**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.**

* Descripción breve del problema y si hay antecedentes.
* Descripción de los datos y su idoneidad para abordar el problema.
* Vista previa de los resultados y las principales conclusiones.

Con el objetivo de comprar la mayor cantidad de propiedades en la localidad de Chapinero de Bogotá gastando lo menos posible, en el presente trabajo se busca desarrollar un modelo predictivo de precios de venta. Debido a que, aunque existe información sobre de las propiedades en Bogotá, hay pocos datos sobre estos inmuebles en Chapinero.

Para esto utilizamos información sobre las propiedades que están a la venta en Properati. Que contiene información sobre el área total del inmueble, área cubierta, número de habitaciones (entre dormitorios y baños), habitaciones, baños, tipo de propiedad (casa o apartamento), ubicación, localidad, área máxima, descripción y precio. Sin embargo, existe un importante número de registros de información incompletos que hacen necesario utilizar información complementaria que nos permitan hacer mejores predicciones como la distancia más corta de cada inmueble a diferentes parques, museos, centros médicos, etc.

1. **Datos.** 
   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](https://datosabiertos.bogota.gov.co/) de Bogotá para descargar la información 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](https://www.openstreetmap.org/#map=13/4.6254/-74.1878) 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 identificar cuáles eran las áreas de estos inmuebles si las personas hacían referencia a metros, mt2, mt o similares en sus descripciones. Adicionalmente, pudimos recolectar información acerca de si el inmueble contaba con parqueadero o no por medio de la tokenización, asignamos un 1 si el inmueble contaba con parqueadero y 0 de lo contrario.

Ahora bien, en los casos en los cuales hacía falta una importante cantidad de información se realizó imputación para así poder tener unas mejores estimaciones como en el área total del inmueble, el área cubierta, ,número de habitaciones y baños.

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.

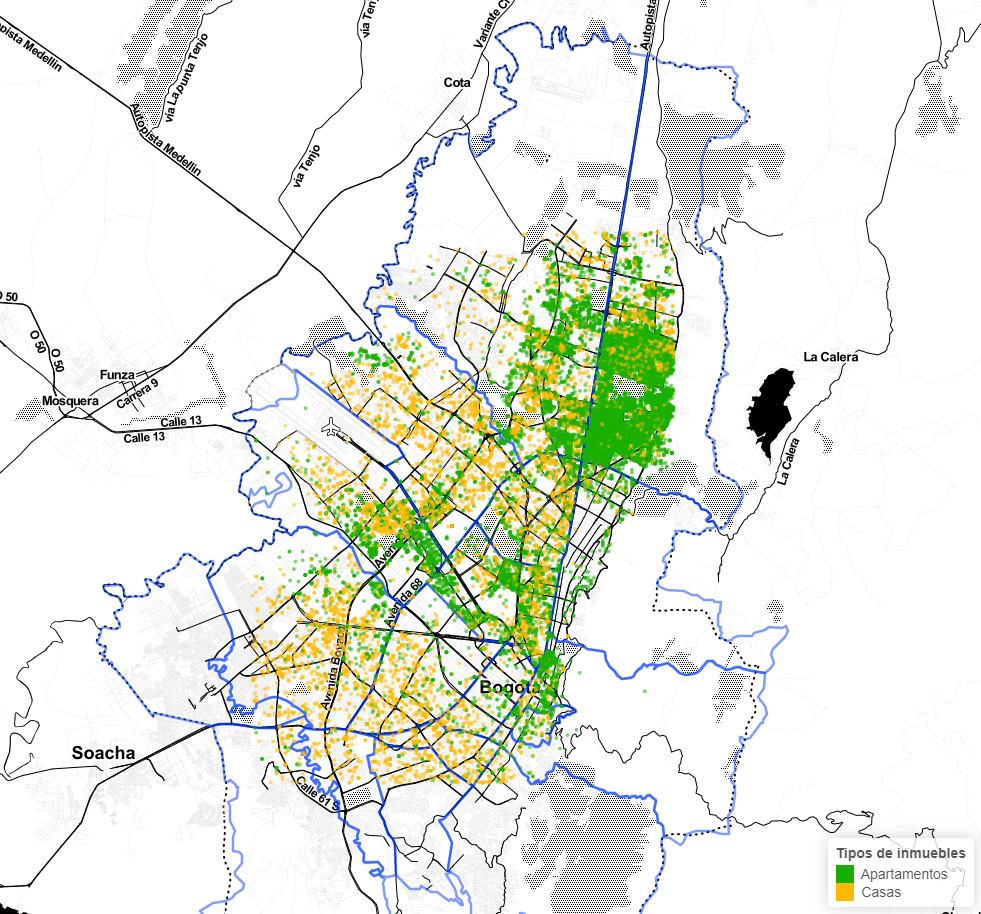
* 1. **Estadísticas descriptivas**

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.

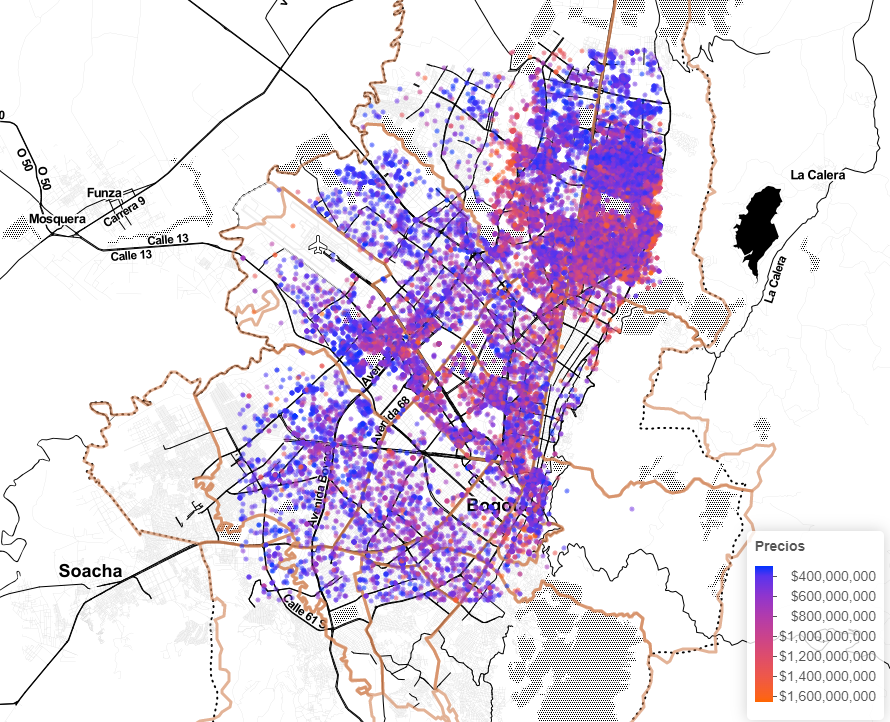
El Mapa 1 muestra la distribución de los inmuebles en la ciudad de Bogotá, en 18 localidades, diferenciando cada inmueble según sea Casa o Apartamento. El Mapa 2 muestra la distribución de los precios de estos inmuebles; de esta manera, se identificó que el 76% de los datos corresponden a Apartamentos y el 24% a Casas, esta variable fue considerada también dentro de los modelos de predicción como un factor.

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.

**Mapa 1. Distribución de inmuebles entre Casas y Apartamentos**

****

**Mapa 2. Distribución de inmuebles según precio**

****

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**

**Gráfico, Histograma

Descripción generada automáticamente**

|  |  |
| --- | --- |
| **Figura 2. Boxplot de distribución de los precios** | **Figura 3. Boxplot de la distribución logarítmica de los precios** |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente |

Ahora bien, se encontró que los inmuebles tenían un promedio de XX m2 de área total y un promedio de XX 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 mdel colegio más cercano, 138 mdel 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, biblioestaciones, paradas de buses, estaciones de Transmilenio, entre otros logares.

**Tabla 1. Estadísticas descriptivas de los datos de entrenamiento**

**Tabla 2. Correlación de variables en datos de entrenamiento**

1. **Modelo y resultados.**

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

La propuesta en este documento es que los precios de venta de los inmuebles se pueden predecir a partir de las siguientes características ():

* 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]](#footnote-1) más cercana
* Distancia a la ESE[[2]](#footnote-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

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 y variando el parámetro alfa con 11 valores entre 0 y 1, de igual manera se generó un modelo Super Learner, utilizando Bosques Aleatorios y Regresión Lineal.

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 se acerca a cero; con este criterio se cargaron X modelos y aquel con el puntaje de calificación más alto en Kaggle fue XXXX.

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

1. **Conclusiones y recomendaciones.**

1. **Bibliografía.**
2. **Apéndice de Anexos e Imágenes.**

1. Institución Prestadora de Servicios de Salud [↑](#footnote-ref-1)
2. Empresa Social del Estado [↑](#footnote-ref-2)