MODELOS PREDICTIVOS PARA EL PRONÓSTICO DE PRIMAS EMITIDAS Y SINIESTROS INCURRIDOS EN EL RAMO DE RESPONSABILIDAD CIVIL PROFESIONAL DE SEGUROS DEL ESTADO S.A.

Julián Esteban Coronado Gil
Código 925982

Cindy Lorena Salas Valencia
Código 925988

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

Director
Luis Eduardo Girón Cruz

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2026

FICHA RESUMEN

TÍTULO DEL PROYECTO: MODELOS PREDICTIVOS PARA EL PRONÓSTICO DE PRIMAS EMITIDAS Y SINIESTROS INCURRIDOS EN EL RAMO DE RESPONSABILIDAD CIVIL PROFESIONAL DE SEGUROS DEL ESTADO S.A.

1. ÁREA DE TRABAJO: "Ciencia de Datos / Seguros"
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Julián Esteban Coronado Gil y Cindy Lorena Salas Valencia
4. CORREO ELECTRÓNICO:
   juliancourse07@javerianacali.edu.co - loresalas24@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO:
   - Bogotá, 3223213150
   - Calle 76ª#4-16, Neiva-Huila, 3222590371
6. DIRECTOR: Luis Eduardo Girón Cruz
7. VINCULACIÓN DEL DIRECTOR: Docente hora cátedra Facultad Ciencias Económicas y Administrativas
8. CORREO ELECTRÓNICO DEL DIRECTOR: legiron@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): N/A
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): N/A
11. OTROS GRUPOS O EMPRESAS: N/A
12. PALABRAS CLAVE (al menos 5): Primas, Siniestros, Pólizas, Inteligencia de Negocio y Machine Learning.
13. FECHA DE INICIO: Mayo 01 de 2025
14. DURACIÓN ESTIMADA (En meses): Catorce meses (14)

1. RESUMEN:

Este proyecto se enfoca en el desarrollo de modelos predictivos basados en modelos econométricos de series de tiempo y técnicas de machine learning para estimar el importe de las primas y el valor de los siniestros incurridos en el ramo de Responsabilidad Civil profesional de Seguros del Estado S.A., incorporando variables espaciales y temporales. La dinámica del sector asegurador colombiano entre 2016 y 2024, junto con eventos inesperados como la pandemia de COVID-19, evidencia la necesidad de herramientas analíticas más precisas para la toma de decisiones. Inicialmente, se realiza la extracción y transformación de la base de datos transaccional, seguida de un análisis exploratorio que garantiza la calidad, consistencia e integridad de la información, así como la identificación de patrones y anomalías relevantes. Posteriormente, se ajustan los modelos SARIMAX, XGBoost y LightGBM para el importe de las primas, evaluando su desempeño mediante las métricas SMAPE, RMSE y MAE bajo validación Walk-Forward, donde SARIMAX presenta el mejor desempeño con error moderado. Para los siniestros incurridos se implementa el modelo ARIMA debido a la no presencia de estacionalidad en la serie, complementado con XGBoost y LightGBM bajo la misma estrategia de validación, destacándose LightGBM por su capacidad para capturar relaciones no lineales. Finalmente, se desarrolla una aplicación en *Streamlit* que integra proyecciones históricas, *forecast,* presupuesto ajustado por IPC y escenarios de sensibilidad, facilitando el análisis estratégico y la toma de decisiones. En conjunto, el sistema propuesto optimiza la gestión del riesgo, la asignación de recursos y la planificación, mejorando el desempeño financiero y comercial, incrementando la suficiencia actuarial y aportando al fortalecimiento del sector asegurador colombiano mediante el uso de inteligencia artificial.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

ANEXOS

Anexo A. Aplicación desarrollada

El acceso a la aplicación se encuentra disponible en el siguiente enlace:

https://jeks4mxlhvrskvssbhjfwx.streamlit.app/

Anexo B. Código fuente del proyecto

El código fuente completo del proyecto, incluyendo el preprocesamiento de los datos, el entrenamiento y evaluación de los modelos SARIMAX, ARIMA, XGBoost y LightGBM, fue desarrollado en Python mediante cuadernos Jupyter Notebook. Dicho código se encuentra disponible bajo solicitud dirigida a los autores del proyecto.

INTRODUCCIÓN

En los últimos años, el sector asegurador colombiano ha experimentado un crecimiento significativo en términos de primas emitidas y siniestralidad, reflejando una expansión dinámica pero también marcada por periodos de alta volatilidad. En los registros emitidos por fasecolda, entre el año 2016 y 2024, el valor total de primas emitidas pasó de 24,31 a 56,13 billones de pesos, mientras que el valor total de los siniestros incurridos casi se duplicó, alcanzando los 25,5 billones [1]. A pesar de este crecimiento, fenómenos como la pandemia del COVID-19 y la incertidumbre económica han generado fluctuaciones abruptas que dificultan la previsión precisa de la demanda de seguros y la ocurrencia de siniestros.

En el marco del sector asegurador, existen diversos ramos que agrupan las principales líneas de negocio, como personas, daños y seguridad social. Dentro de los seguros de daños, uno de los ramos de mayor relevancia es el de Responsabilidad Civil, destacándose dentro de este la Responsabilidad Civil Profesional, el cual protege a los asegurados frente a los perjuicios ocasionados a terceros. Este ramo es el foco del presente proyecto aplicado, orientado al análisis y predicción del comportamiento del importe de las primas y los siniestros incurridos en la compañía Seguros del Estado S.A. En los últimos años, se ha observado una dinámica cambiante en la siniestralidad y en la demanda de pólizas, lo que evidencia la importancia de implementar herramientas de analítica avanzada y modelos de Machine Learning que permitan anticipar tendencias, optimizar la gestión del riesgo y fortalecer la toma de decisiones estratégicas dentro del portafolio de seguros del ramo de Responsabilidad Civil profesional.

Frente a este panorama, la compañía Seguros del Estado S.A. se enfrenta al reto de tomar decisiones estratégicas bajo condiciones de alta variabilidad y complejidad. Si bien ha mostrado avances importantes en su crecimiento y diversificación de productos, también ha experimentado periodos en los que la siniestralidad ha aumentado de manera considerable, generando desequilibrios entre el valor de las primas y los costos asociados a los siniestros. En este contexto, el estudio se enfoca en el riesgo técnico del seguro, particularmente en los riesgos de tarificación y siniestralidad, mediante la predicción del número de primas, pólizas emitidas y de siniestros incurridos, con el propósito de fortalecer la gestión técnica y contribuir a la estabilidad financiera del ramo de Responsabilidad Civil Profesional.

En este contexto, el uso de técnicas de inteligencia artificial, y en particular de *Machine Learning*, representa una oportunidad clave para optimizar los procesos de toma de decisiones. Los modelos de *Machine Learning* permiten procesar grandes volúmenes de datos históricos y actuales, identificar patrones ocultos y generar predicciones con altos niveles de precisión. Al integrar variables espaciales (como ubicación geográfica y distribución demográfica) y temporales (como ciclos económicos y estacionalidad), estos modelos ofrecen un enfoque integral y adaptativo para anticipar comportamientos en contextos regionales y temporales específicos.

Este proyecto propone el diseño y desarrollo de un sistema predictivo basado en modelos econométricos de series temporales y técnicas de Machine Learning, orientado a la estimación del comportamiento del importe de las primas y de siniestros en el ramo de Responsabilidad Civil Profesional, pertenecientes a la clase de seguros de daños de la compañía Seguros del Estado S.A. Para ello, se realiza un análisis exploratorio de datos, se evalúa la calidad de las variables espaciales y temporales, y se ajustan modelos comparativos con validación walk-forward. Además, se construye una aplicación interactiva en Streamlit, que permite visualizar en tiempo real las predicciones y facilita el monitoreo de los indicadores clave por región y periodo.

La implementación de esta solución permite mejorar la planificación de recursos, reducir la incertidumbre asociada al riesgo de seguro, en particular los riesgos de suscripción y siniestralidad en el ramo por afectaciones en pólizas de Responsabilidad Civil Profesional, aumentar la competitividad frente a otras aseguradoras y contribuir a la estabilidad financiera de la compañía. En última instancia, esta propuesta aporta valor estratégico tanto para Seguros del Estado S.A. como para el fortalecimiento de las capacidades analíticas del sector asegurador colombiano.

I. CONTEXTUALIZACIÓN DEL PROYECTO

A. Definición del problema

1) *Planteamiento del problema*

En Colombia, el crecimiento sostenido del valor total de primas emitidas por el sector asegurador entre los años 2016 y 2024, muestra un aumento de 24,31 billones en 2016 a 56,13 billones en 2024, lo que evidencia una expansión constante a lo largo del tiempo. En particular, durante los años 2021 y 2022 se registró un incremento significativo del valor total en las tasas de crecimiento interanual, con un 15,83 % en 2021 y un 33,91 % en 2022, posiblemente impulsado por la recuperación económica tras la pandemia y una mayor demanda de seguros. Por el contrario, en 2020 el crecimiento fue mínimo, con apenas un 1,02 %, probablemente como consecuencia de la pandemia de COVID-19. Asimismo, se observa una moderación en las variaciones durante los años 2017 (6,98 %), 2018 (4,87 %) y 2019 (7,05 %) [1] .

En el sector asegurador, los siniestros se clasifican en pagados, incurridos e incurridos netos de XL (Excess of Loss o Exceso de Pérdida, de ahora en adelante "XL".), donde este último corresponde a los siniestros incurridos una vez descontada la parte cubierta por el reaseguro de exceso de pérdida. Según las estadísticas publicadas por Fasecolda [1], los siniestros incurridos muestran comportamientos diferenciados por ramo (Daños, Personas y Seguridad Social). Daños presenta un crecimiento sostenido y estable, pasando de 3,92 billones en 2016 a 7,76 billones en 2024; Personas evidencia una evolución más irregular, con un incremento notable de 38,8 % en 2021 y variaciones moderadas posteriores; y Seguridad social concentra el mayor volumen y volatilidad, al aumentar de 4,40 billones en 2016 a 11,55 billones en 2024, destacándose el crecimiento de 46,2 % en 2023. En los tres ramos, los siniestros pagados crecen hasta 2021 y luego caen en 2025 (−39 % en Daños, −29 % en Personas y −27 % en Seguridad social). De igual forma, los siniestros incurridos netos de XL aumentan hasta 2023 y descienden en 2025 (−36 %, −28 % y −31 %, respectivamente), siendo Seguridad social el de mayor volatilidad reciente. Dado el comportamiento representativo y la relevancia del ramo de Daños dentro del sector, el análisis se centra en este ramo, específicamente en la línea de Responsabilidad Civil profesional.

En este contexto, la compañía Seguros del Estado S.A., en el ramo de Responsabilidad Civil Profesional, presenta un comportamiento fluctuante tanto en las primas emitidas como en los diferentes tipos de siniestros. Las primas evidencian una tendencia general de crecimiento, pasando de 0,10 billones en 2016 a 0,18 billones en 2024, con incrementos destacados en 2018 (20,8 %) y 2021 (24,2 %), reflejando una mayor colocación de pólizas y dinamismo comercial. En cuanto a los siniestros, los pagados disminuyeron de 36,30 mil millones en 2016 a 21,76 mil millones en 2020, para luego recuperarse sostenidamente hasta 32,82 mil millones en 2024, en línea con la expansión del mercado y la normalización postpandemia. Los siniestros incurridos mostraron descensos marcados en 2018 (−78,16 %), asociados a un menor nivel de reclamaciones y ajustes técnicos en la suscripción, seguidos de crecimientos sostenidos entre 2021 y 2024, impulsados por la reactivación económica, el incremento de la exposición al riesgo y el mayor dinamismo en la contratación de pólizas empresariales. Por su parte, los incurridos netos de XL presentaron una caída pronunciada entre 2016 y 2018, pero posteriormente repuntaron hasta alcanzar 42,47 mil millones en 2024. En conjunto, este comportamiento refleja una recuperación progresiva del ramo acompañada de un entorno económico en expansión y una mayor intensidad en la frecuencia y severidad de los siniestros [1].

En la nueva era de la tecnología, todo rige con la inteligencia artificial (IA), la cual es implementada en diferentes sectores, realizando tareas desde el análisis de imágenes hasta el reconocimiento de voz por medio de algoritmos. Los modelos de Machine Learnig son una rama de la inteligencia artificial, utilizados para identificar patrones y predecir comportamientos futuros para la toma de decisiones.

Un modelo de Machine Learning aprende mediante los datos para realizar predicciones, y a medida que pasa el tiempo el modelo se hace más preciso y eficaz con mayor número de datos. Estos modelos se pueden clasificar en modelos supervisados que aprenden con patrones identificados de los datos y modelos no supervisados que revelan patrones generales en los datos [2].

En el sector asegurador es muy indispensable contar con modelos predictivos para predecir el importe de primas y el valor de los sisniestros incurridos para poder garantizar la sostenibilidad financiera y una adecuada gestión del riesgo. Los patrones de compra y riesgo están influenciados por variables espaciales (distribución demográfica, condiciones socioeconómicas regionales) y temporales (estacionalidad, ciclos económicos), generando incertidumbre en la asignación de recursos y estrategias de mitigación de riesgos.

En particular, Seguros del Estado S.A. enfrenta la dificultad de anticipar el comportamiento global del importe de las primas y el valor de los siniestros incurridos para el ramo de responsabilidad Civil profesional, incorporando variables espaciales y temporales asociadas a sus sucursales, lo que constituye un desafío dual en la modelación predictiva orientada a la toma de decisiones. Los síntomas de esta situación se reflejan en proyecciones inexactas del comportamiento de las primas y el valor de los siniestros incurridos, coberturas insuficientes frente a eventos de riesgo, inadecuada asignación de recursos humanos y financieros, influenciada por la variabilidad espacial entre sucursales ubicadas en diferentes zonas del país. Esta problemática dentro de la compañía aseguradora es causada por la falta de modelos predictivos avanzados y la alta variabilidad en factores espaciales y temporales. En este sentido, la implementación de modelos de predicción permite proyectar el comportamiento del importe de las primas y del valor de los siniestros incurridos, incorporando variables espaciales y temporales, facilitando la construcción de escenarios para la planificación financiera y la estimación de reservas técnicas, apoyando la toma de decisiones sobre estas variables. De igual forma, posibilitará anticipar tendencias de siniestralidad, estimar la exposición futura al riesgo y optimizar la distribución del modelo comercial, fortaleciendo la toma de decisiones estratégicas y la eficiencia en la gestión de portafolios dentro de la compañía Seguros del Estado S.A. Por ende, si persiste la ausencia de modelos predictivos que incorporen la variabilidad espacial y temporal, aumenta la incertidumbre en la estimación del comportamiento de las primas y la siniestralidad. En este escenario podrían generarse consecuencias significativas en la compañía Seguros del Estado S.A. y en el sector asegurador colombiano, como desajustes entre la oferta y demanda de pólizas, incremento en la siniestralidad no cubierta, pérdida de competitividad frente a compañías que implementan herramientas de inteligencia artificial, ineficiencia operativa por asignación inadecuada de recursos, debilitamiento de la estabilidad financiera y estancamiento o retroceso del crecimiento empresarial.

2) *Formulación del problema*

¿Qué modelos econométricos y de aprendizaje automático presentan el mejor desempeño predictivo para estimar la producción de primas y el valor de siniestros incurridos en los ramos de Responsabilidad Civil Profesional de Seguros del Estado S.A., considerando variables espaciotemporales?

3) Sistematización

- ¿Qué características espaciotemporales y patrones estacionales presentan los datos históricos del importe de primas y el valor de los siniestros incurridos en el ramo de Responsabilidad Civil profesional y cómo puede evaluarse su calidad y consistencia para el análisis predictivo de Seguros del Estado S.A.?
- ¿Qué diferencias presentan los modelos econométricos y de aprendizaje automático en la estimación del importe de primas y del valor de los siniestros incurridos en los ramos de Responsabilidad Civil Profesional de Seguros del Estado S.A., considerando variables espaciotemporales?
- ¿Cómo integrar las proyecciones obtenidas en una herramienta interactiva que facilite el análisis del comportamiento histórico, las proyecciones y la toma de decisiones estratégicas en Seguros del Estado S.A.?

B. *Objetivos del proyecto*

1) *Objetivo general*

Desarrollar modelos predictivos basados en modelos econométricos y en técnicas de *Machine Learning* para la estimación de la producción de primas y el valor de siniestros incurridos, mediante el análisis de variables espaciales y temporales, utilizando SARIMAX/ARIMA, XGBoost y LightGBM, con el fin de optimizar la toma de decisiones estratégicas y la gestión eficiente de los recursos en el ramo de Responsabilidad Civil Profesional de la compañía de Seguros del Estado S.A.

2) *Objetivos específicos:*

- ✓ Analizar los datos históricos de la producción de primas y el valor de siniestros incurridos, aplicando técnicas de análisis exploratorio (EDA) y preprocesamiento para garantizar su calidad, consistencia espaciotemporal y trazabilidad, identificando patrones estacionales y tendencias relevantes.
- ✓ Comparar el desempeño predictivo de modelos econométricos y de aprendizaje automático para el pronóstico de primas emitidas y el valor de los siniestros incurridos en el ramo de Responsabilidad Civil Profesional de Seguros del Estado S.A., incorporando variables espaciotemporales y utilizando validación temporal tipo walk-forward y métricas de evaluación como RMSE, SMAPE y MAE.
- ✓ Diseñar una aplicación interactiva en Streamlit que integre las proyecciones obtenidas y permita visualizar en tiempo real el comportamiento histórico, el forecast mensual, el presupuesto sugerido con ajuste por IPC, y escenarios de sensibilidad, facilitando el análisis estratégico de la información por parte de la dirección comercial y de inteligencia de negocios.

C. *Marco de referencia*

1) *Marco teórico*

a) *Conceptos básicos del contrato de seguro*

En el contexto del análisis de datos para el sector asegurador, resulta esencial comprender ciertos conceptos básicos que permiten interpretar correctamente la información utilizada en los modelos predictivos.

Uno de los conceptos fundamentales en el ámbito asegurador es la **prima**, definida como el precio pactado por el seguro contratado. Según Seguros del Estado, “es la remuneración que recibe la aseguradora para hacerle frente a los riesgos que está amparando en la póliza y es la contraprestación que está obligando a ambas partes a cumplir con lo establecido en el contrato. Es el pago que se hace por adelantado para iniciar el contrato de seguro y en ocasiones puede ser demandada legalmente cuando la aseguradora ha iniciado la cobertura en ciertos riesgos” [3].

Dentro del contrato de seguros, el **siniestro** representa la materialización del riesgo amparado en la póliza. En otras palabras, es el suceso que activa la obligación de indemnización por parte de la aseguradora. Según Seguros del Estado “es la realización del riesgo. Es cuando sucede lo que se está amparando en la póliza y es motivo de indemnización, por ejemplo, un robo, un choque, una enfermedad o accidente, un incendio, etc.” [3]. Por otro lado, los siniestros son medidos en siniestros pagados, siniestros incurridos y siniestros incurridos XL. A continuación, se describen cada una de estas mediciones:

i. Siniestros pagados

De acuerdo con Fasecolda, los siniestros pagados corresponden al valor neto de los siniestros efectivamente pagados por las aseguradoras durante un período determinado [4].

ii. Siniestros incurridos

De acuerdo con la Superintendencia Financiera, los siniestros incurridos corresponden al registro detallado del costo total de los siniestros, incluyendo los valores pagados, las reservas de siniestros pendientes y los ajustes realizados en el tiempo, así como elementos de recuperación como salvamentos, recobros y reintegros. Esta información permite estimar de manera integral la obligación real de la aseguradora frente a los siniestros, considerando además su participación en esquemas de coaseguro [5] .

iii. Siniestros incurridos XL

El reaseguro de exceso de pérdida (Excess of Loss, XL) es un esquema no proporcional en el cual el reasegurador asume únicamente la porción del siniestro que excede un nivel de retención previamente definido por la aseguradora. En este modelo, la aseguradora cubre las pérdidas hasta dicho umbral, mientras que el reasegurador cubre el exceso hasta un límite establecido, reduciendo así la exposición neta de la entidad frente a eventos de alta severidad [6]

En este contexto, los siniestros incurridos neto XL corresponden al valor total de los siniestros incurridos una vez descontada la participación del reaseguro bajo este esquema, reflejando el costo real asumido por la aseguradora.

La **póliza** es el instrumento probatorio por excelencia del contrato de seguro. Es aconsejable leer todas las cláusulas contenidas en la póliza para tener una información completa de sus términos y condiciones. En ella se reflejan las normas que, de forma general, particular o especial, regulan la relación contractual convenida entre el Asegurador y el Asegurado [3]

b) ¿Qué es Machine Learning?

Machine Learning conocido por sus siglas en inglés (ML) hace referencia al aprendizaje automático, siendo este una técnica que permite descubrir conexiones desconocidas dentro de los datos, analizando big data para identificar patrones y tendencias que no se pueden percibir con un análisis estadístico básico. Esta técnica se basa en algoritmos complejos que aprenden de los datos para construir modelos capaces de predecir resultados o clasificar datos [7].

El aprendizaje automático o Machine Learnig es una rama de la inteligencia artificial que da respuesta al comportamiento de los datos, es importante tener en cuenta la diferencia entre estos dos conceptos como lo menciona Chen es su escrito “todo aprendizaje automático es IA, pero no toda IA es aprendizaje automático” [7].

c) Tipos de modelos

Los modelos de aprendizaje automático se pueden clasificar en dos categorías principales: supervisados y no supervisados. En cada caso, se aplican diferentes métricas de evaluación y validación para determinar qué tan bien funciona el modelo según el comportamiento de los datos.

i. Modelos supervisados

Los modelos de aprendizaje supervisado se utilizan para predecir el comportamiento de una variable objetivo, basándose en un conjunto de datos de entrenamiento. Estos modelos aprenden a partir de ejemplos históricos, identificando patrones que les permiten predecir o clasificar resultados cuando se enfrentan a nuevos datos. Entre los

modelos más representativos de aprendizaje supervisado se encuentran:

- *Regresión lineal*

Una de las características de este modelo, es que la variable independiente sea continua y la dependiente sea continua o discreta. Otra característica de la regresión lineal es que la predicción de la variable objetivo se usa con una línea recta. La ecuación de la línea recta es

$$y = ax + b + e \quad (1)$$

siendo $a{-}Intersección$, $b{-}\ Pendiente$, $e{-}Error$**.** Para implementar este modelo, es necesario que ambas variables tengan una relación lineal, y tener en cuenta que este modelo es sensible a valores atípicos [8]. Por esta razón, es la importancia de realizar una limpieza adecuada de los datos antes de correr cualquier modelo. Por otro lado, es importante tener en cuenta que la variable dependiente e independiente pueden ser discretas, dependiendo del enfoque de la investigación.

Es importante señalar que los modelos supervisados utilizan la regresión lineal cuando se busca explicar o predecir una variable dependiente en función de un conjunto de variables independientes.

- *Redes neuronales LONG SHORT-TERM MEMORY (LSTM)*

El modelo de redes neuronales conocido como Memoria a Largo y Corto Plazo (LSTM) por sus siglas en inglés, se utiliza principalmente en el aprendizaje supervisado, aunque también puede aplicarse en contextos no supervisados o auto-supervisados, dependiendo del tipo de investigación.

Las redes neuronales LSTM (Long Short-Term Memory) representan una variante de las redes recurrentes que incorporan celdas de memoria diseñadas para conservar información durante periodos prolongados, abordando así el problema del desvanecimiento del gradiente característico de las RNN tradicionales. Estas redes gestionan la información tanto a corto como a largo plazo: por un lado, las funciones de activación procesan la memoria reciente del modelo, y por otro, los pesos sinápticos evolucionan con la experiencia a través del entrenamiento. Inicialmente, estos pesos se asignan de forma aleatoria y se ajustan mediante la optimización de una función de pérdida, como la entropía cruzada, aunque sus actualizaciones son más lentas en comparación con las funciones de activación [9].

- *Series temporales:*

Una serie temporal es una secuencia de datos recolectados en puntos equidistantes en el tiempo. Estas observaciones pueden ser de una sola variable, en cuyo caso se denominan series univariantes, o de múltiples variables registradas simultáneamente, conocidas como series multivariantes. En ambos casos, el objetivo es modelar el comportamiento de dichas observaciones a lo largo del tiempo para analizar patrones, tendencias o realizar predicciones [10]. En el caso del presente proyecto se centra en los modelos de series de tiempo univariado.

En términos matemáticos, una serie univariante puede representarse como $y_1, y_2, \dots, y_N$, donde y_t es el valor observado en el instante t, y N es el número total de observaciones [10].

- *Modelos ARIMA(p,d,q)*

Los modelos *ARIMA (Autoregressive Integrated Moving Average)* y *SARIMA (Seasonal ARIMA)* constituyen herramientas fundamentales en el análisis y pronóstico de series temporales. Desarrollados por *Box y Jenkins* [11] , estos modelos permiten capturar las dependencias dinámicas de los datos en el tiempo, incluyendo patrones estacionales y tendencias. Según Hyndman y Athanasopoulos [12], su utilidad radica en su flexibilidad para describir procesos estocásticos y en su capacidad predictiva comprobada en campos económicos, industriales y científicos.

Modelo

El modelo autorregresivo de orden , denotado como , se define como:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \quad (2)$$

donde:

X_t: valor actual de la serie,
ε_t: error aleatorio con media cero y varianza constante,
ϕ_i: coeficientes autorregresivos.

En notación de operador de rezago B , el modelo se expresa como:

$$\phi_1(B) X_t = \varepsilon_t \quad (3)$$

con

$$\phi_1(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_P B^P \quad (4)$$

Modelo $MA(q)$

El modelo de media móvil de orden q, $MA(q)$, se representa por:

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{(t-1)} - \theta_2 \varepsilon_{(t-2)} - \cdots - \theta_q \varepsilon_{(t-q)} \quad (5)$$

donde θ_j son los parámetros de media móvil y ε_t los errores aleatorios (ruido blanco).

Modelo ARIMA(p, d, q)

El modelo ARIMA combina las partes autorregresivas (AR) y de media móvil (MA), incorporando una diferenciación d para eliminar la tendencia y volver estacionaria la serie. Su expresión general es:

$$\phi(B)(1 - B)^d X_t = \theta(B)\varepsilon_t \quad (6)$$

Donde $(1 - B)^d$ representa el operador de diferenciación aplicado veces.

- *Modelo SARIMA*$(p, d, q)(P, D, Q)_s$

El modelo SARIMA extiende ARIMA incorporando la estacionalidad mediante parámetros adicionales:

$$\Phi(B^s)\phi(B)(1 - B)^d(1 - B^s)^D X_t = \Theta(B^s)\theta(B)\varepsilon_t \quad (7)$$

donde:

s : periodicidad estacional (por ejemplo, 12 para datos mensuales),
(P, D, Q): órdenes estacionales de las partes autorregresiva, de diferenciación y de media móvil.

- *Metodología Box–Jenkins*

El método *Box–Jenkins* propone un ciclo iterativo para ajustar modelos ARIMA y SARIMA, compuesto por cuatro etapas principales [11], [12]:

- *Identificación:* determinar los valores de p, d, q, P, D, Q mediante el análisis de los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF).
- *Estimación:* calcular los parámetros del modelo utilizando métodos de máxima verosimilitud o mínimos cuadrados.
- *Diagnóstico:* verificar si los residuos del modelo se comportan como ruido blanco mediante pruebas como el test de *Ljung–Box*.
- *Pronóstico:* generar predicciones y construir intervalos de confianza basados en el modelo ajustado.

Esta metodología, conocido como el ciclo *Box–Jenkins*, permite una evaluación sistemática y refinamiento iterativo del modelo para lograr un ajuste adecuado a los datos temporales. La Figura 1 describe el proceso de identificación, estimación, diagnóstico y predicción de modelos de series temporales, incorporando los parámetros $\boldsymbol{d}$ (diferenciación para estacionariedad en media), $\boldsymbol{D}$ (diferenciación estacional) y $\boldsymbol{\lambda}$ (parámetro de transformación para estabilizar la varianza).

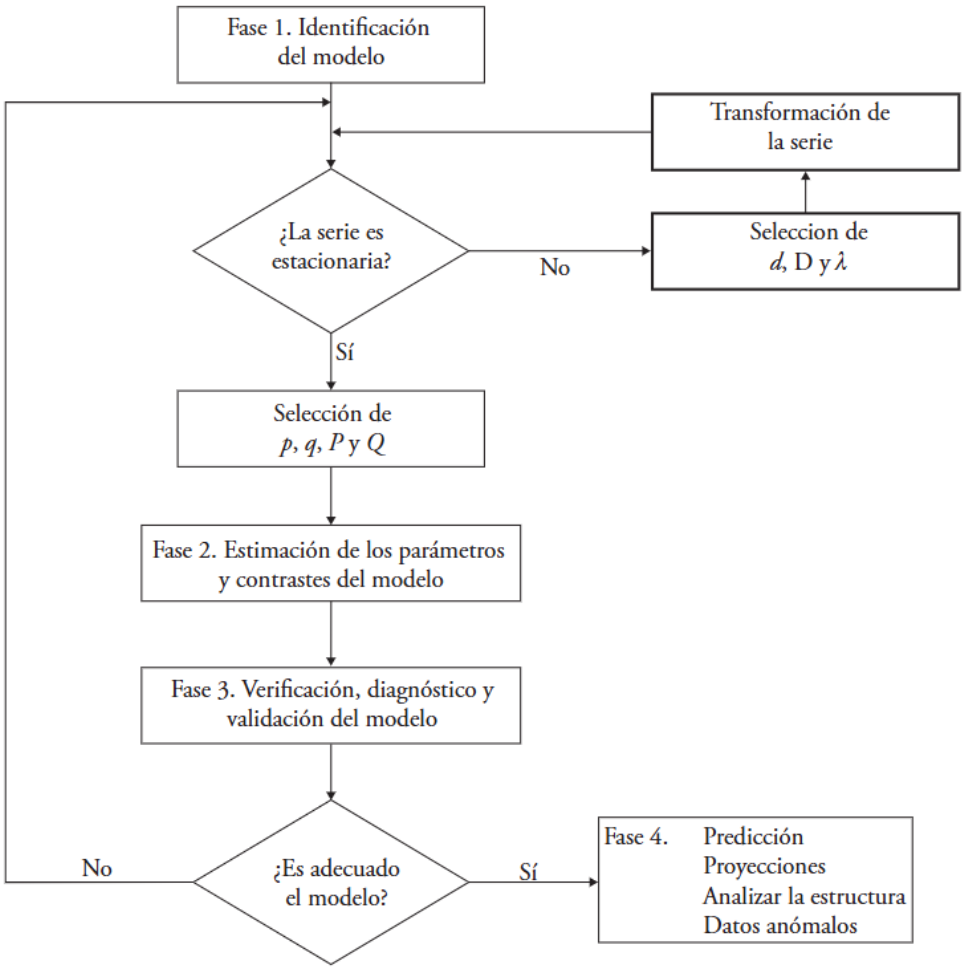


Fig.1. Ciclo del método *Box–Jenkins* para modelos ARIMA/SARIMA. Fuente: Adaptado de S. Luis-Rojas et al [13].

- *Identificación, estimación y diagnóstico:*

Identificación

Se busca la estacionariedad mediante diferencias; se observan las funciones *ACF* y *PACF* para determinar los órdenes.

Estimación

Los parámetros se obtienen mediante Máxima Verosimilitud *(MLE)* o Método de Mínimos Cuadrados.

Diagnóstico

Se evalúan los residuos del modelo; deben comportarse como ruido blanco (media ≈ 0, varianza constante).

El modelo con mejor *AIC (Akaike Information Criterion)* o *BIC (Bayesian Information Criterion)* se considera el más adecuado [11], [12]

- *Ejemplo de formulación de modelos*

$$ARIMA(1,1,1):$$

$$(1-\phi_1 B)(1-B)X_t = (1-\theta_1 B)\,\varepsilon_t \tag{8}$$

$$SARIMA(0,1,1)\times(0,1,1)_{12}$$

$$(1-B)(1-B^{12})X_t = (1-\theta_1 B)(1-\Theta_1 B^{12})\varepsilon_t \tag{9}$$

Este modelo es común para datos mensuales con estacionalidad anual.

- *Buenas prácticas y advertencias:*

Los modelos *ARIMA* y *SARIMA* suponen que la estructura temporal de la serie es estable; por tanto, si existen cambios estructurales o eventos atípicos, las predicciones pueden verse afectadas [11], [12]. En presencia de heterocedasticidad en los residuos, se recomienda emplear modelos *ARCH* o *GARCH*, los cuales permiten modelar la varianza condicional variable en el tiempo [14] Asimismo, para datos con relaciones no lineales o con múltiples variables exógenas, los modelos híbridos o de *Machine Learning* pueden complementar el análisis tradicional de series temporales [15].

- *XGBoost*

Hace referencia a *(Extreme Gradient Boosting)*, es método de aprendizaje automático de código abierto y diseñado para funcionar en entornos distribuidos. Utiliza árboles de decisión combinados mediante la técnica de *gradient boosting*, un enfoque de aprendizaje supervisado que se basa en la optimización por descenso de gradiente. Destaca por su alta velocidad, eficiencia y excelente rendimiento al trabajar con grandes volúmenes de datos [16].

- *Función objetivo regularizada en XGBoost*

En un conjunto de datos compuesto por n observaciones y m variables explicativas, representado como $D = \{x_i, y_i\}$, donde cada x_i corresponde a un vector de características y y_i al valor objetivo, los modelos basados en árboles tipo ensemble, como *XGBoost*, generan predicciones mediante la suma de múltiples funciones aditivas. En este enfoque, el modelo final se construye a partir de la combinación de K árboles de decisión, donde cada uno contribuye

progresivamente a mejorar la precisión del modelo [17].

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \epsilon \mathcal{F} \tag{10}$$

En el modelo XGBoost, el conjunto

$$\mathcal{F} = \{f(x) = w_{q(x)}(q: \mathbb{R}^m \to \mathrm{T}, w \epsilon \mathbb{R}^{\mathrm{T}})\} \tag{11}$$

representa el espacio de árboles de regresión tipo CART. En esta expresión, q define la estructura del árbol encargada de asignar cada observación a una hoja específica, mientras que w corresponde al conjunto de pesos asociados a las hojas terminales del árbol. Asimismo, T representa el número total de hojas, donde cada una contiene un valor continúo utilizado para generar la predicción final de cada observación de acuerdo con las reglas de decisión establecidas por el árbol [17].

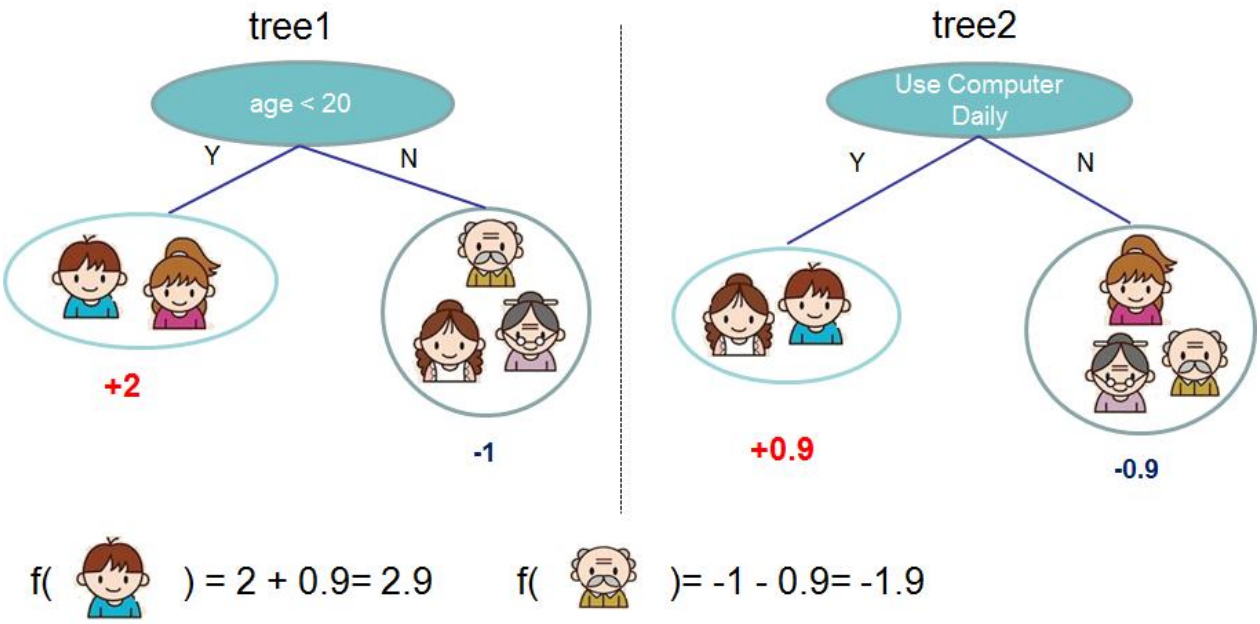


Fig.2. Modelo de Conjunto de Árboles. La predicción final para un ejemplo dado es la suma de las predicciones de cada árbol. Fuente: Tianqi Chen y Carlos Guestrin [17].

- *LightGBM*

LightGBM es un algoritmo de *gradient boosting* desarrollado por Microsoft, perteneciente a la familia de modelos basados en árboles de decisión potenciados (*Gradient Boosted Decision Trees* o *GBDT*), junto con algoritmos como *XGBoost* y *CatBoost*. Este modelo se destaca por su alta eficiencia computacional, rapidez de entrenamiento y buen desempeño en diversas tareas de modelado predictivo. Además, *LightGBM* puede emplearse tanto en problemas de clasificación como de regresión [18].

LightGBM es un algoritmo de *gradient boosting* que construye modelos de forma secuencial, donde cada árbol de decisión busca corregir los errores del anterior. Se caracteriza por utilizar un aprendizaje basado en histogramas y una estrategia de crecimiento por hojas (*leaf-wise*), lo que permite acelerar el entrenamiento, reducir el consumo de memoria y mejorar la precisión predictiva, especialmente en grandes volúmenes de datos [18].

Este modelo puede emplearse tanto en tareas de clasificación como de regresión y ha demostrado un alto desempeño en aplicaciones relacionadas con predicción, análisis de series temporales, detección de fraudes y sistemas de recomendación. Entre sus principales ventajas destacan su rapidez, eficiencia computacional y capacidad para manejar grandes conjuntos de datos. Sin embargo, debido a la complejidad de sus árboles, puede presentar riesgo de sobreajuste cuando el modelo aprende patrones específicos del conjunto de entrenamiento y pierde capacidad de generalización sobre nuevos datos. Para reducir este problema, se suelen aplicar técnicas como validación cruzada, regularización, limitación de profundidad de los árboles y *early stopping* [18].

El uso nativo de *LightGBM* ofrece ventajas relacionadas con una mayor eficiencia en el manejo de grandes volúmenes de datos, tiempos de entrenamiento más rápidos y un control más detallado sobre los parámetros del modelo. Además, permite acceder a funcionalidades avanzadas, como el entrenamiento continuo y configuraciones específicas que no siempre están disponibles en el *API* de *scikit-learn*. No obstante, el uso de *LightGBM* mediante el entorno de *scikit-learn* puede resultar más práctico para la integración con herramientas como *pipelines* y procesos de optimización de hiperparámetros [18].

d) Evaluación de modelos

Una de las fases más importantes, es la evaluación del modelo en el cual se implementan diferentes métricas entre ellas tenemos:

i. SMAPE

Es una métrica, traducida por Error Porcentual Absoluto Medio Simétrico. Es utilizada para medir con precisión los modelos de predicción, especialmente en modelos de series temporales [19]. Esta métrica se calcula mediante la ecuación:

$$SMAPE = \frac{100\%}{N}\sum_{i=0}^{N-1}\frac{2 * |y_i - \hat{y}_i|}{|y| + |\hat{y}|} \tag{12}$$

ii. RMSE Y MAE

Corresponden a métricas de evaluación que cuantifican el error del modelo, representando la diferencia entre los valores predichos y los valores reales de la variable analizada. Estas métricas se calculan de la siguiente manera:

$$RMSE = \sqrt{\frac{\sum (y_i - y_p)^2}{n}} \tag{13}$$

$$MAE = \frac{|(y_i - y_p)|}{n} \tag{14}$$

iii. Validación

En el contexto de machine learning o modelos predictivos, la validación es el proceso mediante el cual se evalúa el desempeño de un modelo con datos que no fueron utilizados durante su entrenamiento, con el objetivo de verificar su capacidad para generalizar y predecir correctamente sobre nuevos datos. Uno de los métodos implementados en series temporales para la validación es:

- *Walk-forward*

Esta técnica comúnmente utilizada en la validación de modelos de series temporales, la cual consiste en entrenar un modelo con un periodo de datos históricos (in-sample, IS) y validarlo con un conjunto siguiente (out-of-sample, OOS), desplazando progresivamente ambos conjuntos a lo largo del tiempo. Esta estrategia permite simular un entorno real de predicción, adaptando el modelo a nuevos datos conforme avanza el tiempo [20].

El aprendizaje automático se ha consolidado como una herramienta fundamental para el análisis y modelado de datos, especialmente en contextos donde la complejidad y el volumen de información superan las capacidades de los métodos estadísticos tradicionales. A través de modelos supervisados como la regresión lineal, las redes neuronales LSTM o el algoritmo XGBoost, es posible generar predicciones precisas basadas en patrones históricos. En particular, las series temporales representan un campo clave dentro del análisis predictivo, al capturar la evolución de variables a lo largo del tiempo. Finalmente, el uso de métricas como SMAPE o el índice de Gini, junto con técnicas de validación como el método walk-forward, garantiza una evaluación rigurosa y realista del desempeño de los modelos desarrollados.

2) Antecedentes

En los últimos años, el mercado asegurador ha atravesado un proceso de endurecimiento significativo, caracterizado por condiciones más estrictas de suscripción y un aumento en las primas, especialmente en sectores con alta presión siniestral. Aunque algunos segmentos han comenzado a percibir una estabilización, otros continúan enfrentando dificultades debido a la baja apetencia de riesgo por parte de las aseguradoras [21].

Según el Informe sobre la Situación del Mercado Asegurador 2023 de WTW, actualmente puede hablarse de un mercado asegurador a dos velocidades: mientras que los clientes con buena calidad de riesgos comienzan a experimentar mejoras, las industrias consideradas agravadas permanecen en un entorno de mercado duro [21]. Esta dualidad plantea nuevos desafíos en la gestión de riesgos, la emisión de pólizas y la toma de decisiones por parte de las aseguradoras, lo cual requiere soluciones técnicas y predictivas más precisas.

Es necesario y rentable implementar modelos estadísticos o de machine learning para predecir el número de siniestros y pólizas emitidas durante un determinado periodo de tiempo, ya que estos modelos permiten anticiparse a la demanda del mercado, optimizar la gestión del riesgo y mejorar la toma de decisiones estratégicas en las aseguradoras. Un claro ejemplo de esta aplicación se encuentra en el proyecto titulado "Modelo Predictivo de Siniestros en Seguros Dentales" [22], desarrollado para la aseguradora AXA Colpatria Falabella. Este estudio buscó resolver la falta de herramientas predictivas que permitieran anticipar el valor total de los siniestros dentales al cierre del mes siguiente. Para ello, se utilizó un enfoque de regresión lineal simple alimentado con datos históricos de pagos realizados durante los 30 días previos, demostrando que incluso modelos estadísticos básicos pueden proporcionar estimaciones útiles para planear y optimizar recursos.

Este antecedente valida la pertinencia de implementar modelos predictivos en el sector asegurador, aunque su alcance se limita a un solo tipo de seguro (odontológico) y a una única variable (valor de siniestros). En contraste, el presente proyecto busca ampliar ese enfoque mediante el uso de modelos de machine learning más avanzados, capaces de predecir tanto el número de siniestros como las pólizas emitidas en diferentes líneas de seguros, considerando múltiples variables explicativas y contextuales.

Otro estudio relevante es el desarrollado por Rivera Jiménez y Pineda Ríos, en el cual se aborda el diseño de un modelo estadístico para pronosticar la tasa de siniestralidad de un seguro de desempleo en Colombia [23]. Dado que el desempleo es una variable clave para la economía nacional y afecta directamente la estabilidad financiera de los hogares, el trabajo propone el uso de métodos estadísticos, como modelos lineales dinámicos y validaciones con *backtesting*, para estimar el nivel de exposición que tendría este tipo de producto asegurador frente a diferentes escenarios, incluyendo crisis económicas. Los indicadores RMSE y $\hat{R}$ (*Rhat*) fueron utilizados para comparar y seleccionar el mejor modelo, y para verificar la convergencia del algoritmo. Este estudio contribuye al análisis predictivo dentro del sector asegurador, aunque enfocado en seguros de desempleo.

El presente proyecto se diferencia principalmente en el tipo de cobertura (siniestros generales vs. desempleo), pero coincide en el objetivo de anticipar el riesgo y mejorar la capacidad de respuesta de las aseguradoras mediante modelos estadísticos y de *machine learning*.

En este contexto, el presente proyecto busca aportar al campo mediante el desarrollo de un modelo predictivo más robusto, capaz de estimar el valor de los siniestros incurridos como el importe de las primas, mediante la implementación de modelos econométricos como SARIMAX/ARIMA y técnicas de aprendizaje automático supervisado como XGBoost y LightGBM. Para la evaluación del desempeño de los modelos se emplean métricas especializadas como SMAPE, RMSE y MAE. Este enfoque representa un avance frente a los antecedentes revisados y una contribución relevante para el fortalecimiento de la gestión estratégica en las entidades aseguradoras.

En las compañías aseguradoras, predecir la demanda de pólizas y siniestros, es muy importante para la toma de decisiones, ya que permite anticipar los ingresos por primas y los egresos por indemnizaciones, facilitando una gestión financiera más eficiente, la adecuada constitución de reservas técnicas, y el diseño de estrategias comerciales y de suscripción basadas en el riesgo real. Además, mejora la capacidad de respuesta ante eventos adversos y contribuye a la estabilidad y sostenibilidad de las aseguradoras en el largo plazo. Es por tal motivo que Ortega [24] establece en su estudio la estimación del número de pólizas de seguro previsional y su frecuencia en el año 2020, vinculada al pago si el asegurado sobrevive al período establecido. Para ello, son utilizados los modelos de regresión cuasi-Poisson y redes neuronales recurrentes.

Por otro lado, el autor menciona el uso de datos históricos de siniestros pagados, pertenecientes a los diez años anteriores al año 2020. Dado que los siniestros del seguro previsional pueden pagarse mucho después de haber ocurrido, fue necesario el cálculo de una estimación intermedia llamada IBNR (por sus siglas en inglés), que se refiere a los siniestros que ya ocurrieron pero que aún no han sido reportados [24].

Un siniestro que haya ocurrido en 2019 podría no ser reportado ni pagado sino hasta el año 2024. Por ello, fue necesario realizar inicialmente la estimación de los montos IBNR correspondientes a las pólizas emitidas entre 2010 y 2019, basándose en un triángulo de siniestros incurridos para llevar a cabo dicho cálculo [24].

En este proyecto Ortega [24] establece una comparación entre los modelos cuasi-Poisson y Redes Neuronales recurrentes RNN, para calcular el número de pólizas emitidas IBNR y su frecuencia. Los resultados son menores en los años más recientes desde el 2015 al 2019, cuando se utiliza el modelo RNN en comparación con cuasi-Poisson, posiblemente a la falta de variables predictoras. Por tal razón, el autor recomienda para estudios futuros la integración de más variables predictoras, como características esenciales de los asegurados.

En el ámbito actuarial, la predicción de primas y siniestros se ha abordado históricamente con modelos de series de tiempo tipo *ARIMA* y *SARIMA*, y su extensión con regresores exógenos (*SARIMAX*) permite incorporar señales macroeconómicas y variables de negocio como el IPC, los ciclos de renovación o la estacionalidad comercial, mejorando la capacidad explicativa ante fluctuaciones de corto plazo. La formulación del *SARIMAX* dentro del marco de modelos de espacio de estados proporciona una estructura estadística sólida para la estimación por máxima verosimilitud, el filtrado y el suavizado, como lo plantean Durbin y Koopman en su obra de referencia sobre series de tiempo y espacio de estados [25]. Su implementación en Python, documentada oficialmente en la librería statsmodels, facilita el ajuste, diagnóstico e interpretación de modelos estacionales con variables exógenas [26]. En el contexto del sector asegurador, estudios como el de Cummins han demostrado la eficacia de los modelos *ARIMA* y *SARIMAX* para pronosticar los costos de siniestros pagados, mostrando su utilidad para la planificación financiera y el control de suficiencia técnica en las reservas [27].

De forma complementaria, la necesidad de operacionalizar la analítica predictiva en herramientas accesibles para los equipos de negocio ha impulsado el uso de aplicaciones web ligeras desarrolladas en Python, entre las que destaca Streamlit por su facilidad para integrar modelos de predicción y visualizaciones interactivas. Investigaciones recientes, como las de Nápoles-Duarte et al. [28] y Fabrizzio et al. [29], destacan el uso de *Streamlit* como plataforma académica y aplicada para el despliegue de modelos científicos, simulaciones y sistemas de soporte a la decisión, gracias a su capacidad para conectar modelos de *Machine Learning* con interfaces web dinámicas. En este sentido, su adopción en proyectos de pronóstico y alertas tempranas en el ámbito asegurador permite trasladar los resultados analíticos como los generados mediante *SARIMAX* o *XGBoost* hacia una visualización interactiva en tiempo real, fortaleciendo la transparencia, la trazabilidad y la utilidad práctica de la analítica en la gestión del riesgo.

En resumen, los trabajos revisados evidencian la creciente importancia del uso de modelos estadísticos y de *machine learning* en el sector asegurador, especialmente para la predicción de siniestros y la evaluación del riesgo. Estudios aplicados en seguros odontológicos y de desempleo han demostrado que estos enfoques permiten anticipar comportamientos del mercado y mejorar la toma de decisiones operativas y estratégicas. Sin embargo, la mayoría de estas investigaciones se han enfocado en líneas de negocio específicas o en modelos de complejidad limitada, lo cual deja espacio para propuestas que amplíen la cobertura y la sofisticación del análisis predictivo.

II. ANÁLISIS EXPLORATORIO Y PREPROCESAMIENTO DE LOS DATOS HISTÓRICOS DEL IMPORTE DE LAS PRIMAS Y DEL VALOR DE LOS SINIESTROS INCURRIDOS EN SEGUROS DEL ESTADO S.A.– OBJETIVO 1

En esta etapa, el análisis se realiza de manera independiente para cada variable, con el propósito de identificar tendencias, patrones temporales, estacionalidad, anomalías y características relevantes de los datos históricos. En consecuencia, no se busca establecer una comparación directa ni una relación de dependencia entre el importe de las primas y el valor de los siniestros incurridos, ya que cada serie temporal será modelada y pronosticada de forma individual conforme a sus propias características estadísticas y temporales.

La base de datos cuenta con variables de segmentación geográfica (sucursal); sin embargo, para efectos del presente estudio, el análisis y modelado se realizan a nivel agregado, construyendo una única serie temporal para cada variable objetivo (primas y siniestros). La variable de sucursal no se utiliza para la estimación de modelos diferenciados, sino únicamente como variable descriptiva del contexto de los datos.

Adicionalmente, este enfoque responde al interés de la compañía Seguros del Estado S.A. por mejorar la capacidad de pronóstico agregado de las primas y la siniestralidad, con el fin de fortalecer la planeación financiera, la gestión del riesgo y la toma de decisiones estratégicas a nivel corporativo, sin desagregar el análisis por sucursal en esta etapa del estudio.

A. Recolección y preprocesamiento de datos

En este capítulo se describe el proceso de recolección y preprocesamiento utilizados en este Proyecto Aplicado. Para ello, se recopilan los registros históricos del importe de las primas y los valores de los siniestros incurridos de la compañía de Seguros del Estado S.A. Posteriormente, se realiza un análisis exploratorio de los datos (EDA) y se aplica técnicas de limpieza y transformación orientadas a garantizar la calidad, la consistencia temporal y trazabilidad de la información, con el fin de preparar el conjunto de datos para la etapa posterior del proyecto que corresponde al modelado.

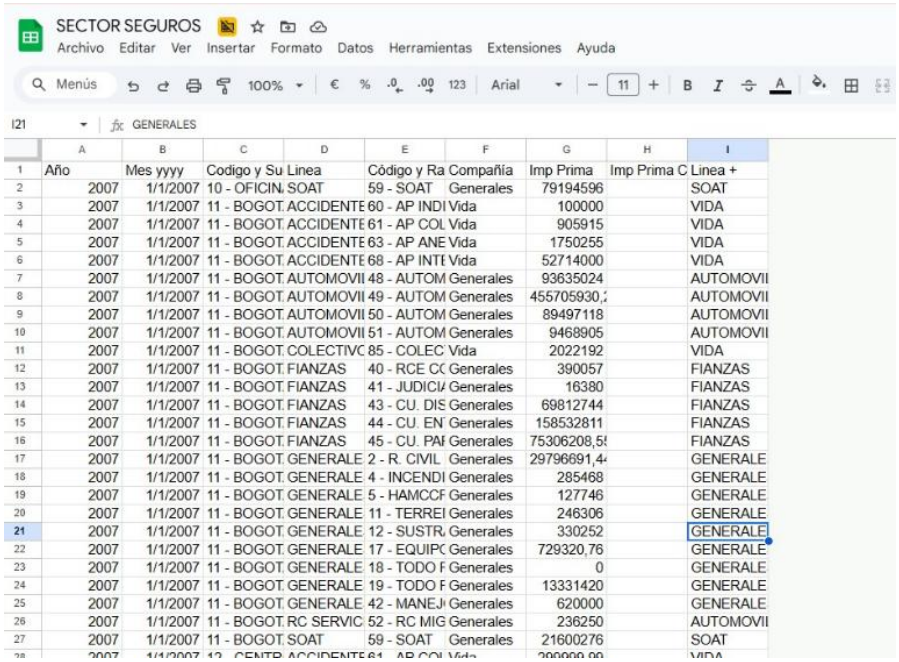
A través de este análisis que se realiza en este capítulo, se busca identificar patrones estacionales, valores atípicos, y tendencias relevantes según los atributos Año, Fecha, Sucursal, Línea, Compañía, Importe prima (Ventas de pólizas de seguros), presupuesto y siniestralidad, proporcionando una base sólida para el modelado predictivo posterior.

1) Análisis descriptivo del Importe de las primas

a) ETL (Extracción, Transformación y Carga de datos)

En este proyecto se construye un conjunto de indicadores de emisión a partir de la base transaccional de pólizas, integrando tres fuentes internas del Data Warehouse: la tabla de operaciones (hechos de emisión histórica), el catálogo de ramos (dimensión de clasificación comercial) y el catálogo de sucursales (dimensión geográfica y organizacional). En el informe académico, estas fuentes se presentan como DW_Operaciones, DW_Ramos y DW_Sucursales, sin hacer referencia a su denominación interna real. A partir de estas fuentes se lleva a cabo un proceso de extracción, transformación y consolidación orientado a generar un dataset analítico con periodicidad mensual, ramos y líneas comerciales estandarizados, integración geográfica por sucursal y cifras de primas emitidas y recaudadas. Este dataset sirve posteriormente como base para el modelado estadístico, el análisis de comportamiento y la construcción de la plataforma analítica del proyecto.

La Figura 3 presenta un extracto del dataset original de primas emitidas, donde se observan variables temporales, comerciales y económicas como Año, Fecha, Línea, Compañía e Importe de Prima. Esta vista corresponde a los datos antes del proceso de limpieza y estandarización realizado durante el preprocesamiento.



	A	B	C	D	E	F	G	H	I
1	Año	Mes yyyy	Codigo y Su	Linea	Código y Ra	Compañía	Imp Prima	Imp Prima C	Linea +
2	2007	1/1/2007	10 - OFICIN	SOAT	59 - SOAT	Generales	79194596		SOAT
3	2007	1/1/2007	11 - BOGOT	ACCIDENTE	60 - AP INDI	Vida	100000		VIDA
4	2007	1/1/2007	11 - BOGOT	ACCIDENTE	61 - AP COL	Vida	905915		VIDA
5	2007	1/1/2007	11 - BOGOT	ACCIDENTE	63 - AP ANE	Vida	1750255		VIDA
6	2007	1/1/2007	11 - BOGOT	ACCIDENTE	68 - AP INTE	Vida	52714000		VIDA
7	2007	1/1/2007	11 - BOGOT	AUTOMOVII	48 - AUTOM	Generales	93635024		AUTOMOVII
8	2007	1/1/2007	11 - BOGOT	AUTOMOVII	49 - AUTOM	Generales	455705930,2		AUTOMOVII
9	2007	1/1/2007	11 - BOGOT	AUTOMOVII	50 - AUTOM	Generales	89497118		AUTOMOVII
10	2007	1/1/2007	11 - BOGOT	AUTOMOVII	51 - AUTOM	Generales	9468905		AUTOMOVII
11	2007	1/1/2007	11 - BOGOT	COLECTIVC	85 - COLEC	Vida	2022192		VIDA
12	2007	1/1/2007	11 - BOGOT	FIANZAS	40 - RCE CC	Generales	390057		FIANZAS
13	2007	1/1/2007	11 - BOGOT	FIANZAS	41 - JUDICI	Generales	16380		FIANZAS
14	2007	1/1/2007	11 - BOGOT	FIANZAS	43 - CU. DIS	Generales	69812744		FIANZAS
15	2007	1/1/2007	11 - BOGOT	FIANZAS	44 - CU. EN	Generales	158532811		FIANZAS
16	2007	1/1/2007	11 - BOGOT	FIANZAS	45 - CU. PAI	Generales	75306208,5		FIANZAS
17	2007	1/1/2007	11 - BOGOT	GENERALE	2 - R. CIVIL	Generales	29796691,4		GENERALE
18	2007	1/1/2007	11 - BOGOT	GENERALE	4 - INCENDI	Generales	285468		GENERALE
19	2007	1/1/2007	11 - BOGOT	GENERALE	5 - HAMCCF	Generales	127746		GENERALE
20	2007	1/1/2007	11 - BOGOT	GENERALE	11 - TERREI	Generales	246306		GENERALE
21	2007	1/1/2007	11 - BOGOT	GENERALE	12 - SUSTR	Generales	330252		GENERALE
22	2007	1/1/2007	11 - BOGOT	GENERALE	17 - EQUIPC	Generales	729320,76		GENERALE
23	2007	1/1/2007	11 - BOGOT	GENERALE	18 - TODO F	Generales	0		GENERALE
24	2007	1/1/2007	11 - BOGOT	GENERALE	19 - TODO F	Generales	13331420		GENERALE
25	2007	1/1/2007	11 - BOGOT	GENERALE	42 - MANEJ	Generales	620000		GENERALE
26	2007	1/1/2007	11 - BOGOT	RC SERVIC	52 - RC MIG	Generales	236250		AUTOMOVII
27	2007	1/1/2007	11 - BOGOT	SOAT	59 - SOAT	Generales	21600276		SOAT

Fig.3. Vista inicial de la base transaccional de primas: Variables temporales, comerciales y Económicas de Seguros Del Estado S.A.
Fuente: Base de datos histórica de Seguros del Estado S.A.

i. Carga del dataset para el análisis

Para el análisis exploratorio se conecta la base de datos a Python mediante la librería pandas, permitiendo importar los datos desde archivos locales en formato Excel o CSV, así como desde hojas de cálculo en Google Sheets, garantizando flexibilidad en la obtención y manejo de la información. Por otro lado, se lleva a cabo la normalización de los atributos para estandarizar y uniformar los datos, con el fin de facilitar el análisis, evitar errores en el procesamiento y asegurar la coherencia entre los atributos al momento de aplicar el modelo.

En la TABLA I se puede visualizar la base de datos original con 528.198 registros y 9 atributos sin normalizar llamados (Año, Mes yyy, Codigo Sucursal, Línea, Código y Ramo, Compañía, Imp_Prima, Imp_Prima_Cuota y Línea.

TABLA I
VISTA PREVIA DE LOS DATOS HISTÓRICOS DE PRIMAS EMITIDAS DE SEGUROS DEL ESTADO S.A.

	Año	Mes yyyy	Codigo y Sucursal	Linea	Código y Ramo	Compañía	Imp Prima	Imp Prima Cuota	Linea +
0	2007	1/1/2007	10 - OFICINA PRINCIPAL SOAT	SOAT	59 - SOAT	Generales	79194596	NaN	SOAT
1	2007	1/1/2007	11 - BOGOTA	ACCIDENTES PERSONALES	60 - AP INDIVIDUAL	Vida	100000	NaN	VIDA
2	2007	1/1/2007	11 - BOGOTA	ACCIDENTES PERSONALES	61 - AP COLECTIVOS	Vida	905915	NaN	VIDA
3	2007	1/1/2007	11 - BOGOTA	ACCIDENTES PERSONALES	63 - AP ANEXO AUTOS	Vida	1750255	NaN	VIDA

4	2007	1/1/2007	11 - BOGOTA	ACCIDENTES PERSONALES	68 - AP INTEGRAL E	Vida	52714000	NaN	VIDA

Fuente: Base de datos histórica de Seguros del Estado S.A.

ii. Normalización de los atributos para el análisis

En la TABLA II se presentan los primeros siete registros del conjunto de datos normalizado. Se observa que el número de registros se reduce a 496.158, debido a que la base de datos corresponde a una fuente transaccional que se encuentra en constante actualización. Por esta razón, se establece un punto de corte temporal hasta el 31 de diciembre de 2025, con el propósito de delimitar el periodo de análisis. Asimismo, el conjunto de datos se reduce a ocho atributos principales: Año, Fecha, Sucursal, Ramo, Línea, Compañía, Imp_prima y Presupuesto.

A partir de este proceso se estandarizan los nombres de los atributos para garantizar coherencia en la estructura del conjunto de datos. De igual forma, se normaliza el formato de las fechas, convirtiendo los distintos tipos de columnas temporales a un formato mensual estándar. Finalmente, los valores numéricos son depurados mediante la eliminación de símbolos de moneda, separadores de miles y otros caracteres que puedan afectar su correcta interpretación en los análisis posteriores.

TABLA II
VISTA PREVIA DE LOS DATOS HISTÓRICOS DE PRIMAS EMITIDAS DE SEGUROS DEL ESTADO S.A. DESPUÉS DEL PROCESO DE NORMALIZACIÓN EN LOS ATRIBUTOS.

AÑO	FECHA	SUCURSAL	RAMO	LINEA	COMPANIA	IMP_PRIMA	PRESUPUESTO
2007	2007-01-01	10 - OFICINA PRINCIPAL SOAT	59 - SOAT	SOAT	GENERALES	79194596.0	NaN
2007	2007-01-01	11 - BOGOTA	60 - AP INDIVIDUAL	ACCIDENTES PERSONALES	VIDA	100000.0	NaN
2007	2007-01-01	11 - BOGOTA	61 - AP COLECTIVOS	ACCIDENTES PERSONALES	VIDA	905915.0	NaN
2007	2007-01-01	11 - BOGOTA	63 - AP ANEXO AUTOS	ACCIDENTES PERSONALES	VIDA	1750255.0	NaN
2007	2007-01-01	11 - BOGOTA	68 - AP INTEGRAL E	ACCIDENTES PERSONALES	VIDA	52714000.0	NaN
2007	2007-01-01	11 - BOGOTA	48 - AUTOMOVILES LIV. IND	AUTOMOVILES	GENERALES	93635024.0	NaN
2007	2007-01-01	11 - BOGOTA	49 - AUTOMOVILES LIV. COL	AUTOMOVILES	GENERALES	455705930.2	NaN
2007	2007-01-01	11 - BOGOTA	50 - AUTOMOVI	AUTOMOVILES	GENERALES	89497118.0	NaN

			LES IND PES.

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

b) Análisis exploratorio (EDA)

i. Análisis de valores únicos y consistencia de categorías para el análisis

En esta sección también se identifican los valores únicos de las categorías de los diferentes atributos no numéricos, para verificar si hay inconsistencias en la escritura las categorías.

Se pueden visualizar los tres atributos categóricos en la TABLA III correspondientes a la Sucursal, Línea y Compañía. El atributo Sucursal contiene 69 valores únicos, lo que indica la existencia de 69 categorías asociadas a las oficinas de Seguros del Estado S.A. distribuidas a nivel nacional, reflejando la amplitud geográfica de la operación. El atributo Línea presenta 10 categorías que agrupan los distintos tipos de productos aseguradores ofrecidos, entre los cuales se incluyen SOAT, Accidentes Personales, Automóvil, Fianzas, Generales, RC Servicio Público, Vida Individual, Vida Grupo y Salud. El análisis se enfoca en el ramo Responsabilidad Civil Profesional, perteneciente a la línea Generales de la categoría Generales de la compañía. Sin embargo, las demás líneas se incluyen de manera descriptiva para proporcionar contexto, verificar la exhaustividad e integridad del dataset y garantizar que no existan registros mal clasificados. Adicionalmente, el atributo Compañía presenta dos categorías, Generales como se mencionó anteriormente y Vida, que representan las principales líneas de operación de Seguros del Estado S.A.

TABLA III
DISTRIBUCIÓN DE VALORES ÚNICOS DE LOS ATRIBUTOS CATEGÓRICOS

Atributo	Tipo	Cantidad de valores únicos	Valores únicos
Surcursal	Categórico	69	'10 - OFICINA PRINCIPAL SOAT' '11 - BOGOTA' '12 - CENTRO INTERNACIONAL' '14 - CHAPINERO' '15 - NORTE' '17 - CORREDORES' '18 - CHICO' '19 - ANTIGUO LAGO' '20 - AVENIDA 19' '21 - ANTIGUO COUNTRY' '25 - IBAGUE' '30 - VILLAVICENCIO' '33 - CALLE 100' '36 - LAGO' '37 - INTEGRA' '39 - TUNJA' '40 - POPAYAN' '41 - PASTO' '42 - MANIZALES' '43 - CENTRAL' '45 - CALI' '55 - PEREIRA' '65 - MEDELLIN' '75 - CARTAGENA' '85 - BARRANQUILLA' '96 - BUCARAMANGA' '44 - AVENIDA JIMENEZ' '60 - ARMENIA' '61 - AGENCIA MANDATARIA - NEIVA' '88 - SOAT CALI' '92 – SOAT PEREIRA' '86 - SOAT BARRANQUILLA' '87 - SOAT BUCARAMANGA' '89 - SOAT CARTAGENA' '90 - SOAT IBAGUE' '91 - SOAT MEDELLIN' '93 - SOAT TUNJA' '1 - OFICINA PRINCIPAL' '62 - AGENCIA MANDATARIA - PARQUE 93' '63 - AGENCIA MANDATARIA - CALLE 98' '64 - NIZA' '46 - AGENCIA SANTA MARTA' '48 - AGENCIA RIO NEGRO' '49 - AGENCIA CUCUTA' '50 - AGENCIA BARRANCABERMEJA' '51 - AGENCIA SOGAMOSO' '52 - AGENCIA TULUA' '47 - AGENCIA VALLEDUPAR' '53 - AGENCIA MONTERIA' '54 - OFICINA PRINCIPAL VIDA' '107 - SOAT AGENCIA TULUA' '99 - SOAT PASTO' '109 - SOAT ARMENIA' '94 - SOAT VILLAVICENCIO' '100 - SOAT MANIZALES' '104 - SOAT AGENCIA CUCUTA' '108 - SOAT AGENCIA MONTERIA' '102 - SOAT AGENCIA VALLEDUPAR' '66 - AGENCIA MANDATARIA SABANA' '2 - OFICINA NEGOCIOS ESPECIALES' '98 - SOAT POPAYAN' '97 - SOAT NEIVA' '105 - SOAT AGENCIA BARRANCABERMEJA' '57 - AGENCIA YOPAL' '110 - SOAT AGENCIA YOPAL' '56 - AGENCIA MANDATARIA BARRANCABERMEJA' '103 - SOAT AGENCIA RIONEGRO' '101 - SOAT AGENCIA SANTA MARTA' '106 - SOAT AGENCIA SOGAMOSO']
Ramo	Categórico	70	'59 - SOAT' '60 - AP INDIVIDUAL' '61 - AP COLECTIVOS' '63 - AP ANEXO AUTOS' '68 - AP INTEGRAL E' '48 - AUTOMOVILES LIV. IND' '49 - AUTOMOVILES LIV. COL' '50 - AUTOMOVILES PES. IND' '51 -

			AUTOMOVILES PES. COL' '85 - COLECTIVO VIDA' '40 - RCE CONTRATOS' '41 - JUDICIALES' '43 - CU. DISP.LEGAL.' '44 - CU. ENTIDAD.EST.' '45 - CU. PARTICULAR.' '2 - R. CIVIL EXTRAC.' '4 - INCENDIO' '5 - HAMCCP / AMIT' '11 - TERREMOTO' '12 - SUSTRACCION' '17 - EQUIPO ELECTRONICO' '18 - TODO RIESGO CONSTRUC' '19 - TODO RIESGO EQUIPO Y' '42 - MANEJO' '52 - RC MIGRACION' '71 - VIDA GRUPO' '80 - VIDA INDIVIDUAL' '82 - VI AHORRO' '15 - SEGURESTADO CARGA' '13 - TRANSP. DE VALORES' '14 - TRANSP DE MERCANCIAS' '30 - RCE PASAJEROS' '77 - VG SERVI. PUBLICOS' '86 - CV USO' '31 - RCC PASAJEROS' '21 - ROTURA DE MAQUINARIA' '72 - VG DEUDORES' '76 - VG COM PENSIONES' '95 - SER FUN FAMILIAR' '3 - R. C. PROFESIONAL' '96 - SER FUN MERCAESTADO' '20 - TODO RIESGO MONTAJE' '90 - SALUD INTEGRAL' '73 - VG PROTECCION ESTADO' '32 - RCE EXCESO PASAJEROS' '9 - VIDRIOS' '1 - R. CIVIL' '16 - SEGURESTADO PROPIETA' '74 - VG MERCAESTADO' '62 - AP EVENTOS' '10 - LUCRO CESANTE' '92 - SERV FUN COLEC FAM' '33 - RCC EXCESO PASAJEROS' '34 - RC EMPRESAS PASAJ.' '75 - VGMASIVO' '93 - SER FUN COLEC AGRUP' '64 - APMASIVOS' '94 - VG FAM GASTOS EMERGENTE' '69 - PIE GASTOS EDUC' '25 - IRF' '35 - RC EXCESO L.U. PASAJ' '65 - A P COMP PENSIONES' ‘46 - CU. ENTIDAD.EST.ONL' '47 - CU. PARTICULAR.ONL' ‘53 - JUDICIALES ONL' '91 - SALUD MASIVOS' '54 - RCE CONTRATOS RAPIE' '78 - VG DESEM O INC TEM' '55 - RCE CONTRATOS PPP' '57 - PA.PPP'
Línea	Categórico	10	'SOAT' 'ACCIDENTES PERSONALES' 'AUTOMOVILES' 'COLECTIVO' 'FIANZAS' 'GENERALES' 'RC SERVICIO PUBLICO' 'VIDA GRUPO' 'VIDA INDIVIDUAL' 'SALUD'
Compañía	Categórico	2	'GENERALES' 'VIDA'

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

ii. Filtrado de registros por ramo: R.C. profesional

Para centrar el análisis en el ramo de Responsabilidad Civil Profesional, se aplicó un filtrado específico del dataset, conservando únicamente los registros correspondientes a este ramo dentro de la línea Generales. Tras el filtrado, el conjunto de datos resultante quedó conformado por 7.932 registros, los cuales representan todas las transacciones históricas disponibles para este ramo. Este procedimiento garantiza que el análisis exploratorio y el modelado predictivo se realicen sobre un dataset coherente y homogéneo, evitando la inclusión de registros de otros ramos que podrían introducir ruido o inconsistencias.

La TABLA IV presenta una muestra de los primeros registros del dataset filtrado correspondiente al ramo Responsabilidad Civil Profesional. El conjunto de datos resultante contiene 7.932 registros como se menciona anteriormente y 8 atributos, incluyendo información temporal (Año y Fecha), variables categóricas (Sucursal, Ramo, Línea y Compañía) y variables numéricas asociadas al negocio asegurador, como el importe de la prima y el presupuesto. Esta estructura permite analizar la evolución de las primas emitidas en este ramo específico dentro de la línea Generales de la compañía.

TABLA IV
PRIMEROS REGISTROS DEL DATASET FILTRADO DEL RAMO RESPONSABILIDAD CIVIL PROFESIONAL.

	AÑO	FECHA	SUCURSAL	RAMO	LINEA	COMPAÑIA	IMP_PRIMA	PRESUPUESTO
18	2007	2007-01-01	25 - IBAGUE	3 - R. C. PROFESIONAL	GENERALES	GENERALES	14706920.00	NaN
09	2007	2007-01-01	40 - POPAYAN	3 - R. C. PROFESIONAL	GENERALES	GENERALES	9409168.99	NaN

72	2007	2007-02-01	19 - ANTIGUO LAGO	3 - R. C. PROFESIONAL	GENERALES	GENERALES	2300000.01	NaN
20	2007	2007-02-01	39 - TUNJA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	342184.00	NaN
38	2007	2007-02-01	40 - POPAYAN	3 - R. C. PROFESIONAL	GENERALES	GENERALES	306262.00	NaN

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. Calidad y consistencia para el análisis

Después de normalizar el dataset, se verifica la calidad de la información: Si hay datos faltantes, duplicados, y cuál es el rango temporal cubierto.

La Tabla V presenta el resumen de la calidad y consistencia de los datos correspondientes al ramo Responsabilidad Civil Profesional, mostrando que el dataset filtrado está compuesto por 7.932 registros y 8 atributos: Año, Fecha, Sucursal, Ramo, Línea, Compañía, Imp_prima y Presupuesto como se mencionó anteriormente en los resultados obtenidos. En cuanto al atributo temporal, la fecha mínima registrada corresponde al 1 de enero de 2007, mientras que la fecha máxima es el 1 de diciembre de 2025, lo cual es consistente con la normalización mensual aplicada al dataset. Asimismo, no se identificaron registros duplicados, lo que garantiza la integridad de la información utilizada para el análisis. En relación con la calidad de los datos, se identificaron 985 valores nulos en el atributo IMP_PRIMA y 192 en PRESUPUESTO, los cuales serán considerados en las etapas posteriores de tratamiento de datos. Por otra parte, se observan valores negativos y montos extremos en los atributos IMP_PRIMA y PRESUPUESTO, situación coherente con la naturaleza transaccional de la base de datos, ya que esta no solo registra emisiones de pólizas, sino también movimientos como revocaciones, cancelaciones y anulaciones, los cuales se reflejan mediante valores negativos para representar la reversión de transacciones previas. En este sentido, los valores positivos y negativos observados forman parte del comportamiento normal del sistema, representando el efecto acumulado de dichas operaciones dentro del flujo del negocio asegurador, por lo que se consideran datos válidos que deben conservarse para reflejar adecuadamente la dinámica real del ramo analizado.

TABLA V
RESUMEN DE CALIDAD Y CONSISTENCIA DE LOS DATOS HISTÓRICOS DE PRIMAS EMITIDAS DE SEGUROS DEL ESTADO S.A.

Métrica	Valor
filas	7932
columnas	[AÑO, FECHA, SUCURSAL, RAMO, LINEA, COMPAÑIA, ...
fechas_min	2007-01-01
fechas_max	2025-12-01
duplicados	0
IMP_PRIMA_nulos	985
IMP_PRIMA_min	-202753990.9
IMP_PRIMA_max	3897004030.0
PRESUPUESTO_nulos	192
PRESUPUESTO_min	-14182099.49
PRESUPUESTO_max	4821858505.0

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iv. Detección de valores faltantes y definición de su tratamiento para el análisis

Luego de la eliminación de las filas duplicadas, se procede a verificar la presencia de valores nulos en cada uno de los atributos con el fin de identificar posibles inconsistencias y definir el tratamiento adecuado para garantizar la calidad del conjunto de datos.

En la TABLA VI se evidencian resultados positivos en cuanto a la completitud de los atributos Año, Fecha, Sucursal, Línea y Compañía, los cuales no presentan valores faltantes. No obstante, se observa la presencia de 985 valores nulos en el atributo *Imp_Pima* y 192 valores nulos en *Presupuesto*, lo que indica la necesidad de aplicar un tratamiento adecuado para garantizar la consistencia y calidad del conjunto de datos.

TABLA VI
NÚMERO DE DATOS FALTANTES O NULOS POR ATRIBUTO

Atributo	Valores nulos
AÑO	0
FECHA	0
SUCURSAL	0
RAMO	0
LINEA	0
COMPAÑIA	0
IMP_PRIMA	985
PRESUPUESTO	192

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Luego, se verifica en la TABLA VII a partir de qué fechas comienzan a presentarse los valores faltantes en los atributos con el fin de identificar patrones temporales y determinar si dichas ausencias corresponden a periodos específicos del registro de transacciones.

A partir de la revisión de los primeros registros de la TABLA VII donde aparecen valores nulos, se observa que estos se presentan principalmente en el atributo PRESUPUESTO, mientras que el atributo IMP_PRIMA mantiene valores registrados en la mayoría de las observaciones. Este comportamiento sugiere que las ausencias no están asociadas a errores en la captura de información, sino posiblemente a periodos en los cuales no se definió o registró un valor presupuestal dentro del sistema. De esta manera, el análisis temporal de los valores faltantes permite comprender mejor la dinámica del registro de la información y facilita la toma de decisiones respecto al tratamiento de estos datos en las etapas posteriores del análisis.

TABLA VII
DATOS FALTANTES POR FECHA: LOS VALORES NULOS EN IMP_PRIMA Y PRESUPUESTO APARECEN DESDE 2008 Y SE MANTIENEN HASTA 2025

	AÑO	FECHA	SUCURSAL	RAMO	LINEA	COMPAÑIA	IMP_PRIMA	PRESUPUESTO
218	2007	2007-01-01	25 - IBAGUE	3 - R. C. PROFESIONAL	GENERALES	GENERALES	14706920.00	0.00
309	2007	2007-01-01	40 - POPAYAN	3 - R. C. PROFESIONAL	GENERALES	GENERALES	9409168.99	0.00
672	2007	2007-02-01	19 - ANTIGUO LAGO	3 - R. C. PROFESIONAL	GENERALES	GENERALES	2300000.01	0.00
820	2007	2007-02-01	39 - TUNJA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	342184.00	0.00
838	2007	2007-02-01	40 - POPAYAN	3 - R. C. PROFESIONAL	GENERALES	GENERALES	306262.00	0.00
...	...	...	...	...	...	...	...	...
8902	2007	2007-12-01	75 - CARTAGENA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	3665000.00	0.00
8940	2007	2007-12-01	85 - BARRANQUILLA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	9627999.81	0.00
8980	2007	2007-12-01	96 - BUCARAMANGA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	4764000.00	0.00
9026	2008	2008-01-01	10 - OFICINA PRINCIPAL SOAT	3 - R. C. PROFESIONAL	GENERALES	GENERALES	0.00	0.00
9067	2008	2008-01-01	11 - BOGOTA	3 - R. C. PROFESIONAL	GENERALES	GENERALES	1990000.00	12660384.02

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Después de identificar los valores faltantes, se aplica imputación con valor cero sobre los atributos IMP_PRIMA y PRESUPUESTO. Esta estrategia se sustenta en que los valores nulos no representan datos desconocidos ni errores de captura, sino la ausencia real de producción o de asignación presupuestal en un periodo determinado; en consecuencia, reemplazarlos por cero constituye una representación semánticamente fiel de la operación del negocio. Este enfoque corresponde al uso de una constante global para reemplazar valores faltantes, estrategia reconocida como válida cuando el valor de reemplazo tiene una interpretación clara dentro del dominio de los datos [30]. Si bien la literatura advierte que el uso de una constante arbitraria puede generar interpretaciones erróneas en el modelo, en el presente caso el valor cero no es arbitrario: representa meses sin emisión de pólizas o sin asignación presupuestal, lo cual constituye un estado operativo válido y esperado dentro de la operación del negocio asegurador [30].

En el contexto de bases de datos transaccionales de seguros, la no emisión de pólizas en una combinación sucursal-ramo-compañía-periodo equivale operativamente a un importe de prima igual a cero, lo que justifica el uso de este valor como imputación por defecto. Se descartó el uso de estimadores estadísticos como la media o la mediana, dado que, como señalan Han, Kamber y Pei, estos métodos están diseñados para distribuciones simétricas o sesgadas de datos numéricos continuos y su aplicación introduciría valores ficticios que no corresponden a ninguna transacción real, distorsionando la distribución real de los datos [30]. De igual forma, se descartó eliminar los registros con valores faltantes, puesto que ignorar las tuplas implica no aprovechar los valores restantes del registro, lo cual resulta ineficiente cuando esos datos son útiles para el análisis [30]. En este caso particular, la eliminación de registros con valores faltantes en el atributo IMP_PRIMA habría suprimido simultáneamente los valores presentes en el atributo

PRESUPUESTO para esas mismas tuplas, generando una pérdida de información válida y disponible, afectando la integridad del conjunto de datos utilizado en el análisis exploratorio previo y produciendo una matriz de datos incompleta e inconsistente [30]. Por esta razón, la imputación con cero permitió conservar la totalidad de los registros, garantizando que todos los atributos permanecieran disponibles y que la matriz de datos mantuviera su estructura rectangular completa.

Ahora bien, el modelo predictivo opera exclusivamente sobre IMP_PRIMA, el atributo PRESUPUESTO fue utilizado únicamente dentro del análisis exploratorio de datos y en la plataforma de visualización desarrollada en Streamlit. Su eliminación parcial habría generado una matriz de datos incompleta e inconsistente, comprometiendo la integridad del análisis y la coherencia entre las variables a lo largo del proyecto.

Cabe señalar que la imputación con cero (*zero imputation*) constituye una técnica reconocida en la literatura para el tratamiento de datos faltantes. De acuerdo con Yi et al., esta estrategia consiste en asignar el valor cero a las observaciones ausentes y es considerada una de las alternativas más simples e intuitivas para manejar la falta de información en conjuntos de datos utilizados en aplicaciones de aprendizaje automático. Aunque los autores advierten que su utilización debe evaluarse de acuerdo con las características del problema y la naturaleza de los datos, reconocen que se trata de un procedimiento ampliamente empleado en distintos contextos de análisis de datos [31].

En la TABLA VIII se puede observar que la imputación se realizó exitosamente, confirmando que no quedan valores faltantes y garantizando la integridad del conjunto de datos para el análisis posterior.

TABLA VIII
VALORES FALTANTES POR ATRIBUTO DESPUÉS DE LA IMPUTACIÓN

Atributo	Valores nulos
AÑO	0
FECHA	0
SUCURSAL	0
RAMO	0
LINEA	0
COMPAÑIA	0
IMP_PRIMA	0
PRESUPUESTO	0

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Una vez finalizado el proceso de limpieza, imputación y consolidación de los datos, el dataset del importe de las primas quedó conformado por 7.932 registros transaccionales correspondientes al ramo de Responsabilidad Civil Profesional, abarcando el periodo enero de 2007 hasta diciembre de 2025.

v. Distribución del volumen del importe de las primas por sucursal

La Figura 4 presenta el *Top 10 de sucursales con mayor volumen de primas* dentro de un total de 44 sucursales analizadas, evidenciando una concentración significativa de la producción en un número reducido de unidades operativas. En particular, la sucursal Medellín se destaca ampliamente al registrar aproximadamente 93.19 mil millones, superando de forma considerable al resto de sucursales. En un segundo nivel se ubican Agencia Mandataria – Parque 93 y Bucaramanga, con valores cercanos a 31.55 mil millones, seguidas por Calle 100, Cartagena y Antiguo Country, cuyos montos se sitúan entre 27 y 28 mil millones. Posteriormente aparecen Cali, Barranquilla, Bogotá y Pasto, con valores entre 15 y 24 mil millones. Este comportamiento evidencia una distribución heterogénea en la generación de primas, donde un número reducido de sucursales concentra una proporción importante del volumen total, lo que puede estar asociado a factores como el tamaño del mercado regional, la presencia de clientes corporativos o la dinámica comercial de cada sede.

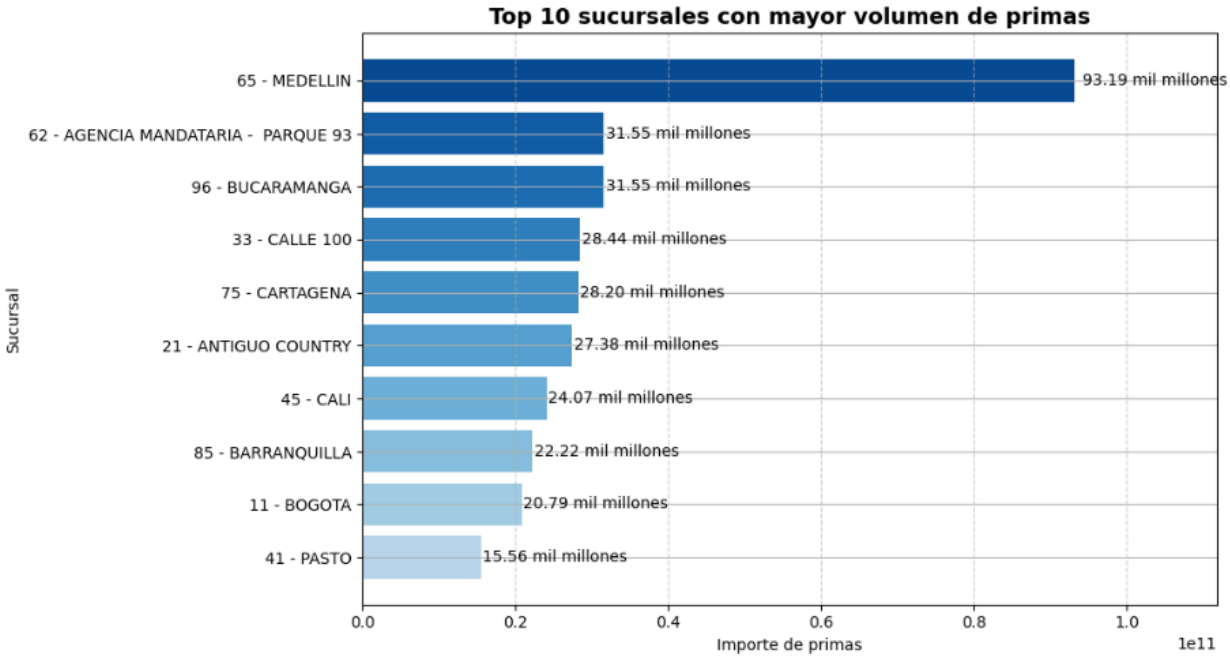


Fig.4. Top 10 de sucursales con mayor volumen de primas del Ramo R.C. Profesional en Seguros del Estado S.A. Elaboración propia.

vi. Análisis de la distribución y valores atípicos en las variables financieras asociadas

En la Figura 5 se presentan los diagramas de caja de las variables IMP_PRIMA y PRESUPUESTO, en los cuales se observa una distribución altamente asimétrica hacia la derecha, caracterizada por una fuerte concentración de valores en rangos bajos cercanos a la base de los gráficos y la presencia de numerosos valores atípicos que se extienden hacia magnitudes considerablemente altas. Esta configuración evidencia una marcada dispersión en los datos, donde la mayoría de los registros corresponde a importes de primas y presupuestos de menor cuantía, mientras que un grupo reducido de observaciones alcanza valores significativamente elevados. Dicho comportamiento es consistente con la naturaleza de los datos del sector asegurador, en el que la mayor parte de las pólizas presentan montos moderados, mientras que algunos casos particulares pueden asociarse a pólizas corporativas, contratos especiales o presupuestos de gran escala.

Por otro lado, desde una perspectiva comparativa, aunque IMP_PRIMA alcanza un valor extremo ligeramente superior, la variable PRESUPUESTO presenta una dispersión más amplia y una mayor concentración de valores atípicos en niveles elevados, lo que evidencia una variabilidad más marcada en los montos presupuestales. Esto sugiere que, mientras las primas tienden a ajustarse dentro de rangos relativamente más definidos, los presupuestos muestran una mayor heterogeneidad y amplitud entre los registros, reflejando escenarios financieros más diversos y posiblemente asociados a decisiones estratégicas y operativas más flexibles. En este contexto, los valores atípicos no se consideran errores ni es apropiado eliminarlos o tratarlos mediante técnicas de imputación, ya que representan situaciones reales propias de la dinámica operativa y financiera de la compañía. Estos registros pueden corresponder a pólizas de alto valor, contratos corporativos o presupuestos estratégicos que, aunque menos frecuentes, forman parte del funcionamiento normal de la empresa. Por tanto, los outliers no distorsionan el análisis, sino que aportan información relevante sobre la diversidad y magnitud de los montos gestionados por Seguros del Estado S.A.

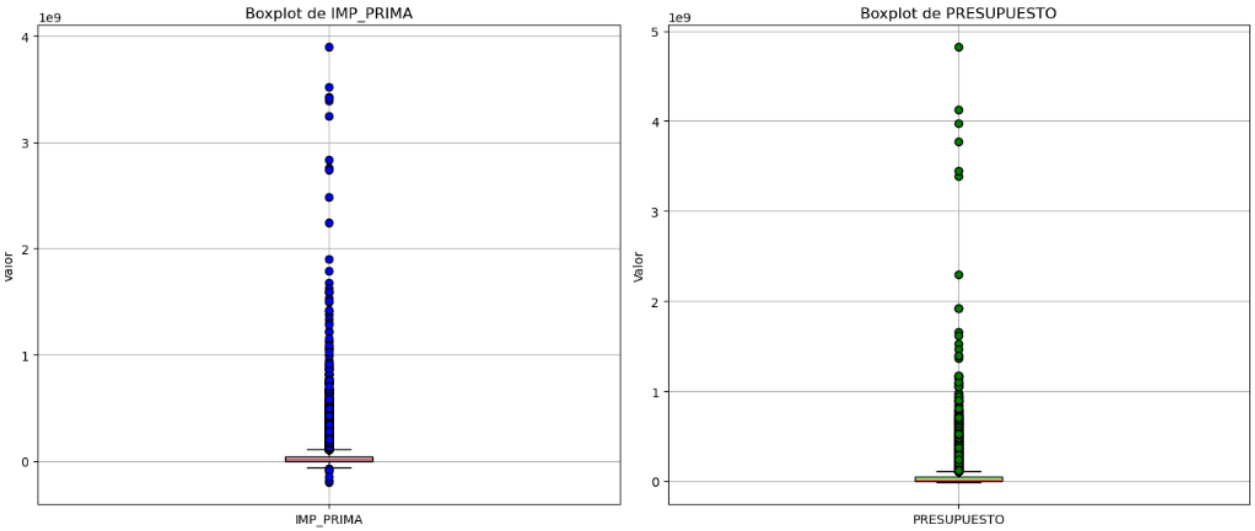


Fig.5. Representación boxplot de las variables financieras IMP_PRIMA y PRESUPUESTO del Ramo R.C. en Seguros del Estado S.A. Elaboración propia.

Para respaldar la interpretación visual obtenida a partir de los gráficos de boxplot, se calculan las métricas estadísticas de coeficiente de asimetría (skewness) y curtosis, las cuales permiten evaluar de manera cuantitativa la forma de la distribución de los datos.

Los resultados del coeficiente de asimetría (skewness) muestran en la Tabla IX valores muy altos para IMP_PRIMA (10.29) y PRESUPUESTO (12.58), lo que indica una fuerte asimetría positiva, es decir, la mayoría de los valores se concentran en montos bajos mientras que existen algunos registros con valores muy elevados. Asimismo, los valores de curtosis (159.79 y 259.73, respectivamente) evidencian distribuciones leptocúrticas, caracterizadas por colas pesadas y una alta presencia de valores extremos, confirmando lo observado en los diagramas de caja.

TABLA IX
ESTADÍSTICAS DE FORMA DE LA DISTRIBUCIÓN PARA LAS VARIABLES IMP_PRIMA Y PRESUPUESTO

	Variable	Skewness	Curtosis
0	IMP_PRIMA	10.289931	159.791272
1	PRESUPUESTO	12.580185	259.728913

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

vii. Tendencia temporal del importe mensual

La Figura 6 muestra la evolución del importe de las primas mensuales a lo largo del periodo analizado, evidenciando una tendencia creciente en el volumen del importe de las primas desde 2007 hasta 2025. Durante los primeros años se observa un comportamiento relativamente estable con valores bajos y menor variabilidad; sin embargo, a partir de aproximadamente 2015 comienza a evidenciarse un incremento gradual tanto en el nivel como en la volatilidad de la serie. En 2020, periodo asociado a la pandemia de COVID-19, se perciben fluctuaciones en la dinámica de las primas, lo que puede relacionarse con los efectos económicos y la incertidumbre generada en los mercados; no obstante, posteriormente la serie retoma su tendencia ascendente, mostrando picos más pronunciados en los años recientes, especialmente entre 2022 y 2025. Este comportamiento sugiere una expansión en la producción de primas y una mayor variabilidad en los montos registrados en los últimos años del periodo analizado.

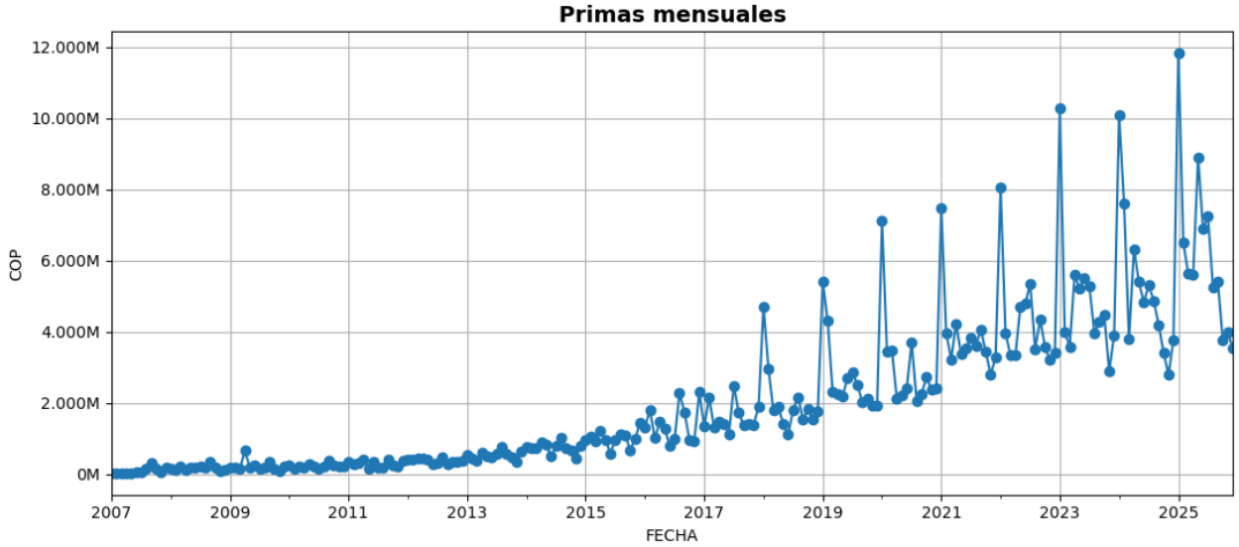


Fig.6. Evolución histórica del importe mensual de primas del ramo de R.C. Profesional (2007–2025) en Seguros del Estado S.A. Elaboración propia a partir de los datos procesados en la herramienta Python.

viii. Comportamiento estacional y evolución anual

En el heatmap o mapa de calor Figura 7 se presenta la estacionalidad mensual de las primas (IMP_PRIMA) a lo largo de los años analizados, donde la intensidad del color representa mayores volúmenes de importe de primas. En los primeros años del periodo se observan valores relativamente bajos y con menor variabilidad; sin embargo, a partir de aproximadamente 2018 se evidencia un incremento progresivo en los montos registrados. Asimismo, se identifica un patrón estacional el cual es propio y característico de los procesos comerciales de renovación de póliza, en el cual los meses iniciales del año, particularmente enero y febrero, tienden a concentrar mayores valores de primas en comparación con otros meses. Este comportamiento se acentúa en los años recientes, especialmente entre 2023 y 2025, donde se observan las tonalidades más intensas del mapa de calor, reflejando un aumento significativo en la producción de primas. En conjunto, la figura sugiere una tendencia creciente acompañada de patrones estacionales, lo cual resulta relevante para el análisis temporal y la posterior modelación mediante técnicas de series de tiempo.

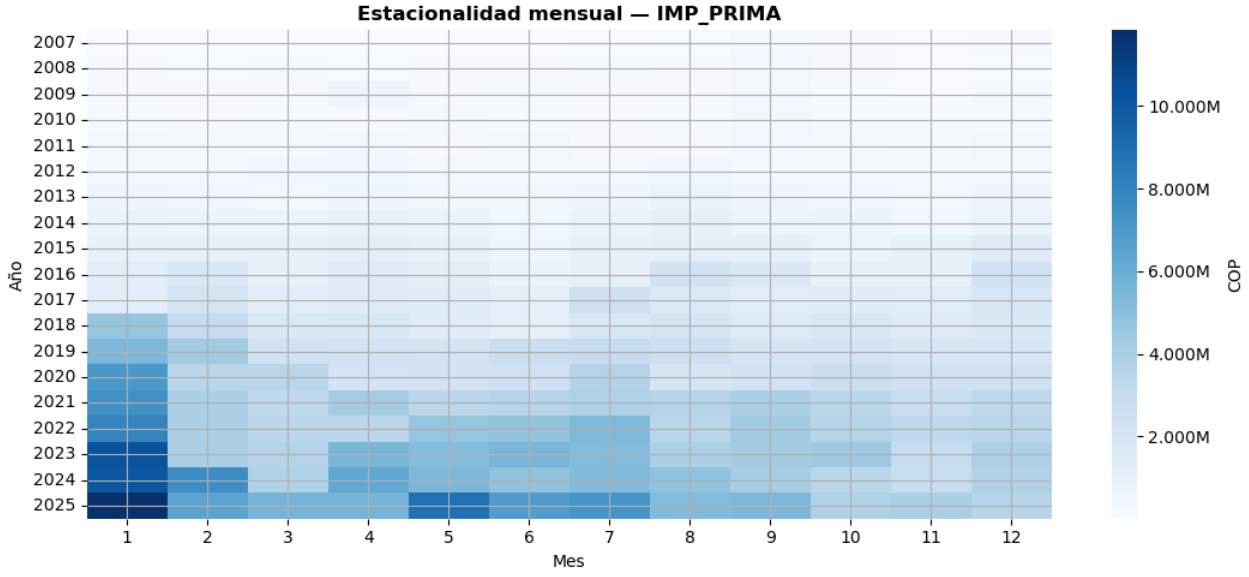


Fig.7. Estacionalidad mensual del atributo Importe de Primas del Ramo R.C. Profesional de la compañía de Seguros del Estado S.A. Elaboración propia a partir de los datos procesados en la herramienta Python.

La Tabla X presenta las observaciones identificadas con valores elevados mediante el estadístico Z-Score, lo que permite detectar registros significativamente superiores al promedio de la serie del importe de las primas. Los resultados evidencian que varios de estos valores corresponden principalmente al mes de enero en diferentes años, especialmente entre 2018 y 2025, lo que sugiere un comportamiento recurrente al inicio de cada periodo anual. Estos valores son consistentes con la dinámica del sector asegurador, ya que en los primeros meses del año se presenta un mayor volumen en el importe de primas debido a procesos de renovación de pólizas, inicio de nuevas coberturas y estrategias comerciales implementadas por la aseguradora al comienzo de cada periodo anual, lo que ocasiona una mayor concentración de emisiones y recaudos en dichos meses. Entre ellos destaca enero de 2022, que registra el mayor Z-Score (3.03), indicando uno de los picos más pronunciados de la serie. Asimismo, los valores observados en 2023, 2024 y 2025 reflejan montos elevados de primas, consistentes con la tendencia creciente identificada previamente en el análisis temporal.

TABLA X
EVOLUCIÓN ANUAL DEL INDICADOR Y SU Z-SCORE (2009–2025) EN SEGUROS DEL ESTADO S.A.

	FECHA	VALOR	Z_SCORE
0	2009-04-01	653.221.454	2,89
1	2018-01-01	4.708.407.106	2,91
2	2019-01-01	5.391.412.966	2,91
3	2020-01-01	7.101.538.060	2,85
4	2021-01-01	7.459.192.115	2,94
5	2022-01-01	8.043.041.887	3,03
6	2023-01-01	10.265.991.983	2,97
7	2024-01-01	10.091.257.691	2,83
8	2025-01-01	1.183.6236.872	2,67

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 8 correspondiente al presupuesto mensual del ramo de Responsabilidad Civil Profesional evidencia una tendencia creciente de largo plazo a lo largo del periodo analizado (2007–2025). En los primeros años, particularmente entre 2007 y 2013, los valores presupuestados se mantienen en niveles relativamente bajos y con poca variabilidad, lo que sugiere una etapa inicial de menor participación del ramo dentro de la cartera o una planificación financiera más conservadora. A partir de 2014 se observa un incremento progresivo del presupuesto, acompañado de una mayor dispersión en los datos, lo que indica una expansión gradual de las metas comerciales y financieras asignadas a este ramo.

A partir de 2017, y con mayor intensidad desde 2020, la serie presenta un comportamiento más dinámico, con fluctuaciones mensuales más pronunciadas y picos de alta magnitud, alcanzando sus valores máximos entre 2023 y 2025. Este comportamiento sugiere que el presupuesto del ramo no solo aumentó en términos absolutos, sino que también se volvió más sensible a ajustes periódicos, posiblemente asociados a estrategias de crecimiento, revisiones comerciales, cambios en la demanda del mercado o actualizaciones en las proyecciones corporativas. En términos generales, la evolución del presupuesto refleja una expectativa sostenida de crecimiento para el ramo de Responsabilidad Civil Profesional dentro de la compañía.

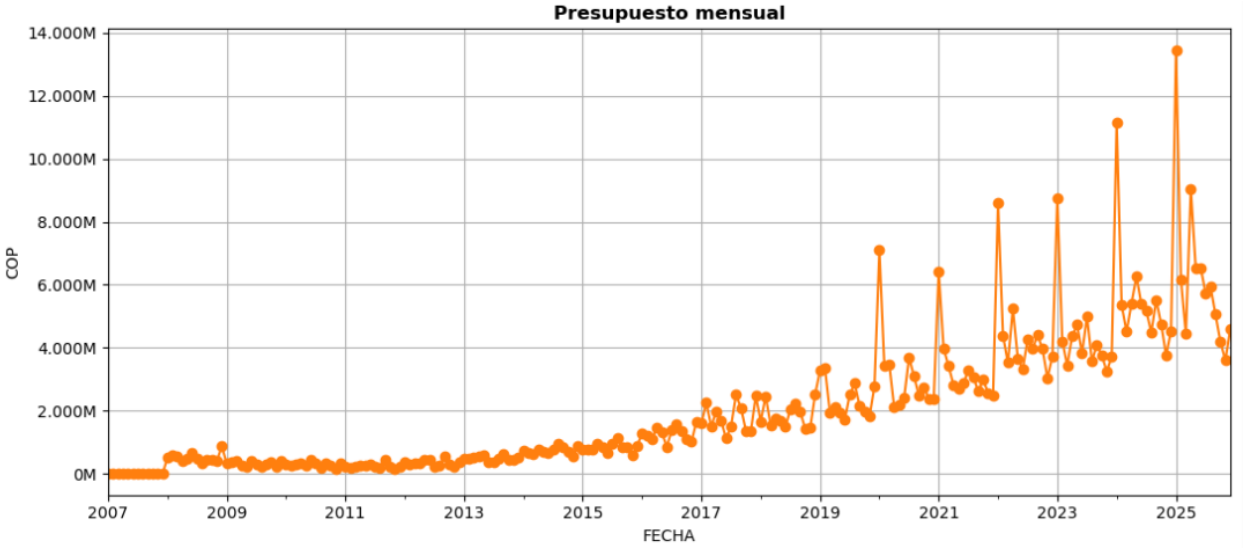


Fig.8. Presupuesto mensual desde 2008 hasta el 2024 en Seguros del Estado S.A. Elaboración propia a partir de los datos procesados en la herramienta Python.

Al comparar la serie del presupuesto mensual con la serie del importe de las primas mensuales (IMP_PRIMA), se observa que ambas presentan una tendencia ascendente similar en el tiempo, lo cual evidencia una relación coherente entre la planeación financiera del ramo y su comportamiento real de producción. En ambos casos, los valores se mantienen relativamente estables durante los primeros años del periodo, pero a partir de aproximadamente 2014 comienza una fase de crecimiento sostenido que se intensifica después de 2017. Esta similitud sugiere que, a nivel general, el comportamiento de la producción de las primas ha seguido una trayectoria alineada con la evolución del presupuesto definido para el ramo.

No obstante, al analizar con mayor detalle, se evidencia que la serie del importe de las primas mensuales presenta una volatilidad más marcada que la del presupuesto, con picos más abruptos y oscilaciones más frecuentes, especialmente entre 2018 y 2025. Esto es esperable, dado que el presupuesto representa una meta o proyección planificada, mientras que el importe de las primas refleja el comportamiento real del mercado, sujeto a factores operativos, comerciales y coyunturales. En este sentido, aunque el presupuesto muestra una evolución creciente y relativamente estructurada, la producción de primas exhiben una dinámica más irregular, lo que sugiere que la ejecución real del ramo puede experimentar periodos de sobrecumplimiento o desviaciones frente a lo proyectado.

En conjunto, ambas Figuras 6 y 8 permiten inferir que el ramo de Responsabilidad Civil Profesional ha mostrado una expansión importante en el tiempo, tanto desde la perspectiva de planeación presupuestal como desde la generación efectiva de primas. Sin embargo, la mayor dispersión observada en la serie del importe de las primas indica la necesidad de profundizar en análisis posteriores sobre estacionalidad, variabilidad y posibles eventos atípicos.

En consecuencia, la comparación entre ambas series sugiere que, si bien existe una alineación general entre la evolución del presupuesto y el comportamiento del importe de las primas, la mayor variabilidad observada en la serie real resalta la importancia de incorporar métodos de modelación que permitan capturar adecuadamente las fluctuaciones temporales y mejorar la capacidad predictiva del ramo.

2) Análisis descriptivo de siniestros incurridos

a) ETL (Extracción, Transformación y Carga de datos)

i. Extracción y selección de datos

A partir de las mismas fuentes utilizadas para el *dataset* de importe de primas, se construye el conjunto de datos

para el análisis del valor de siniestros incurridos. Para ello, se realiza un proceso de extracción, transformación y consolidación de la información, orientado a generar un *dataset* con periodicidad mensual que incluye variables como principales ciudades, año, compañía, ramo, fecha, valor y departamento.

Adicionalmente, es importante señalar que se tomó específicamente el atributo del valor de los siniestros incurridos, ya que en la fuente original esta columna contiene tanto información del valor de los siniestros incurridos como del importe de las primas. Por esta razón, se efectuó una validación adicional sobre dicho campo, con el fin de asegurar que la información extraída corresponde efectivamente a el valor de los siniestros incurridos, tal como se evidencia en la Tabla XI.

TABLA XI
EXTRACCIÓN Y SELECCIÓN DE DATOS DEL VALOR DE LOS SINIESTROS INCURRIDOS DEL RAMO DE RESPONSABILIDAD CIVIL PROFESIONAL POR CIUDAD Y PERIODO

	Principales Ciudades	Año	COMPAÑÍA	CIUDAD	RAMOS	Siniestros	FECHA	Valor	DEPARTAMENTO
0	BOGOTA	2019	ESTADO	BOGOTA	RESPONSABILIDAD CIVIL	Siniestros	2019-01-31	1472306	BOGOTÁ DC
1	BOGOTA	2024	ESTADO	BOGOTA	RESPONSABILIDAD CIVIL	Siniestros	2024-01-31	824557	BOGOTÁ DC
2	BOGOTA	2022	ESTADO	BOGOTA	RESPONSABILIDAD CIVIL	Siniestros	2022-01-31	787872	BOGOTÁ DC
3	BOGOTA	2021	ESTADO	BOGOTA	RESPONSABILIDAD CIVIL	Siniestros	2021-01-31	1014766	BOGOTÁ DC
4	BOGOTA	2015	ESTADO	BOGOTA	RESPONSABILIDA CIVIL	Siniestros	2015-01-31	1800430	BOGOTÁ DC

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Nota. En la base de datos de siniestros incurridos, el ramo se encuentra registrado como “Responsabilidad Civil”; por ello, esta denominación se emplea en el documento para referirse al mismo objeto de estudio analizado.

ii. Transformación y depuración del dataset

Una vez realizada la extracción de los datos, se lleva a cabo la fase de transformación y depuración con el objetivo de garantizar la calidad, consistencia y estructura adecuada del dataset para su posterior análisis y modelamiento. En primer lugar, se efectúa la limpieza de los nombres de las columnas, estandarizando su formato para facilitar su manipulación y comprensión. Posteriormente, se realiza el renombramiento de variables clave, con el fin de asegurar coherencia semántica y alineación con los objetivos del análisis, destacándose la variable VALOR_SINIESTRO_INCURRIDO como medida principal de estudio. Adicionalmente, se construye la variable SUCURSAL, mediante la combinación de las variables ciudad y departamento, lo que permite generar una dimensión geográfica más robusta para el análisis territorial del valor de los siniestros incurridos.

En cuanto a las variables temporales, se verificó y estandarizó el formato de la variable fecha, asegurando su correcta interpretación como dato tipo fecha y validando la ausencia de registros inválidos. Posteriormente, se lleva a cabo la limpieza de variables numéricas, eliminando inconsistencias y garantizando que los valores asociados a los siniestros incurridos fueran correctamente interpretados como datos cuantitativos. De igual forma, se realiza la depuración de variables categóricas, estandarizando sus valores para evitar duplicidades o inconsistencias en la información. Como parte del proceso de optimización del dataset, se eliminan variables redundantes o no relevantes

para el análisis, y se define un conjunto final de variables clave, compuesto por el año, la fecha, la sucursal, el ramo y el valor del siniestro incurrido. Finalmente, se obtiene un dataset limpio y estructurado como se evidencia en la Tabla XII, con un total de 2.395 registros y 5 atributos (variables), listo para su uso en el análisis exploratorio y el desarrollo de los modelos predictivos.

TABLA XII
PRIMEROS REGISTROS DEL DATASET FINAL LIMPIO DE LOS SINIESTROS INCURRIDOS

	AÑO	FECHA	SUCURSAL	RAMO	VALOR_SINIESTRO_INCURRIDO
0	2015	2015-01-31	PEREIRA - RISARALDA	RESPONSABILIDAD CIVIL	4843900
1	2015	2015-01-31	NEIVA - HUILA	RESPONSABILIDAD CIVIL	2900600
2	2015	2015-01-31	IBAGUE - TOLIMA	RESPONSABILIDAD CIVIL	2495400
3	2015	2015-01-31	PASTO - NARIÑO	RESPONSABILIDAD CIVIL	388100
4	2015	2015-01-31	CARTAGENA - BOLÍVAR	RESPONSABILIDAD CIVIL	1272100

Fuente: Base de datos histórica de la compañía de Seguros del Estado S.A.

b) Análisis exploratorio (EDA)

i. Exploración de variables categóricas

En la Tabla XIII se presenta el atributo SUCURSAL, el cual cuenta con 34 categorías que representan la distribución geográfica de los valores de siniestros incurridos a nivel nacional. Por su parte, la variable RAMO evidencia una homogeneidad total, al estar conformada únicamente por la categoría “Responsabilidad Civil”, dentro de la cual se encuentran los registros analizados en esta investigación.

TABLA XIII
DISTRIBUCIÓN DE VARIABLES CATEGÓRICAS DEL DATASET DEL VALOR DE LOS SINIESTROS INCURRIDOS

Variable	Cantidad de valores únicos	Valores únicos (resumen)
SUCURSAL	34	'PEREIRA - RISARALDA' 'NEIVA - HUILA' 'IBAGUE - TOLIMA' 'PASTO - NARIÑO' 'CARTAGENA - BOLÍVAR' 'CALI - VALLE DEL CAUCA' 'BUCARAMANGA - SANTANDER' 'MEDELLIN - ANTIOQUIA' 'MANIZALES - CALDAS' 'TUNJA - BOYACÁ' 'BOGOTA - BOGOTÁ DC' 'VILLAVICENCIO - META' 'BARRANQUILLA - ATLÁNTICO' 'POPAYAN - CAUCA' 'ARMENIA - QUINDÍO' 'VALLEDUPAR - CESAR' 'SANTA MARTA - MAGDALENA' 'MONTERIA - CÓRDOBA' 'CUCUTA - NORTE DE SANTANDER' 'TULUA - VALLE DEL CAUCA' 'YOPAL - CASANARE' 'RESTO PAIS - RESTO DEL PAIS' 'MOCOA - PUTUMAYO' 'FLORENCIA - CAQUETÁ' 'SJGUAVIARE - GUAVIARE' 'PALMIRA - VALLE DEL CAUCA' 'PTO. INIRIDA - GUAINÍA' 'GIRARDOT - CUNDINAMARCA' 'RIOHACHA - LA GUAJIRA' 'LETICIA - AMAZONAS' 'QUIBDO - CHOCÓ' 'SINCELEJO - SUCRE' 'SAN ANDRES - SAN ANDRÉS Y PROVIDENCIA' 'BUENAVENTURA - VALLE DEL CAUCA'
RAMO	1	RESPONSABILIDAD CIVIL

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

ii. Evaluación de calidad, consistencia y completitud del dataset

La Tabla XIV muestra el conjunto de datos del valor de los siniestros incurridos compuesto por 2.395 registros y 5 variables como se mencionó anteriormente, con cobertura temporal desde enero de 2015 hasta octubre de 2025. No se evidencian valores duplicados ni datos nulos, lo que garantiza una completitud del 100%. La variable VALOR_SINIESTRO_INCURRIDO presenta alta variabilidad, con valores que oscilan entre -10.832.800 COP y 571.097.400 COP, y un promedio de 14.278.738.2881 COP, lo que refleja la heterogeneidad en la magnitud de los valores de los siniestros incurridos registrados. Adicionalmente, la presencia de valores negativos puede estar asociada a procesos de ajuste contable, recuperación de siniestros o liberación de reservas, lo cual resulta consistente con la

naturaleza de este tipo de información financiera.

TABLA XIV
RESUMEN DE LA CALIDAD Y CONSISTENCIA

	valor
filas	2395
columnas	[AÑO, FECHA, SUCURSAL, RAMO, VALOR_SINIESTRO_INCURRIDOS
fechas_min	2015-01-31
fechas_max	2025-10-31
duplicados	0
VALOR_SINIESTRO_INCURRIDO_nulos	0
VALOR_SINIESTRO_INCURRIDO_min	-10836800.0
VALOR_SINIESTRO_INCURRIDO_max	571097400.0
VALOR_SINIESTRO_INCURRIDO_media	14278738.2881
completitud_%	100.0

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. Valores faltantes en el dataset

Como se observa en la Tabla XV, generada mediante Python, no se presentan valores faltantes en ninguna de las variables de la base de datos del valor de los siniestros incurridos, lo que evidencia la integridad y completitud de la información disponible.

TABLA XV
NÚMERO DE VALORES FALTANTES POR ATRIBUTO

Atributo	Valores faltantes
AÑO	0
FECHA	0
SUCURSAL	0
RAMO	0
SINIESTRO_NETO	0
dtype:	Int64

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

El dataset del valor de los siniestros incurridos quedó compuesto por 2.395 registros transaccionales correspondientes al ramo de Responsabilidad Civil Profesional, abarcando el periodo enero de 2015 hasta diciembre de 2024, sin valores faltantes y con la estructura necesaria para garantizar la calidad del análisis posterior.

iv. Distribución de siniestros por sucursal

La Figura 9 muestra la distribución del valor de los siniestros incurridos por sucursal, evidenciando una alta concentración en Bogotá – Bogotá D.C., la cual supera ampliamente al resto de ciudades, con un valor cercano a los 22.519,9 millones de COP. En segundo lugar, se encuentra Medellín, seguida por Bucaramanga y Cali; sin embargo, estas ciudades presentan valores considerablemente inferiores en comparación con Bogotá, lo que evidencia una marcada diferencia en la magnitud de los siniestros. Este comportamiento refleja una distribución altamente asimétrica, en la que una única ciudad concentra la mayor proporción del valor total, mientras que el resto de las sucursales presenta niveles significativamente menores. Si bien algunas de las principales ciudades del país registran valores relativamente más altos frente a otras regiones, es importante destacar que la concentración del riesgo no es homogénea, sino que se encuentra fuertemente centralizada en Bogotá, es asociado a su mayor tamaño económico, densidad poblacional y nivel de exposición al riesgo asegurado.

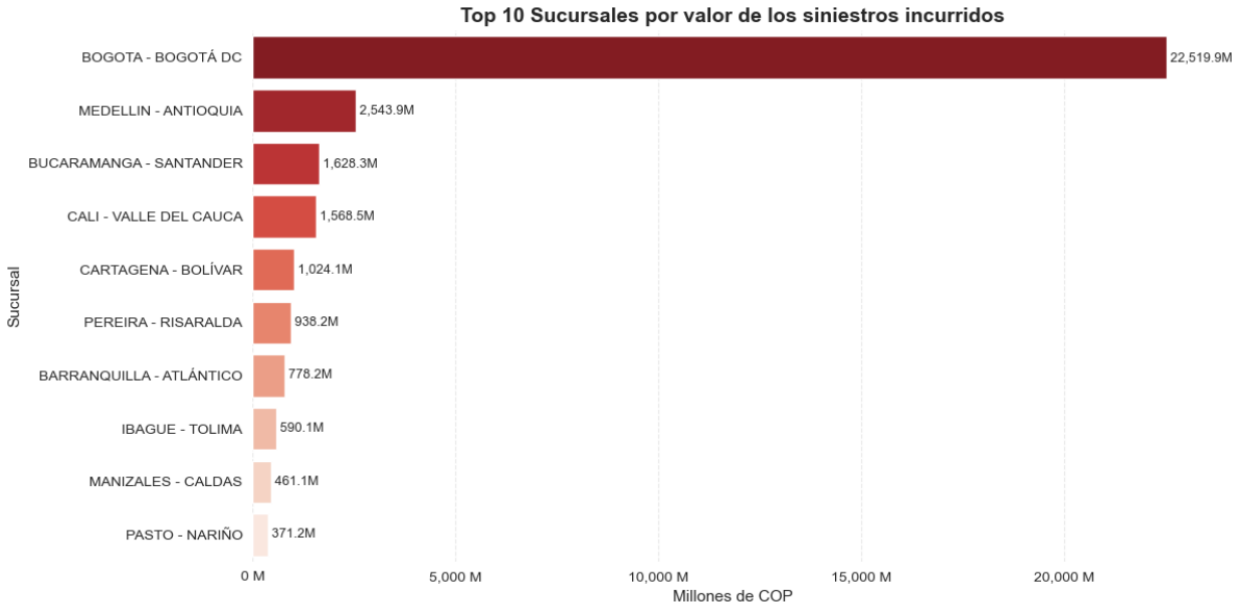


Fig.9. Top 10 sucursales con mayor valor de siniestros incurridos en Colombia en la compañía de Seguros del Estado S.A. Elaboración propia mediante la herramienta de Python.

v. Distribución y valores atípicos

El boxplot de la Figura 10 del valor de los siniestros incurridos evidencia una distribución fuertemente asimétrica hacia la derecha, donde la mayoría de los valores se concentran en montos relativamente bajos, mientras que existe una gran cantidad de valores atípicos que se extienden hacia cifras muy elevadas. La mediana se ubica cerca del extremo inferior de la caja, lo que confirma que la mayor parte de los siniestros son de bajo valor, pero la media (≈ 14.3 millones COP) se ve significativamente afectada por la presencia de estos outliers de alta severidad. La amplitud de los datos y la dispersión observada sugieren un comportamiento típico de riesgos con eventos poco frecuentes, pero de alto impacto, lo cual es clave para el análisis actuarial y la gestión del riesgo.

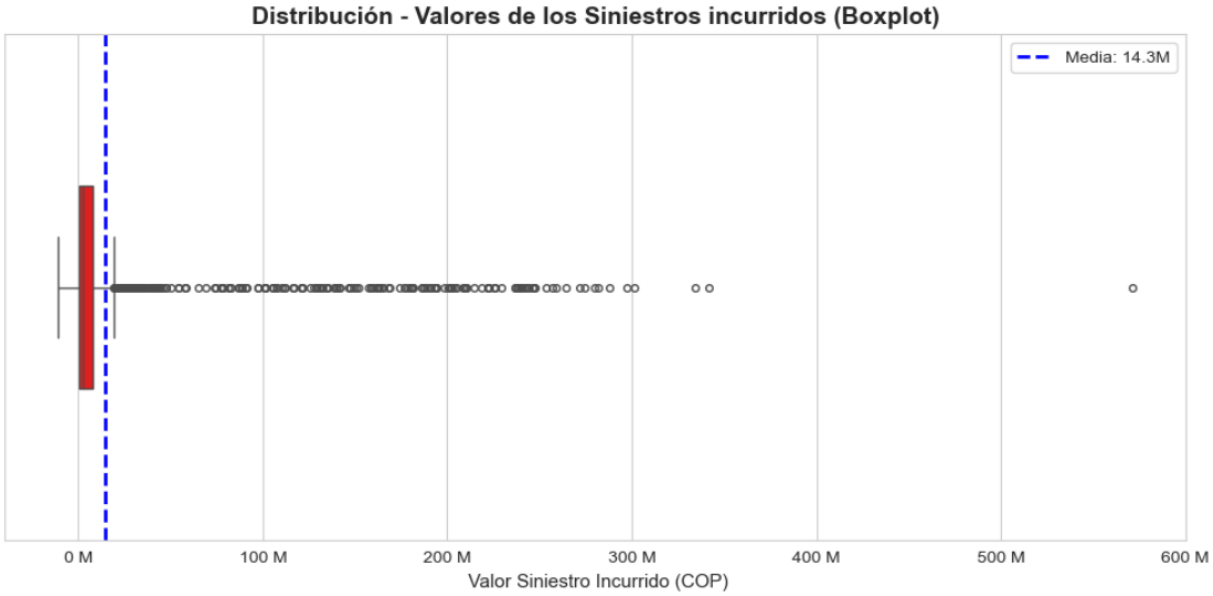


Fig.10. Distribución del valor del siniestro incurrido de la compañía de Seguros del Estado S.A. mediante Boxplot. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

Cabe señalar que los valores atípicos no son objeto de tratamiento ni depuración, ya que representan eventos reales de alta severidad. Su inclusión es fundamental para capturar adecuadamente la distribución del riesgo y evitar sesgos en el análisis.

El análisis de la Tabla XVI de asimetría y curtosis es totalmente coherente con lo observado en el boxplot del valor del siniestro incurrido. La asimetría positiva elevada (4.9779) confirma cuantitativamente la fuerte inclinación hacia la derecha que ya se evidenciaba en la gráfica, donde la mayoría de los datos se concentran en valores bajos y una cola larga se extiende hacia montos altos. Por su parte, la curtosis extremadamente alta (30.2007) indica una distribución leptocúrtica, caracterizada por una alta concentración de valores alrededor de la mediana y una presencia significativa de valores extremos. Esto coincide con la gran cantidad de outliers observados en el boxplot, los cuales no solo están presentes, sino que tienen un impacto considerable en la forma de la distribución y en medidas como la media. En conjunto, tanto la gráfica como la tabla reflejan un comportamiento típico del valor de los siniestros incurridos: alta frecuencia de eventos de baja cuantía y frecuencia de eventos de muy alto impacto, lo que refuerza la importancia de no excluir estos valores atípicos en el análisis.

TABLA XVI
ASIMETRÍA Y CURTOSIS DEL VALOR DEL SINIESTRO INCURRIDO

Métrica	Valor
Asimetría	4.9779
Curtosis	30.2007

Fuente: Elaboración propia.

vi. Tendencias / estacionalidad / anomalías

La Figura 11 evidencia la evolución del valor de los siniestros incurridos, mostrando un comportamiento altamente variable con picos significativos a lo largo del tiempo. En el periodo previo a la pandemia (2015–2019), los valores se mantienen relativamente estables, aunque con algunos eventos atípicos de alta magnitud. A partir de 2020, coincidiendo con la pandemia de COVID-19, se observa una caída marcada en los niveles de siniestralidad, alcanzando mínimos durante 2020 y manteniéndose en niveles bajos durante 2021, lo cual puede asociarse a la reducción de la actividad económica y la exposición al riesgo. Posteriormente, desde 2022 se evidencia una recuperación progresiva, acompañada de un incremento en la volatilidad y en la magnitud de los valores, especialmente hacia 2024 y 2025, lo que sugiere un retorno a condiciones normales del mercado asegurador. Este comportamiento refleja un cambio estructural en la serie, destacando el impacto temporal de la pandemia y la posterior reactivación económica.

Fig.11. Evolución temporal del valor del siniestro incurrido de la compañía de Seguros del Estado S.A. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

La Figura 12 muestra el patrón estacional del valor de los siniestros incurridos por mes y año, permitiendo identificar posibles comportamientos repetitivos a lo largo del tiempo. En términos generales, no se observa una estacionalidad fuerte y consistente, ya que los valores no siguen un patrón claramente repetitivo en los mismos meses para todos los años. Sin embargo, se identifican ciertas tendencias parciales. En los años previos a la pandemia (2015–2019), se evidencian valores relativamente más altos hacia los últimos meses del año, especialmente en noviembre y diciembre, donde se presentan algunos de los picos más significativos, lo que podría sugerir un leve efecto estacional asociado al cierre de periodos o mayor actividad económica. No obstante, este comportamiento no se repite de manera uniforme en todos los años. Durante el periodo de pandemia (2020–2021), se observa una reducción generalizada en los valores de los siniestros a lo largo de todos los meses, con una menor variabilidad y ausencia de picos marcados, lo que confirma el impacto de la COVID-19 en la dinámica del sector asegurador. Posteriormente, en el periodo postpandemia (2022–2025), se evidencia una recuperación en los niveles de siniestralidad progresiva de los valores, con incrementos en algunos meses específicos como septiembre, octubre y diciembre. A pesar de esto, la distribución sigue siendo irregular, sin un patrón estacional claramente definido. En conjunto, el análisis sugiere la presencia de una estacionalidad débil o no concluyente, predominando un comportamiento influenciado por factores externos y eventos específicos más que por ciclos periódicos regulares. Por otro lado, se evidencia la ausencia de registros correspondientes a los meses de octubre y diciembre del año 2025. No obstante, es importante señalar que para las etapas de entrenamiento y prueba de los modelos se consideraron únicamente los periodos comprendidos entre 2015 y 2024, los cuales cuentan con información completa.

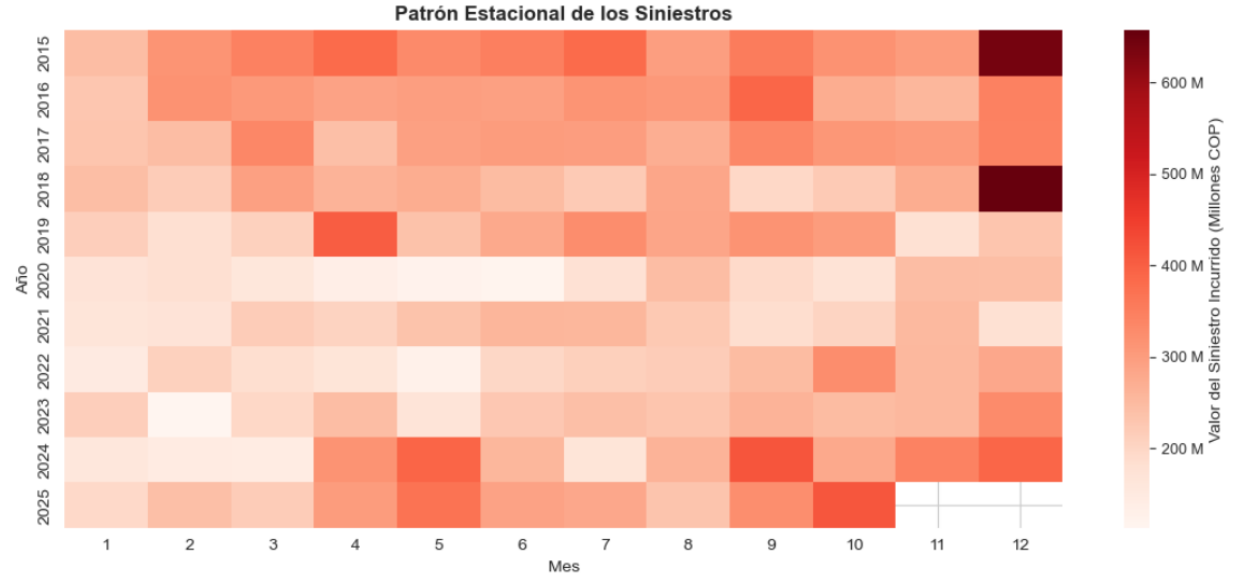


Fig.12. Mapa de calor de la estacionalidad mensual del valor del siniestro incurrido con comparación interanual. Fuente: Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

El análisis de valores atípicos mediante el método de Z-score permitió identificar observaciones con desviaciones significativas respecto al comportamiento promedio de la serie. En este sentido, se detectaron dos registros correspondientes a diciembre de los años 2015 y 2018, con valores de 642.102.200 COP y 657.039.100 COP, respectivamente, los cuales presentan puntuaciones Z de 4,53 y 4,71. Estos valores superan ampliamente el umbral común de $|Z| \geq 2{,}5$, lo que indica que se trata de eventos atípicos de alta magnitud dentro de la serie. La ocurrencia de estos picos puede estar asociada a siniestros de alta severidad o eventos extraordinarios, característicos del sector asegurador. Adicionalmente, se observa que ambos valores atípicos se concentran en el mes de diciembre, lo que podría sugerir una posible relación con cierres de periodo, ajustes contables o incrementos en la actividad económica al final del año. No obstante, dado que este comportamiento no se repite de forma consistente en todos los años, no puede considerarse como un patrón estacional, sino como eventos aislados.

TABLA XVII
ANOMALÍAS DETECTADAS

	FECHA	VALOR	Z_SCORE
0	2015-12-31	642102200	4.53
1	2018-12-31	657039100	4.71

Fuente: Elaboración propia.

III. MODELADO AVANZADO DE SERIES DE TIEMPO CON SARIMAX/ARIMA, XGBOOST Y LIGHTGBM - OBJETIVO 2

Los resultados del análisis exploratorio permiten caracterizar la estructura temporal de la serie del importe de las primas y el valor de siniestros incurridos, identificando componentes de tendencia, estacionalidad y variabilidad. Dado que estas características influyen directamente en la capacidad predictiva de los modelos, se procede al desarrollo de técnicas avanzadas de series de tiempo. En particular, se implementan los modelos SARIMA/ARIMA, XGBoost y LightGBM con el objetivo de evaluar su desempeño predictivo y comparar enfoques estadísticos tradicionales con métodos basados en aprendizaje automático.

Con el fin de garantizar la reproducibilidad de los resultados, todos los modelos de aprendizaje automático son implementados con una semilla aleatoria fija (random_state=42), asegurando que los resultados sean consistentes ante cualquier nueva ejecución del código. Las implementaciones se realizaron en Python, utilizando las librerías statsmodels para el modelo SARIMAX/ARIMA, XGBoost para el modelo de gradient boosting, LightGBM para el modelo de árboles potenciados y pandas, numpy y scikit-learn para el preprocesamiento y evaluación de los modelos. El código fuente completo se encuentra disponible en el Anexo B del presente documento.

A. Justificación metodológica de los modelos para el importe de primas y el valor de siniestros incurridos del ramo de Responsabilidad Civil Profesional

La selección de los modelos SARIMA/ARIMA, XGBoost y LightGBM se fundamenta en la naturaleza temporal de la variable analizada, correspondiente al importe de las primas y el valor de los de los siniestros incurridos, el cual presenta dependencia temporal, tendencia y posibles patrones estacionales. En este contexto, SARIMA/ARIMA resulta apropiado al tratarse de modelos estadísticos ampliamente utilizado en el análisis de series de tiempo, capaz de capturar la estructura temporal mediante componentes autorregresivos, de medias móviles y estacionales, lo que facilita la interpretación del comportamiento histórico de la serie.

Por otra parte, se incorporan los modelos XGBoost y LightGBM como enfoques de aprendizaje automático supervisado, reconocidos por su alta capacidad predictiva y su eficiencia en la modelación de relaciones no lineales. Mientras XGBoost se destaca por su robustez y precisión, LightGBM ofrece ventajas en términos de velocidad de entrenamiento y manejo eficiente de grandes volúmenes de datos. La inclusión de estos modelos permite evaluar si técnicas de machine learning pueden mejorar la precisión de las proyecciones frente a los métodos tradicionales de series temporales.

En consecuencia, la utilización conjunta de estos enfoques permite realizar una comparación metodológica basada en métricas de desempeño, con el fin de identificar el modelo que ofrezca mayor precisión en la estimación del comportamiento futuro del importe de las primas y el valor de los siniestros incurridos. La evaluación se realiza mediante estrategias de validación walk-forward y métricas de error como SMAPE y RMSE, garantizando así una selección objetiva del modelo más adecuado para la generación de pronósticos.

A partir del análisis comparativo presentado en la Tabla XVIII, se observa que los modelos ARIMA/SARIMAX, XGBoost y LightGBM ofrecen características adecuadas para el análisis del importe de las primas y el valor de los siniestros incurridos. Mientras que ARIMA/SARIMAX permite capturar la estructura temporal de la serie mediante componentes autorregresivos y estacionales, XGBoost y LightGBM aportan una mayor capacidad para identificar patrones no lineales presentes en los datos. En particular, XGBoost se destaca por su robustez y precisión predictiva, mientras que LightGBM ofrece ventajas en términos de eficiencia computacional y rapidez en el entrenamiento. Por esta razón, estos enfoques son seleccionados para el proceso de modelado y posteriormente comparados mediante métricas de desempeño, con el fin de determinar el modelo que proporcione las proyecciones más precisas.

Si bien ambos modelos de aprendizaje automático parten de la misma serie temporal y el mismo periodo de entrenamiento y prueba, existe una diferencia en el número de variables explicativas utilizadas: **XGBoost emplea 22 variables** mientras que **LightGBM incorpora 27 variables**. Esta diferencia se debe a que LightGBM, por su estrategia de crecimiento por hojas (*leaf-wise*) y su capacidad para manejar mayor dimensionalidad sin incrementar

significativamente el riesgo de sobreajuste, permite incorporar variables adicionales como diferencias de primer y segundo orden (diff_1 y diff_2) y máximos móviles en ventanas de 3, 6 y 12 periodos (rolling_max), las cuales enriquecen la representación de la volatilidad y los valores extremos de la serie. A pesar de esta diferencia, la comparación entre ambos modelos sigue siendo válida dado que ambos son evaluados sobre el **mismo periodo de prueba** y con las **mismas métricas de desempeño** (SMAPE, RMSE y MAE), lo que garantiza una evaluación objetiva e imparcial del rendimiento de cada enfoque.

TABLA XVIII
ANÁLISIS COMPARATIVO DE MODELOS PARA LA MODELACIÓN DEL IMPORTE DE PRIMAS

Modelo	Tipo de enfoque	Ventajas principales	Limitaciones	Pertinencia para el estudio
ARIMA	Estadístico (series de tiempo)	Modela tendencia y autocorrelación temporal; fácil interpretación; adecuado para series sin estacionalidad fuerte.	No considera componente estacional que es lo que refleja el valor de los siniestros.	Alto
SARIMAX	Estadístico (series de tiempo)	Captura tendencia, estacionalidad y dependencia temporal; alta interpretabilidad.	Supone relaciones lineales; requiere estacionariedad	Alto
XGBoost	Machine Learning (boosting)	Alta precisión predictiva; modela relaciones no lineales; robusto ante overfitting	Mayor complejidad; requiere ajuste de hiperparámetros.	Alto
LightGBM	Machine Learning (boosting)	Entrenamiento más rápido; eficiente con grandes volúmenes de datos; buen rendimiento.	Menor interpretabilidad; sensible a datos pequeños o mal estructurados	Alto

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Con base en lo anterior, en las siguientes secciones se presentan las pruebas estadísticas necesarias para evaluar las propiedades de la serie temporal, así como el procedimiento de ajuste, validación y comparación de los modelos de predicción propuestos.

B. Desarrollo de los modelos predictivos para la estimación de la producción de primas del ramo de Responsabilidad Civil Profesional

1) Ajuste de modelos predictivos

a) Modelo SARIMAX

i. Test de estacionariedad (Dickey-Fuller)

Con esta prueba se evalúa la estacionariedad de la serie mensual del importe de las primas una vez aplicada la diferenciación regular de orden uno y la diferenciación estacional de periodo 12, correspondientes a los parámetros $d = 1$ y $D = 1$ del modelo SARIMA. El objetivo es verificar que la serie transformada cumple el supuesto de estacionariedad requerido para la estimación del modelo. En el marco de la prueba Dickey-Fuller Aumentada (ADF), se plantean las siguientes hipótesis:

.

H_0: La serie tiene raíz unitaria (NO es estacionaria)

H_1**:** La serie NO tiene raíz unitaria (ES estacionaria)

Por otro lado, tenemos los criterios de decisión:

Si $p - value < 0.05 \rightarrow$ Rechazamos $H_0 \rightarrow$ Serie es estacionaria
Si $p - value \geq 0.05 \rightarrow$ No rechazamos $H_0 \rightarrow$ Serie NO es estacionaria

Nota. La variable FECHA fue verificada previamente y presenta una única observación por mes, por lo que la serie se encuentra correctamente estructurada en frecuencia mensual

La inspección visual de la serie temporal evidenció la presencia de variaciones recurrentes con periodicidad anual, consistentes con un comportamiento estacional. Dado que la magnitud de estas fluctuaciones no permanece constante entre periodos y presenta variaciones a lo largo del tiempo, se asume una componente estacional de naturaleza estocástica (s = 12). En consecuencia, se aplica una diferenciación regular de primer orden (d = 1) para remover la tendencia y una diferenciación estacional de primer orden (D = 1) para eliminar la componente estacional antes de realizar la prueba Dickey-Fuller Aumentada (ADF), con el propósito de verificar el cumplimiento del supuesto de estacionariedad requerido para la estimación del modelo SARIMA.

Los resultados de la prueba Dickey-Fuller Aumentada (ADF) evidencian que la serie transformada mediante diferenciación regular y estacional cumple con el supuesto de estacionariedad requerido para la modelación. El estadístico de prueba obtenido (-7.0874) es considerablemente menor que los valores críticos al 10% (-2.5746), 5% (-2.8762) y 1% (-3.4636), lo que proporciona una fuerte evidencia para rechazar la hipótesis nula de presencia de raíz unitaria. Asimismo, el p-value de 0.0000 confirma que la serie es estadísticamente estacionaria con un alto nivel de confianza. La prueba utilizó 15 rezagos, seleccionados automáticamente mediante el criterio de información de Akaike (AIC), con el propósito de controlar la autocorrelación residual y garantizar la validez estadística del contraste. Estos resultados indican que las transformaciones aplicadas logran eliminar adecuadamente los componentes de tendencia y estacionalidad presentes en la serie original, permitiendo que sus propiedades estadísticas permanezcan estables en el tiempo y justificando su utilización en la estimación del modelo SARIMA.

TABLA XIX
RESULTADOS DEL TEST DE DICKEY-FULLER AUMENTADA (ADF)

Estadístico	Valor	Interpretación
ADF Statistic	-7.0874	Estadístico de prueba
p-value	0.0000	☑ Estacionaria
Lags	15	Número de rezagos
Observaciones	199	Observaciones utilizadas
Valor crítico 10%	-2.5746	Umbral al 10%
Valor crítico 5%	-2.8762	Umbral al 5%
Valor crítico 1%	-3.4636	Umbral al 1%

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 13 constituye un soporte gráfico de los resultados obtenidos mediante la prueba Dickey-Fuller Aumentada (ADF) aplicada a la serie transformada mediante diferenciación regular y estacional. La representación visual muestra el criterio de decisión del contraste, donde la región sombreada a la izquierda del valor crítico al 5% (-2.8762) corresponde a la zona de rechazo de la hipótesis nula de presencia de raíz unitaria. Se observa que el estadístico calculado (ADF = -7.0874) se ubica claramente dentro de dicha región, corroborando el resultado estadístico previamente obtenido y proporcionando evidencia suficiente para rechazar la hipótesis nula de no estacionariedad. En consecuencia, se confirma que la serie transformada es estacionaria y cumple el supuesto requerido para la estimación del modelo SARIMA.

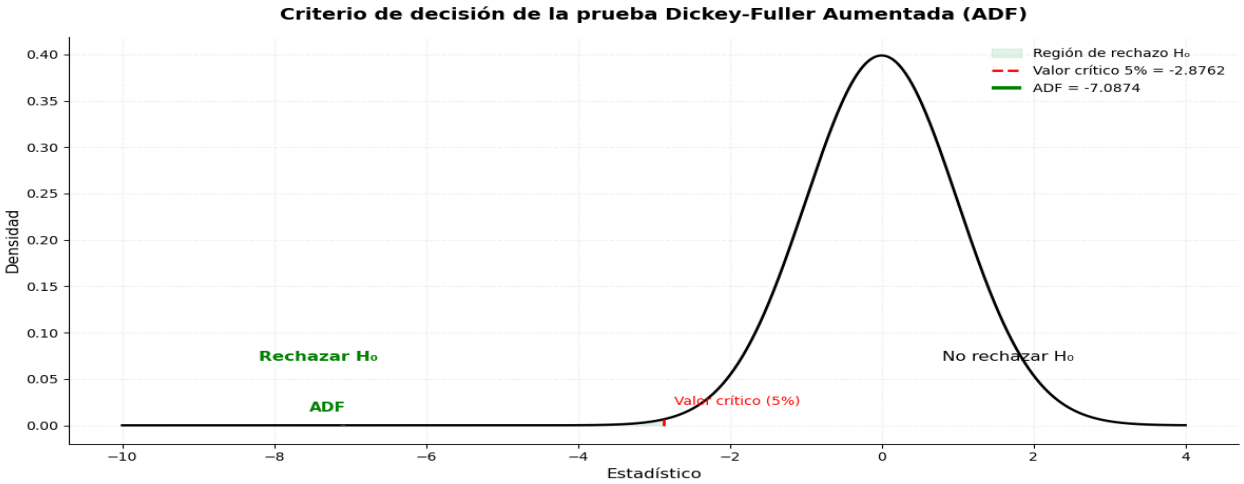


Fig.13. Criterio de decisión de la prueba Dickey-Fuller Aumentada (ADF) aplicada a la serie diferenciada regular y estacionalmente del importe de primas en el Ramo de Responsabilidad Civil Profesional. Elaboración propia.

ii. Análisis y visualización de transformaciones de la serie temporal

Con el fin de analizar las propiedades estadísticas de la serie temporal del importe de primas (IMP_PRIMA) y evaluar su comportamiento en términos de tendencia, varianza y estacionalidad, se aplican diferentes transformaciones con el objetivo de aproximarla a condiciones de estacionariedad requeridas por modelos como SARIMAX. La transformación logarítmica log(IMP_PRIMA+1) permite reducir la heterocedasticidad observada en la serie original, evidenciada en la Figura 14 por el crecimiento de la amplitud de las fluctuaciones a lo largo del tiempo.

Posteriormente, la primera diferencia elimina la tendencia ascendente, centrando la serie alrededor de cero, mientras que la diferencia estacional con rezago de 12 meses permite evidenciar patrones recurrentes anuales. Finalmente, la combinación de logaritmo, diferenciación y diferenciación estacional genera una serie más estable, con reducción de tendencia, estacionalidad y variabilidad, lo cual se aproxima a los supuestos de estacionariedad requeridos para el modelado SARIMAX. En conjunto, las transformaciones confirman que la serie original no es estacionaria, lo que justifica su tratamiento previo al modelamiento.

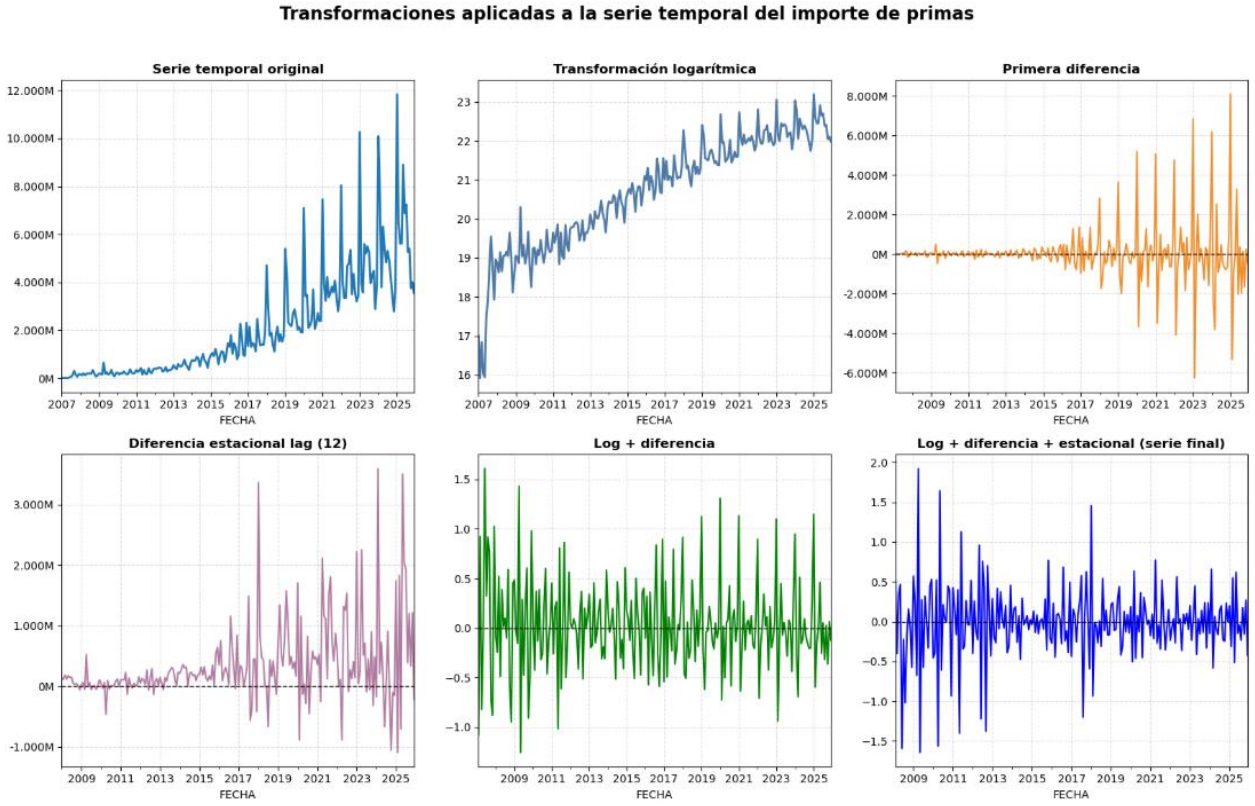


Fig.14. Transformaciones de la serie temporal del importe de las primas del ramo de Responsabilidad Civil profesional en Seguros del Estado S.A. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

iii. Análisis ACF y PACF para identificación del modelo

Las funciones de autocorrelación permiten identificar los posibles órdenes del modelo **SARIMA(p,d,q)** en su parte no estacional:

- **ACF (*Autocorrelation Function*)** → Sugiere el orden **q** del componente de **media móvil (MA)**
- **PACF (*Partial Autocorrelation Function*)** → Sugiere el orden **p** del componente **autorregresivo (AR)**

El análisis se realiza sobre la serie transformada mediante logaritmo log(IMP_PRIMA + 1), diferenciación de primer orden y diferenciación estacional, con el fin de garantizar la estacionariedad de la serie.

En la Figura 15 se presenta el análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie del importe de las primas, una vez aplicada la transformación logarítmica, la diferenciación de primer orden y la diferenciación estacional. En la ACF se observa que la mayoría de las autocorrelaciones se encuentran dentro de los intervalos de confianza, con algunos picos significativos, especialmente en el primer rezago y alrededor del rezago 12, lo que sugiere la presencia de componentes de media móvil y un posible efecto estacional. Por su parte, la PACF muestra un pico claramente significativo en el primer rezago, mientras que los rezagos posteriores se mantienen en su mayoría dentro de los límites, lo que indica la presencia de un componente autorregresivo de bajo orden, particularmente AR(1). En conjunto, estos resultados evidencian que la serie ha sido adecuadamente estacionarizada y respaldan la consideración de un modelo SARIMA(1,1,1)(1,1,1,12), adecuado para capturar la dinámica temporal y los patrones estacionales de la serie analizada.

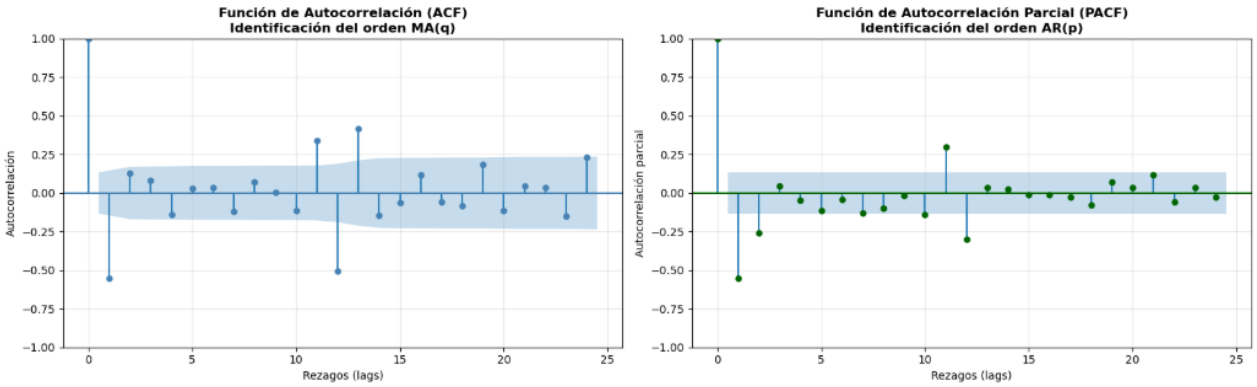


Fig.15. Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie log-diferenciada de la producción de primas. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

iv. Entrenamiento del modelo SARIMA y estimación de parámetros mediante SARIMAX en Python

Aunque el modelo especificado corresponde a un SARIMA (1,1,1) (1,1,1,12), en el entorno de Python se utiliza la implementación SARIMAX disponible en la librería *statsmodels*, ya que esta clase generaliza el modelo SARIMA al permitir la inclusión de variables exógenas. En este caso, al no incorporar variables externas, el modelo SARIMAX se comporta exactamente como un SARIMA, por lo que su uso no altera la estructura ni la interpretación del modelo. La elección de SARIMAX no responde a un cambio metodológico, sino a una limitación del lenguaje, dado que Python no dispone de una clase independiente llamada "SARIMA", sino que todo se implementa a través de SARIMAX por su mayor flexibilidad y capacidad de extensión.

El conjunto de datos fue dividido en 216 observaciones para entrenamiento (enero 2007 – diciembre 2024) y 12 observaciones para prueba (año 2025), equivalente al 95% y 5% respectivamente. El modelo fue entrenado exitosamente, evidenciado por la correcta convergencia del algoritmo de estimación sin errores, lo que indica que la especificación es adecuada para capturar la dinámica y estacionalidad de la serie.

El modelo muestra que los parámetros autorregresivos no son significativos, con AR(1) $\varphi_1 = -0.0605$ ($p = 0.4826$) y SAR(1) $\Phi_1 = 0.0129$ ($p = 0.7623$), lo que indica baja dependencia con valores pasados. En contraste, los componentes de medias móviles sí son significativos: MA(1) $\theta_1 = -0.9026$ ($p = 0.0000$) y SMA(1) $\Theta_1 = -0.5585$ ($p = 0.0000$), evidenciando que el modelo se ajusta principalmente corrigiendo errores previos, especialmente a nivel estacional.

TABLA XX
INTERPRETACIÓN DE PARÁMETROS CLAVE SARIMAX(1,1,1)(1,1,1,12)

Parámetro	Coeficiente	p-value	Significancia	Interpretación
AR(1) - φ_1	-0.0605	0.4826	❌ No significativo	Indica que las variaciones de la serie presentan dependencia con las variaciones del periodo inmediatamente anterior.
MA(1) - θ_1	-0.9026	0.0000	☑ Significativo	Los errores del periodo anterior influyen significativamente en la serie.
SAR(1) - Φ_1	0.0129	0.7623	❌ No significativo	No existe evidencia de dependencia estacional autorregresiva.
SMA(1) - Θ_1	-0.5585	0.0000	☑ Significativo	Indica que el modelo corrige errores estacionales pasados, lo que sugiere una estructura estacional relevante en la serie.

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

En términos generales, los resultados del modelo SARIMAX(1,1,1)(1,1,1,12) evidencian que la dinámica de la serie está principalmente explicada por el componente de media móvil tanto en su parte no estacional como estacional, mientras que el componente autorregresivo no presenta significancia estadística. Esto sugiere que la serie depende más de los errores pasados que de sus propios valores históricos, lo cual es característico de procesos con alta variabilidad y memoria en los shocks. Adicionalmente, la significancia del componente estacional de media móvil indica la presencia de una estructura estacional importante que el modelo logra capturar parcialmente. Sin embargo, la falta de

significancia del componente SAR(1) sugiere que la estructura estacional autorregresiva no está siendo completamente representativa, lo que motiva la exploración de modelos alternativos con mayor flexibilidad en la componente estacional.

Debido a los resultados anteriores, especialmente la falta de significancia en el componente autorregresivo estacional y la presencia de patrones de autocorrelación en los residuos se procede a ajustar la estructura del modelo incrementando el orden de la componente estacional, con el objetivo de mejorar la captura de la dependencia temporal.

La Tabla XXI presenta el modelo finalmente ajustado SARIMAX (1,1,1)(2,1,2,12), el cual está dominado principalmente por la componente de media móvil y la estacionalidad. El término AR(1) no es significativo, lo que indica baja dependencia con el pasado inmediato, mientras que el MA(1) resulta altamente significativo, evidenciando que los errores recientes explican gran parte de la dinámica de la serie. En la estructura estacional, los efectos en el rezago 12 son relevantes tanto en SAR como en SMA, lo que confirma una estacionalidad anual bien capturada. Aunque el SAR(24) no es significativo, el SMA(24) sí aporta ajuste estacional, sugiriendo que la estructura principal está explicada por el primer ciclo estacional y complementada parcialmente por el segundo. Por lo tanto, los análisis de diagnóstico, validación y pronóstico presentados a continuación se realizan con base en este modelo ajustado.

TABLA XXI
INTERPRETACIÓN DE PARÁMETROS CLAVE SARIMAX (1,1,1)(2,1,2,12)

Parámetro	Coeficiente	p-value	Significancia	Interpretación
AR(1) - φ_1	0.0505	0.5801	❌	No significativo → baja dependencia temporal
MA(1) - θ_1	-0.9212	0.0000	☑	Significativo → influencia de errores pasados
SAR (ar.S.L12)	-0.6782	0.0000	☑	Efecto estacional autorregresivo
SAR (ar.S.L24)	-0.0533	0.2076	❌	Efecto estacional autorregresivo
SMA (ma.S.L12)	0.2554	0.0301	☑	Corrección de errores estacionales
SMA (ma.S.L24)	-0.2172	0.0382	☑	Corrección de errores estacionales

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

v. Diagnostico residuos

La Tabla XXII muestra que los residuos del modelo presentan una media cercana a cero (0.007), lo que indica ausencia de sesgo sistemático, y una desviación estándar estable, sugiriendo una adecuada estabilidad en la varianza. Sin embargo, los altos valores de skewness (7.73) y kurtosis (112.63) evidencian una distribución fuertemente asimétrica y con colas pesadas, lo que refleja la presencia de valores atípicos y desviaciones respecto a la normalidad, aunque en términos generales el comportamiento del modelo se mantiene estable.

TABLA XXII
ESTADÍSTICAS DESCRIPTIVAS DE LOS RESIDUOS DEL MODELO SARIMAX (1,1,1)(2,1,2,12)

	Métrica	Valor	Ideal	Evaluación
0	Media	0.007156	≈ 0	☑
1	Desviación Estándar	1.379373	Pequeña	☑
2	Skewness	7.730800	≈ 0	⚠
3	Kurtosis	112.637500	≈ 0	⚠

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La versión resumida del test de Ljung-Box mostrada en la Tabla XXIII se construyó considerando la metodología de Box-Jenkins, en la cual la independencia de los residuos se evalúa mediante los correlogramas ACF y PACF, utilizando un número de rezagos equivalente a $\frac{n}{4}$. En este caso, el análisis se realizó con 216 observaciones, lo que dio lugar a 54 rezagos evaluados. Para facilitar la interpretación, los resultados se organizaron de forma resumida a partir del comportamiento general de los p-values en dicho conjunto de rezagos. Se observó que en todos los rezagos

evaluados los p-values son superiores a 0.05, por lo que no se rechaza la hipótesis nula de ausencia de autocorrelación. En consecuencia, los residuos del modelo SARIMAX (1,1,1)(2,1,2,12) se comportan como ruido blanco.

Test de Ljung-Box (Autocorrelación de Residuos)

Hipótesis

H_0: Los residuos NO tienen autocorrelación ☑

H_1: Los residuos SÍ tienen autocorrelación ❌

TABLA XXIII
RESULTADOS DEL TEST DE LJUNG-BOX PARA LOS RESIDUOS DEL MODELO SARIMAX (1,1,1)(2,1,2,12)

	lb_stat	lb_pvalue	Evaluación
1	2.9173	0.0876	☑ Ruido blanco
2	4.4424	0.1085	☑ Ruido blanco
3	4.8970	0.1795	☑ Ruido blanco
4	5.0046	0.2868	☑ Ruido blanco
5	6.6085	0.2514	☑ Ruido blanco
…	…	…	…
50	40.7136	0.8226	☑ Ruido blanco
51	40.7273	0.8479	☑ Ruido blanco
52	41.1612	0.8601	☑ Ruido blanco
53	41.2364	0.8798	☑ Ruido blanco
54	41.2607	0.8984	☑ Ruido blanco

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 16 muestra que los residuos presentan un comportamiento mayormente centrado alrededor de cero a lo largo del tiempo, lo que indica que el modelo logra capturar adecuadamente la dinámica de la serie; sin embargo, al inicio del período (aproximadamente entre 2007 y 2009) se observan valores atípicos importantes, con residuos positivos y negativos de gran magnitud, evidenciando episodios de alta variabilidad o posibles choques no explicados por el modelo. El histograma confirma que la mayoría de los residuos se concentran cerca de cero, aunque la distribución presenta asimetría debido a estos valores extremos. Por otra parte, los correlogramas ACF y PACF muestran que la mayoría de las autocorrelaciones permanecen dentro de las bandas de confianza, sugiriendo que los residuos se aproximan a un ruido blanco y que no existe una estructura temporal fuerte remanente; no obstante, algunos rezagos aislados superan ligeramente los límites, indicando la posible presencia de dependencias residuales menores que podrían asociarse con componentes no capturados completamente por el modelo SARIMAX (1,1,1)(2,1,2,12).

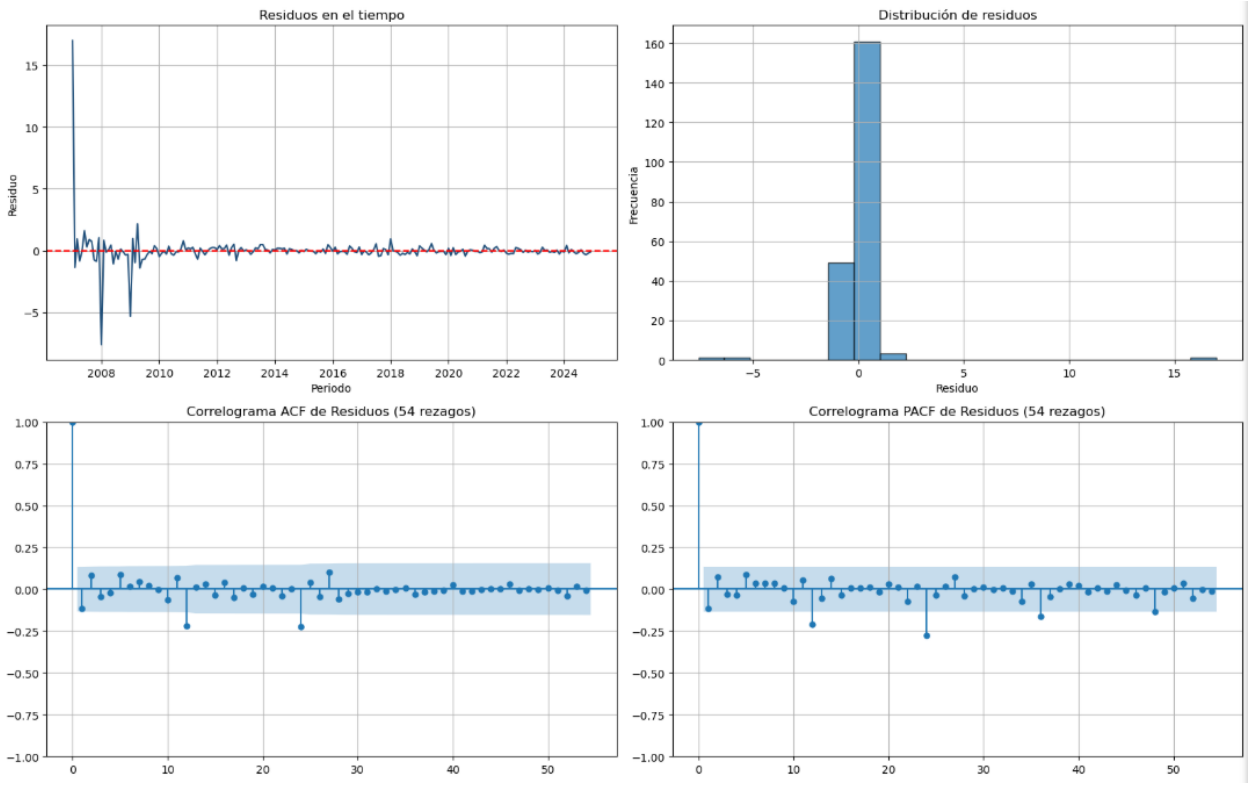


Fig.16. Gráficos de diagnóstico de residuos del modelo SARIMAX (1,1,1) (2,1,2,12). Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

vi. Generación de pronósticos con el modelo

La Tabla XXIV evidencia que el modelo presenta una capacidad aceptable para seguir el comportamiento de la serie durante 2025, ya que los valores reales se mantienen dentro de los intervalos de confianza del 95%. Sin embargo, se observan mayores desviaciones en algunos meses, especialmente en mayo (-33,50%), marzo (-24,48%) y diciembre (21,66%), indicando una tendencia del modelo a subestimar los valores reales en varios periodos. A pesar de ello, los errores más bajos se presentan en enero (-5,08%), febrero (3,09%) y agosto (-4,18%), reflejando un mejor ajuste en estos meses.

TABLA XXIV
COMPARACIÓN ENTRE VALORES REALES Y PRONOSTICADOS DEL MODELO SARIMAX(1,1,1)(2,1,2,12) CON INTERVALO DE CONFIANZA (95%)

Fecha	Valor real	Pronóstico	IC 95% Inferior	IC 95% Superior	Error %
2025-01	$11.836.236.872	$11.234.936.638	$6.901.379.286	$18.289.648.491	-5,08%
2025-02	$6.506.747.459	$6.708.069.264	$4.103.948.283	$10.964.610.210	3,09%
2025-03	$5.618.530.254	$4.243.391.558	$2.591.523.520	$6.948.180.009	-24,48%
2025-04	$5.612.586.021	$6.403.222.081	$3.904.091.382	$10.502.124.313	14,09%
2025-05	$8.898.015.045	$5.917.363.264	$3.601.914.058	$9.721.272.476	-33,50%
2025-06	$6.874.746.803	$5.654.302.804	$3.436.136.058	$9.304.387.152	-17,75%
2025-07	$7.238.561.833	$6.042.614.524	$3.666.103.376	$9.959.672.856	-16,52%
2025-08	$5.261.808.673	$5.041.739.123	$3.053.873.469	$8.323.571.240	-4,18%
2025-09	$5.399.232.258	$4.852.885.881	$2.934.701.285	$8.024.837.652	-10,12%
2025-10	$3.752.360.005	$4.256.853.509	$2.570.087.818	$7.050.654.716	13,44%
2025-11	$4.001.648.443	$3.302.881.018	$1.990.901.985	$5.479.437.512	-17,46%
2025-12	$3.540.110.800	$4.306.785.409	$2.591.851.485	$7.156.428.778	21,66%

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

vii. Evaluación del modelo

Validación Hold Out forecasting

El SMAPE de 16.10% indica que el modelo presenta un error porcentual moderado, lo que refleja una buena capacidad de ajuste en términos relativos, ya que las predicciones se desvían en promedio alrededor de ese porcentaje respecto a los valores reales. El RMSE de 1.169.036 evidencia que, en términos absolutos, los errores son elevados, lo que sugiere variaciones significativas entre los valores pronosticados y los observados, especialmente en la escala original de la serie. Por su parte, el MAE de 925.149.79 muestra que el error promedio absoluto también es considerable, indicando que, en promedio, las predicciones se alejan de manera importante de los valores reales. En conjunto, las métricas reflejan un desempeño aceptable en términos porcentuales, pero con errores absolutos de gran magnitud.

TABLA XXV
EVALUACIÓN DEL DESEMPEÑO DEL MODELO SARIMAX MEDIANTE SMAPE, RMSE Y MAE

Métrica	Valor	Interpretación
SMAPE	16.10%	Error porcentual moderado (bueno)
RMSE	1.169.036.902	Error alto en escala absoluta
MAE	925.149.279	Error promedio absoluto elevado

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 17 muestra que el importe de prima presenta una tendencia creciente sostenida a lo largo del tiempo, con un aumento más acelerado desde aproximadamente 2016 y una alta volatilidad reflejada en picos pronunciados. Al iniciar el conjunto de prueba, se observa una disminución en los valores reales, aunque mantienen variabilidad. El modelo SARIMAX (1,1,1)(2,1,2,12) que se ajusta logra capturar la tendencia general y el cambio a la baja, ajustándose de forma razonable a los datos reales, aunque subestima algunos picos y no reproduce completamente las fluctuaciones extremas. El intervalo de confianza al 95% es amplio, lo que indica incertidumbre en el pronóstico, pero en general contiene los valores observados, sugiriendo un buen desempeño global del modelo, aunque con limitaciones ante cambios bruscos.

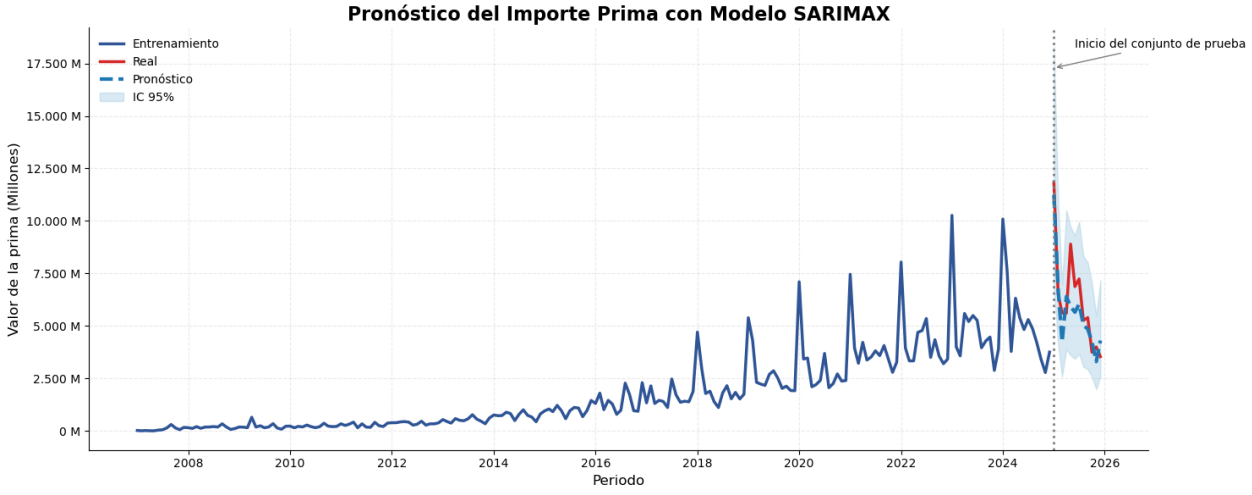


Fig.17. Evolución y Pronóstico del Importe de Prima. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

b) Modelo XGBoost

En esta etapa se realiza la ingeniería de características *(features)* con el objetivo de capturar patrones temporales,

estacionalidad y variabilidad en la serie de datos, mejorando así la capacidad predictiva del modelo.

Se generan las siguientes variables:

- *Rezagos (lags):* Se utilizan 12 variables de rezago porque la serie presenta una estacionalidad anual, es decir, patrones que se repiten cada 12 meses. Incluir los lags de 1 a 12 permite capturar tanto la dependencia de corto plazo como el efecto estacional completo dentro de un ciclo anual, lo cual es consistente con la metodología de Box-Jenkins para series mensuales.
- *Medias móviles:* se calculan medias móviles con ventanas de 3, 6 y 12 meses, con el fin de suavizar la serie y resaltar tendencias subyacentes.
- *Desviación estándar móvil:* se estima la desviación estándar en ventanas de 3, 6 y 12 meses, lo que permite capturar la variabilidad y volatilidad de los datos en distintos horizontes temporales.
- *Variables temporales:* se incorporan variables asociadas al mes, trimestre y año, con el propósito de modelar efectos estacionales y patrones recurrentes.
- *Tendencia:* se añade una variable de tendencia para representar el comportamiento general de crecimiento o decrecimiento de la serie a lo largo del tiempo.

En total, se construyen 22 características, las cuales se utilizan como variables explicativas en el entrenamiento del modelo predictivo.

i. Ingeniería de características (Feature engineering)

La Tabla XXVI presenta un subconjunto de las variables generadas a partir de la serie temporal de IMP_PRIMA, incluyendo rezagos y métricas móviles que se utilizan como variables explicativas del modelo. Se observa que la serie presenta fluctuaciones en el tiempo, evidenciando un comportamiento variable. Los rezagos (lag_1, lag_2, …) reflejan los valores históricos y permiten capturar la dependencia temporal entre observaciones, mientras que la media móvil de corto plazo suaviza las variaciones y facilita la identificación de la tendencia. Por su parte, la desviación estándar móvil evidencia cambios en la volatilidad, mostrando periodos de mayor o menor dispersión. En conjunto, estas variables enriquecen la información del modelo al incorporar tanto la memoria temporal como patrones de tendencia y variabilidad, lo que contribuye a mejorar su capacidad predictiva.

TABLA XXVI
VARIABLES DERIVADAS PARA EL MODELO DE SERIES TEMPORALES

FECHA	IMP_PRIMA	lag_1	lag_2	…	rolling_mean_3	rolling_std_3	…
2008-01-01	156,659,100	170,952,200	61,217,140		129,609,500	59,659,080	
2008-02-01	122,694,900	156,659,100	170,952,200		150,102,000	24,787,850	
2008-03-01	206,757,400	122,694,900	156,659,100		162,037,100	42,288,530	
2008-04-01	126,651,000	206,757,400	122,694,900		152,034,400	47,432,750	
2008-05-01	186,548,800	126,651,000	206,757,400		173,319,100	41,659,670	

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

ii. Selección de variables predictoras

Para el modelo XGBoost se evalúa un enfoque experimental mediante la comparación de diferentes configuraciones de variables predictoras (22, 25 y 27 variables), con el objetivo de analizar el impacto de la ingeniería de características en el desempeño del modelo. Los resultados evidencian un comportamiento no monótono en las métricas de evaluación, observándose que el incremento en el número de variables no siempre se traduce en una mejora en el error de predicción. En particular, la configuración de 25 variables presenta un incremento en el error respecto a la configuración de 22 variables, mientras que la configuración de 27 variables muestra una reducción del error, lo que evidencia sensibilidad del modelo frente a la selección de características. En este contexto experimental, se selecciona

la configuración de 22 variables debido a que presenta un desempeño estable y competitivo, sin depender de variaciones en la incorporación de variables adicionales. Esta configuración permite una representación más parsimoniosa de la serie temporal, conservando las variables más representativas del comportamiento histórico y estacional.

TABLA XXVII
COMPARACIÓN DEL DESEMPEÑO DEL MODELO XGBOOST SEGÚN EL NÚMERO DE VARIABLES PREDICTORAS

Modelo	Variables predictoras	SMAPE (%)	RMSE	MAE
XGBoost	22	21,45	1.620.642.000	1.322.542.000
XGBoost	25	21,85	1.701.958.000	1.357.825.000
XGBoost	27	18,12	1.394.252.000	1.093.309.000

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. División Train / Test

Se realiza la división del conjunto de datos en dos subconjuntos: entrenamiento y prueba. Estos datos corresponden a una serie temporal agregada mensualmente a partir de los 7.932 registros transaccionales del ramo de Responsabilidad Civil Profesional, que abarca el periodo enero de 2007 hasta diciembre de 2025. Sin embargo, debido a la construcción de variables de rezago de hasta 12 periodos durante el proceso de ingeniería de características, las primeras 12 observaciones correspondientes al año 2007 no pueden ser utilizadas al no contar con historia previa suficiente, por lo que el conjunto de datos efectivo inicia en enero de 2008. El conjunto de entrenamiento está compuesto por 204 observaciones mensuales que van desde enero de 2008 hasta diciembre de 2024 y 22 variables (*features*), las cuales se utilizan para ajustar el modelo predictivo. Por su parte, el conjunto de prueba contiene 12 observaciones correspondientes al año 2025 y las mismas 22 variables, empleándose para evaluar el desempeño del modelo XGBoost en datos no utilizados durante el entrenamiento. Esta partición representa aproximadamente el 94% de los datos para entrenamiento y el 6% para prueba.

iv. Entrenamiento del modelo

En esta sección se implementa el modelo basado en el algoritmo XGBoost, el cual corresponde a un método de ensamblado que construye múltiples árboles de decisión de manera secuencial, optimizando los errores residuales en cada iteración.

El modelo se desarrolla utilizando la clase *XGBRegressor*, orientada a problemas de regresión, con una función objetivo de tipo regresión de error cuadrático *(reg:squarederror)*, adecuada para modelar variables continuas como la serie analizada.

La estructura del modelo se define mediante los siguientes hiperparámetros:

TABLA XXVIII
CONFIGURACIÓN DE HIPERPARÁMETROS DEL MODELO XGBOOST

Parámetro	Valor	Descripción
objective	reg:squarederror	Función objetivo para problemas de regresión (minimiza el error cuadrático)
n_estimators	100	Número de árboles que se construyen en el modelo
max_depth	5	Profundidad máxima de cada árbol de decisión
learning_rate	0.1	Tasa de aprendizaje que controla el impacto de cada árbol
subsample	0.8	Proporción de datos utilizada en cada iteración
colsample_bytree	0.8	Proporción de variables utilizadas para construir cada árbol
random_state	42	Semilla para garantizar la reproducibilidad del modelo

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Los resultados presentados en la Tabla XXIX muestran que el modelo XGBoost otorgó mayor importancia a las variables asociadas con la volatilidad y el comportamiento histórico de la serie temporal. La característica más influyente fue *rolling_std_12* (0.297890), indicando que la variabilidad observada durante los últimos 12 períodos tuvo un papel clave en la predicción. Asimismo, *lag_12* (0.245465) y *rolling_mean_3* (0.223340) evidencian que el modelo considera relevante tanto el comportamiento registrado un año atrás como la tendencia reciente de corto plazo. Variables como *lag_11*, *month* y *quarter* también aportan información al modelo, aunque con menor peso relativo, mientras que características como lag_6 presentan una influencia reducida en el proceso predictivo.

TABLA XXIX
IMPORTANCIA DE CARACTERÍSTICAS (FEATURE IMPORTANCE)

	Feature	Importance
17	rolling_std_12	0.297890
11	lag_12	0.245465
12	rolling_mean_3	0.223340
10	lag_11	0.089055
14	rolling_mean_6	0.032158
18	month	0.022745
13	rolling_std_3	0.018598
15	rolling_std_6	0.015809
19	quarter	0.012916
5	lag_6	0.006326

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 18 evidencia que el modelo *XGBoost* asignó mayor relevancia a las variables relacionadas con la variabilidad y el comportamiento histórico reciente de la serie, destacándose *rolling_std_12*, *lag_12* y *rolling_mean_3* como las características más influyentes en el proceso de predicción. Esto sugiere que el modelo basa gran parte de sus pronósticos en patrones temporales previos y en la dinámica de corto y mediano plazo de la serie.

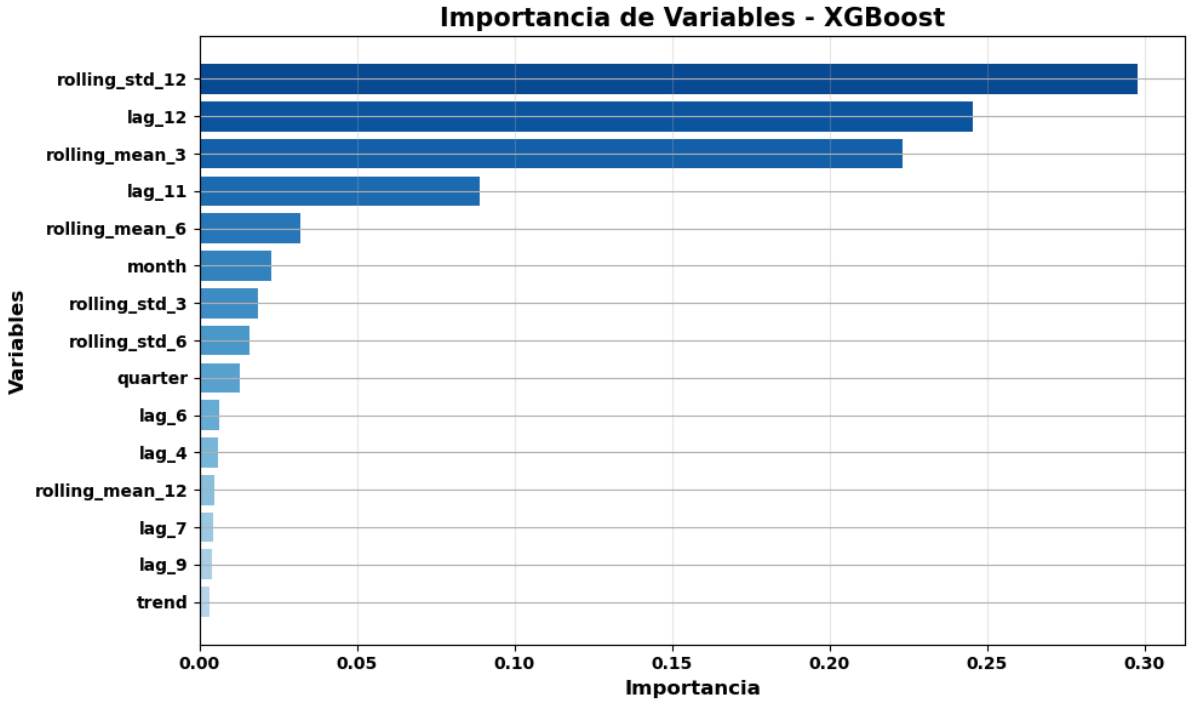


Fig.18. Importancia de variables del modelo XGBoost. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

v. Evaluación del modelo

Validación Hold Out forecasting

Las métricas obtenidas en la Tabla XXX evidencian el desempeño del modelo en la predicción de la serie temporal. El valor de SMAPE (21,45%) indica que, en promedio, el error relativo entre los valores reales y los pronosticados es moderado, lo que sugiere una capacidad aceptable de ajuste del modelo. Por su parte, el RMSE (1.620.642.253) muestra la magnitud del error penalizando en mayor medida las desviaciones grandes, lo que evidencia la presencia de algunas diferencias significativas entre los valores reales y estimados. En complemento, el MAE (1.322.542.018) refleja el error absoluto promedio en las mismas unidades de la variable, indicando que, en términos generales, las predicciones se desvían de los valores reales en una magnitud considerable. En conjunto, estos resultados sugieren que el modelo logra capturar la tendencia general de la serie, aunque presenta oportunidades de mejora en la precisión de los pronósticos.

TABLA XXX
MÉTRICAS DE DESEMPEÑO DEL MODELO XGBOOST

Métrica	Valor	Descripción
SMAPE	21,45 %	Error porcentual medio simétrico
RMSE	$ 1.620.642.253	Raíz del error cuadrático medio
MAE	$ 1.322.542.018	Error absoluto medio

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

vi. Pronóstico con modelo XGBoost

En la Figura 19 muestra que durante el periodo de entrenamiento se observan fluctuaciones marcadas con incrementos abruptos, mientras que en la fase de pronóstico el modelo logra captar la tendencia general, aunque suaviza los picos más extremos. Esto indica que el modelo reconoce la dinámica global de la serie, pero presenta limitaciones para anticipar cambios bruscos, reflejando una buena aproximación en tendencia, pero menor precisión en la variabilidad.

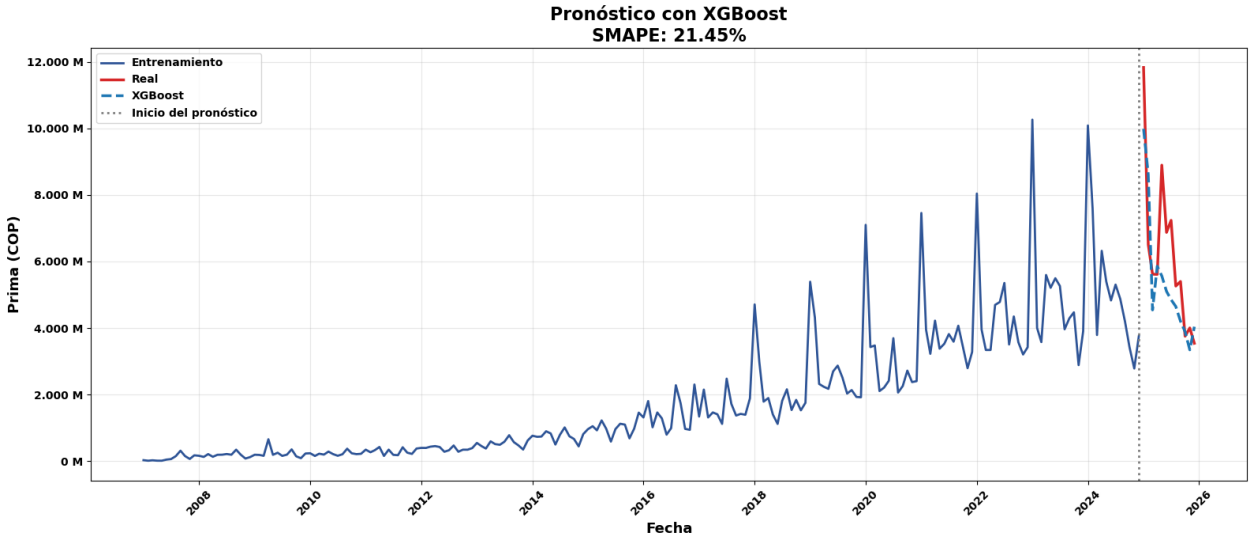


Fig.19. Pronóstico de la serie temporal del importe de prima, incluyendo datos de entrenamiento, valores reales y predicción del modelo. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

c) Modelo LightGBM

Para el desarrollo del modelo predictivo se utiliza el algoritmo LightGBM, un método de aprendizaje basado en árboles de decisión que permite capturar relaciones no lineales y patrones complejos en series de tiempo.

Previamente, se realiza un proceso de ingeniería de características *(feature engineering)* con el objetivo de transformar la serie original en un conjunto de variables explicativas más informativas. En este proceso se generaron diferentes tipos de variables. En primer lugar, se incluyen rezagos *(lags*) desde el periodo 1 hasta el 12, los cuales permiten incorporar información histórica de la variable objetivo y capturar dependencias temporales tanto de corto como de mediano plazo. Adicionalmente, se calculan diferencias de primer y segundo orden *(diff_1 y diff_2)* con el fin de modelar los cambios entre periodos consecutivos, facilitando la identificación de variaciones y tendencias en la serie.

Asimismo, se construyen variables de tipo ventana móvil *(rolling),* incluyendo medias y desviaciones estándar para ventanas de 3, 6 y 12 periodos. Estas variables permiten resumir el comportamiento reciente de la serie, aportando información relevante sobre su nivel y volatilidad. Como complemento, se incorporan máximos móviles *(rolling_max)* para ventanas de 3, 6 y 12 periodos, con el propósito de capturar picos o valores extremos en diferentes horizontes de tiempo, lo que contribuye a mejorar la capacidad del modelo para identificar comportamientos atípicos.

Finalmente, se agregan variables temporales, tales como el mes, el trimestre y el año, junto con una variable de tendencia, con el fin de capturar patrones estacionales y la evolución general de la serie a lo largo del tiempo.

i. Ingeniería de características (Feature engineering) del conjunto de datos

La Tabla XXXI presenta una muestra representativa de las variables derivadas utilizadas en el modelo LightGBM, construidas a partir de la serie temporal de la variable IMP_PRIMA. Estas variables incluyen rezagos, estadísticas móviles, diferencias y componentes temporales que permiten enriquecer la información de la serie. Debido a la alta dimensionalidad del conjunto de datos, solo se muestra una parte de la tabla como referencia ilustrativa del proceso de transformación aplicado.

TABLA XXXI
INGENIERÍA DE CARACTERÍSTICAS (FEATURE ENGINEERING) DEL CONJUNTO DE DATOS

FECHA	IMP_PRIMA	lag_1	lag_2		rolling_mean_3	rolling_std_6	rolling_max_6	month	quarter	year	trend
2008-01-01	156.659.100	170.952.165	61.217.138	…	129.609.500	80.421.562	309.092.997	1	1	008	2
2008-02-01	122.694.875	156.659.100	170.952.165	…	150.102.000	82.025.646	309.092.997	2	1	008	3
2008-03-01	206.757.404	122.694.875	156.659,100	…	162.037.100	49.289.448	309.092.997	3	1	008	4
2008-04-01	126.651.020	206.757.404	122.694.875	…	152.034.400	49.744.269	309.092.997	4	2	008	5
2008-05-01	186.548.831	126.651.020	206.757.404	…	173.319.100	33.191.798	206.757.404	5	2	008	6
2008-06-01	189.200.256	186.548.831	126.651.020	…	167.466.700	34.995.012	206.757.404	6	2	008	7
2008-07-01	208.950.016	189.200.256	186.548.831	…	194.899.700	38.872.894	208.950.016	7	3	008	8
2008-08-01	188.497.889	208.950.016	189.200.256	…	195.549.400	29.939.763	208.950.016	8	3	008	9
2008-09-01	340.172.970	188.497.889	208.950.016	…	245.873.600	71.093.681	340.172.970	9	3	008	0
2008-10-01	190.721.587	340.172.970	188.497.889		239.797.500	60.726.730	340.172.970	10	4	008	1

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Tabla XXXII presenta el Top 10 de variables más importantes utilizadas por el modelo LightGBM, evidenciando que las variables de tipo diferencia (diff_1 y diff_2) son las más influyentes en la predicción de la serie temporal. Asimismo, se destaca la relevancia de algunos rezagos como lag_6, lag_10, lag_11 y lag_12, lo que indica la importancia de la información histórica en diferentes horizontes temporales. También aparecen variables de tipo estadístico como rolling_std_3 y rolling_mean_3, las cuales capturan la variabilidad y el comportamiento reciente de la serie. Finalmente, la variable month aporta un componente estacional, aunque con menor impacto relativo dentro del modelo. En conjunto, estos resultados muestran que el modelo combina información de tendencia, estacionalidad y memoria temporal para generar sus predicciones.

TABLA XXXII
TOP 10 VARIABLES MÁS IMPORTANTES EN EL MODELO LIGHTGBM

	Variable	Importancia
12	diff_1	137
13	diff_2	125
15	rolling_std_3	119
5	lag_6	102
23	month	98
10	lag_11	89
11	lag_12	88
9	lag_10	69
7	lag_8	68
14	rolling_mean_3	67

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 20 es una visualización que complementa los resultados anteriores al mostrar que LightGBM asignó mayor importancia a las variables relacionadas con cambios recientes de la serie (diff_1 y diff_2), seguidas de medidas de variabilidad (rolling_std_3) y rezagos temporales (lag_6, lag_11, lag_12). Esto evidencia que el modelo basa sus predicciones principalmente en el comportamiento reciente e histórico de la serie temporal.

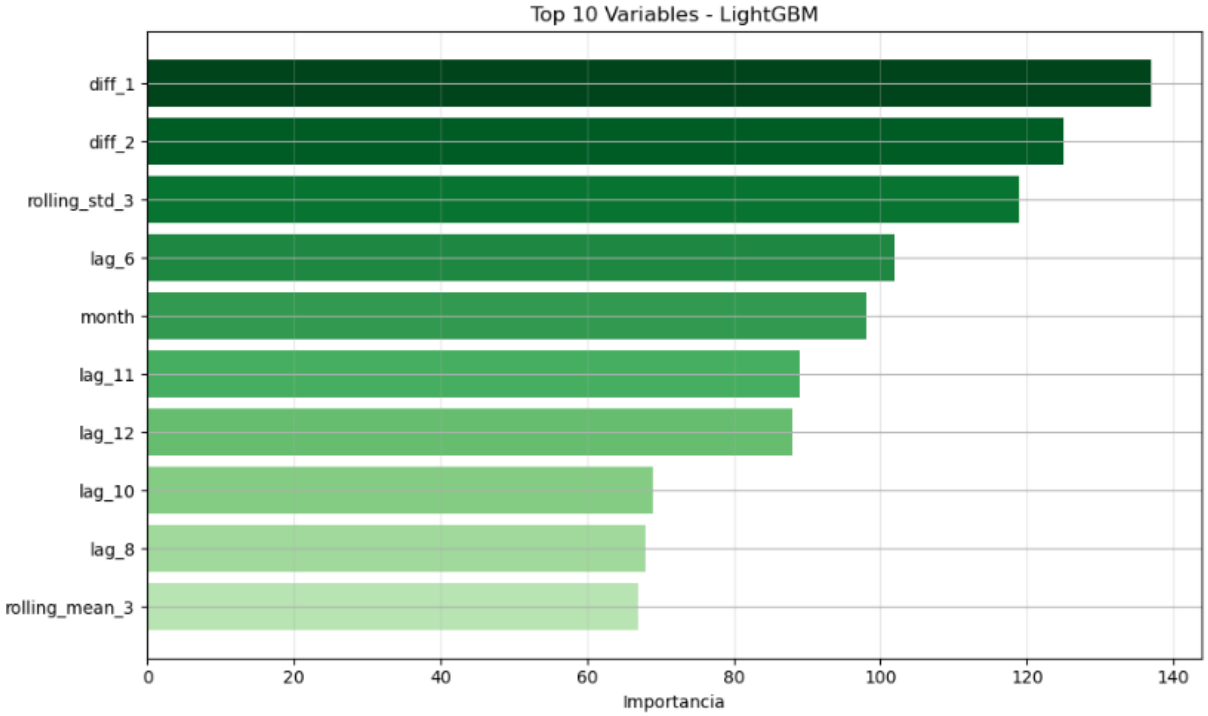


Fig.20. Top 10 de las variables más importantes del modelo LightGBM. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

ii. Selección de Variables predictoras

Para determinar la cantidad adecuada de variables de entrada del modelo LightGBM, se realiza un proceso experimental evaluando diferentes configuraciones de características predictoras. Inicialmente se generaron 27

variables mediante ingeniería de características, incluyendo rezagos temporales, variables estadísticas móviles y componentes temporales. Posteriormente, se entrena el modelo con tres configuraciones de entrada (22, 25 y 27 variables) manteniendo las mismas condiciones de entrenamiento y evaluación. El desempeño fue comparado mediante las métricas SMAPE, RMSE y MAE, seleccionando la configuración con mejor desempeño general considerando las métricas evaluadas.

TABLA XXXIII
COMPARACIÓN DEL DESEMPEÑO DEL MODELO LIGHTGBM SEGÚN EL NÚMERO DE VARIABLES PREDICTORAS

Modelo	Variables	SMAPE (%)	RMSE	MAE
LightGBM	22	23,14	5.466.938.000	1.401.423.000
LightGBM	25	13,78	5.660.397.000	911.725.500
LightGBM	27	13,37	5.669.581.076	888.132.519

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Finalmente, la configuración con 27 variables fue seleccionada dentro de las configuraciones evaluadas debido a que presentó el menor valor de SMAPE y MAE, logrando un mejor equilibrio entre precisión porcentual y error absoluto promedio, además de conservar más información temporal de la serie.

iii. División Train / Test

El conjunto de datos utilizado para el modelo LightGBM mantiene la misma partición temporal definida anteriormente, con 204 observaciones para entrenamiento (enero 2008 – diciembre 2024) y 12 observaciones para prueba (año 2025). A diferencia del modelo XGBoost, este modelo incorpora 27 variables explicativas, dado que LightGBM, por su estrategia de crecimiento por hojas (*leaf-wise*) y su capacidad para manejar mayor dimensionalidad, puede aprovechar un conjunto más amplio de características sin incrementar significativamente el riesgo de sobreajuste, lo que justifica la incorporación de variables adicionales como diferencias (diff_1 y diff_2) y máximos móviles en ventanas de 3, 6 y 12 periodos para enriquecer la representación de la volatilidad y los valores extremos de la serie.

iv. Entrenamiento del modelo

La Tabla XXXIV presenta la configuración de hiperparámetros utilizada para el modelo LightGBM, los cuales controlan su complejidad, velocidad de aprendizaje y capacidad de generalización. En conjunto, estos parámetros permiten equilibrar el ajuste del modelo, reduciendo el riesgo de sobreajuste mediante la limitación de la profundidad de los árboles y el uso de submuestreo de datos y variables. Asimismo, la tasa de aprendizaje moderada favorece un entrenamiento más estable, mientras que la semilla de aleatoriedad garantiza la reproducibilidad de los resultados.

TABLA XXXIV
CONFIGURACIÓN DE HIPERPARÁMETROS DEL MODELO LIGHTGBM

Parámetro	Valor	Descripción breve
n_estimators	300	Número de árboles del modelo
max_depth	5	Profundidad máxima de cada árbol
learning_rate	0.08	Tasa de aprendizaje
subsample	0.9	Proporción de muestras usadas por árbol
colsample_bytree	0.9	Proporción de variables usadas por árbol
random_state	42	Semilla para reproducibilidad

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

v. Evaluación del modelo

Validación Hold Out forecasting

Las métricas obtenidas como se evidencia en la TABLA XXXV indican un buen desempeño del modelo

LightGBM en la predicción de la serie temporal. El SMAPE de 13,37% refleja un error porcentual relativamente bajo, lo que sugiere una alta precisión en términos relativos. Por su parte, el RMSE y el MAE evidencian el nivel de desviación en unidades monetarias, mostrando que el modelo mantiene errores controlados frente a los valores reales. En conjunto, estos resultados indican una capacidad predictiva consistente y adecuada para el objetivo del modelo.

TABLA XXXV
MÉTRICAS DE DESEMPEÑO DEL MODELO LIGHTGBM

Métrica	Valor	Descripción
SMAPE	13,37 %	Error porcentual medio simétrico
RMSE	$5.669.581.076	Raíz del error cuadrático medio
MAE	$888.132.519	Error absoluto medio

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

vi. Pronóstico con el modelo LightGBM

La figura 3.4 muestra que el modelo LightGBM logra capturar adecuadamente la tendencia creciente de la serie, manteniendo un comportamiento consistente en el periodo de prueba; sin embargo, presenta dificultades para representar los picos más altos, donde tiende a subestimar los valores reales, evidenciando una suavización de la serie; a pesar de esto, el modelo alcanza un SMAPE de 13.37%, lo que indica un desempeño bueno, con capacidad predictiva sólida aunque con oportunidades de mejora en la captura de la volatilidad.

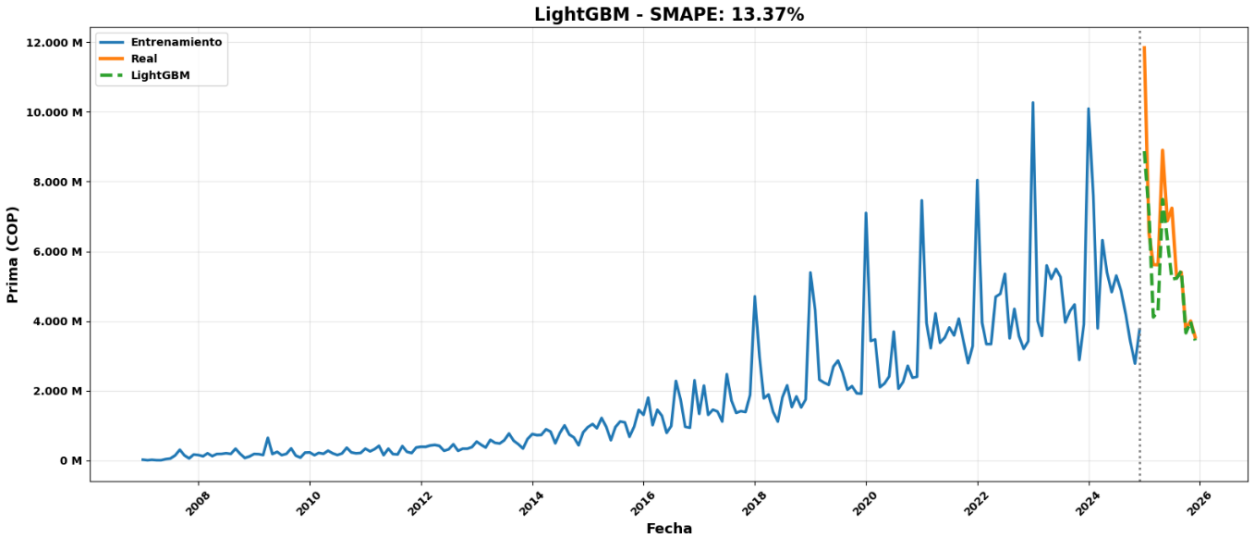


Fig.21. Pronóstico de la serie temporal del importe de prima, incluyendo datos de entrenamiento, valores reales y predicción del modelo LightGBM. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

2) Validación Walk-Forward de los modelos predictivos SARIMAX, XGBoost y LigthGBM

Los resultados presentados en la Tabla XXXVI de validación *Walk-Forward* muestran que el modelo *SARIMAX* presentó el mejor desempeño predictivo, al registrar los valores más bajos en todas las métricas evaluadas: un *SMAPE* de 15,50%, un *RMSE* de 1.143.983.646 y un MAE de 889.153.776. Considerando los criterios de interpretación del SMAPE, este resultado puede clasificarse como un desempeño bueno, indicando que los pronósticos obtenidos fueron cercanos a los valores reales y con un nivel de error moderado. Por su parte, LightGBM obtuvo un rendimiento intermedio, con un SMAPE de 17,82%, también considerado dentro de un rango bueno, aunque con errores superiores a los de SARIMAX. En contraste, *XGBoost* presentó el mayor nivel de error, con un SMAPE de 20,92%, clasificándose como un desempeño aceptable. En general, los resultados sugieren que la estructura temporal y estacional de la serie

fue capturada de manera más eficiente por *SARIMAX*.

TABLA XXXVI
MÉTRICAS DE DESEMPEÑO EN VALIDACIÓN WALK-FORWARD

	Modelo	SMAPE (%)	RMSE	MAE
0	SARIMAX	15,50%	1.143.983.646	889.153.776
1	XGBoost	20,92%	1.588.940.070	1.325.797.927
2	LightGBM	17,82%	1.309.555.727	1.078.989.888

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 22 presenta la comparación de los resultados obtenidos mediante la validación walk-forward entre los modelos *SARIMAX*, *XGBoost* y *LightGBM* frente a los valores reales de los siniestros incurridos durante el año 2025. En términos generales, se observa que los tres modelos logran capturar la tendencia temporal de la serie, aunque con diferencias en la precisión de los pronósticos a lo largo de los meses. *XGBoost* y *LightGBM* presentan mayores desviaciones en algunos periodos, especialmente en los meses de mayo y octubre, mientras que *SARIMAX* muestra un comportamiento más estable y cercano a la dinámica observada en varios puntos de la serie. Asimismo, los modelos de aprendizaje automático evidencian una mejor capacidad para aproximarse a ciertos cambios bruscos de la serie, aunque también presentan sobreestimaciones y subestimaciones más marcadas. En conjunto, la gráfica permite identificar las diferencias en el ajuste predictivo de cada enfoque y evaluar visualmente su capacidad para representar el comportamiento temporal del valor de los siniestros incurridos.

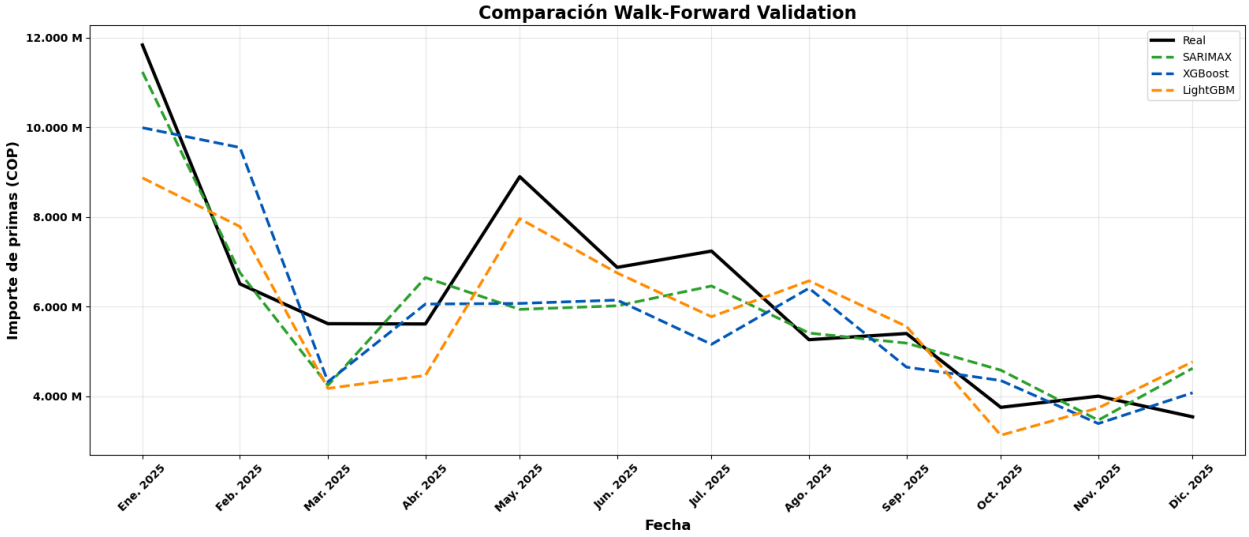


Fig.22. Desempeño de Modelos en Validación Walk-Forward. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

3) Comparación final: SARIMAX vs XGBoost vs LightGBM para el importe de primas

La Figura 23 muestra la comparación de los pronósticos obtenidos mediante los modelos SARIMAX, XGBoost y LightGBM frente a los valores reales de las primas, así como los errores absolutos registrados en cada mes de la validación. En términos generales, se observa que los errores varían considerablemente entre los modelos y entre los distintos periodos evaluados, lo que evidencia la complejidad y volatilidad de la serie temporal. XGBoost presentó errores elevados en algunos meses específicos, especialmente en febrero, mayo y julio, indicando dificultades para

ajustarse a cambios bruscos de la serie. Por su parte, LightGBM mostró errores altos al inicio del periodo y en ciertos meses intermedios, aunque en otros logró aproximarse de forma más precisa a los valores reales. SARIMAX, aunque presentó un comportamiento más estable en varios meses, también registró desviaciones importantes, particularmente alrededor de mayo y diciembre. En conjunto, los resultados reflejan que ningún modelo mantiene el menor error en todos los periodos, pero sí existen diferencias claras en la capacidad de adaptación de cada enfoque frente a las fluctuaciones de la serie temporal.

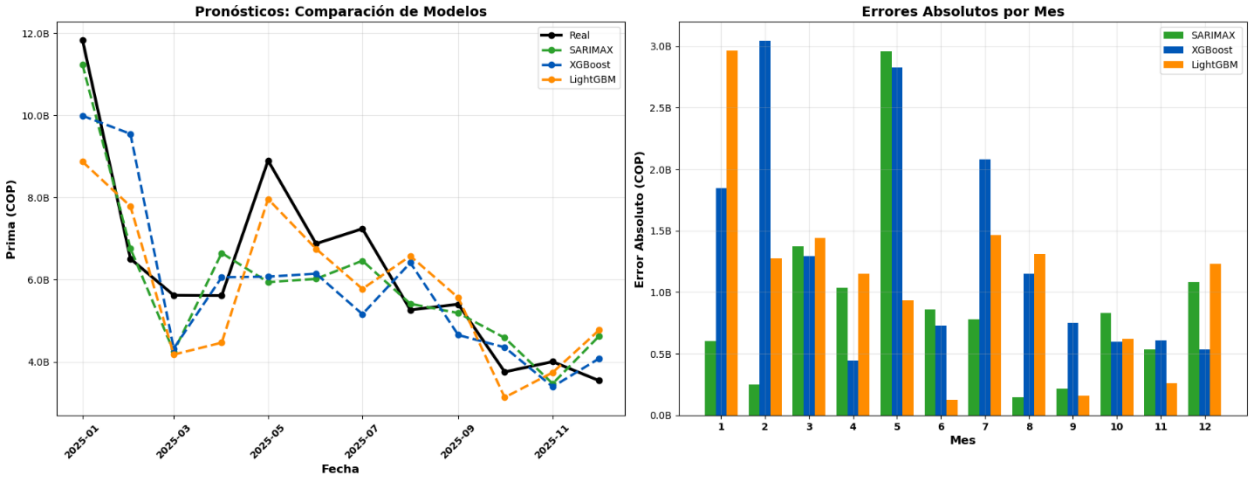


Fig.23. Comparación de pronósticos y errores absolutos de los modelos SARIMAX, XGBoost y LightGBM en la predicción mensual del Importe prima. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

4) Nowcast: pronóstico del mes en curso

El resultado de la Tabla XXXVII del nowcast para enero de 2026 muestra una diferencia importante entre los modelos. El modelo SARIMAX estima un valor de $13.156.057.927, acompañado de un intervalo de confianza amplio (entre $8.137.458.735 y $21.269.768.095), lo que indica alta incertidumbre, pero también capacidad para capturar posibles picos en la serie. En contraste, los modelos de machine learning predicen valores considerablemente más bajos: XGBoost con $4.503.904.256 y LightGBM con $4.845.233.482, ambos muy cercanos entre sí, lo que sugiere estabilidad, pero posible subestimación. Cabe resaltar que en estos modelos aparece "N/A" en los intervalos de confianza, ya que algoritmos como XGBoost y LightGBM no generan de forma directa intervalos probabilísticos, a diferencia de SARIMAX, que es un modelo estadístico y sí permite estimar la incertidumbre del pronóstico. En conjunto, esto evidencia que SARIMAX está captando una dinámica más volátil o estacional, mientras que los modelos basados en árboles tienden a ser más conservadores en el pronóstico.

TABLA XXXVII
COMPARACIÓN DE PRONÓSTICOS PARA ENERO DE 2026

Modelo	Pronóstico	IC 95% Inferior	IC 95% Superior
SARIMAX	$13.160.057.927	$8.137.458.735	$21.269.768.095
XGBoost	$4.503.904.256	N/A	N/A
LightGBM	$4.845.233.482	N/A	N/A

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Tabla XXXVIII evidencia que existen diferencias significativas entre los modelos, ya que tanto XGBoost como LightGBM presentan desviaciones superiores al 60% respecto a SARIMAX. Esto confirma que los modelos de machine learning están subestimando el valor proyectado en comparación con el enfoque estadístico.

TABLA XXXVIII
COMPARACIÓN DE DIFERENCIAS ABSOLUTAS Y PORCENTUALES DEL NOWCAST ENTRE MODELOS

Comparación	Diferencia (COP)	Diferencia (%)
SARIMAX vs XGBoost	$8.652.153.671	65.77%
SARIMAX vs LightGBM	$8.310.824.446	63.17%

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

5) Cierre anual proyectado

Los resultados presentados en la Tabla XXXIX de la proyección anual muestran diferencias importantes entre los modelos utilizados para estimar el comportamiento futuro de la serie temporal. El modelo SARIMAX presentó la estimación más alta, con una proyección anual de $81.822.291.503 COP y un promedio mensual cercano a $6.818 millones, además de registrar el mayor valor máximo mensual proyectado ($13.156 millones), lo que sugiere una mayor sensibilidad frente a posibles incrementos extremos o picos estacionales. Por su parte, XGBoost generó las estimaciones más conservadoras, tanto en el acumulado anual como en el promedio mensual, indicando un comportamiento más estable y moderado de la serie. En contraste, LightGBM presentó resultados intermedios, cercanos a los obtenidos por SARIMAX, aunque con valores máximos menos elevados. En conjunto, los resultados evidencian que los modelos estadísticos y de aprendizaje automático capturan dinámicas distintas de la serie, especialmente en la magnitud de los valores extremos proyectados.

TABLA XXXIX
PROYECCIÓN DEL CIERRE ANUAL 2026 POR MODELO PREDICTIVO

Concepto	SARIMAX	XGBoost	LightGBM
Proyección anual	$81.822.291.503	$65.865.326.592	$79.818.753.596
Promedio mensual proyectado	$6.818.524.292	$5.488.777.216	$6.651.562.800
Máximo mensual proyectado	$13.156.057.927	$7.094.520.832	$9.229.123.792

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

C. Desarrollo de los modelos predictivos para la estimación de siniestros incurridos del Ramo de Responsabilidad Civil profesional

1) Ajustes de modelos predictivos

a) Modelo ARIMA

i. Test de estacionariedad (Dickey-Fuller)

Se evalúa la estacionariedad de la serie temporal mensual del valor del siniestro incurrido como se realizó en el importe de las primas, como requisito previo para el ajuste de modelos de series de tiempo como *ARIMA*.

H_0**:** La serie tiene raíz unitaria (NO es estacionaria)
H_1: La serie NO tiene raíz unitaria (ES estacionaria)

Por otro lado, tenemos los criterios de decisión:

Si $p - value < 0.05 \rightarrow$ Rechazamos H₀ → Serie ES estacionaria
Si $p - value \geq 0.05 \rightarrow$ No rechazamos H₀ → Serie NO es estacionaria

Los resultados de la prueba Dickey-Fuller Aumentada (ADF) evidencian en la Tabla XL que la serie temporal presenta estacionariedad en nivel, dado que el p-value obtenido (0.0192) es inferior al nivel de significancia del 5%, permitiendo rechazar la hipótesis nula de presencia de raíz unitaria. Asimismo, el estadístico de prueba ADF (-3.2142) es menor que los valores críticos al 5% (-2.8848) y 10% (-2.5792), lo que respalda la estabilidad estadística de la serie; sin embargo, no supera el umbral más estricto del 1% (-3.4833). La prueba empleó 3 rezagos, seleccionados automáticamente para controlar la autocorrelación, utilizando 126 observaciones en el análisis. En conjunto, estos resultados indican que la serie no requiere diferenciación adicional y puede utilizarse directamente en la estimación de modelos ARIMA con d = 0.

TABLA XL
RESULTADOS DEL TEST DE DICKEY-FULLER AUMENTADO (ADF) PARA EL VALOR DEL SINIESTRO INCURRIDO

Estadístico	Valor	Interpretación
ADF Statistic	-3.2142	Estadístico de prueba
p-value	0.0192	☑ Estacionaria
Lags utilizados	3	Número de rezagos usados
Observaciones	126	Cantidad de observaciones
Valor crítico (1%)	-3.4833	Umbral al 99% de confianza
Valor crítico (5%)	-2.8848	Umbral al 95% de confianza
Valor crítico (10%)	-2.5792	Umbral al 90% de confianza

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 24 permite comparar el comportamiento de la serie del valor del siniestro incurrido en su escala original y bajo transformación logarítmica. En la serie original se observa una alta volatilidad, con picos pronunciados en distintos periodos (especialmente alrededor de 2016 y 2019), lo que evidencia una varianza inestable y la presencia de eventos extremos propios del comportamiento de siniestros. En contraste, la transformación logarítmica comprime la escala de los datos y reduce de forma significativa la dispersión, suavizando los picos y haciendo más homogénea la variabilidad a lo largo del tiempo, aunque aún se mantienen fluctuaciones que reflejan la naturaleza aleatoria del fenómeno. En conjunto, la comparación muestra que la transformación mejora la estabilidad de la serie sin eliminar completamente su comportamiento irregular, lo cual es esperable en datos de siniestralidad.

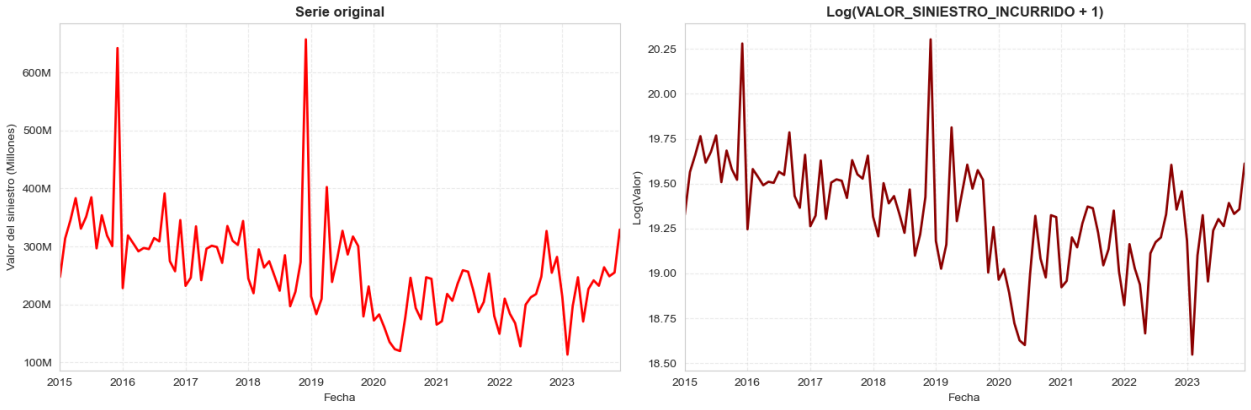


Fig.24. Transformaciones de la serie temporal del valor del siniestro incurrido para el análisis de estacionariedad. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

La Figura 25 muestra los correlogramas ACF y PACF de la serie original del valor del siniestro incurrido, utilizados para identificar la estructura autorregresiva y de medias móviles del modelo. En el correlograma ACF se observan autocorrelaciones positivas significativas en los primeros rezagos y una disminución gradual sin un corte abrupto, lo que sugiere una persistencia temporal moderada y la posible presencia de componentes autorregresivos. Por su parte, el PACF presenta un pico significativo en el rezago 1 y una disminución en los rezagos posteriores, indicando que la dependencia directa se concentra principalmente en el primer periodo. En conjunto, estos resultados

evidencian una estructura temporal de corto plazo compatible con modelos ARIMA de bajo orden y permiten proponer inicialmente un modelo ARIMA(1,0,1), donde el parámetro autorregresivo (p = 1) captura la dependencia temporal observada en el PACF, el componente de medias móviles (q = 1) incorpora el efecto de los errores pasados evidenciado en la ACF y el parámetro de diferenciación (d = 0) se justifica debido a que la serie fue identificada previamente como estacionaria en nivel mediante la prueba Dickey-Fuller Aumentada.

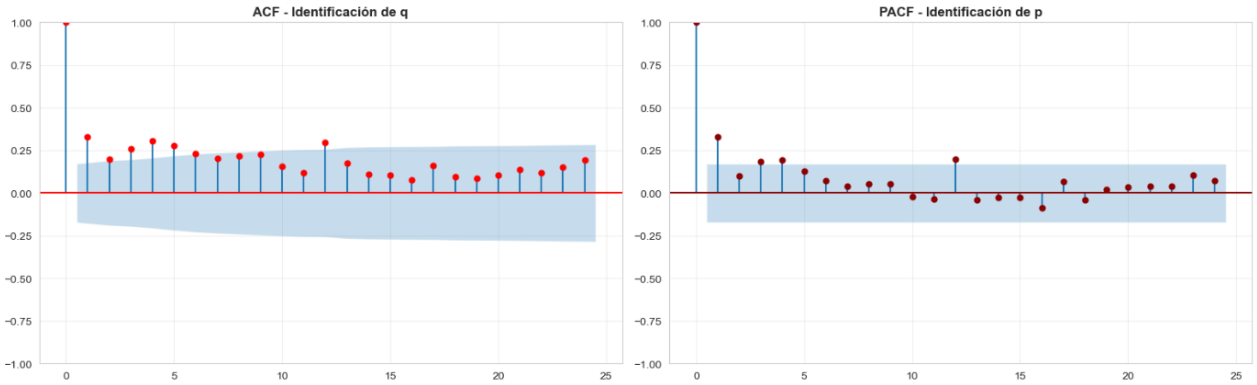


Fig.25. Autocorrelación (ACF) y PACF. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

ii. Entrenamiento del modelo ARIMA y estimación de parámetros mediante Python

El conjunto de datos del valor de los siniestros incurridos, compuesto por 2.395 registros transaccionales agregados mensualmente, fue dividido en 108 observaciones para entrenamiento (enero 2015 – diciembre 2023) y 12 observaciones para prueba (año 2024), equivalente al 90% y 10% respectivamente. Se tomó un año como periodo de prueba respetando el orden cronológico de la serie para evitar fuga de información (*data leakage*), y dado que el dataset cuenta con únicamente 10 años de historia, destinar más años a la prueba habría reducido considerablemente el entrenamiento, limitando la capacidad del modelo para aprender los patrones históricos de la serie.

La Tabla XLI de parámetros del modelo ARIMA(1,0,1) evidencia que tanto el componente autorregresivo como el de media móvil son altamente significativos estadísticamente, dado que ambos presentan p-values de 0.0000, lo que indica una fuerte evidencia para rechazar la hipótesis nula de no significancia. El coeficiente AR(1) positivo (0.9315) sugiere una alta persistencia en la serie, es decir, el valor actual del siniestro incurrido depende fuertemente del comportamiento del periodo inmediatamente anterior. Por otro lado, el coeficiente MA(1) negativo (-0.6315) refleja la corrección de choques aleatorios pasados, indicando que las perturbaciones recientes tienen un efecto inverso en la dinámica actual de la serie. En conjunto, estos resultados muestran que el modelo captura adecuadamente tanto la dependencia temporal como la estructura de errores, lo que respalda su capacidad explicativa sobre el comportamiento de los siniestros incurridos.

TABLA XLI
PARÁMETROS DEL MODELO ARIMA (1,0,1)

	Parámetro	Coeficiente	p-value	Significancia	Interpretación
0	AR(1) - φ_1	0.9133	0.0000	☑ Significativo	Captura la dependencia del valor actual con respecto al periodo anterior.
1	MA(1) - θ_1	-0.6315	0.0000	☑ Significativo	Modela el efecto de choques aleatorios del periodo anterior sobre el valor actual.

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. Diagnóstico de residuos ARIMA

El diagnóstico de residuos del modelo ARIMA(1,0,1) indica un desempeño adecuado en términos de ruido

blanco, ya que la media de los residuos es prácticamente cero (-0.0039), lo que sugiere ausencia de sesgo sistemático en las predicciones. La desviación estándar (0.2445) es baja, evidenciando una dispersión controlada de los errores. En cuanto a la forma de la distribución, tanto la asimetría (skewness = 0.3337) como la curtosis (2.7523) se encuentran cercanas a los valores teóricos de una distribución normal, lo que indica que los residuos presentan una distribución aproximadamente simétrica y sin colas excesivamente pesadas. En conjunto, estos resultados sugieren que los residuos cumplen razonablemente con los supuestos de independencia, homocedasticidad y normalidad aproximada, por lo que el modelo captura adecuadamente la estructura de la serie temporal..

TABLA XLII
MÉTRICAS DE RESIDUOS

	Métrica	Valor	Ideal	Evaluación
0	Media	-0.039	≈ 0	☑
1	Desviación Estándar	0.2445	Baja	☑
2	Skewness	0.3337	≈ 0	☑
3	Kurtosis	2.7523	≈ 0	☑

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

El resultado del test de Ljung-Box indica en la Tabla XLIII que los residuos del modelo ARIMA no presentan evidencia de autocorrelación significativa en ninguno de los rezagos evaluados (1 a 27), ya que todos los valores p son consistentemente mayores a 0.49 y en su mayoría superiores a 0.60, lo que lleva a no rechazar la hipótesis nula de independencia serial. Esto sugiere que los residuos se comportan como ruido blanco, es decir, no contienen estructura temporal remanente que el modelo no haya capturado. Además, la estabilidad de los p-valores a lo largo de los rezagos refuerza la conclusión de que no existen patrones sistemáticos ni dependencia en los errores, lo que respalda la adecuada especificación del modelo ARIMA(1,0,1) para la serie analizada.

Es de tener en cuenta que los 27 rezagos seleccionados para el test de Ljung-Box se determinan a partir de la regla empírica $\frac{n}{4}$, donde $n = 108$ que corresponde al número de observaciones de la serie, lo que permite definir un número adecuado de rezagos para evaluar la autocorrelación de los residuos de manera robusta.

H_0: los residuos NO tienen autocorrelación (ruido blanco)
H_1: los residuos SÍ tienen autocorrelación

TABLA XLIII
TEST DE LJUNG-BOX PARA EL MODELO ARIMA(1,0,1)

	lb_stat	lb_pvalue	Evaluación
1	0.2647	0.6069	☑ Ruido blanco
2	1.4177	0.4922	☑ Ruido blanco
3	2.0894	0.5541	☑ Ruido blanco
4	2.7365	0.6028	☑ Ruido blanco
5	3.2232	0.6656	☑ Ruido blanco
25	20.8573	0.7005	☑ Ruido blanco
26	22.1131	0.6825	☑ Ruido blanco
27	22.6159	0.7055	☑ Ruido blanco

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 26 de diagnóstico de residuos del modelo ARIMA(1,0,1) permiten evaluar visualmente si se cumplen los supuestos de ruido blanco, evidenciando resultados mixtos que deben interpretarse con cuidado.

La Figura 26 muestra que los residuos del modelo se distribuyen alrededor de cero y no presentan patrones sistemáticos en el tiempo, lo que sugiere un comportamiento aleatorio. El histograma evidencia una concentración de los errores cerca de la media, mientras que el gráfico Q-Q indica que la normalidad se cumple de manera aproximada en la parte central de la distribución, aunque existen desviaciones en las colas asociadas a algunos valores atípicos. Por

su parte, el correlograma ACF revela que las autocorrelaciones residuales se encuentran dentro de los límites de confianza, indicando ausencia de autocorrelación significativa. En conjunto, estos resultados permiten concluir que los residuos se comportan como ruido blanco y que el modelo captura adecuadamente la estructura temporal de la serie, aunque persisten ligeras desviaciones respecto al supuesto de normalidad.

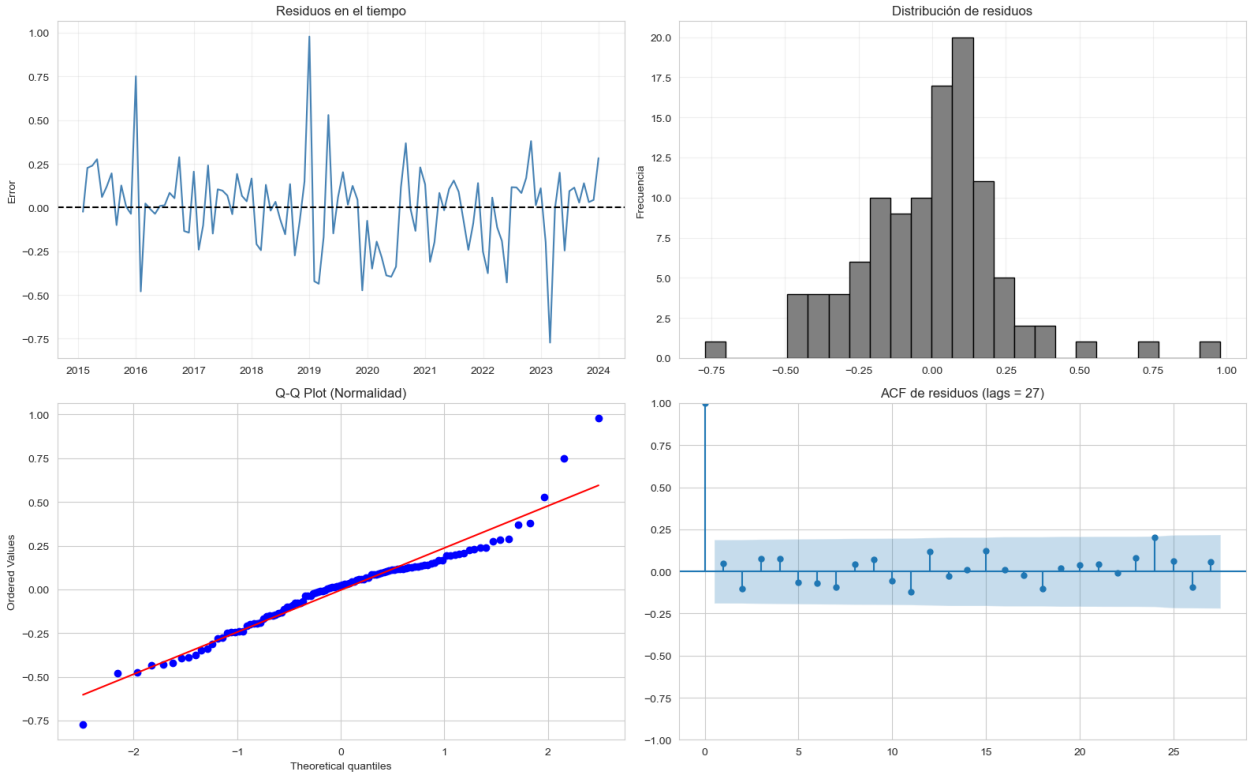


Fig.26. Diagnóstico de residuos del modelo ARIMA(1,0,1). Elaboración propia en Python a partir de la base de datos del valor de siniestros incurridos de la compañía de Seguros del Estado S.A.

iv. Generación de pronósticos con el modelo

La Tabla XLIV muestra que el modelo presenta una tendencia a sobreestimar los valores observados durante los primeros meses de 2024, registrando errores porcentuales elevados en enero (81,17 %), febrero (95,67 %) y marzo (95,84 %). A partir de abril, los errores disminuyen y alternan entre sobreestimaciones y subestimaciones, aunque se observan diferencias importantes en algunos meses, especialmente septiembre (-34,80 %) y diciembre (-31,54 %). A pesar de estas variaciones, todos los valores reales se encuentran dentro de los intervalos de confianza del 95 %, lo que indica que el modelo logra capturar adecuadamente la incertidumbre asociada a los pronósticos. En general, el modelo reproduce la tendencia global de la serie, aunque presenta limitaciones para reflejar con precisión los cambios bruscos observados en determinados periodos.

TABLA XLIV
COMPARACIÓN ENTRE VALORES REALES Y PRONOSTICADOS DEL MODELO ARIMA (1,0,1) CON INTERVALO DE CONFIANZA (95%)

Fecha	Valor real	Pronóstico	IC 95% Inferior	IC 95% Superior	Error %
2024-01	$160.060.800	$289.989.697	$173.853.299	$483.706.814	81,17%
2024-02	$146.559.000	$286.775.907	$169.273.689	$485.842.907	95,67%
2024-03	$144.950.100	$283.867.629	$165.448.263	$487.045.494	95,84%
2024-04	$315.514.200	$281.233.422	$162.221.714	$487.556.422	-10,87%
2024-05	$393.590.400	$278.845.472	$159.478.339	$487.557.108	-29,15%
2024-06	$257.792.800	$276.679.105	$157.129.864	$487.185.096	7,33%
2024-07	$168.420.900	$274.712.392	$155.107.633	$486.545.359	63,11%
2024-08	$263.847.100	$272.925.793	$153.357.388	$485.718.296	3,44%
2024-09	$416.123.800	$271.301.868	$151.835.676	$484.765.544	-34,80%
2024-10	$281.251.500	$269.825.016	$150.507.295	$483.734.291	-4,06%
2024-11	$343.915.500	$268.481.263	$149.343.453	$482.660.523	-21,93%
2024-12	$390.384.300	$267.258.068	$148.320.394	$481.571.506	-31,54%

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

v. Evaluación del modelo

Validación Hold Out forecasting

Las métricas obtenidas en la Tabla XLV evidencian que el modelo ARIMA(1,0,1) presenta un bajo desempeño predictivo. En primer lugar, el SMAPE de 33,30% indica un error porcentual elevado, superando ampliamente el umbral del 30%, lo que clasifica el modelo como deficiente en términos de precisión. Este resultado confirma que las predicciones se desvían considerablemente de los valores reales. Por su parte, el RMSE de $101.601.056 refleja una alta magnitud del error, lo que implica que el modelo presenta desviaciones significativas frente a los valores observados, especialmente en presencia de picos o valores extremos. De manera complementaria, el MAE de $87.262.868 indica que, en promedio, las predicciones se alejan cerca de 87 millones de pesos de los valores reales, lo cual es considerablemente alto para fines predictivos. En conjunto, estas métricas confirman que el modelo no logra capturar adecuadamente la dinámica de la serie, particularmente su alta volatilidad y la presencia de valores atípicos, lo que conduce a estimaciones poco precisas. Este comportamiento es consistente con el análisis previo del pronóstico, donde se observó una tendencia del modelo a generar valores cercanos a un promedio, sin adaptarse a fluctuaciones abruptas.

TABLA XLV
MÉTRICAS DE DESEMPEÑO DEL MODELO ARIMA (1,0,1)

Métrica	Valor	Interpretación
SMAPE	33,30%	Error porcentual promedio alto
RMSE	$101.601.056	Alta desviación respecto a valores reales
MAE	$87.262.868	Error promedio elevado en términos absolutos
Evaluación	❌ POBRE	El modelo presenta bajo desempeño predictivo

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 27 presenta la serie histórica del valor de los siniestros incurridos junto con los pronósticos generados por el modelo ARIMA (1,0,1) para el año 2024. Se observa que la serie histórica exhibe una alta variabilidad, con picos

pronunciados en algunos periodos y fluctuaciones importantes a lo largo del tiempo. El modelo proyecta una trayectoria relativamente estable y ligeramente decreciente durante 2024, con valores cercanos a los 270–290 millones de pesos. Sin embargo, los valores reales presentan oscilaciones significativas, alcanzando máximos superiores a 400 millones y mínimos cercanos a 145 millones, lo que evidencia que el modelo no logra capturar adecuadamente los cambios bruscos observados en la serie. Aunque los valores reales permanecen dentro de los intervalos de confianza del 95 %, la diferencia entre la serie observada y los pronósticos sugiere una capacidad limitada para representar la volatilidad.

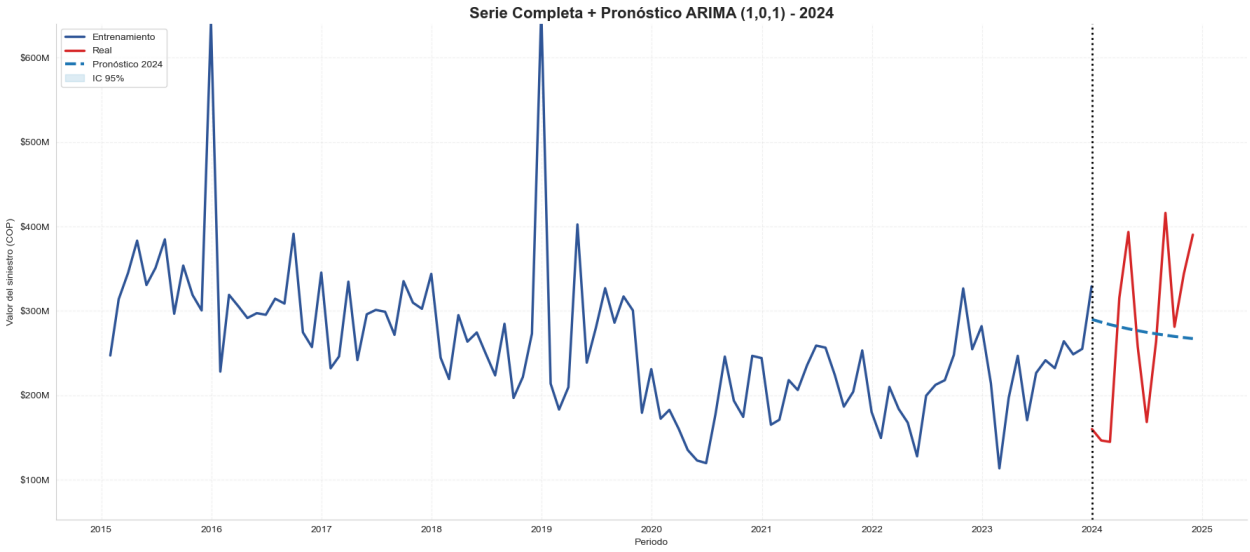


Fig.27. Comparación entre valores reales y pronosticados mediante ARIMA (1,0,1). Elaboración propia en Python a partir de la base de datos del valor de siniestros incurridos de la compañía de Seguros del Estado S.A.

b) Modelo XGBoost

Se utiliza la librería XGBoost para la implementación de modelos de aprendizaje automático orientados a mejorar la capacidad predictiva de la serie.

i. Ingeniería de características (Feature engineering)

La Tabla XLVI muestra la serie mensual del valor del siniestro incurrido junto con el conjunto de variables explicativas creadas mediante ingeniería de características. Los *lags (lag_1 a lag_12)* representan los valores de siniestros en meses anteriores, permitiendo al modelo capturar dependencia temporal directa. Las variables de *rolling mean y rolling std* resumen el comportamiento reciente del sistema, mostrando tanto la tendencia (promedios móviles a 3, 6 y 12 meses) como la volatilidad (desviación estándar), lo cual ayuda a identificar estabilidad o cambios bruscos en los siniestros. Adicionalmente, las variables temporales como *month, quarter y year* introducen estacionalidad y estructura calendario, mientras que *trend* incorpora una progresión lineal del tiempo para capturar tendencias de largo plazo. En conjunto, este *dataset* convierte la serie temporal original en un problema supervisado donde el modelo XGBoost puede aprender patrones históricos, estacionales y de variabilidad para realizar predicciones más robustas.

TABLA XLVI
RESUMEN (PRIMERAS OBSERVACIONES DEL DATASET XGBOOST)

Fecha	Valor Siniestro	Lag 1	Lag 2	Lag 3	Rolling Mean 3	Rolling Std 3	Month	Year
2016-01-31	228,202,700	642,102,200	300,647,800	318,841,900	390,317,600	221,040,000	1	2016
2016-02-29	319,069,500	228,202,700	642,102,200	300,647,800	396,458,100	217,531,500	2	2016
2016-03-31	305,512,400	319,069,500	228,202,700	642,102,200	284,261,500	49,019,300	3	2016
2016-04-30	291,637,400	305,512,400	319,069,500	228,202,700	305,406,400	13,716,360	4	2016
2016-05-31	297,384,100	291,637,400	305,512,400	319,069,500	298,178,000	6,971,483	5	2016

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

ii. Selección de variables predictoras

Aunque la configuración de 27 variables presenta el menor error en las métricas de desempeño en la Tabla XLVII, la configuración de 22 variables fue seleccionada debido a su comportamiento más estable y consistente frente a las demás configuraciones evaluadas. En las configuraciones de 22 y 25 variables se observa un desempeño similar, mientras que la configuración de 27 variables presenta una mayor sensibilidad en la reducción del error, lo cual indica que el modelo XGBoost no mantiene un comportamiento estable frente al incremento de características. Este mismo comportamiento ya se había evidenciado en el modelo de primas, donde la inclusión de un mayor número de variables no necesariamente implica una mejora proporcional en el desempeño, reforzando la importancia de priorizar configuraciones más estables y parsimoniosas. Por lo tanto, la selección de 22 variables responde a un criterio de estabilidad del modelo y consistencia metodológica, manteniendo un equilibrio entre simplicidad y capacidad predictiva.

TABLA XLVII
COMPARACIÓN DEL DESEMPEÑO DEL MODELO XGBOOST SEGÚN EL NÚMERO DE VARIABLES PREDICTORAS

Modelo	Variables	SMAPE (%)	RMSE	MAE
XGBoost	22	23,05	80.315.885	61.986.804
XGBoost	25	23,32	78.099.756	61.784.204
XGBoost	27	11,70	35.954.274	31.678.226

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. División Train / Test

El *split* temporal es la división del conjunto de datos en dos partes: entrenamiento y prueba, respetando el orden cronológico de la serie. Como se explica en los modelos de importe de primas, la construcción de variables de rezago de hasta 12 periodos genera una pérdida de las primeras 12 observaciones, por lo que el conjunto de datos efectivo inicia en enero de 2016. El modelo se entrena con datos de enero 2016 a diciembre 2023 (96 observaciones) y se evalúa con datos de 2024 (12 observaciones), manteniendo 22 variables explicativas en ambos conjuntos, equivalente a una partición del 89% para entrenamiento y 11% para prueba.

iv. Entrenamiento del modelo

La Tabla XLVIII presenta la configuración del modelo *XGBoost* utilizado. Este fue definido como un regresor con función de pérdida basada en el error cuadrático medio, lo cual es apropiado para la predicción de variables continuas como el valor del siniestro incurrido. El modelo se entrenó con 300 árboles y una profundidad máxima de 4 niveles, lo que le permite capturar relaciones no lineales en los datos sin incrementar de forma excesiva el riesgo de sobreajuste. El *learning rate* de 0.05 controla el ritmo de aprendizaje, favoreciendo una convergencia más estable y progresiva. Adicionalmente, los parámetros de *subsample* y *colsample_bytree*, ambos fijados en 0.8, incorporan aleatoriedad en las observaciones y variables utilizadas por cada árbol, lo que contribuye a mejorar la capacidad de generalización del modelo. Finalmente, la fijación de la semilla asegura la reproducibilidad de los resultados. En conjunto, esta configuración busca un equilibrio entre precisión predictiva y control del sobreajuste en el contexto de series de tiempo.

TABLA XLVIII
CONFIGURACIÓN DE HIPERPARÁMETROS DEL MODELO XGBOOST

Parámetro	Valor	Descripción
objective	reg:squarederror	Función de pérdida para regresión (minimiza error cuadrático)
n_estimators	300	Número de árboles del modelo
max_depth	4	Profundidad máxima de cada árbol (controla complejidad)
learning_rate	0.05	Tasa de aprendizaje (controla qué tanto aprende por árbol)

subsample	0.8	Porcentaje de muestras usadas por árbol (regularización)
colsample_bytree	0.8	Porcentaje de variables usadas por árbol (reduce sobreajuste)
random_state	42	Semilla para reproducibilidad

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Tabla XLIX de importancia de variables del modelo *XGBoost* evidencia que el comportamiento del *valor del siniestro incurrido* está explicado principalmente por componentes de tendencia suavizada, en particular el *rolling_mean_12*, que es la variable más relevante con una importancia de 0.3083. Esto indica que el modelo depende en mayor medida del comportamiento promedio de largo plazo de la serie. En segundo nivel de relevancia se encuentran los promedios móviles de 6 y 3 meses, lo que refuerza la importancia de las tendencias de mediano y corto plazo en la predicción. Las variables temporales (*month*, *year* y *quarter*) también aportan información relevante, sugiriendo la presencia de estacionalidad en la serie, aunque con menor peso relativo frente a las tendencias. Por otro lado, variables como *lag_12*, *lag_5* y *lag_3* presentan una importancia baja, lo que indica que la estacionalidad directa y las dependencias puntuales tienen una influencia limitada en el modelo. En conjunto, los resultados muestran que el modelo está principalmente guiado por tendencias de largo plazo, mientras que la estacionalidad y la memoria reciente actúan como factores complementarios de menor relevancia.

TABLA XLIX
IMPORTANCIA DE CARACTERÍSTICAS (FEATURE IMPORTANCE)

	Feature	Importance
16	rolling_mean_12	0.308306
14	rolling_mean_6	0.101642
12	rolling_mean_3	0.100870
18	month	0.078161
20	year	0.074682
19	quarter	0.054223
11	lag_12	0.039122
4	lag_5	0.029360
13	rolling_std_3	0.028011
2	lag_3	0.025166

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 28 confirma de forma visual que el modelo está fuertemente dominado por las variables de tendencia suavizada, especialmente el *rolling_mean_12*, que sobresale claramente por encima del resto. También se observa que los promedios móviles de 6 y 3 meses mantienen una relevancia intermedia, mientras que las variables estacionales (*month, quarter, year*) tienen un peso moderado. En contraste, las variables de memoria directa como los *lags* y las medidas de volatilidad presentan una contribución baja, lo que refuerza que la dinámica del modelo está guiada principalmente por tendencias de largo y mediano plazo más que por efectos puntuales o estacionales.

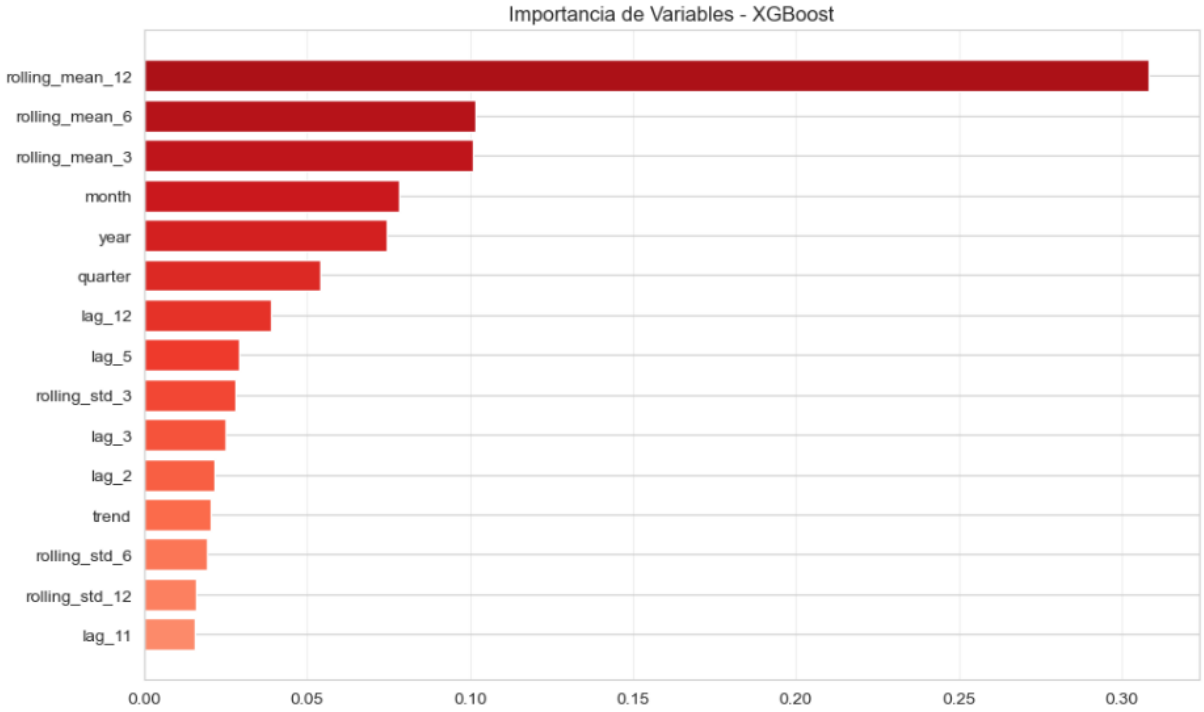


Fig.28. Importancia de las variables en el modelo XGBoost para la predicción del valor del siniestro incurrido. Elaboración propia en Python a partir de la base de datos del valor de siniestros incurridos de la compañía de Seguros del Estado S.A.

v. Evaluación del modelo

Validación Hold Out forecasting

Las métricas obtenidas en la Tabla L del modelo *XGBoost* indican un desempeño razonable en la predicción del valor del siniestro incurrido. El SMAPE de 23.05% sugiere un nivel de error porcentual moderado, lo que indica que las predicciones del modelo se desvían en promedio en aproximadamente un 23% respecto a los valores reales. Por otro lado, el RMSE de 80.3 millones refleja la presencia de errores significativos en algunos puntos, ya que esta métrica penaliza fuertemente las desviaciones grandes. Finalmente, el MAE de 61.9 millones indica que, en promedio, las predicciones del modelo se alejan en esa magnitud de los valores reales, ofreciendo una medida más estable del error general. En conjunto, los resultados muestran que el modelo tiene un desempeño aceptable, aunque aún presenta margen de mejora en la reducción de errores, especialmente en valores extremos de la serie.

TABLA L
MÉTRICAS DE DESEMPEÑO DEL MODELO XGBOOST

Métrica	Valor	Interpretación
SMAPE	23,05%	Error porcentual simétrico del modelo
RMSE	80.315.885	Error cuadrático medio (sensibilidad a errores grandes)
MAE	61.986.804	Error absoluto medio promedio

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

i. Pronóstico con modelo XGBoost

La Figura 29 muestra el desempeño del modelo *XGBoost* en la predicción del valor del siniestro incurrido, comparando el periodo de entrenamiento (azul), los valores reales de 2024 (rojo) y el pronóstico del modelo (línea

punteada). Se observa que durante el entrenamiento la serie presenta alta volatilidad, con picos pronunciados como el de 2019, lo que evidencia un comportamiento irregular difícil de modelar. En el periodo de prueba, el modelo logra capturar la tendencia general de crecimiento y algunos movimientos de la serie; sin embargo, presenta rezagos en la respuesta frente a cambios bruscos, especialmente en los picos más altos, donde tiende a subestimar los valores reales. A pesar de ello, el modelo sigue adecuadamente la dirección global de la serie, lo que es consistente con el SMAPE de 23.05%, indicando un ajuste razonable, pero con limitaciones para capturar variaciones extremas.

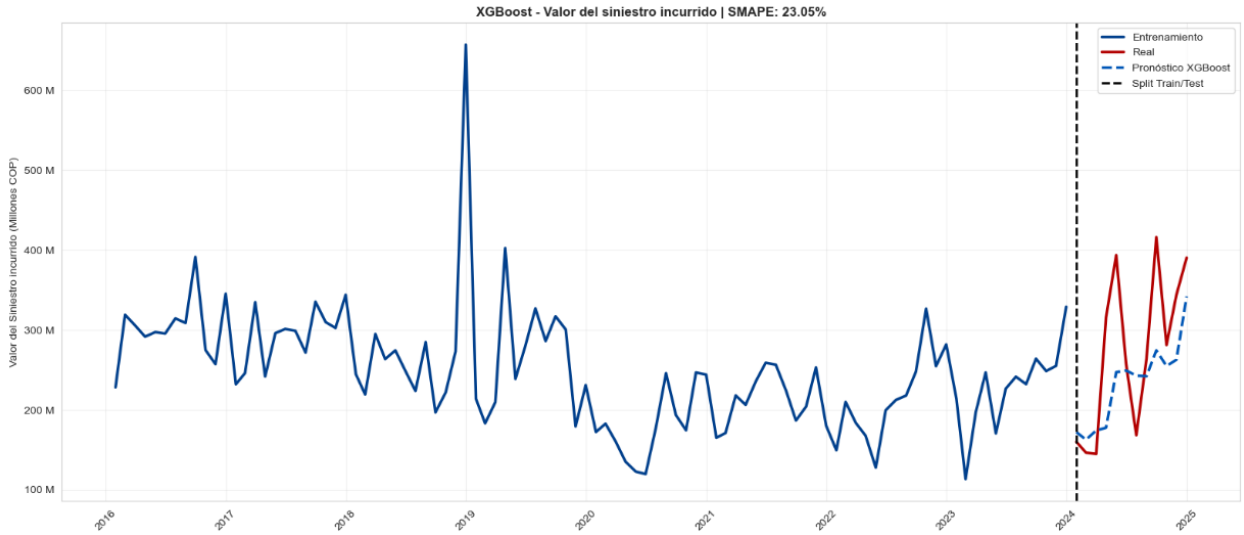


Fig.29. Comparación entre los valores reales y el pronóstico del modelo XGBoost para el valor del siniestro incurrido, incluyendo el periodo de entrenamiento y prueba. Elaboración propia en Python a partir de la base de datos del valor de siniestros incurridos de la compañía de Seguros del Estado S.A.

c) Modelo LightGBM

Previo al desarrollo del modelo, se lleva a cabo la instalación de la librería *LightGBM*, indispensable para implementar este segundo enfoque de aprendizaje automático. A través del comando *pip install lightgbm*, la herramienta se incorpora al entorno de trabajo *jupyter*, permitiendo posteriormente la construcción, entrenamiento y evaluación de modelos basados en árboles de decisión, especialmente adecuados para tareas de regresión como la estimación del valor de los siniestros.

i. Ingeniería de características (Feature engineering) del conjunto de datos

La Tabla LI evidencia la transformación de la serie original mediante la incorporación de múltiples variables derivadas que fortalecen el análisis predictivo. Se parte de la variable objetivo VALOR_SINIESTRO_INCURRIDO, a partir de la cual se generan 12 rezagos (lag_1 a lag_12) que capturan la dependencia temporal de los valores pasados. Adicionalmente, se incluyen diferencias (diff_1 y diff_2) que permiten modelar cambios recientes en la serie. También se incorporan estadísticas de ventanas móviles como rolling_mean_3, rolling_std_3, rolling_max_3, rolling_mean_6, rolling_std_6, rolling_max_6, rolling_mean_12, rolling_std_12 y rolling_max_12, las cuales reflejan el comportamiento promedio, la variabilidad y los valores máximos en distintos horizontes temporales. Finalmente, se añaden variables temporales como month, quarter y year, junto con la variable trend, que representa la tendencia de la serie en el tiempo. En conjunto, estas variables permiten capturar la dinámica temporal, la estacionalidad, la volatilidad y la evolución de los siniestros, mejorando la capacidad del modelo para generar predicciones más precisas. La tabla presentada corresponde a un extracto representativo del conjunto de datos transformado, ya que la versión completa resulta extensa debido al número de variables generadas. Por ello, se incluye únicamente una muestra con fines ilustrativos, permitiendo visualizar la estructura y composición de las variables utilizadas sin sobrecargar la

presentación del documento.

TABLA LI
INGENIERÍA DE CARACTERÍSTICAS (FEATURE ENGINEERING) DEL CONJUNTO DE DATOS

Fecha	Valor del siniestro incurrido	Lag_1	Lag_3	Mean_6	Std_6	Mean_12	Month	qarter	Trend
2016-01-31	228.202.700	642.102.200	318.841.900	356.715.200	145.702.000	354.118.600	1	1	12
2016-02-29	319.069.500	228.202.700	300.647.800	360.437.400	144.140.200	354.533.000	2	1	13
2016-03-31	305.512.400	319.069.500	642.102.200	352.396.100	145.922.000	351.230.500	3	1	14
2016-04-30	291.637.400	305.512.400	228.202.700	347.862.00	147.586.300	343.594.000	4	2	15
2016-05-31	297.384.100	291.637.400	319.069.500	347.318.000	147.800.900	340.811.000	5	2	16

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

ii. Selección de variables predictoras

Para el caso del modelo del valor de los siniestros incurridos se evalúa las mismas tres configuraciones de variables predictoras (22, 25 y 27) mediante el enfoque experimental, con el fin de analizar su impacto en el desempeño del modelo LightGBM. Los resultados muestran que la configuración de 27 variables presenta el menor error de predicción, mientras que las configuraciones de 22 y 25 variables mantienen un comportamiento más estable en las métricas evaluadas. En este sentido, la selección de variables se basa en el equilibrio entre desempeño predictivo y estabilidad del modelo, considerando el comportamiento global de las métricas. Por lo tanto, para este análisis se selecciona la configuración de 27 variables.

TABLA LII
COMPARACIÓN DEL DESEMPEÑO DEL MODELO LIGHTGBM SEGÚN EL NÚMERO DE VARIABLES PREDICTORAS

Modelo	Variables	SMAPE (%)	RMSE	MAE
LightGBM	22	25,51	76.145.397	65.704.817
LightGBM	25	25,33	73.082.332	64.988.725
LightGBM	27	13,21%	39.59.388	35.823.724

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

iii. División Train / Test

La división del conjunto de datos mantiene la misma partición temporal definida en el modelo XGBoost para el valor de los siniestros incurridos, con 96 observaciones para entrenamiento (enero 2016 – diciembre 2023) y 12 observaciones para prueba (año 2024), equivalente al 89% y 11% respectivamente. A diferencia del modelo anterior, LightGBM incorpora 27 variables explicativas en ambos conjuntos, tal como se explica en el modelo LightGBM del importe de las primas.

iv. Entrenamiento del modelo

El modelo *LightGBM* se configura con un conjunto de hiperparámetros orientados a equilibrar capacidad de aprendizaje y control del sobreajuste. Se utiliza 300 estimadores (*n_estimators=300*), lo que permite un aprendizaje progresivo a través de múltiples árboles, mientras que *learning_rate=0.05* regula la contribución de cada uno, favoreciendo un entrenamiento más estable. La complejidad de los árboles se controla principalmente mediante *num_leaves=31*, ya que *max_depth=-1* no impone un límite de profundidad, permitiendo mayor flexibilidad estructural. Adicionalmente, se incorporan mecanismos de regularización como *subsample=0.9* y *colsample_bytree=0.9*, que introducen aleatoriedad al utilizar el 90% de las muestras y variables en cada árbol, reduciendo la varianza del modelo. Finalmente, *random_state*=42 asegura la reproducibilidad de los resultados. En conjunto, esta configuración representa un modelo con capacidad de capturar relaciones no lineales, aunque su desempeño puede ser sensible al tamaño reducido del conjunto de datos, por lo que es recomendable complementar

con técnicas de validación como Walk- Forward que veremos más adelante para evitar sobreajuste.

TABLA LIII
CONFIGURACIÓN DE HIPERPARÁMETROS DEL MODELO LIGHTGBM PARA EL VALOR DE SINIESTROS INCURRIDOS

Parámetro	Valor	Descripción
n_estimators	300	Número total de árboles que construye el modelo. A mayor cantidad, mayor capacidad de aprendizaje, pero también mayor riesgo de sobreajuste y tiempo de entrenamiento.
max_depth	-1	Profundidad máxima de cada árbol. -1 significa sin límite, lo que permite árboles más complejos, aunque puede aumentar el overfitting si no hay regularización adecuada.
num_leaves	31	Número máximo de hojas por árbol. Controla la complejidad del modelo: más hojas permiten aprender patrones más complejos, pero aumentan el riesgo de sobreajuste.
learning_rate	0.05	Tasa de aprendizaje que determina cuánto aporta cada árbol al modelo final. Valores bajos hacen el entrenamiento más lento pero más estable.
subsample	0.9	Porcentaje de filas usadas para construir cada árbol. Introduce aleatoriedad y ayuda a reducir el sobreajuste.
colsample_bytree	0.9	Porcentaje de variables usadas en cada árbol. Reduce la correlación entre árboles y mejora la generalización del modelo.
random_state	42	Semilla aleatoria que garantiza que los resultados del modelo sean reproducibles.

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

El modelo LightGBM, presentado en la Tabla LIV, evidencia que las variables con mayor influencia en la predicción de los siniestros incurridos están asociadas principalmente con la dinámica de los cambios y la volatilidad de la serie temporal. En particular, destacan diff_1, rolling_std_12 y diff_2 como los predictores más relevantes, lo que indica que el modelo otorga mayor peso a las variaciones en el tiempo y al nivel de incertidumbre anual que a los valores absolutos de los siniestros. Por otro lado, variables como lag_3, lag_11 y lag_12 presentan una menor importancia, sugiriendo que la memoria de largo plazo aporta información limitada en comparación con los cambios recientes y la volatilidad. En conjunto, estos resultados reflejan que el comportamiento de la serie es altamente dinámico, donde las fluctuaciones y los cambios estructurales tienen un mayor poder explicativo que los niveles históricos.

TABLA LIV
TOP 10 VARIABLES MÁS IMPORTANTES EN EL MODELO LIGHTGBM

	Variable	Importancia
12	diff_1	124
21	rolling_std_12	120
13	diff_2	91
2	lag_3	69
15	rolling_std_3	66
14	rolling_mean_3	62
18	rolling_std_6	51
10	lag_11	28
17	rolling_mean_6	26
11	lag_12	23

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 30 confirma claramente el resultado anterior: las variables con mayor poder predictivo son *diff_1*, *rolling_std_12* y *diff_2*, que aparecen como las barras más largas del gráfico, indicando que el modelo prioriza los cambios y la volatilidad de la serie por encima de los niveles históricos. Asimismo, variables como los *lags* más lejanos (*lag_11* y *lag_12*) y medias móviles tienen menor importancia relativa, lo que refuerza la conclusión de que el comportamiento del siniestro está dominado por su dinámica reciente más que por su memoria de largo plazo.

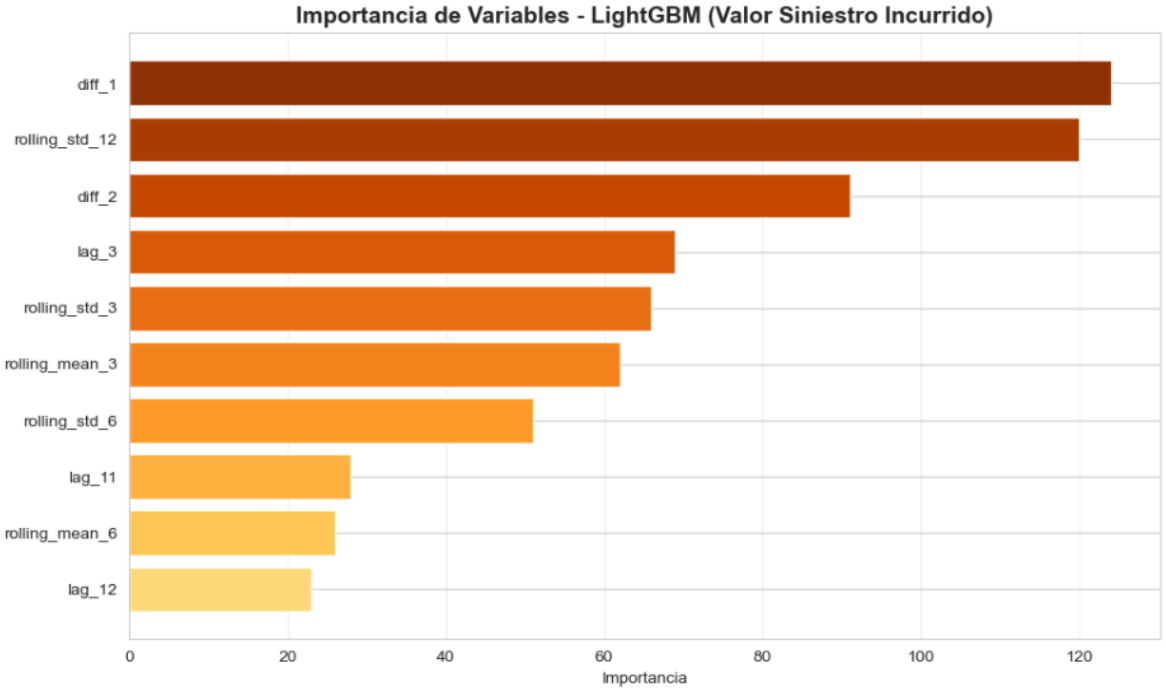


Fig.30. Top 10 de las variables más importantes del modelo LightGBM. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

v. Evaluación del modelo

Validación Hold Out forecasting

Los resultados que se presentan en la Tabla LV evidencian un desempeño adecuado y consistente del modelo en la predicción del valor de los siniestros incurridos. En particular, el SMAPE de 13.21% indica que las predicciones del modelo se desvían, en promedio, alrededor de un 13% respecto a los valores reales, lo cual refleja una buena precisión relativa dentro del contexto del análisis actuarial. En términos absolutos, el MAE (34,937,610) representa la magnitud promedio del error en el valor de los siniestros, mientras que el RMSE (39,559,388), al ser ligeramente superior, sugiere la presencia de algunos siniestros de alta cuantía que generan desviaciones más elevadas, lo cual es coherente con la naturaleza propia de los siniestros incurridos, caracterizados por su alta variabilidad y presencia de valores extremos. En conjunto, los resultados indican que el modelo presenta un comportamiento estable, con una capacidad adecuada para estimar el valor de los siniestros incurridos dentro del portafolio analizado, siendo útil para fines de análisis y toma de decisiones en el contexto del riesgo asegurador.

TABLA LV
MÉTRICAS DE DESEMPEÑO DEL MODELO LIGHTGBM PARA EL VALOR DE SINIESTROS INCURRIDOS

Métrica	Valor	Descripción
SMAPE	13,21%	Error porcentual simétrico que mide qué tan lejos están las predicciones de los valores reales en términos relativos.
RMSE	39.559.388	Raíz del error cuadrático medio. Penaliza fuertemente los errores grandes, por lo que refleja la magnitud de los errores más extremos.
MAE	34.937.610	Error absoluto medio. Indica el error promedio en las mismas unidades de la variable objetivo.

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

vi. Pronóstico con modelo LightGBM

La Figura 31 muestra la evolución histórica de los siniestros incurridos en millones de COP, junto con la comparación entre los valores reales y las predicciones del modelo LightGBM en el periodo de prueba. En la serie de entrenamiento (línea azul), se observa un comportamiento altamente variable de los siniestros a lo largo del tiempo, con fluctuaciones significativas y la presencia de un pico atípico alrededor del año 2019, lo cual refleja la naturaleza volátil del riesgo asegurador y la existencia de eventos de alta severidad. En la etapa de prueba, se evidencia que las predicciones del modelo (línea verde punteada) siguen de manera consistente la tendencia de los valores reales (línea roja), capturando adecuadamente el comportamiento general de la serie. Aunque se presentan algunas desviaciones puntuales, especialmente en los picos locales, el modelo mantiene una trayectoria cercana a los valores observados, lo cual indica una buena capacidad de generalización. La línea de separación entre entrenamiento y prueba muestra claramente la correcta división temporal del conjunto de datos, permitiendo validar el desempeño del modelo en escenarios futuros no vistos. En conjunto, la gráfica sugiere que el modelo LightGBM logra capturar la dinámica general de los siniestros incurridos, incluyendo su tendencia y variabilidad, con un ajuste razonable en la fase de predicción.

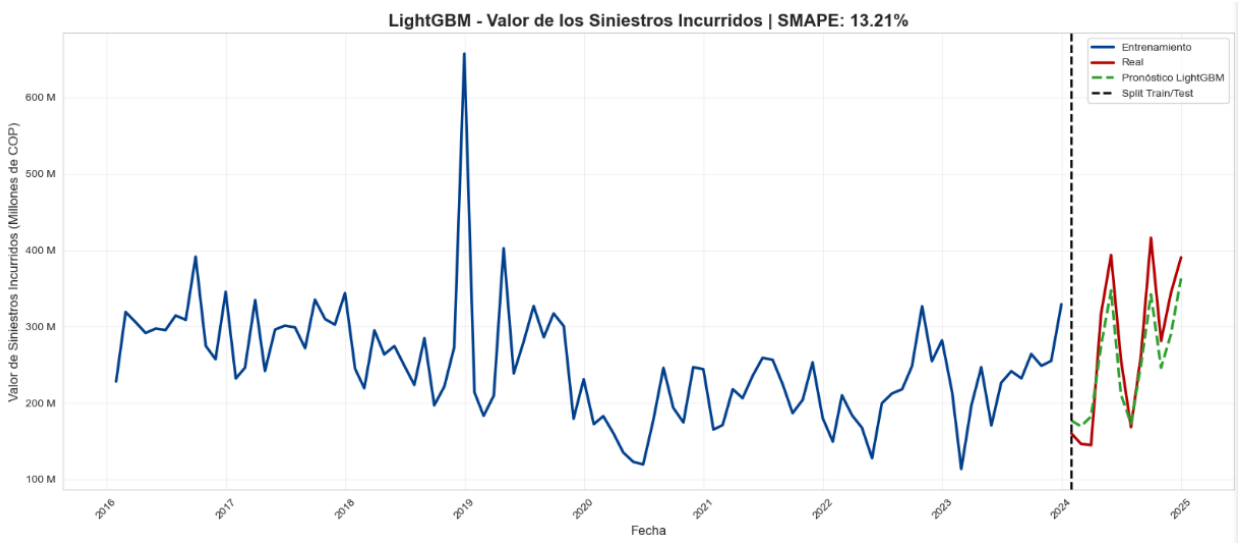


Fig.31. Comparación entre los valores reales y el pronóstico del modelo LightGBM para el valor del siniestro incurrido, incluyendo el periodo de entrenamiento y prueba. Elaboración propia en Python a partir de la base de datos del valor de siniestros incurridos de la compañía de Seguros del Estado S.A

2) Validación Walk-Forward de los modelos predictivo ARIMA, XGBoost y LigthGBM

Los resultados presentados en la Tabla LVI evidencian que el modelo LightGBM obtuvo el mejor desempeño global, al registrar los menores valores de SMAPE (11.07%), RMSE (31.029.292) y MAE (27.332.138), lo que indica una mayor capacidad para capturar la dinámica compleja y no lineal de la serie correspondiente al valor de los siniestros incurridos. Por su parte, el modelo XGBoost presentó un desempeño intermedio, con un SMAPE de 21.39%, evidenciando una capacidad aceptable de predicción, aunque con menor precisión frente a LightGBM. En contraste, el modelo ARIMA registró los mayores errores en todas las métricas evaluadas, con un SMAPE de 33.69%, reflejando limitaciones para representar adecuadamente la alta volatilidad y los cambios abruptos presentes en la serie temporal. En conjunto, los resultados sugieren que el comportamiento de los siniestros incurridos presenta patrones complejos y no lineales, favoreciendo el uso de modelos basados en técnicas de boosting, particularmente LightGBM.

TABLA LVI
MÉTRICAS DE DESEMPEÑO EN VALIDACIÓN WALK-FORWARD

	Modelo	SMAPE (%)	RMSE	MAE
0	ARIMA	33,69%	101.975.505	86.582.502
1	XGBoost	21,39%	70.923.657	57.424.180
2	LightGBM	11,07%	31.029.292	27.332.138

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Figura 32 confirma los resultados obtenidos en las métricas de evaluación, evidenciando que LightGBM es el modelo con mejor desempeño al presentar los menores errores (SMAPE 11,07%, RMSE 31.029.292 y MAE 27.332.138), lo cual se refleja en su capacidad para seguir con mayor precisión los picos y caídas de la serie real. Por su parte, XGBoost muestra un ajuste intermedio, con una respuesta más suavizada y menor sensibilidad a los extremos (SMAPE 21,39%, RMSE 70.923.657 y MAE 57.424.180). En contraste, ARIMA presenta el peor desempeño tanto en términos métricos (SMAPE 33,69%, RMSE 101.975.505 y MAE 86.582.502) como visuales, evidenciando una curva excesivamente suavizada que no logra capturar adecuadamente la volatilidad ni los cambios abruptos de la serie. En conjunto, estos resultados refuerzan que los modelos basados en árboles, especialmente LightGBM, son más adecuados para representar la dinámica irregular y no lineal de los siniestros incurridos.

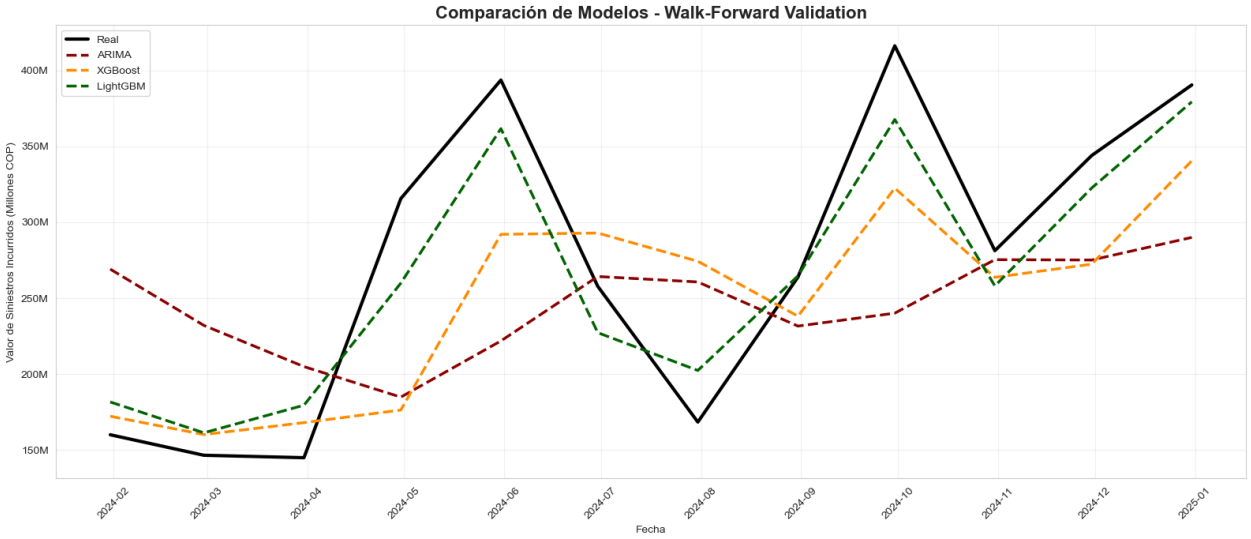


Fig.32. Comparación del desempeño de los modelos ARIMA, XGBoost, y LightGBM en Validación Walk-Forward para la predicción del valor de los siniestros incurridos. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

3) Comparación final: ARIMA vs XGBoost vs LightGBM para el valor de los siniestros incurridos

La Figura 33 de error absoluto promedio por mes confirma de manera cuantitativa las diferencias entre modelos, evidenciando que LightGBM presenta los errores más bajos y estables en la mayoría de los periodos, con valores que oscilan aproximadamente entre 10M y 55M COP, lo que refleja una alta precisión y consistencia incluso en meses de mayor volatilidad. En comparación, XGBoost muestra un comportamiento intermedio, con errores más variables que alcanzan picos cercanos a los 140M COP (abril) y 105M COP (julio), indicando una menor capacidad para ajustarse a cambios abruptos. Por su parte, ARIMA registra los errores más elevados y dispersos, superando los 150M COP en meses como mayo y alcanzando aproximadamente 175M COP en septiembre, además de mantener niveles altos en la mayoría de los periodos. Esta diferencia no solo confirma que LightGBM es el modelo más preciso en términos promedio, sino que también es el más robusto frente a fluctuaciones, mientras que ARIMA evidencia una alta inestabilidad y menor capacidad de adaptación a la dinámica de la serie.

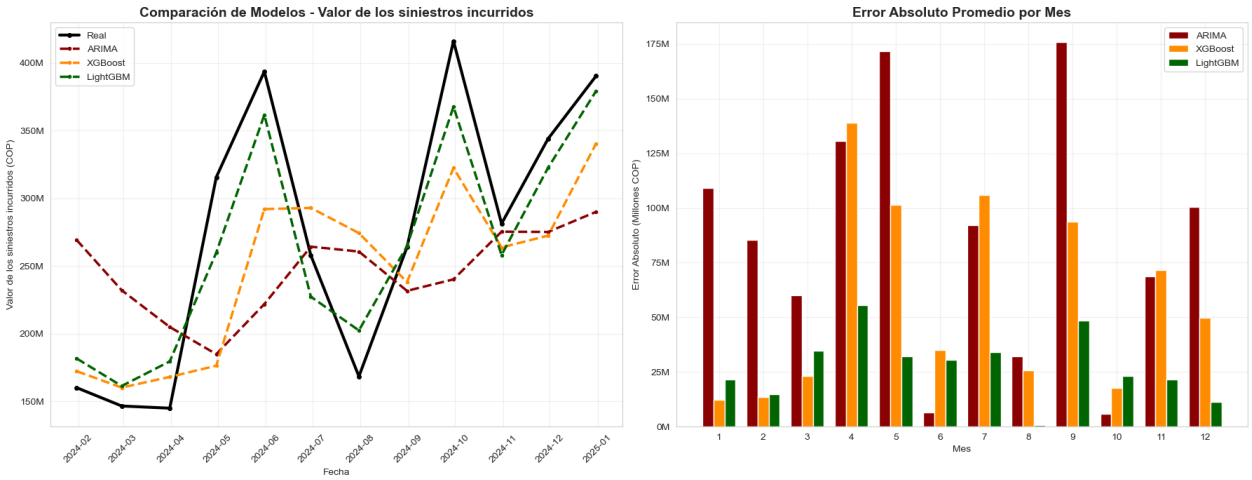


Fig.33. Comparación de pronósticos y errores absolutos de los modelos ARIMA, XGBoost y LightGBM en la predicción mensual del valor del siniestro incurrido. Elaboración propia en Python a partir de la base de datos de Seguros del Estado S.A.

4) Nowcast: pronóstico del mes en curso

Los resultados del nowcast presentados en la Tabla LIII para enero de 2025 evidencian diferencias importantes entre los modelos utilizados. El modelo ARIMA proyectó un valor de siniestros incurridos de $311.820.111 COP, acompañado de un intervalo de confianza al 95% entre $186.840.175 y $520.400.827, lo que refleja la incertidumbre inherente al modelo estadístico. Por su parte, XGBoost estimó un valor superior de $378.490.624 COP, mientras que LightGBM presentó el pronóstico más alto con $388.498.842 COP. La cercanía entre los resultados obtenidos por XGBoost y LightGBM sugiere una tendencia consistente identificada por los modelos de aprendizaje automático, mientras que ARIMA mostró una estimación más conservadora. Adicionalmente, considerando que LightGBM obtuvo el mejor desempeño durante la validación walk-forward, su pronóstico puede interpretarse como el más robusto para representar el comportamiento esperado de los siniestros incurridos en el periodo proyectado.

TABLA LVII
COMPARACIÓN DE PRONÓSTICOS PARA ENERO DE 2025

Modelo	Pronóstico	IC 95% Inferior	IC 95% Superior
ARIMA	$311.820.111	$186.840.175	$520.400.827
XGBoost	$378,490,624	N/A	N/A
LightGBM	$388,498,842	N/A	N/A

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

Los resultados comparativos de la Tabla LVIII evidencian diferencias considerables entre el modelo ARIMA y los modelos de aprendizaje automático utilizados en el pronóstico de los siniestros incurridos. En particular, la diferencia entre ARIMA y XGBoost fue de $76.033.283 COP, equivalente al 25,14%, mientras que la diferencia entre ARIMA y LightGBM alcanzó los $86.041.500 COP, correspondiente al 28,45%. Estos resultados indican que los modelos basados en machine learning proyectan valores más altos frente al enfoque estadístico tradicional, posiblemente debido a su mayor capacidad para capturar patrones no lineales y variaciones complejas presentes en la serie temporal. Además, la diferencia ligeramente mayor observada en LightGBM es coherente con su mejor desempeño durante la validación walk-forward, consolidándolo como el modelo más robusto dentro del análisis realizado.

TABLA LVIII
COMPARACIÓN DE DIFERENCIAS ABSOLUTAS Y PORCENTUALES DEL NOWCAST ENTRE MODELOS

Comparación de Modelos	Diferencia Absoluta (COP)	Diferencia Relativa
ARIMA vs XGBoost	$66.670.513	21,38%
ARIMA vs LightGBM	$76.678.731	24.59%

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

5) Cierre anual proyectado

Los resultados de la Tabla LIX de la proyección anual para 2025 muestran diferencias relevantes entre los modelos utilizados para estimar el comportamiento futuro de los siniestros incurridos. El modelo ARIMA presentó la mayor proyección acumulada anual con un valor de $3.479.994.683 COP, seguido muy de cerca por LightGBM con $3.313.780.559 COP, mientras que XGBoost estimó un cierre anual más conservador de $2.922.081.280 COP. En cuanto al promedio mensual proyectado, ARIMA y LightGBM mantuvieron valores similares, superiores a los obtenidos por XGBoost, lo que sugiere una tendencia de mayor estabilidad en las estimaciones de estos dos modelos. Sin embargo, al analizar los máximos mensuales proyectados, los modelos de machine learning evidenciaron una mayor capacidad para capturar posibles picos o incrementos abruptos en los siniestros, especialmente LightGBM con un máximo estimado de $388.498.842 COP. En conjunto, los resultados indican que LightGBM logra un equilibrio entre una proyección anual consistente y una mejor sensibilidad frente a variaciones extremas de la serie, coherente con su mejor desempeño obtenido durante la validación walk-forward.

TABLA LIX
PROYECCIÓN DEL CIERRE ANUAL 2025 POR MODELO PREDICTIVO

concepto	ARIMA	XGBoost	LigthGBM
Proyección anual 2025	$3.479.994.683	$2.922.081.280	$3.313.780.559
Promedio mensual proyectado	$289.999.557	$243.506.768	$276.148.380
Máximo mensual proyectado	$311.820.111	$378.490.624	$388.498.842

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

La Tabla LX evidencia diferencias importantes entre las proyecciones generadas por los modelos evaluados. La mayor discrepancia se presenta entre ARIMA y XGBoost, con una diferencia absoluta de $557.913.403, lo que indica comportamientos de predicción considerablemente distintos. Por su parte, ARIMA y LightGBM muestran la menor diferencia ($166.214.124), sugiriendo una mayor similitud en sus estimaciones. En conjunto, estos resultados reflejan que la elección del modelo influye significativamente en las proyecciones obtenidas, especialmente al comparar enfoques tradicionales de series temporales con técnicas de aprendizaje automático.

TABLA LX
DIFERENCIAS ABSOLUTAS EN LAS PROYECCIONES ENTRE LOS MODELOS (COP)

Comparación entre Modelos	Diferencia Absoluta (COP)
ARIMA vs XGBoost	$557.913.403
ARIMA vs LightGBM	$166.214.124
XGBoost vs LightGBM	$391.699.279

Fuente: Elaboración propia a partir de datos de Seguros del Estado S.A., procesados en Python.

IV. APLICACIÓN INTERACTIVA EN STREAMLIT PARA PROYECCIONES, PRESUPUESTO Y ESCENARIOS DE SENSIBILIDAD – OBJETIVO 3

En cumplimiento del tercer objetivo específico, se diseña e implementa una aplicación interactiva en Streamlit, consolidada en el repositorio de GitHub Predicción Fasecolda, con el propósito de integrar en un solo entorno los resultados del análisis predictivo de primas y siniestros del ramo de Responsabilidad Civil Profesional. Esta herramienta permitió trasladar los hallazgos obtenidos en la fase de modelación a una interfaz visual, dinámica y comprensible, orientada a facilitar la consulta de información y apoyar la toma de decisiones estratégicas dentro del contexto asegurador.

La aplicación es concebida como un medio para transformar los resultados técnicos en una solución de uso práctico. En lugar de presentar únicamente tablas o salidas numéricas aisladas, se desarrolló una interfaz capaz de mostrar de forma organizada el comportamiento histórico de las variables, los resultados de los modelos predictivos y los insumos de planeación presupuestal. En ese sentido, el desarrollo de la herramienta cumplió una función integradora, al conectar el procesamiento de datos, la predicción y la visualización en un mismo espacio de trabajo.

De acuerdo con lo implementado en el repositorio, la aplicación incorpora los modelos SARIMAX, XGBoost y LightGBM, los cuales se utilizan para analizar y proyectar el comportamiento de las series temporales. Además, la comparación entre estos modelos se realiza mediante métricas como SMAPE, RMSE y MAE, lo que permite seleccionar la alternativa con mejor desempeño para apoyar las proyecciones. Este aspecto resulta especialmente importante, ya que fortalece la confiabilidad de la herramienta al no depender de un único enfoque, sino de un proceso comparativo que facilita elegir el modelo más adecuado según la evidencia empírica.

La estructura general de la aplicación se organizó en tres pestañas principales: Análisis de Primas, Análisis de Siniestros y Comparación de Modelos. Esta organización permitió distribuir de manera clara los distintos componentes del análisis. En la pestaña de primas, el usuario puede revisar el comportamiento histórico de la producción y contrastarlo con los valores pronosticados por los modelos seleccionados. En la pestaña de siniestros, se incluye igualmente la evolución histórica y la comparación de pronósticos, complementada con un sistema de alertas y un mapa interactivo por municipio. Finalmente, en la pestaña de comparación de modelos se presentan de forma consolidada las métricas de desempeño, facilitando una evaluación más objetiva de los resultados.

Uno de los aportes más importantes en relación con el objetivo 3 fue la construcción de una visualización interactiva del comportamiento histórico y de los pronósticos. La aplicación permite observar en gráficas comparativas los datos de entrenamiento, los datos reales de prueba y las predicciones generadas por cada modelo. Esta funcionalidad aporta valor tanto académico como práctico, ya que facilita la interpretación del ajuste de los modelos y permite identificar de manera más intuitiva la cercanía entre los valores estimados y los observados. De esta forma, la interfaz no solo muestra resultados, sino que contribuye a su análisis e interpretación.

Adicionalmente, se integró una sábana de presupuesto para el año objetivo, elaborada a partir del modelo con mejor desempeño en la predicción de primas. Este componente constituye una respuesta directa al objetivo planteado, pues convierte las proyecciones en un insumo operativo para la planeación. La sábana presenta el pronóstico mensual, el valor acumulado y un presupuesto sugerido, lo cual resulta útil para visualizar la evolución esperada durante el período proyectado. Esta funcionalidad representa un paso importante en la traducción de la modelación predictiva hacia decisiones concretas de asignación y seguimiento presupuestal.

Un elemento clave dentro de esta sábana de presupuesto es el ajuste por IPC, el cual fue implementado en la aplicación mediante un control interactivo que permite modificar el porcentaje aplicado a las proyecciones. Gracias a esta funcionalidad, el usuario puede observar cómo cambian los valores presupuestados bajo diferentes supuestos de ajuste inflacionario. Desde una perspectiva aplicada, esto permite que la herramienta no se limite a generar proyecciones nominales, sino que ofrezca una aproximación más realista a la planeación financiera. En consecuencia, el ajuste por IPC aporta pertinencia al análisis y fortalece la utilidad de la aplicación como apoyo a la toma de decisiones.

En cuanto a los escenarios de sensibilidad, el desarrollo realizado en el repositorio evidencia una aproximación

funcional a través de la modificación dinámica del IPC. Aunque no se observa un módulo independiente con escenarios formalmente etiquetados como optimista, moderado o pesimista, sí existe la posibilidad de variar el porcentaje de ajuste y examinar cómo este impacta el presupuesto sugerido. Por ello, puede afirmarse que el proyecto incorpora una sensibilidad básica pero útil, enfocada en la variación inflacionaria de las proyecciones. Este resultado es coherente con el alcance real de la aplicación y demuestra que la herramienta ya cuenta con una base para futuros desarrollos de escenarios más complejos.

Otro aspecto que fortalece el desarrollo del objetivo 3 es la incorporación de un componente espacial en el análisis de siniestros. La presencia de un mapa interactivo de alertas por municipio amplía la utilidad de la aplicación, ya que permite asociar el comportamiento de los siniestros con una lectura territorial. Esto resulta especialmente valioso en una investigación centrada en la predicción espaciotemporal, pues complementa la proyección temporal con un recurso visual que facilita identificar posibles focos de atención geográfica. Así, la aplicación no solo resume resultados numéricos, sino que también aporta una representación más integral del fenómeno estudiado.

En términos generales, la aplicación desarrollada en Predicción Fasecolda cumple con el propósito de integrar las proyecciones obtenidas y presentarlas en un entorno visual e interactivo. Su diseño permite consultar el comportamiento histórico de primas y siniestros, comparar modelos predictivos, construir presupuestos ajustados por IPC y explorar una primera aproximación a escenarios de sensibilidad. Además, la herramienta incorpora elementos gráficos y espaciales que facilitan la comprensión de los resultados y mejoran su aprovechamiento por parte de los usuarios.

En consecuencia, puede afirmarse que el tercer objetivo específico fue alcanzado satisfactoriamente. La aplicación en Streamlit logró consolidar una solución funcional que articula análisis histórico, pronóstico, comparación de modelos y apoyo presupuestal en una sola plataforma. Esto representa un aporte significativo para la investigación, ya que demuestra que los resultados del análisis predictivo pueden traducirse en una herramienta tecnológica útil, accesible y orientada a fortalecer la toma de decisiones estratégicas en Seguros del Estado S.A.

Fig.34. Interfaz de la aplicación interactiva para el análisis predictivo del importe de las primas y la comparación de modelos. Fuente: Elaboración propia en Streamlit, implementada en el repositorio GitHub Predicción Fasecolda.

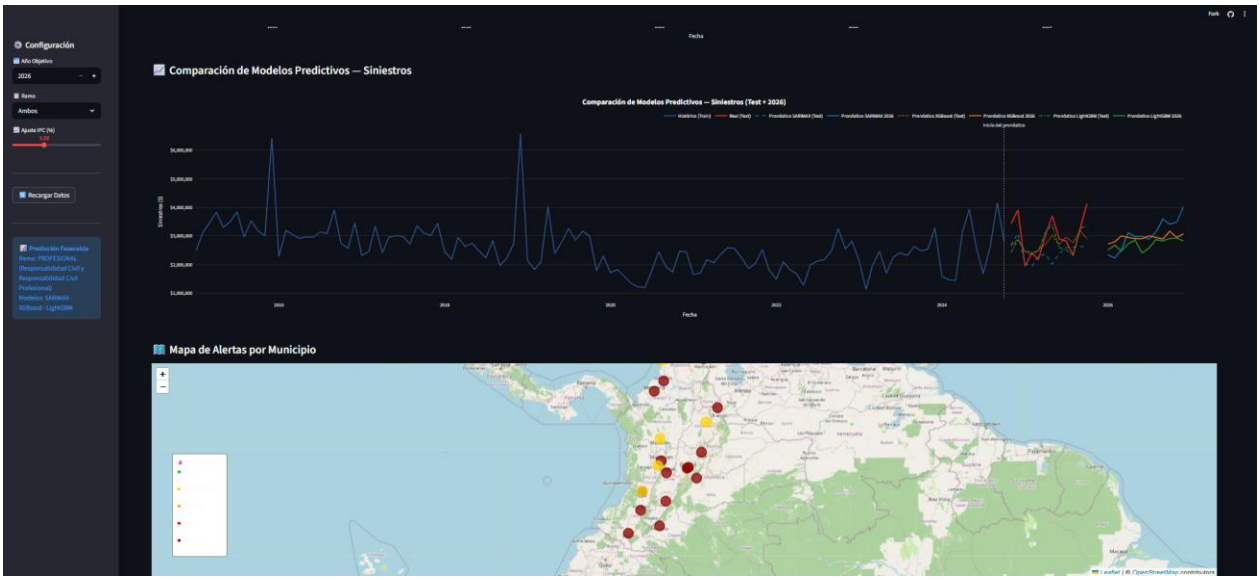


Fig.35. Interfaz de la aplicación interactiva para el análisis predictivo del valor de los siniestros incurridos y la comparación de modelos. fuente: elaboración propia en streamlit, implementada en el repositorio github predicción fasecolda

V. CONCLUSIONES Y TRABAJOS FUTUROS

A. Conclusiones

La compañía de Seguros del Estado S.A. puede aplicar técnicas de machine learning mediante la implementación de modelos predictivos como el SARIMAX/ARIMA, XGBoost y LightGBM para estimar el comportamiento futuro de la producción de primas y el valor de los siniestros incurridos, incorporando variables espaciales relacionadas con la ubicación de las sucursales distribuidas en diferentes ciudades y departamentos del territorio colombiano, junto con la evolución temporal del comportamiento de las primas y los siniestros incurridos optimizando la toma de decisiones estratégicas y operativas en la gestión financiera y administrativa de los recursos de la compañía.

Los datos históricos presentan una periodicidad mensual, correspondiente a los registros realizados el último día de cada mes. Para el importe de las primas, se dispone de información desde el año 2007 hasta el 2025, provenientes de las diferentes sucursales de la compañía distribuidas a nivel nacional. Por otro lado, el importe de las primas presenta una tendencia creciente y un comportamiento estacional desde el año 2007 hasta el 2025, evidenciando un crecimiento relativamente estable durante los primeros años. A partir del año 2015 se observa un incremento más pronunciado en el importe de las primas, destacándose variaciones significativas especialmente durante los meses de enero y febrero, comportamiento asociado a la dinámica comercial y operativa que suele presentarse al inicio de cada año. Los datos históricos correspondientes a los siniestros incurridos presentan igualmente una periodicidad mensual, con registros realizados el primer día de cada mes por las diferentes sucursales de la compañía desde el año 2015 hasta el 2025. Asimismo, la serie evidencia un comportamiento estacional moderado, sin patrones estacionales altamente pronunciados. Estas características temporales permiten identificar esos patrones que son los relevantes para la generación de los modelos predictivos SARIMAX/ARIMA, XGBoost y LightGBM. Previo al análisis de las características espaciotemporales y los patrones estacionales, se evaluó la calidad y consistencia de los datos mediante un proceso de validación y depuración de la información histórica. En este proceso se realizó la normalización de los atributos, el análisis de valores únicos para verificar la consistencia de las categorías utilizadas en el análisis, la identificación de valores faltantes junto con la definición de su tratamiento, y el análisis de la distribución del volumen del importe de las primas por sucursal, con el fin de garantizar la calidad de los datos utilizados en el análisis predictivo. Estos procedimientos permiten garantizar la calidad de la información utilizada, asegurando su adecuación para el desarrollo del análisis predictivo.

El modelo SARIMAX presenta el mejor desempeño frente a los modelos XGBoost y LigthGBM para la estimación temporal del “Importe de las primas” mediante la validación de *Walk-Forward* reflejando un (SMAPE) de 15,50%, porcentaje de error promedio respecto a los valores reales, (RMSE) un error promedio de 1.143.983.646 COP y un MAE de 889.153.776 indicando que, en promedio, el modelo se equivoca en 824.824.246 millones de pesos COP por predicción. Por otro lado, el modelo LightGBM muestra el mejor desempeño en comparación con ARIMA y XGBoost para la estimación del “Valor de los siniestros incurridos” igualmente con la validación de Walk- Forward evidenciando un SMAPE de 11,07%, RMSE de 31.029.292 COP y un MAE DE 27.332.138 COP. En conjunto, los resultados muestran que los modelos de tipo SARIMAX son más adecuados para capturar la dinámica de la producción de primas, mientras que los modelos en *boosting* como LightGBM presentan un mejor desempeño en la predicción de siniestros, lo cual aporta valor a la toma de decisiones en la compañía aseguradora. Cabe resaltar que el objetivo de la compañía es predecir, a nivel agregado, la producción de primas con el fin de estimar de manera más precisa el presupuesto, optimizar la planeación financiera y fortalecer la toma de decisiones estratégicas y operativas en la gestión de sus recursos. El modelo SARIMAX presenta la proyección del importe de las primas, estimando un cierre anual proyectado para el año 2026 de 81.801 millones de COP, lo que permite fortalecer la planeación presupuestal y financiera de la compañía.

Finalmente, se desarrolla la aplicación interactiva Streamlit integrada en el repositorio GitHub Predicción Fasecolda, permitiendo consolidar en una sola plataforma el análisis histórico, la comparación de modelos predictivos y las proyecciones del ramo de Responsabilidad Civil profesional. La herramienta incorpora los modelos SARIMAX/ARIMA, XGBoost y LigthGBM , evaluados mediante las métricas SMAPE, RMSE y MAE, facilitando la selección del modelo con mejor desempeño para las proyecciones. La aplicación permite visualizar de forma dinámica el comportamiento histórico y los pronósticos de primas y siniestros, incorporando una sábana presupuestal

ajustada por IPC y la representación espacialmente del valor de los siniestros incurridos mediante mapas interactivos por municipio. En conjunto, estos resultados demuestran que el análisis predictivo puede transformarse en una herramienta tecnológica funcional y aplicada, orientada a fortalecimiento de la planeación financiera, el seguimiento presupuestal y la toma de decisiones estratégicas y operativas en Seguros del Estado S.A.

B. Trabajos futuros

Se propone ampliar el análisis predictivo mediante la incorporación de variables macroeconómicas, financieras y territoriales, que permitan fortalecer la capacidad explicativa de los modelos. De igual manera, se plantea extender el desarrollo de la aplicación interactiva hacia otros ramos del sector asegurador, incorporando nuevos módulos de análisis, actualización automática de datos y escenarios de sensibilidad más robustos, incluyendo proyecciones optimistas, moderadas y pesimistas. Finalmente, se sugiere fortalecer el componente espaciotemporal mediante técnicas geoespaciales avanzadas y sistemas de alertas tempranas que permitan mejorar el monitoreo y la toma de decisiones estratégicas en Seguros del Estado S.A.

VI. REFERENCIAS

[1] Fasecolda (Federación de Aseguradores Colombianos), «Estadísticas del sector asegurador,» Fasecolda, 24 marzo 2025. [En línea]. Available: https://www.fasecolda.com/fasecolda/estadisticas-del-sector/visualizador-inteligente-de-cifras/dashboard/. [Último acceso: 12 mayo 2025].

[2] M. Crabtree, «DataCamp,» DataCamp, 25 Abril 2024. [En línea]. Available: https://www.datacamp.com/es/blog/what-is-machine-learning.

[3] SEGUROS DEL ESTADO S.A., «SEGUROS DEL ESTADO,» [En línea]. Available: https://www.segurosdelestado.com/pages/GlosarioOP. [Último acceso: 01 06 2025].

[4] Fasecolda-Federación de aseguradores colombianos, «Indicadores del sector asegurador: datos que cuentan historias,» s.f.. [En línea]. Available: https://www.fasecolda.com/definicion-de-los-indicadores-del-sector/.

[5] Superintendencia Financiera de Colombia, «Instrucciones generales relativas a las operaciones de las entidades aseguradoras, capitalización e intermediarios de seguros, Parte II, Título IV, Capítulo II.,» 2017.

[6] Regure By The Algorithm, «Excess of Loss Reinsurance,» s.f.. [En línea]. Available: https://www.getregure.com/glossary/excess-of-loss/?utm_source. [Último acceso: 2026].

[7] M. Chen, «OCI,» OCI, 25 Noviembre 2024. [En línea]. Available: https://www.oracle.com/ar/artificial-intelligence/machine-learning/what-is-machine-learning/.

[8] F. Z. Maksood y G. Achuthan , «Analysis of Data Mining Techniques and its Applications,» *International Journal of Computer Applications (0975 – 8887) ,* vol. Volume 140, nº 3, p. 9, 2016.

[9] M. R. B. d. Quirós, «Análisis de Líneas de Costa con Redes Neuronales LSTM,» UNIVERSITAT OBERTA DE CATALUNYA, https://openaccess.uoc.edu/bitstream/10609/119692/6/mrivasbeTFM0620memoria.pdf, 2020.

[10] J. A. Mauricio, Análisis de Series Temporales, Madrid: Universidad Complutense de Madrid, 2007.

[11] G. E. Box, G. M. Jenkins, G. C. Reinsel y G. M. Ljung, Análisis de series de tiempo: pronóstico y control, 5.ª edición, Hoboken, New Jersey, EE. UU: John Wiley & Sons, Inc., 2015.

[12] R. J. Hyndman y G. Athanasopoulos, Pronóstico: principios y práctica (3.ª ed.), Melbourne: OTexts, 2021.

[13] S. L. Rojas, R. C. García Sánchez, R. García Mata, O. A. Arana Coronado y A. González Estrada, «Metodología Box- Jenkins para pronosticar los precios de huevo blanco pagados al productor en México,» *Agrociencia,* vol. 53, nº 6, p. 16, 2019.

[14] R. F. Engle, «Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,» *The Econometric Society,* vol. Vol. 50, p. pp. 987–1007, 1982.

[15] P. G. Zhang, «Pronóstico de series temporales utilizando un modelo híbrido ARIMA y de red neuronal,» *ScienceDirect (Elsevier),* vol. Volumen 50, pp. páginas 159-175, Enero 2003.

[16] E. Kavlakoglu y E. Russi, «IBM,» [En línea]. Available: https://www.ibm.com/mx-es/think/topics/xgboost. [Último acceso: 01 06 2025].

[17] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System,» p. 13, 2016.

[18] A. Velazquez Bustamante, «Medium,» 9 Enero 2024. [En línea]. Available: https://antonio-velazquez-bustamante.medium.com/entendiendo-a-fondo-el-lighgbm-ed157f601535. [Último acceso: 10 Mayo 2026].

[19] R. J. Hyndman y A. B. Koehler, «Another look at measures of forecast accuracy,» *International Journal of Forecasting 22 (2006) 679 – 688,* vol. 22, nº 4, p. 679 – 688, 2006.

[20] M. Castany, «Quantdemy,» 18 Abril 2023. [En línea]. Available: https://quantdemy.com/metodo-walk-forward/. [Último acceso: 2025 06 01].

[21] W. T. Watson, «WTW,» 11 01 2024. [En línea]. Available: https://www.wtwco.com/es-es/insights/2024/01/un-mercado-a-dos-niveles-tendencias-y-predicciones-sobre-seguros-a-seguir-de-cerca-en-2024. [Último acceso: 01 06 2025].

[22] C. D. G. González, J. O. Garay Rey, J. C. Robayo Ávila, L. V. González Castaño y N. C. Niño Calixto, «Modelo Predictivo de Siniestros en Seguros Dentales,» Fundación Universitaria Compensar, Bogotá D.C., 2023.

[23] E. V. Rivera Jimenez y W. . D. Pineda Rios, «Pronóstico de la tasa de siniestralidad de un seguro de desempleo en Colombia mediante modelos lineales dinámicos.,» Universidad Santo Tomas - Facultad de Estadística, Maestría en Estadística Aplicada.

[24] S. O. Alzate, «Exploración de las redes neuronales para la proyección de la máxima pérdida esperada de una póliza de seguros: aplicación para un seguro previsional,» Universidad Nacional de Colombia - Facultad de Ciencias, Escuela de Estadística - Maestría en Estadística, Medellín, 2021.

[25] J. Durbin y S. J. Koopman, Time Series Analysis by State Space Methods, Oxford: Oxford University Press, 2012.

[26] J. Perktold, S. Seabold y J. Taylor, «statsmodels-developers.,» [En línea]. Available: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html?utm_source. [Último acceso: 12 Octubre 2025].

[27] J. Cummins y G. L. Griepentrog, «Forecasting automobile insurance paid claim costs using econometric and ARIMA models,» *ScienceDirect (Elsevier),* vol. 1, pp. Pages 203-215, 1985.

[28] J. Nápoles-Duarte, A. Biswas, M. I. Parker, J. Palomares-Baez, M. A. Chávez-Rojo y L. M. Rodríguez-Valdez, «Stmol: A component for building interactive molecular visualizations within streamlit web-applications,» *Frontiers in Molecular Biosciences,* vol. 9, p. Art. 990846, 22.

[29] G. C. Fabrizzio, A. L. Erdmann y L. M. d. Oliveira, «Web App for prediction of hospitalisation in Intensive Care Unit by covid-19,» *Revista Brasileira de Enfermagem,* vol. 76, nº 6, p. e20220740, Diciembre 2023.

[30] J. Han, M. Kamber y J. Pei, Data Mining: Concepts and Techniques, Waltham, Massachusetts, Estados Unidos: Morgan Kaufmann (sello de Elsevier), 2011.

[31] J. Yi, J. Lee, K. J. Kim, S. J. Hwang y E. Yang, WHY NOT TO USE ZERO IMPUTATION? CORRECTING, 2020.