# Problem Set 3: Making Money with ML?

Big Data y Machine Learning para Economía Aplicada

Gustavo Adolfo Castillo Álvarez (201812166),
Alexander Almeida Ramírez (202225165),
Jorge Luis Congacha Yunda (201920042) y
Jaime Orlando Buitrago González (200612390)

2024-05-27

## 1 Introducción

En su *paper*, Rosen (1974) proporciona una aproximación metodológica (econométrica) para estimar precios hedónicos. Más específicamente, explora los equilibrios del mercado y las decisiones de los agentes y los efectos en el bienestar para evidenciar la relación entre los precios de los bienes y sus diferentes características. Esta fue una contribución importante para estimar los precios (Hedónicos) de los bienes a partir de las características o atributos (precios implícitos) a partir de la observación de los precios de bienes heterogéneos sus características asociadas.

Como uno de los *papers* más citados del *Journal of Political Economy* (Greenstone, 2017), se obtiene una aproximación al comportamiento de la oferta y demanda (equilibro) a partir del proceso generador de datos entre los bienes y sus características. Para el mercado inmobiliario -en el que los comportamientos de los agentes, oferta y demanda, y equilibrios no se pueden explicar en su totalidad con la teoría microeconómica convencional-, esta aproximación ha sido de singular importancia, más aún en el contexto de la ciudad de Bogotá.

Así por ejemplo, Perdomo (2011) utiliza los precios hedónicos cómo método econométrico para estimar el precio de las viviendas y posteriormente el de la cercanía a Transmileno; Carriazo et al. (2013) identifican cómo de no incluir la calidad del aire en las estimación a través de precio hedónicos conlleva un sesgo en los precios de las viviendas en Bogotá. Lozano & Anselin (2012) utilizan variables espaciales en los submercados de la viviendas para estimar los precios en Bogotá, identificando una gran capacidad predictiva, pero afectada (sobre estimación) por los estratos socioeconómicos. También se encuentran Medina et al. (2007) quienes demuestran como los subsidios aplicados a las tarifas de los servicios públicos se transfieren al precio de las viviendas.

Esta breve exploración bibliográfica refleja los usos y potencialidades de la contribución de Rosen (1974) al mercado inmobiliario con énfasis en la ciudad de Bogotá. Claro está, como la capital de Colombia, Bogotá es una ciudad heterogénea con dinámicas poblacionales, sociales y económicas distribuidas a por todo el territorio urbano, las cuales pueden afectar los precios del mercado inmobiliario. Más aun, en una localidad como Chapinero, la cual es una de las áreas centrales de la ciudad; con una gran catnidad de población flotante; donde se cuenta con vías que se conectan con el sistema de movilidad de la ciudad; se tienen usos del suelo asociados a actividades comerciales y residenciales; y se tienen parques, canales y corredores ecológicos (Decreto 468, 2006).

Por estas características, tal vez Chapinero sea la localidad con la mayor complejidad de formas de habitar el territorio de la ciudad del país, lo cual la hace un espacio geográfico donde convergen múltiples factores que afectan los precios de las viviendas. Por tal razón, construir uno o varios modelos predictivos de los precios es desafiante y una necesidad, ante una potencial inversión que compre la mayor cantidad de predios, gastando lo menos posible.

Ahora bien, para construir el o los modelos se utilizó una muestra de los anuncios publicados en la página web Properati. Este sitio web de origen argentino y con presencia en varios países de Latinoamérica, le permite a propietarios publicar un bien inmueble con la mayor cantidad de información (precio, características y amenidades, ubicación y descripción), y a personas interesadas en adquirir o arrendar un bien contactarse con el comprador. Es decir, se utilizó una página web que ofrece servicios de intermediación, pero que permite caracterizar la oferta de bienes en Bogotá.

De igual manera, se utilizó la información de Open Street Map (transporte público, comercio, equipamentos colectivos, y vías y trasnporte público), la cual permitió asociar información geográfica relevante a la muestra de Properati. Así como se utilizó también la información de la página web Infraestructura de Datos Espaciales de Bogotá (Ideca), un portal de libre acceso y consulta de la Unidad Administrativa Especial de Catastro Distrital (UAECD), de donde se utilizaron las bases de datos (capas) de las localidades de Bogotá

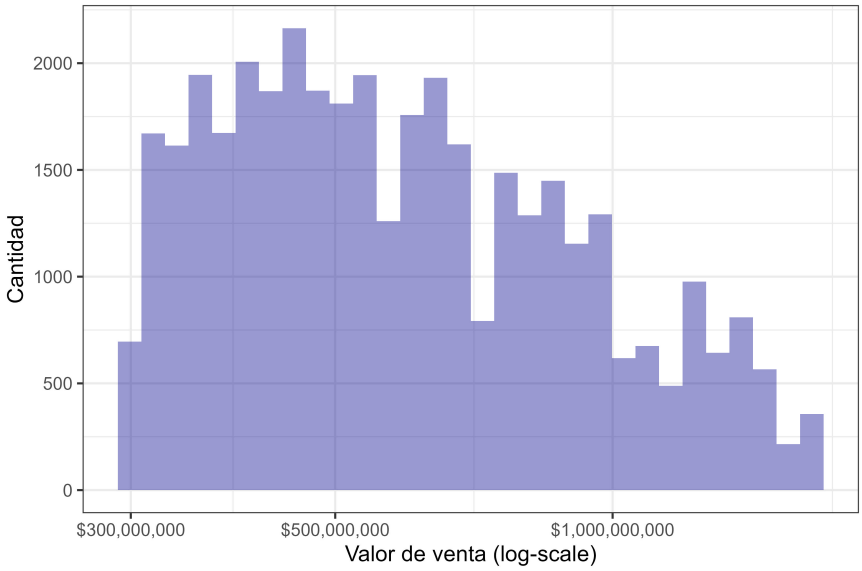
Con estas fuentes de información se construyeron diferentes modelos para predecir el precio de las viviendas en Chapinero. Estos modelos aplican los conceptos abordados en la clase *Big Data y Machine Learning para Economía Aplicada*; más concretamente, se utilizaron técnicas *Ramdon forest*, regresión lineal, *Elastic Net*, *Boosting* y *Decission tree*, utilizándose el error absoluto medio (MAE por sus siglas en inglés) como criterio de selección del modelo fuera de muestra. En cuyo caso, dados los resultados obtenidos en *kaggle* se optó por el modelo *Random Forest*, como se explicará más adelante.

En las siguientes secciones, se presentan las bases de datos utilizadas y las trasnformaciones, presentando a su vez estadísticas descriptivas que facilitan su comprensión; la descripción del modelo utilizado; los resultados obtenidos luego de su implementación; y unas conclusiones de todo el ejercicio.

## 2 Datos y Métodos

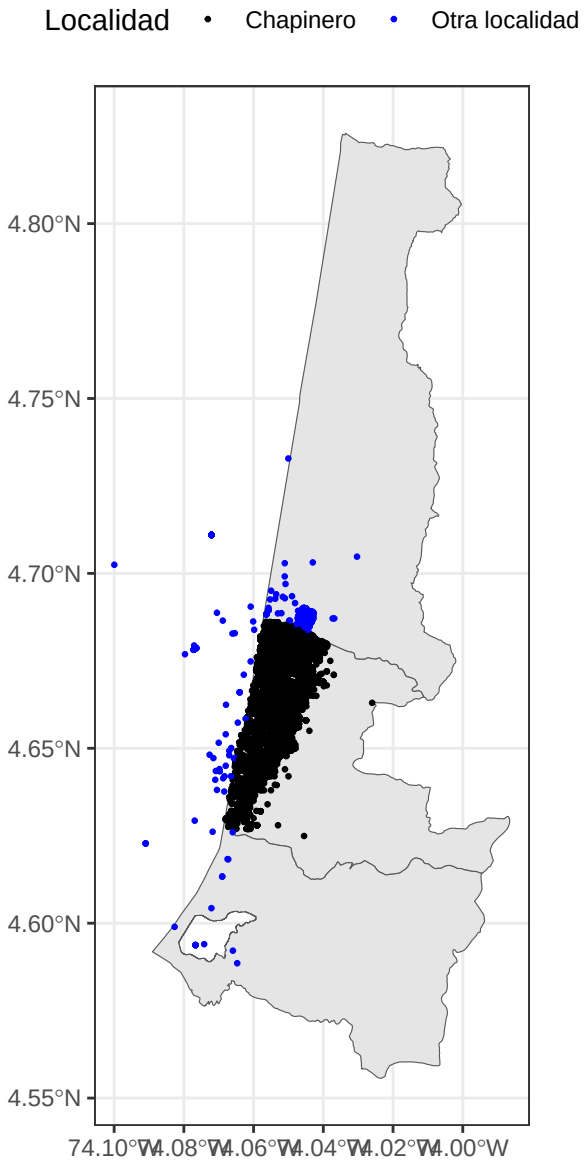
Los datos principales consistieron en los archivos descargados de la página de datos de la Competencia en Kaggle. Esta base de datos contiene 48.930 observaciones, distribuidas en 10.286 observaciones para prueba (21%) y 38.644 para entrenamiento (79%). Cada observación representa un inmueble para venta publicado en Properati entre 2019 y 2021 en la ciudad de Bogotá. Dentro de las características de cada inmueble se tienen las áreas (construida y cubierta), el número de habitaciones y baños, el tipo de propiedad (casa o apartamento), y el título y descripciónn con el que se hizo la publicación en la página web y localización (latitud y longitud). Dada la estructura de la base de datos es muy probable que se haya construido a través de un ejercicio de *Web Scrapping* y luego se haya dispuesto para su descarga y uso en el Problem Set 3.

Figura 1: Distribución del precio de las viviendas (log)




A partir de las descripciones, se utilizó el análisis de texto para constrastar el tipo de inmueble publicado con el que se encuentra en la descripción, extraer el número de pisos de las casas, e identificar el piso en el que se encuentran los apartamentos. Como se tienen todos los inmuebles para Bogotá, en función de su localización (latitud y longitud), se extrajeron aquellas casas y apartamentos dentro de la localidad de Chapinero. Este ejercio se realizó mediante la extracción los de puntos (capa construida con la latitud y longitud) dentro del polígono de Cahpinero (capa del Ideca). Al cruzar la capa de los polígonos las localidades con la tabla de *test* observamos que 593 propiedades se encuentran fuera de Chapinero (Figura 2), que es la localidad en la cuál se predecirán los precios de las viviendas.

Figura 2: Ubicación datos de validación




Utilizando el mismo método de análisis espacial, se utilizaron las capas de Open Street Maps para medir las distancias del inmueble a gimnasios, estaciones de Transmilenio, bares, supermercados, colegios, hospitales, principales avenidas, centros comerciales y universidades.Las principales estasdísticas descriptivas según los tipos de bienes inmuebles se ven a continuación:

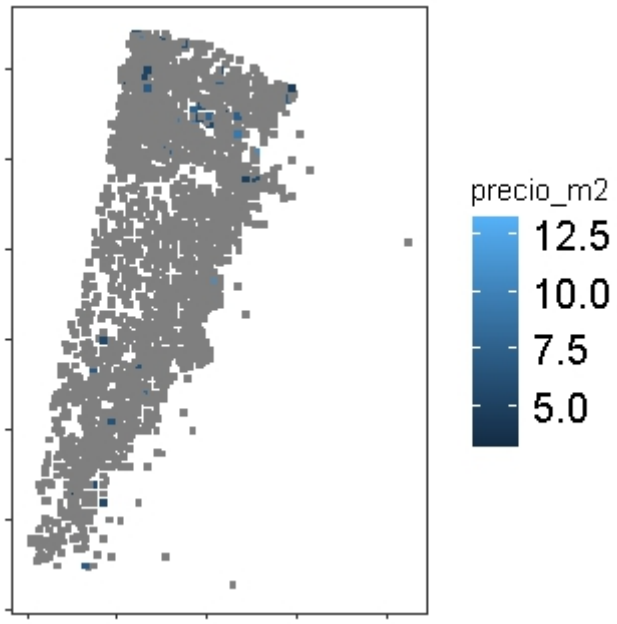
Para la localidad de Chapinero, luego del procedimiento anterior, se obtuvo una base de datos con 10.000 observaciones. En total, se obtuvieron 9.702 (97%) apartmentos y 298 (3%) casas en venta. En promedio, los inmuebles fueron publicados a un precio de 847,1 millones, tienen 137,4$mt^2$ de área construida, 128,3$mt^2$ de área habitables, 2,4 habitaciones, 2,6 baños y 2,3 dormitorios. Las amenidades más cercanas son los parques a una distancia promedio de 159,9 metros, centros comerciales a de 246,9 metros y las avenidas a 230,2 metros. Las más lejanas son los bares a 1,1 km, seguido por los hospitales (830 m) y las estaciones de transmileno (801 m).

| Variable | Promedio | Desv. est. | Mín | Máx. |
|---|---|---|---|---|
| Precio (millones) | 847.2 | 375.8 | 300.0 | 1650.0 |
| Área cubierta (km2) | 125.9 | 57.7 | 31.0 | 505.0 |
| Área total (km2) | 185.2 | 2447.5 | 15.0 | 108800.0 |
| Habitaciones | 2.3 | 0.9 | 1.0 | 11.0 |
| Baños | 2.7 | 1.0 | 1.0 | 7.0 |
| Dromitorios | 2.4 | 1.0 | 0.0 | 11.0 |
| Número de pisos | 1.0 | 0.1 | 1.0 | 6.0 |
| Piso | 2.1 | 3.0 | 1.0 | 60.0 |
| Dist. parque (km) | 159.9 | 95.5 | 4.7 | 1948.5 |
| Dist. Transmilenio (km) | 801.1 | 476.3 | 0.3 | 2968.8 |
| Dist. bar (km) | 1149.1 | 470.1 | 1.5 | 2297.6 |
| Dist. colegio (km) | 481.3 | 300.7 | 0.0 | 2271.6 |
| Dist. avenida (km) | 230.2 | 228.4 | 0.0 | 2074.1 |
| Dist. universidad (km) | 725.4 | 411.7 | 0.0 | 2768.8 |
| Dist. gimnasio (km) | 730.1 | 402.4 | 10.2 | 2685.6 |
| Dist. CAI(km) | 560.3 | 312.6 | 3.8 | 1680.2 |
| Dist. supermercados (km) | 531.5 | 289.6 | 0.0 | 2142.9 |
| Dist. hospitales (km) | 830.2 | 415.6 | 0.0 | 3408.1 |
| Dist. centro comercial (km) | 247.0 | 188.6 | 1.1 | 2428.6 |

Cuadro 1: Estadísticas descriptivas

Sobre este último hallazgo, esta oferta de bienes inmuebles puede estar caracterizada por casas o apartamentos más alejadas de las zonas con mayor flujo de población flotante. Al ser una localidad en la que en su centralidad alberga la mayor oferta de universidades de Bogotá, así como donde se desarrollan actividades comerciales, los inmuebles destinados a usos residenciales que se ofrecen en el mercado se encuentran más alejados de esta centralidad. Claro está, se tiene una heterogénea composición, pues como se observa en la tabla anterior aproximadante las desviaciones estándar de las distancias son la mitad de los valores promedio, y se obtienen distancia en un rango de entre 0 m 3 km.

Figura 3: Precio en M2 en Chapinero (precio por millones)




# 3 Modelo y resultados

En esta sección se exponen los modelos que se utilizaron para predecir los precios de la vivienda en la localidad de Chapinero, las variables utilizadas en los modelos y la selección de hiperparámetros. En particular, se utilizaron los modelos de bosques, bosques aleatorios, regresión lineal, elasticnet y boosting para realizar la predicción. Finalmente, se presenta una comparación del rendimiento de estos modelos. El mejor resultado se obtuvo con el modelo de Random Forest utilizando como métrica el Mean Absolute Error metric (MAE).

## 3.1 Modelo y descripción de las variables

Para todos los modelos utilizados para la predicción, se utilizó el siguiente modelo:

$$Precio = f(\text{características propias}, \text{características del vecindario}, \text{variables temporales})$$

Siguiendo a Mendieta López & Perdomo (2007), se utilizaron tres categorías de variables que capturan atributos que explican el precio de la vivienda. El primero, corresponde a variables que capturan características convencionales de las viviendas como el número de habitaciones, el número de baños, la superficie total y el tipo de propiedad. Los valores faltantes de estas variables se imputaron como la mediana de la muestra con el fin de realizar la predicción. El segundo grupo está conformado por características del vecindario que corresponde a criterios de seguridad, por ejemplo, la distancia de la vivienda al Comando de Atención Inmediata(CAI), a criterios de servicios, como la distancia a parques, centros comeciales, instituciones educativas, y de accesibilidad de transporte, por ejemplo, distancia a la estación del transmilenio. Las variables de distancia fueron calculadas utilizando la paqueteria Osmadata del programa R. El tercero, corresponde a factores

de geolocalización,como la longitud y la latitud) y temporales, como el mes y el año, debido a que los precios pueden variar por el ciclo económico. En el anexo A, se puede encontrar la tabla con la descripción de las variables utilizadas.

### 3.2 Descripción del modelo

Utilizando el modelo y las variables mencionadas en el Anexo A, se dividió la base de datos en dos conjuntos(entrenamiento y validación). Los modelos se entrenaron con el conjunto de entranamiento para, posteriormente, realizar la predicción en el conjunto de validación. El criterio para escoger al mejor modelo fue el Error absoluto medio(MAE) que mide la magnitud promedio de los errores de predicción del modelo:

$$MAE = \frac{1}{N} \sum |precio - \text{precio predicho}|$$

Por lo tanto, entre menor sea el MAE, mejor desempeño tiene el modelo. Debido a que no se contaba en el conjunto de validación con la información del precio observado, se escogió como mejor modelo aquel con menor MAE reportado en la competencia Kaggle.

El mejor modelo utilizando esta métrica fue el Random Forest. Este algorito de aprendizaje funciona creando "selvas" de árboles de decisión y promediando las predicciones de los árboles individuales para obtener una predicción final. Este enfoque permite mejorar el desempeño de las predicciones debido a que cada árbol se contruye utilizando un conjunto de datos de entrenamiento y un conjunto aleatorio de variables predictoras, lo que disminuye el sobreajuste y mejora la generalización del modelo.

Para este informe, se implementó el modelo Random Forest utilizando el paquete "caret" del softaware R. Se incluyeron como variables predictoras las del Anexo A. La estrategia usada para seleccionar los hiperparámetros se basó en la definición de una grilla con diferentes combinaciones posibles, en la que probara de forma aleatoria la cantidad de variables en cada división del árbol(2,3,4,5,8) y el criterio de "variance" para evaluar la reducción de la varianza de los nodos resultantes de cada árbol. Además, se estableció un tamaño mínimo de los nodos terminales de cada árbol con el fin de prevenir el sobreajuste(1,2,3,6). Finalmente, los parámetros de control para el proceso de entrenamiento fue una configuración de validación cruzada con 10 particicones para evaluar el rendimiento del modelo.

### 3.3 Comparativa con otros modelos

Además del modelo finalmente escogido, se ejecutaron otras técnicas para realizar la predicción de los precios de la vivienda en Chapinero. Se utilizaron para todos los otros modelos las mismas variables explicativas descritas en la tabla Anexo. Se estimaron las siguientes técnicas adicionales al Random Forest: regresión lineal, Elastic net, Boosting y árboles de decisión. Para todos estos modelos, se utiliza un cross-validation con 10 particiones con el fin de mejorar la predicción y de evitar el sobreajuste.

En la tabla anterior se puede observar el Error Cuadrático Medio (RMSE) para la muestra de entrenamiento, así como el Error absoluto medio (MAE) de la muestra de validación que registró los envíos a la competencia kaggle(debido a que la muestra no tenía la información del precio para estimarla directamente) y de los datos de entrenamiento. Se aprecia que el Random Forest tiene

| Estadística | Random Forest | Linear Regression | Elastic Net | Boosting | Decision tree |
|---|---|---|---|---|---|
| MAE (Kaggle) | 223834975 | 308749889 | 260870545 | 289260591 | 304528484 |
| RMSE | 185392301 | 258160495 | 210281690 | 266505969 | 282794569 |
| MAE(train) | 125757655 | 191380688 | 143548002 | 200055282 | 216060881 |

Cuadro 2: Comparativa con otros modelos

un desempeño sistemáticamente superior en todas las medidas con respecto a los otros modelos. Los siguientes mejores modelos son el Elastic Net, Boosting, regresión lineal y árboles de decisión respectivamente. Estos resultados son esperados en la medida en que el Random Forest permite capturar relaciones no lineales(a diferencia de la regresión lineal y Elastic Net), además de que es más robusto a sobreajuste con respecto a modelos más complejos como el Boosting.

# 4 Conclusiones

El presente estudio realizó un ejercicio de predicción del precio de la vivienda para la localidad de Chapinero en Bogotá Utilizando datos de https://www.properati.com.co. Para construir el modelo, se empleaó información de esta base de datos e información de distancia calculadas a partir de Open Source Maps (OSM). Se ejecutaron cinco algoritmos de machine learning (Random Forest, Árboles de decisión, regresión lineal, boosting y Elastict Net) y se seleccionó el mejor utilizando como métrica el Error Absoluto Medio (MAE).

Se encontró que la oferta de bienes inmubles de Chapinero alberga la mayor oferta de universidades en Bogotá, así como dónde se desarrollan mayor actividad comercial. Asimismo, se encontró que el mejor modelo que predice los precios es el Random Forest debido a que permite capturar relaciones no lineales, además de que es más robusto a sobreajuste por la aleatorización de las variables y del conjunto de datos que utiliza para entrenar en la construcción de los árboles.

Sin embargo, se recomienda seguir estimando modelos que permitan estimar los precios de las viviendas debido a la naturaleza cambiante de este sector. Es importante la predicción de precios debido a que puede servir a las inmobiliarias con el fin de evitar sobreestimación de precios que puedan causar pérdidas económicas.

# 5 Referencias bibliográficas


Alcaldía Mayor de Bogotá. (20 de noviembre de 2006). *Por el cual se reglamenta la UPZ 99, Chapinero, ubicada en la loclaidad Chapinero* (Decreto 468 de 2006, Ed.).

Carriazo, F., Ready, R., & Shortle, J. (2013). Using stochastic frontier models to mitigate omitted variable bias in hedonic pricing models: A case study for air quality in Bogotá, Colombia. *Ecological Economics*, *91*, 80-88.

Greenstone, M. (2017). The Continuing Impact of Sherwin Rosen's «Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition». *Journal of Political Economy*, *125*(6), pp. 1891-1902.

Lozano, n., & Anselin, L. (2012). Is the price right?: Assessing estimates of cadastral values for Bogotá, Colombia. *Regional Science. Policy & Practice*, *4*(4), 495-508.

Medina, C., Morales, L., Bernal, r., & Torero, M. (2007). Stratification and Public Utility Services in Colombia: Subsidies to Households or Distortion of Housing Prices? [with Comments].

*Economía*, *7*(2), 41-99.
Mendieta López, J. C., & Perdomo, J. A. (2007). *Especificación y estimación de un modelo de precios hedónico espacial para evaluar el impacto de Transmilenio sobre el valor de la propiedad en Bogotá*.
Perdomo, J. (2011). A methodological proposal to estimate changes of residential property value: case study developed in Bogotá. *Applied Economics Letters*, *18*(16), 1577-1581.
Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34-55.

# 6 Repositorio

El respositorio se puede consultar en el siguiente enlace:

- Problem Set 3: Making Money with ML?

# 7 Anexo A: Variables utilizadas en el modelo.

| Nombre de la variable | Descripción |
|---|---|
| rooms_median | Número de baños de la vivienda |
| bathrooms_median | Número de baños de la vivienda |
| surface_total_median | Área total de la vivienda |
| property_type | =1 si la vivienda corresponde a una casa |
| distancia_parque | Distancia de la vivienda al parque más cercano |
| distancia_comercial | Distancia de la vivienda al cento comercial más cercano |
| distancia_avenida_principal | Distancia de la vivienda a la avenida principal |
| distancia_universidad | Distancia de la vivienda a la universidad más cercana |
| distancia_cai | Distancia de la vivienda al CAI de policía más cercano |
| distancia_bar | Distancia de la vivienda al bar más cercano |
| distancia_gimnasio | Distancia de la vivienda al gimnasio más cercano |
| distancia_transmi | Distancia de la vivienda a la estación de transmilenio más cercana |
| distancia_SM | Distancia de la vivienda al Supermercado más cercano |
| distancia_colegio | Distancia de la vivienda al colegio más cercano |
| distancia_hospitales | Distancia de la vivienda al hospital más cercano |
| month,year | Año-mes |
| lat,lon | Coordenadas de geolocalización de la casa |

Cuadro 3: Nombres de las variables y su descripción