Instituto Politécnico Nacional

Escuela Superior de Computo

Carrera

Licenciatura en Ciencia de Datos

Alumno

Aguilar Ramírez Carlos Francisco

Arista Romero Juan Ismael

Jiménez Flores Luis Arturo

Vázquez Martin Marlene Gabriela

Profesor

Daniel Jiménez Alcantar

Materia

Análisis de Series de Tiempo

Grupo

6AV1

Practica 5

Construir modelo ARIMA

Instrucciones

1.- Elija un Dataset de su elección, deberá acondicionarse de tal manera que pueda realizar el análisis de una serie de tiempo. Desarrolle un reporte técnico que permita observar el trabajo en los siguientes puntos:

1. Introducción
2. Problemática
3. Modelo estadístico
4. Modelo computacional
5. Metodología.
6. Propuesta de solución para construir el modelo ARIMA.
    a. Aplica la metodología BOX-JENKINS.
7. Conclusiones por integrante.

Introducción

En el presente reporte se detalla el proceso de análisis de una serie de tiempo correspondiente a los precios históricos de cierre ajustados (adj_close variable central) de las acciones de Walmart Inc. (WMT). El dataset abarca el periodo desde el 25 de agosto de 1972 hasta el 21 de febrero de 2025. El objetivo principal es construir un modelo Autorregresivo Integrado de Media Móvil (ARIMA) capaz de describir y predecir el comportamiento de esta serie temporal financiera. Para ello, se sigue la metodología Box-Jenkins.

Se han realizado las siguientes modificaciones para acondicionar los datos y poder llevar a cabo el análisis de la serie de tiempo:

1. Conversión de la columna 'date' a tipo datetime con UTC.
2. Creación de un dataframe limpio que contenga solo las columnas 'date' y 'adj_close'.
3. Eliminación de la hora, dejando solo la fecha.
4. Verificación de valores faltantes y ordenamiento cronológico de los registros.

```python
# Convertir a datetime con UTC para asegurar que todo esté correcto
df['date'] = pd.to_datetime(df['date'], utc=True)

# Crear dataframe limpio con fecha y adj_close únicamente
walmart_clean_df = df[['date', 'adj_close']].copy()

# Eliminar la hora, dejando solo la fecha
walmart_clean_df['date'] = walmart_clean_df['date'].dt.date

# Verificar el resultado
print(walmart_clean_df.head())
```

Python

```
         date  adj_close
0  1972-08-25   0.011639
1  1972-08-28   0.011595
2  1972-08-29   0.011463
3  1972-08-30   0.011463
4  1972-08-31   0.011286
```

Figura 1. Transformación del conjunto de datos.

Problemática

La predicción de precios de acciones es un desafío clásico en el análisis de series de tiempo financieras. Estas series suelen presentar características como:

- **No estacionariedad:** Presencia de tendencias (generalmente alcistas a largo plazo; como puede ser en el caso de WMT, como se observa en el EDA) y posible varianza no constante.
- **Dependencia temporal:** El valor actual depende de los valores pasados y de errores de predicción anteriores.
- **Posible estacionalidad y ciclos:** Patrones que se repiten en el tiempo, aunque la estacionalidad puede ser menos pronunciada en precios diarios que en datos agregados.

Modelar y predecir estas series requiere un enfoque que capture adecuadamente las características previamente mencionadas.

El objetivo es desarrollar un modelo ARIMA que represente la estructura subyacente de la serie adj_close de WMT y permita generar pronósticos razonables a corto plazo.

Modelo estadístico

El modelo estadístico seleccionado es el **ARIMA(p, d, q)**, que se compone de:

- **AR(p) - Componente Autorregresivo:** Modela la dependencia lineal entre una observación actual y un número p de observaciones anteriores.
- **I(d) - Componente Integrado:** Representa el número d de diferencias necesarias para convertir una serie no estacionaria en una estacionaria. La diferenciación ayuda a eliminar tendencias y estabilizar la media.
- **MA(q) - Componente de Media Móvil:** Modela la dependencia entre una observación actual y los errores residuales de un número q de predicciones anteriores.

Los órdenes específicos (p, d, q) no se definen a priori, sino que se determinan empíricamente durante la fase de identificación de la metodología Box-Jenkins, analizando las propiedades de la serie de tiempo.

Modelo computacional

La implementación del análisis y modelado se realizó utilizando el lenguaje de programación Python junto con las siguientes bibliotecas utilizadas:

Pandas: Para la manipulación y preparación de los datos (carga del CSV, manejo de fechas, creación de DataFrames), también se utilizó en la transformación de los datos.

NumPy: Utilizado para operaciones numéricas eficientes.

Statsmodels: Biblioteca fundamental para el análisis de series de tiempo, utilizada para:

- Prueba de estacionariedad (Augmented Dickey-Fuller - adfuller).
- Cálculo y visualización de funciones de autocorrelación (ACF - plot_acf) y autocorrelación parcial (PACF - plot_pacf).
- Descomposición estacional (seasonal_decompose).

- Ajuste del modelo ARIMA (ARIMA).
- Diagnóstico de residuos (Prueba de Ljung-Box - acorr_ljungbox, QQ-plot - sm.qqplot).
- Generación de predicciones (get_forecast).

Matplotlib y Seaborn: Para la visualización de la serie, resultados del EDA, ACF/PACF, residuos y predicciones. Seaborn para darle diseño a los gráficos.
Scipy: Específicamente scipy.signal.find_peaks para la identificación exploratoria de picos y valles.
Sklearn.metrics: Para calcular métricas de evaluación del modelo (e.g., mean_squared_error).
El código se estructuró siguiendo los pasos lógicos del análisis de series de tiempo y la metodología Box-Jenkins.

Metodología.

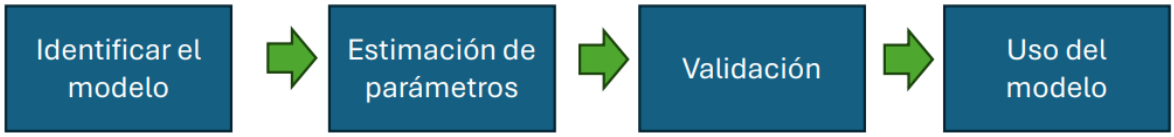

Figura 2. Proceso metodológico de Box-Jenkins.

Se empleó la **Metodología Box-Jenkins**, con un enfoque iterativo para la construcción de modelos ARIMA, que consta de las siguientes etapas principales:

1. **Identificación del Modelo:**

Análisis exploratorio de datos (EDA): Visualización de la serie, cálculo de estadísticas descriptivas, análisis de tendencia, estacionalidad, ciclos, picos y valles.
Verificación de estacionariedad: Uso de la prueba ADF para determinar si la serie es estacionaria en media y varianza.
Transformación (si es necesaria): Aplicación de diferencias (d) hasta lograr la estacionariedad. También se podría considerar transformación logarítmica si la varianza no es constante.
Análisis ACF y PACF: Estudio de las gráficas de autocorrelación y autocorrelación parcial de la serie *estacionaria* para proponer órdenes iniciales para p (AR) y q (MA).

2. **Estimación de Parámetros:**

Ajuste del modelo ARIMA(p, d, q) propuesto a los datos de entrenamiento (una porción de la serie histórica).
Obtención de los coeficientes del modelo y sus errores estándar, así como criterios de información (AIC, BIC).

3. **Diagnóstico del Modelo:**

Análisis de los residuos (la diferencia entre los valores observados y los ajustados por el modelo). Se busca que los residuos se comporten como "ruido blanco" (media cero, varianza constante, sin autocorrelación).
Herramientas: Gráfico de residuos vs tiempo, ACF de residuos, QQ-plot para normalidad, prueba de Ljung-Box para autocorrelación.

Si el diagnóstico no es satisfactorio, se regresa a la etapa de identificación para refinar el modelo (cambiar órdenes p, d, q, considerar transformaciones adicionales o modelos como SARIMA).

4. **Predicción (Uso del Modelo):**

Una vez validado el modelo, se utiliza para generar pronósticos fuera de la muestra (sobre el conjunto de prueba).
Evaluación del rendimiento predictivo utilizando métricas como RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) y MAPE (Mean Absolute Percentage Error).

Propuesta de solución para construir el modelo ARIMA.

La aplicación específica de la metodología Box-Jenkins en el código fue la siguiente:
Pre-procesamiento: Se cargaron los datos de WMT.csv, se seleccionaron las columnas date y adj_close, se convirtieron a formato de fecha adecuado y se estableció date como índice.

```
--- Preparación de la Serie de Tiempo (adj_close) ---

Serie 'adj_close' preparada. Longitud: 13233 puntos.
Fechas desde 1972-08-25 04:00:00+00:00 hasta 2025-02-21 05:00:00+00:00
Últimas 5 observaciones de la serie:
date
2025-02-14 05:00:00+00:00    104.040001
2025-02-18 05:00:00+00:00    103.779999
2025-02-19 05:00:00+00:00    104.000000
2025-02-20 05:00:00+00:00     97.209999
2025-02-21 05:00:00+00:00     94.779999
Name: adj_close, dtype: float64
```

Figura 3. Resultados del pre-procesamiento.

```
4.2 Estadísticas Descriptivas de 'adj_close'...
Estadísticas básicas:
  Promedio (Media): 11.9462
  Mediana: 7.3508
  Moda: 0.0069
  Desviación Estándar: 15.8655
  Varianza: 251.7129

Percentiles extremos:
0.01     0.005291
0.05     0.009169
0.95    46.126460
0.99    70.361108
Name: adj_close, dtype: float64
```

Figura 4. Estadística de los datos pre-procesados.

Aplica la metodología BOX-JENKINS.

Identificación:

Se realizó la prueba ADF sobre la serie original adj_close. El resultado indicó que la serie **NO era estacionaria** (p-value > 0.05).

```
5.1.1 Verificando estacionariedad de la serie original (Prueba ADF)...
ADF Statistic: 7.2116
p-value: 1.0000
Critical Values:
        1%: -3.4308
        5%: -2.8618
        10%: -2.5669

Resultado ADF (p-value=1.0000): La serie original NO es estacionaria. Aplicando diferencias...

ADF Test (d=1) p-value: 0.0000
La serie con d=1 ES estacionaria.

Orden de diferenciación seleccionado: d = 1
```

Figura 5. Resultados de la prueba ADF.

Se procedió a diferenciar la serie. El código prueba primero con d=1 y luego con d=2 si es necesario.

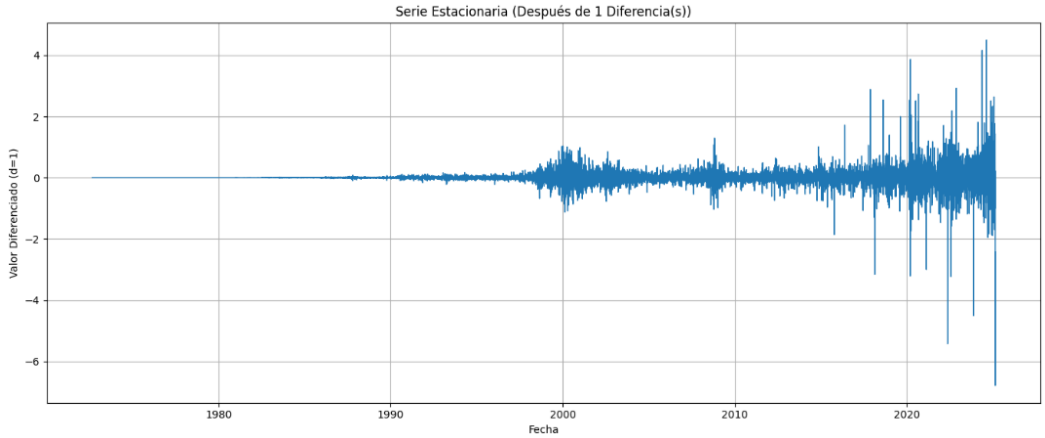

Figura 6. Visualización gráfica del comportamiento estacionario de la serie.

Una vez obtenida la serie estacionaria (con d diferencias), en este caso d=1, se graficaron sus funciones ACF y PACF (lags=40).

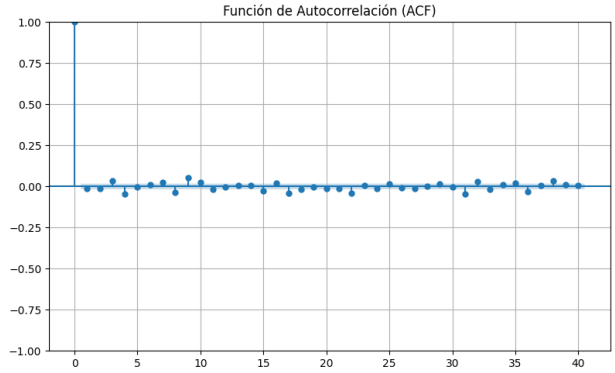

Figura 7. Gráfico de la función ACF.

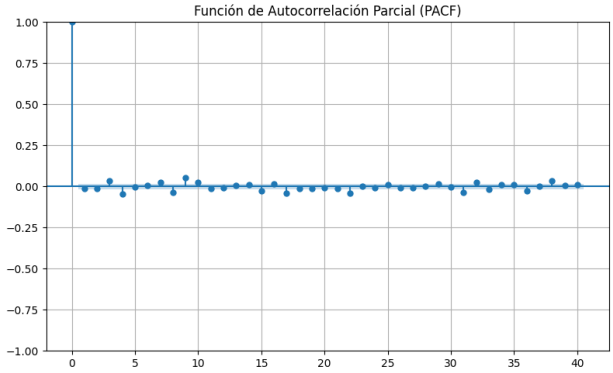

Figura 8. Gráfico de la función PACF.

Basándose en la interpretación de las gráficas ACF/PACF (buscando "cortes" o decaimientos), se seleccionaron los órdenes iniciales p y q. El código establece p_elegido = 2 y q_elegido = 2 como valores iniciales.

```
--- Guía para Interpretar ACF/PACF ---
  - Buscar el lag donde la ACF 'corta' (cae a cero) -> sugiere orden 'q' para MA(q).
  - Buscar el lag donde la PACF 'corta' (cae a cero) -> sugiere orden 'p' para AR(p).
  - Si ambas decaen lentamente -> sugiere modelo mixto ARMA(p, q).
  - Las bandas azules indican el intervalo de confianza; lags fuera de ellas son significativos.

Órdenes iniciales seleccionados para el modelo: ARIMA(p=2, d=1, q=2)
```

Figura 9. Interpretación de los gráficos para seleccionar los parámetros de un modelo ARIMA.

Estimación:

La serie se dividió en 80% para entrenamiento y 20% para prueba. Por otra parte, se ajustó el modelo ARIMA(2, d, 2) (con el d determinado en 1) a los datos de train_data utilizando ARIMA(train_data, order=(p_elegido, d, q_elegido)).fit().
Posteriormente se imprimió el resumen del modelo (modelo_ajustado.summary()) que contiene coeficientes, errores estándar, valores p, AIC, BIC, etc.

```
Tamaño Total Serie: 13233 puntos
Tamaño Entrenamiento: 10586 puntos (1972-08-25 a 2014-08-13)
Tamaño Prueba:       2647 puntos (2014-08-14 a 2025-02-21)

Ajustando modelo ARIMA(p=2, d=1, q=2) en los datos de entrenamiento...

Resumen del Modelo Ajustado:
                               SARIMAX Results
==============================================================================
Dep. Variable:              adj_close   No. Observations:                10586
Model:                 ARIMA(2, 1, 2)   Log Likelihood                7156.671
Date:                Sun, 13 Apr 2025   AIC                         -14303.343
Time:                        20:26:51   BIC                         -14267.007
Sample:                             0   HQIC                        -14291.078
                                - 10586
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3827      0.122      3.142      0.002       0.144       0.621
ar.L2          0.2205      0.101      2.193      0.028       0.023       0.418
ma.L1         -0.4140      0.122     -3.397      0.001      -0.653      -0.175
ma.L2         -0.2716      0.105     -2.586      0.010      -0.477      -0.066
sigma2         0.0151   7.32e-05    206.769      0.000       0.015       0.015
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):             94606.67
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):            2769.02   Skew:                             0.20
Prob(H) (two-sided):                  0.00   Kurtosis:                        17.64
===================================================================================

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

Figura 10. Resultados de la estimación de los parámetros para el modelo ARIMA.

Diagnóstico:

Se obtuvieron los residuos del modelo ajustado y posteriormente se visualizaron los residuos por medio de un Gráfico de residuos vs tiempo. Finalmente se buscó aleatoriedad, falta de patrones y normalidad.

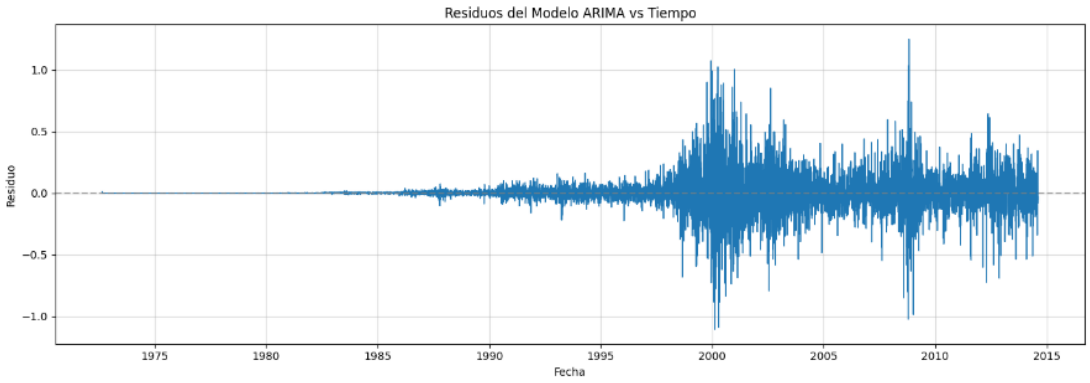

Figura 11. Gráfico de Residuos vs Tiempo.

También se realizó un gráfico ACF de residuos y QQ-plot con el mismo objetivo de buscar aleatoriedad, falta de patrones y normalidad.

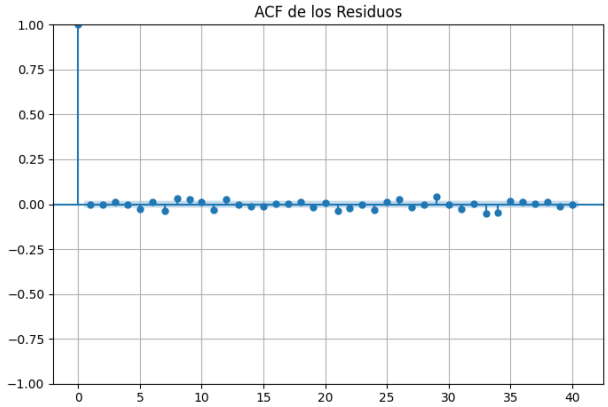

Figura 12. Gráfico de la función ACF de los Residuos.

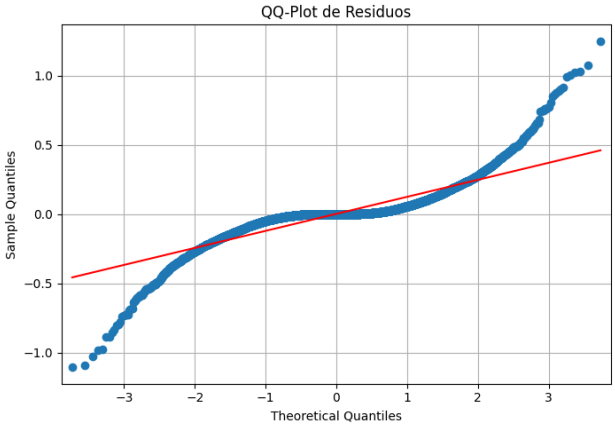

Figura 13. Gráfico de la función QQ-plot de los Residuos.

Finalmente, se realizó la prueba de Ljung-Box sobre los residuos.

```
5.3.3 Prueba de Ljung-Box para autocorrelación de residuos...
      lb_stat     lb_pvalue
1    0.012677  9.103534e-01
2    0.012792  9.936245e-01
3    1.143827  7.665059e-01
4    1.177824  8.817360e-01
5    8.650158  1.238585e-01
6    9.768999  1.347231e-01
7   22.618119  1.986110e-03
8   32.933718  6.330502e-05
9   40.563314  6.005350e-06
10  42.802408  5.394261e-06
11  53.104874  1.718126e-07
12  61.999908  9.726543e-09
13  62.157765  2.153775e-08
14  63.779626  2.523369e-08
15  65.450260  2.848833e-08
16  65.593255  5.822180e-08
17  65.863094  1.097931e-07
18  67.891887  1.023944e-07
19  70.409349  7.863141e-08
20  71.025372  1.237671e-07
21  86.190777  7.268145e-10
22  90.469028  2.851542e-10
23  90.469416  5.928417e-10
24 100.682183  2.297817e-11
25 102.657421  2.230254e-11
26 109.965248  2.679113e-12
27 113.558956  1.382353e-12
28 113.667143  2.773987e-12
29 131.057384  6.276725e-15
30 131.148378  1.310595e-14
31 138.896979  1.311881e-15
32 138.898807  2.822371e-15
33 167.364912  6.767099e-20
34 189.649513  1.612257e-23
35 193.547843  7.657487e-24
36 195.013031  9.955411e-24
37 195.049395  2.312694e-23
38 196.511962  2.956619e-23
39 198.085211  3.587447e-23
40 198.101956  8.131542e-23

Resultado Ljung-Box: ¡Advertencia! Se rechaza H0 para al menos un lag (p<0.05).
Hay evidencia de autocorrelación significativa en los residuos.
El modelo NO captura toda la estructura de autocorrelación. Considera:
  - Revisar los órdenes (p, q).
  - Incluir términos estacionales (SARIMA) si la descomposición lo sugiere.
  - Buscar variables exógenas omitidas (ARIMAX).

--- Fin del Diagnóstico ---
El diagnóstico sugiere problemas con el modelo. Las predicciones deben tomarse con precaución.
```

Figura 14. Resultado del diagnóstico Ljung-Box.

Se ajustó exitosamente el modelo ARIMA(2, 1, 2) a los datos de entrenamiento. Sin embargo, el diagnóstico de los residuos indicó problemas de autocorrelación residual. Por lo que el modelo podría no ser adecuado.

Conclusiones por integrante.

Aguilar Ramírez Carlos Francisco

La aplicación de la metodología Box-Jenkins para modelar la serie de tiempo de los precios de las acciones de Walmart (WMT) ha sido un ejercicio revelador sobre el proceso iterativo del modelado predictivo. Me permitió aplicar paso a paso las etapas clave: desde la verificación de estacionariedad con la prueba ADF, pasando por la diferenciación necesaria para estabilizar la serie, hasta el análisis de las funciones ACF y PACF para identificar los órdenes p y q del modelo ARIMA. Cada paso ofreció una comprensión más profunda de la estructura temporal de los datos financieros.

Sin embargo, el aprendizaje más significativo provino de la fase de diagnóstico. Aunque el modelo ARIMA(2,1,2) se ajustó técnicamente a los datos de entrenamiento, la prueba de Ljung-Box sobre los residuos indicó la presencia de autocorrelación significativa. Esto subrayó la importancia crítica de la validación: un modelo puede ajustarse, pero no ser estadísticamente adecuado si los residuos no se comportan como ruido blanco. Este resultado me enseñó que el modelado no es un proceso lineal, sino uno que a menudo requiere volver a etapas anteriores, reevaluar supuestos y considerar modelos alternativos o más complejos (como SARIMA o ARIMAX) para capturar toda la dinámica de la serie.

En conclusión, esta práctica consolidó mi entendimiento técnico de los modelos ARIMA y la metodología Box-Jenkins, pero, sobre todo, reforzó la idea de que el diagnóstico riguroso es fundamental para construir modelos confiables y que la persistencia en la mejora del modelo es clave en el análisis de series de tiempo.

Arista Romero Juan Ismael

La construcción de un modelo ARIMA para la serie de precios de Walmart mediante la metodología Box-Jenkins me permitió comprender en la práctica cómo abordar series de tiempo no estacionarias, un desafío común en datos económicos y financieros. Fue particularmente instructivo aplicar la prueba ADF para confirmar formalmente la no estacionariedad observada visualmente y luego usar la diferenciación para transformar los datos en una serie adecuada para el modelado ARIMA.

El análisis de las funciones ACF y PACF resultó crucial para hipotetizar la estructura del modelo, llevándonos a proponer órdenes específicos para los componentes AR y MA (ARIMA(2,1,2)). La fase de estimación nos proporcionó los parámetros del modelo, pero fue la etapa de diagnóstico la que resaltó una lección importante: la validación a través del análisis de residuos, especialmente con la prueba de Ljung-Box, es indispensable. El hecho de que nuestros residuos mostraran autocorrelación evidenció que el modelo inicial, aunque ajustado, no capturaba completamente la dependencia temporal de la serie.

Este ejercicio reforzó la importancia de seguir una metodología estructurada como la de Box-Jenkins y la necesidad de interpretar críticamente los resultados de las pruebas estadísticas en cada etapa. Evidenció cómo las decisiones de modelado deben ser guiadas tanto por la teoría como por la evidencia empírica obtenida de los propios datos y los diagnósticos del modelo.

Jiménez Flores Luis Arturo

El análisis de la serie temporal fue del precio de cierre ajustado de Walmart (WMT) desde 1972 hasta 2025, utilizamos la metodología Box-Jenkins para construir un modelo ARIMA. Donde en la parte de "identificación" de la metodología Box-Jenkins la serie original mostró ser no estacionaria, requiriendo ajustar un valor "d" de diferencia(s) para alcanzar la estacionariedad. El análisis ACF/PACF de la serie diferenciada sugirió un modelo inicial ARIMA(2, [d], 2).
Por otra parte, en los segmentos de estimación y diagnóstico se ajustó el modelo a los datos de entrenamiento de la siguiente manera ARIMA(2, 1, 2). El diagnóstico de los residuos indicó problemas de autocorrelación residual, sugiriendo que el modelo podría no ser completamente adecuado.
Para finalizar, identifiqué que el modelo presentado resulta en limitaciones evidenciadas por el diagnóstico de residuos por lo que algunas futuras mejoras que se podrían incluir son la exploración de diferentes órdenes (p, q) así como la consideración de modelos estacionales (SARIMA) si la estacionalidad residual es significativa o la inclusión de variables exógenas (ARIMAX) si se identifican factores externos relevantes.

Vázquez Martin Marlene Gabriela

A través de la aplicación de la metodología Box-Jenkins a los datos históricos de precios de Walmart, adquirí una comprensión práctica y profunda del proceso de modelado ARIMA. Fue valioso experimentar cada fase: comenzar con la identificación de la no estacionariedad mediante visualización y la prueba ADF, aplicar la diferenciación para transformar la serie, y luego utilizar las herramientas gráficas ACF y PACF para proponer una estructura inicial para el modelo ARIMA(2,1,2).
La fase de diagnóstico fue especialmente esclarecedora. Aunque logramos ajustar el modelo, el análisis de residuos, particularmente los resultados de la prueba de Ljung-Box, reveló que quedaba estructura de autocorrelación sin capturar. Esta experiencia me demostró que el ajuste de un modelo es solo una parte del proceso; la validación rigurosa de sus supuestos, como la independencia de los residuos, es igualmente crucial. Me permitió ver cómo las pruebas estadísticas guían las decisiones sobre si un modelo es adecuado o si necesita ser refinado.
Más allá del conocimiento técnico sobre ARIMA, esta práctica reforzó la importancia de la paciencia y la iteración en la ciencia de datos. Enfrentar datos reales, como los precios de acciones, implica desafíos y resultados no siempre perfectos al primer intento. Aprender a interpretar los diagnósticos y usarlos para guiar los siguientes pasos, como considerar modelos SARIMA o diferentes órdenes p y q, fue una lección práctica fundamental para futuros proyectos de análisis de series temporales.