# *La relación entre IPC (Índice de Precios y Cotizaciones) con variables macroeconómicas: Actividad Económica, Inflación, Oferta Monetaria M2 y Tasa de Interés aplicada a la eficiencia y predictibilidad del mercado financiero mexicano.*

Un contraste analítico entre la Econometría y la Ciencia de Datos

Por: Gabriel González Méndez

*"La aversión a la pérdida es más poderosa que la atracción por la ganancia."*

- *Daniel Kahneman*

# Contenido


Una breve introducción 3

Marco teórico y revisión de la literatura 6

Metodología 15

Desarrollo 18

Conclusiones 41

Bibliografía 45

Anexos 45

# Una breve introducción

## I. Contexto y justificación del tema:

El uso de modelos econométricos y la aplicación de técnicas estadísticas ha ayudado a la sociedad mexicana al diseño de políticas económicas, al tiempo que dichos modelos y métodos dotan a los tomadores de decisiones de una perspectiva del comportamiento de variables de interés a futuro.

Actualmente, con el auge de la Ciencia de Datos, el uso de algoritmos para la resolución de problemas (en este caso problemas sociales aplicando métodos cuantitativos) y la creciente demanda de lenguajes de programación orientados a potenciar el análisis económico; se plantea una integración metodológica entre modelos econométricos y modelos propios de la ciencia de datos.

En este trabajo, dicha integración, de primera instancia, tiene una naturaleza comparativa entre los modelos. Se buscan identificar fortalezas y debilidades dentro los modelos utilizados, para posteriormente encontrar un punto convergente adecuado al problema de investigación planteado en este documento. Una integración de este calibre no es cosa aleatoria, ya que los métodos fundamentales de la ciencia de datos descansan en los mismos métodos fundamentales usados en el análisis econométrico, esta convergencia metódica y matemática dota a los economistas de una riqueza en función a su perspectiva y arsenal de herramientas.

La célula de este documento responde a una naturaleza predictiva (a través del trabajo de series temporales), el comportamiento del Índice de Precios y Cotizaciones (IPC), tomando cuatro variables macroeconómicas en perspectiva para la propia explicación. Las variables macroeconómicas son la actividad económica (IGAE), la tasa de interés, la oferta monetaria (M2), y la inflación. Existen distintas aplicaciones al realizar una predicción robusta del comportamiento del IPC, y encontrar propiamente el peso que tienen las variables macroeconómicas en dicho comportamiento; a saber, que sirve tanto al consumidor que busca hacer inversiones con capital propio en la BMV, tanto al productor que tiene un stock de capital y desea tener rendimientos fructíferos haciendo esta última estrategia de inversión.

La esencia de esta investigación sirve tanto a los agentes económicos inversores como a los economistas que buscamos una integración técnica.

## II. Planteamiento del problema

En economía aplicada, la relación que existe entre una economía real con los comportamientos decisivos que ejercen los agentes económicos, los mercados financieros, la política económica y el comportamiento de las variables macroeconómicas representa una dinámica compleja se enmarca en los debates fundamentales de las finanzas modernas, así como en los enfoques alternativos de la teoría económica.

La base teórica para el análisis univariado del Índice de Precios y Cotizaciones (IPC) es la Hipótesis del Mercado Eficiente (HME), formalizada por Fama (1970). En su forma débil, la HME postula que los precios actuales de los activos ya incorporan toda la información contenida en los datos históricos de precios. Como consecuencia, no es posible obtener rendimientos superiores de manera consistente mediante el análisis de patrones pasados.

El corolario empírico de esta hipótesis es que los precios de los activos (o sus logaritmos) siguen una caminara aleatoria. Como explican textos de econometría de referencia como Gujarati y Porter (2010), un paseo aleatorio es una serie de tiempo no estacionaria con una raíz unitaria, cuya primera diferencia es ruido blanco. La implicación directa es que el mejor pronóstico para el precio de mañana es el precio de hoy, más una deriva aleatoria.

En contraposición a la HME, la teoría de valuación de activos argumenta que el valor intrínseco de los mercados bursátiles debe estar anclado a los fundamentos macroeconómicos de la economía real. Modelos como el Modelo de Descuento de Dividendos establecen un vínculo directo entre los precios de las acciones y dos variables clave: los flujos de efectivo futuros esperados (dividendos) y la tasa de descuento utilizada para traerlos a valor presente (Bodie, Kane, & Marcus, 2018).

Este enfoque justifica el uso de modelos multivariados (como el Modelo Autorregresivo de Rezagos Distribuidos) para buscar relaciones de cointegración, un concepto desarrollado por Engle y Granger (1987), que sugiere que, aunque las series puedan ser paseos aleatorios individualmente, están ligadas por un equilibrio a largo plazo dictado por la teoría económica.

Por otro lado, la dinámica de los mercados financieros es inherentemente compleja, caracterizada por cambios de régimen, efectos de retroalimentación y comportamiento adaptativo de los agentes. Esto sugiere que las relaciones entre las variables macroeconómicas y el IPC pueden ser no lineales (McMillan, 2001).

Esta posible no linealidad justifica la aplicación de modelos de Machine Learning. Algoritmos como XGBoost o redes neuronales LSTM son estimadores universales de funciones, capaces de capturar patrones complejos y no lineales sin que el investigador deba especificarlos a priori. Su superioridad predictiva sobre los modelos lineales, obtenida en numerosos estudios empíricos (Henrique, Sobreiro, & Kimura, 2019), puede ser interpretada como evidencia de la existencia de estas dinámicas no lineales en los datos.

## III. Objetivos e hipótesis

a) **Objetivo general:** Analizar y comparar la capacidad de los modelos econométricos (ARIMA y Modelo Autorregresivo de Rezagos Distribuidos), de Machine Learning (XGBoost y Random Forest) y un modelo de redes perteneciente al aprendizaje profundo LSTM (Long Short-Term Memory) para explicar y pronosticar el comportamiento del Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores, utilizando un conjunto de variables macroeconómicas clave.

b) **Objetivos específicos:**

- Analizar las propiedades de las series de tiempo, determinando el orden de integración del IPC y de las variables macroeconómicas (IGAE, M2, inflación, tasa de interés) mediante pruebas de raíz unitaria de Dickey-Fuller Aumentada.
- Construir un modelo ARIMA univariado para el IPC como un modelo de referencia que represente la Hipótesis del Mercado Eficiente en su forma débil.
- Implementar un modelo autorregresivo de rezagos distribuidos multivariado para investigar la existencia de una relación de cointegración de largo plazo entre el IPC y al menos una de las variables macroeconómicas, así como estudiar una relación que corresponda al corto plazo.
- Entrenar y evaluar modelos de Machine Learning (XGBoost y Random Forest) y el modelo de Aprendizaje Profundo (LSTM) para pronosticar el IPC, capturando posibles relaciones no lineales y dinámicas temporales complejas.
- Comparar el desempeño predictivo de todos los modelos a través de métricas de error (como RMSE y MAE), para concluir qué enfoque metodológico ofrece el pronóstico más preciso para el mercado bursátil mexicano en el horizonte de estudio.
- Hacer una comparativa entre modelos de análisis y el comportamiento real del IPC.

c) **Hipótesis:**

- La variable oferta monetaria es el principal componente en cuanto al comportamiento del IPC.
- Los resultados que se obtienen cuando se usan métodos econométricos, contrastados con la metodología de la ciencia de datos, puede divergir por la misma naturaleza metodológica que le es propia a las dos áreas. Y en este sentido, el planteamiento resulta ser un periplo metodológico, ya que en primer lugar se planteó la relevancia de la oferta monetaria $M_2$, bajo un principio de causalidad y de explicación con respecto al comportamiento del IPC.
- Y por el otro lado, a la Ciencia de Datos le compete una calidad únicamente predictiva, y es por tal motivo que se contrastan los resultados ente estas dos grandes perspectivas.
- Teniendo el punto anterior en cuenta, se enuncia que el modelo ARIMA es el modelo más conveniente para predecir el comportamiento del IPC con respecto a otros modelos.

## IV. Alcances y limitaciones

1. Selección de Variables: El análisis no considera otras variables que podrían influir en el IPC, como factores políticos, flujos de capital extranjero o el desempeño de mercados internacionales, lo que podría generar un sesgo de variable omitida.
2. Interpretabilidad de modelos: Mientras que los modelos econométricos ofrecen coeficientes directamente interpretables en términos económicos, los modelos de Machine Learning y de Aprendizaje Profundo actúan como una especie de cajas negras, dificultando la extracción de implicaciones de política económica directas a partir de sus parámetros internos.
3. Frecuencia de Datos: El uso de datos mensuales suaviza la volatilidad diaria y semanal, por lo que las conclusiones no aplican a estrategias de inversión de alta frecuencia.

# Marco teórico y revisión de la literatura

## I. Conceptos clave y definiciones

### Sobre el Índice de Precios y Cotizaciones

El IPC es el principal indicador bursátil que se tiene dentro la Bolsa Mexicana de valores, este mide la variación los precios de las acciones con mayor liquidez y mejor capitalización dentro del mercado mexicano. La utilización de este indicador en este trabajo es importante y con gran relación a los objetivos de estudio, ya que, en general, si el IPC sube, en promedio las acciones que lo componen han aumentado de valor; si baja, el mercado ha perdido valor. Es por tal motivo que este documento es de interés para los agentes económicos que toman decisiones de inversión.

En términos de análisis, en el IPC se pueden identificar los ciclos, tendencias y volatilidad que le corresponden, a raíz de estos elementos, es que los economistas podemos estimar modelos predictivos con ayuda de modelos econométricos, y como principal interés a esta investigación, ayudar al análisis de datos en series temporales con modelos propios de la Ciencia de Datos.

Además, de que de acuerdo con el trabajo del profesor Hernández Mota, J. L. (2015), el sistema no es solamente una variable explicativa del desempeño económico, sino que es un factor productivo fundamental que puede tener efectos permanentes sobre la tasa de crecimiento de una economía.

## Modelo ARIMA

Este modelo autorregresivo de medias móviles es angular en el desarrollo de este documento, porque, además de formar parte de una de las hipótesis planteadas, este es uno de los modelos más utilizados y enseñados para el trabajo con este tipo de datos.

Gujarati, D. N., & Porter, D. C. (2010) especifican a este modelo como ARIMA (p, d, q), componiéndose este de tres componentes:

- Su componente autorregresivo AR(p) que asume que la serie temporal depende linealmente de sus componentes pasados, en este caso específico, que el IPC depende linealmente de sus valores en el pasado. Este componente en cuestión tiene la estructura:

$$Y_t = c + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + u_t;$$

  Dónde:

  - $Y_{t-p}$: Son los valores que adopta la serie en periodos anteriores.
  - $\Phi_p$: Representa los coeficientes que miden el impacto de cada rezago en el valor actual de la serie temporal.
  - $c$: Es la constante del modelo.
  - $u_t$: Es el término de error estocástico.
- Componente integrado I(d): Que es el componente que se encarga de hacer que los datos sean estacionarios, y evitar que el análisis sea nulo, por motivos de correr una regresión espuria. El parámetro descrito como I(d) indica el número de veces que la serie se diferenció, hasta alcanzar la estacionariedad.
- Componente de media móvil MA(q): Que se refiere a que el valor que se obtiene en la serie actual depende de los errores pronosticados en los periodos anteriores, o sea, el parámetro descrito como MA(q) indica cuantos errores rezagados se incluyen en el modelo. En otras palabras, este componente del modelo captura los efectos aleatorios en el corto plazo. Su estructura corresponde a la siguiente:

$$Y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q};$$

  Dónde:

  - $\mu$: Es la media de la serie temporal.
  - $u_t$: Es el error estocástico.
  - $\theta_q$: Son los coeficientes que miden de los errores pasados en el modelo.

## Modelo Autorregresivo de Rezagos Distribuidos

La justificación del uso de este modelo en el presente análisis, y teniendo en perspectiva el trabajo de Nkoro, E., & Uko, A. K. (2016), radica en su capacidad para estudiar las relaciones dinámicas que existen en las series de tiempo tanto en el corto plazo, como en el largo plazo. En este caso, la motivación principal es encontrar una relación que tenga una función explicativa en el comportamiento el IPC.

Siguiendo a los autores citados en el párrafo anterior, un Modelo Autorregresivo de Rezagos Distribuidos nos permite llegar a estos objetivos planteados a partir de la determinación de la cointegración, al tiempo que se analiza el o los impactos en el corto y en el largo plazo por parte de la una de las variables macroeconómicas (dentro del análisis se hace la selección de la variable más óptima) para con el IPC.

La estructura matemática del modelo tiene la siguiente forma:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \Phi_i \; Y_{t-i} + \sum_{j=0}^{q} \delta_j \, Y_{t-j} + u_t$$

Dónde:

- $Y_{t-i}$: Son los rezagos de la variable dependiente.
- $Y_{t-j}$: Son los rezagos de la variable independiente.
- $p \wedge q$ : Son los números de rezagos óptimos.
- $\Phi \wedge \delta$: Son coeficientes por estimar.
- $u_t$: Es el error estocástico.

Nkoro, E., & Uko, A. K. (2016) señalan una robustez y flexibilidad en la práctica de este modelo, incluso en muestras pequeñas, estos motivos refuerzan la necesidad de su uso en este documento.

## El análisis de datos usando algoritmos de Machine Learning

- ### Bosque aleatorio[1] (Random Forest)

El uso es pertinente, ya que VanderPlas, J. (2016) le da un uso dual en el sentido de que tiene aplicaciones tanto para realizar trabajo de clasificación de datos, como de regresión; característica que se ajusta perfectamente a los objetivos del trabajo.

Este modelo puede ser entendido como el perfeccionamiento e integración del análisis con un solo árbol de decisión, haciéndolo más potente y preciso. Siguiendo este planteamiento, lo que hace el modelo Random Forest es aplicar múltiples modelos simples, entrenándolos de manera diferente a cada uno de ellos.

[1] La base estructural del Bosque aleatorio es un árbol de decisión, mismo que consiste en una secuencia de reglas binarias o nodos que dividen el conjunto de datos en ramas.

En este contexto, y al hacer un análisis de regresión con respecto a los datos, el modelo hace un promedio de las predicciones, esto en función a la profundidad[2] propia del mismo bosque aleatorio. Así mismo, también se promedian los errores, esto deviene en una buena práctica en términos de sobreajuste del modelo, y en teoría, mejora la calidad de los datos trabajados.

El siguiente diagrama ilustra la estructura del modelo de bosque aleatorio:

*Esquema 1. Estructura visual del modelo Random Forest*

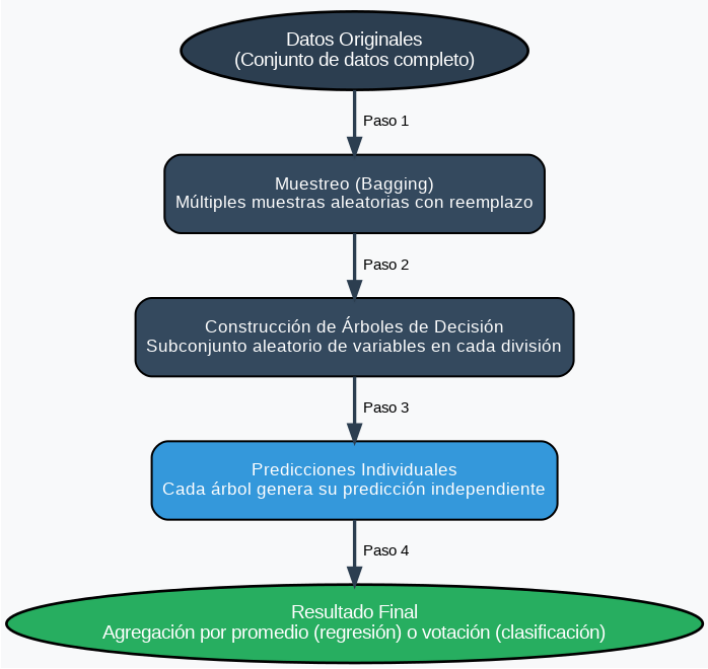



Diagrama de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

- **Extreme Gradient Boosting**

De acuerdo con la documentación oficial de la librería *Scikit-learn*, Scikit-learn developers. (2025), este modelo es una forma optimizada y escalable de la potenciación del gradiente tradicional[3], sirviendo tanto como para clasificación, como para regresión (como es el caso de esta investigación. La optimización mencionada, consiste en existe una regularización que evita el sobreajuste del modelo, está computacionalmente optimizado, este modelo tiene funcionalidad con valores faltantes y tiene validación en cada iteración para que se use el número de árboles óptimo. Este modelo se usa mucho en proyectos propios de la ciencia de datos para la predicción de precios según ciertas características, por tal motivo, es que en este documento se ejecuta dicho algoritmo.

[2] La profundidad del modelo Random Forest se especifica en los hiperparámetros dentro del código.
[3] El concepto detrás de esto es que el modelo usa modelos más débiles para construir un modelo más fuerte, esto quiere decir que es un modelo de ensamble.

La diferencia con el modelo de Random Forest es que en lugar de que se construyan muchos modelos independientes, los modelos se construyen de manera secuencial y aditiva.

Siguiendo mi propia práctica en proyectos de machine learning, mismos que abarcan la implementación de este modelo, y contrastándolo con la documentación oficial, se identifican las siguientes características precedentes:

- El modelo es sensible al ruido y a valores atípicos (característica que se toma en cuenta al momento de predecir el comportamiento del IPC).
- Tiene la capacidad de manejar datos heterogéneos.
- Es flexible en términos de los hiperparámetros, propios a su programación y personalización analítica.

La formulación matemática deriva de la idea central, que corresponde a construir un modelo predictivo de forma aditiva, de modo que cada árbol haga el intento de corregir los errores que tuvieron árboles anteriores:

$$\Rightarrow \hat{Y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \boldsymbol{F} \qquad (1)$$

Donde:

- $\boldsymbol{F}$ es el espacio donde habitan todos los posibles árboles regresores.
- $f_k$ representa la función del k-ésimo árbol de regresión.

Como se puede observar en el planteamiento del modelo, $\exists$ una predicción $\hat{Y}_i$ $\forall$ observación $x_i$.

La naturaleza de este modelo es minimizar el error, y para eso se debe de aplicar una minimización de la función objetivo[4].

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\, f_k \qquad (2)$$

Se tiene que la función de pérdida: $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$, que mide que mide qué tan lejos están las predicciones de los valores reales del modelo (se usa el error cuadrático medio).

Y que la regularización: $\sum_{k=1}^{K} \Omega\, f_k$, que penaliza la complejidad de los árboles[5]. La función de complejidad de un árbol individual es: $\Omega(f)$.

[4] La función objetivo en el modelo XGBoost representa el error de predicción y la complejidad del modelo, lo que es un concepto clave para no caer en sobreajuste.

[5] La complejidad de los árboles se refiere a qué tan grande, ramificado y específico es un árbol regresor, si se tiene una alta complejidad, se tiene riesgo de sobreajuste. Siendo más específicos, cuando se trabajan con árboles simples, la regla creada es muy general y aplicable para los otros árboles; cuando se tiene un árbol complejo, esa regla no aplica para todos los árboles, se tiene que ser específico y memorizar cada caso.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (3)$$

Dónde:

- T: Representa el número de hojas del árbol regresor.
- $w_j^2$: Es la puntuación de cada hoja.
- $\gamma \wedge \lambda$: Son parámetros que dictaminan la gravedad referente a la penalización.

Tomando en cuenta que el modelo tiene un entrenamiento iterativo, en cada iteración $t$ se añade un nuevo árbol llamado $f_t$, esto para corregir los errores del modelo anterior.

$$Def:$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (4)$$

Para obtener la función la cual se quiere minimizar en $t$, se tiene que sustituir (4) en (2):

$$\Rightarrow Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) + \Omega f_t \quad (5)$$

En términos de optimización de la función (5), el modelo utiliza dentro de su algoritmo una aproximación de Taylor[6] de segundo orden de la función de pérdida (para que cualquier algoritmo funciones con cualquier función de perdida, siempre que esta sea diferenciable).

En este sentido, se necesitan dos elementos:

- El gradiente:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \qquad (6)$$

- El hessiano:

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{(\partial \hat{y}_i^{(t-1)})^2} \qquad (7)$$

Teniendo ahora definida la función objetivo, se aplica el método descrito, y tenemos que:

Sea $F(x+a) \approx F(x) + F'(x)a + \frac{1}{2} F''(a)^2$ la identidad de la fórmula general de una expansión de Taylor de segundo orden para $F(x+a)$ alrededor de P(x) y que nos servimos para esta aplicación de optimización del modelo, tenemos que:

[6] En este caso, se usa una aproximación, ya que se corta el método después del término de la segunda derivada, o sea el Hessiano.

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx l\left(y_i, \hat{y}_i^{(t-1)}\right) + \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \; f_t(x_i) + \frac{1}{2}\frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{(\partial \hat{y}_i^{(t-1)})^2} f_t(x_i)^2$$

Donde:

- El punto P(x) en el cual se llevará la expansión corresponde a $\hat{y}_i^{(t-1)}$
- $a$, corresponde al nuevo árbol que se añade $f_t(x_i)$

Sustituyendo las definiciones del gradiente y del Hessiano, se obtiene:

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2$$

$$\therefore Obj^{(t)} \approx \sum_{i=1}^{n} [l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega f_t$$

Dónde en cada paso representado por $t$, el término $y_i, \hat{y}_i^{(t-1)}$ tiene valor de constante, ya que representa un paso anterior del modelo que se mantiene fijo.

Entonces tenemos una nueva función a optimizar más sencilla, y que depende del gradiente, del Hessiano y de las predicciones del nuevo árbol llamado $f_t$:

$$Obj^{(t)} = \sum_{i=1}^{n} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega f_t \qquad (8)^7$$

¿Cómo se encuentra la mejor puntuación para cada hoja?

Para eso, se tiene que recordar que la puntuación está dada por $w_j \; \wedge \; f_t(x_i) = w_j \; \Leftrightarrow \; x_i \in$ a la hoja "j" ($I_j$). Entonces la puntuación óptima que minimiza al objetivo es:

$$w^* = \frac{\sum_i^n \in \; I_j g_i}{\sum_i^n \in \; I_j h_i + \lambda} \qquad (9)$$

Teniendo el valor de la función objetivo, usando las puntuaciones óptimas, sustituimos la ecuación (9) en (8) para obtener la función de puntuación de T hojas:

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j}^n g_i\right)^2}{\sum_{i \in I_j}^n h_i + \lambda} + \gamma T \qquad (10)^8$$

[7] La ecuación (8) se usa para medir la calidad del árbol usado en el análisis

[8] Se tiene que un valor de Obj* más bajo tiene el significado de ser un mejor árbol (se debe minimizar la función).

Ahora, la ganancia se calcula:

$$G = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}^{n} g_i\right)^2}{\sum_{i\in I_L}^{n} h_i + \lambda} + \frac{\left(\sum_{i\in I_L}^{n} g_i\right)^2}{\sum_{i\in I_R}^{n} h_i + \lambda} - \frac{\left(\sum_{i\in I}^{n} g_i\right)^2}{\sum_{i\in I}^{n} h_i + \lambda}\right] - \gamma \quad (11)$$

Dónde:

- $I_L \wedge I_R$: Son las observaciones capturadas en las hojas izquierda y derecha después de la división.
- $I$: Son todas las observaciones antes de la división.

El modelo elige de entre todas las opciones, aquella que maximiza la función de ganancia.

## Redes Neuronales[9] un esbozo del Aprendizaje profundo

- **El modelo LSTM (Long Short-Term Memory)**

He planteado el siguiente diagrama, basándome en el trabajo de Hochreiter, S., & Schmidhuber, J. (1997), así como en Gers, F. A., Schmidhuber, J., & Cummins, F. (2000) para describir la arquitectura de una celda propia de una red neuronal[10], así como la actualización del modelo:

*Esquema 2. Diagrama de una celda del LSTM y de la estructura de una red neuronal en general*

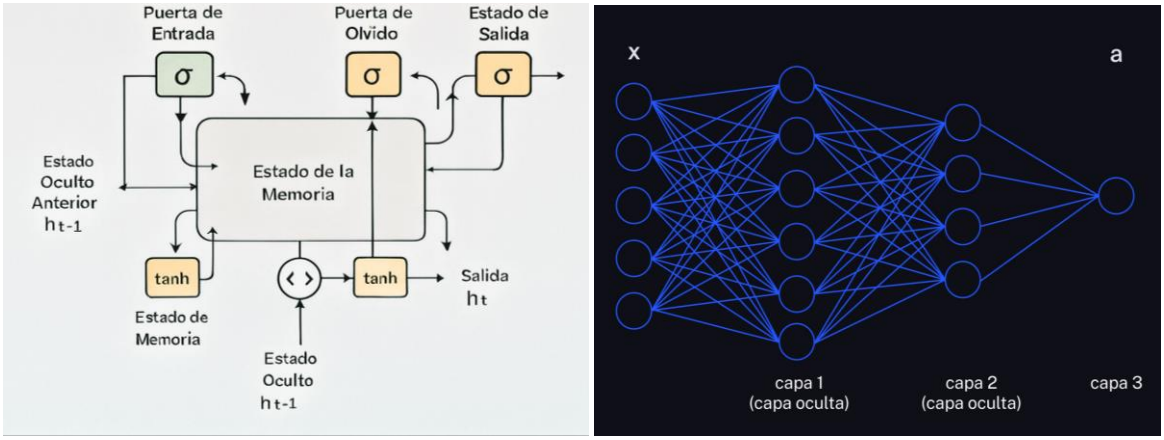



Diagrama de una celda LSTM de elaboración propia, el diagrama de la estructura de una red neuronal fue recuperado de la red.

[9] Página recomendada para visualizar la estructura, tasa de aprendizaje, función de activación, tasa de regularización y tipo de problema a resolver de una red neuronal: https://playground.tensorflow.org/#activation=sigmoid&batchSize=10&dataset=xor&regDataset=reg-plane&learningRate=0.00001&regularizationRate=0&noise=0&networkShape=5&seed=0.34965&showTestData=false&discretize=false&percTrainData=50&x=true&y=true&xTimesY=false&xSquared=false&ySquared=false&cosX=false&sinX=false&cosY=false&sinY=false&collectStats=false&problem=regression&initZero=false&hideText=false

[10] Dentro de una red neuronal tenemos tres elementos fundamentales: una celda se refiere a una unidad mínima dentro de esta arquitectura, una capa es una secuencia de celdas conectadas a lo largo del tiempo y una red se forma por una o varias capas. Las redes neuronales tienen una función de activación dependiendo del modelo.

El análisis con este tipo de arquitectura radica en la potencia de este modelo. Este pertenece a un tipo de red llamado red neuronal recurrente, mismo que fue diseñado para el análisis de datos secuenciales, en este caso, las series temporales.

Siguiendo su propia utilización para la que fueron diseñadas estas redes, la idea de la aplicación de este modelo de red neuronal es aprovechar la capacidad que tienen en memoria, para que el modelo identifique y aprenda a partir de patrones dados en la serie de tiempo.

Además, este modelo tiene la capacidad de modelar relaciones no lineales, y ese análisis de linealidad lo cubre propiamente el modelo autorregresivo de rezagos distribuidos. Normalmente se usa este tipo de modelos en el sector financiero para analizar patrones no lineales, ya que, en estos mercados, el comportamiento del IPC tiene una memoria de eventos pasados, y naturalmente bajo este modelo, se pueden identificar tendencias en el largo plazo.

## II. Estado del arte del análisis de datos basado en la Econometría y modelos propios de la Ciencia de Datos

La revisión y selección de los siguientes artículos fue minuciosamente realizada, con el objetivo de justificar, sustentar y validar el proceder técnico y trabajo con los datos realizado.

- Mukherjee, T. K., & Naka, A. (1995). Dynamic relations between macroeconomic variables and the Japanese stock market: an application of a vector error correction model. *The Journal of Financial Research*, *18*(2), 223–237. Recuperado de: https://doi.org/10.1111/j.1475-6803.1995.tb00563.x

En este artículo Mukherjee, T. K., & Naka, A. (1995) encontraron un equilibrio a largo plazo entre variables macroeconómicas a través del trabajo con de investigación que corresponde a su dinámica. Para estos efectos, usan un modelo de correcciones vectorial, así como técnicas de cointegración. Lo que encuentran los autores, es fundamentalmente que, en el largo plazo las variables macroeconómicas influyen en el comportamiento del mercado de acciones.

La relevancia en este documento radica en la selección de variables macroeconómicas, elemento que pertenece a la hipótesis del trabajo, en este caso, no se usa un modelo VECM, se usa un modelo autorregresivo de rezagos distribuidos.

- Al-Jafari, M. K. (2016). The macroeconomic determinants of the Omani stock market. *International Journal of Economics and Financial Issues*, *6*(1), 221–228.

Este artículo fue revisado y anexado al documento una vez realizado el análisis bajo el modelo autorregresivo de rezagos distribuidos, en aras de sustentar la buena práctica en términos de la aplicación del modelo. A saber, que en este trabajo se encontró una mezcla de órdenes de integración, (I(1) y I(2)) respectivamente.

- Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications*, *36*(3), 5932–5941.

A lo largo de este artículo Atsalakis, G. S., & Valavanis, K. P. (2009) aplican el enfoque tradicional de modelos econométricos (un modelo ARIMA) con lo que ellos llaman *soft computing*, que dicha arquitectura del modelo pertenece al Aprendizaje Profundo. Lo que sustenta y valida las intenciones de análisis de este trabajo de investigación, ya que modelos que podrían pasar por alto elementos de no linealidad que los algoritmos propios del Machine Learning y las redes neuronales no despreciarían.

# Metodología

## I. Tipo de investigación

Siguiendo a Hernández Sampieri, Fernández Collado, & Baptista Lucio, (2014) este trabajo engloba en su totalidad de un enfoque cuantitativo, aunque tiene trascendencia de corte cualitativa, ya que este busca influir en el comportamiento de los agentes económicos. La finalidad que le corresponde es la investigación es de tipo correlacional-explicativa, de acuerdo con estos autores. Es correlacional porque busca medir el grado de asociación entre el Índice de Precios y Cotizaciones y un conjunto de variables macroeconómicas. Simultáneamente, explicativa al intentar determinar en qué medida estas variables fundamentales influyen y pueden predecir el comportamiento del mercado bursátil, yendo más allá de la simple descripción de la relación.

El diseño de la investigación es no experimental de tipo longitudinal de tendencia, puesto que se analizan los cambios y la evolución de las variables a través del tiempo (desde enero de 2010 hasta mayo de 2023) para hacer inferencias sobre su relación dinámica, sin manipular deliberadamente ninguna de las variables.

Luego se contrastan los resultados predichos con el comportamiento real del Índice de Precios y Cotizaciones a lo largo del año 2024, para guardar la calidad de predicción.

## II. Procedimientos y técnicas utilizadas

Este proceder corresponde a:

1. **Recolección y preparación de los datos.** Como herramienta fundamental de análisis se utilizó el lenguaje de programación orientada al tratamiento de datos *Python*, así como la librería *Pandas*, para la preparación de los datos.
2. **Análisis de las series temporales.** Siguiendo a Gujarati & Porter, (2010), aplicó la prueba de raíz unitaria de Dickey-Fuller Aumentada a cada serie para diagnosticar su estacionariedad y determinar su orden de integración, en aras de no caer en una regresión espuria.

3. **Modelado econométrico**:
    a. Se ajustó un modelo ARIMA (Autorregresivo Integrado de Medias Móviles) univariado sobre la serie del IPC como modelo de referencia, para evaluar la Hipótesis del Mercado Eficiente.
    b. Se usó un modelo autorregresivo de rezagos distribuidos para analizar la existencia de una relación de cointegración de largo plazo entre el IPC y las variables macroeconómicas, dada la presencia de distintos órdenes de integración en las series.
4. **Modelado usando algoritmos de Machine Learning:**
    a. Se transformó el problema de pronóstico de series de tiempo a uno de aprendizaje supervisado mediante la creación de variables rezagadas, posteriormente se implementaron dos modelos: Random Forest (que estudia las relaciones no lineales) y XGBoost (que mejora la capacidad predictiva). En cada modelo se dividieron los datos en una proporción de 80% para el entramiento del modelo y 20% en concepto de prueba.
5. **Uso de redes neuronales, análisis propio al Aprendizaje Profundo:**
    a. Se desarrolló un modelo LSTM para capturar relaciones complejas no detectables por los modelos lineales.
6. Se generaron pronósticos a 12 meses con cada uno de los modelos desarrollados y se evaluó su precisión utilizando métricas de error estándar.
7. Se comparó cada modelo con el comportamiento real del Índice de Precios y Cotizaciones.

## III. Herramientas y materiales

El análisis de datos y la implementación de los modelos se realizaron utilizando el lenguaje de programación Python (versión 3.11.9) en un entorno de Jupyter Notebook. Se emplearon las siguientes librerías especializadas:

- **Para la manipulación y análisis de datos:** *Pandas* y *Numpy*.
- **Para la visualización de datos:** *Matplotlib* y *Seaborn*.
- **Para el análisis econométrico:** *statsmodels* (para las pruebas ADF, ARIMA y ARDL) y *pmdarima* (para la selección automática de órdenes en el modelo ARIMA).
- **Para los modelos de machine learning:** *Scikit-learn* (para la preparación de datos y métricas de evaluación), *XGBoost* (para el modelo de Gradient Boosting) y *TensorFlow* con la API *Keras* (para la construcción del modelo LSTM).

## IV. Criterios de análisis

Como a lo largo de este trabajo de investigación, nos enfrentamos a un problema de análisis continuo, la comparación del desempeño de pronóstico de todos los modelos (ARIMA, ARDL, Random Forest, XGBoost y LSTM) se basó en las siguientes métricas de error, calculadas sobre un conjunto de prueba:

- **Error Cuadrático Medio (RMSE):** Que mide tanto la magnitud del promedio de los errores de predicción como los valores reales. Penaliza más fuertemente los errores grandes.

Planteamiento matemático:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}$$

En donde:

- $y_i$: Es el valor del dato observado.
- $\widehat{y_i}$: Es el valor predicho por el modelo.
- $n$: El número de observaciones dentro del modelo.

- **Error Absoluto Medio (EAM):** Proporciona la magnitud promedio de los errores, siendo una métrica más directa e intuitiva de la desviación promedio.

Planteamiento matemático:

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_1}|$$

En donde:

- $y_i$: Es el valor del dato observado.
- $\widehat{y_i}$: Es el valor predicho por el modelo.
- $n$: El número de observaciones dentro del modelo.

# Desarrollo

## I. Presentación de la información

### a) Estadística descriptiva

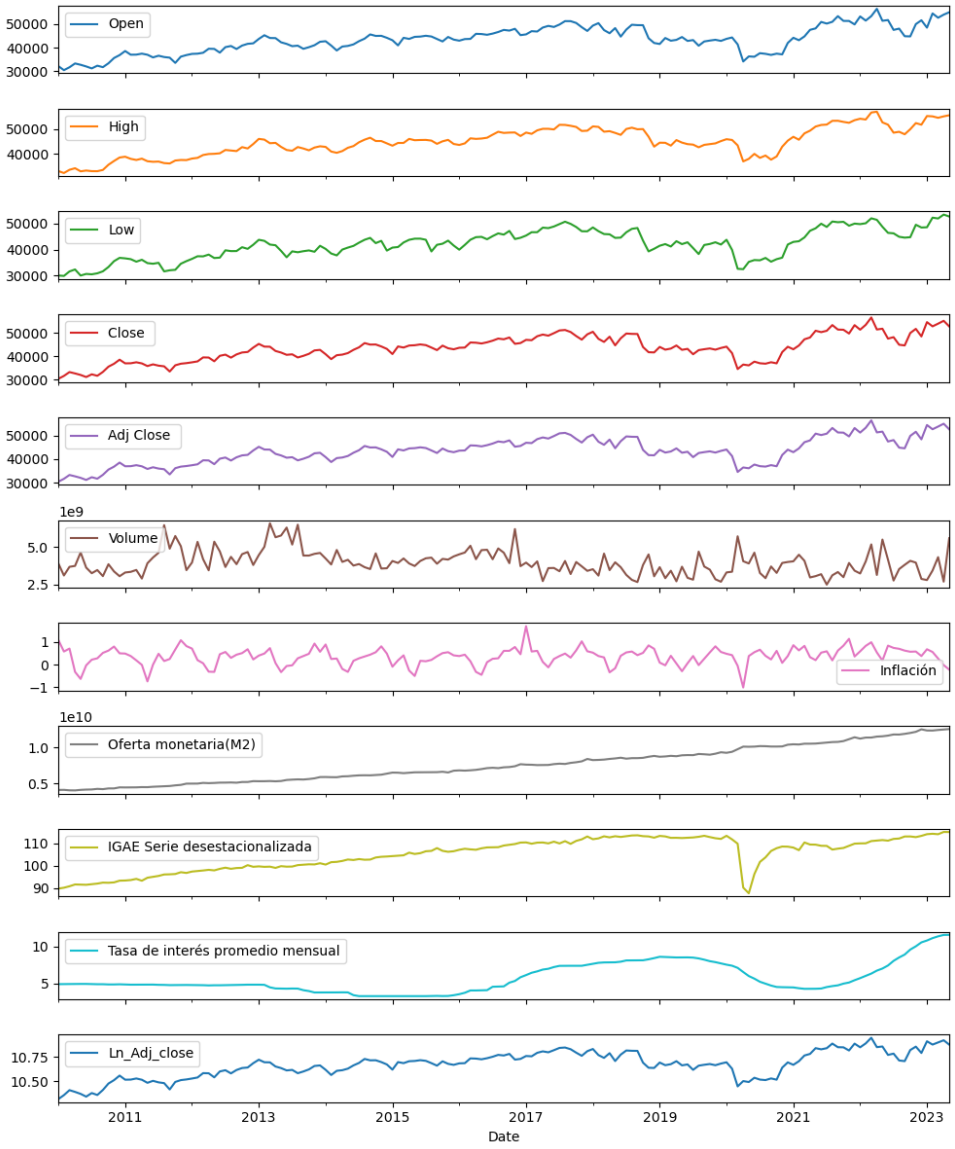
Para efectos de una buena práctica, en términos de presentación de la información relevante que conlleva este trabajo, se empieza esclareciendo que tipo de variables se tomaron en cuenta. Como ya se citó con anterioridad, la naturaleza de las variables analizadas corresponde a una serie temporal, y complementando dicha información, se puntualiza que se trabaja con un modelo semilogarítmico, ya que se trabaja con el precio de cierre ajustado inherente al IPC, entonces en la base de datos se tienen las siguientes variables:

- ln(Precio de cierre ajustado IPC) mensual.
- Tasa de interés, se tomó un promedio mensual del desempeño diario de la variable.
- Actividad económica con datos mensuales (IGAE), que mide la actividad global de la economía.
- Inflación mensual
- Oferta monetaria (M2) mensual.

Es propio citar que, estas variables se trabajaron desde un conjunto de variables más grande, incluyendo esta los precios de apertura y cierre y el volumen, correspondientes todos estos al IPC. Pero posteriormente se acota la base de datos y solo se toman en cuenta las variables descritas. No obstante, se ofrece un esbozo del comportamiento de cada variable perteneciente a la base de datos matriz y natural.

Además, el análisis consta de trece años, de enero de 2010 hasta mayo del 2023, este intervalo temporal fue para cuidar la consistencia de los datos en todas las variables, evitar valores nulos, etc.

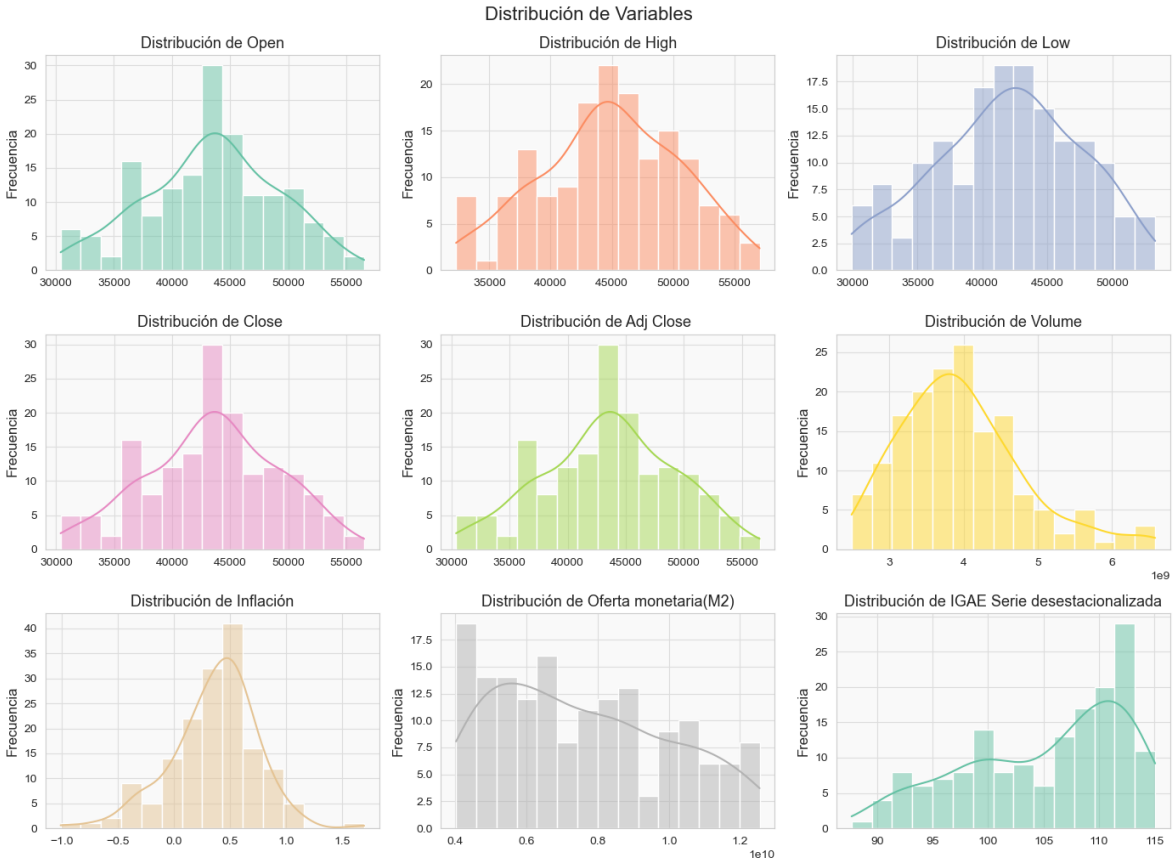
*Gráfica 1. Comportamiento de las variables a lo largo del tiempo*




Elaboración propia con datos de Yahoo Finance y Banxico

Siguiendo estos efectos, se presenta la matriz de correlación entre todas las variables, así como la distribución de cada variable inscrita en la base de datos madre:

*Gráfica 2. Conjunto de distribuciones*




Elaboración propia con datos de Yahoo Finance y Banxico

El esbozar todo el conjunto de variables es importante en este análisis, ya que, al conjuntar toda la narrativa de datos, se puede llegar a un mejor desarrollo econométrico; mejores interpretaciones, en términos de análisis económico en relación con el comportamiento de las variables; así como una mejor práctica en el planteamiento de los modelos. La distribución de las variables de precio de apertura, máximo, mínimo y precio de cierre operan con un rango estable, pero sujetos a shocks en el mercado financiero (mismos que justifican su volatilidad), o bien, factores de carácter social.

En relación con el precio de cierre ajustado, este goza de una distribución ligeramente más compacta, misma que sugiere una menor dispersión en los datos. En un sentido económico, el ajuste elimina distorsiones propias de la serie de tiempo, esto ayuda a modelar los rendimientos históricos, así como a hacer comparaciones intertemporales y evitar sesgo en los datos, esto por cambios estructurales.

El volumen tiene un fuerte sesgo a la derecha, lo que indica que la mayoría de las observaciones se capturaron en un volumen bajo, lo que indica que la liquidez del activo es estable en tiempos ordinarios, pero este comportamiento también está sujeto a shocks en el mercado.

El gráfico que modela la distribución de los datos correspondientes a la inflación revela una distribución centrada en la media, con esto, se tiene una baja dispersión en las colas.

Ahora, refiriéndonos a la oferta monetaria, la distribución que le precede está concentrada, además el gráfico de comportamiento temporal revela que esta es creciente en el tiempo. El comportamiento de la oferta monetaria es en exceso interesante y relevante, ya que como es creciente, y siguiendo la teoría económica, esta puede ejercer presiones inflacionarias.

Para la actividad económica, se tiene que, en el primer gráfico, el de comportamiento temporal, una captura precisa de la actividad económica durante la pandemia del COVID-19, durante los años de dicho fenómeno, pero posteriormente se observa una recuperación sostenida.

De la misma manera, se ofrece la tabla que comunica las estadísticas descriptivas de las variables de interés:

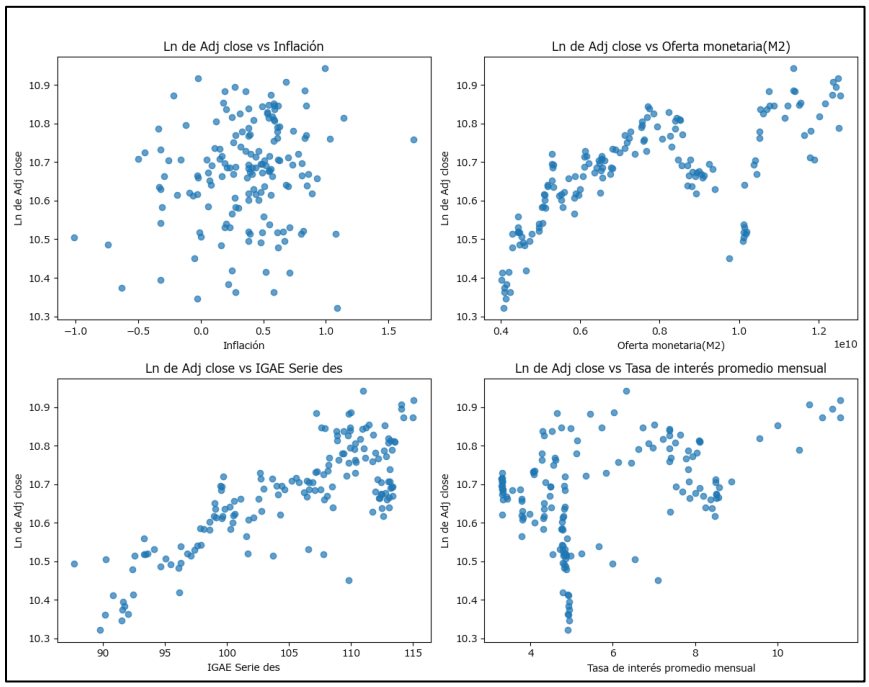
*Tabla 1. Estadística descriptiva*

| | Open | High | Low | Close | Adj Close | Volume | Inflación | Oferta monetaria | Actividad económica | Tasa de interés | Ln(Adj Close) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 161 | 161 | 161 | 161 | 161 | 161 | 161 | 161 | 161 | 161 | 161 |
| Mean | 43380.1612 | 44816.1794 | 41936.4806 | 43484.9798 | 43484.9798 | 3967615799 | 0.36111801 | 7591898004 | 105.1291497 | 5.69503106 | 10.6715429 |
| Std. | 5676.17394 | 5722.03815 | 5593.92673 | 5638.86586 | 5638.86586 | 838180628 | 0.3837333 | 2461847707 | 7.158946832 | 2.00708188 | 0.13303979 |
| Min | 30416.64 | 32283.14 | 29926.06 | 30391.61 | 30391.61 | 2487695900 | -1.01 | 4019684815 | 87.6963 | 3.29 | 10.3219219 |
| 25% | 39523.32 | 41167.3 | 38062.14 | 40185.23 | 40185.23 | 3405984800 | 0.17 | 5326721542 | 99.5842 | 4.32 | 10.6012548 |
| 50% | 43707.02 | 44714.41 | 42066.08 | 43714.93 | 43714.93 | 3899846400 | 0.39 | 7231920684 | 107.2155 | 4.85 | 10.685445 |
| 75% | 47543.35 | 48983.93 | 46143.81 | 47524.45 | 47524.45 | 4429287200 | 0.59 | 9326244070 | 111.2217 | 7.38 | 10.7689996 |
| Max | 56530.56 | 57064.16 | 53308.07 | 56536.68 | 56536.68 | 6585330200 | 1.7 | 12552103031 | 115.0737 | 11.54 | 10.9426449 |

Elaboración propia con datos de Yahoo Finance y Banxico

Teniendo en cuenta un esbozo de la relación estática entre el logaritmo del precio de ajuste y las variables macroeconómicas, se tienen los siguientes gráficos de dispersión, como se puede apreciar, esa relación no es lineal, por tal motivo, se introdujo al análisis los modelos propios de la ciencia de datos citados anteriormente.

*Gráfica 3. Relación ln (Precio de cierre Ajustado) vs variables macroeconómicas*



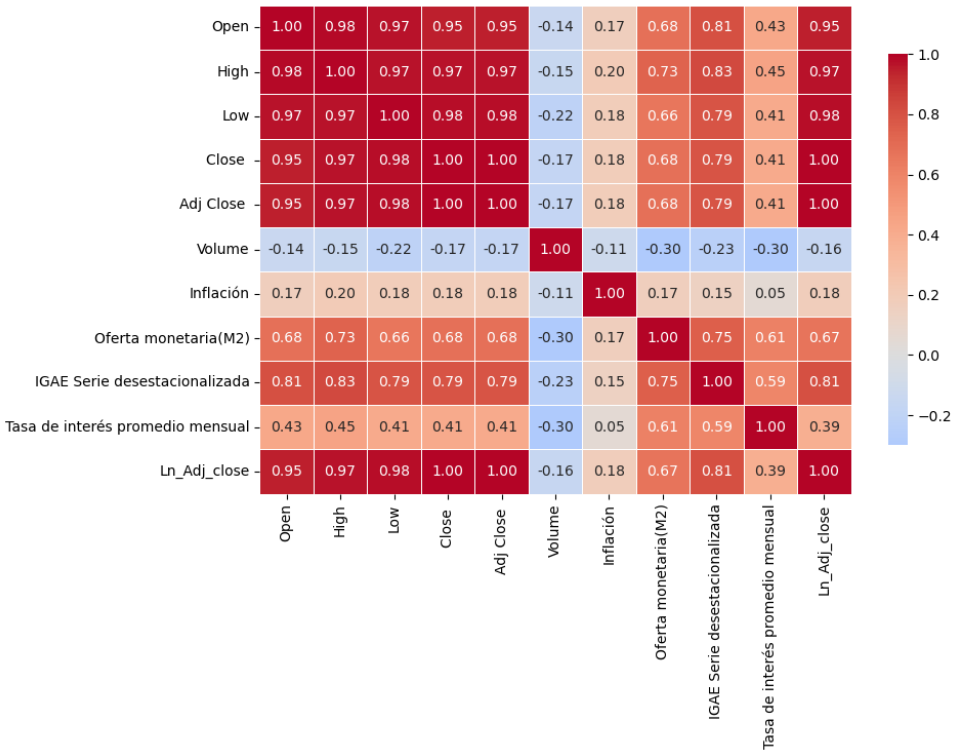

Elaboración propia con datos de Yahoo Finance y Banxico

Sin más, es importante destacar la relación que pareciera tener la oferta monetaria con el logaritmo del IPC (aunque no es lineal), esta tiene una tendencia positiva en un tramo del gráfico, y se tiene al mismo tiempo que el logaritmo del precio de cierre ajustado también crece, (una pendiente positiva). Financieramente podría implicar que a razón que la oferta monetaria sufre aumentos, existe una mayor liquidez en el sistema, por lo que, siguiendo a la teoría económica, esto estimula a la demanda de activos financieros. Además, la relación positiva sugiere que la expansión monetaria ha coincidido con un ciclo de crecimiento económico o con políticas de estímulo que favorecen los mercados. Pero estas sugerencias a bote pronto se desenmascaran en el Gráfico 5.

En cuanto a la relación entre la actividad económica y el logaritmo del precio de cierre ajustado del IPC, esta parece ser una relación lineal, mientras que, con la tasa de interés y la inflación, esta relación parece ser nula.

Para concluir la presentación de información, en esta narrativa de las variables de interés, se adjunta la matriz de correlación siguiente:

*Gráfica 4. Matriz de correlación entre todas las variables*




Elaboración propia con datos de Yahoo Finance y Banxico

Como se puede observar en la matriz, se tiene que variables de como actividad económica como el IGAE (con 0.81) y la Oferta Monetaria (con 0.67) gozan de una fuerte correlación positiva, como se observó en el gráfico anterior. Por otro lado, la correlación con la Inflación (0.18) es muy débil.

Para descartar una regresión espuria por no estacionariedad, se va a evitar correr una regresión por mínimos cuadrados ordinarios, en su lugar se evalúa la estacionariedad de las series de tiempo con la prueba Dickey-Fuller Aumentada.

### b) **Planteamiento de estacionariedad para el análisis de series de tiempo**

Siguiendo a Gujarati y Porter (2010) se tiene el siguiente planteamiento para la estacionariedad de una serie de tiempo:

$$\text{Sea } Y_t = \alpha + \rho Y_{t-1} + u_t \text{ una serie temporal;}$$

- $\alpha$: Es una constante.
- $\rho$: Un coeficiente autorregresivo.
- $u_t$: Un término de error.

Se tiene que:

- Si $\rho < 1$, entonces la serie en cuestión es estacionaria.
- Si $\rho = 1$ y que $\alpha = 0$, entonces la serie es una caminata aleatoria pura.
- Si $\rho = 1$ y $\alpha \neq 0$, entonces la serie temporal es una caminata con deriva.

c) **Planteamiento de la prueba Dickey-Fuller Aumentada**

La Prueba de Dickey-Fuller Aumentada (ADF) es un contraste estadístico diseñado para detectar la presencia de raíces unitarias en series de tiempo, esto es equivalente a verificar si una serie es no estacionaria. Siguiendo a Gujarati y Porter (2010) y Wooldridge (2013), se puede afirmar que la presencia de una raíz unitaria implica que los choques en la serie tienen efectos permanentes, y, por tanto, la media y la varianza no son constantes en el tiempo.
Partimos de un proceso autorregresivo de primer orden:

$$Y_t = \rho Y_{t-1} + u_t$$

Dónde:

- $Y_t$: Representa la variable de interés.
- $\rho$: Es el coeficiente autorregresivo.
- $u_t$: Es el término de error con $[E(u_t) = 0], [Var(u_t) = \sigma^2]$, sin autocorrelación.

Para la transformación de la prueba desarrollamos lo siguiente:

Restamos $(Y_{t-1})$ en ambos lados de la ecuación, y tenemos que:

$$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t$$

$$\Delta Y_t = (\rho - 1) Y_{t-1} + u_t$$

Y definimos que:

$$\delta = \rho - 1$$

Y así se obtiene:

$$\Delta Y_t = \delta Y_{t-1} + u_t$$

Con todo esto, tenemos las siguientes hipótesis correspondientes a la prueba:

- $H_0$: $(\delta = 0) \Leftrightarrow (\rho = 1) \Rightarrow$ Existe una raíz unitaria ⇛ La serie es no estacionaria.
- $H_1$: $(\delta < 0) \Leftrightarrow (|\rho| < 1) \Rightarrow$ La serie es estacionaria.

De acuerdo con Gujarati y Porter (2010) y con Wooldridge (2013), al trabajar con una serie de tiempo, el error estocástico puede estar correlacionado, para poder corregir este efecto, se añaden rezagos de la variable dependiente en diferencias. Tenemos los casos:

a) Sin constante ni tendencia:

$$\Delta Y_t = \gamma Y_{t-1} + \sum_{i=1}^{\rho} \alpha_i \Delta Y_{t-i} + u_t$$

b) Con constante:

$$\Delta Y_t = \alpha + \gamma Y_{t-1} + \sum_{i=1}^{\rho} \alpha_i \Delta Y_{t-i} + u_t$$

c) Con constante y tendencia:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{\rho} \alpha_i \Delta Y_{t-i} + u_t$$

Dónde:

- $\rho$ : Es el número de rezagos incluidos para eliminar la autocorrelación.
- $\beta t$ : Representa una tendencia determinista.

Para Gujarati y Porter (2010), el número de rezagos $(\rho)$ se elige para que los residuos $(u_t)$ sean ruido blanco:

$$u_t \sim N(0, \sigma^2)$$

Con respecto con el estadístico de prueba para la prueba ADF, este se basa en la prueba t, pero sus valores críticos no siguen la distribución t de Student estándar, por tal motivo, se deben usar tablas específicas para la prueba, construida por los autores de la prueba ADF. Se tienen las siguientes reglas de decisión:

- Regla de decisión $ADF_\tau$:

Si $\tau_{ADF} < \tau_\alpha \Rightarrow$ Rechazamos $H_0$ (Serie estacionaria).

Si $\tau_{ADF} \geq \tau_\alpha \Rightarrow$ No rechazamos $H_0$ (Serie no estacionaria).

- Regla de decisión usando el valor p:

Si $p \leq \alpha \Rightarrow$ Se rechaza $H_0$

Si $p > \alpha \Rightarrow$ Aceptamos $H_0$

Ahora, se procede a llevar a cabo ambos criterios de decisión, trabajando con un nivel de significancia del 5%

- Resultados con valor P:

```
Variables estacionarias:
   -  Volume (p-valor = 0.0044)

Variables NO estacionarias:
- Open (p-valor = 0.2599)
- High (p-valor = 0.3031)
- Low (p-valor = 0.2543)
- Close (p-valor = 0.1288)
- Adj Close (p-valor = 0.1288)
- Inflación (p-valor = 0.2969)
- Oferta monetaria(M2) (p-valor = 1.0000)
- IGAE Serie desestacionalizada (p-valor = 0.2536)
- Tasa de interés promedio mensual (p-valor = 0.2283)
- Ln_Adj_close (p-valor = 0.0874)
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

- Resultados con criterio $\tau$:

```
Variables estacionarias:
 - Volume: τ = -3.6827, τ_crit = -2.8800

Variables no estacionarias:
- Open: τ = -2.0624, τ_crit = -2.8798
- High: τ = -1.9628, τ_crit = -2.8799
- Low: τ = -2.0758, τ_crit = -2.8798
- Close: τ = -2.4475, τ_crit = -2.8798
- Adj Close: τ = -2.4475, τ_crit = -2.8798
- Inflación: τ = -1.9767, τ_crit = -2.8811
- Oferta monetaria(M2): τ = 3.1183, τ_crit = -2.8815
- IGAE Serie desestacionalizada: τ = -2.0776, τ_crit = -2.8800
- Tasa de interés promedio mensual: τ = -2.1412, τ_crit = -2.8801
- Ln_Adj_close: τ = -2.6278, τ_crit = -2.8798
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Se procede a hacer la diferenciación sobre todas las variables de interés con un bucle iterativo hasta que sean estacionarias.

```
--- Diferenciando series hasta alcanzar estacionariedad ---

Orden de integración (veces que se diferenció cada serie):
{'Ln_Adj_close': 1, 'Tasa de interés promedio mensual': 1, 'IGAE Serie desestacionalizada': 1, 'Oferta monetaria(M2)': 2, 'Inflación': 1}
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

## II. Análisis y discusión de resultados

Una vez aplicada la diferenciación y habiendo obtenido una nueva base de datos lista para trabajar, un dataframe que contiene las variables de interés en su forma estacionaria, se procede a la construcción de los modelos comentados en el documento. Se obtienen los siguientes resultados con respecto a los órdenes de integración:

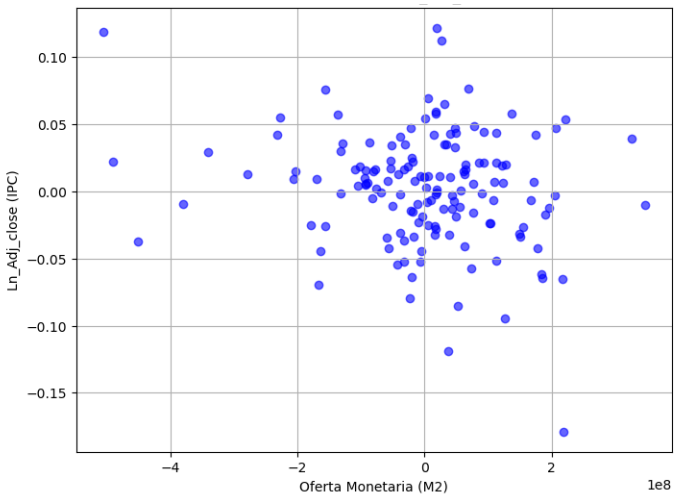
- Logaritmo del precio de cierre ajustado: Orden 1.
- Tasa de interés: Orden1.
- Actividad económica: Orden 1.
- Oferta monetaria: Orden 2.
- Inflación: Orden 1.

Se observa un cambio en términos de la relación entre la oferta monetaria y el logaritmo del precio de cierre ajustado del IPC, ya en series estacionarias, lo que refuerza la idea de la importancia de trabajar con datos estacionarios, ya que se puede interpretar el gráfico 3, en la parte de la relación entre oferta monetaria y el logaritmo del precio ajustado de cierre del IPC como una ilusión óptica que termina en una regresión espuria. En el gráfico 5 se puede apreciar una nube de puntos sin patrón establecido, que es la verdadera relación estática de las variables.

Al tener este tipo de relaciones entre variables, se justifica el uso de modelos más refinados para el análisis que compete a este documento, ya que el modelo autorregresivo de rezagos distribuidos podrá hallar una relación de cointegración en el largo plazo (ya que en el análisis estático esta relación es nula).

Y en el contexto de los modelos propios de la Ciencia de Datos, este análisis reclama modelos que supervisan relaciones no lineales y que su vez, dependen de otras variables a lo largo del tiempo, en este sentido, los modelos de Ciencia de Datos fungen como el complemento perfecto a los modelos econométricos construidos.

*Gráfica 5. Relación estacionaria: Oferta monetaria y ln del precio de cierre IPC*




Elaboración propia con datos de Yahoo Finance y Banxico

## a) Análisis econométrico

### a.1) Modelo ARIMA sobre el logaritmo del precio de cierre ajustado del IPC

- Primer análisis ARIMA:

```
ARIMA(1,0,1)(0,0,0)[0]             : AIC=-514.218, Time=0.14 sec
ARIMA(0,0,0)(0,0,0)[0]             : AIC=-515.935, Time=0.06 sec
ARIMA(1,0,0)(0,0,0)[0]             : AIC=-516.079, Time=0.07 sec
ARIMA(0,0,1)(0,0,0)[0]             : AIC=-516.189, Time=0.08 sec
ARIMA(0,0,2)(0,0,0)[0]             : AIC=-514.231, Time=0.10 sec
ARIMA(1,0,2)(0,0,0)[0]             : AIC=-512.928, Time=0.28 sec
ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=-514.853, Time=0.19 sec

Best model:  ARIMA(0,0,1)(0,0,0)[0]
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Tabla 2. Resumen del primero modelo ARIMA*

```
--- Resumen del Mejor Modelo ARIMA sobre Datos Estacionarios ---
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  148
Model:               SARIMAX(0, 0, 1)   Log Likelihood                 260.094
Date:                Mon, 22 Sep 2025   AIC                           -516.189
Time:                        11:51:29   BIC                           -510.194
Sample:                    02-01-2011   HQIC                          -513.753
                         - 05-01-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
...
==============================================================================
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Este resultado **ARIMA (0,0,1)** nos habla de que los valores pasados en cuanto al logaritmo del precio de cierre ajustado del IPC no son útiles para encontrar una tendencia o cambio en la serie de tiempo. Además de que el componente de integración "d" está bien calculado, ya que la serie se había diferenciado con anterioridad, por eso arroja un resultado "d = 0".

En cuanto al componente de media móvil "q=1", se identifica un error en el periodo t-1, lo que indica que la serie de tiempo no depende de su valor en el mes anterior, sino del "shock" (el error de pronóstico) del mes anterior. Se realizará una segunda construcción del modelo ARIMA sobre IPC, con datos sin diferenciar para contrastar los resultados.

- Segundo ARIMA:

```
Performing stepwise search to minimize aic
      ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-516.320, Time=0.10 sec
      ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-517.832, Time=0.09 sec
      ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-518.115, Time=0.10 sec
      ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-518.261, Time=0.14 sec
      ARIMA(0,1,0)(0,0,0)[0]             : AIC=-519.459, Time=0.06 sec

      Best model:  ARIMA (0,1,0)(0,0,0)[0]
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Tabla 3. Segundo Modelo ARIMA*

```
--- Resumen del Mejor Modelo ARIMA (desde un enfoque estándar) ---
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  150
Model:               SARIMAX(0, 1, 0)   Log Likelihood                 260.730
Date:                Mon, 22 Sep 2025   AIC                           -519.459
Time:                        11:51:30   BIC                           -516.455
Sample:                    12-01-2010   HQIC                          -518.239
                         - 05-01-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0018      0.000     12.399      0.000       0.001       0.002
==============================================================================
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

**En este caso el resultado es ARIMA (0,1,0)**

Y podemos contrastar una mejora en el modelo, ya que tenemos un resultado de -519.459 en Criterio de Información de Akaike (AIC) en contra de -516.189 en el modelo anterior, y siguiendo la regla general con el Criterio de Información de Akaike (AIC) es que un valor más bajo indica un mejor modelo, y, por ende, un modelo estadísticamente más robusto. Con este resultado, se encuentra una caminata aleatoria con deriva, ya que:

- *p=0 (AR)*: No hay componente autorregresivo. El cambio en el IPC de hoy no depende de los cambios de los meses pasados.
- *d=1 (I)*: La serie fue diferenciada una vez para ser estacionaria. Esto confirma que el logaritmo del IPC tiene una raíz unitaria, esto confirma la diferenciación que se obtuvo en la parte correspondiente de diferenciación, que el logaritmo del IPC tenía que ser diferenciada una vez, para convertirse en serie estacionaria.
- *q=0 (MA)*: No hay componente de media móvil. El cambio en el IPC de hoy no depende de los errores de pronóstico de los meses pasados.

La ecuación de este modelo (incluyendo el intercepto que se ajustó) es:

$$\Delta \ln(IPC_t) = c + u_t$$

Que es lo mismo que:

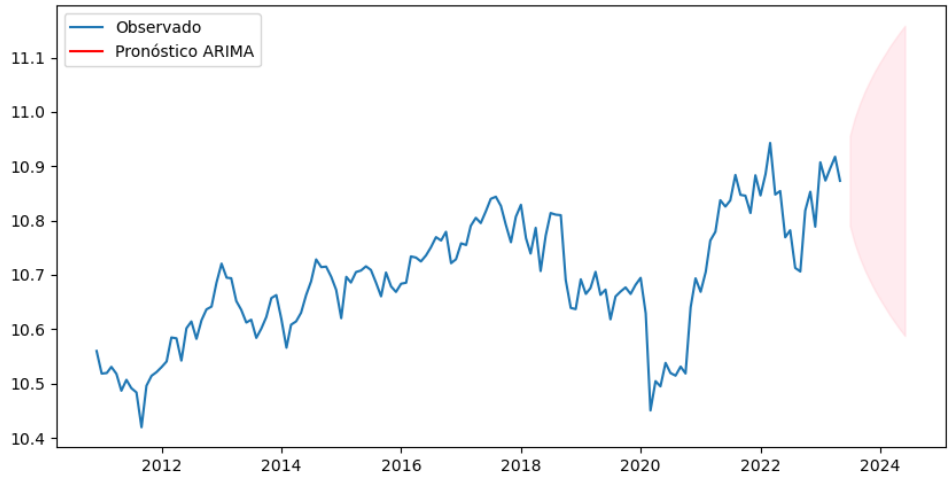
$$\ln(IPC_t) - ln(IPC_{t-1}) = c + u_t$$

Y que reordenando tenemos que:

$$\ln(IPC_t) = c + ln(IPC_{t-1}) + u_t;$$

- $\ln(IPC_t)$: Es el valor del logaritmo del IPC hoy.
- $c$ : Es una constante, conocida como la deriva.
- $ln(IPC_{t-1})$: Es el valor del logaritmo del IPC en el periodo anterior.
- $u_t$: Es un término de error aleatorio (ruido blanco).

*Gráfica 6. Pronóstico de IPC usando el segundo modelo ARIMA*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Por lo tanto, para este modelo se puede afirmar que la mejor predicción para el valor del IPC de mañana es simplemente el valor de hoy, más una pequeña tendencia constante, más un componente aleatorio.

Teóricamente, este comportamiento de la serie temporal respalda la hipótesis de los mercados eficientes al encontrar que los componentes AR (p=0) y MA (q=0). Pero siguiendo la hipótesis del mercado eficiente, tenemos que la constante, sino que está representada por la tendencia de crecimiento a largo plazo del mercado, el crecimiento económico promedio, la inflación esperada, la prima de riesgo que los inversionistas exigen por mantener activos bursátiles en lugar de activos libres de riesgo, así como otros componentes, que en este documento se abarcan con las variables macroeconómicas propuestas.

Ahora, después de ver el comportamiento individual del logaritmo de los precios de cierre de ajuste en IPC a través del modelo ARIMA, se construye un modelo autorregresivo de rezagos distribuidos tomando en cuenta una variable macroeconómica fundamental.

En cuanto al pronóstico que se hace a partir de modelo ARIMA:

El pronóstico se manifiesta como una línea recta con pendiente positiva. Esto es una consecuencia directa del modelo de caminata aleatoria, el cual predice que el cambio en el logaritmo del IPC será constante en cada periodo. Económicamente, un incremento constante en la escala logarítmica se traduce en una tasa de crecimiento porcentual constante en el nivel original del índice. De esta manera, el modelo sugiere que, a pesar de la volatilidad de corto plazo, la expectativa fundamental del mercado es continuar con su tendencia histórica de crecimiento. La pendiente de esta línea representa la deriva, un término constante que captura el rendimiento promedio a largo plazo del mercado, impulsado por factores fundamentales como el crecimiento económico y la prima de riesgo (Gujarati & Porter, 2010).

Por otro lado, el intervalo de confianza estimado se ensancha a medida que el pronóstico se aleja en el tiempo. Esta creciente incertidumbre es una característica fundamental de los modelos de paseo aleatorio y es consistente con la Hipótesis del Mercado Eficiente en su forma débil. Como lo describen Gujarati y Porter (2010) y Wooldridge (2013), si los precios de los activos siguen una caminata aleatoria, los movimientos futuros son inherentemente impredecibles a partir de la información pasada. El ensanchamiento del intervalo de confianza cuantifica esta incertidumbre acumulada, mostrando que la certeza de la predicción disminuye significativamente a medida que nos adentramos en el futuro.

Este pronóstico ARIMA sugiere un mercado con una tendencia alcista subyacente y constante, pero cuyos movimientos específicos a futuro son cada vez más inciertos, lo cual valida el comportamiento de paseo aleatorio identificado en el análisis de la serie.

### a.2) Modelo Autorregresivo de Rezagos Distribuidos (con datos diferenciados) para puntualizar la dinámica a corto plazo

```
Orden de rezagos óptimo seleccionado: [1] para Ln_Adj_close y {'Oferta
monetaria(M2)': [0]} para las exógenas.
```

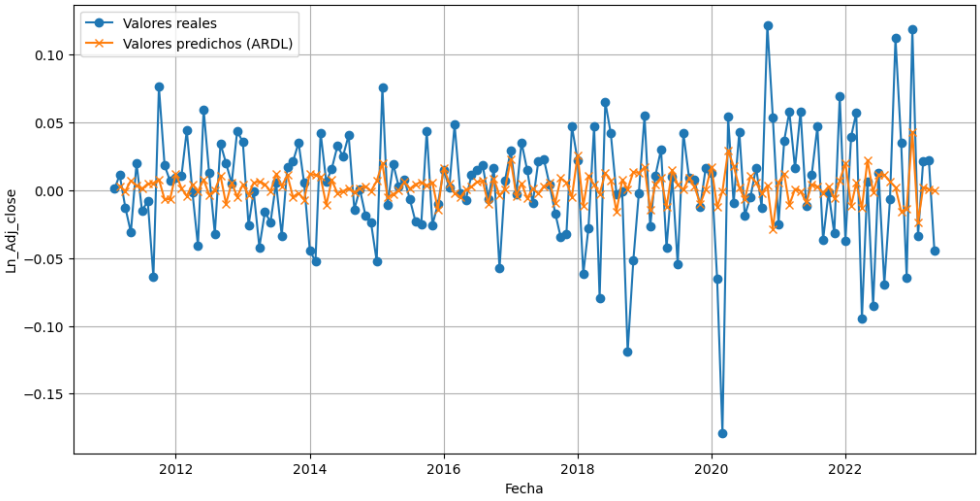
*Tabla 4. Modelo Autorregresivo de Rezagos Distribuidos*

```
                            ARDL Model Results
==============================================================================
Dep. Variable:           Ln_Adj_close   No. Observations:                  148
Model:                     ARDL(1, 0)   Log Likelihood                260.938
Method:               Conditional MLE   S.D. of innovations              0.041
Date:                Mon, 22 Sep 2025   AIC                           -513.875
Time:                        11:51:35   BIC                           -501.913
Sample:                    03-01-2011   HQIC                          -509.015
                         - 05-01-2023
================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          0.0028      0.003      0.826      0.410      -0.004       0.010
Ln_Adj_close.L1               -0.1470      0.082     -1.794      0.075      -0.309       0.015
Oferta monetaria(M2).L0 -6.062e-11   2.55e-11     -2.378      0.019   -1.11e-10   -1.02e-11
================================================================================================

--- Prueba de Límites para Cointegración (Bounds Test) ---
                            ARDL Model Results
==============================================================================
Dep. Variable:           Ln_Adj_close   No. Observations:                  148
Model:                     ARDL(1, 0)   Log Likelihood                260.938
...
const                          0.0028      0.003      0.826      0.410      -0.004       0.010
Ln_Adj_close.L1               -0.1470      0.082     -1.794      0.075      -0.309       0.015
Oferta monetaria(M2).L0 -6.062e-11   2.55e-11     -2.378      0.019   -1.11e-10   -1.02e-11
================================================================================================
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

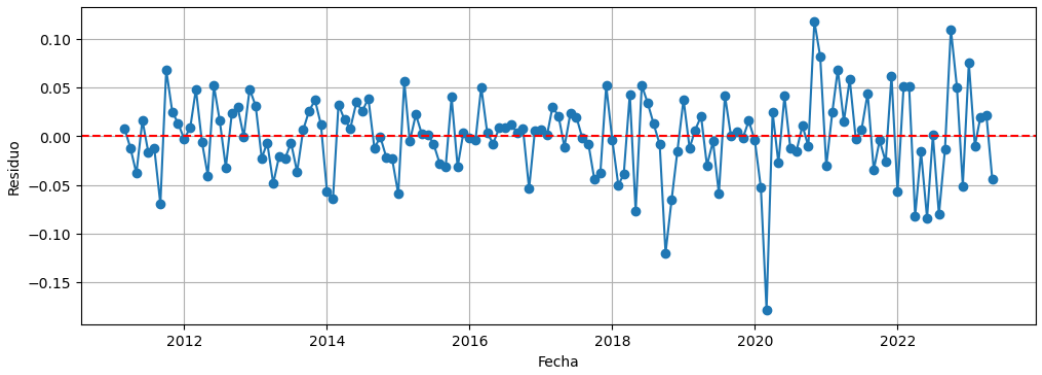
Gráficamente:

*Gráfica 7. Modelo ADRL-Valores reales del IPC vs valores predichos*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 8. Residuos del modelo autorregresivo de rezagos distribuidos*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Se tiene que, de acuerdo con el modelo planteado, que denota un análisis en el corto plazo, la ecuación correspondiente es:

$$\Delta ln(IPC)_t = 0.0028 - 0.1470 \cdot \Delta \ln(IPC)_{t-1} - 6.062 \times 10^{-11} \Delta OM_t + u_t$$

Donde:

- 0.0028: Es la contante del modelo y no es estadísticamente significativo, con valor de p= 0.410, o sea que no hay tendencia de crecimiento promedio mensual.
- $\Delta ln(IPC)_t$: Es el cambio en ln(IPC) de acuerdo con el mes actual.
- $\Delta ln(IPC)_{t-1}$: Es el cambio en el ln(IPC) en el mes anterior, que es marginalmente significativo con valor p = 0.075, esto indica que si el IPC sufrió un cambio positivo, para el mes corriente se espera un cambio menor.
- $\Delta OM_t$: Es el cambio en la oferta monetaria existente en el mes actual, y es significativo con un valor de p= 0.019 e indica la relación a corto plazo de estas variables, o sea que si cambia la oferta monetaria en el mes actual habrá una disminución en el IPC de ese mismo mes.

## A.2.1) Análisis con modelo autorregresivo de rezagos distribuidos (en niveles)

*Tabla 5. Resultados del modelo autorregresivo de rezagos distribuidos (en niveles)*

ARDL Model Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Ln_Adj_close | No. Observations: | 150 |
| Model: | ARDL(1, 1, 0, 1) | Log Likelihood | 273.971 |
| Method: | Conditional MLE | S.D. of innovations | 0.038 |
| Date: | mar., 30 sep. 2025 | AIC | -531.943 |
| Time: | 18:45:26 | BIC | -507.911 |
| Sample: | 01-01-2011 - 05-01-2023 | HQIC | -522.179 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.6393 | 0.564 | 4.682 | 0.000 | 1.525 | 3.754 |
| Ln_Adj_close.L1 | 0.7338 | 0.058 | 12.735 | 0.000 | 0.620 | 0.848 |
| Oferta monetaria(M2).L0 | -6.401e-11 | 3.5e-11 | -1.826 | 0.070 | -1.33e-10 | 5.27e-12 |
| Oferta monetaria(M2).L1 | 7.151e-11 | 3.54e-11 | 2.020 | 0.045 | 1.51e-12 | 1.42e-10 |
| IGAE Serie desestacionalizada.L0 | 0.0019 | 0.001 | 1.991 | 0.048 | 1.37e-05 | 0.004 |
| Tasa de interés promedio mensual.L0 | 0.0558 | 0.028 | 1.982 | 0.049 | 0.000 | 0.112 |
| Tasa de interés promedio mensual.L1 | -0.0655 | 0.029 | -2.260 | 0.025 | -0.123 | -0.008 |

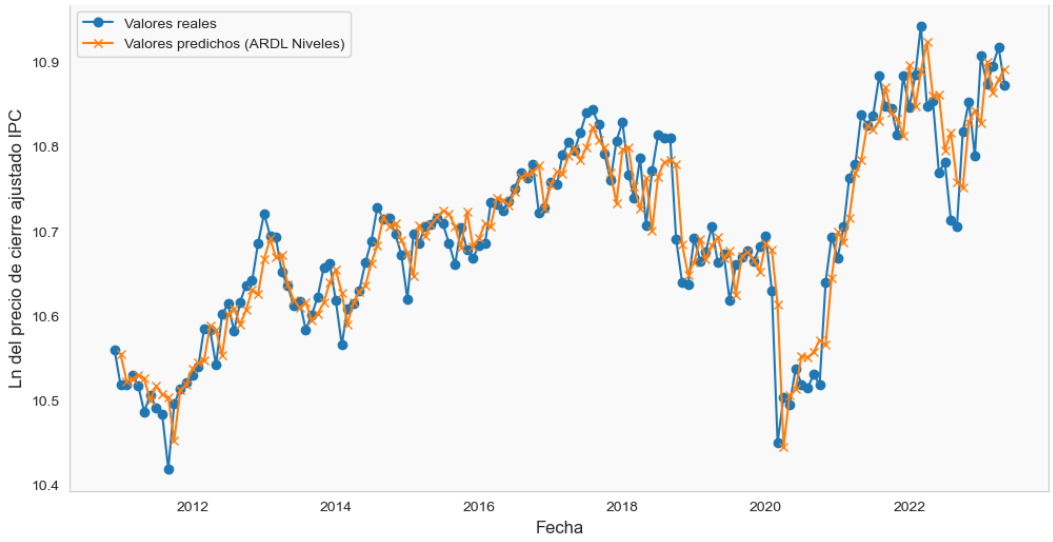
--- Prueba de Límites para Cointegración en Niveles (Bounds Test) ---

...

| | | | | | | |
|---|---|---|---|---|---|---|
| IGAE Serie desestacionalizada.L0 | 0.0019 | 0.001 | 1.991 | 0.048 | 1.37e-05 | 0.004 |
| Tasa de interés promedio mensual.L0 | 0.0558 | 0.028 | 1.982 | 0.049 | 0.000 | 0.112 |
| Tasa de interés promedio mensual.L1 | -0.0655 | 0.029 | -2.260 | 0.025 | -0.123 | -0.008 |

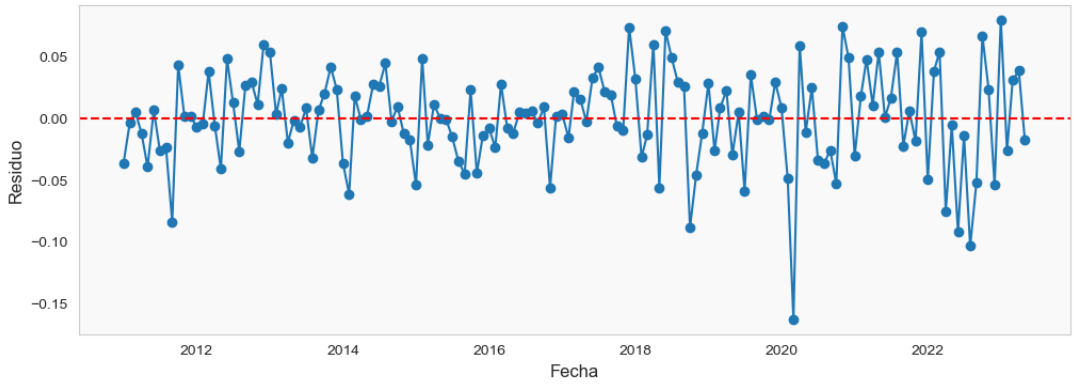
Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 9. Valores reales vs predichos (en niveles) bajo el modelo autorregresivo de rezagos distribuidos*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 10. Residuos del modelo autorregresivo de rezagos distribuidos (en niveles)*




En este contexto, se tiene una métrica de AIC más robusta de -531.943, y además se tiene que el modelo en niveles halló como variables significativas a la oferta monetaria, a la actividad económica y a la tasa de interés. Este resultado sustenta la elección de las variables macroeconómicas, al tiempo que confirma la cointegración, ya que cuando el IPC se desvía del equilibrio en el largo plazo, este se corrige en un 26.62%, ya que teniendo en cuenta el coeficiente el logaritmo del precio de cierre ajustado:

$$1 - 0.7338 = .2662 * 100 = 26.62\%$$

Con el modelo (1,1,0,1) se está incluyendo propiamente:

- Un rezago de la variable dependiente.
- Un rezago que corresponde a la oferta monetaria.
- Ningún rezago de la actividad económica.
- Un rezago de la tasa de interés.

Se tiene también que el modelo tiene un buen ajuste, y que es preciso, además de que la prueba bonus confirma la evidencia de cointegración ya que varios de los coeficientes resultaron ser significativos con valor $p < 0.05$

Y tenemos la siguiente ecuación:

$$\ln(IPC) = 2.6393 + 0.7338\ln(IPC)_{t-1} - 6.401x10^{-11}OM_t + 7.151x10^{-11}OM_{t-1} + 0.0019IGAE_t + 0.0558TI_t - 0.0TI_{t-1} + u_t$$

Con este resultado se puede decir que, la caminata aleatoria no es nada simple, ya que, en el comportamiento, está contenida la actividad económica, oferta del dinero y su propio costo.

Así, los agentes económicos interesados en invertir podemos obtener una ventaja muy significativa al consultar los datos públicos existentes en Banco de México para tomar decisiones financieras. A diferencia de Enders (2014), en este resultado se obtiene que el mercado financiero tiene un sustento macroeconómico, que es dinámico y que se ajusta en el largo plazo.

Al encontrar una relación de cointegración significativa, se captura empíricamente el efecto que el profesor Hernández Mota, J. L. (2015) describe teóricamente en su artículo. Es decir, se demuestra que el desarrollo y la liquidez del sistema financiero, junto con la actividad económica real y la tasa de interés, tienen un efecto real y cuantificable en la economía, tal como predice el modelo teórico.

### b) Análisis desde el Machine Learning

En esta parte del análisis, ofrezco una visión integrativa, ya que voy a trabajar con esta serie temporal con los algoritmos citados. Además de que se añade al análisis, en aras de conducir el documento por una buena práctica, el desarrollo de un modelo Lasso, ya que se busca de aplicarlo como selector de variables.

Siguiendo la literatura correspondiente, como en Géron, A. (2022) y Brownlee, J. (2018), este problema de trabajo con series temporales se convierte en un análisis de aprendizaje supervisado, esto para la correcta construcción de los modelos ofrecidos para la predicción del comportamiento del IPC.

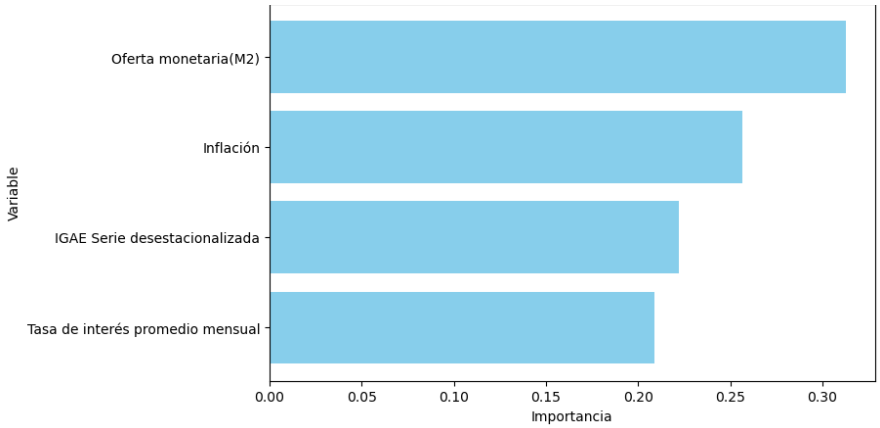
#### b.1) Modelo Random Forest

Resultados obtenidos:

```
--- Evaluación 1 del Modelo Random Forest ---
Puntaje R² en entrenamiento: 0.6294
Puntaje R² en prueba: -0.2757

Importancia de las características: [0.25636429 0.31290641 0.22182905
0.20890024]

Características: ['Inflación', 'Oferta monetaria(M2)', 'IGAE Serie
desestacionalizada', 'Tasa de interés promedio mensual']

Importancia de las características (con nombres): {'Inflación':
np.float64(0.25636429472550964), 'Oferta monetaria(M2)':
np.float64(0.3129064101306756), 'IGAE Serie desestacionalizada':
np.float64(0.22182905241336837), 'Tasa de interés promedio mensual':
np.float64(0.20890024273044644)}

--- Evaluación 2 del Modelo Random Forest ---
MSE: 0.0035
RMSE: 0.0595
MAE: 0.0479
R²: -0.2757
```

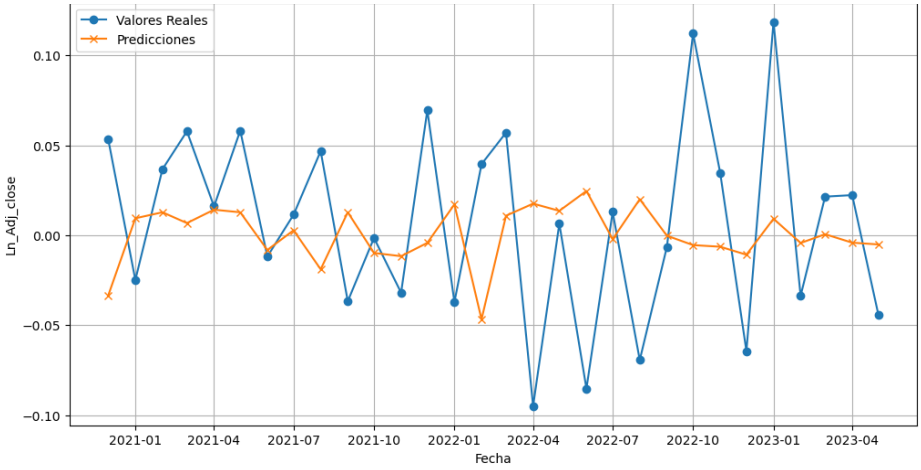
Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 11: Importancia de características en el modelo Random Forest*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 12. Predicciones del modelo Random Forest en contra los valores reales del IPC*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Se efectúa un segundo análisis, en aras de evitar el sobreajuste en los resultados, este segundo análisis de modelo optimizado funge como resultado definitivo, y como base para el contraste de resultados.

```
--- Evaluación del Modelo Random Forest OPTIMIZADO ---
Puntaje R² en entrenamiento: 0.2673
Puntaje R² en prueba: -0.0660
RMSE en prueba: 0.0546
MAE en prueba: 0.0450
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

### b.2) Modelo Extreme Gradient Boosting (XGBoost)

```
--- Evaluación del Modelo XGBoost ---
R² Entrenamiento: 0.8505
R² Prueba: -0.3270
MSE (Prueba): 0.0037
RMSE (Prueba): 0.0607
MAE (Prueba): 0.0502
```

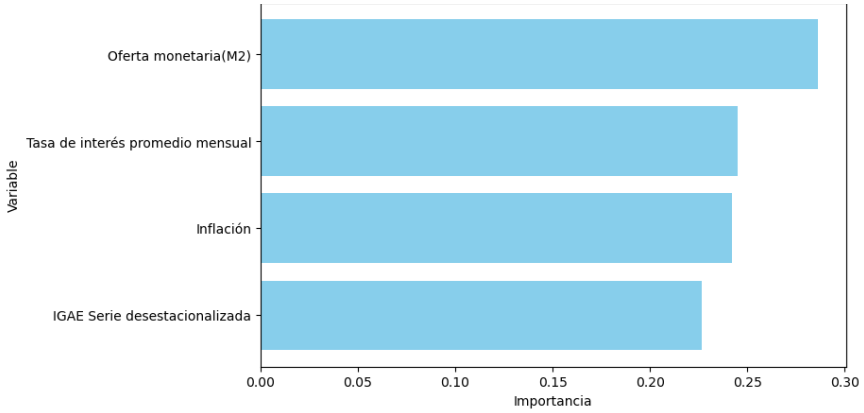
Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

De la misma manera, procedo a llevar a cabo el mismo procedimiento de optimización en el código para el modelo XG Boost, esta optimización consiste en la optimización de hiperparámetros dentro del código, y la validación cruzada para series de tiempo, así como característica rezagadas del logaritmo de el precio de cierre ajustado para que el modelo tenga una memoria optimizada de los cambios recientes

```
--- Evaluación del Modelo XGBoost OPTIMIZADO ---
Puntaje R² en entrenamiento: 0.9436
Puntaje R² en prueba: -0.1454
RMSE en prueba: 0.0566
MAE en prueba: 0.0456
```

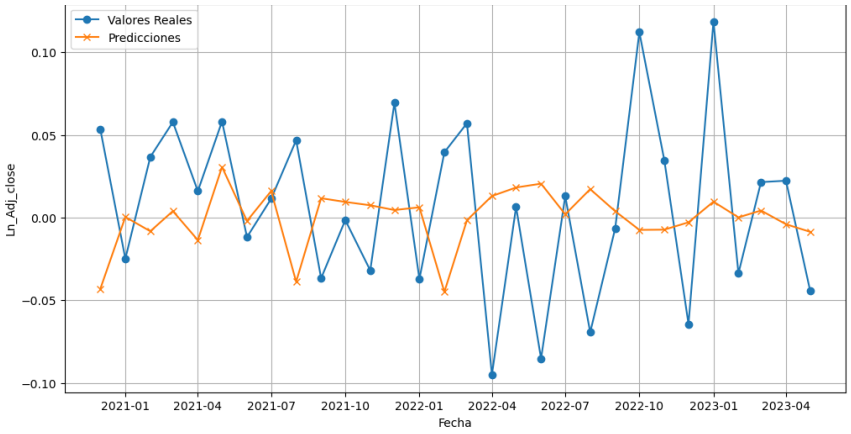
Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 13. Importancia de características en el modelo XGBoost*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 14. Predicciones del Modelo XGBoost en contra de valores reales del IPC*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

### c) Análisis desde el Aprendizaje Profundo

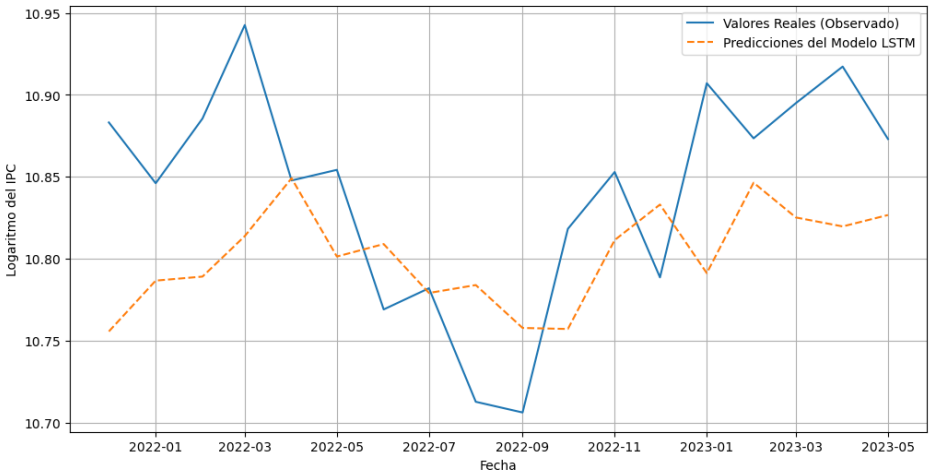
### c.1) Modelo LSTM (Long Short-Term Memory)

A continuación, se usará un modelo llamado LSTM (Long Short-Term Memory) el entrenamiento de este modelo le da un giro importante a esta investigación dada la relevancia dinámica que se halló entre las variables, y propongo el uso del LSTM, ya que su arquitectura introduce un mecanismo para decidir qué información guardar, cuál olvidar y qué información usar para la predicción.

```
Métricas de Desempeño del Modelo LSTM (sobre datos de prueba) ---
Error Absoluto Medio (MAE): 0.0631
Raíz del Error Cuadrático Medio (RMSE): 0.0731
```

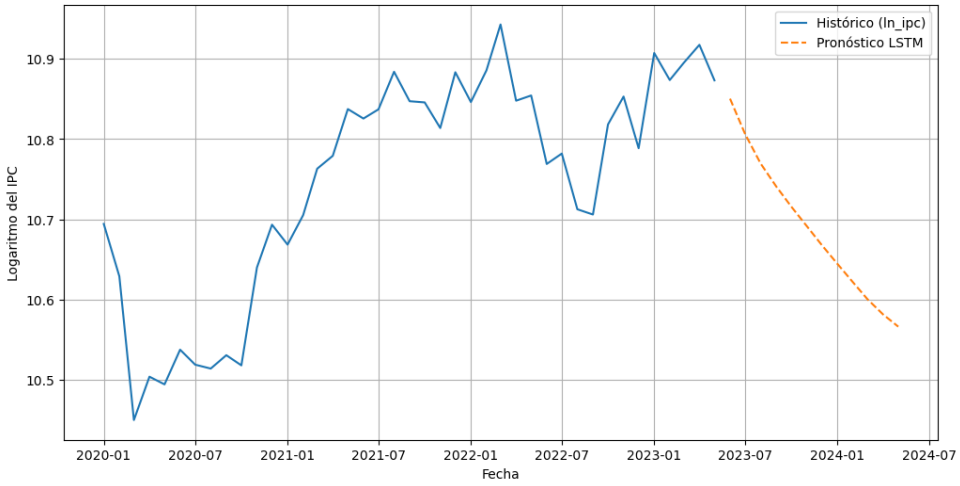
Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 15. Predicciones de la red neuronal LSTM contra los valores reales del IPC*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

*Gráfica 16. Pronóstico del comportamiento del IPC con red neuronal LSTM*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

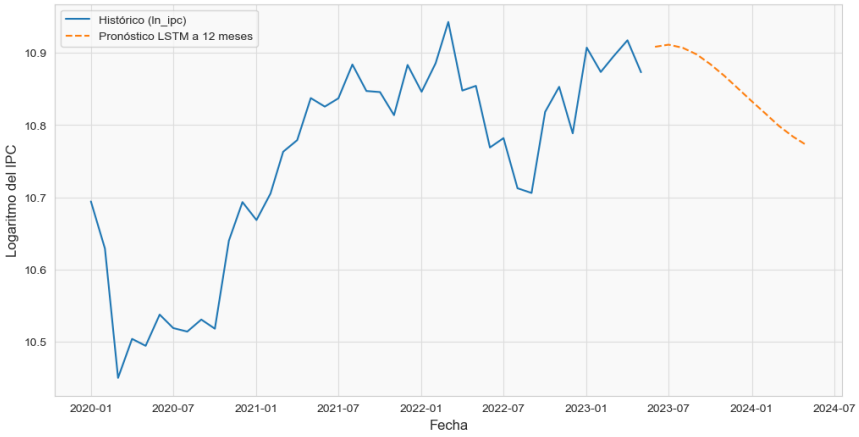
En esta parte del análisis, también se optimiza el modelo correspondiente a redes neuronales.

```
--- Resultados de la Validación Cruzada ---
Scores RMSE de cada fold: [np.float64(0.0381), np.float64(0.0359),
np.float64(0.0876), np.float64(0.0885), np.float64(0.0738)]
RMSE Promedio: 0.0648
Desviación Estándar del RMSE: 0.0233
```

Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

Gráficamente, la predicción correspondiente:

*Gráfica 17. Pronóstico del IPC, bajo modelo LSTM optimizado*




Análisis de elaboración propia, en entorno Jupyter (lenguaje Python 3.11.9)

# Conclusiones

Para poder establecer las conclusiones pertinentes, es imperativo contrastar las métricas de los modelos involucrados, pero también contrastar las predicciones con el comportamiento real del IPC, y comparar los resultados con el comportamiento real del Índice de Precios y cotizaciones.

Siguiendo esta idea de análisis, presento la serie de tiempo completa, para de primera instancia, compararla con la predicción que hizo el modelo ARIMA. Este análisis captura quince años de comportamiento, desde la primera fecha, enero 2010 hasta septiembre 2025.

*Gráfica 18. Comportamiento de la serie de tiempo 2010 -2025, Precio de Cierre Ajustado del IPC*


Análisis de elaboración propia, en Excel, con datos de Yahoo Finance

# I. Resumen de hallazgos

Este trabajo de investigación naturalmente busca el modelo óptimo para el comportamiento del IPC bajo un análisis comparativo entre modelos econométricos y modelos pertenecientes a la ciencia de datos.

Siguiendo la narrativa de datos que corresponde a los resultados obtenidos, el modelo ARIMA (0,1,0) fue seleccionado como el más parsimonioso y con el mejor ajuste estadístico (AIC de -519.459). Este modelo, corresponde a una caminata aleatoria con deriva, sugiere que la mejor predicción para el IPC futuro es su valor actual más una tendencia constante, como se modelo en la parte del análisis correspondiente a este modelo. Este resultado respalda la Hipótesis del Mercado Eficiente en el sentido de que los movimientos del mercado son estocásticos y no pueden ser pronosticados consistentemente a partir de su propia historia, en el gráfico 6, se ve la tendencia con pendiente positiva existente, comportamiento que se corrobora en el comportamiento real del Índice de Precios y Cotizaciones hasta septiembre 2025.

Posteriormente, se siguió el análisis en aras de complementar el resultado obtenido con el modelo ARIMA (0,1,0) con un modelo autorregresivo de rezagos distribuidos, y se encontró que la oferta monetaria tiene una relación a largo plazo estadísticamente significativa en el corto plazo. Luego se replanteó el análisis, pensado en el largo plazo y se encontró que el comportamiento del IPC tiene un sustento macroeconómico en su mayoría, y que bajo este análisis, se tiene una velocidad de ajuste mensual del 26.6% al mes, en términos regreso al equilibrio, y que, como se mencionó anteriormente, los agentes económicos tenemos la ventaja en términos de consulta de información para hacer decisiones financieras

Ahora, bajo la hipótesis de que las relaciones entre las variables macroeconómicas y el comportamiento del índice de precios y cotizaciones, podría tener una dinámica no lineal, se usaron modelos propios del machine learning y el aprendizaje profundo. Y aunque se identificó y validó el resultado del modelo autorregresivo de rezagos distribuidos, en el sentido de que la oferta monetaria es la variable más importante en relación con el IPC, estos modelos tuvieron un severo sobreajuste, a pesar de que se hizo la optimización propia y métodos de validación pertinentes (en los cuales se añadieron rezagos del logaritmo del precio de cierre ajustado del IPC, dentro del modelo XG Boost). Este resultado demuestra, que a pesar de que los modelos de machine learning tienen una gran capacidad para aprender patrones complejos, estos no lograron hacer una buena predicción del IPC.

Finalmente, se entrenó un modelo LSTM (tanto en su versión optimizada, como en la primer versión), y la red neuronal predijo una tendencia bajista, mientras en la realidad sucedió lo contrario. Económicamente, esto sugiere que la red neuronal identificó un patrón complejo en la configuración de las variables macroeconómicas que, basado en la historia, señalaba una posible corrección del mercado, un tipo de advertencia que los otros modelos no pudieron generar.

Entonces:

- El modelo autorregresivo de rezagos distribuidos confirma una relación de equilibrio a largo plazo con fundamentos macroeconómicos.
- Los modelos de machine learning fallan en la predicción de los movimientos de corto plazo, ya que en el conjunto de prueba (con datos no vistos) es donde el modelo falló obteniendo un $R^2$ de -0.0660 para Random Forest y -0.1454 para XGBoost, es decir, que estos modelos se quedaron sin poder predictivo. Lo que nos dice que los modelos complejos no funcionan correctamente para ser usados en el corto plazo.
- El Modelo ARIMA resultó un modelo acertado para la predicción del IPC, y el modelo autorregresivo de rezagos distribuidos fue contundente en las predicciones. Lo que deja claro que el uso del modelado econométrico, por su precisión y metodología aún es superior a los modelos propios de la ciencia de datos que se utilizaron.

## II. Respuesta a los objetivos planteados

1. En primer lugar, se demostró que el modelo ARIMA es un modelo, en términos de desempeño, para predecir el comportamiento del IPC dentro un intervalo de un año.
2. A partir de una selección de variables macroeconómicas, se encontró una relación a largo plazo entre el IPC y la oferta monetaria, actividad económica y tasa de interés, a partir del análisis realizado bajo el modelo autorregresivo de rezagos distribuidos.
3. Se encontró una relación significativa en el corto plazo entre el IPC y la oferta monetaria, bajo el análisis que se presenta en el modelo autorregresivo de rezagos distribuidos tanto como en el corto, como en el largo plazo.
4. Se demostró también que los modelos econométricos pueden complementar el análisis predictivo con modelos propios de machine learning, aunque estos modelos no resultaron satisfactorios en un intervalo tan corto, propio de un año.
5. La oferta monetaria fue la variable más importante para ambos modelos, tanto el modelo autorregresivo de rezagos distribuidos, como en los modelos XGBoost y Random Forest, esto puede ser el punto de partida para el diseño de políticas económicas, medidas y programas que estimulen la inversión por parte de los agentes económicos, y de acuerdo con el profesor Hernández Mota, J. L. (2015), fortalecer el sistema financiero mexicano, al tiempo que se mejora la tasa de crecimiento de la economía.
6. En consecuencia, del punto anterior, y siguiendo a Tobin, J. (1969)[11] robustece el análisis del desempeño de la oferta monetaria.

[11] Tobin J. (1969) mantiene la hipótesis que, bajo una política monetaria expansiva, los agentes económicos diversifican su portafolio de inversión teniendo en cuenta la liquidez existente, es así como existe un nuevo equilibrio. A menor liquidez, los agentes económicos usan ese ingreso para destinarlo al consumo.

## III. Observaciones y posibles mejoras en función de los resultados obtenidos

Referente a las recomendaciones y observaciones:

1. A lo largo de este documento se demostró que la aplicación de algoritmos propios del machine learning se tiene que usar con cuidado, dado su desempeño, esto puede llevar a interpretaciones erróneas de la dinámica del mercado, y, por consiguiente, a decisiones sesgadas. La recomendación para los analistas interesados es usar ambas perspectivas, y usar el criterio económico para tomar decisiones basadas en datos.
2. Modelar la volatilidad del IPC a través de un modelo GARCH.
3. En la medida de lo posible, usar modelos híbridos.
4. Se podría aprovechar el poder predictivo de los modelos de machine learning usando un intervalo más amplio, y más variables macroeconómicas, haciendo más rico el análisis, o bien para darle continuidad a la predicción que hace el modelo ARIMA, es decir, hacer el trabajo predictivo a un año con el modelo ARIMA, y a partir de ahí seguir el análisis con los modelos de machine learning y de aprendizaje profundo.

# Bibliografía


- Boschetti, A., & Massaron, L. (2023). *Python data science essentials* (3rd ed.). Packt Publishing.
- Brownlee, J. (2018). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.
- Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10), 2451–2471.
- Gujarati, D. N., & Porter, D. C. (2010). *Econometría* (5ª ed.). McGraw-Hill.
- Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251.
- Hernández Mota, J. L. (2015). *El papel del desarrollo financiero como fuente del crecimiento económico*. Revista finanzas y política económica, 7(2), 235–256. https://doi.org/10.14718/revfinanzpolitecon.2015.7.2.2
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill.
- Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
- McMillan, D. G. (2001). Non-linear predictability of UK stock market returns. *Oxford Bulletin of Economics and Statistics*, *63*(5), 639–657.
- Nkoro, E., & Uko, A. K. (2016). *Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation*. *Journal of Statistical and Econometric Methods*, *5*, 1–3. https://www.scienpress.com/Upload/JSEM/Vol%205_4_3.pdf
- Scikit-learn developers. (2025). *Supervised learning*. En *Scikit-learn 1.7.2 Documentation: https://scikit-learn.org/stable/supervised_learning.html*
- Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking*, *1*(1), 15–29.
- VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media.


# Anexos

Repositorio de Git & GitHub en donde se encuentra el cuaderno de trabajo y se construyen los modelos a partir de código en Python, usando las librerías descritas:

https://github.com/GabrielGM153