Introducción y Objetivos

Este proyecto tiene como objetivo aplicar técnicas de análisis de datos y machine learning para la empresa Airbnb, con el fin de predecir los precios de las propiedades en función de diversas variables. Esto se lleva a cabo mediante la exploración de los datos, la identificación de patrones relevantes y el entrenamiento de modelos predictivos. En síntesis, la empresa busca desarrollar un modelo que permita estimar con precisión los precios de las propiedades en función de sus características específicas.

Descripción del Dataset

El dataset cuenta con múltiples variables que se analizaron para entender su impacto en la variable objetivo.

Diccionario dataset Airbnb	
Variable	Significado
id	Identificado de la publicación
property_type	Tipo de Propiedad
room_type	Tipo de Habitación
amenities	Amenities que tiene la propiedad
accommodates	Cantidad de comodidades de la publicación
bathrooms	Cantidad de baños
bed_type	Tipo de cama
cancellation_policy	Politica de cancelación
cleaning_fee	Si tiene un recargo por limpieza o no
city	Ciudad
description	Descripción de la publicación
first_review	Fecha de la primera review
host_has_profile_pic	Si el host tiene foto de perfil o no
host_identity_verifie	Si el host es verificado por la página
host_response_rate	Frecuencia de respuesta del host
host_since	Fecha desde que el host se inició en Airbnb
instant_bookable	Si la propiedad se puede reservar de manera instantanea o requiere aprobación del dueño
last_review	Fecha de la última review
latitude	Latitud geográfica de la propiedad
longitude	Longitud geográfica de la propiedad
name	Nombre de la publicación
neighbourhood	Barrio de la propiedad
number_of_reviews	Cantidad de reviews de la publicación
review_scores_ratin	Puntaje de la publicación
thumbnail_url	URL de la publicación
zip code	Código postal de la propiedad
bedrooms	Cantidad de cuartos
beds	Cantidad de camas
price	Precio por noche de la propiedad

Las principales características del dataset son las siguientes:

- Contiene 19,309 filas y 29 columnas en su estado inicial.
- Incluye tanto datos categóricos como numéricos.
- Presenta un total de 20,500 valores nulos distribuidos entre diversas columnas.

Dado este escenario, se llevó a cabo un proceso de preprocesamiento para manejar los datos faltantes, reemplazándolos según su tipo:

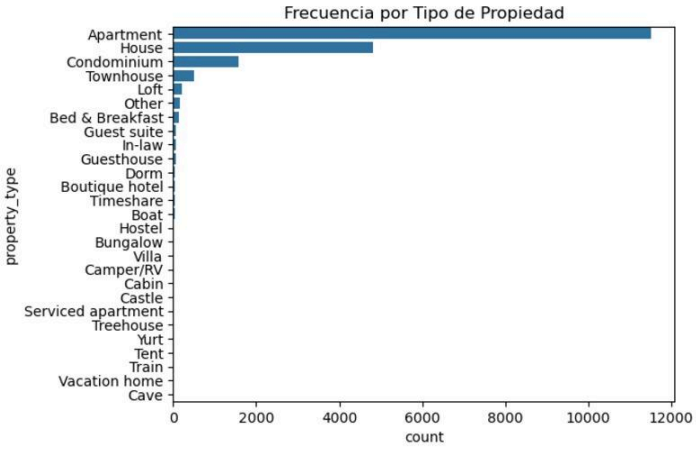
- ➔ Datos categóricos: Se imputaron utilizando la moda
- ➔ Datos numéricos simétricos: Se reemplazaron con la media.
- ➔ Datos numéricos asimétricos: Se utilizó la mediana

Este enfoque permitió conservar la mayor cantidad de información posible mientras se preparaban los datos para el análisis y modelado.

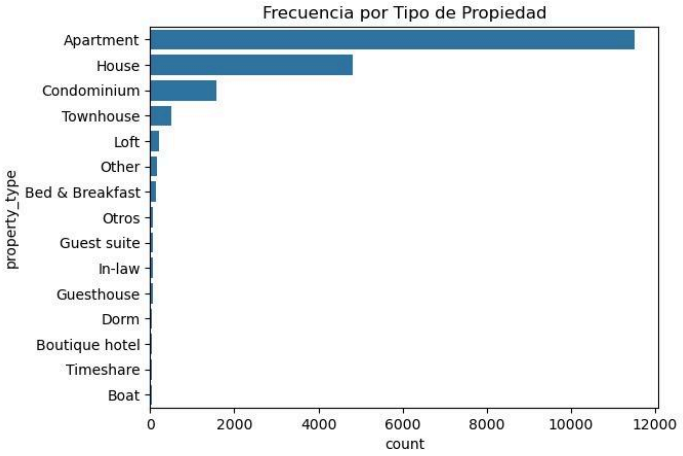
	Tipo de datos	Nulos	Duplicados
id	int64	0	0
property_type	object	0	0
room_type	object	0	0
amenities	object	0	0
accommodates	int64	0	0
bathrooms	float64	35	0
bed_type	object	0	0
cancellation_policy	object	0	0
cleaning_fee	bool	0	0
city	object	0	0
description	object	0	0
first_review	object	3954	0
host_has_profile_pic	object	3	0
host_identity_verified	object	3	0
host_response_rate	object	4296	0
host_since	object	3	0
instant_bookable	object	0	0
last_review	object	3954	0
latitude	float64	0	0
longitude	float64	0	0
name	object	0	0
neighbourhood	object	1458	0
number_of_reviews	int64	0	0
review_scores_rating	float64	4134	0
thumbnail_url	object	2402	0
zipcode	object	225	0
bedrooms	float64	17	0
beds	float64	24	0
price	float64	0	0

Análisis Exploratorio de Datos

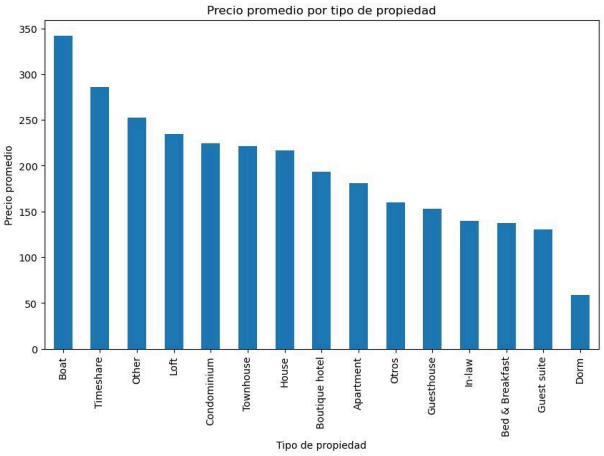
Iniciamos el análisis exploratorio de datos con un gráfico que muestra la frecuencia por tipo de propiedad(*para una mejor visualización de gráficos ir a notebook*):

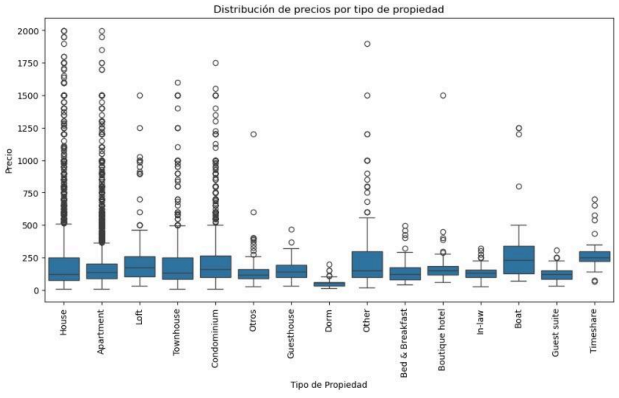


El gráfico muestra que la mayoría de los tipos de propiedades tienen una frecuencia muy baja, lo que dificulta un análisis significativo. Por esta razón, se decidió agrupar en la categoría "Otros" aquellos tipos de propiedades con menos de 20 apariciones en el dataset, representando menos del 1% del total de los datos. Esta agrupación simplifica el análisis al reducir la dispersión de las categorías y permite enfocarse en los tipos de propiedades más relevantes.



Siguiendo con el análisis, analizamos qué tan relacionado está el precio con el tipo de propiedades e hicimos un boxplot para poder visualizar la distribución de estos precios



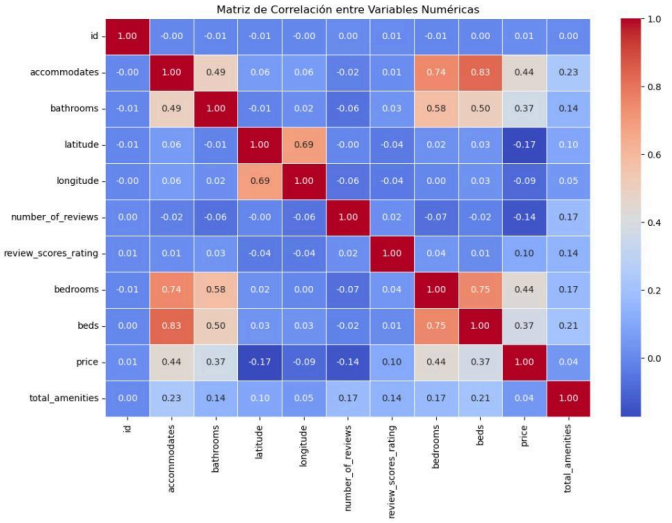


El tipo de propiedad se identifica como un factor clave en la determinación del precio.

Propiedades exclusivas como Boat y Timeshare dominan en los rangos de precios más altos.
Por otro lado, opciones más económicas como Dorm y Guest Suite representan alternativas accesibles.

Los outliers en las propiedades más comunes reflejan variaciones significativas debidas a características específicas como la ubicación o servicios adicionales. Sin embargo, la gran cantidad de outliers en los precios podría distorsionar o sesgar el modelo predictivo.
Por esta razón, se decidió filtrar los precios eliminando los datos en los cuantiles por encima del 90% y por debajo del 10%.

Al analizar los gráficos anteriores, se observaron posibles correlaciones entre variables como el número de amenities, tipos de propiedad, número de reseñas y cantidad de habitaciones. Para verificar esta hipótesis, se construyó una matriz de correlación, lo que permitió explorar y corroborar relaciones significativas entre las variables.



Como conclusión de la matriz, previo a iniciar el modelo de predicción decidimos eliminar todas aquellas columnas que no tienen correlación alguna con el análisis y en aquellas que mostraban una alta correlación entre sí se dejó sólo una de ellas para alivianar el dataset. También se transformó las variables categóricas de utilidad en variables dummies para poder ingresarlas en el modelo.

Materiales y métodos

Modelos implementados:

<u>Regresión Lineal:</u>
busca modelar la relación entre las variables independientes y la variable dependiente usando una función lineal.

$$y = w0 + w1x1 + w2x2 + \ldots + wnxn$$

- y: Es la etiqueta, el precio de las propiedades.
- xi: Representa las features, como tipo de propiedad, número de habitaciones, etc.
- wi: Parámetros del modelo

La Regresión Lineal utiliza el método de mínimos cuadrados para minimizar el RSS

StandardScaler
Para poder usar SVR se necesita tener datos con escalas similares ya que sino se le dará prioridad a variables con datos de mayor dimensión como precio comparado con cantidad de camas. es por eso que se usa standard scaler.

$$z = \frac{x-\mu}{\sigma}$$

Donde:

- μ: Media.
- σ: Desviación estándar.

SVR
Busca maximizar margen con un output numérico real.
lo realiza generando un hiperplano que predice valores maximizando el margen entre predicciones y datos reales
determina un margen de error epsilon como función de costo y trata que las muestras queden dentro de este.

- ➢ Si la diferencia entre el valor ypred y el valor real es menor que epsilon el modelo no lo penaliza.
- ➢ Puntos fuera del margen de tolerancia se penalizan

Optimización con Randomized Search CV:
Es un método para encontrar la mejor combinación de hiperparámetros de un modelo.

- ➢ Define un espacio de hiperparámetros
- ➢ Genera combinaciones de hiperparámetros.
- ➢ Entrena el modelo
- ➢ Cross validation para medir el desempeño
- ➢ Selecciona el mejor modelo

Experimentos y resultados

División del Dataset
El conjunto de datos se dividió en dos subconjuntos: uno para entrenamiento y otro para prueba (70/30%).

- Tamaño del conjunto de entrenamiento: 13.516 muestras.
- Tamaño del conjunto de prueba: 5.793 muestras.

Preprocesamiento de los Datos
Antes de entrenar los modelos, se aplicó un proceso de escalado utilizando StandardScaler para garantizar que todas las características tengan una escala uniforme y no se le dé prioridad a ciertos valores solo por su escala.

Modelos
Se probaron dos modelos principales: Regresión Lineal y SVR. Usando MSE como métrica principal para su comparacion.

Regresión Lineal:
Se entrenó utilizando los datos escalados.
Resultado del **MSE: 162,85.**

SVR con Búsqueda de Hiperparámetros:

Se utilizó RandomizedSearchCV para buscar los mejores hiperparámetros en un espacio definido.
Resultado del MSE: 27.120,18

SVR con Parámetros Ajustados:
Se amplió el espacio de búsqueda para explorar más combinaciones:
Resultado del **MSE: 29.352,81**

El modelo de Regresión Lineal mostró un desempeño aceptable, con un MSE significativamente más bajo (162.85) que el modelo SVR sin optimizar.

Discusión y Conclusiones

Este proyecto se centró en el análisis de datos de propiedades de Airbnb con el fin de predecir precios en función de diversas características. A través de técnicas de EDA, preprocesamiento y machine learning, se logró estimar los precios de manera precisa y eficiente.

Los resultados permitieron identificar patrones clave en los datos, factores como el tipo de propiedad, la ubicación y la

cantidad de amenities son los principales determinantes del precio, mostrando una correlación significativa con el valor final.

En conclusión, este trabajo demostró cómo el uso combinado de técnicas de machine learning y un análisis detallado de datos puede ser una herramienta altamente efectiva para predecir precios en un mercado dinámico como el de Airbnb.

Referencias

-*Scikit-Learn:*
https://scikit-learn.org/stable/modules/svm.html#regression
https://scikit-learn.org/stable/modules/preprocessing.html
- *Pandas:*
https://pandas.pydata.org/docs/user_guide/index.html
- *Repositorio académico*
Github:https://github.com/clusterai/Ciencia-de-Datos-UTN-FRBA
-Seaborn tutorial:
https://seaborn.pydata.org/tutorial/introduction.html
-introduction to statistical learning
https://www.statlearning.com/