Problem Set 3: Making money with ML?

Sofia Charry Tobar
s.charry@uniandes.edu.co

Laura Manuela Rodriguez Morales
lm.rodriguezm@uniandes.edu.co

Nicol Valeria Rodríguez Rodríguez
nv.rodriguezr1@uniandes.edu.co

Brahyan Alexander Vargas Rojas
ba.vargas@uniandes.edu.co

El repositorio del ejercico es: https://github.com/nvrr2028/Taller-3-BDML.git

1. Introducción

El objetivo de este ejercicio es crear un modelo acertado y confiable para predecir los precios de venta de viviendas en Chapinero, con el fin de apoyar a la compañía en la toma de decisiones de compra y rentabilidad. Se seguirá la metodología de métodos hedónicos que utiliza las características de las propiedades como predictores del precio de venta, ya que describen un bien diferenciado (Rosen, 1974). En la literatura, se ha utilizado consistentemente la metodología de métodos hedónicos para predecir precios de vivienda (Rosen, 1974; Harrison y Rubinfeld, 1978; Bourassa et al., 2010), y se ha destacado la utilidad de los algoritmos de *Machine Learning* en la predicción de precios de vivienda en la literatura reciente (Park y Bae, 2015; Varma et al., 2018; Yu et al., 2018; Truong et al., 2020; Mora-Garcia et al., 2022).

Siguiendo la literatura mencionada, se utilizan predictores de las características físicas del inmueble, provenientes de anuncios de venta del inmueble, y su distancia a servicios, provenientes de datos geográficos. Dado que los vendedores deciden el precio en función de comodidades valiosas del inmueble, como número de habitaciones, y la cercanía a servicios importantes, como a un hospital, los datos son útiles para predecir los precios de venta. Sin embargo, los datos son de 2019 a 2021, por lo que pueden no reflejar las condiciones actuales del mercado inmobiliario. Entonces se recomienda actualizar los datos a medida en

que estén disponibles para revisar caídas en la precisión del modelo escogido independientes de los predictores, como por ejemplo un posible enfriamiento del mercado inmobiliario.

Después de comparar varios modelos, el modelo con el mayor poder de predicción es *Gradient Boosting (GBM)*. por lo que la estrategia de aprendizaje lenta de errores es especialmente útil para la prediccion de precios. Además, se concluye que las variables asociadas a las características físicas de una vivienda cuentan con un elevado poder predictivo sobre los precios, frente a las variables de distancia.

2. Data

2.1. Base de datos

La base de datos combina información de dos fuentes distintas. La primera recopila precios de inmuebles en Bogotá y sus características entre enero de 2019 y diciembre de 2021 provenientes de publicaciones de vendedores en `https://www.properati.com.co`. La segunda fuente es Open Source Maps (OSM), que proporcionó datos geográficos sobre la ubicación de las facilidades relevantes en Chapinero, Bogotá. A partir de estos datos, se crearon variables para medir la distancia entre el inmueble y las facilidades correspondientes. La base se dividió en un 70 % de entrenamiento (38.644 observaciones) y un 30 % de prueba (10.286 observaciones).

La base de datos es útil para predecir el precio de venta de las propiedades ya que los vendedores establecen el precio en función de las características del inmueble y los servicios cercanos, según Rosen (1974) y Harrison y Rubinfeld (1978). Sin embargo, al ser los datos de 2019 a 2021, pueden no reflejar las condiciones actuales del mercado inmobiliario, lo que puede afectar la precisión del modelo de predicción. Por lo tanto, es importante actualizar los datos regularmente para detectar cambios en el mercado y mejorar la capacidad predictiva del modelo de ML, según Williams (2021).

Las variables predictoras fueron seleccionadas en función de la unicidad de las características y de si la característica o servicio responde a necesidades importantes como salud, seguridad, educación, abastecimiento, transporte y recreación, con base en trabajos previos (Rosen, 1974; Park y Bae, 2015; Varma et al., 2018; Yu et al., 2018; Truong et al., 2020; Mora-Garcia et al., 2022). Por ejemplo, más adelante se destaca la unicidad de los inmuebles con terraza en la muestra. También, se escogieron la distancia a un hospital, CAI, parque o colegio porque responden a necesidades importantes. Específicamente, se incluyeron 10 predictores de distancia usando las ubicaciones de las siguientes facilidades disponibles en OSM: parque, gimnasio, Transmilenio, CAI, centro comercial, bar, colegios, supermercados (SM), universidades y hospitales. También, se incluyeron 4 variables del título o descripción de las propiedades usando extracción de texto: si el inmueble pertenece a Chapinero, super-

ficie cubierta, si tiene terraza, parqueadero y acceso a espacios sociales como salón comunal o piscina. Favor ver Anexo 1 para mayor detalle sobre las variables escogidas.

Se recuperaron muchas observaciones faltantes en el conjunto de datos. La variable de área cubierta del inmueble no tenía el 44 % de las observaciones y la variable cantidad de baños no tenía el 25,5 % de las observaciones. Para estimar las observaciones faltantes, se extrajo el valor numérico de la descripción de los inmuebles, de acuerdo con múltiples formas de representar "metros" de los inmuebles y se adicionaron en los valores faltantes. Para las demás observaciones sin información, se realizaron cuatro regresiones. Cada regresión estimó las observaciones faltantes de *surface_covered* o *bathrooms* en función de la cantidad de habitaciones y el tipo de propiedad para las bases de entrenamiento y prueba. Las nuevas variables son *surface_covered2* y *bathrooms2* para las bases de entrenamiento y prueba.

2.2. Análisis descriptivo

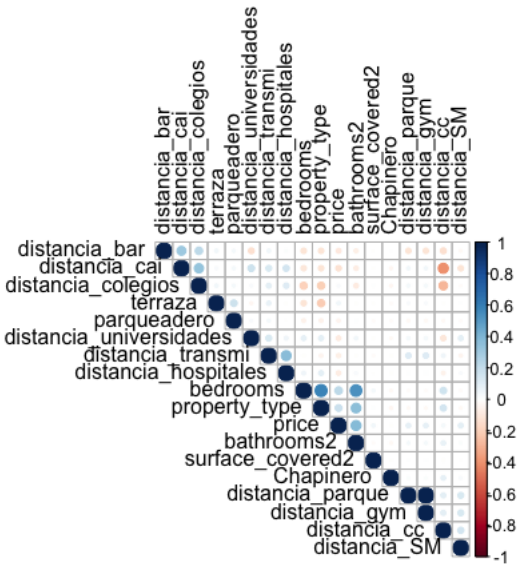
La Tabla 1 presenta las estadísticas descriptivas de las variables que se consideraron relevantes para predecir el precio de venta de las propiedades en la localidad de Chapinero. La base de entrenamiento contiene 38.644 observaciones. La variable de interés *price* tiene un valor promedio $654'534.675 COP, con un máximo de $1'650.000 COP y un mínimo de $300.000 COP. Adicionalmente, el promedio de habitaciones por inmueble es de aproximadamente 3,14 habitaciones, con una desviación estándar de 1,53 habitaciones. Además, hay 2,89 baños en promedio por inmueble. Por otra parte, se consideran 4 variables dummy para este estudio. Estas variables son *property_type*, *terraza*, *parqueadero* y *Chapinero* y toman el valor de 1 si el hogar cuenta con la característica descrita o 0 en caso contrario. Finalmente, se presentan las estadísticas descriptivas de la distancia medida en metros a varios servicios e instituciones de la ciudad. Se realizará un análisis más minucioso en las próximas estadísticas descriptivas.

Tabla 1: Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
price	38,644	654,534.675	311,417.887	300,000	1,650,000
month	38,644	5.665	3.289	1	12
year	38,644	2,020.294	0.760	2,019	2,021
surface_total	7,854	153.950	274.370	16	17,137
surface_covered	21,009	1,075.533	3,744.153	1	79,692
rooms	20,384	3.008	1.372	1	11
bedrooms	38,644	3.145	1.535	0	11
bathrooms	28,573	2.884	1.093	1	13
property_type	38,644	0.245	0.430	0	1
lat	38,644	4.691	0.037	4.577	4.765
lon	38,644	-74.063	0.032	-74.170	-74.026
Chapinero	38,644	0.025	0.157	0	1
terraza	38,644	0.458	0.498	0	1
social	38,644	0.381	0.486	0	1
parqueadero	38,644	0.722	0.448	0	1
distancia_parque	38,644	160.676	100.981	0.991	3,344.619
distancia_gym	38,644	160.676	100.981	0.991	3,344.619
distancia_transmi	38,644	949.702	684.761	3.559	6,296.792
distancia_cai	38,644	1,021.334	501.982	2.416	2,957.094
distancia_cc	38,644	659.464	383.784	0.555	4,706.625
distancia_bar	38,644	1,384.312	697.225	4.664	3,968.576
distancia_SM	38,644	414.853	257.882	2.052	3,647.073
distancia_colegios	38,644	541.660	302.754	4.862	1,725.294
distancia_universidades	38,644	1,061.549	568.663	3.195	4,338.044
distancia_hospitales	38,644	922.406	535.573	9.743	3,565.820
surface_covered2	38,644	708.619	2,790.971	-65.024	79,692.000
bathrooms2	38,644	2.442	1.235	0.000	13.000

El Gráfico 1 muestra la correlación entre las variables seleccionadas. El precio del inmueble *price* tiene una correlación negativa con la distancia a algunas instituciones de seguridad y colegios, así como con la presencia de terraza o parqueadero en el inmueble. Por otro lado, *price* presenta una correlación positiva con el tipo de propiedad, cuartos y baños y con la distancia del inmueble a centros comerciales, gimnasios, universidades, hospitales, parques y estaciones de Transmilenio. No se encontró una correlación significativa al 5 % entre el precio y la ubicación en Chapinero, ni con la superficie de la propiedad.

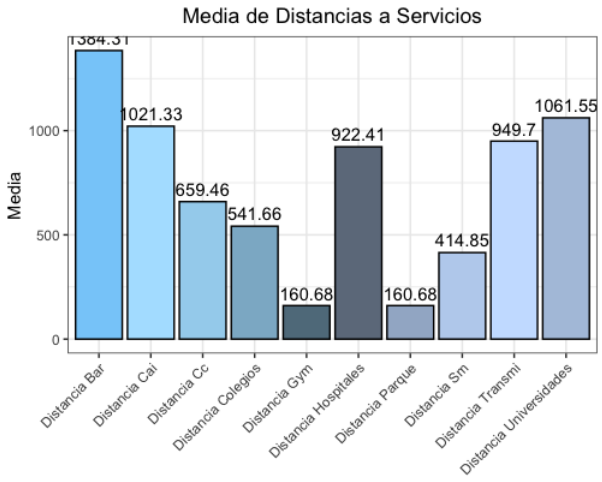
Gráfico 1: Mapa de correlaciones del conjunto de variables seleccionadas



Nota: El tamaño de los círculos indica la magnitud de la correlación entre las variables. Además, se excluyen los coeficientes de correlación que no sean signficativos al 5 %.

El Gráfico 2 muestra la media de la distancia a los servicios e instituciones analizados en el estudio. Se observa que la distancia promedio a los bares es la mayor en comparación con otros servicios, lo que sugiere que, en promedio, los inmuebles están más alejados de los bares. En contraste, la distancia promedio a los parques y gimnasios es similar, indicando que todos los hogares en Bogotá tienen acceso a estos servicios a una distancia promedio de 160,68 metros. Las instituciones educativas de educación superior se encuentran relativamente lejos de las propiedades, con una media de 1.061,55 metros. Las estaciones de Transmilenio tienen una distancia promedio de 949,7 metros de las propiedades. Por último, los comandos de atención inmediata (CAI), que pueden ser un indicador de la seguridad de la propiedad, se encuentran alejados de los inmuebles en promedio.

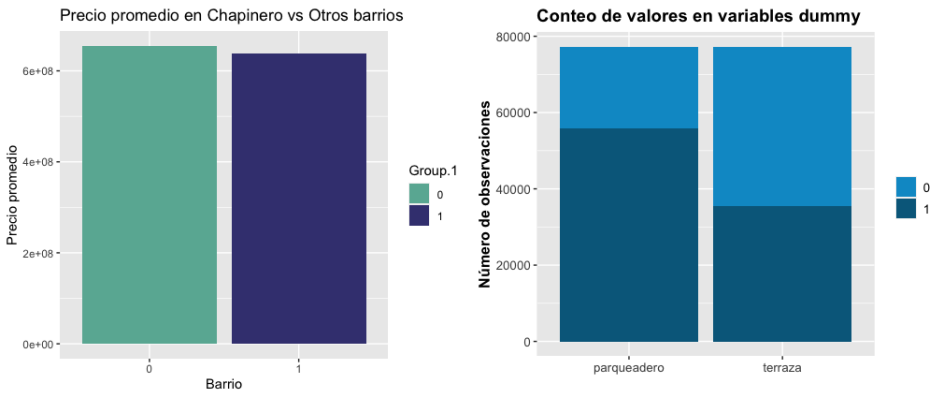
Gráfico 2



Fuente: cálculos de los autores.

Se utilizan variables creadas a partir de las descripciones de las propiedades para un análisis más detallado. Se ha observado que la mayoría de las propiedades tienen estacionamiento, pero la mayoría no tiene terraza. Estas variables son importantes para predecir el valor de una propiedad, ya que son características únicas en algunos bienes. Además, se promediaron los precios entre Chapinero y otras localidades de Bogotá para realizar un análisis más preciso de la localidad de Chapinero. En promedio, se encontró que las propiedades en otras localidades son ligeramente más caras que en Chapinero. Sin embargo, este promedio podría estar subestimado porque Chapinero contiene a Chicó, uno de los barrios más prestigiosos y costosos de Bogotá.

Gráfico 3



(a) (b)

3. Modelo y resultados

El modelo con el mejor desempeño para predecir el precio de las viviendas en Chapinero es *Gradient Boosting (GBM)*. Por lo tanto, esta estrategia de aprendizaje lento de los errores resulta muy utilidad para la compañía. A continuación, se detallará el proceso de estimación y se compara frente a otras técnicas.

3.1. Variables utilizadas

La variable a predecir *Price* se explica a partir de *surface_covered2*, *bedrooms*, *bathrooms2*, *Chapinero*,*property_type*,*terraza*,*social*,*parqueadero*, *distancia_parque*, *distancia_gym*, *distancia_transmi*, *distancia_cai*, *distancia_cc*, *distancia_bar*, *distancia_SM*, *distancia_colegios*, *distancia_universidades*, *distancia_hospitales*. Remitirse al Anexo 1 para los detalles de las variables.

3.2. Entrenamiento del modelo

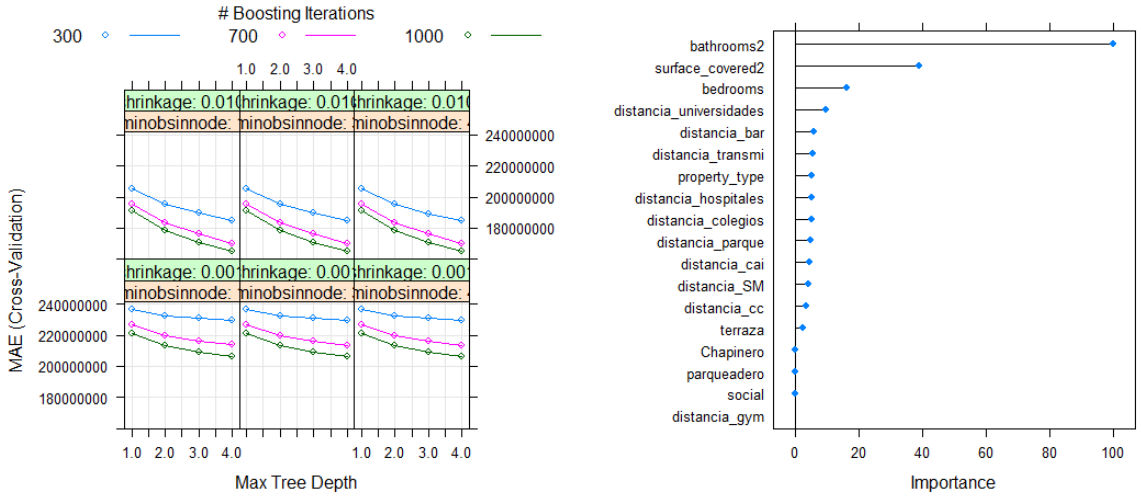
La evaluación del modelo se realizo de la siguiente manera:

1. La muestra de entrenamiento **train** se dividió en un dos conjuntos adicionales, de tal manera que el 70 % se utilizara en el entrenamiento del modelo, y el 30 % restante en su evaluación. Este procedimiento se debe a que la base de **test** no contiene la variable de *Price*.

2. Una vez se hayan entrenado los modelos y verificado que no hay problemas de *overfitting*, se predicen los precios de las viviendas en Chapinero a partir de la base de **test**, para luego introducir los resultados en Kaggle y obtener la medida de evaluación *MAE*.

Teniendo en cuenta lo anterior, el modelo GBM pretende estimar el precio a partir de las variables especificadas en la sección 3.2. Con respecto a la selección de los hiperparámetros, el modelo se entrenó de acuerdo con: (i) ***n.trees***: 300, 700, 1000 (ii) ***interaction.depth***: 1, 2, 3, 4, (iii) ***shrinkage***: 0.01, 0.001 y (iv) ***n.minobsinnode***: 10, 30, 40.

Gráfico 4: Resultados del módelo GBM



(a) (b)

Como se observa en el Panel (a) del Gráfico 4, el modelo evidencia una reducción significa de la medida de desempeño, MAE, conforme incrementa el número de árboles, la profundidad de los mismos y para niveles de regularización cercanos a 0.01. De esta manera, el mejor modelo GBM seleccionado de acuerdo con los hiperparámetros es n.trees = 1000, interaction.depth = 4, shrinkage = 0.01 y n.minobsinnode = 10. Adicional a lo anterior, se resalta que las variables de mayor importancia o utilidad en el trabajo de predicción son, en orden: (i) el número de baños en la residencia (*bathrooms2*), (ii) el área de la vivienda (*surface_covered2*), (iii) el número de cuartos (*rooms*) y (iv) la distancia con universidades (*distancia_universidades*).

3.3. Comparación

Antes de hacer la elección del modelo final también se corrieron otras técnicas para realizar la predicción. El objetivo se mantiene en predecir el *Precio* de la vivienda a partir del conjunto de variables explicativas descritas en la Sección 2.2. Teniendo en cuenta lo anterior se estimaron las siguientes técnicas: regresión lineal, Lasso, Ridge, Elastic net, Gradient Boosting y Superlearnear. En particular, la regresión utilizada para todos los modelos es como se muestra en el Anexo 2. Se utiliza la técnica de remuestreo *cross-validation* con k=10, con el objetivo de tener una mejor medida del desempeño de los modelos. Junto a lo anterior, los hiperparámetros de los mejores modelos y sus resultados son:

```
# Regresión Lineal:
Precio = lm(x)
```

```
X vector de características. Fórmula Anexo 2
# Lasso:
alpha = 1, lambda = 0.01
# Ridge:
alpha = 0 and lambda = 11403996.
# Elastic Net:
alpha = 1   , lambda = 0.01
# Superlearner
# GBM
n.trees= 1000,interaction.depth= 4,
shrinkage= 0.01, n.minobsinnode= 10
```

Tabla 2: Comparación del modelo principal con otras técnicas

Estadística	GBM	RL	Lasso	Ridge	Elastic Net	Superlearner
MAE (Kaggle)	263245024	327258864	327259725	299411431	298404785	279848874
RMSE	22480235	276669470	276670870	276656809	276670870	240176517
RSQ	0.485	0.208	0.208	0.208	0.208	0.416
MAE	165167704	209889663	209937182	210206122	209937182	178784207

Por lo tanto, de acuerdo con la Tabla 2, se observa que el modelo GBM evidencia mejores resultados en todas las medidas de evaluación de forma consistente. Por otro lado, los 3 siguientes mejores modelos son, en orden, Superlearner, Ridge y Elastic Net. Además, las técnicas de regresión lineal y Lasso evidencian el menor desempeño en el ejercicio de pronóstico de precios de la vivienda.

4. Conclusiones y recomendaciones

El modelo de aprendizaje automático de mayor utilidad para la compañía seria *Gradient Boosting (GBM)*. Según nuestro modelo para la predicción de precios en propiedades de Bogotá podemos darnos cuenta que: Los bogotanos valoramos en gran medida la presencia de un número de baños acorde a las personas que hay en el hogar, lo cuál debe ser muy diferente a si estuvieramos hablando de países europeos o quizá estados unidos. Valoramos que las viviendas sean amplias, en términos de superficie cubierta y el número de habitaciones. Estos resultados no nos sorprenden pues a mayor área, habitaciones y baños, es normal que el precio aumente.

Nos damos cuenta que la mayoría de atributos de cercanía, sea a colegios, supermercado, transporte público, etc, son factores que pueden usarse para variar el precio pero no van a ser sus principales determinantes. Algo que para nosotros resultó sorpresivo fue la poca importancia de la variable de la presencia u ausencia de un parqueadero.

Dado que en nuestra muestra de datos para el entrenamiento del modelo no teníamos muchas observaciones que correspondieran a *Chapinero*, nuestra localidad de predicción de interés, fue importante aplicar métodos de Spatial Cross Validation y de esta forma garantizar que las caracteristicas correlacionadas entre observaciones que puedan afectar nuestra predicción fueran tenidas en cuenta.

El fiasco de Zillow (Williams, 2021) es un ejemplo que muestra la importancia de monitorear la precisión de los modelos de machine learning en la predicción del precio de venta de propiedades. En este caso, los modelos no fueron actualizados a tiempo, lo que llevó a una sobreestimación de los precios debido a las condiciones cambiantes del mercado inmobiliario. Es crucial realizar un seguimiento continuo de la precisión de los modelos y actualizarlos en caso de que se produzcan cambios en las condiciones del mercado que afecten los precios de las propiedades. Aunque este estudio se centra en una predicción específica, se sugiere que la actualización de los modelos sea una práctica habitual en el futuro.

Referencias

Bourassa, S. C., Cantoni, E. y Hoesli, M. (2010). "Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods". En: *The Journal of Real Estate Research* 32.2, págs. 139-160. ISSN: 08965803. URL: http://www.jstor.org/stable/24888337 (visitado 08-03-2023).

Harrison, D. y Rubinfeld, D. (mar. de 1978). "Hedonic housing prices and the demand for clean air". En: *Journal of Environmental Economics and Management* 5, págs. 81-102. DOI: 10.1016/0095-0696(78)90006-2.

Mora-Garcia, R.-T., Cespedes-Lopez, M.-F. y Perez-Sanchez, V. R. (nov. de 2022). "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times". En: *Land* 11.11, págs. 1-32. URL: https://ideas.repec.org/a/gam/jlands/v11y2022i11p2100-d979663.html.

Park, B. y Bae, J. K. (2015). "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data". En: *Expert Systems with Applications* 42.6, págs. 2928-2934. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.11.040. URL: https://www.sciencedirect.com/science/article/pii/S0957417414007325.

Rosen, S. (1974). "Hedonic prices and implicit markets: product differentiation in pure competition". En: *The Journal of Political Economy* 82.1, págs. 34-55.

Truong, Q. et al. (2020). "Housing Price Prediction via Improved Machine Learning Techniques". En: *Procedia Computer Science* 174. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things, págs. 433-442. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.06.111. URL: https://www.sciencedirect.com/science/article/pii/S1877050920316318.

Varma, A. et al. (abr. de 2018). "House Price Prediction Using Machine Learning and Neural Networks". En: págs. 1936-1939. DOI: 10.1109/ICICCT.2018.8473231.

Williams, D. (dic. de 2021). *The $500MM Debacle at Zillow Offers: What Went Wrong with the AI Models?* URL: https://insidebigdata.com/2021/12/13/the-500mm-debacle-at-zillow-offers-what-went-wrong-with-the-ai-models/.

Yu, L. et al. (2018). "Prediction on Housing Price Based on Deep Learning". En: *International Journal of Computer and Information Engineering* 12.2, págs. 90-99. ISSN: eISSN: 1307-6892. URL: https://publications.waset.org/vol/134.

5. Anexos

Anexo 1

- **Price:** es el precio de venta del ofertante. Está tomada de la base de los datos recopilados de https://www.properati.com.co.
- **Month:** es el mes de publicación de la oferta en https://www.properati.com.co.
- **Year:** es el año de publicación de la oferta en https://www.properati.com.co.
- **Surface_total:** es el área total del terreno en el que se construyó la casa, incluyendo áreas cubiertas y descubiertas, por lo que incluye cualquier área exterior, como un jardín, una entrada de autos o un garaje. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co.
- **Surface_covered:** es el área total cubierta de los espacios construidos dentro de la casa. Da cuenta del espacio habitable real dentro de la casa. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co.
- **Rooms:** es la cantidad de habitaciones dentro de la propiedad. Es parte de las variables pre-existentes en la base de los datos recopilados de https://www.properati.com.co.
- **Bedrooms:** es la cantidad de cuartos para dormir dentro de la propiedad. Es parte de las variables pre-existentes en la base de los datos recopilados de https://www.properati.com.co.
- **Bathrooms:** es la cantidad de baños dentro de la propiedad. Es parte de las variables pre-existentes en la base de los datos recopilados de https://www.properati.com.co.
- **Property_type:** es el tipo de vivienda. Toma el valor de 1 si es una casa y de 0 si es un apartmento. Es parte de las variables pre-existentes en la base de los datos recopilados de https://www.properati.com.co.
- **Lat:** es la latitud de la coordenada de la vivienda. Es parte de las variables de Open Source Maps (OSM).
- **Lon:** es la longitud de la coordenada de la vivienda. Es parte de las variables de Open Source Maps (OSM).
- **Chapinero**: toma el valor de 1 si la vivienda hace parte de Chapinero y 0 de lo contrario. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co. La extracción se programó con múltiples posibles variaciones de los nombres de los barrios de Chapinero.

- **Terraza**: toma el valor de 1 si la vivienda tiene terraza y 0 de lo contrario. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co. La extracción se programó con múltiples posibles variaciones para referirse a una terraza, como balcon, balcn o mirador.

- **Social**: toma el valor de 1 si la vivienda cuenta con espacios sociales y 0 de lo contrario. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co. La extracción se programó con múltiples posibles variaciones para referirse a espacios sociales, como BBQ, piscina, salón comunal, entre otros.

- **Parqueadero**: toma el valor de 1 si la vivienda cuenta con parqueadero y 0 de lo contrario. Se construyó con extracción de texto de los datos recopilados de https://www.properati.com.co. La extracción se programó con múltiples posibles variaciones para referirse a parqueaderos, como cochera o garaje.

- **Distancia_parque**: es la distancia en metros de la vivienda al parque más cercano. Se construyó utilizando la ubicación de parques de Open Source Maps (OSM).

- **Distancia_gym**: es la distancia en metros de la vivienda al gimnasio más cercano. Se construyó utilizando la ubicación de gimnasios de Open Source Maps (OSM).

- **Distancia_transmi**: es la distancia en metros de la vivienda a la estación de bus más cercana. Se construyó utilizando la ubicación de las estaciones de bus de Open Source Maps (OSM).

- **Distancia_cai**: es la distancia en metros de la vivienda al Centro de Atención Inmediata (CAI) más cercano. Se construyó utilizando la ubicación de los police ammenities de Open Source Maps (OSM).

- **Distancia_gym**: es la distancia en metros de la vivienda al centro de entrenamiento físico más cercano. Se construyó utilizando la ubicación de centro de entrenamiento físico de Open Source Maps (OSM).

- **Distancia_bar**: es la distancia en metros de la vivienda al bar más cercano. Se construyó utilizando la ubicación de bares de Open Source Maps (OSM).

- **Distancia_SM**: es la distancia en metros de la vivienda al supermercado más cercano. Se construyó utilizando la ubicación de supermercados de Open Source Maps (OSM).

- **Distancia_colegios**: es la distancia en metros de la vivienda al colegio más cercano. Se construyó utilizando la ubicación de colegios de Open Source Maps (OSM).

- **Distancia_universidades**: es la distancia en metros de la vivienda a la universidad más cercana. Se construyó utilizando la ubicación de universidades de Open Source Maps (OSM).

- **Distancia_hospitales**: es la distancia en metros de la vivienda al hospital más cercano. Se construyó utilizando la ubicación de hospitales de Open Source Maps (OSM).

- **Surface_covered2:** es la superficie cubierta luego de recuperar los valores faltantes con las estimaciones de regresiones lineales en función del número de cuartos para dormir y el tipo de propiedad (si es o no una casa). Se creó añadiendo las estimaciones de las regresiones a la variable de superficie cubierta pre-existente en la base de los datos recopilados de https://www.properati.com.co utilizando dos variables independientes de esta misma base de datos.

- **Bathrooms2:** es la cantidad de baño luego de recuperar los valores faltantes con las estimaciones de regresiones lineales en función del número de cuartos para dormir y el tipo de propiedad (si es o no una casa). Se creó añadiendo las estimaciones de las regresiones a la variable de bañ pre-existente en la base de los datos recopilados de https://www.properati.com.co utilizando dos variables independientes de esta misma base de datos.

Anexo 2

Ecuación modelos de regresión

$$\begin{aligned} Precio_i = \beta_1 surfacecovered_i + \beta_2 bedrooms_i + \beta_3 bathrooms2_i + \beta_4 Chapinero_i \\ + \beta_5 property\_type_i + \beta_6 terraza_i + \beta_7 social_i + \beta_8 parqueadero_i \\ + \beta_9 distancia\_parque_i + \beta_{10} distancia\_gym_i \\ + \beta_{11} distancia\_transmi_i + \beta_{12} distancia\_cai_i + \beta_{13} distancia\_cc_i + \beta_{14} distancia\_SM_i \\ + \beta_{15} distancia\_colegios_i + \beta_{16} distancia\_universidades_i + \beta_{17} distancia\_hospitales_i \end{aligned} \tag{1}$$