Instituto Politécnico Nacional
Carrera de Licenciatura en Ciencia de Datos
"La técnica al Servicio de la Patria".

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 6

Inferencia de parámetros en modelos ARIMA

Instrucciones Practica 6

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 la **Inferencia de parámetros en modelos ARIMA**
    a. Aplica la metodología BOX-JENKINS.
7. Conclusiones por integrante.

Introducción

En esta práctica realizaremos un análisis de series de tiempo aplicado a datos de temperaturas mínimas diarias en Australia (Melbourne). El objetivo es inferir los parámetros de un modelo ARIMA utilizando la metodología Box-Jenkins. Se utilizan diversas librerías de Python como pandas, numpy, matplotlib y statsmodels para llevar a cabo el análisis.

Problemática

La problemática abordada es modelar y potencialmente predecir el comportamiento de las temperaturas mínimas diarias. Las series de tiempo de temperatura suelen presentar características como estacionalidad (ciclos anuales) y autocorrelación, lo que requiere técnicas específicas para su análisis. El desafío es identificar un modelo estadístico (ARIMA en este caso) que capture adecuadamente la estructura temporal de los datos para realizar inferencias o predicciones válidas.

Modelo estadístico

El modelo estadístico central utilizado en esta práctica es el *AutoRegressive Integrated Moving Average* (ARIMA). Un modelo ARIMA se define por tres órdenes: (p, d, q).

- p: Orden de la parte Autorregresiva (AR), que modela la dependencia de un valor actual con valores pasados.
- d: Orden de la Diferenciación (Integrated - I), que indica cuántas veces se debe diferenciar la serie para hacerla estacionaria.

- q: Orden de la parte de Medias Móviles (MA), que modela la dependencia de un valor actual con errores de predicción pasados.

La selección de estos órdenes (p, d, q) es el objetivo principal de la inferencia de parámetros mediante la metodología Box-Jenkins.

Modelo computacional

El modelo computacional se implementa en Python utilizando un conjunto de librerías especializadas:

Pandas: Para la manipulación y carga de datos (DataFrames y Series).

NumPy: Para operaciones numéricas.

Matplotlib y Seaborn: Para la visualización de datos (gráficos de series, ACF/PACF, residuos, etc.).

Statsmodels: Librería clave que proporciona las herramientas para el análisis de series de tiempo, incluyendo:

- Prueba de Dickey-Fuller Aumentada (adfuller) para estacionariedad.
- Funciones para graficar ACF y PACF (plot_acf, plot_pacf).
- Implementación del modelo ARIMA (ARIMA).
- Descomposición estacional (seasonal_decompose).
- Pruebas de diagnóstico de residuos (ej., Ljung-Box a través de acorr_ljungbox).

Scipy: Utilizada para encontrar picos y valles (find_peaks) en el EDA.

Sklearn: Utilizada para métricas de evaluación como el Error Cuadrático Medio (mean_squared_error).

El código está estructurado en funciones para mejorar la modularidad y reutilización, cubriendo desde la carga de datos hasta la evaluación del modelo.

Metodología.

La metodología general empleada en el documento sigue un enfoque estructurado para el análisis de series de tiempo:

1. **Carga y Preparación de Datos:** Lectura del archivo CSV, conversión de tipos de datos (fechas, numéricos), manejo de valores faltantes, ordenación y renombrado de columnas.

```
--- 2. Cargando y Preparando Datos desde: /content/daily-minimum-temperatures-in-me.csv ---
Datos cargados exitosamente.
Se encontraron 3 NaNs en 'temp_min'. Rellenando con interpolación lineal...

Inspección inicial del DataFrame procesado:
Primeras 5 filas:
| fecha               | temp_min   |
|:--------------------|:-----------|
| 1981-01-01 00:00:00 | 20.7       |
| 1981-01-02 00:00:00 | 17.9       |
| 1981-01-03 00:00:00 | 18.8       |
| 1981-01-04 00:00:00 | 14.6       |
| 1981-01-05 00:00:00 | 15.8       |

Información general:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3650 entries, 0 to 3649
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   fecha     3650 non-null   datetime64[ns]
 1   temp_min  3650 non-null   float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 57.2 KB

Estadísticas descriptivas:
|       | temp_min   |
|:------|:-----------|
| count | 3650       |
| mean  | 11.1793    |
| std   | 4.06813    |
| min   | 0          |
| 25%   | 8.3        |
| 50%   | 11         |
| 75%   | 14         |
| max   | 26.3       |
```

Figura 1. Carga y preparación de los datos

```
--- 3. Preparando la Serie de Tiempo Final ('temp_min') ---

Serie 'Temperatura Mínima Diaria' preparada. Longitud: 3650 puntos.
Fechas desde 1981-01-01 hasta 1990-12-31
Últimas 5 observaciones:
| fecha               | Temperatura Mínima Diaria   |
|:--------------------|:----------------------------|
| 1990-12-27 00:00:00 | 14                          |
| 1990-12-28 00:00:00 | 13.6                        |
| 1990-12-29 00:00:00 | 13.5                        |
| 1990-12-30 00:00:00 | 15.7                        |
| 1990-12-31 00:00:00 | 13                          |

Serie original creada:
fecha
1981-01-01    20.7
1981-01-02    17.9
1981-01-03    18.8
1981-01-04    14.6
1981-01-05    15.8
Name: Temperatura Mínima Diaria, dtype: float64
```

Figura 2. Preparación de la Serie de Tiempo

2. **Análisis Exploratorio de Datos (EDA):** Visualización de la serie temporal, cálculo de estadísticas descriptivas, análisis de tendencia (medias móviles), descomposición estacional, identificación de picos/valles y análisis de cambios anuales.

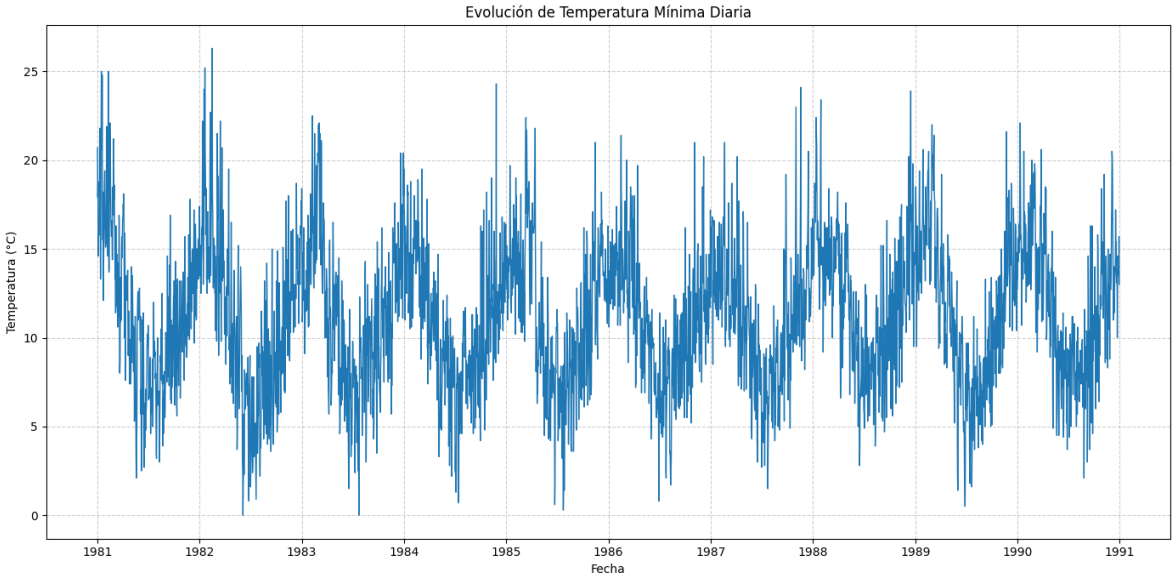

Figura 3. Análisis exploratorio de la evolución de la temperatura.

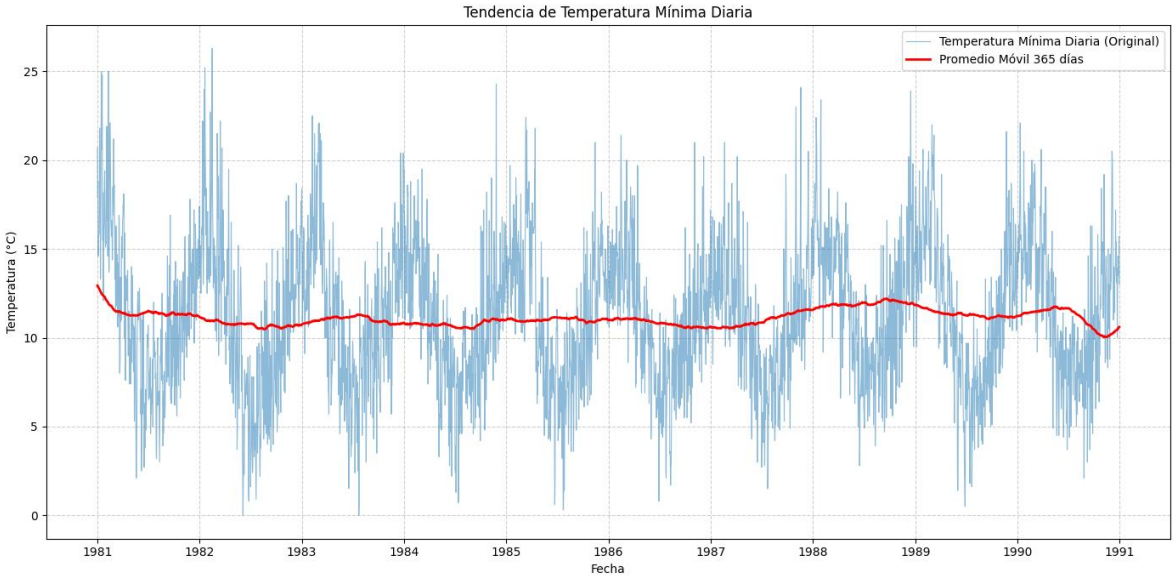

Figura 4. Análisis exploratorio de la evolución de la temperatura Promedio móvil.

3. **Modelado ARIMA (Metodología Box-Jenkins):** Aplicación de los pasos de identificación, estimación, validación y uso del modelo.
4. **Conclusiones y Recomendaciones:** Resumen de los hallazgos y sugerencias para futuros análisis (como el uso de SARIMA).

Propuesta de solución para la Inferencia

La propuesta de solución implementada en el documento consiste en aplicar rigurosamente la Metodología Box-Jenkins para inferir los parámetros (p, d, q) del modelo ARIMA más adecuado para la serie de temperaturas mínimas diarias.

Aplica la metodología BOX-JENKINS

Esta metodología se aplica de la siguiente manera:

Identificación del Modelo:

Se realiza la prueba de Dickey-Fuller Aumentada (ADF) sobre la serie original. Si la serie no es estacionaria (p-value > 0.05), se aplica diferenciación (d=1, d=2, ...) y se repite la prueba ADF hasta lograr estacionariedad. En este caso particular, la serie original resultó ser estacionaria según la prueba ADF (p-value ≈ 0.0003), por lo que se determinó d=0.

```
--- 5. Metodología Box-Jenkins (ARIMA) - Identificación ---

5.1.1 Resultados de la Prueba Dickey-Fuller Aumentada para Temperatura Mínima Diaria:
|                           | 0           |
|:--------------------------|:------------|
| Test Statistic            | -4.44052    |
| p-value                   | 0.000251472 |
| #Lags Used                | 20          |
| Number of Observations Used | 3629      |
| Critical Value (1%)       | -3.43215    |
| Critical Value (5%)       | -2.86234    |
| Critical Value (10%)      | -2.56719    |
  Conclusión: La serie es probablemente estacionaria (p-value=0.0003).

La serie original 'Temperatura Mínima Diaria' ES estacionaria (d=0).
```

Figura 5. Parámetros identificados de los datos.

Se grafican la Función de Autocorrelación (ACF) y la Función de Autocorrelación Parcial (PACF) de la serie (posiblemente diferenciada, aquí d=0). Se analizan estos gráficos para obtener una estimación inicial de los órdenes p (basado en dónde corta la PACF) y q (basado en dónde corta la ACF). Para este dataset, se seleccionaron inicialmente p=3 y q=1.
Órdenes seleccionados (p,d,q): (3, 0, 1)

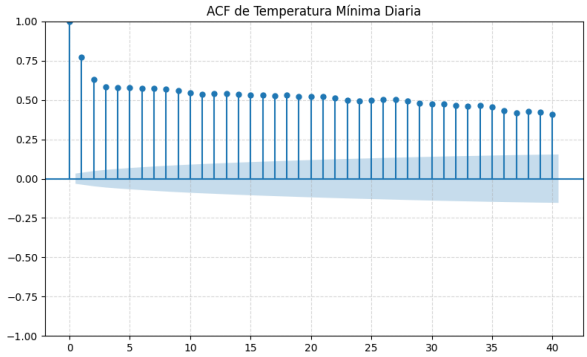

Figura 6. ACF de Temperaturas.

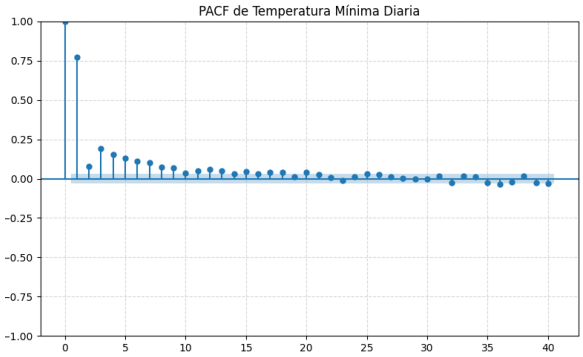

Figura 7. PACF de Temperaturas.

Estimación de Parámetros:

Se divide la serie en conjuntos de entrenamiento (90%) y prueba (10%). Posteriormente se ajusta el modelo ARIMA con los órdenes identificados (ARIMA(3, 0, 1)) utilizando los datos de entrenamiento. La librería statsmodels calcula los coeficientes del modelo (constante, términos AR, términos MA) y sus errores estándar, junto con métricas como Log Likelihood, AIC, BIC.

```
--- 5. Metodología Box-Jenkins (ARIMA) - Estimación ---

--- División de Datos ---
Tamaño Total: 3650, Ratio Entrenamiento: 90.0%
Tamaño Entrenamiento: 3285 (1981-01-01 a 1989-12-31)
Tamaño Prueba:      365 (1990-01-01 a 1990-12-31)

--- 5.2 Ajustando modelo ARIMA(3, 0, 1) ---

Resumen del Modelo Ajustado:
                               SARIMAX Results
==========================================================================================
Dep. Variable:     Temperatura Mínima Diaria   No. Observations:                 3285
Model:                        ARIMA(3, 0, 1)   Log Likelihood               -7563.961
Date:                       Sat, 26 Apr 2025   AIC                          15139.923
Time:                               07:27:12   BIC                          15176.506
Sample:                                    0   HQIC                         15153.021
                                        - 3285
Covariance Type:                          opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.4361      0.838     13.642      0.000       9.793      13.079
ar.L1          1.4827      0.019     76.140      0.000       1.445       1.521
ar.L2         -0.6131      0.028    -21.873      0.000      -0.668      -0.558
ar.L3          0.1254      0.019      6.716      0.000       0.089       0.162
ma.L1         -0.8931      0.012    -72.111      0.000      -0.917      -0.869
sigma2         5.8516      0.136     43.058      0.000       5.585       6.118
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                16.29
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.85   Skew:                             0.09
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.29
===================================================================================
```

Figura 8. Parámetros estimados para el modelo ARIMA.

Validación (Diagnóstico del Modelo):

Se analizan los residuos del modelo ajustado para verificar si se comportan como "ruido blanco" (es decir, si son aleatorios, sin autocorrelación y preferiblemente normales).

Gráfico de Residuos vs Tiempo: Se busca que no haya patrones evidentes y que la varianza sea constante.

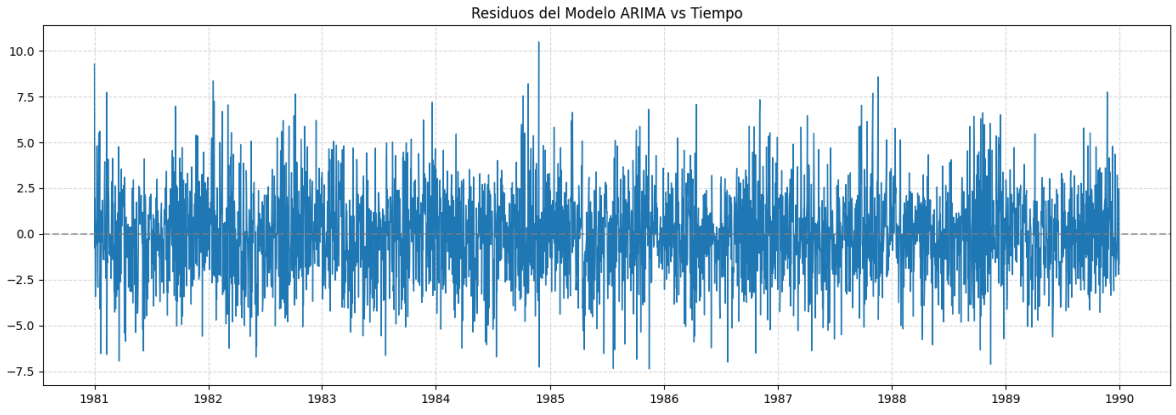

Figura 9. Gráfico del Modelo ARIMA a través del Tiempo.

ACF/PACF de Residuos: Se espera que ninguna autocorrelación sea significativa (los lags deben caer dentro de las bandas de confianza).

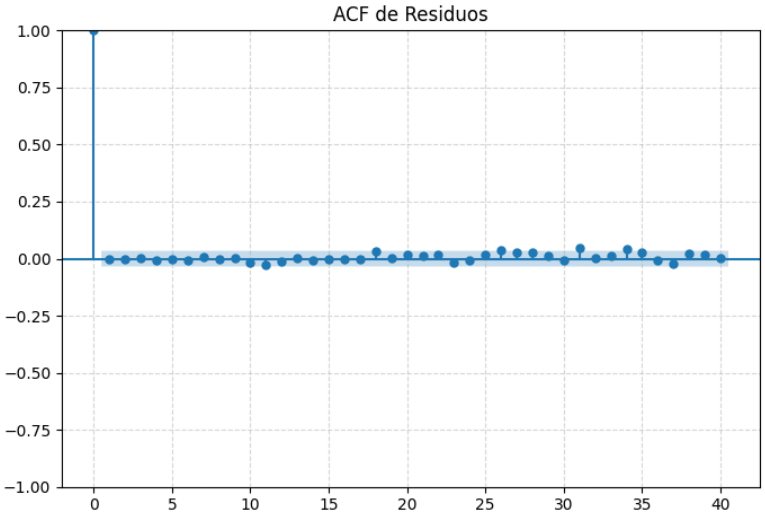

Figura 10. ACF de Residuos.

QQ-Plot: Se busca que los puntos sigan la línea diagonal, indicando normalidad (aunque esto es menos crítico que la independencia).

Figura 11. QQ-Plot de Residuos.

Prueba de Ljung-Box: Prueba formal para la autocorrelación residual. Se espera que los p-values sean mayores que 0.05 para todos los lags probados, indicando ausencia de autocorrelación. En este caso, la prueba Ljung-Box indicó que los residuos parecían independientes, llevando a un diagnóstico "Satisfactorio" para el modelo ARIMA(3,0,1) ajustado.

```
5.3.3 Prueba de Ljung-Box para autocorrelación de residuos...
|    | lb_stat    | lb_pvalue   |
|:---|:-----------|:------------|
| 1  | 0.00523084 | 0.942344    |
| 2  | 0.0194077  | 0.990343    |
| 3  | 0.0492275  | 0.997138    |
| 4  | 0.245835   | 0.993037    |
| 5  | 0.297012   | 0.997699    |
| 6  | 0.430225   | 0.998587    |
| 7  | 0.645321   | 0.998722    |
| 8  | 0.648926   | 0.999643    |
| 9  | 0.731433   | 0.999847    |
| 10 | 1.84348    | 0.997405    |
| 11 | 3.99754    | 0.969989    |
| 12 | 4.38287    | 0.975496    |
| 13 | 4.39742    | 0.986183    |
| 14 | 4.62991    | 0.990326    |
| 15 | 4.6364     | 0.994783    |
| 16 | 4.63659    | 0.997285    |
| 17 | 4.6371     | 0.998624    |
| 18 | 7.95664    | 0.979275    |
| 19 | 7.98332    | 0.986838    |
| 20 | 9.28837    | 0.97932     |
| 21 | 9.82576    | 0.981087    |
| 22 | 10.6229    | 0.979724    |
| 23 | 11.4004    | 0.978719    |
| 24 | 11.531     | 0.984694    |
| 25 | 12.524     | 0.981818    |
| 26 | 16.5617    | 0.921728    |
| 27 | 19.1522    | 0.86439     |
| 28 | 21.3634    | 0.809875    |
| 29 | 21.838     | 0.826836    |
| 30 | 22.1057    | 0.850171    |
| 31 | 29.8801    | 0.523494    |
| 32 | 29.8828    | 0.574093    |
| 33 | 30.3738    | 0.598525    |
| 34 | 36.8778    | 0.337242    |
| 35 | 39.7142    | 0.267989    |
| 36 | 39.7965    | 0.304817    |
| 37 | 41.4298    | 0.283465    |
| 38 | 43.1907    | 0.259135    |
| 39 | 44.2681    | 0.259031    |
| 40 | 44.3159    | 0.29454     |

Resultado Ljung-Box: Residuos parecen independientes (ruido blanco).

Resultado del Diagnóstico: Satisfactorio
```

Figura 12. Resultados obtenidos de la Prueba de Ljung-Box.

Uso del Modelo (Predicción):

Se utiliza el modelo ajustado para generar predicciones sobre el conjunto de prueba.
Se visualizan las predicciones junto con los datos reales y los intervalos de confianza.
Se evalúa numéricamente el rendimiento del modelo usando métricas como RMSE, MAE y MAPE sobre el conjunto de prueba. Para el modelo ARIMA(3,0,1) se obtuvieron RMSE=3.6937, MAE=3.0705 y MAPE=34.35%.

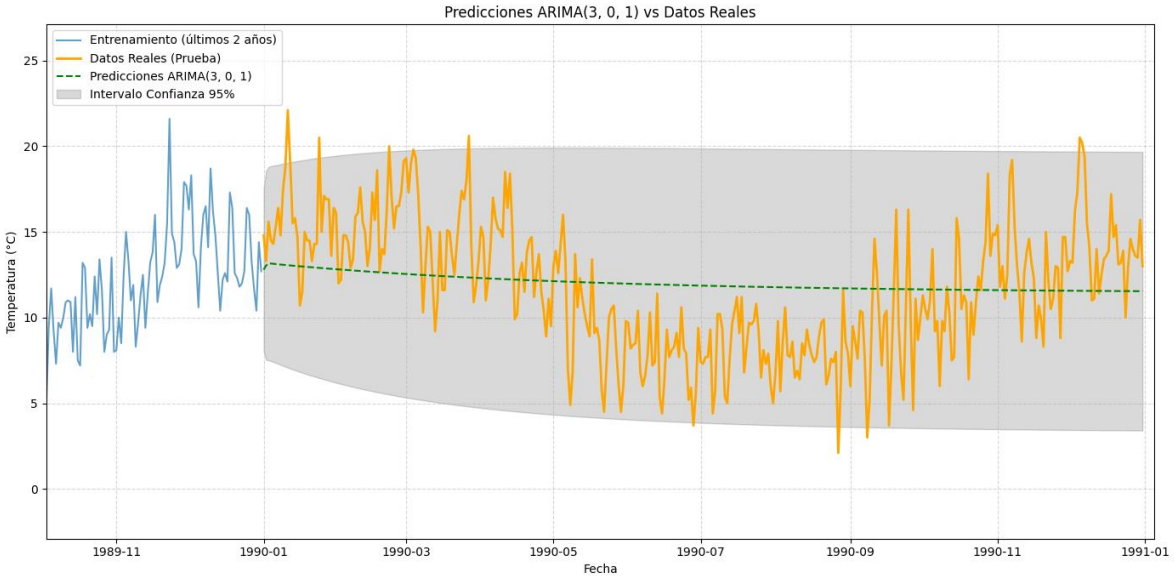

Figura 13. Predicción realizada del Modelo ARIMA a los datos de Temperatura.

Conclusiones por integrante

Aguilar Ramírez Carlos Francisco

La aplicación de la metodología Box-Jenkins para modelar la serie de tiempo de las temperaturas mínimas diarias ha sido un ejercicio revelador sobre el proceso de modelado predictivo. Me permitió aplicar paso a paso las etapas clave: desde la verificación de estacionariedad con la prueba ADF, que sorpresivamente indicó que la serie original era estacionaria (d=0), hasta el análisis de las funciones ACF y PACF para identificar los órdenes p y q del modelo ARIMA, sugiriendo un ARIMA(3,0,1) inicial. Cada paso ofreció una comprensión más profunda de la estructura temporal de los datos climáticos.

El aprendizaje más significativo provino de la fase de diagnóstico. Aunque el modelo ARIMA(3,0,1) se ajustó técnicamente y la prueba de Ljung-Box sobre los residuos indicó que estos parecían independientes (ruido blanco), el análisis exploratorio previo y la visualización de las predicciones mostraron claramente que el modelo simple no capturaba la fuerte estacionalidad anual inherente a los datos. Esto subrayó la importancia crítica de no solo validar estadísticamente los residuos, sino también de evaluar si el modelo captura los patrones dominantes observados en los datos. Este resultado me enseñó que un diagnóstico estadístico "satisfactorio" no siempre implica un modelo útil para la predicción si ignora características clave como la estacionalidad, y que a menudo se requieren modelos más complejos (como SARIMA).

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 debe

considerar tanto las pruebas estadísticas como la capacidad del modelo para replicar los patrones visuales de la serie, y que la elección del modelo adecuado (en este caso, probablemente SARIMA) es crucial para el análisis de series de tiempo con componentes estacionales fuertes.

Arista Romero Juan Ismael

La construcción de un modelo ARIMA para la serie de temperaturas mínimas diarias mediante la metodología Box-Jenkins me permitió comprender en la práctica cómo abordar series de tiempo, aunque en este caso, el desafío fue diferente al esperado. Fue particularmente instructivo aplicar la prueba ADF y encontrar que la serie original ya era estacionaria (d=0), contrario a lo que a menudo ocurre con datos económicos.

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(3,0,1)). 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 (gráficos y prueba de Ljung-Box), aunque resultó estadísticamente satisfactoria indicando residuos como ruido blanco, no fue suficiente por sí sola. La fuerte estacionalidad vista en el EDA y en la pobre capacidad predictiva del modelo evidenció que el modelo simple, aunque estadísticamente "válido" en sus residuos, no capturaba la dinámica esencial 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 no solo de las pruebas estadísticas sino también del comportamiento predictivo del modelo en el contexto de los patrones observados en los datos. Evidenció cómo las decisiones de modelado deben ser guiadas tanto por los diagnósticos formales como por la evaluación cualitativa de si el modelo refleja la realidad subyacente, llevando a la recomendación de usar SARIMA para estos datos.

Jiménez Flores Luis Arturo

En esta práctica realizamos un análisis sobre una serie de tiempo que registra las temperaturas mínimas diarias en Melbourne ubicado en Australia.

Usamos la metodología Box-Jenkins para evaluar un modelo ARIMA. La fase inicial de Identificación reveló, mediante la función realizar_prueba_adf, que la serie original presentaba características de estacionariedad (p-value ≈ 0.0003), determinándose un orden de diferenciación d=0 a través de la función identificar_orden_diferenciacion. El análisis posterior de las gráficas ACF y PACF, generadas por visualizar_acf_pacf, sugirió una estructura autorregresiva y de medias móviles que llevó a la proposición de un modelo inicial ARIMA(3,0,1).

Posteriormente, en la etapa de Estimación, se procedió a dividir los datos utilizando dividir_datos_entrenamiento_prueba (90% entrenamiento, 10% prueba) y se ajustó el

modelo ARIMA(3,0,1) propuesto mediante la función ajustar_modelo_arima. El resumen del modelo proporcionado por statsmodels detalló los coeficientes estimados y las métricas de ajuste.

La fase de Validación (Diagnóstico), ejecutada por la función diagnosticar_modelo, incluyó el análisis gráfico de los residuos (vs. tiempo, ACF, QQ-plot) y la prueba de Ljung-Box (sm.stats.acorr_ljungbox). Esta última arrojó un p-values consistentemente superiores a 0.05 para los *lags* evaluados, indicando que los residuos se comportaban estadísticamente como ruido blanco (independientes), lo que condujo a calificar el diagnóstico formal como "Satisfactorio".

No obstante, la evaluación del Uso del Modelo en la predicción sobre el conjunto de prueba continuó presentado limitaciones a pesar del diagnóstico estadístico favorable de los residuos, también la visualización generada por visualizar_prediccion mostró una incapacidad manifiesta del modelo ARIMA(3,0,1) para replicar la pronunciada estacionalidad anual observada durante el EDA (evidente en descomponer_serie_estacional). Las predicciones tendían a aplanarse, desviándose significativamente de los ciclos reales de temperatura. Las métricas cuantitativas obtenidas mediante evaluar_prediccion (RMSE=3.6937, MAE=3.0705, MAPE=34.35%) confirmaron este desempeño predictivo deficiente.

Por lo tanto, este análisis indica que las pruebas diagnósticas sobre los residuos del ARIMA(3,0,1) fueron satisfactorias, indicando corrección estadística del modelo ajustado, sin embargo, su incapacidad para capturar la fuerte componente estacional lo vuelve inadecuado para fines predictivos en esta serie. Se concluye que un modelo ARIMA simple es insuficiente.

Vázquez Martin Marlene Gabriela

A través de la aplicación de la metodología Box-Jenkins a los datos de temperaturas mínimas diarias, adquirí una comprensión práctica del proceso de modelado ARIMA. Fue valioso experimentar cada fase: comenzar con la identificación de la estacionariedad, donde la prueba ADF indicó que la serie original ya lo era (d=0), y luego utilizar las herramientas gráficas ACF y PACF para proponer una estructura inicial para el modelo ARIMA(3,0,1).

La fase de diagnóstico fue especialmente esclarecedora. Logramos ajustar el modelo y el análisis de residuos, particularmente los resultados de la prueba de Ljung-Box, indicó que estos se comportaban como ruido blanco. Sin embargo, esta experiencia me demostró que un diagnóstico estadísticamente "satisfactorio" no garantiza un modelo útil si esta falla en capturar patrones dominantes evidentes en los datos, como la fuerte estacionalidad anual en este caso. Ver las predicciones comparadas con los datos reales dejó claro que el modelo simple era insuficiente.

Más allá del conocimiento técnico sobre ARIMA, esta práctica reforzó la importancia de una evaluación integral del modelo. Enfrentar datos reales implica desafíos, y aprendí que

interpretar los diagnósticos formales junto con la capacidad predictiva visual del modelo es fundamental. Reconocer las limitaciones del ARIMA simple y entender por qué se necesita un modelo más avanzado como SARIMA para estos datos fue una lección práctica clave para futuros análisis de series temporales.

Fuentes y contenido relacionado