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).

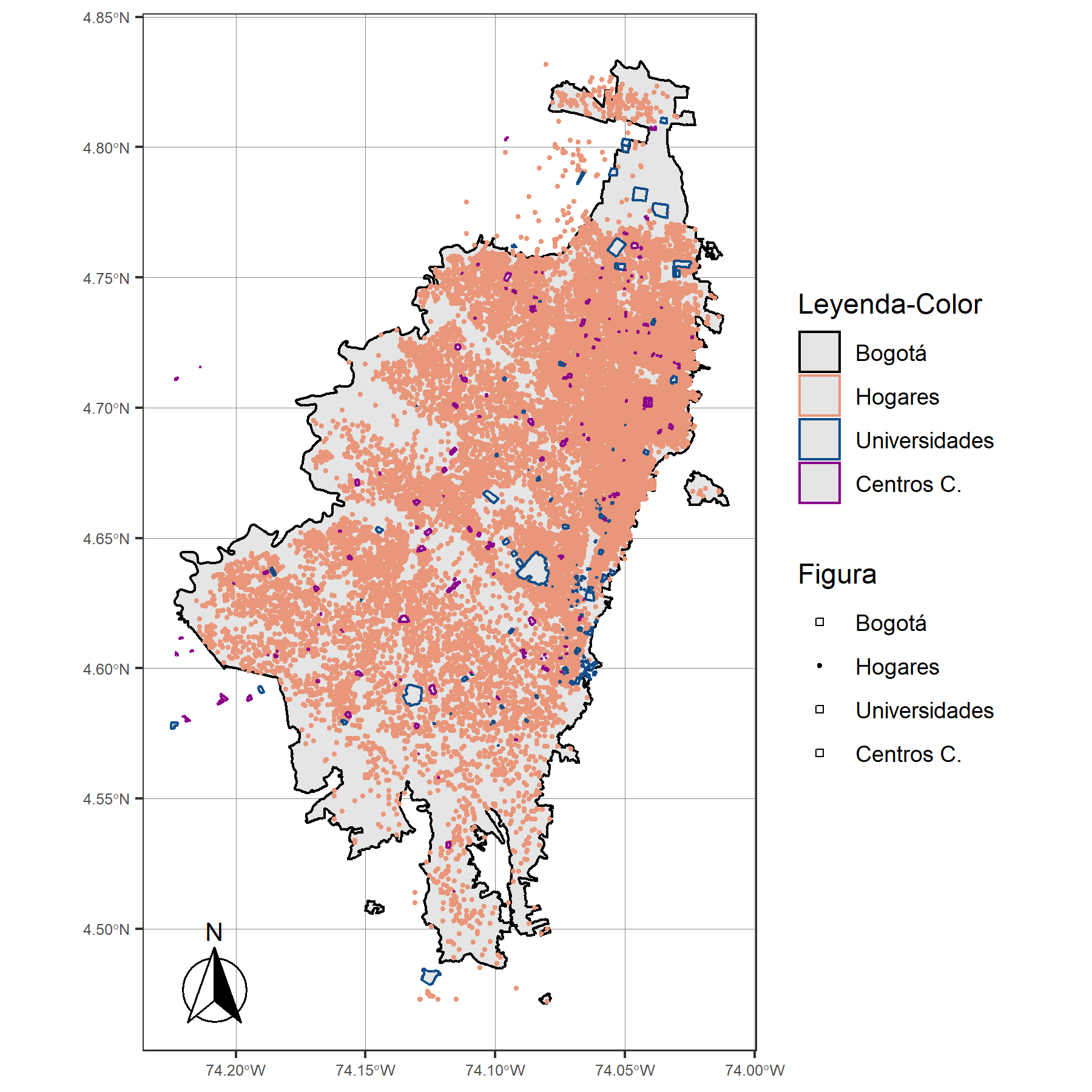
**Tabla 1. Estadísticas descriptivas (base *train) \****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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 |  |

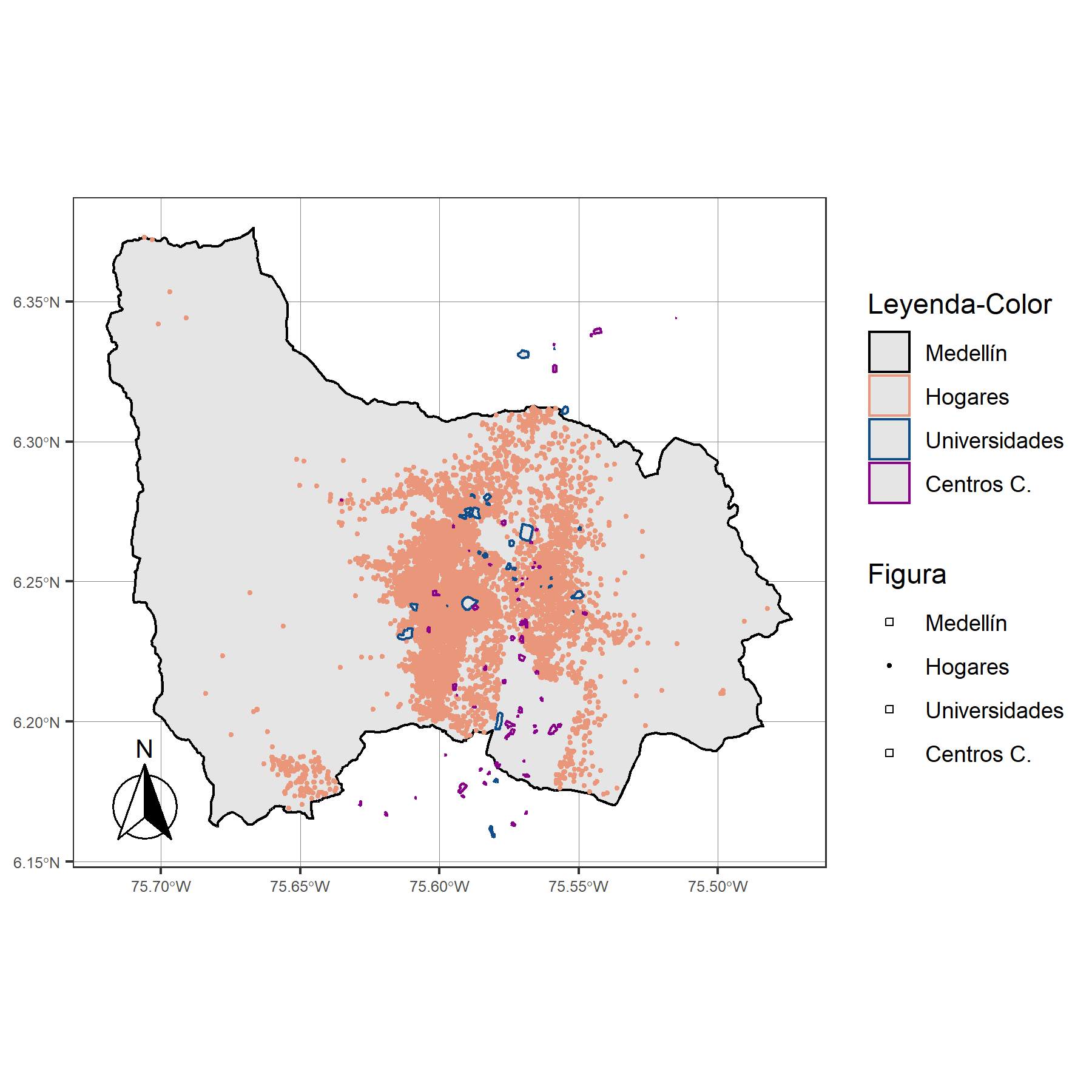
\* El dato corresponde al promedio y el valor entre paréntesis a la desviación estándar. Para las variables dummies de parqueadero y terrazaPatio, el dato corresponde al número de viviendas que cumplen con dicha característica, y el valor entre paréntesis indica la proporción sobre el total.

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

**Fuente:** Properati, cálculos propios.

****Gráfica 1.** Mapa Bogotá (*train)*

**Fuente:** Construcción propia.

****Gráfica 2.** Mapa Medellín (*train)*

**Fuente:** Construcción propia.

En la tabla 2 se caracterizan las viviendas a las cuales buscamos predecir su precio (*test),* y las gráficas 3 y 4, corresponden al mapa de las viviendas, centros comerciales y universidades en dicha muestra*.* La muestra se compone de 11,150 viviendas, 793 en la localidad de Chapinero, en Bog, y 10,357 en la comuna El Poblado, en Med. En cuanto a la superficie, en Chapinero el promedio es de 94 mts2, y en El Poblado, de 187 mts2. Las viviendas cuentan con un promedio de 3 baños y 3 habitaciones en El Poblado, superior al promedio de Chapinero. En cuanto a tenencia de parqueadero y terraza, El Poblado cuenta con una mayor proporción de propiedad con dichas características (68% y 64%, respectivamente). La distancia promedio a la universidad más cercana es inferior en Chapinero (aprox. 200 mts), y a un centro comercial es menor en El Poblado (aprox. 370 mts).

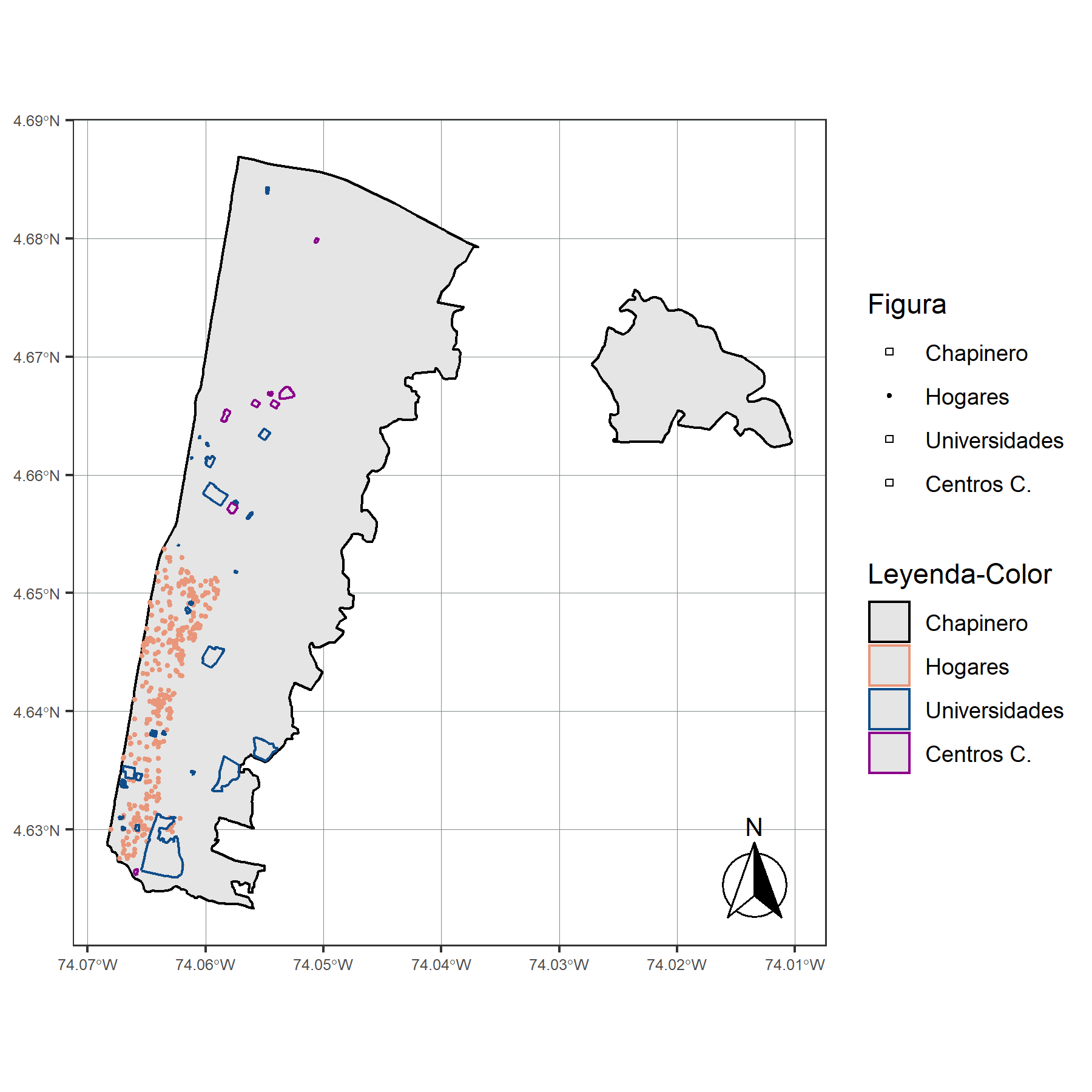
**Tabla 2. Estadísticas descriptivas (base *test) \****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Ciudad** | |  |
| **Variables** | **Bogotá D.C.** | **Medellín** | **p-valor \*\*** |
| baños | 1.90  (0.90) | 3.14  (1.14) | <0.001 |
| habitaciones | 1.91  (1.27) | 3.02  (0.91) | <0.001 |
| superficie | 94.64  (97.66) | 187.59  (228.67) | <0.001 |
| universidad | 207.54  (135.58) | 1663.02 (713.38) | <0.001 |
| centroComercial | 1036.54 (437.75) | 373.83  (252.46) | <0.001 |
| parqueadero (=1) | 490  (61.8) | 7088  (68.4) | <0.001 |
| terrazaPatio (=1) | 422  (53.2) | 6684  (64.5) | <0.001 |
| N | 793 | 10,357 |  |

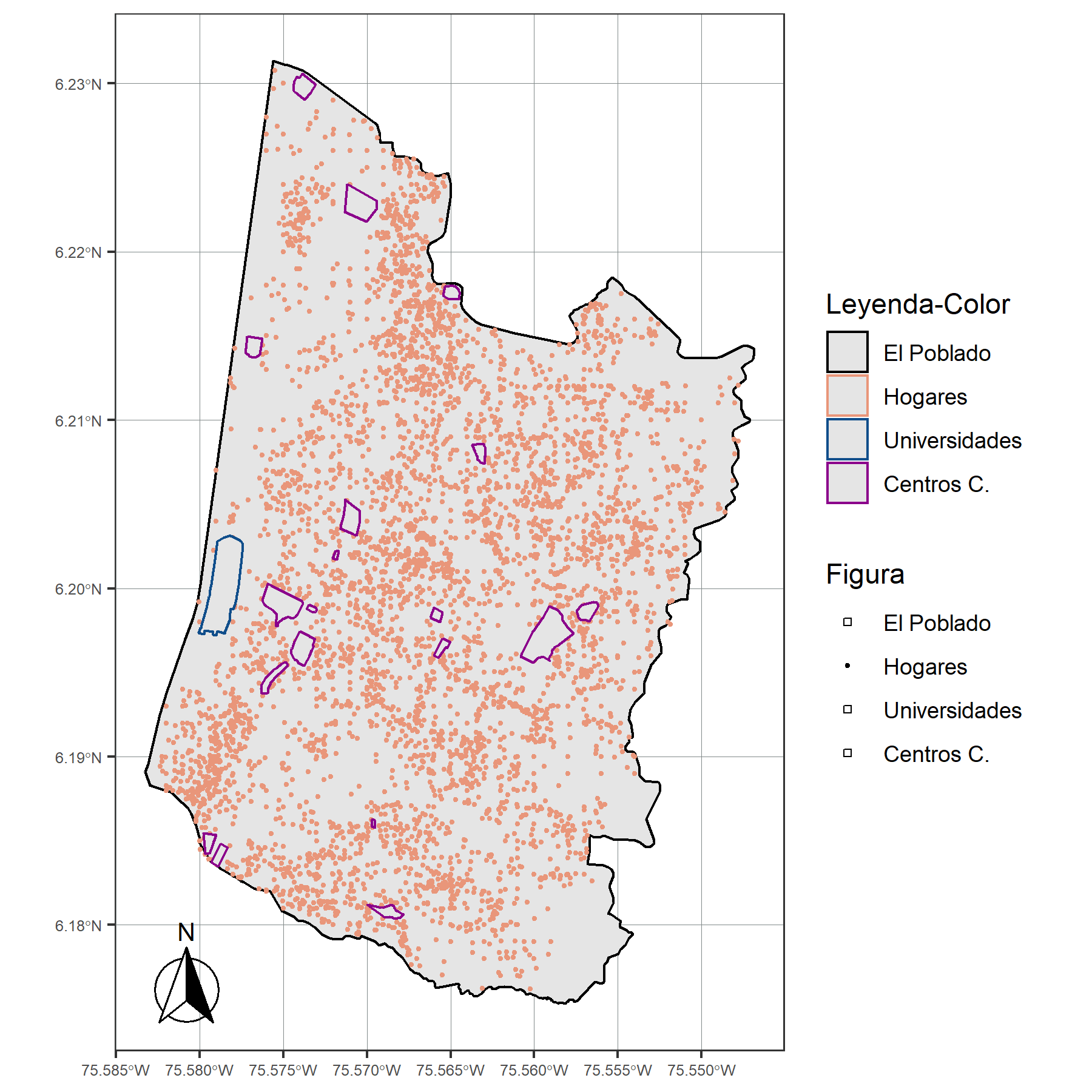
\* El dato corresponde al promedio y el valor entre paréntesis a la desviación estándar. Para las variables dummies de parqueadero y terrazaPatio, el dato corresponde al número de viviendas que cumplen con dicha característica, y el valor entre paréntesis indica la proporción sobre el total.

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

**Fuente:** Properati, cálculos propios.

****Gráfica 3.** Mapa Chapinero, Bogotá (*test)*

**Fuente:** Construcción propia.

****Gráfica 4.** Mapa El poblado, Medellín (*test)*

**Fuente:** Construcción propia.

**3. Modelo y resultados**

Se busca hacer una aproximación de un modelo predictivo del precio de una vivienda, para lo cual utilizamos 8 diferentes estimaciones. En concreto, exploramos un modelo de regresión lineal con validación cruzada en 5 conjuntos; 2 métodos de selección de modelos: *Best Subset Selection* (*BSuS)* y *Backward Stepwise Selection (BSwS)[[3]](#endnote-3)*; un modelo usando el método de regularización *Elastic Net[[4]](#endnote-4)* con validación cruzada en 5 conjuntos; un modelo de *Extreme Gradient Boosting* (*XGBoost*)[[5]](#endnote-5); una estimación con *Random Forest[[6]](#endnote-6)*; y una estimación con la técnica de *Superlearners[[7]](#endnote-7)*

**3.1 Variables**

Las variables que incluimos para entrenar el modelo seleccionado fueron aquellas descritas previamente, ya que luego de hacer las estimaciones, el modelo más robusto y que presenta el mejor ajuste es aquel que incluye todas las variables. Consideramos que incorporar todos los predictores es importante puesto que sabemos que el precio de una vivienda depende de diversos factores. Depende de características propias de la propiedad, por lo que incluimos número de habitaciones y baños, área total, y tenencia de terraza-parqueadero. Adicionalmente, incorporamos una variable de tipo de vivienda, que diferencia si corresponde a casa o apartamento (en este sentido varía la valoración de un inmueble).

De otro lado, el precio depende características del entorno, como la zona donde se encuentra ubicada la vivienda, y por esto se incluye una variable que identifica si la ciudad es Bogotá o Medellín, así como de la cercanía a universidades, ya que en ciudades principales es donde se encuentran las mejores universidades del país, y, por ende, es el destino de una cantidad importante de estudiantes y docentes. Igualmente, se incorpora la cercanía a centros comerciales, ya que estas zonas, por su importancia para dinamizar el comercio y el consumo pueden estar en mejores condiciones, y tener un mayor potencial de valorización.

**3.2 Descripción del modelo**

El modelo seleccionado corresponde a una estimación con *Random Forest*, donde la variable dependiente es el precio de la vivienda en logaritmo, y las variables independientes son el número de habitaciones, número de baños, superficie, universidad, centro comercial, parqueadero, terraza, la variable dummy de ciudad y de tipo de propiedad. La estimación utilizó validación cruzada en 5 conjuntos y se entrenó con la base *train,* con 107,567 observaciones (86,211 localizados en Bogotá y 21,356 en Medellín). Debido a la naturaleza del método, este no necesita de hiperparámetros que sensibilizan los resultados, sin embargo, se definió el número de árboles de decisión para considerar (# árboles).

**3.3 Medida de evaluación**

Para seleccionar el modelo utilizamos dos medidas de evaluación, por un lado, el **indicador precio promedio por vivienda (precioXpropiedad)**. La medición fue construida calculando para cada modelo el dinero total invertido si se compran las propiedades al precio estimado por cada modelo, considerando que únicamente se realiza la transacción si el precio estimado no subestima el precio real en más de COP40 mn, posteriormente, se encontró el número de propiedades compradas y se calculó la razón . Desde este enfoque, encontramos que el modelo que permite minimizar el indicador, es decir, adquirir el mayor número de viviendas por el menor precio, corresponde al modelo de *Random Forest (RF)* (Tabla 3)*.*

**Tabla 3. Indicador precio por vivienda**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **Dinero gastado** | **Propiedades compradas** | **Precio por propiedad** |
| M. 1 | 3.969646E+13 | 64,169 | 618,623,576 |
| M. 2 | 3.873203E+13 | 62,652 | 618,208,986 |
| M. 3 | 3.883675E+13 | 62,968 | 616,769,677 |
| M. 4 | 3.945605E+13 | 64,160 | 614,963,435 |
| M. 5 | 4.026321E+13 | 70,554 | 570,672,181 |
| M. 6 |  |  |  |
| M. 7 |  |  |  |

*\* M.1 es la estimación con regresión lineal y validación cruzada, M.2 con Best Subset Selection, M.3 con Backward Stepwise Selection, M.4 con Elastic Net y validación cruzada, M.5 con XGBoost, M.6 con Random Forest, M. 7 al SuperLearner con regresión lineal y Random Forest.*

**Fuente:** Cálculos propios.

De otro lado, se calculó el error cuadrático medio de cada modelo (Tabla 4), y se encontró, en línea con la comparación previa, que el modelo que reporta el mejor ajuste es *Random Forest*, pues reporta el menor error cuadrático medio. En la Gráfica 5 se observa que 4 modelos estimados tienen un error de ajuste significativamente más alto (regresión lineal, *BSwS, BSuS* y *Elastic Net)*, y dentro de las demás estimaciones, el modelo (6) es el mejor.

**Tabla 4. Error cuadrático medio**

|  |  |
| --- | --- |
| **Modelo** | ***MSE*** |
| (1) | 0,569070 |
| (2) | 0,618127 |
| (3) | 0,579928 |
| (4) | 0,569338 |
| (5) | 0,340858 |
| (6) |  |
| (7) | 0,248742 |

*\* M.1 es la estimación con regresión lineal y validación cruzada, M.2 con Best Subset Selection, M.3 con Backward Stepwise Selection, M.4 con Elastic Net y validación cruzada, M.5 con XGBoost, M.6 con Random Forest, M. 7 al SuperLearner con regresión lineal y Random Forest*

**Fuente:** 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)
3. *Best Subset Selection* combina los predictores disponibles y *Backward Stepwise Selection* (introduce todas las variables en la ecuación y excluye secuencialmente una tras otra) [↑](#endnote-ref-3)
4. Este método es una combinación de las penalizaciones que imponen sobre los coeficientes *lasso* y *ridge*, dos métodos de regularización, este modelo fuerza a que los coeficientes del modelo tiendan a cero, minimizado el riesgo de sobreajuste (overfitting), reduciendo la varianza, mitigando el efecto de correlaciones entre predictores y mejorando la estimación. [↑](#endnote-ref-4)
5. Es un método de estimación supervisado de *Machine Learning* que utiliza el principio de *boosting*, es decir, genera múltiples modelos de predicción “débiles”, árboles de decisión individuales, secuencialmente, y va generando un modelo más robusto y con mejor poder predictivo. [↑](#endnote-ref-5)
6. Una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento, y luego combina los resultados para obtener un modelo más robusto. Vale la pena resaltar que en este modelo los árboles crecen hasta su máxima extensión, mientras que en *XGBoost* esta es limitada. [↑](#endnote-ref-6)
7. Es un algoritmo de ensamble de *Machine Learning* que utilizando validación cruzada estima el desempeño de un conjunto de modelos especificados, para hacer la mejor predicción posible, a partir de un promedio ponderado óptimo de los modelos. [↑](#endnote-ref-7)