Universidad de los Andes

Maestría Economía Aplicada: *Big Data and Machine Learning for Applied Economics*

**Grupo:** Laura Natalia Capacho, Sebastián David Beltrán y Yurani Gonzalez[[1]](#endnote-1)

**GitHub URL:** <https://github.com/sbeltro/G10_PS3>

***Problem Set 3: Making Money with ML?***

**1. Introducción:**

La estimación del precio de una vivienda es una tarea retadora pues es un bien cuyo valor se determina por un conjunto heterogéneo de factores. En este trabajo, se hicieron diversas aproximaciones para encontrar el mejor ajuste y además la mejor inversión. El modelo seleccionado de *Random Forest* se compone de un conjunto de árboles de decisión, y es un método que no utiliza hiperparámetros. Una de sus grandes ventajas es que no está sujeto a la sensibilidad de hiperparámetros, tiene una estimación interna relativamente exacta dada su forma de validación cruzada, funciona muy bien con grandes bases de datos y puede usarse para clasificación o predicción (Cánovas et al., 2016; Espinosa, 2020). No obstante, es difícil de interpretar gráficamente, puede sobreajustar si hay ruido, y si los predictores tienen distintos niveles puede generar sesgos (Espinosa, 2020).

**2. Datos**

Se utilizaron 2 bases de datos de Properati[[2]](#endnote-2), de estas seleccionamos 3 variables: habitaciones, baños y superficie (para *missings* se imputaron valores de información extraída de descripción y título, y la media en las manzanas correspondientes). Por otro lado, adicionamos 2 variables provenientes del título y la descripción: parqueadero y terraza/patio;y 2 variables de *Open Street Maps*: universidad y centro comercial (Anexo 1 para detalles).

En la tabla 1 se presentan las estadísticas descriptivas de la base *train,* y las gráficas 1 y 2, corresponden al mapa de las viviendas, centros comerciales y universidades en dicha muestra, en Bogotá D.C (Bog) y Medellín (Med), respectivamente. Contamos con una muestra de 107,567 viviendas, 86,211 en Bog, y 21,356 en Med. En particular, podemos ver que el precio promedio de las viviendas es superior en Bog (aprox. COP760 mn, frente al COP400 mn en Med), destacamos que la desviación estándar es alta, lo cual es comprensible si tenemos en cuenta que la muestra de propiedades es bastante heterogénea, desde pequeñas casas, hasta grandes edificios. En términos de la superficie o área total de la vivienda, en Bog el promedio es de 146 metros cuadrados (mts2), y en Med de 123 mts2.

Resaltamos que en ambas ciudades se tiene un promedio de 2 baños y 3 habitaciones por vivienda. En Bog el 67% de viviendas cuenta con parqueadero y el 53% con terraza o patio, frente al 62% y 63% en Med, respectivamente. Por último, la distancia a la universidad más cercana es similar en ambas ciudades, aproximadamente 1,000 mts, mientras la distancia al centro comercial más cercano es mayor en Med (aprox. 870 mts, frente a 670 mts en Bog).

**Tabla 1. Estadísticas descriptivas (base *train) \****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Ciudad** | |  |
| **Variables** | **Bogotá D.C.** | **Medellín** | **p-valor \*\*** |
| precio | 764,283,984 (713601204) | 405,774,604 (392863012) | <0.001 |
| baños | 2.68  (1.19) | 2.24  (1.00) | <0.001 |
| habitaciones | 3.08  (1.45) | 3.08  (1.08) | 0.002 |
| superficie | 146.68  (201.07) | 123.50  (213.46) | <0.001 |
| universidad | 1,033.15 (821.50) | 1,140.34 (1079.50) | <0.001 |
| centroComercial | 677.90  (751.93) | 873.81  (655.37) | <0.001 |
| parqueadero (=1) | 57,465  (66.7) | 13,209  (61.9) | <0.001 |
| terrazaPatio (=1) | 45,744  (53.1) | 13,396  (62.7) | <0.001 |
| N | 86,211 | 20,356 |  |

\* El dato corresponde al promedio y el valor entre paréntesis a la desviación estándar. Para las variables dummies de parqueadero y terrazaPatio, el dato corresponde al número de viviendas que cumplen con dicha característica, y el valor entre paréntesis indica la proporción sobre el total.

\*\* El p-valor corresponde a una prueba de diferencia de medias entre los dos grupos

**Fuente:** Properati, cálculos propios.

1. **Notas**

   Los códigos de los estudiantes del grupo son 202121025, 202121021 y 201212100, respectivamente. [↑](#endnote-ref-1)
2. https://www.properati.com.co [↑](#endnote-ref-2)