Taller 3.
¿Ganar dinero con ML?
"Se trata de la ubicación...!"

Integrantes: Isabella Mendez Pedraza. Cód.: 201814239
Manuela Ojeda Ojeda. Cód.: 201814476
Juan Sebastián Tellez Melo. Cód.: 201513710
Andrés Mauricio Palacio Lugo. Cód.: 201618843

Link del repositorio: https://github.com/AndresMPL/Repositorio_PS3.git

1. Introducción.

El problema abordado en el presente documento consiste en estimar un modelo que permita predecir los precios de compra de casas y apartamentos en la localidad de Chapinero en Bogotá, con información de ventas previas obtenidas de https://www.properati.com.co. Esta información contiene datos de los inmuebles como área total del inmueble, área cubierta, número de habitaciones (entre dormitorios y baños), habitaciones, baños, tipo de propiedad (casa o apartamento), ubicación, descripción y precio. Sin embargo, este ejercicio plantea un reto adicional y es que los datos de entrenamiento carecen en un gran porcentaje de la información de áreas, principalmente, y que no existen muchos datos de la Localidad de Chapinero.

Este reto planteó la necesidad de integrar nuevas variables al modelo que fueron extraídas de páginas de internet como Datos Abiertos de Bogotá y Open Street Maps y de la información disponible en la descripción de cada inmueble. Inicialmente, se buscó integrar al modelo determinantes geográficos para el precio de la vivienda, a partir de lo que plantean *Glaeser y Gyourko (2018)* en *The Economic Implications of Housing Supply*, al señalar que las condiciones geográficas pueden diferenciar los costos de producción de viviendas. Posteriormente, se buscó extraer datos como las áreas y si contaba con parqueadero o no a partir de la descripción usada para la venta de dicho inmueble.

Los resultados mostraron que la ubicación de los inmuebles cerca a lugares como colegios, parques o estaciones de transporte tienen un gran impacto en la determinación de los precios de venta de los inmuebles.

2. Datos.

2.1. Descripción de los datos

El modelo de predicción de precios en este ejercicio utilizó para su entrenamiento una base de datos ("*Train*") de casas y apartamentos que se encuentran en venta y que consta de 38.644 registros de información con variables como como el precio del inmueble, áreas del inmueble, número de habitaciones, baños y su ubicación geográfica.

Al analizar la base de datos se identificó que solamente el 20% de los registros contenían información del área total del inmueble, 22% contenía información del área cubierta, ,52%

contenía información del número de habitaciones (entre dormitorios y baños) y 73% contenía información del número de baños de cada inmueble, lo que evidenció un gran número de registros de información incompletos y generó la necesidad de agregar nuevas variables con el fin de intentar predecir los precios de venta de estos inmuebles. Esto obedece también al impacto que tienen características como los datos de las áreas de cada inmueble y su ubicación geográfica en la determinación del precio de venta.

En primer lugar, se utilizó la página de Datos Abiertos de Bogotá para descargar la información de ubicación de Parques, Museos, Centros Médicos (entre IPS privadas y Hospitales Públicos), Colegios, Centros de Atención Inmediata (CAI-Policía), Biblioestaciones, Centros Financieros, Número de delitos por localidad (incluye todo tipo de delitos) y ubicación de los cuadrantes de policía, y de la página de Open Street Maps se descargó la información de Paradas de Buses y Estaciones de Transmilenio.

Posteriormente, se calculó la distancia más corta de cada inmueble a cada una de estas ubicaciones descargadas, utilizando sus puntos de ubicación o centroides de los polígonos, según correspondía, y cada distancia se incluyó en la base de datos como una variable adicional, es decir, fueron incluidas 11 nuevas variables del entorno geográfico de cada inmueble.

En segundo lugar, a partir de las descripciones se buscó obtener la información sobre los metros cuadrados de cada inmueble. Para esto se generaron bigramas por medio de los cuales pudimos aproximarnos a la identificación de las áreas de estos inmuebles cuando las descripciones hacían referencia a metros, mt2, mt o similares. Adicionalmente, pudimos recolectar información acerca de si la propiedad contaba con parqueadero o no por medio de la *tokenización*, asignando un 1 si el inmueble contaba con parqueadero y 0 de lo contrario y esta variable fue utilizada como un factor en el modelo.

Ahora bien, en los casos en los cuales hacía falta una importante cantidad de información de habitaciones, dormitorios, baños o áreas se realizó imputación con el fin de generar estimaciones considerando estos ajustes y su impacto.

El proceso de limpieza de la muestra incluyó, adicionalmente, la verificación de ubicación de los inmuebles, asegurando que todos contenían los datos de longitud y latitud y que todos estuviesen ubicados en Bogotá, así mismo, que todas las operaciones correspondían a ventas y en este proceso no fueron eliminados registros.

El proceso mencionado aquí se desarrolla en los scripts "1_Cleaning", "2_Geography" y "3_Distances", que se encuentran en el repositorio de Git Hub.

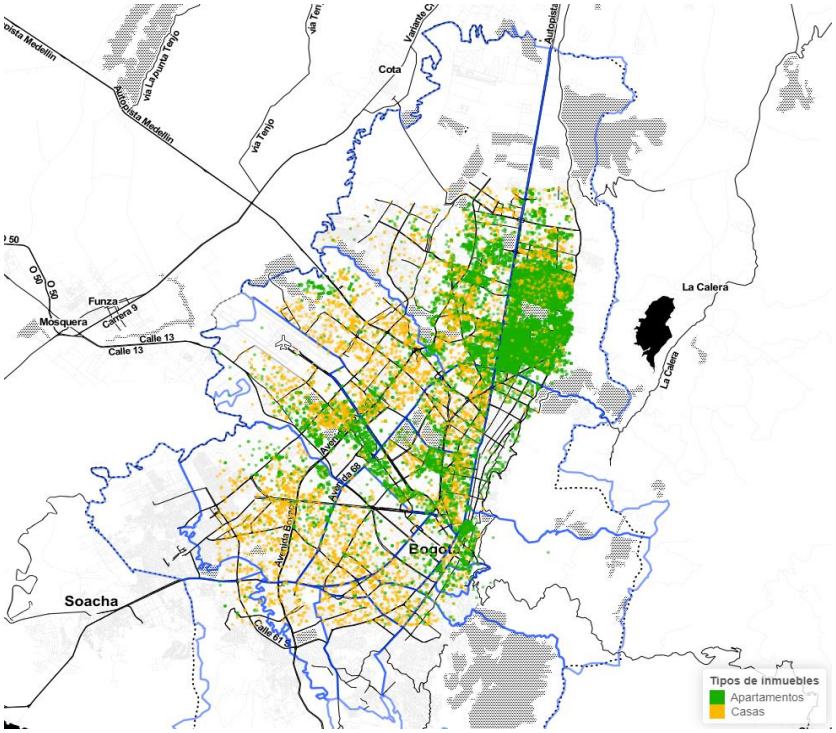
2.2. Estadísticas descriptivas

Luego de terminar la limpieza de los datos, se procedió a analizar la distribución de la información disponible, así como la ubicación geográfica de dichos datos.

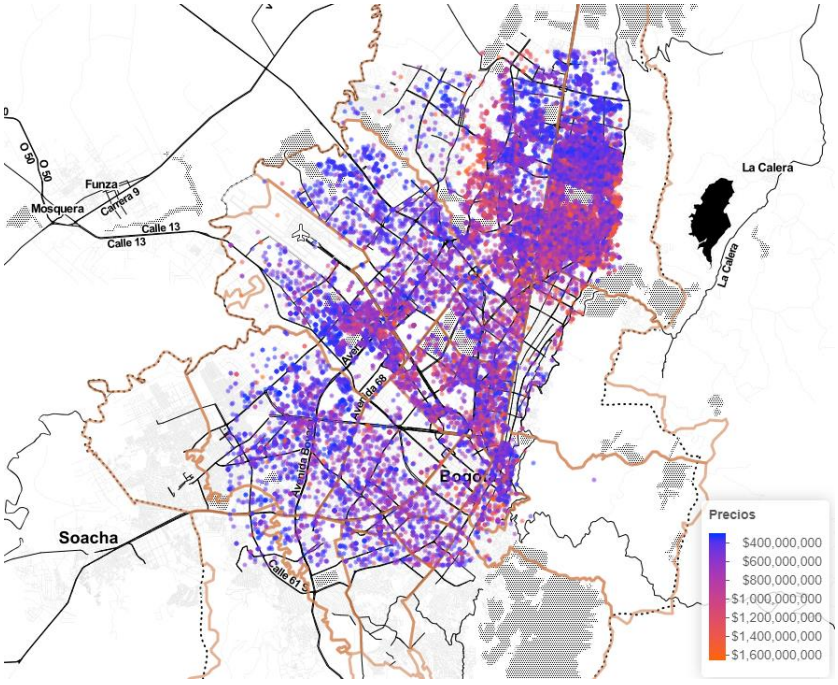
El Mapa 1 muestra la distribución de los inmuebles en la ciudad de Bogotá, en 18 localidades, diferenciando cada inmueble según sea Casa o Apartamento. El Mapa 2 muestra la

distribución de los precios de estos inmuebles; de esta manera, se identificó que el 76% de los datos corresponden a Apartamentos y el 24% a Casas, esta variable fue considerada también dentro de los modelos de predicción como un factor.

Mapa 1. Distribución de inmuebles entre Casas y Apartamentos



Mapa 2. Distribución de inmuebles según precio



Así mismo, se identificó que los precios de venta oscilan entre $300 millones y $1.650 millones, con un valor medio de $655 millones, tal como se detalla en la Tabla 1, y que los inmuebles con los mayores precios de venta se encuentran ubicados alrededor de la localidad de Chapinero, como se observa en el Mapa 2.

Por otra parte, se identificó que los precios presentan una alta distribución hacia el valor medio, tal como se observa en la figura 1, sin embargo, para efectos del presente ejercicio no se consideró la eliminación de *outliers* puesto que, al analizar el logaritmo de los precios, se encontró que los datos estaban distribuidos alrededor de la mediana y entre el primer y tercer cuartil, tal como se evidencia en la figura 3.

Figura 1. Histograma de distribución de los precios

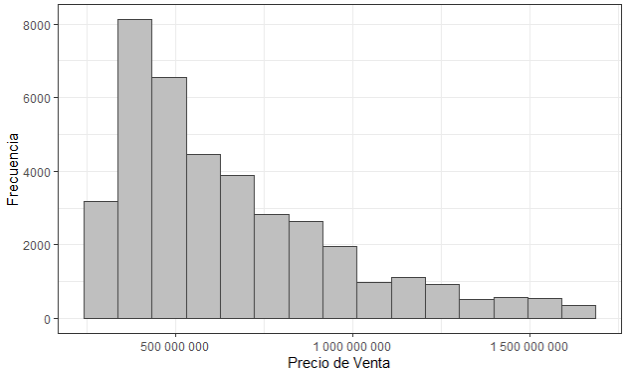


Figura 2. Boxplot de distribución de los precios

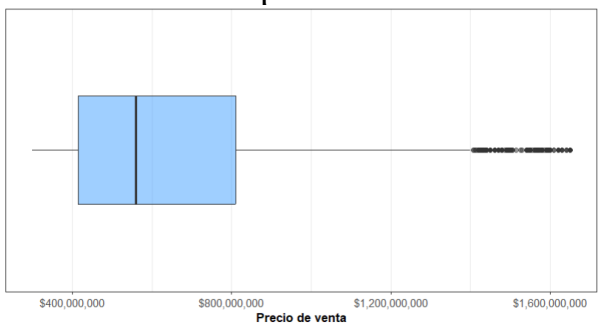
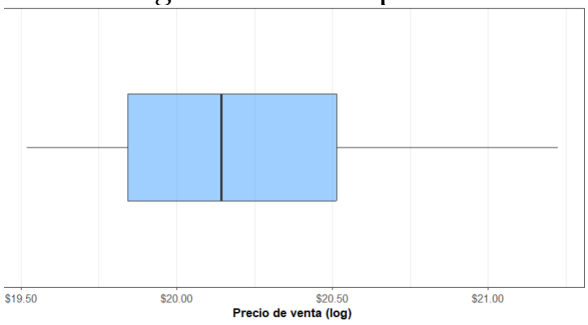


Figura 3. Boxplot de la distribución logarítmica de los precios



Ahora bien, el ejercicio de identificación de áreas a partir de las descripciones genera un promedio de 1.350 m^2 de área total y el ejercicio de imputación generó un promedio de 3 habitaciones (entre dormitorios y baños), lo que se puede detallar en la tabla 1. En esta misma tabla, se encuentran los valores medios de la distancia de cada inmueble a las nuevas variables geográficas de su entorno. Por ejemplo, se identificó que, en promedio, cada inmueble se encuentra a 313 m del colegio más cercano, 138 m del parque más cercano, 200 m del centro de salud privado más cercano y a 1.261 m del hospital público más cercano, allí mismo se pueden consultar las distancias medias a los CAI, paradas de buses, estaciones de Transmilenio, entre otros lugares.

Tabla 1. Estadísticas descriptivas de los datos de entrenamiento

Statistic	N	Mean	St. Dev.	Min	Max
Precio	38,644	654,534,675.00	311,417,887.00	300,000,000	1,650,000,000
Área total	38,644	153.99	123.69	16	17,137
Área cubierta	38,644	131.99	36.07	2	1,336
Habitaciones	38,644	3.00	1.00	1	11
Dormitorios	38,644	3.14	1.53	0	11
Baños	38,644	2.91	0.94	1	13
Distancia parque	38,644	138.72	84.40	1	2,311
Distancia museo	38,644	1,548.41	1,035.61	10	6,548
Distancia IPS	38,644	205.06	158.11	1	3,024
Distancia ESE	38,644	1,260.68	549.39	15	3,572
Distancia colegios	38,644	313.12	203.03	2	3,074
Distancia CAI	38,644	767.07	374.72	2	2,957
Distancia b.est.	38,644	3,885.35	1,712.44	32	7,764
Distancia centrof	38,644	3,451.07	2,367.86	10	12,983
Distancia cuadrantes	38,644	390.29	188.39	4	2,152
Distancia buses	38,644	250.84	155.56	0	2,673
Distancia tm	38,644	887.24	632.35	4	4,114
Parqueadero	38,644	0.69	0.46	0	1
Mts2	13,098	1,350.05	4,205.84	0	76,610
Superficie total imp	7,854	153.95	274.37	16	17,137
Superficie cubierta imp	8,565	131.93	76.62	2	1,336
Habitaciones imp	38,644	3.14	1.53	0	11
Baños imp	28,573	2.88	1.09	1	13
Cuartos imp	20,384	3.01	1.37	1	11

3. Modelo y resultados.

El ejercicio de predicción de los precios de venta obedeció al siguiente modelo:

$$P_i = \beta_0 + \beta_p X_p + \varepsilon$$

La propuesta en este documento es que los precios de venta de los inmuebles se pueden predecir a partir de las características (X_p) que se detallan en el Anexo 1 - Listado de variables incluidas en los modelos.

Con estas variables se generaron modelos de regresión lineal y cuadrática con diferentes interacciones entre las variables, se regularizaron estos modelos utilizando Elastic Net, con una grilla de 20 lambdas , variando el parámetro alfa con 11 valores entre 0 y 1, utilizando validación cruzada con 5 y 10 folds y se generaron modelos utilizando Random Forest.

Cada uno de estos modelos se generó con el 70% de los datos de la BD Train y se evaluó en primer lugar en el 30% restante de los datos. Posteriormente, cada modelo se implementó en la base de datos de "*Test*", proporcionada para este ejercicio, la cual contaba con 10.286 registros y con las mismas variables de la BD Train, excepto la variable de Precio.

Con el fin de determinar los modelos que se cargarían en la plataforma Kaggle, se estimó en cada predicción el *Mean Absolute Error* (*MAE*), buscando que este resultado fuera lo más cercano a cero posible; con este criterio se cargaron 10 modelos y aquel con el puntaje de calificación más alto en Kaggle fue el que se hizo mediante Random Forest y no utilizó los datos estimados de las áreas generadas de las descripciones ni los datos imputados.

Este modelo utilizó 26 variables, incluyendo interacciones entre las variables iniciales, generó 500 árboles y el mejor resultado tuvo como parámetros mtry = 5 y min.node.size = 10. Esto nos muestra que los principales determinantes, según este ejercicio, se basan en las condiciones geográficas del inmueble, es decir, su ubicación a lugares de interés como parques, bibliotecas, centros de atención médica, paradas de buses y el número de eventos de delito en la localidad que se encuentran ubicados .

El proceso aquí descrito se ejecutó mediante los scripts denominados "4_Models" y "5_Prediction".

4. Conclusiones y recomendaciones.

En este documento se buscó estimar un modelo que permitiera predecir los precios de compra de casas y apartamentos en la localidad de Chapinero de Bogotá, de la cual existía poca información. Una de las principales limitaciones que se observaron fue que existía un gran número de registros de información incompletos, lo que hizo necesario integrar determinantes geográficos que pudieran afectar el precio de la vivienda como la distancia de los inmuebles a parques, museos, centros médicos, entre otros.

A partir de modelos principalmente de regresión lineal, Elastic Ret y Random Forest , se estimó en cada predicción el *Mean Absolute Error* (MAE), buscando que este resultado fuera lo más cercano a cero. Los resultados mostraron que el modelo de Random Forest fue con el cual obtuvimos la mejor predicción, utilizando 500 árboles.

Los resultados obtenidos muestran que los principales determinantes de los precios de venta de los inmuebles se basan en las condiciones geográficas del inmueble, es decir, su ubicación a lugares de interés como parques, bibliotecas, centros de atención médica, paradas de buses, el número de eventos de delito en la localidad que se encuentran ubicados, entre otras .

5. Referencias.

Edward Glaeser and Joseph Gyourko (2018). *The Economic Implications of Housing Supply*. The Journal of Economic Perspectives , Vol. 32, No. 1 (Winter 2018), pp. 3-30 Published by: American Economic Association URL: https://www.jstor.org/stable/10.2307/26297967.

6. Apéndice.

Anexo 1. Listado de variables incluidas en los modelos.

- Área total del inmueble
- Área cubierta del inmueble
- Número de habitaciones
- Número de baños
- Número de dormitorios
- Distancia al parque más cercano
- Distancia al museo más cercano
- Distancia a la IPS[1] más cercana
- Distancia a la ESE[2] más cercana
- Distancia al colegio más cercano
- Distancia al CAI más cercano
- Distancia a la biblioestación más cercana
- Distancia al centro de referencia más cercano
- Distancia a la parada de bus más cercana
- Distancia a la estación de Transmilenio más cercana
- Área total y área cubierta
- Número de habitaciones
- Número de baños
- Número de dormitorios

[1] Institución Prestadora de Servicios de Salud

[2] Empresa Social del Estado