INSTITUTO POLITECNICO

NACIONAL

Escuela Superior de Cómputo

(ESCOM)

Licenciatura en Ciencias de Datos.

Nombre de la unidad de aprendizaje:

Análisis de Series de Tiempo.

Grupo: 6AV1.

Nombre de la Actividad:

“Práctica 5 construir Modelo ARIMA”.

Nombre del alumno(a):

Arteaga Gonzalez Edwin Yahir.

Juarez Gaona Erick Rafael.

Rico Gaytan Diana Andrea.

Ruiz Merino Wendy Ivonne.

Fecha:

13/04/2025.

1. Introducción.

El Índice de Calidad del Aire (AQI, por sus siglas en inglés) es un indicador crítico para evaluar la salud ambiental de una región, sintetizando en una única métrica la concentración de contaminantes clave como material particulado (PM2.5 y PM10), dióxido de nitrógeno (NO_2), ozono (O_3) y monóxido de carbono (CO). En contextos urbanos con alta densidad poblacional y actividad industrial acelerada, como Bengaluru (India), monitorear y predecir este índice se convierte en una herramienta esencial para la gestión pública, la salud comunitaria y la sostenibilidad ambiental.

Bengaluru, conocida como la "Capital Tecnológica de la India", ha experimentado un crecimiento exponencial en las últimas dos décadas, impulsado por su ecosistema de startups, parques industriales y una población que supera los 12 millones de habitantes. Este desarrollo, sin embargo, ha generado desafíos ambientales significativos: un parque vehicular en aumento (más de 8 millones de vehículos registrados en 2020), expansión de obras de infraestructura y emisiones industriales no reguladas. Como resultado, la ciudad ha enfrentado episodios recurrentes de contaminación atmosférica, con AQI frecuentemente en el rango "pobre" (201-300) durante los meses de invierno, según datos del Central Pollution Control Board (CPCB) de India.

Este estudio se centra en el período comprendido entre marzo de 2015 y julio de 2020, un intervalo que captura transformaciones estructurales en la ciudad, incluyendo la implementación de la fase inicial del Sistema de Transporte Rápido (BMTC) y eventos exógenos como la reducción temporal de emisiones durante los confinamientos por COVID-19 en 2020.

2. Problemática

Bengaluru, conocida como el Silicon Valley de la India, ha experimentado un crecimiento rápido en las últimas décadas. Esto ha incrementado el tráfico vehicular y la actividad industrial, afectando la calidad del aire. Comprender la evolución del AQI puede apoyar decisiones de mitigación, alertas tempranas y planificación sostenible.

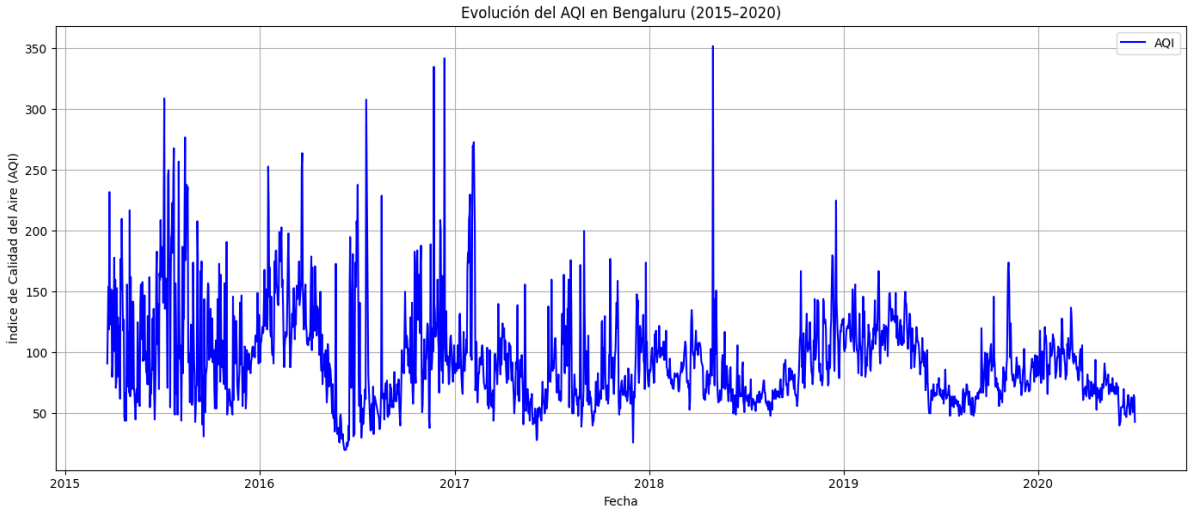


Figura 1. Evolución del AQI en Bengaluru (2015-2020).

3. Metodología Box-Jenkins

1. Obtención de los datos

Los datos fueron obtenidos de la pagina kaggle

2.Preprocesamiento de datos

Se eliminaron los valores nulos y se convirtió fecha en la variable data time y se segmentaron los datos para la ciudad de Bengaluru

3. Identificación del Modelo

a. Prueba ADF

La prueba Augmented Dickey-Fuller (ADF) es un test estadístico utilizado para determinar si una serie temporal tiene una raíz unitaria (es decir, si es no estacionaria). La estacionariedad es un requisito clave para modelos como ARIMA, ya que garantiza que las propiedades estadísticas de la serie (media, varianza) no cambien con el tiempo.

Los valores obtenidos son:

- ADF Statistic = -6.077
- p-valor = 1.11e-07

Hipótesis de la Prueba

- Hipótesis nula (H0): La serie tiene una raíz unitaria (es no estacionaria).
- Hipótesis alternativa (H1): La serie no tiene una raíz unitaria **(es estacionaria).**

Regla de Decisión

- Si el p-valor < 0.05 (nivel de significancia del 5%), se rechazaH0, concluyendo que la serie es estacionaria.
- Si e p-valor ≥ 0.05, no se rechaza H0, asumiendo que la serie es no estacionaria.

Resultado Obtenido

- p-valor = 1.11e-07(0.000000111), que es mucho menor que 0.05.
- Conclusión: Se rechazaH0. La serie temporal NO tiene una raíz unitariay, por lo tanto, es estacionaria.

Análisis del Estadístico ADF

- El valor crítico del estadístico ADF para un nivel de significancia del 5% suele ser aproximadamente -2.86 (varía según el tamaño de la muestra).
- Tu resultado: ADF Statistic = -6.077, que es más negativo que -2.86.
- Interpretación: Cuanto más negativo sea el estadístico ADF, mayor es la evidencia contra H0. En este caso, -6.077 es una señal fuerte de estacionariedad.

b. Gráfica de la ACF

Eje horizontal (eje de las "lag" o retrasos)

- Se extiende aproximadamente desde lag 0 hasta lag 40. Cada punto en la gráfica indica la correlación de la serie de tiempo con ella misma en un retraso específico.

Eje vertical (valor de la autocorrelación)

- Va desde valores negativos (por debajo de -0.25) hasta valores positivos (por encima de 0.75, en la parte superior).
- La línea de correlación 0 (marcada en el eje horizontal) sirve como referencia para evaluar si hay autocorrelación positiva o negativa en cada retraso.

Barras o puntos que representan la autocorrelación para cada lag

- Se ve un pico muy pronunciado y positivo en lag 0 (que suele ser igual a 1 porque es la autocorrelación con la propia serie), aunque a veces se omite o se muestra muy arriba.
- El primer retraso (lag 1) parece ser fuertemente positivo (por encima de 0.7) o cercano a ese valor.
- A partir de lag 2 y en adelante, las barras van tomando valores negativos y luego se acercan a 0, lo cual indica que la correlación de la serie en lags superiores (2, 3, etc.) es menor o casi nula.

- Algunas barras (puntos) caen por debajo de la línea de significancia (las bandas azules que se marcan en muchos gráficos de ACF/PACF). Cuando la barra cae fuera de esas bandas, se considera “significativa” (es decir, estadísticamente distinta de cero).

Interpretación general

- Que la ACF caiga rápidamente cerca de 0 después del primer o segundo retraso sugiere que, tras diferenciar la serie, no hay mucha autocorrelación en lags mayores, lo que suele ser un indicio de que la serie se ha vuelto más estacionaria.

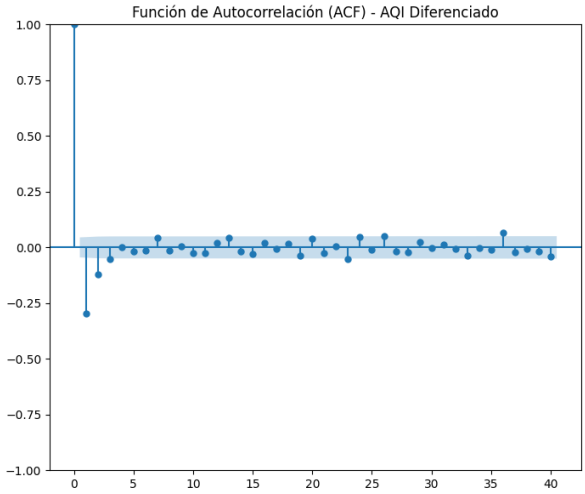


Figura 2. Función de Autocorrelación (ACF) - AQI Diferenciado.

- El pico positivo en lag 1 sugiere que hay cierta dependencia con el valor inmediatamente anterior, incluso después de la diferenciación.

c. Gráfica de la PACF

Eje horizontal (lag o retrasos)

- También va de 0 a aproximadamente 40, al igual que la ACF.

Eje vertical (valor de la autocorrelación parcial)

- Representa cuánta correlación adicional hay en ese lag una vez que se han tenido en cuenta los efectos de los retrasos anteriores.

Barras o puntos para cada lag

- Se observa un pico inicial (lag 1) que suele ser significativo, es decir, está por encima o por debajo de las bandas de confianza.
- Posteriormente, las barras tienden a dispersarse alrededor de 0 y la mayoría no parecen sobrepasar por mucho las líneas de significancia.

Interpretación general

- Un pico significativo en lag 1 y valores cercanos a 0 en lags posteriores (o que caen dentro de las bandas de confianza) implica que la única dependencia notable se da en el primer retraso.
- Esto a menudo sugiere que, después de diferenciar la serie, podrías estar frente a un modelo ARIMA donde el componente AR (autoregresivo) podría ser de orden 1 (o tal vez muy bajo), ya que la PACF normalmente nos indica el orden del componente AR en la serie.

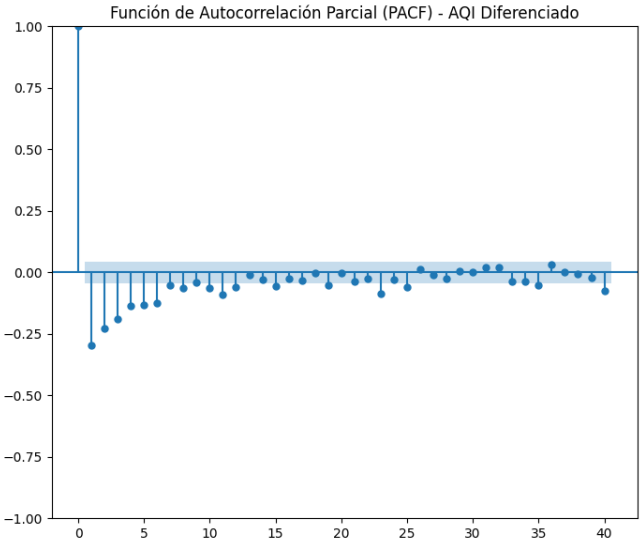


Figura 3. Función de Autocorrelación Parcial (PACF) - AQI Diferenciado.

e. Aplicació una diferenciación (d=1*d*=1)

- **Precaución metodológica**: Aunque la serie es estacionaria según ADF, podrían existir **tendences residuales o heterocedasticidad** que justifiquen una diferenciación para mejorar el ajuste del modelo.
- **Análisis gráfico**: Si la serie original mostraba una **tendencia visual o fluctuaciones no constantes**, la diferenciación ayuda a suavizarla.
- **ACF/PACF**: Las gráficas de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie original podrían haber sugerido la necesidad de integrar (d=1*d*=1) para identificar mejor p y q.

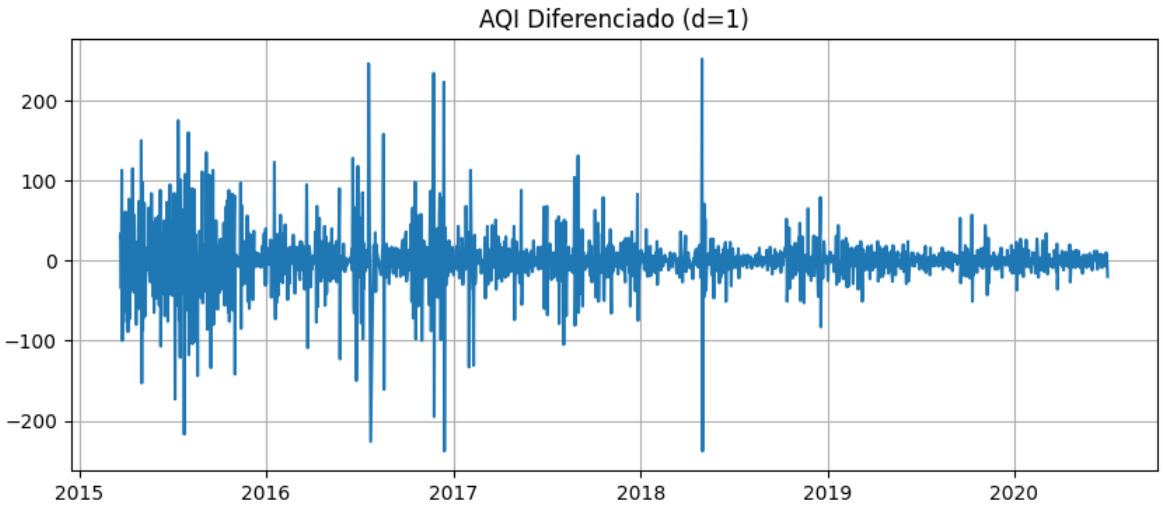


Figura 4. AQI Diferenciado.

Los resultados de las pruebas indican que la serie original del AQI en Bengaluru es estacionaria, por lo que procederemos a emplear el modelo ARIMA.

4. Estructura del modelo

Modelo: ARIMA(2, 1, 2)

Esto significa que se asume un proceso autoregresivo (AR) de orden 2, luego se diferencia la serie una vez (d = 1) para volverla estacionaria, y se incluye un componente de promedio móvil (MA) de orden 2.

Parámetros estimados

La tabla muestra los coeficientes de cada término AR y MA, así como su error estándar, valor z, valor p y el intervalo de confianza:

1. **AR.L1 y AR.L2**
    - Son los coeficientes del componente autoregresivo de orden 1 y 2, respectivamente.
    - Ejemplo (valores aproximados según la captura):
        - AR.L1 = -0.2947, $p < 0.05$
        - AR.L2 = 1.5227, $p < 0.05$
    - Ambas estimaciones son estadísticamente significativas ($p < 0.05$), lo que sugiere que los valores pasados (en lags 1 y 2) tienen una influencia significativa en el valor presente del AQI diferenciado.
2. **MA.L1 y MA.L2**
    - Son los coeficientes del componente de promedio móvil de orden 1 y 2, respectivamente.
    - Ejemplo (valores aproximados):
        - MA.L1 = -1.2070, $p < 0.05$
        - MA.L2 = 0.9188, $p < 0.05$

- También resultan significativos, indicando que los “shocks” o errores (residuos) en lags 1 y 2 influyen significativamente en la dinámica de la serie.

3. **Sigma2**
    - Es la varianza estimada del error (ruido) del modelo.
    - Ejemplo (valor aproximado): sigma2 = 887.5197
    - Representa la magnitud de la variabilidad residual. Un valor más bajo suele indicar que el modelo captura mejor la variabilidad de la serie, pero hay que compararlo con otros modelos o con la escala original de los datos.

Métricas de calidad de ajuste

- Log Likelihood (Log verosimilitud): -9188.82
- AIC (Criterio de Información de Akaike): 18387.664
- BIC (Criterio de Información Bayesiano): 18419.76
- HQIC (Criterio de Información de Hannan-Quinn): 18397.885

Generalmente, para comparar modelos es útil contrastar AIC y BIC de diferentes configuraciones (por ejemplo, ARIMA(2,1,2) vs. ARIMA(1,1,1), etc.). El modelo con valores más bajos de estos criterios es preferible, pues indica un mejor equilibrio entre el ajuste del modelo y la penalización por el número de parámetros.

5. Verificación del modelo – Validación

Análisis de residuos

Se analizaron los residuos del modelo ajustado para evaluar si cumplen con los supuestos necesarios:

- **Gráfico temporal de residuos** muestra que no hay patrones visibles ni tendencias, especialmente en los últimos años.
-

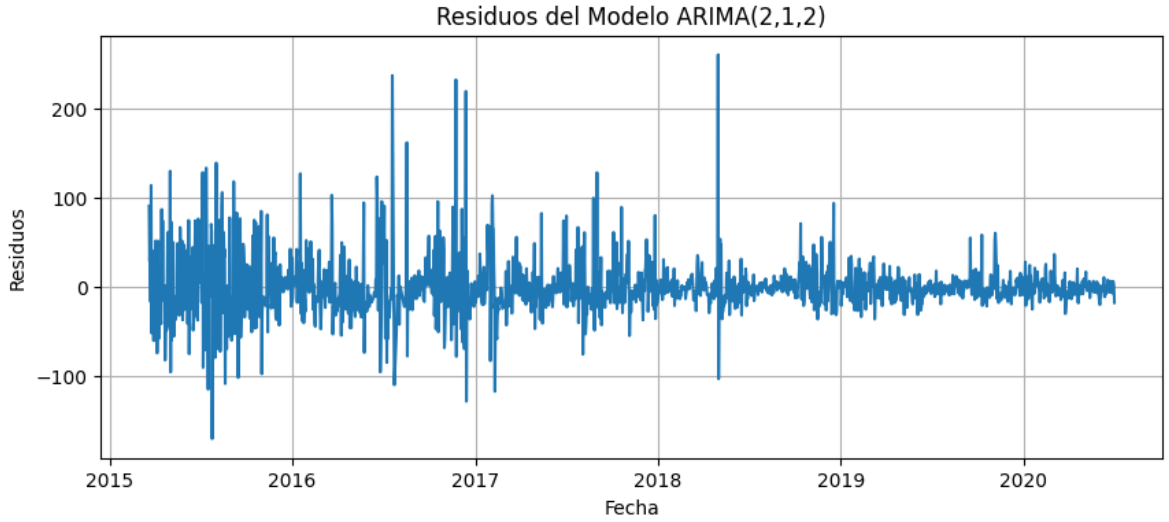
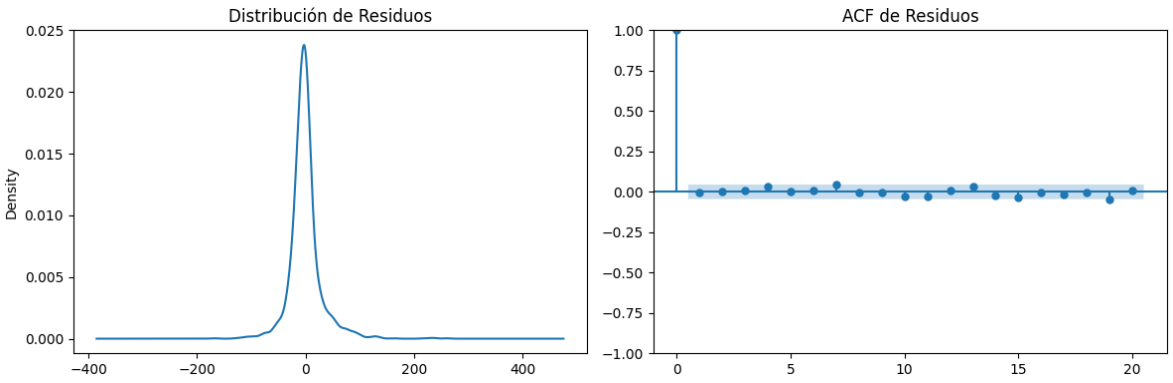


Figura 5. Residuos del Modelo ARIMA.

- **Distribución de residuos** centrada en cero, aunque con colas largas, lo cual es común en series con eventos extremos (picos de AQI).

- **ACF de residuos** no presenta autocorrelación significativa → los residuos se comportan como ruido blanco.



Figuras 6 y 7. Distribución de Residuos y ACF de Residuos.

RMSE (Root Mean Squared Error)

El RMSE se calculó con los datos de prueba, y resultó en: 9.114006124280083.

Este valor indica que, en promedio, el error de predicción es de alrededor de 9 puntos AQI, lo cual es razonable para una serie que varía en rangos de 40 a 100+.

Prueba de Ljung-Box

La prueba se aplicó a los primeros 10 rezagos de los residuos. Los resultados fueron:

Rezago	**Estadístico (LB)**	**p-valor**
1	0.0036	0.9524
2	0.0046	0.9977
3	0.0814	0.9940
4	2.0654	0.7237
5	2.0847	0.8373
6	2.1559	0.9048
7	6.4555	0.4877
8	6.4883	0.5927
9	6.5313	0.6858
10	8.4653	0.5835

Todos los p-valores son mucho mayores a 0.05, por lo que no se rechaza la hipótesis nula de independencia → los residuos no están autocorrelacionados. Esto valida que el modelo ARIMA(2,1,2) es estadísticamente adecuado.

6. Pronóstico del Modelo ARIMA(2,1,2)

Pronóstico sobre datos de prueba

Se dividió la serie en:

- 1600 observaciones para entrenamiento.
- El resto (~158 observaciones) se utilizaron para validar el pronóstico.
- Con un RMSE de 27.02.

El valor de RMSE representa el error promedio entre los valores reales y pronosticados en la serie de prueba. Aunque es más alto que el RMSE calculado anteriormente con el conjunto completo (≈9.11), sigue siendo razonable dado que se trata de una validación sobre datos no vistos.

Pronóstico de los próximos 30 días

El modelo generó un pronóstico de 30 días a partir del último dato disponible. En la gráfica, se muestran los últimos 100 días reales de AQI.Se visualiza el pronóstico futuro (línea punteada) generado por el modelo ARIMA(2,1,2).

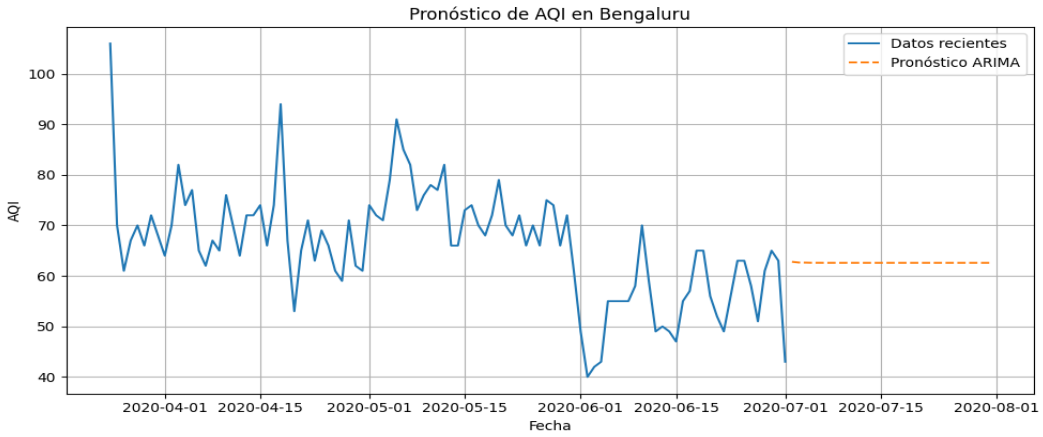


Figura 8. Pronóstico de AQI en Bengaluru.

Aunque el modelo genera un pronóstico válido, este tiende a estabilizarse en torno a un valor constante (≈62), lo cual se debe a que:

- El modelo no detecta una tendencia fuerte ni estacionalidad reciente.
- El ARIMA(2,1,2) predice una continuación del comportamiento promedio reciente.

4.Conclusiones por integrante:

Arteaga González Edwin Yahir:

Durante esta práctica entendí cómo aplicar correctamente la metodología Box-Jenkins para la construcción de un modelo ARIMA. Además, aprendí a interpretar las gráficas de ACF y PACF, lo cual es fundamental para seleccionar adecuadamente los parámetros del modelo.

Juárez Gaona Erick Rafael: Aunque el ARIMA(2,1,2) es adecuado para patrones históricos, su limitación ante eventos atípicos (como la caída abrupta de emisiones durante el COVID-19) subraya la necesidad de enfoques híbridos. Combinar modelos SARIMA (para estacionalidad), machine learning (para capturar no linealidades) o incorporar datos en tiempo real (sensores IoT) mejorarían la precisión. Además, el RMSE de 27.02 en datos de prueba sugiere sobreajuste potencial, requiriendo validación cruzada con múltiples particiones.

Rico Gaytan Diana Andrea: Con esta práctica comprendí la importancia de transformar la serie para volverla estacionaria y cómo esta transformación impacta en la calidad del modelo. Me siento más segura al trabajar con series temporales y análisis de residuos.

Ruiz Merino Wendy Ivonne: Pude observar cómo un modelo ARIMA bien ajustado puede generar pronósticos confiables y útiles para la toma de decisiones. También comprendí la importancia de las métricas de error para evaluar la efectividad del modelo y su aplicabilidad real.: