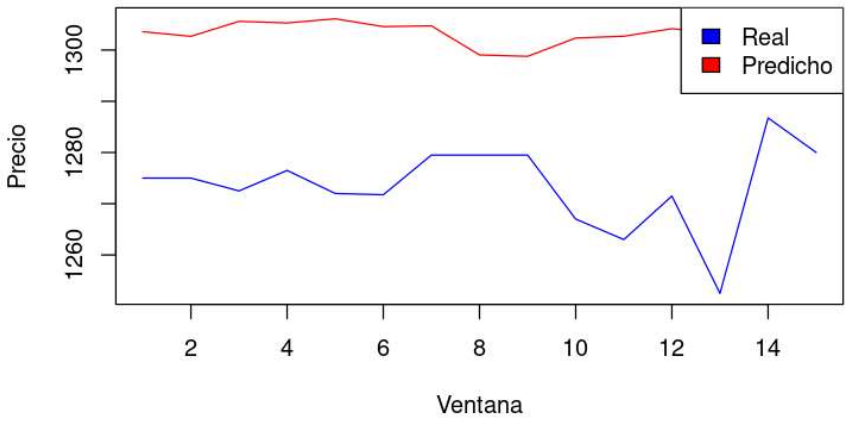
El modelo ARIMA, aplicado al análisis univariado de la serie temporal del precio del aceite de soya, muestra un Mean Squared Error (MSE) de 64.3602, un Mean Absolute Error (MAE) de 5.78735, y un Root Mean Squared Error (RMSE) de 8.022481. Estas métricas reflejan un aceptable nivel de precisión, indicando que el modelo es capaz de predecir los precios con errores moderadamente bajos. El MSE sugiere que, en promedio, las diferencias cuadradas entre los precios reales y predichos no son muy significativas, mientras que el MAE apunta a una desviación media aceptable en las predicciones del modelo. Por su parte, el RMSE, al ser ligeramente superior al MAE, señala que el modelo maneja adecuadamente los errores más grandes, aunque estos contribuyen a aumentar el promedio de la desviación. En conjunto, las métricas muestran que el modelo ARIMA es efectivo para capturar la dinámica de la serie temporal del aceite de soya,

ofreciendo predicciones con un grado razonable de precisión.

4.4.1.2. Prophet

A continuación, en la Figura 11 se presenta el pronóstico de la serie del precio de futuros de soya empleando el modelo desarrollado con Prophet.

Figura 11. Pronóstico de la serie del precio de futuros de soya empleando Prophet.



Fuente: Elaboración propia.

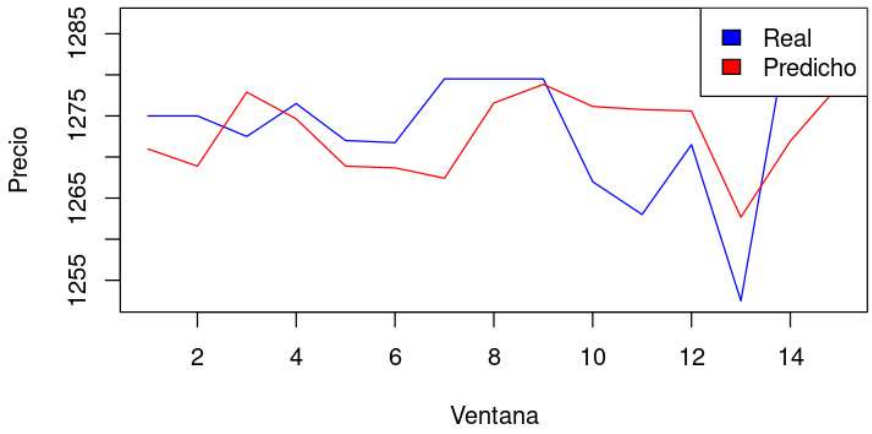
La gráfica compara los precios reales y predichos del aceite de soya usando Prophet de forma univariada, donde la línea roja de la predicción muestra que los pronósticos generados con el modelo Prophet están por encima del precio real del aceite de soya en la ventana de tiempo evaluada entre el 30 de septiembre al 14 de octubre del 2023.

4.4.1.2. Elman

Por otro lado, se emplean las redes neuronales Elman, esta es una red neuronal recurrente, lo que significa que tiene conexiones que retroceden en el tiempo, por esta razón es un algoritmo muy interesante al momento de estudiar el comportamiento de los pronósticos para los del precio de futuros de soya.

A continuación, en la Figura 12 se presenta el Pronóstico de la serie del precio de futuros de soya empleando redes neuronales Elman.

Figura 12. Pronóstico de la serie del precio de futuros de soya empleando redes neuronales Elman.



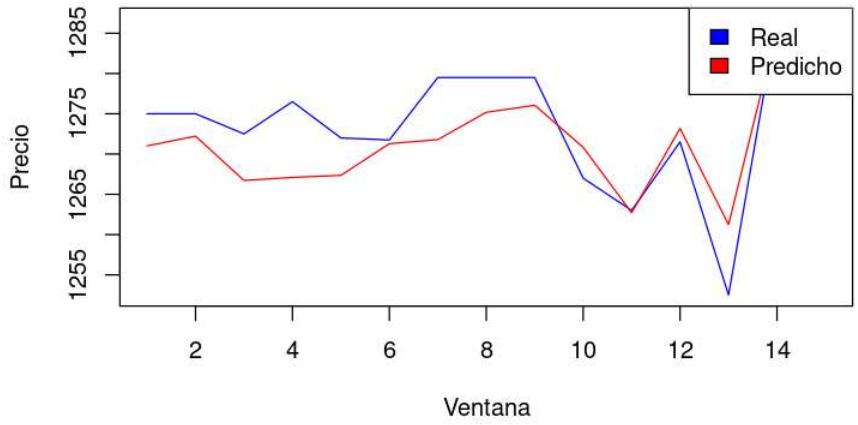
Fuente: Elaboración propia.

La gráfica muestra una comparación entre los valores reales y las predicciones realizadas con el modelo Elman univariado para el precio del aceite de soya. La línea azul muestra los valores reales del precio, y la línea roja representa los valores predichos por el modelo Elman. Las líneas se mueven juntas en varios puntos, lo que indica momentos en los que las predicciones del modelo están cercanas a los valores reales. Sin embargo, hay áreas donde las líneas divergen, lo que indica diferencias entre las predicciones y los valores reales. La cercanía de las dos líneas sugiere que el modelo tiene una cierta capacidad para seguir la tendencia de los precios reales del aceite de soya; aunque hay fluctuaciones y puntos donde las predicciones no coinciden con los valores reales, lo cual es normal en cualquier modelo de predicción.

4.4.1.3. Jordan

A continuación, en la Figura 13 se presenta el Pronóstico de la serie del precio de futuros de soya empleando redes neuronales Jordan.

Figura 13. Pronóstico de la serie del precio de futuros de soya empleando redes neuronales Jordan.



Fuente: Elaboración propia.

La gráfica representa la actuación de un modelo Jordan univariado en la predicción de precios del aceite de soya, mostrando una alineación general entre los valores reales (línea azul) y las predicciones (línea roja) a lo largo del eje del índice, que denota puntos de tiempo sucesivos. A pesar de la tendencia general bien capturada por el modelo, se identifican desviaciones notables, en particular cerca del índice 10, donde las predicciones se apartan temporalmente de los valores reales.

4.4.1.4. Comparación y elección del modelo univariado óptimo para precios de soya

A continuación, en la Tabla 1 se presenta resultados de los modelos según las métricas de evaluación.

Tabla 1. Resultados de los modelos univariados según las métricas de evaluación MSE, MAE y RMSE para serie de tiempo del precio de futuros de soya

Modelos	Métricas		
	MSE	MAE	RMSE
ARIMA	64.3602	5.78735	8.022481
Prophet	950.6263	29.579	30.83223
Elman	56.56388	6.096727	7.52089
Jordan	23.49676	3.897841	4.847345

Fuente: Elaboración propia.

Entre los cuatro modelos univariados analizados para predecir la serie temporal del precio del aceite de soya, el modelo Jordan demuestra ser el más efectivo, evidenciado por sus métricas de rendimiento más bajas (MSE de 23.49676, MAE de 3.897841, y RMSE de 4.847345), lo que indica la mayor precisión y el menor error en sus predicciones. Aunque el modelo Elman también muestra un buen rendimiento, no supera al Jordan, mientras que el modelo Prophet registra los valores más altos en todas las métricas, sugiriendo ser el menos adecuado para este conjunto de datos en específico. El modelo ARIMA, por su parte, presenta un rendimiento intermedio, superando a Prophet pero sin alcanzar la precisión de los modelos basados en redes neuronales. La superioridad del modelo Jordan se atribuye a su capacidad para capturar de manera efectiva las dependencias temporales complejas y no lineales en los datos del aceite de soya, lo que lo convierte en la opción más recomendable para predecir esta serie temporal en particular.

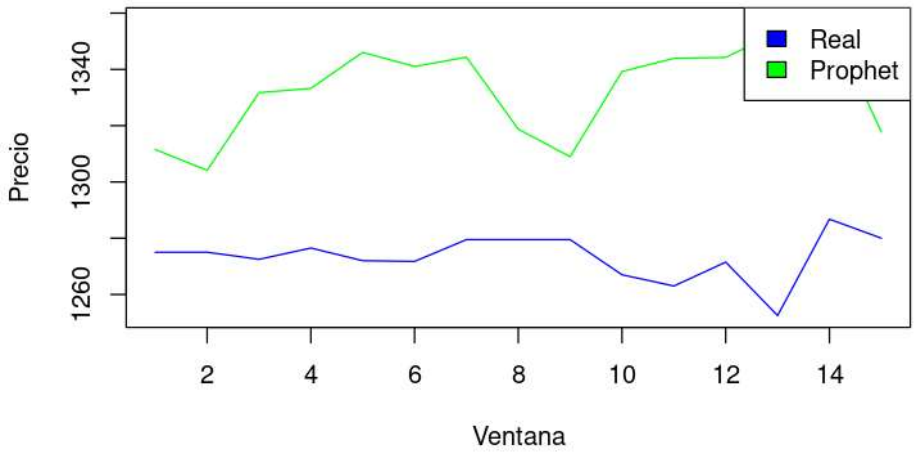
4.4.2 Modelación precio futuro del aceite de soya con variables exógenas

Para analizar la influencia de diversas variables externas en la serie temporal del precio del aceite de soya, se desarrollan una serie de algoritmos entrenados y evaluados en el mismo rango de tiempo que se empleó para los modelos univariados. Sin embargo, para la modelación multivariada se construye un modelo Prophet, unas redes neuronales convolucionales y redes neuronales recurrentes LSTM. Donde en el entrenamiento de estos modelos se incorporaron variables externas que son significativas con la variable objetivo, como el precio del petróleo crudo, la cotización del dólar, el precio del maíz, el índice de precios al consumidor (IPC) de Argentina, y factores contextuales como la situación política en el país productor y la situación social derivada del conflicto Rusia-Ucrania. Se decide excluir la variable económica relacionada con el IPC de China, a pesar de ser un importante país consumidor, tras identificar que la correlación de esta con el precio del dólar es negativamente débil, sugiriendo una influencia limitada en el contexto del modelo para predecir el precio del aceite de soya. Esta selección estratégica de variables busco optimizar la capacidad predictiva del modelo, centrándose en aquellos factores que tienen una relación más directa o significativa con los precios del aceite de soya dentro del periodo establecido.

4.4.2.1 Prophet

El modelo Prophet, ampliado con variables externas, se empleó para prever la serie temporal del precio de futuros de soya, enfocándose en el impacto de factores adicionales que hay en el mercado de futuros. La Figura 14 muestra este pronóstico detallado.

Figura 14. Pronóstico de la serie del precio de futuros de soya empleando Prophet con variables externas.



Fuente: Elaboración propia.

La gráfica ilustra una comparativa entre los datos reales y las predicciones hechas por el modelo Prophet multivariado para el precio del aceite de soya, donde la línea verde representa las predicciones y la azul los valores reales. Observamos cierta correspondencia en las tendencias capturadas por el modelo, pero también notables discrepancias en los valores de los precios, ya que los pronósticos están por encima de los valores reales.

4.4.2.2 Redes Convolucionales

Este proceso comienza con la definición de una función esencial para crear ventanas temporales, lo que permite estructurar secuencias de observaciones destinadas al entrenamiento del modelo. Se procede luego a la configuración de estas ventanas de datos para el entrenamiento, enfocándose en una ventana temporal específica de 15 días para la predicción.

Posteriormente, se integran variables exógenas en la matriz de datos, enriqueciendo así el contexto analítico y mejorando la capacidad predictiva del modelo. Este conjunto de datos se transforma después a un formato óptimo para su procesamiento por el modelo de Redes Convolucionales (CNN).

Una vez definido el modelo CNN, se procede a su entrenamiento exhaustivo, ajustando sus parámetros para adaptarse de manera precisa a las tendencias y patrones intrínsecos de los datos.

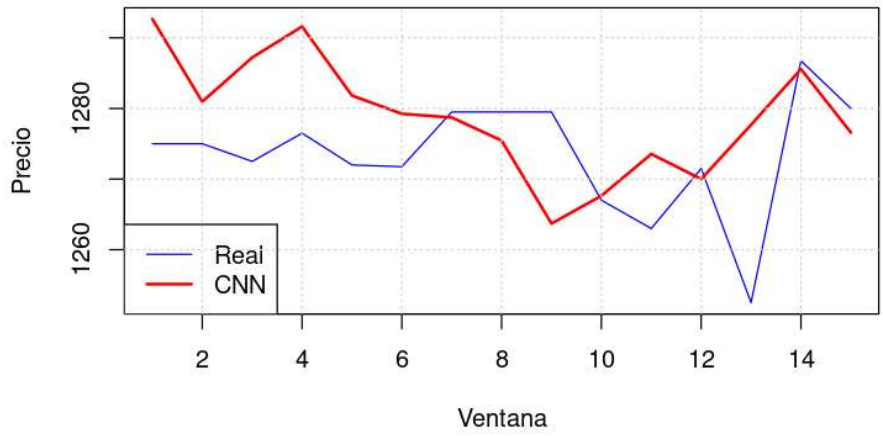
En el entrenamiento del modelo de red convolucional (CNN) para predecir la serie temporal del precio del aceite de soya, se ajustaron varios parámetros críticos para optimizar su rendimiento.

Se estableció una capa convolucional inicial con 64 filtros y un tamaño de kernel de 2, seguida de

una tasa de dropout del 20% para mitigar el sobreajuste y una capa de max pooling con pool size de 2 para reducir la dimensionalidad y destacar las características esenciales. El modelo se aplanó y se conectó a una capa densa de 50 unidades con activación "relu", seguida de otra capa de dropout y una capa densa final con activación lineal para la salida. Se compiló utilizando el optimizador Adam con una tasa de aprendizaje de 0.001 y el RMSE como función de pérdida, entrenándose a lo largo de 110 épocas con un tamaño de batch de 32. Estos ajustes fueron diseñados cuidadosamente para mejorar la capacidad del modelo de capturar las dinámicas subyacentes en los datos del aceite de soya, permitiendo predicciones detalladas y fiables que son fundamentales para la toma de decisiones en el mercado.

Finalmente, se utilizan estas redes entrenadas para realizar predicciones detalladas, proporcionando así insights valiosos y pronósticos fiables para la toma de decisiones en el mercado del aceite de soya. En la Figura 15 se puede observar el comportamiento de las predicciones con el modelo CNN.

Figura 15. Pronóstico de la serie del precio de futuros de soya empleando Redes CNN con variables externas.

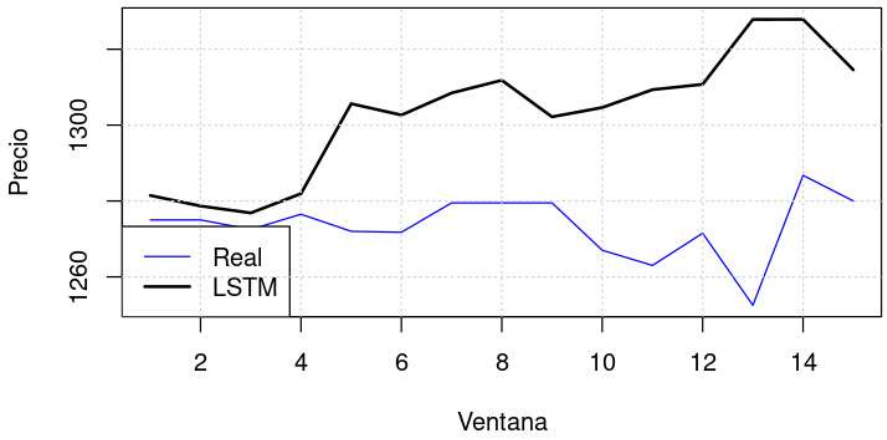


La gráfica muestra la actuación de un modelo de redes convolucionales (CNN) multivariado en la tarea de predecir el precio del aceite de soya. Las líneas azul y roja representan los valores reales y las predicciones del modelo, respectivamente. La superposición de ambas líneas en varios tramos sugiere que el modelo es capaz de seguir las tendencias de los precios a lo largo del tiempo con una precisión razonable. Sin embargo, se observan fluctuaciones y períodos donde las predicciones se desvían de los precios reales, lo cual es un comportamiento típico en modelos de predicción, y más aún en mercados volátiles como los de commodities.

4.4.2.3 Redes RNN-LSTM

En este proyecto, se implementó un modelo de redes neuronales recurrentes LSTM, para predecir la serie temporal del precio del aceite de soya, iniciando con la adecuación de los datos a una estructura de serie temporal compatible con análisis secuenciales. Para asegurar la integridad del conjunto de datos frente a posibles valores faltantes en variables críticas, se aplicó una técnica de imputación basada en el promedio. El modelo se diseñó con capas LSTM que contienen 64 unidades para capturar las dependencias temporales, integrando además capas de dropout con el objetivo de reducir el sobreajuste y capas densas para la predicción final. La compilación del modelo se efectuó con un optimizador Adam y una tasa de aprendizaje de 0.001, utilizando el RMSE como función de pérdida. El entrenamiento se realizó durante 200 épocas con un tamaño de batch de 32, buscando afinar la capacidad de aprendizaje del modelo sobre la complejidad de los datos. En la Figura 16 se puede observar el comportamiento de las predicciones con el modelo de redes neuronales recurrentes.

Figura 16. Pronóstico de la serie del precio de futuros de soya empleando una red neuronal RNN-LSTM con variables externas.



Fuente: Elaboración propia.

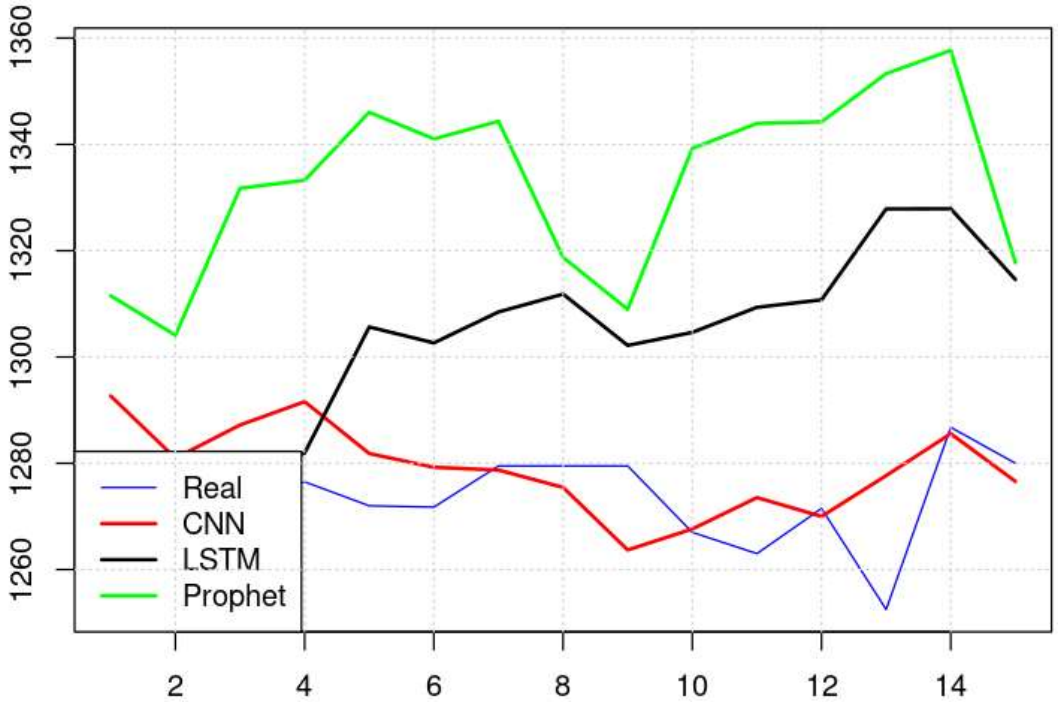
La gráfica muestra la comparación entre los valores reales (en azul) y las predicciones (en negro) de un modelo de redes neuronales recurrentes LSTM para el precio del aceite de soya. La línea de predicción sigue de cerca a la línea de valores reales, lo que indica que el modelo LSTM ha aprendido patrones subyacentes en los datos y puede predecir la tendencia del precio con cierta precisión. Aunque hay puntos donde las líneas se separan, indicando errores en la predicción, en general, la proximidad de las dos líneas sugiere un modelo adecuado para pronósticos en periodos de tiempo corto. Las áreas donde las predicciones no coinciden exactamente con los valores reales podrían ser momentos de volatilidad en el mercado o puntos donde el modelo necesita ajustes adicionales, ya que conforme aumenta la ventana, las predicciones y los valores reales divergen

mucho más.

4.4.2.4 Comparación y elección del modelo multivariado óptimo para precios de soya

Para evaluar la precisión de los modelos Prophet, Convolucional y RNN-LSTM en la predicción de precios del aceite de soya, empleamos tres métricas estadísticas esenciales que son el Medio MSE, MAE y el RMSE. Estas métricas brindan una evaluación detallada comparando las predicciones de cada modelo con los valores reales de la serie temporal. En la Figura 17 se muestra la comparación de los pronósticos de los diferentes modelos multivariados entrenados.

Figura 17. Pronóstico de la serie del precio de futuros de soya empleando Propeth, CNN, LSTM



Fuente: Elaboración propia.

Al aplicar el MSE, MAE y RMSE, pudimos realizar una comparación objetiva entre los modelos Prophet, Convolucional y RNN-LSTM, determinando cuál se alinea más estrechamente con los valores reales observados en la serie de tiempo del precio del aceite de soya. Esta comparación facilita la identificación del modelo más preciso y confiable para predecir los precios futuros en este mercado.

A continuación, en la Tabla 2, se establecen las métricas de error de los 15 días de la ventana de evaluación.

Tabla 2. Resultados de modelos con variables externas según las métricas de evaluación MSE, MAE y RMSE para serie de tiempo del precio de futuros de soya

Modelos	Métricas		
	MSE	MAE	RMSE
Prophet	3965.884	59.5942	62.97526
Convolucionales	131.4521	8.921419	11.46526
RNN-LSTM	1217.395	29.51497	34.89119

Fuente: Elaboración propia.

Basándonos en las métricas de Mean Squared Error (MSE), Mean Absolute Error (MAE), y Root Mean Squared Error (RMSE) de los tres modelos multivariados (CNN, LSTM, y Prophet), el modelo de red convolucional (CNN) presenta el mejor rendimiento para la predicción de la serie temporal del precio del aceite de soya. Esto se evidencia por tener los valores más bajos en todas las métricas evaluadas: un MSE de 131.4521, un MAE de 8.921419, y un RMSE de 11.46526. Estos resultados sugieren que el modelo CNN ha sido capaz de capturar de manera más efectiva las complejidades, patrones y variaciones en los datos multivariados, resultando en predicciones más precisas y con menor error en comparación con los modelos LSTM y Prophet, donde se observan errores significativamente mayores. El modelo LSTM muestra un rendimiento mejor que el modelo Prophet, pero ambos son superados por el modelo CNN, lo que indica la eficacia de las redes convolucionales en el manejo de este tipo específico de datos de series temporales.

4.4.3 Comparación modelos univariados y multivariados

Al comparar las métricas de rendimiento entre el modelo Jordan (Univariado) y las redes convolucionales CNN (Multivariado) para predecir la serie temporal del precio del aceite de soya, el modelo Jordan, siendo univariado, muestra una superioridad en precisión, reflejada en menores valores de los errores. Esto sugiere que Jordan ha capturado más efectivamente las tendencias y patrones en los datos de la serie de tiempo del aceite de soya. Sin embargo, el valor del enfoque multivariado, como el utilizado en el modelo CNN, reside en su capacidad para integrar múltiples variables y capturar la complejidad de los factores que influyen en los precios, ofreciendo una perspectiva más amplia y detallada que puede ser crucial para comprender y predecir dinámicas de mercado complejas. Aunque en este caso el modelo univariado Jordan supera en precisión, el uso de modelos multivariados como el CNN subraya la importancia de considerar múltiples factores y la interacción entre ellos en análisis predictivos más completos y contextualizados.

5. HERRAMIENTA WEB PARA SEGUIMIENTO Y ANÁLISIS EN TIEMPO REAL DEL MERCADO DE SOYA

En este contexto, uno de los objetivos es poder visualizar los resultados obtenidos mediante una aplicación web, el cual ayude a los decisores en el mercado de futuros a actuar de manera eficiente y hacer seguimiento al comportamiento de los precios en tiempo real, ya que para cerrar los contratos de futuros se deben tener en cuenta diferentes factores que estamos abordando en los pronósticos generados por el modelo. Dada esta necesidad, se desarrolló una interfaz gráfica usando la herramienta de visualización Power BI. El tablero de control permitió disponer los pronósticos del precio de los futuros de soya, donde se genera un resumen sobre el comportamiento de los precios con respecto al día, semana y mes. Los elementos gráficos utilizados para construir y representar la información en dashboard son los siguientes:

Tarjetas

Contienen el precio de los futuros de soya semanal, el precio promedio del mes en curso y el pronóstico semanal, el cual varía dependiendo el día, también en las tarjetas se tiene la variación porcentual mensual del precio de los futuros de soya con respecto al mes anterior, para este elemento gráfico que muestra la variación porcentual, se emplean reglas de colores que representan situaciones positivas (verde) o negativas (rojo) según el comportamiento del precio cuando aumenta o disminuye.

Figura 18. Tarjetas de resumen del precio de los futuros de soya, variación porcentual y pronostico.

Fuente: Elaboración propia.

Gráficos

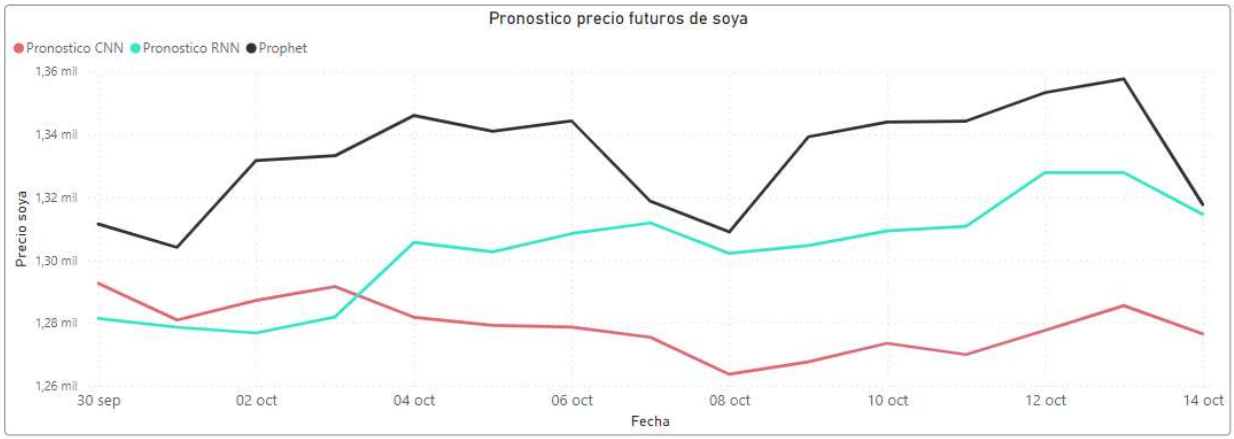
Para el gráfico de la Figura 19 se emplea un gráfico de líneas que discriminara la serie histórica y el pronóstico para los siguientes 15 días. La serie histórica se representa con el color azul los pronósticos de los modelos Prophet (gris), Redes neuronales convolucionales (rojo) y Redes neuronales recurrentes (verde).

Figura 19. Gráfico de líneas para representar los precios históricos de los futuros de soya y sus pronósticos.

Fuente: Elaboración propia.

En la Figura 20 también se emplea un gráfico, al igual que en la Figura 19 se elige un gráfico de líneas, que ayuda a representar el valor del pronóstico para los siguiente 15 días, donde cada línea representa el valor del precio pronosticado en el transcurso de los días, de esta forma, la proyección del modelo Prophet se visualiza con el color (gris), Redes neuronales convolucionales (rojo) y Redes neuronales recurrentes (verde).

Figura 20. Gráfico de líneas para representar los pronósticos del precio de futuros de soya en un horizonte de 15 días.



Fuente: Elaboración propia.

En la Figura 21, se presenta el Dashboard completo para futuros de soya el cual permite el análisis

del precio de este bien agrícola para tomar decisiones en el mercado de futuros.

Figura 21. Dashboard completo para la serie de precios de futuros de soya.

Fuente: Elaboración propia.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

- La comparativa entre modelos univariados y multivariados para predecir el precio del aceite de soya muestra que el modelo univariado Jordan supera en rendimiento a los multivariados. Este resultado resalta que la adición de variables externas no garantiza automáticamente mejoras en el rendimiento de los modelos en cuanto a la precisión de los pronósticos.

- El análisis realizado a través de la comparación de los modelos ARIMA, Prophet, Redes neuronales Elman y redes neuronales Jordan permite tomar decisiones informadas sobre qué modelo de pronóstico e identificar qué variables adicionales son más efectivas para predecir el precio del aceite de soya.

- El análisis detallado de los modelos predictivos para el precio del aceite de soya ha evidenciado que el modelo Convolucional sobresale significativamente. Su superioridad se refleja en su precisión tanto a corto como a largo plazo, con valores notablemente bajos en métricas clave como MSE, MAE y RMSE. Este rendimiento lo convierte en la opción más robusta y fiable, destacando su eficacia y precisión en la modelación del precio del aceite de soya en los modelos multivariados usados.

- El estudio abre caminos para investigaciones futuras, como la incorporación de datos adicionales, el uso de técnicas de aprendizaje profundo más avanzadas o la exploración de enfoques de modelado alternativos.

- Las metodologías y lecciones aprendidas podrían aplicarse a otros mercados y productos, ampliando el alcance y la utilidad de este enfoque basado en machine learning.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] A. Hecht, “Precios de los granos: Podrían avecinarse subidas explosivas,” *Investing.com*. https://es.investing.com/analysis/precios-de-los-granos-podrian-avecinarse-subidas-explosivas-200450265.

[2] J. E. Hanke and D. W. Wichern, *PRONÓSTICOS EN LOS NEGOCIOS*, 9th ed. México: PEARSON EDUCACIÓN, 2010.

[3] J. Garcia, J. M. Molina, and A. Berlanga, *Ciencia De Datos. Técnicas Analíticas Y Aprendizaje Estadístico*, Un enfoque. Alfaomega, 2018.

[4] L. A. O. Quintero, “Las redes neuronales artificiales como herramienta para la predicción.” IBM, “Machine Learning,” *IBM - Deutschland*. https://www.ibm.com/mx-es/analytics/machine-learning.

[5] G. Miner et al., Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Elsevier, 2012. [Online]. Available: https://www.sciencedirect.com/book/9780123869791/practical-text-mining-and-statistical-analysis-for-non-structured-text-data-applications.

[6] R. S. Pontoh, S. Zahroh, H. R. Nurahman, R. I. Aprillion, A. Ramdani, and D. I. Akmal, “Applied of feed-forward neural network and facebook prophet model for train passengers forecasting,” *J. Phys. Conf. Ser.*, vol. 1776, no. 1, pp. 0–9, 2021, doi: 10.1088/1742-6596/1776/1/012057.

[7] R. H. Shumway and D. S. Stoffer, *Time Series: A Data Analysis Approach Using R*. 2019.

[8] P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, 3rd ed. Springer, 2016. ISBN: 978-3319298528

[9] J. L. Elman, "Finding Structure in Time," in Cognitive Science, vol. 14, no. 2, pp. 179-211, 1990.

[10] M. I. Jordan, “Serial order: a parallel distributed processing approach. (Tech. Rep. No. 8604). San Diego: University of California, Institute for Cognitive Science.,” no. 667, 1986.

[11] CME Group Inc, “Guía de Futuros para los Operadores,” pp. 20–21, 2011, [Online]. Available: http://www.cmegroup.com/trading/files/traders-guide-to-futures-spn.pdf.

[12] Stonex, “Mercado mundial de aceites vegetales,” 2020.

[13] M. Cecilia García, A. M. Jalal, L. A. Garzón, and J. M. López, “Métodos para predecir índices Bursátiles,” *Ecos Econ.*, vol. 17, no. 37, pp. 51–82, 2013, doi: 10.17230/ecos.2013.37.3.

[14] C. Rodriguez, “Modelos no lineales de pronóstico de series temporales basados en inteligencia computacional para soporte en la toma de decisiones agrícolas,” p. 164, 2016, [Online]. Available: https://rdu.unc.edu.ar/handle/11086/4604.

[15] S. J. Taylor y B. Letham, “Forecasting at Scale”, Amer. Statistician, vol. 72, n.º 1, pp. 37–45, enero de 2018. Accedido el 9 de diciembre de 2023. [En línea]. Disponible: https://doi.org/10.1080/00031305.2017.1380080

[16] Pérez Paredes, A., Cruz de los Ángeles, J. A., Guatemala Villalobos, A. M. D. J., & Juárez

Fonseca, V. (2018). Importancia de los pronósticos en la toma de decisiones en las MIPYMES. *Revista GEON (Gestión, Organizaciones y Negocios)*, *5*(1), 97–114. https://doi.org/10.22579/23463910.17

[17] I. Goodfellow, Y. Bengio, y A. Courville, Deep Learning. MIT Press, 2016.

[18] J. De Lucio, "Estimación adelantada del crecimiento regional mediante redes neuronales LSTM", *Investigaciones Regionales - J. Regional Res.*, vol. 49, pp. 45–64, mayo de 2021. Accedido el 9 de diciembre de 2023. [En línea]. Disponible: https://doi.org/10.38191/iirr-jorr.21.007

[19] G. E. P. Box, G. M. Jenkins, y G. C. Reinsel, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015.