UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Trabajo Práctico Final

Ciencia de Datos

“Análisis exploratorio y desarrollo de modelo de Machine Learning para predecir el precio de propiedades en Capital Federal”

Docentes:

- Ing. Palazzo, Martín
- Ing. Aguirre, Nicolás
- Ing. Chas, Santiago

Alumnos:

- Salvadó, Marco
- Elbaum, Juan

Año 2022

Índice

Introducción y objetivos

El presente trabajo es parte del proceso de evaluación académica del curso de Ciencia de Datos de la UTN.

El objetivo es el análisis de un dataset extraído de la plataforma Properati, el cual contiene información de publicaciones de propiedades de Capital Federal, y posterior desarrollo de un modelo de Machine Learning para la predicción del precio de dichas propiedades.

Descripción del dataset

El dataset cuenta con un total de 38.656 registros y 26 variables con distintas características de las propiedades, como el tipo de propiedad, su ubicación geográfica, cantidad de dormitorios, baños y ambientes, superficie cubierta y total, y precio de la misma, entre otras.

Análisis exploratorio de datos

Limpieza y preprocesamiento

En primer lugar, comenzamos verificando el tipo de dato de cada variable y la existencia de nulos que puedan distorsionar los análisis posteriores. A partir de esto se eliminaron 3 variables las cuales no tenían ningún valor no nulo y se profundizó el análisis sobre otras variables que tenían cierto porcentaje de valores nulos.

Además, se verificó que no hayan publicaciones duplicadas y en el caso de haberlas, se procedió a su eliminación. También, habían variables como el país o la localidad, en este caso Capital Federal, Argentina, que se repetía para todos los registros por lo tanto se decidió desestimar porque no aportan variabilidad al modelo.

Se detectó que las variables superficie cubierta, superficie total y cantidad de dormitorios tenían un alto nivel de valores nulos y dado que se consideran variables importantes en la predicción del precio, se tomó la decisión de imputar ambas superficies por la media y a la cantidad de dormitorios por la mediana.

Finalmente, nos quedamos con un dataset que contiene 24.084 registros y 10 dimensiones. Ahora sí, podemos comenzar con el análisis exploratorio propiamente.

Análisis exploratorio

En primer lugar, realizamos un histograma para observar la distribución de la variable precio. Además, para ver la distribución en otra escala, se hizo una transformación de los precios mediante la función logaritmo y nuevamente se realizó el histograma con esta columna.

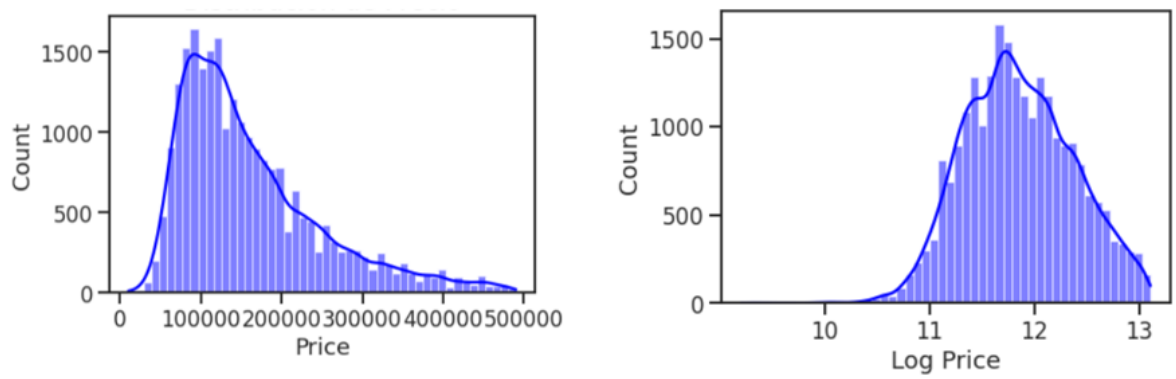


Gráfico 1: Distribución de la variable precio. Elaboración propia.

En la gráfica izquierda podemos observar que el precio se distribuye con una tendencia sesgada hacia la izquierda, en donde el precio con mayor frecuencia se encuentra alrededor de los U$S100.000. A la derecha, a partir de la conversión a escala logarítmica, la distribución se asemeja a una normal.

Por otro lado, realizamos un BoxPlot de la variable “Barrio” a partir del cual observamos que Puerto Madero es claramente el barrio donde se concentran los mayores precios, mientras que en Villa Soldati ocurre lo opuesto. Además, podemos observar que tanto en Caballito, Palermo como en Floresta, encontramos la mayor cantidad de outliers, es decir, es posible encontrar precios muy económicos que distan de la gran mayoría de los casos.

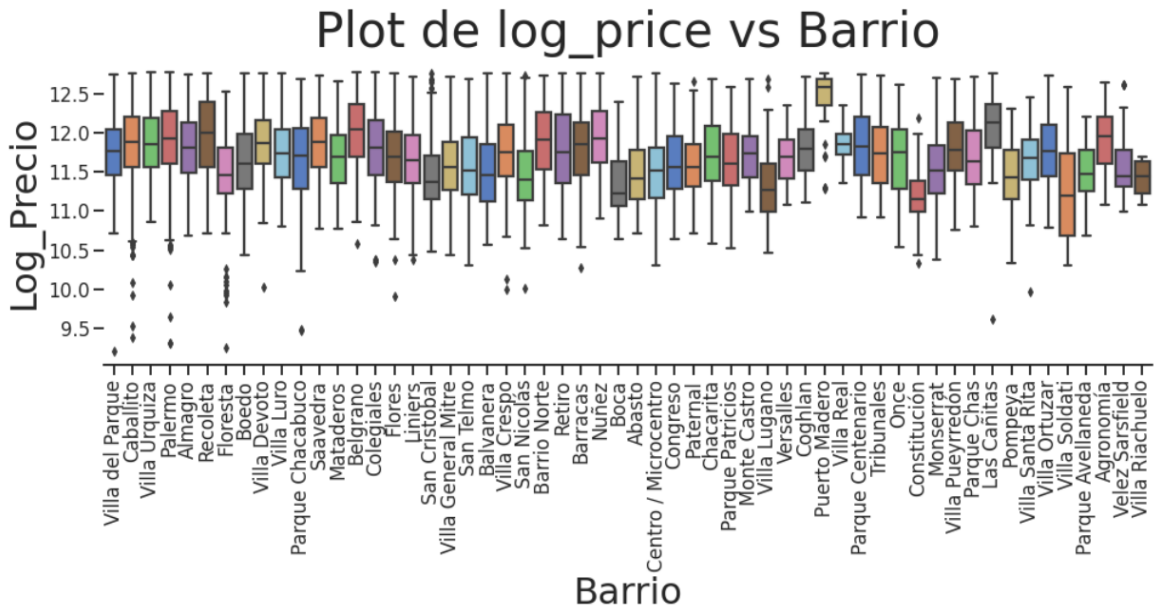
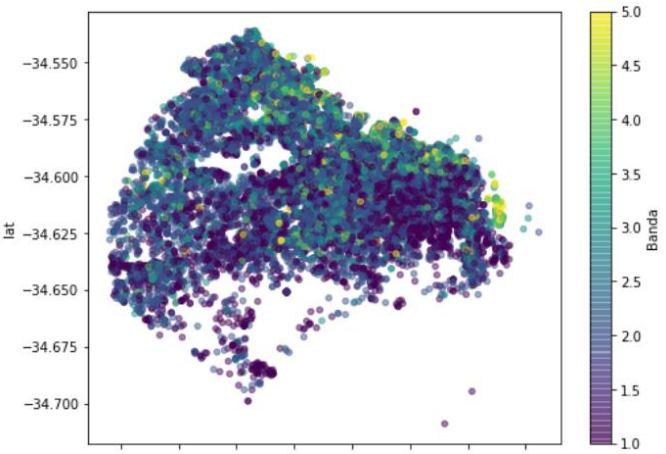


Gráfico 2: Box Plot en función de la variable “Barrio”.. Elaboración propia.

Continuamos agrupando los precios en 7 bandas, y luego con dichas bandas, y los datos de latitud y longitud, representamos las publicaciones en un mapa, con una escala de colores para visualizar las bandas.



Más allá de casos aislados, se observa que los precios más elevados se encuentran concentrados en el corredor norte de la ciudad y Puerto Madero, mientras que las propiedades más económicas están en la zona sur.

Sin embargo, cada zona internamente es heterogénea y es posible encontrar precios que pertenecen a todas las bandas.

Gráfico 3: Scatter plot en función de “Banda”.. Elaboración propia.

A continuación, observamos el top 10 de cantidad de publicaciones por barrio, siendo Palermo el líder en dicho ranking, seguido muy parejo por Caballito y Belgrano.

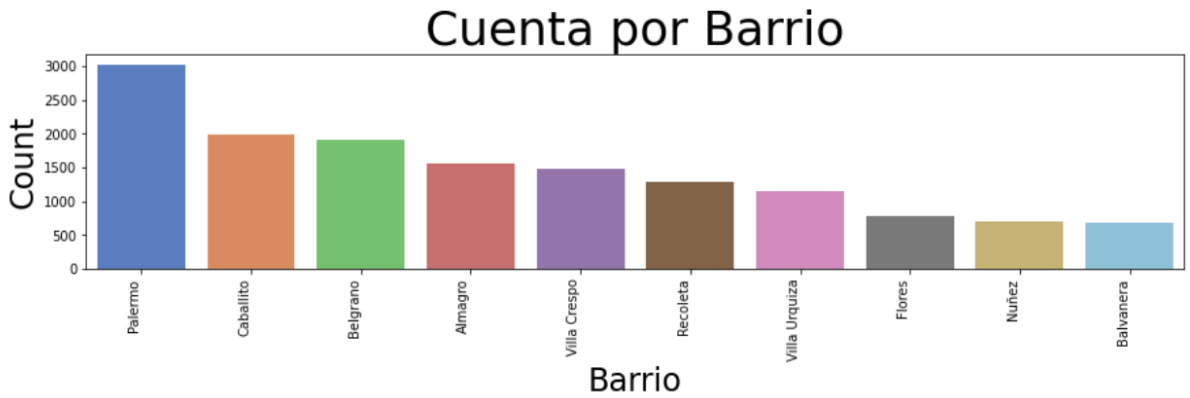
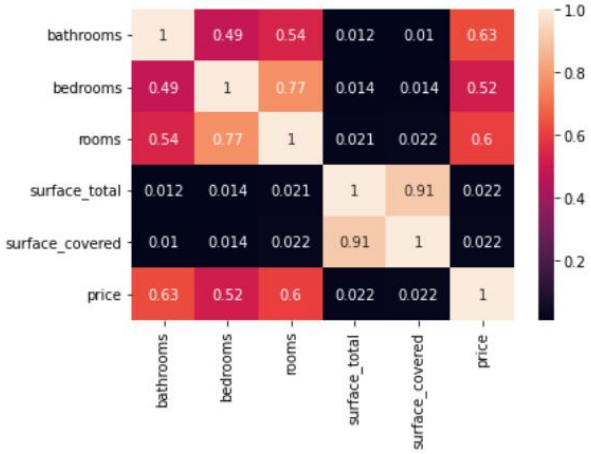


Gráfico 4:Count plot de la variable “Barrio”. Elaboración propia.

Por último, realizamos una matriz de correlación para analizar cómo se correlacionan las variables numéricas.



Es posible observar que las correlaciones más altas se encuentran en cantidad de baños, dormitorios y ambientes, respecto a la variable precio.

Esto nos permite inferir que a mayor cantidad de dichas variables, el precio de la propiedad será mayor.

Gráfico 5: Matriz de correlación. Elaboración propia.

Materiales y métodos

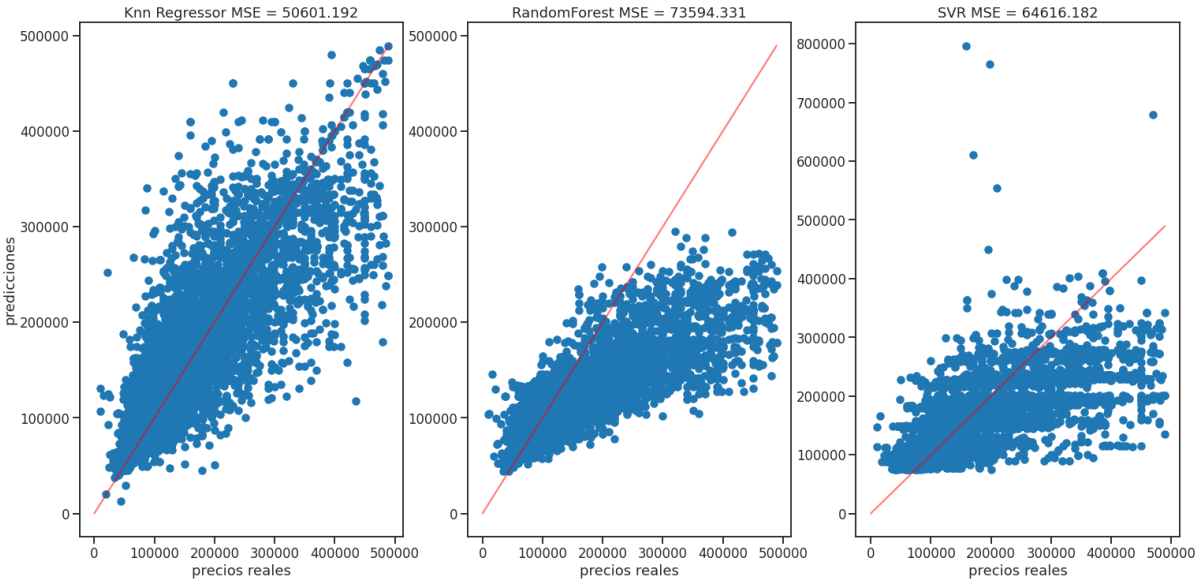
Los algoritmos utilizados para predecir la variable precio fueron los siguientes:

- **Knn Regression**: k-nearest-neighbour es un modelo que busca los datos más similares por cercanía y clasifica los mismos basado en la mayoría de los datos que lo rodean. Luego predice nuevos puntos a partir de dicha clasificación/etiquetado.
- **Random Forest Regression**: algoritmo que se basa en realizar preguntas binarias, obteniendo árboles de decisión hasta llegar a la decisión final siguiendo las features que maximicen la ganancia.
- **Support Vector Regression**: es un modelo que construye una función lineal (hiperplano) y determina un margen como función de costo, tratando de que todas las muestras estén dentro del mismo.

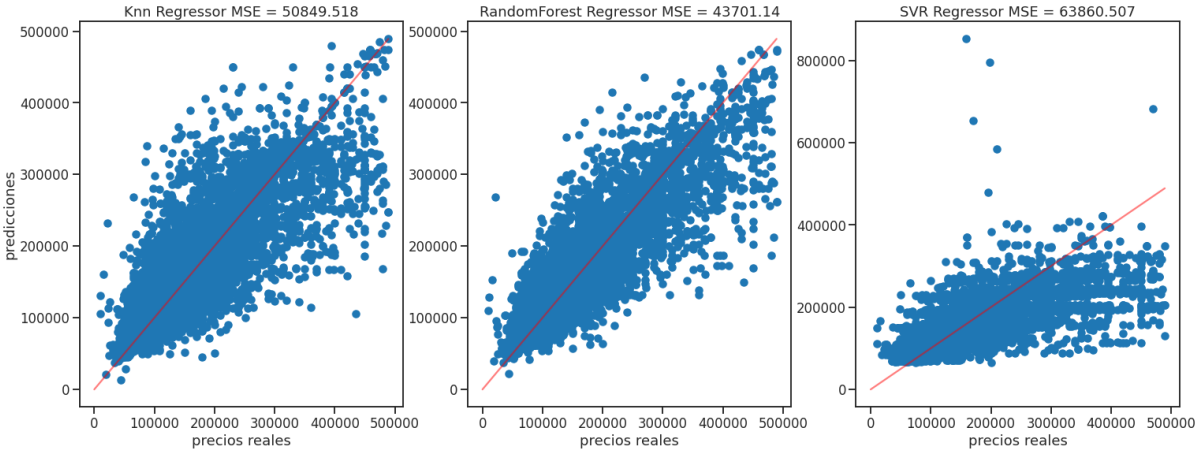
Además, se utilizó la técnica Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del dataset y luego volver a predecir.

Experimentos y resultados

El primer experimento se realizó generando las variables dummies para las variables categóricas, luego se dividió el dataset en train y test (70% train, resto test) y finalmente se aplicaron los 3 modelos mencionados con sus respectivos resultados:



Luego aplicamos la técnica de reducción de la dimensionalidad, que se conoce como PCA, y nos quedamos con las dos componentes principales que acumulan el 97% de la variabilidad. Finalmente volvimos a ejecutar los modelos anteriores con los siguientes resultados:



Resultados consolidados:

Modelo	MSE	MSE con PCA
Knn Regressor	50601,192	50849,52
RandomForest Regressor	73594,331	43616,75
SVM Regressor	64616,182	63860,51

Discusión y conclusiones

En primer lugar, en cuanto al error con las reducciones de la dimensionalidad de los modelos Knn y SVR, no se distinguen muchos cambios. Se observa en el modelo Knn, luego de aplicar PCA el aumento del error de la predicción, lo cual implica que al reducir la dimensionalidad se han perdido variables representativas de los registros que permitan una predicción precisa. En contraposición el modelo de Support Vector logra achicar un poco el error con sus variables representativas.

Se concluye que el modelo más adecuado para predecir el precio es Random Forest Regressor, ya que luego de aplicar PCA entrenó el modelo en un espacio latente dando mejor resultado que el Knn Regressor, dado que el modelo de Random Forest es el que mejor se ajusta a los datos y el cual permite generalizar de la mejor forma el problema, ya que minimiza el error al recibir nuevos registros no conocidos, teniendo la menor distancia entre la predicción y el valor real.

Referencias

- Tibshirani et. al (2021), An Introduction to Statistical Learning (Springer, 2° edición, pg n° 252, “Principal Components Regression”)
- Vanderplas, Jake (2016), Python Data Science Handbook (O’Reilly)
- Liaw, Andy et. al (2001), Classification and Regression by RandomForest