Problem Set 3: *Making Money with ML?* "It's all about location location location!!!"

Luisa Cuellar-201613942
Daniel Mendivelso-201513296
Isabella Riveros – 201923015
Link del repositorio: https://github.com/iriverosu/Problem-set-3.git

Introducción

En los últimos años la capital del valle se ha consolidado como centro inmobiliario para inversionistas y hogares, debido a las favorables dinámicas de mercado que han permitido la consolidación de una amplia oferta de proyectos, sobre todo en el norte y centro-sur de la ciudad. De acuerdo con algunos estudios de mercado, el aumento de la oferta y demanda de inmuebles en la ciudad se debe al incremento de la calidad de vida y mejor relación costo beneficio de las viviendas. Un modelo de predicción del precio de la vivienda en este contexto es útil porque, por un lado, facilita el emparejamiento entre vendedores y compradores al disminuir las asimetrías de información que puedan estar generando ineficiencias en este mercado, y, por otro lado, contribuye a predecir las tendencias futuras de los precios de la vivienda, y el comportamiento de este mercado después de ciertos períodos de tiempo o luego de picos económicos como por ejemplo, el ocasionado por una pandemia. Dicho modelo predictivo, contribuye a estos objetivos al identificar qué condiciones geográficas y características de la vivienda son las que están determinando su precio comercial.

Teniendo en cuenta lo anterior, el presente proyecto es una respuesta al desafío asociado a la predicción de los precios de la vivienda en la ciudad de Cali, a través del uso de un modelo Random Forest entrenado con datos espaciales para las ciudades de Bogotá y Medellín. Los datos utilizados contienen información de algunas características de los inmuebles y del espacio donde están ubicados (13 variables en total). Los resultados obtenidos muestran que el modelo RF predice con un error cuadrático medio menor en comparación con la predicción de modelos de regresión lineal (OLS), regularización (Elastic net), XG-boots y backward/forward. Así mismo, las predicciones obtenidas para los precios de 5000 casas en la ciudad de Cali muestran como el precio por metro cuadrado se comporta acorde a lo observado en los últimos años, donde en el sur de la ciudad se aglomeran las viviendas con mayor precio comercial, las cuales están asociadas a una mayor área superficial construida, un mayor número de baños y una ubicación en zonas con menores índices de inseguridad, cercanas a universidades, colegios de mayor nivel y/o zonas de trabajo, pero lejanas a zonas industriales y algunos centros comerciales. Este modelo cuenta con la ventaja de predecir el precio de las viviendas eficientemente con pocas variables, pero tiene la desventaja de entrenarse con datos de otras ciudades, limitando los resultados al supuesto de que los mercados inmobiliarios son semejantes y comparables entre ciudades.

Datos

La muestra de entrenamiento está compuesta por 2 ciudades principales de Colombia (Bogotá, Medellín), donde las observaciones se pueden encontrar en dos categorías especiales, urbano y rural. En total, la base reúne 13 variables y 51,437 observaciones. Para el caso de la muestra de testeo, corresponden a 5000 casas ubicadas en la zona urbana de la ciudad de Cali, las cuales cuentan con las mismas 13 variables de la base de entrenamiento. Para este caso, se dividió la base de entrenamiento entre muestra de entrenamiento y validación, resultando en una muestra de 41.150 y 10.287 observaciones respectivamente.

Las variables de interés utilizadas para predecir el precio de las viviendas en Cali se dividen en dos grupos, por un lado, las correspondientes a las características propias de la vivienda, donde se encuentran: el número de habitaciones, tipo de propiedad, la superficie total y el número de baños, y

por otro lado, las correspondientes a la zona donde está ubicada la casa. Particularmente, por medio de la aplicación Open Street Maps se construyeron 9 variables que miden la distancia de cada vivienda a algunos "amenities" que se agrupan en las siguientes dimensiones: transporte, educación, comercio, industria, y seguridad.

Particularmente, en la dimensión de educación se tomaron en cuenta las distancias de las viviendas a universidades, colegios y jardines escolares, pues la literatura ha evidenciado que las casas cercanas a colegios y universidades de alta calidad impactan positivamente el precio de las viviendas. Asimismo, en la dimensión de transporte se determinó la distancia a estaciones de bus y vías principales, pues el tener mayor acceso a estas puede ser un determinante en la decisión de los compradores de vivienda y por ende en su precio final. No obstante, cabe resaltar que la dirección de dicho impacto depende del tipo de transporte y de las dinámicas particulares de cada ciudad. Además, para el caso de la dimensión de comercio e industria se calcularon las distancias de las viviendas a los centros de oficinas, a las zonas industriales y a los centros comerciales, ya que estar cerca a estos puntos puede aumentar o disminuir el valor de la vivienda, dependiendo de otros factores socioeconómicos. Finalmente, en términos de seguridad, se utilizó la distancia al Centro de Atención Inmediata más cercano, como proxi que indique el nivel de seguridad de la zona.

Para la elección de dichas variables se hizo uso de la literatura donde se ha demostrado en distintos contextos su relevancia. Por ejemplo, Caracas (2021) muestra que la cercanía a centros comerciales, a centros de esparcimiento o recreación y a establecimientos educativos son determinantes del precio promedio por metro cuadrado de una vivienda. Asimismo, Alzate (2019) define el área construida, el tipo de vivienda, el estrato, el número de baños y habitaciones como determinantes del precio de la vivienda en Rionegro. De la misma manera, Martínez (2018), determina los factores que influyen en el precio de la vivienda en Cali, mostrando que existe una agrupación espacial del precio, es decir, proyectos con precios altos están rodeados por proyectos con precios altos, explicado principalmente por las características propias de la vivienda, su cercanía con el centro tradicional de empleo, el acceso a transporte y las vías principales de la ciudad. Por último, en el trabajo de Peña (2005) se identifica que la seguridad ciudadana, medida como la diferencia entre las tasas de denuncia de delitos de cada mes. es clave en la determinación del precio de las viviendas en Santiago de Chile.

Teniendo en cuenta lo anterior, la estructura de este análisis descriptivo evalúa la interacción entre las variables de interés y los amenities mencionados sobre el comportamiento en el precio de las viviendas. Particularmente, como se observa en la tabla 1, la base de datos muestra, que en promedio los inmuebles que tienen la categoría de "casa" valen cerca de un 30% más frente a aquellos inmuebles que son clasificados como "apartamentos". De igual forma, el comportamiento en el tamaño del área de las casas y apartamentos es heterogéneo, ya que los apartamentos en promedio cuentan con una superficie total más pequeña que los apartamentos (196m2, frente a 264m2, en su mismo orden). En cuanto al número de habitaciones, observamos el mismo comportamiento que los datos ya mencionados anteriormente, los apartamentos tienen en promedio un menor número de habitaciones frente a las casas (2.71, 4.43, respectivamente).

Por otra parte, en cuanto a las variables asociadas a las distancias de las viviendas a estaciones de buses, Centros de Atención Inmediata (CAI), colegios, universidades, centros comerciales, zonas industriales y zonas de oficinas, se encuentra que los colegios y los puntos de comercio tienen la mayor proximidad a las propiedades (391.748 mts, 426.534 mts, en su mismo orden). En cambio, la distancia hacia zonas industriales presenta el mayor tramo con 1467 mts. Si bien es cierto, que estas distancias se agrupas en general tanto para Bogotá y Medellín, los datos varían a un mayor nivel de especificidad.

Paralelamente, si observamos la Figura 1, encontramos que las propiedades que concentran el mayor valor en la ciudad de Bogotá se encuentran ubicadas entre el Norte hacia el Centro de la ciudad, con una mayor proximidad hacia los Cerros Orientales. Dentro de este espacio urbano se concentra la mayor oferta de oficinas, centros comerciales, y estaciones de bus. Sin embargo, el comportamiento de otros

amenities como colegios, y CAI es muy homogéneo en Bogotá debido a que está distribuido por toda la ciudad. En cuanto a las zonas industriales, es claro que en el caso de la ciudad de Bogotá parece castigar el precio de las viviendas que son circundantes de estos puntos.

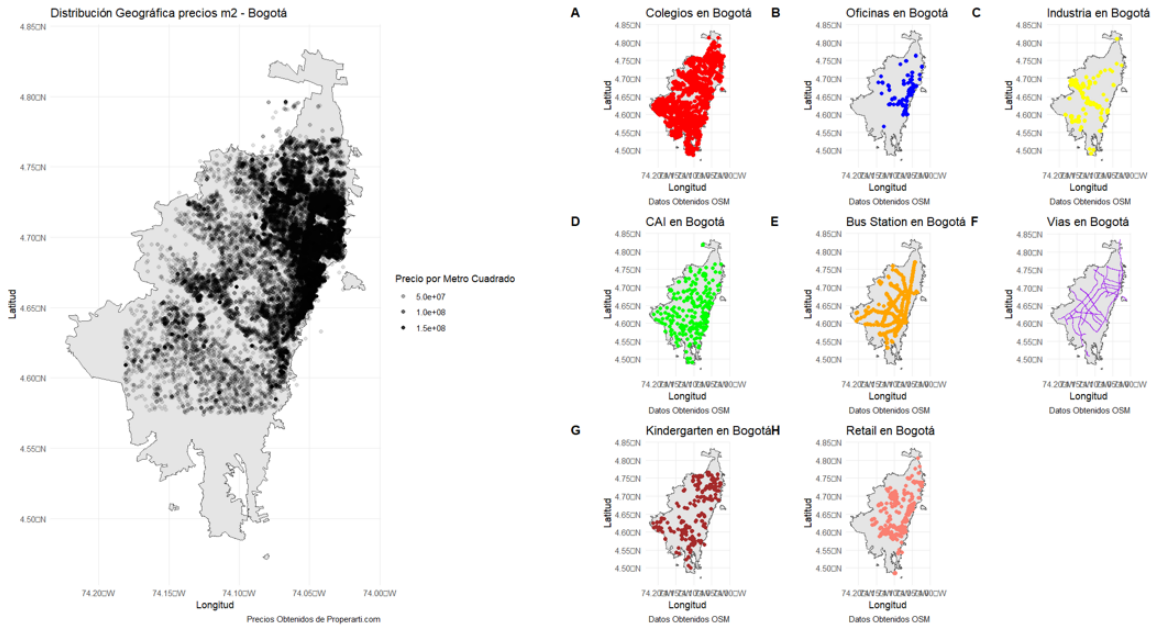
Entre tanto, la ciudad de Medellín no concentra una zona particular hacia un punto cardinal en específico en el valor de las propiedades, ya que a diferencia de Bogotá el precio incrementa tanto en el occidente como en el oriente de la ciudad, sin embargo, en este caso el Sur de la ciudad parece favorecer el precio de las viviendas a diferencia de Bogotá. En cuanto a los amenities, la ubicación de colegios, oficinas y vías principal parecen impulsar el precio de las viviendas, de igual forma, los puntos de retail también impulsan el valor de las propiedades. Entre tanto, la ubicación de CAI en el Norte de la ciudad corresponde a los lugares en los que se concentra un menor valor. A diferencia de Bogotá los puntos de industria no parecen castigar el precio de las propiedades, quizá esto se deba más a la naturaleza local.

Tabla 1. Estadísticas Descriptivas de distancias promedio a Amenities

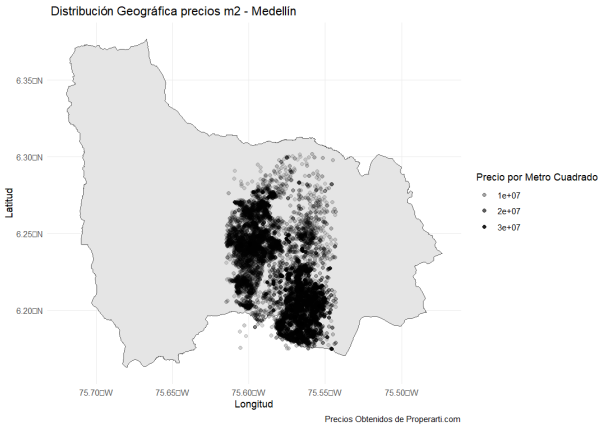
Variable	Observaciones	Promedio	Desv.Est.	Min	Max
Distancia a Universidad	51437	947.312	653.617	1.296	453.934
Distancia a Colegios	51437	391.748	264.272	0	192.532
Distancia a Kindergarten	51437	799.754	513.828	1.931	411.234
Distancia a Estaciones de Bus	51437	1074.935	865.432	4.42	445.469
Distancia a Via Principal	51437	426.534	414.28	0	138.032
Distancia a Oficinas	51437	1078.784	878.323	3.227	454.443
Distancia a Industria	51437	1467.699	758.491	5.452	880.584
Distancia a Comercio	51437	759.035	556.434	0.124	372.318
Distancia a Cai	51437	776.441	500.516	1.529	419.766

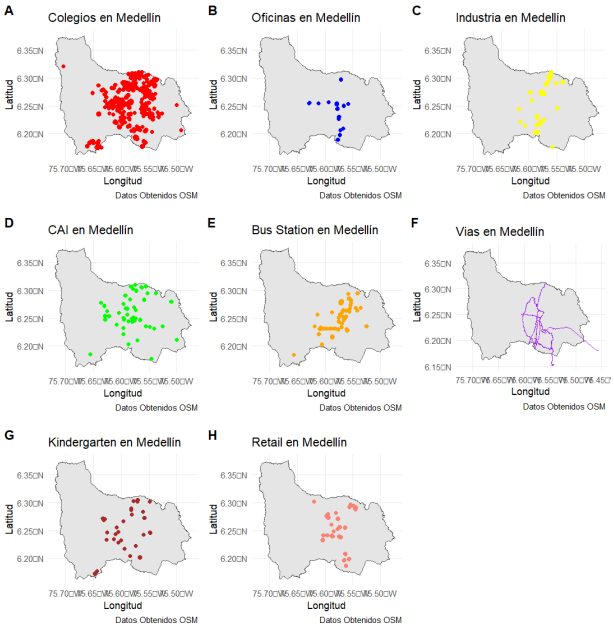
Figura 1. Ubicación geográfica de amenities para Bogotá y Medellín y distribución de viviendas de acuerdo al precio

A. Bogotá



B. Medellín





Modelo y Resultados

Dado que el objetivo principal es construir un modelo de regresión que prediga correctamente el precio de la vivienda en Cali, se realizó un ejercicio de predicción en tres fases. En primer lugar, con los datos de entrenamiento correspondientes a las ciudades de Bogotá y Medellín, se entrenaron 6 modelos que exploran relaciones lineales y no lineales entre las variables de interés y el precio de la vivienda. Previamente, dicha muestra de entrenamiento se dividió en dos, donde el 70% de la muestra se utilizó para el entrenamiento de los modelos y el 30% para su validación. Entre los modelos seleccionados se exploraron metodologías como: regresión lineal con Mínimos Cuadrados Ordinarios, selección óptima de submuestra con Backward y Forward Selection, regularización bajo el modelo de Elastic Net y modelos de predicción como Random Forest y XG-boost. Para el caso de los modelos de Elastic net, Random Forest y XG-boost, la selección de los hiperparámetros se realizó a través de validación cruzada con un k-fold de 10. Los resultados obtenidos dentro de muestra evidencian que el modelo XG-boost es el que mejor predice, con un RMSE de 0.348. para siete (7) hiperparámetros calculados por validación cruzada para maximizar su capacidad predictiva (tabla 1).

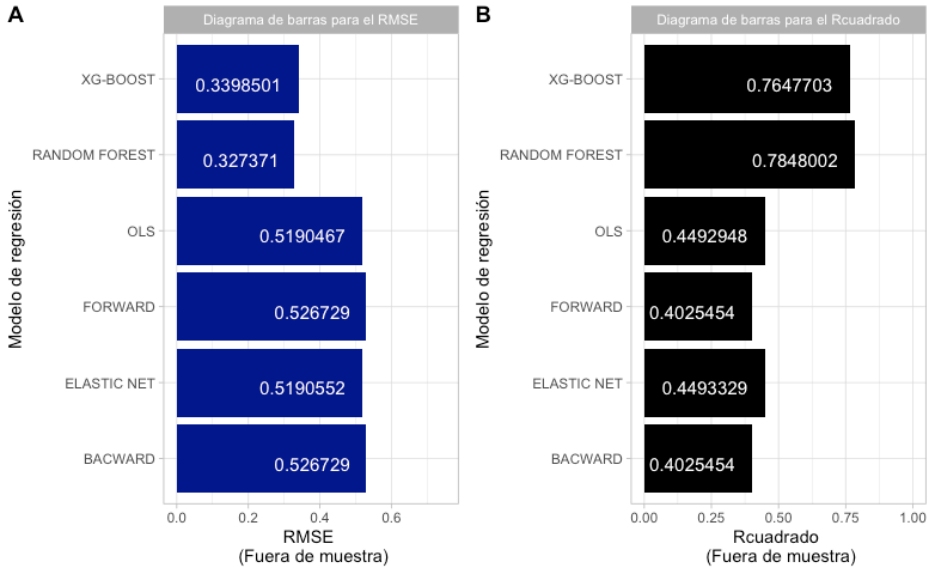
En segundo lugar, una vez finalizada la etapa de entrenamiento se estimó cada modelo en la muestra de validación, con el propósito de elegir el que mejor prediga fuera de muestra, basándose en el error cuadrático medio (RMSE) como métrica de evaluación, dado que en este caso se enfrenta un problema de regresión. En la tabla 2 se observa que el modelo ganador fue el correspondiente al Random Forest el cual obtuvo un RMSE de 0.327 y un R-cuadrado de 78,5% (figura 2). Los hiperparámetros óptimos de este modelo fueron obtenidos a través de validación cruzada y corresponden a: 3 variables utilizadas en cada resampleo (mtry =3), 10 observaciones mínimo en cada nodo terminal (Min node size =10) y una regla de separación basada en la máxima varianza. (anexo 1)

Tabla 2. Comparación modelos estimados para problema de regresión

		Errores predictivos			***Parámetros***
Modelo	***Muestra***	***RMSE***	***Rsquared***	***MAE***	
OLS	*Dentro muestra*	*0.514107*	*0.4458714*	*0.3935926*	
	Fuera muestra	*0.5190467*	*0.4492948*	*0.3972433*	
Backward	*Dentro muestra*	*0.5329517*	*0.4044860*	*0.4070588*	

	Fuera muestra	*0.526729*	*0.4025454*	*0.4327536*							
Forward	*Dentro muestra*	*0.5329273*	*0.4043518*	*0.4070229*							
	Fuera muestra	*0.526729*	*0.4025454*	*0.4327536*							
Elastic Net	*Dentro muestra*	*0.5148004*	*0.4453253*	*0.3946089*	***lambda***	***alpha***					
	Fuera muestra	*0.5190552*	*0.4493329*	*0.3972199*	*0.0001*	*1*					
Random Forest	*Dentro de muestra*	*0.363424*	*0.728282*	*0.2631535*	***mtry***	***Split rule***			***Min node size***		
	Fuera de muestra	*0.327371*	*0.7848002*	*0.224281*	*3*	*varianza*			*10*		
XG-Boost	*Dentro de muestra*	*0.347947*	*0.7478363*	*0.2450082*	***Max depth***	***eta***	***subsample***	***Min child wight***	***gamma***	***rondas***	***Col sample by tree***
	Fuera de muestra	*0.3398501*	*0.7647703*	*0.2392797*	*8*	*0.3*	*0.6*	*50*	*0*	*500*	*0.7*

Figura 2. RMSE y Rsquared fuera de muestra para modelos seleccionados



Básicamente, este modelo ganador utiliza la técnica de bagging, en donde se hacen múltiples copias de la base de entrenamiento utilizando bootstrap. Allí se ajusta un árbol de decisión independiente a cada copia y al final se combinan todos los árboles para crear un único modelo predictivo.

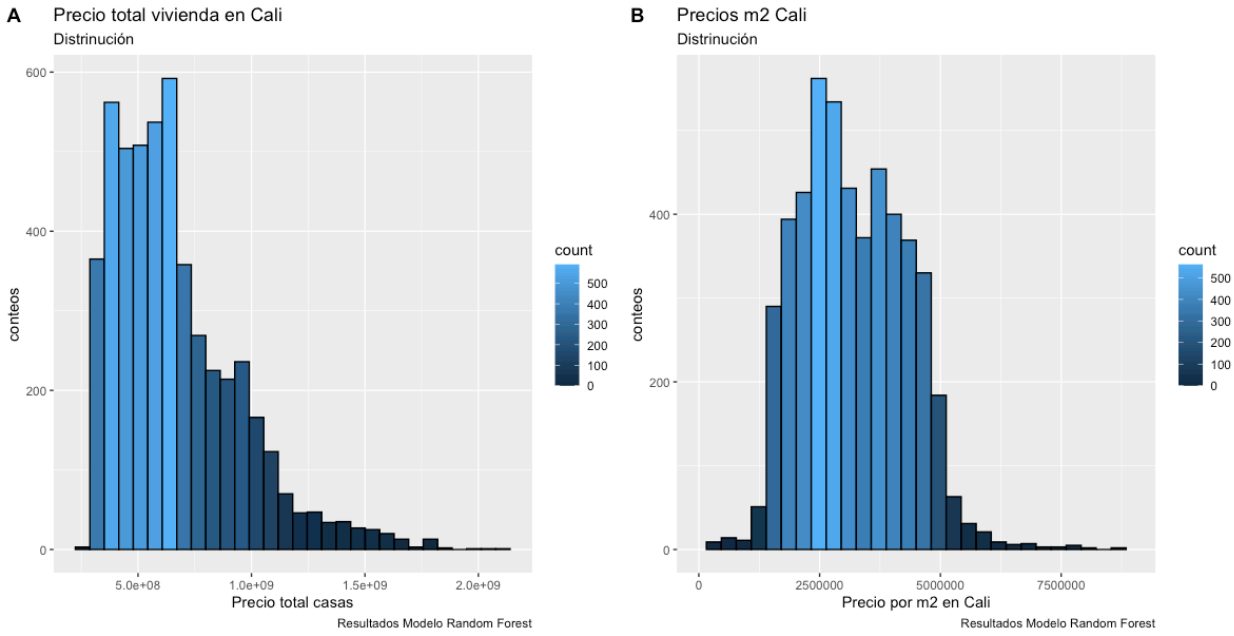
$$\hat{f}_{bag} = \frac{1}{B} \; b \sum_{b=1}^{B} \hat{f}^{b}(x)$$

Quiere decir que el modelo suaviza las predicciones y da mejores métricas de desempeño debido a que garantiza que ninguna de las variables sea dominante en el ajuste del modelo, disminuyendo el *overfitting* a los datos de entrenamiento y fomentando una menor varianza en comparación. No obstante, cabe resaltar que este modelo hereda el problema asociado con el bagging, donde si hay un predictor fuerte, diferentes arboles serán muy similares entre sí, por lo que habría alta correlación entre estos y por ende no se garantizaría la reducción de la varianza. (anexo 3). Así mismo, si existe correlación entre los atributos, es posible que estos pierdan relevancia, no obstante, en este caso no se observan correlaciones altas (anexo 5)

Finalmente, en la tercera fase se realizó la predicción de los precios de la vivienda en Cali (figura 3) donde se evidencia que, el precio promedio por metro cuadrado en la ciudad es de $3’229.684, con una desviación estándar de 1’488.997 y unos niveles mínimo y máximo de $2’522.303 y $45.857.671 respectivamente. Adicionalmente, el valor comercial total promedio de la vivienda en Cali es de $665.592.640 con una desviación de $279.256.012. Esto demuestra que, como es sabido el avalúo de

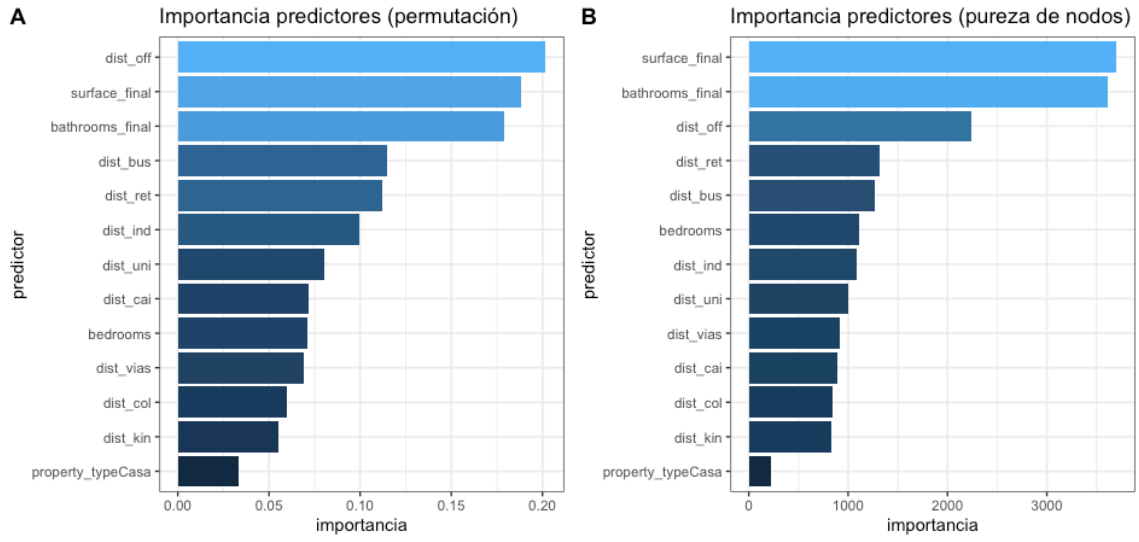
la tierra en esta ciudad es menor en comparación con los valores observados en ciudades como Medellín y Bogotá. (Figura 3)

Figura 3. Distribución precio total y por metro cuadrado predicho de la vivienda en Cali



Ahora bien, en cuanto a los predictores usados en este caso, se usaron las 13 variables descritas previamente y se entrenó y predijo el modelo para el logaritmo del salario (anexo 4). La distribución de la importancia de las variables de acuerdo con el modelo se presenta en la figura 4, en esta se ve que el área de la superficie total, el número de baños, la distancia a la oficina, la distancia a las estaciones de buses, la distancia a centros comerciales y la distancia a zonas industriales, tienen una importancia superior a 10% en la predicción del precio de la vivienda, ya sea bajo una metodología de pureza de nodos o de permutación. Para el caso de la superficie total, número de baños y distancia a las oficinas (CBD) la importancia es superior 17%.

Figura 4. Importancia de variables en la predicción modelo Random Forest

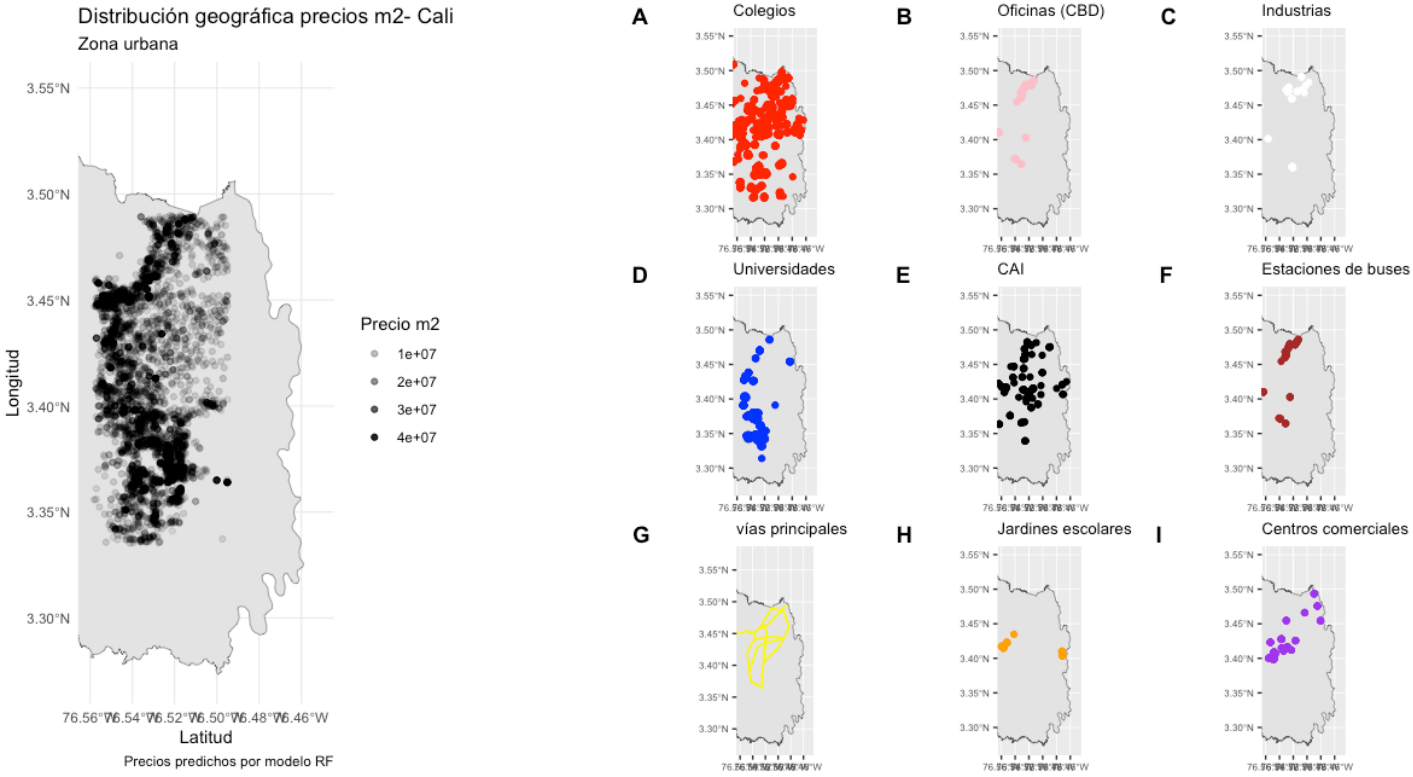


Así mismo, en cuanto al comportamiento del precio por metro cuadrado y los amenities seleccionados, se observa que los precios más elevados corresponden a las casas ubicadas en la zona centro-sur de la ciudad y en una pequeña parte del noroeste. Particularmente, este comportamiento se asocia a una ubicación en zonas con menores índices de inseguridad (menos cantidad de CAI en la zona), cercanas a universidades y zonas de trabajo (CBD), pero lejanas a zonas industriales y algunos centros comerciales. (Figura 5). Con respecto a la distancia en colegios en general no se encuentra un patrón claro, ni tampoco con respecto a la distancia a vías principales, ni jardines de niños. Esto último puede deberse a que la participación de estas variables ya esté capturada por otras variables, como, por ejemplo, la participación de jardines de niños puede estar acotada por colegios, o la de vías principales por la de estaciones de buses. En el anexo 6 se observa el comportamiento del logaritmo del precio de las viviendas con respecto a las distancias como variables y se observan los patrones previamente descritos.

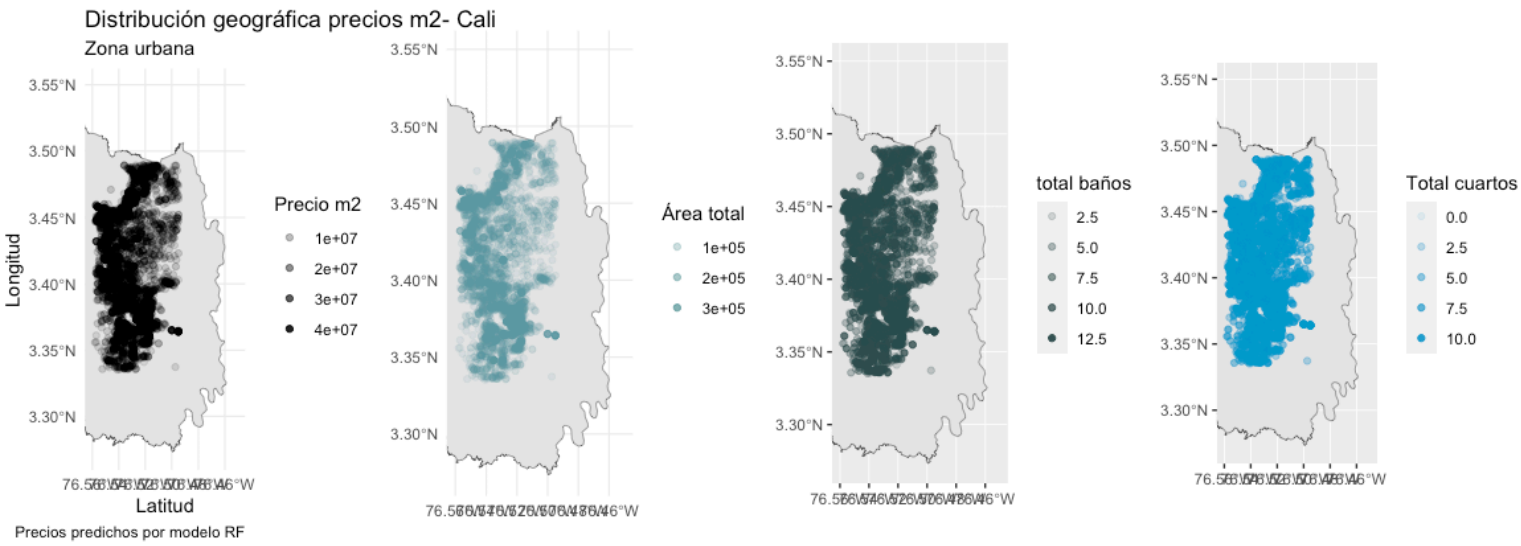
Adicionalmente, se ve que existe una mayor varianza en el precio a medida que estos aumentan, es decir en la cola derecha de la distribución. Esto puede resultar en que la participación de las variables aumente o disminuya para las viviendas más costosas. Esto tiene sentido, pues en casas más caras, es donde características como, por ejemplo, estar cerca de una estación de bus o estar cerca de un centro comercial pierde participación en la predicción del precio, debido a que por ejemplo, en estas zonas viven personas con carro propio o que tienen hábitos de consumo que no demandan tener acceso a centros de comercio frecuentemente, o como se ha visto en los últimos años, los puestos de teletrabajo tienden a estar en cargos altos, lo cual hace innecesario el hecho de vivir cerca del CBD u oficinas para personas que pueden acceder a viviendas de mayor valor. Por lo anterior, es que un modelo que explote estas no linealidades será el óptimo para predecir el precio.

Figura 5. Ubicación geográfica de amenities para Cali y distribución de viviendas de acuerdo al precio

A. Amenities



B. Atributos de la vivienda



Conclusiones y Recomendaciones

La estrategia de predicción a través de un modelo de Random Forest para predecir el precio de la vivienda en la ciudad de Cali es el mejor instrumento para predecirlo correctamente. Particularmente, este modelo permitió identificar la importancia de algunos atributos de la vivienda y de la zona donde está ubicada en el incremento del error del modelo, permitiendo así su fácil interpretación. Particularmente en este caso, se observó que la superficie total de la vivienda, el número de baños, la distancia al CBD o zona de trabajo, la distancia a estación de buses, a centros comerciales, universidades y a zonas industriales son aspectos relevantes en la determinación del precio de la vivienda. Más particularmente, vivir cerca de universidades aumenta su precio, así como vivir en zonas industriales lo disminuye. Paralelamente, se observa que variables como la distancia a vías principales, a jardines de niños y colegios en general tienen una participación menor en dicha predicción en la ciudad de Cali.

Más particularmente, el sur de Cali tiene una buena representación en la captación de plusvalía con las zonas de Bochalema y Ciudad Pacífica que quedan cerca de la Universidad Autónoma. Esta zona en los últimos años se ha desarrollado inmobiliariamente y corresponde a zonas de expansión de la ciudad según su POT. El precio por metro cuadrado más alto es el de Bochalema, le sigue la zona de Alfaguara y luego Ciudad Pacífica, esta última en las zonas de expansión de la ciudad.

Vale la pena abordar en futuras investigaciones el impacto de estas última variables desagregándolas por sus heterogeneidades, por ejemplo, desagregar la variable colegios por tipo de colegio de acuerdo a su nivel de calidad o una proxi más certera respecto a la seguridad a nivel de barrio. Esto porque, como se ha evidenciado en la literatura, el precio de la vivienda se ve afectado de forma diferencial dependiendo de los mecanismos que operan tras estos elementos.

Bibliografia

Alzate, y. V. (2019). *Modelo de predicción de precios de viviendas en el municipiode rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz* . Medellín: universidad pontificia bolivariana.

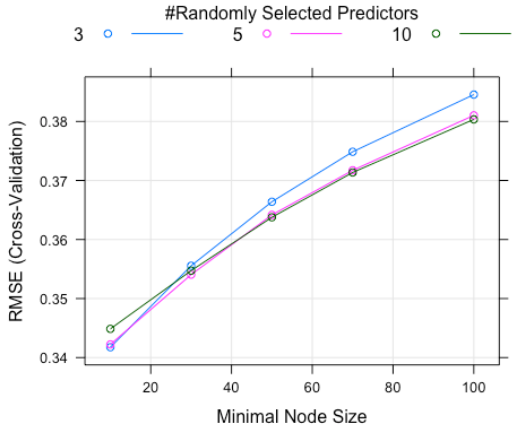
Caracas, S. B. (2021). *Modelación del precio de la oferta de vivienda en venta de la ciudad de Cali,considerando variables propias del activo y covariables de su entorno.* Santiago de Cali: Universidad del Valle.

Martínez, M. A. (2018). *Determinantes y distribución del precio de la vivienda nueva encali a través de un modelo gwr.* Cali: universidad del valle.

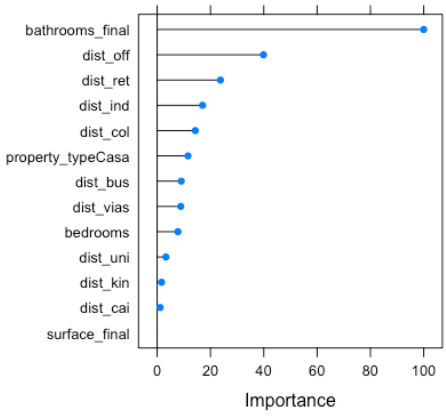
Peña, J. M. (2005). *Efectos de la seguridad ciudadana en el precio de las viviendas: un analisis de precios hedonicos.* Santiago: universidad de chile.

Anexos:

1. **Validación cruzada de modelo Random Forest y elección de hiperparámetros**

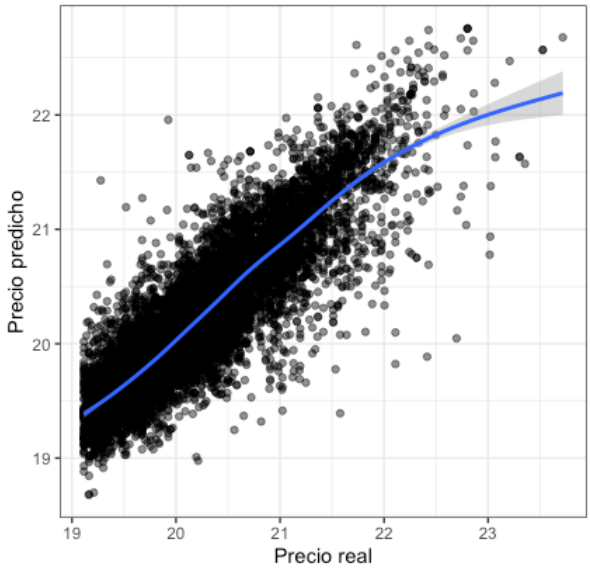


2. **Importancia de las variables modelo Elastic Net**

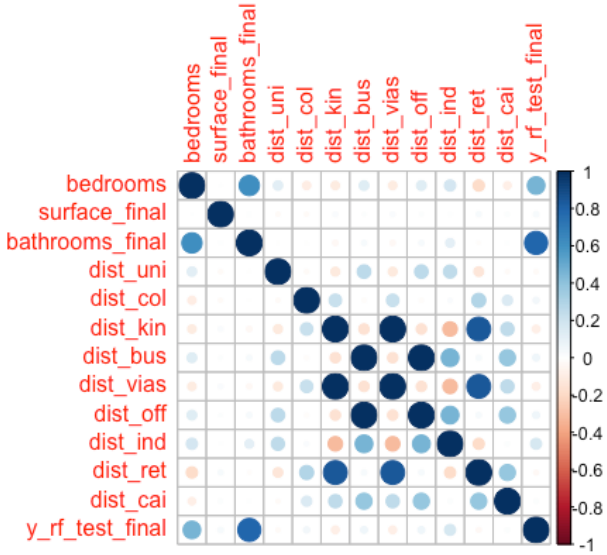


3. **Precio predicho frente a precio real (error de predicción)**

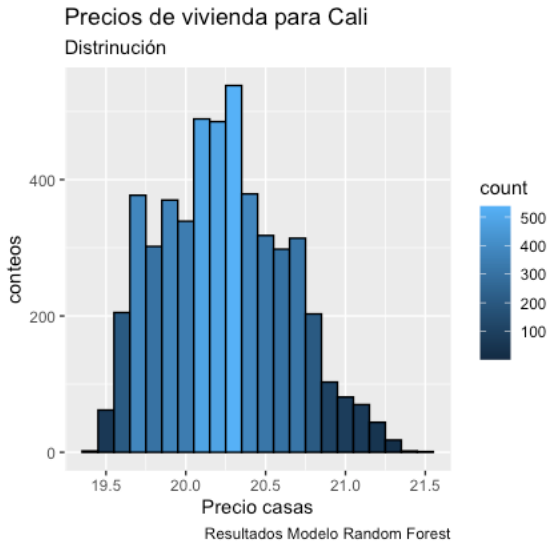
.



4. Distribución del logaritmo del precio comercial total estimado para la vivienda en Cali



5. Correlación entre variables del modelo RF de predicción



6. Relación de los amenities con el logaritmo del precio de la vivienda