Predicción de la inflación en Colombia mediante un análisis comparativo de los modelos ARIMA y Lag-Llama

Shirley Andrea Cala Pérez

Trabajo de grado para optar al título de Ingeniera Industrial

Director

Vlakxmir Robles Marín

Magíster en Ingeniería

Universidad Industrial de Santander

Facultad de Ingenierías Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2025

Dedicatoria

A mi madre, que siempre soñó con verme convertida en profesional. Este logro también le pertenece, con su amor y sacrificio me enseñó a luchar por aquello que ella no tuvo la oportunidad de vivir. A ti, mamá, dedico cada esfuerzo y cada página de este trabajo, con la esperanza de que sientas el mismo orgullo por mí que yo siento por ti.

Tabla de contenido

Lista de tablas

Lista de figuras

Lista de Apéndices

Los apéndices están adjuntos y puede visualizarlos en la base de datos de la biblioteca UIS

Apéndice A. Conjunto de datos IPC interanual

Apéndice B. Cuadernos Jupyter en Python de la metodología implementada

Apéndice C. Artículo científico de carácter publicable

Resumen

Título: Predicción de la inflación en Colombia mediante un análisis comparativo de los modelos ARIMA y Lag-Llama[*]

Autor: Shirley Andrea Cala Pérez[**]

Palabras clave: Inflación, Series temporales, ARIMA, Lag-Llama, Modelos de pronóstico, Aprendizaje automático.

Descripción:

La inflación constituye uno de los fenómenos macroeconómicos más relevantes debido a su impacto en la estabilidad económica, el poder adquisitivo de los hogares y la toma de decisiones. Su predicción ha sido objeto de estudio durante décadas mediante modelos estadísticos tradicionales y, más recientemente, a través de técnicas de aprendizaje profundo orientadas a capturar dinámicas no lineales en escenarios de alta volatilidad. Este estudio realiza un análisis comparativo de ambos enfoques tomando la inflación interanual en Colombia durante el período 2004–2024. Se desarrolló una revisión de la literatura y un análisis bibliométrico para contextualizar el uso de modelos clásicos y emergentes en el pronóstico económico, seguido de la implementación de los modelos ARIMA y Lag-Llama en el lenguaje de programación Python bajo dos esquemas de predicción: Multi-step (12 meses simultáneos) y Rolling-step (un mes con actualización continua). El desempeño se evaluó mediante métricas como MAE, RMSE y MAPE. Los resultados muestran que ARIMA alcanzó un mejor desempeño en ambos esquemas, destacándose en el pronóstico rolling con un MAPE de 2,61 %, mientras que Lag-Llama presentó errores más altos, lo cual sugiere la necesidad de procesos de fine-tuning para mejorar su precisión en el contexto colombiano. En conclusión, los hallazgos confirman que ARIMA se consolida como un modelo confiable y aplicable en el corto plazo, mientras que Lag-Llama representa una alternativa prometedora a futuro, aportando además un antecedente novedoso en la aplicación de modelos de inteligencia artificial al análisis inflacionario en Colombia.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico – Mecánicas. Escuela de Estudios Industriales y Empresariales, Ingeniería
Industrial. Director: Vlakxmir Robles Marín, M.Sc. Ingeniería Industrial.

Abstract

Title: *Forecasting Inflation in Colombia through a Comparative Analysis of ARIMA and Lag-Llama Models*[*]

Author: Shirley Andrea Cala Pérez[**]

Keywords: Inflation, Time series, ARIMA, Lag-Llama, Forecasting models, Machine learning.

Description:
Inflation is one of the most relevant macroeconomic phenomena due to its impact on economic stability, household purchasing power, and decision-making. Its prediction has been studied for decades using traditional statistical models and, more recently, deep learning techniques designed to capture nonlinear dynamics in highly volatile scenarios. This study conducts a comparative analysis of both approaches using year-on-year inflation in Colombia over the period 2004–2024. A literature review and a bibliometric analysis were carried out to contextualize the use of classical and emerging models in economic forecasting, followed by the implementation of ARIMA and Lag-Llama models in Python under two prediction schemes: multi-step (12 months simultaneously) and Rolling-step (one month with continuous updating). Performance was evaluated using metrics such as MAE, RMSE, and MAPE. The results show that ARIMA achieved better performance in both schemes, particularly in rolling forecasts with a MAPE of 2.61 %, while Lag-Llama exhibited higher errors, suggesting the need for fine-tuning to improve its accuracy in the Colombian context. In conclusion, the findings confirm that ARIMA remains a reliable and applicable model in the short term, while Lag-Llama represents a promising alternative for the future, also providing a novel precedent in the application of artificial intelligence models to inflation analysis in Colombia.

---

[*] Undergraduate Thesis

[**] Faculty of Physical-Mechanics Engineering. School of Industrial and Business Studies, Industrial Engineering. Advisor: Vlakxmir Robles Marín M.Sc. Industrial Engineering.

Introducción

Históricamente, la inflación ha sido objeto de estudio en el ámbito económico, a causa de su influencia directa en las tasas de interés, la capacidad de adquisición de los hogares, la rentabilidad de las empresas y la estabilidad macroeconómica de los países (Bernanke & Mishkin, 1997). De acuerdo con el Banco de la República (2023) en Colombia, las presiones inflacionarias provienen tanto de factores internos (como del aumento en la demanda) y de influencias externas, sobre todo en productos como el petróleo y los alimentos. Esta situación ha puesto la necesidad de contar con herramientas más precisas y flexibles que nos ayuden a monitorear y prever las variaciones en los precios.

En este contexto, Lag-Llama, una arquitectura propuesta por Rasul et al. (2023) se presenta como una alternativa innovadora. Esta herramienta, construida sobre modelos de tipo *transformer*[1], permite integrar información tanto de corto como de largo plazo. Su principal ventaja es su capacidad de adaptarse a distintos escenarios sin necesidad de ser entrenada previamente con datos específicos del problema, lo que se conoce como enfoque *zero-shot*, es decir, la capacidad del modelo para realizar predicciones razonables sobre nuevas tareas sin requerir entrenamiento adicional. Esta característica la convierte en una opción eficaz para entornos con alta variabilidad. A pesar de que este modelo ha sido estudiado en otras disciplinas como la ingeniería, la medicina, la energía y las ciencias sociales (Ali et al., 2024; Amjad et al., 2024; Saravanan et al., 2024) aún

[1] Un modelo transformer es una arquitectura de redes neuronales que reemplaza la recurrencia por mecanismos de autoatención, lo que le permite procesar secuencias de datos de manera paralela y eficiente. Vaswani et al. (2017)

no se ha registrado una investigación científica sobre su uso en contextos económicos específico de Colombia. Esto ofrece la oportunidad de evaluar su efectividad frente a técnicas tradicionales como ARIMA y determinar si puede ofrecer ventajas más relevantes en la predicción inflacionaria.

Este estudio pretende realizar un contraste entre los modelos ARIMA y Lag-Llama con el fin de evaluar su capacidad para determinar la inflación mensual en Colombia. Se emplearán a las series históricas del Índice de Precios al Consumidor (IPC) y se usaran indicadores de evaluación como el MAE, MSE y MAPE, que facilitan la medición del margen de error en las proyecciones. La investigación abarca desde la revisión teórica y la preparación de la información, hasta la ejecución computacional en Python y el análisis comparativo de los hallazgos. Con esto, se pretende aportar acerca de la capacidad de los modelos de inteligencia artificial para el estudio económico del país.

Este documento incluye un marco de referencia (Sección 4) que establece los antecedentes y fundamentos conceptuales de la investigación, seguido de la metodología (Sección 5), organizada en cuatro fases: revisión de la literatura, preparación de los datos, implementación de los modelos y evaluación de resultados. En la Sección 10 se presenta la documentación del proyecto, que contempla la elaboración de un artículo académico, la consolidación del documento final y el desarrollo de un repositorio en GitHub. Finalmente, las Secciones 11 y 12 recogen las conclusiones y recomendaciones, así como posibles líneas de investigación futura, junto con las referencias bibliográficas que respaldan el estudio.

1. Planteamiento del problema

La inflación en Colombia ha sido un problema persistente, con episodios de alta volatilidad como el pico del 13.12 % en 2022 en el Índice de Precios al Consumidor (IPC), impulsada por factores estructurales como la alta dependencia de importaciones, rigideces en la oferta y cuellos de botella productivos, sumados a perturbaciones externas derivadas del conflicto en Ucrania, fenómenos climáticos extremos y tensiones en los mercados internacionales (DANE, 2023). Además a esto, sus efectos se evidencian en la vida diaria: se encarece el crédito, se pierde poder adquisitivo y se amplían las brechas sociales (Banco Mundial, 2024).

En este contexto, predecir el comportamiento de la inflación es clave tanto para el diseño de políticas monetarias como para la planificación de empresas y hogares. Sin embargo, el desfase en las publicaciones del IPC por parte del DANE, limita la capacidad de reacción oportuna, lo que ha hecho del nowcasting una herramienta indispensable para instituciones como el Banco de la República y para actores económicos en contextos volátiles (Modugno, 2013).

Frente a este desafío, los modelos tradicionales como ARIMA presentan limitaciones importantes en contextos de alta variabilidad. Por ejemplo, estudios recientes muestran que tienden a subestimar la inflación de alimentos en Colombia en hasta 2 puntos porcentuales durante el fenómeno de El Niño, debido a su incapacidad para capturar saltos repentinos en los precios agrícolas (Bejarano Salcedo et al., 2022). Frente a ello, surgen alternativas basadas en deep learning, como Lag-Llama, una arquitectura de tipo transformer diseñada para realizar predicciones incluso sin entrenamiento específico con los datos del problema. Su flexibilidad le permite adaptarse mejor a dinámicas no lineales y entornos con alta incertidumbre (Rasul et al., 2023).

A diferencia de estudios previos centrados en economías desarrolladas, esta investigación aplica por primera vez el modelo Lag-Llama al caso colombiano, caracterizado por una informalidad laboral de 58 %, rezagos en la disponibilidad de datos y una alta sensibilidad a eventos disruptivos (DANE, 2023). Evaluar su desempeño en este entorno aporta evidencia novedosa para la literatura econométrica y puede fortalecer el diseño de herramientas predictivas en economías emergentes.

Además de su valor académico, esta comparación entre ARIMA y Lag-Llama tiene implicaciones prácticas ya que pretende determinar cuál de los dos modelos ofrece mayor precisión y adaptabilidad en la predicción de la inflación puede contribuir a modernizar las estrategias de política monetaria, y también a mejorar la toma de decisiones en sectores especialmente sensibles a las variaciones de precios.

A partir de esto, se formula la siguiente pregunta de investigación:

¿Cuál de los modelos, ARIMA o Lag-Llama, ofrece mayor precisión y adaptabilidad en la predicción de la inflación en Colombia?

2. Objetivos

2.1 Objetivo general

Analizar y comparar la eficacia de los modelos ARIMA y Lag-Llama en la predicción de la inflación en Colombia, con el propósito de determinar cuál de los dos modelos ofrece mejores resultados en términos de precisión y aplicabilidad en el contexto económico del país.

2.2 Objetivo Específicos

Realizar una revisión bibliográfica sobre los fundamentos teóricos y la aplicabilidad de los modelos ARIMA y Lag-Llama en la predicción de series de tiempo de la inflación.

Caracterizar los datos históricos de inflación en Colombia para su implementación en ambos modelos de predicción.

Aplicar los modelos ARIMA y Lag-Llama a los datos recopilados para analizar y estimar el comportamiento mensual de la inflación en el país.

Comparar el desempeño de ambos modelos mediante métricas de error para determinar su precisión y aplicabilidad.

Elaborar un artículo académico de carácter publicable que sintetice los hallazgos obtenidos de la investigación realizada.

2.3 Resultados esperados

Revisión teórica sobre los modelos ARIMA y Lag-Llama, que permitirá entender sus fundamentos, cómo han sido utilizados en la predicción de la inflación y qué ventajas y desventajas presentan en diferentes contextos económicos.

Base de datos estructurada con información histórica de la inflación en Colombia, correspondiente al periodo 2004-2024 (252 observaciones mensuales), recopilada de fuentes

oficiales, asegurando su correcta organización, segmentación y preparación para el análisis de series temporales.

Implementación de los modelos ARIMA y Lag-Llama en el lenguaje de programación Python, considerando una división diferenciada de los datos para entrenamiento y prueba, según las características metodológicas propias de cada enfoque.

Comparación del desempeño de ambos modelos, mediante métricas estadísticas como MAE (Error Absoluto Medio), MSE (Error Cuadrático Medio) y MAPE (Error Porcentual Absoluto Medio), con el fin de determinar cuál ofrece mayor precisión en el pronóstico inflacionario.

Análisis e interpretación de los resultados obtenidos, destacando qué modelo se ajusta mejor a la dinámica inflacionaria de Colombia y en qué situaciones podría ser más útil que el otro.

Artículo académico con los hallazgos de la investigación, donde se expongan los resultados y su posible impacto en estudios futuros sobre predicción económica.

3. Cumplimiento de objetivos

A continuación, en la tabla 1, se detalla el cumplimiento de los objetivos establecidos en el presente trabajo.

Cumplimiento de objetivos

Tabla 1

Cumplimiento de objetivos

Objetivo	Página / Apéndice
Realizar una revisión bibliográfica sobre los fundamentos teóricos y la aplicabilidad de los modelos ARIMA y Lag-	Página 40

Llama en la predicción de series de tiempo de la inflación.	
Caracterizar los datos históricos de inflación en Colombia para su implementación en ambos modelos de predicción.	Página 55
Aplicar los modelos ARIMA y Lag-Llama a los datos recopilados para analizar y estimar el comportamiento mensual de la inflación en el país.	Página 61
Comparar el desempeño de ambos modelos mediante métricas de error para determinar su precisión y aplicabilidad.	Página 84
Elaborar un artículo académico de carácter publicable que sintetice los hallazgos obtenidos de la investigación realizada.	Página 87 / Apéndice C

4. Marco de referencias

4.1 Marco de Antecedentes

En Colombia se han llevado a cabo diversas investigaciones que utilizan modelos de aprendizaje automático y estadísticos con el fin de incrementar la precisión en el pronóstico de la inflación. Dentro de estas contribuciones se encuentra el trabajo realizado por Carmona Restrepo (2022), desarrolla un modelo de redes neuronales regularizadas para pronosticar la inflación en Colombia, con el objetivo de contrastar tres modelos de penalización (LASSO, Ridge y Elastic Net) frente a uno clásico como el ARIMA. Recopila de 24 series económicas obtenidas de Banco de la República de Colombia, en las cuales estandariza una periodicidad mensual y tras un proceso de selección mediante un modelo aditivo generalizado (GAM), reduce el conjunto a siete variables con mayor aporte marginal. Adicional, el horizonte de predicción fijado es de un mes, yendo en coherencia con aplicaciones de corto plazo en decisiones financieras y de política monetaria. La

evaluación se realiza con validación cruzada y métricas fuera de muestra (MSE y R^2), eligiendo mejor especificación una red MLP regularizada con Elastic Net, que exhibe el menor error y el mayor coeficiente de determinación, posteriormente al comparar con un ARIMA $(3,2,0)(2,0,0)_{12}$, reporta un MSE superior para el ARIMA, concluyendo que la regularización mejora la capacidad de generalización del modelo neuronal.

Por otro lado, Loaiza Zapata (2022) plantea el pronóstico de la inflación mensual en Colombia a partir de determinantes macroeconómicos mediante un enfoque comparativo entre modelos de series temporales tradicionales y de machine learning. Para ello, selecciona 24 variables económicas y financieras del DANE y el Banco de la República, y desarrolla ocho modelos: cinco de regresión múltiple (regresión lineal, KNN, SVR, MLP y árboles de decisión) y tres de series temporales (ARIMA, LSTM univariable y LSTM multivariable). Los resultados aluden que el ARIMA es el modelo con menor error de predicción, seguido por el LSTM multivariable, el cual aporta una mejor interpretación al identificar qué variables explican el comportamiento de la inflación aunque presenta limitaciones derivadas de la alta correlación de las variables y la complejidad de su arquitectura. El estudio concluye que aunque el ARIMA sigue siendo competitivo, los modelos basados en redes neuronales pueden capturar mejor cambios drásticos en los determinantes, siempre que se optimice su diseño.

Con un matiz distinto, Peña Ordóñez (2019) desarrolla un enfoque mixto para el pronóstico de la inflación colombiana en el corto plazo, combinando un modelo ARIMA desagregado con un algoritmo de machine learning Random Forest. El autor parte de la construcción de una serie de entre la canasta base 2009-2018 y la base 2018, lo que permite analizar las 12 divisiones de gasto que conforman el IPC. La estrategia consistió en aplicar un modelo ARIMA para cada división y luego ponderar los resultados para obtener la inflación nacional mensual, mientras que el Random

Forest se aplicó directamente sobre la serie de inflación mensual. Ambos modelos se evaluaron en el periodo noviembre 2017 - octubre 2019. Los hallazgos muestran que el Random Forest mejora el desempeño del ARIMA desagregado bajo diferentes métricas de error, evidenciando la capacidad de los métodos de machine learning para capturar no linealidades y choques exógenos que los modelos lineales tradicionales no logran reflejar de forma adecuadamente.

Un trabajo de grado desarrollado en la Universidad Industrial de Santander (UIS), titulado Modelo Predictivo de la Variable TRM: Análisis de Box-Jenkins vs. Filtro de Kalman, buscó comparar la eficacia de dos metodologías para pronosticar la Tasa Representativa del Mercado (TRM) colombiana. Arciniegas Hernández y Castaño Arévalo (2022) presentan un modelo predictivo aplicado a la Tasa Representativa del Mercado (TRM) en Colombia, comparando la metodología Box-Jenkins ARIMA con la técnica de Kalman Filters. El estudio se centra en evaluar la eficiencia de ambas metodologías para pronosticar la serie cambiaria en un horizonte de cinco años, empleando información histórica del Banco de la República. La metodología incluyó una fase de análisis bibliométrico, procesamiento de datos, implementación en Python y validación de desempeño mediante métricas como RMSE, MAD y MAPE. Los resultados evidencian que, si bien el modelo ARIMA ofrece una estructura clásica y probada en el análisis de series económicas, el filtro de Kalman logra captar dinámicas latentes y volatilidad aportando estimaciones más flexibles en escenarios de alta incertidumbre.

A partir de estos antecedentes, se evidencia que el uso de técnicas de aprendizaje automático en el pronóstico de la inflación en Colombia ha tenido relevancia en los últimos años, y que su desempeño, en muchos casos ha superado al de modelos econométricos clásicos como ARIMA.

4.2 Marco Teórico

4.2.1 Inflación

La inflación es uno de los temas más tratados en economía, debido al impacto que puede generar sobre el poder adquisitivo de las personas, las decisiones en cuando a política monetaria y, en última instancia, sobre el bienestar general de la sociedad. De manera sencilla, puede entenderse como un aumento sostenido y generalizado en los precios de los bienes y servicios de una economía, lo que implica que el dinero pierde valor con el tiempo. (Banco de la República, 2023a)

Desde el enfoque monetarista, una de las interpretaciones más influyentes proviene de los estudios de Friedman (1968), quien sostenía que la inflación es esencialmente, un fenómeno monetario. Según esta visión, cuando la cantidad de dinero en circulación crece de manera más rápida que la producción de bienes y servicios, se generan presiones sobre los precios. Friedman & Schwartz (1963), concluyeron que las variaciones en la oferta monetaria explicaban en gran medida, los episodios inflacionarios de largo plazo en Estados Unidos.

No obstante, existen otras corrientes del pensamiento económico que explican la inflación desde diferentes ángulos. La teoría estructuralista, particularmente aplicada al contexto de América Latina, considera que la inflación en países en desarrollo está fuertemente influenciada por problemas estructurales como la rigidez de los mercados laborales y productivos, la escasez de ciertos insumos, los cuellos de botella en la producción junto con la fuerte dependencia de las importaciones (Pinto, 1970; Prebisch, 1996). Desde esta óptica, los procesos inflacionarios responden a fallas profundas en la estructura económica, más que a un exceso de demanda monetaria.

Por su parte, la teoría keynesiana pone énfasis en el rol de la demanda agregada y en las expectativas que los agentes económicos tienen respecto al comportamiento futuro de la economía. En este enfoque, un incremento sostenido en la demanda, sin un aumento proporcional en la oferta, puede generar presiones sobre los precios. Asimismo, las expectativas inflacionarias pueden actuar como un factor autónomo de inflación, pues si los agentes anticipan un aumento en los precios, ajustan su comportamiento en función de esa expectativa, provocando que la inflación efectivamente ocurra (Keynes, 1963)

Otros enfoques han ampliado aún más las posibles causas de este fenómeno. Algunos análisis provenientes de la teoría del ciclo político-económico sugieren que los gobiernos en determinados contextos, pueden inducir presiones inflacionarias deliberadas en períodos previos a elecciones, con el objetivo de estimular temporalmente el crecimiento y favorecer su imagen pública (Nordhaus, 1975). A su vez, la teoría del pass-through advierte sobre cómo las devaluaciones del tipo de cambio, más alusivamente en economías dependientes de importaciones, pueden trasladarse rápidamente a los precios internos (Goldfajn & Werlang, 2000). También se han desarrollado argumentos en torno a la llamada inflación por costos, que ocurre cuando aumentos en los precios de los factores de producción como los salarios, las materias primas o la energía, son trasladados a los consumidores finales, impulsando al alza el nivel general de precios (Blanchard & Johnson, 2013).

De forma más reciente, algunas corrientes modernas han combinado elementos de distintos enfoques tradicionales. Por ejemplo, la teoría neokeynesiana incorpora expectativas racionales y rigideces nominales, explicando que los precios y salarios no se ajustan de manera inmediata a las condiciones económicas cambiantes, lo que puede generar inflación persistente incluso con políticas monetarias estables (Mankiw, 2006).

En el caso colombiano, el Banco de la República define la inflación como “el aumento sostenido del nivel general de precios de los bienes y servicios que forman parte del consumo de los hogares” (Banco de la República, 2023a). Esta se mide oficialmente a través del Índice de Precios al Consumidor (IPC), calculado por el Departamento Administrativo Nacional de Estadística (DANE). Además de reflejar el aumento del costo de vida, el IPC es un instrumento esencial para la toma de decisiones en política monetaria, así como para la negociación de salarios, tarifas, subsidios y contratos. Disponer de estimaciones razonablemente precisas permite mejorar decisiones clave como la fijación de tasas de interés, aumentos salariales o planificación contractual. Según el Banco Mundial (2024) mantener la inflación bajo control no solo favorece la estabilidad económica, sino que también crea un entorno más propicio para el crecimiento sostenible y la atracción de inversión extranjera.

4.2.2 Series Temporales

Las series temporales son una herramienta indispensable para el estudio de datos que evolucionan en el tiempo. Para Montenegro García (2011), una serie temporal es una sucesión cronológicamente ordenada, donde el tiempo no es sólo una variable auxiliar, sino que se trata de una dimensión que determina la estructura de los propios datos y las técnicas de análisis susceptibles de ser empleadas.

4.2.2.1 Tipos de Series Temporales. Las series temporales pueden ser determinísticas y estocásticas. Las series determinísticas son aquellas que muestran patrones marcados y permanentes, como la tendencia o la estacionalidad. Las series estocásticas llevan un componente aleatorio; los valores futuros dependerán parcialmente de los valores previos, pero a la vez también incluirán un grado de incertidumbre.

4.2.2.2 Componentes de las Series Temporales. De acuerdo con Montenegro García (2011), los cuatro componentes que se identifican en una serie temporal son:

Tendencia: hace referencia a la evolución a largo plazo, asociada a los cambios estructurales o un crecimiento sostenido. Se puede modelizar con funciones lineales, polinómicas o exponenciales, dependiendo de su forma.

Estacionalidad: son patrones que se repiten periódicamente en intervalos regulares como un ciclo anual del consumo o de la producción. Reúne efectos predecibles del calendario como son las estaciones o los acontecimientos recurrentes.

Ciclos: fluctúan en períodos irregulares y suelen estar asociadas a ciertas variables del proceso macroeconómico tal como son las recesiones o expansiones. A diferencia de la estacionalidad, los ciclos no siguen una regularidad fija.

Componente aleatorio o irregular: corresponde la variabilidad no capturada por los componentes anteriores y que es generado por la aparición de eventos inesperados, errores en la medición o perturbaciones externas.

4.2.2.3 Enfoques de modelado. El análisis de series temporales permite construir modelos que capturan relaciones históricas con el fin de proyectar comportamientos futuros. Estos enfoques pueden agruparse en dos grandes categorías:

Modelos estadísticos clásicos, como ARIMA, SARIMA o VAR, que requieren supuestos de estacionariedad y linealidad (Box & Jenkins, 1976).

Modelos de aprendizaje automático, que permiten capturar relaciones complejas sin necesidad de suposiciones rigurosas, como las redes neuronales recurrentes (RNN, LSTM) o algoritmos de árboles (Random Forest, XGBoost). (Goodfellow et al., 2016)

4.2.3 Modelo ARIMA

El modelo ARIMA (AutoRegressive Integrated Moving Average) fue propuesto por Box y Jenkins (1976) como una técnica estadística para modelar y pronosticar series temporales. Su objetivo principal consiste en buscar modelar la correlación entre los rezagos de la serie temporal y los términos de error asociados a la predicción, permitiendo realizar proyecciones a futuro con base en la estructura histórica de los datos. Este modelo integra tres componentes: autorregresivo (AR), integrado (I) y de medias móviles (MA), y se expresa como ARIMA *(p, d, q),* donde:

p: Número de términos autorregresivos (AR).

d: Número de diferenciaciones necesarias para hacer la serie estacionaria.

q: Número de términos de medias móviles (MA).

4.2.3.1 Componente Autorregresivo (AR). Este componente implica que los valores actuales de la serie dependen linealmente de valores pasados. Es decir, la variable a predecir se modela en función de sus propios valores anteriores, con un término de error aleatorio:

$$y_t = \emptyset_1 y_{t-1} + \ \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \epsilon_t \quad (1)$$

En esta Ecuación 1, y_t representa el valor actual de la serie, $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ son los coeficientes autorregresivos que indican la influencia de cada rezago en la predicción, y ϵ_t es el término de error aleatorio, que refleja factores externos o ruido en los datos. Este componente resulta útil cuando se observan patrones persistentes de dependencia entre los valores presentes y pasados de la serie.

4.2.3.2 Componente Integrado (I). El componente Integrado (I) permite transformar una serie no estacionaria en estacionaria mediante la diferenciación de los datos. La diferenciación se calcula restando el valor actual con el valor anterior:

$$y'_t = y_t - y_{t-1} \quad (2)$$

Aquí, en la Ecuación 2, y'_t representa la nueva serie transformada. Si después de una diferenciación la serie aún presenta tendencias o varianza inestable, puede aplicarse el proceso nuevamente. Este componente es clave para estabilizar la media y la varianza de la serie, permitiendo un modelado más preciso de su comportamiento.

4.2.3.3. Componente de Medias Móviles (MA). El componente Medias Móviles (MA) modela la relación entre los valores actuales y los errores de predicción pasados. A diferencia del componente AR, que se basa en los valores pasados de la serie, el MA se centra en los residuos o errores que se han generado al estimar dichos valores. La expresión general del modelo es:

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \ \theta_1 \epsilon_{t-1} - \cdots + \theta_q \epsilon_{t-q} \quad (3)$$

En esta Ecuación 3, y_t es el valor actual de la serie, mientras que ϵ_t representa el error aleatorio del periodo actual. Los coeficientes$\theta_1,\ \theta_2 \ldots \theta_q$ indican el impacto de los errores pasados en la predicción de los valores futuros.

4.2.3.4. Metodología Box-Jenkis. La Metodología Box-Jenkins es un enfoque estructurado para la construcción de modelos ARIMA, que se compone de tres fases principales, qu ese detalla a continuación bajo lo expuesto por Hyndman y Athanasopoulos (2018):

Identificación. En esta primera etapa se prepara la serie y se selecciona un modelo candidato verificando su estacionariedad, condición indispensable para la aplicación de ARIMA. Para este fin se emplean pruebas como la Dickey–Fuller Aumentada (ADF), que contrasta la hipótesis de existencia de raíz unitaria y permite determinar si es necesario aplicar diferenciación (Said & Dickey, 1984). Asimismo, se analizan los correlogramas de la función de autocorrelación (ACF), que mide el grado de correlación entre la serie y sus rezagos, y de la función de autocorrelación parcial (PACF), que estima la correlación con un rezago específico eliminando el efecto de los rezagos intermedios. Estas funciones sirven como guía para proponer los posibles

órdenes de los componentes autorregresivos (p) y de medias móviles (q), respectivamente (Box et al., 2016; Hyndman & Athanasopoulos, 2018).

Estimación. Una vez identificado el modelo, se procede a estimar sus parámetros. En la práctica, los programas estadísticos emplean rutinas iterativas que optimizan los coeficientes para minimizar los errores de predicción. La comparación entre modelos candidatos se apoya en criterios de información, entre ellos el Akaike Information Criterion (AIC), que evalúa el equilibrio entre ajuste y complejidad, (Akaike, 1974) y el Bayesian Information Criterion (BIC), que introduce una penalización más estricta a medida que aumenta el número de parámetros (Schwarz, 1978).

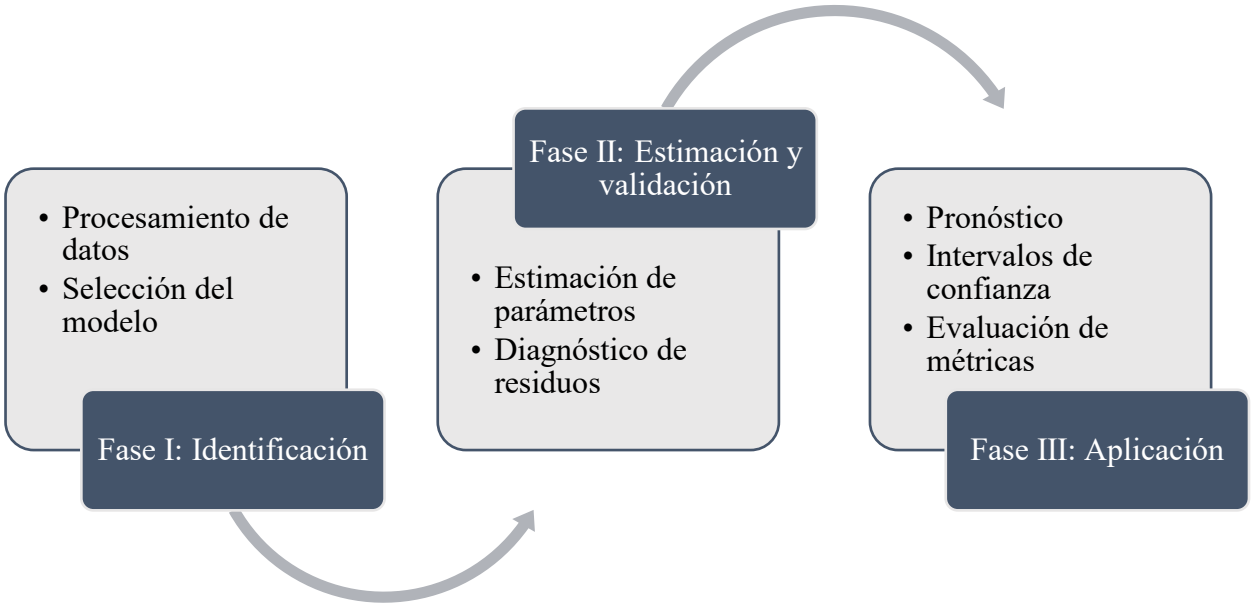
Validación. Busca determinar la adecuación del modelo estimado, para ello se revisa que los residuos se asemejen a un proceso de ruido blanco, es decir, que presenten media cero, varianza constante y ausencia de autocorrelación significativa (Box et al., 2016). Este diagnóstico se complementa con pruebas de significancia, entre ellas la Ljung–Box, que evalúa si las autocorrelaciones de los residuos difieren de cero y, por lo tanto, si el modelo ha logrado capturar la estructura de la serie (Ljung & Box, 1978). Además, se exige que el modelo sea parsimonioso y estable, cumpliendo los criterios de estacionariedad e invertibilidad. En caso contrario, es necesario retornar a la fase de identificación y ajustar un modelo alternativo.

Pronóstico: Una vez validado, el modelo se utiliza para generar proyecciones a corto, mediano y largo plazo. Estas predicciones se presentan junto con sus intervalos de confianza, lo que permite cuantificar la incertidumbre asociada a los resultados. La calidad de los pronósticos se evalúa mediante métricas de error como el Mean Absolute Error (MAE), la Root Mean Squared Error (RMSE) y el Mean Absolute Percentage Error (MAPE), ampliamente empleadas para

comparar el desempeño predictivo de modelos de series temporales (Hyndman & Athanasopoulos, 2018).

Figura 1

Metodología clásica Box-Jenkis



Nota. Adaptado de *Time series analysis: Forecasting and control* (3.ª ed.), por G. E. P. Box y G. M. Jenkins, 1976, Prentice-Hall.

4.2.4 Modelo Lag-Llama

El modelo Lag-Llama representa un enfoque innovador para el pronóstico probabilístico de series temporales. Fue desarrollado por Rasul et al. (2023) presentada como una solución a las limitaciones que surgen de los modelos tradicionales, como ARIMA o las redes neuronales recurrentes, los cuales suelen requerir entrenamiento específico para cada conjunto de datos. A diferencia de estos enfoques, Lag-Llama ha sido preentrenado sobre una amplia variedad de series temporales provenientes de múltiples fuentes, lo que le permite generalizar eficazmente y adaptarse a nuevos contextos sin necesidad de ajustes extensivos. Este modelo surge como

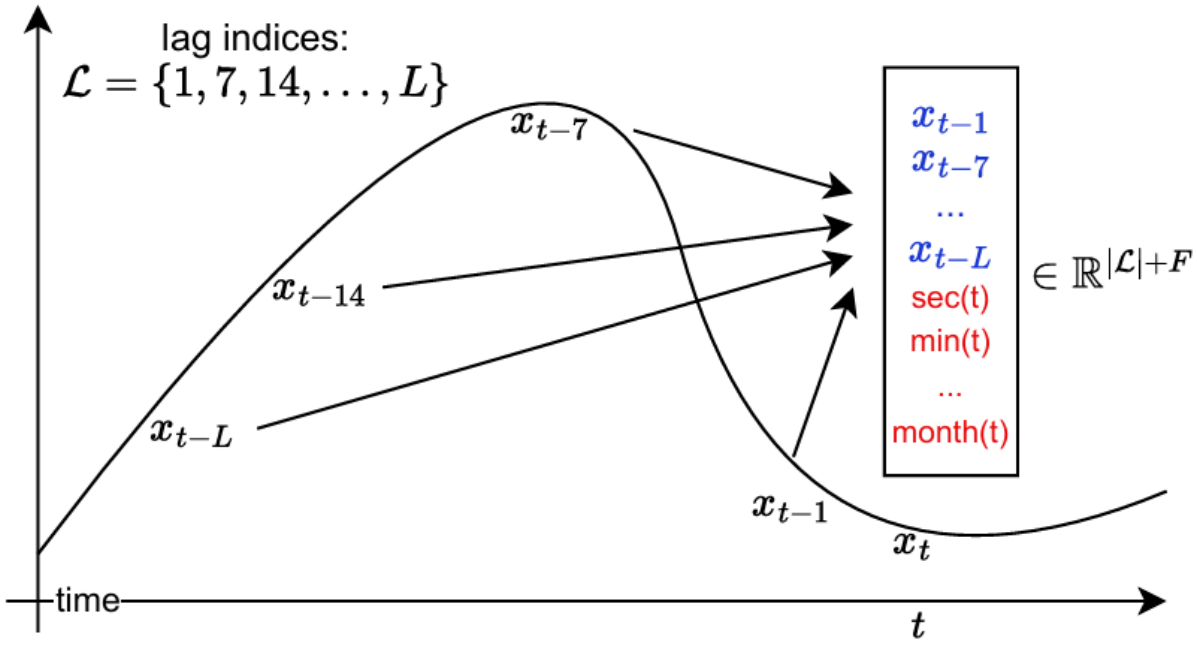
respuesta a la necesidad de contar con herramientas más flexibles y generalizables, especialmente frente a las limitaciones de modelos clásicos como ARIMA y enfoques basados en redes neuronales recurrentes (RNN).

Lag-Llama ha demostrado un rendimiento superior en la predicción de series univariadas, manteniendo su eficacia incluso ante conjuntos de datos no vistos previamente, lo cual lo convierte en una alternativa prometedora dentro del campo del modelado temporal.

4.2.4.1 Tokenización y Características de Rezago (Lag Features). Una de las particularidades de Lag-Llama radica en su proceso de tokenización, el cual se basa en la generación de características de rezago. Para ello, se define un conjunto de índices de rezago $\mathcal{L} = \{1,2,\dots,L\}$, que determina qué valores históricos serán utilizados. Cada observación actual x_t se transforma en un vector de tokens que incluye los valores pasados $x_{t-L}, \dots, x_{t-1}$, junto con información adicional como covariables temporales, como lo son la hora, día y mes del año. Este método permite al modelo adaptarse a diferentes frecuencias y patrones en los datos sin necesidad de depender de un conjunto de datos específico (Rasul et al., 2023).

Figura 2

Esquema de tokenización de Lag-Llama



Nota. Reproducido de *Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting*, por K. Rasul et al., 2023, arXiv. http://arxiv.org/abs/2310.08278

4.2.4.2. Arquitectura de Lag-Llama. El proceso comienza con una proyección lineal, que transforma los tokens resultantes de la tokenización en vectores densos de dimensión fija. Esta etapa asegura una representación homogénea de las series, incluso cuando provienen de dominios o escalas distintas, facilitando que las capas posteriores procesen la información de forma coherente.

A continuación, se incorpora la codificación posicional rotatoria (RoPE), cuyo propósito es preservar la información relativa de la posición en la secuencia. A diferencia de las codificaciones absolutas tradicionales, RoPE emplea rotaciones que mejoran la capacidad del modelo para reconocer tanto rezagos cercanos como lejanos. Esta característica resulta útil en

contextos económicos, donde los choques pasados pueden seguir influyendo en la dinámica actual (Su, Zhang, & Sennrich, 2021; Rasul et al., 2023).

Otro componente clave es la normalización de raíz media cuadrática (RMSNorm), aplicada bajo un esquema de pre-norm, es decir, antes de los bloques de atención. Esta decisión difiere de los transformers originales que aplicaban la normalización después (post-norm). El enfoque pre-norm contribuye a mantener una escala adecuada en los cálculos internos, estabilizando los gradientes y facilitando un entrenamiento más eficiente y con mejor convergencia (Zhang & Sennrich, 2019; Rasul et al., 2023).

El núcleo de la arquitectura lo constituyen los módulos de autoatención, organizados en capas de transformer. Gracias a este mecanismo, el modelo asigna diferentes ponderaciones a los rezagos, identificando automáticamente qué intervalos temporales aportan más información. Así, puede capturar tanto patrones estacionales recurrentes como variaciones irregulares, superando las limitaciones de modelos más rígidos como ARIMA.

Finalmente, Lag-Llama incluye una cabeza de distribución paramétrica, que en lugar de generar valores puntuales produce distribuciones probabilísticas basadas en la t de Student. Este enfoque no solo entrega un pronóstico central, sino también intervalos de confianza, incorporando explícitamente la incertidumbre en la predicción. Esto constituye una ventaja frente a los métodos deterministas, ya que proporciona información más rica y confiable para la toma de decisiones en escenarios económicos y financieros.

En conjunto, estos componentes dotan a Lag-Llama de una arquitectura robusta y flexible, capaz de aprender relaciones temporales complejas y adaptarse a distintos contextos. Por ello, se le reconoce como un modelo de vanguardia dentro del campo de los foundation models aplicados a series temporales (Bommasani et al., 2021; Rasul et al., 2023).

4.2.4.3 Modelado de la Distribución de Salida. A diferencia de otros modelos que producen predicciones puntuales, Lag-Llama genera una distribución de probabilidad sobre los posibles valores futuros. Para ello utiliza la distribución t de Student, cuyos parámetros (grados de libertad) permiten modelar tanto la tendencia central como la dispersión y la forma de la distribución. La función de probabilidad se

$$p(x_t|\mu_t, \sigma_t, v_t) = \frac{\Gamma\left(\frac{v_t+1}{2}\right)}{\sigma_t\sqrt{v_t\pi}\,\Gamma\left(\frac{v_t}{2}\right)}\left(1 + \frac{(x_t-\mu_t)^2}{v_t{\sigma_t}^2}\right)^{-\frac{v_t+1}{2}} \quad (4)$$

Este enfoque permite capturar la incertidumbre en la predicción y adaptarse a diferentes distribuciones de datos (Rasul et al., 2023).

4.2.4.4 Normalización de Valores y Estrategias de Preprocesamiento. Dado que las series de tiempo pueden presentar escalas y magnitudes muy diversas, es fundamental aplicar estrategias de normalización que permitan compararlas de manera adecuada. En Lag-Llama se emplean dos técnicas principales, la primera es el escalado estandarizado, en el que a cada ventana de datos se le resta la media y se divide por la desviación estándar, cómo se expresa en la Ecuación 5. La segunda, es la Normalización robusta basada en el rango intercuartil (IQR), en este caso, se centra la serie restando la mediana y se escala dividiendo por el IQR, como se observa en la Ecuación 6, donde $IQR(x_{1:C})$ es la diferencia entre el tercer y el primer cuartil de la ventana de datos.

$$x'_t = \frac{x_t - \mu}{\sigma}, \quad (5)$$

$$x'_t = \frac{x_t - Med(x_{1:C})}{IQR(x_{1:C})}, \quad (6)$$

Estas estrategias permiten al modelo trabajar de manera eficiente con series que varían en escala, rango y distribución sin necesidad de ajustes manuales en cada conjunto de datos (Rasul et al., 2023).

4.2.5 Métricas de evaluación

La evaluación de modelos de regresión constituye un paso clave en el desarrollo de sistemas predictivos, ya que permite estimar su capacidad de ajuste y su precisión al realizar pronósticos. Existen diversas métricas que cuantifican el error y valoran la calidad del modelo, aportando información esencial para interpretar su rendimiento en distintos contextos (Hyndman & Athanasopoulos, 2018). Estas herramientas no solo facilitan la comparación entre enfoques distintos, sino que también ayudan a detectar fortalezas y debilidades en el comportamiento predictivo de cada modelo.

4.2.5.1 Error Absoluto Medio (MAE). El Mean Absolute Error (MAE) cuantifica la magnitud promedio de los errores cometidos, sin tener en cuenta su dirección, lo cual lo convierte en una métrica fácil de interpretar. Su fórmula es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

Esta métrica resulta especialmente útil en aplicaciones donde los errores pequeños y grandes se consideran con igual importancia. No obstante, una de sus limitaciones es que no penaliza fuertemente los errores grandes, lo cual puede ser problemático en escenarios donde las desviaciones significativas tienen un mayor impacto (Chai & Draxler, 2014).

4.2.5.2 Error Cuadrático Medio (MSE) y Raíz del Error Cuadrático Medio (RMSE). El Error Cuadrático Medio (MSE) se define como el promedio de los errores elevados al cuadrado, calculados a partir de la diferencia entre los valores observados y los pronosticados:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

Al elevar los errores al cuadrado, el MSE penaliza con mayor severidad aquellos que son más grandes, lo cual puede ser deseable en contextos donde es necesario evitar grandes desviaciones.

Por otro lado, el Root Mean Squared Error (RMSE) se define como la raíz cuadrada del MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

El RMSE conserva las unidades originales de la variable objetivo, lo que facilita su interpretación en muchos casos prácticos. Su sensibilidad a los errores grandes lo hace útil cuando se busca minimizar este tipo de desviaciones, aunque puede ser menos intuitivo que el MAE (Hyndman & Athanasopoulos, 2018).

4.2.5.3 Coeficiente de Determinación (R^2). El coeficiente de determinación, conocido como R^2, cuantifica qué porcentaje de la varianza de la variable dependiente logra ser explicado por el modelo, en relación con la variabilidad total observada:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{10}$$

Cuando el estadístico R^2 se aproxima a 1, el modelo logra capturar la mayor parte de la variación de la variable analizada, mientras que los valores próximos a 0 indican que la proporción explicada es mínima. No obstante, es importante tener en cuenta que un R^2 alto no garantiza una buena capacidad de predicción, ya que un modelo sobreajustado puede obtener un valor elevado sin generalizar bien en nuevos datos (Chai & Draxler, 2014).

4.2.5.4 Error Porcentual Absoluto Medio (MAPE). El Mean Absolute Percentage Error (MAPE) expresa el error en términos relativos, facilitando la comparación entre modelos que operan sobre distintas escalas:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{11}$$

Debido a su interpretación como porcentaje, el MAPE es una métrica muy intuitiva y popular en entornos empresariales y financieros. Sin embargo, presenta limitaciones cuando los valores reales se aproximan a cero, ya que puede generar errores desproporcionadamente grandes o indefinidos. En tales casos, alternativas como el SMAPE (Symmetric Mean Absolute Percentage Error) son preferidas, ya que corrigen parte de esta inestabilidad (Hyndman & Athanasopoulos, 2018)

7.2.6 Python

Python es un lenguaje de programación de alto nivel que ha ganado gran popularidad en el ámbito académico y científico, especialmente en proyectos relacionados con análisis de datos, modelado predictivo y machine learning. Desde su creación por Guido van Rossum y el apoyo de otros desarrolladores, como Tim Peters, Python se ha consolidado como una herramienta esencial en la investigación aplicada, en parte gracias a su sintaxis sencilla y su versatilidad (Van Rossum & Drake, 2009). Este lenguaje se destaca por su amplia gama de bibliotecas especializadas, como pandas, desarrollada por McKinney (2017) ,que facilita la manipulación de datos, y otras similares como NumPy, scikit-learn y matplotlib, que enriquecen su funcionalidad en diversas áreas de análisis y modelado.

Entre sus bibliotecas más utilizadas, se encuentran NumPy para operaciones numéricas de alto rendimiento, pandas para la manipulación de datos tabulares y scikit-learn para la implementación de modelos de aprendizaje automático (Pedregosa et al., 2011). Estas

herramientas permiten abordar de manera eficiente tareas como el análisis exploratorio de datos, la construcción de modelos predictivos y la evaluación de resultados, lo que ha permitido que Python sea el lenguaje preferido en la ciencia de datos y en la investigación económica.

En cuanto a los entornos de desarrollo, uno de los más utilizados en el ámbito académico y de investigación es Jupyter Notebook, una herramienta interactiva que integra en un mismo entorno la escritura de código, la generación de gráficos y la inclusión de explicaciones en un mismo documento. Esta característica facilita la documentación del proceso analítico, favorece la reproducibilidad y mejora la interpretación de los resultados (Kluyver et al., 2016). Además, se destaca por su capacidad para ejecutar celdas de código de manera secuencial y permitir la visualización inmediata de los resultados, lo que lo convierte en una herramienta ideal para el análisis interactivo y exploratorio.

Además, para la gestión del código fuente y el trabajo colaborativo, se utiliza GitHub, una plataforma basada en el sistema de control de versiones Git. Permite almacenar el historial de cambios, organizar proyectos en repositorios y colaborar con otros investigadores o desarrolladores de forma remota. En contextos académicos, esta herramienta es especialmente valorada por su contribución a la transparencia, la trazabilidad del proceso analítico y la capacidad de realizar revisiones de código de manera eficiente (Perkel, 2016).

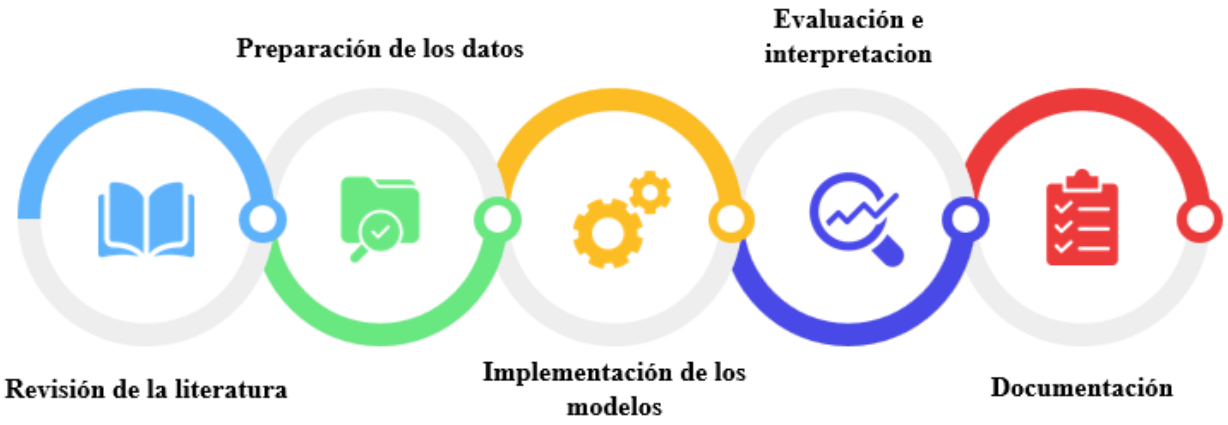
5. Metodología

Esta investigación se apoya conceptualmente en el enfoque Knowledge Discovery in Databases, KDD), el cual ofrece una estructura general para extraer información útil y relevante a partir de grandes volúmenes de datos (Maimon & Rokach, 2010). Dicho enfoque resulta

pertinente, tanto para organizar de manera organizada el proceso analítico, desde la preparación de los datos hasta la evaluación de los modelos predictivos. Con base en este marco, las fases metodológicas adoptadas en este trabajo se resumen en la Figura 2 y se describen a continuación.

Figura 3

Fases metodológicas del proyecto de investigación



Fase 1: Revisión de la literatura

En esta primera fase se identificaron los enfoques metodológicos y antecedentes aplicables a la predicción de la inflación en economías emergentes, considerando los modelos a utilizar, tanto clásico como ARIMA como enfoques recientes basados en aprendizaje automático y profundo. Para ello se diseñó una ecuación de búsqueda aplicada en Scopus. Los documentos seleccionados fueron analizados mediante herramientas bibliométricas de Scopus y VOSviewer, con el fin de identificar tendencias de publicación, áreas disciplinares, países y coocurrencias de palabras clave. Dada la reciente aparición de Lag-Llama y su limitada indexación en las bases de datos, se incluyeron fuentes complementarias como artículos técnicos y publicaciones en repositorios abiertos.

Fase 2: Preparación de los datos

En esta fase se recolectó la serie de inflación interanual para Colombia (DANE) y se verificó la calidad de los datos, confirmando la ausencia de valores faltantes o atípicos relevantes. Se estandarizaron formatos de fechas y valores, transformando la serie a frecuencia mensual (MS) y tipificándola en *float32* para garantizar consistencia y eficiencia en el procesamiento. Asimismo, se definió un enfoque univariado sin variables exógenas y se dividieron los datos en un conjunto de entrenamiento (2004-2023) y otro de validación y pronóstico (2024). Finalmente, se efectuó un análisis exploratorio que incluyó estadísticas descriptivas, descomposición estacional, correlogramas ACF/PACF y pruebas de estacionariedad, insumos fundamentales para orientar el modelado posterior.

Fase 3: Implementación de modelos

Tras la fase anterior, se procedió a la implementación de los modelos en Python 3.10 utilizando el entorno Jupyter Notebook, lo cual permitió la ejecución de los códigos, visualizaciones y documentación.

En el caso del modelo ARIMA, se aplicó la metodología Box-Jenkins, que comprende cuatro etapas fundamentales: identificación, estimación, diagnóstico y pronóstico. Para ello, se hizo uso de las librerías *pmdarima* y *statsmodels*, que facilitan tanto la automatización en la selección de parámetros como la validación de los supuestos del modelo.

Por otro lado, para la implementación de Lag-Llama se realizó sobre *PyTorch*, mediante la interfaz de *GluonTS*, inicializando el modelo a partir de un checkpoint disponible en Hugging Face Hub. Este procedimiento permitió evaluar el modelo bajo un esquema zero-shot, es decir, sin necesidad de un ajuste o entrenamiento adicional sobre los datos locales.

Para valorar su comportamiento en diferentes escenarios, se definieron dos esquemas de predicción: uno de tipo *Multi-step* con horizonte de 12 meses, que permite proyectar de forma

simultánea el conjunto completo de un año, y otro de tipo *Rollig-step,* en el cual se generan predicciones de un mes adelante que se van actualizando de manera progresiva conforme se incorporan nuevos datos. Esta dualidad facilitó comparar tanto la capacidad de pronóstico extendido como la de seguimiento continuo, aspectos relevantes para aplicaciones prácticas en política económica.

Fase 4: Evaluación e interpretación

En esta fase se definió la estrategia de evaluación de los modelos, la cual consistió en comparar su desempeño mediante métricas de error reconocidas en la literatura (MAE, RMSE y MAPE). La evaluación se aplicó en los dos esquemas de pronóstico ya mencionadas en la fase anterior, con el propósito de valorar tanto horizontes de predicción largos como escenarios de actualización secuencial mes a mes. Los resultados de estas comparaciones fueron posteriormente interpretados en función de su pertinencia para el análisis inflacionario en Colombia.

Fase 5: Documentación

En la etapa final, se elaboró un artículo académico publicable en el que se presentaron los principales hallazgos del proyecto. De manera complementaria, se consolidó un documento final que integró los resultados obtenidos, incluyendo los análisis exploratorios, la aplicación de los modelos y la comparación de desempeños.

Por último, se creó un repositorio público como parte de la estrategia de documentación y difusión, en el cual se dispuso el código de la modelación empleada, con el propósito de garantizar la transparencia, la reproducibilidad y el acceso abierto a la comunidad académica.

6. Fase 1: Revisión de la literatura

En esta fase se llevó a cabo una revisión de literatura enfocada en estudios relacionados con la predicción de la inflación en Colombia y en otras economías emergentes. El análisis contempló tanto modelos tradicionales de series temporales, como el ARIMA, como enfoques recientes de aprendizaje automático y profundo, con especial énfasis en Lag-Llama por su carácter innovador y su potencial de aplicación en pronósticos económicos.

Para estructurar el proceso, se inició con un análisis bibliométrico que permitió caracterizar la producción científica en torno al tema, identificar tendencias de publicación, principales países y áreas de investigación, así como visualizar las coocurrencias de palabras clave.

6.1 Análisis Bibliométrico

Se empleó la base de datos Scopus para obtener la información necesaria para el análisis bibliométrico, debido a su amplia disponibilidad de literatura científica.

6.1.1 Análisis Bibliométrico inicial

Esta estrategia permitió abarcar tanto el enfoque tradicional representado por ARIMA como los modelos más modernos de predicción basados en Machine Learning y Deep Learning. La fórmula de búsqueda utilizada fue la siguiente:

TITLE-ABS-KEY (econom*) AND TITLE-ABS-KEY (inflation) AND TITLE-ABS-KEY ("Autoregressive Integrated Moving Average" OR arima) AND TITLE-ABS-KEY ("Lag-Llama" OR "Long Short-Term Memory" OR LSTM OR "Deep Learning" OR "Machine Learning") AND TITLE-ABS-KEY (forecast* OR projection* OR estimation*)

Esta fórmula se construyó mediante el uso de palabras clave conectadas con operadores como “AND” y “OR”, permitiendo ampliar los criterios de búsqueda. Además, se emplearon términos como "econom*", para incluir variantes de una misma raíz, como economy, economic o

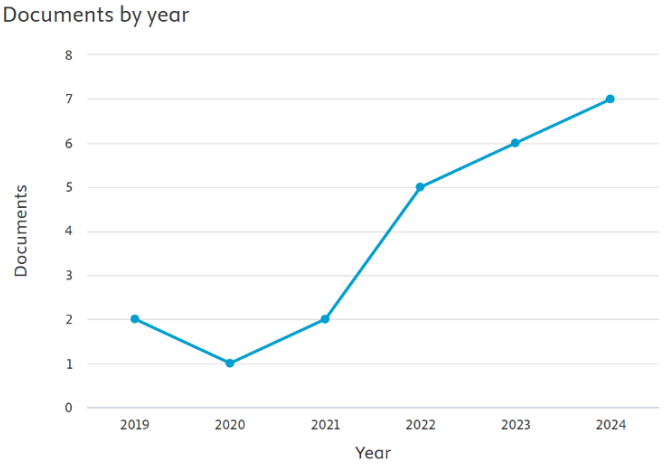
economics. También se incluyeron términos relacionados con el pronóstico, como forecast, projection y estimation, para delimitar la búsqueda.

Como resultado, se identificaron 23 documentos hasta el año 2024. Para el análisis de esta producción científica se utilizaron las herramientas bibliométricas integradas en Scopus, complementadas con el software VOSviewer.

6.1.1.1 Publicaciones por año. La producción científica relacionada con la investigación inició en 2019, con la publicación de dos documentos, lo que representó el 8,7 % del total. En 2020 se registró una disminución, con solo un documento publicado (4,3 %). Posteriormente, en 2021, se publicaron dos documentos, equivalentes al 8,7 %. A partir de 2022 se evidenció un crecimiento, registrándose cinco documentos, que representaron el 21,7 % del total anual. En 2023 el número ascendió a seis documentos, correspondientes al 26,1 %, y en 2024 se alcanzó el punto más alto de producción con siete documentos, equivalentes al 30,4%. Este comportamiento, ilustrado en la Figura 1, refleja un interés creciente en la temática de estudio, especialmente a partir de 2022, consolidándose 2024 como el año de mayor actividad científica en el área.

Figura 4

Producción científica publicada por año



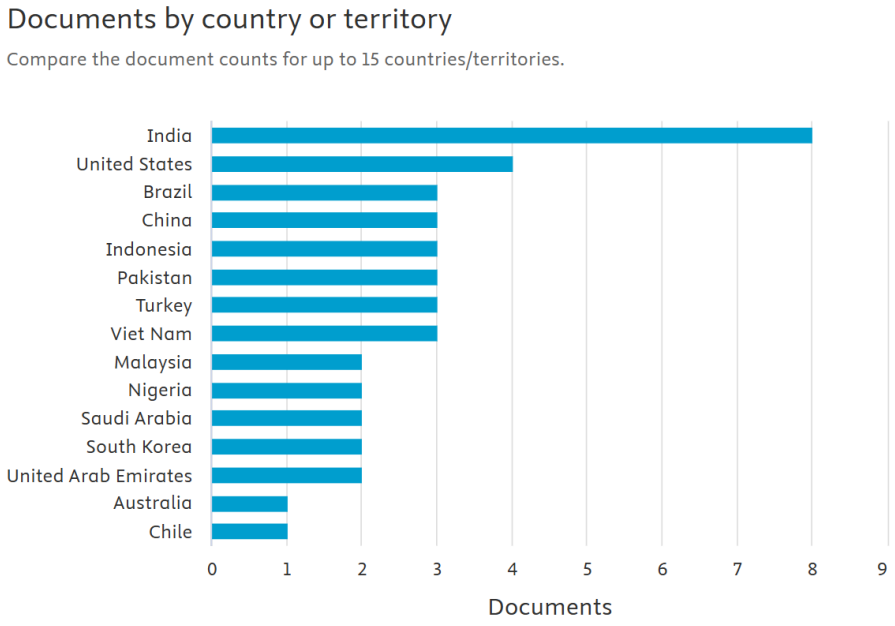
Nota. Reproducido de Scopus (2024), a partir de la ecuación de búsqueda utilizada.

6.1.1.2 Países / Regiones. La Figura 2 muestra la distribución de la producción científica por país o territorio. Estados Unidos lidera con cinco documentos, representando aproximadamente el 21,7 % del total. Le sigue India con cuatro documentos, equivalentes al 17,4 %. Turquía y Vietnam comparten la tercera posición, cada uno con tres documentos (13 %). China contribuye con dos documentos (8,7 %), mientras que Bangladesh, Brasil, Chile, Hungría e Indonesia participan con un documento cada uno, representando individualmente el 4,3 % de la producción.

Es importante destacar que Brasil y Chile son los únicos países latinoamericanos presentes en el ranking, lo cual resulta relevante para esta investigación. Su participación científica puede aportar perspectivas metodológicas al contexto colombiano, fortaleciendo el marco teórico del estudio.

Figura 5

Producción científica por país

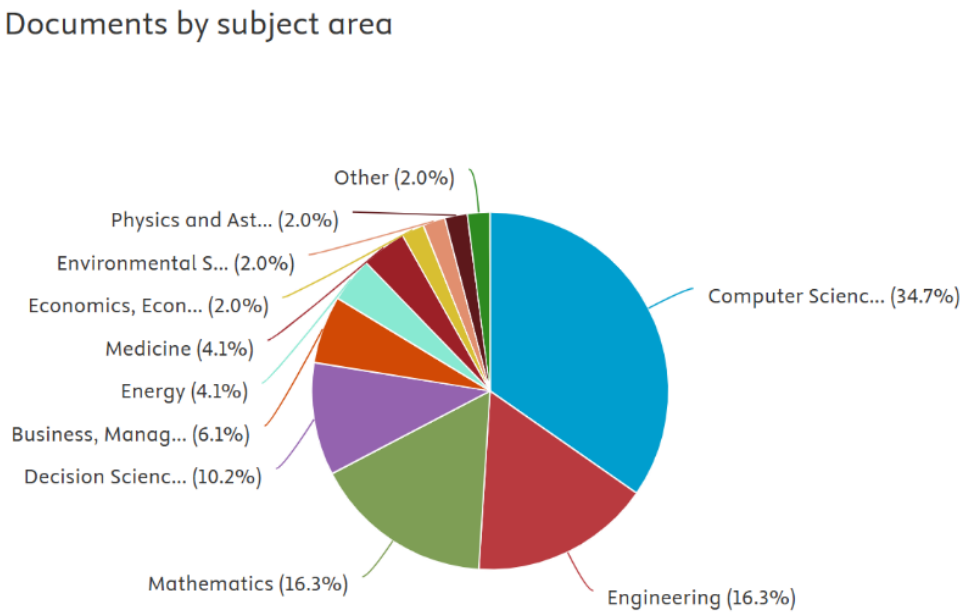


Nota. Reproducido de Scopus (2024), a partir de la ecuación de búsqueda utilizada.

6.1.1.3 Áreas de investigación. La Figura 4 permite visualizar que la gran parte de las investigaciones proviene del sector de *Ciencias de la Computación* (34.7%), seguido por Ingeniería (16.3%), Matemáticas (16.3 %) y *Ciencias de la Decisión* (10.2 %), lo que refleja un enfoque técnico y analítico en la temática. A demás, áreas como *Negocios y Gestión* (6.1%) también representan un porcentaje significativo, destacando su relación con el objeto de estudio.

Figura 6

Producción científica por área de investigación



Nota. Reproducido de Scopus (2024), a partir de la ecuación de búsqueda utilizada.

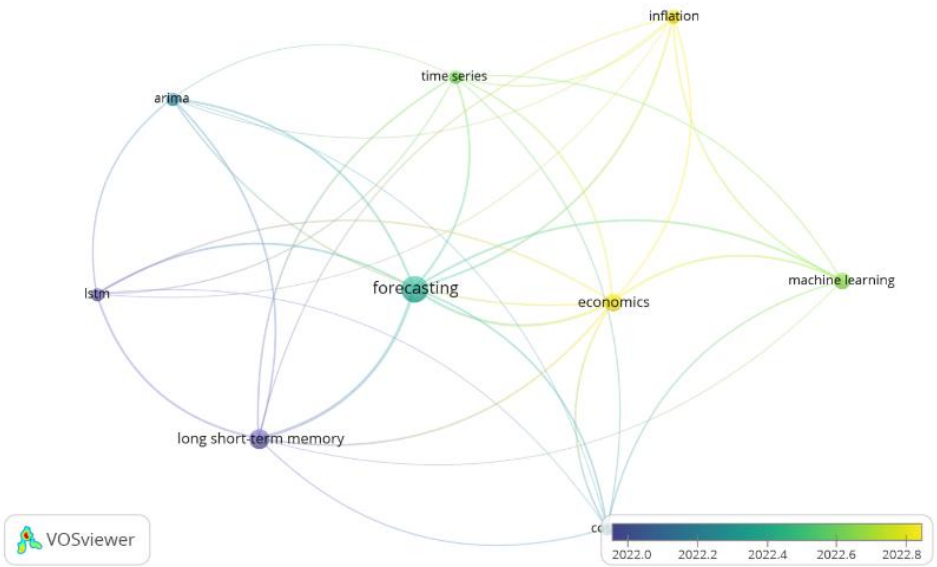
6.1.1.4 Palabras clave. El mapa de red, ilustrado en la Figura 4, representa las relaciones entre las palabras clave identificadas en el proceso de búsqueda bibliográfica. Se observa que "forecasting" actúa como el nodo central del análisis, estableciendo vínculos con términos relevantes como "time series", "economics", "inflation" y "machine learning". Esto refleja la importancia del concepto de pronóstico en el contexto investigativo actual.

Asimismo, se destaca que los modelos "ARIMA" y "long short-term memory (LSTM)" presentan conexiones directas con "forecasting", lo cual evidencia su rol fundamental en los enfoques de predicción económica y en el análisis comparativo de modelos. Cabe señalar que "machine learning" también emerge como un término relevante, indicando un creciente interés por la aplicación de técnicas de aprendizaje automático en la predicción de variables económicas como la inflación.

Finalmente, la escala de colores aplicada en el mapa sugiere una evolución temporal de los estudios, donde los temas más recientes (representados en tonos más claros) se concentran en la intersección entre predicción, aprendizaje automático y fenómenos económicos.

Figura 7

Mapa de red por palabras clave



Nota. Gráfico adaptado en VOSviewer a partir de la ecuación de búsqueda utilizada.

6.1.2 Análisis bibliométrico específico

Considerando la ausencia de estudios centrados en Lag-Llama en la búsqueda inicial, se diseñó una segunda estrategia de búsqueda orientada exclusivamente a esta técnica emergente. La fórmula aplicada fue la siguiente:

TITLE-ABS-KEY ("Lag-Llama" OR "Lag Llama") AND TITLE-ABS-KEY (forecast*)

Esta búsqueda se estructura únicamente con el término forecast, con el fin de identificar publicaciones que mencionen explícitamente el modelo Lag-Llama en contextos de pronóstico, independiente del área de estidio.

El resultado de esta búsqueda fue de 6 documentos, reflejando la limitada visibilidad académica de este modelo, probablemente atribuible a su reciente desarrollo. En la Tabla 1, sintetiza la información recopilada a partir de esta búsqueda específica.

Tabla 2

Documentos por autores, áreas de investigación y país.

Autor(es)	País	Área de investigación	N° de documentos
Ali M.; Moore P.W.; Barkouki T.; Brattain L.J.	Estados Unidos	Computer Science, Engineering, Medicine, Physics and Astronomy	1
Sun Y.; Cheng B.; Li J.; Jiang G.; Zhang T.; Zhou H.	China	Computer Science	1
Saravanan H.K.; Dwivedi S.; Praveen P.; Arjunan P.	India	Computer Science, Engineering, Energy	2
Gupta D.; Bhatti A.; Parmar S.	Canadá	Computer Science	1

Amjad F.; Korotko T.; Rosin A.	Estonia	Computer Science, Engineering, Energy, Mathematics, Social Sciences	1

Nota. Elaboración propia a partir de datos extraídos de Scopus.

Dado que el modelo Lag-Llama es una propuesta reciente, su principal publicación científica aún no se encuentra indexada en bases como Scopus. Por tal motivo, se recurrió a fuentes complementarias, como el repositorio arXiv, donde Rasul et al. (2023) publicaron el primer artículo de este modelo titulado "Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting".

Este análisis refuerza la justificación de este trabajo de investigación, al evidenciar la necesidad de explorar y evaluar el desempeño de Lag-Llama en comparación con modelos tradicionales como ARIMA, dada su baja representación en la literatura académica actual.

6.2. Análisis preliminar de la literatura

A continuación, se presentan trabajos recientes que comparan modelos estadísticos tradicionales como ARIMA y enfoques más modernos basados en aprendizaje automático:

Chuku et al. (2019) analizaron si los métodos de inteligencia computacional podían superar a los modelos tradicionales en la predicción del crecimiento económico de Kenia, Nigeria y Sudáfrica. Para ello compararon ARIMA y un modelo estructural con dos enfoques no lineales, el de regresión no paramétrica y redes neuronales artificiales (ANN), empleando datos trimestrales de 1970 a 2016 e incorporando covariables como inflación, tasas de interés, comercio exterior y precios de materias primas. La validación incluyó pronósticos dentro y fuera de muestra, también métricas de error como RMSE y SMAPE, y la prueba de Diebold-Mariano para contrastar diferencias significativas. Los resultados mostraron que los métodos no lineales superaron a los enfoques lineales, especialmente al integrar variables macroeconómicas, reduciendo los errores de

pronóstico. Los autores destacan que la capacidad de ANN y de la regresión no paramétrica para capturar dinámicas complejas y no lineales son más adecuadas en entornos volátiles. Este hallazgo resulta relevante para la presente investigación, pues refuerza la idea de que el pronóstico económico se beneficia de modelos más flexibles, lo que conecta directamente con el potencial de enfoques recientes como Lag-Llama, diseñados para superar las limitaciones métodos convencionales.

De manera similar, Özgür y Akkoç (2022) pronosticaron la inflación mensual en Turquía, el cual es un caso de economía emergente con alta inflación, mediante algoritmos de Machine Learning con métodos de contracción (shrinkage) para la selección de variables, en comparación con enfoques tradicionales como ARIMA y VAR. La base de datos abarca 229 variables macroeconómicas, entre ellas, tipos de cambio, indicadores bursátiles, cifras fiscales y de balanza de pagos, índices de producción y comercio, tasas del mercado monetario, precios, entre otros. A diferencia de los modelos econométricos convencionales, las técnicas *LASSO* y *Elastic Net*, seleccionan de forma automática las variables más relevantes, lo que hace el modelo más sencillo sin perder precisión. Con estos últimos se alcanzaron los menores errores de pronóstico (RMSE) frente a ARIMA, VAR y un modelo de referencia como *Prophet*, Significativamente, el modelo lasso redujo las entradas a solo ocho variables significativas de las 229 posibles, demostrando el valor añadido de la reducción de dimensionalidad que es simplificar el modelo sin sacrificar exactitud predictiva. En un contexto emergente como el turco, caracterizado por inflación alta y volátil, este enfoque aporta importantes avances metodológicos, es la primera aplicación de técnicas de machine learning para el pronóstico inflacionario de Turquía, mejorando sustancialmente la calidad de las previsiones respecto a los métodos tradicionales.

En otro contexto, Peirano et al. (2021) proponen un modelo híbrido que combina SARIMA y LSTM con el fin de mejorar la predicción de la inflación en cinco economías latinoamericanas emergentes: Brasil, México, Chile, Colombia y Perú. El planteamiento se justifica en que la inflación presenta tanto componentes lineales y estacionales, que pueden ser capturados por SARIMA, como relaciones no lineales derivadas de choques económicos y cambios de régimen, que son más adecuadamente modeladas por las redes LSTM. Para validarlo, los autores utilizaron series mensuales de inflación comprendidas entre 1958 y 2019, dividiendo los datos en conjuntos de entrenamiento y prueba mediante un esquema de ventanas deslizantes con horizonte de un mes. El modelo híbrido fue contrastado frente a SARIMA y LSTM por separado, así como contra otros enfoques de referencia como redes neuronales artificiales, sistemas de inferencia difusa, modelos neuro-difusos y un híbrido SARIMA-ANN. Los resultados evidencian que la combinación SARIMA-LSTM logra en promedio la mayor precisión predictiva, reduciendo significativamente el error de pronóstico en México, Colombia y Perú, y mostrando también mejoras, aunque no estadísticamente concluyentes, en Brasil y Chile. En conjunto, el estudio demuestra que la integración de métodos lineales y no lineales potencia la capacidad de capturar tanto la estructura estacional como las dinámicas complejas de la inflación, ofreciendo pronósticos más robustos en contextos de alta volatilidad macroeconómica.

Por su parte, Adyatma y Alamsyah (2022) realizaron un estudio enfocado en pronosticar el Índice Compuesto de la Bolsa de Indonesia (IDX Composite) utilizando variables macroeconómicas clave como la inflación, la oferta monetaria y el tipo de cambio, esto a raíz de la relación estrecha de estas con las fluctuaciones del mercado bursátil. Por ello, los autores compararon un modelo clásico ARIMA versus enfoques basados en redes neuronales (una red neuronal artificial, ANN, y una red de memoria a corto y largo plazo, Long Short-Term Memory

o LSTM) para determinar cuál ofrecía una mayor precisión predictiva sobre el índice. Los resultados mostraron que el modelo LSTM alcanzó el menor error cuadrático medio (MSE) en las predicciones, seguido por la red neuronal ANN, y finalmente el modelo ARIMA resultó con el error más elevado. Este resultado resalta la importancia de emplear modelos avanzados para mejorar el desempeño predictivo, especialmente en series de tiempo de corto plazo con comportamiento marcadamente no lineal, donde los enfoques tradicionales tienden a ser insuficientes.

Más recientemente, Ismail et al. (2023) llevaron a cabo una comparación métodos de pronóstico de la inflación a partir del IPC de Bangladesh, evaluando tanto modelos tradicionales (ARIMA junto con su variante estacional SARIMA) como algoritmos modernos de aprendizaje automático (incluyendo *Prophet*, SVR, redes neuronales tipo LSTM y GRU, además de un esquema AutoML basado en *XGBoost*). La precisión de cada modelo se evaluó mediante métricas de error comunes MAE, RMSE y MAPE aplicando además validación cruzada para asegurar la robustez del análisis comparativo. Resultados: Los experimentos mostraron que la red neuronal LSTM y el modelo automatizado alcanzaron mejor pronóstico de la inflación. Sin embargo, el estudio resalta que el desempeño superior de estos modelos avanzados dependía en gran medida de la calidad y cantidad de datos disponibles, evidenciando que con datos limitados o poco fiables su ventaja se reducía. Estos hallazgos subrayan la necesidad de contar con datos confiables para aprovechar plenamente las ventajas de modelos complejos, al tiempo que destacan la pertinencia de emplear enfoques de machine learning en economías emergentes como Bangladesh, donde la dinámica inflacionaria responde a factores estructurales particulares

Asimismo, Xavier et al. (2023) compararon trece enfoques de pronóstico, incluyendo modelos lineales tradicionales como ARIMA y ETS, técnicas de aprendizaje automático como

SVR y redes neuronales MLP, el modelo de pronóstico *Prophet*, entre otros, con el fin de evaluar su eficacia en la proyección de la inflación brasileña. Los resultados indicaron que si bien ETS demostró un buen rendimiento individual en ciertos indicadores (por ejemplo, obtuvo uno de los menores errores porcentuales), el modelo híbrido propuesto por los autores (ETS+Bagging) logró reducir de forma significativa el error de predicción frente a cualquiera de los modelos individuales evaluados. La metodología empleada incluyó el uso de técnicas de ensemble (combinación de múltiples modelos) y la validación cruzada para evaluar la robustez del enfoque y garantizar que la mejora en la precisión no se debiera a un sobreajuste a los datos de entrenamiento. Como resultado, este enfoque híbrido aporta una mayor estabilidad y confiabilidad al pronóstico de la inflación en entornos económicos complejos como el de Brasil, al aprovechar las fortalezas complementarias de modelos lineales y no lineales para capturar patrones diversos en la dinámica inflacionaria.

Lakshmi Narayanaa et al. (2023) exploraron la predicción de la inflación empleando un modelo estadístico tradicional, ARIMA, frente a una red neuronal profunda, LSTM. Este análisis realizado sobre la serie temporal de la inflación mensual del Reino Unido correspondiente al período 1947-2021, reveló que el modelo ARIMA mostró un mejor desempeño en pronósticos de corto plazo, mientras que el modelo LSTM alcanzó mayor precisión en predicciones a mediano plazo. Las diferencias observadas son a raíz de la naturaleza de cada enfoque, los datos mensuales con oscilaciones breves favorecen a ARIMA, que captura con eficacia patrones transitorios de corto alcance, en tanto que la arquitectura LSTM aprovecha su capacidad de memoria para reflejar tendencias subyacentes de más larga duración. Por último, los autores sugieren que futuros trabajos consideren modelos multivariantes, que se incorporen indicadores económicos adicionales y

enfoques híbridos que combinen técnicas tradicionales con redes neuronales, como vías promisorias para mejorar la precisión de las proyecciones de inflación.

Rasul et al. (2023) introdujeron el primer estudio sobre el modelo Lag-Llama referenciado como modelo fundacional generalista para la predicción probabilística de series temporales univariadas. Su arquitectura se basa en un *transformer decoder-only* que utiliza rezagos (*lags*) como covariables y fue preentrenado sobre un amplio conjunto de series de distintos dominios. Este diseño le otorga una notable capacidad de generalización: en escenarios zero-shot, el modelo logra pronósticos competitivos sin necesidad de entrenamiento adicional, y cuando se somete a ajuste fino (*fine-tuning*) alcanza desempeños de nivel estado del arte, superando tanto a modelos estadísticos clásicos como ARIMA y a enfoques profundos como LSTM. Estos resultados muestran que Lag-Llama combina eficiencia computacional con flexibilidad, posicionándose como un candidato prometedor para tareas complejas como la predicción de la inflación, donde los métodos convencionales suelen enfrentar limitaciones para capturar dinámicas no lineales o abruptas.

No obstante, los propios autores reconocen limitaciones y desafíos. En su versión actual, Lag-Llama está restringido a la predicción univariada, lo que limita su capacidad para capturar relaciones entre múltiples variables económicas. Además, su rendimiento puede verse comprometido en dominios muy distintos a los incluidos en el entrenamiento, lo que obliga a disponer de suficientes datos históricos para lograr una buena adaptación. Rasul et al. (2023) también destacan que el acceso a corpus diversos y de gran tamaño es todavía un obstáculo en la construcción de modelos fundacionales para series de tiempo, y sugieren la necesidad de ampliar la escala del modelo y explorar variantes multivariantes en futuras investigaciones.

Después de la publicación del trabajo base, Saravanan et al. (2024) evaluaron el potencial de los Modelos de Fundación para Series Temporales (TSFMs) en la predicción de carga energética a corto plazo, un campo con alta volatilidad y patrones estacionales marcados. El estudio comparó modelos de última generación como Chronos y Lag-Llama con métodos tradicionales como *LightGBM* y transformadores entrenados desde cero. Los resultados mostraron que Chronos sobresalió en escenarios zero-shot, evidenciando la capacidad de un modelo fundacional para transferir aprendizajes entre dominios y aplicarlos en contextos energéticos sin entrenamiento adicional. En contraste, Lag-Llama presentó un desempeño más moderado, por debajo de Chronos pero aún competitivo frente a otros enfoques. No obstante, el análisis señaló que la capacidad de generalización de Lag-Llama resulta más limitada y depende en mayor medida del fine-tuning para alcanzar su máximo potencial. Los autores destacan que, aunque estos modelos fundacionales representan un avance significativo frente a métodos convencionales, todavía enfrentan el desafío de adaptarse con mayor eficiencia a dominios específicos, lo que sugiere la necesidad de incorporar variables externas para mejorar la precisión.

Por último, tenemos el análisis de Ali et al. (2024) en el que valuaron tres modelos fundacionales Chronos, TimesFM y Lag-Llama, en la predicción zero-shot de señales ECG. Los hallazgos evidenciaron que Lag-Llama obtuvo el desempeño más bajo, sin superar siquiera el método ingenuo de proyectar el valor anterior, además de presentar un elevado costo computacional: cada predicción tomó en promedio 609 ms frente a solo 0.05 ms del LSTM clásico. Esta lentitud lo hace poco viable en series de alta frecuencia como las biomédicas. Asimismo, se observó que sin procesos de ajuste, el modelo tiende a sobregeneralizar, convergiendo hacia la media y perdiendo capacidad para reflejar cambios abruptos. Aunque el ajuste fino mejoró significativamente a otros modelos como Chronos, en Lag-Llama no corrigió los errores e incluso

mostró riesgos de sobreentrenamiento. Estas limitaciones permiten anticipar implicaciones en contextos económicos: aplicado directamente a series de inflación, podría generar pronósticos centrados en la media histórica, subestimando choques o variaciones extremas, además de requerir altos recursos computacionales y una calibración cuidadosa para lograr mayor precisión.

En conjunto, la literatura revisada muestra una línea desde modelos estadísticos tradicionales hacia enfoques híbridos, de aprendizaje profundo y fundacionales, cada uno con aportes y limitaciones. Mientras los métodos clásicos como ARIMA resultan útiles para patrones lineales de corto plazo, los enfoques basados en machine learning y modelos fundacionales destacan por su capacidad de capturar dinámicas complejas y generalizar a nuevos contextos. Sin embargo, en el caso particular de la inflación en Colombia, aún no se han reportado aplicaciones con modelos de última generación como Lag-Llama, lo que refuerza la pertinencia de la presente investigación.

7. Fase 2: Preparación de los datos

7.1 Recolección y descarga de los datos

La fuente principal de información utilizada en este estudio corresponde a la serie histórica de la tasa de inflación interanual en Colombia, publicada por el Departamento Administrativo Nacional de Estadística (DANE) en su portal web oficial. Para acceder a la información, se empleó la herramienta de consulta dinámica que permite seleccionar rangos temporales específicos y exportar la información en distintos formatos. En este caso, se seleccionó el periodo comprendido entre enero de 2004 y diciembre de 2024, abarcando más de dos décadas de datos, lo que garantiza la inclusión de diferentes ciclos económicos, periodos de estabilidad, episodios inflacionarios y

coyunturas internacionales que afectan el comportamiento de los precios en el país. Esta amplitud temporal ofrece una base sólida para el análisis al permitir que los modelos capturen patrones dinámicos.

La exportación de los datos se realizó en formato Excel (.xlsx) el 15 de junio de 2024, con el fin de asegurar la trazabilidad y replicabilidad del proceso, el cual se encuentra disponible en el Apéndice A.

7.2 Preparación de los datos

El proceso de preparación y estandarización de la serie se desarrolló en varias etapas:

7.2.1 Revisión inicial de la base de datos

Se verificó la ausencia de valores faltantes, duplicados o inconsistencias, y no se identificaron valores atípicos relevantes. Por este motivo, no fue necesario aplicar procesos de imputación o depuración.

7.2.2 Estandarización de formatos

Se ajustaron los formatos decimales y de fechas para garantizar compatibilidad con el lenguaje de programación Python y con las funciones de análisis de series temporales.

La serie se transformó a una frecuencia mensual con inicio de mes (MS), lo que aseguró la consistencia temporal de los registros.

Los valores fueron tipificados al formato float32, optimizando el uso de memoria y mejorando la eficiencia de los cálculos sin sacrificar precisión.

7.2.3 Definición del diseño metodológico

Se adoptó un enfoque univariado, sin variables exógenas, con el fin de garantizar condiciones comparables entre los modelos en estudio. Con esta decisión se buscó que la

evaluación del desempeño se centrara exclusivamente en la capacidad predictiva de cada modelo sobre la serie de inflación.

7.2.4 División de los datos

La muestra se dividió en dos periodos: el primero, comprendido entre 2004 y 2023, se destinó al entrenamiento de los modelos, mientras que el año 2024 se reservó para la validación y el pronóstico fuera de muestra. Esta estrategia permitió evaluar la eficacia de los modelos en un horizonte reciente y relevante, así como comprobar su capacidad de generalización frente a información nueva.

7.3 Análisis exploratorio

Antes de proceder al modelado, se llevó a cabo un análisis exploratorio de la serie temporal de inflación mensual en Colombia con el propósito de caracterizar su comportamiento histórico, identificar patrones relevantes y establecer insumos que orienten la selección de modelos adecuados.

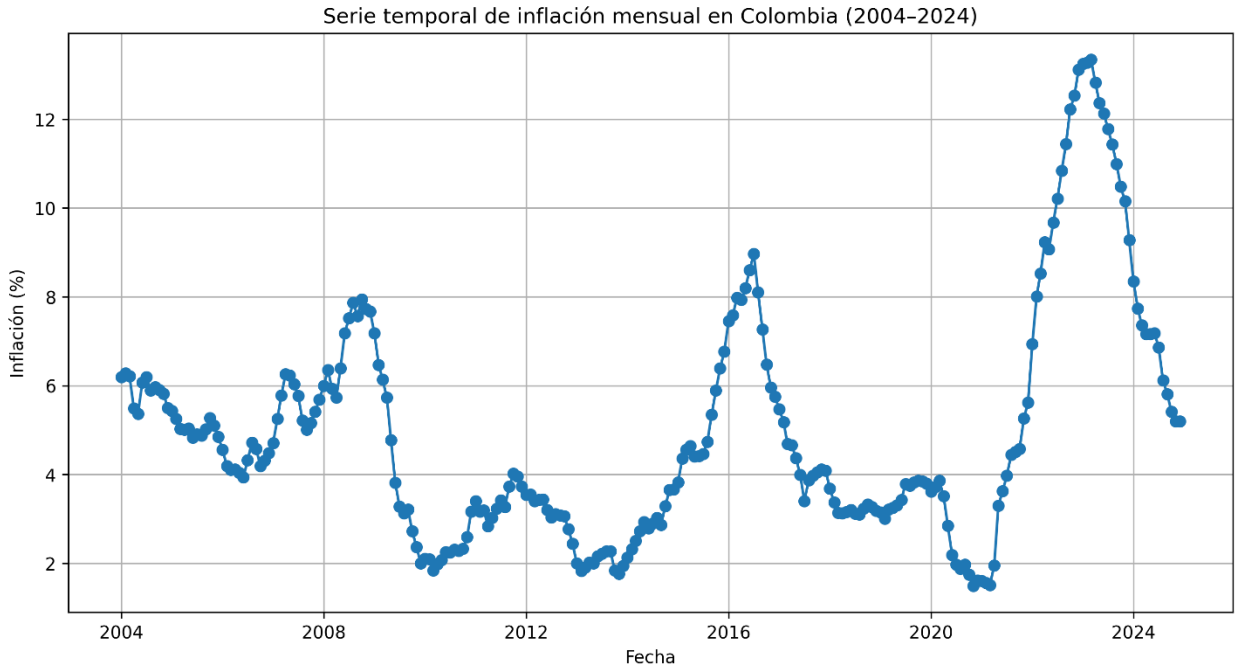
7.3.1 Visualización de la serie temporal

La Figura 8 muestra la evolución mensual de la inflación en Colombia durante el periodo 2004-2024. A lo largo de la serie no se distingue una tendencia única, sino fases con comportamientos diferenciados. Entre 2009 y 2014, por ejemplo, se observa un lapso de relativa estabilidad con niveles moderados de inflación. En contraste, en 2016 se registra un incremento abrupto, probablemente vinculado a factores externos e internos, entre ellos la depreciación del peso frente al dólar. A partir de 2020, los efectos derivados de la pandemia y las alteraciones en las cadenas globales de suministro dieron lugar a una dinámica más inestable, con una trayectoria creciente. También se aprecia que la magnitud de las fluctuaciones no es uniforme: en algunos años los cambios fueron contenidos, mientras que en otros alcanzaron oscilaciones marcadas. Este

comportamiento evidencia que la serie no es estacionaria y que, para su modelado, será necesario aplicar procesos de diferenciación.

Figura 8

Serie temporal de la inflación interanual en Colombia (2004–2024)



7.3.2 Cálculo de estadísticas descriptivas

La Tabla 3 resume las principales medidas descriptivas de la serie. En total se analizaron 252 datos correspondientes a observaciones mensuales. El promedio de inflación en el período fue 4,95 %, lo que indica que en términos generales los precios se mantuvieron en niveles moderados frente al rango meta del Banco de la República. El valor mínimo registrado fue de 1,49 %, mientras que el máximo alcanzó 13,34 %, reflejando episodios de choques económicos tanto internos como externos. La desviación estándar de 2,62 % muestra que la serie presenta una variabilidad considerable, lo que confirma que no se trata de un fenómeno estable. Los percentiles indican que el 25 % de los valores estuvo por debajo de 3,17 % y el 75 % por debajo de 6,04 %, de modo que

la mayor parte de las observaciones se concentran en niveles intermedios, aunque con episodios de inflación alta en años puntuales.

Tabla 3

Estadísticas descriptivas de la inflación interanual en Colombia (2004–2024)

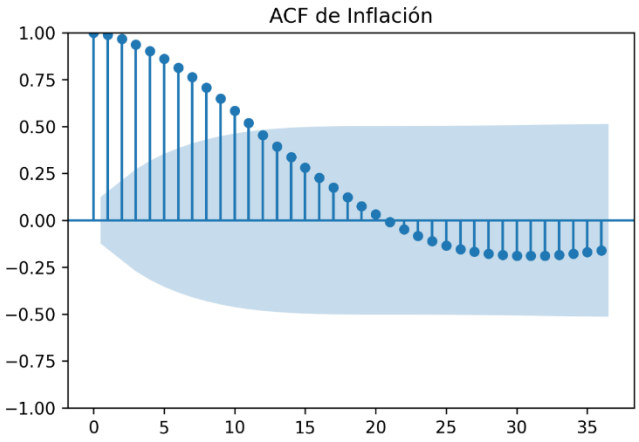
Medida	Valor
Número de datos	252
Promedio	4,95%
Desviación estándar (%)	2,62%
Mínimo	1,49%
Percentil 25	3,17
Mediana	4,34
Percentil 75 (%)	6,04
Máximo	13,34

7.3.3 Gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF)

La función de autocorrelación (ACF) en la Figura 9, de la serie de inflación muestra un descenso lento y sostenido a medida que aumentan los rezagos. Este comportamiento refleja que la inflación depende en gran medida de sus propios valores pasados, lo que resulta lógico en variables macroeconómicas que suelen presentar persistencia. La permanencia de la autocorrelación en varios rezagos indica que la serie no es estacionaria, ya que los efectos de los choques no se disipan con rapidez, sino que permanecen en el tiempo.

Figura 9

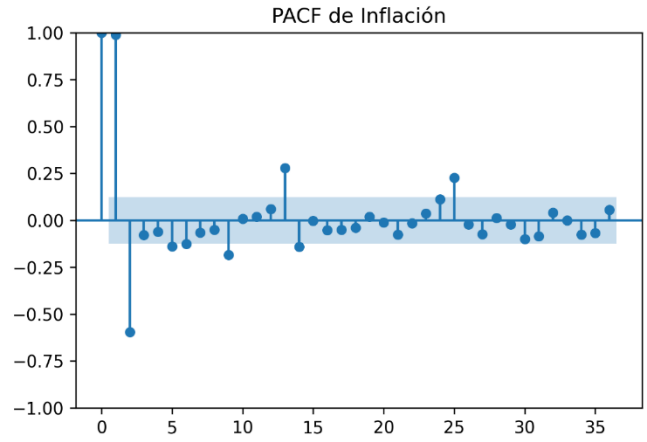
Función de autocorrelación (ACF) de la inflación interanual en Colombia



Por su parte, la función de autocorrelación parcial (PACF) en la Figura 10, evidencia que los rezagos uno y dos presentan valores significativos, mientras que en el tercer rezago aparece un valor negativo, seguido de correlaciones parciales pequeñas y sin relevancia estadística en los rezagos posteriores. Este patrón sugiere que la dinámica de la inflación puede explicarse en buena medida con los primeros rezagos, lo que orienta a considerar modelos autorregresivos de bajo orden, como un AR(1) o un AR(2), como punto de partida para la modelación.

Figura 10

Función de autocorrelación parcial (PACF) de la inflación mensual en Colombia



7.3.4 Descomposición de la serie

La Figura 11 presenta la descomposición clásica de la inflación mensual en Colombia entre 2004 y 2024, diferenciando tres componentes: tendencia, estacionalidad y residuos.

En la tendencia, se aprecian ciclos de largo plazo con repuntes claros en 2008, 2016 y 2022. Estos episodios coinciden con momentos de tensión económica, como la crisis financiera internacional, choques de oferta en alimentos y energía, y más recientemente los efectos de la pandemia y la presión inflacionaria global. También se observan fases de moderación, como entre 2010 y 2013, que reflejan etapas de relativa estabilidad en los precios. Esto sugiere que la inflación en Colombia no responde únicamente a dinámicas monetarias inmediatas, sino que incorpora factores estructurales y coyunturales que generan oscilaciones más prolongadas.

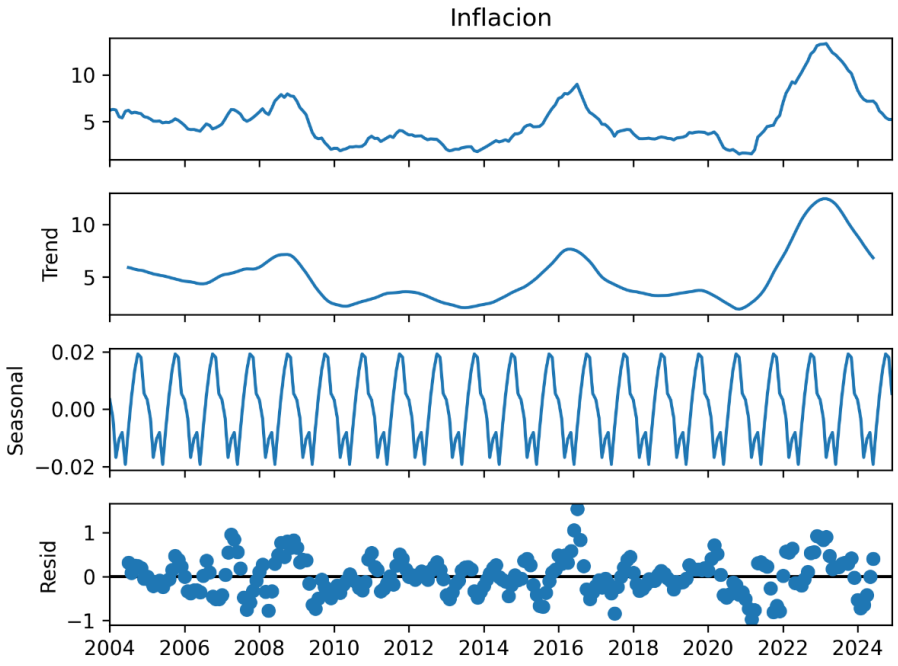
El componente estacional revela patrones recurrentes que se repiten de forma anual. Entre ellos destacan los incrementos en alimentos en épocas de baja cosecha y el aumento del consumo de los hogares hacia fin de año. Esta periodicidad muestra que la inflación tiene elementos predecibles que conviene considerar en modelos como ARIMA o SARIMA, diseñados para capturar estacionalidad.

En cuanto al residual, este concentra la variación no explicada por la tendencia ni la estacionalidad. La mayor parte se agrupa en torno a cero, aunque sobresalen picos en 2009, 2015 y 2020, que coinciden con choques inesperados como la recesión global, la caída en el precio del petróleo y los efectos de la pandemia. Aunque no exhiben un patrón sistemático, estos extremos sugieren la presencia de eventos exógenos que escapan a los supuestos de linealidad.

En conjunto, la descomposición muestra que la inflación en Colombia combina una tendencia de mediano plazo, estacionalidad recurrente y choques irregulares.

Figura 11

Descomposición de la serie de inflación interanual en Colombia (2004–2024)



8. Fase 3: Implementación de modelos

La implementación de los modelos se documentó en cuadernos de programación desarrollados en Python, los cuales se incluyen en el Apéndice B.

8.1 Herramientas y entorno de programación

La implementación de los modelos se desarrolló en Python 3.10 dentro del entorno interactivo Jupyter Notebook, lo que permitió integrar código, gráficos y documentación en un mismo flujo de trabajo.

En el caso de ARIMA, se utilizaron las librerías *statsmodels* y *pmdarima*, con apoyo de *pandas* para la manipulación de datos y de *matplotlib* y *seaborn* para la visualización gráfica. La verificación de estacionariedad se efectuó mediante la prueba de Dickey–Fuller aumentada (ADF),

mientras que la identificación de dependencias temporales se apoyó en los gráficos ACF y PACF realizados por *statsmodels*, herramienta para determinar los parámetros del modelo.

Para Lag-Llama, se implementó la librería oficial basada en *PyTorch* y compatible con GluonTS, lo que permitió configurar un enfoque probabilístico de series temporales. El modelo se ejecutó a partir de un checkpoint preentrenado disponible en *Hugging Face Hub*, estrategia que posibilitó una evaluación en modalidad zero-shot y redujo los tiempos de configuración. La preparación de los datos se gestionó en *pandas*, mientras que la definición del modelo incluyó parámetros clave como la distribución de salida (StudentTOutput) y la función de pérdida (NegativeLogLikelihood), elementos fundamentales para capturar la incertidumbre en los pronósticos. Finalmente, Las predicciones se generaron con la función *make_evaluation_predictions* de GluonTS y la evaluación se realizó mediante métricas estándar (MAE, MSE, RMSE y MAPE), calculadas manualmente con NumPy. Los resultados se representaron con matplotlib y se complementaron con rutinas de manipulación en pandas y NumPy.

En la Tabla 4 se presenta un resumen de las principales herramientas, librerías y procedimientos empleados en la implementación de los modelos ARIMA y Lag-Llama.

Tabla 4

Recursos utilizados en la implementación de los modelos ARIMA y Lag-Llama

ÍTEM	ARIMA	Lag-Llama
Librerías principales	*statsmodels, pmdarima, pandas, matplotlib, seaborn*	*PyTorch, GluonTS, pandas, numpy, checkpoint en Hugging Face*
Preparación de datos	Manipulación con *pandas*	Manipulación con *pandas*

Configuración/Particularidades	• Prueba Dickey–Fuller Aumentada (ADF) para estacionariedad • Gráficos ACF y PACF para identificación de parámetros	• Enfoque probabilístico • Zero-shot desde checkpoint preentrenado • Distribución Student-t (StudentTOutput) • Función de pérdida NegativeLogLikelihood
Evaluación y visualización	Gráficos en *matplotlib* y *seaborn*	• make_evaluation_predictions y Evaluator de *GluonTS* • Visualización en *matplotlib*

8.2 Configuración Modelo Arima

8.2.1. Identificación del modelo

Inicialmente, se verificó la estacionariedad mediante la prueba Dickey-Fuller Aumentada (ADF) aplicada a la serie original. El estadístico obtenido fue -2.6881 con $p = 0.0761$, al compararlo con el valor crítico al 5 % (-2.8738) no fue posible rechazar la hipótesis nula de raíz unitaria, por lo que la serie no es estacionaria en niveles, como se observa en la Figura 12.

Figura 12

Resultados de la prueba ADF en Python sobre la serie de inflación interanual diferenciada (2004–2024)

```
#identificación
#Prueba ADF
adf_test = adfuller(serie)
print("Prueba ADF:")
print(f" Estadístico ADF: {adf_test[0]:.4f}")
print(f" p-valor: {adf_test[1]:.4f}")
print(" Valores críticos:", adf_test[4])
```

```
Prueba ADF:
 Estadístico ADF: -2.6881
 p-valor: 0.0761
 Valores críticos: {'1%': -3.458128284586202, '5%': -2.873761835239286, '10%': -2.5732834559706235}
```

Con base en ello, se aplicó una diferenciación de primer orden ($d = 1$) para estabilizar la media y remover tendencias. La Figura 13 muestra la serie diferenciada, la cual oscila alrededor de cero con varianza más estable, confirmando visualmente la transformación. Sobre la serie diferenciada, la ADF arrojó un estadístico de -4.2119 con $p = 0.00063$, claramente por debajo del umbral del 5 %, lo que confirma la estacionariedad tras la transformación y justifica el uso de $d = 1$ en la familia ARIMA. Dicho resultado se aprecia en la Figura 14.

Figura 13

Serie de inflación interanual diferenciada de primer orden (2004–2024)

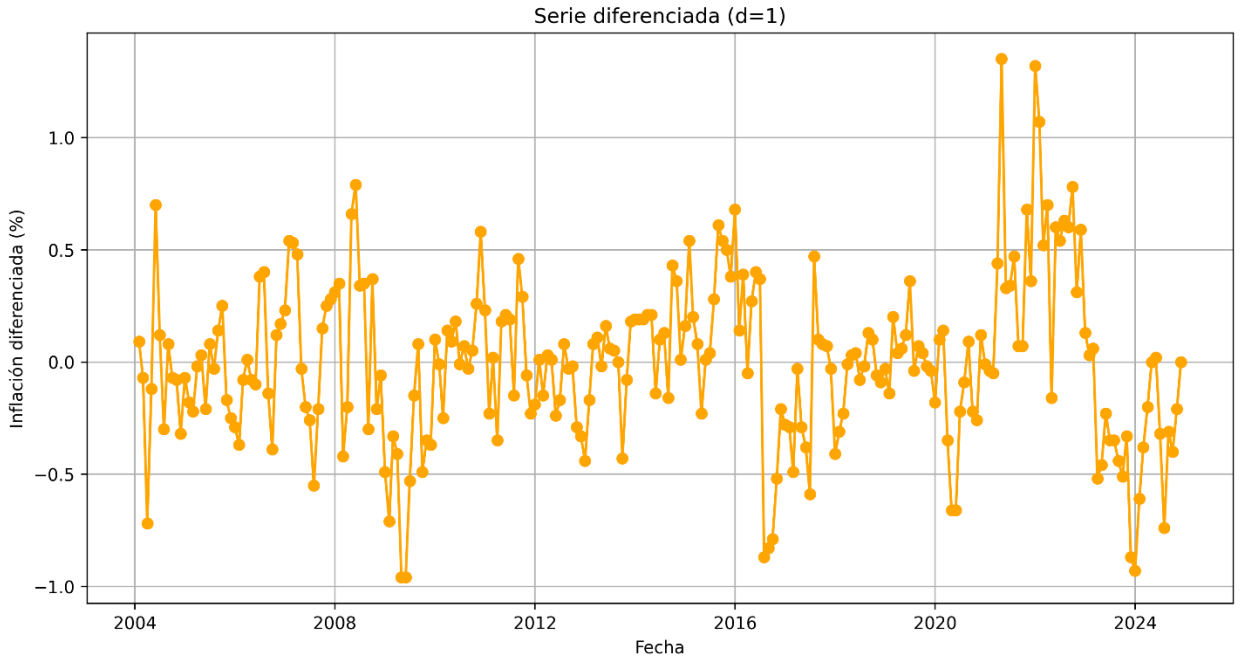


Figura 14

Resultados de la prueba ADF en Python sobre la serie de inflación interanual diferenciada (2004–2024)

```
#Prueba ADF sobre la serie diferenciada
adf_test_diff = adfuller(serie_diff)
print("Prueba ADF sobre serie diferenciada (d=1):")
print(f" Estadístico ADF: {adf_test_diff[0]}")
print(f" p-valor: {adf_test_diff[1]}")
print("Valores críticos:", adf_test_diff[4])
```

```
Prueba ADF sobre serie diferenciada (d=1):
 Estadístico ADF: -4.211882623167407
 p-valor: 0.0006290390462698629
Valores críticos: {'1%': -3.458128284586202, '5%': -2.873761835239286, '10%': -2.5732834559706235}
```

A continuación, se inspeccionaron las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) hasta 36 rezagos. La Figura 15 presenta la PACF, en la que se observa un pico significativo en el rezago 1, mientras que los rezagos posteriores se mantienen dentro de las bandas de confianza, lo que sugiere que un componente autorregresivo de bajo orden

capta la dependencia remanente. Por su parte, la Figura 16 muestra la ACF, caracterizada por una caída rápida hacia cero después de los primeros rezagos y con la mayoría de coeficientes dentro de las bandas, patrón consistente con una serie ya estacionaria.

Figura 15

Resultados de la prueba ADF sobre la serie de inflación interanual diferenciada (2004–2024)

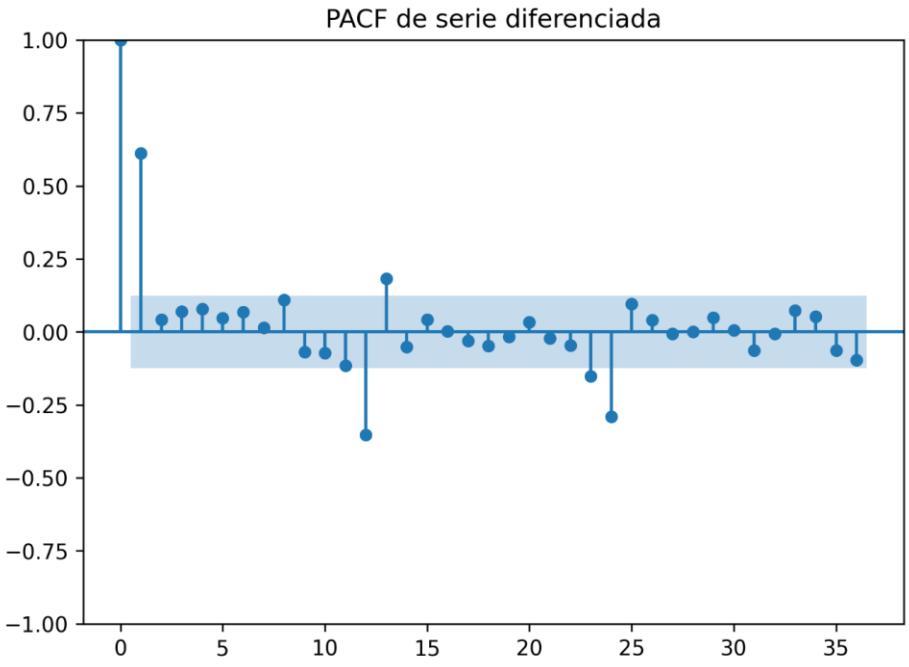
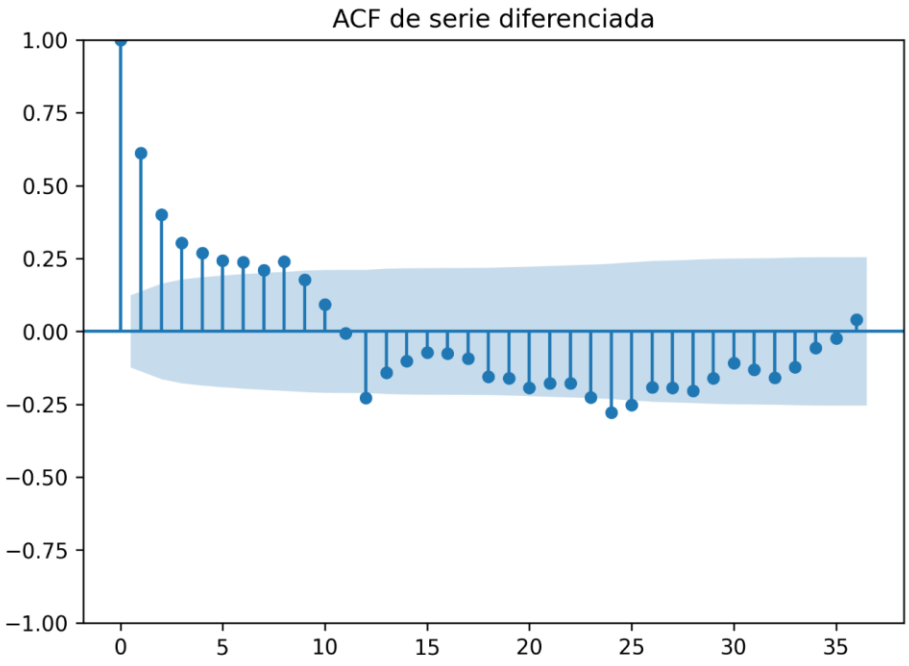


Figura 16

Función de autocorrelación (ACF) de la serie de inflación diferenciada en Colombia (2004–2024)



De forma complementaria, se verificó ausencia de estacionalidad marcada: no se observaron picos persistentes en múltiplos de 12 en ACF/PACF ni señales claras que justificaran una diferenciación estacional, por lo que no se consideró necesario un término SARIMA. Tampoco se aplicaron transformaciones logarítmicas, dado que se trabaja con tasas interanuales (no niveles) y no se evidenciaron problemas severos de varianza dependiente del nivel.

Con estos elementos $d = 1$ confirmado por ADF, AR(1) sugerido por PACF y sin estacionalidad clara, se fijó la base para la siguiente fase, donde se contrastó esta especificación con alternativas cercanas mediante criterios de información.

8.2.2. Estimación

Para definir los parámetros óptimos del modelo se empleó la función *auto_arima,* la cual permite identificar combinaciones de p, d y q que minimizan el criterio de información AIC. En este caso, se establecieron valores máximos de cinco para p y q con el fin de garantizar parsimonia

y evitar la sobreparametrización, mientras que $d = 1$ se fijó de acuerdo con la diferenciación realizada previamente para lograr estacionariedad.

Figura 17

Código en Python para la selección automática de parámetros ARIMA

```
modelo_auto = pm.auto_arima(
    serie,
    start_p=0, start_q=0,
    max_p=5, max_q=5,
    d=1,
    seasonal=False,
    stepwise=True,
    suppress_warnings=True,
    information_criterion='aic'
)

print("Modelo sugerido por auto_arima:")
print(modelo_auto.summary())
```

El modelo estimado mostró un coeficiente autorregresivo de primer rezago $(ar.L1 = 0.6093)(ar.L1 = 0.6093)(ar.L1 = 0.6093)$, el cual resultó estadísticamente significativo al 1 %. Este valor refleja una persistencia moderada de la inflación, indicando que la variación mensual depende en parte del comportamiento observado en el mes anterior. La varianza de los errores se estimó en 0.0845, mientras que el AIC obtenido (96.47) fue menor al de otras especificaciones candidatas, como el ARIMA(0,1,1) y el ARIMA(2,1,0), lo que respalda su elección como modelo más adecuado. La salida generada por statsmodels se muestra en la Figura 18, mientras que los resultados resumidos se presentan en la Tabla 5.

Figura 18

Salida de statsmodels para la estimación del modelo ARIMA(1,1,0)

```
Modelo sugerido: ARIMA(1,1,0)

Resumen del modelo:
                               SARIMAX Results
==============================================================================
Dep. Variable:              Inflacion   No. Observations:                  252
Model:                 ARIMA(1, 1, 0)   Log Likelihood                -46.239
Date:                Wed, 01 Oct 2025   AIC                             96.477
Time:                        23:11:57   BIC                            103.528
Sample:                    01-01-2004   HQIC                            99.315
                         - 12-01-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.6093      0.046     13.223      0.000       0.519       0.700
sigma2         0.0845      0.005     15.470      0.000       0.074       0.095
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                37.56
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               1.09   Skew:                             0.26
Prob(H) (two-sided):                  0.71   Kurtosis:                         4.82
===================================================================================

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

Tabla 5

Resultados de estimación del modelo ARIMA(1,1,0)

Parámetro	**AR(1)**	**σ^2**
Coeficiente	0.6093	0.0845
Error estándar	0.046	0.005
Z	13.223	15.470
p-valor	0.000	0.000
IC 95 % Inferior	0.519	0.074
IC 95 % Superior	0.700	0.095

Posteriormente, el modelo fue ajustado manualmente mediante la función ARIMA de la librería *statsmodels*, confirmando la consistencia de los resultados obtenidos con *auto_arima*. El código de estimación se muestra en la Figura 19 y la salida generada en la Figura 20. La salida de *statsmodels* incluye también algunas pruebas automáticas de residuos (como Ljung-Box en lag = 1 y Jarque-Bera), que se discuten en la sección de diagnóstico.

Figura 19

Código en Python para la estimación manual del modelo ARIMA(1,1,0)

```
# Extraer el orden (p,d,q) sugerido por auto_arima
p, d, q = modelo_auto.order
print(f"Modelo sugerido: ARIMA({p},{d},{q})")

# Ajustar modelo en statsmodels
from statsmodels.tsa.arima.model import ARIMA

modelo = ARIMA(serie, order=(p, d, q))
resultado = modelo.fit()

print("\nResumen del modelo:")
print(resultado.summary())
```

Figura 20

Salida de statsmodels para la estimación manual del modelo ARIMA(1,1,0)

```
Modelo sugerido: ARIMA(1,1,0)

Resumen del modelo:
                               SARIMAX Results
==============================================================================
Dep. Variable:              Inflacion   No. Observations:                  252
Model:                 ARIMA(1, 1, 0)   Log Likelihood                -46.239
Date:                Thu, 02 Oct 2025   AIC                             96.477
Time:                        00:22:41   BIC                            103.528
Sample:                  01-01-2004   HQIC                            99.315
                       - 12-01-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.6093      0.046     13.223      0.000       0.519       0.700
sigma2         0.0845      0.005     15.470      0.000       0.074       0.095
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                37.56
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               1.09   Skew:                             0.26
Prob(H) (two-sided):                  0.71   Kurtosis:                         4.82
===================================================================================

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

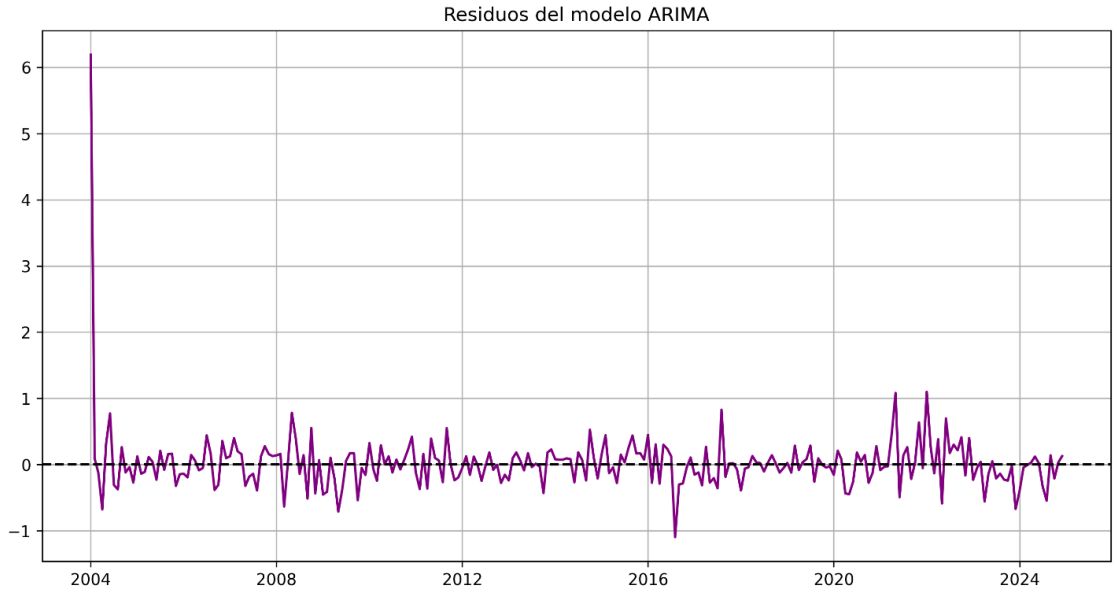
8.2.3 Diagnóstico del modelo.

El diagnóstico del modelo ARIMA(1,1,0) se llevó a cabo mediante el análisis de los residuos, con el objetivo de verificar el cumplimiento de los supuestos de la metodología Box-Jenkins. En primer lugar, la inspección gráfica (Figura 21) evidenció que los residuos se distribuyen de manera aleatoria alrededor de cero, sin presentar tendencias sistemáticas ni variaciones heterocedásticas visibles. Este comportamiento respalda un ajuste adecuado en términos visuales, ya que sugiere que el modelo logró eliminar la mayor parte de la estructura presente en la serie original.

Residuos del modelo ARIMA de la inflación interanual en Colombia (2004–2024)

Figura 21

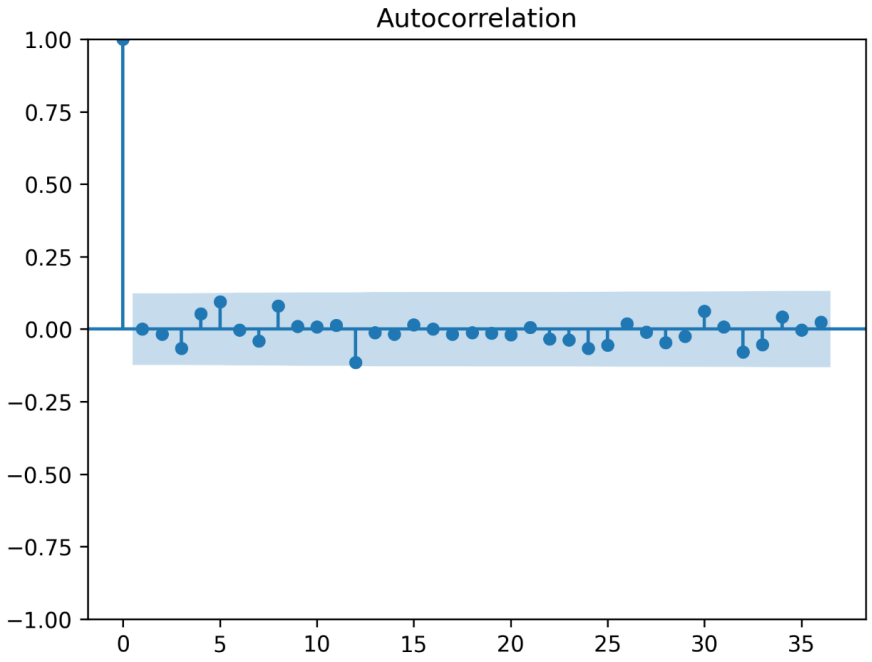
Residuos del modelo ARIMA de la inflación interanual en Colombia (2004–2024)



Posteriormente, se evaluó la función de autocorrelación de los residuos (Figura 22). La mayoría de los coeficientes se ubicaron dentro de los intervalos de confianza al 95 %, lo que indica ausencia de correlaciones significativas. Este resultado es consistente con la independencia temporal de los errores, condición fundamental para considerar que el modelo capturó la dinámica de la inflación de manera satisfactoria.

Figura 22

Función de autocorrelación (ACF) de los residuos del modelo ARIMA



Por último, tal como se mencionó en la sección 8.2.2, la salida de statsmodels reportó de manera automática la prueba de Ljung–Box en un rezago (lag = 1), cuyo resultado ($p = 0.71$) sugiere ausencia de autocorrelación inmediata en los residuos. De manera complementaria, en esta sección se aplicó la prueba de Ljung–Box a 12 rezagos, obteniéndose un estadístico de 9.93 con un valor p de 0.622. Ambos resultados permiten no rechazar la hipótesis nula de independencia serial, confirmando que los residuos se comportan como ruido blanco y no contienen información adicional que el modelo no haya explicado previamente.

En conjunto, estos resultados permiten concluir que el modelo ARIMA(1,1,0) cumple satisfactoriamente con los criterios de diagnóstico, validándose como una representación estadísticamente adecuada de la serie de inflación analizada.

8.2.4 Pronóstico

8.2.4.1 Pronóstico Multi-step. El modelo ARIMA(1,1,0) fue estimado con la información histórica comprendida entre enero de 2004 y diciembre de 2023, definida como conjunto de entrenamiento. A partir de este ajuste, se generó una proyección multi-step de 12 meses para el año 2024 con intervalos de confianza al 95 %.El procedimiento de estimación se implementó en Python mediante la librería statsmodels. El fragmento de código utilizado para calcular el pronóstico se presenta en la Figura 23.

Figura 23

Código para pronóstico Multi-step en Python con ARIMA(1,1,0)

```
# Multi-step
train = serie.loc[:'2023-12-01']

modelo = ARIMA(train, order=(p, d, q))
resultado = modelo.fit()

forecast_2024 = resultado.get_forecast(steps=12)

mean_pred = forecast_2024.predicted_mean
conf_int = forecast_2024.conf_int(alpha=0.05)
```

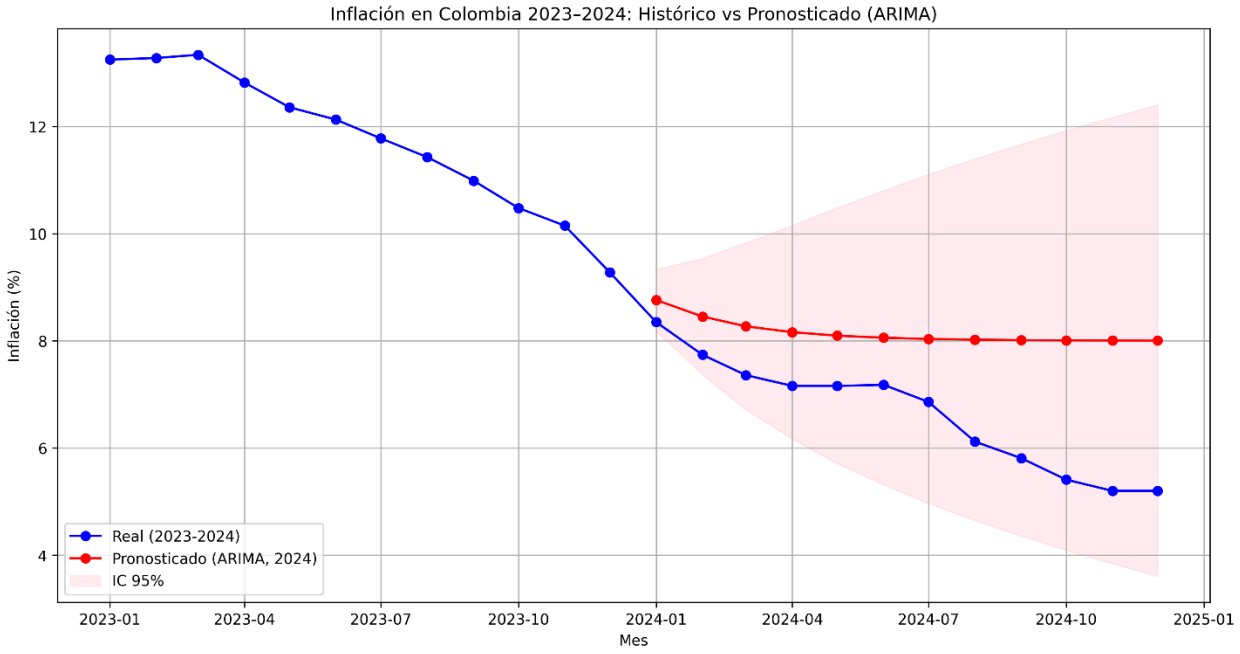
A partir de esto, se obtuvo la proyección de la inflación mensual para 2024. En la Figura 24 se ilustra el comportamiento de las predicciones: la línea roja corresponde al pronóstico para 2024 y la azul representa los valores observados junto con la trayectoria histórica de referencia. Los resultados muestran que el modelo logró capturar parcialmente la tendencia descendente registrada a inicios de 2024, sin embargo, conforme avanza el horizonte de proyección, la serie estimada se estabiliza en un comportamiento relativamente plano. Esto indica que el modelo reproduce la dinámica general de la inflación, pero tiene limitaciones para anticipar las fluctuaciones de corto plazo.

Asimismo, se aprecia un ensanchamiento progresivo de los intervalos de confianza, lo cual refleja la acumulación de incertidumbre característica de este tipo de proyecciones largas. Este

comportamiento es esperable dentro del marco ARIMA y confirma que, aunque el modelo es válido para describir tendencias generales, su precisión disminuye a medida que se incrementa la distancia temporal respecto al conjunto de entrenamiento.

Figura 24

Pronóstico Multi-step de la inflación interanual en Colombia con el modelo ARIMA (2023–2024)



8.2.4.2 Pronóstico Rolling-step. Con el propósito de aproximar el ejercicio a un escenario más cercano al uso práctico, se implementó un esquema de pronóstico dinámico o rolling forecast. Bajo esta estrategia, en cada periodo del año 2024 se proyectó un paso adelante, se incorporó el valor observado más reciente y se recalibró el modelo antes de emitir la siguiente estimación.

El procedimiento fue programado en Python mediante la librería *statsmodels*. El fragmento de código utilizado se presenta en la Figura 25.

Figura 25

Código para pronóstico Rolling-step en Python con ARIMA(1,1,0)

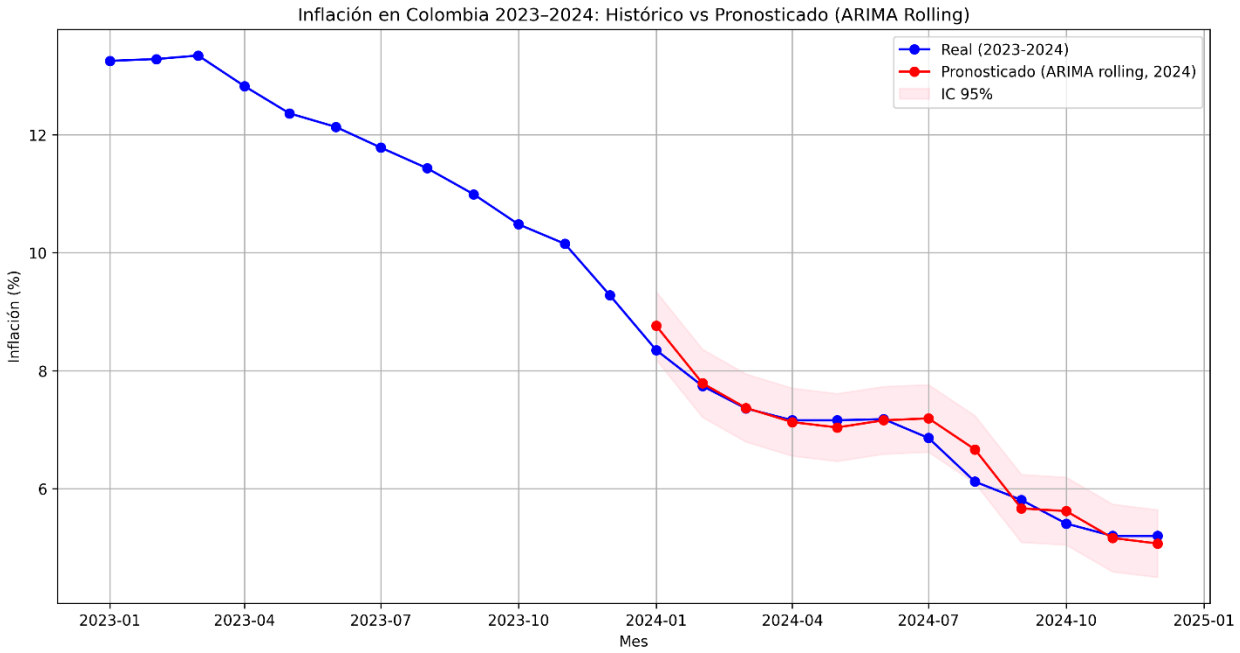
```python
# Rolling-step
train = serie.loc[:'2023-12-01']
test = serie.loc['2024-01-01':]

modelo = ARIMA(train, order=(p,d,q))
resultado = modelo.fit()

predicciones = []
indices = []
lower_bounds = []
upper_bounds = []


for fecha_real in test.index:

    pred = resultado.get_forecast(steps=1)
    mean_pred = pred.predicted_mean.iloc[0]
    conf_int = pred.conf_int(alpha=0.05).iloc[0]

    predicciones.append(mean_pred)
    indices.append(fecha_real)
    lower_bounds.append(conf_int["lower Inflacion"])
    upper_bounds.append(conf_int["upper Inflacion"])

    train = pd.concat([train, serie.loc[[fecha_real]]])

    # Reajustar modelo
    modelo = ARIMA(train, order=(p,d,q))
    resultado = modelo.fit()
```

Dado lo anterior, la Figura 26 presenta la comparación entre las predicciones y los datos reales. En este caso, la línea roja (pronóstico) muestra un ajuste mucho más cercano a la trayectoria de la inflación observada (línea azul), en contraste con lo evidenciado en el enfoque estático. Además, la mayoría de las proyecciones se mantuvo dentro de los intervalos de confianza al 95 %, cuyos rangos resultaron más estrechos que en la estrategia anterior.

Estos hallazgos indican que el pronóstico dinámico mejora la capacidad predictiva *out-of-sample,* permitiendo capturar con mayor precisión los movimientos de corto plazo. En consecuencia, este enfoque resulta más adecuado para contextos de seguimiento continuo y apoyo a la toma de decisiones en tiempo real, donde la actualización constante de la información disponible es fundamental para mantener la calidad de las estimaciones.

Figura 26

Pronóstico Rolling-step de la inflación internaual en Colombia con el modelo ARIMA (2023–2024)



8.3. Configuración Modelo Lag-Llama

8.3.1 Datos de entrada

Los datos utilizados corresponden a la serie ya preparada en la Fase 2, es decir, transformada a frecuencia mensual (MS), tipificada en formato float32 y dividida en conjuntos de entrenamiento (2004–2023) y validación (2024). A partir de esta base, se procedió a la configuración específica del modelo.

8.3.2 Configuración de parámetros

El modelo se configuró en modalidad zero-shot, empleando el checkpoint oficial sin procesos de ajuste adicionales. Esta decisión se tomó porque Lag-Llama, al estar preentrenado sobre un conjunto muy amplio y diverso de series temporales, ofrece la ventaja de generalizar su desempeño sin necesidad de un reentrenamiento específico para el caso colombiano. De esta

manera, se pudo evaluar su capacidad de transferencia hacia un dominio económico distinto al de su entrenamiento original.

Con el fin de capturar patrones de distinta naturaleza, se utilizó una ventana de contexto de 60 meses (equivalente a cinco años de observaciones). Este horizonte resultó apropiado para representar tanto las dinámicas de corto plazo como los ciclos de mediano plazo en la inflación. A partir de este contexto, el modelo generó 300 trayectorias de pronóstico, número que fue definido manualmente mediante el parámetro num_samples en la implementación. La elección de este valor respondió a consideraciones metodológicas y prácticas: con un número reducido de trayectorias (por ejemplo, 100) los intervalos de confianza tienden a ser inestables, mientras que un número excesivo (como 1000) incrementa de forma significativa el costo computacional sin aportar mejoras sustanciales en la calidad de las estimaciones. Por tanto, se adoptó 300 como un balance adecuado entre precisión estadística y eficiencia computacional.

A partir de estas trayectorias se construyeron intervalos de confianza al 90 %, lo que permitió cuantificar de manera explícita la incertidumbre asociada a las proyecciones. De este modo, además de contar con una estimación puntual de la inflación, se obtuvo una representación probabilística más completa del fenómeno, aspecto crucial en el análisis de variables macroeconómicas caracterizadas por su alta volatilidad.

En términos de implementación, se fijó un tamaño de lote de 32, lo que permitió balancear la eficiencia computacional con la estabilidad de los cálculos. Asimismo, se activó la restricción de no negatividad para garantizar que los pronósticos se mantuvieran dentro de valores plausibles, coherentes con la naturaleza del fenómeno inflacionario. Finalmente, dado que la longitud total de la secuencia (contexto más horizonte) superaba lo considerado en el preentrenamiento original, se

aplicó un ajuste de rope scaling en las representaciones posicionales, con el fin de mantener la estabilidad del mecanismo de atención en horizontes más amplios.

Por tratarse de un modelo probabilístico, es importante señalar que las predicciones pueden variar levemente entre ejecuciones, ya que se generan mediante muestreo. En consecuencia, los resultados presentados corresponden a una ejecución representativa, para valorar su funcionamiento y las limitaciones del modelo en el contexto analizado. El fragmento de código empleado para la configuración de parámetros del modelo Lag-Llama en modalidad zero-shot se presenta en la Figura 27.

Figura 27

Código de configuración de parámetros en Python para el modelo Lag-Llama en modalidad zero-shot

```
prediction_length = 12
context_length = 60
num_samples = 300
batch_size = 32
device = "cuda" if torch.cuda.is_available() else "cpu"

ckpt = torch.load("lag-llama.ckpt", map_location=device)
estimator_args = ckpt["hyper_parameters"]["model_kwargs"]

estimator = LagLlamaEstimator(
    ckpt_path="lag-llama.ckpt",
    prediction_length=prediction_length,
    context_length=context_length,
    device=torch.device(device),


    input_size=estimator_args["input_size"],
    n_layer=estimator_args["n_layer"],
    n_embd_per_head=estimator_args["n_embd_per_head"],
    n_head=estimator_args["n_head"],
    scaling=estimator_args["scaling"],
    time_feat=estimator_args["time_feat"],

    nonnegative_pred_samples=True,
    batch_size=batch_size,
    num_parallel_samples=num_samples,

    rope_scaling={
        "type": "linear",
        "factor": max(1.0, (context_length + prediction_length) / estimator_args["context_length"]),
    }
)
```

8.3.3 Construcción del estimador y predictor

La construcción del modelo se inició a partir de la instanciación del LagLlamaEstimator, empleando los hiperparámetros originales del checkpoint preentrenado (número de capas, cabezas de atención, tamaño de embeddings y time-features). Esta decisión permitió garantizar que el modelo conservara su diseño fundacional, manteniendo la coherencia con la arquitectura propuesta en sus estudios iniciales.

Con esta configuración, se generó el predictor, encargado de automatizar procesos esenciales en la preparación de los datos. Entre ellos se incluyó la creación de ventanas de rezagos (lag features), que permiten al modelo aprovechar la dependencia temporal de la serie; la incorporación de variables de calendario, como el mes del año representado en codificación circular, para capturar patrones estacionales implícitos; y la aplicación de una normalización robusta basada en la mediana y el rango intercuartílico (IQR), lo que contribuyó a reducir la influencia de valores extremos sin distorsionar la estructura general de la serie.

Un aspecto diferenciador de Lag-Llama radica en la naturaleza probabilística de su salida. En lugar de ofrecer un único valor puntual, el modelo estima los parámetros de una distribución Student-t (media, escala y grados de libertad), lo que permite capturar tanto la tendencia más probable de la inflación como la amplitud de la incertidumbre asociada. Este enfoque no solo genera trayectorias centrales, sino también múltiples simulaciones posibles, aportando una visión más completa del riesgo y la variabilidad inherentes al fenómeno inflacionario. El fragmento de código correspondiente a la construcción del estimador y del predictor se presenta en la Figura 28.

Figura 28

Código en Python para la construcción del estimador y predictor del modelo Lag-Llama

```
predictor = estimator.create_predictor(
    estimator.create_transformation(),
    estimator.create_lightning_module()
)


train_ds = PandasDataset(train_data, freq="M", target="Inflacion")
```

8.3.4 Generación de pronósticos

8.3.4.1 Pronóstico Multi-step. Con la configuración descrita, se generó un pronóstico multi-step de 12 meses, correspondiente a la proyección simultánea de la inflación para todo el año 2024. A partir de 300 trayectorias simuladas, se construyeron intervalos de confianza al 90 %, lo que permitió cuantificar la incertidumbre asociada a las proyecciones.

El procedimiento de implementación en Python se presenta en la Figura 29.

Figura 29

Código para pronóstico multi-step en Python con Lag-Llama

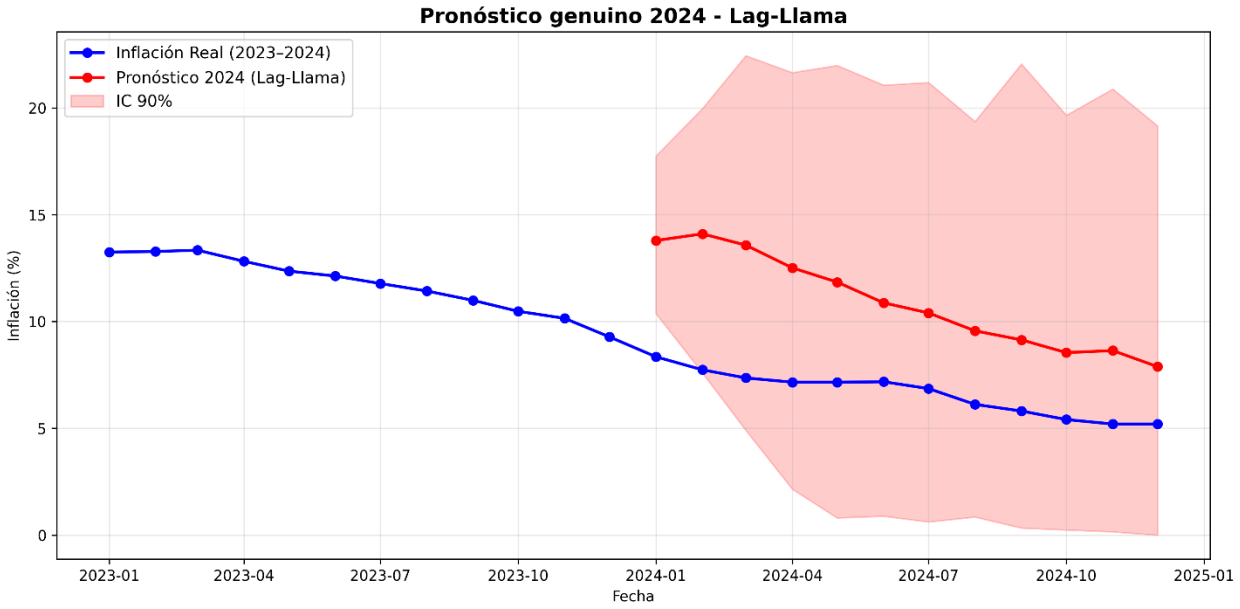
```
forecast_it, ts_it = make_evaluation_predictions(
    dataset=train_ds,
    predictor=predictor,
    num_samples=num_samples
)

forecasts = list(forecast_it)
forecast = forecasts[0]


forecast_index = pd.date_range(start="2024-01-01", periods=prediction_length, freq="MS")
forecast_df = pd.DataFrame({
    "Prediccion_Media": forecast.mean,
    "Percentil_5": forecast.quantile(0.05),
    "Percentil_95": forecast.quantile(0.95)
}, index=forecast_index)
```

Los resultados del pronóstico se muestran en la Figura 30, el modelo Lag-Llama estimó una senda descendente de la inflación, aunque con una pendiente más moderada que la observada

en la serie real. La amplitud creciente de las bandas de confianza a medida que avanza el horizonte refleja la dificultad inherente al esquema multi-step, en el cual los errores de predicción tienden a acumularse y amplificarse. Este enfoque constituye, no obstante, una evaluación estricta de la capacidad predictiva del modelo, dado que no se beneficia de observaciones reales posteriores.

Figura 30

Pronóstico Multi-step de la inflación interanual en Colombia con el modelo Lag-Llama (2023–2024)



8.3.4.3 Pronóstico Rolling-step. Con el fin de simular un escenario de predicción en tiempo real, se aplicó un pronóstico rolling de un mes adelante para todo el año 2024. En cada iteración, el modelo generó una predicción de un paso, incorporó la observación real más reciente y recalibró sus parámetros antes de emitir el siguiente pronóstico.

El procedimiento de implementación en Python se muestra en la Figura 31.

Figura 31

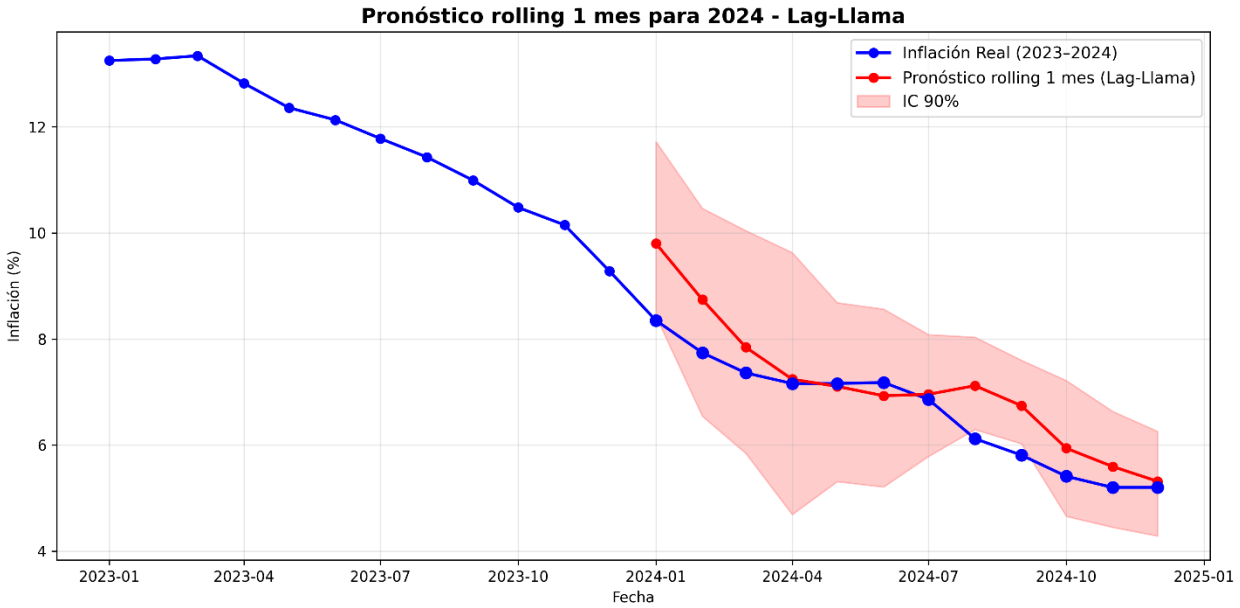
Código para pronóstico Rolling-step en Python con Lag-Llama

```
forecast_it, ts_it = make_evaluation_predictions(
    dataset=train_ds,
    predictor=predictor_roll,
    num_samples=num_samples
)

forecasts = list(forecast_it)
if len(forecasts) == 0:
    print(f"No se obtuvo forecast para {month_start.date()}")
    continue

f = forecasts[0]
pred_mean = float(f.mean[0])
p05 = float(f.quantile(0.05)[0])
p95 = float(f.quantile(0.95)[0])

real_val = None
if month_start in df.index:
    real_val = float(df.loc[month_start, 'Inflacion'])

results.append({
    "Fecha": month_start,
    "Prediccion_Media": pred_mean,
    "Percentil_5": p05,
    "Percentil_95": p95,
    "Real": real_val
})
```

Los resultados se aprecian en la Figura 32, este esquema permitió que Lag-Llama se adaptara de manera más eficaz a los cambios recientes en la dinámica inflacionaria, siguiendo de cerca la trayectoria real. La mayoría de las observaciones se mantuvo dentro de los intervalos de confianza al 90 %, que en este caso resultaron más estrechos que en el multi-step.

Figura 32

Pronóstico Rolling-step de la inflación interanual en Colombia con el modelo Lag-Llama (2023–2024)



La comparación entre ambos enfoques resalta que, mientras el esquema multi-step ofrece una visión de conjunto sobre el comportamiento esperado de la inflación, el rolling proporciona una mayor precisión out-of-sample y es particularmente útil para aplicaciones prácticas que requieren seguimiento continuo y ajustes frecuentes, como la planificación económica y la toma de decisiones en política monetaria.

9. Fase 4: Evaluación e interpretación

La evaluación de los modelos se realizó a partir de tres métricas de error ampliamente utilizadas en series temporales: el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Porcentual Absoluto Medio (MAPE).

Para su cálculo se usaron funciones en Python, apoyadas en la librería *scikit-learn* y en operaciones numéricas con NumPy. En ambos modelos, las métricas se estimaron a partir de la comparación entre los valores reales de la inflación y los pronósticos obtenidos en los esquemas multi-step y rolling-step.

En este caso no se utilizó la clase *Evaluator* de GluonTS, ya que está pensada para evaluaciones con múltiples series o ventanas de validación. Al trabajar con una única serie y un horizonte fijo de 12 meses, resultó más práctico y consistente calcular las métricas de forma directa, lo que permitió aplicar el mismo procedimiento de evaluación tanto a ARIMA como a Lag-Llama.

Este procedimiento permitió cuantificar la magnitud de los errores absolutos, penalizar en mayor medida las desviaciones grandes y expresar el error relativo en términos porcentuales, facilitando la interpretación comparativa entre modelos. Los resultados obtenidos se resumen en la Tabla 6.

Tabla 6

Métricas de error en el pronóstico de inflación modelos ARIMA y Lag-Llama

Estrategia	Modelo	MAE	RMSE	MAPE
Multi-step	Arima	1.5292	1.7447	25.67%
	Lag-Llama	4.2781	4.4482	63.71%
Rolling-step	Arima	0.1696	0.2379	2.61%
	Lag-Llama	0.5326	0.6917	7.97%

En el escenario multi-step, el ARIMA alcanzó un MAE de 1.53 y un MAPE de 25.67 %. Si bien este resultado presenta que refleja dificultades para anticipar fluctuaciones de corto plazo,

representa un desempeño considerablemente mejor que el obtenido por Lag-Llama, cuyos errores resultaron entre dos y tres veces mayores (MAE de 4.29 y MAPE de 63.71 %).

El contraste se vuelve aún más evidente bajo el enfoque rollin-step. En este caso, el ARIMA redujo sus errores de manera significativa, con un MAE de apenas 0.17 y un MAPE de 2.61 %, lo que demuestra una alta precisión predictiva al incorporar observaciones reales de manera secuencial. Lag-Llama también mejoró respecto al escenario multi-step alcanzando un MAE de 0.53 y un MAPE de 7.97 %, aunque sus resultados se mantuvieron por debajo del desempeño de ARIMA.

En conjunto los hallazgos sugieren que el modelo ARIMA es más robusto y confiable para el pronóstico de la inflación mensual en Colombia, particularmente bajo un esquema de actualización continua, donde mostró un ajuste muy cercano a los valores observados. No obstante, debe reconocerse que en horizontes largos, como el multi-step, incluso el ARIMA presenta limitaciones importantes.

Una lectura detenita de estas métricas permite dimensionar su impacto práctico. En el caso del ARIMA multi-step, un MAPE de 25.67 % implica que, si la inflación real fuese de 10 %, el modelo podría errar en aproximadamente ±2.6 puntos porcentuales, proyectando valores entre 7.4 % y 12.6 %. Este nivel de desviación resulta problemático en escenarios de política monetaria, donde diferencias de uno o dos puntos pueden modificar decisiones clave. Por el otro lado, bajo el enfoque rolling, el error relativo de 2.61 % se traduce en un rango mucho más ajustado, de 9.7 % a 10.3 % sobre una inflación real de 10 %, lo cual confirma su pertinencia en contextos de seguimiento mensual y apoyo a la toma de decisiones económicas. Para Lag-Llama, los errores fueron con mayor magnitud: en multi-step, un MAPE de 63.71 % implica proyecciones que podrían variar entre 3.6 % y 16.4 % ante una inflación real de 10 %, rango demasiado amplio para

ser operativo. Incluso en rolling, aunque el error bajó a 7.97 %, esto aún representa cerca de ±0.8 p.p., manteniéndose por debajo del nivel de precisión alcanzado por ARIMA.

10. Fase 5: Documentación

En esta fase final se consolidaron y organizaron los resultados obtenidos, junto con los recursos generados durante el desarrollo del proyecto, con el propósito de garantizar su accesibilidad, reproducibilidad y transparencia. A continuación, se presentan los componentes realizados:

10.1 Elaboración de artículo académico

En esta etapa se realizó un artículo de carácter publicable, en el cual se sintetizaron los principales hallazgos del proyecto. La redacción se realizó siguiendo los estándares de estilo y rigor científico que exigen las revistas académicas, lo que implicó una estructuración clara de la introducción, el marco teórico, la metodología aplicada y los resultados obtenidos. Este se encuentra disponible en el Apéndice C.

10.2 Consolidación del documento final

Se elaboró el presente documento que recopiló todo el proceso investigativo. En este informe se plasmaron los análisis exploratorios iniciales, la implementación de los modelos ARIMA y Lag-Llama, la comparación de métricas de desempeño y la discusión crítica de los resultados, buscando tener una visión completa y ordenada del trabajo, facilitando la consulta futura por parte de otros estudiantes, investigadores o profesionales interesados en la predicción de la inflación.

10.3 Repositorio Git Hub

Como parte de la estrategia de documentación y difusión, se creó un repositorio público en GitHub. En este espacio se encuentran los *notebooks* correspondientes al análisis exploratorio y a la modelación con los modelos ARIMA y Lag-Llama, esto como propósito de garantizar la transparencia y la reproducibilidad del estudio, permitiendo poder replicar los resultados o adaptar la metodología a otros contextos.

El repositorio está disponible en: https://github.com/shirleyacala-coder/Prediccion_Inflacion_Colombia_ARIMA_LagLlama.git

11. Conclusiones

El presente trabajo tuvo como objetivo predecir la inflación mensual en Colombia mediante un análisis comparativo de los modelos ARIMA y Lag-Llama, considerando el período comprendido entre 2004 y 2024. Para ello, se desarrollaron procesos de recolección, limpieza y preparación de datos, seguidos de la implementación de cada modelo bajo dos enfoques de predicción: Multi-step y Rolling-step

Los resultados permiten concluir que ARIMA continúa siendo modelo sólido y confiable para la predicción de series económicas como la inflación. En contraste, el modelo Lag-Llama, aun cuando proyectó trayectorias descendentes coherentes con la tendencia observada, presentó errores significativamente mayores, estos resultados sugieren que, en su configuración básica de zero-shot, Lag-Llama no logra superar la robustez de los modelos clásicos, aunque muestra potencial como marco probabilístico capaz de generar intervalos de confianza y cuantificar la incertidumbre de manera más flexible.

Desde el punto de vista práctico, este estudio aporta evidencia de que los modelos tradicionales de series temporales continúan siendo una herramienta de valor para el monitoreo y la toma de decisiones económicas en Colombia. Su fácil manejo y capacidad de actualización periódica los convierten en aliados estratégicos para instituciones como el Banco de la República, entidades gubernamentales y empresas privadas que requieren anticipar cambios en el costo de vida.

En el ámbito académico, la investigación constituye la primera aplicación documentada de Lag-Llama para la predicción de inflación en Colombia. Aunque sus resultados fueron menos precisos, la experiencia abre la puerta a la exploración de modelos fundacionales de series temporales en contextos macroeconómicos locales. De esta manera, la tesis sienta un precedente y un punto de comparación para futuros estudios que deseen incorporar enfoques de aprendizaje profundo en la predicción económica.

No obstante, es importante reconocer algunas limitaciones. La investigación se restringió al uso de datos históricos de inflación sin la incorporación de variables exógenas, lo que podría reducir la capacidad predictiva en escenarios de choques externos o cambios estructurales. Adicionalmente, el modelo Lag-Llama no fue sometido a procesos exhaustivos de optimización de hiperparámetros como el fine-tuning, debido a restricciones de tiempo y recursos computacionales. Finalmente, la evaluación se concentró en un único horizonte de pronóstico (12 meses), lo que limita la generalización de los resultados a otros plazos.

En síntesis, esta tesis confirma que, para la predicción de la inflación en Colombia, los modelos clásicos ARIMA mantienen un desempeño superior, mientras que Lag-Llama representa una alternativa emergente con potencial de desarrollo. El contraste entre ambos enfoques evidencia que la combinación de métodos tradicionales y técnicas modernas de aprendizaje automático

puede constituir un camino prometedor para fortalecer la analítica económica en un entorno cada vez más incierto y dinámico.

12. Recomendaciones

A partir de los resultados obtenidos en este estudio, se plantean las siguientes recomendaciones tanto para la aplicación práctica de los modelos como para futuras investigaciones:

En primer lugar, se sugiere el uso preferente de ARIMA bajo un esquema rolling-step. Esta estrategia, al permitir la actualización mensual del modelo con datos reales, asegura un alto nivel de precisión y mitiga la tendencia a la convergencia hacia valores constantes que caracteriza a los pronósticos multi-step. Asimismo, en series con patrones recurrentes es aconsejable incorporar estacionalidad explícita, evaluando de manera sistemática combinaciones de parámetros estacionales.

En segundo lugar, se recomienda la inclusión de variables exógenas en futuras aplicaciones, extendiendo los modelos hacia versiones ARIMAX o SARIMAX. La incorporación de indicadores macroeconómicos como tasas de interés, tipo de cambio, salarios o expectativas de inflación podría aumentar la capacidad predictiva, especialmente en contextos de choques estructurales.

Respecto a Lag-Llama, se plantea la necesidad de un proceso exhaustivo de optimización (fine tuning). Ajustar de manera sistemática parámetros como el tamaño de ventana, el número de rezagos, la tasa de aprendizaje, la profundidad de la red o la regularización, mediante validación cruzada y técnicas de búsqueda de hiperparámetros (grid search, random search o Bayesian optimization), permitiría mejorar su precisión, en especial en pronósticos de horizonte extendido.

De igual manera, resulta conveniente explorar arquitecturas híbridas que integren lo mejor de ambos enfoques: la capacidad interpretativa de los modelos estadísticos ARIMA y SARIMA con la potencia de representación de técnicas de aprendizaje profundo, como redes neuronales recurrentes o transformadores. Esta línea de investigación puede contribuir a capturar tanto patrones lineales como no lineales en la dinámica inflacionaria.

Otra recomendación es ampliar la evaluación a diferentes horizontes de predicción. Además del horizonte anual (12 meses) utilizado en este estudio, se sugiere analizar plazos de 3, 6 y 9 meses, con el fin de identificar en qué escenarios cada modelo ofrece mejor desempeño y facilitar así su adopción práctica según las necesidades de los usuarios.

Finalmente, se recomienda dar continuidad a la transferencia metodológica. La estrategia comparativa desarrollada puede aplicarse a otras series económicas relevantes para el país, como el desempleo, el crecimiento del PIB o los precios de materias primas. Además, sería conveniente que las entidades responsables del análisis económico consideren la implementación de estos modelos como complemento a sus herramientas actuales, aprovechando su bajo costo computacional y facilidad de actualización. En paralelo, se sugiere fomentar la capacitación institucional en técnicas emergentes de predicción de series temporales, fortaleciendo así la capacidad de análisis y la toma de decisiones basadas en evidencia en el ámbito económico y de política pública.

Referencias Bibliográficas

Adyatma, S. F., & Alamsyah, A. (2022). The Indonesia Stock Exchange Composite Prediction based on Macroeconomic Indicators using ARIMA, LSTM, and ANN. *Proceedings - 2022 8th International Conference on Science and Technology, ICST 2022*. https://doi.org/10.1109/ICST56971.2022.10136262

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6)*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ali, M., Moore, P. W., Barkouki, T., & Brattain, L. J. (2024a). Zero-Shot Forecasting for ECG Time Series Data Using Generative Foundation Models. *2024 IEEE 20th International Conference on Body Sensor Networks, BSN 2024 - Proceedings*. https://doi.org/10.1109/BSN63547.2024.10780646

Ali, M., Moore, P. W., Barkouki, T., & Brattain, L. J. (2024b). Zero-Shot Forecasting for ECG Time Series Data Using Generative Foundation Models. *2024 IEEE 20th International Conference on Body Sensor Networks, BSN 2024 - Proceedings*. https://doi.org/10.1109/BSN63547.2024.10780646

Amjad, F., Korotko, T., & Rosin, A. (2024). Forecasting PV Energy Generation Using Transformer-Based Architectures: A Comparative Study of Lag-Llama, TFT, and DeepAR. *2024 IEEE 65th Annual International Scientific Conference on Power and*

Electrical Engineering of Riga Technical University, RTUCON 2024 - Proceedings. https://doi.org/10.1109/RTUCON62997.2024.10830763

Arciniegas Hernández, D., & Castaño Arévalo, A. (2022). Modelo para la predicción de la variable Tasa Representativa del Mercado colombiano basado en un análisis de Box Jenkins versus la técnica de Kalman Filters [Tesis de pregrado, Universidad Industrial de Santander]. *NOESIS*. https://noesis.uis.edu.co/items/7d8ce12d-0963-47a5-a62c-7aab25f63b96

Banco de la República. (2023a). *Econo-cimientos: ¿Qué tanto sabe sobre la inflación?* https://www.banrep.gov.co/es/banrep-educa/econo-cimientos/que-tanto-sabe-sobre-inflacion

Banco de la República. (2023b). *Informe de política monetaria – julio 2023*. https://repositorio.banrep.gov.co/server/api/core/bitstreams/da1ab590-6d75-4a8f-834f-f62ce92f0098/content

Banco Mundial. (2024). *Tendencias recientes de pobreza y desigualdad en América Latina y el Caribe*.

Bejarano Salcedo, V., Julio Román, J., Caicedo García, E., & Cárdenas Cárdenas, J. (2022). Entendiendo, Modelando y Pronosticando el Efecto de "El Niño" Sobre los Precios de los Alimentos: El Caso Colombiano (Borradores de Economía No. 1102). *Banco de la República de Colombia*. https://repositorio.banrep.gov.co/server/api/core/bitstreams/bd1adbda-f383-4d69-adc5-a9dc549b4afc/content

Bernanke, B. S., & Mishkin, F. S. (1997). *Inflation targeting: A new framework for monetary policy?* (NBER Working Paper No. 5893). National Bureau of Economic Research. https://www.nber.org/papers/w5893

Bisong, E. (2019). *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. https://doi.org/10.1007/978-1-4842-4470-8/COVER

Blanchard, O., & Johnson, D. R. (2013). *Macroeconomía* (6.ª ed.). Pearson Educación.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (3rd ed.). Prentice-Hall.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/GMD-7-1247-2014

Chuku, C., Simpasa, A., & Oduor, J. (2019). Intelligent forecasting of economic growth for developing economies. *International Economics*, *159*, 74–93. https://doi.org/10.1016/j.inteco.2019.06.001

Friedman, M. (1968). The Role of Monetary Policy. *The American Economic Review*, *58*(1), 1–17. https://www.jstor.org/stable/1831652

Friedman, M., & Schwartz, A. J. (1963). Money and Business Cycles. *The American Economic Review*, *45*(1), 32–64. https://www.jstor.org/stable/1927148

Goldfajn, I., & Werlang, S. R. da C. (2000). *Working Paper Series The Pass-through from Depreciation to Inflation: A Panel Study* (Working Paper Series No. 5). Banco Central do Brasil. http://www.bcb.gov.br

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

Ismail, L. B., Joytu, M. I., Plabon, T. I., & Oshman, M. S. (2023). Evaluation of Machine Learning Models to Forecast Inflation: Bangladesh as a Case Study. *Proceedings of the 2023 International Symposium on Networks, Computers and Communications, ISNCC 2023*. https://doi.org/10.1109/ISNCC58260.2023.10323827

Keynes, J. M. (1963). *The general theory of employment, interest, and money*. Springer International Publishing. https://doi.org/10.1007/978-3-319-70344-2

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. *Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Lakshmi Narayanaa, T., Skandarsini, R. R., Ida, S. J., Sabapathy, S. R., & Nanthitha, P. (2023). Inflation Prediction: A Comparative Study of ARIMA and LSTM Models Across Different Temporal Resolutions. *3rd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2023 - Proceedings*, 1390–1395. https://doi.org/10.1109/ICIMIA60377.2023.10425970

Loaiza Zapata, J. F. (2022). Pronóstico de la inflación colombiana: una aproximación desde los modelos machine learning [Tesis de pregrado, Universidad EAFIT]. *Repositorio institucional de la Universidad EAFIT.* https://repository.eafit.edu.co/items/2d096c80-d07f-4fc3-bd5c-325bf6c7f5b2

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65(2),* 297–303. https://doi.org/10.1093/biomet/65.2.297

Mankiw, N. G. (2006). The Macroeconomist as Scientist and Engineer. *The Journal of Economic Perspectives, 20*(4), 1–46. https://doi.org/10.1257/jep.20.4.29

McKinney, W. (2017*). Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (2nd ed.). O'Reilly Media

Modugno, M. (2013). Now-casting inflation using high frequency data*. International Journal of Forecasting, 29*(4)*,* 664–675. https://doi.org/10.1016/j.ijforecast.2012.12.003

Montenegro García, Á. (2011). *Análisis de series de tiempo.* Editorial Pontificia Universidad Javeriana. https://www-digitaliapublishing-com.bibliotecavirtual.uis.edu.co/a/19504

Nordhaus, W. D. (1975). The Political Business Cycle. *The Review of Economic Studies, 42*(2), 169–190. https://www.jstor.org/stable/2296528

Özgür, Ö., & Akkoç, U. (2022). Inflation forecasting in an emerging economy: selecting variables with machine learning algorithms. *International Journal of Emerging Markets, 17*(8), 1889–1908. https://doi.org/10.1108/IJOEM-05-2020-0577

Palma, W. (2016). *Time series analysis*. John Wiley & Sons.

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., & Perrot, M. (2011). Scikit-learn: Machine

learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Peirano, R., Kristjanpoller, W., & Minutolo, M. C. (2021). Forecasting inflation in Latin American countries using a SARIMA–LSTM combination. *Soft Computing, 25(16),* 10851–10862. https://doi.org/10.1007/s00500-021-06016-5

Peña Ordóñez, A. C. (2019). Pronóstico de la Inflación Colombiana: una aproximación desde un modelo Arima desagregado y Machine Learning [Tesis de pregrado, Universidad de Los Andes]. *Repositorio Institucional Séneca.* https://repositorio.uniandes.edu.co/entities/publication/56a8464b-b8e2-4ac3-818e-871a97712841

Perkel, J. M. (2016). GitHub: The software that builds software. *Nature, 538*(7623), 127–128. https://doi.org/10.1038/538127a

Pinto, A. (1970). Naturaleza e implicaciones de la "heterogeneidad estructural" de la América Latina. *El Trimestre Económico, 37*(145), 83–100. https://www.jstor.org/stable/20856116

Prebisch, R. (1996). El desarrollo económico de la América Latina y alguno de sus principales problemas. *El Trimestre Económico, 63*(249), 175–245. https://www.jstor.org/stable/45406431

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Bayazi, M. J. D., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2023). Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv*. http://arxiv.org/abs/2310.08278

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika, 71*(3), 599–607. https://doi.org/10.1093/biomet/71.3.599

Saravanan, H. K., Dwivedi, S., Praveen, P., & Arjunan, P. (2024). Analyzing the Performance of Time Series Foundation Models for Short-term Load Forecasting. *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation,* 346–349. https://doi.org/10.1145/3671127.3699536

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y., & Ji, R. (2021). RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv*. https://doi.org/10.48550/arXiv.2104.09864

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace Independent Publishing Platform

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS 2017), 30*, 5998–6008. https://arxiv.org/pdf/1706.03762

Xavier, A. L. S., Fernandes, B. J. T., & de Oliveira, J. F. L. (2023). Hybrid Model and Ensemble for Inflation Forecasting: A Machine Learning Approach. *2023 IEEE Latin American Conference on Computational Intelligence, LA-CCI 2023*. https://doi.org/10.1109/LA-CCI58595.2023.10409426

Zhang, B., & Sennrich, R. (2019). Root Mean Square Layer Normalization. *arXiv*. https://doi.org/10.48550/arXiv.1910.07467