Trabajo práctico 2

Taller de programación
Maestría en Economía Aplicada
Universidad de Buenos Aires

Granly Jiménez

9 de noviembre de 2024

Objetivo

El objetivo de este trabajo es realizar un análisis exhaustivo de una base de datos de Airbnb en la ciudad de Nueva York para entender las características de los alojamientos y sus precios, así como desarrollar un modelo de predicción que permita estimar dichos precios en función de Múltiples. variables.

A lo largo de todas las etapas, se emplean técnicas estadísticas y de visualización, así como metodologías de limpieza y modelado, con el fin de obtener insights valiosos sobre los factores que afectan los precios y construir una herramienta predictiva útil para la toma de decisiones en este mercado.

Introducción

El análisis de datos es clave para comprender patrones y tomar decisiones informadas. En este trabajo, se estudia una base de datos de Airbnb en Nueva York, evaluando cómo factores como ubicación, tipo de habitación, disponibilidad y precios influyen en los valores de los hospedajes. El proceso incluye limpieza de datos, imputación de valores faltantes, manejo de variables categóricas mediante codificación y la aplicación de modelos predictivos como la regresión lineal. Este modelo permite interpretar la relación entre variables y precios, complementado con métricas como el error cuadrático medio (MSE) para evaluar su precisión. El análisis no solo prevé precios, sino que identifica las características más valoradas por los usuarios, optimizando decisiones en el mercado.

Limpieza, modelización, visualización, predicción y validación de datos

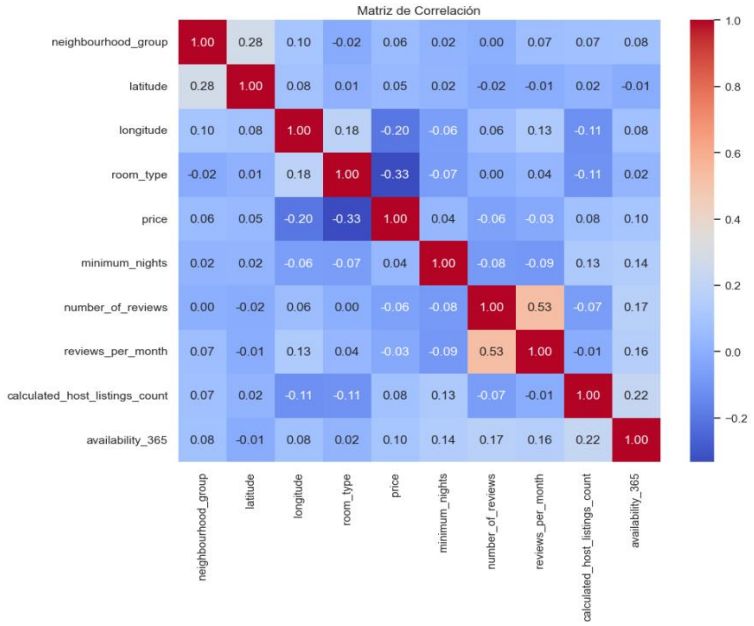
Para realizar un análisis de datos preciso, es esencial seguir varios pasos clave. Primero, la limpieza de datos asegura que el análisis no esté contaminado por valores duplicados, faltantes o atípicos (outliers), los cuales, de no tratarse, pueden afectar los resultados. Transformar variables categóricas mediante técnicas de codificación permite convertir datos cualitativos, como tipos de

habitación o vecindarios, en valores numéricos interpretables por el modelo. Durante el análisis exploratorio, se visualizan patrones y relaciones entre variables a través de gráficos, facilitando la identificación de posibles correlaciones.

En el modelado, una técnica útil es la regresión lineal, que permite interpretar cómo variables como la ubicación o las reseñas afectan el precio. También, el Análisis de Componentes Principales (PCA) reduce la cantidad de variables, manteniendo la mayor parte de la información para simplificar el análisis. Finalmente, la evaluación del modelo con métricas como el MSE o el MAE mide la precisión de las predicciones, indicando su utilidad para futuros datos.

En el caso de reviews_per_month, los valores faltantes se completan con 0, interpretando que propiedades sin reseñas no han tenido actividad en ese aspecto. Esta práctica es común cuando se manejan variables de frecuencia, permitiendo que los análisis continúen sin interrupciones y preservando la integridad del conjunto de datos. Del mismo modo last_review, que indica la última reseña recibida, se llena con 0, denotando una falta de revisión. Aunque esta imputación es práctica para evitar obstáculos en el análisis, también puede influir en la interpretación de la antigüedad de las reseñas, ya que no refleja con precisión la fecha real.

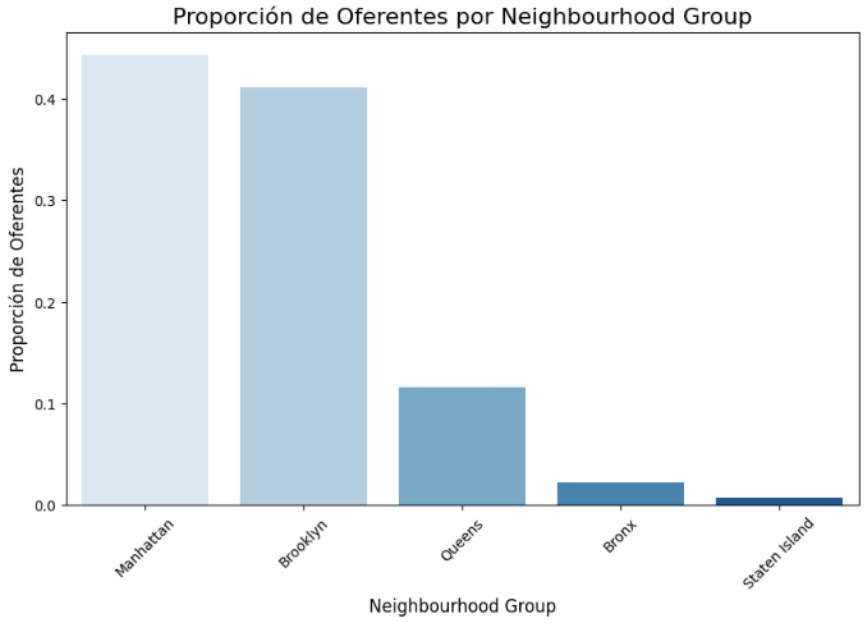
La imputación tiene varios beneficios. Permite conservar todos los registros, lo cual es fundamental en análisis con grandes volúmenes de datos; evita errores en los algoritmos de aprendizaje automático que requieren datos completos; y facilitar la identificación de propiedades sin reseñas, aportando conocimientos sobre patrones de interacción de los usuarios. Según la literatura sobre imputación de datos (Little & Rubin, 2002), completar valores faltantes es clave para mantener la estructura y mejorar la eficiencia de los modelos predictivos. Sin embargo, la imputación debe realizarse con cuidado para no distorsionar la naturaleza de los datos ni introducir sesgos.



La matriz de evaluación revela las relaciones lineales entre las variables del conjunto de datos de Airbnb. En relación al precio, la variable con mayor rentabilidad negativa es room_type(-0.33), lo que indica que el tipo de habitación influye significativamente en los precios, siendo las habitaciones privadas o compartidas generalmente más económicas que las propiedades completas. Las demás variables presentan correlaciones débiles con el precio, lo que sugiere que factores como la ubicación (latitude, longitude) y la disponibilidad (availability_365) tienen un impacto limitado en la determinación de los precios.

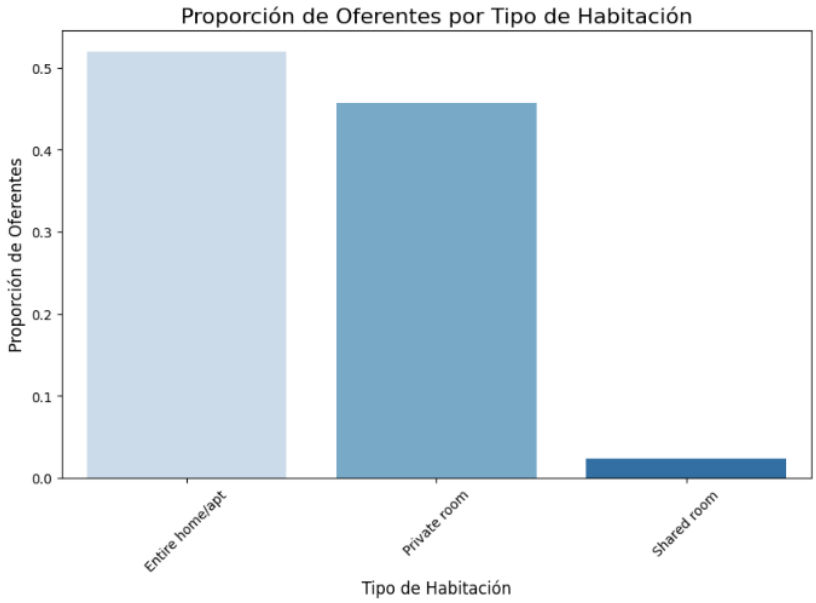
Por otro lado, se observa una valoración moderada (0.53) entre number_of_reviewsy reviews_per_month, algo esperado, ya que un mayor número de reseñas suele asociarse con una mayor frecuencia de evaluaciones. Sin embargo, la mayoría de las correlaciones entre las demás variables son débiles o insignificantes, como la relación entre neighbourhood_group y características como latitude o longitude, lo que sugiere que el grupo de vecindario tiene poca conexión directa con otras variables.

En general, la matriz muestra que las correlaciones entre las variables son mayormente bajas (menores a 0.2), lo que podría indicar que muchas relaciones no son lineales o tienen una influencia limitada. Este panorama resalta la importancia de explorar interacciones o transformaciones más complejas para capturar patrones significativos en un modelo predictivo del precio.

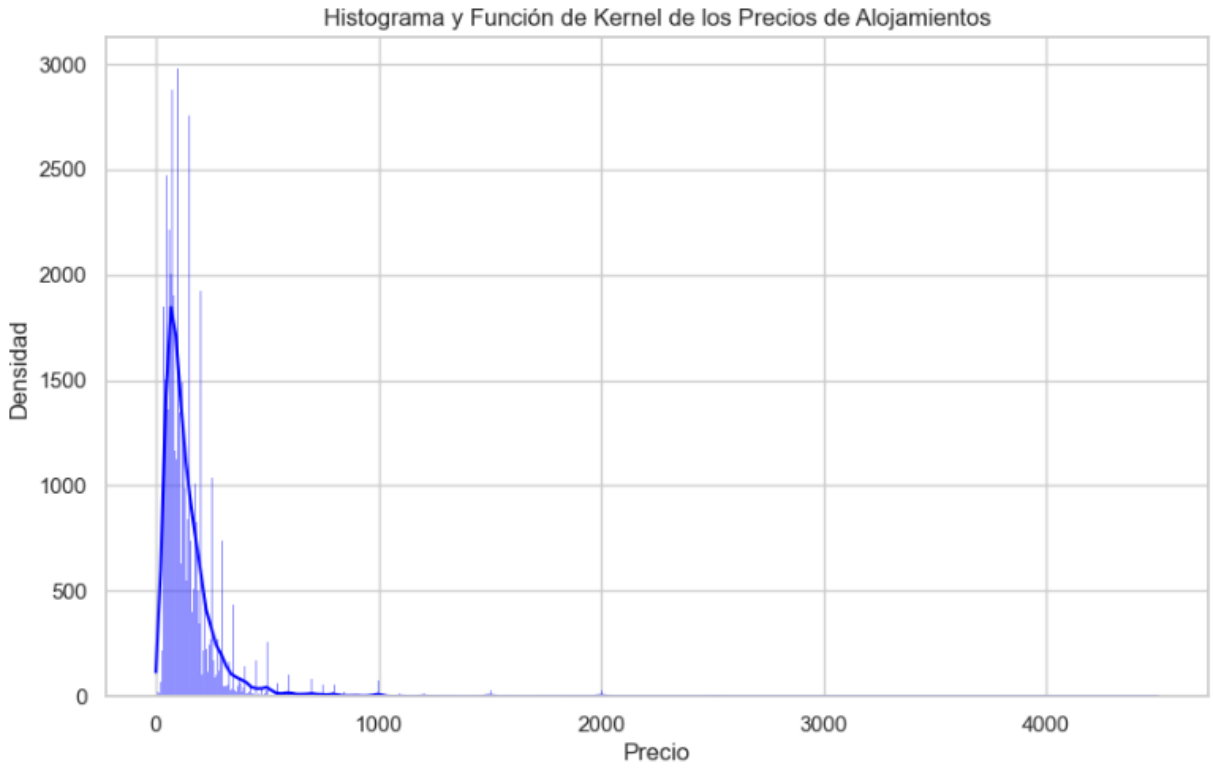


La concentración de oferentes en Manhattan y Brooklyn, que juntos representan cerca del 85% de la oferta total, sugiere una alta preferencia y competitividad en estas áreas, que son conocidas tanto por su atractivo turístico como por su desarrollo inmobiliario. Manhattan lidera con una proporción superior al 40%, seguida de cerca por Brooklyn. Esto es indicativo de la demanda turística que se concentra en zonas urbanas céntricas y bien conectadas, y también puede reflejar las políticas locales de alquileres y las normas de zonificación que permiten y fomentan la oferta en estos barrios. Queens, el Bronx y Staten Island representan una proporción mucho menor del

total de oferentes, posiblemente debido a su menor atractivo turístico o restricciones en cuanto a licencias para alquileres a corto plazo. Además, las diferencias en accesibilidad y en la infraestructura turística también podrían ser factores que influyan en esta distribución, sugiriendo una demanda que está influenciada tanto por la percepción de valor de cada zona como por la disponibilidad de espacios.

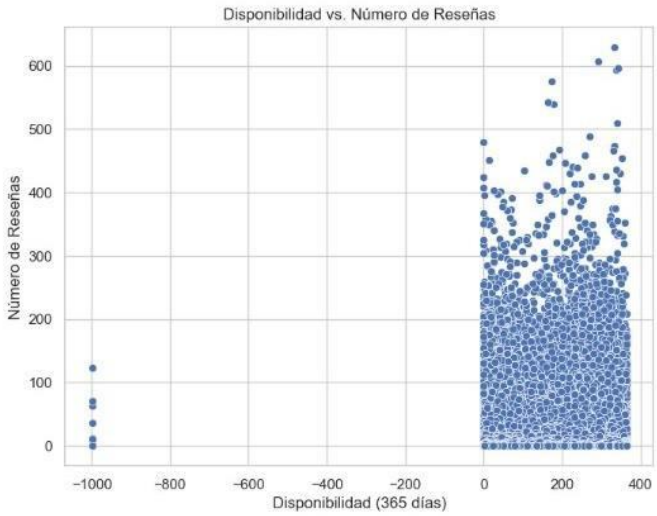


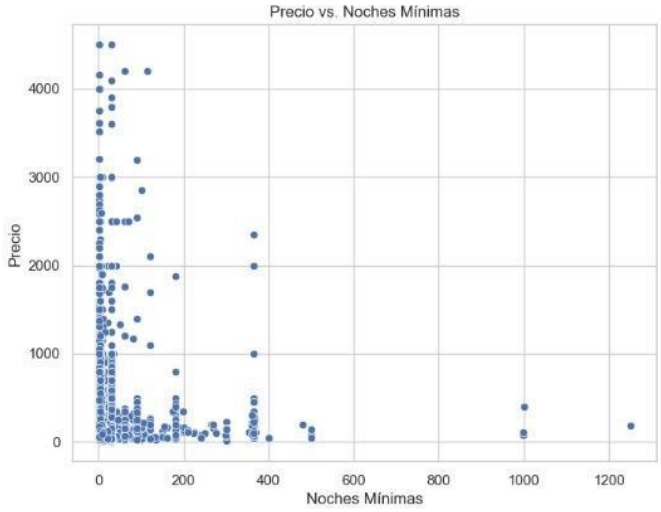
En cuanto a la composición por tipo de habitación, el predominio de ofertas de viviendas completas (más del 50%) indica que los anfitriones están adaptando sus espacios para satisfacer la demanda de privacidad de los huéspedes, quienes en su La mayoría prefiere alquilar propiedades completas en lugar de compartir espacios comunes. Esto puede ser una respuesta a las expectativas de los usuarios de Airbnb, quienes pueden valorar la independencia que ofrece un apartamento o casa completa en una ciudad tan dinámica y transitada como Nueva York. La alta proporción de habitaciones privadas (alrededor del 45%) también es relevante, ya que sugiere que existe un mercado significativo de viajeros que buscan una experiencia más económica pero que aún valoran un cierto grado de privacidad, probablemente turistas o trabajadores que buscan estancias más. largas y económicas. Las habitaciones compartidas representan la menor proporción, lo que podría estar relacionado con la reticencia de los viajeros a compartir espacios con extraños, especialmente en contextos urbanos de alta densidad y donde el valor de la privacidad es fundamental.



El histograma y la función de kernel muestran que la distribución de los precios de los alojamientos está sesgada a la derecha, con una alta concentración en valores cercanos a cero. Esto indica que la mayoría de los alojamientos tienen precios bajos, mientras que los valores más altos son excepcionales. La curva KDE, generada automáticamente, utiliza un ancho de banda adecuado para capturar detalles locales, ya que refleja picos precisos en las frecuencias. Sin embargo, el sesgo y la presencia de valores atípicos (como precios superiores a 2000) podrían influir en el promedio y en el análisis general del mercado.

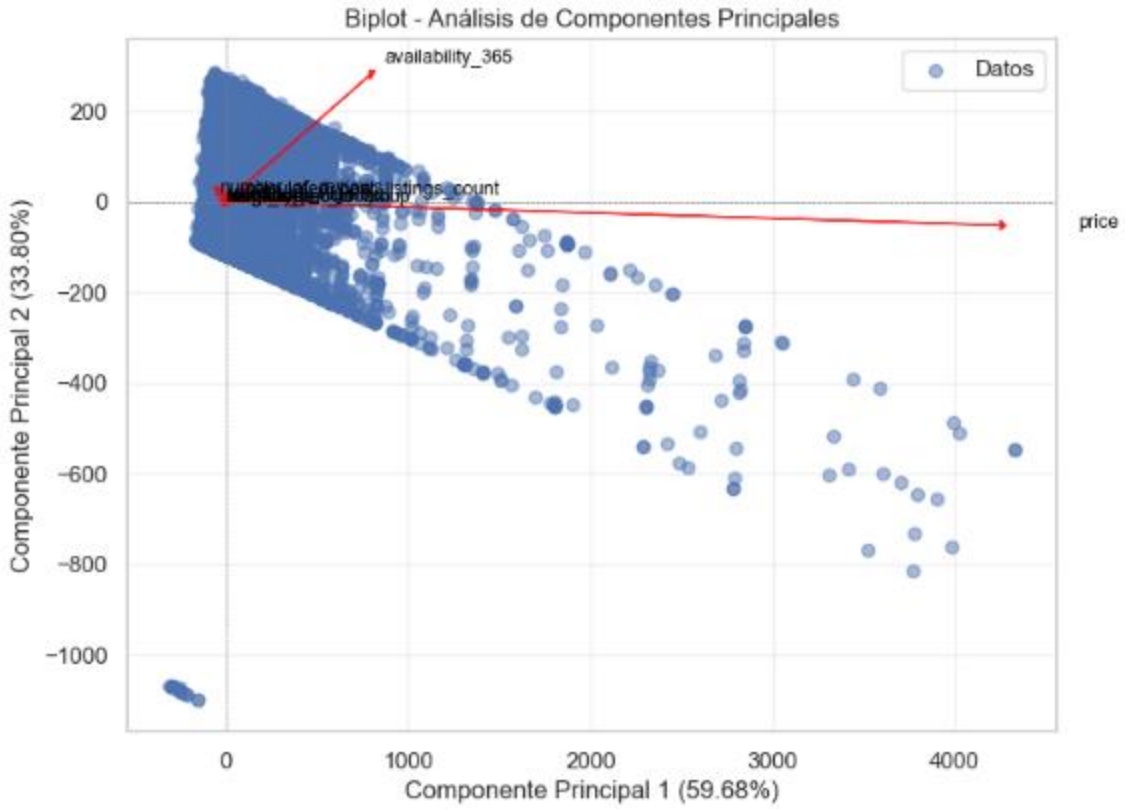
En cuanto a las estadísticas generales, el precio mínimo observado es de 0.0, el precio máximo es de 4500.0, y el promedio se encuentra en 149.04. Por otro lado, el análisis por grupos muestra que los precios promedio varían significativamente entre los Neighborhood groups, con valores que oscilan entre 87,58 y 191,37. Por tipo de habitación, los alojamientos privados tienen el precio promedio más alto (206,24), mientras que los compartidos y habitaciones individuales muestran precios considerablemente más bajos (88,08 y 70,13, respectivamente).





Primero, he decidido analizar la relación entre el precio y las noches mínimas requeridas, donde se observa que la mayoría de las propiedades tienen precios bajos y requieren pocas noches mínimas, concentrándose en un rango menor a 200 noches y con precios generalmente inferiores a 1000 dolares. Sin embargo, hay valores atípicos, como propiedades que requieren más de 1000 noches mínimas, aunque estas son extremadamente raras. Esto sugiere que, en general, los alojamientos más accesibles no imponen requisitos de estadía prolongada.

Posteriormente, analizo la relación entre la disponibilidad (en días al año) y el número de reseñas, evidencia que la mayoría de los alojamientos tienen disponibilidades positivas y se concentran entre 0 y 200 días disponibles. A medida que aumenta la disponibilidad, parece haber una mayor cantidad de reseñas, lo que podría indicar una relación entre mayor disponibilidad y mayor popularidad. Sin embargo, se observan puntos negativos en la disponibilidad que podrían representar errores o valores inusuales en los datos, lo que merece una revisión adicional. Ambos gráficos revelan tendencias claras, pero también indican la presencia de datos atípicos que podrían distorsionar el análisis.



En el análisis de componentes principales, obtenemos que los dos primeros componentes explican el 93.49% de la varianza, lo cual nos indica que estas capturan casi toda la información relevante de los datos originales. Este alto porcentaje confirma que el PCA es una representación eficaz para reducir la dimensionalidad sin perder detalles significativos. Las variables más influyentes están bien diferenciadas a través de los loadings, que reflejan su contribución a cada componente. El gráfico sugiere patrones claros y relaciones lineales entre los datos proyectados, lo que facilita su interpretación en este espacio reducido.

	MSE	**RMSE**	**MAE**
Train	6.609	81,2965	46,1919
Test	24.051,0266	155,0839	72,1117
R^2 de train		0.4643	
R^2 de test		0.0890	

El modelo diseñado para predecir precios nos muestra resultados que, aunque son aceptables en el conjunto de train, se encuentran algunos problemas importantes cuando se aplica a datos

nuevos. Durante el entrenamiento, el modelo logra un error promedio de 46 dólares, lo que indica que, en general, no se desvía demasiado de los valores reales. Además, logra explicar casi el 46% de las variaciones en los precios, lo cual sugiere que es capaz de identificar algunos patrones relevantes en los datos con los que fue entrenado.

Sin embargo, al evaluar su desempeño en el conjunto de prueba, el modelo pierde precisión de manera significativa. El error promedio sube a 72 dólares, lo que refleja una mayor incertidumbre en las predicciones. Más preocupante aún, el modelo apenas puede explicar el 9% de las variaciones en los precios de los datos nuevos, lo que evidencia que no logra capturar patrones generales ni adaptarse a datos diferentes de los del entrenamiento.

Esto indica un problema de sobreajuste: el modelo se ha ajustado demasiado a los detalles específicos del conjunto de entrenamiento y no logra captar lo suficiente sobre las relaciones generales en los datos.

Conclusiones

En este trabajo se presenta un análisis exhaustivo de una base de datos relacionados con alojamientos en la ciudad de Nueva York, cuyo objetivo es entender cómo distintas características influyen en los precios de los hospedajes. A lo largo del proceso, se emplearon metodologías como la regresión lineal y el Análisis de Componentes Principales (PCA) para identificar patrones significativos y reducir la dimensionalidad de los datos, facilitando la interpretación y la predicción precisa de los precios.

Los resultados obtenidos revelaron que ciertas variables, como el tipo de habitación y el número de reseñas, tienen una relación considerable con el precio, mientras que otras, como la ubicación exacta, muestran una influencia más débil. El análisis también identificó una alta concentración de oferta en áreas como Manhattan y Brooklyn, lo que refleja la demanda turística y la competitividad de estas zonas. Además, se encontró que las propiedades completas, que ofrecen mayor privacidad, son las más valoradas, aunque también hay una proporción significativa de habitaciones privadas. Sin embargo, la distribución de los precios mostró una tendencia hacia valores bajos, con pocos casos de precios extremadamente altos.

En cuanto al modelo predictivo, aunque el rendimiento fue adecuado en el conjunto de datos de entrenamiento, al evaluarlo con nuevos datos, se observará un notable descenso en su precisión. Esto sugiere un problema de sobreajuste, donde el modelo se ajustó demasiado a las especificidades del conjunto inicial y no logró generalizar bien con datos no vistos.