Problem Set 3 – Grupo 2

Alejandro García 201810340 - Andrés Torres 202115992 - Ignacio Serrano 201618789

https://github.com/AndresT2022/ProblemSet3.git

Introducción

El presente documento pretende realizar la predicción del precio de las viviendas en Chapinero (Bogotá) y El Poblado (Medellín) por medio de datos espaciales suministrados por el docente y datos espaciales obtenidos en OpenStreetMap.

Se realizaron pruebas con 5 modelos (dos XGboost, dos Random Forest y un OSL) del cual se definió escoge el modelo bajo la metodología XGboost que utiliza todas las variables planteadas.

Datos

De los datos suministrados por el docente, se filtró la información para El Poblado y Chapinero, obteniendo las siguientes cantidades de observaciones:

Tabla 1. Cantidades de observaciones

Lugar	**Test**	**Train**
Chapinero	478	2100
El Poblado	4463	757
Total	**4941**	**2857**

Revisando la literatura asociada a la predicción del precio de viviendas, se encontró que el área de la vivienda y el número de baños son variables que explican muy bien el precio de la vivienda, por lo cual, se realizó una imputación de estas variables con ayuda de la descripción que trae la base de datos y realizando buffer´s, lo anterior debido a que estas variables presentaban muchos missing values. Adicional a las dos variables mencionadas, se utilizó el número de habitaciones, la cual no tenía missing values.

Por medio de OpenStreetMap se incorporaron tres variables al modelo, las cuales son la distancia mas cercana al parque, al bar y a las zonas de comercio. En ese orden de ideas, se utilizaron cinco variables en total para los modelos de predicción propuestos.

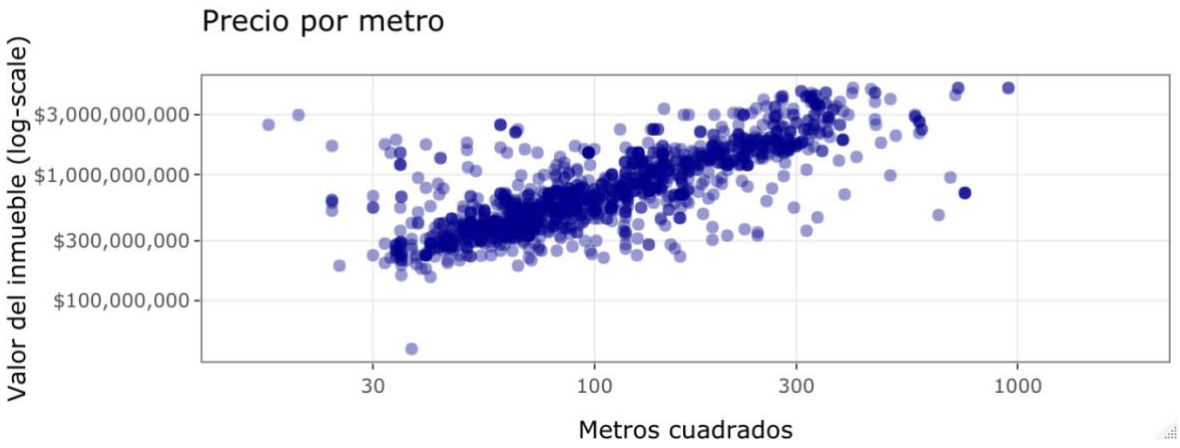
Tabla 2. Estadísticas descriptivas - Base de datos suministrada

	Test		**Train**	
Num observaciones:		4941		2857
Variables:	Cantidad NA´s	Media	Cantidad NA´s	Media
bedrooms	0%	2.93	0%	2.68
bathrooms	50%	3.13	25%	2.9
surface_total	82%	207	47%	199
surface_covered	87%	157	82%	129

En la siguiente gráfica se muestra el precio contra el metro cuadrado del inmueble, en donde se observa una relación lineal entre ellos. Sin embargo, él área de un inmueble no es la única variable que explica la totalidad del precio. Por lo anterior, se incorporan las variables de

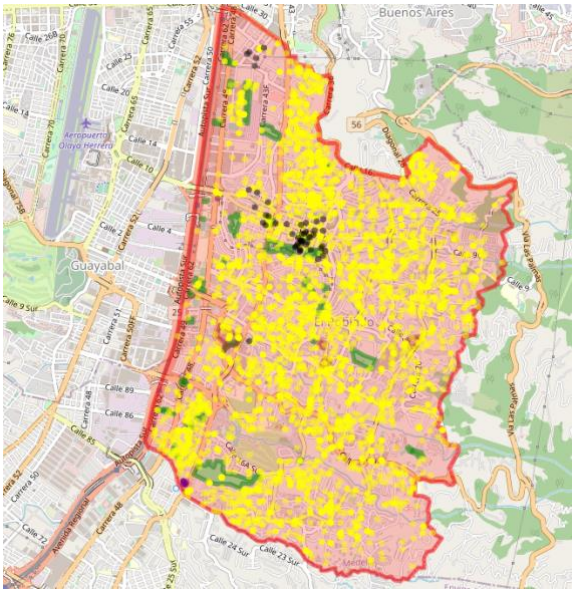
OpenStreetMap y las otras dos variables de la base de datos suministrada, ya que consideramos que pueden afectar el precio de las viviendas por espectativas del mercado.

Gráfica 1. Comparación precio y área vivienda

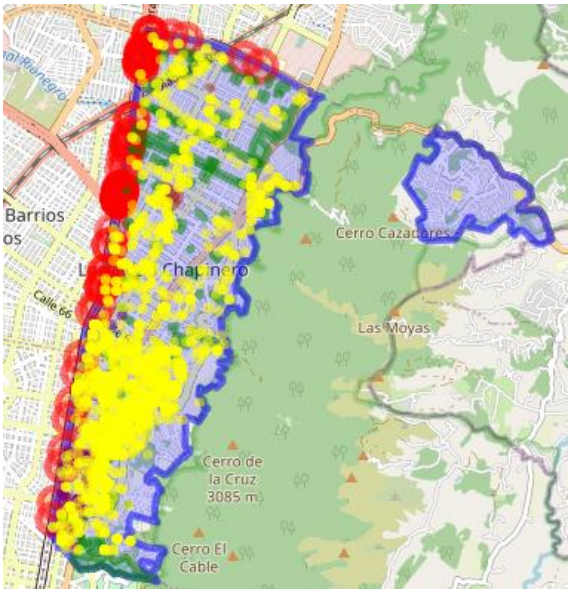


En el mapa 1, se muestra la distribución de la muestra (amarillo), los parques (verde), los bares (en negro) y las zonas de comercio (café) para El Poblado. En mapa 2, se muestra la distribución de la muestra (amarillo), los parques (verde), los bares (en negro), las estaciones de buses (rojo) y las zonas de comercio (café) para Chapinero.

Mapa 1. El Poblado



Mapa 2. Chapinero



Modelos y resultados.

$$\boldsymbol{Precio = Area + Num\_Baños + Num\_habitaciones + Dist\_parque + Dist\_bares + Dist\_comercio}$$

Se corrieron cinco modelos, uno OLS, dos Random Forest y dos XGboost. El modelo escogido corresponde a uno de los modelos evaluados con XGboost, el cual contiene todas las variables de la ecuación descrita anteriormente.

La regresión por OSL nos permitió identificar la dirección del efecto de las variables, que es coherente con las premisas del mercado inmobiliario, indicando que, a mayor cantidad de área, de baños y de habitaciones, se tiene un precio mayor, a mayor distancia de los bares el precio aumenta, y a mayor distancia de parques y comercio, el precio disminuye

Se compararon los RSM de cada modelo, siendo el de menor valor el modelo Random Forest, pero se decide escoger el modelo XGboost bajo el criterio de que el modelo Random Forest puede sobreajustarse sobre los datos de entrenamiento.

Conclusión y recomendaciones.

Los resultados del modelo seleccionado, que evita un sobreajuste, nos muestran que, dadas las distancias a puntos estratégicos, y las amenidades dentro de los inmuebles, el precio de los mismos aumenta conforme mas de estas amenidades tengan y menos distancia a puntos comerciales. Las distancias a parques, aunque muestran cierta relación con el precio, no ponderan tanto como las distancias a centros comerciales.

En general los precios se ajustan a la tendencia al mercado inmobiliario, las predicciones realizadas presentan un valor mínimo de 199 Mill COP, un máximo de 3,709 Mill COP y un promedio de 1,284 Mill COP. Se resalta que el modelo se guio por el indicador MSE dada la naturaleza de las variables (numéricas).