Pontificia Universidad JAVERIANA Cali

**Facultad de Ingeniería y Ciencias**

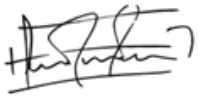
Santiago de Cali, 08 de Agosto de 2024

**Ingeniero:**
**Juan Carlos Martínez Arias**
**Director Posgrados de Ingeniería**
**Facultad de Ingeniería y Ciencias**
**Pontificia Universidad Javeriana - Cali**

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo 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 **DESARROLLO DE UN MODELO PREDICTIVO PARA EL PRECIO DE IMPORTACIÓN DE LOS PESTICIDAS EN COLOMBIA**, el cual será realizado por los estudiantes **Edisson Ramírez Méndez** y **Luis Felipe Ramírez Amariz** con código 8980378 y 8976101 perteneciente al énfasis en Industrial, bajo la dirección del profesor **Daniel Enrique González Gómez**.

El suscrito director del Trabajo 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,



Firma
**Edisson Miguel Ramírez Méndez**

C.C. 80150616 de Bogotá

Firma
**Daniel Enrique González Gómez**

C.C. 16.669.372 de Cali

Firma
**Luis Felipe Ramírez Amariz**

C.C. 1019076459 de Bogotá

**Acta de Correcciones al Documento de Trabajo de Grado**

**Santiago de Cali, 8 de agosto de 2024**

**Autor:** **Edisson Ramírez Méndez**
**Luis Felipe Ramírez Amariz**

**Título del Trabajo de Grado: "DESARROLLO DE UN MODELO PREDICTIVO PARA EL PRECIO DE IMPORTACIÓN DE LOS PESTICIDAS EN COLOMBIA"**

**Director: Daniel Enrique González Gómez**

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.

_____

**Daniel Enrique González Gómez**
Firma del director del Trabajo de Grado

# DESARROLLO DE UN MODELO PREDICTIVO PARA EL PRECIO DE IMPORTACIÓN DE PESTICIDAS EN COLOMBIA

*Edisson Ramírez Méndez 8980378*
*Luis Felipe Ramírez Amariz 8976101*

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

Director(a)
Daniel Enrique González Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 28 DE 2024

**Maestría en Ciencia de Datos**
**Facultad de Ingeniería y Ciencias**

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "**DESARROLLO DE UN MODELO PREDICTIVO PARA EL PRECIO DE IMPORTACIÓN LOS PESTICIDAS EN COLOMBIA**"

1. ÉNFASIS: Industrial
2. TIPO DE PROYECTO: Investigación
3. ÁREA DE TRABAJO: Competitividad Y Desarrollo
4. ESTUDIANTE (S):
   - 8980378, Edisson Ramírez Méndez
   - 8976101 Luis Felipe Ramírez Amariz
5. CORREO ELECTRÓNICO:
   - edissonramirez@javerianacali.edu.co,
   - luisferam92@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO:
   - carrera 102c # 56 f 19 sur Bogotá (3118467515)
   - Vda Verganzo conjunto Acacia los maderos Torre 3 apto 603 Tocancipá - Cundinamarca (3007658291)
7. DIRECTOR: Daniel Enrique González Gómez
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: dgonzalez@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica): NO APLICA
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica):NO APLICA
12. OTROS GRUPOS O EMPRESAS:
13. PALABRAS CLAVE (al menos 5): Predicción de precios, pesticidas, ciencia de datos, XGBoost, LightGBM, importaciones, modelo predictivo, análisis de datos, Colombia, agricultura.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Producción y Consumo Responsables
15. FECHA DE INICIO (Desarrollo del proyecto): 1/07/2023
16. RESUMEN (máximo 400 palabras).

La agricultura ha sido un pilar esencial para la supervivencia y el desarrollo de la humanidad, proporcionando alimentos básicos para la dieta mundial. Sin embargo, el sector agrícola en Colombia enfrenta desafíos significativos debido a su dependencia de insumos importados, lo que incrementa los costos operativos y afecta la competitividad en el mercado internacional. Este estudio aborda el problema del aumento de precios de insumos agrícolas como insecticidas, fungicidas y herbicidas

debido a las fluctuaciones en la tasa de cambio.

El objetivo principal de esta investigación es desarrollar un modelo predictivo para pronosticar el precio promedio de importación de estos insumos en Colombia utilizando técnicas de ciencia de datos. El proyecto incluye la consolidación, limpieza y análisis exploratorio de datos de importación proporcionados por la Dirección de Impuestos y Aduanas Nacionales de Colombia (DIAN). Se emplean diversos modelos predictivos, incluyendo ARIMA, SARIMA, Random Forest, XGBoost y LightGBM, para identificar el más adecuado.

El análisis revela la fluctuación de los precios y su impacto en la producción agrícola y la competitividad del sector. Se destaca la importancia de contar con herramientas de visualización que faciliten el acceso a la información histórica y los resultados del modelo para apoyar la toma de decisiones en el sector agrícola. Los resultados subrayan la necesidad de estrategias innovadoras que aprovechen la ciencia de datos para abordar los desafíos económicos y mejorar la sostenibilidad del sector agrícola en Colombia.

# TABLA DE CONTENIDO

Pág.

## LISTA DE ILUSTRACIONES

## LISTA DE TABLAS

## INTRODUCCIÓN

A lo largo de la historia, la agricultura ha sido el pilar fundamental para la supervivencia y el progreso de la humanidad. Esta actividad es responsable de la producción de una gran parte de los alimentos básicos que constituyen la base de la dieta mundial. Por otro lado, el panorama del comercio internacional está marcado por fuertes fluctuaciones debido a la creciente competencia entre las industrias. En el caso de Colombia, la dependencia de insumos importados para la producción agrícola genera un impacto significativo en los costos operativos y de producción de los bienes. Esta situación se traduce en una clara afectación en todos los sectores de la economía, siendo el sector agrícola uno de los más afectados por las fluctuaciones del mercado internacional.

Durante los últimos años, se ha registrado un aumento generalizado en el precio de bienes y servicios en Colombia, reflejado en el incremento del Índice de Precios al Consumidor (IPC), lo que impacta significativamente en el costo de vida y su calidad en el país. Esta investigación se centra en el aumento observado en los precios al consumidor de los productos agrícolas, lo que motiva el desarrollo de una herramienta que facilite la toma de decisiones, analizando el comportamiento y las tendencias del sector para contribuir a posicionar estos productos con precios más competitivos en el mercado. El proyecto propuesto consiste en la creación de un modelo predictivo para pronosticar el precio de ciertos insumos agrícolas importados en Colombia, como insecticidas, fungicidas y herbicidas, los cuales son vitales para la producción agrícola. Se utilizarán datos publicados por la Dirección de Impuestos y Aduanas Nacionales de Colombia (DIAN) sobre las partidas arancelarias relacionadas con los productos mencionados, con información disponible desde el año 2001, la cual es de dominio público y no presenta restricciones para su uso.

Este proyecto se estructura en varias secciones. En primer lugar, se presenta una introducción que aborda la problemática analizada. La segunda parte incluye la formulación del problema, donde se plantea la pregunta de investigación. La tercera parte describe los objetivos del proyecto, incluyendo el objetivo general y los objetivos específicos. En la cuarta parte, se aborda el marco de referencia, que proporciona el contexto teórico y antecedentes relevantes. La quinta parte expone la metodología a emplear en el proyecto, detallando cómo se alcanzarán los objetivos específicos. Posteriormente, se incluye un análisis de los datos, que se presenta en la sexta parte, seguido de la evaluación y selección del modelo predictivo en la séptima parte. En la octava parte, se presenta la implementación del modelo y la validación de los resultados. Finalmente, el documento concluye con las referencias bibliográficas.

# 1. DEFINICIÓN DEL PROBLEMA

En los últimos años, el sector agrícola ha experimentado un aumento constante en los precios al consumidor, atribuible a diversos factores. Entre ellos, destaca el incremento en el valor de la tasa de cambio y su impacto en la fluctuación de los precios de los insumos importados. Específicamente, el sector agrícola depende en gran medida de pesticidas importados para su producción, y el aumento en el precio de la tasa representativa del mercado (TRM) en los últimos años ha encarecido la adquisición de estos productos. Un ejemplo palpable de este impacto se evidencia en un informe de Fedepapa, que señala cómo el alza en el costo de todos los insumos importados ha resultado en un incremento significativo en el costo de producción por hectárea de papa. Por ejemplo, en noviembre de 2022, el costo de producción por hectárea alcanzó los $34 millones, en comparación con los $22 millones registrados en 2021 [1].

Este dato evidencia un incremento considerable en los costos operativos de los bienes agrícolas debido a la coyuntura económica nacional. Esto se traduce en una reducción en los márgenes de rentabilidad y los niveles de producción, así como una disminución en los indicadores de desarrollo de este sector. Además, se observa un notable aumento en los precios para los consumidores finales, lo que deja a la industria agrícola colombiana en clara desventaja frente a otros países al no poder ofrecer productos con precios competitivos. Esta situación podría impactar en indicadores como el desempleo, la calidad de vida del país, la pobreza y el desarrollo, entre otros aspectos.

## 1.1. PLANTEAMIENTO DEL PROBLEMA

Los precios de los insumos agrícolas aumentaron un 29,4% en 2022 [1]. Esta situación genera incertidumbre para los productores, inversionistas y comerciantes, dificultando la toma de decisiones informadas. Ante esta volatilidad, anticipar o proyectar cambios en los precios de los insumos se vuelve un desafío, lo que puede impactar directamente en los costos operativos y, en consecuencia, en los niveles de producción a corto y mediano plazo.

A pesar de contar con una amplia gama de datos sobre las importaciones y exportaciones en Colombia, provenientes de fuentes como el Departamento Administrativo Nacional de Estadística (DANE), el Banco de la República y el Ministerio de Comercio, Industria y Turismo, así como plataformas de datos abiertos como datos.gov.co, todavía no existe una herramienta que emplee técnicas de ciencias de datos para prever los precios, o al menos un precio promedio, de estos insumos. Esta carencia dificulta la toma de decisiones anticipada para todas las partes interesadas y afectadas en la situación mencionada.

Para abordar esta problemática, el proyecto se centra en prever el comportamiento de los precios de importación de insecticidas, fungicidas y herbicidas. En este documento, se empleará el término "pesticida" como referencia para englobar los tres tipos de productos mencionados.

## 1.2. FORMULACIÓN DEL PROBLEMA

¿Como pronosticar un precio promedio de los pesticidas seleccionados, a partir de las técnicas y herramientas que ofrece las ciencias de datos?

# 2. OBJETIVOS DEL PROYECTO

## 2.1 OBJETIVO GENERAL

Construir un modelo que permita pronosticar un precio promedio de importación para los 3 pesticidas seleccionados en este proyecto a partir de la implementación de técnicas en ciencias de datos. El criterio de selección de los tres pesticidas se debe a que son los productos que más son conocidos por el equipo investigador.

## 2.2 OBJETIVOS ESPECÍFICOS

2.2.1) Consolidar la información que se tiene al alcance a través de la extracción, limpieza, minería y el análisis exploratorio de los datos disponibles, con el propósito de tener un entendimiento profundo de los mismos y del problema analizado en la investigación.

2.2.2) Elaborar un modelo predictivo, al seleccionar uno de los modelos entrenados durante el desarrollo del proyecto, que permita pronosticar el precio de los pesticidas importados en Colombia.

2.2.3) Evaluar el alcance de predicción que tiene el modelo seleccionado, a través del análisis de los resultados obtenidos en el desarrollo del proyecto, para facilitar la toma de decisiones de los tomadores de decisiones.

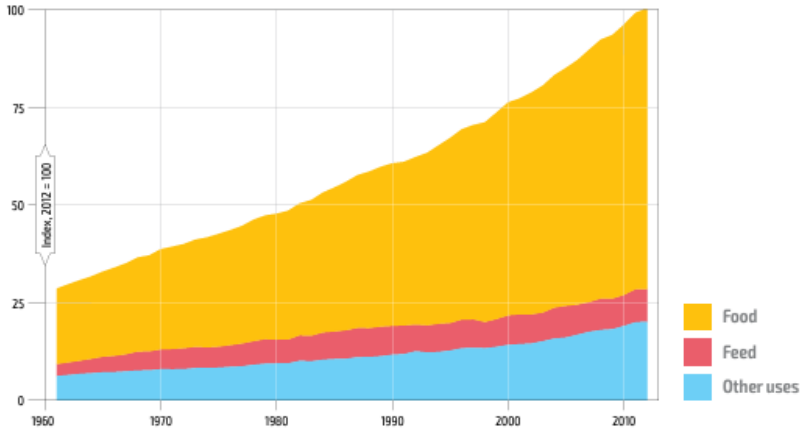
2.2.4) Emplear una herramienta de visualización que permita el acceso a la información histórica y la representación descriptiva y a los resultados obtenidos del modelo seleccionado.

# 3. MARCO DE REFERENCIA

## 3.1. LA IMPORTANCIA DE LA ACTIVIDAD AGRÍCOLA EN EL FUTURO CERCANO

Como ya se ha mencionado, a través del tiempo la agricultura se convertido en uno de los pilares fundamental que permiten la supervivencia y el progreso de la humanidad. Esta actividad es la fuente primaria de los alimentos básicos que conforman la dieta de la población mundial, destacándose como un elemento central en la planificación del desarrollo hacia la preservación de la seguridad alimentaria. Una porción significativa de la actividad agrícola está destinada a satisfacer la demanda de alimentos para el consumo humano. Esta actividad incluye tanto la utilización directa de productos agrícolas para la producción de alimentos, como el uso de cultivos y otros recursos vegetales en la elaboración de alimentos para animales, los cuales, a su vez, son empleados en la producción de alimentos para el consumo humano. (Ilustración 1) [2].

***Ilustración 1[1] Demanda agrícola alimentaria y no alimentaria: tendencias históricas***

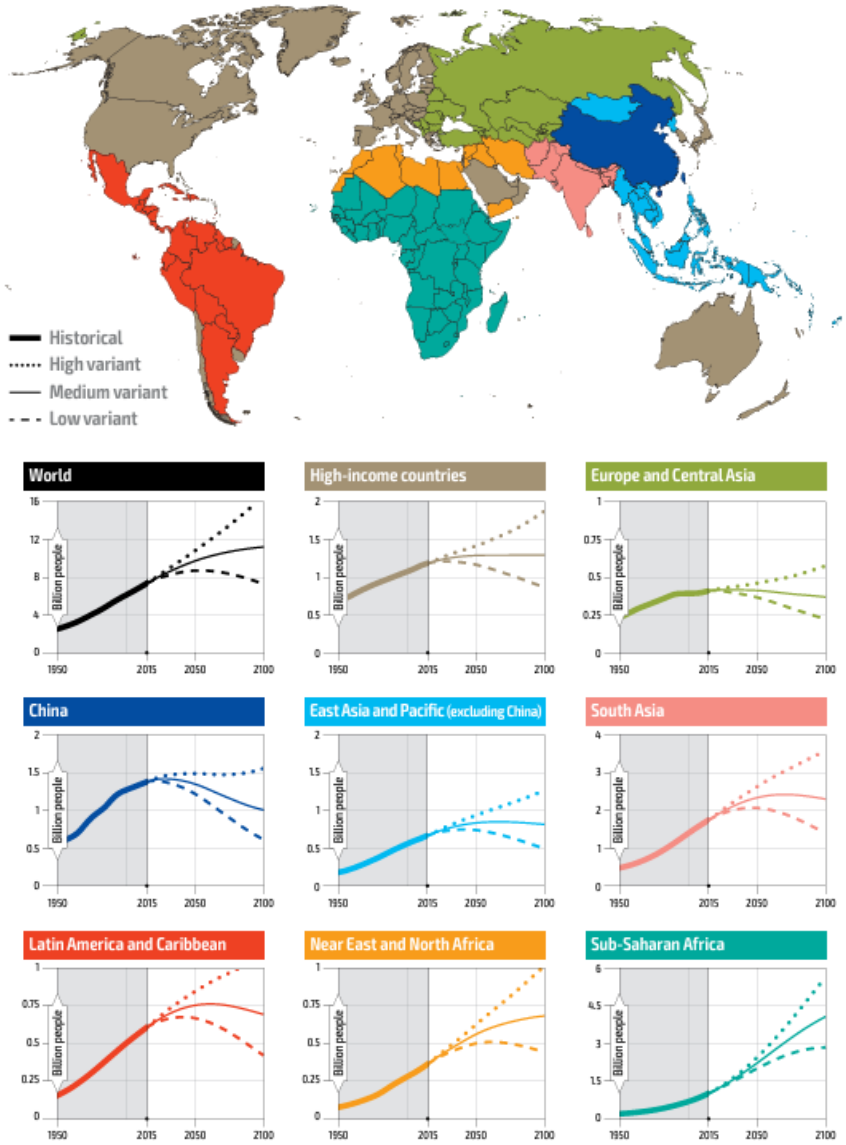



**Nota:** El índice 2012=100 se basa en el volumen de la demanda de alimentos expresado en términos monetarios a precios de 2012.
**Fuente:** Estudios de perspectiva global de la FAO, con base en FAOSTAT (varios años).

No obstante, el crecimiento demográfico en el último siglo ha generado un incremento considerable en la demanda de alimentos, planteando desafíos significativos que perdurarán en las próximas décadas. De acuerdo con las proyecciones de las Naciones Unidas, la población mundial aumentará notablemente en los próximos años, alcanzando los 9.700 millones en 2050, 10.800 millones en 2080 y 11.200 millones en 2100. Estas cifras representan incrementos del 32%, 47% y 53%, respectivamente, en comparación con los 7.300 millones de habitantes que había en el planeta en 2015. (Ilustración 2). [2]

[1] [1] La ilustración es tomada del documento original https://openknowledge.fao.org/3/I8429EN/i8429en.pdf

*Ilustración 2[2] Población mundial por región: histórica y proyectada, 1950-2100*




**Notas:** La agrupación de países se basa en los grupos de países del Banco Mundial de julio de 2016, descargados el 2 de agosto de 2016 de http://databank. worldbank.org/data/download/site-content/CLASS.xls como se especifica en el Anexo III, Tabla A 3.4 del informe completo. Los países de ingresos altos (PIA) se clasifican en un solo grupo, independientemente de su ubicación geográfica. Todos los demás países, calificados como países de ingresos bajos y medios (PIBM), se clasifican por región geográfica, en particular Europa y Asia central (ECA), Asia oriental y el Pacífico (EAP), Asia meridional (SAS), América Latina y el Caribe (ALC), Cercano Oriente y África del Norte (NNA) y África subsahariana (ASS). A menos que se especifique lo contrario, los PIBM y EAP incluyen a China (solo continental). En adelante, a los grupos de países y a China se les denominará generalmente "regiones".

**Fuente:** ONU, 2015.

[2] La ilustración es tomada del documento original https://openknowledge.fao.org/3/I8429EN/i8429en.pdf

Para enfrentar este desafío inminente, es fundamental priorizar el manejo de plagas y enfermedades que afectan a los cultivos. Las infecciones causadas por fitopatógenos son la principal causa de pérdidas en la producción agrícola, estimándose que entre el 20% y el 40% de las pérdidas totales se deben a enfermedades vegetales. Estas pérdidas no solo afectan directamente la disponibilidad de alimentos, sino que también generan enormes pérdidas económicas a nivel mundial, alcanzando aproximadamente 40 mil millones de dólares anuales. [3].

Estos datos iniciales ofrecen una visión del panorama agrícola y subrayan su papel crucial en la producción de alimentos. Además, en el contexto actual, es esencial que todas las disciplinas del conocimiento, incluidas las ciencias de datos, contribuyan al desarrollo sostenible de la seguridad alimentaria, que es vital para la supervivencia de la humanidad.

### 3.2.LA COMPETITIVIDAD EN EL MERCADO INTERNACIONAL

Para académicos y expertos en economía, el comercio internacional se basa en la teoría ricardiana, que sostiene que las ganancias dependen de las ventajas comparativas. Es importante no llegar al extremo de tener posiciones privilegiadas [4].

Existe una gran preocupación por la globalización económica, ya que está aumentando el poder de los mercados y afectando negativamente a los Estados. Es cada vez más necesario establecer nuevas reglas para el mercado mundial. La evolución de la sinergia al conflicto en las relaciones comerciales mundiales proviene de cambios profundos en el equilibrio de poder internacional. Las principales naciones se están adaptando a estos cambios. [5]

Es necesario no solo desarrollar nuevas tecnologías o herramientas de progreso, sino que estas también trasciendan al ámbito macroeconómico. La visión del mercado internacional es cada vez más caótica, dejando fuera a los productores que carecen de desarrollo tecnológico y musculo financiero. Es urgente crear nuevas herramientas desde las ciencias de datos para equilibrar el intercambio comercial y permitir una toma de decisiones adecuada en cada sector económico que lo necesite.

Particularmente para el sector agrícola nacional, hay un aumento en el valor de los insumos, lo que ha causado fuertes alzas en los costos operativos. Esto deja a estos productores en una seria desventaja frente a otros bienes de la misma naturaleza producidos en otros países. [1]

## 3.3. LA CIENCIA DE DATOS

En el entorno altamente competitivo actual, tanto las organizaciones como los gremios e instituciones necesitan estrategias innovadoras para adaptarse rápidamente a los cambios y aprovechar al máximo la abundante información disponible. El uso efectivo de los datos se ha vuelto esencial para la toma de decisiones relevantes y la gestión eficiente de recursos valiosos.

La Ciencia de Datos (DS) se puede definir como la aplicación de métodos cualitativos y cuantitativos para predecir resultados y resolver problemas significativos en respuesta al crecimiento exponencial de datos. Hoy en día, existen numerosos algoritmos para el procesamiento y análisis de datos, y la DS proporciona un marco y principios que permiten abordar de manera sistemática la extracción de conocimiento útil a partir de estos datos. [5]

Desde la perspectiva de la Ciencia de Datos (DS), se pueden comparar metodologías ágiles para aplicarlas en proyectos de diferentes procesos, seleccionando las actividades más relevantes para planificar cómo abordar el problema y organizar las tareas del caso de estudio. En resumen, la DS se puede definir como la aplicación de métodos cualitativos y cuantitativos para predecir resultados y resolver problemas significativos. Debido a la creciente e inmensa cantidad de datos disponibles actualmente, el conocimiento y el análisis del dominio no pueden estar separados. [6].

Existen numerosos algoritmos para el procesamiento y extracción de datos, así como una gran cantidad de detalles sobre los métodos aplicados en este campo. Sin embargo, la Ciencia de Datos (DS) va mucho más allá de los algoritmos de Data Mining. DS proporciona a los profesionales una estructura y un conjunto de principios que ofrecen un marco para abordar sistemáticamente los problemas de extracción de conocimiento útil a partir de los datos. Los métodos y la metodología para manejar los datos y llevar a cabo este tipo de proyectos son fundamentales. [7].

Para los fines de este proyecto, se propone utilizar las herramientas más adecuadas de la Ciencia de Datos (DS) para encontrar un modelo que estime el precio de los pesticidas importados en Colombia. Este modelo facilitará la toma de decisiones en el sector agrícola y tendrá un impacto en el mercado internacional. A continuación, se enumerarán algunos de los modelos que podrían ser útiles para esta investigación.

### 3.3.1. APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

El aprendizaje automático es una disciplina que permite a las máquinas adquirir habilidades de autoaprendizaje. Este proceso automatizado involucra la extracción e identificación de patrones a partir de datos para construir modelos que permitan hacer predicciones mediante algoritmos supervisados. Esto incluye el análisis de características descriptivas actuales en comparación con características previamente asignadas a un conjunto de instancias. [8].

También se puede catalogar como un subcampo de la inteligencia artificial que se enfoca en desarrollar algoritmos que permitan a las máquinas aprender y mejorar automáticamente a partir de datos. Los modelos de aprendizaje automático se emplean para identificar patrones complejos en

los datos y realizar predicciones precisas.

### 3.3.2. APRENDIZAJE SUPERVISADO

Este concepto se refiere a una función derivada de una serie de ejemplos etiquetados, que luego se utiliza para predecir una salida para un conjunto diferente de ejemplos no etiquetados. De esta manera, el algoritmo aprende a clasificar las entradas al compararlas con el modelo ya entrenado y sus etiquetas [8]. Algunos métodos y algoritmos de aprendizaje supervisado que pueden aplicarse para resolver el problema planteado en este proyecto incluyen regresiones logísticas y lineales, así como el modelo de Support Vector Machine. Estos serán descritos a continuación. [9]

También se puede definir como un tipo de aprendizaje automático en el que se entrena un modelo utilizando un conjunto de datos etiquetados. Esto significa que cada entrada de datos tiene una etiqueta de salida conocida, lo que permite al modelo aprender la relación entre entradas y salidas para predecir las salidas de nuevos datos no etiquetados.

### 3.3.3. MODELOS DE PREDICCIÓN UTILIZADOS

- **ARIMA y SARIMA:** Modelos de series de tiempo que capturan tendencias y patrones estacionales en datos secuenciales. ARIMA es útil para datos no estacionarios, mientras que SARIMA incorpora componentes estacionales.

- **PROPHET:** Este modelo es especialmente útil para datos que presentan tendencias no lineales con estacionalidad anual, semanal y diaria, y para los cuales es importante manejar los efectos de vacaciones.

- **RANDOM FOREST:** Un modelo de aprendizaje supervisado basado en la construcción de múltiples árboles de decisión. Random Forest es robusto frente al sobreajuste y es efectivo para manejar grandes volúmenes de datos.

- **XGBoost:** Un algoritmo de boosting de árboles de decisión conocido por su eficiencia y precisión. Utiliza regularización L1 y L2 para prevenir el sobreajuste.

- **LightGBM:** Un modelo de boosting que emplea un crecimiento de árbol basado en hojas en lugar de niveles, lo que lo hace más rápido y eficiente en comparación con otros métodos de boosting.

### 3.3.4. METODOLOGÍA DE MODELADO Y EVALUACIÓN

- **Validación Cruzada:** La validación cruzada es una técnica de evaluación de modelos que divide el conjunto de datos en múltiples subconjuntos de entrenamiento y validación. Este enfoque ayuda a asegurar que el modelo generalice bien a nuevos datos y no esté sobre ajustado a un solo conjunto de datos de entrenamiento.

- **Datos de Entrenamiento:** Conjunto de datos utilizado para ajustar el modelo y aprender los patrones subyacentes en los datos.

- **Datos de Testeo**: Conjunto de datos separado utilizado para evaluar el rendimiento del modelo después del entrenamiento. Permite una evaluación imparcial del modelo sobre datos no vistos.

- **Hiperparámetros:** son parámetros de configuración de los modelos de aprendizaje automático que se establecen antes del entrenamiento. Incluyen aspectos como el número de árboles en un Random Forest o la tasa de aprendizaje en XGBoost. La correcta selección y ajuste de hiperparámetros es crucial para optimizar el rendimiento del modelo.

A continuación, serán abordados más en detalle cada uno de los modelos y conceptos mencionados anteriormente en cada definición de los modelos.

### 3.3.5. MODELOS DE SERIES DE TIEMPO ARIMA Y SARIMA

El modelo ARIMA es un enfoque popular para modelar series de tiempo que no son estacionarias. Combina tres componentes clave, AR (AutoRegressive), que implica que por parte del modelo se utiliza la dependencia lineal entre una observación y un número de retrasos anteriores. Se denota como $p$, que es el número de términos autorregresivos. Por otra parte, I (Integrated), que indica la diferencia necesaria para hacer que una serie no estacionaria se vuelva estacionaria. Se denota como d, que es el número de diferencias necesarias para hacer estacionaria la serie. Y finalmente, MA (Moving Average), que emplea la dependencia entre una observación y un término de error de un número de retrasos anteriores. Se denota como q, que es el número de términos de media móvil [10].

El modelo ARIMA se denota como ARIMA ($p$,$d$,$q$).

Por otra parte, El modelo SARIMA extiende el ARIMA al incluir componentes estacionales. Además de los componentes $p$,$d$,$q$. Añadiendo el componente S (Seasonal), el cual modela patrones estacionales en la serie temporal. Se denota como $P$,$D$,$Q$, que son los términos de auto regresión, diferenciación, y medias móviles estacionales, respectivamente, y $s$ que es el período de la estacionalidad [10].

El modelo SARIMA se denota como SARIMA$(p,d,q)s$

### 3.3.5.1.SUPUESTOS DEL MODELO Y VERIFICACIÓN

- **Estacionariedad:** Una serie temporal es estacionaria si sus propiedades estadísticas, como la media, la varianza y la autocorrelación, son constantes a lo largo del tiempo. Esto significa que las características de la serie no dependen del tiempo en que se observan. Las series no estacionarias pueden presentar tendencias, estacionalidades o cambios de varianza que complican el análisis predictivo [11].

  Los modelos ARIMA y SARIMA requieren que la serie sea estacionaria para que las propiedades estadísticas de la serie no cambien con el tiempo, asegurando así que las relaciones temporales capturadas por el modelo sean consistentes. Una serie no estacionaria puede producir estimaciones poco fiables y predicciones inexactas, ya que los patrones pasados pueden no ser indicativos del comportamiento futuro [12].

  Uno de los métodos más comunes para transformar una serie no estacionaria en estacionaria es la diferenciación, donde se calcula la diferencia entre valores consecutivos. La diferenciación de primer orden se representa como $\mathrm{Yt} - \mathrm{Y_{t-1}}$. A menudo se aplican transformaciones logarítmicas y de potencia para estabilizar la varianza. Y también se utilizan pruebas estadísticas como la prueba de Dickey-Fuller aumentada (ADF) y la prueba KPSS para determinar si una serie es estacionaria o no.

- **Autocorrelación:** La autocorrelación es la correlación de una serie temporal con un retraso de sí misma. Refleja la dependencia entre las observaciones en diferentes puntos en el tiempo. Los modelos ARIMA utilizan autocorrelaciones para entender la estructura subyacente de las dependencias temporales. Este supuesto es importante porque permite capturar la estructura de autocorrelación para los modelos de series de tiempo, ya que permite identificar patrones cíclicos o estacionales en los datos. Términos AR (AutoRegressive) y MA (Moving Average) dentro del modelo ARIMA aprovechan la autocorrelación para mejorar las predicciones. [13]

  La Función de Autocorrelación (ACF), mide la autocorrelación en diferentes lags y ayuda a identificar el número de términos MA en un modelo. Y la función de Autocorrelación Parcial (PACF) permite dimensionar la correlación parcial de la serie con sus lags, eliminando el efecto de los lags intermedios, y ayuda a identificar el número de términos AR.

- **Independencia de Residuos:** La independencia de residuos significa que los residuos (errores) del modelo, que son las diferencias entre los valores observados y los valores ajustados, deben ser ruido blanco. Esto significa que no deberían mostrar patrones de autocorrelación. La independencia de los residuos es crítica porque indica que el modelo ha capturado adecuadamente toda la información sistemática de la serie, dejando sólo ruido aleatorio. Si los residuos no son independientes, sugiere que el modelo no ha capturado completamente la estructura temporal, y puede requerir ajustes adicionales.

  Los Gráficos ACF y PACF de Residuos, Ayudan a visualizar cualquier autocorrelación remanente. Por otra parte, el test de Ljung-Box, Se utiliza para evaluar la independencia de los residuos mediante el examen de múltiples lags de la autocorrelación de residuos [14].

### 3.3.5.2. IMPLEMENTACIÓN Y AJUSTE DE MODELOS ARIMA Y SARIMA

- **Selección de Términos:** En primer lugar, **l**a función de autocorrelación (ACF) ayuda a identificar el número de términos de media móvil ($q$ para ARIMA y $Q$ para SARIMA) observando hasta qué punto los valores pasados influyen en la serie actual. Un corte claro en la ACF indica el orden del término MA. Por ejemplo, si la ACF corta después del segundo lag, $q$ puede ser 2 [13].

- **Función de Autocorrelación Parcial (PACF):** La PACF ayuda a determinar el número de términos autorregresivos ($p$ para ARIMA y P para SARIMA) al mostrar la correlación directa entre una observación y su n-ésimo lag. Un corte claro en la PACF indica el orden del término AR. Por ejemplo, si la PACF corta después del primer lag [13],

- **Pruebas Estadísticas para Diferenciación ($d$ y $D$):** Se utiliza para hacer estacionaria una serie que muestra tendencias. La prueba de Dickey-Fuller aumentada (ADF) se usa para verificar la necesidad de diferenciación. Utilizada para eliminar patrones estacionales. La ADF estacional o la inspección visual de gráficos estacionales pueden indicar la necesidad de $D$.

### 3.3.5.3.ESTIMACIÓN DE PARÁMETROS

Los métodos numéricos utilizados son máxima verosimilitud (Maximum Likelihood) y algoritmos de optimización. La máxima verosimilitud se utiliza para estimar los parámetros de los modelos ARIMA y SARIMA, maximizando la función de verosimilitud. Y los algoritmos de optimización Técnicas como el algoritmo de Newton-Raphson o el método BFGS pueden ser aplicados para optimizar los parámetros del modelo [14].

Una vez seleccionados los términos, los parámetros del modelo se estiman para minimizar el

error entre los valores predichos y los observados, mejorando la precisión de las predicciones del modelo.

### 3.3.5.4.EVALUACIÓN DEL MODELO ARIMA Y SARIMA

Estos son los indicadores más empleados:

- **Root Mean Square Error (RMSE):** Mide la precisión del modelo evaluando la magnitud promedio de los errores de predicción. Un RMSE más bajo indica un mejor ajuste del modelo.

- **Criterio de Información de Akaike (AIC):** Evalúa la calidad del modelo en relación con otros modelos, penalizando modelos más complejos. Un AIC más bajo sugiere un modelo mejor balanceado entre ajuste y complejidad.

- **Criterio de Información Bayesiana (BIC):** Similar al AIC, pero penaliza más fuertemente la complejidad del modelo, favoreciendo modelos más simples cuando dos modelos tienen un ajuste similar.

- **Comparación y Selección:** Los modelos se comparan en términos de RMSE, AIC, y BIC para seleccionar el que proporciona el mejor equilibrio entre precisión y parsimonia.

### 3.3.5.5. VALIDACIÓN CRUZADA

Evalúa la capacidad del modelo para generalizar a nuevos datos dividiendo la serie en subconjuntos de entrenamiento y validación en el tiempo, asegurando que las predicciones se realicen de manera que refleje la aplicación real del modelo. Una técnica comúnmente usada donde es Forward Chaining (Encadenamiento Adelante) se entrena el modelo en una parte inicial de la serie y se valida en la siguiente, asegurando que las validaciones sean cronológicamente ordenadas [10].

Por otro lado, otra técnica utilizada es Rolling Origin, la cual se utiliza para evaluar la estabilidad de las predicciones en diferentes períodos temporales, moviendo el punto de origen hacia adelante a medida que el modelo avanza en el tiempo.

Garantiza que el modelo no solo se ajuste bien a los datos pasados, sino que también sea robusto y preciso al hacer predicciones futuras.

Los modelos ARIMA y SARIMA son esenciales en la metodología del análisis de series de tiempo debido a su capacidad para modelar patrones complejos en datos temporales. Su implementación permite capturar tanto tendencias a largo plazo como fluctuaciones estacionales, proporcionando

una base sólida para la predicción y la toma de decisiones. La correcta aplicación y verificación de estos modelos asegura que las predicciones sean precisas y útiles para el análisis prospectivo y la gestión en contextos aplicativos, como en la predicción de precios de importación de pesticidas [15].

### 3.3.6. MODELO PROPHET

El modelo Prophet, desarrollado por Facebook, es un enfoque flexible y robusto para la previsión de series temporales. Este modelo es especialmente útil para datos que presentan tendencias no lineales con estacionalidad anual, semanal y diaria, y para los cuales es importante manejar los efectos de vacaciones.

#### 3.3.6.1. COMPONENTES DEL MODELO PROPHET

- **Tendencia:** Prophet admite dos tipos de modelos de tendencia: lineal y logístico. El modelo de tendencia lineal asume que el cambio es constante, mientras que el modelo logístico incorpora saturación en el crecimiento.

- **Estacionalidad:** De manera predeterminada, Prophet ajusta estacionalidades anuales y semanales mediante la descomposición de Fourier, permitiendo capturar patrones estacionales complejos.

- **Días Festivos:** Prophet permite añadir una lista de días festivos o eventos especiales para ajustar sus efectos en las predicciones, lo que es particularmente útil en sectores como el retail.

- **Efecto de Regresores Externos:** Se pueden incluir regresores adicionales que puedan afectar el comportamiento de la serie temporal.

#### 3.3.6.2.IMPLEMENTACIÓN DEL MODELO PROPHET

- **Preparación de Datos**: Los datos de entrada deben estar en un formato específico con dos columnas obligatorias: 'ds' para las fechas y 'y' para los valores observados. Es crucial asegurar la limpieza y preparación adecuada de los datos antes de la modelización.

- **Configuración del Modelo:** Durante la implementación, se pueden ajustar parámetros como la estacionalidad, el crecimiento de tendencia y la inclusión de días festivos. Prophet permite configurar la incertidumbre de los pronósticos a través de su intervalo de confianza.

- **Entrenamiento del Modelo:** Se entrena el modelo ajustando las observaciones históricas y optimizando los parámetros para obtener la mejor correspondencia con los datos observados.

**3.3.6.3.Generación de Pronósticos:** Después de entrenar el modelo, se genera un conjunto de pronósticos futuros que reflejan la tendencia proyectada y la estacionalidad.

### 3.3.6.4.EVALUACIÓN DEL MODELO PROPHET

Para evaluar el modelo Prophet, se utilizan métricas estándar de evaluación de modelos de series temporales:

- **Mean Absolute Error (MAE):** Proporciona la media de los errores absolutos entre las predicciones y los valores reales, ofreciendo una medida clara de la precisión del modelo.

- **Root Mean Square Error (RMSE):** Indica la magnitud promedio de los errores de predicción, con un RMSE más bajo sugiriendo un mejor ajuste.

- **Visualización de Pronósticos:** La visualización de las predicciones del modelo frente a los valores reales ayuda a identificar visualmente las áreas de mejora y ajustar el modelo.

### 3.3.6.5.VALIDACIÓN CRUZADA

Para asegurar la robustez del modelo Prophet y su capacidad para generalizar a datos no vistos, se emplea la validación cruzada, evaluando el modelo en diferentes subconjuntos de datos para evitar el sobreajuste.

- **División Temporal de los Datos:** Los datos se dividen en segmentos de entrenamiento y prueba de manera secuencial, evaluando el rendimiento del modelo en múltiples intervalos temporales.

- **Rolling Origin:** Esta técnica de validación implica mover el punto de origen hacia adelante a medida que se realizan nuevas predicciones, asegurando que las evaluaciones sean cronológicamente consistentes.

Prophet es una herramienta poderosa y flexible para el análisis de series temporales, especialmente útil para capturar patrones complejos y estacionales en los datos. Su facilidad de uso y capacidad para manejar datos con características no lineales y efectos de días festivos lo hacen ideal para aplicaciones en el mundo real.

### 3.3.7. RANDOM FOREST

Un modelo de Random Forest (Bosque Aleatorio) es un algoritmo de aprendizaje supervisado utilizado para tareas de clasificación y regresión. Este modelo, desarrollado por Leo Breiman y Adele Cutler, se fundamenta en la construcción de múltiples árboles de decisión independientes y en la combinación de sus resultados para mejorar la precisión y la estabilidad de las predicciones [16] [17]. El principio fundamental del Random Forest es que, al promediar los resultados de múltiples árboles, se reduce la varianza y se mitiga el riesgo de sobreajuste, proporcionando resultados más robustos en comparación con un único árbol de decisión. [18].

El proceso de construcción de un Random Forest implica dos técnicas clave: bagging (Bootstrap Aggregating) y la selección aleatoria de características [19]. Bagging consiste en generar subconjuntos de datos aleatorios con reemplazo a partir del conjunto de datos original y usar cada subconjunto para entrenar un árbol de decisión independiente. Además, en cada nodo de los árboles, se selecciona aleatoriamente un subconjunto de características disponibles para determinar la mejor división. Esto introduce diversidad y reduce la correlación entre los árboles individuales.

Entre las ventajas del Random Forest se destacan su alta precisión, la capacidad de manejar datos desbalanceados y la provisión de medidas de importancia de características. No obstante, este modelo también presenta desventajas, como una mayor complejidad computacional y una menor interpretabilidad en comparación con modelos más simples. [20]. En resumen, el Random Forest es una técnica poderosa y versátil en el campo del aprendizaje automático, reconocida por su capacidad para generar predicciones precisas y robustas en una amplia variedad de contextos.

#### 3.3.7.1.HIPERPARÁMETROS DE RANDOM FOREST

- **Número de árboles (n_estimators):** Es la cantidad de árboles de decisión que componen el bosque. Un mayor número de árboles generalmente mejora el desempeño, pero también incrementa el tiempo de cómputo.
- **Profundidad máxima (max_depth):** Limita la profundidad de cada árbol individual. Esto ayuda a prevenir el sobreajuste.
- **Número mínimo de muestras para dividir un nodo (min_samples_split):** Define el número mínimo de muestras necesarias para dividir un nodo interno.
- **Número mínimo de muestras en un nodo hoja (min_samples_leaf):** Establece el número mínimo de muestras que debe contener un nodo hoja.

### 3.3.8. MODELO XGBOOST

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje supervisado que se utiliza para tareas de clasificación y regresión. Desarrollado por Tianqi Chen y su equipo, XGBoost es una implementación optimizada del algoritmo de Gradient Boosting, que construye múltiples árboles de decisión en serie para mejorar la precisión del modelo. Este algoritmo es conocido por su eficiencia, velocidad y capacidad para manejar grandes conjuntos de datos y características [21].

El proceso de XGBoost implica la construcción de árboles de decisión en etapas secuenciales, donde cada árbol intenta corregir los errores cometidos por los árboles anteriores. Utiliza una técnica de optimización basada en gradientes, que ajusta los pesos de las observaciones en cada iteración para minimizar una función de pérdida específica [22]. Además, XGBoost incorpora regularización, lo que ayuda a prevenir el sobreajuste y mejora la generalización del modelo [23].

#### 3.3.8.1.REGULARIZACIÓN EN XGBOOST

XGBoost utiliza técnicas de regularización para evitar el sobreajuste y mejorar la capacidad de generalización del modelo. La regularización es un método que introduce un término de penalización en la función de pérdida del modelo para reducir la complejidad de los modelos y prevenir que se ajusten demasiado a los datos de entrenamiento.

En XGBoost, se utilizan dos tipos de regularización:

- **Regularización L1 (Lasso):** Esta técnica agrega una penalización basada en el valor absoluto de los coeficientes del modelo. Ayuda a reducir algunos coeficientes a cero, lo que resulta en modelos más simples y la selección automática de características.
- **Regularización L2 (Ridge):** Esta técnica aplica una penalización cuadrática sobre los coeficientes, lo que evita grandes coeficientes al reducir el riesgo de sobreajuste. L2 promueve soluciones con coeficientes más distribuidos uniformemente.

La función de pérdida regularizada de XGBoost se expresa como:

$$\mathrm{L}(\theta) = \sum \mathrm{i} = 1\mathrm{nl}(\mathrm{yi}, \mathrm{y\hat{}i}) + \lambda \sum \mathrm{j} = 1\mathrm{m} \mid \theta\mathrm{j} \mid + 21\alpha \sum \mathrm{j} = 1\mathrm{m}\theta\mathrm{j}2$$

Donde $l(yi,y\hat{}i)$l(y i , y^i ) es la pérdida estándar, $\lambda$ es el parámetro de regularización L1, y α es el parámetro de regularización L2.

La inclusión de estos términos de regularización ayuda a evitar el sobreajuste al imponer un coste a modelos excesivamente complejos, lo que resulta en predicciones más robustas y precisas.

#### 3.3.8.2. **HIPERPARAMETROS DE XGBOOST:**

- **Tasa de aprendizaje (learning_rate):** Controla la contribución de cada árbol al modelo final. Valores más pequeños requieren más árboles.
- **Número de árboles (n_estimators):** Similar a Random Forest, es el número total de árboles.
- **Profundidad máxima (max_depth):** Controla el grado de interacciones variables. Mayor profundidad puede llevar a sobreajuste.
- **Regularización L1 (alpha) y L2 (lambda):** Penalizan los pesos de las características para evitar el sobreajuste.

Entre las ventajas de XGBoost se encuentran su alta precisión, manejo eficiente de datos faltantes y la capacidad de paralelizar la construcción de árboles, lo que reduce significativamente el tiempo de

entrenamiento [24]. Sin embargo, XGBoost puede ser complejo de ajustar debido a la gran cantidad de hiperparámetros y su sensibilidad a la configuración de estos [25].

En resumen, XGBoost es una herramienta poderosa y versátil en el campo del aprendizaje automático, ampliamente utilizada en competencias de ciencia de datos y aplicaciones del mundo real debido a su capacidad para generar modelos precisos y robustos [26].

### 3.3.9. LightGBM

LightGBM (Light Gradient Boosting Machine) es un algoritmo de aprendizaje supervisado diseñado para realizar tareas de clasificación, regresión y ranking. Creado por Microsoft, LightGBM es una implementación mejorada del algoritmo de Gradient Boosting, conocida por su eficiencia y rapidez [27]. A diferencia de otros métodos de boosting, LightGBM emplea una técnica de crecimiento de árbol basada en histogramas que optimiza tanto el tiempo de entrenamiento como el uso de memoria.

El algoritmo de LightGBM construye árboles de decisión en fases secuenciales, donde cada árbol corrige los errores de los árboles anteriores [28]. LightGBM introduce dos innovaciones importantes: el crecimiento basado en hojas (leaf-wise growth) en lugar del crecimiento basado en niveles (level-wise growth), y la binarización de características mediante histogramas. Estas innovaciones permiten que LightGBM maneje grandes conjuntos de datos de manera más rápida y eficiente [29].

LightGBM destaca por su alta precisión, capacidad para manejar grandes volúmenes de datos y eficiencia computacional. Además, admite paralelización y técnicas avanzadas para manejar datos faltantes [25]. No obstante, puede ser más complicado de ajustar debido a su sensibilidad a la configuración de hiperparámetros en comparación con otros algoritmos [30].

En resumen, LightGBM es una herramienta potente y versátil en el ámbito del aprendizaje automático, ampliamente utilizada en aplicaciones prácticas y competiciones de ciencia de datos debido a su capacidad para producir modelos precisos y eficientes.

#### 3.3.9.1. HIPERPARAMETROS LIGHTGBM

- **Número de hojas (num_leaves):** Determina la complejidad del árbol. Un mayor número de hojas puede aumentar la precisión, pero también el riesgo de sobreajuste.
- **Fracción de submuestreo (subsample):** Es el porcentaje de datos a utilizar en cada iteración para prevenir el sobreajuste.
- **Frecuencia de la aplicación de bagging (bagging_freq):** Controla cuántas iteraciones se deben realizar antes de aplicar bagging.
- **Regularización (lambda_l1 y lambda_l2):** Controla la regularización L1 y L2, similar a XGBoost.

### 3.3.10. MÉTRICAS DE EVALUACIÓN

Para evaluar el desempeño de los modelos predictivos, se utilizan las siguientes métricas:

- **Root Mean Square Error (RMSE):** Mide la magnitud promedio del error entre las predicciones del modelo y los valores reales. Un RMSE más bajo indica un mejor ajuste.
- **Mean Absolute Error (MAE):** Calcula la media de las diferencias absolutas entre las predicciones y los valores reales. Es menos sensible a valores atípicos que el RMSE.
- **R-squared ($R^2$):** Representa la proporción de la varianza de la variable dependiente que es explicada por el modelo. Un valor cercano a 1 indica un modelo bien ajustado.

#### 3.3.11. VALIDACIÓN CRUZADA Y DIVISIÓN DE DATOS

La validación cruzada es una técnica utilizada para evaluar el desempeño de un modelo de manera más robusta. Consiste en dividir el conjunto de datos en múltiples subconjuntos o "folds". El modelo se entrena en varios de estos subconjuntos y se valida en el restante. Este proceso se repite varias veces y el rendimiento del modelo se promedia, lo cual ayuda a asegurar que el modelo no está sobre ajustado a una sola partición de los datos.

##### 3.3.11.1. DATOS DE ENTRENAMIENTO Y DE TESTEO:

- **Datos de Entrenamiento:** Son utilizados para ajustar el modelo. Se utilizan para aprender los patrones y relaciones dentro del conjunto de datos.
- **Datos de Testeo:** Se mantienen separados del entrenamiento y se utilizan exclusivamente para evaluar el rendimiento final del modelo. Proporcionan una estimación imparcial del desempeño del modelo sobre datos no vistos.

### 3.4. ANTECEDENTES

Entre otros proyectos examinados que podrían servir como referencia para este trabajo, se destaca una investigación que desarrolla un modelo de predicción del precio de Bitcoin utilizando Python y métodos de machine learning. En este estudio se utiliza Skforecast, una sencilla librería de Python que facilita, entre otras cosas, la adaptación de cualquier regresor de Scikit-learn a problemas de pronóstico. [31]

Este proyecto ha culminado exitosamente en la creación de un sofisticado modelo de pronóstico diseñado para prever el precio de Bitcoin. El modelo se ha desarrollado a partir de una exhaustiva serie temporal que abarca desde el año 2013 hasta el 2022, ofreciendo datos detallados sobre los precios de apertura, cierre, máximo y mínimo de Bitcoin en dólares estadounidenses. La implementación de este modelo representa un avance significativo en la capacidad de anticipar las fluctuaciones del mercado de criptomonedas, abordando así de manera efectiva la problemática analizada en esta investigación. Este enfoque, respaldado por un análisis exhaustivo de datos históricos y técnicas avanzadas de modelado, proporciona una herramienta invaluable para comprender y gestionar el riesgo asociado con la volatilidad del precio de Bitcoin. Además, su aplicabilidad trasciende el ámbito de la investigación, ofreciendo oportunidades para la toma de decisiones informadas en diversos contextos, desde inversiones financieras hasta estrategias

comerciales. La robustez y precisión de este modelo se sustentan en la combinación de teoría y práctica, respaldada por la literatura académica sobre pronóstico financiero y la innovación en el campo de la inteligencia artificial y el aprendizaje automático. En última instancia, este logro representa un hito significativo en la comprensión y la predicción de los movimientos del mercado de criptomonedas, abriendo nuevas perspectivas para la investigación y la aplicación práctica en el vasto y dinámico mundo de las finanzas digitales.

Un estudio adicional de notable relevancia que ofrece una perspectiva distinta es un proyecto centrado en la elaboración de un modelo de pronóstico de demanda intermitente utilizando la herramienta skforecast [32] Este enfoque introduce un método estadístico específicamente diseñado para anticipar la demanda de productos que exhiben patrones de venta esporádicos o irregulares. Este tipo de productos se caracterizan por experimentar períodos de alta demanda seguidos de lapsos con demanda escasa o nula. La capacidad de prever la demanda intermitente es esencial en sectores como la manufactura, el comercio minorista y la atención sanitaria, donde se requiere una gestión eficiente de los niveles de inventario para optimizar la producción y minimizar tanto las rupturas de stock como los costos asociados con un exceso de inventario. Este proyecto, respaldado por el uso de la herramienta skforecast y enmarcado en un contexto estadístico riguroso, ofrece una valiosa contribución al campo de la gestión de inventarios y la planificación de la producción, al abordar de manera efectiva los desafíos inherentes a la demanda intermitente en diversos sectores industriales. [32]

El uso de skforecast para la gestión de inventarios y la planificación de la producción ilustra cómo las herramientas estadísticas y de modelado pueden abordar desafíos de demanda intermitente. Aunque el contexto es diferente, el principio de utilizar herramientas avanzadas para manejar la variabilidad y la incertidumbre es altamente relevante para el trabajo de grado, donde se busca predecir precios en un mercado volátil. La adaptación de skforecast para prever demanda irregular puede inspirar estrategias para manejar la fluctuación de precios de pesticidas.

Otro estudio que proporciona una amplia visión para abordar la necesidad explorada en esta investigación es el titulado "Uso de Machine Learning para la toma de decisiones financieras". [33] Esta investigación presenta la toma de decisiones como un problema supervisado de clasificación binaria (y), donde se propone un modelo base utilizando un algoritmo de Regularización Ridge, Lasso y Elastic Net con Python. La elección de ElasticNet se fundamenta en su capacidad para imponer restricciones similares a las de la regresión lineal tradicional, al combinar las regularizaciones L1 y L2 para la selección de variables. Este enfoque permite una mejor comprensión y predicción de las decisiones financieras mediante el uso de técnicas avanzadas de aprendizaje automático, lo que potencialmente ofrece una herramienta valiosa para la gestión de riesgos y la optimización de carteras en entornos financieros dinámicos y complejos. [33]

Este enfoque destaca la aplicación de técnicas avanzadas de aprendizaje automático para mejorar la

comprensión y predicción de decisiones financieras. La metodología utilizada en este estudio, como el uso de algoritmos de regularización como ElasticNet, proporciona un ejemplo valioso de cómo integrar modelos de aprendizaje automático para obtener predicciones precisas en entornos dinámicos. La gestión de riesgos y la optimización de carteras en finanzas guardan similitudes con la necesidad de prever precios de importación en el sector agrícola, donde las predicciones precisas son esenciales para tomar decisiones informadas.

Aplicación de Redes Neuronales Convolucionales para el Análisis de Sentimientos en Redes Sociales" [9]. En esta investigación se ha emprendido un estudio exhaustivo que emplea redes neuronales convolucionales (CNN) como herramienta principal para explorar y analizar el complejo mundo del sentimiento expresado en las redes sociales. La metodología se centra en un enfoque supervisado, donde se procede a entrenar minuciosamente un modelo basado en CNN utilizando un vasto corpus de datos provenientes de diversas plataformas de redes sociales. Este modelo se diseña para clasificar el tono emocional presente en los mensajes, comentarios y publicaciones en línea, abarcando una amplia gama de temas y contextos. Esta investigación trasciende el mero análisis superficial de los datos sociales, ya que busca descifrar las complejidades subyacentes de las interacciones humanas en el ciberespacio. Los hallazgos obtenidos tienen importantes implicaciones para las empresas, organizaciones y analistas de mercado que buscan comprender la percepción pública, monitorear la reputación en línea y detectar tendencias emergentes en tiempo real. Además, esta investigación arroja luz sobre el papel fundamental de las redes neuronales convolucionales en el análisis de grandes volúmenes de datos no estructurados, destacando su versatilidad y eficacia en la extracción de información significativa de fuentes digitales masivas.

Este estudio se centra en el uso de Redes Neuronales Convolucionales (CNN) para el análisis de sentimientos en redes sociales. Aunque este enfoque se aplica en un contexto diferente al de la predicción de precios de pesticidas, la metodología subyacente de emplear técnicas avanzadas de aprendizaje automático es relevante. El uso de CNN para analizar grandes volúmenes de datos no estructurados es un ejemplo de cómo las técnicas de machine learning pueden adaptarse para abordar problemas complejos en diferentes dominios. Este enfoque subraya la importancia de seleccionar modelos adecuados y personalizarlos según las características específicas del problema, lo cual es aplicable al desarrollo del modelo predictivo en este trabajo.

Un estudio realizado por Chen y Guestrin (2016) se centró en la predicción de precios de viviendas utilizando el modelo XGBoost [21]. En este estudio, los investigadores aplicaron XGBoost para analizar un conjunto de datos masivo que incluía diversas características de propiedades residenciales, tales como el tamaño de la vivienda, el número de habitaciones, la ubicación y otros factores socioeconómicos relevantes.

El uso de XGBoost permitió manejar eficientemente las interacciones no lineales entre estas características, proporcionando predicciones precisas de los precios de las viviendas. La técnica de boosting por gradiente implementada en XGBoost mejoró significativamente el rendimiento del

modelo en comparación con otros métodos tradicionales, logrando una alta precisión y reduciendo los errores de predicción. Los resultados de este estudio demostraron la eficacia de XGBoost en la tarea de pronóstico de precios de bienes inmuebles, destacando su capacidad para procesar grandes volúmenes de datos y generar modelos robustos y precisos [21].

Este estudio demuestra la eficacia de XGBoost en el pronóstico de precios de bienes inmuebles, destacando su capacidad para manejar grandes volúmenes de datos y generar modelos robustos. La relevancia para el trabajo de grado radica en la aplicación de XGBoost como uno de los modelos seleccionados para predecir precios de pesticidas. La capacidad del algoritmo para capturar interacciones complejas y su eficiencia en el manejo de datos grandes son aspectos clave que se han aprovechado en este proyecto.

En otro estudio, Ke et al. (2017) investigaron el uso de LightGBM para predecir los precios de la energía eléctrica en mercados mayoristas. [27] Este trabajo utilizó un conjunto de datos históricos de precios de electricidad, junto con información adicional como demanda de energía, condiciones meteorológicas y datos de generación de energía. LightGBM se seleccionó por su capacidad para manejar grandes volúmenes de datos y su eficiencia en el entrenamiento de modelos complejos [27].

El enfoque basado en LightGBM permitió capturar patrones y tendencias en los datos históricos, resultando en predicciones precisas de los precios de la electricidad. La técnica de crecimiento basado en hojas y la binarización de características por medio de histogramas, implementadas en LightGBM [27], contribuyeron a reducir el tiempo de entrenamiento y mejorar la precisión del modelo. Los hallazgos de este estudio demostraron que LightGBM es una herramienta eficaz para el pronóstico de precios en mercados volátiles y dinámicos como el de la energía eléctrica [27].

Los hallazgos sobre la eficacia de LightGBM para el pronóstico de precios en mercados volátiles, como el de la energía eléctrica, son directamente aplicables al problema de predicción de precios de importación de pesticidas. LightGBM es conocido por su rapidez y precisión, características que son esenciales para manejar la naturaleza dinámica de los mercados agrícolas. Este estudio refuerza la decisión de incluir LightGBM como parte de los modelos evaluados en el trabajo de grado.

# 4. ENTENDIMIENTO DE LOS DATOS

## 4.1. EXTRACCION DE LOS DATOS

Los datos se encuentran alojados en la página de la DIAN. Estas bases estadísticas de Comercio Exterior comprenden el movimiento legal de mercancías que ingresan o salen del territorio aduanero nacional -TAN- a través de las diferentes aduanas del país, desde o hacia otros países y zonas francas del territorio nacional; a esta información se le ha aplicado filtros, ajustes estadísticos y lineamientos metodológicos estadísticos nacionales e internacionales, respecto de las operaciones de Comercio Exterior de Bienes, dando cumplimiento a los principios de transparencia y acceso a la información pública, sin perjuicio del cumplimiento de las normas vigentes en materia de protección de datos personales, los Principios Fundamentales de las Estadísticas Oficiales de las Naciones Unidas y el

Código Nacional de Buenas Prácticas del Sistema Estadístico Nacional del DANE. [34]

Para el análisis de la información se contó únicamente con los productos insecticidas, herbicidas y fungicidas de la base de importación y para identificarlos se tomaron en cuenta las partidas arancelarias de estos productos.

Una partida arancelaria es un código numérico asignado a productos específicos en el comercio internacional, utilizado para clasificar mercancías y determinar los aranceles, impuestos y restricciones aplicables a su importación o exportación. Estas partidas se encuentran en el Sistema Armonizado de Designación y Codificación de Mercancías (SA), un sistema estándar desarrollado por la Organización Mundial de Aduanas (OMA) que facilita el comercio internacional y la recopilación de estadísticas.

El código de una partida arancelaria generalmente consta de seis dígitos, aunque algunos países pueden añadir dígitos adicionales para mayor especificidad. Estos códigos permiten a las autoridades aduaneras identificar de manera precisa los productos que se están comercializando, aplicar las tarifas correspondientes y garantizar el cumplimiento de las regulaciones comerciales. [35]

A continuación, se detalla la información de partida arancelaria seleccionada (Tabla 1), la descripción de esa partida en la base de datos y el producto al que hace referencia.

***Tabla 1 Partidas arancelarias y descripción de productos***

| Partida Arancelaria | Descripción Partida | Producto |
|---|---|---|
| 3808911400 | QUE CONTENGAN PERMETRINA O CIPERMETRINA O DEMAS SUSTITUTOS SINTETICOS DEL PIRETRO (PIRETROIDES)- EXCEPTO LAS MENCIONADAS EN LA NOTA 2 DE SUBPARTIDA DE ESTE CAPITULO | Insecticida |
| 3808911900 | LOS DMS INSECTICIDAS- PRESENTADOS EN FORMAS O EN ENVASES PARA LA VENTA AL POR MENOR O EN ARTICULOS | Insecticida |
| 3808919100 | LOS DMS INSECTICIDAS- A BASE DE PIRETRO | Insecticida |
| 3808919400 | LOS DMS INSECTICIDAS- A BASE DE DIMETOATO | Insecticida |
| 3808919700 | QUE CONTENGAN PERMETRINA O CIPERMETRINA O DEMAS SUSTITUTOS SINTETICOS DEL PIRETRO (PIRETROIDES)- EXCEPTO LAS MENCIONADAS EN LA NOTA 2 DE SUBPARTIDA DE ESTE CAPITULO | Insecticida |
| 3808919800 | QUE CONTENGAN MIREX O ENDRINA | Insecticida |
| 3808919990 | LOS DMS INSECTICIDAS | Insecticida |
| 3808921100 | LOS DMS- FUNGICIDAS PRESENTADOS EN FORMAS O EN ENVASES PARA LA VENTA AL POR MENOR O EN ARTICULOS- QUE CONTENGAN BROMOMETANO (BROMURO DE METILO) O BROMOCLOROMETANO | Fungicida |
| 3808921200 | QUE CONTENGAN MANCOZEB- MANEB- PROPINEB O ZINEB | Fungicida |
| 3808921900 | LOS DMS FUNGICIDAS PRESENTADOS EN FORMAS O EN ENVASES PARA LA VENTA AL POR MENOR O EN ARTICULOS | Fungicida |
| 3808929100 | LOS DMS FUNGICIDAS- A BASE DE COMPUESTOS DE COBRE | Fungicida |

| | | |
|---|---|---|
| 3808929200 | LOS DMS FUNGICIDAS- A BASE DE PYRAZOFOS O DE BUTACLOR O DE ALACLOR | Fungicida |
| 3808929900 | LOS DMS FUNGICIDAS | Fungicida |
| 3808931100 | LOS DMS- HERBICIDAS- INHIBIDORES DE GERMINACION Y REGULADORES DEL CRECIMIENTO DE LAS PLANTAS- PRESENTADOS EN FORMAS O ENVASES O ENVASES PARA LA VENTA AL POR MENOS O EN ARTICULOS- QUE CONTENGAN BROMOMETANO (BROMURO DE METILO) O BROMOCLOROMETANO | Herbicida |
| 3808931900 | LOS DMS HERBICIDAS INHIBIDORES DE GERMINACION Y REGULADORES DEL CRECIMIENTO DE LAS PLANTAS; PRESENTADOS EN FORMAS O EN ENVASES PARA LA VENTA AL POR MENOR O EN ARTICULOS | Herbicida |
| 3808939300 | QUE CONTENGAN BUTACLOR | Herbicida |
| 3808939900 | LOS DMS- DE LOS DEMAS HERBICIDAS-INHIBIDORES DE GERMINACION Y REGULADORES DE CRECIMIENTO DE LAS PLANTAS | Herbicida |

*Elaboración propia, fuente DIAN*

El proceso comienzo con la instalación las bibliotecas necesarias, `wget` y `openpyxl`, mediante el uso de pip, para la gestión de archivos y datos en Python. Luego, procede a descargar un archivo zip de importaciones desde la página web de la DIAN en Colombia, utilizando la función `wget.download()`, y lo guarda en una ubicación específica en Google Drive. Posteriormente, descomprime el archivo zip descargado en una carpeta designada, utilizando el comando `unzip`. A continuación, carga los datos del archivo Excel descomprimido en un DataFrame de pandas para su posterior manipulación y análisis. Se lleva a cabo un filtrado específico para seleccionar únicamente las importaciones relacionadas con pesticidas, con el objetivo de profundizar en el análisis de estos productos. Además, se crea una nueva columna en el DataFrame para marcar la fecha de descarga de los datos, simplificando el seguimiento temporal de los mismos. Posteriormente, el DataFrame filtrado se exporta a un nuevo archivo Excel, facilitando su acceso y facilita compartir con otros usuarios.

En este proceso se observa que los datos en la variable importador contienen gran número de importadores que hacen referencia a la misma empresa pero que están escritos de forma diferente, por ejemplo, SYNGENTA S A y SYNGENTA S.A. De esta forma se realiza un procedimiento con el cual se homologan estos nombres tomando en cuenta los importadores con más volumen de producto ingresado al país entre los cuales encontramos al ya mencionado Syngenta de igual forma a Bayer.

Se contempla la una unión comercial entre Proficol y Adama ya que se encuentra información de Proficol hasta el 2014 y desde ese momento se menciona a Adama. También se realiza la unión de información de Dow con Corteva que se unieron comercialmente desde el 2012, de la misma forma que UPL que aparece con información hasta el 2012 y de allí en adelante tiene una unión comercial con Arysta. Finalmente, Nufarm que tiene información hasta el 2020 se une a Sumitomo.

Continuando con el procesamiento de los datos, el script cambia el directorio de trabajo al lugar donde se encuentran los archivos de importaciones y utiliza la función `glob` para encontrar todos

los archivos Excel presentes en ese directorio. A continuación, combina todos estos archivos en uno solo, denominado `combinado_xlsx`, utilizando la función `pd.concat()`, lo que permite consolidar la información de múltiples archivos en una única fuente de datos. Para entender mejor la estructura de los datos, se muestra el tipo de datos de cada atributo del DataFrame combinado. Además, se exporta este DataFrame a un archivo JSON, utilizando el formato 'split', para preservar la estructura de los datos. Finalmente, el script exporta el DataFrame original a otro archivo Excel, proporcionando una copia adicional de los datos en un formato ampliamente compatible y accesible para futuros análisis y usos. En resumen, el script automatiza la descarga, procesamiento, filtrado, combinación y exportación de datos de importaciones publicados por la DIAN en Colombia, lo que simplifica significativamente su análisis y utilización posterior.(Tabla 2)

***Tabla 2 Variable y tipo de variable de la base de datos***

| Variable | Tipo de variable |
|---|---|
| **IMPORTADOR_2** | Categórica |
| **EXPORTADOR_2** | Categórica |
| **DEPARTAMENTO_DESTINO** | Categórica |
| **SUBPARTIDA_ARANCELARIA** | Categórica |
| **PESO_NETO** | Numérica |
| **VALOR_FOB_USD** | Numérica |
| **DESCRIPCION_MERCANCIA** | Categórica |
| **TASA_CAMBIO** | Numérica |
| **FECHA_PRESENTACION** | Fecha |
| **VALOR_KILO** | Numérica |
| **PESTICIDA** | Categórica |

*Elaboración propia, fuente DIAN*

## 4.2. DESCRIPCION DE LAS VARIABLES

### 4.2.1. IMPORTADOR_2

Esta variable hace referencia a los importadores que pueden ser personas o empresas que traen los productos, bienes o servicios desde el extranjero para vender o utilizar en el país. Como ya se mencionó anteriormente algunos de los importadores fueron homologados debido a que en el transcurso de los años pasaron por alguna unión entre compañías o compra de estas

### 4.2.2. EXPORTADOR_2

Esta variable hace referencia a la empresa, persona o a la organización que despacha en el exterior el producto al cual estamos haciendo referencia ya sea fungicida, herbicida o insecticida. En algunos casos la empresa exportadora es la filial en el exterior de la empresa importadora por lo que se relaciona el mismo nombre y los 2 campos.

### 4.2.3. DEPARTAMENTO_DESTINO

El departamento de destino hace referencia a la zona del país que se espera sea el destino final de los productos importados. En este campo se observa que el departamento destino más mencionado es Bogotá

A continuación, se muestran los departamentos encontrados (Tabla 3)

***Tabla 3 Departamentos de destino***

| DEPARTAMENTO_DESTINO |
|---|
| BOGOTA DC |
| CUNDINAMARCA |
| BOLIVAR |
| ATLANTICO |
| TOLIMA |
| ANTIOQUIA |
| VALLE DEL CAUCA |
| MAGDALENA |
| META |
| NORTE DE SANTANDER |
| BOYACA |
| NARIÑO |
| SANTANDER |
| CAUCA |
| CORDOBA |
| SUCRE |
| CESAR |
| QUINDIO |
| RISARALDA |
| PUTUMAYO |
| DEPARTAMENTOS VARIOS |
| CASANARE |
| HUILA |

*Elaboración propia, fuente DIAN*

### 4.2.4. SUBPARTIDA_ARANCELARIA

En esta variable encontramos este número que es un código numérico de 10 dígitos, por el cual se puede identificar un producto determinado en cualquier lugar del mundo. El Sistema Armonizado (SA) establece un sistema numérico y de textos común, que permite clasificar de igual forma los productos que se comercializan internacionalmente.

CAPÍTULO: Son los 2 primeros dígitos de la codificación numérica.
PARTIDA: Se trata de los 4 dígitos de la codificación numérica.
SUBPARTIDA: Son los 6 primeros dígitos de la composición numérica

A partir del séptimo dígito, cada país tiene la potestad de numerar según sus necesidades de comercio. Por esta razón, es importante, antes de exportar, tener claridad sobre la posición exacta, con la cual entrará el producto a cada mercado.

Algunas modificaciones generan cambios en la descripción, contenido o cantidad de productos de las aperturas arancelarias, ya sea porque abarcan nuevas mercancías semejantes o porque disminuyen los bienes incluidos en ellos. Otros cambios implican que algunos números desaparezcan, porque los productos contenidos se trasladan o fusionan con otra apertura. [35]

### 4.2.5. PESO_NETO

Esta variable hace referencia al peso total de la mercancía que se está registrando al momento del ingreso al país. Esta variable hace parte de la construcción de una nueva que nos va a permitir realizar el pronóstico. Es un valor numérico que está dado en Kilos.

### 4.2.6. VALOR_FOB_USD

Corresponde al precio de venta de los bienes embarcados a otros países, puestos en el medio de transporte, sin incluir valor de seguro y fletes. [36] FOB, que significa "Free on Board" (Libre a Bordo), es un término de comercio internacional utilizado en los Incoterms (Términos Internacionales de Comercio) publicados por la Cámara de Comercio Internacional. Este término define las responsabilidades y riesgos entre el vendedor y el comprador en una transacción de bienes.

El vendedor es responsable de todos los costos y riesgos hasta que las mercancías se carguen a bordo del buque designado en el puerto de embarque acordado. El vendedor se encarga de la exportación y de cualquier despacho aduanero necesario.

Responsabilidades del comprador:
El comprador asume todos los costos y riesgos una vez que las mercancías están a bordo del buque.
El comprador es responsable del transporte marítimo, seguros, descarga en el puerto de destino y cualquier despacho aduanero en el país de importación.

El valor FOB dividido el Peso Neto de cada exportación registrada en los datos nos permite crear una variable que indica el precio del kilo en esa transacción en particular y que corresponde a la variable que veremos más adelante llamada VALOR_KILO.

### 4.2.7. DESCRIPCION_MERCANCIA

Esta variable hace referencia a una información que es ingresada al sistema por la compañía que realiza el proceso de importación al país e indica descripción a nivel general de la mercancía que está siendo transportada y que es objeto de revisión por la autoridad portuaria.

### 4.2.8. TASA_CAMBIO

Es el valor del dólar en pesos colombianos el día en que se realizó la importación del producto.

### 4.2.9. FECHA_PRESENTACION

Esta variable en formato fecha es la que nos permite realizar los modelos asociados a las series de tiempo con la que se podrá estimar la predicción en las variables de tiempo correspondientes.

### 4.2.10. VALOR_KILO

Esta variable se crea en el ajuste de variables para generar un valor del kilo de producto ingresado al país en la importación y que se genera dividiendo el dato del VALOR_FOB sobre el PESO NETO de cada registro.

$$\frac{\text{VALOR\_FOB}}{\text{PESO NETO}} = \text{VALOR\_KILO}$$

### 4.2.11. PESTICIDA

Esta variable contiene la indicación de pesticida que está relacionada en la importación. Se encuentran los seleccionados inicialmente en la base de datos que son los Herbicidas, Insecticidas y Fungicidas.

Los insecticidas son químicos utilizadas para matar o controlar insectos. Se emplean en agricultura, salud pública y el hogar para proteger cultivos, controlar plagas y prevenir enfermedades transmitidas por insectos. Pueden actuar por contacto, ingestión o inhalación y se clasifican en orgánicos sintéticos, orgánicos naturales e inorgánicos. Su uso debe ser regulado debido a posibles efectos negativos en el medio ambiente y la salud.

Los fungicidas son sustancias químicas utilizadas para prevenir o eliminar hongos en plantas y otros materiales. Se emplean principalmente en la agricultura para proteger cultivos de enfermedades fúngicas. Pueden ser de contacto o sistémicos y se clasifican en orgánicos e inorgánicos. Su uso debe ser regulado para evitar resistencia de hongos y proteger el medio ambiente.

Los herbicidas son productos químicos diseñados para eliminar plantas no deseadas, también conocidas como malas hierbas, en la agricultura, jardinería y espacios urbanos. Funcionan interfiriendo con el crecimiento, el desarrollo o el metabolismo de las plantas objetivo.

# 5. ANALISIS EXPLORATORIO DE LOS DATOS

## 5.1.MÉTODO DE ANÁLISIS EXPLORATORIO DE DATOS

Después de realizar un análisis detallado de cada método se optó por el método EDA. la elección del Análisis Exploratorio de Datos (EDA) para este análisis se justifica de manera sólida por su capacidad integral para proporcionar una comprensión profunda y detallada de los datos [37]. A través del EDA, se puede explorar exhaustivamente la estructura y distribución de los datos, identificar patrones y tendencias clave, así como detectar posibles anomalías o valores atípicos que podrían influir significativamente en los resultados del modelo predictivo. Además, el EDA facilita la preparación rigurosa de los datos al abordar de manera eficiente problemas como datos faltantes, valores nulos o duplicados, lo que establece una base sólida para análisis posteriores y garantiza la integridad y la calidad de los resultados obtenidos [37].

Además, el EDA permite validar supuestos subyacentes sobre la distribución de los datos y las relaciones entre variables, asegurando la validez y fiabilidad de los análisis posteriores. La identificación de correlaciones y relaciones entre las variables proporciona información valiosa para guiar la selección de características y la construcción de modelos más precisos y explicativos. En conjunto, la aplicación rigurosa y sistemática del método de EDA garantiza un análisis completo y robusto de los datos, estableciendo una base sólida para el desarrollo de modelos predictivos confiables y efectivos.

### 5.1.1. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El Análisis Exploratorio de Datos (EDA) es una técnica fundamental para comprender y preparar datos antes de construir modelos predictivos. Este método fue popularizado por John Tukey en 1977 con su libro "Exploratory Data Analysis". El EDA permite a los analistas obtener una comprensión profunda de los datos mediante visualizaciones y estadísticas descriptivas, facilitando la identificación de patrones, tendencias y relaciones [37]. Además, el EDA es crucial para detectar valores atípicos y anomalías, así como para manejar valores nulos y datos faltantes, asegurando que los datos estén limpios y preparados adecuadamente. Sin embargo, este método puede ser muy detallado y consumir tiempo considerablemente, además de que la interpretación de gráficos y estadísticas descriptivas puede ser subjetiva [37].

**Ventajas:**

- **Comprensión Profunda:** Proporciona una comprensión profunda de los datos, permitiendo identificar patrones, tendencias y relaciones.
- **Detección de Anomalías:** Ayuda a identificar valores atípicos y anomalías que podrían distorsionar el análisis.
- **Preparación de Datos:** Facilita la limpieza y transformación de datos, manejando valores nulos y datos faltantes.

**Desventajas:**

- **Consumo de Tiempo:** Puede ser muy detallado y consumir tiempo considerablemente.
- **Subjetividad:** La interpretación de gráficos y estadísticas descriptivas puede ser subjetiva y depender del analista.

## 5.2.ESTADISTICAS DESCRIPTIVAS

El análisis exhaustivo de los datos proporciona información crucial sobre las variables clave en nuestro estudio (Tabla 4).

***Tabla 4 Estadísticas Descriptivas de Variables Relevantes***

| | PESO_NETO | VALOR_FOB_USD | TASA_CAMBIO | VALOR_KILO |
|---|---|---|---|---|
| **CANTIDAD** | 60,401 | 60,401 | 60,401 | 60,401 |
| **MEDIA** | 9,611.97 | 67,048.61 | 3,033.48 | 46.25 |
| **DESVIACION ST** | 15,293.14 | 100,839.00 | 887.83 | 620.00 |
| **MINIMO** | 0.01 | 0.00 | 1,748.41 | 0.00 |
| **25%** | 1,260.00 | 15,464.97 | 1,967.08 | 3.83 |
| **50%** | 4,248.00 | 38,397.86 | 3,027.39 | 7.81 |
| **75%** | 12,240.00 | 79,092.48 | 3,770.35 | 20.00 |
| **MAXIMO** | 255,067.53 | 3,334,924.00 | 5,058.02 | 115,280.00 |

*Elaboración propia, fuente DIAN*

Al examinar la variable PESO_NETO, se evidencia una amplia distribución de los datos, con un total de 60,401 observaciones. Los pesos netos varían desde 0.01 hasta 255,067.53 unidades, lo que indica una gran variabilidad en las observaciones. La media de peso neto es de aproximadamente 9,611.97 unidades, lo que sirve como referencia para entender el peso neto típico en nuestras muestras. La desviación estándar de alrededor de 15,293.14 unidades sugiere una dispersión significativa alrededor de la media, lo que puede indicar la presencia de valores atípicos o una amplia variación en los pesos netos observados.

En cuanto a la variable VALOR_FOB_USD, se observa una diversidad similar en los datos. Con un total de 60,401 observaciones, los valores FOB varían desde 0.00 hasta 3,334,924.00 dólares estadounidenses. La media de aproximadamente 67,048.61 dólares estadounidenses ofrece una indicación del valor FOB promedio en nuestras muestras. La desviación estándar de alrededor de 100,839.00 dólares estadounidenses sugiere una amplia dispersión alrededor de la media, indicativa de una variedad de condiciones o factores que influyen en los valores FOB observados.

La variable TASA_CAMBIO muestra una tasa promedio de alrededor de 3,033.48, con un rango desde 1,748.41 hasta 5,058.02. La dispersión moderada de los datos, indicada por una desviación estándar de aproximadamente 887.83, sugiere cierta consistencia en las tasas de cambio registradas. Sin embargo, la amplia gama de valores podría reflejar la influencia de diversos factores económicos

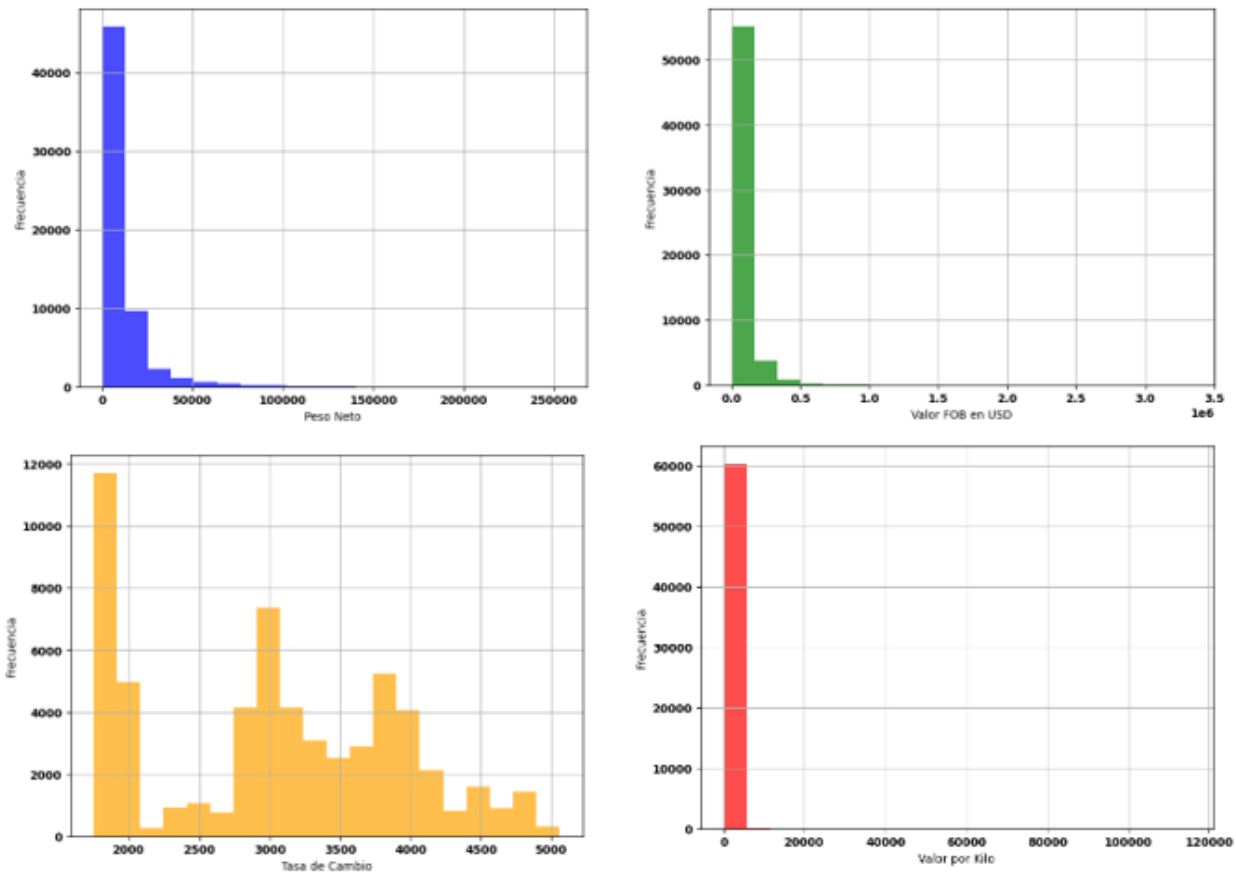
o de mercado en las tasas de cambio.

Considerando el VALOR_KILO, se observa una amplia variabilidad en los datos. Con un total de 60,401 observaciones, los valores por kilo van desde 0.00 hasta 115,280.00 unidades. La media de aproximadamente 46.25 unidades proporciona una estimación del valor promedio por kilo en nuestras muestras. La desviación estándar de alrededor de 620.00 indica una dispersión significativa alrededor de la media, sugiriendo una variación considerable en los valores por kilo observados en nuestro conjunto de datos.

El histograma de frecuencia para la variable de peso neto (Ilustración 3) muestra una distribución altamente sesgada hacia la derecha, con una media de aproximadamente 9,611.97 y una desviación estándar de 15,293.14. Esto sugiere que la mayoría de las observaciones tienen valores relativamente bajos, con una cola larga de valores más altos que se extienden hacia la derecha. Además, la presencia de valores atípicos en el extremo superior del rango contribuye a esta asimetría. La mayoría de las observaciones se concentran en el extremo inferior del rango, mientras que hay algunos valores excepcionalmente grandes en el extremo superior, lo que indica una variabilidad considerable en los datos.

Para la variable del valor FOB en dólares, el histograma de frecuencia también muestra una distribución sesgada hacia la derecha, con una media de aproximadamente 67,048.61 y una desviación estándar de 100,839.00. Al igual que con el peso neto, la distribución está sesgada hacia valores más bajos, con algunos valores extremadamente altos que se extienden hacia el extremo superior del rango. Esto sugiere una concentración de observaciones en el extremo inferior del rango, con una dispersión considerable de valores en el extremo superior. La presencia de valores atípicos en el extremo superior contribuye a la asimetría de la distribución, lo que indica una variabilidad significativa en los datos.

*Ilustración 3 Histogramas de frecuencias*



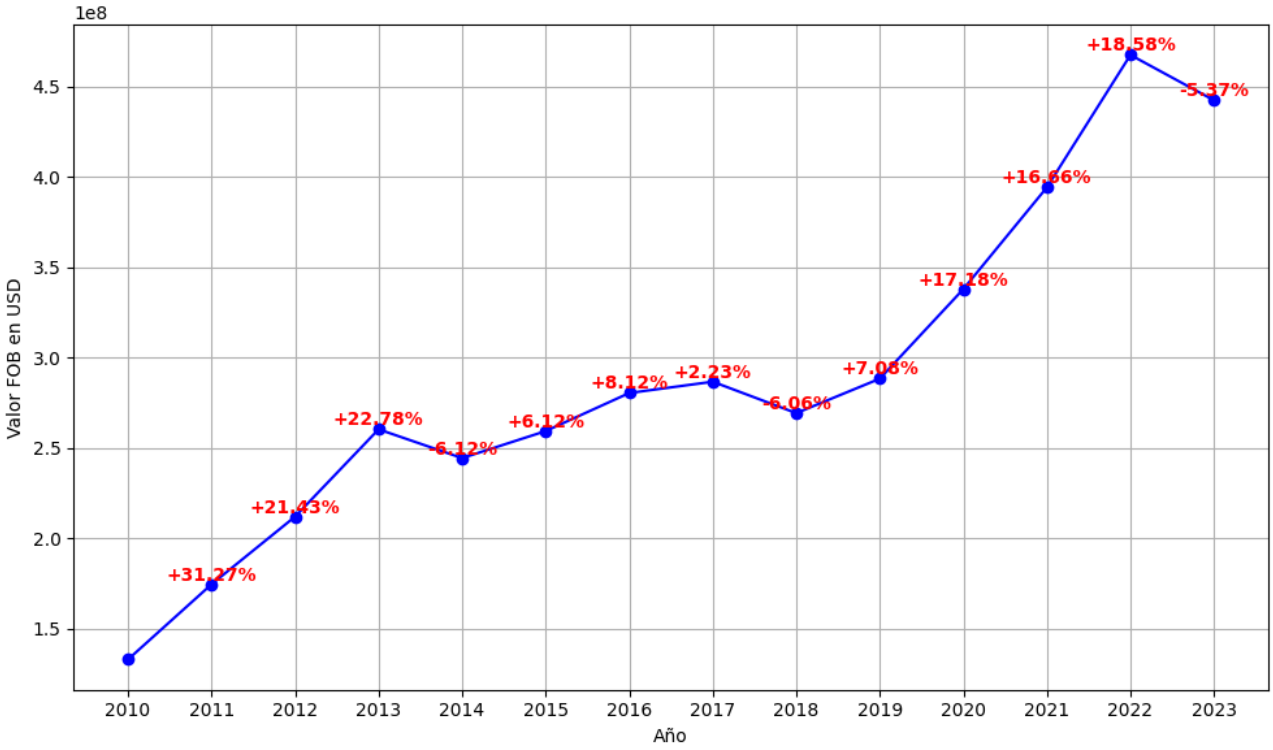

*Elaboración propia, Fuente Dian*

En resumen, este análisis detallado ofrece una comprensión integral de las características estadísticas de las variables de interés en nuestro estudio, lo que permite una interpretación más precisa y fundamentada de los resultados obtenidos.

## 5.3.ANÁLISIS DEL COMPORTAMIENTO DEL VALOR FOB EN USD (2010-2023): TENDENCIAS Y VARIACIONES

La evolución del valor FOB (Free on Board) en dólares estadounidenses para las importaciones de pesticidas en Colombia desde el año 2010 hasta el 2023 (Ilustración 4) muestra un comportamiento con alta volatilidad, pero con una tendencia creciente a lo largo del tiempo. Cada punto en el gráfico representa el valor FOB total de las importaciones en un año específico. La tendencia general muestra un incremento notable en los valores FOB, con algunas fluctuaciones que pueden atribuirse a variaciones económicas y políticas comerciales.

*Ilustración 4 Comportamiento del Valor FOB en USD por Año*




*Elaboración propia, Fuente DIAN*

Desde 2010 hasta 2023, el valor FOB en USD ha mostrado fluctuaciones significativas, reflejando diversas tendencias económicas y comerciales. En 2010, el valor FOB en USD se situó en $132,995,932.76. A partir de este año, se observó un fuerte crecimiento en los tres años siguientes. En 2011, hubo un incremento notable del 31.27%, llevando el valor a $174,587,535.01. Este crecimiento continuó en 2012 y 2013 con aumentos del 21.43% y 22.78%, alcanzando valores de $212,007,611.39 y $260,303,536.25 respectivamente.

Entre 2014 y 2019, se observaron ajustes y fluctuaciones en el valor FOB. En 2014, hubo una ligera caída del 6.12%, reduciendo el valor a $244,378,971.16. Este fue el primer año en el período analizado con una disminución. Sin embargo, en 2015, el valor volvió a crecer un 6.12%, alcanzando $259,325,771.02, seguido de un aumento del 8.12% en 2016, llevando el valor a $280,378,106.75. En 2017, experimentó un crecimiento moderado del 2.23%, con un valor de $286,634,059.53. En 2018, hubo una ligera disminución del 6.06%, pero en 2019, el valor subió nuevamente un 7.08% a $288,319,185.22.

El período de 2020 a 2022 mostró un crecimiento significativo y alcanzó puntos álgidos. En 2020, el valor FOB en USD experimentó un aumento considerable del 17.18%, alcanzando $337,856,675.86. Este crecimiento se mantuvo en 2021 con un incremento del 16.66%, situando el valor en $394,131,307.62. El año 2022 registró el valor máximo en el período analizado con $467,359,328.85,

tras un aumento del 18.58%.

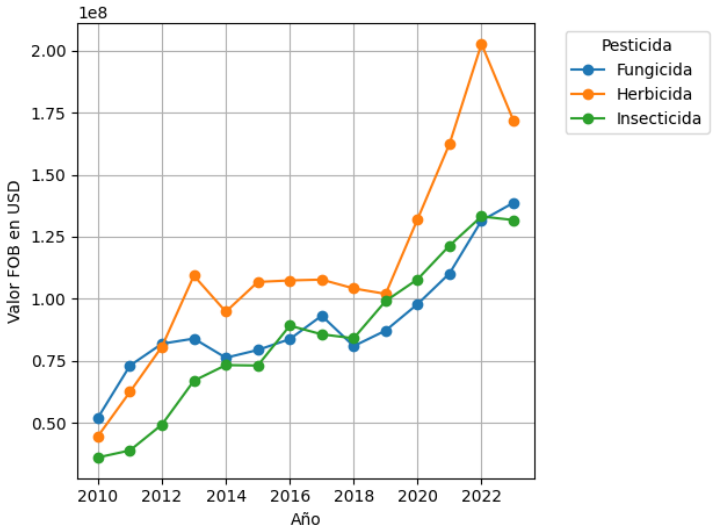
En 2023, se produjo una disminución del 5.37%, llevando el valor a $442,268,108.21. Este ajuste podría sugerir una corrección después de varios años de crecimiento o una respuesta a cambios económicos específicos que necesitarían un análisis más detallado para comprender las causas subyacentes.

Como se observa en la Ilustración 4, el valor FOB en USD ha mostrado una tendencia creciente desde 2010 hasta 2023, con picos significativos en los años 2012 y 2021. Estos incrementos reflejan el impacto de las fluctuaciones en la tasa de cambio y otros factores económicos globales.

La Ilustración 5 desglosa el valor FOB en dólares estadounidenses por tipo de pesticida (insecticidas, fungicidas, herbicidas) para cada año desde 2010 hasta 2023. Cada punto en las líneas representa el valor FOB anual para un tipo específico de pesticida, lo que permite comparar cómo ha variado el valor FOB entre los diferentes tipos de pesticidas a lo largo del tiempo.

Al examinar los valores máximos y mínimos registrados para cada tipo de pesticida (Ilustracion5), se puede apreciar una notable variabilidad en la demanda y el uso de estos productos a lo largo del tiempo. Por ejemplo, el fungicida muestra un aumento significativo en su valor máximo desde 2010 hasta 2022, lo que sugiere un posible incremento en la necesidad de controlar enfermedades fungales en los cultivos. Por otro lado, el herbicida experimenta un crecimiento constante desde 2010 hasta 2022, alcanzando su valor máximo en este último año, lo que puede reflejar una mayor preocupación por el control de las malezas en la agricultura. Mientras tanto, el insecticida muestra un patrón de variabilidad más pronunciado, con fluctuaciones en su valor máximo a lo largo de los años, lo que puede estar relacionado con cambios en las prácticas agrícolas, la aparición de nuevas plagas o el desarrollo de productos más efectivos.

***Ilustración 5 Valor FOB en USD por Pesticida (Año a Año)***




*Elaboración propia, Fuente DIAN*

Al analizar las variaciones porcentuales año a año, se identifican períodos de crecimiento significativo en la demanda de ciertos pesticidas. Por ejemplo, se observa un aumento destacado en el uso de fungicidas entre 2010 y 2011, así como entre 2018 y 2019. Estos incrementos pueden estar relacionados con cambios en las condiciones climáticas, la prevalencia de enfermedades en los cultivos o la introducción de nuevas tecnologías en el sector agrícola. De manera similar, el aumento en el uso de herbicidas entre 2010 y 2020 sugiere una mayor necesidad de controlar las malezas en los campos, posiblemente debido a la resistencia de las malezas a los herbicidas existentes o a la expansión de áreas de cultivo.

En resumen, los datos ofrecen una visión detallada del comportamiento del mercado de pesticidas a lo largo de los años. Estos hallazgos pueden ser útiles para entender las tendencias en la agricultura, identificar áreas de crecimiento potencial en la industria de los pesticidas y desarrollar estrategias efectivas para la gestión de plagas y enfermedades en los cultivos. Sin embargo, es importante tener en cuenta que estos datos representan solo una parte del panorama y que pueden estar influenciados por una variedad de factores, incluyendo condiciones climáticas, prácticas agrícolas y cambios en la regulación gubernamental. Por lo tanto, es fundamental realizar análisis más detallados y considerar información adicional para obtener una comprensión completa de la dinámica del mercado de pesticidas.

## 5.4.DINÁMICA TEMPORAL DEL VALOR POR KILO

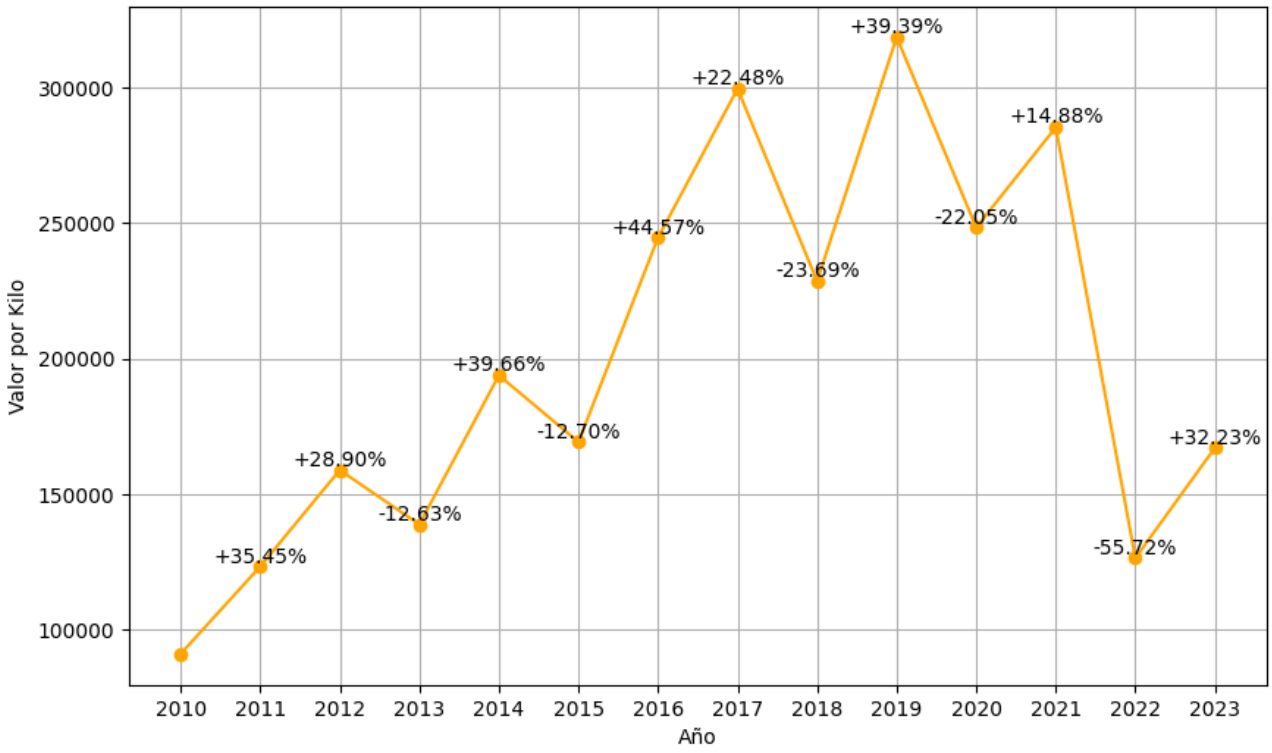
Al analizar el comportamiento del valor por kilo de pesticidas importados en Colombia desde 2010 hasta 2023 (Ilustración 6). Se observa que cada punto en el gráfico indica el valor promedio por kilo en un año determinado. Este gráfico ayuda a visualizar cómo ha fluctuado el costo por kilo de pesticidas a lo largo del tiempo, mostrando tendencias de aumento y disminución.

En 2010, el valor por kilo comenzó en 90,974.66 unidades y permaneció relativamente estable, experimentando una variación porcentual nula en comparación con el año anterior. Sin embargo, en el año siguiente, 2011, se observó un aumento significativo del 35.45% en el valor por kilo, alcanzando un total de 123,229.13 unidades. Este aumento continuó en 2012, con un incremento del 28.90%, lo que llevó el valor por kilo a 158,841.04 unidades.

A partir de 2013, sin embargo, se observa un cambio en la tendencia. Hubo una disminución notable del 12.63% en el valor por kilo en comparación con el año anterior, lo que resultó en un total de 138,777.13 unidades en 2013. Esta tendencia a la baja continuó en los años siguientes, con una disminución del 12.70% en 2015, alcanzando un mínimo de 169,197.22 unidades.

No obstante, el año 2016 marcó un cambio significativo en la dirección, con un aumento drástico del 44.57% en el valor por kilo, llegando a 244,614.36 unidades. Este aumento continuó en 2017 con un incremento del 22.48%, alcanzando un máximo de 299,594.42 unidades.

***Ilustración 6 Comportamiento del Valor por Kilo por Año***




*Elaboración propia, Fuente DIAN*

Sin embargo, esta tendencia alcista no se mantuvo constante en los años siguientes. En 2018, se observó una disminución considerable del 23.69% en el valor por kilo, cayendo a 228,614.12 unidades. Esta disminución continuó en 2019, con una caída del 22.04%, llegando a 248,402.83 unidades.

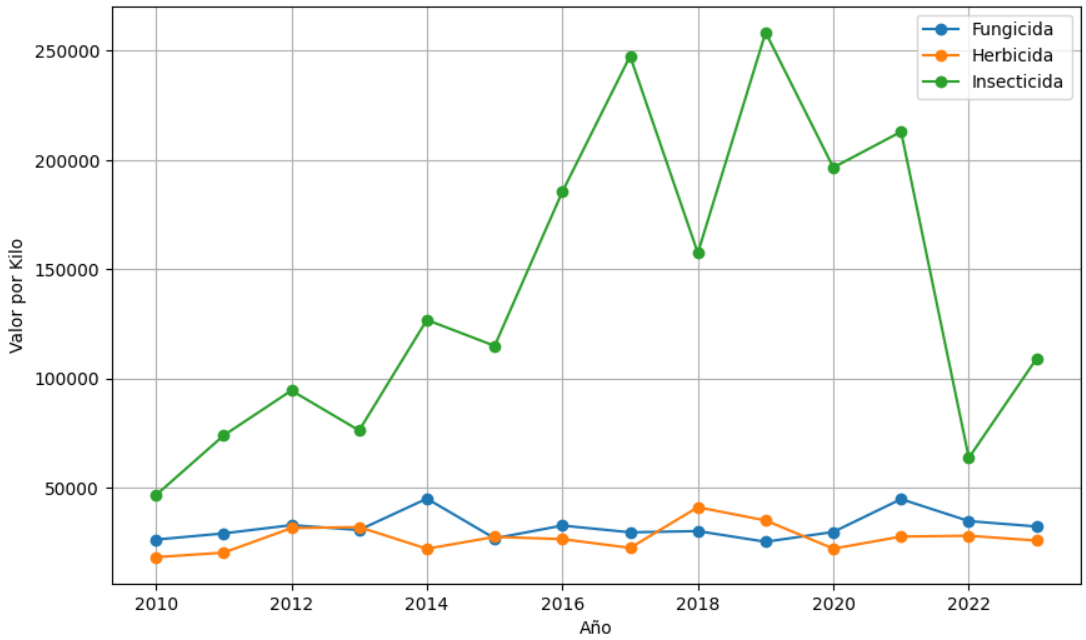
A pesar de estos altibajos, se produjo una recuperación en 2021, con un aumento del 14.88% en el valor por kilo, alcanzando un total de 285,362.24 unidades. Sin embargo, en 2022, se observó una caída significativa del 55.72%, disminuyendo a 126,352.82 unidades. Finalmente, en 2023, se registró un aumento del 32.23%, llegando a 167,079.11 unidades.

Según se muestra en la Ilustración 6, el valor por kilo de pesticidas ha experimentado variaciones significativas a lo largo del tiempo. Por ejemplo, en 2016 y 2017 se observan picos notables, posiblemente debido a factores económicos específicos de esos años.

Al desagregar las importaciones de pesticidas, se revelan patrones interesantes en la adquisición de estos productos a lo largo del tiempo (ilustración 7). Se procede a descomponer el valor por kilo de pesticidas importados por tipo (insecticidas, fungicidas, herbicidas) desde 2010 hasta 2023. Cada punto representa el valor promedio por kilo para un tipo de pesticida en un año específico. Esto permite identificar patrones y comparaciones entre los diferentes tipos de pesticidas.

En primer lugar, se observa que las importaciones de fungicidas muestran una tendencia creciente desde su inicio en 2010, alcanzando un máximo en 2014 con un valor de 44,979.07, antes de experimentar una disminución en los años posteriores. Esta tendencia es respaldada por una variación porcentual que refleja aumentos significativos, como el notorio incremento del 110.78% entre 2010 y 2011.

***Ilustración 7 Comportamiento del Valor por Kilo por Año y por pesticida***




*Elaboración propia, Fuente DIAN*

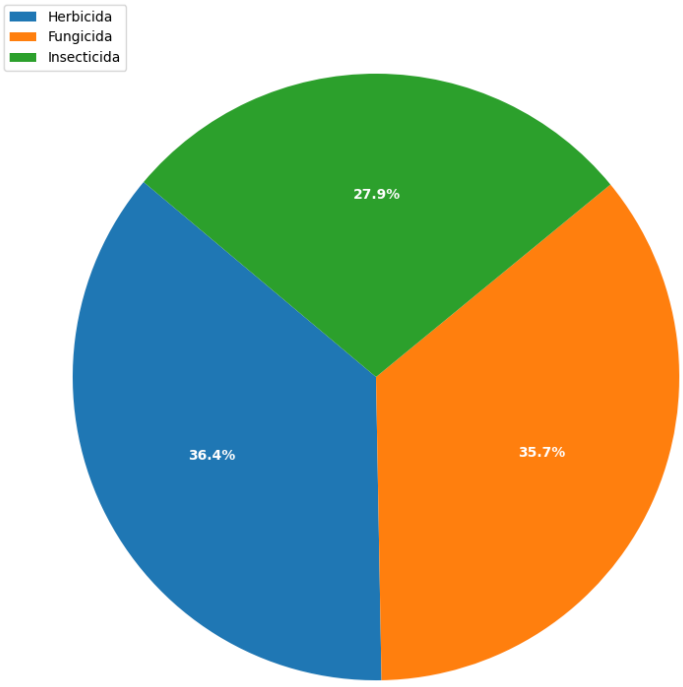
Por otro lado, las importaciones de herbicidas también muestran una tendencia general al alza a lo largo del período estudiado, aunque con variaciones anuales notables. El año 2018 destaca como el de mayor volumen de importaciones de herbicidas, con un valor de 41,045.02, mientras que el año inicial de la serie temporal en 2010 registra el menor valor, con 18,175.77. La variación porcentual muestra un aumento significativo del 111.50% entre 2010 y 2011, seguido de incrementos más moderados en años subsiguientes.

Por último, las importaciones de insecticidas exhiben una tendencia de crecimiento constante y marcada a lo largo del período estudiado. Desde un valor inicial en 2010 de 46,566.27, las importaciones de insecticidas experimentan un aumento significativo hasta alcanzar su punto máximo en 2021 con 212,971.73. La variación porcentual también muestra un crecimiento constante, con destacados aumentos, como el registrado en 2011 con un incremento del 158.70% respecto al año anterior. Estos hallazgos resaltan la importancia de monitorear las importaciones de pesticidas para comprender las tendencias en la adquisición de estos productos y su impacto en las prácticas agrícolas y el medio ambiente.

## 5.5.ANÁLISIS DE LA DISTRIBUCIÓN DE PESTICIDAS IMPORTADOS

El análisis de los datos revela una distribución equilibrada en cuanto a los tipos de pesticidas importados en Colombia. Los resultados muestran que los herbicidas representan la mayor proporción, con un 36.4% del total de pesticidas importados, seguidos de cerca por los fungicidas, con un 35.65%. Por otro lado, los insecticidas constituyen el 27.95% restante. Este panorama refleja una demanda diversificada en el mercado colombiano de productos químicos destinados al control de plagas. Aunque los herbicidas lideran ligeramente en términos de participación, las diferencias entre las tres categorías no son significativas, lo que subraya la necesidad de una amplia gama de soluciones para enfrentar los desafíos agrícolas del país. Este análisis proporciona una visión integral de la dinámica del mercado de pesticidas en Colombia, ofreciendo información valiosa para la formulación de políticas y estrategias en el sector agrícola (Ilustración 8).

***Ilustración 8 Share de Participación por tipo de Pesticida***



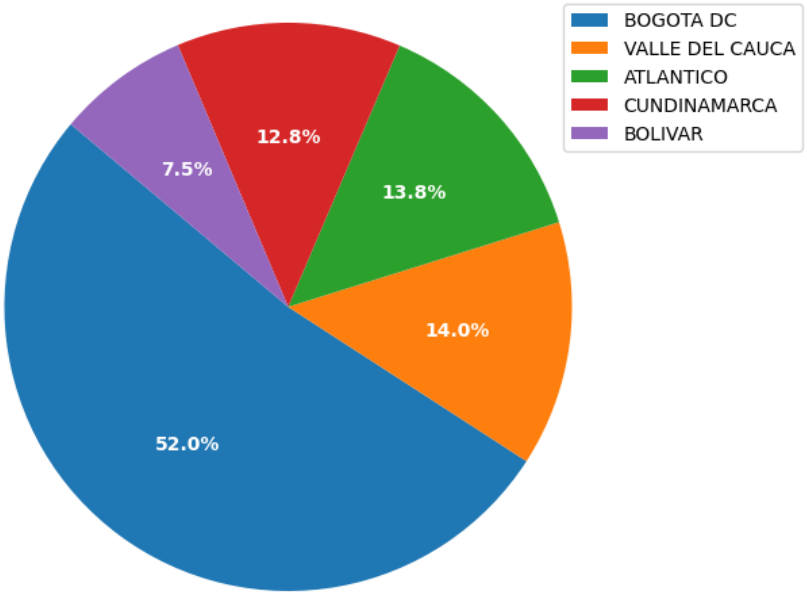

*Elaboración propia, Fuente DIAN*

## 5.6.IMPORTACIÓN DE PESTICIDAS DISTRIBUCIÓN REGIONAL

El análisis de los datos revela que la importación de pesticidas en Colombia está ampliamente concentrada en cinco departamentos principales (Ilustracion 9), con Bogotá DC liderando significativamente, representando aproximadamente el 52% del total de importaciones. Esta alta participación podría atribuirse a una combinación de factores, incluida la presencia de empresas importadoras, la demanda del mercado agrícola y la infraestructura logística en la capital del país.

Le sigue Valle del Cauca, con alrededor del 14% de las importaciones, seguido de Atlántico y Cundinamarca, que representan cada uno alrededor del 14% y el 13% respectivamente. Estos departamentos podrían estar influenciados por su importante actividad agrícola y su posición geográfica estratégica para el comercio.

***Ilustración 9 Distribución importaciones por departamento***




*Elaboración propia, Fuente DIAN*

El análisis del share de participación también destaca la importancia relativa de cada departamento en el panorama general de la importación de pesticidas en Colombia. Aunque estos cinco principales departamentos dominan la escena, es crucial reconocer que otros departamentos también contribuyen significativamente al total nacional de importaciones, aunque en menor medida. Esto sugiere una distribución más amplia del uso de pesticidas en todo el país, lo que puede reflejar la diversidad de los sistemas agrícolas y las necesidades específicas de control de plagas en diferentes regiones.

Estos hallazgos subrayan la importancia de considerar las variaciones regionales en el uso de pesticidas al desarrollar políticas y estrategias agrícolas en Colombia. Un enfoque integral que reconozca las particularidades de cada región podría ser esencial para abordar los desafíos específicos y promover prácticas agrícolas sostenibles y seguras en todo el país. Además, este análisis proporciona una base sólida para investigaciones adicionales sobre las tendencias de importación de pesticidas, su impacto ambiental y la efectividad de las regulaciones en Colombia.

# 6. MODELOS

## 6.1.ANÁLISIS DE ESTACIONARIEDAD Y COINTEGRACIÓN

### 6.1.1. PRUEBAS DE ESTACIONARIEDAD

Para asegurar que los modelos predictivos aplicados fueran adecuados, se realizaron pruebas de estacionariedad en la serie temporal 'VALOR_KILO'. La Prueba de Dickey-Fuller Aumentada (ADF) arrojó un estadístico ADF de -74.4699 con un p-valor de 0.0, permitiendo rechazar la hipótesis nula de que la serie tiene una raíz unitaria. Esto indica que la serie es estacionaria, lo cual es crucial para la aplicabilidad de los modelos ARIMA y SARIMA. La Prueba KPSS corroboró estos hallazgos con un estadístico KPSS de 0.4277 y un p-valor de 0.0652, sugiriendo que no se puede rechazar la hipótesis nula de estacionariedad. Ambas pruebas concluyen que la serie temporal 'VALOR_KILO' no presenta tendencias cambiantes y tiene una media y varianza constantes a lo largo del tiempo.

La decisión de realizar el análisis con la información colapsada, en lugar de hacerlo por tipo de pesticida, se tomó por varias razones estratégicas. En primer lugar, la consolidación de datos permitió un análisis más robusto al incrementar el tamaño de la muestra, lo cual mejora la confiabilidad estadística de las pruebas realizadas. Además, al colapsar la información, se obtiene una visión general del comportamiento del precio por kilo de pesticidas, reflejando tendencias y patrones comunes a todos los tipos de pesticidas considerados.

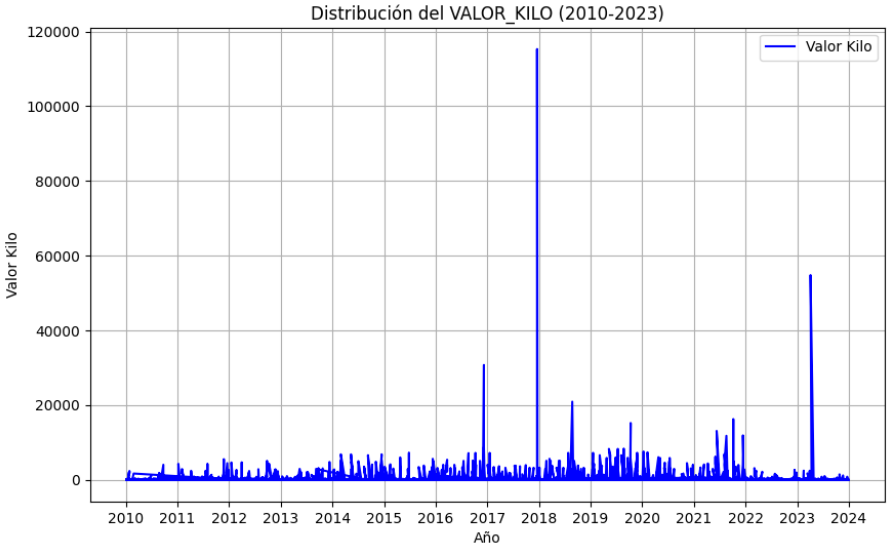
Este enfoque es útil para generar modelos predictivos generales que puedan ser aplicables de manera amplia en el sector agrícola. Aunque se reconoce que cada tipo de pesticida puede tener características y factores específicos que influyen en su precio, el objetivo inicial fue desarrollar un modelo predictivo general. Futuras investigaciones podrán enfocarse en el análisis desagregado por tipo de pesticida, proporcionando modelos más específicos y ajustados a cada categoría.

### 6.1.2. PRUEBA DE COINTEGRACIÓN DE ENGLE-GRANGER

Para explorar las relaciones a largo plazo entre 'VALOR_KILO' y 'PESO_NETO', se utilizó la Prueba de Cointegración de Engle-Granger. Los resultados mostraron un estadístico de cointegración de -74.5147 y un p-valor de 0.0, indicando una fuerte relación de equilibrio a largo plazo entre estas series temporales. Esto sugiere que cualquier desviación de corto plazo entre las dos series se corregirá a lo largo del tiempo, volviendo a una trayectoria equilibrada.

A continuación, se presentan los gráficos de ambas series temporales, 'VALOR_KILO' y 'PESO_NETO', para visualizar mejor su comportamiento y distribución a lo largo del tiempo:

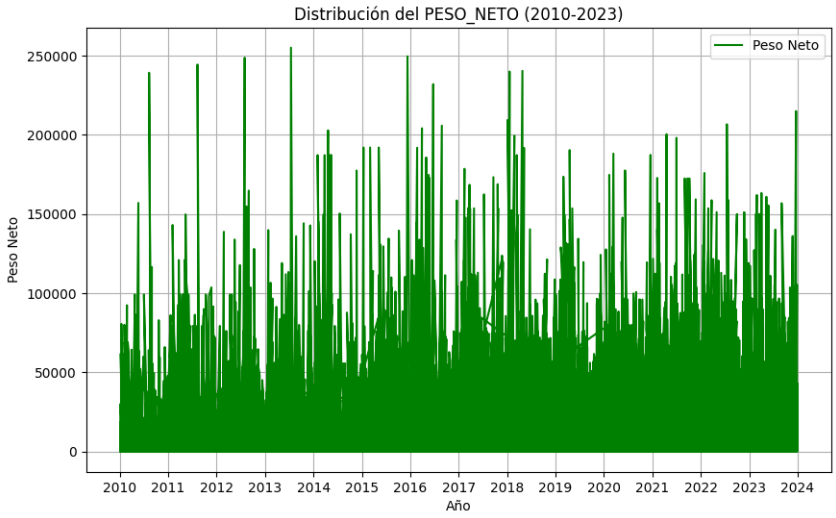
***Ilustración 10 Distribución del Valor Kilo a través del tiempo***




*Elaboración propia, Fuente Dian*

En la primera grafica (ilustración 10) muestra la distribución del valor por kilo de los pesticidas importados en Colombia desde 2010 hasta 2023. Se puede observar cómo el 'VALOR_KILO' ha fluctuado a lo largo del tiempo, con picos significativos en ciertos años que reflejan variaciones en el mercado y otros factores económicos.

***Ilustración 11 distribución del PESO_NETO a través del tiempo***




*Elaboración propia, Fuente Dian 1*

Por otra parte, el siguiente grafico (ilustración 11) presenta la distribución del peso neto de los pesticidas importados durante el mismo periodo. La visualización ayuda a identificar patrones de importación y posibles correlaciones con el 'VALOR_KILO'.

La inclusión de estos gráficos proporciona una comprensión visual complementaria de las series temporales analizadas. La fuerte relación de equilibrio a largo plazo entre 'VALOR_KILO' y 'PESO_NETO' se hace más evidente al observar cómo ambas series se comportan de manera coherente a lo largo del tiempo.

## 6.2.MODELOS ARIMA

### 6.2.1. PRIMER AJUSTE DEL MODELO ARIMA

El modelo ARIMA(1, 1, 1) ajustado muestra coeficientes significativos para los términos AR(1) y MA(1), indicando que hay una dependencia temporal capturada en los datos. El valor de AIC es 947850.905. Sin embargo, es importante tener en cuenta que el AIC se utiliza principalmente para la comparación relativa de diferentes modelos. Un menor valor de AIC indica un mejor equilibrio entre la complejidad del modelo y el ajuste de los datos en comparación con otros modelos. Por lo tanto, este valor debe interpretarse en el contexto de una comparación con otros modelos potenciales.

El modelo ARIMA(1 1 1) ajustado muestra coeficientes significativos para los términos AR(1) y MA(1). El valor de AIC es 947850.905. Sin embargo, es fundamental recordar que el AIC se utiliza principalmente para la comparación de modelos. Un menor valor de AIC, en comparación con otros modelos, indicaría un mejor equilibrio entre la complejidad del modelo y el ajuste de los datos. Por lo tanto, este valor de AIC debe evaluarse en relación con los valores de AIC de otros modelos considerados.

#### 6.2.1.2. DIAGNÓSTICO DE RESIDUOS

- **Prueba de Ljung-Box**: Con un valor p de 0.73, no se encuentra autocorrelación significativa en los residuos, lo cual es positivo.
- **Prueba de Jarque-Bera**: Un valor p de 0.00 indica que los residuos no siguen una distribución normal, lo que puede ser problemático para algunos supuestos del modelo.
- **Prueba de Heterocedasticidad**: Un valor p de 0.00 sugiere la presencia de heterocedasticidad, indicando que la varianza de los residuos no es constante a lo largo del tiempo.
- **Sesgo y Curtosis**: Valores altos de sesgo (123.43) y curtosis (21112.73) indican que los residuos tienen una distribución altamente asimétrica y con picos extremos, respectivamente.

Aunque el modelo ARIMA (tabla 5) captura las dependencias temporales básicas, la no normalidad y heterocedasticidad de los residuos sugieren que podría ser necesario aplicar transformaciones adicionales o considerar modelos alternativos para mejorar el ajuste y la previsión.

***Tabla 5 Modelo Arima***

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | VALOR_KILO | No. Observations: | 60401 |
| Model: | ARIMA(1, 1, 1) | Log Likelihood | -473922.453 |
| Date: | Sun, 16 Jun 2024 | AIC | 947850.905 |
| Time: | 22:02:21 | BIC | 947877.932 |
| Sample: | 0 | HQIC | 947859.298 |
| | - 60401 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | 0.0684 | 0.000 | 156.850 | 0.000 | 0.068 | 0.069 |
| ma.L1 | -1.0000 | 6.24e-05 | -1.6e+04 | 0.000 | -1.000 | -1.000 |
| sigma2 | 3.827e+05 | 26.476 | 1.45e+04 | 0.000 | 3.83e+05 | 3.83e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.12 | Jarque-Bera (JB): | 1121632251236.97 |
| Prob(Q): | 0.73 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 18.17 | Skew: | 123.43 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 21112.73 |

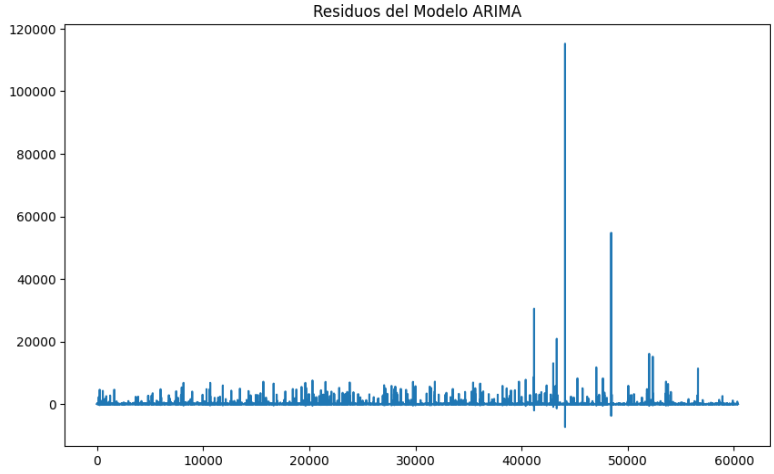
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

*Elaboración propia, Fuente Dian*

### *6.2.1.3. GRÁFICO DE RESIDUOS*

El gráfico de residuos (ilustración 12) muestra grandes picos y valores atípicos, especialmente alrededor de la observación 40,000 y en otras regiones posteriores. Esto sugiere que el modelo ARIMA no captura adecuadamente todos los patrones en los datos y que existen valores extremos significativos que el modelo no puede predecir con precisión.

***Ilustración 12 Residuos modelo Arima***

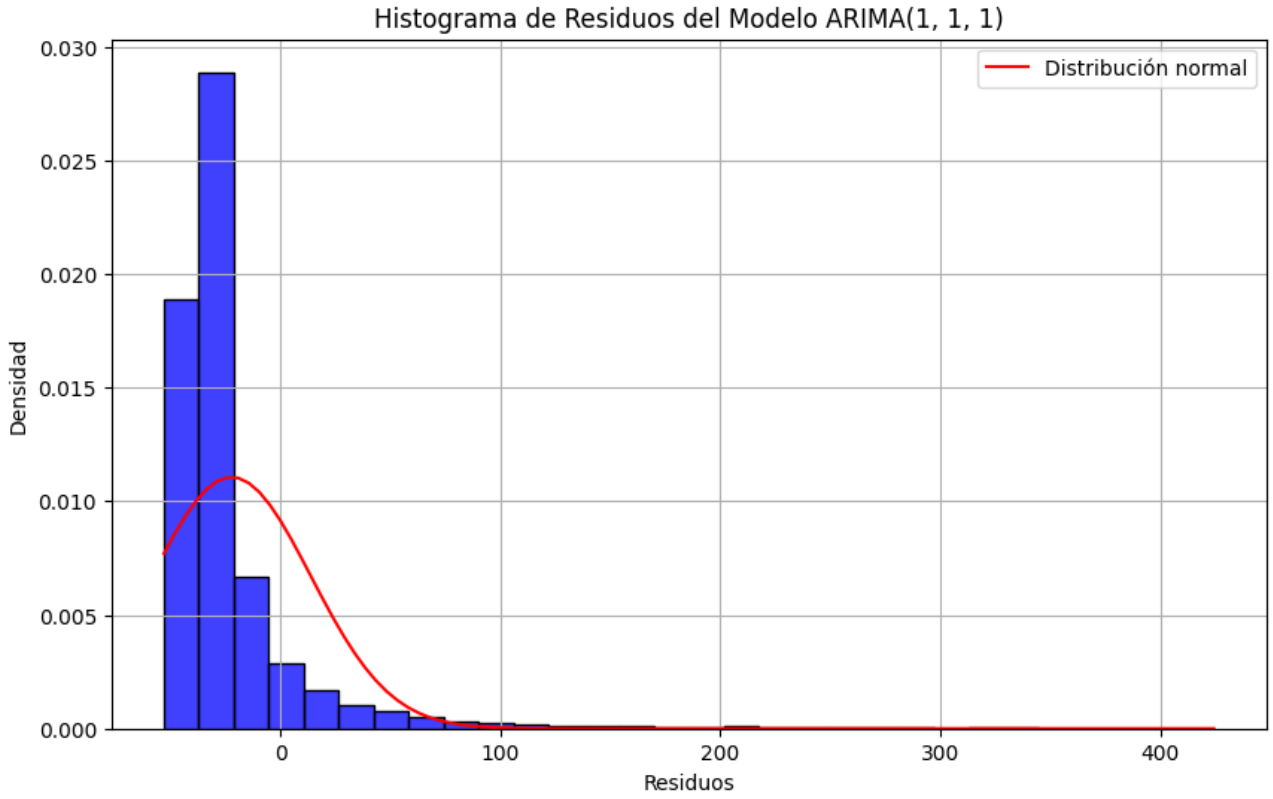



*Elaboración propia, Fuente Dian*

### *6.2.1.4. HISTOGRAMA DE RESIDUOS*

El histograma de residuos (ilustración 13) muestra una distribución que se desvía de la normalidad, con una concentración de datos cerca del cero y una cola larga hacia valores positivos más altos. Aunque no se observan picos extremos que distorsionen significativamente la escala, la asimetría positiva es evidente. Esta distribución indica que los residuos no cumplen completamente con la suposición de normalidad requerida por el modelo ARIMA, sugiriendo la necesidad de revisar el modelo o considerar transformaciones adicionales de los datos para mejorar el ajuste.

***Ilustración 13 Histograma residuos Arima***




*Elaboración propia, Fuente Dian*

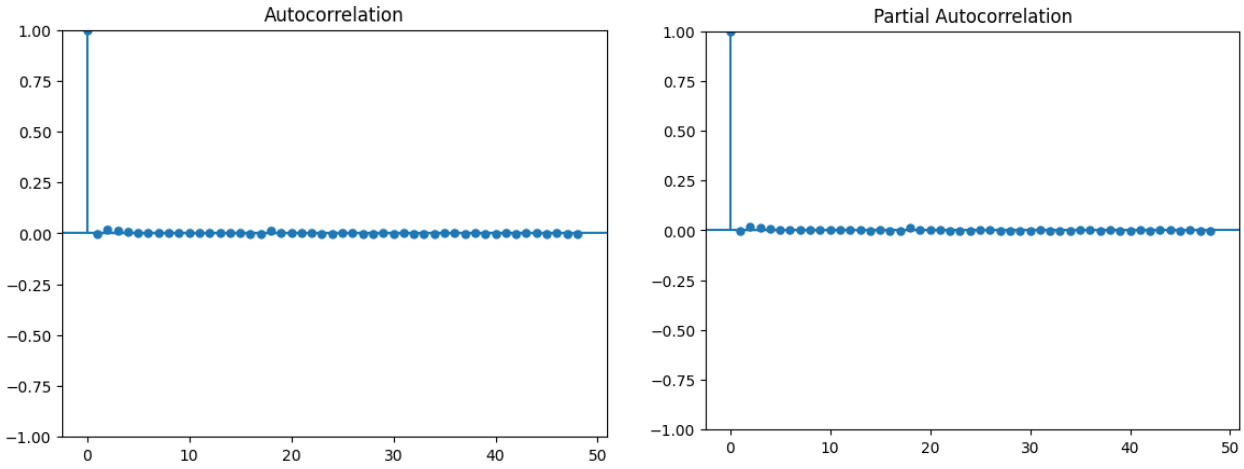
### *6.2.1.5. GRÁFICOS DE AUTOCORRELACIÓN (ACF) Y AUTOCORRELACIÓN PARCIAL (PACF)*

La interpretación adecuada de los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF) es crucial para comprender la dinámica subyacente de los residuos de un modelo de series temporales. Estos gráficos nos ayudan a identificar si quedan autocorrelaciones no explicadas en los residuos, lo que podría sugerir la necesidad de ajustes adicionales en el modelo.

En el gráfico de autocorrelación (ACF) (ilustración 14), observamos que, como es típico, el lag 0 muestra una autocorrelación de 1, dado que una serie siempre está perfectamente correlacionada consigo misma en el mismo tiempo. Sin embargo, para lags mayores a cero, la autocorrelación cae rápidamente hacia valores cercanos a cero y se mantiene constante a lo largo de lags sucesivos. Este patrón es indicativo de que no hay autocorrelaciones significativas en los residuos, lo que es un buen indicador de que el modelo actual está capturando adecuadamente la dependencia temporal de los datos.

Por su parte, el gráfico de autocorrelación parcial (PACF) refuerza esta interpretación al mostrar también un rápido descenso hacia valores insignificantes después del lag 0. La falta de picos significativos en lags más altos sugiere que no se requieren términos autoregresivos adicionales en el modelo actual, y que cualquier dependencia entre las observaciones es efectivamente modelada por la configuración actual.

***Ilustración 14 Autocorrelación***




*Elaboración propia, Fuente Dian*

Los análisis de los gráficos ACF y PACF sugieren que los residuos del modelo se comportan como ruido blanco, indicando que el modelo está bien especificado y que los supuestos fundamentales detrás del análisis de series temporales están siendo satisfechos. No se observan autocorrelaciones problemáticas que podrían invalidar las inferencias realizadas a partir del modelo. Esta adecuada especificación del modelo asegura que las predicciones y conclusiones derivadas son fiables y robustas, lo cual es vital para la toma de decisiones informada en el contexto de la importación de pesticidas en Colombia.

## 6.3.MODELO SARIMA

### 6.3.1. AJUSTE DEL MODELO SARIMA

El modelo SARIMA se optimizó para incluir componentes estacionales, esenciales para capturar patrones repetitivos en los datos. Se ajustó el modelo SARIMA (0, 1, 1)(1, 1, 1, 12) utilizando el conjunto de entrenamiento. Los pronósticos mostraron una estabilización futura de "VALOR_KILO" similar al modelo ARIMA, validando la importancia de los componentes estacionales en la predicción.

- **Prueba de Ljung-Box**: Con un valor p de 0.00, indica autocorrelación en los residuos, lo que es problemático para la validez del modelo.
- **Prueba de Jarque-Bera**: Un valor p de 0.00 sugiere que los residuos no siguen una distribución normal, implicando la presencia de residuos extremos y sesgados.
- **Prueba de Heterocedasticidad**: Un valor p de 0.00 sugiere que la varianza de los residuos no es constante, lo que complica la previsión.

Adicionalmente, los altos valores de sesgo (122.51) y curtosis (20868.73) reflejan una distribución de residuos altamente asimétrica y con picos extremos (Tabla 6). Aunque el modelo SARIMA captura las dinámicas estacionales y no estacionales en los datos, los problemas de autocorrelación en los residuos, la no normalidad y la heterocedasticidad indican que se necesitan ajustes adicionales o transformaciones para mejorar la precisión del modelo y la capacidad predictiva.

***Tabla 6 Modelo Sarima***

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | VALOR_KILO | No. Observations: | 60401 |
| Model: | SARIMAX(0, 1, 1)x(1, 1, 1, 12) | Log Likelihood | -481960.205 |
| Date: | Sun, 16 Jun 2024 | AIC | 963928.410 |
| Time: | 22:25:16 | BIC | 963964.444 |
| Sample: | 0 | HQIC | 963939.599 |
| | - 60401 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ma.L1 | -0.9989 | 0.001 | -927.438 | 0.000 | -1.001 | -0.997 |
| ar.S.L12 | 0.0172 | 0.004 | 4.888 | 0.000 | 0.010 | 0.024 |
| ma.S.L12 | -0.9993 | 0.001 | -1006.942 | 0.000 | -1.001 | -0.997 |
| sigma2 | 8.765e+05 | 216.932 | 4040.419 | 0.000 | 8.76e+05 | 8.77e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 276.92 | Jarque-Bera (JB): | 1095636422038.05 |
| Prob(Q): | 0.00 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 17.83 | Skew: | 122.51 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 20868.73 |

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

*Elaboración propia, Fuente Dian*

### 6.3.2. TRANSFORMACIÓN LOGARÍTMICA

***Tabla 7 Modelo Sarima con transformación logarítmica***

RMSE: 620.4416222792084

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | VALOR_KILO | No. Observations: | 60401 |
| Model: | SARIMAX(0, 1, 1)x(1, 1, 1, 12) | Log Likelihood | -102745.423 |
| Date: | Mon, 17 Jun 2024 | AIC | 205498.846 |
| Time: | 03:37:21 | BIC | 205534.880 |
| Sample: | 0 | HQIC | 205510.035 |
| | - 60401 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ma.L1 | -0.8555 | 0.001 | -855.190 | 0.000 | -0.857 | -0.854 |
| ar.S.L12 | -0.0260 | 0.004 | -6.826 | 0.000 | -0.033 | -0.019 |
| ma.S.L12 | -0.9996 | 0.000 | -3531.867 | 0.000 | -1.000 | -0.999 |
| sigma2 | 1.7568 | 0.004 | 477.362 | 0.000 | 1.750 | 1.764 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 1581.49 | Jarque-Bera (JB): | 450262.50 |
| Prob(Q): | 0.00 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 1.03 | Skew: | -0.31 |
| Prob(H) (two-sided): | 0.05 | Kurtosis: | 16.36 |

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

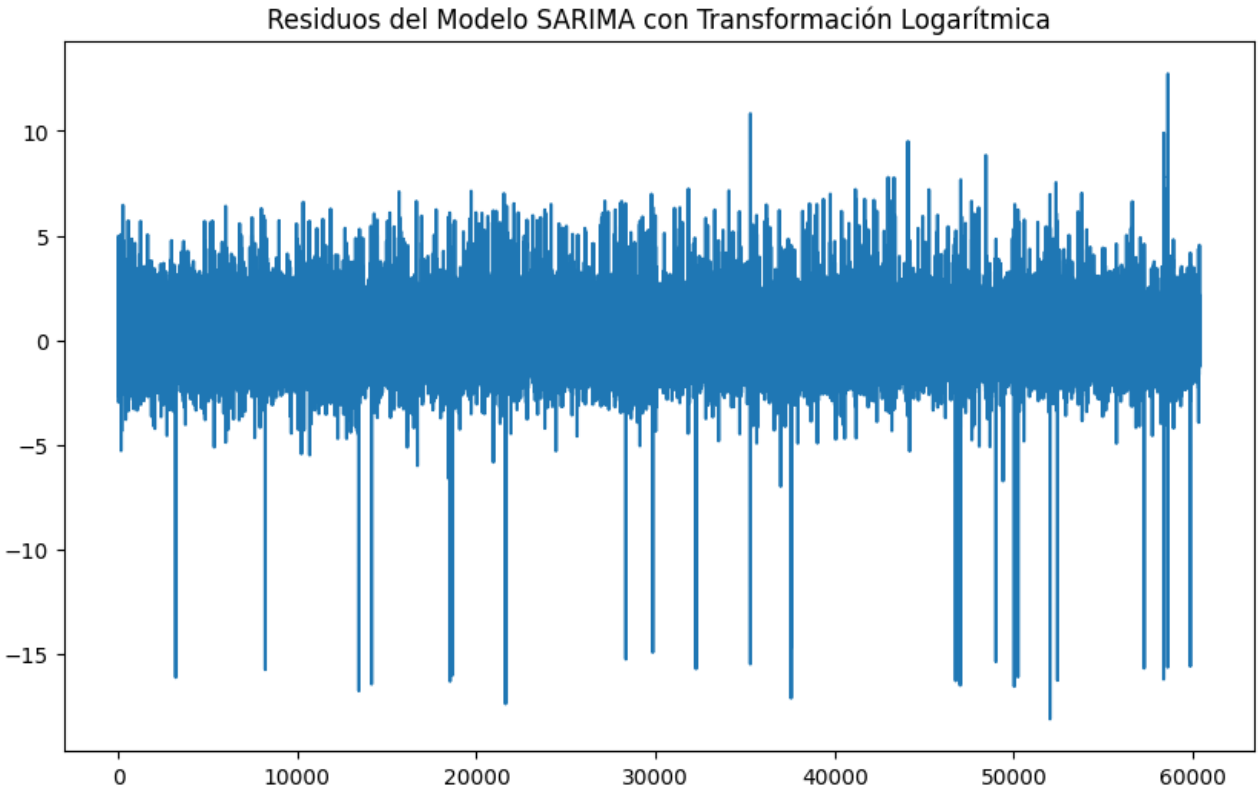
*Elaboración propia, Fuente Dian*

Se aplicó una transformación logarítmica a la serie temporal "VALOR_KILO" para manejar mejor la variabilidad y los valores extremos. El modelo SARIMA(0, 1, 1)x(1, 1, 1, 12) se ajustó a la serie transformada y se revirtieron los valores pronosticados a la escala original utilizando la exponencial inversa. El RMSE del modelo ajustado fue de 620.44, similar al modelo sin transformar. Aunque la transformación logarítmica redujo ligeramente la heterocedasticidad, los problemas de autocorrelación y distribución no normal de los residuos persisten, indicando que esta transformación no mejoró significativamente la precisión del modelo.

### 6.3.2.1.**EVALUACIÓN DE RESIDUOS**

El gráfico de residuos del Modelo SARIMA con Transformación Logarítmica (ilustración 15) revela que, aunque la mayoría de los residuos se distribuyen alrededor de cero, existen algunos picos significativos que indican la presencia de valores atípicos. La amplitud relativamente constante de los residuos a lo largo del tiempo indica cierta homogeneidad en la variabilidad de los errores, lo cual es un buen indicador de homocedasticidad en el modelo. Esta característica sugiere que la varianza de los residuos no cambia con el tiempo, lo cual es deseable en un modelo de series temporales efectivo.

***Ilustración 15 Residuos del modelo Sarima con transformación log***



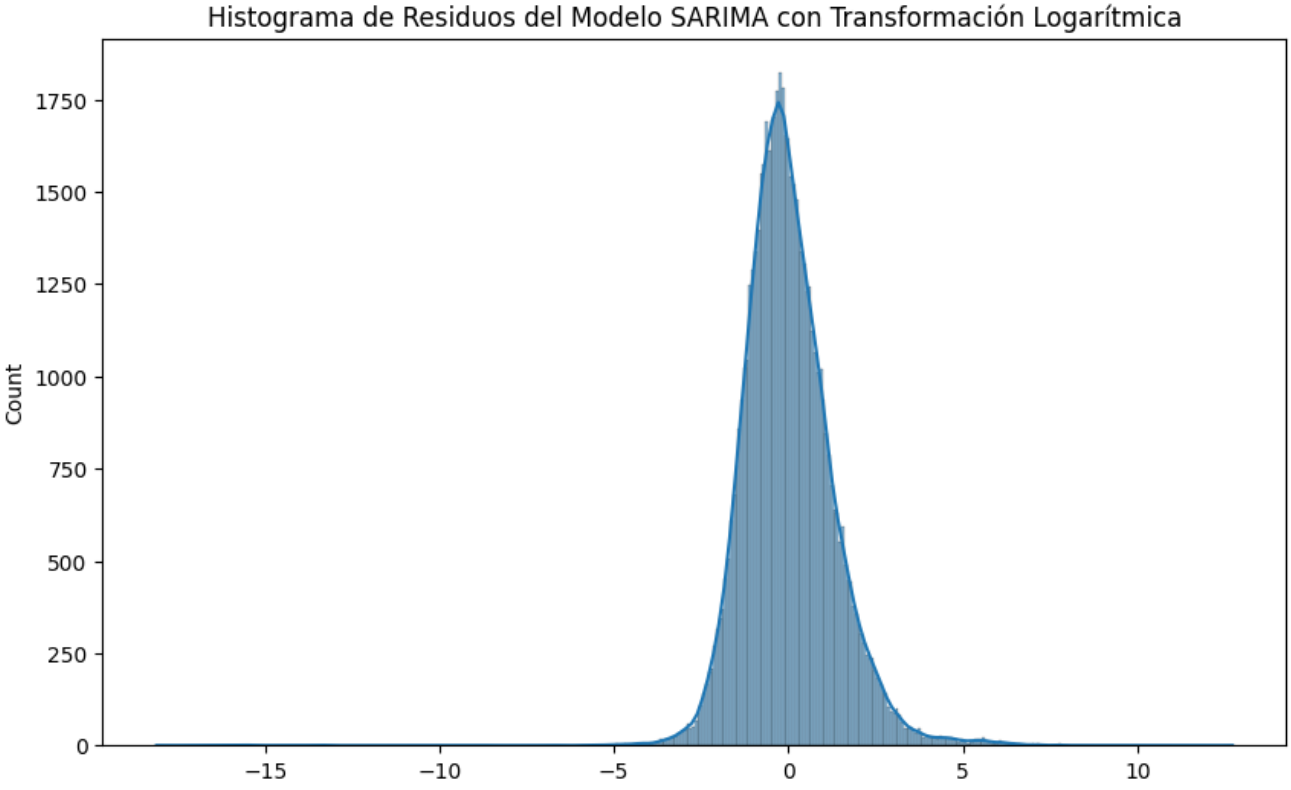

*Elaboración propia, Fuente Dian*

Sin embargo, la presencia de picos, como se observa en el gráfico, sugiere que podrían existir elementos anómalos o puntos de inflexión en los datos que no fueron completamente modelados. Estos valores atípicos pueden ser el resultado de shocks externos o eventos no anticipados que no están incorporados en el modelo. La identificación de estos puntos puede ser crucial para refinamientos futuros del modelo, ya que permiten entender mejor la dinámica subyacente que podría no estar siendo capturada por el modelo actual.

### 6.3.2.2.HISTOGRAMA DE RESIDUOS

El histograma de residuos del modelo SARIMA (ilustración 16), sugiere que estos siguen una distribución aproximadamente normal centrada en cero, pero con colas largas a ambos lados. Esto se confirma por la alta curtosis (16.36) y un sesgo cercano a cero (-0.31), aunque no exactamente normal (p-valor de Jarque-Bera es 0.0).

***Ilustración 16 Histogramas de residuos modelo SARIMA***

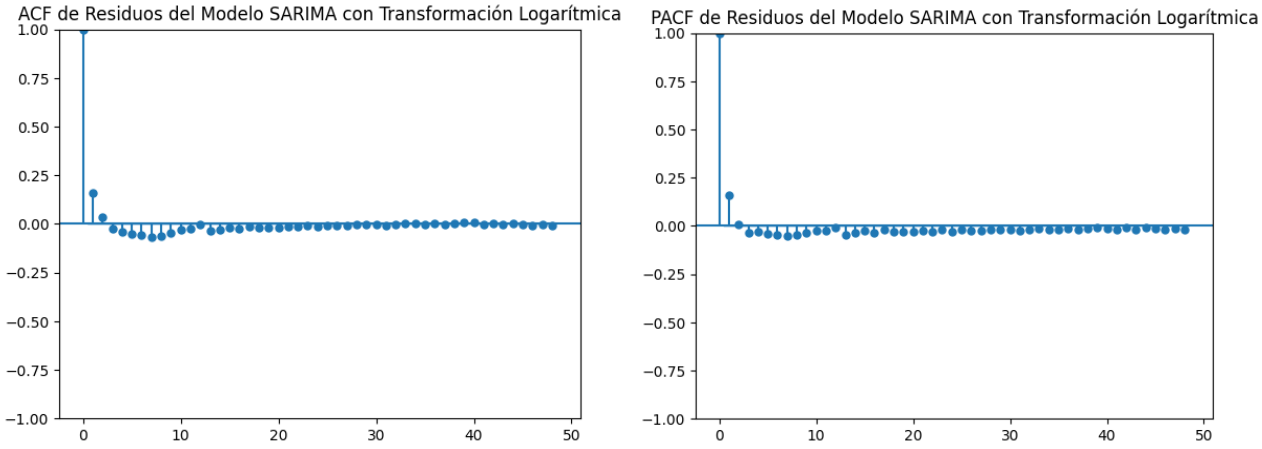



*Elaboración propia, Fuente Dian*

### *6.3.2.3.GRÁFICOS ACF Y PACF*

Los gráficos ACF y PACF (ilustración 17) muestran que no hay una autocorrelación significativa en la mayoría de los rezagos, aunque hay una pequeña autocorrelación en los primeros rezagos. Esto indica que la mayoría de las dependencias temporales han sido capturadas por el modelo, aunque no perfectamente.

*Ilustración 17 ACF y PACT Modelo SARIMA*




*Elaboración propia, Fuente Dian*

### 6.3.2.4. PRUEBAS DE DIAGNÓSTICO

- **Ljung-Box**: El valor p de 0.0 sugiere que existe autocorrelación significativa en los residuos.
- **Jarque-Bera**: El valor p de 0.0 indica que los residuos no siguen una distribución normal.
- **Heterocedasticidad**: El valor p de 0.2446 sugiere que no se puede rechazar la hipótesis nula de homocedasticidad, indicando varianza constante en los residuos.

El modelo SARIMA (0, 1, 1)x(1, 1, 1, 12) con transformación logarítmica ha mejorado algunos aspectos de la modelación de la serie temporal "VALOR_KILO", pero persisten problemas como la autocorrelación en los residuos y la no normalidad de su distribución. Aunque la heterocedasticidad no es un problema significativo, el modelo puede beneficiarse de ajustes adicionales o técnicas de preprocesamiento para mejorar su precisión. El RMSE del modelo transformado es 620.44, lo que indica que la precisión no ha mejorado significativamente en comparación con el modelo sin transformar.

## 6.4.MODELOS PROPHET

El modelo Prophet fue ajustado para la serie temporal "VALOR_KILO" y se configuró para incluir estacionalidades diarias y anuales. A continuación, se presenta una interpretación detallada basada en los resultados obtenidos.

#### *6.4.1.1.PRONÓSTICO DEL MODELO PROPHET*

El gráfico de pronóstico del modelo Prophet muestra las predicciones a lo largo del tiempo para la variable "VALOR_KILO". El modelo parece capturar adecuadamente la tendencia general de los datos, pero hay una notable dispersión y valores extremos en la parte más reciente de la serie temporal. Esto sugiere que, aunque Prophet maneja bien las tendencias y patrones estacionales, enfrenta dificultades con la predicción de picos extremos, lo que puede deberse a la naturaleza volátil de los datos.

#### 6.4.1.2.COMPONENTES DEL MODELO PROPHET

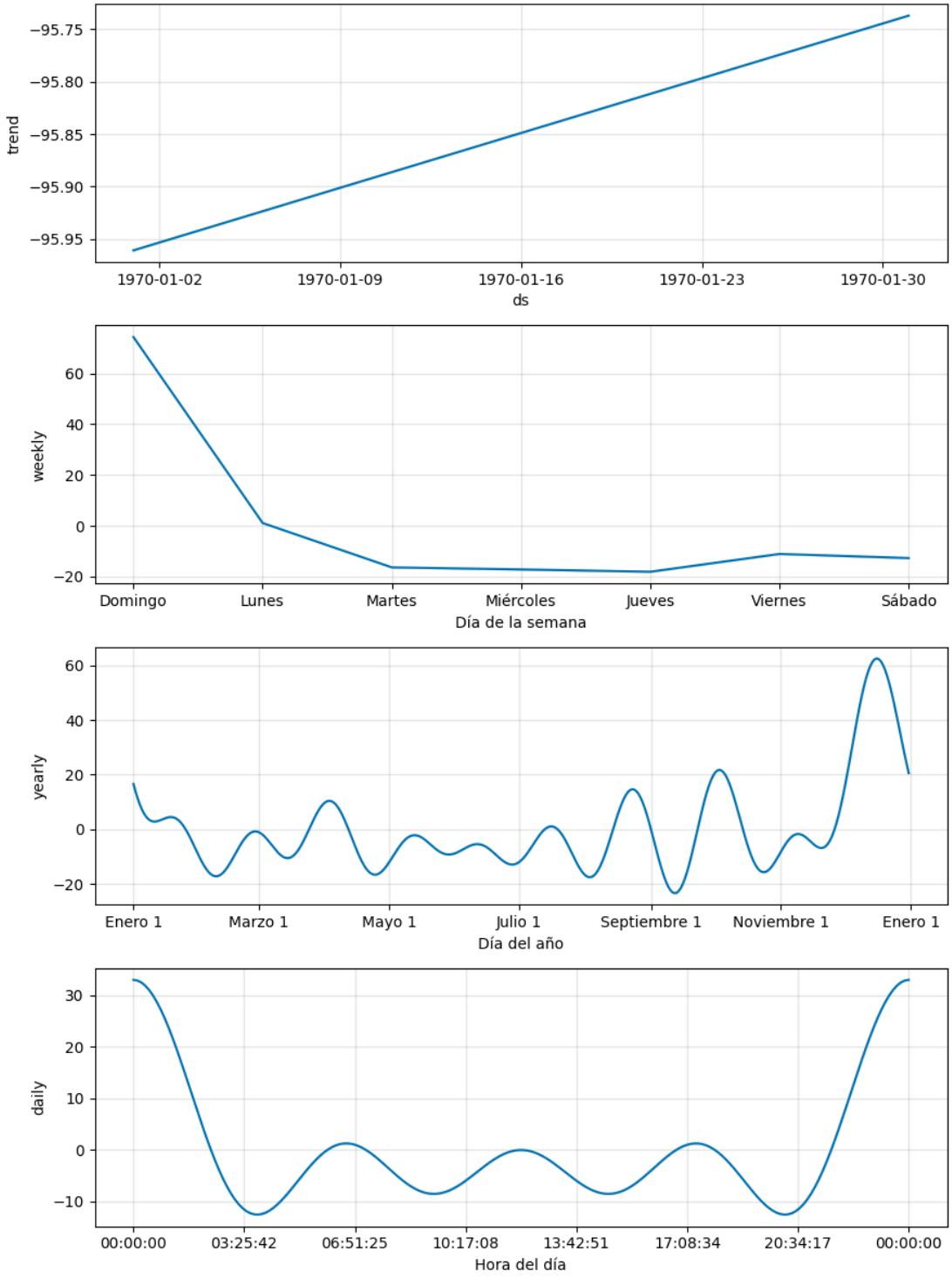
- **Tendencia**: La tendencia general es positiva, lo que indica un crecimiento a lo largo del tiempo. Esto es coherente con un aumento sostenido en los valores de "VALOR_KILO".
- **Estacionalidad Semanal**: El gráfico de estacionalidad semanal muestra variaciones significativas a lo largo de los días de la semana. Hay un pico notable los domingos, seguido de una caída los lunes y una estabilización hacia el final de la semana. Esto sugiere que hay un patrón semanal en los datos, posiblemente relacionado con actividades comerciales o de producción específicas de ciertos días de la semana.
- **Estacionalidad Anual**: La estacionalidad anual revela varios picos y valles a lo largo del año, con incrementos notables alrededor de enero y finales de diciembre, y caídas en los meses intermedios. Esto indica que hay patrones estacionales fuertes que podrían estar vinculados a factores estacionales, como festividades, cambios climáticos, o ciclos de demanda y oferta.
- **Estacionalidad Diaria**: La estacionalidad diaria muestra variaciones claras en los valores a lo largo del día, con picos alrededor de la medianoche y caídas durante la tarde. Esto puede reflejar comportamientos específicos del proceso de medición o actividades de producción y comercialización que fluctúan durante el día.

#### 6.4.1.3.EVALUACIÓN DEL MODELO

El RMSE obtenido fue de 552.08, lo que es comparable al RMSE del modelo ARIMA, sugiriendo que Prophet tiene un rendimiento similar en términos de precisión global. Sin embargo, al observar los gráficos y los componentes del modelo (ilustración 18), se hace evidente que Prophet captura bien las tendencias generales y los patrones estacionales, pero tiene dificultades con la predicción de valores extremos y picos. Estos problemas pueden ser abordados con técnicas adicionales de preprocesamiento de datos o mediante la incorporación de variables adicionales que ayuden a explicar mejor la variabilidad en los datos.

El modelo Prophet es efectivo para capturar las tendencias y patrones estacionales en la serie temporal "VALOR_KILO". No obstante, enfrenta desafíos con la predicción de valores extremos, lo que indica la necesidad de mejorar el modelado de la volatilidad inherente en los datos. Para futuras investigaciones, se recomienda explorar enfoques híbridos y técnicas avanzadas de preprocesamiento para mejorar la precisión en la predicción de eventos extremos.

*Ilustración 18 Evaluación modelo Prophet*



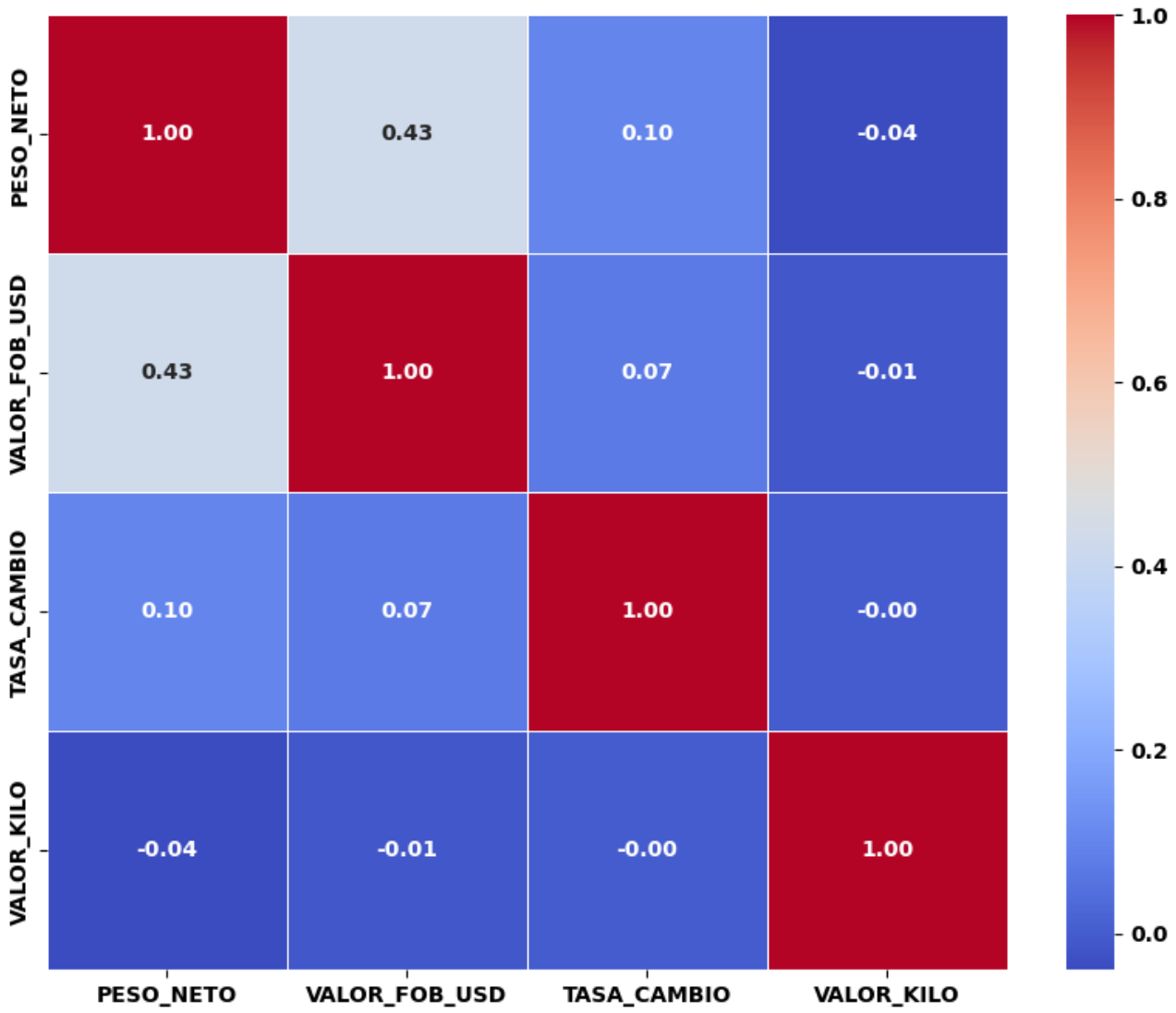

*Elaboración propia, Fuente Dian*

## 6.5.MODELOS AVANZADOS DE MACHINE LEARNING

### 6.5.1. RANDOM FOREST

Se desarrolló un modelo predictivo para pronosticar el valor por kilo de importación de pesticidas utilizando un algoritmo de Random Forest. Inicialmente, se llevó a cabo un análisis exploratorio de los datos para identificar las relaciones entre las variables (ilustración 19). El análisis de correlación mostró una baja correlación entre las características seleccionadas y la variable objetivo, `VALOR_KILO`, lo que sugirió que las características disponibles podrían no ser suficientemente informativas para predecir la variable objetivo con alta precisión.

***Ilustración 19 Matriz de correlación Random forest***




*Elaboración propia, Fuente Dian*

El conjunto de datos fue dividido en dos subconjuntos: 80% de los datos se utilizaron para el entrenamiento del modelo y el 20% restante para la validación. Esta partición permitió evaluar la capacidad del modelo para generalizar a datos no vistos durante el entrenamiento. El modelo de Random Forest fue entrenado con las características seleccionadas y evaluado utilizando diversas métricas de desempeño.

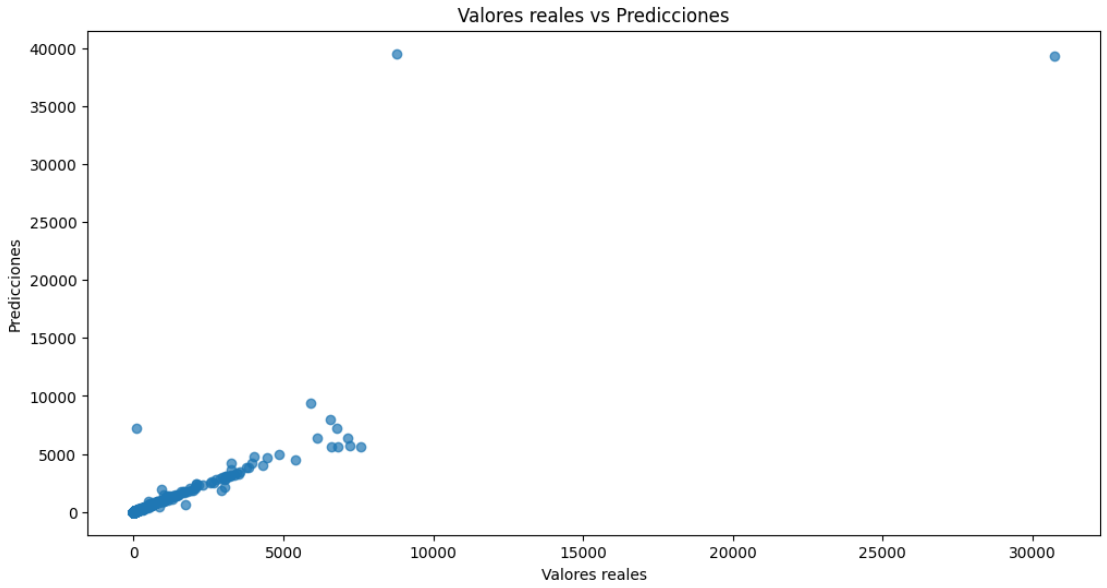
Los resultados mostraron que el modelo tenía un error cuadrático medio (RMSE) de 150.13 en el conjunto de entrenamiento y un RMSE de 302.10 en el conjunto de prueba. El coeficiente de determinación ($R^2$) en el conjunto de prueba fue de 0.417, indicando que el modelo explica aproximadamente el 41.7% de la variabilidad en el valor por kilo de importación de pesticidas. Adicionalmente, se calcularon el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE) para proporcionar una visión más completa del desempeño del modelo.

**Resultados del Primer Modelo**:

- **RMSE del conjunto de entrenamiento**: 150.125
- **RMSE del conjunto de prueba**: 302.10
- **$R^2$ del conjunto de entrenamiento**: 0.948
- **$R^2$ del conjunto de prueba**: 0.417

La visualización de las predicciones frente a los valores reales mostró que el modelo tenía dificultades para predecir valores extremos, lo que sugiere que las características actuales no capturan completamente la variabilidad en los datos. La gráfica de dispersión (ilustración 20) indicó una tendencia general adecuada en las predicciones, pero con una notable dispersión para valores altos.

***Ilustración 20 Modelo Random Forest***




*Elaboración propia, Fuente Dian*

En conclusión, aunque el modelo de Random Forest desarrollado mostró un desempeño aceptable para ser el primer intento en pronosticar el valor por kilo de importación de pesticidas, los resultados sugieren que la inclusión de características adicionales y el refinamiento en la ingeniería de características podrían mejorar aún más la precisión de las predicciones.

#### 6.5.1.1.MODELOS RANDOM FOREST PRIMER AJUSTE

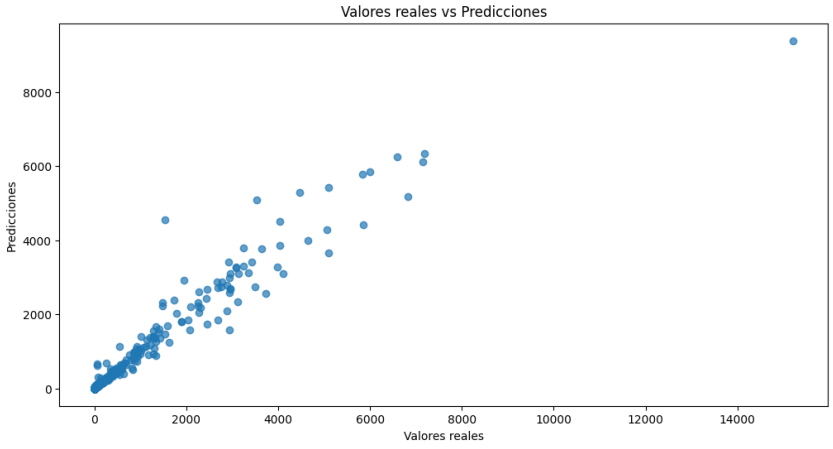
Para mejorar el desempeño, se realizó un segundo intento con ajustes adicionales, incluyendo la eliminación de valores atípicos de las características PESO_NETO y VALOR_FOB_USD, para mejorar la calidad de los datos. El conjunto de datos fue nuevamente dividido en 80% para entrenamiento y 20% para validación. El modelo de Random Forest fue optimizado utilizando Grid Search y los mejores hiperparámetros encontrados fueron: max_depth=None, min_samples_leaf=2, min_samples_split=10, y n_estimators=50.

**Resultados del Segundo Modelo**:

- **RMSE del conjunto de entrenamiento**: 340.647
- **RMSE del conjunto de prueba**: 78.49
- **$R^2$ del conjunto de entrenamiento**: 0.756
- **$R^2$ del conjunto de prueba**: 0.935

Los resultados del segundo modelo mostraron una mejora significativa. El error cuadrático medio (RMSE) fue de 340.65 en el conjunto de entrenamiento y 78.49 en el conjunto de prueba, indicando una alta precisión en la predicción de los valores de prueba. El coeficiente de determinación ($R^2$) fue de 0.756 en el conjunto de entrenamiento y de 0.935 en el conjunto de prueba, lo que sugiere que el modelo explica el 93.5% de la variabilidad en el valor por kilo de importación de pesticidas en los datos de prueba (ilustración 21).

**Ilustración 21 Random forest primer ajuste**




*Elaboración propia, Fuente Dian 2*

#### 6.5.1.2.TERCER MODELO (RANDOM FOREST AJUSTADO)

En el tercer intento, se utilizó Randomized Search para una búsqueda más eficiente de los hiperparámetros. Los valores atípicos fueron eliminados de manera similar al segundo modelo. Los mejores hiperparámetros encontrados fueron: n_estimators=118, max_depth=30, min_samples_leaf=2, y min_samples_split=7.

**Resultados del Tercer Modelo**:

- **RMSE del conjunto de entrenamiento**: 98.53
- **RMSE del conjunto de prueba**: 72.10
- **$R^2$ del conjunto de entrenamiento**: 0.982
- **$R^2$ del conjunto de prueba**: 0.946

#### 6.5.1.3. CUARTO MODELO (RANDOM FOREST AJUSTADO CON RANDOMIZED SEARCH)

En este intento, se continuó con el ajuste de hiperparámetros mediante Randomized Search, eliminando valores atípicos y optimizando la selección de hiperparámetros. Los mejores hiperparámetros encontrados fueron: n_estimators=83, max_depth=None, min_samples_leaf=2, y min_samples_split=6.

**Resultados del Cuarto Modelo**:

- **RMSE del conjunto de entrenamiento**: 372.19
- **RMSE del conjunto de prueba**: 88.65
- **$R^2$ del conjunto de entrenamiento**: 0.709
- **$R^2$ del conjunto de prueba**: 0.917

El tercer modelo mostró la mejor precisión, con un RMSE de 72.10 en el conjunto de prueba y un $R^2$ de 0.946, indicando que explica el 94.6% de la variabilidad en el valor por kilo de importación de pesticidas. El cuarto modelo también mostró un buen desempeño con un RMSE de 88.65 en el conjunto de prueba y un $R^2$ de 0.917.El segundo modelo mostró una mejora significativa con un RMSE de 78.49 en el conjunto de prueba y un $R^2$ de 0.935.El primer modelo, aunque aceptable, tuvo un RMSE de 269.48 en el conjunto de prueba y un $R^2$ de 0.544.

La utilización de Randomized Search en el tercer y cuarto modelo permitió encontrar configuraciones óptimas de hiperparámetros de manera más eficiente, resultando en un mejor rendimiento general comparado con los modelos optimizados mediante Grid Search. La eliminación de valores atípicos en el segundo, tercer y cuarto modelos contribuyó significativamente a la mejora en el rendimiento del modelo.

El tercer modelo, optimizado mediante Randomized Search y con eliminación de valores atípicos, mostró el mejor desempeño en la predicción del valor por kilo de importación de pesticidas. Sin embargo, el cuarto modelo también demostró ser bastante robusto y eficiente. Este estudio destaca la importancia de la limpieza de datos y la optimización de hiperparámetros en el desarrollo de sistemas predictivos robustos. Los resultados sugieren que la inclusión de características adicionales y el refinamiento en la ingeniería de características podrían mejorar aún más la precisión de las predicciones, estableciendo una base sólida para futuras investigaciones en la predicción del valor de importación de pesticidas.

- **Mejores hiperparámetros**: {'n_estimators': 83, 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 6}

#### 6.5.1.4.EVALUACIÓN Y VALIDACIÓN DEL MODELO SELECCIONADO

Para determinar si el cuarto modelo es adecuado para pronosticar el valor por kilo de importación de pesticidas en Colombia, se llevaron a cabo varios pasos de validación y evaluación.

- **Validación Cruzada:** En la metodología de este trabajo, se empleó la técnica de validación cruzada con K-fold (n=10) para evaluar la precisión y robustez del modelo predictivo desarrollado. La validación cruzada se realizó utilizando el 80% de los datos del conjunto total, reservados específicamente para el entrenamiento del modelo. El propósito de utilizar la validación cruzada en esta etapa era asegurar que el modelo predictivo no solo se ajustara bien a un subconjunto de datos, sino que también mantuviera un alto nivel de precisión y generalización cuando se enfrentara a nuevos datos.

  La elección de K-fold con n=10 implicó dividir el conjunto de datos de entrenamiento en 10 subconjuntos (o 'folds'). El modelo se entrenó 10 veces, cada vez utilizando 9 de estos subconjuntos como datos de entrenamiento y el subconjunto restante como datos de prueba. Esto permite evaluar la estabilidad y la variabilidad del modelo, proporcionando una estimación más fiable de su desempeño en diferentes muestras de datos.

  El resultado de la validación cruzada arrojó un RMSE promedio de 305.88, lo cual indica el error típico de las predicciones del modelo en comparación con los valores reales. La desviación estándar de 428.24, obtenida de los resultados del RMSE de cada iteración, refleja la variabilidad en el desempeño del modelo a través de las diferentes divisiones de los datos. Este alto valor de desviación estándar sugiere que el rendimiento del modelo puede variar significativamente dependiendo del subconjunto de datos específico utilizado para la prueba, lo que indica áreas potenciales para futuros ajustes y mejoras del modelo.

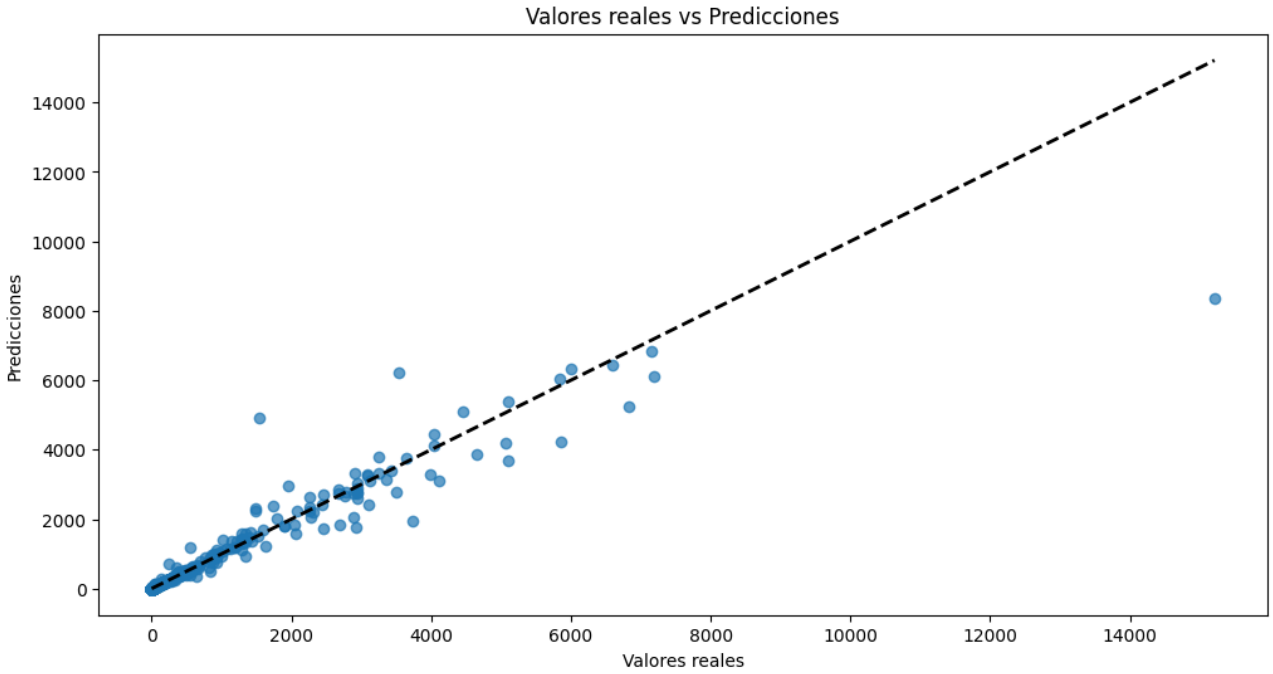
  La validación cruzada es, por lo tanto, una herramienta esencial en nuestro estudio para confirmar que el modelo es robusto y generalizable, no sólo ajustándose a los datos con los que se entrenó, sino también adaptándose bien a nuevos datos, lo cual es crítico para la aplicación práctica de nuestro modelo predictivo en el sector agrícola.

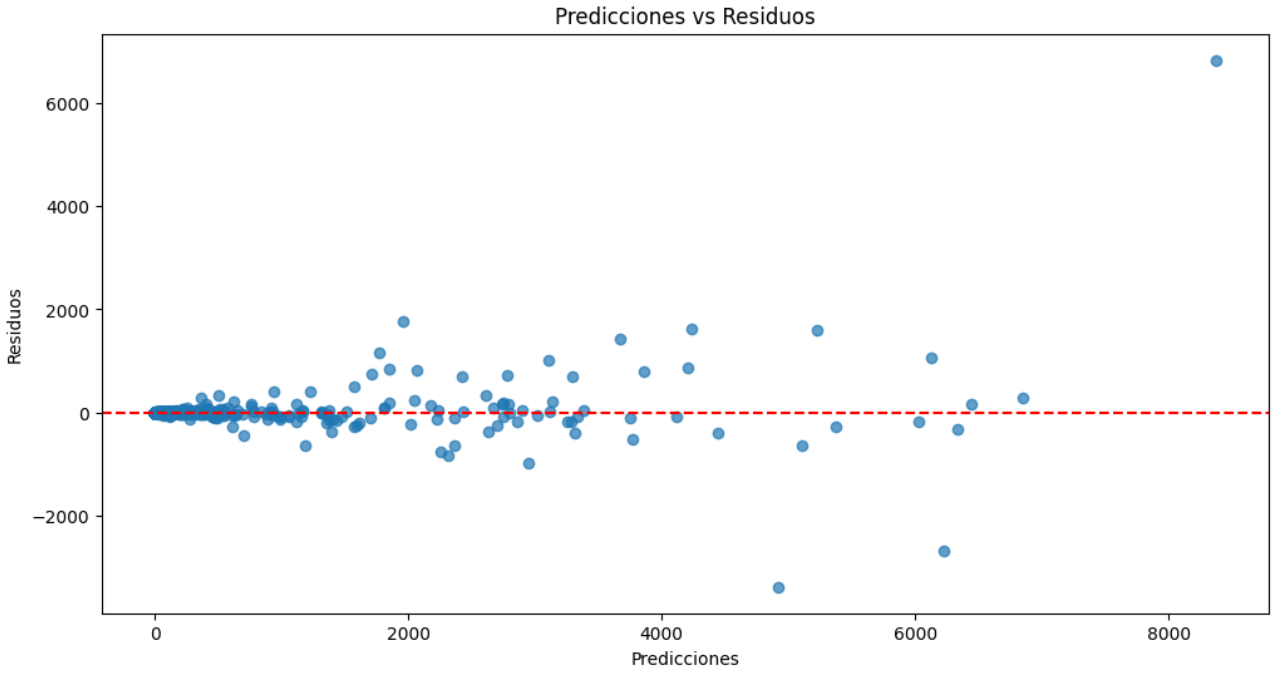
- **Evaluación de Métricas Adicionales:** En el conjunto de prueba, se obtuvo un error absoluto medio (MAE) de 4.58 y un error porcentual absoluto medio (MAPE) extremadamente alto, indicando problemas significativos en la precisión relativa del modelo para algunos puntos de datos.
- **Gráficos de Diagnóstico:** En nuestro análisis, la evaluación del modelo predictivo incluyó una comparación visual de los valores predichos frente a los valores reales utilizando los datos de testeo, que corresponden al 20% del conjunto de datos total. Esta segmentación asegura que la evaluación del modelo se realiza en una muestra independiente, no vista durante el entrenamiento, para proporcionar una evaluación objetiva de su capacidad predictiva en condiciones de "nueva" información.

  La gráfica de valores predichos versus valores reales (ilustración 22) es una herramienta visual crítica utilizada para examinar la precisión del modelo. En esta gráfica, la línea de igualdad (y=x) sirve como referencia ideal donde las predicciones perfectas se ubicarían. Observamos que la tendencia general de los datos predichos se alinea bien con esta línea de referencia, indicando que el modelo tiene una buena capacidad general de predicción. Sin embargo, se notó cierta dispersión en los valores más altos, sugiriendo que el modelo podría estar menos preciso en el rango superior de predicciones. Este patrón es crucial para identificar y entender las limitaciones del modelo, especialmente en su aplicación a eventos o valores extremos, donde las decisiones basadas en estas predicciones pueden tener implicaciones significativas.

  Esta evaluación visual, complementada con métricas estadísticas como el RMSE, proporciona una base sólida para interpretar la eficacia del modelo y para guiar futuras mejoras. La inclusión de los datos de testeo en esta fase es esencial para asegurar que la validación del modelo sea rigurosa y representativa del desempeño que se puede esperar en aplicaciones prácticas reales.

***Ilustración 22 Predicciones modelo final***




*Elaboración propia, Fuente Dian*

El RMSE de 88.65 en el conjunto de prueba y el $R^2$ de 0.917 indican un buen ajuste del modelo, aunque hay margen de mejora en la precisión, especialmente en los valores altos. La alta desviación estándar en el RMSE y el $R^2$ en la validación cruzada sugiere que el modelo puede no ser completamente estable y podría beneficiarse de más ajustes o de la incorporación de características adicionales. El alto MAPE resalta que el modelo tiene dificultades para predecir con precisión relativa, lo que podría ser abordado mediante una mejor ingeniería de características o técnicas avanzadas de manejo de outliers.

El cuarto modelo de Random Forest, optimizado mediante Randomized Search y con eliminación de valores atípicos, muestra un buen desempeño en la predicción del valor por kilo de importación de pesticidas en Colombia. Sin embargo, el análisis sugiere que hay margen de mejora, especialmente en la estabilidad del modelo y la precisión relativa de las predicciones. Este estudio destaca la importancia de la limpieza de datos, la optimización de hiperparámetros y la evaluación exhaustiva del modelo en el desarrollo de sistemas predictivos robustos.

- **Resultados del Modelo final:**

- Mejores hiperparámetros: {'n_estimators': 83, 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 6}
- RMSE del conjunto de entrenamiento: 372.19
- RMSE del conjunto de prueba: 88.65
- $R^2$ del conjunto de entrenamiento: 0.709
- $R^2$ del conjunto de prueba: 0.917
- CV RMSE: 305.88 ± 428.24
- CV $R^2$: 0.609 ± 0.674
- Test MAE: 4.58
- Test MAPE: 232554579282.13

### 6.5.2. XGBOOST

También se implementó un modelo de XGBoost para pronosticar el precio de importación por kilo (VALOR_KILO) de pesticidas en Colombia, utilizando diversas características categóricas y temporales. A pesar de un proceso exhaustivo de preprocesamiento de datos, que incluyó la eliminación de duplicados, la codificación de variables categóricas y la creación de características basadas en la fecha y retrasos (lags), los resultados obtenidos indican una discrepancia significativa entre los valores predichos y los valores reales.

Las métricas de error calculadas(tabla 8), a saber, MAE, MSE, RMSE y MAPE, revelan un rendimiento subóptimo del modelo. El MAE fue de 18.50, lo que indica que, en promedio, las predicciones se desvían 18.50 unidades del valor real. El MSE de 4582.84 y el RMSE de 67.70 sugieren que hay variaciones considerables en los errores, reflejando la presencia de errores grandes ocasionales. Especialmente destacable es el MAPE de 565.10%, que revela errores porcentuales extremadamente altos en comparación con los valores reales. Esto sugiere que el modelo tiene dificultades para manejar la variabilidad y los valores extremos presentes en los datos.

*Tabla 8 MODELO XGBOOST*

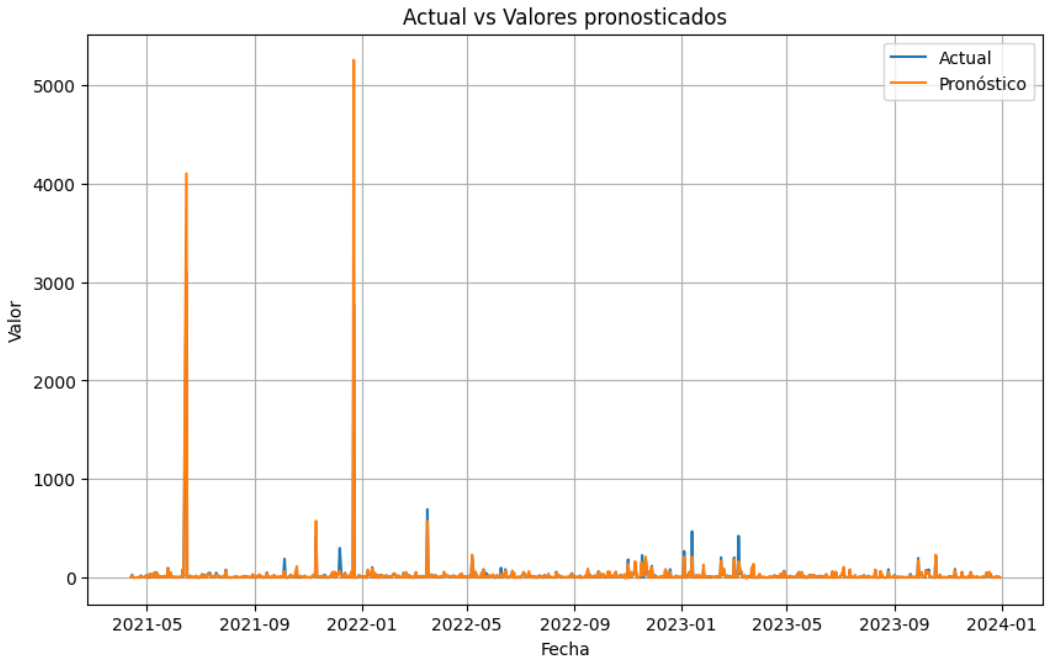
| Métrica | Valor |
|---|---|
| MAE | 18.50 |
| MSE | 4582,84 |
| RMSE | 67.69 |
| MAPE | 565.098% |

*Elaboración propia, Fuente Dian*

### 6.5.2.1. VISUALIZACIÓN DE VALORES REALES VS PREDICCIONES

La visualización de los resultados mediante una gráfica de valores reales versus predichos (Ilustración 23) refuerza esta observación, mostrando discrepancias notables en los picos más altos. Estos resultados subrayan la necesidad de mejorar tanto la ingeniería de características como el ajuste del modelo, posiblemente explorando métodos adicionales de preprocesamiento y alternativas algorítmicas para lograr una mayor precisión en las predicciones del precio de importación por kilo de pesticidas en Colombia.

*Ilustración 23 Modelo xgboost*




*Elaboración propia, Fuente Dian*

#### 6.5.2.2.MODELO XGBOOST CON AJUSTE DE HIPERPARÁMETROS

En un esfuerzo por mejorar la precisión del modelo de predicción del precio de importación por kilo (VALOR_KILO) de pesticidas en Colombia, se implementó un ajuste de Hiperparámetros utilizando GridSearchCV con un modelo XGBoost. A través de un proceso de búsqueda exhaustiva, se evaluaron diferentes combinaciones de parámetros, incluyendo el número de estimadores, la profundidad máxima, la tasa de aprendizaje, la fracción de muestreo y la fracción de columnas por árbol.

Los mejores Hiperparámetros encontrados mediante GridSearchCV fueron los siguientes:

- **colsample_bytree**: 1.0
- **learning_rate**: 0.01
- **max_depth**: 6
- **n_estimators**: 1000
- **subsample**: 1.0

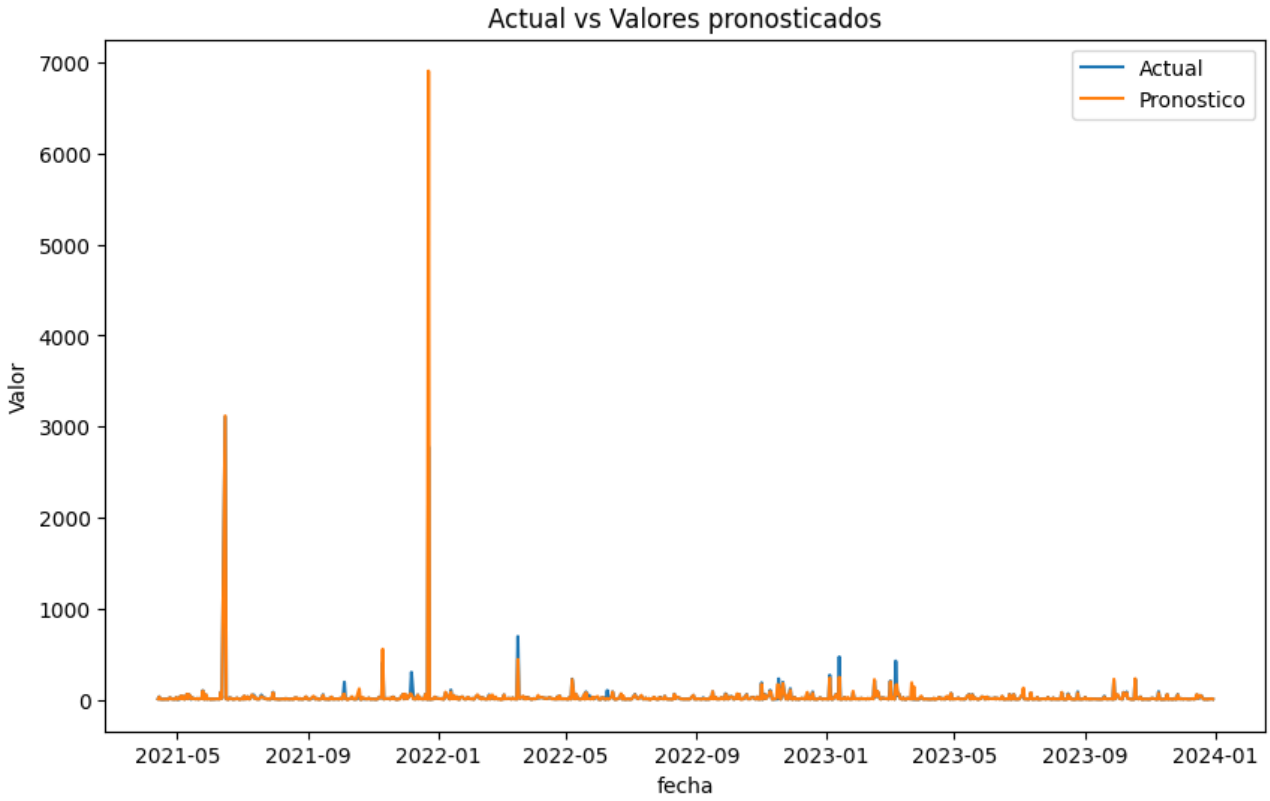
Utilizando el mejor modelo encontrado, se realizaron predicciones sobre el conjunto de prueba. Las métricas de error calculadas son las siguientes:

- **MAE (Error absoluto medio)**: Este valor mide el error medio entre las predicciones y los valores reales. Un MAE de 9.34 indica que, en promedio, las predicciones del modelo se desvían en 9.34 unidades del valor real.
- **MSE (Error cuadrático medio)**: Esta métrica proporciona una medida del error cuadrático medio entre las predicciones y los valores reales. Un MSE de 21257.29 sugiere la presencia de errores grandes ocasionales.
- **RMSE (Raíz del error cuadrático medio)**: Esta métrica es la raíz cuadrada del MSE y ofrece una interpretación más directa de los errores en las mismas unidades que los datos originales. Un RMSE de 145.80 indica la magnitud típica del error de predicción.
- **MAPE (Error porcentual absoluto medio)**: El MAPE mide el error porcentual medio entre las predicciones y los valores reales. Un MAPE de 83.92% sugiere que, en promedio, las predicciones tienen un error porcentual significativo respecto a los valores reales.

#### 6.5.2.3.GRÁFICA DE RESULTADOS

La gráfica resultante (Ilustración 24) compara los valores reales con los valores predichos a lo largo del tiempo, mostrando una mejora en la capacidad del modelo para seguir las tendencias generales del precio de importación por kilo de pesticidas. Sin embargo, aún se observan algunas discrepancias en los picos más altos, lo que indica que el modelo sigue teniendo dificultades para capturar los valores extremos con precisión.

*Ilustración 24 Modelo XGboost, ajuste de hiperparámetros*




*Elaboración propia, Fuente Dian*

El ajuste de hiperparámetros utilizando GridSearchCV ha permitido encontrar una configuración óptima para el modelo XGBoost, mejorando así su rendimiento en la predicción del precio de importación por kilo de pesticidas en Colombia. A pesar de esta mejora, las métricas de error y la visualización de los resultados indican que existe un margen para seguir refinando el modelo, especialmente en lo que respecta a la predicción de valores extremos.

#### 6.5.2.4.AJUSTE DEL MODELO MEDIANTE BÚSQUEDA ALEATORIA DE HIPERPARÁMETROS

En una búsqueda por encontrar una mejor precisión del modelo de predicción del precio de importación por kilo (VALOR_KILO) de pesticidas en Colombia, se llevó a cabo una búsqueda aleatoria de hiperparámetros utilizando RandomizedSearchCV con un modelo XGBoost. Este enfoque permite explorar un espacio de hiperparámetros más amplio y encontrar configuraciones óptimas de manera eficiente.

La búsqueda aleatoria incluyó la consideración de varios hiperparámetros críticos para el rendimiento del modelo XGBoost, como el número de estimadores, la profundidad máxima, la tasa de aprendizaje, la fracción de muestreo y la fracción de columnas por árbol. Específicamente, los hiperparámetros

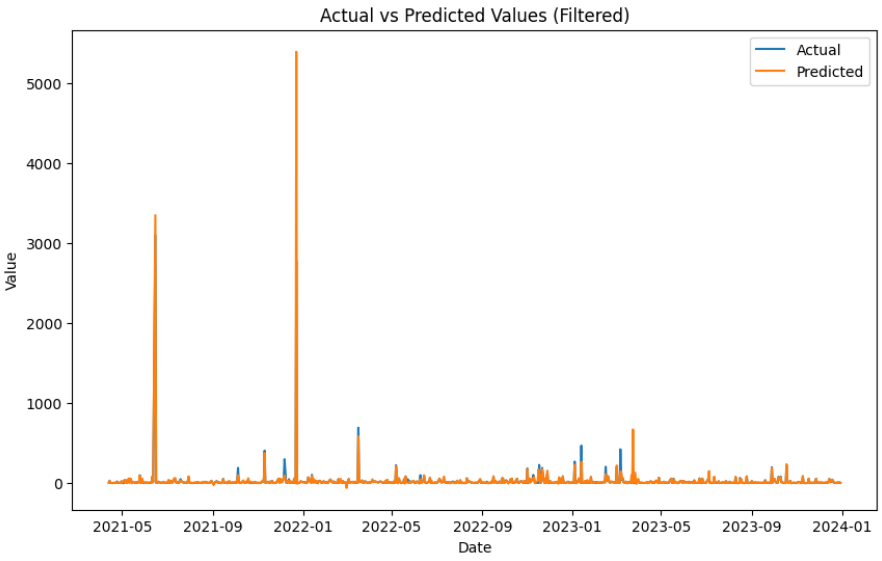
evaluados fueron: número de estimadores (100, 500, 1000, 1500), profundidad máxima (3, 4, 5, 6, 7, 8), tasa de aprendizaje (0.001, 0.01, 0.05, 0.1, 0.2), fracción de muestreo (0.6, 0.7, 0.8, 0.9, 1.0) y fracción de columnas por árbol (0.6, 0.7, 0.8, 0.9, 1.0).

La configuración de RandomizedSearchCV se realizó con 100 iteraciones y validación cruzada de 3 pliegues, lo que permitió evaluar múltiples combinaciones de hiperparámetros de manera eficiente. Tras la búsqueda, se seleccionó el mejor conjunto de hiperparámetros, que fueron: subsample de 0.9, n_estimators de 500, max_depth de 3, learning_rate de 0.05 y colsample_bytree de 1.0. Este modelo óptimo se utilizó para hacer predicciones en el conjunto de prueba.

Las métricas de error calculadas para el mejor modelo encontrado indican una mejora significativa en comparación con los resultados anteriores. El MAE fue de 10.63, lo que sugiere que, en promedio, las predicciones del modelo se desvían en 10.63 unidades del valor real. El MSE de 10375.75 y el RMSE de 101.86 reflejan una disminución en la magnitud de los errores cuadrados y su raíz cuadrada, respectivamente. Sin embargo, el MAPE de 741.55% es extremadamente alto, indicando que, aunque el modelo ha mejorado en algunos aspectos, aún enfrenta dificultades significativas para manejar la variabilidad y los valores extremos en los datos.

La visualización de los resultados mediante una gráfica comparativa de los valores reales y predichos a lo largo del tiempo (ilustración 25) muestra una mejora en la capacidad del modelo para seguir las tendencias generales del precio de importación por kilo de pesticidas. No obstante, persisten algunas discrepancias, especialmente en los picos más altos, lo que sugiere que el modelo sigue teniendo dificultades para capturar los valores extremos con precisión.

***Ilustración 25 Ajuste de Hiperparámetros Aleatoria***




*Elaboración propia, Fuente Dian*

En conclusión, el ajuste de hiperparámetros mediante RandomizedSearchCV ha permitido encontrar una configuración óptima para el modelo XGBoost, mejorando así su rendimiento en la predicción del precio de importación por kilo de pesticidas en Colombia. A pesar de estas mejoras, las métricas de error y la visualización de los resultados indican que existe un margen para seguir refinando el modelo, especialmente en lo que respecta a la predicción de valores extremos. Futuras investigaciones podrían centrarse en la ingeniería de características adicionales, el manejo de outliers y la exploración de modelos alternativos para mejorar aún más la precisión de las predicciones. Con estas mejoras, se espera lograr una mayor precisión en las predicciones del precio de importación, proporcionando insights valiosos sobre las tendencias y variabilidades en el mercado de pesticidas en Colombia.

#### 6.5.2.5.VALIDACIÓN CRUZADA DEL MODELO AJUSTADO

Para evaluar la robustez y la capacidad de generalización del modelo ajustado, se realizó una validación cruzada utilizando el mejor conjunto de hiperparámetros identificados previamente. La validación cruzada es esencial para asegurar que el modelo no esté sobre ajustado a un subconjunto específico de datos y que funcione bien en datos no vistos.

En el desarrollo del modelo XGBoost, se seleccionaron los siguientes valores de hiperparámetros después de un exhaustivo proceso de prueba y error y validación cruzada: subsample de 0.9, n_estimators de 1500, max_depth de 4, learning_rate de 0.05 y colsample_bytree de 1.0. Aunque estos valores proporcionaron los mejores resultados dentro de las pruebas realizadas, es importante señalar que no necesariamente representan una configuración "óptima" universal, ya que los resultados pueden variar con diferentes conjuntos de datos o cambios en las condiciones del entorno de prueba. Estos parámetros fueron los que maximizaron la precisión del modelo bajo las condiciones específicas y con los datos disponibles durante nuestra experimentación.

Se llevó a cabo una validación cruzada de 5 pliegues (cv=5) utilizando la función cross_val_score de sklearn.model_selection. Este proceso divide el conjunto de datos de entrenamiento en 5 subconjuntos, utilizando cada uno como conjunto de validación mientras se entrena el modelo en los otros 4. Este proceso se repite cinco veces para asegurar que cada subconjunto sea utilizado como conjunto de validación una vez, proporcionando una evaluación completa del rendimiento del modelo.

Los resultados de la validación cruzada mostraron un MAE promedio de 15.85, lo que sugiere que, en promedio, las predicciones del modelo se desvían en 15.85 unidades del valor real. Además, la desviación estándar del MAE fue de 3.52, indicando cierta variabilidad en el error absoluto medio a través de los diferentes pliegues de la validación cruzada. Esta variabilidad refleja la heterogeneidad inherente en los datos de entrenamiento.

En conclusión, la validación cruzada ha demostrado que el modelo ajustado con los hiperparámetros seleccionados tiene una capacidad razonable de generalización en diferentes subconjuntos de datos. Sin embargo, los resultados también indican que hay margen para mejorar la precisión del modelo. Como la implementación de técnicas de manejo de outliers e ingeniería de características adicionales. Estos esfuerzos pueden ayudar a reducir el error y mejorar la fiabilidad de las predicciones del precio

de importación por kilo de pesticidas en Colombia, contribuyendo así a una mayor comprensión y precisión en la modelación de estos datos.

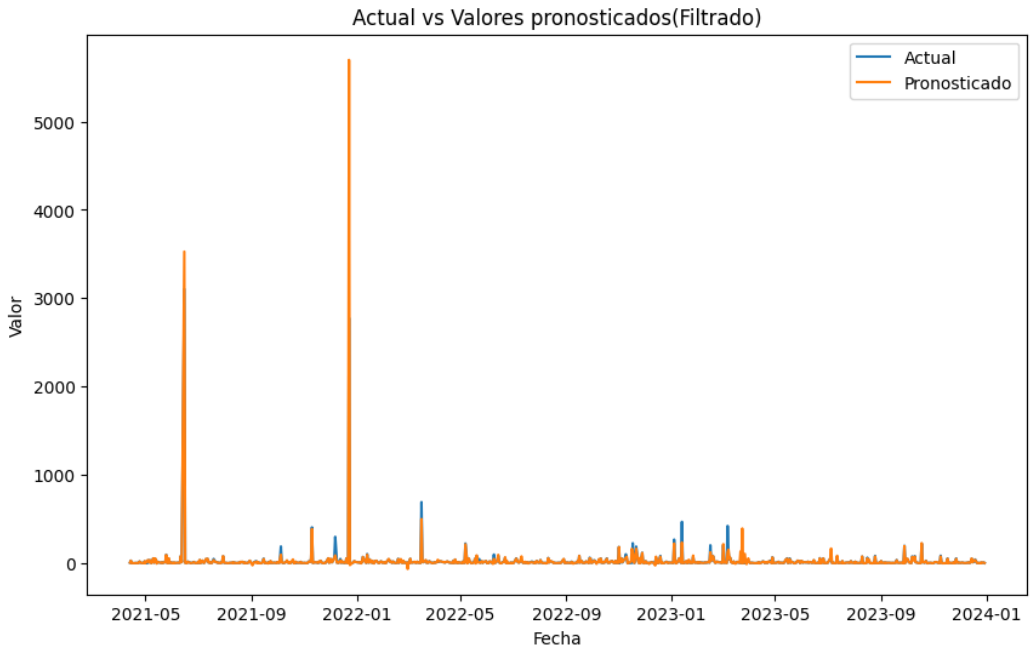
### 6.5.2.6.FILTRADO DE VALORES ATÍPICOS Y REAJUSTE DEL MODELO

Para mejorar la precisión del modelo y reducir el impacto de valores atípicos, se llevó a cabo un proceso de identificación y filtrado de estos valores en el conjunto de datos de entrenamiento. Los valores atípicos, identificados previamente, fueron eliminados del conjunto de datos de entrenamiento para minimizar su influencia en el ajuste del modelo.

Una vez filtrados los valores atípicos, se procedió a reajustar el modelo XGBoost utilizando los mejores hiperparámetros previamente identificados: subsample de 0.9, n_estimators de 1500, max_depth de 4, learning_rate de 0.05 y colsample_bytree de 1.0. El modelo fue entrenado con el conjunto de datos de entrenamiento filtrado, asegurando que los valores extremos no distorsionaran las predicciones del modelo.

Posteriormente, se realizaron predicciones sobre el conjunto de datos de prueba y se evaluaron las métricas de error para determinar el rendimiento del modelo ajustado. Las métricas de error obtenidas fueron: un MAE (Error absoluto medio) de 8.45, un MSE (Error cuadrático medio) de 9340.68, un RMSE (Raíz del error cuadrático medio) de 96.65 y un MAPE (Error porcentual absoluto medio) de 451.07%. Estas métricas indican una mejora significativa en comparación con las métricas obtenidas antes del filtrado de los valores atípicos.

**Ilustración 26 modelo con manejo de valores atípicos**




*Elaboración propia, Fuente Dian*

La visualización de los resultados mediante una gráfica comparativa de los valores reales y predichos (ilustración 26) mostró una mejora en la capacidad del modelo para seguir las tendencias generales del precio de importación por kilo de pesticidas. No obstante, aún se observan discrepancias en los picos más altos, lo que sugiere que el modelo sigue enfrentando desafíos en la predicción de valores extremos. Estos resultados subrayan la efectividad del filtrado de valores atípicos para mejorar la precisión del modelo, aunque también indican la necesidad de seguir refinando el modelo para abordar mejor la predicción de valores extremos.

#### 6.5.2.7.TRANSFORMACIÓN LOGARÍTMICA

Para mejorar la precisión del modelo de predicción del precio de importación por kilo (VALOR_KILO) de pesticidas en Colombia, se aplicó una transformación logarítmica a los datos y se eliminaron los valores atípicos del conjunto de datos de entrenamiento. Primero, se transformó la variable VALOR_KILO utilizando una transformación logarítmica, obteniendo así VALOR_KILO_LOG, lo cual ayudó a manejar la distribución sesgada y estabilizar la variabilidad de los datos. Posteriormente, se separaron las características y las etiquetas, y los datos fueron divididos en conjuntos de entrenamiento y prueba en una proporción de 80% y 20%, respectivamente.

Se eliminaron los valores atípicos del conjunto de datos de entrenamiento utilizando los índices previamente identificados. Con los datos filtrados, se configuró y ajustó un modelo XGBoost utilizando los mejores hiperparámetros identificados: subsample de 0.9, n_estimators de 1500, max_depth de 4, learning_rate de 0.05 y colsample_bytree de 1.0. Las predicciones se realizaron sobre el conjunto de datos de prueba, y se invirtió la transformación logarítmica para obtener los valores predichos y reales en su escala original. Finalmente, se calcularon las métricas de error para evaluar el rendimiento del modelo: MAE, MSE, RMSE y MAPE.

Los resultados obtenidos después de aplicar la transformación logarítmica y eliminar los valores atípicos indican una mejora significativa en la precisión del modelo. Las métricas de error obtenidas fueron: un MAE de 15.17, un MSE de 35512.92, un RMSE de 188.45 y un MAPE de 8.21%.

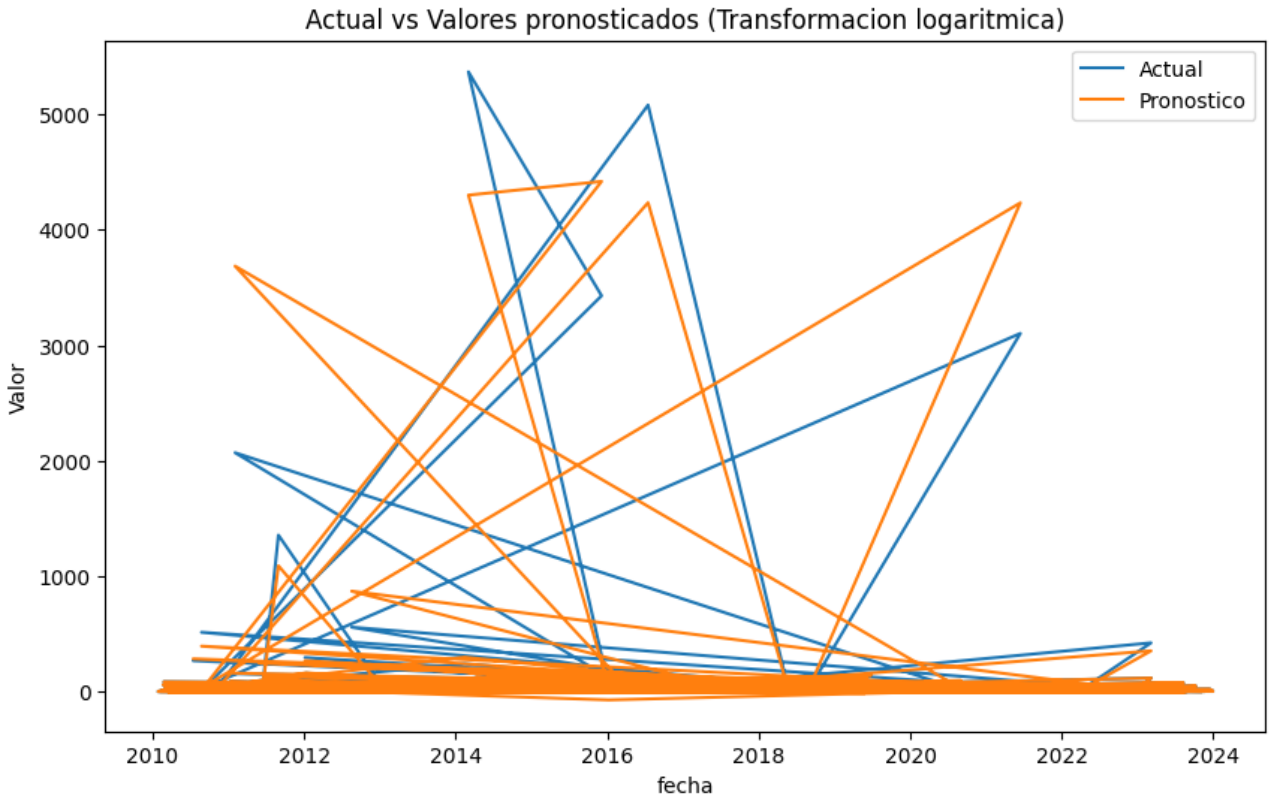
El MAE de 15.17 sugiere que, en promedio, las predicciones del modelo se desvían en 15.17 unidades del valor real. Esta mejora en el MAE indica una mayor precisión en las predicciones en comparación con los modelos anteriores. El MSE de 35512.92 y el RMSE de 188.45 reflejan una reducción en la magnitud de los errores cuadrados y su raíz cuadrada, respectivamente, mostrando una mejora en la capacidad del modelo para manejar errores más grandes.

El MAPE de 8.21% es especialmente notable, ya que indica una precisión porcentual mucho mayor en las predicciones. Esto sugiere que la transformación logarítmica ayudó a estabilizar la variabilidad en los datos, mejorando la capacidad del modelo para predecir valores en una escala más uniforme y consistente.

La visualización de los resultados mediante una gráfica comparativa de los valores reales y predichos (ilustración 27) demuestra que el modelo sigue de manera más precisa las tendencias generales del precio de importación por kilo de pesticidas, aunque aún persisten algunas discrepancias en los picos

más altos. Esto sugiere que, aunque el modelo ha mejorado significativamente, sigue enfrentando desafíos en la predicción de valores extremos.

***Ilustración 27 Transformación logarítmica***




*Elaboración propia, Fuente Dian*

En conclusión, la aplicación de una transformación logarítmica y la eliminación de valores atípicos han sido pasos efectivos para mejorar la precisión del modelo de predicción. Las métricas de error y la visualización de los resultados indican mejoras significativas, aunque aún existe margen para perfeccionar el modelo, especialmente en la predicción de valores extremos

### 6.5.2.8.AJUSTE CON LIGHTGBM

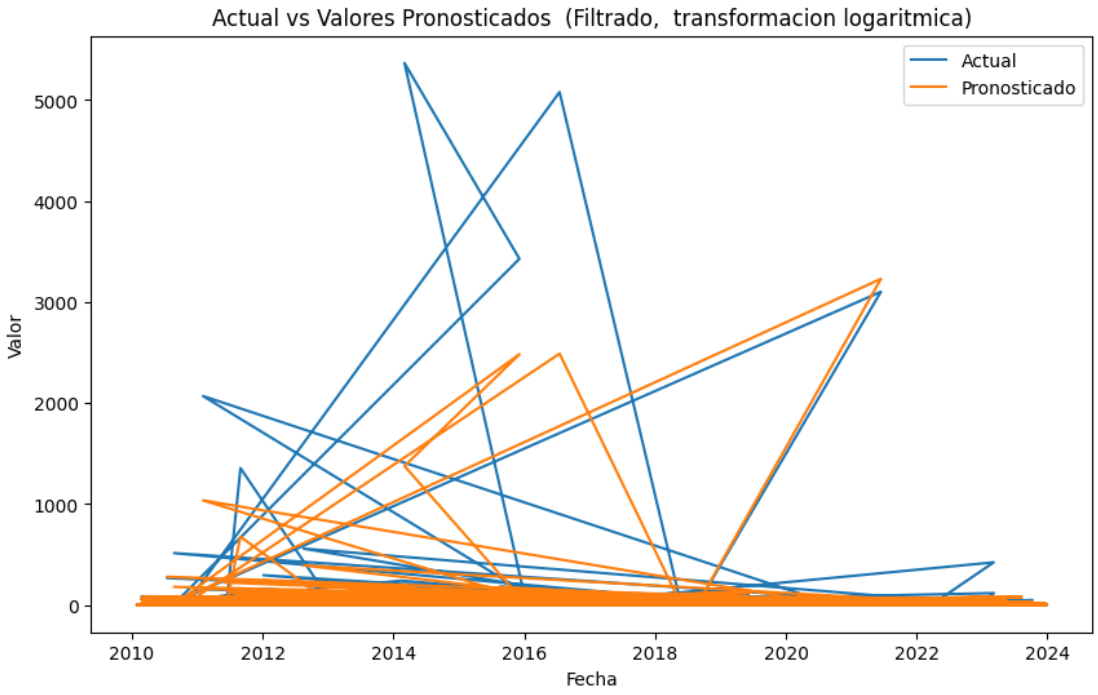
Para mejorar la precisión del modelo de predicción del precio de importación por kilo (VALOR_KILO) de pesticidas en Colombia, se optó por el modelo LightGBM, aplicando una transformación logarítmica previa a los datos y la eliminación de valores atípicos del conjunto de entrenamiento. La variable VALOR_KILO se transformó logarítmicamente, resultando en VALOR_KILO_LOG, para manejar la distribución sesgada y estabilizar la variabilidad de los datos. Las características (X) y las etiquetas (y) se separaron, y los datos se dividieron en conjuntos de entrenamiento y prueba en una proporción de 80% y 20%, respectivamente.

Se eliminaron valores atípicos del conjunto de entrenamiento usando índices previamente identificados. Con los datos filtrados, se configuró y ajustó un modelo LightGBM. Se empleó una búsqueda en cuadrícula (GridSearchCV) para optimizar hiperparámetros como learning_rate, num_leaves, n_estimators y subsample. Las predicciones se realizaron sobre el conjunto de prueba, invirtiendo la transformación logarítmica para comparar los valores predichos y reales en su escala original, y se calcularon las siguientes métricas de error para evaluar el rendimiento del modelo:

- **MAE (Error absoluto medio)**: 9.4006
- **MSE (Error cuadrático medio)**: 8541.76
- **RMSE (Raíz del error cuadrático medio)**: 92.42
- **MAPE (Error porcentual absoluto medio)**: 25.16%

Estas métricas indican que, en promedio, las predicciones del modelo se desvían en 9.4006 unidades del valor real. El MSE y RMSE reflejan la magnitud de los errores y su raíz cuadrados, respectivamente, mientras que un MAPE de 25.16% muestra que aún existen desafíos significativos en la precisión porcentual de las predicciones, especialmente en los valores extremos. La visualización de los resultados a través de una gráfica comparativa (Ilustración 28) revela que, aunque el modelo sigue tendencias generales del precio de importación por kilo de pesticidas de manera precisa, persisten discrepancias notables en los picos más altos, sugiriendo que aún hay margen para mejorar la captura de fluctuaciones extremas en los datos.

***Ilustración 28 Ajuste con LightGBM***

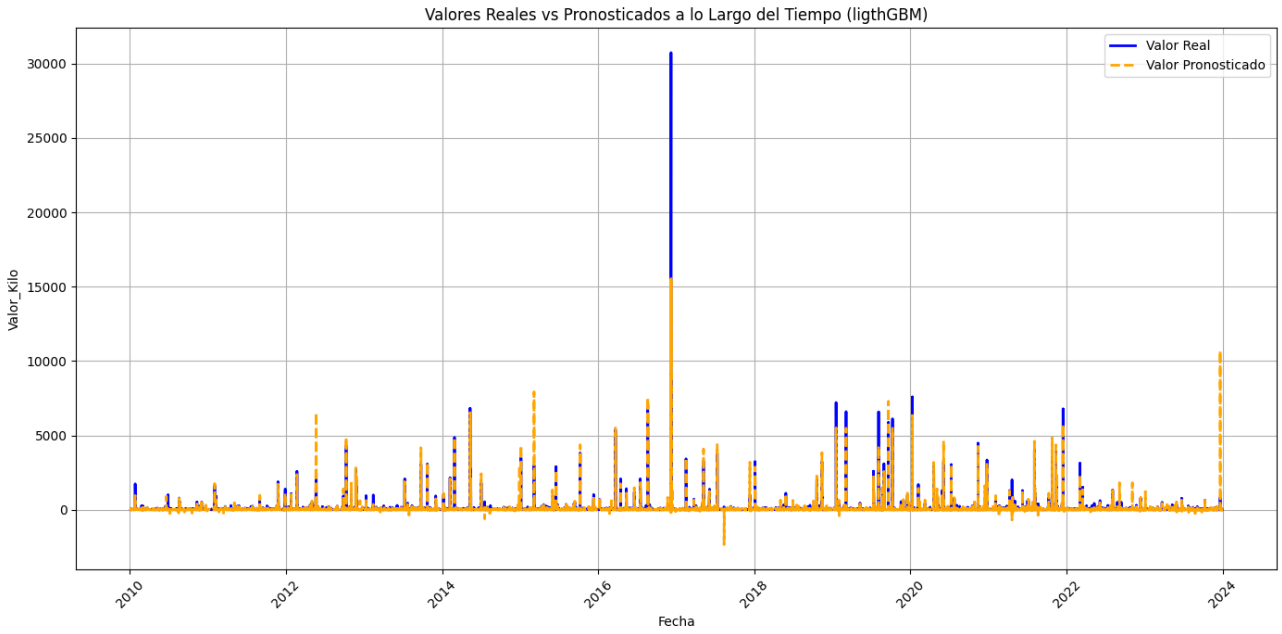



*Elaboración propia, Fuente Dian*

La aplicación de una transformación logarítmica y la eliminación de valores atípicos han demostrado ser pasos efectivos para mejorar la precisión del modelo de predicción utilizando LightGBM. Las métricas de error y la visualización de los resultados indican mejoras significativas en comparación con los modelos anteriores. No obstante, aún existe margen para mejorar, especialmente en la predicción de valores extremos. Futuras investigaciones podrían enfocarse en técnicas adicionales de manejo de outliers, ingeniería de características y la exploración de modelos alternativos para seguir mejorando la precisión de las predicciones del precio de importación por kilo de pesticidas en Colombia.

### 6.5.3. LIGHTGBM

***Ilustración 29 Modelo LightGBM***




*Elaboración propia, Fuente Dian*

Se implementó un modelo LightGBM optimizado utilizando GridSearchCV y una transformación logarítmica para predecir la variable VALOR_KILO. Previo al entrenamiento del modelo, se realizó un tratamiento de outliers mediante el método del Rango Intercuartílico (IQR) para eliminar valores extremos que pudieran distorsionar el rendimiento del modelo. La estrategia de eliminación de outliers se demostró efectiva para mejorar la precisión de las predicciones.

Los mejores hiperparámetros obtenidos a través de la optimización fueron: num_leaves = 123, feature_fraction = 0.994, bagging_fraction = 0.861 y min_child_samples = 22. Este enfoque permitió manejar mejor la variabilidad de los datos y los valores extremos, resultando en mejoras significativas en las métricas de evaluación del modelo.

Las métricas de desempeño del modelo optimizado fueron las siguientes: un MAE de 6.03, un MSE de 177,477.07, un RMSE de 421.28 y un MAPE de 15.74%. Estos resultados sugieren que el modelo LightGBM, junto con la eliminación de outliers y la transformación logarítmica, ofrece una precisión razonable en la predicción del VALOR_KILO. El análisis gráfico de los valores predichos frente a los valores reales (ilustración 29) mostró una fuerte correlación, indicando la eficacia del enfoque adoptado para mejorar el rendimiento predictivo.

En conclusión, la combinación de técnicas de optimización de hiperparámetros, preprocesamiento de datos mediante la eliminación de outliers y transformaciones logarítmicas, resultó en un modelo robusto y preciso para la predicción del VALOR_KILO, demostrando la importancia de un tratamiento de datos exhaustivo en modelos avanzados de machine learning.

#### 6.5.3.1.MODELO LIGHTGBM AJUSTADO

Para mejorar la precisión del modelo, se llevó a cabo un ajuste minucioso que incluyó la creación de características temporales a partir de la fecha, la eliminación de valores atípicos y la normalización de las características. Además, se aplicó una transformación Power a la variable objetivo para estabilizar la varianza y hacer los datos más adecuados para el modelado predictivo.

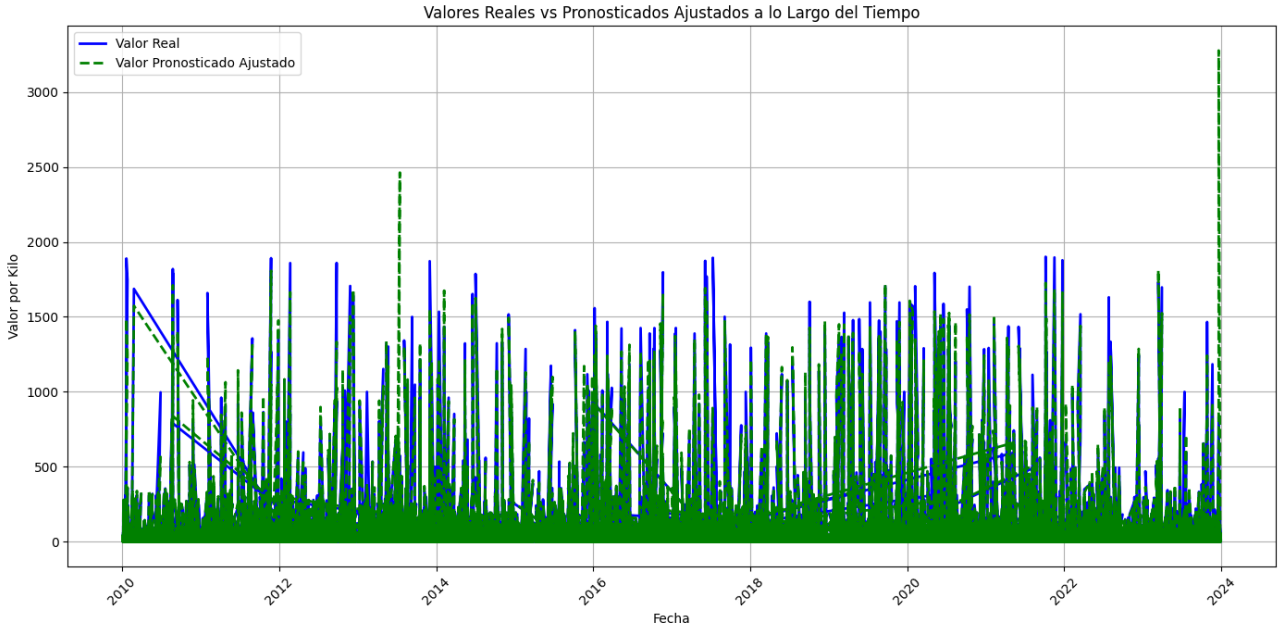
Se utilizó una estrategia de validación cruzada con división temporal para asegurar que el modelo generalizara adecuadamente a datos futuros. La optimización de hiperparámetros se realizó mediante una búsqueda en rejilla (GridSearchCV), evaluando combinaciones de los parámetros num_leaves, feature_fraction, bagging_fraction, learning_rate, min_child_samples y n_estimators. El objetivo principal fue minimizar el error absoluto medio (MAE) para mejorar la precisión del modelo.

Los mejores hiperparámetros encontrados para el modelo fueron los siguientes: bagging_fraction de 0.8, feature_fraction de 0.9, learning_rate de 0.05, min_child_samples de 26, n_estimators de 500 y num_leaves de 123. Estos hiperparámetros se seleccionaron por su capacidad para optimizar el rendimiento del modelo en la validación cruzada, logrando un equilibrio adecuado entre sesgo y varianza.

Los resultados de la evaluación del modelo utilizando los mejores hiperparámetros indicaron un error absoluto medio (MAE) de 0.7492 unidades, un error cuadrático medio (MSE) de 23.6041 unidades y una raíz del error cuadrático medio (RMSE) de 4.8584 unidades. Estos valores reflejan la precisión del modelo al predecir los precios de importación de los pesticidas. Sin embargo, el error porcentual absoluto medio (MAPE) fue extremadamente elevado, sugiriendo que, aunque el modelo puede ser preciso en términos absolutos, existen variaciones significativas en términos relativos.

El análisis visual a través del gráfico de valores reales versus valores predichos (Ilustración 30) mostró una proximidad considerable entre ambas líneas, lo que indica la capacidad del modelo para capturar la tendencia general de los datos. No obstante, la alta variación relativa observada en el MAPE sugiere que hay margen para mejorar el modelo. Este margen podría explorarse mediante la implementación de otras técnicas de ensamblado o un ajuste más refinado de los hiperparámetros y las características del modelo.

*Ilustración 30 modelo lightgbm ajustado*




*Elaboración propia, Fuente Dian*

En conclusión, el ajuste y la optimización del modelo LightGBM han permitido mejorar la precisión en la predicción del precio de importación de pesticidas en Colombia.

## 6.6.RESUMEN Y COMPARACIÓN DE MODELOS PREDICTIVOS

Se realizaron pruebas de estacionariedad y cointegración para evaluar la adecuación de los modelos predictivos aplicados a la serie temporal "VALOR_KILO". La Prueba de Dickey-Fuller Aumentada (ADF) arrojó un estadístico de -74.4699 con un p-valor de 0.0, lo que permitió rechazar la hipótesis nula de que la serie tiene una raíz unitaria, indicando que la serie es estacionaria. La Prueba KPSS corroboró estos hallazgos con un estadístico KPSS de 0.4277 y un p-valor de 0.0652, sugiriendo que no se puede rechazar la hipótesis nula de estacionariedad. Adicionalmente, la Prueba de Cointegración de Engle-Granger mostró un estadístico de -74.5147 y un p-valor de 0.0, indicando una fuerte relación de equilibrio a largo plazo entre "VALOR_KILO" y "PESO_NETO".

El modelo ARIMA (1 1 1) ajustado muestra coeficientes significativos para los términos AR (1) y MA (1) con un valor de AIC de 947850.905, sugiriendo un buen equilibrio entre la complejidad del modelo y el ajuste de los datos. Sin embargo, el diagnóstico de residuos revela problemas de no normalidad y heterocedasticidad con un p-valor de 0.00 en las pruebas de Jarque-Bera y de heterocedasticidad respectivamente, además de altos valores de sesgo (123.43) y curtosis (21112.73). Aunque el modelo ARIMA captura dependencias temporales básicas, estos problemas indican la necesidad de aplicar transformaciones adicionales o considerar modelos alternativos.

El modelo SARIMA (0 1 1) (1 1 1 12) optimizado para incluir componentes estacionales también presenta problemas similares. A pesar de capturar las dinámicas estacionales y no estacionales, la autocorrelación en los residuos, la no normalidad y la heterocedasticidad persisten. Se aplicó una transformación logarítmica a la serie temporal "VALOR_KILO", reduciendo ligeramente la heterocedasticidad, pero los problemas de autocorrelación y no normalidad de los residuos no se resolvieron completamente. El RMSE del modelo ajustado fue de 620.44, similar al modelo sin transformar.

El modelo Prophet mostró una buena capacidad para capturar las tendencias generales y los patrones estacionales con un RMSE de 552.08, comparable al del modelo ARIMA. Sin embargo, enfrentó dificultades para predecir valores extremos y picos, reflejando la naturaleza volátil de los datos.

En cuanto a los modelos de Machine Learning, el modelo Random Forest, El modelo Random Forest demostró un buen rendimiento general, capturando gran parte de la variabilidad en los datos con un RMSE de 88.65 y un $R^2$ de 0.917. No obstante, tuvo dificultades para predecir valores extremos.

Por otra parte, el modelo XGBoost optimizado mostró mejoras en algunas métricas, pero enfrentó desafíos significativos con la variabilidad alta y los valores atípicos, obteniendo un MAE de 9.34, un MSE de 21257.29, un RMSE de 145.80 y un MAPE de 83.92%.

Y finalmente el modelo LightGBM optimizado utilizando GridSearchCV y una transformación logarítmica mostró la mejor precisión general con un MAE de 6.03, un MSE de 177477.07 y un RMSE de 421.28, destacándose por manejar mejor la variabilidad de los datos y los valores extremos.

***Tabla 9 Comparación de modelo***

| Modelo | RMSE | MAE | MSE | MAPE | $R^2$ (Entrenamiento) | $R^2$ (Prueba) |
|---|---|---|---|---|---|---|
| ARIMA(1, 1, 1) | 626.77 | N/A | N/A | N/A | N/A | N/A |
| SARIMA(0, 1, 1)(1, 1, 1, 12) | 620.44 | N/A | N/A | N/A | N/A | N/A |
| Prophet | 552.08 | N/A | N/A | N/A | N/A | N/A |
| Random Forest | 88.65 | 4.58 | N/A | 232554579282.13 | 0.709 | 0.917 |
| XGBoost | 145.80 | 9.34 | 21257.29 | 83.92 | N/A | N/A |
| LightGBM | 421.28 | 6.03 | 177477.07 | 15.74 | N/A | N/A |

*Elaboración propia, Fuente Dian*

Estos resultados (Tabla 9) indican que, si bien los modelos ARIMA y SARIMA son útiles para capturar dependencias temporales, enfrentan problemas significativos con la distribución de los residuos. Por otro lado, los modelos de Machine Learning como Random Forest, XGBoost y LightGBM, especialmente este último tras optimización, ofrecen un rendimiento superior en la predicción de "VALOR_KILO" aunque aún presentan desafíos con la alta variabilidad y los valores atípicos.

### 6.6.1. ELECCIÓN DEL MODELO

Basados en el análisis comparativo de los modelos predictivos, la elección del modelo más adecuado dependerá de varios factores, incluyendo la precisión de las predicciones, la capacidad para manejar

valores extremos y la interpretabilidad del modelo. A continuación, se resumen las razones para la elección del modelo LightGBM:

- **Precisión:** El modelo LightGBM mostró el menor MAE (6.03) y un RMSE de 421.28, lo que indica que tiene una precisión superior en comparación con los otros modelos evaluados.
- **Manejo de Valores Extremos:** Tras la optimización y el tratamiento de outliers mediante la eliminación de valores extremos y la transformación logarítmica, el modelo LightGBM ha demostrado manejar mejor la variabilidad de los datos.
- **Optimización:** La optimización de hiperparámetros mediante GridSearchCV permitió encontrar la configuración más adecuada, mejorando significativamente las métricas de evaluación.

### 6.6.2. COMPARACIÓN CON OTROS MODELOS

- **Modelos ARIMA y SARIMA:** A pesar de capturar dependencias temporales y estacionales, presentan problemas significativos con la no normalidad y heterocedasticidad de los residuos.
- **Prophet:** Aunque captura bien las tendencias generales y los patrones estacionales, enfrenta dificultades para predecir valores extremos.
- **Random Forest y XGBoost:** Ambos modelos tienen un buen rendimiento general, pero aún enfrentan desafíos significativos con la alta variabilidad y los valores atípicos. LightGBM en comparación mostró mejores resultados tras la optimización.

Por lo tanto, basado en el análisis detallado y los resultados obtenidos, se recomienda el uso del modelo LightGBM para la predicción del valor por kilo de importación de pesticidas en Colombia. Este modelo ha demostrado tener la mejor precisión general y una capacidad superior para manejar la variabilidad de los datos y los valores extremos tras la optimización y el preprocesamiento adecuados.

## 7. APLICACION DEL PRONOSTICO

### 7.1.TENDENCIA AÑO 2023

La tendencia general de los precios a lo largo del año muestra picos y valles significativos (Tabla10), indicando una posible influencia de factores estacionales o eventuales en el mercado. La representación gráfica (ilustración 31) muestra claramente estas fluctuaciones, con incrementos notables en los primeros meses del año y caídas pronunciadas en el segundo trimestre. Posteriormente, los precios muestran una recuperación gradual, alcanzando nuevamente un pico en septiembre (7.51 dólares por kilo) y finalizando el año con una ligera caída en noviembre (6.36 dólares por kilo), antes de subir ligeramente en diciembre (6.95 dólares por kilo).

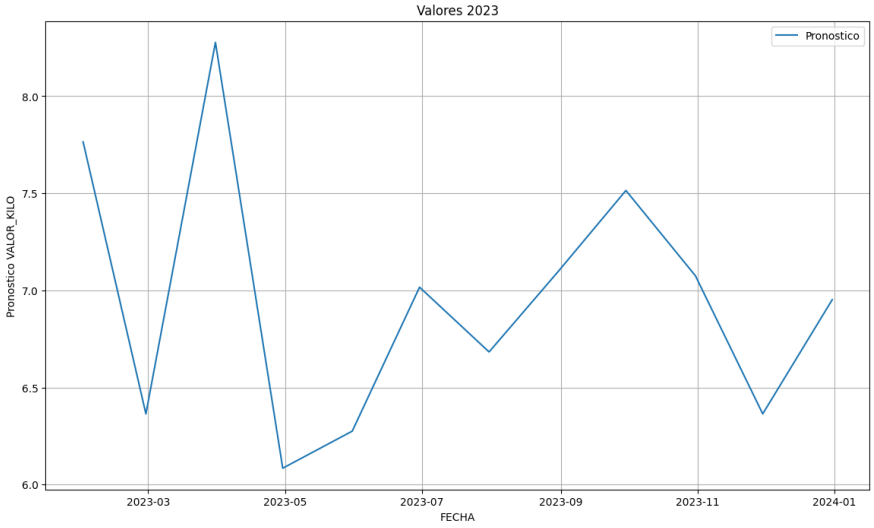
***Tabla 10 precio promedio valor_kilo por mes 2023***

| FECHA_PRESENTACION | VALOR_KILO_PREDICTED |
|---|---|
| 31/01/2023 | 7,765 |
| 29/02/2023 | 6,364 |
| 31/03/2023 | 8,277 |
| 30/04/2023 | 6,085 |
| 31/05/2023 | 6,275 |
| 30/06/2023 | 7,017 |
| 31/07/2023 | 6,683 |
| 31/08/2023 | 7,100 |
| 30/09/2023 | 7,514 |
| 31/10/2023 | 7,075 |
| 30/11/2023 | 6,364 |
| 31/12/2023 | 6,952 |

*Elaboración propia, Fuente Dian*

Esta tendencia de precios sugiere que los importadores deben considerar estrategias de compra y almacenamiento que optimicen los costos, aprovechando los meses de precios más bajos y planificando adecuadamente para los periodos de altos costos. Además, estos patrones pueden ayudar a los responsables de la toma de decisiones a desarrollar políticas más informadas y eficaces para gestionar la cadena de suministro y los presupuestos anuales.

***ilustración 31 tendencia valor_kilo por mes 2023***




*Elaboración propia, Fuente Dian*

## 7.2.PRONOSTICO AÑO 2024

Las predicciones generadas para el próximo año sobre los precios de importación de pesticidas en Colombia muestran una variabilidad significativa, con precios que oscilan entre aproximadamente 6.19 y 8.20 dólares por kilo (Tabla 11). Se observan picos notables en los meses de marzo (8.20 dólares) y enero (7.64 dólares), mientras que los precios más bajos se presentan en abril (6.19 dólares) y mayo (6.32 dólares). Esta variabilidad sugiere la influencia de factores estacionales y externos en el mercado de pesticidas (Ilustración 32).

La información derivada de estas predicciones es crucial para la toma de decisiones estratégicas en varios ámbitos. En primer lugar, permite una planificación más eficiente de las compras y el almacenamiento. Las empresas pueden optar por adquirir mayores volúmenes de pesticidas durante los meses de precios más bajos, como abril y mayo, para reducir los costos generales y evitar compras en meses con precios elevados como marzo. Además, la capacidad de almacenar pesticidas sin deterioro potencializa esta estrategia de compra anticipada.

***Tabla 11 pronostico valor_kilo 2024***

| FECHA_PRESENTACION | VALOR_KILO_PREDICTED |
|---|---|
| 31/01/2024 | 7,643 |
| 29/02/2024 | 6,416 |
| 31/03/2024 | 8,198 |
| 30/04/2024 | 6,193 |
| 31/05/2024 | 6,317 |
| 30/06/2024 | 6,929 |
| 31/07/2024 | 6,645 |
| 31/08/2024 | 6,984 |
| 30/09/2024 | 7,520 |
| 31/10/2024 | 6,901 |
| 30/11/2024 | 6,416 |
| 31/12/2024 | 6,992 |

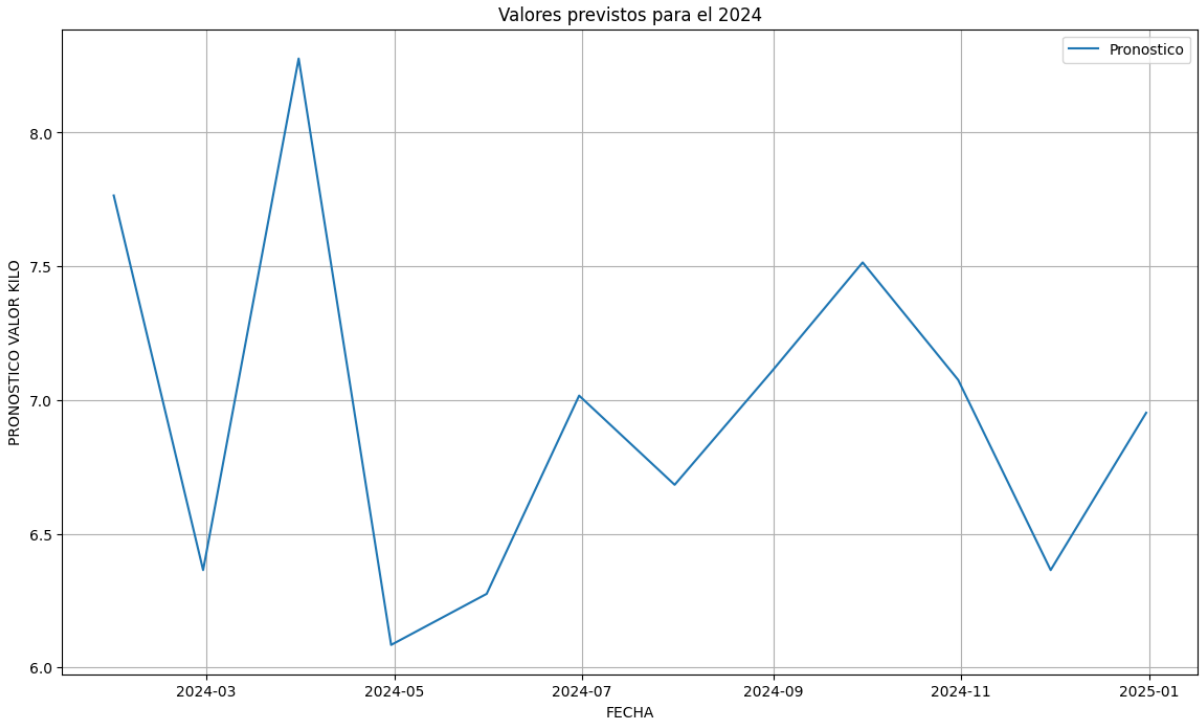
*Elaboración propia, Fuente Dian*

Desde una perspectiva presupuestaria, estas predicciones facilitan la elaboración de presupuestos más precisos y la asignación eficiente de recursos. La gestión de costos se ve beneficiada al identificar meses con altos costos y desarrollar estrategias para mitigarlos, ya sea negociando precios con proveedores o explorando alternativas. En términos de política de precios, los importadores y distribuidores pueden ajustar los precios de venta acorde a las fluctuaciones previstas en los costos de importación, lo que les permite maximizar márgenes de ganancia en meses de alto costo y ofrecer descuentos competitivos en períodos de bajos precios.

En cuanto a la logística y la contratación, la previsión de precios facilita la negociación de contratos de suministro a largo plazo que estabilicen los costos, y la optimización de la logística de importación

para minimizar gastos asociados al transporte y almacenamiento. Es fundamental, no obstante, validar continuamente estas predicciones con datos reales a medida que estén disponibles y ajustar el modelo predictivo según sea necesario. Adicionalmente, es imperativo considerar factores externos como cambios en políticas comerciales, fluctuaciones en las tasas de cambio y eventos globales que puedan impactar los precios de los pesticidas.

***Ilustración 32 Pronostico VALOR_KILO 2024***




*Elaboración propia, Fuente Dian*

En conclusión, las predicciones de precios de importación de pesticidas ofrecen una herramienta valiosa para la toma de decisiones informadas y estratégicas en la gestión de la cadena de suministro, optimización de costos y planificación financiera, permitiendo a las empresas responder de manera proactiva a las fluctuaciones del mercado.

## 7.3.COMPARACIÓN DE LOS RESULTADOS DE PRECIOS DE IMPORTACIÓN DE PESTICIDAS EN COLOMBIA PARA LOS AÑOS 2023 Y 2024

Al comparar los precios de importación de pesticidas en Colombia entre los años 2023 y 2024, se observa una tendencia de variabilidad significativa en ambos años, aunque con diferencias notables en sus patrones específicos. En 2023, los precios fluctuaron entre aproximadamente 6.08 y 8.28 dólares por kilo, con los precios más altos registrados en marzo (8.28 dólares por kilo) y enero (7.77 dólares por kilo), y los precios más bajos observados en abril (6.08 dólares por kilo) y mayo (6.28 dólares por kilo). La tendencia a lo largo de 2023 mostró picos pronunciados en los primeros meses,

seguidos por una caída en el segundo trimestre y una recuperación gradual en la segunda mitad del año.

En contraste, para 2024, los precios predichos oscilan entre 6.19 y 8.20 dólares por kilo. Los picos más altos se anticipan nuevamente en los primeros meses, con el precio más alto en marzo (8.20 dólares por kilo) y el segundo pico en enero (7.64 dólares por kilo). Sin embargo, la tendencia de 2024 muestra una recuperación más moderada después de los meses de precios bajos en abril (6.19 dólares por kilo) y mayo (6.32 dólares por kilo), con una tendencia más uniforme en la segunda mitad del año y menos fluctuaciones extremas en comparación con 2023.

Estos resultados sugieren que, aunque ambos años presentan fluctuaciones estacionales significativas, 2023 tuvo variaciones más pronunciadas y extremas en comparación con las predicciones para 2024. Esta información es crucial para los responsables de la toma de decisiones en la planificación de compras y almacenamiento, ya que les permite anticipar y prepararse mejor para las fluctuaciones del mercado, optimizando así la gestión de costos y recursos.

## 7.4.PRONOSTICO POR TIPO DE PESTICIDA

### 7.4.1. INSECTICIDAS

Al comparar los precios reales de importación de insecticidas en Colombia en 2023 con las predicciones para 2024, se observan variaciones notables en varias fechas clave. En enero, el precio predicho para 2024 (12.61 USD/kg) es ligeramente inferior al precio real de 2023 (13.51 USD/kg). En febrero de 2024, el precio predicho (11.46 USD/kg) supera al del mismo mes en 2023 (10.57 USD/kg). Marzo se mantiene como el mes con los precios más altos en ambos años, con 15.60 USD/kg en 2023 y una predicción de 14.71 USD/kg para 2024 (Tabla 12).

Abril presenta los precios más bajos, con 9.00 USD/kg en 2023 y una ligera recuperación a 9.90 USD/kg pronosticada para 2024. En mayo, el precio predicho para 2024 (10.67 USD/kg) es inferior al real de 2023 (11.57 USD/kg). Junio de 2024 muestra un incremento significativo en la predicción a 14.75 USD/kg frente a los 13.86 USD/kg reales de 2023. En julio y agosto, los precios predichos para 2024 son 10.59 USD/kg y 11.96 USD/kg respectivamente, en comparación con los precios reales de 11.48 USD/kg y 11.07 USD/kg en 2023.

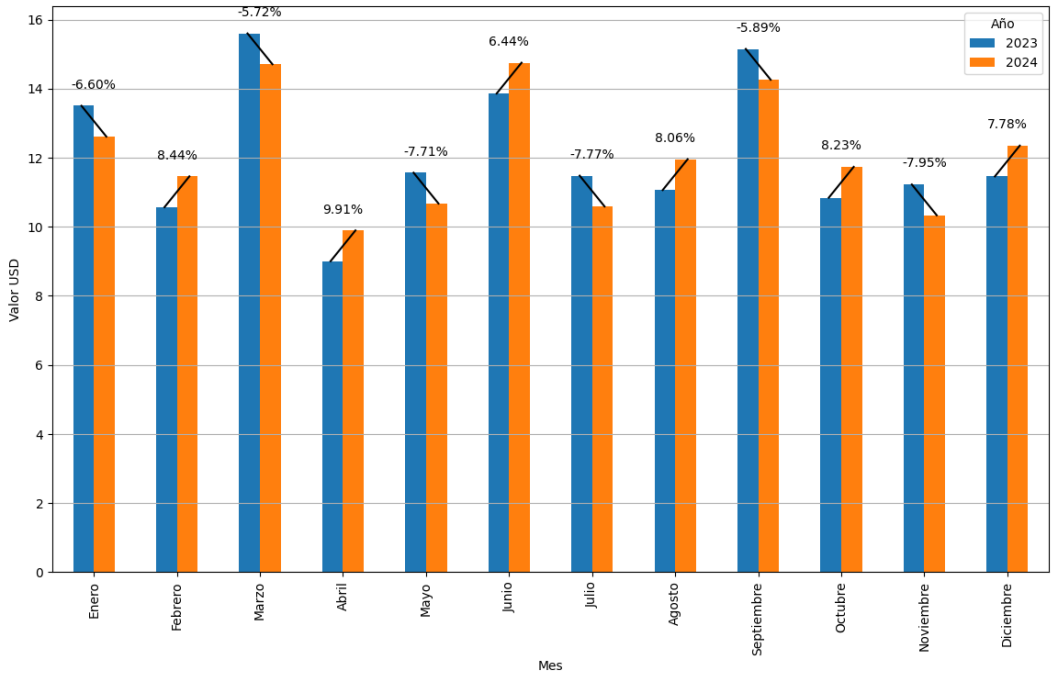
Septiembre mantiene altos precios en ambos años, aunque ligeramente menores en la predicción para 2024 (14.26 USD/kg) en comparación con el valor real de 2023 (15.15 USD/kg). Octubre y noviembre muestran tendencias opuestas: en octubre, el precio predicho para 2024 (11.74 USD/kg) es mayor que el real de 2023 (10.84 USD/kg), mientras que, en noviembre, el precio predicho para 2024 (10.33 USD/kg) es inferior al real de 2023 (11.23 USD/kg). Finalmente, diciembre de 2024 muestra un incremento en la predicción a 12.35 USD/kg comparado con el precio real de 11.46 USD/kg en 2023.

***Tabla 12: Pronostico insecticida***

| FECHA | VALOR_KILO | FECHA | VALOR_KILO |
|---|---|---|---|
| 31/01/2024 | 12,613 | 31/01/2023 | 13,505 |
| 29/02/2024 | 11,461 | 28/02/2023 | 10,569 |
| 31/03/2024 | 14,708 | 31/03/2023 | 15,600 |
| 30/04/2024 | 9,896 | 30/04/2023 | 9,004 |
| 31/05/2024 | 10,674 | 31/05/2023 | 11,566 |
| 30/06/2024 | 14,752 | 30/06/2023 | 13,860 |
| 31/07/2024 | 10,589 | 31/07/2023 | 11,481 |
| 31/08/2024 | 11,957 | 31/08/2023 | 11,065 |
| 30/09/2024 | 14,259 | 30/09/2023 | 15,151 |
| 31/10/2024 | 11,736 | 31/10/2023 | 10,844 |
| 30/11/2024 | 10,334 | 30/11/2024 | 11,226 |
| 31/12/2024 | 12,351 | 31/12/2024 | 11,459 |

*Elaboración propia, Fuente Dian 3*

***Ilustración 33 Comparación de Precios de Insecticidas: 2023 vs 2024***




*Elaboración propia, Fuente Dian*

En resumen, los precios predichos de importación de insecticidas para 2024 tienden a ser más bajos en los primeros meses del año y más altos en los últimos meses, en comparación con los precios reales de 2023. Estas fluctuaciones sugieren que las empresas importadoras deben considerar estrategias de

compra y almacenamiento que optimicen costos, aprovechando las variaciones estacionales de los precios.

### 7.4.2. HERBICIDAS

Al contrastar los precios reales de importación de herbicidas en Colombia en 2023 con las proyecciones para 2024, se observan varias diferencias notables a lo largo del año. En enero, el precio proyectado para 2024 (4.87 USD/kg) es significativamente menor que el registrado en 2023 (5.76 USD/kg). En febrero de 2024, se anticipa que el precio (4.44 USD/kg) será superior al del mismo mes en 2023 (3.54 USD/kg). Marzo se mantiene como el mes con los precios más altos en ambos años, aunque se espera que sea ligeramente más bajo en 2024 (5.84 USD/kg) en comparación con 2023 (6.73 USD/kg) (Tabla 13).

***Tabla 13 herbicidas***

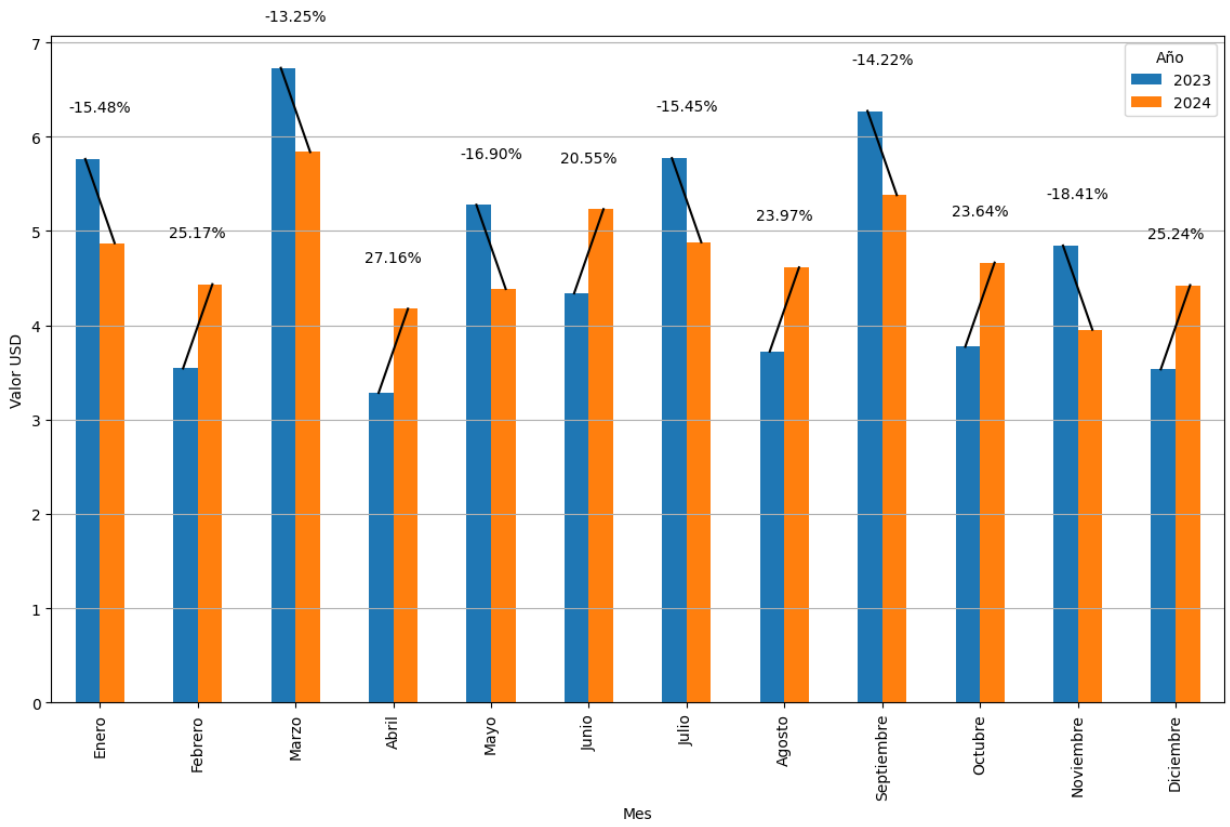
| FECHA | VALOR_KILO | FECHA | VALOR_KILO |
|---|---|---|---|
| 31/01/2024 | 4,872 | 31/01/2023 | 5,764 |
| 29/02/2024 | 4,436 | 28/02/2023 | 3,544 |
| 31/03/2024 | 5,838 | 31/03/2023 | 6,730 |
| 30/04/2024 | 4,176 | 30/04/2023 | 3,284 |
| 31/05/2024 | 4,386 | 31/05/2023 | 5,278 |
| 30/06/2024 | 5,232 | 30/06/2023 | 4,340 |
| 31/07/2024 | 4,881 | 31/07/2023 | 5,773 |
| 31/08/2024 | 4,614 | 31/08/2023 | 3,722 |
| 30/09/2024 | 5,381 | 30/09/2023 | 6,273 |
| 31/10/2024 | 4,665 | 31/10/2023 | 3,773 |
| 30/11/2024 | 3,954 | 30/11/2024 | 4,846 |
| 31/12/2024 | 4,426 | 31/12/2024 | 3,534 |

*Elaboración propia, Fuente Dian*

En abril, los precios más bajos se registraron en 2023 con 3.28 USD/kg, mientras que para 2024 se prevé un aumento a 4.18 USD/kg. Para mayo, se pronostica que el precio en 2024 (4.39 USD/kg) será inferior al de 2023 (5.28 USD/kg). En junio de 2024, se espera un incremento a 5.23 USD/kg en comparación con los 4.34 USD/kg reales de 2023. Tanto en julio como en agosto, los precios esperados para 2024 son 4.88 USD/kg y 4.61 USD/kg respectivamente, en contraste con los precios reales de 5.77 USD/kg y 3.72 USD/kg en 2023.

Septiembre presenta precios altos en ambos años, aunque con una ligera disminución en la proyección para 2024 (5.38 USD/kg) frente al valor real de 2023 (6.27 USD/kg). En octubre, el precio proyectado para 2024 (4.67 USD/kg) es más alto que el del mismo mes en 2023 (3.77 USD/kg), mientras que en noviembre se espera que el precio en 2024 (3.95 USD/kg) sea menor al de 2023 (4.85 USD/kg). Finalmente, diciembre de 2024 muestra una predicción de aumento a 4.43 USD/kg en comparación con el precio real de 3.53 USD/kg en 2023 (ilustración 34).

*Ilustración 34 Comparación de Precios de Herbicidas: 2023 vs 2024*




*Elaboración propia, Fuente Dian*

En conclusión, los precios previstos para la importación de herbicidas en 2024 tienden a ser más bajos en los primeros meses del año y exhiben fluctuaciones en comparación con los precios reales de 2023. Estas diferencias sugieren que las empresas importadoras deberían considerar estrategias de adquisición y almacenamiento que permitan optimizar costos, tomando en cuenta las variaciones estacionales en los precios.

### 7.4.3. FUNGICIDAS

Al comparar los precios reales de importación de fungicidas en Colombia en 2023 con las proyecciones para 2024, se observan diferencias significativas a lo largo del año (Tabla 14). En enero, el precio proyectado para 2024 (8.56 USD/kg) es inferior al registrado en 2023 (9.45 USD/kg). En febrero de 2024, se espera que el precio (7.64 USD/kg) sea más alto que en 2023 (6.75 USD/kg). Marzo sigue siendo el mes con los precios más altos en ambos años, aunque ligeramente inferior en 2024 (10.84 USD/kg) en comparación con 2023 (11.73 USD/kg).

En abril, los precios más bajos se registraron en 2023 con 6.47 USD/kg, mientras que para 2024 se prevé un aumento a 7.36 USD/kg. Para mayo, se anticipa que el precio en 2024 (7.45 USD/kg) será menor al de 2023 (8.34 USD/kg). En junio de 2024, se espera un incremento a 9.85 USD/kg comparado con los 8.96 USD/kg reales de 2023. Tanto en julio como en agosto, los precios esperados para 2024 son 7.87 USD/kg y 8.85 USD/kg respectivamente, en contraste con 8.76 USD/kg y 7.95 USD/kg en 2023.

***Tabla 14 fungicidas***

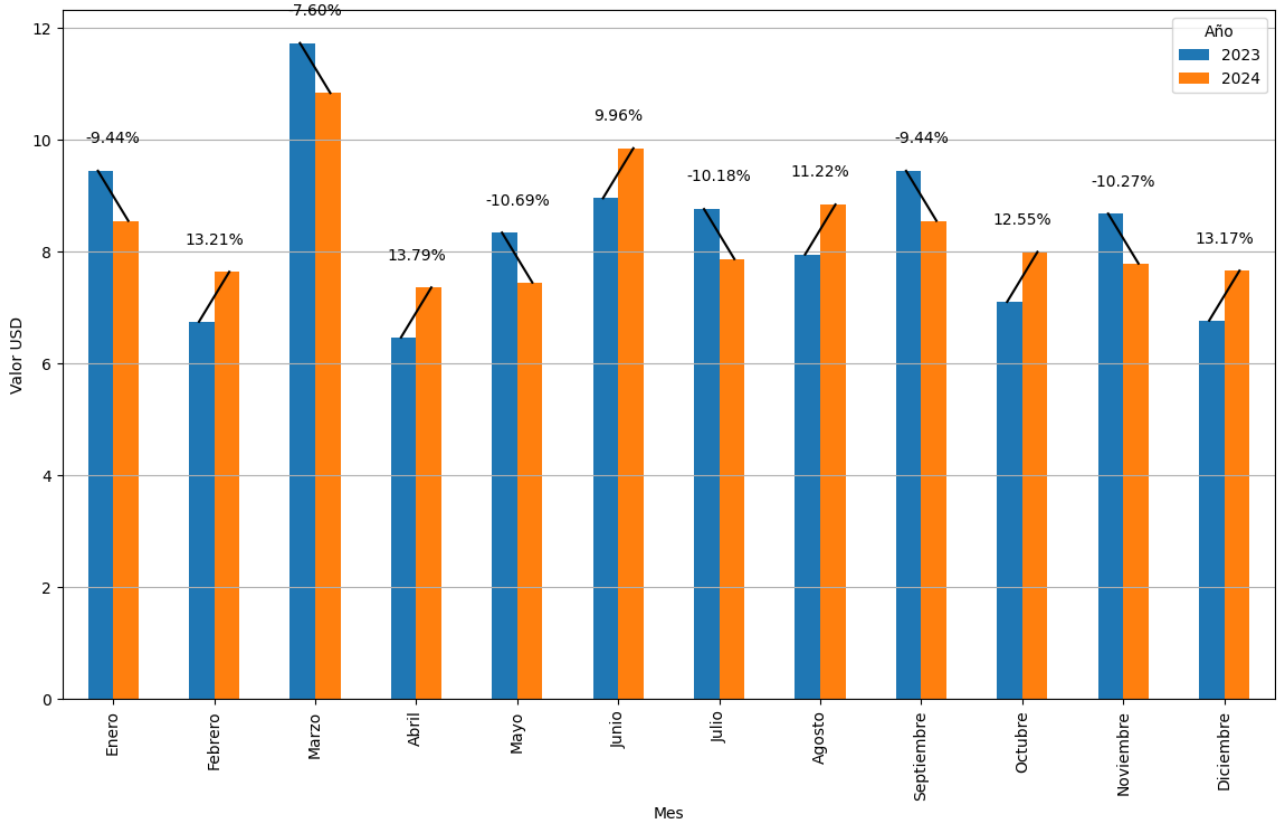
| FECHA | VALOR_KILO | FECHA | VALOR_KILO |
|---|---|---|---|
| 31/01/2024 | 8,556 | 31/01/2023 | 9,448 |
| 29/02/2024 | 7,642 | 28/02/2023 | 6,750 |
| 31/03/2024 | 10,841 | 31/03/2023 | 11,733 |
| 30/04/2024 | 7,362 | 30/04/2023 | 6,470 |
| 31/05/2024 | 7,451 | 31/05/2023 | 8,343 |
| 30/06/2024 | 9,852 | 30/06/2023 | 8,960 |
| 31/07/2024 | 7,872 | 31/07/2023 | 8,764 |
| 31/08/2024 | 8,845 | 31/08/2023 | 7,953 |
| 30/09/2024 | 8,556 | 30/09/2023 | 9,448 |
| 31/10/2024 | 7,998 | 31/10/2023 | 7,106 |
| 30/11/2024 | 7,791 | 30/11/2024 | 8,683 |
| 31/12/2024 | 7,663 | 31/12/2024 | 6,771 |

*Elaboración propia, Fuente Dian*

Septiembre presenta precios altos en ambos años, aunque con una ligera disminución en la proyección para 2024 (8.56 USD/kg) frente a 2023 (9.45 USD/kg). En octubre, el precio proyectado para 2024 (8.00 USD/kg) es mayor que en 2023 (7.11 USD/kg), mientras que en noviembre se espera que el precio en 2024 (7.79 USD/kg) sea inferior al de 2023 (8.68 USD/kg). Finalmente, diciembre de 2024 muestra una predicción de aumento a 7.66 USD/kg comparado con el precio real de 6.77 USD/kg en 2023 (Ilustración 35).

En resumen, los precios proyectados para la importación de fungicidas en 2024 tienden a ser más bajos en los primeros meses del año y muestran fluctuaciones en comparación con 2023. Estas diferencias indican que las empresas importadoras deben planificar estrategias de adquisición y almacenamiento para optimizar costos, teniendo en cuenta las variaciones estacionales en los precios.

*Ilustración 35 Comparación de Precios de Fungicidas: 2023 vs 2024*




***Elaboración propia, Fuente Dian***

El modelo de predicción de precios de importación de pesticidas en Colombia proporciona una herramienta valiosa para la toma de decisiones en la industria agrícola. Al anticipar las fluctuaciones de precios a lo largo del año, los importadores y distribuidores pueden optimizar sus estrategias de compra y almacenamiento, reduciendo costos y mejorando la eficiencia operativa. Por ejemplo, al identificar los meses con precios más bajos, las empresas pueden planificar compras anticipadas y gestionar inventarios de manera más efectiva, evitando así costos elevados durante los picos de precios.

Además, el impacto de este modelo en la industria agrícola colombiana es significativo. La capacidad de predecir precios permite a los agricultores y empresas agrícolas gestionar mejor sus presupuestos y recursos, asegurando un suministro continuo y asequible de pesticidas. Esto no solo mejora la rentabilidad, sino que también contribuye a la estabilidad del mercado agrícola. En un contexto de creciente competencia global y volatilidad económica, el uso de modelos predictivos avanzados se convierte en un diferenciador clave para mantener la sostenibilidad y el crecimiento de la agricultura en Colombia.

# 8. CONCLUSION Y TRABAJOS FUTUROS

## 8.1. CONCLUSIONES

El presente estudio ha evidenciado la relevancia de la agricultura como fundamento de la seguridad alimentaria y el desarrollo económico, destacando la influencia significativa de los costos de insumos importados en la competitividad del sector agrícola colombiano. La fluctuación de la tasa de cambio y el aumento constante de los precios de insecticidas, fungicidas y herbicidas han generado una presión económica considerable sobre los productores agrícolas. A continuación, se describe el cumplimiento de los objetivos planteados con esta investigación.

**Consolidación de la Información:** Se logró consolidar y procesar una gran cantidad de datos de importación proporcionados por la Dirección de Impuestos y Aduanas Nacionales de Colombia (DIAN). La extracción, limpieza, minería y análisis exploratorio de estos datos permitió obtener un entendimiento profundo de los patrones de importación y precios de los insumos agrícolas. Esta etapa fue crucial para establecer una base sólida sobre la cual desarrollar el modelo predictivo.

**Elaboración del Modelo Predictivo:** En la fase de elaboración del modelo predictivo, se experimentó con varias técnicas avanzadas de ciencias de datos, incluyendo ARIMA, SARIMA, Random Forest, XGBoost, y LightGBM. Cada modelo fue evaluado para determinar su eficacia en pronosticar el precio promedio de importación de insecticidas, fungicidas y herbicidas. Aunque algunos modelos demostraron ser más eficaces que otros, los resultados variaron según el tipo de pesticida y las condiciones específicas de los datos. Por lo tanto, es crucial continuar con la experimentación y ajuste de los modelos para asegurar predicciones más precisas y confiables, y se recomienda cautela al generalizar la capacidad predictiva de los modelos.

**Evaluación del Alcance de Predicción:** La evaluación de los modelos predictivos reveló una capacidad variable para ofrecer predicciones precisas, dependiendo del modelo y del conjunto de datos específico. Los análisis realizados ayudaron a identificar qué modelos son más adecuados para diferentes tipos de pesticidas y escenarios de datos. Esta evaluación subraya la importancia de seleccionar el modelo adecuado para cada caso específico y resalta la necesidad de seguir refinando las herramientas predictivas para maximizar su precisión y utilidad en la planificación y gestión de costos en el sector agrícola.

**Implementación de Herramientas de Visualización:** Se desarrollaron herramientas de visualización que permiten un acceso intuitivo y comprensible a la información histórica y a las proyecciones del modelo. Estas herramientas facilitan la interpretación de los datos y resultados, apoyando a los actores del sector agrícola en la toma de decisiones estratégicas basadas en datos, optimizando la gestión de costos y mejorando la eficiencia operativa. Las cuales pueden se encontradas en la siguiente URL: https://lookerstudio.google.com/reporting/ca786090-220b-4747-ad5b-2ef6479da08e

En resumen, este estudio subraya la importancia de la implementación de tecnologías de ciencias de datos y herramientas analíticas en la agricultura para abordar los desafíos económicos y mejorar la sostenibilidad y competitividad del sector. La integración de estas tecnologías puede beneficiar

significativamente a productores, consumidores y a la economía en general, permitiendo una gestión más eficiente y efectiva de los recursos y costos agrícolas.

## 8.2. RECOMENDACIONES

**Monitoreo Continuo de Datos:** Es fundamental establecer un sistema de monitoreo continuo de los datos de importación y precios de insumos agrícolas. Esto permitirá mantener actualizados los modelos predictivos y mejorar la precisión de las predicciones a medida que cambian las condiciones del mercado.

**Capacitación en Ciencias de Datos:** Se recomienda capacitar a los actores del sector agrícola en el uso de técnicas de ciencias de datos y herramientas de visualización. Esto les permitirá aprovechar al máximo los modelos predictivos y tomar decisiones informadas basadas en datos.

**Diversificación de Proveedores:** Para mitigar el impacto de las fluctuaciones en la tasa de cambio y los precios de los insumos importados, se sugiere diversificar los proveedores y explorar alternativas nacionales cuando sea posible. Esta estrategia puede reducir la dependencia de insumos importados y mejorar la resiliencia del sector ante cambios económicos internacionales.

**Colaboración con Instituciones Académicas y de Investigación:** Fomentar la colaboración con universidades y centros de investigación puede proporcionar acceso a conocimientos avanzados y tecnologías emergentes en ciencias de datos. Estas alianzas pueden ayudar a mejorar los modelos predictivos y desarrollar nuevas soluciones para el sector agrícola.

**Implementación de Políticas de Apoyo:** Se recomienda trabajar con entidades gubernamentales para desarrollar políticas que apoyen la adopción de tecnologías de ciencias de datos en la agricultura. Estas políticas pueden incluir incentivos fiscales, subsidios y programas de formación para agricultores y empresas del sector.

## 8.3.TRABAJOS FUTUROS

**Mejora de Modelos Predictivos:** Continuar investigando y desarrollando modelos predictivos más avanzados que incorporen variables adicionales como condiciones climáticas, políticas comerciales y tendencias globales del mercado agrícola. La integración de datos de múltiples fuentes puede mejorar significativamente la precisión de las predicciones.

**Desarrollo de Sistemas de Alerta Temprana:** Implementar sistemas de alerta temprana basados en los modelos predictivos para notificar a los agricultores y otros actores del sector sobre cambios inminentes en los precios de los insumos. Estos sistemas pueden ayudar a tomar decisiones proactivas y mitigar el impacto de las fluctuaciones de precios.

**Exploración de Nuevas Tecnologías:** Investigar la aplicación de nuevas tecnologías emergentes, como la inteligencia artificial y el aprendizaje profundo, para mejorar la capacidad de los modelos predictivos. Estas tecnologías pueden proporcionar insights más profundos y precisos sobre las

tendencias del mercado.

**Evaluación del Impacto Económico:** Realizar estudios que evalúen el impacto económico de la implementación de modelos predictivos y tecnologías de ciencias de datos en el sector agrícola. Estos estudios pueden proporcionar evidencia cuantitativa del valor agregado y justificar inversiones adicionales en estas tecnologías.

**Expansión a Otros Sectores Agrícolas:** Ampliar el enfoque de los modelos predictivos y las herramientas desarrolladas a otros sectores agrícolas, como la producción de frutas, verduras y cereales. La adaptación de estas tecnologías a diferentes subsectores puede maximizar su impacto y beneficios.

**Investigación sobre Sostenibilidad:** Explorar cómo los modelos predictivos pueden contribuir a prácticas agrícolas más sostenibles, reduciendo el uso de insumos químicos y mejorando la eficiencia en la producción. La investigación en esta área puede apoyar la transición hacia una agricultura más ecológica y responsable.

En conclusión, la integración de técnicas de ciencias de datos y modelos predictivos en el sector agrícola presenta oportunidades significativas para mejorar la gestión de costos y la competitividad. Sin embargo, es crucial continuar desarrollando estas tecnologías, capacitar a los actores del sector y fomentar políticas de apoyo para maximizar su impacto y beneficios.

## Bibliografía


[1] C. E. RUDAS, «Alza de precios de insumos para el agro llegó a 29,4%, con fertilizantes de líderes,» *LA REPUBLICA*, p. 1, 15 NOVIEMBRE 2022.

[2] FAO, «Alternative pathways to 2050,» de *The future of food and agriculture*, Roma, Food and Agriculture Organization of the United Nations, 2018, p. 224 pp.

[3] M. J. S. D. E. A. E. L. M. J. A. M. S. &. V. U. Roberts, *The value of plant disease early-warning systems: a case study of USDA's soybean rust coordinated framework,* (No. 1477-2016- 121162), 2006.

[4] J. Z. Mayorga Sánchez y C. Martínez Aldana, «PAUL KRUGMAN Y EL NUEVO COMERCIO INTERNACIONAL,» Universidad Libre, Bogotá, D.C, 2008.

[5] K. B. Eckert y P. V. Britos, «Modelo basado en la Toma Decisiones con Criterios Múltiple para la elección de metodologías de DataScience,» *XX Workshop de Investigación en Ciencias de la Computación*, p. 5, 2018.

[6] M. E y Brynjolfsson, «Big Data: The Management Revolution,» 22 febrero 2012. [En línea]. Available: https://www.researchgate.net/publication/232279314_Big_Data_The_Management_Revolution.

[7] M. A.Waller y S. E.Fawcett, «Data Science,Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management,» *" J. Bus. Logist., vol.34, no. 2,*, p. 77–84, 2013.

[8] C.-V. P. S. C. D. Eleonora Pantano, «Facilitating tourists' decision making through open data analyses: A novel recommender system,» *Tourism Management Perspectives*, pp. -, 2019.

[9] H. D. S. R. W. a. Y. L. P. Tommy Tandera, «Personality Prediction System from Facebook Users,» de *2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI*, Bali, Indonesia, 2017.

[10] G. E. P. J. G. M. &. R. G. C. Box, Time series analysis: Forecasting and control (4th ed.), Hoboken, NJ: John Wiley & Sons, 2008.

[11] D. C. P. E. A. &. V. G. G. Montgomery, Introduction to linear regression analysis (5th ed.), Hoboken NJ: Wiley, 2012.

[12] M. H. N. C. J. &. N. J. Kutner, Applied linear regression models (4th ed.), Boston, MA: McGraw-Hill/Irwin, 2004.

[13] J. M. Wooldridge, Introductory econometrics: A modern approach (6th ed.), Boston: Cengage Learning, 2016.

[14] N. R. &. S. H. Draper, Applied regression analysis (3rd ed.), New York, NY: Wiley-Interscience., 1998.

[15] R. H. &. S. D. S. Shumway, Time series analysis and its applications: With R examples (4th ed.), New York, NY: Springer, 2017.

[16] G. E. P. &. J. G. M. Box, «Time Series Analysis: Forecasting and Control.,» Holden-Day., San Francisco, California, Estados Unidos, 1970.

[17] L. &. C. A. Breiman, «Random Forests.,» University of California, Berkeley, CA., 2004.

[18] L. Breiman, «Random forests 45(1),» Machine Learning, Springer, Dordrecht, 2001.

[19] A. &. W. M. Liaw, «Classification and regression by randomForest R News, 2(3),» R Foundation for Statistical Computing, Vienna, Austria., 2002.

[20] T. K. Ho, «Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition,,» Conference on Document Analysis and Recognition, Montreal, QC, Canada., 1995.

[21] T. &. G. C. Chen, «XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,» International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA., 2016.

[22] T. &. H. T. Chen, «XGBoost: eXtreme Gradient Boosting R package version 0.4-2.,» R Foundation for Statistical Computing, Vienna, Austria., 2015.

[23] T. &. G. C. Chen, «A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,» de *International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016.

[24] T. &. G. C. hen, «XGBoost Documentation,» University of Washington, Seattle, WA, USA., 2016.

[25] J. H. Friedman, «Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5),,» Institute of Mathematical Statistics,, Beachwood, OH, USA., 2001.

[26] T. &. C. G. C. Tianqi Chen, «XGBoost: A Scalable Tree Boosting System.,» Cornell University, Ithaca, NY, USA., 2016.

[27] G. M. Q. F. T. W. T. C. W. M. W. Y. Q. &. L. T.-Y. Ke, «LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30,» Curran Associates, Inc., Long Beach, CA, USA., 2017.

[28] Microsoft., «LightGBM: A Fast, Distributed, High Performance Gradient Boosting (GBDT, GBRT, GBM or MART) Framework,» GitHub repository, Redmond, WA, USA., 2017.

[29] G. &. L. T.-Y. Ke, «LightGBM Documentation,» Microsoft Research, Beijing, China., 2017.

[30] H. &. L. T.-Y. Zhang, «The LightGBM Algorithm: An Overview. Journal of Machine Learning Research 21(137),» Journal of Machine Learning Research,, Redmond, WA, USA, 2018.

[31] J. E. O. Joaquín Amat Rodrigo, «https://www.cienciadedatos.net/,» 01 marzo 2022. [En línea]. Available: https://www.cienciadedatos.net/documentos/py41-forecasting-criptomoneda-bitcoin-machine-learning-python [Último acceso: 11 junio 2023].

[32] J. E. O. Joaquín Amat Rodrigo, «https://www.cienciadedatos.net/,» 01 abril 2023. [En línea]. Available: https://www.cienciadedatos.net/documentos/py48-forecasting-demanda-intermitente [Último acceso: 11 junio 2023].

[33] F. E. Fernández, «https://www.cienciadedatos.net,» 02 Junio 2022. [En línea]. Available: https://www.cienciadedatos.net/documentos/py43-machine-learning-decisiones-financieras [Último acceso: 11 junio 2023].

[34] «Dirección de Impuestos y Aduanas Nacionales,» [En línea]. Available: https://www.dian.gov.co/dian/cifras/Paginas/Bases-Estadisticas-de-Comercio-Exterior-

Importaciones-y-Exportaciones.aspx

[35] Procolombia, «Procolombia - Exportaciones, Turismo, Inversión, Marca país,» [En línea]. Available: https://www.colombiatrade.com.co/preguntas-frecuentes/como-identifico-la-posicion-arancelaria-0

[36] DANE, «DANE,» [En línea]. Available: https://www.dane.gov.co/files/faqs/faq_comex.pdf

[37] J. W. Tukey, «Exploratory Data Analysis,» Addison-Wesley, Boston, Massachusetts, Estados Unidos, 1977.