PREDICCIÓN DEL COMPORTAMIENTO DEL INDICE DE PRECIOS AL CONSUMIDOR EN COLOMBIA MEDIANTE ALGORITMOS MACHINE LEARNING

TRABAJO DE GRADO
Opción seminario-diplomado

Corporación Universitaria Remington.
Facultad de Ingenierías
pregrado:
Especialización en dirección de operaciones en mejoramiento continuo.

Estudiante:
Linda Maria Heredia Castillo
CC 1090371691

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.

2025

Dedicatoria

A Dios, a mamita Maria, a mi esposo y mi familia a quienes siento presente en cada momento de mi vida.

Agradecimientos

A Dios, a la universidad a mis compañeras de especialización Kathe y Jessi, al entusiasmo del profe Juan Carlos Briñez de León para dirigir este seminario.

Tabla de contenidos

1. Resumen

Este trabajo de investigación se centra en la predicción del comportamiento del Índice de Precios al Consumidor (IPC) en Colombia utilizando algoritmos de machine learning. Se analizan datos históricos de variación mensual del IPC desde 2014 hasta 2024, proporcionados por el DANE. Se exploran diferentes modelos predictivos, incluyendo regresión lineal, regresión polinómica, y varios algoritmos de machine learning como Random Forest, XGBoost y SVR. Los resultados indican que los modelos basados en Random Forest y XGBoost ofrecen el mejor rendimiento predictivo. Además, se observa que ni la ciudad ni la categoría tienen una relación lineal fuerte con el IPC mensual, sugiriendo la necesidad de explorar modelos no lineales. Este estudio destaca la importancia de utilizar datos limpios y modelos adecuados para mejorar la precisión de las predicciones económicas.

2. Palabras clave

Índice de Precios al Consumidor (IPC), Machine Learning, Predicción, Modelos no lineales, Random Forest, Análisis de datos

3. Marco conceptual y contextual

El IPC, o índice de precios al consumidor mide la disminución o el aumento de los precios, esta cifra va relacionada con la opción de compra de los consumidores en alrededor de doce categorías, tales como: alimentos y bebidas, prendas de vestir, arriendos, servicios públicos, artículos para el hogar, salud, transportes, servicios de información y comunicación, recreación, cultura, educación, restaurantes y hoteles, y otros bienes o múltiples servicios diversos, tal como indica el Dane, El IPC mide la variación de precios de los bienes y servicios más consumidos por los hogares colombianos, agrupados en doce categorías, y se calcula a partir de más de 55.000 fuentes en 38 ciudades del país (Departamento Administrativo Nacional de Estadística – DANE, s.f.).

A todos los sectores del país les afecta la inflación y el cierre de IPC anual, dado que este representa el índice de aumento para el año siguiente para sistemas de arrendamiento, salarios contractuales, vivienda, contratos en general, entre otros y de acuerdo como lo menciona el (Banco de la República, 2024), para lo que resta del año y para 2026, la inflación total continuaría reduciéndose en un contexto de una actividad económica en recuperación, pero con excesos de capacidad productiva, la indexación de algunos precios a una inflación más baja, y presiones moderadas de la tasa de cambio sobre los precios. De esta manera, se espera que la inflación converja al 3% a finales de 2026, con lo anterior se infiere que siempre que se habla de este tema e requieren ciertas predicciones o algoritmos, y aunque nada es seguro, es un tema relativo en este trabajo construiremos modelos a basados en machine learning,

para predecir un modelo que busque basado en datos históricos predecir un comportamiento del IPC.

La minería de datos o los conjuntos de datos son muy importantes en la actualidad, unos datos depurados y limpios son un oro líquido, para cualquier empresa o sector, y si nos centramos encontrar patrones ocultos, y lograr a través de los datos predecir comportamientos del mercado y tomar decisiones estratégicas basadas en la evidencia de los datos.

Como indica (Caro & Peña, 2021), el paradigma clásico de la estadística es encontrar el mejor modelo. El concepto de modelo óptimo está bien definido en entornos simples, bajo fuertes hipótesis sobre el proceso generador de los datos, pero empieza a desdibujarse cuando admitimos incertidumbre sobre este proceso o prescindimos de un modelo generador único, con esto si queda en manifiesto que no existe un único modelo, y que no sirve un único modelo, y es razonable pensar que un modelo que funciona hoy, en un futuro se desdibuje, y es necesario comprender señales de alerta en modelo, al igual que comparar modelos para encontrar el que se ajuste a los datos.

4. Pregunta problema

¿una estrategia computacional basada en algoritmos de machine learning puede predicción del comportamiento del índice de precios al consumidor en Colombia mediante algoritmos machine learning?

5. Acercamiento a los datos:

Estos datos fueron extraídos de la página de kaggle, (Córdova Rosas, 2024), comprende un conjunto de datos del DANE, con alrededor de 12 categorias, total (Sin categorizar), Alimentos y bebidas no alcohólicas, Bebidas alcohólicas y tabaco, Prendas de vestir y calzado, Alojamiento, agua, electricidad, gas y otros combustibles, Muebles, artículos para el hogar y para la conservación ordinaria del hogar, Salud, Transporte, Información y comunicación, Recreación y cultura, Educación, Restaurantes y hoteles, Bienes y servicios diversos, todas estas categorías distribuidas para 23 ciudades de Colombia, los datos representan la variación mensual del IPC desde el año 2014 al 2024.

6. Objetivos

6.1 Objetivo general

- Desarrollar un modelo predictivo basado en algoritmos de *machine learning* que permita estimar el comportamiento del Índice de Precios al Consumidor (IPC) en Colombia, utilizando datos históricos de variación mensual entre 2014 y 2024.

6.2 Objetivo específicos

- Analizar el comportamiento histórico del IPC por categorías de gasto y ciudades en Colombia entre 2014 y 2024, a partir de datos proporcionados por el DANE.
- Depurar el conjunto de datos para garantizar su calidad y adecuación a modelos de aprendizaje automático.
- Seleccionar el modelo con mayor precisión y capacidad de generalización mediante métricas estadísticas como el RMSE, ARIMA y Random forest.
- Validar el modelo final con datos recientes y generar visualizaciones que muestren su desempeño y posibles aplicaciones para decisiones económicas.

7. Desarrollo e implementación del aprendizaje

7.1 Preparación y análisis de los datos

Para iniciar con un análisis preliminar de los datos, se han generado dos gráficos básicos: la distribución de la variable de salida y una matriz de correlación de las categorías. A continuación, se presentan los gráficos obtenidos en Google Colab. La matriz de correlación indica que ni la ciudad ni la categoría tienen una relación lineal fuerte con el IPC mensual, lo que sugiere la necesidad de explorar modelos no lineales o técnicas de predicción más avanzadas. Sin embargo, la variable 'year' muestra una leve correlación positiva, sugiriendo una tendencia ascendente del IPC a lo largo del tiempo. El IPC se compone de varias categorías que reflejan los hábitos de consumo de los hogares, tales como alimentos y bebidas no alcohólicas, vestido y calzado, vivienda, salud, transporte, comunicaciones, recreación y cultura, educación, restaurantes y hoteles, entre otros. Estas categorías se ponderan según su importancia en el gasto total de los hogares y se actualizan periódicamente para reflejar cambios en los patrones de consumo.

ilustración 1. distribución de la variable IPC

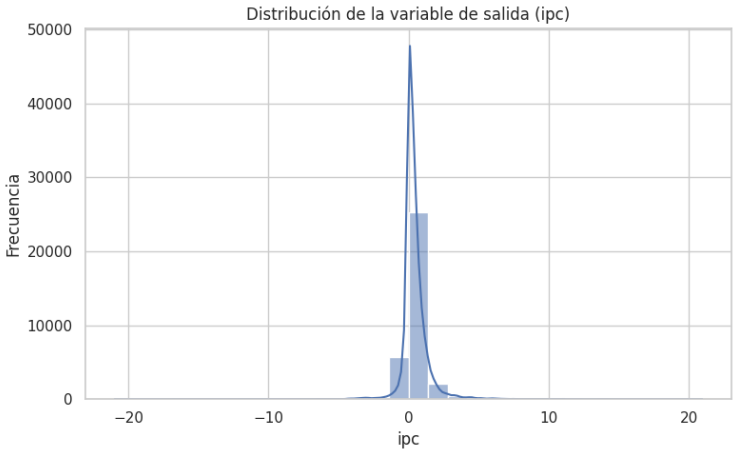
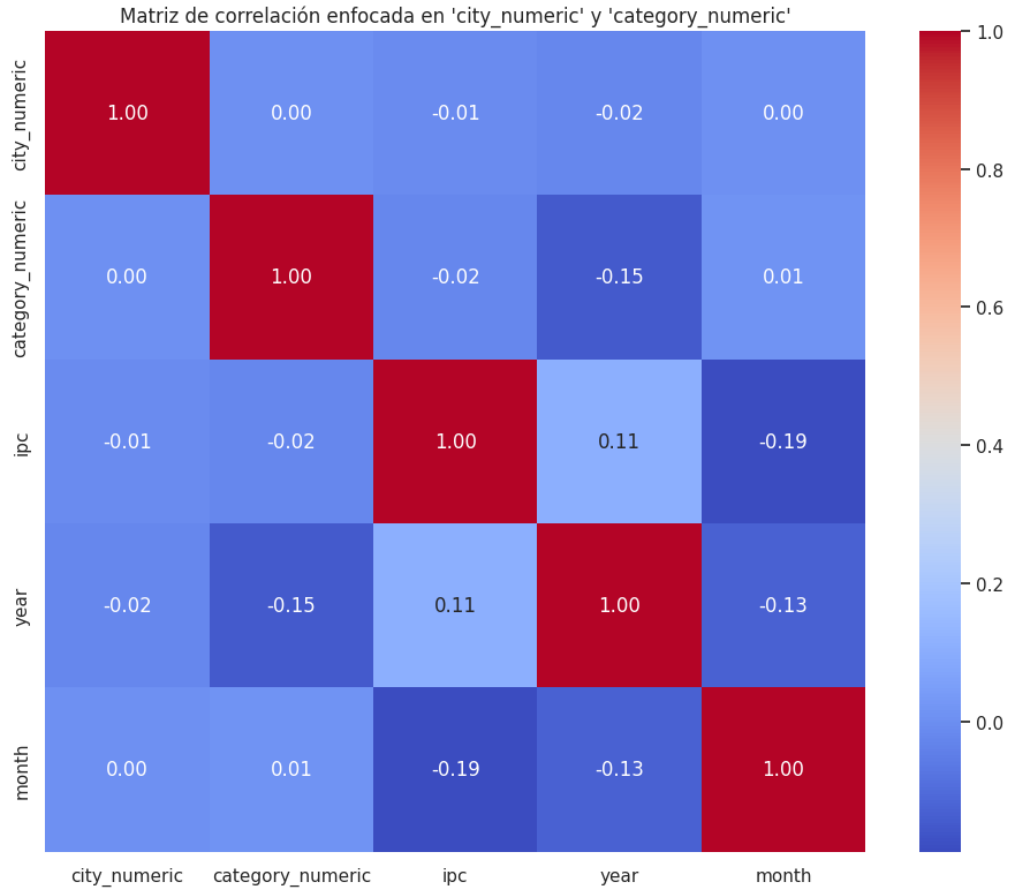
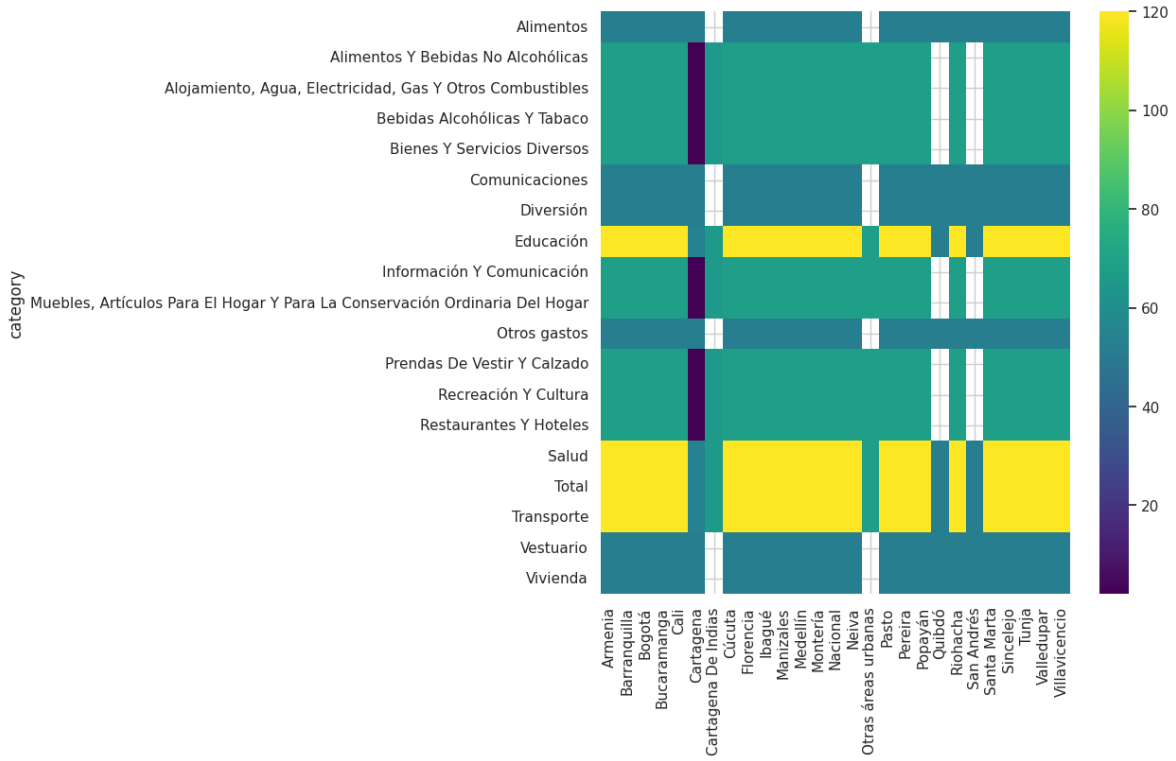


Ilustración 2.Correlación categorías versus IPC



En la siguiente grafica de Comparación de consumos por Categoría en Diferentes Ciudades se explicará la influencia de las diversas categorías versus las ciudades, en el mapa de color compara diversas categorías de gastos (como "Alimentos", "Educación" y "Salud") en diferentes ubicaciones (como "Armenia", "Barranquilla" y "Bogotá"). Los colores varían de púrpura oscuro a amarillo, representando valores que oscilan entre 0 y 120. Este tipo de visualización permite identificar rápidamente las áreas con mayores y menores gastos en cada categoría y ubicación, facilitando el análisis de patrones de consumo y posibles disparidades regionales en el costo de vida.

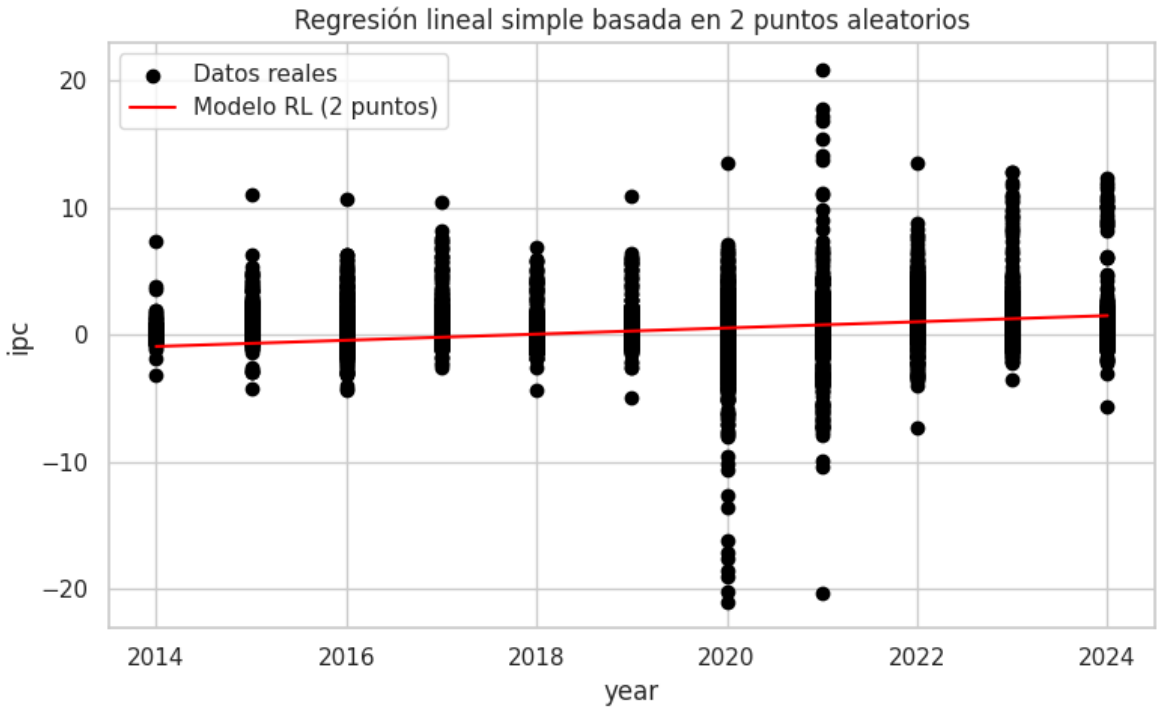
Ilustración 3. Comparación de consumos por Categoría en Diferentes Ciudades



7.2 Toma de Decisiones

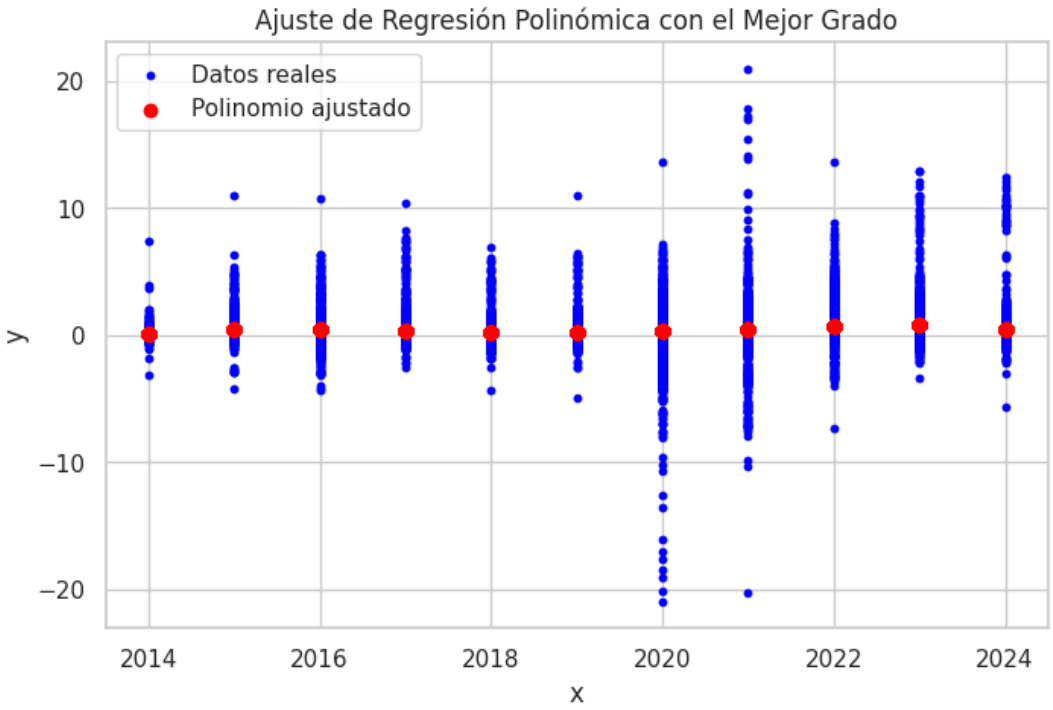
Como punto de partida para modelos de toma de decisiones se inicia con regresión lineal, basada en dos puntos aleatorios, sin embargo, aunque la línea roja parece capturar una tendencia general creciente, no se ajusta bien a las desviaciones de los valores reales, lo que reafirma que una regresión puede parecer confiable, si solo se analiza de forma visual, pero se puede sesgar porque los datos no se ajustan a este modelo.

Ilustración 4. Regresión lineal, basada en dos puntos aleatorios



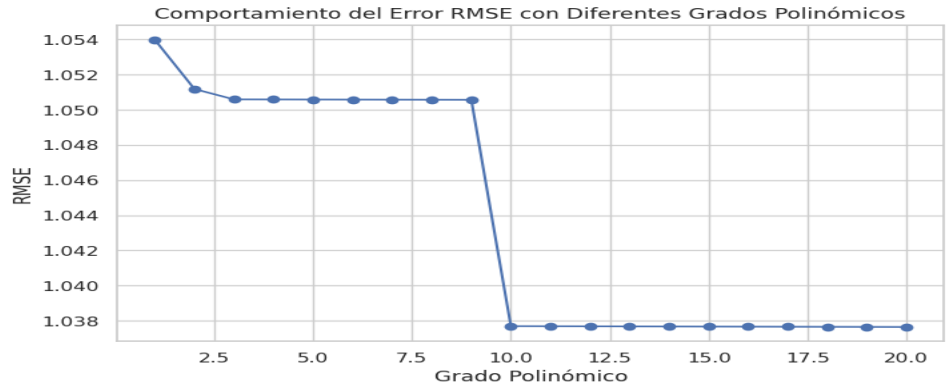
Como segundo punto de partida se expondrá el siguiente modelo, aunque se buscó el mejor grado polinómico para ajustar el IPC mensual, el modelo seleccionó una línea constante, lo que indica que no encontró una tendencia significativa. Esto evidencia que los cambios mensuales en el IPC pueden ser demasiado irregulares para ser capturados por un polinomio simple y empezamos a declarar la necesidad de trabajar sobre modelos de predicción.

Ilustración 5. Ajuste de Regresión Polinómica con el Mejor Grado



Al observar el comportamiento con diferentes grados polinómicos, se observa que el error RMSE se mantiene casi constante con grados bajos, pero a partir del grado 10 se observa una mejora notoria. Esto justifica el uso de un polinomio de grado medio-alto, ya que mejora el ajuste de los datos.

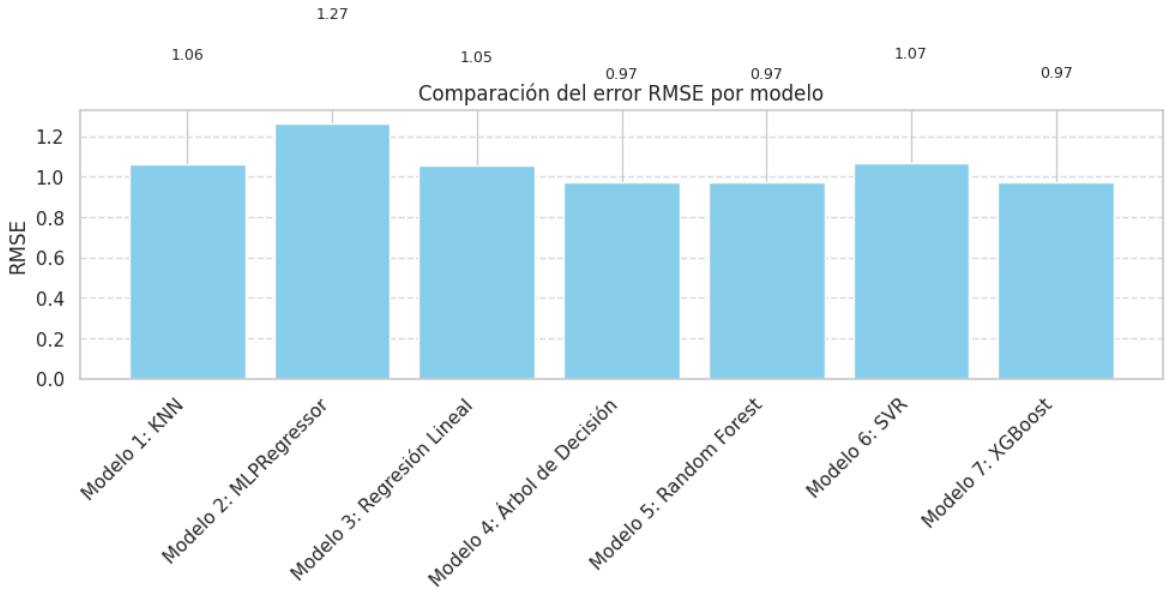
Ilustración 6. Comportamiento del Error RMSE con Diferentes Grados



7.3 Análisis de desempeño

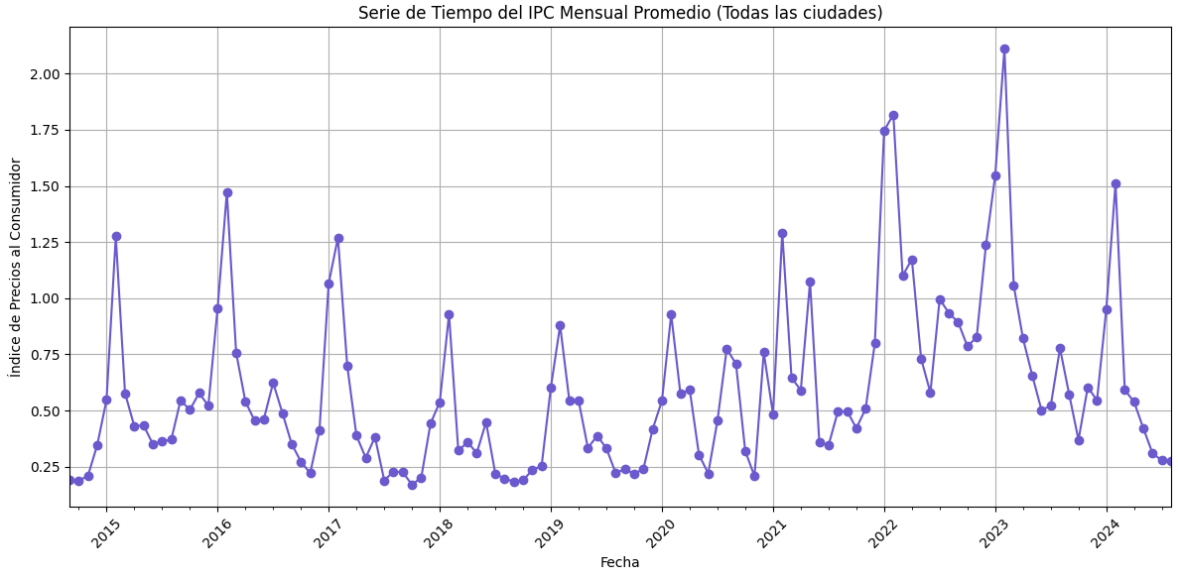
a continuación, se compararan 7 modelos usando machine learning, donde encontramos los valores de rmse, de la grafica se aprecia que los los modelos como el árbol de decisión, random forest y xgboost— ofrecen el mejor rendimiento predictivo para esta serie, alcanzando un rmse mínimo de 0.97, de igual forma y de manera contraria los modelos como mlpregressor y svr presentaron mayor error, lo cual puede deberse a parámetros no óptimos, de manera general se asevera que los modelos como Random forest o xgboost detectan fácilmente patrones, las interacciones entre variables y logran predicciones más precisas, lo que refuerza el uso de machine learning en el análisis de datos confiables.

Ilustración 7. Comparación por modelos



A continuación, se presenta la gráfica de serie tiempo en la que se observa fluctuaciones en el Índice de Precios al Consumidor (IPC) a lo largo del tiempo, con picos notables alrededor de principios de 2016, mediados de 2017, finales de 2021 y principios de 2023.

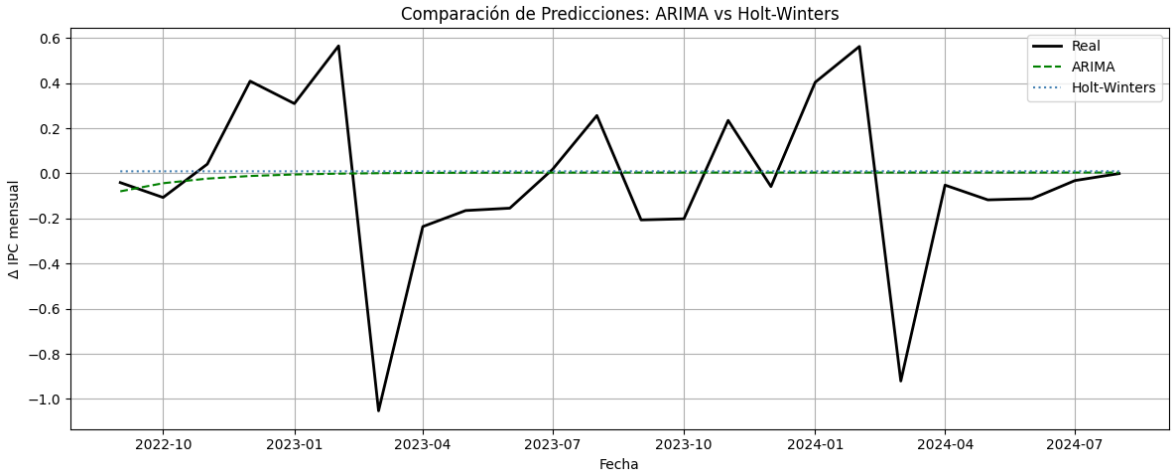
Ilustración 8. Serie de tiempo



7.4 Modelo de predicción

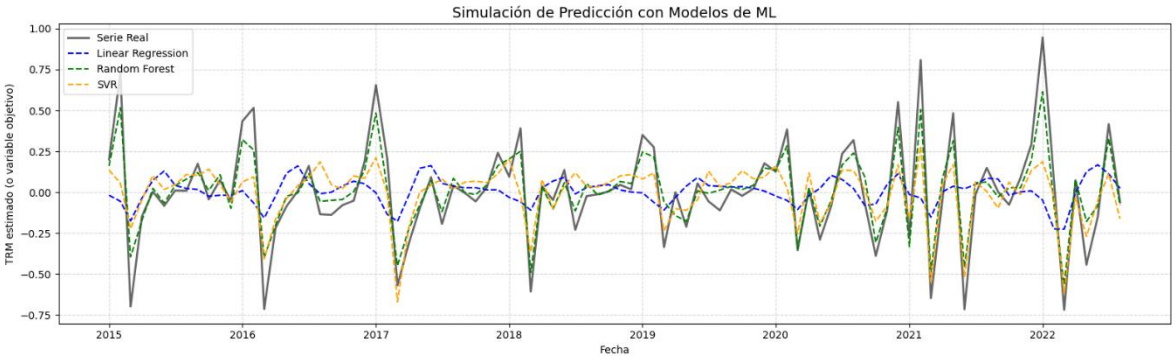
A continuación, compararemos dos modelos ARIMA y Holt-Winter, donde encontramos que logran una aproximación general, pero no muestran una precisión de la variabilidad de la serie real. La línea negra (IPC real) muestra oscilaciones marcadas, mientras que las predicciones (líneas verdes punteada y azul punteado) tienden a suavizar los cambios y a mantenerse cerca de cero. Esto indica que ninguno de los dos modelos se adapta a los patrones encontrados.

Ilustración 9.Comparación de predicciones



En a la siguiente gráfica mostraremos distintos modelos de *machine learning* predicen; Random Forest (verde) es el que más se aproxima a la forma general de la serie real: sigue los picos y caídas con mayor precisión. SVR (naranja) se mantiene relativamente estable, y responde menos a cambios bruscos, lo que podría indicar sub-ajuste o excesiva suavización; Regresión Lineal (azul) muestra una respuesta intermedia: capta cierta tendencia, pero no alcanza a modelar bien las fluctuaciones.

Ilustración 10. Simulación de Predicción con Modelos de ML



8. Conclusiones y trabajos posteriores

Los modelos de machine learning, especialmente Random Forest y XGBoost, fueron los más efectivos para predecir el IPC en Colombia. Estos modelos detectan patrones complejos y manejan bien las interacciones entre variables.

No encontramos una relación lineal fuerte entre el IPC y las categorías de gasto o ciudades. Esto sugiere que los modelos no lineales son más adecuados para este tipo de predicciones.

La variable 'year' muestra una correlación positiva con el IPC, indicando que el IPC tiende a aumentar con el tiempo. Esto es importante para la planificación económica y la toma de decisiones.

La calidad y limpieza de los datos son fundamentales para el éxito de los modelos predictivos. Datos bien depurados y estructurados permiten obtener resultados más precisos y confiables.

Futuras Investigaciones: Se recomienda explorar otros algoritmos de machine learning y técnicas de predicción para mejorar aún más la precisión de los modelos. También sería útil considerar factores externos que puedan influir en el IPC, como políticas económicas y eventos globales.

9. Bibliografía

- Banco de la República. (2024). Inflación en Colombia: evolución reciente y perspectivas. Informe de Inflación, m. d.-i. (s.f.).Caro, Á. &.–4. (s.f.).
- https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc, D. A.–D. (s.f.).
- Córdoba Rosas, A. (2024). *IPC from September 2022 to August 2024 by DANE* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/alejandrocrdobaros/ipc-from-september-2022-to-august-2024-by-dane
- Departamento Administrativo Nacional de Estadística – DANE. (s.f.). *Índice de Precios al Consumidor (IPC)*. https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc (https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc)