

Análisis Predictivo de Precios de Airbnb en la Ciudad de México Utilizando Modelos de Aprendizaje Supervisado



Israel Cervantes J Bulmaro Juárez H José Castro A Francisco Tajonar S
XVII Semana Internacional de la Estadística y la Probabilidad (2024)

Resumen

Este proyecto se centra en el análisis predictivo de precios de Airbnb en CDMX utilizando modelos de aprendizaje supervisado. Se realiza un análisis exploratorio de datos para comprender la estructura y características del conjunto de datos. Posteriormente, se aplican modelos de regresión lineal y regresión múltiple, ajustando los modelos para mejorar su rendimiento. Además, se implementa el clasificador K-Nearest Neighbors para comparar su efectividad frente a los modelos de regresión.

Introducción

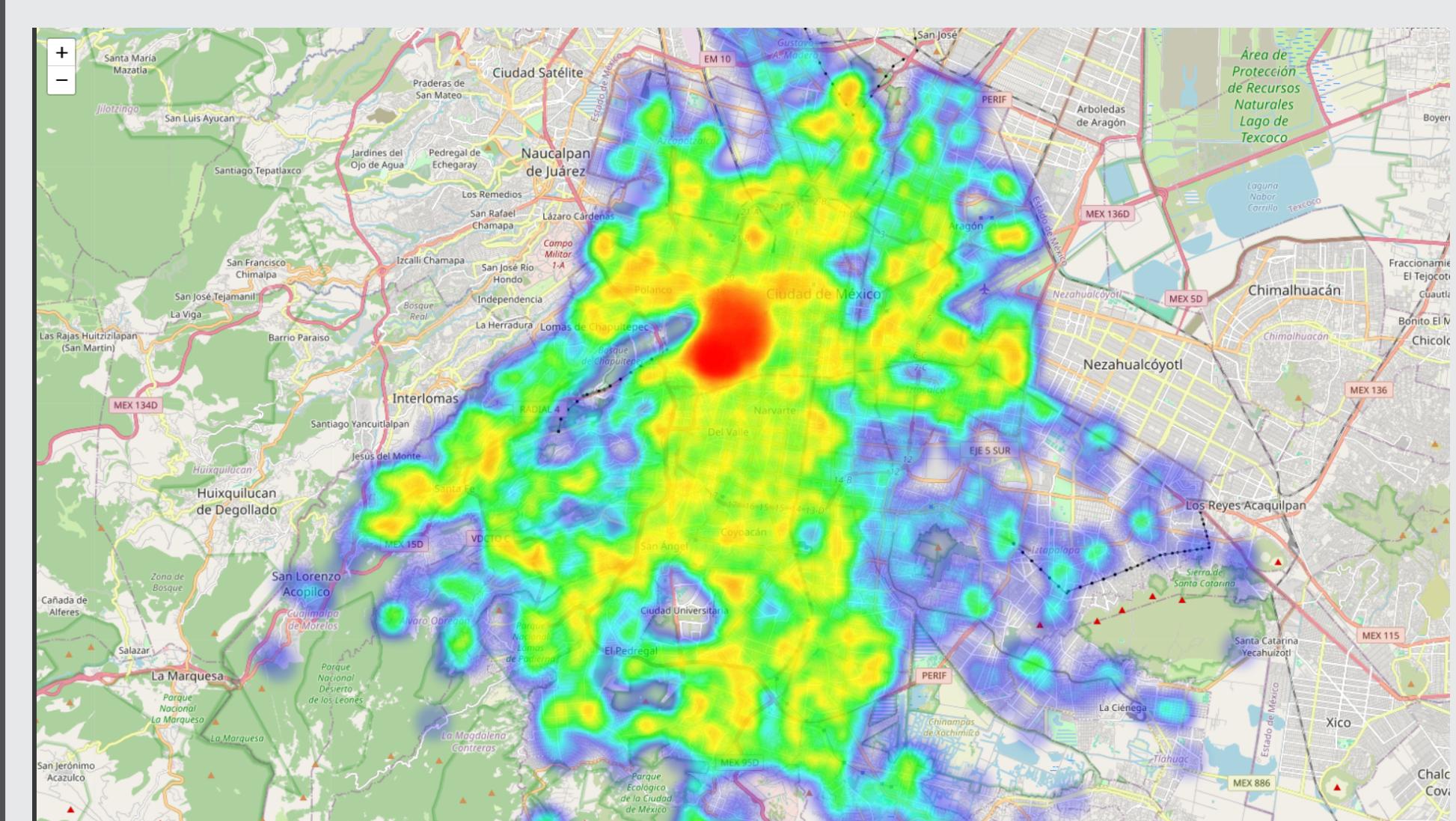


Figura 1: Mapa de calor CDMX

En las últimas décadas, el mercado de alquileres a corto plazo ha crecido notablemente, especialmente en ciudades con alta demanda turística y residencial, como la Ciudad de México. Este estudio se centra en aplicar técnicas de aprendizaje supervisado para estimar los precios de Airbnb, buscando no solo mejorar la precisión de las predicciones en un entorno de mercado dinámico, sino también proporcionar una visión detallada de los factores que afectan los precios y la disponibilidad de las propiedades. El objetivo es ofrecer una comprensión más profunda del mercado que permita a propietarios y arrendatarios tomar decisiones más informadas en un entorno competitivo.

Conclusion

El análisis de técnicas de aprendizaje supervisado para predecir precios de propiedades en Airbnb en la Ciudad de México revela que, a pesar de la complejidad del modelo de regresión lineal múltiple, este no siempre supera al modelo de regresión lineal simple en precisión. En cambio, el modelo K-Nearest Neighbors (KNN) ofrece una mayor estabilidad y precisión, especialmente con un valor de k de 3, que equilibra el riesgo de sobreajuste y mejora la confiabilidad. Estos hallazgos destacan la importancia de ajustar adecuadamente los parámetros del modelo y seleccionar la técnica más efectiva para mejorar la precisión en la predicción de precios en un mercado dinámico.

Referencias

- [1] W. III Mendenhall D. D. Wackerly and R. L. Scheaffer. *Estadística matemática con aplicaciones*. Cengage Learning, 7th edition, 2010.
- [2] W. Wang S. Nasiriany, G. Thomas and A. Yang. *A Comprehensive Guide to Machine Learning*. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2019.
- [3] T. Hastie R. Tibshirani G. James, D. Witten and J. Taylor. *An Introduction to Statistical Learning with Applications in Python*. Springer Texts in Statistics. Springer Nature Switzerland AG, 2023.

Análisis Exploratorio y Método

En el análisis de datos para predecir precios de Airbnb en Ciudad de México, se utilizó un conjunto de datos de **26,536** entradas y **18** columnas, encontrando la necesidad de limpieza debido a *datos inválidos y valores nulos*. Se detectaron *valores atípicos* en variables como el precio y el número de reseñas, y la distribución de residencias mostró alta densidad en áreas centrales como Polanco y la Condesa (Véase figura 1). Las *correlaciones* revelaron relaciones significativas entre variables de reseñas, mientras que el precio y la disponibilidad anual mostraron una *dispersión* sin una relación clara. Los *gráficos de densidad* indicaron concentraciones y ses-

gos en variables, especialmente en el precio, que mostró una distribución sesgada a la derecha con valores extremos.

Se utilizaron los siguientes modelos:

◆ El modelo de **regresión lineal** estima la relación entre una variable dependiente y y una independiente x mediante

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Se busca minimizar la suma de los errores cuadráticos para ajustar los parámetros.

◆ La **regresión lineal múltiple** extiende la regresión lineal a múltiples variables independientes x_1, x_2, \dots, x_n con

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon,$$

optimiza los coeficientes β_i para reducir el error cuadrático [1].

◆ El algoritmo de **k-vecinos más cercanos** clasifica valores basándose en los k vecinos más cercanos. La ecuación es:

$$\text{Clase}(z) = \arg \max_y \sum_{i=1}^k \mathbb{I}_{\{y_i=y\}},$$

donde $\mathbb{I}_{\{y_i=y\}}$ es 1 si la clase del i -ésimo vecino es y , y 0 en caso contrario. La selección del valor de k es crucial: un k pequeño puede causar sobreajuste, mientras que un k grande puede llevar a subajuste. Se utilizan diversas métricas de distancia, como Euclídea, Manhattan, Minkowski [2].

Resultados

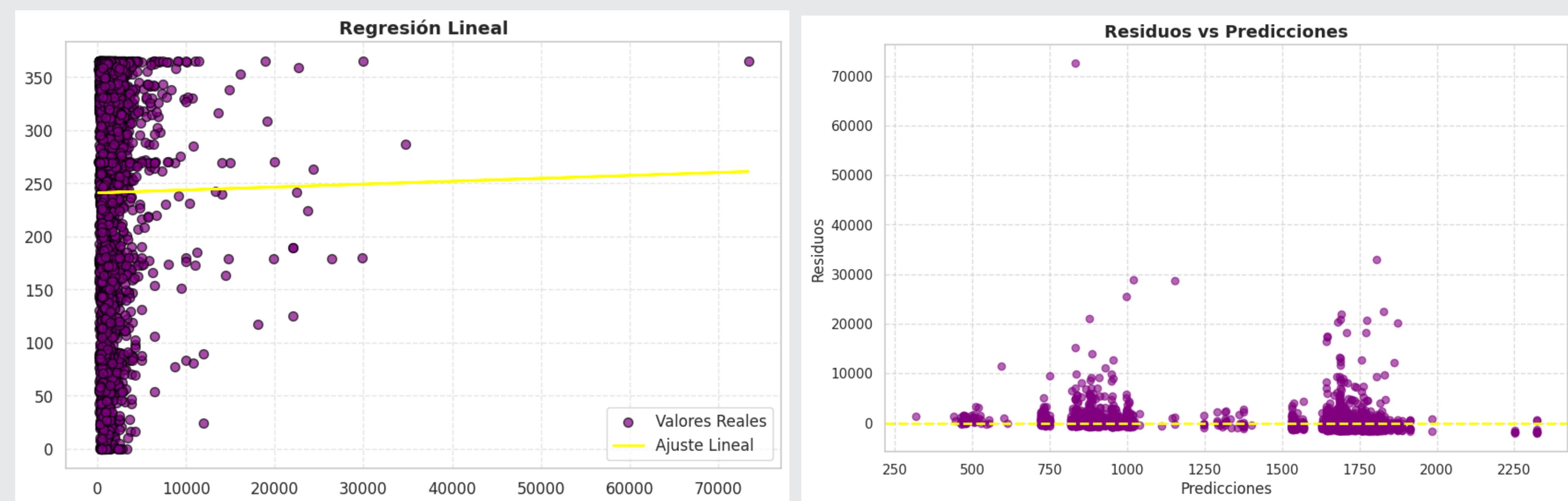


Figura 2: Valores reales vs predichos (izquierda); Errores regresión lineal múltiple (derecha)

La tabla muestra los resultados de las métricas de evaluación para los modelos de regresión:

Tabla 1: Comparación de resultados entre modelos de regresión

Métrica	Lineal	Lineal Múltiple
MAE	96.2408	898.5457
MSE	12020.6596	16153599.0411
RMSE	109.6387	4019.1540

Estos resultados muestran que el modelo de regresión lineal múltiple presenta errores mayores en comparación con el modelo de regresión lineal simple, lo que sugiere que el modelo más complejo no mejora la precisión de las predicciones de manera efectiva en este caso.

Mientras que para el modelo KNN la siguiente gráfica sugiere que valores bajos de k pueden causar *sobreajuste y reducir la precisión*, mientras que un k ligeramente mayor ofrece una precisión más estable.

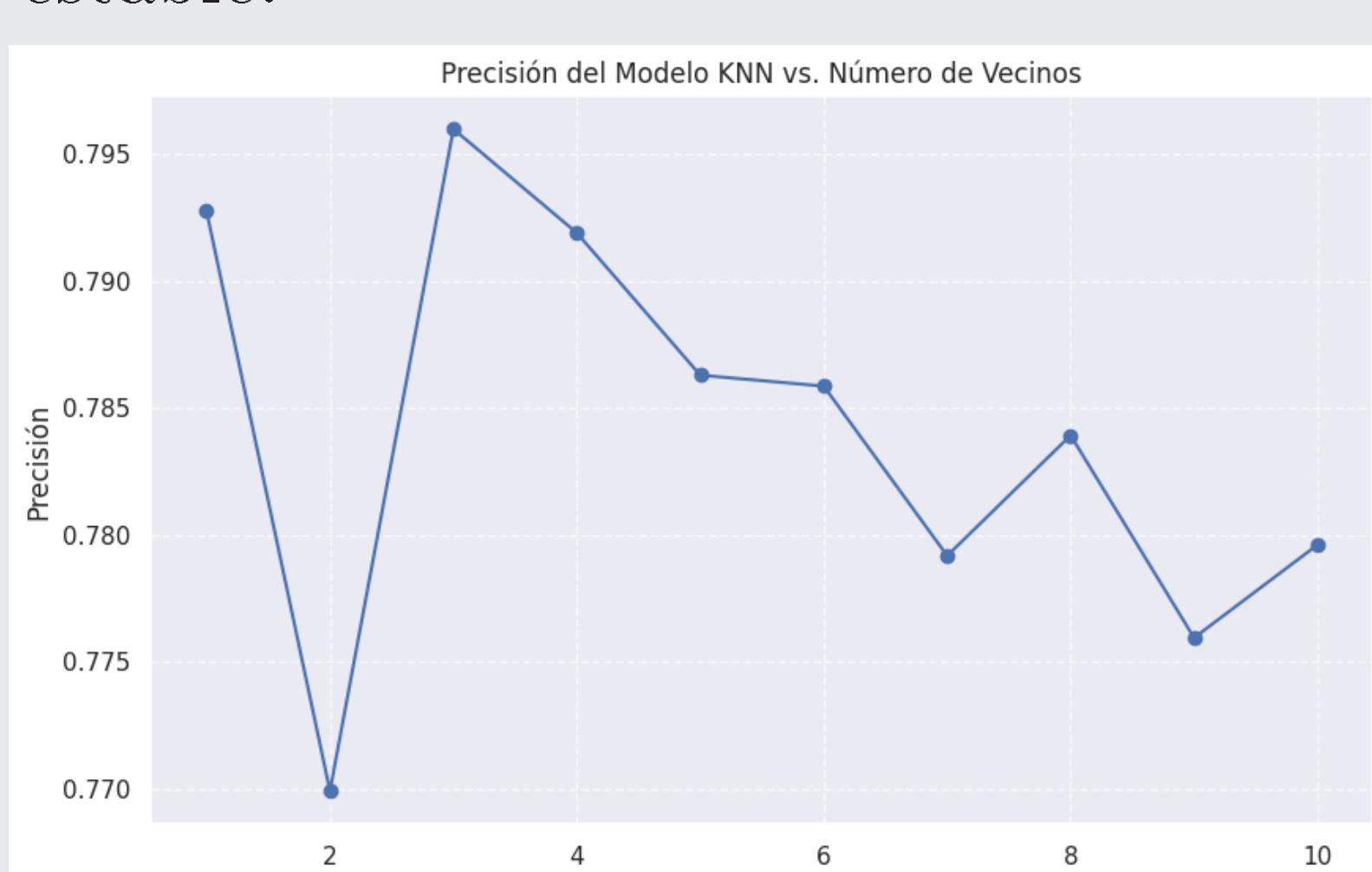


Figura 3: Precisión para diferentes valores de k

Un $k = 3$ proporciona un equilibrio óptimo entre ajuste y generalización, resultando en un rendi-

miento más confiable y menos susceptible a errores.

La *precisión* del modelo es aproximadamente el **79.60%** de las predicciones positivas son correctas. El *F1-score* de refleja un equilibrio entre precisión y *recall*, siendo **78.89%** la medida combinada de la exactitud de las predicciones y la capacidad de capturar todos los positivos relevantes [3]. A continuación, se muestra la tabla con la especificidad para cada clase:

Tabla 2: Especificidad para cada clase

Clase	Especificidad
Barato	0.9381
Razonable	0.9752
Costoso	0.9991
Exclusivo	0.6071