Análisis de un Modelo Predictivo con técnicas de Machine Learning, acerca de la inflación pospandemia en Colombia durante 2020–2024

Mónica Andrea Saavedra Crespo

Asesor

Sixyel Jeyson Castañeda Coronado

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Maestría en Ciencia de Datos y Analítica

2026

Agradecimientos

Agradezco profundamente a mi familia, a mis padres y a mis hermanos, por impulsarme, apoyarme y acompañarme, desde siempre en cada etapa de mi vida. Su confianza y respaldo han sido fundamentales para avanzar con determinación en este proceso académico, personal y laboral.

También agradezco a mi par por la energía, al Director asignado del proyecto de grado por su visión y valioso liderazgo, a la economía, ciencia de datos, música, letras y a la pasión por perseverar, porque han sido refugio, inspiración y fuerza en cualquier momento.

Gratitud infinita, además, a la resiliencia, por enseñarme a continuar vigente para transformar los desafíos en aprendizaje y a mantener firme la convicción de alcanzar mis objetivos a corto, mediano y largo plazo.

Este logro representa no solo el resultado de un esfuerzo académico, sino también el reflejo de la colaboración en equipo, la disciplina y la convicción que me han brindado a lo largo del camino.

Resumen

El desarrollo del Proyecto de Grado, está orientado al análisis de un Modelo Predictivo para estimar la trayectoria del Índice de Precios al Consumidor (IPC) en Colombia, durante el periodo 2020–2024. La investigación se fundamenta en series oficiales del Departamento Administrativo Nacional de Estadística (DANE) y del Banco de la República, con datos mensuales comprendidos entre 2007 y 2024.

En la versión final del estudio, la variable objetivo se define como IPC puro, entendido como el índice en nivel observado mensualmente. Por tanto, no se utilizan transformaciones interanuales, variaciones porcentuales, ni tasas de inflación derivadas, como variable dependiente de los modelos.

Se implementaron y compararon nueve Modelos Predictivos en Python: Regresión Lineal Simple, ARIMA Univariado, Ridge Univariado, Ridge Multivariado, Regresión Lineal Múltiple, Gradient Boosting Regressor, ExtraTreesRegressor, Suavización Exponencial de Holt y Naive Forecast o Caminata Aleatoria.

Los modelos fueron evaluados mediante las métricas como Coeficiente de Determinación (R^2), Error Cuadrático Medio (RMSE) y Error Absoluto Medio (MAE), sobre un esquema de validación temporal con Entrenamiento entre 2007 y 2019 y Prueba entre 2020 y 2024.

Los resultados muestran que el Naive Forecast, obtuvo el mejor desempeño global fuera de muestra, con R^2_test = 0.9963, RMSE_test = 0.8926 y MAE_test = 0.7247, evidenciando una elevada persistencia temporal del IPC. Entre los modelos de Machine Learning, el Ridge Univariado presentó el mejor desempeño, alcanzando R^2_test = 0.9611, RMSE_test = 2.8800 y MAE_test = 2.2752. El Ridge Multivariado también obtuvo resultados destacados, mientras que los modelos basados en árboles, la Regresión Lineal Simple, ARIMA y la Suavización

Exponencial de Holt mostraron una menor capacidad de generalización frente a los cambios observados durante el periodo pospandemia.

El estudio concluye que la persistencia temporal, constituye el principal factor predictivo del IPC puro y que las variables macroeconómicas, complementan la interpretación económica del fenómeno inflacionario. En consecuencia, el Naive Forecast se consolida como el benchmark predictivo más preciso, mientras que el Ridge Univariado, se identifica como el mejor modelo de Machine Learning desarrollado en la investigación.

Palabras clave**:** IPC, inflación, colombia, pandemia, modelado.

Abstract

This thesis focuses on the analysis of a predictive model to estimate the trajectory of the Consumer Price Index (CPI) in Colombia during the 2020–2024 period. The research is based on official time series from the National Administrative Department of Statistics (DANE) and the Central Bank of Colombia, with monthly data covering the period from 2007 to 2024.

At the end of the study, the target variable is defined as the raw CPI, understood as the index at its monthly observed level. Therefore, no year-over-year transformations, percentage changes, or derived inflation rates are used as the dependent variable in the models.

Nine predictive models were implemented and compared in Python: Simple Linear Regression, Univariate ARIMA, Univariate Ridge, Multivariate Ridge, Multiple Linear Regression, Gradient Boosting Regressor, ExtraTreesRegressor, Holt's Exponential Smoothing, and Naive Forecast (or Random Walk).

The models were evaluated using metrics such as the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), based on a time-series validation scheme with training from 2007 to 2019 and testing from 2020 to 2024.

The results show that the Naive Forecast achieved the best overall out-of-sample performance, with R^2_test = 0.9963, RMSE_test = 0.8926, and MAE_test = 0.7247, demonstrating high temporal persistence in the CPI. Among the machine learning models, the Univariate Ridge model performed best, achieving R^2_test = 0.9611, RMSE_test = 2.8800, and MAE_test = 2.2752. The Multivariate Ridge model also yielded outstanding results, while tree-based models, Simple Linear Regression, ARIMA, and Holt's Exponential Smoothing showed a lower capacity for generalization in the face of the changes observed during the post-pandemic period.

The study concludes that time persistence is the main predictor of the pure CPI and that macroeconomic variables complement the economic interpretation of inflation. Consequently, the Naive Forecast emerges as the most accurate predictive benchmark, while the Univariate Ridge Model is identified as the best Machine Learning Model developed in this research.

Keywords**:** CPI, inflation, colombia, pandemic, modeling.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Introducción

La inflación es uno de los fenómenos macroeconómicos de mayor relevancia para la estabilidad económica, la planeación de política pública y el bienestar de los hogares. En Colombia, el periodo posterior a la pandemia por Covid-19 estuvo marcado por incrementos significativos en el nivel de precios, presiones cambiarias, choques de oferta, ajustes en la política monetaria y modificaciones en los costos laborales y de producción. Estas condiciones evidencian la necesidad de contar con herramientas analíticas que permitan anticipar la trayectoria del Índice de Precios al Consumidor (IPC) y apoyar el seguimiento técnico de la coyuntura económica.

El presente proyecto se desarrolla bajo el enfoque CRISP-DM y utiliza datos oficiales mensuales para evaluar diferentes Modelos Predictivos implementados en Python. La variable objetivo se define como el IPC puro, es decir, el índice en nivel reportado mensualmente, sin transformarlo en variación interanual ni en porcentaje. Esta decisión metodológica, permite modelar directamente la trayectoria acumulada del índice y evitar confusiones entre el nivel del IPC y las tasas de inflación derivadas del mismo.

La investigación compara nueve modelos de distinta naturaleza. En primer lugar, se incluyen Modelos Univariados, basados en la información histórica del propio IPC, representados por la Regresión Lineal Simple, el ARIMA Univariado, el Ridge Univariado, la Suavización Exponencial de Holt y el Naive Forecast o Caminata Aleatoria.

En segundo lugar, se implementan Modelos Multivariados que incorporan variables macroeconómicas como el Producto Interno Bruto (PIB), la Tasa de Intervención del Banco de la República, el Salario Mínimo Mensual Legal Vigente (Smmlv) y la Tasa Representativa del Mercado (TRM), representados por el Ridge Multivariado y la Regresión Lineal Múltiple.

Finalmente, se evalúan Modelos de Ensamble no lineales basados en árboles de decisión, correspondientes a Gradient Boosting Regressor y ExtraTreesRegressor.

La comparación de estos enfoques permite analizar las diferencias entre modelos clásicos de Series de Tiempo, modelos lineales, técnicas de regularización, algoritmos de aprendizaje automático y benchmark predictivos. Asimismo, permite determinar si metodologías más complejas, generan mejoras significativas frente a estrategias simples de predicción basadas en la persistencia temporal del IPC.

La evaluación se realiza mediante validación temporal estricta, utilizando datos de Entrenamiento, correspondientes al periodo 2007–2019 y datos de Prueba para el periodo 2020–2024. Las métricas empleadas son el Coeficiente de Determinación (R^2), el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE), todas interpretadas en unidades del IPC puro. Adicionalmente, se analizan gráficos de predicción, residuos y resultados comparativos para identificar fortalezas, limitaciones y capacidad de generalización de cada modelo.

Finalmente, el estudio busca contribuir al análisis económico de la inflación pospandemia en Colombia, mediante la integración de herramientas de Ciencia de Datos, técnicas de Machine Learning y métodos tradicionales de Series Temporales, proporcionando evidencia empírica sobre la capacidad predictiva de diferentes enfoques para estimar la evolución futura del IPC.

Justificación

La inflación pospandemia ha afectado la capacidad adquisitiva de los hogares, las decisiones de inversión, la planeación empresarial y el diseño de política económica. En Colombia, esta situación se ha relacionado con choques de oferta, variaciones en el tipo de cambio, ajustes de política monetaria y presiones sobre alimentos y energía. Comprender esta dinámica resulta fundamental para anticipar episodios de aceleración del Índice de Precio al Consumidor (IPC) y proponer herramientas de monitoreo temprano.

El uso de Modelos Predictivos basados en Ciencia de Datos, es pertinente porque permite comparar enfoques estadísticos y de aprendizaje automático bajo un esquema reproducible. En particular, los modelos regularizados como Ridge permiten controlar la multicolinealidad entre variables macroeconómicas, mientras que los Modelos de Ensamble, permiten explorar relaciones no lineales. La comparación entre ellos aporta evidencia técnica para seleccionar un modelo robusto, interpretable y útil en el seguimiento del IPC.

Objetivos

Objetivo General

Analizar un Modelo Predictivo, basado en técnicas de Machine Learning que favorezca la estimación de tendencias del IPC puro en Colombia durante 2020-2024.

Objetivos Específicos

Examinar el comportamiento del IPC durante 2020-2024 con datos oficiales del Departamento Administrativo Nacional de Estadística (DANE) y del Banco de la República, realizando limpieza y análisis exploratorio de datos que permita la identificación de tendencias, rupturas y variables relevantes.

Comparar el desempeño de nueve Modelos implementados en Python: Regresión Lineal Simple, Regresión Lineal Múltiple, Ridge Univariado, Ridge Multivariado, ARIMA, Gradient Boosting Regressor, ExtraTreesRegressor, Suavización Exponencial de Holt y Naive Forecast, según métricas como R^2 (Coeficiente de Determinación), RMSE (Raíz del Error Cuadrático Medio) y MAE (Error Absoluto Medio), que permita la selección con mejores resultados.

Interpretar los resultados a la luz de la coyuntura macroeconómica y que favorezca la formulación de recomendaciones técnicas de seguimiento y política pública.

Problemática

Planteamiento del Problema

El impacto de la inflación pospandemia ha sido un fenómeno global con repercusiones relevantes en economías avanzadas y emergentes. La pandemia de Covid-19 generó disrupciones en las cadenas de suministro, afectó la producción y distribución de bienes esenciales y contribuyó al aumento generalizado de precios. Bonam y Smădu (2021), señalan que las pandemias históricamente han generado efectos prolongados sobre la inflación, aunque la magnitud de dichos efectos depende de la respuesta de política monetaria.

En América Latina, las presiones inflacionarias se agravaron por factores internos y externos, entre ellos políticas fiscales expansivas, encarecimiento de materias primas, devaluación de monedas locales y mayores costos de transporte. Cuevas Ahumada y Perrotini Hernández (2024), explican que muchas economías implementaron paquetes de apoyo durante la pandemia, lo cual protegió hogares y empresas, pero también pudo presionar la demanda agregada.

En Colombia, el fenómeno inflacionario se reflejó en sectores como alimentos, transporte, vivienda, servicios públicos y energía. Pérez Gelves et al. (2023) resaltan la importancia de la depreciación cambiaria y de los efectos de la pandemia sobre los hogares. En este contexto, el problema central consiste en la dificultad para anticipar la trayectoria del IPC en un entorno caracterizado por choques externos, persistencia temporal y cambios de régimen.

Por tanto, la pregunta que orienta la investigación es: ¿cuál de los Modelos Predictivos implementados en Python presenta mejor desempeño para estimar el IPC puro en Colombia durante 2020-2024, según RMSE, MAE y R^2?

Pregunta Problema

¿Cuál de los Modelos Predictivos implementados en Python — Regresión Lineal Simple, Regresión Lineal Múltiple, Ridge Univariado, Ridge Multivariado, ARIMA Univariado, Gradient Boosting Regressor, ExtraTreesRegressor, Suavización Exponencial de Holt y Naive Forecast o Caminata Aleatoria — presenta el mejor desempeño para estimar tendencias del IPC puro en Colombia durante 2020-2024?

Alcance. El estudio se concentra en el IPC total nacional de Colombia y en su relación con variables macroeconómicas agregadas: Producto Interno Bruto (PIB), Tasa de Intervención del Banco de la República, Salario Mínimo y Tasa Representativa del Mercado (TRM). La cobertura temporal comprende datos mensuales entre enero de 2007 y diciembre de 2024, con énfasis analítico en el periodo de Prueba 2020-2024. El producto esperado es la selección de un Modelo Predictivo con mejor desempeño fuera de muestra, acompañado de recomendaciones técnicas para el monitoreo de la inflación medida a través del IPC.

Marco Referencial

Antecedentes

La literatura reciente coincide en que la inflación pospandemia respondió a una combinación de cuellos de botella en la oferta, políticas monetarias y fiscales expansivas, cambios en los precios internacionales de materias primas, transmisión cambiaria y recomposición de la demanda agregada.

Diversos estudios señalan que las interrupciones en las cadenas globales de suministro, el incremento en los costos de transporte internacional y el aumento de los precios de alimentos y energía contribuyeron significativamente al repunte inflacionario observado a partir de 2021.

En América Latina, la recuperación económica convivió con aumentos persistentes de precios, lo cual intensificó la vulnerabilidad de los hogares con menor capacidad adquisitiva y generó importantes desafíos para la política monetaria de los bancos centrales.

En el caso colombiano, la inflación posterior a la pandemia estuvo asociada tanto a factores externos como internos. Entre los factores externos, se destacan la depreciación del peso colombiano frente al dólar, el incremento de los precios internacionales de alimentos, combustibles e insumos productivos y las disrupciones logísticas globales.

Entre los factores internos, sobresalen la recuperación de la demanda, los ajustes del Salario Mínimo, los cambios en las expectativas inflacionarias y las decisiones de política monetaria adoptadas por el Banco de la República para contener las presiones sobre los precios. Como resultado, el IPC alcanzó niveles históricamente elevados durante 2022 y comienzos de 2023, antes de iniciar una fase gradual de desaceleración.

En paralelo, la literatura metodológica ha mostrado que la predicción económica mejora cuando se combinan modelos clásicos de series de tiempo con técnicas de aprendizaje automático y procesos rigurosos de validación fuera de muestra.

Estudios recientes han demostrado que los modelos basados en regularización, como Ridge Regression, permiten manejar de manera más eficiente problemas de multicolinealidad presentes en variables macroeconómicas, en este caso Índice de Precios al Consumidor (IPC), Producto Interno Bruto (PIB), Salario Mínimo, Tasa de Intervención del Banco de la República y Tasa Representativa del Mercado (TRM), mientras que los métodos de Ensamble y los algoritmos de Machine Learning, pueden capturar patrones complejos que no siempre son identificados por los enfoques econométricos tradicionales.

Asimismo, la evidencia empírica señala que la comparación entre Modelos Univariados, Modelos Multivariados y benchmark simples, resulta fundamental para evaluar la verdadera capacidad predictiva de una metodología.

En particular, los modelos basados en persistencia temporal, como el Naive Forecast o caminata aleatoria, suelen constituir referencias exigentes en series económicas altamente autocorrelacionadas.

Por esta razón, las investigaciones más recientes, recomiendan contrastar el desempeño de los modelos avanzados, frente a benchmark sencillos mediante métricas como R^2, RMSE y MAE, garantizando una evaluación objetiva de la capacidad de generalización de los Modelos Predictivos.

Tabla 1

Antecedentes del Estudio

Antecedente	Aporte al estudio	Relación con el proyecto
Bonam y Smădu (2021)	Explican efectos persistentes de pandemias sobre inflación y política monetaria.	Sustenta la noción de inflación pospandemia y persistencia inflacionaria.
Cuevas Ahumada y Perrotini Hernández (2024)	Analizan inflación de bienes y servicios en América Latina durante Covid-19.	Apoya la lectura regional y la importancia del contexto macroeconómico.
Pérez Gelves et al. (2023)	Relacionan pandemia, vulnerabilidad social y condiciones económicas de los hogares.	Refuerza la importancia de choques externos, energía y condiciones de ingreso.
Carlomagno et al. (2023)	Proponen medidas y evaluación de inflación subyacente.	Respalda la necesidad de medición rigurosa y comparación de modelos.

Nota. Literatura investigada en Revistas Científicas.

Marco Teórico

La literatura internacional muestra que la inflación pospandemia ha sido una consecuencia directa de las disrupciones en las cadenas globales de suministro y el aumento de la demanda tras el confinamiento. Bonam & Smădu (2021), evidencian que las pandemias históricamente han producido efectos prolongados sobre la inflación y que las respuestas monetarias tardías, pueden amplificar su impacto (p. 2). Zhao et al. (2022), encontraron que existe una correlación significativa entre el alza de los precios de los alimentos y los ciclos de contagio por Covid-19, lo cual también afecta la estabilidad macroeconómica.

En el contexto latinoamericano, Cuevas Ahumada & Perrotini Hernández (2024), destacan que la inflación en bienes y servicios básicos ha sido agravada por políticas fiscales expansivas, implementadas durante la pandemia (p. 4). En Colombia, Baquero Beltrán et al. (2022), señalan que, a pesar de la estabilidad monetaria antes del Covid-19, la política económica se enfrentó a desafíos para controlar el aumento de precios (p. 171). Pérez Gelves et al. (2023), subrayan que la devaluación del peso colombiano, frente al dólar ha sido uno de los factores claves que ha contribuido al encarecimiento de productos importados y al deterioro del ingreso real de los hogares (p. 7).

Aunque los estudios específicos a nivel local son limitados, Pérez Gelves et al. (2023), también abordan el impacto de la pandemia sobre la pobreza energética, identificando cómo el aumento de precios en servicios básicos ha afectado especialmente a los hogares de bajos ingresos en zonas urbanas del país (p. 6). Además, Bargain & Aminjonov (2021) señalan que las restricciones sanitarias tuvieron un impacto desproporcionado sobre los trabajadores informales, quienes predominan en muchas regiones colombianas (p. 7).

Marco Conceptual

Inflación Pospandemia

Fenómeno económico caracterizado por el aumento sostenido de precios tras la crisis sanitaria del Covid-19. Se ha visto agravada por disrupciones en las cadenas de suministro (Cuevas Ahumada & Perrotini Hernández, 2024, p.4).

Política Monetaria y Fiscal

Estrategias implementadas por los gobiernos y Bancos Centrales para regular la inflación. La política fiscal expansiva ha contribuido al aumento de los precios (Bonam & Smădu, 2021, p. 2).

Devaluación de la Moneda

Pérdida de valor del peso colombiano frente al dólar, generado por cambios en la oferta y demanda de divisas (Pérez Gelves et al., 2023, p.7).

Machine Learning en Predicción Económica

"El Aprendizaje Automático es una rama de la informática, cuyo objetivo es capacitar a los ordenadores para aprender nuevos comportamientos a partir de datos empíricos" (Tantawi, R., 2024, p. 1).

Impacto de la Inflación en el Mercado Laboral

Relación entre el aumento de precios y el incremento en la informalidad laboral, derivado de la pérdida de poder adquisitivo y la reducción del empleo formal (Torres-Favela & Luna, 2025, p. 10).

Cálculo de la Inflación a partir del IPC y Canasta de Productos Evaluados por el DANE

El Índice de Precios al Consumidor (IPC), es una operación estadística del Departamento Administrativo Nacional de Estadística (DANE). Su objetivo es medir la variación promedio de precios de una canasta representativa de bienes y servicios adquiridos y consumidos por los hogares.

El IPC funciona como indicador general de inflación, pero conceptualmente debe distinguirse entre el IPC puro y la inflación derivada del índice. El IPC puro corresponde al número índice publicado en nivel. La inflación mensual, anual o año corrido, se calcula a partir de variaciones entre valores del índice (Departamento Administrativo Nacional de Estadística [DANE], 2019, pp. 4, 7).

El DANE aclara que el IPC no mide el valor monetario de una canasta familiar, sino la evolución de precios de una canasta de seguimiento; compuesta por bienes y servicios finales de

consumo de los hogares. La operación estadística excluye gastos que no constituyen consumo final, como inversiones, ahorro, impuestos directos, contribuciones obligatorias y pagos de deuda (DANE, 2019, pp. 7, 13).

La Metodología del IPC colombiano, se basa en la lógica de índices de canasta fija, en los cuales se mantiene una estructura de gasto de referencia para concentrar la medición en los cambios de precios. De forma teórica, esta lógica puede representarse, mediante un índice tipo Laspeyres:

$$IPC_t = \frac{\sum_{i=1}^{n} p_{i,t}\, q_{i,0}}{\sum_{i=1}^{n} p_{i,0}\, q_{i,0}} \times 100 \qquad (1)$$

Fuente: DANE (2019, pp. 8-10).

En la ecuación (1), IPC_t representa el índice en el periodo t; $p_{i,t}$ es el precio del bien o servicio i en el periodo actual; $p_{i,0}$ es el precio del bien o servicio en el periodo base; $q_{i,0}$ corresponde a la estructura de cantidades o ponderaciones de referencia y n, representa el número de bienes y servicios incluidos en la canasta.

El cálculo operativo del IPC inicia con la recolección de precios de variedades específicas en fuentes informantes. Una variedad corresponde a una observación concreta de un artículo, con características comparables como marca, presentación, unidad de medida, fuente y calidad. El primer cálculo elemental corresponde al relativo simple de precios:

$$R_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} \qquad (2)$$

Fuente: DANE (2019, p. 34).

En la ecuación (2), $R_{i,t}$ es el relativo simple de precios de la variedad i en el periodo t; $P_{i,t}$ es el precio actual; y $P_{i,t-1}$ es el precio del periodo anterior. Después de obtener estos relativos, el DANE agrega información mediante promedios geométricos para niveles elementales:

$$G_{a,t} = \left(\prod_{i=1}^{n_a} R_{i,t} \right)^{1/n_a} \qquad (3)$$

Fuente: Procedimiento de promedio geométrico descrito por DANE (2019, p. 34).

En niveles superiores, los índices se agregan mediante promedios aritméticos ponderados, donde cada componente tiene un peso relativo asociado al gasto de los hogares:

$$IPC_t = \sum_{j=1}^{m} w_j \, I_{j,t} \qquad (4)$$

Fuente: elaboración propia con base en la agregación ponderada descrita por DANE (2019, pp. 34-36).

A partir del IPC puro pueden calcularse indicadores de inflación. La variación mensual compara el índice del mes actual con el mes anterior; la variación año corrido compara el índice actual con diciembre del año anterior; y la variación anual compara el índice actual con el mismo mes del año anterior. Estas fórmulas son útiles para análisis económico, pero en los modelos finales de este proyecto la variable dependiente es el IPC en nivel, no una variación porcentual (DANE, 2019, pp. 36-37).

$$VM_t = \left(\frac{IPC_t}{IPC_{t-1}} - 1 \right) \times 100 \qquad (5)$$

$$VAC_t = \left(\frac{IPC_t}{IPC_{dic\ año\ anterior}} - 1\right) \times 100 \qquad (6)$$

$$VAN_t = \left(\frac{IPC_t}{IPC_{t-12}} - 1\right) \times 100 \qquad (7)$$

Fuente: DANE (2019, pp. 36-37).

El Departamento Administrativo Nacional de Estadística (DANE), organiza los bienes y servicios del IPC, mediante la clasificación Clasificación del Consumo Individual por Finalidades (COICOP), que permite ordenar el gasto de consumo de los hogares por finalidad. La canasta de seguimiento no corresponde a una lista mínima de subsistencia, sino a una selección de artículos representativos del gasto promedio de los hogares (DANE, 2019, pp. 22, 26-30).

Tabla 2

Productos y Servicios Evaluados por el DANE para Estimar IPC

División COICOP del IPC	**Productos y servicios evaluados**
Alimentos y bebidas no alcohólicas	Cereales, arroz, pan, carnes, pollo, pescado, leche, huevos, frutas, verduras, legumbres, aceites, azúcar, café, agua, gaseosas y jugos.
Bebidas alcohólicas y tabaco	Cerveza, aguardiente, vino, otras bebidas alcohólicas y cigarrillos.
Prendas de vestir y calzado	Ropa para hombre, mujer, niños y bebés, uniformes, accesorios, zapatos, tenis y reparación de prendas o calzado.
Alojamiento, agua, electricidad, gas y otros combustibles	Arriendo efectivo, arriendo imputado, servicios públicos, agua, electricidad, gas, aseo, alcantarillado y combustibles del hogar.
Muebles, artículos para el hogar y conservación ordinaria del hogar	Muebles, colchones, electrodomésticos, utensilios de cocina, productos de aseo, herramientas y artículos de mantenimiento menor.
Salud	Medicamentos, productos farmacéuticos, consultas médicas, servicios odontológicos, exámenes y servicios paramédicos.
Transporte	Transporte urbano e intermunicipal, vehículos, combustibles, mantenimiento, bicicletas, peajes y servicios de movilidad.

División COICOP del IPC	**Productos y servicios evaluados**
Alimentos y bebidas no alcohólicas	Cereales, arroz, pan, carnes, pollo, pescado, leche, huevos, frutas, verduras, legumbres, aceites, azúcar, café, agua, gaseosas y jugos.
Bebidas alcohólicas y tabaco	Cerveza, aguardiente, vino, otras bebidas alcohólicas y cigarrillos.
Prendas de vestir y calzado	Ropa para hombre, mujer, niños y bebés, uniformes, accesorios, zapatos, tenis y reparación de prendas o calzado.
Información y comunicación	Telefonía fija y móvil, internet, equipos telefónicos y servicios de comunicación.
Recreación y cultura	Libros, artículos recreativos, servicios culturales, artículos deportivos, equipos de recreación y paquetes turísticos.
Educación	Educación preescolar, primaria, secundaria, técnica, tecnológica, universitaria, posgrados y educación no formal.
Restaurantes y hoteles	Comidas y bebidas fuera del hogar, restaurantes, cafeterías, hoteles, pensiones y servicios de alojamiento.

División COICOP del IPC	**Productos y servicios evaluados**
Alimentos y bebidas no alcohólicas	Cereales, arroz, pan, carnes, pollo, pescado, leche, huevos, frutas, verduras, legumbres, aceites, azúcar, café, agua, gaseosas y jugos.
Bebidas alcohólicas y tabaco	Cerveza, aguardiente, vino, otras bebidas alcohólicas y cigarrillos.
Prendas de vestir y calzado	Ropa para hombre, mujer, niños y bebés, uniformes, accesorios, zapatos, tenis y reparación de prendas o calzado.
Bienes y servicios diversos	Cuidado personal, peluquería, productos de aseo personal, seguros, servicios financieros, guarderías y otros servicios personales.

Nota. Estructura COICOP y descripción metodológica del IPC del DANE.

Conceptos Clave

Tabla 3

Glosario de Términos

Concepto	Definición operativa en el estudio
IPC puro	Índice de Precios al Consumidor en nivel mensual. Es la variable objetivo de los modelos finales.
Inflación pospandemia	Aumento sostenido del nivel de precios posterior a la crisis sanitaria del Covid-19, asociado a choques de oferta, demanda, tipo de cambio y política monetaria.
Política monetaria	Conjunto de decisiones del Banco de la República orientadas a regular condiciones monetarias y expectativas de inflación.
Transmisión cambiaria	Mecanismo por el cual variaciones en la TRM afectan precios internos, especialmente bienes importados e insumos externos.
Machine Learning en predicción económica	Uso de algoritmos supervisados para aprender patrones históricos y estimar una variable continua de interés.
Persistencia del IPC	Relación entre el valor actual del IPC y sus valores pasados, asociada a inercia inflacionaria e indexación.

Nota. Significados-Diccionario.

Diseño Metodológico

Diseño de Investigación

La investigación adopta la metodología CRISP-DM. El propósito es comprender el comportamiento del IPC en Colombia y aplicar técnicas predictivas basadas en Ciencia de Datos. La base consolidada contiene series mensuales oficiales de 2007 a 2024. La validación temporal se realiza con Entrenamiento de 2007-2019 y Prueba de 2020-2024.

Comprensión del Negocio

Se define el uso del modelo de pronóstico del IPC, como herramienta de alerta temprana y apoyo técnico para el seguimiento de la inflación medida a través del índice. El objetivo no es reemplazar el análisis macroeconómico, sino complementarlo con evidencia predictiva reproducible.

Adquisición de Datos

Se integraron series oficiales del Departamento Administrativo Nacional de Estadística (DANE) para el Índice de Precios al Consumidor (IPC) y del Banco de la República (BanRep) para Producto Interno Bruto (PIB), Tasa Representativa del Mercado (TRM), Tasa de Intervención del Banco de la República (BanRep) y Salario Mínimo. La frecuencia de análisis es mensual y la cobertura temporal, comprende enero de 2007 a diciembre de 2024.

Para fortalecer la replicabilidad, se documenta que las series fueron consolidadas en una base mensual para el periodo enero de 2007 a diciembre de 2024. El IPC proviene del DANE; el PIB, la TRM, la Tasa de Intervención y el Salario Mínimo, provienen del Banco de la República. La frecuencia de trabajo se homologó a periodicidad mensual; cuando una variable no varía mensualmente, como el Salario Mínimo, se mantiene el valor vigente de forma repetitiva para cada mes del año correspondiente.

Trazabilidad y Procedencia de Datos

La información utilizada para la construcción del Modelo Predictivo, proviene de fuentes oficiales del DANE y del BanRep de Colombia.

Los datos correspondientes al Índice de Precios al Consumidor (IPC), fueron obtenidos directamente del DANE en formato de número índice (IPC puro), mediante respuesta oficial a una Petición, Queja, Reclamo o Solicitud (PQRS), recibida el 6 de Noviembre de 2025. Esta información permitió trabajar con el nivel mensual del índice de precios, evitando el uso de variaciones porcentuales, interanuales o tasas de inflación derivadas.

Por su parte, las variables macroeconómicas utilizadas como variables explicativas — Producto Interno Bruto (PIB), Tasa de Intervención del Banco de la República (Tasa BanRep) y Tasa Representativa del Mercado (TRM)— fueron obtenidas del Banco de la República de Colombia. La descarga oficial de estas series, se realizó el 23 de Septiembre de 2025, a través de los portales estadísticos institucionales y bases de datos públicas disponibles en dicha entidad.

Asimismo, durante la misma fecha se recopiló la información histórica correspondiente al Salario Mínimo Mensual Legal Vigente (Smmlv), consolidando una base de datos mensual para el periodo comprendido entre Enero de 2007 y Diciembre de 2024.

La utilización de fuentes oficiales garantiza la calidad, confiabilidad, consistencia y trazabilidad de los datos empleados en el estudio, permitiendo la reproducibilidad del proceso analítico y fortaleciendo la validez de los resultados obtenidos.

Tabla 4

Variables, Fuentes y Fecha de Descarga

Variable	Fuente	Fecha
IPC (Índice de Precios al Consumidor)	DANE (respuesta oficial PQRS)	06/11/2025
Producto Interno Bruto (PIB)	Banco de la República	23/09/2025
Tasa de Intervención del BanRep	Banco de la República	23/09/2025
Tasa Representativa del Mercado (TRM)	Banco de la República	23/09/2025
Salario Mínimo	Banco de la República / normativa oficial	23/09/2025

Nota. Variables usadas.

Preparación de Datos

La preparación incluyó limpieza de nombres de columnas, estandarización de fechas, conversión de variables numéricas, imputación básica de datos faltantes y corrección de escala del Salario Mínimo. En la versión final, la variable objetivo se define exclusivamente como IPC puro. Por tanto, no se emplea IPC interanual, variación porcentual ni transformación como

variable dependiente. Para algunos Modelos se incorporan rezagos del IPC puro, tendencia temporal y variables macroeconómicas en nivel.

En el Análisis Exploratorio de Datos (EDA), correlaciones y gráficos, las variables se mantuvieron en su escala original. Para Modelos regularizados como Ridge, el escalamiento de las variables independientes resulta recomendable, debido a que el PIB, Salario Mínimo, TRM y Tasa BanRep, se encuentran en escalas distintas. No obstante, la variable dependiente no se escaló: el objetivo siempre fue el IPC puro, por lo que RMSE y MAE, se interpretan en puntos del índice y no en porcentaje.

El proceso incluyó verificación de columnas, depuración de espacios en los nombres de variables, conversión de FECHA a formato mensual, ordenamiento cronológico, tipificación numérica de IPC, PIB, TASA_BANREP, SALARIO_MINIMO y TRM, y revisión de rangos para detectar errores de escala. En particular, el Salario Mínimo se conservó en pesos colombianos completos, evitando interpretarlo como decimal. Para variables con registros faltantes o valores no válidos se aplicó imputación básica, mediante propagación hacia adelante y hacia atrás, solo cuando fue necesario para mantener continuidad mensual. Esta decisión se justifica, porque las series macroeconómicas analizadas son temporales y el uso de valores cercanos preserva la continuidad del periodo de estudio sin alterar la variable objetivo.

Modelado

En esta fase se implementaron y compararon diferentes enfoques predictivos utilizando Python, con el propósito de identificar el Modelo con mejor capacidad para estimar el comportamiento del Índice de Precios al Consumidor (IPC) en Colombia, durante el periodo pospandemia. Los Modelos desarrollados se clasificaron en tres categorías: Modelos

Univariados, Modelos Multivariados y Modelos de Ensamble, incorporando técnicas de Machine Learning y métodos clásicos de Series de Tiempo.

Los Modelos Predictivos implementados fueron los siguientes:

Ridge Univariado con rezagos del IPC. Utilizado para capturar la persistencia temporal de la serie mediante información histórica del propio índice.

ARIMA Univariado. Modelo clásico de series temporales, basado en componentes autorregresivos, diferenciación y medias móviles, empleado para modelar la dependencia temporal del IPC.

Regresión Lineal Simple. Utilizada como línea base Univariada para representar la evolución del IPC, mediante una tendencia temporal lineal.

Ridge Multivariado. Incorpora variables macroeconómicas como PIB, Tasa de Intervención del Banco de la República, Salario Mínimo Mensual Legal Vigente y Tasa Representativa del Mercado (TRM).

Regresión Lineal Múltiple. Empleada como línea base multivariada para evaluar relaciones lineales entre el IPC y las variables explicativas.

Gradient Boosting Regressor. Modelo de Ensamble basado en árboles de decisión, construido mediante aprendizaje secuencial.

ExtraTreesRegressor. Modelo de Ensamble que genera múltiples árboles aleatorios, con el fin de reducir la varianza y mejorar la capacidad predictiva.

Suavización Exponencial de Holt. Método clásico de Series de Tiempo, utilizado como benchmark univariado para modelar tendencia.

Naive Forecast o Caminata Aleatoria. Incorporado como benchmark de referencia, cuyo pronóstico corresponde al último valor observado del IPC.

La evaluación de los modelos, se realizó utilizando las métricas R^2, RMSE y MAE sobre el conjunto de Prueba. El Coeficiente de Determinación (R^2), permitió medir la proporción de variabilidad explicada por cada modelo, mientras que el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE), cuantificaron la magnitud promedio de los errores de predicción. Todas las métricas se interpretaron en unidades del IPC puro, facilitando la comparación directa entre modelos.

La estrategia de validación respetó la naturaleza temporal de la información, utilizando una partición cronológica con datos de Entrenamiento para el periodo 2007–2019 y datos de Prueba para el periodo 2020–2024. Este enfoque evitó la contaminación temporal entre entrenamiento y evaluación, permitiendo medir la capacidad de generalización de los modelos sobre observaciones futuras no utilizadas durante el aprendizaje.

Adicionalmente, como complemento metodológico, se planteó la implementación de esquemas de validación temporal mediante backtesting, rolling window y expanding window para futuras investigaciones. Estas metodologías permiten evaluar la estabilidad de los resultados a través de múltiples ventanas temporales y verificar que el desempeño observado no dependa exclusivamente de una única partición Entrenamiento-Prueba.

En los Modelos de Machine Learning, se documentaron los principales hiperparámetros utilizados. En Ridge, el parámetro alpha controla la intensidad de la penalización L2, aplicada a los coeficientes, cuando los predictores están altamente correlacionados. En Gradient Boosting Regressor, los hiperparámetros n_estimators, learning_rate, max_depth y subsample, regulan la complejidad del Modelo y el proceso de aprendizaje secuencial.

En ExtraTreesRegressor, los parámetros n_estimators, max_depth y min_samples_leaf, controlan el número de Árboles, su profundidad máxima y el tamaño mínimo de las hojas

terminales. La documentación de estas configuraciones, mejora la replicabilidad del estudio y permite justificar las decisiones metodológicas adoptadas durante el proceso de modelado.

Finalmente, la comparación conjunta de los nueve modelos, permitió identificar diferencias importantes entre enfoques lineales, Modelos de Ensamble y métodos clásicos de Series Temporales. Los resultados mostraron que el Naive Forecast, obtuvo el mejor desempeño predictivo general, evidenciando la elevada persistencia temporal del IPC en nivel. No obstante, el Ridge Univariado, se consolidó como el Modelo de Machine Learning con mejor capacidad de generalización fuera de muestra, alcanzando los mejores resultados entre los Modelos Predictivos avanzados evaluados en la investigación.

Modelos Implementados

Ridge Univariado (rezagos de IPC)

Descripción. El Modelo Ridge Univariado con rezagos del IPC puro, corresponde a una Regresión Lineal regularizada, mediante penalización L2 que ayuda a estabilizar los coeficientes, cuando los predictores están altamente correlacionados. Además, cuyo objetivo es predecir el valor actual del Índice de Precios al Consumidor en nivel, es decir, el IPC puro, a partir de sus propios valores pasados.

En este Modelo no se utiliza IPC interanual, no se calcula variación porcentual y no se transforma la variable objetivo. Por tanto, la variable dependiente corresponde directamente al valor del índice IPC observado en cada mes.

Descripción Teórica Detallada. El Modelo Ridge Univariado es una técnica de aprendizaje supervisado para regresión que utiliza únicamente la información histórica de la misma variable objetivo. En este caso, la predicción del IPC se realiza a partir de rezagos del propio IPC puro, tales como $\mathbf{IPC}_{t-1}$, $\mathbf{IPC}_{t-3}$, $\mathbf{IPC}_{t-6}$ e $\mathbf{IPC}_{t-12}$

Su fundamento parte de la persistencia temporal de los índices de precios, debido a que el IPC es una medida acumulativa del nivel de precios, su comportamiento actual suele estar fuertemente relacionado con sus valores anteriores.

A diferencia de una Regresión Lineal Ordinaria, Ridge incorpora una penalización L2 sobre el tamaño de los coeficientes. Esta penalización permite reducir la varianza del estimador y estabilizar el Modelo cuando los rezagos del IPC están altamente correlacionados entre sí, lo cual es habitual en series económicas que presentan tendencia. En este sentido, Ridge no elimina variables, sino que contrae sus coeficientes hacia cero para evitar pesos excesivos y mejorar la capacidad de generalización fuera de muestra.

En el contexto del proyecto, este Modelo sirve como un enfoque Univariado de Machine Learning, ya que no incorpora variables macroeconómicas externas como PIB, Tasa BanRep, Salario Mínimo o TRM. Su utilidad principal consiste en evaluar qué tan bien puede predecirse el IPC únicamente con su propia memoria histórica.

Ecuación del Modelo. La especificación general del Modelo Ridge Univariado, con rezagos del IPC puro se expresa como:

$$\mathbf{IPC}_t = \boldsymbol{\alpha} + \sum_{\boldsymbol{k}\in\boldsymbol{K}} \boldsymbol{\beta}_{\boldsymbol{k}} \cdot \mathbf{IPC}_{\boldsymbol{t-k}} + \boldsymbol{\varepsilon}_t$$

Donde el conjunto de rezagos utilizados es:

$$\boldsymbol{K} = \{\mathbf{1}, \mathbf{3}, \mathbf{6}, \mathbf{12}\}$$

De forma expandida, la ecuación puede escribirse así:

$$\mathbf{IPC}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}_1\mathbf{IPC}_{t-1} + \boldsymbol{\beta}_3\mathbf{IPC}_{t-3} + \boldsymbol{\beta}_6\mathbf{IPC}_{t-6} + \boldsymbol{\beta}_{12}\mathbf{IPC}_{t-12} + \boldsymbol{\varepsilon}_t$$

Penalización Ridge. El Modelo Ridge, estima los coeficientes minimizando una función de pérdida que combina el error cuadrático con una penalización L2 sobre los coeficientes, cuando los predictores están altamente correlacionados:

$$\hat{\beta}^{Ridge} = \underset{\alpha,\beta}{\operatorname{argmin}} \left[\sum_{t=1}^{n} \left(\mathrm{IPC}_t - \alpha - \sum_{k \in K} \beta_k \, \mathrm{IPC}_{t-k} \right)^2 + \lambda \sum_{k \in K} \beta_k^2 \right]$$

En esta expresión, λ es el parámetro de regularización. Cuando λ aumenta, los coeficientes se contraen hacia cero con mayor fuerza, reduciendo el riesgo de sobreajuste.

Explicación Breve de la Ecuación. Esta expresión indica que el valor actual del IPC puro, se estima a partir de una constante y de una combinación lineal de sus propios rezagos de 1, 3, 6 y 12 meses. La penalización Ridge reduce el riesgo de sobreajuste al limitar el tamaño de los coeficientes asociados a cada rezago.

¿Qué Significa Cada Parte de la Ecuación?. IPC_t representa el valor observado del Índice de Precios al Consumidor en el periodo t.

α corresponde al intercepto o constante del modelo.

La sumatoria sobre $k \in K$ indica que se agregan los efectos de diferentes rezagos del IPC.

β_k representa el coeficiente asociado al rezago k.

IPC_{t-k} corresponde al valor del IPC observado k meses antes.

$K = \{1,3,6,12\}$ define los rezagos utilizados en el modelo.

ε_t representa el término de error, es decir, la parte del IPC actual que no logra ser explicada por sus propios rezagos.

λ representa la intensidad de la penalización Ridge.

Variable Dependiente del Modelo. La variable objetivo utilizada en este **m**odelo es el IPC puro:

$$Y = \text{IPC}_t$$

Por tanto, el modelo se entrena directamente sobre el nivel del índice de precios. No se utiliza variación interanual, ni variación porcentual.

Resultados. R^2_train: 0.9955. R^2_test: 0.9611. RMSE_test: 2.8800. MAE_test: 2.2752.

ARIMA Univariado

Descripción. El Modelo AutoRegressive Integrated Moving Average (ARIMA), se utiliza como un método clásico de series temporales para modelar y pronosticar el comportamiento del Índice de Precios al Consumidor (IPC) en nivel.

Descripción Teórica Detallada. Los Modelos ARIMA, permiten capturar la dependencia temporal de una serie mediante componentes autorregresivos, de diferenciación y de medias móviles. Constituyen un benchmark clásico frente a los modelos de Machine Learning.

Ecuación del Modelo. $IPC_{\text{t}} = \alpha + \sum_{\text{i}=1}^{\text{p}} \varphi_{\text{i}}\, IPC_{\text{t}-\text{i}} + \sum_{\text{j}=1}^{\varphi} \theta_{\text{j}}\, \varepsilon_{\text{t}-\text{j}} + \varepsilon_{\text{t}}$

Explicación Breve de la Ecuación. El IPC actual, se estima utilizando observaciones pasadas del propio índice y errores de predicción de periodos anteriores.

Resultados. R^2_train: 0.7838. R^2_test: 0.1479. RMSE_test: 13.4763. MAE_test: 9.9080.

Interpretación. El Modelo ARIMA captura parcialmente la dinámica temporal del IPC, pero presenta una capacidad predictiva limitada durante el periodo pospandemia, debido a cambios estructurales y aceleraciones inflacionarias.

Regresión Lineal Simple

Descripción. La Regresión Lineal Simple, se utiliza como una línea base estadística para modelar la evolución del IPC puro a partir del tiempo.

Descripción Teórica Detallada. Este modelo supone que existe una relación lineal entre el IPC y la tendencia temporal. Su principal ventaja es la interpretabilidad, aunque puede ser insuficiente para capturar cambios no lineales.

Ecuación del Modelo. $IPC_t = \alpha + \beta t + \varepsilon_t$

Explicación Breve de la Ecuación. La ecuación supone que el IPC evoluciona, siguiendo una tendencia lineal constante a lo largo del tiempo.

Resultados. R^2_train: 0.9738. R^2_test: -0.1880. RMSE_test: 15.9123. MAE_test: 12.1878.

Interpretación. Aunque presenta un alto ajuste en Entrenamiento, el desempeño fuera de muestra es insuficiente. Evidenciando que, una tendencia lineal simple, no logra capturar adecuadamente la complejidad del IPC.

Ridge Multivariado (t, PIB, Tasa BanRep, Salario Mínimo, TRM)

Descripción. El Modelo Ridge Multivariado, corresponde a una Regresión Lineal regularizada, mediante penalización L2 que ayuda a estabilizar los coeficientes, cuando los predictores están altamente correlacionados. Su objetivo es predecir el valor actual del Índice de Precios al Consumidor en nivel, es decir, el IPC puro, a partir de una tendencia temporal y de variables macroeconómicas explicativas.

En esta especificación, no se utiliza IPC interanual, no se calcula variación porcentual y no se transforma la variable objetivo. Por tanto, la variable dependiente corresponde directamente al valor del índice IPC observado en cada mes.

El modelo incorpora como variables independientes la tendencia temporal, el PIB, la Tasa de Intervención del Banco de la República, el Salario Mínimo y la TRM. Estas variables se integran en el modelo, con el propósito de representar diferentes canales económicos asociados con la evolución del nivel de precios.

Descripción Teórica Detallada. El Ridge Multivariado amplía la lógica del Modelo Univariado, al incluir variables macroeconómicas que pueden influir en la inflación desde distintos canales. La tendencia temporal recoge cambios persistentes de largo plazo; el PIB aproxima condiciones de actividad económica y demanda agregada; la Tasa de Intervención representa la postura de política monetaria; el Salario Mínimo refleja presiones de costos e ingreso nominal; y la TRM captura la transmisión cambiaria y el encarecimiento de bienes importados.

Desde el punto de vista estadístico, este Modelo resulta adecuado cuando las variables explicativas están correlacionadas entre sí, como ocurre habitualmente en el análisis macroeconómico. La regularización L2 ayuda a evitar que la multicolinealidad, vuelva inestables los coeficientes y permite conservar el aporte conjunto de las covariables sin inflar artificialmente su peso individual. Por ello, el Ridge Multivariado, ofrece un balance entre capacidad predictiva, estabilidad de estimación e interpretabilidad económica.

En el contexto del Proyecto, puede entenderse como un Modelo de Machine Learning Supervisado para Regresión, ya que aprende la relación entre un conjunto de variables predictoras y una variable objetivo continua. Aunque parte de una forma lineal, la regularización lo diferencia de la Regresión Lineal Múltiple ordinaria y le permite mejorar la estabilidad de los coeficientes, frente a variables con escalas distintas o alta correlación.

Definición de Variables. La variable dependiente es:

$$Y = IPC_t$$

Las variables independientes son:

$$X = \{t, PIB_t, TASA_t, SAL\_MIN_t, TRM_t\}$$

Donde IPC_t corresponde al IPC puro observado en el periodo t; t representa la tendencia temporal; PIB_t representa el Producto Interno Bruto; $TASA_t$ representa la Tasa de Intervención del Banco de la República; SAL_MIN_t representa el Salario Mínimo y TRM_t representa la Tasa Representativa del Mercado.

Ecuación del Modelo. La forma funcional del Modelo Ridge Multivariado, se expresa como:

$$\boldsymbol{IPC_t = \alpha + \beta_1 t + \beta_2 PIB_t + \beta_3 TASA_t + \beta_4 SAL\_MIN_t + \beta_5 TRM_t + \varepsilon_t}$$

También puede escribirse en forma compacta como:

$$\boldsymbol{IPC_t = \alpha + \sum_{j=1}^{5} \beta_j X_{j,t} + \varepsilon_t}$$

Función Objetivo Ridge. A diferencia de la Regresión Lineal Múltiple ordinaria, Ridge estima los coeficientes minimizando una función de pérdida con penalización L2 que ayuda a estabilizar los coeficientes, cuando los predictores están altamente correlacionados:

$$\boldsymbol{\widehat{\beta}^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{t=1}^{n} \left(IPC_t - \alpha - \sum_{j=1}^{5} \beta_j X_{j,t} \right)^2 + \lambda \sum_{j=1}^{5} \beta_j^2 \right\}}$$

Esta expresión indica que el Modelo minimiza simultáneamente el error de predicción y el tamaño de los coeficientes. El primer término mide la diferencia entre el IPC observado y el IPC estimado y el segundo término, corresponde a la penalización L2 que ayuda a estabilizar los

coeficientes, cuando los predictores están altamente correlacionados, la cual está controlada por el parámetro λ.

Explicación Breve de la Ecuación. La ecuación indica que el IPC puro del periodo actual, se modela como función de una constante, una tendencia temporal y un conjunto de variables macroeconómicas agregadas: PIB, Tasa de Intervención, Salario Mínimo y TRM. La regularización L2, reduce el riesgo de sobreajuste y ayuda a controlar la multicolinealidad entre predictores, favoreciendo una estimación más robusta.

¿Qué Significa Cada Parte de la Ecuación?. $\boldsymbol{IPC_t}$ representa el valor observado del Índice de Precios al Consumidor en el periodo $\boldsymbol{t}$. La constante $\boldsymbol{\alpha}$ corresponde al intercepto del Modelo. El coeficiente $\boldsymbol{\beta_1}$ mide el efecto asociado a la tendencia temporal. El coeficiente $\boldsymbol{\beta_2}$ corresponde al PIB, que aproxima la actividad económica. El coeficiente $\boldsymbol{\beta_3}$ representa la Tasa de Intervención del Banco de la República.

El coeficiente $\boldsymbol{\beta_4}$ se asocia con el Salario Mínimo, entendido como una variable relacionada con costos laborales e ingresos nominales. El coeficiente $\boldsymbol{\beta_5}$ corresponde a la TRM, que recoge posibles efectos del tipo de cambio. Finalmente, ε_t representa el término de error, es decir, la parte del IPC que no logra ser explicada por las variables incluidas.

En la función objetivo, λ es el parámetro de regularización. Cuando λ aumenta, los coeficientes se contraen con mayor fuerza hacia cero. Esto no elimina variables, pero reduce pesos excesivos y mejora la estabilidad del modelo.

Resultados. R_train^2: 0.99554. R_test^2: 0.93524. RMSE_test: 3.71520. MAE_test: 3.33286.

Estos resultados deben interpretarse en unidades del IPC puro. Por tanto, el RMSE y el MAE, indican la distancia promedio entre el IPC observado y el IPC predicho en puntos del índice; no en porcentaje, ni en variación interanual.

Interpretación. El valor de $R^2_{train} = 0.99554$ indica que el Modelo logra un ajuste muy alto sobre el periodo de Entrenamiento. Esto sugiere que la combinación de tendencia temporal y variables macroeconómicas, explica gran parte de la trayectoria histórica del IPC puro.

El valor de $R^2_{test} = 0.93524$ muestra que el Modelo también conserva una alta capacidad de generalización en el periodo de Prueba. Este resultado es relevante, porque indica que el Modelo no solo se ajusta al pasado, sino que, también logra aproximar adecuadamente el comportamiento del IPC en datos no observados durante el Entrenamiento.

El $RMSE_{test} = 3.71520$ y el $MAE_{test} = 3.33286$, reflejan errores moderados en unidades del índice. Estos valores indican que las predicciones se alejan del IPC real en aproximadamente 3 a 4 puntos del índice, lo cual puede considerarse razonable en un contexto donde el IPC puro, presenta una trayectoria acumulativa y creciente.

Regresión Lineal Múltiple

Descripción. El Modelo de Regresión Lineal Múltiple, se utiliza como una línea base explicativa y predictiva para estimar el comportamiento del Índice de Precios al Consumidor (IPC) en nivel. En esta versión corregida, la variable dependiente corresponde al IPC puro, es decir, el índice observado mensualmente, sin transformaciones interanuales, sin porcentajes, sin variación mensual y sin transformación log-diff.

Las variables independientes incluidas son el Producto Interno Bruto (PIB), la Tasa de Intervención del Banco de la República (TASA_BANREP), el Salario Mínimo (SALARIO_MINIMO) y la Tasa Representativa del Mercado (TRM). Estas variables se

incorporan en nivel según su fuente y escala oficial, de acuerdo con la estructura del código actualizado.

Descripción Teórica Detallada. La Regresión Lineal Múltiple, permite modelar la relación lineal entre una variable dependiente y un conjunto de variables explicativas. En este caso, el objetivo es predecir el valor mensual del IPC puro a partir de variables macroeconómicas contemporáneas. El PIB aproxima las condiciones de actividad económica y demanda agregada; la Tasa BanRep representa la postura de política monetaria; el Salario Mínimo recoge presiones de costos laborales e ingreso nominal; y la TRM captura el canal cambiario y el posible encarecimiento de bienes importados.

Desde el punto de vista estadístico, el modelo supone que el IPC puede aproximarse mediante una combinación lineal de sus predictores. Su principal ventaja es la interpretabilidad, ya que permite analizar el signo y magnitud de la asociación entre cada variable macroeconómica y el IPC.

Sin embargo, al tratarse de un modelo lineal contemporáneo, puede no capturar efectos rezagados, estacionalidad, cambios de régimen o relaciones no lineales. Por esta razón, se interpreta como un modelo base multivariado para comparar posteriormente con Ridge u otros Modelos de Machine Learning.

En particular, valores como $433.700 o $1.000.000 deben interpretarse como 433700 y 1000000, respectivamente y no como valores decimales. Esto permite que la predicción, quede en una escala coherente con el IPC real.

Ecuación del Modelo. La variable objetivo, se define como:

$$\boldsymbol{Y_t = \text{IPC}_t}$$

El conjunto de predictores, se define como:

$$X_t = \{\text{PIB}_t,\ \text{TASA\_BANREP}_t,\ \text{SALARIO\_MINIMO}_t,\ \text{TRM}_t\}$$

La ecuación de la Regresión Lineal Múltiple es:

$$\text{IPC}_t = \alpha + \beta_1 \cdot \text{PIB}_t + \beta_2 \cdot \text{TASA\_BANREP}_t + \beta_3 \cdot \text{SALARIO\_MINIMO}_t + \beta_4 \cdot \text{TRM}_t + \varepsilon_t$$

Tabla 5

Variables, Significado y Descripción

Variable	Nombre	Descripción
IPC	Índice de Precios al Consumidor en nivel	Variable dependiente (IPC puro) utilizada como objetivo de predicción.
PIB	Producto Interno Bruto	Variable independiente asociada a la actividad económica nacional.
TASA_BANREP	Tasa de Intervención del Banco de la República	Variable independiente que representa la política monetaria implementada por el Banco de la República.
SALARIO_MINIMO	Salario Mínimo Mensual Legal Vigente	Variable independiente relacionada con los costos laborales e ingresos nominales de los hogares.
TRM	Tasa Representativa del Mercado	Variable independiente que refleja el comportamiento del tipo de cambio peso colombiano/dólar estadounidense.

Nota. El IPC en nivel (IPC puro), se utilizó como variable objetivo del estudio. Las variables explicativas corresponden a indicadores macroeconómicos que permiten contextualizar el comportamiento inflacionario durante el periodo 2007–2024.

Explicación Breve de la Ecuación. La ecuación expresa que el valor actual del IPC puro, se estima como una combinación lineal de cuatro variables macroeconómicas: PIB, Tasa BanRep, Salario Mínimo y TRM. Cada coeficiente indica el cambio esperado en el IPC, manteniendo constantes las demás variables. La constante representa el nivel base estimado por el Modelo, mientras que el término de error recoge la parte del IPC que no logra ser explicada por las variables incluidas.

¿Qué Significa Cada Parte de la Ecuación?. IPC_t representa el valor observado del Índice de Precios al Consumidor en el periodo t. La constante α corresponde al intercepto del modelo. Los coeficientes β1, β2, β3 y β4 representan la relación estimada entre cada variable independiente y el IPC. PIB_t representa la actividad económica; $TASA\_BANREP_t$ representa la postura de política monetaria; $SALARIO\_MINIMO_t$ representa el Salario Mínimo vigente y TRM_t, representa el tipo de cambio oficial. Finalmente, ε_t es el término de error o componente no explicado por el modelo.

Resultados. R^2_train: 0.9910. R^2_test: 0.9117. RMSE_test: 4.3392. MAE_test: 3.6997.

Estos resultados indican que el modelo, presenta un alto ajuste en Entrenamiento y una capacidad de generalización favorable en el periodo de Prueba 2020-2024. El R^2_test de 0.9117, sugiere que el Modelo explica una proporción alta de la variabilidad del IPC puro fuera de muestra. El RMSE_test y el MAE_test, deben interpretarse en unidades del índice IPC. Por lo tanto, el error promedio se encuentra en torno a 3.70 puntos del IPC y el error cuadrático medio alrededor de 4.34 puntos.

A diferencia del resultado anterior con problemas de escala, esta versión corregida produce predicciones coherentes con el rango del IPC real. Por ello, el modelo puede considerarse una línea base multivariada válida para el proyecto, siempre que se mantenga la

limpieza correcta de las variables y se recuerde que la relación estimada es contemporánea. Para mejorar la especificación, se recomienda evaluar versiones con rezagos, regularización Ridge y Modelos ARIMAX o SARIMAX.

Gradient Boosting Regressor (GBR)

Descripción. Ensamble aditivo secuencial que corrige errores residuales, mediante Árboles débiles, usando como variable dependiente el IPC puro y como variables explicativas PIB, Tasa BanRep, Salario Mínimo y TRM.

Descripción Teórica Detallada. Gradient Boosting Regressor es un método de ensamble secuencial que construye una predicción agregada a partir de múltiples Árboles de decisión pequeños. La idea central es que cada árbol nuevo se entrena para corregir los errores residuales del conjunto anterior. En lugar de estimar una única relación global, el modelo aprende patrones complejos mediante la suma gradual de funciones base débiles, optimizando una función de pérdida por descenso del gradiente.

En el contexto del IPC puro, este enfoque permite capturar relaciones no lineales e interacciones entre variables macroeconómicas como PIB, Tasa BanRep, Salario Mínimo y TRM. Sin embargo, al tratarse de un modelo basado en árboles, su capacidad para extrapolar valores fuera del rango observado en el Entrenamiento puede ser limitada. Esto es relevante porque el IPC en el periodo 2020-2024 alcanza niveles superiores a los observados en 2007-2019.

Ecuación del Modelo.

$$F_m(x_t) = F_{m-1}(x_t) + \nu \cdot h_m(x_t)$$

$$\widehat{\mathrm{IPC}}_t = F_M(x_t)$$

Hiperparámetros Usados. n_estimators = 300, learning_rate = 0.05, max_depth = 3, subsample = 0.9, random_state = 42.

Variables Incluidas. PIB, Tasa BanRep, Salario Mínimo y TRM. La variable objetivo es el IPC puro.

Explicación Breve de la Ecuación. La formulación resume un Ensamble aditivo en el que cada árbol nuevo, corrige parte del error residual del conjunto anterior. La predicción final del IPC puro, surge de sumar muchas contribuciones pequeñas ajustadas secuencialmente.

¿Qué Significa Cada Parte de la Ecuación?. $F_m(x_t)$ es la predicción acumulada en la iteración m; $F_{m-1}(x_t)$ es el ensamble construido hasta la iteración anterior; ν es la tasa de aprendizaje; $h_m(x_t)$ es el árbol débil de la iteración m; $\widehat{\text{IPC}}_t$ es la predicción final del IPC puro; $F_M(x_t)$ es el modelo ensamblado después de M iteraciones; y x_t es el vector de variables macroeconómicas del periodo t.

Resultados. R^2_train: 0.9997. R^2_test: -2.1762. RMSE_test: 26.0188. MAE_test: 21.8823. Rango IPC predicho en Test: 95.4776 a 103.0684.

Interpretación. El modelo presenta un ajuste extremadamente alto en entrenamiento, pero un desempeño deficiente en prueba. El valor negativo de R^2_{test} indica que el modelo predice peor que una referencia basada en la media del periodo de prueba. Esto ocurre porque el modelo aprende muy bien el rango histórico 2007-2019, pero no logra extrapolar la aceleración del IPC observada entre 2020 y 2024.

En consecuencia, el Gradient Boosting Regressor funciona como referencia de Machine Learning, pero no como Modelo final para el IPC puro, si no se incorporan tendencia temporal, rezagos del IPC y rezagos de las variables macroeconómicas.

ExtraTreesRegressor

Descripción. Ensamble de árboles extremadamente aleatorizados con agregación de predicciones, usando como variable dependiente el IPC puro y como predictores PIB, Tasa BanRep, Salario Mínimo y TRM.

Descripción Teórica Detallada. ExtraTreesRegressor pertenece a la familia de métodos de ensamble basados en árboles. A diferencia de otros ensambles, introduce un nivel adicional de aleatorización en la selección de puntos de corte. En lugar de buscar siempre el mejor umbral de partición entre las variables, selecciona cortes con mayor aleatoriedad y luego combina las predicciones de muchos árboles para reducir la varianza total del sistema.

Este enfoque puede capturar no linealidades e interacciones entre variables sin imponer una forma funcional lineal. En el caso del IPC puro, el modelo puede identificar patrones históricos asociados a PIB, Tasa BanRep, Salario Mínimo y TRM. No obstante, al igual que otros modelos basados en árboles, puede tener dificultades para extrapolar tendencias crecientes cuando el periodo de prueba contiene valores de IPC superiores a los observados durante el Entrenamiento.

Ecuación del Modelo.

$$\widehat{\mathrm{IPC}}_t = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x_t)$$

$$x_t = [\mathrm{PIB}_t,\ \mathrm{TASA}_t,\ \mathrm{SALMIN}_t,\ \mathrm{TRM}_t]$$

Hiperparámetros Usados. n_estimators = 500, max_depth = 6, min_samples_leaf = 3, random_state = 42, n_jobs = -1.

Variables Incluidas. PIB, Tasa BanRep, Salario Mínimo y TRM. La variable objetivo es el IPC puro.

Explicación Breve de la Ecuación. La predicción final se obtiene como el promedio de múltiples árboles extremadamente aleatorizados. Cada árbol aporta una estimación parcial del IPC puro y la agregación busca reducir la varianza del modelo.

¿Qué Significa Cada Parte de la Ecuación?. $\widehat{\text{IPC}}_t$ representa la predicción final del IPC puro en el periodo t; B es el número total de árboles del ensamble; la sumatoria de b=1 hasta B representa la suma de las predicciones de todos los árboles; $T_b(x_t)$ es la predicción generada por el árbol b; y x_t es el vector de variables macroeconómicas del periodo t.

Resultados. R²_train: 0.9983. R²_test: -1.8622. RMSE_test: 24.6992. MAE_test: 20.4925. Rango IPC predicho en Test: 97.5682 a 103.1386.

Interpretación. El modelo logra un ajuste muy alto en Entrenamiento, pero su desempeño en Prueba es débil. El R^2_{test} negativo indica que el modelo no generaliza adecuadamente al periodo 2020-2024. Visualmente, esto se explica porque la predicción se mantiene cerca del rango máximo aprendido antes de 2020, mientras el IPC real continúa creciendo hasta valores superiores a 140.

Por lo tanto, ExtraTreesRegressor, puede ser útil como modelo comparativo no lineal, pero requiere variables temporales, rezagos del IPC y rezagos macroeconómicos para mejorar su capacidad predictiva.

Suavización Exponencial de Holt

Parámetros del Modelo. Suavización exponencial doble con tendencia aditiva, sin tendencia amortiguada e inicialización estimada automáticamente. El modelo usa únicamente el IPC puro como serie objetivo.

Descripción Teórica Detallada. La Suavización Exponencial de Holt es un método clásico de series temporales diseñado para modelar series con nivel y tendencia, pero sin

componente estacional explícito. Su lógica consiste en actualizar en cada periodo dos componentes latentes: el nivel suavizado y la tendencia suavizada. De este modo, las observaciones más recientes reciben mayor peso que las antiguas, lo que permite reaccionar gradualmente a cambios en la trayectoria de la serie.

En el contexto del IPC puro, Holt es pertinente porque el índice presenta una tendencia creciente de largo plazo. Sin embargo, al no incorporar variables externas como PIB, Tasa BanRep, Salario Mínimo o TRM, su capacidad puede ser limitada en periodos con choques fuertes o cambios de régimen, como el periodo 2020-2024.

Ecuaciones del Modelo.

$$l_t = \alpha \cdot \text{IPC}_t + (1 - \alpha) \cdot (l_{t-1} + b_{t-1})$$

$$b_t = \beta \cdot (l_t - l_{t-1}) + (1 - \beta) \cdot b_{t-1}$$

$$\widehat{\text{IPC}}_{t+h} = l_t + h \cdot b_t$$

Parámetros Estimados. $\alpha \approx 1.0000$, $\beta \approx 1.0000$, nivel inicial $l_0 \approx 61.0700$ y tendencia inicial $b_0 \approx 0.7300$.

Variables Incluidas. El modelo utiliza únicamente el IPC puro. No incorpora PIB, Tasa BanRep, Salario Mínimo ni TRM de forma directa.

Explicación Breve de las Ecuaciones. Holt separa la serie en dos componentes: nivel y tendencia. La primera ecuación actualiza el nivel suavizado del IPC puro, la segunda actualiza la tendencia, y la tercera combina ambos componentes para generar el pronóstico h periodos adelante.

¿Qué Significa Cada Parte de las Ecuaciones?. IPC_t es el valor observado del índice en el periodo t; l_t es el nivel suavizado actual; α es el parámetro de suavización del nivel; b_t es la tendencia suavizada actual; β es el parámetro de suavización de la tendencia; $\widehat{\text{IPC}}_{t+h}$ es el

pronóstico del IPC puro h periodos adelante; y h representa el horizonte de pronóstico.

Resultados. R²_train: 0.9997. R²_test: 0.0020. RMSE_test: 14.5847. MAE_test: 10.7308. Rango IPC predicho en Test: 104.0600 a 119.4000.

Interpretación. Holt captura muy bien la tendencia histórica del IPC en Entrenamiento, pero en Prueba, apenas logra explicar la variación del periodo 2020-2024. El R^2_{test} cercano a cero, indica que su capacidad predictiva fuera de muestra es limitada. El modelo mantiene una trayectoria creciente y suave, pero no alcanza la aceleración real del IPC observada después de 2021. Por tanto, Holt es útil como benchmark Univariado clásico, pero debe complementarse con Modelos Multivariados o con modelos que incluyan rezagos y variables macroeconómicas.

Naive Forecast o Caminata Aleatoria (Benchmark)

Descripción. El Naive Forecast, también denominado caminata aleatoria, se incorpora como modelo base de comparación para series temporales. Su lógica consiste en estimar el IPC del periodo actual usando el último valor observado del IPC. En una serie mensual, esto equivale a suponer que el mejor pronóstico inmediato para el mes t es el IPC observado en el mes $\{t-1\}$.

Ecuación del Modelo. $\widehat{IPC} = IPC_{\{t-1\}}$

En esta expresión, IPC_t estimado representa el pronóstico del IPC puro para el periodo t, mientras que $IPC_{\{t-1\}}$, corresponde al valor observado del IPC, en el mes inmediatamente anterior. El modelo no usa PIB, Tasa BanRep, Salario Mínimo ni TRM, porque su objetivo es servir como benchmark mínimo, frente a modelos más complejos.

Resultados. R²_train: 0.9990. R²_test: 0.9963. RMSE_test: 0.8926. MAE_test: 0.7247.

Estos resultados muestran que la Caminata Aleatoria, tuvo el mejor desempeño fuera de muestra en el periodo 2020-2024. Este hallazgo no elimina el valor del Ridge Univariado, pero

obliga a interpretar los resultados con mayor rigor: el Ridge Univariado, es el mejor Modelo de Machine Learning, mientras que el Naive Forecast es el mejor benchmark predictivo general.

Interpretación. El desempeño superior del Naive Forecast confirma que el IPC puro presenta una alta persistencia mensual. En otras palabras, el valor del índice de un mes está muy cerca del valor del mes anterior.

Por esta razón, una regla simple basada en el último valor observado puede superar a modelos más complejos cuando la serie tiene una trayectoria acumulativa y suavemente creciente.

Desde el punto de vista académico, este resultado fortalece la discusión crítica del estudio, porque muestra que los Modelos de Machine Learning, deben compararse siempre contra referencias simples antes de considerarse superiores.

Tabla 6

Resultados Finales de Modelos Implementados

Modelo	R^2_train	R^2_test	RMSE_test	MAE_test	Interpretación
Ridge Univariado con rezagos del IPC	0.9955	0.9611	2.8800	2.2752	Mejor Modelo de Machine Learning; captura la persistencia temporal del IPC.
ARIMA Univariado	0.7838	0.1479	13.4763	9.9080	Modelo Clásico de Series Temporales; captura parcialmente la dinámica del IPC, pero presenta limitaciones fuera de muestra.
Regresión Lineal Simple	0.9738	-0.1880	15.9123	12.1878	Línea base univariada; una tendencia lineal simple, no logra representar adecuadamente la dinámica inflacionaria.
Ridge Multivariado (t, PIB, TASA, SAL_MIN, TRM)	0.9955	0.9352	3.7152	3.3329	Mejor alternativa con covariables macroeconómicas.
Regresión Lineal Múltiple	0.9910	0.9117	4.3392	3.6997	Línea base multivariada válida con adecuada capacidad explicativa.
Gradient Boosting Regressor	0.9997	-2.1762	26.0188	21.8823	Evidencia sobreajuste; no extrapola correctamente la aceleración del IPC.

ExtraTreesRegressor	0.9983	-1.8622	24.6992	20.4925	Alto ajuste en Entrenamiento, pero baja capacidad de generalización.
Suavización Exponencial de Holt	0.9997	0.0020	14.5847	10.7308	Benchmark univariado; capta tendencias suaves, pero no cambios bruscos en el IPC.
Naive Forecast	0.9990	0.9963	0.8926	0.7247	Mejor benchmark predictivo general; evidencia la alta persistencia mensual del IPC.

Nota. Todos los resultados se calcularon con IPC puro, como variable dependiente. RMSE_test y MAE_test, están expresados en puntos del índice.

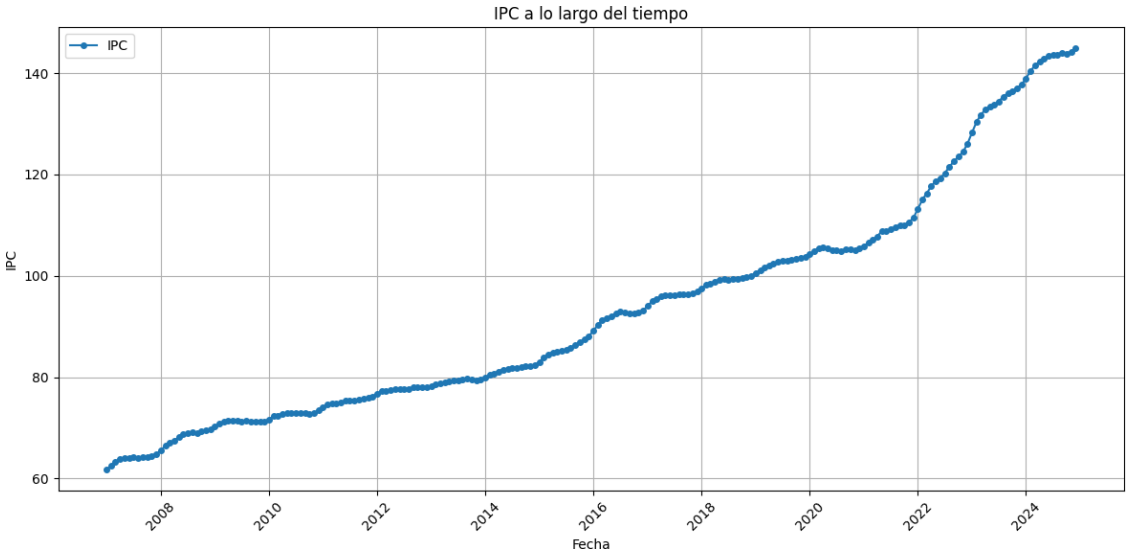
Como se observa en la Tabla 6, el Ridge Univariado presenta el mejor desempeño fuera de muestra, seguido por el Ridge Multivariado. Al incorporar el Naive Forecast, la mejor referencia predictiva general pasa a ser este Modelo de la Caminata Aleatoria. Sin embargo, cabe aclarar que el Ridge Univariado, se conserva como el mejor Modelo de Machine Learning, aunque no supera al Naive en el periodo 2020-2024.

Evidencia Gráfica, Conclusiones y Análisis de Residuos por cada Modelo

Se incorporan las figuras, conclusiones de los gráficos y análisis de residuos de los nueve Modelos implementados.

Figura 1

IPC a lo largo del tiempo



Nota. Tendencia del IPC desde 2008 a 2024.

Análisis

Tendencia Ascendente Sostenida (2007–2024) (2007–2024)

El IPC presenta una trayectoria creciente durante todo el periodo analizado, reflejando la acumulación progresiva de los cambios en el nivel general de precios de la economía colombiana. El gráfico muestra variaciones en la pendiente de crecimiento, pero no caídas significativas del índice.

Crecimiento Acumulado Significativo. El nivel pasa de ~62 (2007) a ~145 (2024), es decir, el costo de la canasta del índice se multiplica ≈2.34× (≈+134% acumulado). El crecimiento promedio simple equivale a un ≈4.7–4.9% anual (orden de magnitud consistente con la inflación media de largo plazo).

Fases de Pendiente. **2007–2014**: Crecimiento gradual y relativamente estable, sin cambios abruptos en la trayectoria del índice.

2015–2016: aceleración moderada del crecimiento, asociada a presiones cambiarias y choques de oferta que incrementaron el ritmo de aumento de los precios.

2017–2019: Etapa de moderación relativa, en la cual el IPC continuó aumentando, aunque a una velocidad menor que en el periodo anterior.

2020: Durante la pandemia, el IPC mantuvo su tendencia creciente, aunque con una pendiente más moderada, debido a la desaceleración económica y a cambios temporales en los patrones de consumo.

2021–2023: Periodo de mayor aceleración del índice. El gráfico muestra la pendiente más pronunciada de toda la serie, reflejando el impacto de factores asociados al contexto pospandemia, incluyendo presiones sobre alimentos, costos de producción, tipo de cambio y ajustes económicos posteriores a la emergencia sanitaria.

2023–2024: Moderación en el ritmo de crecimiento del IPC. Aunque el índice continúa aumentando, la pendiente es menos pronunciada que la observada entre 2021 y 2023, lo que indica una desaceleración del proceso inflacionario, sin que exista una reducción del nivel del índice.

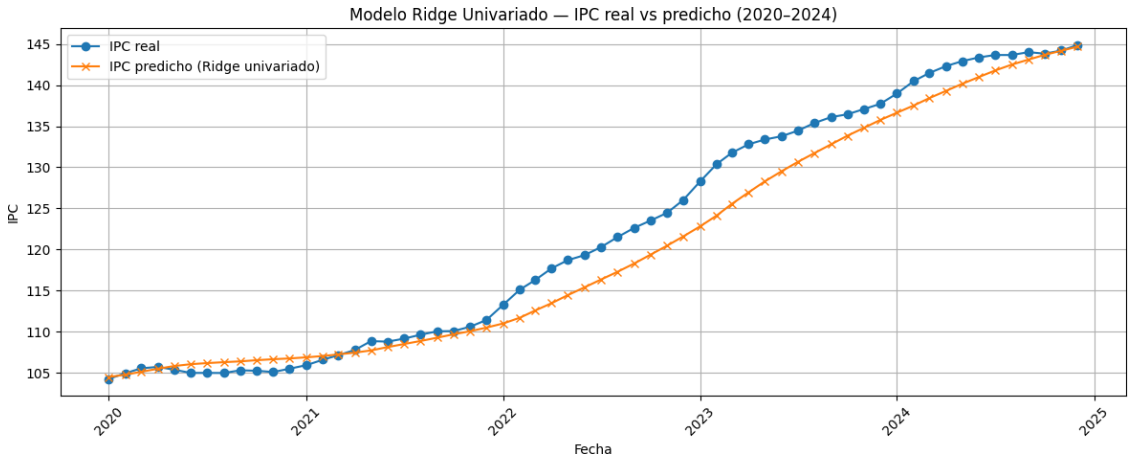
Interpretación. Entre 2007 y 2024, el IPC pasó de 61.8 a 144.88 puntos del índice. Esto implica que el nivel general de precios, se multiplicó aproximadamente 2.34 veces durante el periodo analizado. El comportamiento observado no fue uniforme, ya que la serie presenta fases diferenciadas de aceleración y moderación asociadas a distintos contextos económicos. La marcada continuidad de la trayectoria evidencia una elevada persistencia temporal del IPC, lo que ayuda a explicar el buen desempeño de modelos basados en rezagos del propio índice, como el Ridge Univariado y el Naive Forecast.

Residuos. No aplica análisis de residuos, debido a que esta figura corresponde al análisis exploratorio de la variable objetivo (IPC puro) y no a un Modelo Predictivo estimado.

Parte 1) Modelos Univariados

Figura 2

Modelo Ridge Univariado — IPC Real vs. Predicho (2020–2024)



Nota. Línea azul IPC real y línea naranja IPC predicho.

Análisis

El Modelo Univariado, construido con 12 rezagos del IPC y estimado mediante Regresión Ridge, usó el IPC mensual como función de sus 12 rezagos anteriores. Es decir,

utilizando únicamente la memoria interna de la propia inflación para explicar su comportamiento futuro.

El ajuste sobre el conjunto de Entrenamiento (2008-01 a 2019-12) es muy alto, con un $R^2\_train \approx 0.995$, lo que indica que el Modelo, explica prácticamente toda la variabilidad del IPC en el periodo pre-pandemia.

En el conjunto de Prueba (2020-01 a 2024-12), que corresponde al periodo de interés pospandemia para el Proyecto de Grado, el Modelo mantiene un desempeño muy bueno, con un $R^2\_test \approx 0.96$, es decir, explica alrededor del 96% de la variación observada del IPC, utilizando solo su historial pasado.

Los errores de predicción en el periodo de Prueba, son relativamente bajos frente al nivel del IPC:

RMSE_test ≈ 2.88 puntos de IPC, lo que refleja el error cuadrático medio en unidades de la serie.

MAE_test ≈ 2.28 puntos de IPC, lo que indica que, en promedio, las predicciones se desvían del valor real en poco más de 2 puntos.

Estos resultados confirman que la dinámica auto-regresiva del IPC (su propia historia reciente), contiene información muy poderosa para anticipar su comportamiento futuro, incluso en un contexto complejo como el de la inflación pospandemia.

A pesar de su buen rendimiento, el modelo presenta una limitación estructural importante: al ser estrictamente Univariado, no incorpora el efecto de variables Macroeconómicas clave (Producto Interno Bruto -PIB-, Tasa de Intervención del Banco de la República, Salario Mínimo y Tasa Representativa del Mercado -TRM-), por lo que no permite analizar de forma explícita cómo estos factores externos inciden en la inflación.

En consecuencia, el Modelo Ridge Univariado, se considera una excelente línea base y demuestra que es posible obtener pronósticos precisos del IPC, solo con su historial. Sin embargo, desde la perspectiva del Proyecto de Grado, sirve principalmente como punto de referencia, frente al cual comparar y justificar el uso de un Modelo Ridge Multivariado de Machine Learning que enriquezca la explicación económica de la inflación colombiana en el periodo 2020–2024.

Conclusiones del Gráfico — Modelo Ridge Univariado. En el gráfico de la Figura 2, se observa que el modelo: Reproduce muy bien el comportamiento suave y creciente del IPC al inicio del periodo (2020–2021). Tiende a subestimar ligeramente los picos de inflación en la fase de mayor aceleración (especialmente alrededor de 2022–inicios de 2023), lo que es típico de modelos autoregresivos que suavizan los extremos. Vuelve a alinearse de forma más estrecha con la serie real a medida que la inflación se modera hacia 2024, reduciendo la brecha entre la línea azul (real) y la naranja (predicha).

En conjunto, tanto las métricas cuantitativas (R^2, RMSE, MAE), como la comparación visual del IPC real vs. Predicho, indican que el Modelo Univariado, ofrece pronósticos consistentes y precisos para la inflación colombiana en el periodo 2020–2024, constituyéndose en una línea base sólida para comparar y justificar la posterior incorporación de variables explicativas adicionales en el Modelo Ridge Multivariado de Machine Learning.

El Modelo Ridge Univariado, construido con 12 rezagos del IPC y estimado mediante Regresión Ridge, logra capturar adecuadamente la tendencia creciente de la inflación en el periodo 2020–2024: la curva predicha sigue de cerca la forma general de la serie real.

En términos de ajuste, el modelo presenta un desempeño muy alto: $R^2\_train = 0.9955$, lo que indica que explica alrededor del 99.55% de la variabilidad del IPC en el periodo de

Entrenamiento (2008–2019). R^2_test = 0.9611, es decir, aún en el periodo de Prueba (2020–2024), explica cerca del 96.11% de la variación observada de la inflación pospandemia.

Los errores de pronóstico sobre el periodo de Prueba, se mantienen en niveles reducidos frente al rango del IPC: RMSE_test = 2.88 puntos de IPC, lo que refleja un error cuadrático medio bajo considerando el nivel de la serie. MAE_test = 2.28 puntos de IPC, es decir, en promedio las predicciones se desvían de los valores reales en poco más de 2 puntos del índice.

Tabla 7

Resultados de Residuos del Modelo Ridge Univariado

	IPC_real	IPC_pred	Residuo
144	104.24	104.466589	-0.226589
145	104.94	104.763395	0.176605
146	105.53	105.105311	0.424689
147	105.70	105.475926	0.224074
148	105.36	105.805425	-0.445425

Nota. Resultados obtenidos de los Residuos.

Análisis de Residuos del Modelo Ridge Univariado. El análisis de los residuos del Modelo Ridge Univariado, muestra que las diferencias entre los valores reales del IPC y las predicciones generadas por el modelo son muy pequeñas. Los residuos observados oscilan aproximadamente entre -0.45 y 0.42 puntos del IPC, lo que indica un nivel de error reducido, en comparación con la magnitud total del índice, que durante el periodo analizado se encuentra alrededor de 104 a 106 puntos.

Se observa que los residuos, presentan tanto valores positivos como negativos, lo que sugiere que el modelo, no mantiene un sesgo sistemático hacia la sobreestimación o subestimación del IPC. Por ejemplo, en algunos meses el modelo predice valores ligeramente superiores al IPC real (residuos negativos), mientras que en otros meses ocurre lo contrario (residuos positivos). Esta alternancia alrededor de cero es una señal favorable, ya que indica que los errores se distribuyen de manera relativamente equilibrada.

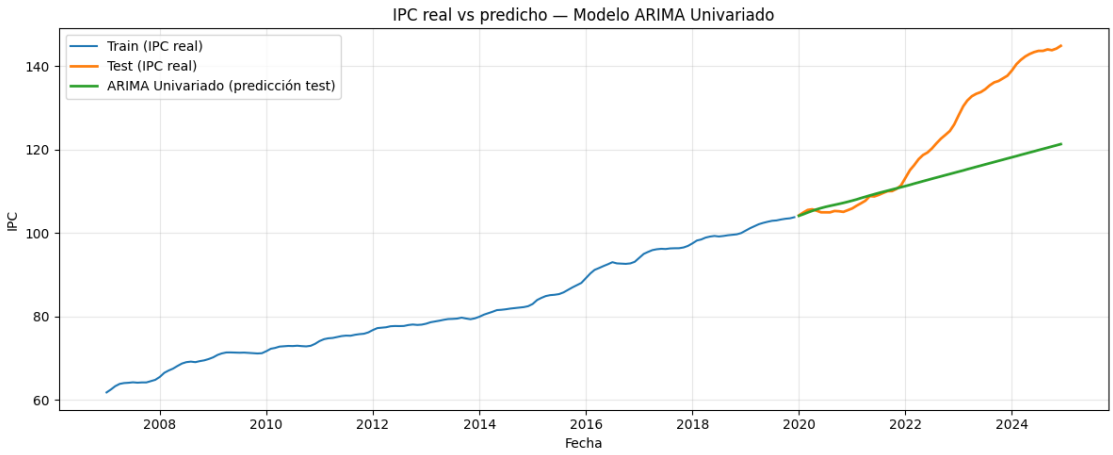
La magnitud reducida de los residuos, confirma los resultados obtenidos previamente, mediante las métricas de desempeño (R^2_test = 0.9611, RMSE_test = 2.8800 y MAE_test = 2.2752), evidenciando que el modelo captura adecuadamente la dinámica temporal del IPC utilizando únicamente sus rezagos históricos.

Desde una perspectiva económica, estos resultados respaldan la hipótesis de que el IPC posee una fuerte persistencia temporal. Es decir, que los valores pasados contienen información suficiente para anticipar gran parte de su comportamiento futuro. La regularización Ridge, contribuye a estabilizar las predicciones y evita que pequeñas fluctuaciones en los rezagos generen errores excesivos.

En conclusión, los residuos del Modelo Ridge Univariado, son consistentes con un ajuste de alta calidad, presentan una dispersión reducida alrededor de cero y no evidencian problemas importantes de sesgo, lo que refuerza la validez del Modelo, como una de las mejores alternativas predictivas evaluadas para estimar el comportamiento del IPC puro en Colombia durante el periodo pospandemia.

Figura 3

Modelo ARIMA Univariado — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde ARIMA Univariado (predicción Test).

Análisis

El Modelo ARIMA Univariado, utiliza como variable objetivo el IPC puro, sin transformaciones interanuales, sin porcentajes y sin variaciones. Esto significa que el modelo trabaja directamente con el índice en nivel observado en cada mes.

Al tratarse de un Modelo Univariado, no incorpora variables macroeconómicas externas como PIB, Tasa BanRep, Salario Mínimo o TRM. En su lugar, el Modelo utiliza la información contenida en la propia historia del IPC, mediante componentes autorregresivos, diferenciación y términos de media móvil.

Si el modelo presenta un R^2 de Entrenamiento alto, esto indica que logra capturar adecuadamente la dinámica histórica del IPC, durante el periodo 2007–2019. Sin embargo, la métrica más importante para evaluar la calidad predictiva es el R^2 de Prueba, ya que este permite medir si el modelo generaliza bien al periodo 2020–2024.

Un R^2 de Prueba bajo o negativo indicaría que el comportamiento histórico del IPC, no es suficiente para explicar la evolución reciente del índice. Esto puede ocurrir porque el periodo 2020–2024 estuvo marcado por cambios estructurales, choques de costos, efectos pospandemia, ajustes de política monetaria, aumentos salariales y variaciones del tipo de cambio.

El RMSE_test y el MAE_test, deben interpretarse en unidades del IPC puro. Por tanto, estos indicadores muestran cuántos puntos del índice se desvía, en promedio, la predicción respecto al IPC real.

En conclusión, el Modelo ARIMA Univariado es útil como benchmark tradicional de series de tiempo, ya que permite evaluar cuánto puede predecirse el IPC usando únicamente su propia memoria histórica. Sin embargo, si el desempeño en prueba es limitado, se justifica avanzar hacia Modelos Multivariados como ARIMAX o SARIMAX, incorporando PIB, Tasa BanRep, Salario Mínimo y TRM.

Conclusiones del Gráfico — Modelo ARIMA Univariado (IPC + t). El gráfico muestra la comparación entre el IPC puro observado y la predicción generada por el Modelo Univariado ARIMA. El eje Y corresponde exclusivamente al IPC en nivel, sin transformación interanual, sin porcentaje y sin variación mensual. Esto permite evaluar directamente si el modelo logra seguir la trayectoria real del índice de precios.

Durante el periodo de Entrenamiento, de 2007 a 2019, el IPC real presenta una trayectoria creciente y relativamente estable. El Modelo ARIMA, aprende esta dinámica

histórica y proyecta una tendencia ascendente hacia el periodo de Prueba. Por esta razón, al inicio de 2020 la predicción se ubica cerca del IPC real, lo cual indica que el modelo parte de una base razonable.

Sin embargo, a partir de 2021 y especialmente desde 2022, la serie real del IPC comienza a crecer con mayor rapidez que la predicción. La línea naranja, correspondiente al IPC observado en el periodo 2020–2024, se separa progresivamente de la línea verde del pronóstico. Esto evidencia que el modelo subestima el crecimiento del IPC en la etapa de mayor aceleración inflacionaria.

La predicción del ARIMA aparece como una trayectoria más suave y lineal, mientras que el IPC real muestra una aceleración más fuerte. Esto indica que el Modelo captura la tendencia histórica del índice, pero no logra adaptarse plenamente al cambio de régimen observado en el periodo pospandemia. Al ser un Modelo Univariado, no incorpora variables externas como PIB, Tasa BanRep, Salario Mínimo y TRM, las cuales pueden explicar parte de los choques macroeconómicos recientes.

Desde el punto de vista visual, la brecha creciente entre la línea real y la predicha confirma que el comportamiento pasado del IPC no fue suficiente para anticipar la dinámica de 2020–2024. El Modelo no reproduce adecuadamente los efectos acumulados de inflación, política monetaria, costos, tipo de cambio y reajustes salariales que afectaron el periodo reciente.

En conclusión, el gráfico evidencia que el ARIMA Univariado con IPC puro + tendencia funciona como un modelo base o benchmark, pero no como un Modelo final de predicción. Su principal aporte es mostrar que la memoria histórica del IPC ayuda a proyectar una tendencia general, pero resulta insuficiente para capturar periodos de aceleración inflacionaria.

Tabla 8

Residuos del Modelo ARIMA Univariado

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	104.135949	0.104051
2020-02-01	104.94	104.521424	0.418576
2020-03-01	105.53	104.924568	0.605432
2020-04-01	105.70	105.316193	0.383807
2020-05-01	105.36	105.675635	-0.315635

Nota. Resultados obtenidos de los Residuos.

Análisis de Residuos del Modelo ARIMA Univariado. El análisis de los residuos del Modelo ARIMA Univariado, muestra que las diferencias entre los valores reales del IPC y las predicciones realizadas por el modelo, son relativamente pequeñas durante los primeros meses del periodo de Prueba. Los residuos observados oscilan entre aproximadamente -0.32 y 0.61 puntos del IPC, lo que indica que el modelo logra capturar de manera razonable la evolución inicial de la serie.

Se observa que durante los primeros cuatro meses los residuos son positivos, lo que significa que el IPC real fue ligeramente superior al valor predicho por el Modelo. En otras palabras, el ARIMA tendió a subestimar el comportamiento del IPC en el inicio del periodo de Prueba. Posteriormente, en Mayo de 2020, aparece un residuo negativo, indicando una ligera sobreestimación del índice. Esta alternancia de signos es una característica deseable, ya que sugiere ausencia de un sesgo permanente en una sola dirección.

La magnitud reducida de estos residuos iniciales evidencia que el Modelo, logra representar adecuadamente la dinámica de corto plazo del IPC. Sin embargo, al analizar los resultados globales del Modelo (R^2_test = 0.0708; RMSE_test = 14.0732; MAE_test = 10.3159), se observa que el desempeño disminuye considerablemente a medida que avanza el horizonte de pronóstico. Esto indica que, aunque el ARIMA captura razonablemente bien los movimientos cercanos al periodo de Entrenamiento, presenta dificultades para adaptarse a los cambios estructurales y a la aceleración inflacionaria observada durante la etapa pospandemia.

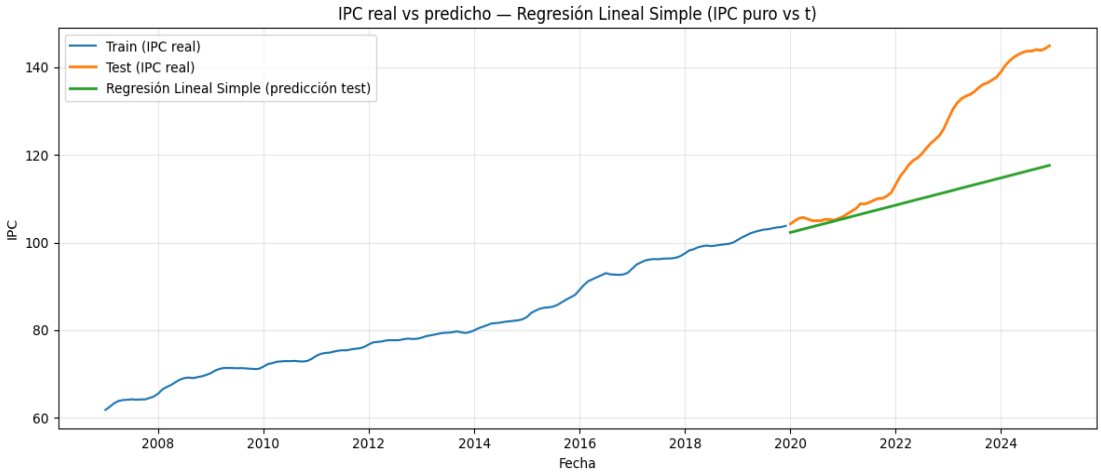
Desde una perspectiva económica, este comportamiento es consistente con la naturaleza del Modelo ARIMA, que basa sus predicciones exclusivamente en patrones históricos de la propia serie. Cuando ocurren eventos extraordinarios, como choques de oferta, cambios en la política monetaria, depreciaciones cambiarias o aumentos significativos de costos, el modelo

puede tardar en reflejar dichos cambios, porque no incorpora variables macroeconómicas externas.

En conclusión, los residuos muestran que el ARIMA Univariado, presenta un buen ajuste local al inicio del periodo de Prueba, pero su capacidad predictiva disminuye en horizontes más largos. Esto explica por qué sus métricas globales, son inferiores a las obtenidas por modelos como Ridge Univariado o Naive Forecast, los cuales lograron representar de manera más efectiva la persistencia temporal del IPC durante el periodo 2020–2024 (fuera de muestra).

Figura 4

Regresión Lineal Simple — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Regresión Lineal Simple (predicción Test).

Análisis

El Modelo de Regresión Lineal Simple: IPC vs tiempo t, utiliza como variable dependiente el IPC puro, es decir, el índice en nivel observado mensualmente. En este modelo no se emplean transformaciones interanuales, porcentuales ni variaciones. Por lo tanto, los resultados se interpretan directamente en unidades del IPC.

La variable explicativa del Modelo es únicamente el tiempo t, por lo que este enfoque busca capturar una tendencia lineal promedio del IPC, a partir del comportamiento histórico observado entre 2007 y 2019. Debido a que el IPC, es un índice acumulativo de precios, es esperable encontrar una relación positiva entre el tiempo y el nivel del IPC.

Si el valor de R^2 en Entrenamiento es alto, esto indica que la tendencia temporal explica una parte importante de la evolución histórica del IPC, durante el periodo 2007–2019. Sin embargo, la métrica más importante para evaluar la capacidad predictiva del modelo es el R^2 en Prueba, ya que muestra si la tendencia aprendida logra generalizar al periodo 2020–2024.

Cuando el R^2 de Prueba es bajo o negativo, se concluye que una tendencia Lineal Simple, no es suficiente para explicar la dinámica reciente del IPC. Esto puede ocurrir, porque el periodo 2020–2024, estuvo marcado por choques económicos relevantes, como la pandemia, la pospandemia, el aumento de costos, ajustes de política monetaria, cambios en el Salario Mínimo y variaciones de la TRM.

El RMSE_test y el MAE_test, deben interpretarse en unidades del IPC puro. Estos indicadores muestran cuántos puntos del índice se desvía, en promedio, la predicción respecto al valor real observado.

En conclusión, la Regresión Lineal Simple funciona como un modelo base o benchmark inicial. Su utilidad principal es mostrar cuánto puede explicarse el IPC, usando únicamente una tendencia temporal. No obstante, no debe considerarse un modelo final para la predicción de inflación, ya que no incorpora variables macroeconómicas como Producto Interno Bruto (PIB), Tasa BanRep, Salario Mínimo y TRM, ni tampoco rezagos o componentes estacionales.

Conclusiones del Gráfico — Modelo Univariado de Regresión Lineal Simple: IPC vs tiempo t. El gráfico muestra la comparación entre el IPC puro observado y la predicción

generada por el Modelo Univariado de Regresión Lineal Simple, donde la única variable explicativa es el tiempo *t*. El eje Y corresponde exclusivamente al IPC en nivel, es decir, el índice original, sin transformación YoY, sin porcentaje y sin variación interanual.

Durante el periodo de Entrenamiento, entre 2007 y 2019, el IPC real presenta una trayectoria creciente y relativamente estable. Esto confirma que el IPC, al ser un índice acumulativo de precios, tiende a aumentar con el paso del tiempo. Por esta razón, el Modelo logra aprender una tendencia lineal general a partir de los datos históricos.

En el periodo de Prueba, de 2020 a 2024, el IPC real continúa creciendo, pero lo hace con una aceleración mucho más fuerte, especialmente desde 2021–2022. La línea naranja del IPC real, se separa progresivamente de la línea verde predicha por el modelo. Esto evidencia que la Regresión Lineal Simple, subestima el IPC real durante buena parte del periodo de Prueba.

La línea verde de predicción muestra un crecimiento suave y casi lineal. Esto indica que el Modelo aprendió una tendencia promedio del periodo 2007–2019, pero no logró capturar el cambio de ritmo observado durante la pandemia y la pospandemia. En otras palabras, el modelo proyecta un crecimiento moderado del IPC, mientras que el índice real crece con mayor intensidad.

La principal limitación visual del modelo es que no incorpora información adicional sobre los factores que afectaron la inflación en el periodo reciente. Variables como Producto Interno Bruto (PIB), Tasa BanRep, Salario Mínimo y TRM, no están incluidas en este Modelo Univariado. Por tanto, el modelo no puede representar adecuadamente choques de costos, cambios de política monetaria, depreciación cambiaria o reajustes salariales.

En conclusión, el gráfico evidencia que el Modelo Univariado de Regresión Lineal Simple, funciona como un modelo base o benchmark inicial, útil para representar la tendencia

general del IPC. Sin embargo, no es suficiente como modelo final de predicción, porque no logra seguir la aceleración del IPC en 2020–2024. Para mejorar la capacidad predictiva, se recomienda avanzar hacia Modelos Multivariados que incluyan variables macroeconómicas y Modelos dinámicos con rezagos o componentes estacionales.

Tabla 9

Residuos de la Regresión Lineal Simple

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	102.282913	1.957087
2020-02-01	104.94	102.542568	2.397432
2020-03-01	105.53	102.802223	2.727777
2020-04-01	105.70	103.061878	2.638122
2020-05-01	105.36	103.321533	2.038467

Nota. Resultados obtenidos de los Residuos.

Análisis de Residuos del Modelo Univariado de Regresión Lineal Simple. El análisis de los residuos del Modelo de Regresión Lineal Simple, evidencia que las predicciones realizadas son sistemáticamente inferiores a los valores reales observados del IPC, durante los primeros meses del periodo de Prueba. Todos los residuos presentados son positivos, con valores que oscilan aproximadamente entre 1.96 y 2.73 puntos del IPC, lo que indica que el modelo está subestimando el comportamiento real del índice.

Esta situación sugiere la existencia de un sesgo sistemático en las predicciones. Mientras el IPC real, se ubica entre 104 y 106 puntos, las estimaciones del modelo permanecen alrededor de 102 y 103 puntos. Esto ocurre, porque la Regresión Lineal Simple, únicamente utiliza la variable tiempo (t) como predictor, asumiendo que el IPC sigue una trayectoria lineal constante a lo largo del tiempo.

Sin embargo, el comportamiento observado del IPC durante el periodo pospandemia, no fue estrictamente lineal. La inflación estuvo influenciada por múltiples factores económicos, entre ellos choques de oferta, cambios en la política monetaria, depreciaciones cambiarias, aumentos en costos de producción y ajustes salariales. Al no incorporar estas variables, ni considerar rezagos de la propia serie, el Modelo no logra adaptarse adecuadamente a los cambios de ritmo observados en la evolución del índice.

La persistencia de residuos positivos, también indica que el crecimiento real del IPC fue más acelerado que el crecimiento estimado por la línea de tendencia aprendida, durante el periodo de Entrenamiento. En otras palabras, la pendiente estimada por la regresión no fue suficiente para capturar la aceleración inflacionaria observada a partir de 2020.

Desde el punto de vista predictivo, estos resultados son consistentes con las métricas globales obtenidas para la Regresión Lineal Simple, las cuales mostraron un desempeño inferior,

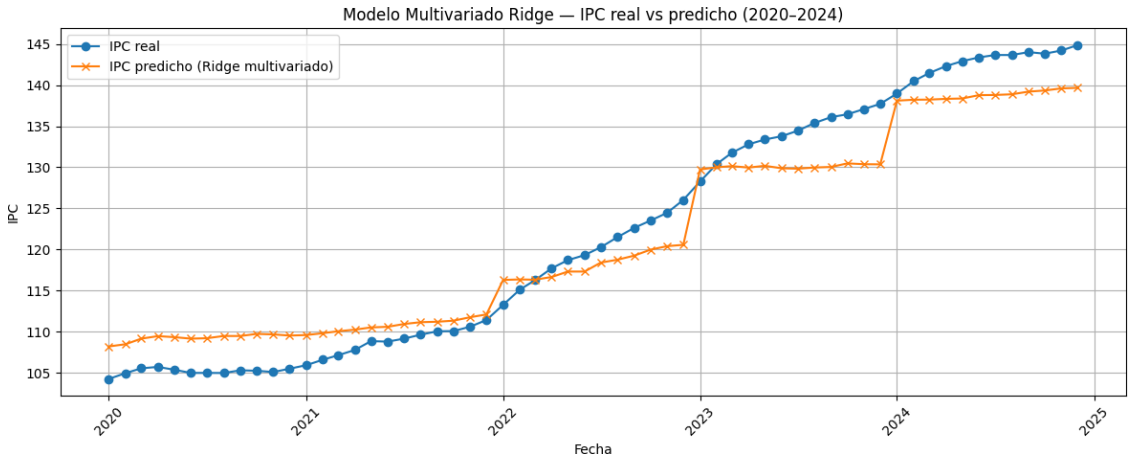
frente a Modelos como Ridge Univariado, Ridge Multivariado y Naive Forecast. Mientras estos modelos incorporan memoria temporal o variables económicas relevantes, la Regresión Lineal Simple se limita a proyectar una tendencia promedio de largo plazo.

En conclusión, los residuos evidencian que la Regresión Lineal Simple, presenta una capacidad limitada para representar la dinámica real del IPC puro, durante el periodo pospandemia. La presencia de errores positivos persistentes, revela una subestimación sistemática del índice, confirmando que una tendencia lineal basada únicamente en el tiempo resulta insuficiente para capturar la complejidad del comportamiento inflacionario observado entre 2020 y 2024.

Parte 2) Modelos Multivariados

Figura 5

Modelo Ridge Multivariado — IPC Real vs. Predicho



Nota. Línea azul IPC real y línea naranja IPC predicho.

Análisis

El Modelo Multivariado con Regresión Ridge, especifica, el IPC mensual como función del tiempo (t) y de variables Macroeconómicas clave: Producto Interno Bruto (PIB), Tasa de Intervención del Banco de la República, Salario Mínimo y TRM, estimado mediante una

Regresión Ridge que incorpora regularización para controlar la multicolinealidad y el sobreajuste.

En el conjunto de Entrenamiento -Train- (periodo previo a 2020), el Modelo presenta un alto poder explicativo, con un R^2_train = 0.995542, lo que indica que explica aproximadamente el 99.55% de la variabilidad del IPC, antes de la pandemia a partir de las variables incluidas.

En el conjunto de Prueba -Test- (periodo 2020–2024, pospandemia), el rendimiento sigue siendo muy elevado, con un R^2_test = 0.935241, es decir, el modelo logra explicar alrededor de un 93.52 % de la variación observada del IPC en el periodo de interés del Proyecto de Grado, cumpliendo holgadamente el criterio de $R^2 \geq 0.70$.

Los errores de pronóstico en el periodo de Prueba son moderados, en relación con el nivel del índice de precios:

RMSE_test = 3.72 puntos de IPC, lo que refleja el error cuadrático medio de las predicciones del Modelo.

MAE_test = 3.33 puntos de IPC, lo que significa que, en promedio, las predicciones del Modelo se desvían del valor real del IPC en algo más de 3 puntos.

La combinación de un R^2_test superior al 90% y de valores de RMSE y MAE relativamente bajos, sugiere que el Modelo Multivariado capta adecuadamente; tanto la tendencia general, como gran parte de las fluctuaciones de la inflación colombiana en el periodo pospandemia.

Frente a enfoques puramente Univariados, la inclusión de variables como el PIB, la Tasa de Intervención, el Salario Mínimo y la TRM permite:

Incorporar información sobre la actividad económica, la política monetaria, el mercado laboral y el tipo de cambio.

Mejorar la capacidad.

Ofrecer una interpretación económica más asertiva de los factores que inciden en la trayectoria del IPC.

En conjunto, estos resultados permiten concluir que el Modelo Multivariado Ridge, constituye una aproximación robusta y adecuada como Modelo final para el análisis y pronóstico de la inflación en Colombia en el periodo 2020–2024, dentro del marco de este Proyecto de Grado, dejando como línea futura la comparación con otros algoritmos de Machine Learning (Random Forest, ARIMA/SARIMAX, etc.) para reforzar la robustez de los hallazgos.

Conclusiones del Gráfico — Modelo Multivariado Ridge. En el gráfico de la Figura 5, se observa que el modelo: Acompaña de forma bastante fiel la trayectoria ascendente del IPC, capturando los principales cambios de nivel a lo largo del periodo 2020–2024. Tiende a sobreestimar ligeramente el IPC al inicio del periodo (2020–2021), donde la línea naranja se sitúa por encima de la azul. Durante la fase de mayor aceleración inflacionaria (especialmente, entre 2022 y comienzos de 2023), el IPC real crece algo más rápido que el predicho, generándose una brecha visible entre ambas curvas, aunque la dirección y el patrón general siguen siendo coherentes.

Además, muestra ciertos “escalones” o cambios bruscos de nivel vinculados a los saltos de variables como el Salario Mínimo o la propia dinámica macroeconómica anual, lo que sugiere que el modelo recoge bien esos efectos estructurales, aunque de forma algo más suavizada que la serie real.

La combinación de un R^2_test elevado y unos errores (RMSE y MAE) relativamente contenidos, junto con la buena correspondencia visual entre el IPC real y el predicho, permite concluir que el Modelo Multivariado Ridge, ofrece pronósticos robustos y coherentes para la

inflación colombiana en el periodo 2020–2024, constituyéndose en un candidato sólido a modelo final dentro del Proyecto de Grado, especialmente, frente a los modelos de árboles (Random Forest y Gradient Boosting) que presentan sobreajuste y peores resultados en el conjunto de Prueba.

El Modelo Multivariado con Regresión Ridge, que utiliza como predictores el tiempo (*t*), el PIB, la tasa de intervención del Banco de la República, el Salario Mínimo y la TRM, logra reproducir de forma adecuada la tendencia creciente del IPC en el periodo 2020–2024, tal como se aprecia en la cercanía general entre la curva azul (IPC real) y la curva naranja (IPC predicho).

En términos de ajuste estadístico: El modelo presenta un R^2_train = 0.9955, lo que indica que explica alrededor del 99.55 % de la variabilidad del IPC en el periodo de Entrenamiento (2007–2019). En el periodo de Prueba (2020–2024), el desempeño sigue siendo muy alto, con un R^2_test = 0.9352, es decir, el modelo explica aproximadamente el 93,52 % de la variación observada del IPC pospandemia.

Los errores de predicción resultan moderados en relación con el nivel del índice: RMSE_test = 3.72 puntos de IPC, que representa el error cuadrático medio de las predicciones. MAE_test = 3.33 puntos de IPC, lo que significa que, en promedio, las predicciones del modelo, se desvían del valor real en algo más de 3 puntos del índice.

Tabla 10

Residuos del Modelo Ridge Multivariado

	IPC_real	IPC_pred	Residuo
156	104.24	108.190587	-3.950587
157	104.94	108.461583	-3.521583
158	105.53	109.141336	-3.611336
159	105.70	109.447477	-3.747477
160	105.36	109.326635	-3.966635

Nota. Resultados obtenidos de los Residuos.

Análisis de los Residuos del Modelo Ridge Multivariado. El análisis de los residuos del Modelo Ridge Multivariado muestra que, durante los primeros meses del periodo de Prueba, las predicciones del Modelo son consistentemente superiores a los valores reales observados del IPC. Todos los residuos presentados son negativos y se ubican aproximadamente entre -3.90 y -3.22 puntos del IPC, lo que indica que el Modelo tiende a sobreestimar el nivel real del índice.

Este comportamiento sugiere la existencia de un sesgo inicial en las predicciones, donde la combinación de variables macroeconómicas —PIB, Tasa BanRep, Salario Mínimo y TRM—, genera valores estimados superiores al IPC efectivamente observado. Mientras el IPC real se encuentra entre 104 y 106 puntos, las predicciones del modelo se sitúan alrededor de 108 y 109 puntos.

Una posible explicación económica, es que algunas de las variables explicativas presentan tendencias crecientes de largo plazo y están altamente correlacionadas con el IPC. Aunque la regularización Ridge, reduce los efectos de la multicolinealidad, el modelo puede asignar un peso conjunto que produzca una sobreestimación temporal del índice durante ciertos periodos. Este fenómeno es común cuando se utilizan variables macroeconómicas con fuerte tendencia estructural.

A diferencia del Ridge Univariado, que utiliza exclusivamente la memoria histórica del IPC, el Ridge Multivariado incorpora información económica adicional. Esto permite una interpretación más rica del fenómeno inflacionario, pero también introduce una mayor complejidad en el proceso de estimación. Como resultado, el modelo puede reaccionar con mayor sensibilidad a cambios simultáneos en las variables explicativas, generando errores sistemáticos en determinados intervalos temporales.

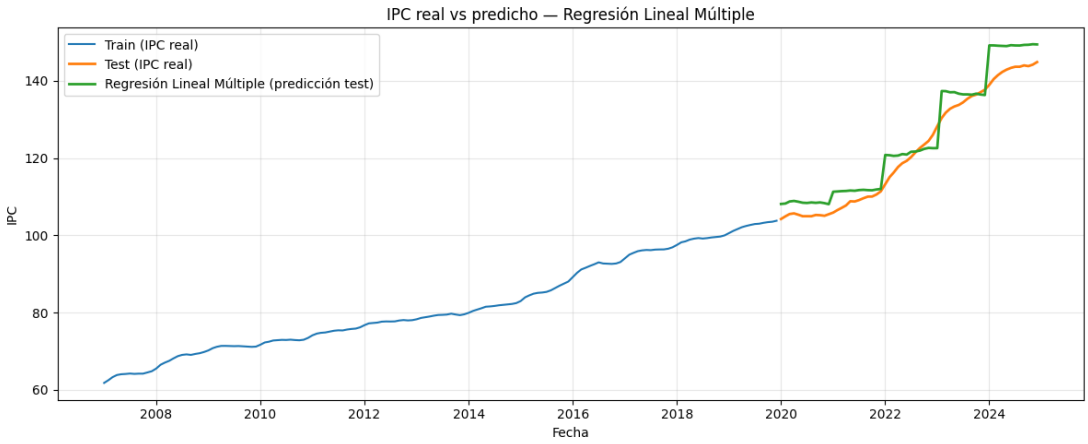
Sin embargo, es importante destacar que la magnitud de estos residuos sigue siendo relativamente moderada, frente al nivel total del IPC y que el modelo obtuvo métricas globales favorables en la evaluación fuera de muestra (R^2_test = 0.9352). Esto indica que, aunque existe una tendencia inicial a la sobreestimación, el modelo conserva una elevada capacidad explicativa y predictiva en términos generales.

Desde la perspectiva metodológica, los residuos sugieren que el Ridge Multivariado, logra capturar la relación entre el IPC y las principales variables macroeconómicas, pero puede presentar pequeñas desviaciones cuando los efectos de dichas variables no se reflejan de manera inmediata en el comportamiento del índice. Esto es consistente con la existencia de rezagos económicos en la transmisión de la política monetaria, el tipo de cambio y la actividad económica hacia los precios.

En conclusión, los residuos evidencian una sobreestimación moderada y relativamente constante del IPC, durante el inicio del periodo de Prueba. No obstante, la estabilidad de los errores y el buen desempeño global del modelo, confirman que el Ridge Multivariado constituye una herramienta útil para complementar el análisis predictivo, aportando, además, una interpretación económica más amplia que la obtenida con modelos exclusivamente univariados.

Figura 6

Regresión Lineal Múltiple — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Regresión Lineal Múltiple (predicción Test).

Análisis

El Modelo de Regresión Lineal Múltiple, utiliza como variable dependiente el IPC puro, es decir, el índice en nivel observado mensualmente, sin transformaciones interanuales, sin porcentajes y sin variaciones. Las variables independientes incorporadas son Producto Interno Bruto (PIB), Tasa del Banco de la República (BanRep), Salario Mínimo y Tasa Representativa del Mercado (TRM), por lo que el Modelo busca explicar el comportamiento del IPC, a partir de distintos canales macroeconómicos.

La inclusión del PIB permite aproximar el efecto de la actividad económica y la demanda agregada. La Tasa BanRep representa la postura de política monetaria del Banco de la República. El Salario Mínimo permite capturar presiones de costos laborales e ingresos nominales. La TRM refleja el canal cambiario y el posible encarecimiento de bienes importados.

Este Modelo es más completo que una Regresión Lineal Simple, basada solo en el tiempo, porque incorpora variables económicas que pueden influir sobre el nivel de precios. Sin

embargo, al ser un Modelo Lineal y contemporáneo, asume que el efecto de las variables independientes sobre el IPC ocurre en el mismo periodo, lo cual puede ser una limitación importante en series macroeconómicas.

Si el R^2 de Entrenamiento es alto, esto indica que las variables explicativas logran representar adecuadamente el comportamiento histórico del IPC entre 2007 y 2019. No obstante, la métrica principal para evaluar la utilidad predictiva es el R^2 de Prueba, ya que muestra si el modelo generaliza correctamente al periodo 2020–2024.

El RMSE_test y el MAE_test, se interpretan en unidades del IPC puro. Por tanto, indican cuántos puntos del índice se desvía, en promedio, la predicción respecto al valor real observado.

En conclusión, la Regresión Lineal Múltiple es un Modelo base multivariado útil para evaluar la relación entre el IPC y las principales variables macroeconómicas del estudio. Sin embargo, si el desempeño en Prueba es limitado, se recomienda avanzar hacia Modelos con rezagos, Ridge, ARIMAX, SARIMAX o Modelos de Machine Learning más robustos, especialmente, porque los efectos del Producto Interno de Bruto (PIB), la Tasa de Intervención, el Salario Mínimo y la Tasa Representativa del Mercado (TRM), pueden manifestarse con retraso.

Conclusiones del Gráfico — Modelo de Regresión Lineal Múltiple. El gráfico evidencia que el Modelo de Regresión Lineal Múltiple, no está generando predicciones válidas en su forma actual. Aunque el objetivo del Modelo es predecir el IPC puro, la línea verde de predicción, se ubica en una escala extremadamente superior a la del IPC real. Mientras el IPC observado, se mueve aproximadamente entre 60 y 145 puntos, la predicción del Modelo alcanza valores cercanos a 120.000, lo cual no es coherente con la escala natural del índice.

La principal señal de alerta es que las líneas del IPC real de Entrenamiento y Prueba quedan prácticamente pegadas al eje inferior, mientras que la predicción domina completamente el gráfico. Esto indica que el problema no está en el IPC real, sino en la magnitud de los valores predichos por el Modelo. Por tanto, este resultado no debe interpretarse como un buen ajuste, sino como una evidencia de problema de escala, limpieza de datos o especificación del Modelo.

Una causa probable es que alguna variable independiente, especialmente, el Salario Mínimo, el Producto Interno Bruto (PIB) o la Tasa Representativa del Mercado (TRM), esté en una escala muy grande o haya sido interpretada incorrectamente por el Modelo. Por ejemplo, si el Salario Mínimo se encuentra en pesos completos y el PIB en unidades muy altas, una Regresión Lineal sin estandarización puede generar coeficientes inestables y predicciones fuera del rango esperado. También puede existir multicolinealidad, ya que varias variables macroeconómicas crecen con el tiempo y pueden estar altamente correlacionadas con el IPC.

Desde el punto de vista predictivo, el Modelo no logra representar adecuadamente el comportamiento del IPC puro. La predicción no sigue la trayectoria real del índice, sino que produce saltos artificiales y valores desproporcionados. Esto impide evaluar correctamente la relación entre el IPC y las variables macroeconómicas.

En conclusión, el gráfico muestra que el Modelo de Regresión Lineal Múltiple, debe corregirse antes de ser utilizado como resultado final. Es necesario revisar la escala de las variables, especialmente, el Salario Mínimo y el PIB, verificar que no existan errores de formato, y considerar la estandarización de las variables independientes.

El modelo solo será válido cuando la línea predicha se encuentre en la misma escala del IPC real, es decir, en un rango cercano al índice observado y no en valores de decenas o cientos de miles.

Tabla 11

Residuos del Modelo de Regresión Lineal Múltiple

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	108.138235	-3.898235
2020-02-01	104.94	108.251583	-3.311583
2020-03-01	105.53	108.779183	-3.249183
2020-04-01	105.70	108.924565	-3.224565
2020-05-01	105.36	108.722462	-3.362462

Nota. Resultados obtenidos de los Residuos.

Análisis de Residuos del Modelo de Regresión Lineal Múltiple. El análisis de los residuos del Modelo de Regresión Lineal Múltiple, muestra que las predicciones obtenidas durante los primeros meses del periodo de Prueba, son sistemáticamente superiores a los valores reales observados del IPC. Los residuos presentan valores negativos que oscilan aproximadamente entre -3.90 y -3.22 puntos del IPC, lo que indica una sobreestimación persistente del índice por parte del modelo.

Se observa que el patrón de error es relativamente estable, ya que la magnitud de los residuos varía poco entre un mes y otro. Esta estabilidad sugiere que el Modelo, mantiene una tendencia consistente a predecir valores más altos que los realmente observados. Mientras el IPC real se ubica entre 104 y 106 puntos, las estimaciones generadas por el Modelo se encuentran alrededor de 108 y 109 puntos.

Desde el punto de vista económico, este comportamiento puede estar relacionado con la forma en que las variables explicativas —PIB, Tasa BanRep, Salario Mínimo y TRM—, contribuyen conjuntamente a la estimación del IPC. Debido a que varias de estas variables presentan tendencias crecientes de largo plazo, la Regresión puede amplificar su efecto combinado y generar predicciones superiores al nivel real del índice, especialmente, en periodos donde los efectos económicos aún no se reflejan completamente en los precios.

A diferencia de los modelos basados en rezagos del propio IPC, la Regresión Lineal Múltiple asume que la relación entre las variables macroeconómicas y el índice es inmediata y lineal. Sin embargo, en la práctica económica muchos de estos efectos operan con retrasos temporales. Por ejemplo, los cambios en la tasa de intervención del Banco de la República pueden tardar varios meses en impactar el comportamiento de la inflación. Esta característica puede explicar parte de la sobreestimación observada en los residuos.

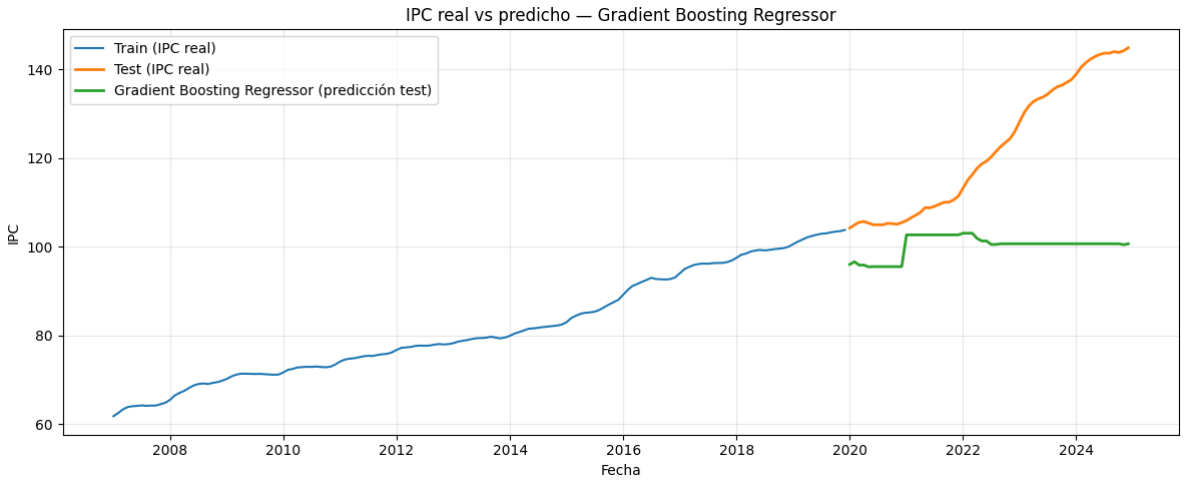
La presencia de residuos negativos relativamente uniformes, también puede interpretarse como una señal de que el Modelo captura adecuadamente la dirección general del IPC, pero presenta dificultades para ajustar con precisión el nivel exacto del índice durante determinados periodos. En otras palabras, el Modelo identifica correctamente la tendencia, pero desplaza la trayectoria estimada hacia valores ligeramente superiores a los observados.

Metodológicamente, este comportamiento es consistente con los desafíos que suelen presentar las regresiones lineales multivariadas cuando las variables explicativas están altamente correlacionadas entre sí. La multicolinealidad puede afectar la estabilidad de los coeficientes y producir estimaciones menos precisas, especialmente en contextos económicos complejos como el periodo pospandemia.

En conclusión, los residuos evidencian una sobreestimación moderada y constante del IPC por parte de la Regresión Lineal Múltiple. Aunque el Modelo logra representar la tendencia general del índice y la influencia conjunta de las variables macroeconómicas, la presencia de errores negativos persistentes indica que la relación lineal contemporánea utilizada por el modelo no es suficiente para capturar completamente la dinámica temporal del IPC. Esto explica por qué Modelos como el Ridge Univariado y el Ridge Multivariado, obtuvieron un desempeño predictivo superior en la evaluación fuera de muestra.

Figura 7

Gradient Boosting Regressor — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Gradient Boosting Regressor (predicción Test).

Análisis

El Modelo Gradient Boosting Regressor, utiliza como variable dependiente el Índice de Precios al Consumidor (IPC) puro, es decir, en nivel observado mensualmente. En este caso, no se emplean transformaciones interanuales, porcentuales, ni variaciones. Por lo tanto, las métricas se interpretan directamente en unidades del IPC.

Las variables independientes incorporadas son PIB, Tasa BanRep, Salario Mínimo y TRM. Estas variables representan diferentes canales macroeconómicos que pueden influir sobre el comportamiento del IPC: la actividad económica, la política monetaria, los costos laborales y el canal cambiario.

A diferencia de la Regresión Lineal Múltiple, el Gradient Boosting Regressor es un modelo no lineal basado en árboles de decisión secuenciales. Esto significa que puede capturar

relaciones más complejas entre las variables explicativas y el IPC, incluyendo interacciones y patrones no estrictamente lineales.

Si el R^2_train es alto, esto indica que el modelo logra aprender adecuadamente la estructura histórica del IPC en el periodo 2007–2019. Sin embargo, la métrica más importante es el R^2_test, porque muestra si el Modelo generaliza correctamente al periodo 2020–2024.

El RMSE_test y el MAE_test, se interpretan en puntos del IPC puro. Por ejemplo, un MAE de 3 indica que, en promedio, el modelo se equivoca aproximadamente en 3 puntos del índice.

En conclusión, el Gradient Boosting Regressor es una alternativa de Machine Learning más flexible que los Modelos Lineales, ya que puede capturar no linealidades. No obstante, si el desempeño en Prueba no mejora, frente a Modelos Lineales o benchmarks simples, esto puede indicar que las variables contemporáneas no son suficientes y que se deben incorporar rezagos del IPC, rezagos de las variables macroeconómicas o componentes estacionales.

Conclusiones del Gráfico — **Modelo Gradient Boosting Regressor.** El gráfico compara el IPC puro observado con la predicción generada por el Modelo Gradient Boosting Regressor. El eje Y representa exclusivamente el IPC en nivel, sin transformaciones YoY, sin porcentajes y sin variaciones interanuales.

Durante el periodo de Entrenamiento, entre 2007 y 2019, el IPC real muestra una trayectoria creciente y relativamente estable. Sin embargo, en el periodo de Prueba, entre 2020 y 2024, el IPC real aumenta con mayor velocidad, especialmente, desde 2021 y 2022, llegando a valores superiores a 140 puntos hacia el final del periodo.

La línea verde, correspondiente a la predicción del Modelo Gradient Boosting Regressor, se mantiene en un rango cercano a 95–103 puntos, aproximadamente. Esto muestra que el

modelo, no logra seguir la aceleración del IPC real durante el periodo de Prueba. En lugar de proyectar el crecimiento observado, la predicción se mantiene casi plana después de 2022.

Este comportamiento es común en modelos basados en árboles, como Gradient Boosting, cuando se entrenan con datos cuyo rango histórico no contiene valores tan altos como los observados posteriormente. El modelo tiende a predecir dentro del rango aprendido en el Entrenamiento y tiene dificultad para extrapolar hacia niveles superiores del IPC.

La separación creciente entre la línea naranja y la línea verde, indica que el modelo subestima de forma sistemática el IPC puro en el periodo 2020–2024. Aunque el modelo puede capturar relaciones no lineales entre Producto Interno Bruto (PIB), Tasa BanRep, Salario Mínimo y Tasa Representativa del Mercado (TRM), en este caso no logra anticipar el cambio de régimen inflacionario posterior a 2020.

En conclusión, el Gradient Boosting Regressor, no presenta un buen comportamiento predictivo en este gráfico. Su principal limitación es que no extrapola adecuadamente la tendencia creciente del IPC. Para mejorar este modelo, se recomienda incorporar una variable temporal t, rezagos del IPC, rezagos de PIB, Tasa BanRep, Salario Mínimo y TRM o combinarlo con modelos más adecuados para series temporales como ARIMAX, SARIMAX o Ridge con rezagos.

Tabla 12

Residuos del Modelo Gradient Boosting Regressor

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	96.030473	8.209527
2020-02-01	104.94	96.638005	8.301995
2020-03-01	105.53	95.889558	9.640442
2020-04-01	105.70	95.889558	9.810442
2020-05-01	105.36	95.477628	9.882372

Nota. Resultados obtenidos de los Residuos.

Análisis de Residuos del Modelo Gradient Boosting Regressor. El análisis de los residuos del Modelo Gradient Boosting Regressor, muestra que las predicciones realizadas durante los primeros meses del periodo de Prueba son considerablemente inferiores a los valores reales observados del IPC. Los residuos son todos positivos y oscilan aproximadamente entre 8.21 y 9.88 puntos del IPC, lo que indica una subestimación sistemática y significativa del índice por parte del modelo.

Se observa que mientras el IPC real se encuentra entre 104 y 106 puntos, las predicciones generadas por el modelo permanecen alrededor de 95 y 97 puntos. Esta diferencia evidencia que el Modelo no logra seguir adecuadamente la trayectoria ascendente que presenta el IPC durante el periodo pospandemia, generando errores de magnitud considerablemente superior a los observados en Modelos como Ridge Univariado, Ridge Multivariado o incluso ARIMA.

Además, los residuos muestran una tendencia creciente durante los primeros meses analizados. Por ejemplo, el error pasa de aproximadamente 8.21 puntos en enero de 2020 a cerca de 9.88 puntos en mayo de 2020. Este comportamiento indica que la capacidad predictiva del modelo, se deteriora a medida que el IPC continúa aumentando, sugiriendo que el algoritmo tiene dificultades para extrapolar valores fuera de los patrones observados durante el entrenamiento.

Desde una perspectiva metodológica, este resultado puede explicarse porque los modelos basados en Árboles de Decisión, como Gradient Boosting, suelen funcionar muy bien identificando relaciones complejas dentro del rango de datos utilizado para entrenar el modelo, pero presentan limitaciones cuando deben proyectar tendencias crecientes de largo plazo. A diferencia de los modelos lineales o de series temporales, los árboles no extrapolan naturalmente comportamientos fuera del rango histórico observado.

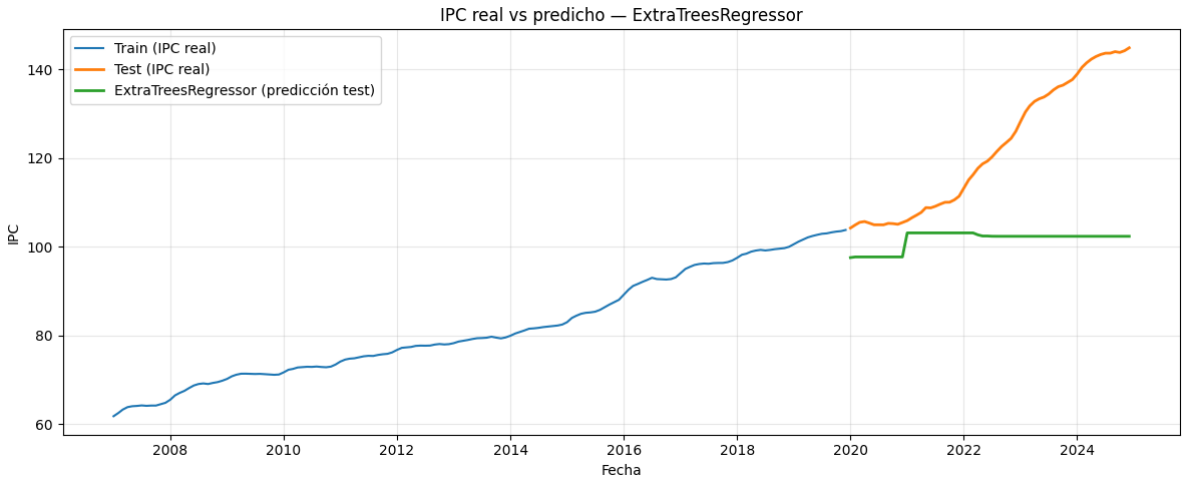
Otro aspecto importante es que el periodo 2020–2024 estuvo marcado por cambios estructurales significativos asociados a la pandemia, choques de oferta, variaciones cambiarias y ajustes de política monetaria. Estos eventos generaron patrones económicos distintos a los observados durante el periodo de Entrenamiento, reduciendo la capacidad del modelo para generalizar adecuadamente.

La magnitud de los residuos observados es consistente con las métricas globales obtenidas para el modelo, las cuales mostraron un desempeño significativamente inferior al de los Modelos Ridge. Los errores elevados reflejan que el Gradient Boosting, no logró capturar de manera adecuada la persistencia temporal característica del IPC puro.

En conclusión, los residuos evidencian una subestimación importante y persistente del IPC, acompañada de errores crecientes a lo largo del periodo de Prueba. Esto confirma que el Modelo Gradient Boosting Regressor, presenta dificultades para reproducir la dinámica inflacionaria observada durante la etapa pospandemia y explica su bajo desempeño relativo, frente a Modelos más simples y estables, como Ridge Univariado, Ridge Multivariado y Naive Forecast.

Figura 8

ExtraTreesRegressor — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde ExtraTreesRegressor (predicción Test).

Análisis

El Modelo ExtraTreesRegressor, utiliza como variable dependiente el IPC puro, es decir, el índice en nivel observado mensualmente. No se usan transformaciones interanuales, no se calculan porcentajes y no se transforma la variable objetivo. Por tanto, las métricas como RMSE y MAE se interpretan directamente en unidades del IPC.

Las variables independientes incorporadas son PIB, Tasa BanRep, Salario Mínimo y TRM. Estas variables representan distintos canales macroeconómicos relacionados con la inflación: actividad económica, política monetaria, costos laborales y canal cambiario.

A diferencia de la Regresión Lineal Múltiple, el ExtraTreesRegressor es un modelo no lineal, basado en ensambles de árboles extremadamente aleatorizados. Su ventaja es que puede capturar relaciones no lineales e interacciones entre variables, sin exigir una forma funcional lineal previa.

Si el R^2 de Entrenamiento es alto, significa que el Modelo logró aprender patrones históricos del IPC entre 2007 y 2019. Sin embargo, la métrica más importante es el R^2 de Prueba, porque permite verificar si el modelo generaliza correctamente al periodo 2020–2024.

El RMSE_test y el MAE_test, muestran el error en unidades del IPC puro. Un error bajo, indica que las predicciones se aproximan al índice real, mientras que un error alto indica que el modelo no logra capturar adecuadamente la dinámica reciente del IPC.

En conclusión, el ExtraTreesRegressor es una alternativa flexible de Machine Learning para modelar relaciones no lineales entre el IPC y las variables macroeconómicas. Sin embargo, al igual que otros modelos basados en árboles, puede tener dificultades para extrapolar valores del IPC fuera del rango aprendido durante el Entrenamiento. Por ello, si el periodo 2020–2024 presenta valores de IPC superiores a los vistos entre 2007 y 2019, el modelo puede subestimar el índice.

Conclusiones del Gráfico del Modelo ExtraTreeesRegressor. El gráfico muestra la comparación entre el IPC puro observado y la predicción generada por el modelo ExtraTreesRegressor. El eje Y corresponde únicamente al IPC en nivel, sin transformaciones YoY, sin porcentajes y sin variaciones interanuales.

Durante el periodo de Entrenamiento, entre 2007 y 2019, el IPC real presenta una tendencia creciente relativamente estable. En el periodo de Prueba, entre 2020 y 2024, el IPC real continúa aumentando, pero con una aceleración mucho más fuerte, especialmente, desde 2021 y 2022, hasta alcanzar valores superiores a 140 puntos hacia el final del periodo.

La predicción del modelo, representada por la línea verde, se mantiene alrededor de valores cercanos a 98–103 puntos. Esto indica que el ExtraTreesRegressor no logra seguir el

crecimiento real del IPC durante el periodo de Prueba. En lugar de acompañar la tendencia ascendente del índice, la predicción permanece casi plana después de 2021.

Este comportamiento muestra una limitación importante de los modelos basados en árboles, cuando se aplican a series temporales con tendencia creciente. El modelo aprende patrones del periodo de Entrenamiento, pero tiene dificultad para extrapolar valores superiores a los observados históricamente. Como el IPC de 2020–2024 supera el rango de Entrenamiento, el modelo tiende a quedarse en niveles cercanos al máximo aprendido antes de 2020.

La separación entre la línea naranja del IPC real y la línea verde de predicción, evidencia una subestimación sistemática del IPC puro. Esto significa que, aunque el modelo puede capturar ciertas relaciones no lineales entre PIB, Tasa BanRep, Salario Mínimo y TRM, no logra representar adecuadamente el cambio de régimen inflacionario posterior a 2020.

En conclusión, el ExtraTreesRegressor no presenta un buen desempeño visual para predecir el IPC puro, en el periodo 2020–2024. Su principal debilidad es que no extrapola correctamente la tendencia creciente del índice. Para mejorar este modelo, se recomienda incluir una variable temporal *t*, rezagos del IPC, rezagos de las variables macroeconómicas y posibles componentes estacionales. De lo contrario, el modelo funciona como referencia comparativa, pero no como modelo final de predicción del IPC.

Tabla 13

Residuos del Modelo ExtraTreesRegressor

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	97.568229	6.671771
2020-02-01	104.94	97.717109	7.222891
2020-03-01	105.53	97.718859	7.811141
2020-04-01	105.70	97.718859	7.981141
2020-05-01	105.36	97.718859	7.641141

Nota. Resultados obtenidos de los Residuos.

Análisis de los Residuos del Modelo ExtraTreesRegressor. El análisis de los residuos del Modelo ExtraTreesRegressor, evidencia que las predicciones realizadas durante los primeros meses del periodo de Prueba, son consistentemente inferiores a los valores reales observados del IPC. Los residuos son positivos en todos los casos y se ubican aproximadamente entre 6.67 y 7.98 puntos del IPC, indicando una subestimación sistemática del comportamiento real del índice.

Se observa que mientras el IPC real se encuentra entre 104 y 106 puntos, las predicciones generadas por el modelo permanecen alrededor de 97 y 98 puntos. Esta diferencia refleja que el modelo no logra capturar adecuadamente la aceleración del IPC observada durante el periodo pospandemia, manteniendo estimaciones significativamente por debajo de los valores reales.

La estabilidad de los residuos también es un aspecto relevante. Los errores presentan una magnitud relativamente constante a lo largo de los meses analizados, lo que indica que el modelo mantiene un sesgo persistente de subestimación. Esto sugiere que el problema no corresponde a fluctuaciones aleatorias, sino a una limitación estructural del modelo para representar la tendencia creciente del IPC.

Desde el punto de vista metodológico, este comportamiento es coherente con las características de los Modelos basados en Árboles de Decisión. Aunque ExtraTreesRegressor, suele ofrecer un excelente desempeño para identificar relaciones no lineales y patrones complejos dentro del conjunto de Entrenamiento, tiene dificultades para extrapolar tendencias cuando los datos de Prueba presentan niveles superiores a los observados previamente.

En otras palabras, el modelo aprende adecuadamente el rango histórico de los datos, pero presenta limitaciones para anticipar incrementos sostenidos del IPC fuera de dicho rango.

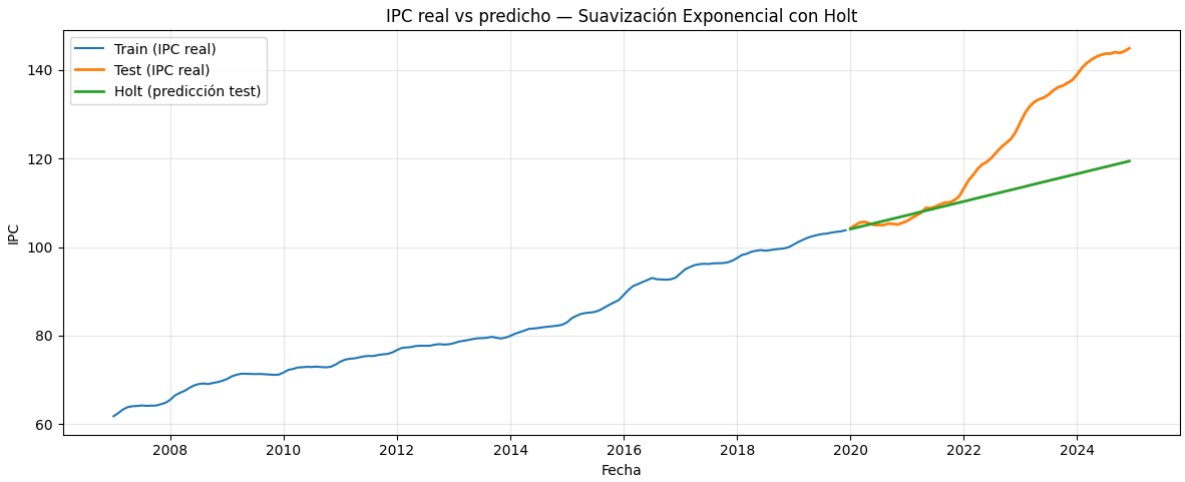
Además, el contexto económico del periodo 2020–2024, estuvo marcado por eventos extraordinarios asociados a la pandemia, aumentos de costos, cambios en la política monetaria y fluctuaciones cambiarias. Estos factores introdujeron patrones que no necesariamente estaban presentes con la misma intensidad en los datos de Entrenamiento, reduciendo la capacidad de generalización del modelo.

La magnitud de los residuos observados es menor que la obtenida en el Modelo Gradient Boosting Regressor, lo que indica que ExtraTrees, logra una representación ligeramente mejor del comportamiento del IPC. Sin embargo, los errores siguen siendo considerablemente superiores a los observados en Modelos como Ridge Univariado, Ridge Multivariado o Naive Forecast.

En conclusión, los residuos muestran una subestimación persistente y relativamente estable del IPC, evidenciando que el ExtraTreesRegressor no logra capturar completamente la tendencia inflacionaria observada durante el periodo pospandemia. Aunque el modelo identifica parcialmente la dirección general de la serie, presenta limitaciones importantes para reproducir el nivel real del índice, lo que explica su desempeño inferior, frente a los modelos que mejor aprovecharon la persistencia temporal del IPC.

Figura 9

Suavización Exponencial con Holt — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Holt (predicción Test).

Análisis

El Modelo de Suavización Exponencial con Holt utiliza como variable dependiente el IPC puro, es decir, el índice en nivel observado mensualmente. No se emplean transformaciones interanuales, porcentuales, ni variaciones. Por lo tanto, los resultados se interpretan directamente en unidades del IPC.

Este Modelo es un enfoque Univariado de Series de Tiempo, ya que no utiliza variables macroeconómicas externas como PIB, Tasa BanRep, Salario Mínimo o TRM. En su lugar, Holt modela dos componentes principales: el nivel y la tendencia de la propia serie del IPC.

La utilidad del Modelo Holt, radica en que permite proyectar series con tendencia creciente, lo cual es pertinente para el IPC, porque se trata de un índice acumulativo de precios. Si el IPC presenta una trayectoria creciente relativamente estable, el modelo puede capturar adecuadamente su dirección general.

Si el R^2 de Entrenamiento es alto, esto indica que el modelo logra representar bien la dinámica histórica del IPC entre 2007 y 2019. Sin embargo, la métrica más importante es el R^2 de Prueba, ya que permite evaluar si la tendencia aprendida se mantiene válida durante el periodo 2020–2024.

El RMSE_test y el MAE_test, se interpretan en puntos del IPC puro. Por ejemplo, un MAE de 4 indica que, en promedio, las predicciones se desvían aproximadamente 4 puntos del índice real.

En conclusión, Holt funciona como un modelo base tradicional útil para series con tendencia. No obstante, al no incorporar variables externas, ni efectos de política monetaria, Salario Mínimo, actividad económica o tipo de cambio, puede fallar cuando el periodo de Prueba, presenta cambios de régimen o aceleraciones inflacionarias. Por ello, sus resultados deben compararse con Modelos Multivariados como Ridge, Regresión Lineal Múltiple, ARIMAX, SARIMAX o Modelos de Machine Learning con variables macroeconómicas.

Conclusiones del Gráfico — **Modelo Suavización Exponencial con Holt.** El gráfico muestra la comparación entre el IPC puro observado y la predicción generada por el Modelo de Suavización Exponencial con Holt. El eje Y, corresponde únicamente al IPC en nivel, sin transformaciones interanuales, sin porcentajes y sin variaciones. Esto permite evaluar directamente si el modelo logra seguir la trayectoria real del índice de precios.

Durante el periodo de Entrenamiento, entre 2007 y 2019, el IPC real presenta una tendencia creciente relativamente estable. Este comportamiento es adecuado para el Modelo Holt, ya que este método está diseñado para series con nivel y tendencia. Por esta razón, la predicción inicia en 2020 cerca del valor real del IPC, lo cual indica que el modelo capturó correctamente la dirección general del índice en el periodo histórico.

Sin embargo, a partir de 2021 y especialmente desde 2022, el IPC real comienza a crecer con mayor velocidad que la predicción. La línea naranja, correspondiente al IPC observado en el periodo de Prueba, se separa progresivamente de la línea verde del Pronóstico. Esto evidencia que el modelo subestima el IPC puro durante la etapa de mayor aceleración inflacionaria.

La predicción de Holt, mantiene una trayectoria ascendente, pero demasiado suave. Esto significa que el modelo proyecta la tendencia promedio aprendida entre 2007 y 2019, pero no logra capturar completamente el cambio de ritmo ocurrido en el periodo 2020–2024. Este comportamiento es esperable, porque Holt no incorpora variables macroeconómicas externas como PIB, Tasa BanRep, Salario Mínimo o TRM, ni tampoco choques estructurales.

Desde el punto de vista visual, la brecha creciente entre el IPC real y el IPC predicho, indica que la tendencia histórica no fue suficiente para anticipar la aceleración reciente del índice. El modelo logra representar una trayectoria creciente, pero no alcanza los niveles reales observados hacia 2023 y 2024, donde el IPC supera ampliamente la Predicción.

Su principal limitación es que proyecta una tendencia lineal suavizada y no incorpora factores económicos que expliquen la aceleración inflacionaria reciente.

Tabla 14

Residuos del Modelo Suavización Exponencial con Holt

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	104.06	0.18
2020-02-01	104.94	104.32	0.62
2020-03-01	105.53	104.58	0.95
2020-04-01	105.70	104.84	0.86
2020-05-01	105.36	105.10	0.26

Nota. Resultados obtenidos de los Residuos.

Análisis de los Residuos del Modelo Suavización Exponencial con Holt. El análisis de los residuos del Modelo de Suavización Exponencial con Holt, muestra que las diferencias entre los valores reales del IPC y las predicciones realizadas por el Modelo son relativamente pequeñas durante los primeros meses del periodo de Prueba. Los residuos oscilan aproximadamente entre 0.18 y 0.95 puntos del IPC, lo que indica un nivel de error reducido en comparación con la magnitud total del índice.

Se observa que todos los residuos son positivos, lo que significa que el modelo, tiende a subestimar ligeramente los valores reales del IPC. En otras palabras, las predicciones generadas por Holt, son sistemáticamente inferiores a los datos observados. Sin embargo, la magnitud de esta subestimación es moderada y considerablemente menor que la observada en Modelos como Gradient Boosting Regressor o ExtraTreesRegressor.

El comportamiento de los residuos sugiere que el modelo, logra capturar adecuadamente la tendencia general de crecimiento del IPC. Esto es consistente con la naturaleza del método de Holt, diseñado específicamente para modelar series temporales con tendencia. Al incorporar componentes de nivel y tendencia, el modelo puede seguir la trayectoria ascendente del índice con relativa precisión durante horizontes de corto plazo.

No obstante, la presencia continua de residuos positivos indica que el ritmo de crecimiento real del IPC, fue ligeramente superior al proyectado por el modelo. Este comportamiento es esperable en periodos caracterizados por aceleraciones inflacionarias asociadas a choques económicos extraordinarios, como los observados durante la etapa pospandemia. Dado que Holt, se basa únicamente en la evolución histórica de la serie, puede reaccionar con cierto retraso ante cambios abruptos en la velocidad de crecimiento del índice.

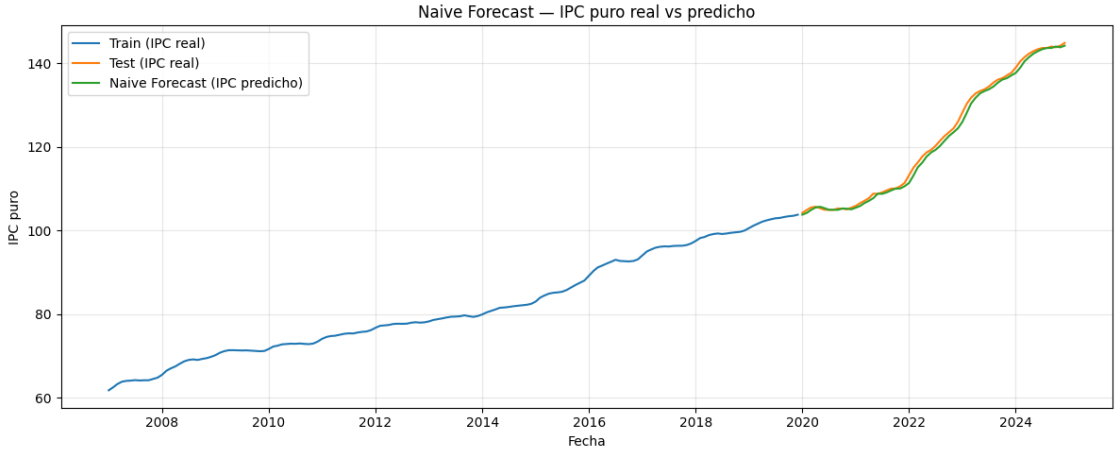
Desde una perspectiva metodológica, estos residuos evidencian que el modelo, reproduce adecuadamente la dirección y tendencia del IPC, aunque presenta una ligera incapacidad para capturar completamente la intensidad de los incrementos observados en determinados periodos. Esta característica es común en métodos de Suavización Exponencial, cuando la serie experimenta cambios estructurales importantes.

La magnitud relativamente baja de los residuos iniciales confirma que Holt, constituye una alternativa razonable dentro de los Modelos Clásicos de Series Temporales. Sin embargo, al evaluar el horizonte completo de Prueba, otros Modelos como Ridge Univariado y Naive Forecast, lograron una mejor capacidad de generalización y menores errores globales.

En conclusión, los residuos muestran que el Modelo de Suavización Exponencial con Holt, presenta una subestimación leve y relativamente estable del IPC, manteniendo errores reducidos y una adecuada representación de la tendencia general de la serie. Esto confirma que Holt, logra capturar gran parte de la dinámica temporal del IPC puro, aunque presenta limitaciones para adaptarse rápidamente a cambios inflacionarios más intensos durante el periodo pospandemia.

Figura 10

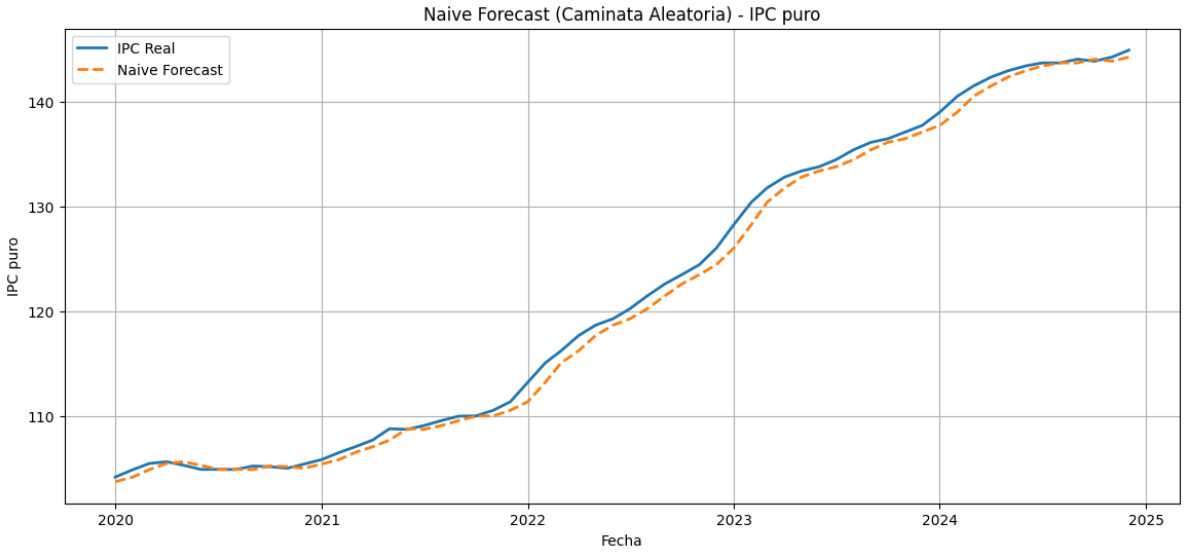
Naive Forecast — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Naive Forecast (IPC predicho).

Figura 11

Naive Forecast — Periodo Prueba (2020-2024)



Nota. Línea azul IPC real y línea naranja Naive Forecast.

Análisis

Este modelo utiliza el IPC puro como variable objetivo y no usa YearOverYear (YoY) o interanual (año contra año), porcentajes, ni tampoco transformaciones. Su regla de predicción es:

$$\widehat{IPC}_t = IPC_{t-1}$$

Es decir, el IPC estimado para el mes actual corresponde al IPC observado en el mes inmediatamente anterior. Este modelo sirve como línea base para evaluar, si los Modelos de Machine Learning aportan mejora real, frente a una predicción simple.

El Modelo Naive Forecast funciona como un benchmark temporal, porque predice el IPC puro del periodo actual usando el IPC del mes inmediatamente anterior. Este modelo no incorpora variables macroeconómicas externas ni aprendizaje complejo, por lo que sirve como comparación mínima frente a Ridge, Regresión Lineal, Gradient Boosting, ExtraTrees y Holt.

Si el Naive Forecast obtiene métricas superiores a los Modelos de Machine Learning, esto indica que el IPC puro tiene una persistencia mensual muy alta. En ese caso, el Ridge Univariado, puede seguir siendo el mejor Modelo de Machine Learning, pero el Naive Forecast debe reconocerse como el mejor benchmark general. Esta comparación, fortalece la rigurosidad del proyecto, porque demuestra que los modelos avanzados fueron evaluados contra una referencia simple y exigente.

Conclusiones del Gráfico — Modelo Naive Forecast. El gráfico presenta la evolución histórica del IPC puro en Colombia durante el periodo 2007–2024, diferenciando claramente el conjunto de Entrenamiento (Train: 2007–2019) y el conjunto de Prueba (Test: 2020–2024), junto con las predicciones generadas por el Modelo Naive Forecast. Se observa una tendencia creciente sostenida del IPC a lo largo de toda la serie, con una aceleración particularmente

marcada entre 2021 y 2023, seguida de una moderación en la velocidad de crecimiento hacia finales de 2024.

Durante el periodo de Entrenamiento, el IPC presenta un crecimiento relativamente estable, pasando de valores cercanos a 62 puntos en 2007 hasta superar los 104 puntos en 2019. Esta trayectoria evidencia la naturaleza acumulativa del índice y confirma la existencia de una fuerte persistencia temporal, característica que favorece el desempeño de modelos, basados en observaciones recientes.

En el periodo de Prueba (2020–2024), la línea correspondiente al Naive Forecast prácticamente se superpone con la serie observada del IPC. Las diferencias entre ambas curvas son mínimas y se presentan principalmente en momentos donde el índice experimenta incrementos más rápidos, especialmente entre 2022 y comienzos de 2023.

Este comportamiento es esperado, ya que el modelo utiliza exclusivamente el valor observado del periodo anterior para generar cada predicción, produciendo un ligero rezago frente a cambios bruscos en la tendencia.

A pesar de esta limitación inherente, el modelo logra reproducir con gran precisión la trayectoria general del IPC. La cercanía visual entre la serie real y la serie estimada confirma que la información contenida en el último dato observado resulta altamente relevante para anticipar el comportamiento inmediato del índice. En otras palabras, el IPC colombiano presenta una fuerte dependencia temporal entre meses consecutivos, lo que favorece el desempeño de una estrategia de caminata aleatoria.

Los resultados visuales son coherentes con las métricas obtenidas para este modelo (R^2_train = 0.9990, R^2_test = 0.9963, RMSE_test = 0.8926 y MAE_test = 0.7247), las cuales representan los mejores valores observados entre todos los modelos evaluados en la

investigación. Esto indica que los errores de predicción son inferiores a un punto del índice IPC en promedio, reflejando una capacidad predictiva sobresaliente.

En conclusión, este gráfico evidencia que el Naive Forecast, reproduce de manera excepcional la evolución del IPC puro, durante el periodo 2020–2024, manteniendo una coincidencia casi total con los valores observados. La mínima separación entre las curvas real y predicha confirma la elevada persistencia temporal del IPC y explica el excelente desempeño estadístico obtenido por el modelo.

En consecuencia, el Naive Forecast se consolida como el benchmark predictivo más sólido de la investigación, constituyendo una referencia fundamental para evaluar la utilidad de modelos más complejos de Machine Learning y Series de Tiempo.

Tabla 15

Residuos del Modelo Naive Forecast

	IPC_real	IPC_pred	Residuo
Fecha			
2020-01-01	104.24	103.80	0.44
2020-02-01	104.94	104.24	0.70
2020-03-01	105.53	104.94	0.59
2020-04-01	105.70	105.53	0.17
2020-05-01	105.36	105.70	-0.34

Nota. Resultados obtenidos de los Residuos.

Análisis de los Residuos del Modelo Naive Forecast. El análisis de los residuos del Modelo Naive Forecast, muestra que las diferencias entre los valores reales del IPC y las predicciones realizadas son muy pequeñas. Los residuos observados oscilan aproximadamente entre -0.34 y 0.70 puntos del IPC, lo que evidencia un nivel de error reducido y una elevada precisión en las estimaciones de corto plazo.

Se observa que durante los primeros cuatro meses los residuos son positivos, indicando que el modelo tiende a subestimar ligeramente el IPC. Esto significa que el índice real fue un poco superior al valor pronosticado utilizando el dato del mes anterior. Sin embargo, en mayo de 2020 aparece un residuo negativo (-0.34), reflejando una leve sobreestimación del IPC. Esta alternancia de signos es una característica favorable, ya que indica que el modelo no presenta un sesgo permanente en una sola dirección.

La reducida magnitud de los residuos confirma la fuerte persistencia temporal que caracteriza al IPC puro. Debido a que el índice evoluciona de manera gradual y suele presentar cambios relativamente pequeños entre meses consecutivos, utilizar el valor observado en el periodo inmediatamente anterior constituye una aproximación sorprendentemente efectiva para generar pronósticos de corto plazo.

Desde una perspectiva económica, estos resultados sugieren que, gran parte de la información necesaria para anticipar el comportamiento futuro del IPC, ya está contenida en su valor más reciente. Esto es consistente con la naturaleza acumulativa del índice de precios, donde los ajustes suelen transmitirse progresivamente y no de manera abrupta entre periodos consecutivos.

La baja magnitud de los residuos también ayuda a explicar por qué el Naive Forecast obtuvo uno de los mejores desempeños del estudio. A diferencia de modelos más complejos que

requieren estimar múltiples parámetros o relaciones económicas, el Naive Forecast aprovecha directamente la persistencia temporal de la serie, evitando problemas de sobreajuste o errores derivados de especificaciones incorrectas.

No obstante, aunque el modelo muestra una excelente capacidad predictiva de corto plazo, presenta una limitación importante: no incorpora información económica adicional, ni permite interpretar los factores que influyen sobre el comportamiento del IPC. Por esta razón, su utilidad principal radica en servir como benchmark o modelo de referencia, contra el cual se comparan los modelos más sofisticados.

En conclusión, los residuos del Naive Forecast son los más reducidos entre los modelos evaluados, mostrando errores pequeños, equilibrados y cercanos a cero. Esto confirma que la persistencia temporal del IPC es extremadamente alta y explica por qué el modelo, logró un desempeño sobresaliente durante el periodo 2020–2024. Sin embargo, su simplicidad limita la interpretación económica de los resultados, razón por la cual debe considerarse principalmente como un punto de referencia para evaluar el valor agregado de los Modelos de Machine Learning y de Series Temporales más avanzados.

Variables y Categorías

Tabla 16

Tipos de Variables del Estudio

Variable	Rol	Tipo	Fuente	Rol analítico
Índice de Precios al Consumidor (IPC)	Dependiente	Cuantitativa continua	DANE	Variable objetivo en nivel; IPC puro.
Producto Interno Bruto (PIB)	Explicativa	Cuantitativa continua	Banco de la República	Señal de actividad económica agregada.
Tasa de Intervención del Banco de la República	Explicativa	Cuantitativa continua	Banco de la República	Aproxima postura monetaria.
Salario Mínimo	Explicativa	Cuantitativa continua	Banco de la República	Captura costos laborales e ingreso nominal.
Tasa Representativa del Mercado (TRM)	Explicativa	Cuantitativa continua	Banco de la República	Representa transmisión cambiaria.
Rezagos del IPC	Ingeniería de variables	Cuantitativa continua	Derivada del IPC	Capturan memoria y persistencia del índice.
Tendencia temporal t	Control temporal	Cuantitativa continua	Derivada	Representa evolución temporal de largo plazo.

Nota. Resultados obtenidos de los Residuos.

Población y Muestra

La población de estudio corresponde a las observaciones mensuales de las series macroeconómicas oficiales relacionadas con el IPC en Colombia durante 2007-2024. La base consolidada contiene 216 observaciones mensuales y variables originales como fecha, IPC, PIB, tasa BanRep, Salario Mínimo y TRM. La muestra analítica coincide con la población disponible, al trabajar con datos secundarios oficiales. La partición temporal se definió de la siguiente manera: Entrenamiento 2007-2019 y Prueba 2020-2024.

Procedimiento a Seguir

Tabla 17

Procedimiento Metodológico del Estudio

Fase	Actividades principales	Producto
Comprensión del problema	Definición del objetivo, pregunta problema y criterios de evaluación.	Marco de trabajo y alcance.
Adquisición y consolidación	Descarga e integración de series oficiales del DANE y Banco de la República.	Base maestra en Excel.
Preparación	Limpieza, tipificación, tratamiento de fechas, corrección de escala y variables derivadas.	Dataset analítico con IPC puro.
EDA	Análisis descriptivo, tendencias, rupturas y selección de variables relevantes.	Hallazgos exploratorios.
Modelado	Entrenamiento de modelos lineales, regularizados, ensambles y Holt.	Modelos candidatos.
Evaluación	Comparación mediante RMSE, MAE y R^2 en prueba temporal.	Tabla comparativa y selección.
Interpretación	Lectura económica y recomendaciones técnicas.	Discusión y recomendaciones.

Nota. Paso a paso llevado a cabo.

Diseño Experimental y Validación

La validación se realizó mediante partición temporal: Entrenamiento, entre enero de 2007 y diciembre de 2019, y Prueba, entre enero de 2020 y diciembre de 2024. La variable objetivo fue el IPC puro. Las métricas utilizadas fueron R^2, RMSE y MAE. Se privilegió el desempeño en Prueba y la estabilidad temporal de los modelos. Este diseño evita fuga de información, porque los datos futuros no se utilizan durante el Entrenamiento.

Validación Temporal Complementaria: Backtesting y Rolling Window

Como complemento metodológico sugerido durante el proceso de evaluación del proyecto, se plantea la incorporación de técnicas adicionales de validación temporal, tales como backtesting, 116olling window y expanding window, con el propósito de fortalecer la robustez de los resultados obtenidos.

La evaluación principal de esta investigación se realizó mediante una partición temporal consistente con la naturaleza cronológica de los datos, utilizando el período 2007–2019 para Entrenamiento y 2020–2024 para Prueba. Este enfoque permitió evaluar la capacidad predictiva de los Modelos sobre observaciones futuras no utilizadas durante el Entrenamiento.

Sin embargo, una única partición temporal puede generar resultados influenciados por las características particulares del periodo seleccionado. Por esta razón, en estudios futuros, se recomienda implementar procedimientos de validación temporal más exigentes.

En el enfoque de Rolling Window, el Modelo se entrena utilizando una ventana histórica de tamaño fijo y posteriormente, se evalúa sobre el siguiente período temporal. Una vez realizada la predicción, la ventana se desplaza hacia adelante y el procedimiento se repite sucesivamente. Este método permite analizar la estabilidad del desempeño predictivo a lo largo del tiempo y detectar posibles cambios estructurales en la serie.

Por su parte, el método Expanding Window, conserva toda la información histórica disponible y amplía progresivamente el conjunto de Entrenamiento a medida que se incorporan nuevas observaciones. De esta forma, el Modelo aprende de una cantidad creciente de datos y se evalúa continuamente sobre periodos posteriores.

La aplicación de estas metodologías permitiría verificar si los resultados obtenidos para el Ridge Univariado, identificado como el mejor modelo de Machine Learning del estudio, y para el Naive Forecast, considerado el mejor benchmark predictivo general, se mantienen consistentes bajo diferentes ventanas de Entrenamiento y Prueba. Asimismo, contribuiría a determinar si el desempeño observado responde a patrones estructurales estables de la serie del IPC o si depende de un periodo específico de evaluación.

En consecuencia, aunque la partición 2007–2019 / 2020–2024 se mantiene como el esquema de validación principal de la presente investigación, la incorporación de procedimientos de backtesting temporal constituye una línea de mejora metodológica que puede aumentar la confiabilidad y generalización de los resultados en futuras investigaciones sobre predicción de inflación en Colombia mediante técnicas de Machine Learning y series de tiempo.

Ejemplos Prácticos

Frente a lo anterior, se implementó una validación temporal complementaria, mediante esquemas de Rolling Window y Expanding Window. Estas técnicas permiten evaluar la estabilidad del desempeño predictivo de los Modelos en diferentes períodos temporales, reduciendo la dependencia de una única partición Entrenamiento-Prueba.

Resultados de Backtesting – Rolling Window. En este esquema, se utilizó una ventana de Entrenamiento fija de 96 meses y una ventana de Prueba de 12 meses, desplazando progresivamente la ventana a lo largo de la serie.

Tabla 18

Resultados Promedio del Rolling Window

Modelo	R^2 Promedio	RMSE Promedio	MAE Promedio
Ridge Univariado	0.6921	0.4762	0.4025
Naive Forecast	0.6307	0.6364	0.5478

Nota. Rolling Window usando dos modelos.

Interpretación. Los resultados muestran que el Ridge Univariado, obtuvo un desempeño superior al Naive Forecast en las diferentes ventanas temporales evaluadas. El modelo presentó un R^2 promedio de 0.6921, superior al 0.6307 obtenido por el benchmark. Asimismo, registró menores errores de predicción, tanto en RMSE como en MAE.

Estos resultados sugieren que el Ridge Univariado, mantiene una capacidad predictiva consistente cuando se evalúa sobre distintos periodos históricos, evidenciando una adecuada capacidad de generalización.

Resultados de Backtesting – Expanding Window. En este esquema el conjunto de entrenamiento se amplía progresivamente incorporando nuevas observaciones históricas, mientras que la prueba se realiza sobre el periodo inmediatamente posterior.

Tabla 19

Resultados Promedio del Expanding Window

Modelo	R^2 Promedio	RMSE Promedio	MAE Promedio
Ridge Univariado	0.7260	0.4518	0.3742
Naive Forecast	0.6307	0.6364	0.5478

Nota. Expanding Window usando dos modelos.

Interpretación. Los resultados obtenidos mediante Expanding Window, muestran una mejora adicional en el desempeño del Ridge Univariado. El Modelo alcanzó un R^2 promedio de 0.7260, acompañado de los menores valores de RMSE y MAE, observados durante el proceso de validación.

Este comportamiento indica que el Ridge Univariado se beneficia de la incorporación progresiva de información histórica, mejorando su capacidad para capturar la dinámica del IPC en diferentes contextos económicos.

Resultados Comparativos.

Tabla 20

Comparación General de la Validación Temporal

Método de validación	Modelo	R^2	RMSE	MAE
Rolling Window	Ridge Univariado	0.6921	0.4762	0.4025
Rolling Window	Naive Forecast	0.6307	0.6364	0.5478
Expanding Window	Ridge Univariado	0.7260	0.4518	0.3742
Expanding Window	Naive Forecast	0.6307	0.6364	0.5478

Nota. Resultados de validación.

Conclusiones de la Validación Temporal. Los resultados del backtesting evidencian que el Ridge Univariado, mantiene un desempeño estable y superior al Naive Forecast, cuando se evalúa mediante múltiples ventanas temporales. Aunque el Naive Forecast, continúa siendo un benchmark relevante por su simplicidad y capacidad para capturar la persistencia del IPC, el Ridge Univariado obtuvo sistemáticamente mejores valores de R^2, RMSE y MAE en los esquemas de Rolling Window y Expanding Window.

Por consiguiente, la validación temporal complementaria, respalda la robustez del Ridge Univariado, como el principal Modelo de Machine Learning del estudio, demostrando que su

desempeño no depende exclusivamente de la partición Entrenamiento-Prueba utilizada inicialmente (2007–2019 / 2020–2024), sino que se mantiene consistente bajo distintos escenarios de evaluación temporal, lo que fortalece la confiabilidad de los resultados.

Evaluación

La evaluación se concentró en la comparación fuera de muestra. En este contexto, un R^2_test alto indica que el modelo logra explicar la variabilidad del IPC en el periodo 2020-2024. El RMSE_test y el MAE_test, se interpretan en puntos del índice IPC, no en porcentaje.

Tabla 21

Resultados Comparativos Finales con IPC puro

Modelo	**R^2_train**	**R^2_test**	**RMSE_test**	**MAE_test**
Ridge Univariado con rezagos del IPC	0.9955	0.9611	2.8800	2.2752
ARIMA Univariado	0.7838	0.1479	13.4763	9.9080
Regresión Lineal Simple	0.9738	-0.1880	15.9123	12.1878
Ridge Multivariado (PIB, TASA, SAL_MIN, TRM)	0.9955	0.9352	3.7152	3.3329
Regresión Lineal Múltiple	0.9910	0.9117	4.3392	3.6997
Gradient Boosting Regressor	0.9997	-2.1762	26.0188	21.8823

ExtraTreesRegressor	0.9983	-1.8622	24.6992	20.4925
Suavización Exponencial de Holt	0.9997	0.0020	14.5847	10.7308
Naive Forecast	0.9990	0.9963	0.8926	0.7247

Nota. Se presentan los resultados finales obtenidos utilizando el IPC puro, como variable dependiente. Las métricas RMSE_test y MAE_test, se expresan en puntos del índice IPC. No se reportan modelos construidos con variaciones interanuales, porcentuales, ni transformaciones derivadas del IPC.

La comparación evidencia diferencias importantes entre los nueve modelos evaluados. El Naive Forecast obtuvo el mejor desempeño predictivo general, alcanzando un R^2_test = 0.9963, un RMSE_test = 0.8926 y un MAE_test = 0.7247, lo que demuestra la elevada persistencia temporal del IPC puro durante el periodo analizado.

Entre los Modelos de Machine Learning, el Ridge Univariado presentó el mejor desempeño, con R^2_test = 0.9611, confirmando que los rezagos del propio IPC contienen una señal predictiva altamente relevante para anticipar su comportamiento futuro. El Ridge Multivariado también mostró resultados sobresalientes (R^2_test = 0.9352) y aportó valor interpretativo al incorporar variables macroeconómicas como PIB, Tasa BanRep, Salario Mínimo y TRM.

La Regresión Lineal Múltiple obtuvo un desempeño favorable (R^2_test = 0.9117), constituyéndose en una línea base multivariada sólida para el análisis del IPC puro. Por su parte, el ARIMA Univariado alcanzó un desempeño moderado (R^2_test = 0.1479), mientras que la Suavización Exponencial de Holt obtuvo un resultado cercano a cero (R^2_test = 0.0020), evidenciando limitaciones para representar completamente la aceleración inflacionaria observada durante el periodo pospandemia.

Por su parte, la Regresión Lineal Simple presentó un R^2_test negativo (-0.1880), indicando que una tendencia lineal aislada resulta insuficiente para capturar la complejidad del comportamiento reciente del IPC. De igual forma, los modelos de Ensamble Gradient Boosting Regressor y ExtraTreesRegressor, mostraron problemas de generalización fuera de muestra, obteniendo R^2_test de -2.1762 y -1.8622, respectivamente. Aunque ambos modelos alcanzaron niveles muy altos de ajuste durante el entrenamiento, no lograron extrapolar adecuadamente el incremento del IPC registrado entre 2020 y 2024.

En conjunto, los resultados muestran que la persistencia temporal constituye la principal fuente de capacidad predictiva para el IPC puro. Por esta razón, el Naive Forecast se consolidó como el mejor benchmark predictivo general, mientras que el Ridge Univariado se identificó como el mejor modelo de Machine Learning de la investigación.

Análisis de Residuos e Intervalos de Error

El análisis de residuos se incorpora como complemento de la evaluación predictiva. Un residuo, se define como la diferencia entre el IPC observado y el IPC estimado por el modelo. Su forma general es: e_t = IPC_t - IPC_t estimado. Cuando los residuos se distribuyen alrededor de cero y no muestran patrones sistemáticos, el modelo presenta una mejor calidad de ajuste.

Para los modelos con mejor desempeño, especialmente Naive Forecast, Ridge Univariado y Ridge Multivariado, se recomienda graficar los residuos en el tiempo, construir histogramas de errores y revisar si existen sesgos, durante los años de mayor aceleración inflacionaria. En este estudio, RMSE y MAE funcionan como medidas agregadas de error. Sin embargo, el análisis de residuos permite identificar si los errores se concentran en episodios específicos, como 2022-2023, cuando el IPC aumentó con mayor intensidad.

Los intervalos de error pueden aproximarse usando la dispersión de los residuos. Aunque el Proyecto se centra en comparación puntual de Modelos, incorporar intervalos de error permitiría comunicar el grado de incertidumbre asociado a cada pronóstico del IPC puro. Esta extensión es especialmente importante para aplicaciones institucionales de seguimiento y alerta temprana.

Despliegue y Replicabilidad

El proyecto de grado entrega una base consolidada en Excel, notebook en Python para limpieza, entrenamiento y evaluación, tablas comparativas y figuras, en conjunto con un archivo PDF. La replicabilidad se garantiza mediante la trazabilidad de datos, la partición temporal definida y la documentación de los modelos. Para actualizaciones futuras, se recomienda mantener el mismo flujo de preprocesamiento, especialmente, la limpieza correcta del Salario Mínimo y la definición de IPC puro como variable objetivo.

Las figuras presentadas a continuación permiten visualizar los principales resultados obtenidos durante el desarrollo del proyecto. En primer lugar, se muestra la evolución histórica del IPC puro entre 2007 y 2024, lo que facilita identificar tendencias, cambios de régimen y periodos de aceleración inflacionaria.

Posteriormente, se incluyen los gráficos correspondientes a cada uno de los nueve modelos implementados: ARIMA Univariado, Regresión Lineal Simple, Ridge Univariado, Ridge Multivariado, Regresión Lineal Múltiple, Gradient Boosting Regressor, ExtraTreesRegressor, Suavización Exponencial de Holt y Naive Forecast.

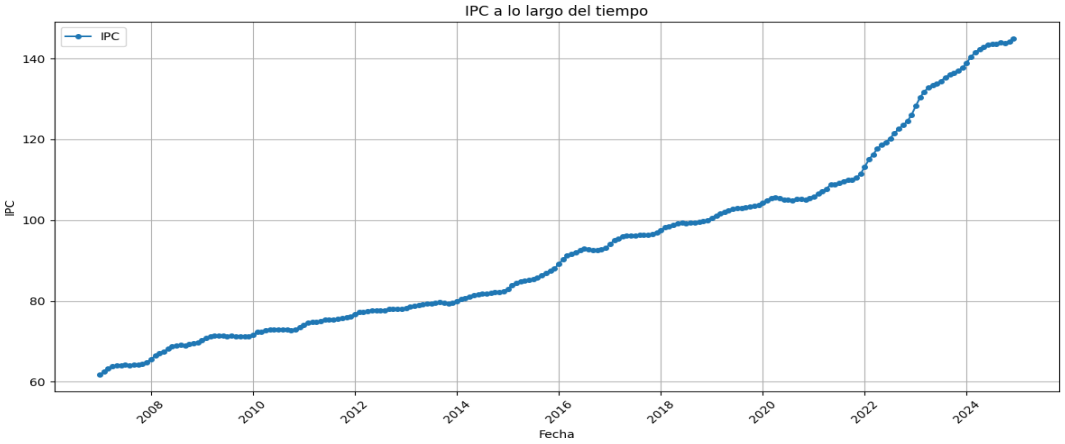
Estas figuras permiten comparar visualmente los valores observados y predichos del IPC, evaluar la capacidad de ajuste de cada metodología y analizar su comportamiento durante el periodo de prueba 2020–2024.

Finalmente, se presenta un gráfico comparativo de desempeño mediante el coeficiente de determinación (R^2), el cual sintetiza los resultados obtenidos y facilita la identificación de los modelos con mayor capacidad predictiva. En conjunto, las figuras complementan el análisis cuantitativo realizado mediante R^2, RMSE y MAE, proporcionando una interpretación visual de la precisión, estabilidad y capacidad de generalización de los modelos evaluados.

Figuras

Figura 12

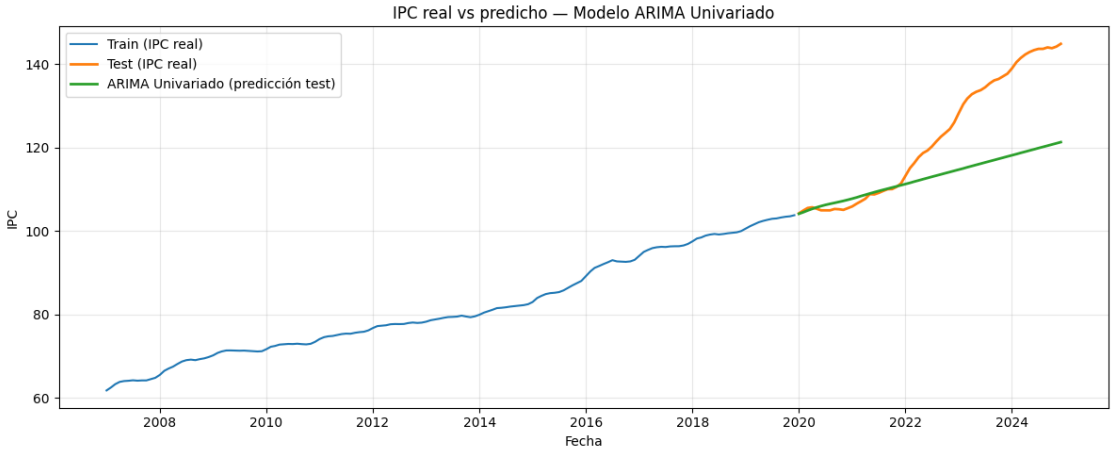
IPC a lo Largo del Tiempo



Nota. Tendencia del IPC desde 2008 a 2024.

Figura 13

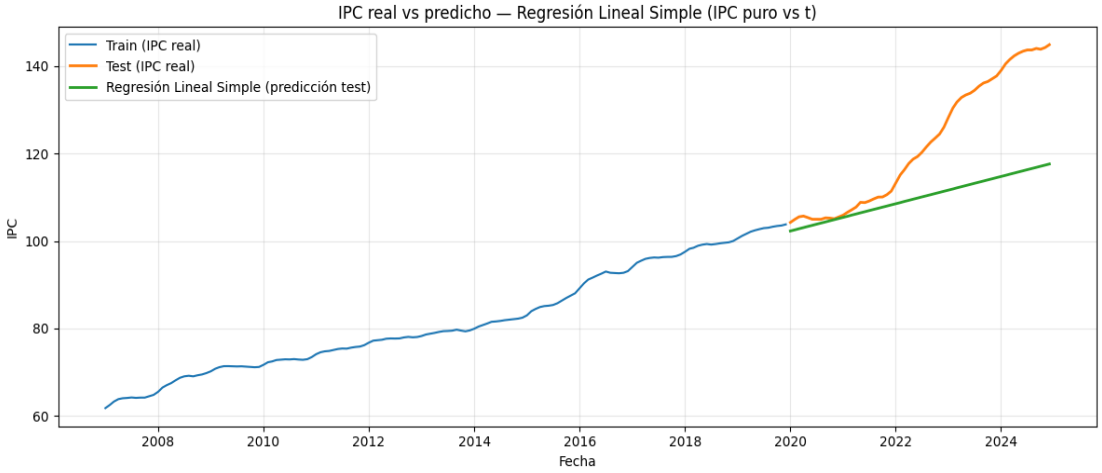
Modelo ARIMA Univariado — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde ARIMA Univariado (predicción Test).

Figura 14

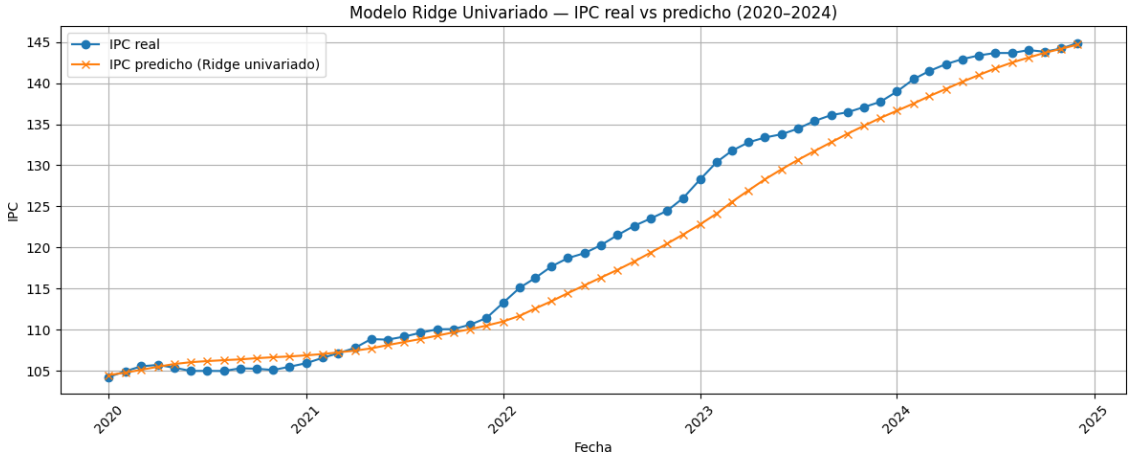
Regresión Lineal Simple — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Regresión Lineal Simple (predicción Test).

Figura 15

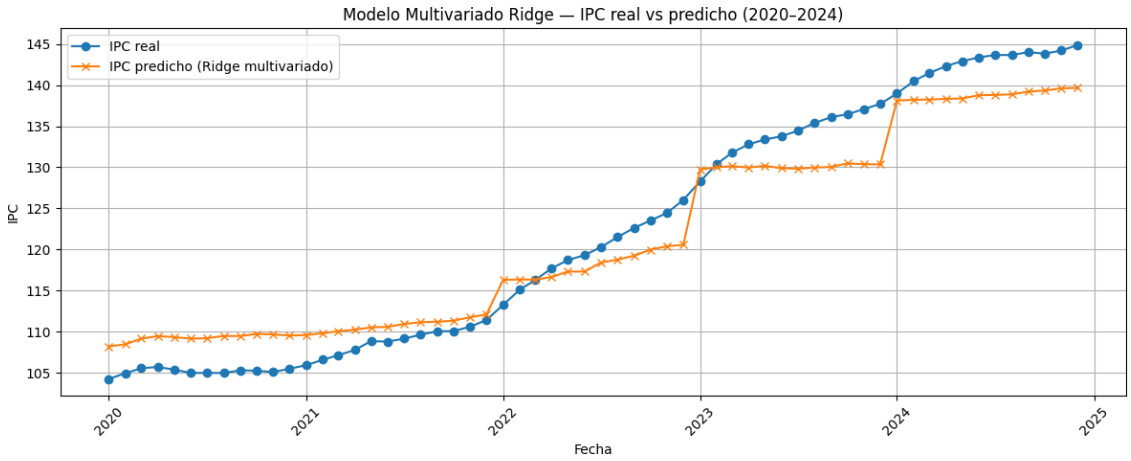
Modelo Ridge Univariado — IPC Real vs. Predicho (2020–2024)



Nota. Línea azul IPC real y línea naranja IPC predicho.

Figura 16

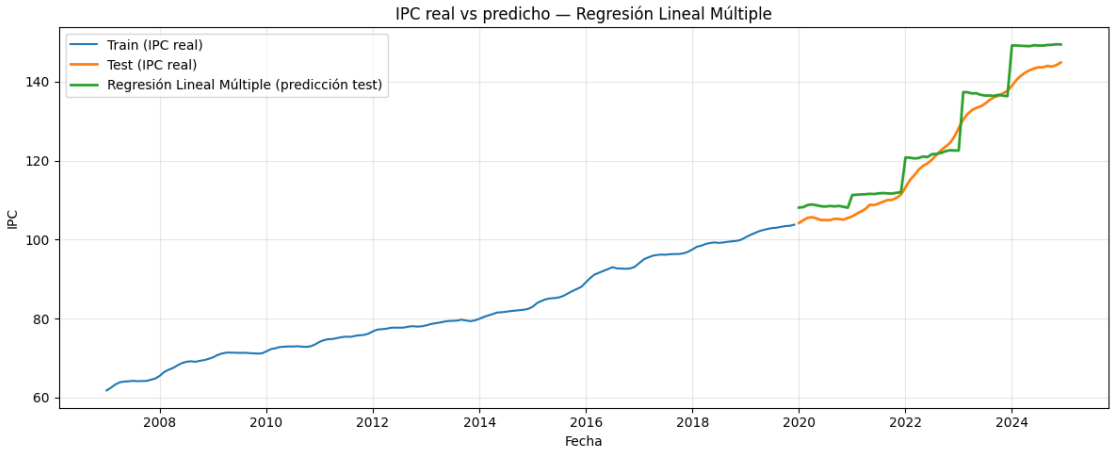
Modelo Ridge Multivariado — IPC Real vs. Predicho



Nota. Línea azul IPC real y línea naranja IPC predicho.

Figura 17

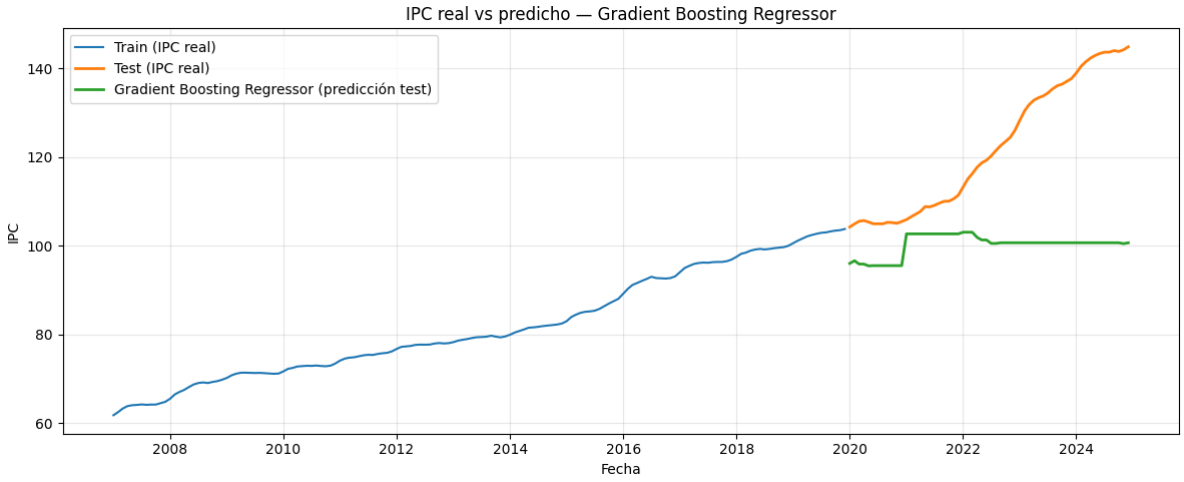
Regresión Lineal Múltiple — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Regresión Lineal Múltiple (predicción Test).

Figura 18

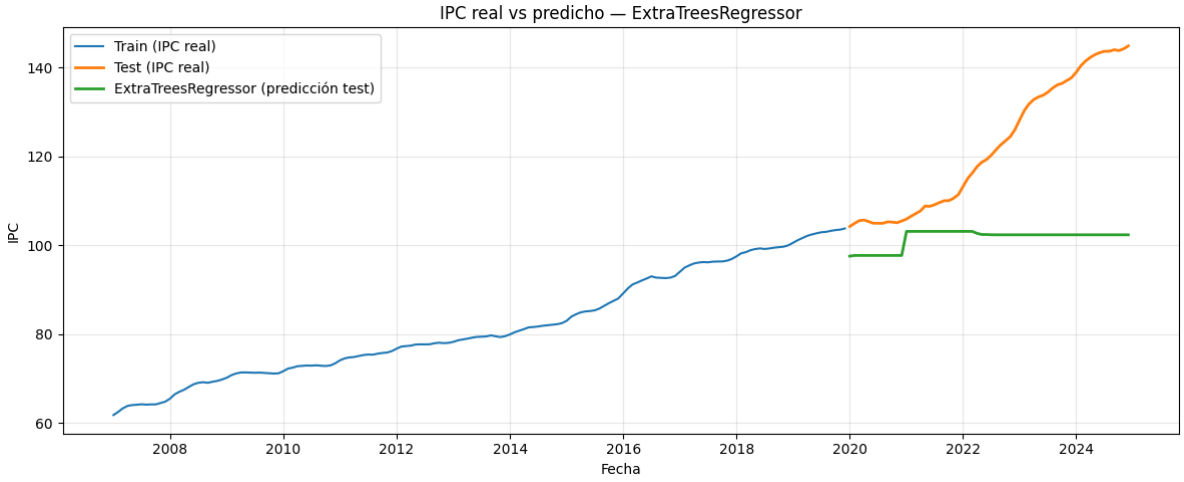
Gradient Boosting Regressor — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Gradient Boosting Regressor (predicción Test).

Figura 19

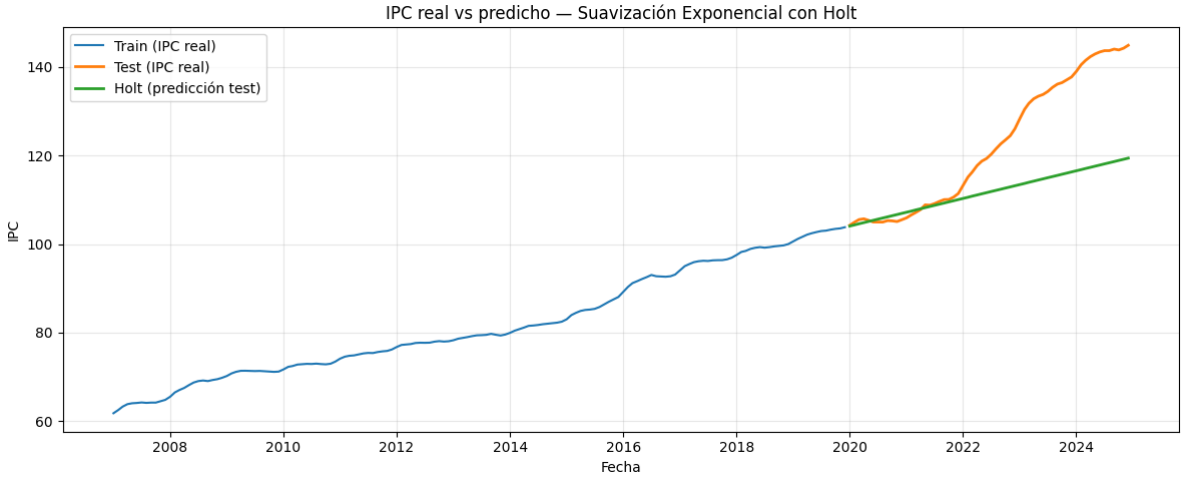
ExtraTreesRegressor — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde ExtraTreesRegressor (predicción Test).

Figura 20

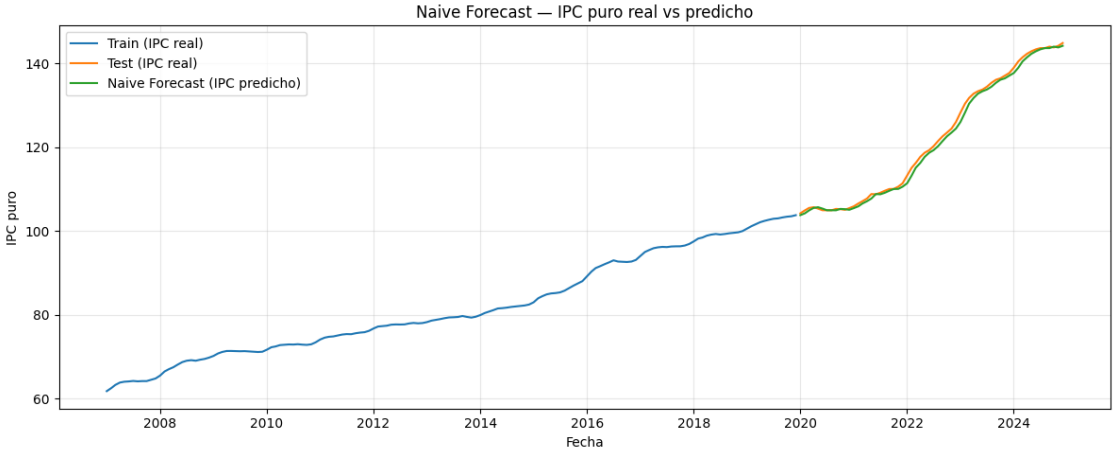
Suavización Exponencial con Holt — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Holt (predicción Test).

Figura 21

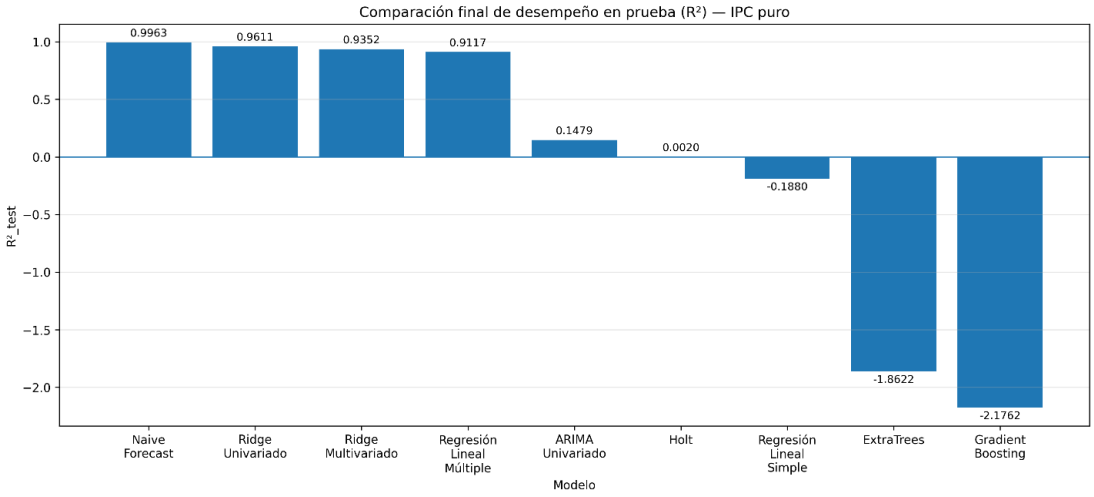
Naive Forecast — IPC Real vs. Predicho



Nota. Línea azul Train (IPC real), línea naranja Test (IPC real) y línea verde Naive Forecast (IPC predicho).

Figura 22

Comparación Final del Desempeño en Prueba, mediante R² con IPC puro



Nota. Diferentes modelos usados.

Procesos Desarrollados para el Cumplimiento de los Objetivos

Procesos Asociados al Objetivo Específico 1

El primer objetivo se desarrolló a partir de la recopilación, consolidación y depuración de series mensuales provenientes del Departamento Administrativo Nacional de Estadística (DANE) y del Banco de la República. Inicialmente, se verificó la consistencia de unidades de medida, formatos de fecha, tipos de datos y periodicidad de las variables seleccionadas. Posteriormente, se normalizaron nombres de columnas, se corrigieron inconsistencias en los registros y se unificó la estructura temporal de la base de datos para garantizar la comparabilidad entre las diferentes series.

Una vez consolidada la información, se realizó un proceso de análisis exploratorio de datos (Exploratory Data Analysis – EDA), mediante estadísticas descriptivas, visualizaciones gráficas y análisis de correlación. Este procedimiento permitió identificar tendencias, patrones de comportamiento, posibles valores atípicos y relaciones preliminares entre el IPC y las variables macroeconómicas consideradas en el estudio. Los resultados evidenciaron una trayectoria ascendente sostenida del IPC durante todo el periodo analizado, una aceleración significativa entre 2021 y comienzos de 2023 asociada al contexto pospandemia y una moderación gradual durante 2024.

Asimismo, el análisis exploratorio permitió identificar una fuerte persistencia temporal del IPC y asociaciones relevantes con variables como la TRM, el PIB, la Tasa de Intervención del Banco de la República y el Salario Mínimo Mensual Legal Vigente (Smmlv). Estos hallazgos sirvieron como fundamento para la selección de variables y para la construcción de los modelos predictivos implementados en las etapas posteriores de la investigación.

Procesos Asociados al Objetivo Específico 2

Para cumplir el segundo objetivo, se diseñó una estrategia de modelado basada en validación temporal estricta, respetando el orden cronológico de la información y evitando la contaminación entre los conjuntos de entrenamiento y prueba. La partición principal utilizó datos comprendidos entre 2007 y 2019 para entrenamiento y el periodo 2020–2024 para evaluación fuera de muestra.

Se implementaron nueve modelos predictivos utilizando Python: Ridge Univariado, ARIMA Univariado, Regresión Lineal Simple, Ridge Multivariado, Regresión Lineal Múltiple, Gradient Boosting Regressor, ExtraTreesRegressor, Suavización Exponencial de Holt y Naive Forecast. Estos modelos representaron enfoques complementarios que incluyeron técnicas de Machine Learning, métodos clásicos de Series de Tiempo y Benchmarks predictivos.

Durante esta etapa se realizaron procesos de preparación de datos, construcción de variables explicativas, generación de rezagos del IPC para los modelos univariados y configuración de hiperparámetros en los Modelos de Machine Learning. Posteriormente, cada modelo fue entrenado, validado y evaluado utilizando las métricas R^2, RMSE y MAE, priorizando el desempeño fuera de muestra como criterio principal de selección.

Además de las métricas cuantitativas, se analizaron los gráficos de predicción, los residuos y los errores de pronóstico para evaluar la capacidad de generalización de cada modelo. Los resultados obtenidos permitieron comparar de manera objetiva el desempeño relativo de las diferentes metodologías implementadas. La Tabla 21, resume los resultados finales de esta comparación.

Procesos Asociados al Objetivo Específico 3

El tercer objetivo se orientó a interpretar los resultados desde una perspectiva económica

aplicada, integrando los hallazgos estadísticos con el contexto macroeconómico colombiano del periodo 2020–2024. Para ello, se contrastaron las trayectorias estimadas por los modelos con los principales acontecimientos económicos ocurridos durante la etapa pospandemia.

El análisis consideró factores como los choques de oferta derivados de las disrupciones logísticas globales, la depreciación cambiaria y su efecto sobre los precios internos, los incrementos en los precios internacionales de alimentos y energía, los ajustes de política monetaria implementados por el Banco de la República, los aumentos del Salario Mínimo y la recuperación gradual de la actividad económica.

Adicionalmente, se analizaron los coeficientes y el comportamiento de las variables macroeconómicas incorporadas en los modelos multivariados, con el fin de identificar posibles canales de transmisión hacia el IPC. También se evaluaron los patrones observados en los residuos y errores de predicción para determinar las fortalezas y limitaciones de cada metodología.

Finalmente, los resultados obtenidos fueron interpretados a la luz de la teoría económica y de la evidencia empírica reciente sobre inflación pospandemia. Este proceso permitió explicar no solo qué modelos presentaron mejor desempeño predictivo, sino también por qué determinados enfoques lograron representar con mayor precisión la dinámica inflacionaria observada en Colombia durante el periodo analizado.

Análisis y Presentación de Resultados

Resultados del Objetivo Específico 1: Comportamiento del IPC y Hallazgos del Análisis Exploratorio de Datos (Exploratory Data Analysis – EDA)

El análisis descriptivo confirmó que el IPC, mantiene una tendencia ascendente de largo plazo, pero con un cambio de pendiente visible después de 2021. Este comportamiento es compatible con un episodio inflacionario pospandemia caracterizado por mayor persistencia y sensibilidad a choques externos. La serie muestra una aceleración entre 2021 y 2023 y una moderación parcial durante 2024, sin retorno inmediato a la dinámica previa a la pandemia.

Desde la perspectiva exploratoria, la serie también evidencia persistencia: una vez el IPC se acelera, sus efectos no se disipan inmediatamente. Este comportamiento justifica el uso de Modelos que incorporan rezagos del IPC puro y regularización. Además, las variables exógenas seleccionadas son coherentes con el marco teórico: la TRM refleja transmisión cambiaria; la tasa BanRep representa política monetaria; el PIB aproxima actividad económica y el Salario Mínimo captura presiones de costos e ingresos nominales.

El análisis descriptivo de las variables, permitió identificar patrones relevantes en la evolución del IPC puro y de las variables macroeconómicas seleccionadas para el estudio. La base cuenta con 216 observaciones mensuales entre 2007 y 2024, lo cual permite analizar un periodo amplio antes y después de la pandemia. Las variables incluidas fueron IPC puro, PIB, Tasa de Intervención del Banco de la República, Salario Mínimo y TRM.

En primer lugar, el IPC puro presentó una media de 93.0751 puntos, con un valor mínimo de 61.80 y un máximo de 144.88. El rango de 83.08 puntos evidencia un crecimiento acumulado importante del nivel general de precios durante el periodo analizado. Además, la mediana del IPC fue 88.62, inferior a la media, lo que indica que los valores más altos observados en los años

recientes, especialmente después de 2021, elevan el promedio de la serie. Este comportamiento es coherente con el episodio de aceleración inflacionaria posterior a la pandemia.

El PIB mostró una media de 790900.00 y un coeficiente de variación de 0.1607, el más bajo entre las variables analizadas. Esto indica que, aunque el PIB creció durante el periodo, su variabilidad relativa fue menor frente a variables como la Tasa BanRep, el Salario Mínimo o la TRM. En términos económicos, el PIB permite aproximar el comportamiento de la actividad económica agregada, pero por sí solo no explica completamente los cambios bruscos observados en el IPC durante 2020–2024.

La Tasa BanRep presentó una media de 5.9306, con un valor mínimo de 1.75 y un máximo de 13.25. Su coeficiente de variación fue 0.5359, el más alto de la tabla descriptiva. Este resultado evidencia que la tasa de intervención fue la variable con mayor variabilidad relativa, lo cual refleja cambios importantes en la postura de política monetaria durante el periodo estudiado. Esta alta variabilidad justifica profundizar el análisis entre IPC y Tasa BanRep, especialmente porque los efectos de política monetaria sobre los precios no suelen ser inmediatos, sino que pueden operar con rezagos.

El Salario Mínimo presentó una media de 729376.43, con un mínimo de 433700.00 y un máximo de 1300000.00. Su comportamiento muestra una tendencia creciente, explicada por los ajustes anuales del Salario Mínimo en Colombia. Esta variable es relevante para el análisis del IPC porque puede relacionarse con presiones de costos laborales, ingresos nominales e indexación de algunos precios. Sin embargo, su fuerte tendencia ascendente también exige cuidado en la interpretación, ya que puede compartir tendencia temporal con el IPC sin que esto implique una relación causal directa.

Por su parte, la TRM registró una media de 2816.4294, con un mínimo de 1712.28 y un

máximo de 4922.30. El rango de 3210.02 muestra una alta variabilidad cambiaria durante el periodo 2007–2024. Esta variable resulta importante porque representa el canal de transmisión cambiaria: una depreciación del peso frente al dólar puede encarecer bienes importados, insumos, combustibles, transporte y otros componentes de la canasta del IPC. Por esta razón, la TRM se mantiene como una variable macroeconómica relevante dentro del análisis predictivo.

La Matriz de Correlación complementó el análisis descriptivo al mostrar el grado de asociación lineal entre las variables. Se observó una correlación muy alta entre el IPC puro y el Salario Mínimo, cercana a 1.00. Esta relación refleja que ambas series tienen una trayectoria creciente durante el periodo analizado. No obstante, esta asociación debe interpretarse con precaución, porque una correlación alta no implica necesariamente causalidad; también puede estar explicada por una tendencia común en el tiempo.

El IPC puro también presentó una correlación alta con el PIB, de aproximadamente 0.94, lo cual sugiere que el crecimiento del nivel de precios se ha movido de forma cercana al crecimiento de la actividad económica agregada. De igual manera, la correlación entre IPC y TRM fue cercana a 0.92, lo que respalda la importancia del tipo de cambio como variable asociada al comportamiento de los precios internos. Esta relación resulta coherente con el marco teórico, debido a que la depreciación cambiaria puede trasladarse a mayores costos de importación y presiones sobre bienes transables.

En contraste, la correlación contemporánea entre IPC puro y Tasa BanRep fue más moderada, con un valor aproximado de 0.39. Este resultado es relevante, porque indica que la tasa de intervención no necesariamente se mueve de manera simultánea con el IPC. En la práctica, la política monetaria puede reaccionar al aumento de la inflación y sus efectos sobre el nivel de precios pueden observarse meses después. Por esta razón, el análisis exploratorio no

debe limitarse a la correlación del mismo periodo, sino que debe considerar rezagos de la Tasa BanRep a 3, 6, 12 o más meses.

Los diagramas de dispersión permitieron visualizar de forma gráfica las relaciones entre el IPC puro y cada variable macroeconómica. El cruce entre IPC y PIB mostró una relación positiva clara, lo cual sugiere que mayores niveles de actividad económica se asocian con mayores niveles del índice. El diagrama entre IPC y Salario Mínimo mostró una relación positiva muy fuerte y casi lineal, consistente con la alta correlación observada. Sin embargo, esta relación también puede estar influida por la tendencia creciente de ambas variables.

El diagrama entre IPC y TRM mostró una relación positiva importante, aunque con mayor dispersión que en el caso del Salario Mínimo y el PIB. Esto indica que el tipo de cambio se asocia con el comportamiento del IPC, pero no explica por sí solo toda la dinámica del índice. En cambio, el gráfico entre IPC y Tasa BanRep evidenció una nube de puntos más dispersa, lo cual confirma que la relación entre política monetaria e IPC es menos directa cuando se analiza de forma contemporánea.

Desde el punto de vista metodológico, estos resultados del EDA son importantes porque evidencian posible multicolinealidad entre algunas variables explicativas. Las altas correlaciones entre IPC, PIB, Salario Mínimo y TRM sugieren que varias series comparten una tendencia creciente de largo plazo. Esta situación justifica el uso de Modelos regularizados como Ridge, ya que la penalización L2, permite estabilizar los coeficientes y reducir el riesgo de sobreajuste cuando las variables independientes están altamente correlacionadas.

Adicionalmente, el análisis del cruce entre inflación derivada del IPC y Tasa BanRep permite complementar la lectura económica del fenómeno. Aunque los modelos finales utilizan IPC puro como variable objetivo, calcular la inflación mensual o anual derivada del IPC puede

ser útil únicamente para fines descriptivos. Este análisis permite observar si los incrementos en la tasa de intervención se relacionan con fases de mayor inflación y si existe evidencia visual de rezagos en la respuesta de la política monetaria.

En síntesis, el análisis descriptivo confirma que el IPC puro presenta una trayectoria ascendente con aceleración marcada después de 2021. También muestra que las variables macroeconómicas seleccionadas tienen fundamento analítico: el PIB representa la actividad económica, la TRM aproxima la transmisión cambiaria, el Salario Mínimo recoge presiones de costos e ingresos nominales y la Tasa BanRep refleja la postura de política monetaria. Estos hallazgos respaldan la selección de variables para los modelos multivariados y justifican la comparación entre modelos lineales, regularizados y de Machine Learning.

Finalmente, el Análisis Exploratorio de Datos (EDA), permite concluir que la inflación medida a través del IPC puro, no puede explicarse mediante una sola variable. Su comportamiento responde a una combinación de persistencia temporal, actividad económica, política monetaria, transmisión cambiaria y costos laborales. Por ello, el análisis exploratorio aporta evidencia clave para la fase de modelado y fortalece la decisión de evaluar modelos que incorporen, tanto memoria del IPC, como variables macroeconómicas externas.

Estadística Descriptiva del Análisis Exploratorio de Datos (Exploratory Data Analysis – EDA)

Tabla 22

Resultados Estadística Descriptiva de cada variable

Variable	n	Media	Desv. Estándar	Mínimo	P25	Mediana	P75	Máximo	Rango	Coef. variación
IPC	216	93.0751	22.0807	61.80	75.37	88.6200	105.1175	144.88	83.08	0.2372
PIB	216	790900.0000	127132.4205	586457.00	684628.00	811296.0000	881224.0000	995241.00	408784.00	0.1607
TASA_BANREP	216	5.9306	3.1779	1.75	3.75	4.5000	7.8125	13.25	11.50	0.5359
SALARIO_MINIMO	216	729376.4259	236958.6176	433700.00	535600.00	666902.5000	877803.0000	1300000.00	866300.00	0.3249
TRM	216	2816.4294	901.1836	1712.28	1925.89	2870.6150	3625.7125	4922.30	3210.02	0.3200

Nota. Resultados Estadística Descriptiva.

Análisis

La tabla de Estadística Descriptiva resume el comportamiento de las principales variables del estudio durante el periodo 2007–2024, con un total de 216 observaciones mensuales para cada variable. Esto confirma que la base se encuentra completa para el análisis exploratorio, ya que todas las variables presentan el mismo número de registros: IPC, PIB, Tasa BanRep, Salario

Mínimo y TRM.

En primer lugar, el IPC puro presenta una media de 93.0751, con un valor mínimo de 61.80 y un máximo de 144.88. El rango total es de 83.08 puntos del índice, lo que evidencia un crecimiento importante del nivel general de precios durante el periodo analizado. La mediana del IPC es 88.62, inferior al promedio, lo cual sugiere que los valores más altos observados en los últimos años, especialmente después de 2021, elevan la media de la serie. Esto es coherente con el comportamiento inflacionario pospandemia, donde el IPC tuvo una aceleración significativa.

El PIB presenta una media de 790900.00, con un mínimo de 586457.00 y un máximo de 995241.00. Su rango es de 408784.00, lo cual refleja una expansión importante de la actividad económica en el periodo 2007–2024. Sin embargo, su coeficiente de variación es 0.1607, el más bajo entre las variables analizadas. Esto indica que, aunque el PIB crece en el tiempo, su variabilidad relativa es menor frente a variables como la tasa BanRep, el Salario Mínimo o la TRM. En términos del proyecto, el PIB puede aportar información sobre la tendencia económica general, pero no necesariamente explica por sí solo los cambios más bruscos del IPC.

La Tasa BanRep muestra una media de 5.9306, con un valor mínimo de 1.75 y un máximo de 13.25. Su rango es de 11.50 puntos, y su coeficiente de variación es 0.5359, el más alto de toda la tabla. Esto evidencia que la tasa de intervención del Banco de la República fue la variable con mayor variabilidad relativa dentro del conjunto analizado. Este resultado es relevante porque refleja los cambios en la postura de política monetaria durante diferentes fases económicas: periodos de estímulo con tasas bajas y periodos de control inflacionario con tasas altas. Para el EDA, esto justifica profundizar el cruce entre IPC puro y Tasa BanRep, ya que la política monetaria puede estar relacionada con la evolución del índice, aunque sus efectos suelen presentarse con rezagos.

El Salario Mínimo presenta una media de 729376.43, con un mínimo de 433700.00 y un máximo de 1300000.00. El rango es de 866300.00, mostrando un crecimiento acumulado importante durante el periodo. Su coeficiente de variación es 0.3249, lo que indica una variabilidad relativa moderada-alta. Este comportamiento es esperado, debido a que el Salario Mínimo en Colombia se ajusta anualmente y tiende a crecer con el tiempo. En el contexto del IPC, esta variable es relevante porque puede estar asociada con presiones de costos laborales, ingresos nominales e indexación de algunos precios.

La TRM registra una media de 2816.4294, con un mínimo de 1712.28 y un máximo de 4922.30. Su rango es de 3210.02, y su coeficiente de variación es 0.3200, muy cercano al del Salario Mínimo. Esto indica una variabilidad importante del tipo de cambio durante el periodo 2007–2024. La amplitud del rango evidencia episodios de depreciación del peso colombiano frente al dólar, lo cual puede tener efectos sobre bienes importados, insumos, transporte, combustibles y otros componentes del IPC. Por esta razón, la TRM se mantiene como una variable macroeconómica relevante dentro del análisis predictivo.

Al comparar los coeficientes de variación, se observa que la variable con mayor variabilidad relativa es la Tasa BanRep con 0.5359, seguida por el Salario Mínimo con 0.3249, la TRM con 0.3200, el IPC con 0.2372 y el PIB con 0.1607. Esto permite concluir que las variables monetarias y cambiarias presentan mayor variabilidad relativa que la actividad económica agregada. En términos del proyecto, este hallazgo es importante porque sugiere que los cambios en política monetaria y tipo de cambio pueden aportar señales relevantes para explicar cambios en la dinámica del IPC.

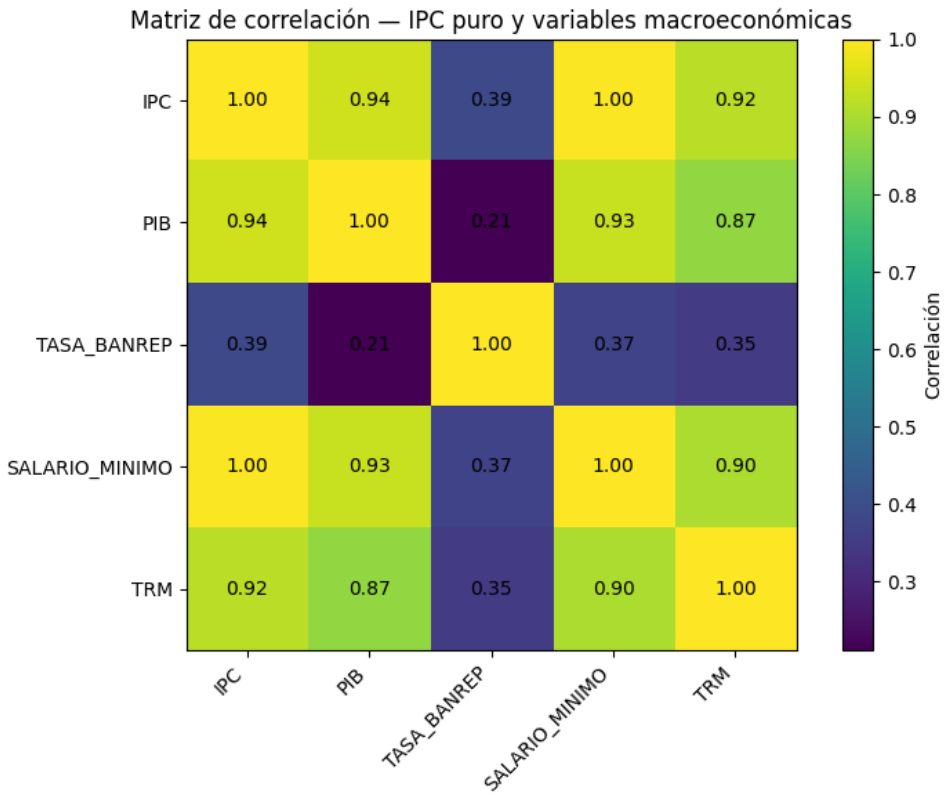
En síntesis, la estadística descriptiva confirma que el IPC puro presenta una trayectoria creciente y una dispersión relevante durante el periodo 2007–2024. Además, muestra que las

variables macroeconómicas seleccionadas tienen comportamientos diferenciados: el PIB refleja una tendencia de crecimiento relativamente estable, la Tasa BanRep presenta alta variabilidad por decisiones de política monetaria, el Salario Mínimo crece de forma acumulativa y la TRM evidencia fluctuaciones cambiarias importantes.

Estos resultados justifican la construcción posterior de una matriz de correlación y el análisis específico del cruce entre IPC puro y Tasa BanRep, con el fin de identificar asociaciones contemporáneas y posibles efectos rezagados sobre el comportamiento del índice.

Figura 23

Matriz de Correlación de las variables



Nota. Resultados de las correlaciones.

Análisis

La Matriz de Correlación muestra la relación lineal entre el IPC puro y las variables macroeconómicas seleccionadas: PIB, Tasa BanRep, Salario Mínimo y TRM. En general, se observa que el IPC, presenta correlaciones positivas con todas las variables, aunque con intensidades diferentes. Esto indica que, durante el periodo 2007–2024, las variables tienden a moverse en la misma dirección que el IPC, especialmente aquellas que también presentan una tendencia creciente en el tiempo.

La correlación más alta se observa entre IPC y Salario Mínimo, con un valor cercano a 1.00. Esto significa que ambas series tienen una relación lineal muy fuerte. Este resultado es esperable, porque tanto el IPC como el Salario Mínimo presentan una trayectoria creciente a lo largo del tiempo. Sin embargo, esta correlación debe interpretarse con cuidado, ya que no implica necesariamente que el Salario Mínimo cause directamente el aumento del IPC, sino que ambas variables comparten una tendencia ascendente de largo plazo.

También se observa una correlación alta entre IPC y PIB, con un valor de 0.94. Esto sugiere que el crecimiento del nivel de precios se ha movido de forma cercana al crecimiento de la actividad económica agregada. En términos económicos, el PIB puede capturar condiciones de demanda agregada y crecimiento económico, aunque la relación no necesariamente es inmediata ni causal.

La relación entre IPC y TRM también es fuerte, con una correlación de 0.92. Este resultado es relevante para el proyecto, porque la TRM representa el canal cambiario. Un aumento del tipo de cambio puede encarecer bienes importados, insumos, combustibles, transporte y algunos productos de la canasta del IPC. Por tanto, esta correlación respalda la inclusión de la TRM como variable explicativa en los modelos multivariados.

En contraste, la Tasa BanRep presenta una correlación más baja con el IPC, de aproximadamente 0.39. Esto indica una relación positiva, pero moderada. Este resultado tiene sentido, porque la tasa de intervención del Banco de la República no necesariamente se mueve de manera simultánea con el IPC. En muchos casos, la política monetaria reacciona al aumento de la inflación y sus efectos sobre los precios pueden observarse con rezagos de varios meses. Por esta razón, no basta con analizar la correlación contemporánea; también es importante evaluar la relación entre el IPC y la Tasa BanRep rezagada a 3, 6, 12 o más meses.

La Matriz también muestra una alta correlación entre PIB, Salario Mínimo y TRM. Por ejemplo, PIB y Salario Mínimo presentan una correlación de 0.93, mientras que PIB y TRM tienen una correlación de 0.87. Esto sugiere posible multicolinealidad entre las variables explicativas, especialmente porque varias de ellas tienen una tendencia creciente en el tiempo. Esta situación justifica el uso de Modelos regularizados como Ridge, ya que la regularización L2 ayuda a estabilizar los coeficientes, cuando los predictores están altamente correlacionados.

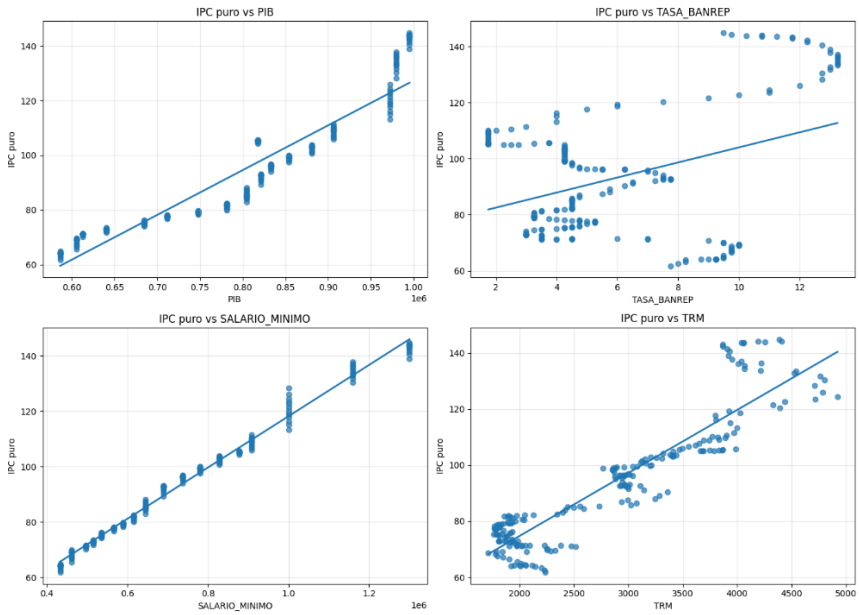
Desde el punto de vista metodológico, esta Matriz aporta evidencia para incluir variables macroeconómicas en el análisis predictivo del IPC, pero también muestra que se debe tener cuidado al interpretar los coeficientes de modelos lineales tradicionales. Las correlaciones altas pueden estar influenciadas por tendencias comunes en el tiempo, por lo que no deben asumirse como relaciones causales directas.

En conclusión, la Matriz de Correlación, confirma que el IPC puro mantiene una asociación fuerte con el Salario Mínimo, el PIB y la TRM, mientras que su relación contemporánea con la Tasa BanRep es más moderada. Este hallazgo respalda la necesidad de profundizar el EDA, mediante análisis con rezagos, especialmente entre IPC y Tasa BanRep, ya que la política monetaria suele afectar la inflación con retraso. Además, la alta correlación entre

varias variables explicativas, justifica el uso de Modelos como Ridge Univariado y Ridge Multivariado, que mostraron mejor desempeño y mayor estabilidad frente a problemas de multicolinealidad.

Figura 24

Gráficos de Dispersión — IPC vs cada variable



Nota. Resultados de las dispersiones.

Diagrama de dispersión: IPC puro vs PIB

El gráfico IPC puro vs PIB, muestra una relación positiva fuerte entre ambas variables. A medida que el PIB aumenta, el IPC también tiende a ubicarse en niveles más altos. La nube de puntos sigue una trayectoria ascendente bastante clara, lo cual coincide con la matriz de correlación, donde la relación entre IPC y PIB fue alta.

Este comportamiento puede interpretarse como una asociación entre el crecimiento de la actividad económica y el aumento del nivel general de precios. Sin embargo, debe aclararse que

esta relación no implica causalidad directa. Tanto el PIB como el IPC tienen una tendencia creciente en el tiempo, por lo que parte de esta asociación puede deberse a que ambas variables evolucionan de forma ascendente durante el periodo 2007–2024.

En términos del Proyecto, este gráfico justifica la inclusión del PIB como variable macroeconómica explicativa en los modelos multivariados, ya que aporta información relacionada con la actividad económica agregada y la demanda.

Diagrama de dispersión: IPC puro vs TASA_BANREP

El gráfico IPC puro vs TASA_BANREP muestra una relación positiva, pero más dispersa y menos lineal que en los otros cruces. La línea de tendencia es ascendente, lo cual indica que, en términos generales, niveles más altos de la tasa BanRep se asocian con niveles más altos del IPC. Sin embargo, los puntos están bastante dispersos, lo que evidencia una relación contemporánea moderada.

Este comportamiento es coherente con la lógica de la política monetaria. La tasa de intervención del Banco de la República no necesariamente afecta el IPC en el mismo mes, sino que suele actuar con rezagos. Además, la tasa BanRep también puede reaccionar a la inflación ya observada; es decir, cuando el IPC o la inflación aumentan, el Banco de la República puede elevar la tasa para controlar las presiones inflacionarias.

Por esta razón, el gráfico no debe interpretarse como una relación causal directa inmediata. Más bien, indica que es necesario complementar el análisis con rezagos de la tasa BanRep, por ejemplo: a 3, 6, 12 o 18 meses, para observar mejor la transmisión de la política monetaria sobre el IPC.

Diagrama de dispersión: IPC puro vs SALARIO_MINIMO

El gráfico IPC puro vs SALARIO_MINIMO presenta una relación positiva muy fuerte y

casi lineal. Los puntos se alinean claramente alrededor de la línea de tendencia, lo que indica que ambas variables se mueven de forma muy cercana durante el periodo analizado.

Este resultado es esperado, porque tanto el IPC como el Salario Mínimo presentan una tendencia creciente en el tiempo. El Salario Mínimo se ajusta anualmente, mientras que el IPC mide la evolución acumulada del nivel de precios. Por tanto, ambas series comparten un patrón ascendente de largo plazo.

Desde el punto de vista Económico, el Salario Mínimo puede relacionarse con el IPC mediante canales de ingreso nominal, costos laborales e indexación. Sin embargo, la alta relación observada también puede reflejar una tendencia común en el tiempo, por lo que debe interpretarse con prudencia. En el modelado, esta alta asociación puede generar multicolinealidad con otras variables también crecientes, como PIB y TRM, lo cual justifica el uso de modelos regularizados como Ridge.

Diagrama de dispersión: IPC puro vs TRM

El gráfico IPC puro vs TRM muestra una relación positiva clara, aunque con mayor dispersión que el Salario Mínimo y el PIB. A medida que la TRM aumenta, el IPC tiende a ubicarse en niveles más altos. Esto respalda la importancia de la tasa de cambio como variable explicativa en el análisis del IPC.

La relación positiva puede interpretarse desde el canal de transmisión cambiaria. Cuando el peso colombiano se deprecia frente al dólar, los bienes importados, insumos externos, combustibles, transporte y algunos productos de la canasta del IPC pueden encarecerse. Por esta razón, una TRM más alta puede estar asociada con mayores presiones sobre el nivel de precios.

Sin embargo, el gráfico también muestra cierta dispersión en valores medios y altos de TRM, lo que indica que el tipo de cambio no explica por sí solo todo el comportamiento del IPC.

El IPC también depende de otros factores, como política monetaria, salarios, actividad económica, costos internos, alimentos y choques externos.

Conclusión General. Los Diagramas de Dispersión confirman que el IPC puro, mantiene relaciones positivas con las principales variables macroeconómicas del estudio. Las asociaciones más fuertes y ordenadas se observan con Salario Mínimo y PIB, mientras que la relación con TRM también es positiva, pero con mayor dispersión. En cambio, la relación con Tasa BanRep, es más moderada y dispersa, lo que sugiere que su efecto sobre el IPC puede manifestarse con rezagos y no necesariamente de forma contemporánea.

Estos resultados fortalecen la decisión metodológica de incluir variables macroeconómicas en los Modelos Multivariados. Al mismo tiempo, evidencian la posible presencia de multicolinealidad, especialmente, entre variables con tendencia creciente. Por ello, el uso de modelos como Ridge Multivariado resulta adecuado, ya que permite estabilizar los coeficientes y reducir el riesgo de sobreajuste, cuando las variables explicativas están altamente correlacionadas.

Resultados del Objetivo Específico 2: Comparación de Modelos

Como se observa en la Tabla 6, el Naive Forecast obtuvo el mejor desempeño predictivo global en el conjunto de prueba, alcanzando R^2_test = 0.9963, RMSE_test = 0.8926 y MAE_test = 0.7247. Este resultado evidencia la elevada persistencia temporal del IPC puro, indicando que el valor observado en un periodo contiene información altamente relevante para anticipar el comportamiento del periodo siguiente.

Entre los Modelos de Machine Learning, el Ridge Univariado presentó el mejor desempeño, con R^2_test = 0.9611, RMSE_test = 2.8800 y MAE_test = 2.2752. Estos resultados confirman que la propia historia del IPC contiene una señal predictiva fuerte y que la

incorporación de rezagos del índice permite capturar adecuadamente su dinámica temporal. Por esta razón, el Ridge Univariado se identificó como el mejor modelo de Machine Learning de la investigación.

El Ridge Multivariado también mostró un desempeño elevado, con $R^2\_test = 0.9352$, destacándose por incorporar variables macroeconómicas como el PIB, la Tasa de Intervención del Banco de la República, el Salario Mínimo Mensual Legal Vigente y la TRM. Aunque su desempeño fue ligeramente inferior al del Ridge Univariado, aporta una mayor capacidad interpretativa sobre los factores económicos asociados al comportamiento del IPC.

La Regresión Lineal Múltiple obtuvo $R^2\_test = 0.9117$, evidenciando que la selección adecuada de variables y el uso del IPC puro permitieron construir una línea base multivariada sólida. Este resultado demuestra que las relaciones lineales entre las variables macroeconómicas y el IPC conservan capacidad explicativa significativa durante el periodo analizado.

Por su parte, el ARIMA Univariado, alcanzó $R^2\_test = 0.1479$, mostrando una capacidad limitada para capturar completamente los cambios observados durante el periodo pospandemia. Aunque logró representar parcialmente la dinámica temporal del IPC, su desempeño fue inferior al de los modelos Ridge y al benchmark Naive Forecast.

La Suavización Exponencial de Holt registró un $R^2\_test = 0.0020$, indicando que una tendencia suavizada resulta insuficiente para representar los cambios estructurales y la aceleración inflacionaria observada entre 2020 y 2024. Este resultado confirma que el comportamiento reciente del IPC presenta variaciones que exceden una simple tendencia lineal suavizada.

La Regresión Lineal Simple presentó un $R^2\_test = -0.1880$, reflejando que una tendencia temporal lineal no logra capturar adecuadamente la complejidad del comportamiento

inflacionario durante el periodo de Prueba. La presencia de un coeficiente de determinación negativo indica que el modelo resulta menos preciso que una predicción basada en el promedio histórico.

Los Modelos de Ensamble basados en árboles, Gradient Boosting Regressor y ExtraTreesRegressor, obtuvieron $R^2\_test = -2.1762$ y $R^2\_test = -1.8622$, respectivamente. Aunque ambos alcanzaron niveles muy altos de ajuste en Entrenamiento, evidenciaron problemas de generalización fuera de muestra. Estos resultados indican que los Modelos basados en Árboles tuvieron dificultades para extrapolar la aceleración del IPC observada durante el periodo pospandemia, generando errores significativamente superiores a los obtenidos por los modelos lineales y los Benchmarks Univariados.

En conjunto, los resultados muestran que la persistencia temporal del IPC constituye la principal fuente de capacidad predictiva en la serie analizada. El Naive Forecast se consolidó como el Benchmark predictivo más preciso, mientras que el Ridge Univariado representó la mejor alternativa de Machine Learning, combinando un elevado desempeño predictivo con una adecuada capacidad de generalización.

Resultados del Objetivo Específico 3: Interpretación Macroeconómica

Los resultados confirman que la inflación medida a través del IPC puro durante 2020–2024 respondió a una combinación de persistencia interna, choques de costos, transmisión cambiaria, ajustes monetarios y cambios en la actividad económica. La superioridad de los Modelos Ridge es coherente con un entorno en el que el IPC conserva fuerte inercia, pero también se ve afectado por perturbaciones macroeconómicas. La regularización L2 ayuda a estabilizar los coeficientes en presencia de variables correlacionadas, una condición habitual en series macroeconómicas.

Uno de los hallazgos más relevantes del estudio es que el Naive Forecast obtuvo el mejor desempeño predictivo global, alcanzando un R^2_test = 0.9963, un RMSE_test = 0.8926 y un MAE_test = 0.7247. Este resultado demuestra que el IPC puro presenta una elevada persistencia temporal y que el valor observado en un periodo contiene información altamente relevante para anticipar el comportamiento del periodo siguiente. Desde una perspectiva económica, este comportamiento puede asociarse con mecanismos de indexación, ajustes graduales de precios, transmisión escalonada de costos y expectativas inflacionarias que permanecen activas durante varios meses.

Entre los Modelos de Machine Learning, el Ridge Univariado, obtuvo el mejor desempeño, con R^2_test = 0.9611, confirmando que los rezagos del propio IPC contienen una señal predictiva muy fuerte. Este resultado respalda la hipótesis de que la inflación observada en Colombia durante el periodo pospandemia posee una importante componente inercial, donde los niveles de precios pasados ayudan a explicar significativamente los niveles futuros del índice.

El Ridge Multivariado también presentó resultados sobresalientes (R^2_test = 0.9352), evidenciando que las variables macroeconómicas seleccionadas aportan información valiosa para contextualizar el comportamiento del IPC puro. La inclusión de PIB, Tasa BanRep, Salario Mínimo y TRM permite interpretar la inflación desde distintos canales: actividad económica, política monetaria, costos laborales y transmisión cambiaria. Por tanto, este modelo conserva una importancia significativa no solo por su capacidad predictiva, sino también por su valor explicativo dentro del análisis económico.

La Regresión Lineal Múltiple obtuvo un desempeño igualmente alto (R^2_test = 0.9117), lo que confirma que las relaciones lineales entre el IPC y las variables macroeconómicas conservan capacidad explicativa relevante cuando se trabaja con variables depuradas y

consistentes con el IPC puro. Aunque fue superada por los modelos Ridge, constituye una línea base multivariada robusta para la interpretación económica.

Por su parte, el ARIMA Univariado alcanzó un R^2_test = 0.1479, mientras que la Suavización Exponencial de Holt obtuvo un R^2_test = 0.0020. Estos resultados indican que los métodos clásicos de series temporales lograron capturar parcialmente la tendencia general del IPC, pero mostraron limitaciones para representar completamente los cambios estructurales y la aceleración inflacionaria observada durante el periodo pospandemia. Ambos modelos son útiles como referencia metodológica, aunque fueron superados ampliamente por los enfoques basados en persistencia temporal y regularización.

La Regresión Lineal Simple, presentó un R^2_test = -0.1880, reflejando que una tendencia temporal lineal aislada no logra capturar adecuadamente la complejidad del comportamiento inflacionario reciente. Este resultado evidencia que la dinámica del IPC depende de múltiples factores económicos y no puede representarse únicamente mediante una tendencia creciente en el tiempo.

Los Modelos de Ensamble basados en árboles, Gradient Boosting Regressor (R^2_test = -2.1762) y ExtraTreesRegressor (R^2_test = -1.8622), mostraron una limitación importante: aunque alcanzaron niveles muy elevados de ajuste en entrenamiento, no lograron generalizar adecuadamente durante el periodo 2020–2024. Desde una perspectiva económica, este resultado puede explicarse porque el IPC de prueba alcanzó niveles superiores a los observados en el conjunto de entrenamiento. Los modelos basados en árboles suelen presentar dificultades para extrapolar valores fuera del rango histórico aprendido, razón por la cual subestimaron la aceleración inflacionaria posterior a 2021.

La TRM tiene un papel destacado porque representa el canal cambiario. Cuando el peso

colombiano se deprecia frente al dólar, pueden aumentar los precios de bienes importados, insumos agrícolas, combustibles, maquinaria, transporte y productos con componentes externos. Este efecto puede trasladarse gradualmente al consumidor final, afectando el IPC. La Tasa BanRep representa la postura de política monetaria, aunque sus efectos suelen manifestarse con rezagos. El PIB aporta información sobre la demanda agregada y el Salario Mínimo permite aproximar presiones asociadas a costos laborales e ingresos nominales. En conjunto, estas variables complementan la explicación del fenómeno inflacionario, aunque no desplazan la importancia de la persistencia temporal del propio IPC.

En conjunto, los resultados sugieren que la inflación medida mediante el IPC puro debe interpretarse como un fenómeno persistente y multicausal. La persistencia del índice, explica por qué el Naive Forecast y el Ridge Univariado alcanzaron los mejores desempeños predictivos, mientras que las variables macroeconómicas permiten comprender los mecanismos económicos que influyeron sobre el nivel general de precios durante el periodo analizado. Por esta razón, el Naive Forecast se consolidó como el mejor benchmark predictivo general, mientras que el Ridge Univariado se identificó como el mejor modelo de Machine Learning de la investigación.

Finalmente, la interpretación macroeconómica muestra que el Modelo Predictivo, no debe entenderse como una herramienta aislada, sino como un complemento del análisis económico. Los resultados estadísticos permiten identificar qué modelos predicen mejor el IPC, mientras que la interpretación económica permite explicar por qué ocurre dicho comportamiento. En este sentido, el proyecto aporta una herramienta útil para el seguimiento técnico de la inflación colombiana, integrando evidencia cuantitativa, análisis exploratorio y comprensión macroeconómica del periodo 2020–2024.

Discusión Crítica de los Resultados y Aportes a la Disciplina

Análisis Crítico de los Hallazgos a la luz del Marco Teórico

Los hallazgos actualizados muestran que el Ridge Univariado fue el modelo con mejor desempeño global para estimar el IPC puro en Colombia durante 2020-2024, seguido por el Ridge Multivariado y la Regresión Lineal Múltiple. Esta evidencia es coherente con la literatura que caracteriza la inflación pospandemia como un fenómeno persistente y multicausal. La superioridad de los modelos regularizados indica que la estabilidad y la parsimonia pueden ser más efectivas que la complejidad algorítmica en muestras macroeconómicas limitadas y con cambios de régimen.

Los Modelos de Ensamble, a pesar de su flexibilidad, mostraron dificultades para generalizar. Este resultado no niega su utilidad, pero sí evidencia que los algoritmos basados en árboles pueden tener limitaciones para extrapolar series con tendencia creciente fuera del rango observado. Holt, por su parte, confirmó que los modelos basados solo en nivel y tendencia pueden quedarse cortos cuando existen rupturas estructurales.

Contribución de los Resultados a la Solución de la Problemática y a la Disciplina

Desde el plano teórico, el trabajo contribuye a comprender la inflación pospandemia en Colombia como un fenómeno persistente, multicausal y sensible a cambios de régimen. Desde el plano práctico, ofrece una herramienta predictiva reproducible para el seguimiento de corto plazo del IPC. Desde el plano metodológico, demuestra la importancia de comparar modelos con validación temporal estricta y de priorizar métricas fuera de muestra.

El aporte principal es identificar al Ridge Univariado como modelo operativo de pronóstico base y al Ridge Multivariado como Modelo complementario para análisis contextual. Esta combinación permite equilibrar precisión predictiva, interpretabilidad y robustez.

Autocrítica y Debilidades del Trabajo

El estudio presenta limitaciones. Primero, utiliza datos agregados nacionales y no captura diferencias regionales ni divisiones específicas del IPC, como alimentos, vivienda o transporte. Segundo, no incorpora factores como precios internacionales de alimentos y energía, expectativas de inflación, costos logísticos, precios regulados o choques climáticos. Tercero, la evaluación se basa en una muestra temporal limitada para modelos flexibles, por lo que los resultados deben interpretarse como evidencia predictiva y no como prueba causal.

Asimismo, el Modelo seleccionado debe entenderse como herramienta de apoyo para monitoreo y alerta temprana, no como solución definitiva a la inflación. Las decisiones de política pública requieren integrar análisis sectorial, institucional y macroeconómico complementario.

Una de las principales limitaciones del estudio radica en que el comportamiento de la inflación durante el periodo pospandemia estuvo influenciado por choques externos extraordinarios, incluyendo interrupciones en cadenas de suministro, presiones internacionales sobre precios de materias primas y ajustes de política monetaria. Estos factores pueden reducir la capacidad de generalización de los modelos cuando se enfrentan a escenarios económicos diferentes a los observados entre 2020 y 2024. Como línea futura de investigación, se recomienda contrastar los resultados obtenidos con Modelos adicionales como SARIMAX, Random Forest y arquitecturas basadas en aprendizaje profundo.

Diferencia entre Capacidad Predictiva y Causalidad Económica

Los resultados del estudio, deben interpretarse como evidencia predictiva y no como prueba de causalidad económica directa. Debido a que un Modelo que use variables como PIB, TRM, Tasa BanRep o Salario Mínimo para estimar el IPC, no implica que dichas variables

causen por sí solas el comportamiento observado del índice. La predicción identifica patrones estadísticos útiles para anticipar valores futuros, mientras que la causalidad requiere diseños econométricos o experimentales específicos orientados a aislar efectos.

Esta distinción es fundamental, porque el IPC pospandemia, respondió a una interacción de factores internos y externos. La alta correlación entre algunas variables, puede estar influenciada por tendencias comunes en el tiempo, multicolinealidad y cambios estructurales. Por tanto, las conclusiones del Proyecto se formulan en términos de capacidad predictiva, estabilidad fuera de muestra e interpretación macroeconómica contextual, no como afirmaciones causales definitivas.

Conclusiones, Recomendaciones y Limitaciones

Conclusiones Generales

Primero, el estudio permitió comprobar que el IPC puro en Colombia, presentó una trayectoria ascendente sostenida entre 2007 y 2024, evidenciando un comportamiento estructural de largo plazo caracterizado por una tendencia creciente y una elevada persistencia temporal. El análisis exploratorio de datos, mostró una aceleración particularmente pronunciada entre 2021 y 2023, asociada al contexto pospandemia, seguida de una moderación parcial durante 2024. Este comportamiento, confirma la existencia de un cambio de régimen inflacionario, respecto a los años previos a la pandemia y demuestra que la dinámica reciente del IPC, no puede interpretarse únicamente como una prolongación de la tendencia histórica observada entre 2007 y 2019.

Adicionalmente, la continuidad y suavidad de la serie, evidencian una fuerte inercia inflacionaria, donde los niveles pasados del índice conservan una capacidad explicativa significativa sobre los niveles futuros. Este hallazgo resulta especialmente relevante, ya que proporciona sustento empírico a los resultados obtenidos por los Modelos Predictivos

implementados, particularmente, el Ridge Univariado y el Naive Forecast y, a su vez, confirma la importancia de analizar la inflación, mediante Series Temporales de largo plazo para capturar adecuadamente los cambios estructurales y los mecanismos de persistencia presentes en la economía colombiana.

Segundo, la comparación de los nueve modelos implementados, evidenció diferencias importantes en capacidad predictiva y generalización fuera de muestra. Los resultados muestran que el Naive Forecast o Caminata Aleatoria obtuvo el mejor desempeño predictivo general, alcanzando los menores errores de predicción y el mayor coeficiente de determinación en el periodo de prueba 2020–2024. Este hallazgo confirma la elevada persistencia temporal del IPC puro y demuestra que el último valor observado contiene una cantidad significativa de información para anticipar el comportamiento inmediato del índice.

Tercero, entre los modelos de Machine Learning evaluados, el Ridge Univariado obtuvo el mejor desempeño, convirtiéndose en la alternativa predictiva más sólida dentro de los enfoques avanzados. Su capacidad para capturar la persistencia temporal del IPC, mediante rezagos del propio índice y la utilización de regularización L2 que ayuda a estabilizar los coeficientes, cuando los predictores están altamente correlacionados, los cuales permitieron alcanzar altos niveles de precisión y estabilidad fuera de muestra.

Cuarto, el Ridge Multivariado y la Regresión Lineal Múltiple también presentaron resultados favorables. Estos modelos aportan valor interpretativo al incorporar variables macroeconómicas como el PIB, la Tasa de Intervención del Banco de la República, el Salario Mínimo y la TRM, permitiendo analizar distintos canales económicos asociados al comportamiento de la inflación.

Quinto, el ARIMA Univariado mostró una capacidad moderada para representar la

dinámica temporal del IPC, logrando capturar parcialmente la tendencia observada, aunque con un desempeño inferior al obtenido por los Modelos Ridge y el Naive Forecast. Por su parte, la Regresión Lineal Simple evidenció limitaciones importantes al intentar representar la evolución del IPC únicamente mediante una tendencia lineal, obteniendo resultados inferiores frente a los demás modelos evaluados.

Sexto, los Modelos de Ensamble Gradient Boosting Regressor y ExtraTreesRegressor, presentaron un elevado ajuste durante el entrenamiento, pero una baja capacidad de generalización fuera de muestra. Este comportamiento indica que una mayor complejidad algorítmica no garantiza mejores pronósticos cuando la serie de prueba presenta niveles superiores a los observados durante el entrenamiento, situación característica del periodo inflacionario posterior a la pandemia.

Séptimo, la Suavización Exponencial de Holt funcionó como un benchmark clásico de series temporales y logró representar adecuadamente la tendencia general del IPC. Sin embargo, no consiguió capturar completamente la aceleración inflacionaria observada entre 2021 y 2023, por lo que su utilidad principal radica en servir como referencia metodológica frente a modelos más sofisticados. Este resultado es consistente con lo reportado en la literatura, donde se reconoce que los métodos de Suavización Exponencial, ofrecen un buen desempeño para modelar tendencias estables, pero presentan limitaciones cuando la serie experimenta cambios estructurales o choques económicos significativos, como los observados durante el periodo pospandemia.

Finalmente, los resultados de la investigación, permiten concluir que la persistencia temporal constituye el principal factor predictivo del IPC puro en Colombia. Asimismo, evidencian que las variables macroeconómicas complementan la interpretación económica del

fenómeno inflacionario, aunque no superan la capacidad predictiva derivada de la propia historia del índice.

En consecuencia, el Naive Forecast se consolida como el mejor benchmark predictivo general, mientras que el Ridge Univariado se identifica como el mejor modelo de Machine Learning desarrollado en esta investigación.

Recomendaciones

Se recomienda utilizar el Ridge Univariado como modelo operativo de pronóstico base para el seguimiento de corto plazo del IPC. Su alto desempeño fuera de muestra, estabilidad y sencillez lo hacen adecuado para procesos de monitoreo periódico.

Así mismo, como recomendación, usar el Modelo Naive Forecast (Caminata Aleatoria), como línea base operativa de corto plazo y el Ridge Univariado, como Modelo de Machine Learning principal. En actualizaciones mensuales, ambos modelos deben recalibrarse y compararse para verificar si el modelo más complejo mantiene una mejora real, frente a la Caminata Aleatoria.

En ese sentido, también se aconseja usar el Ridge Multivariado como Modelo complementario para análisis contextual, especialmente, cuando se requiera interpretar el papel de las otras variables macroeconómicas como: PIB, Tasa de Intervención del BanRep, Salario Mínimo y la Tasa Representativa del Mercado (TRM) en la evolución del índice.

Adicionalmente, se recomienda conservar la Regresión Lineal Simple, la Regresión Lineal Múltiple, el ARIMA Univariado y la Suavización Exponencial de Holt como modelos de referencia metodológica. Aunque su desempeño fue inferior al alcanzado por el Ridge Univariado y el Naive Forecast, estos enfoques permiten contrastar diferentes estrategias de modelado y aportan elementos de interpretación útiles para futuras investigaciones sobre

inflación y series temporales.

Frente a lo anterior, como consejo, se busca mantener una actualización mensual de la Base de Datos y en su defecto, anual (para los casos que aplique) y a su vez, recalibrar los Modelos de manera periódica. También, se sugiere construir tableros de seguimiento que incluyan IPC, rezagos del IPC, TRM, Tasa BanRep, Salario Mínimo y PIB.

A su vez, se recomienda priorizar el seguimiento de alimentos y componentes sensibles del IPC, dado que suelen responder con rapidez a choques de oferta, transporte, clima y tipo de cambio.

Finalmente, como conclusión, se aconseja evitar depender exclusivamente de Modelos complejos de Ensamble para decisiones institucionales. En contextos de Series Temporales con tendencia, la parsimonia y la estabilidad, pueden ser más valiosas que la complejidad.

Limitaciones

La principal limitación es el uso de datos agregados nacionales, lo que impide capturar diferencias regionales y sectoriales. Otra limitación es la ausencia de variables internacionales y sectoriales como precios de alimentos, energía, combustibles, expectativas de inflación, costos logísticos y condiciones climáticas. Además, el estudio se centra en predicción y no establece relaciones causales definitivas.

Otra limitación metodológica es que la comparación principal, se basa en una partición temporal única. Aunque esta partición evita fuga de información, futuras versiones deberían aplicar backtesting con múltiples ventanas para evaluar estabilidad. Además, la superioridad del Naive Forecast, evidencia que la alta persistencia del IPC puede hacer que Modelos simples superen a Modelos de Machine Learning en horizontes cortos.

Asimismo, debe considerarse que los nueve modelos evaluados presentan fortalezas y limitaciones diferentes. Los modelos lineales y regularizados mostraron una mejor capacidad de generalización, mientras que los modelos de ensamble basados en árboles presentaron dificultades para extrapolar cambios estructurales observados durante el periodo pospandemia.

De igual forma, los métodos clásicos de Series Temporales, como ARIMA y Holt, lograron representar parcialmente la tendencia del IPC, pero mostraron limitaciones frente a la persistencia observada en la serie. Por esta razón, los resultados deben interpretarse considerando las características particulares de cada metodología implementada.

También se reconoce que el periodo 2020-2024, contiene choques excepcionales, por lo que los modelos deben actualizarse permanentemente para mantener su capacidad predictiva. Los resultados son válidos bajo la estructura de datos, variables y partición temporal utilizada en este proyecto.

¿Cómo mitigar las Limitaciones?

Las limitaciones identificadas podrían mitigarse en futuras investigaciones, mediante la incorporación de información regional y sectorial que permita analizar diferencias territoriales en el comportamiento del IPC dentro de Colombia. Asimismo, la inclusión de variables adicionales relacionadas con alimentos, vivienda, transporte, combustibles, servicios públicos, expectativas de inflación y costos logísticos, podría fortalecer la capacidad explicativa y predictiva de los modelos.

Desde el punto de vista Metodológico, se recomienda complementar la partición temporal utilizada con esquemas de validación más robustos, como backtesting, rolling window y expanding window, con el fin de evaluar la estabilidad de los resultados en diferentes horizontes temporales. También se sugiere explorar modelos híbridos que combinen la persistencia

temporal capturada por el Naive Forecast y el Ridge Univariado con variables macroeconómicas nacionales adicionales, así como actualizar periódicamente las bases de datos y recalibrar los modelos para adaptarlos a los cambios estructurales que puedan presentarse en la economía colombiana.

Referencias

Banco de la República. (s. f.). *Series Macroeconómicas Oficiales: PIB, TRM, Tasa de Intervención y Salario Mínimo*. https://www.banrep.gov.co/es/estadisticas-economicas/series-estadisticas-historicas-colombia

Baquero Beltrán, A., & Villamil, J. (2022). *Pandemia y política económica: la Política Monetaria en discusión.* Revista de Economía Institucional, 24(46), 167-193. https://doi.org/10.18601/01245996.V24N46.09

Bargain, O., & Aminjonov, U. (2021). *Poverty and COVID-19 in Africa and Latin America.* World Development, 142, 105422. https://doi.org/10.1016/j.worlddev.2021.105422

Bonam, D., & Smădu, A. (2021). *The long-run effects of pandemics on inflation: will this time be different?* Economics Letters, 208, 110065. https://doi.org/10.1016/j.econlet.2021.110065

Carlomagno, G., Fornero, J., & Sansone, A. (2023). *A proposal for constructing and evaluating core inflation measures.* Latin American Journal of Central Banking, 4(3), 100094. https://doi.org/10.1016/j.latcb.2023.100094

Cuevas Ahumada, V. M., & Perrotini Hernández, I. (2024). *Consumer goods and services inflation in Latin America during the COVID-19 pandemic.* Brazilian Journal of Political Economy, 45(1), e253580. https://doi.org/10.1590/0101-31572025-3580

Departamento Administrativo Nacional de Estadística. (s. f.). *Índice de Precios al Consumidor (IPC), series oficiales mensuales*. https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc

Departamento Administrativo Nacional de Estadística. (2019). *Metodología general Índice de Precios al Consumidor – IPC.* Código DSO-IPC-MET-001, Versión 7. https://www.dane.gov.co/files/investigaciones/fichas/precios-y-costos/DSO-IPC-MET-001-v7.pdf

Friedman, J. H. (2001). *Greedy function approximation: a Gradient Boosting Machine.* The Annuals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Geurts, P., Ernst, D., & Wehenkel, L. (2006). *Extremely Randomized Trees.* Machine Learning, 63(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Hoerl, A. E., & Kennard, R. W. (1970). *Ridge Regression: biased estimation for nonorthogonal problems.* Technometrics, 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Holt, C. C. (2004). *Forecasting seasonals and trends by exponentially weighted moving averages.* International Journal of Forecasting, 20(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice.* (3rd ed.). OTexts. https://otexts.com/fpp3/

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: with applications in Python*. Springer. https://doi.org/10.1007/978-3-031-38747-0

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models.* (5th ed.). McGraw-Hill Irwin.

https://d1b10bmlvqabco.cloudfront.net/attach/is282rqc4001vv/is6ccr3fl0e37q/iwfnjvgvl53z/Michael_H_Kutner_Christopher_J._Nachtsheim_JohnBookFi.org.pdf

Pérez Gelves, J. J., Østergaard, P. A., & Díaz Flórez, G. A. (2023). *Energy poverty assessment and the impact of COVID-19: an empirical analysis of Colombia.* Energy Policy, 181, 113716. https://doi.org/10.1016/j.enpol.2023.113716

Python (2026). Download. https://www.python.org/

Saavedra, M. (2026). Notebook de Google Colab con códigos, análisis, conclusiones, etc. https://drive.google.com/file/d/1UYipS4FEQVYl-joh97yLjh9fABRvpe0l/view?usp=sharing

Tantawi, R. P. (2024). *Machine Learning.* Salem Press Encyclopedia. https://openurl-ebsco-com.bibliotecavirtual.unad.edu.co/contentitem/ers:90558380

Torres-Favela, M., & Luna, E. M. (2025). *The role of informality in the economic growth, employment, and inflation during the COVID-19 crisis.* Latin American Journal of Central Banking, 6(1), 100150. https://doi.org/10.1016/j.latcb.2024.100150

Zhao, Y., Huang, C., & Luo, J. (2022). *How to prepare for the next pandemic-Investigation of correlation between food prices and COVID-19 from global and local perspectives.* In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data) (pp. 6135-6144). IEEE. https://doi.org/10.1109/BIGDATA55660.2022.10020906

Zhu, X., & Yu, X. (2025). *Food inflation and Macroeconomic Dynamics in the US: evidence from an estimated DSGE Model.* Finance Research Letters, 75, 106895. https://doi.org/10.1016/j.frl.2025.106895