Cali is Cali, the rest is hill...

Camilo Bonilla Rafael Santofimio Nicolás Velásquez

November 14, 2022

1 Introduction

En Colombia la fijación del precio de la vivienda varia de región a región, en buena parte depende del grado de desarrollo en términos económicos de cada ciudad, de los planes de ordenamiento territorial - POT (tipo uso de suelo), servicios que pueda prestar (transporte, ambiente, ocio y recreación, etc.), el estado del mercado (la oferta y demanda de inmuebles), entre otros. Así entonces, resulta retador establecer el precio de la vivienda de una ciudad específica y extrapolarlo a otra. Por lo tanto, este trabajo se centra en crear modelos que permiten predecir el precio de la vivienda en Cali, a partir de información del mercado inmobiliario de Medellín y Bogotá. Esta comprende datos relacionados con número de habitaciones, área construida, servicios disponibles para cada barrio, y factores relacionados con el transporte, seguridad y ambiente. Con base en esto, ofrecer el servicio técnico a una empresa interesada en el sector inmobiliario de Cali.

En aras de estructurar modelos de predicción basados en machine learning (ML), que eviten el fisco de Zillow, se prioriza la identificación de la degradación de los precios en el mercado, a través del reentrenamiento de modelos con información nueva, es decir, se da mayor peso a la información actualizada, en lugar de los datos históricos. Así entonces, se desarrollan siete (7) modelos de predicción los cuales son: Regresión Lineal, Lasso, Ridge, Elastic Net, Random Forest, XG Boost y finalmente una combinación de todos en una consolidación del modelo Super Learner. Para definir los hiperparámetros se corrió una grilla de búsqueda muy grande, y así definir las mejores penalizaciones, mixtura, numero de árboles, profundidad, numero de predictores, taza de aprendizaje y tamaño del nodo.

Se encontró que variables como área, numero de habitaciones y baños, distancia más cercana a ríos, universidades, estaciones de policía, parques y hospitales, e índice de luminosidad (aprox. Productividad) son variables determinantes a la hora de predecir el precio de la vivienda. A su vez, se pudo establecer que los modelos básicos de regresión lineal, ridge, lasso y elasctic si bien son rápido en el procesamiento, su poder de predicción con datos espaciales es bajo comparado con random forest (RF) y Xgboost (xgb), pues en el mejor de los casos el error absoluto medio no disminuyó más allá de 30%. Por su parte, RF y xgb alcanzaron a bajarlo a 24.1 y 25.2% respectivamente. Por otro lado, se estableció que si bien el super learner ofrece mayor capacidad de predicción, si los modelos por separadas presentan mal desempeño, sus resultados son correspondiente a este (garbage in, garbage out).

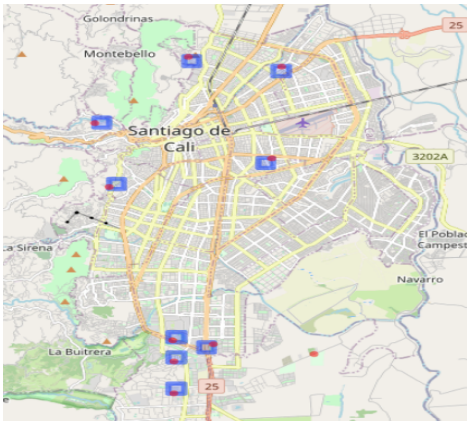
El documento está estructurado así: II. Data, se presenta la fuente, estructura y tipo de información, estadísticas descriptivas y justificación de los datos de entrenamiento y testeo; III. Modelos y resultados, se reporta las variables, modelos, hiperparámetros y principales resultados; IV. Conclusiones y recomendaciones, se describe los principales hallazgos y aportes de los modelos en la predicción de los precios de la vivienda em Cali, y sus respectivas limitaciones; V. Apéndice, contiene las referencias empleadas para el trabajo, y algunas tablas, gráficos y mapas que apoyan la descripción de los datos. Los archivos, tablas, bases de datos y códigos totalmente replicables se pueden obtener en el siguiente link: ProblemSet3 el cual redirige al repositorio en Github.

2 Data

Para la consolidación de la base de datos se recurrió a diversas fuentes. Por un lado, del portal web PROPERATI se extrajo información relacionada con precio de venta, área, número dormitorios, baños, tipo de vivienda (casa o apartamento), coordenadas (WGS84) y una descripción del inmueble. Con esta última, a través de análisis de texto (expresiones regulares) se obtuvo valores faltantes para algunas de estas variables, así mismo, reportes correspondientes a servicios interiores y exteriores adicionales que presta el inmueble como por ejemplo balcón, terraza, piscina, gimnasio, etc.

Por otro lado, con el ánimo de ampliar las variables que pueden explicar y/o predecir el precio de la vivienda en estas ciudades, se recurrió a Open Street Map. De este se obtuvo información correspondiente a los servicios ofrecidos por las ciudades tales como: supermercados, centros comerciales, estaciones de servicio de combustible, estaciones policía, hospitales, estadios, ríos, universidades, estaciones y paradas del transporte público. Con base en estas, se calculó el número de localizaciones que estaban en un radio de 150 metros a cada vivienda, y la distancia más cercana a estos. A su vez, fue necesario establecer el estrato socioeconómico de cada inmueble a partir de los servicios públicos domiciliarios. Por lo que se recurrió al Ceso Nacional de 2018 (DANE, 2018) y al geoportal del (DANE, 2021) para descargar los registros a nivel de municipio y manzana.

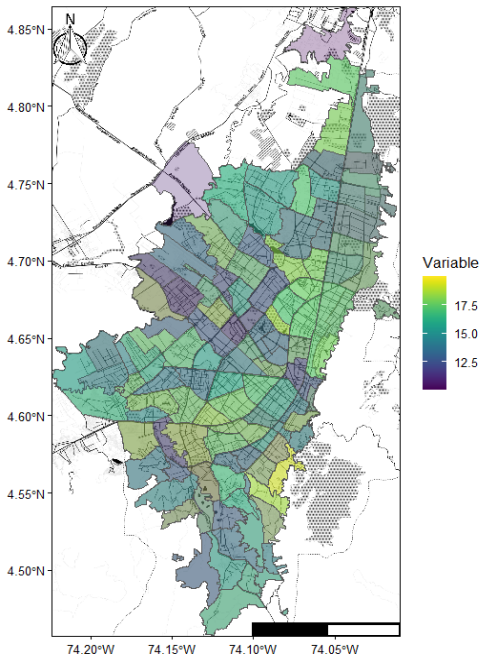
Figure 2.1: Luminosidad de Cali imputada por pixel



Nota: Se imputan el valor de cada pixel lumínico a cada vivienda que se encuentra dentro de este *Fuente*: Elaborado a partir de (EOG, 2021)

Adicionalmente, para obtener información correspondiente a la productividad (actividad económica) se construyó una aproximación de esta a partir de la actividad lumínica para cada ciudad. Los datos se descargaron del portal Earth Observation Group, los cuales corresponde a una capa tipo raster del 2021. Con base en estos, se imputó la iluminación para las viviendas según el valor del píxel (mapa) que se sobrepone a cada observación especifica. En la figura 2.1. se muestra la aproximación al píxel para una muestra aleatoria de 10 observaciones en Cali.

Figure 2.2: Actividad Lumínica Bogota por UPZ



Nota: Esta figura muestra el promedio de la actividad lumínica para Bogotá en el año 2021 a nivel de UPZ. *Fuente*: Elaborado a partir de (EOG, 2021)

Este mismo tratamiento se aplicó a para Bogotá y Medellín a nivel de UPZ y barrio respectivamente, la diferencia radica en la disponibilidad de información para estas subdivisiones administrabas en la capa tipo raster. En la figura 2.1 se observa que el centro de la ciudad y zonas como Puente Aranda, el Centro y Chapinero tienen mayor iluminación, y, por lo tanto, mayor actividad económica. En apéndice se muestra los mapas para Cali y Medellín.

Finalmente, se obtuvieron 30 variables, de las cuales casi todas tenían datos faltantes (N/A), para esto fue necesario imputar con base a los atributos que tenían los vecinos en un radio de 50 metros implementado la estrategia tipo "queen", a pesar de esto, en algunos casos el problema persistía, por lo que se tuvo que ampliar el radio a 150 metros. En el caso puntual del estrato se implementó esta misma técnica, pero fijando únicamente como criterio la moda entre los vecinos. Despuesta de estas depuraciones, se conformó una base para Bogotá, Medellín y Cali con 37985, 13452 y 5000 observaciones respectivamente.

Según los datos reportados en estas bases, se encuentra que en promedio una vivienda en Bogotá cuesta $ 230.509.186,50 más que en Medellín, el 75% de los inmuebles están entre 180 a 220 M2 para las tres ciudades, el máximo número de habitaciones y baños son 11 y

13 respectivamente. Por otro lado, una vivienda en Medellín en promedio cuenta con 18 parques más que Bogotá en un radio de 500 metros, y 50 más que Cali. La luminosidad (aproximación actividad económica) en el Distrito Capital es 2.1 veces mayor que Cali y 0.5 veces más que en Medellín. Esto da cuenta la heterogeneidad entres estas ciudades que se comentaba en la introducción, por lo que resulta un reto predecir el precio de la vivienda. Las tablas con estadísticas descriptivas y algunos histogramas se reportan en al apéndice.

3 Models and results

Se diseñaron y calibraron siete (7) modelos distintos, dentro de estos se encuentran Regresión Lineal, Lasso, Ridge, Elastic Net, Random Forest, XG Boost y finalmente una combinación de todos en una consolidación del modelo Super Learner. Para definir los hiperparámetros se corrió una grilla de búsqueda muy grande (ver Tabla 1), y así definir las mejores penalizaciones, mixtura, numero de árboles, profundidad, numero de predictores, taza de aprendizaje y tamaño del nodo según el caso.

Table 1: Grillas inicial de búsqueda

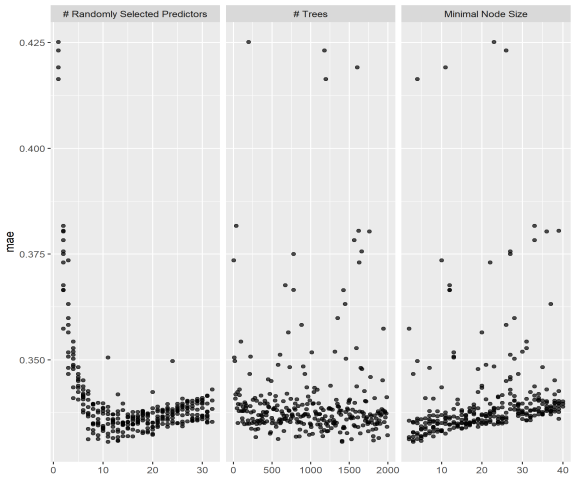
Hiperparámetros	Min.	Max.	Particiones
Penalizacion	0.01	0.8	100
Mixtura	0.02	0.05	100
Árboles	10	2000	1000
Profundidad	2	16	8
Folds	5	8	3
Tamaño nodo	2	40	20
Predictores	5	35	30
Tasa aprendizaje	0.01	0.1	100
Reducción pérdida	10ˆ-8	10ˆ-1	100

Para medir el desempeño de los modelos se utilizó la medida del MAE, o promedio del error absoluto el cual es una medida de errores entre observaciones emparejadas que expresan el mismo fenómeno, en este caso la variable continua del precio, al cual queremos que sea un valor lo más pequeño posible sin llegar a sobreajustar sobre la muestra.

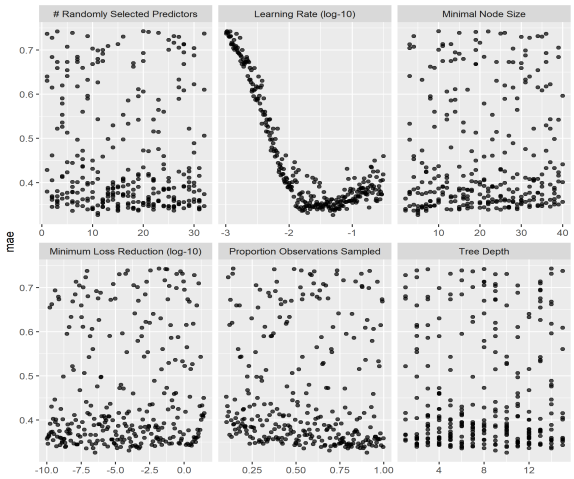
Particularmente se encuentra que los modelos de ridge, lasso, elasticnet su tiempo de procesamiento es sustancialmente inferior a los otros dos modelos, sin embargo, su poder de predicción con los datos espaciales es relativamente bajo (error absoluto medio superior a 30%), al menos para este problema. Con base en esto se establece que variables como área, número de habitaciones y baños, distancia más cercana a ríos, universidades, estaciones de

policía, parques y hospitales, e índice de luminosidad (aprox. Productividad) son variables determinantes a la hora de predecir el precio de la vivienda.

Figure 3.1: Grilla de busqueda Random Forest y Xgboost



(a) RF



(b) XGB

Fuente: Elaborado a partir de (DANE, 2018), (DANE, 2021), (EOG, 2021), (OSM, 2022) y (PROPERATI, 2022)

Así entonces, en la tabla 2 se pueden ver los resultados de la métrica MAE (Mean Absolut Error) para los mejores modelos en la base de train y su desempeño, en esta podemos observar que los modelos con mejores predicciones fueron bosques aleatorios con un MAE de 0.254, este modelo se calibró con 11 predictores al azar que se tomaron en cada iteración, 790 árboles y 3 en tamaño mínimo del último nodo. Por otro lado, xgBoost mejoró un poco en la predicción reportando un MAE de 0.244. Este modelo se obtuvo con 8 predictores al azar que se tomaron en cada iteración, 5 en tamaño mínimo del último nodo, 11 profundidad máxima del árbol, una tasas de aprendizaje de 0.04, reducción de pérdida de 0.000000004 y un tamaño de la muestra de 0.92.

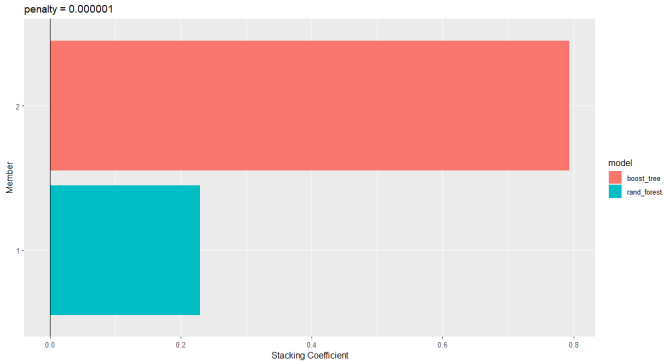
Table 2: Desempeño de modelos

Modelo	MAE
Linear Model	0.442
Lasso	0.372
Ridge	0.376
Elastic Net	0.365
Random Forest	0.254
XgBoost	0.244

Adicionalmente se generó un modelo que combinaba a cada uno de los otros modelos. El Superlearner le dió pesos a cada una de las predicciones para generar un mejor modelo. En este caso como se observa en la figura 3.2, el Super Learner solo escogió los modelos random

forest y XgBoost, dandole un peso apróximado de 80% al último y el 20% restante al bosque aleatorio.

Figure 3.2: SuperLearner importancia de modelos



4 Conclusions and recommendations

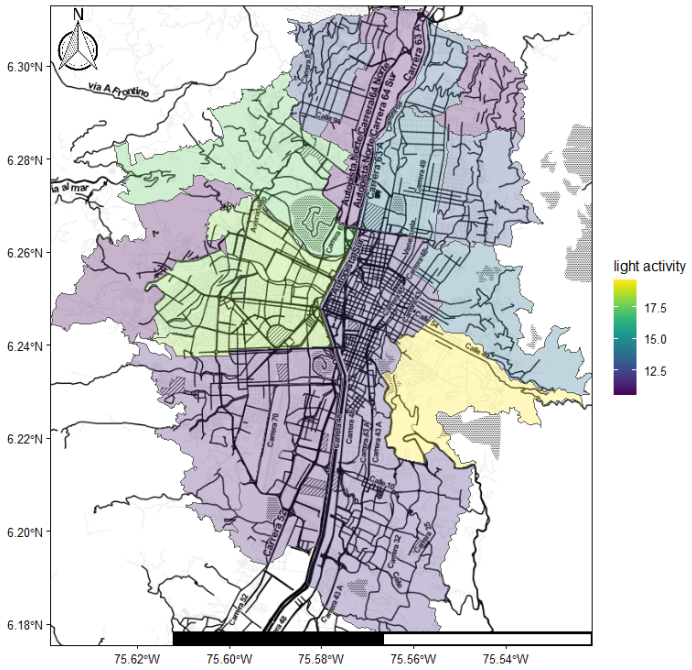
A la luz de los principales reportados en la sección de "Models and results", se considera que los siete (7) modelos en general tienen un poder predictivo relativamente bajo, pues el MAE no llega al menos al 10%, esto puede deberse a omisión de una o más variables que explican la fijación de los precios de la vivienda en Cali, pese a que se consolidó una base de datos con diversas fuentes de información, tales como distancias de los inmuebles a servicios prestado por la ciudad, estratificación, áreas, y datos ráster de luminosidad. Las variables más relevantes de los modelos fueron área total, número de baños y habitaciones, distancia a río, hospital, estación de policía y parques. Esto puede ser indicador de la relativa poca importancia de las demás variables, ya que, al inicio del proyecto se esperaba que variables como estrato o actividad económica fueran que tuvieran mayor peso.

Adicionalmente, se está lidiando con la predicción de una ciudad como Cali la cual no estaba en la base de train. Para poder ajustar las predicciones a esta base se ajustó bajo la media y desviación estándar que se reportaba en el enunciado del programa siendo una media de precios para Cali $555'314.430 de y desviación estándar de $601'842.533. Teniendo en cuenta que el proceso generador de datos per se, viene de una población distinta y pude generar grandes desalineaciones entre los valores predichos y los reales. El valor esperado de la varianza de esa población es distinto y lidiar con esto, junto con el ruido de la propia muestra de entrenamiento y el ruido de la muestra sobre la cual se quiere predecir genera desafíos sumamente importantes.

Se considera que aproximación a la predicción de los precios de las viviendas en Cali puede estar desfasada al objetivo general que era en función de una empresa que quiere comprar y sacar rentabilidad de la diferencia entre el precio estimado y el precio reportado por el propietario. Se espera que haya rentabilidad en lo grueso de la masa de la observación entre los estratos 2 y 4. Para variables atípicas o muy lejos de la distribución las predicciones pueden estar lejanas a los precios reales.

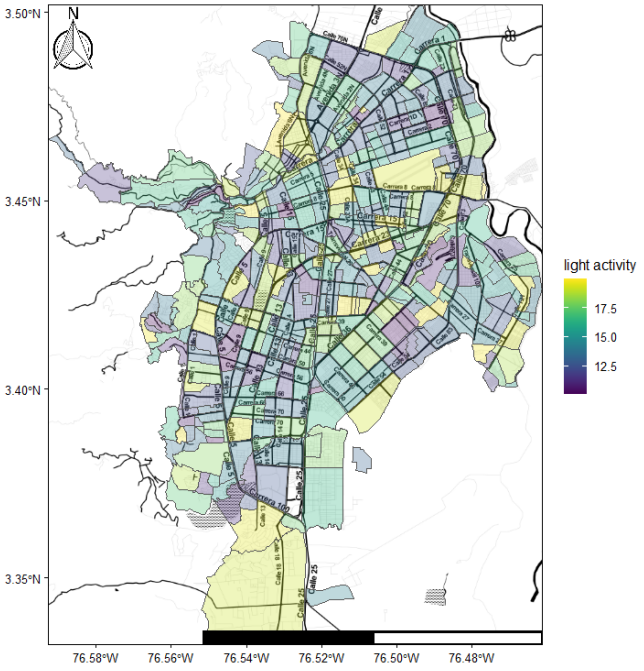
5 Appendix

Figure 5.1: Actividad Lumínica Medellín por Comuna.



Note: Esta figura muestra el promedio de la actividad lumínica para Medellín en el año 2021 a nivel de comuna.
Fuente: Elaborado a partir de (EOG, 2021)

Figure 5.2: Actividad Lumínica Cali por barrio.



Note: Esta figura muestra el promedio de la actividad lumínica para Cali en el año 2021 a nivel de Barrio.
Fuente: Elaborado a partir de (EOG, 2021)

Table 3: Estadísticas descriptivas datos vivienda Bogotá

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Precio	37985	869755897.438	899818885.944	2e+08	3.8e+08	9.9e+08	2e+10
area M2	37922	156.854	173.559	2	82.5	188	10125
Numero de habitaciones	37985	3.079	1.463	0	2	3	11
Numero de baños	37942	2.738	1.241	1	2	3	13
Tipo de vivienda (1 si es casa)	37985	0.231	0.422	0	0	0	1
Sala	37943	0.979	0.145	0	1	1	1
Servicios interiores adicional	37922	0.963	0.188	0	1	1	1
Servicios exteriores adicional	37835	0.635	0.482	0	0	1	1
garage	37931	0.978	0.146	0	1	1	1
Luz natural	37834	0.416	0.493	0	0	1	1
estrato	37722	3.991	1.262	1	3	5	6
Numero de gasolineras a 500 mts	37985	0.914	1.178	0	0	1	8
Distancia gasolinería mas cercana	37985	525.627	312.128	0	289.996	711.981	3348.896
Numero de hospitales a 500 mts	37985	0.415	1.1	0	0	0	10
Distancia hospital mas cercano	37985	970.174	612.976	0	501.227	1343.77	5733.663
Numero de estaciones policia a 500 mts	37985	0.26	0.517	0	0	0	6
Distancia estacion policia mas cercana	37985	918.331	499.528	0	527.926	1275.097	3480.005
Numero de parques a 500 mts	37985	10.47	6.713	0	6	13	53
Distancia parque mas cercano	37985	119.551	113.553	0	43.744	162.813	1962.966
Numero de rios a 500 mts	37985	0.008	0.093	0	0	0	2
Distancia rio mas cercano	37985	4241.51	1765.144	10.71	2930.463	5370.541	8174.077
Numero de universidades a 500 mts	37985	0.537	1.641	0	0	1	20
Distancia universidad mas cercana	37985	931.828	575.882	0	494.811	1277.47	4455.668
Numero de estaciones transporte a 500 mts	37985	0.907	1.795	0	0	1	13
Distancia estacion transporte mas cercana	37985	953.325	655.652	0	439.946	1344.13	5945.504
Numero de Supermercado a 500 mts	37985	1.437	1.614	0	0	2	12
Distancia supermercado mas cercano	37985	457.706	306.537	0	241.71	598.315	3906.213
Numero de C.comercial a 500 mts	37985	0.896	1.374	0	0	1	11
Distancia C.Comercial mas cercano	37985	633.854	392.36	0	340.796	873.14	4601.649
Indice de iluminacion	37673	51.494	11.907	11.065	42.871	58.56	136.119

Fuente: (DANE, 2018), (DANE, 2021), (EOG, 2021), (OSM, 2022) y (PROPERATI, 2022)

Table 4: Estadísticas descriptivas datos vivienda Medellín

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
price	13452	639246710.963	623222205.048	2e+08	3.2e+08	7.3e+08	1.8e+10
area M2	13430	243.671	3261.103	15	86.267	182.175	198000
Numero de habitaciones	13452	3.147	1.024	0	3	3	11
Numero de baños	13418	2.536	1.244	1	2	3	20
Tipo de vivienda (1 si es casa)	13452	0.202	0.401	0	0	0	1
Cuenta con sala o comedor	13438	0.996	0.06	0	1	1	1
Servicios interiores adicional	13439	0.994	0.076	0	1	1	1
Servicios exteriores adicional	13416	0.681	0.466	0	0	1	1
garage	13426	0.989	0.104	0	1	1	1
Luz natural	13371	0.202	0.402	0	0	0	1
estrato	13321	4.193	1.268	1	3	5	6
Numero de gasolineras a 500 mts	13452	0.848	1.253	0	0	1	10
Distancia gasolinería mas cercana	13452	616.736	364.048	0	326.179	867.088	2589.78
Numero de hospitales a 500 mts	13452	0.501	0.824	0	0	1	5
Distancia hospital mas cercano	13452	752.104	456.127	0	412.423	1023.552	2751.782
Numero de estaciones policia a 500 mts	13452	0.132	0.381	0	0	0	2
Distancia estacion policia mas cercana	13452	1969.271	1351.954	0	757.323	3209.722	5568.588
Numero de parques a 500 mts	13452	4.933	4.793	0	1	7	72
Distancia parque mas cercano	13452	261.989	242.921	0	93.608	348.333	2047.942
Numero de rios a 500 mts	13452	0.145	0.423	0	0	0	3
Distancia rio mas cercano	13452	1565.377	815.826	0	929.983	2174.549	4567.176
Numero de universidades a 500 mts	13452	0.431	1.008	0	0	0	8
Distancia universidad mas cercana	13452	1182.716	756.079	0	533.225	1771.706	3958.562
Numero de estaciones transporte a 500 mts	13452	0.354	1.026	0	0	0	8
Distancia estacion transporte mas cercana	13452	1175.889	711.632	0	650.103	1503.194	4570.299
Numero de Supermercado a 500 mts	13452	0.567	0.874	0	0	1	5
Distancia supermercado mas cercano	13452	765.78	511.496	0	359.64	1088.31	3126.206
Numero de C.comercial a 500 mts	13452	0.829	1.133	0	0	1	8
Distancia C.Comercial mas cercano	13452	590.173	397.157	0	302.481	834.107	3362.437
Indice de iluminacion	13092	34.275	12.916	8.012	23.261	44.771	76.717

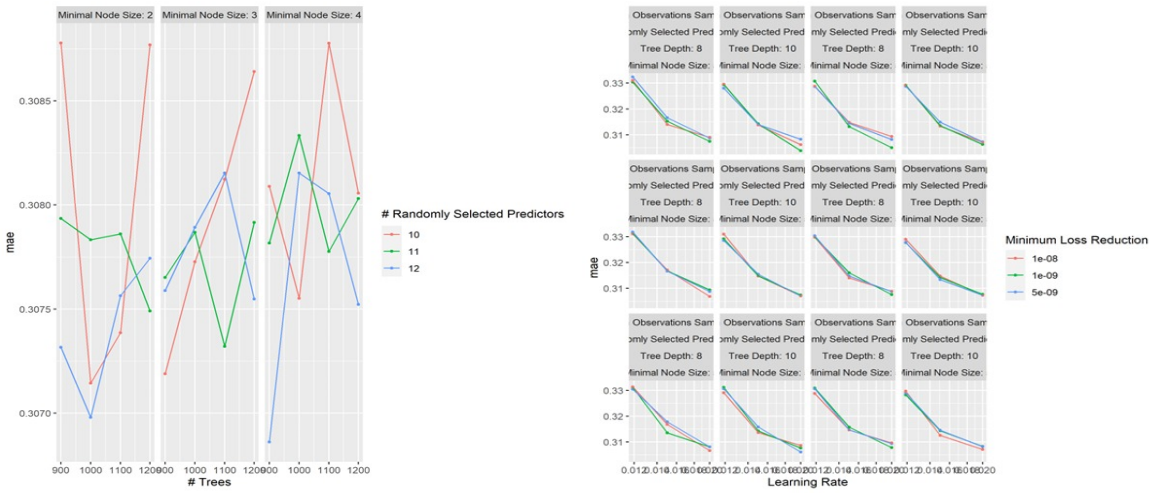
Fuente: (DANE, 2018), (DANE, 2021), (EOG, 2021), (OSM, 2022) y (PROPERATI, 2022)

Table 5: Estadísticas descriptivas datos vivienda Cali

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
price	0	NaN		Inf			-Inf
area M2	4970	221.707	604.454	10	92	220	26000
Numero de habitaciones	5000	3.748	1.603	0	3	4	11
Numero de baños	4985	3.135	1.501	1	2	4	13
Tipo de vivienda (1 si es casa)	5000	0.445	0.497	0	0	1	1
Cuenta con sala o comedor	4994	0.991	0.092	0	1	1	1
Servicios interiores adicional	4981	0.992	0.089	0	1	1	1
Servicios exteriores adicional	4953	0.897	0.304	0	1	1	1
garage	4973	0.99	0.099	0	1	1	1
Luz natural	4882	0.289	0.453	0	0	1	1
estrato	4877	4.059	1.257	1	3	5	6
Numero de gasolineras a 500 mts	5000	1.008	1.542	0	0	2	11
Distancia gasolinería mas cercana	5000	654.862	467.485	0	325.689	859.277	3473.14
Numero de hospitales a 500 mts	5000	0.344	0.842	0	0	0	6
Distancia hospital mas cercano	5000	1155.598	893.968	6.639	548.71	1468.897	5734.853
Numero de estaciones policia a 500 mts	5000	0.137	0.383	0	0	0	2
Distancia estacion policia mas cercana	5000	1126.372	560.733	12.609	712.861	1481.776	3412.63
Numero de parques a 500 mts	5000	1.284	2.892	0	0	1	18
Distancia parque mas cercano	5000	612.431	407.097	0	301.786	870.304	2177.362
Numero de rios a 500 mts	5000	0.097	0.381	0	0	0	4
Distancia rio mas cercano	5000	1957.714	1062.956	16.82	1125.625	2637.253	6126.695
Numero de universidades a 500 mts	5000	0.033	0.178	0	0	0	1
Distancia universidad mas cercana	5000	3750.244	1945.245	3.772	2114.839	5439.197	7627.336
Numero de estaciones transporte a 500 mts	5000	0.37	1.164	0	0	0	8
Distancia estacion transporte mas cercana	5000	1326.193	741.823	3.78	762.677	1741.712	4797.262
Numero de Supermercado a 500 mts	5000	0.351	0.574	0	0	1	3
Distancia supermercado mas cercano	5000	767.269	464.871	0	431.921	998.6	3721.16
Numero de C.comercial a 500 mts	5000	0.26	0.889	0	0	0	9
Distancia C.Comercial mas cercano	5000	1125.504	619.346	0	673.704	1529.296	4167.574
Indice de iluminacion	4545	17.679	9.706	0	11.568	23.141	56.289

Fuente: (DANE, 2018), (DANE, 2021), (EOG, 2021), (OSM, 2022) y (PROPERATI, 2022)

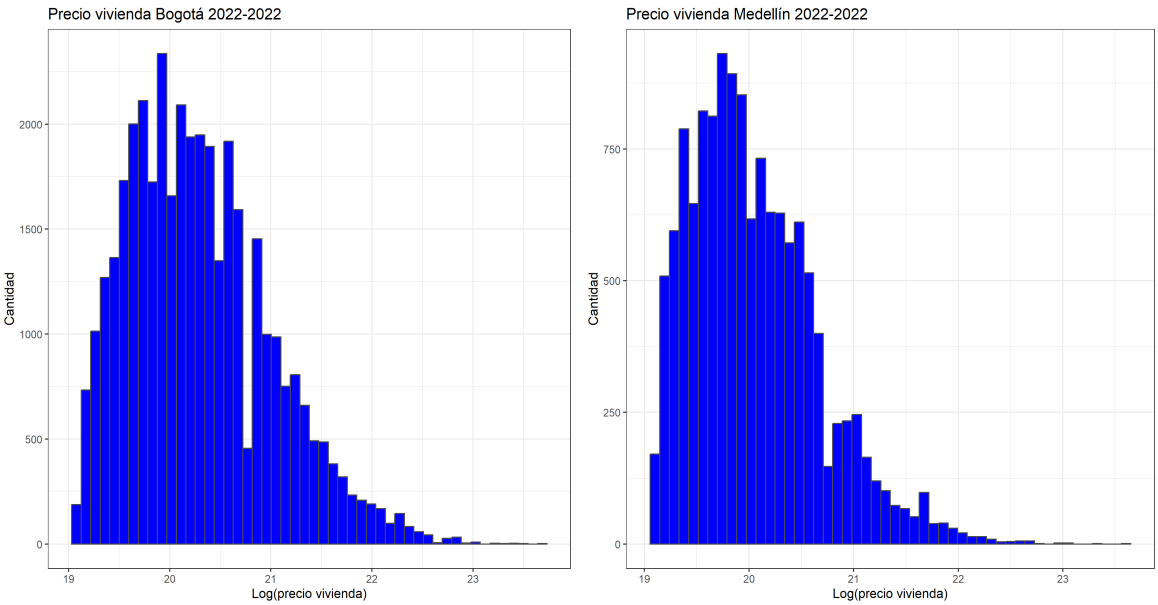
Figure 5.6: Grilla final Random Forest y Xgboost



(a) RF (b) XGB

Fuente: Elaborado a partir de (DANE, 2018), (DANE, 2021), (EOG, 2021), (OSM, 2022) y (PROPERATI, 2022)

Figure 5.3: Log. precio de la vivienda 2020-2022



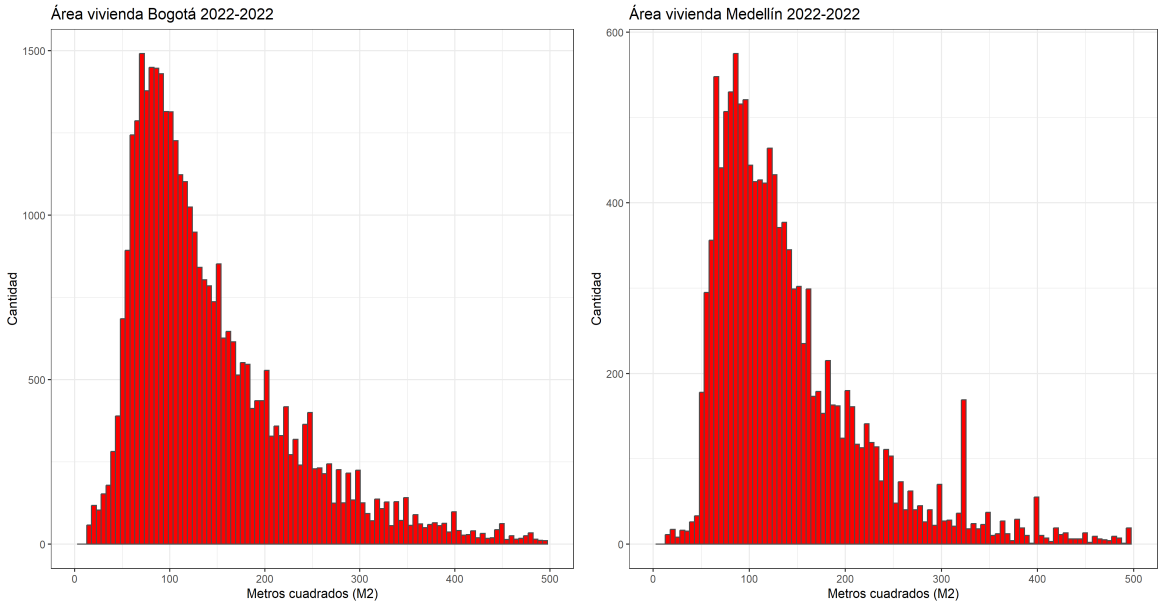
(a) Bogotá (b) Medellín

Fuente: Elaborado a partir de (PROPERATI, 2022)

Table 6: Importancia de variables de modelos

Variable	Var importance
surface_total	0.274
bathrooms	0.134
closest_river	0.110
bedrooms	0.0513
closest_police	0.0457
closest_mall	0.0361
closest_supermarket	0.0338
closest_university	0.0337
lum_val	0.0333
closest_station	0.0332
estrato_X6	0.0318
closest_park	0.0312
closest_fuel	0.0309
closest_hospital	0.0288
less_500m_park	0.0236

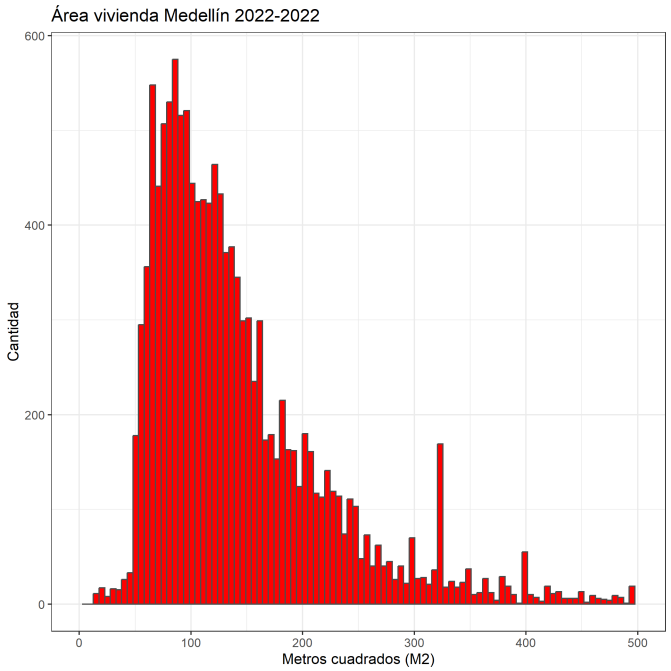
Figure 5.4: Área (M2) vivienda Bogotá y Medellín 2020-2022



(a) Bogotá (b) Medellín

Fuente: Elaborado a partir de (PROPERATI, 2022)

Figure 5.5: Área (M2) vivienda Cali 2020-2022



Fuente: Elaborado a partir de (PROPERATI, 2022)

References

DANE (2018). Censo nacional de población y vivienda. page np. Disponible en: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018.

DANE (2021). Marco geoestadistico nacional (mgn). page np. Disponible en: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/.

EOG (2021). Earth observation group. *Earth Observation Group*, page np. Disponible en: https://eogdata.mines.edu/products/vnl/.

OSM (2022). Openstreetmap. *OpenStreetMap*, page np. Disponible en: https://wiki.openstreetmap.org/wiki/ES:P%C3%A1gina_principal.

PROPERATI (2022). properati.com portal web. page np. Disponible en: https://www.properati.com.co/.