

**Universidad Internacional de La Rioja**

**Escuela Superior de Ingeniería y  
Tecnología**

**Máster Universitario en Análisis y Visualización  
de Datos Masivos**

# Predicción de los precios de vivienda en la ciudad de Medellín y el Área Metropolitana

**Trabajo Fin de Máster**

**Tipo de trabajo:** Desarrollo de software

**Presentado por:** Medina Giraldo, Angie Tatiana

**Director/a:** Granada Mejía, Juan David

## Resumen

La estimación precisa del precio de un inmueble es una actividad que tradicionalmente ha sido compleja, debido al alto número de variables que intervienen en este valor. Históricamente se ha empleado un método simple basado en la comparación con propiedades similares, localizadas aproximadamente en la misma área de mercado. Sin embargo, los avances tecnológicos han originado nuevas técnicas para resolver este tipo de problemática, por ejemplo, la aplicación de métodos de aprendizaje automático que consideran mayor número de variables para una estimación más precisa del valor de venta de los inmuebles.

En este documento se presenta el desarrollo e implementación de algoritmos de Machine Learning para la predicción del precio de venta de viviendas en la ciudad de Medellín y el Área Metropolitana. Los datos utilizados corresponden a anuncios de ventas de inmuebles publicadas en el portal inmobiliario Properati, a partir de los cuales se realizó el análisis de los datos y se aplicaron diferentes técnicas de aprendizaje automático, entre ellos: regresión lineal, árboles de decisión, Random Forest, k vecinos más cercanos y Extreme Gradient Boosting (XGBoost) para la regresión. Como medidas de desempeño para la evaluación de los modelos, se utilizaron el error porcentual absoluto medio (MAPE), el error medio absoluto (MAE) y el coeficiente de determinación  $R^2$  y así determinar el modelo con menor margen de error y mayor precisión en la estimación. Se obtuvieron mejores resultados con el modelo Random Forest, con un  $R^2$  de 0,81 y un MAPE de 14,3%. Los resultados de este documento pueden ser utilizados por todas las personas con el interés de vender y comprar vivienda en la ciudad, y también por los portales inmobiliarios o autoridades fiscales para una mejor estimación del valor de venta de los inmuebles.

**Palabras Clave:** Inteligencia artificial, Machine learning, Vivienda, Regresión, Predicción

## Abstract

Accurate estimation of the properties prices is an activity that has traditionally been complex due to the high number of variables involved in this value. Historically, a simple method has been employed, based on comparison with equivalent properties located in approximately the same market area. However, technological advances have given rise to new techniques to solve this type of problem, for example, the application of machine learning methods that consider a greater number of variables, for a better estimate of the sale value of real estate.

This document presents the development and implementation of Machine Learning algorithms for prediction of the sale price of homes in the city of Medellín and the Área Metropolitana. The data used corresponds to real estate sales announcements published on the Properati real estate portal, The data was analyzed, and different machine learning techniques were applied: linear regression, Decision Trees, Random Forest, k Nearest Neighbors and Extreme Gradient Boosting (XGBoost) for the regression. As performance measures for the evaluation of the models, Mean Absolute Percentage Error (MAPE), the Absolute Mean Error (MAE) and the coefficient of determination  $R^2$  were used to determine the model with the lowest margin of error and the highest precision in the estimate. The best results were obtained with the Random Forest model, with an  $R^2$  of 0,81 and MAPE of 14,3%. The results of this document can be used by all people with an interest in selling and buying housing in the city and by real estate portals or tax authorities for a better estimate of the sale value of the properties.

**Keywords:** Artificial intelligence, Machine learning, Housing, Regression, Predictive model

# Índice de contenidos

1.	Introducción .....	8
1.1	Justificación .....	9
1.2	Planteamiento del trabajo.....	10
1.3	Estructura de la memoria .....	11
2.	Contexto y estado del arte .....	12
2.1.	Mercado inmobiliario .....	12
2.1.1.	Mercado Inmobiliario en la ciudad de Medellín.....	12
2.2.	Avalúo inmobiliario .....	16
2.3.	Machine Learning.....	18
2.3.1.	Técnicas de Machine Learning para la predicción de precios de vivienda.....	19
2.3.2.	Métricas para evaluación de los modelos.....	22
2.4.	Estudios previos.....	24
2.4.1.	Estudios previos alrededor del mundo .....	24
2.4.2.	Estudios previos en Colombia .....	28
2.5.	Conclusiones de los estudios previos.....	29
3.	Objetivos concretos y metodología de trabajo.....	30
3.1.	Objetivo general .....	30
3.2.	Objetivos específicos .....	30
3.3.	Metodología del trabajo.....	31
4.	Desarrollo específico de la contribución .....	34
4.1.	Extracción y obtención de los datos .....	34
4.1.1.	Fuente de los datos.....	34
4.1.2.	Protección de los datos .....	35
4.1.3.	Herramientas .....	35
4.1.4.	Descripción del conjunto de datos.....	36
4.2.	Limpieza y preparación de los datos .....	37
4.3.	Análisis exploratorio de los datos .....	42
4.3.1.	Precio de Venta .....	42
4.3.2.	Localización .....	45
4.3.3.	Características .....	49
4.3.4.	Matriz de correlaciones .....	52
4.4.	Transformaciones y selección de variables.....	54

4.5.	Desarrollo de los Modelos.....	55
4.5.1.	Selección de los algoritmos.....	55
4.5.2.	Implementación de los algoritmos.....	56
4.5.3.	Configuración de hiperparámetros.....	56
4.6.	Evaluación de los modelos.....	61
4.7.	Predicciones.....	62
4.8.	Análisis de los resultados.....	64
5.	Conclusiones y trabajo futuro.....	65
5.1.	Conclusiones.....	65
5.2.	Líneas de trabajo.....	66
6.	Bibliografía.....	67
Anexos.....		71
Anexo I.	Notebook.....	71
Anexo II.	Resultados de los modelos para diferentes conjuntos de datos.....	72

## Índice de tablas

Tabla 1. Distribución de unidades de vivienda en el departamento de Antioquia.....	15
Tabla 2. Tipos de avalúo Inmobiliario .....	16
Tabla 3. Resumen de estudios relacionados en el mundo.....	27
Tabla 4. Resumen de estudios relacionados en Colombia .....	29
Tabla 5. Descripción de las etapas de la metodología CRISP-DM .....	31
Tabla 6. Objetivos de las fases de la metodología de trabajo.....	33
Tabla 7. Fuente de los datos .....	34
Tabla 8. Variables del conjunto de datos original .....	36
Tabla 9. Variables del conjunto de datos por grupos .....	40
Tabla 10. Variables del conjunto de datos modificado.....	41
Tabla 11. Estadísticas descriptivas de las variables precio y precio por m <sup>2</sup> .....	43
Tabla 12. Métricas de evaluación para los algoritmos .....	61
Tabla 13. Métricas de evaluación para los algoritmos, unidades dólares americanos USD ..	62

# Índice de figuras

Figura 1. Mapa del Área Metropolitana del Valle de Aburrá (Área Metropolitana del Valle de Aburrá, 2020) ..... 13

Figura 2. Evolución de la oferta de viviendas en el Valle de Aburrá (Medellín cómo vamos, 2020)..... 15

Figura 3. Distribución de la oferta de viviendas VIS y NO VIS en el Valle de Aburrá (Medellín cómo vamos, 2020)..... 15

Figura 4. Métodos de avalúo inmobiliario (Elaboración propia a partir de Instituto Geográfico Agustín Codazzi, 2018) ..... 17

Figura 5. Metodología CRISP-DM (Rodríguez, 2010)..... 31

Figura 6. Diagrama de procesos de la metodología de trabajo (Elaboración propia) ..... 32

Figura 7. Valores nulos en el conjunto de datos (Elaboración propia) ..... 38

Figura 8. Distribución de tipo de propiedad en el conjunto de datos original (Elaboración propia)..... 40

Figura 9. Distribución de la variable precio (Elaboración propia) ..... 43

Figura 10. Distribución de la variable precio m<sup>2</sup> (Elaboración propia) ..... 43

Figura 11. Evolución de los precios de venta (Elaboración propia)..... 44

Figura 12. Evolución de los precios por m<sup>2</sup> (Elaboración propia)..... 45

Figura 13. Distribución de los datos según municipio (Elaboración propia) ..... 46

Figura 14. Mapa de Calor con Distribución de Precios (Elaboración propia) ..... 46

Figura 15. Boxplot distribución de precios por municipio (Elaboración propia) ..... 47

Figura 16. Boxplot distribución de precios por municipio según tipo de propiedad (Elaboración propia)..... 47

Figura 17. Boxplot distribución de precios m<sup>2</sup> por municipio (Elaboración propia) ..... 48

Figura 18. Boxplot distribución de precios m<sup>2</sup> por municipio según tipo de propiedad (Elaboración propia) ..... 48

Figura 19. Distribución de la variable número de habitaciones (Elaboración propia) ..... 49

Figura 20. Distribución de la variable número de baños (Elaboración propia) ..... 49

Figura 21. Distribución de la variable superficie total (Elaboración propia) ..... 50

Figura 22. Distribución de la variable tipo de propiedad (Elaboración propia) ..... 50

Figura 23. Boxplot de precio y número de Habitaciones (Elaboración propia) ..... 51

Figura 24. Boxplot de precio y número de baños (Elaboración propia)..... 51

Figura 25. Boxplot de precio y Tipo de propiedad (Elaboración propia)..... 52

Figura 26. Boxplot de precio por m<sup>2</sup> y Tipo de propiedad (Elaboración propia)..... 52

Figura 27. Matriz de coeficiente de Correlación entre las variables (Elaboración propia) ..... 52

Figura 28. Relación entre las variables (Elaboración propia) .....53

Figura 29. Métricas de error RMSE y max\_depth para Árbol de decisión (Elaboración propia) .....57

Figura 30. Métricas de error  $R^2$  y max\_depth para Árbol de decisión (Elaboración propia) ..57

Figura 31. Métricas de error RMSE y n\_estimators para Random Forest (Elaboración propia) .....58

Figura 32. Métricas de error  $R^2$  n\_estimators para Random Forest (Elaboración propia) .....58

Figura 33. Métricas de error RMSE y max\_features para Random Forest (Elaboración propia) .....58

Figura 34. Métricas de error  $R^2$  y max\_features para Random Forest (Elaboración propia) .58

Figura 35. Métricas de error RMSE y k\_neighbors para k vecinos más cercanos (Elaboración propia) .....59

Figura 36. Métricas de error para y k\_neighbors para k vecinos más cercanos (Elaboración propia) .....59

Figura 37. Métricas de error RMSE y n\_estimators para XGBoost (Elaboración propia) .....60

Figura 38. Métricas de error  $R^2$  y n\_estimators para XGBoost (Elaboración propia) .....60

Figura 39. Métricas de error RMSE y eta para XGBoost (Elaboración propia) .....60

Figura 40. Métricas de error  $R^2$  y eta para XGBoost (Elaboración propia) .....60

Figura 41. Predicción del precio en relación con el precio real del modelo Regresión Lineal (Elaboración propia) .....62

Figura 42. Predicción del precio en relación con el precio real del modelo Árboles de decisión (Elaboración propia) .....62

Figura 43. Predicción del precio en relación con el precio real de del modelo Random Forest (Elaboración propia) .....63

Figura 44. Predicción del precio en relación con el precio real del modelo K Vecino más cercano (Elaboración propia) .....63

Figura 45. Predicción del precio en relación con el precio real de del modelo XGBoost regresor (Elaboración propia) .....63



# 1. Introducción

El sector inmobiliario es uno de los más importantes de la economía de Colombia, con una contribución significativa al PIB (Producto Interno Bruto) del país. Según el Departamento Nacional de Estadística (DANE), el sector inmobiliario representó aproximadamente el 8,7% del PIB Nacional, en el año 2022 (DANE, Producto Interno Bruto -PIB- nacional trimestral, 2022). Adicionalmente, el sector inmobiliario es de gran importancia debido a su encadenamiento con los demás sectores de la economía nacional.

De acuerdo con las proyecciones del último censo nacional realizado por el DANE, en la ciudad de Medellín y el Área Metropolitana habitan aproximadamente 4,1 millones de personas, las cuales viven en 1,3 millones de viviendas, generando cerca de 5 mil transacciones de compra y venta de inmuebles al año (DANE, Censo Nacional de Población y Vivienda, 2018). La vivienda es considerada una necesidad básica para los seres humanos, por esta razón, el estudio del comportamiento del mercado inmobiliario residencial y la valoración de los precios de venta de los inmuebles tiene un alto impacto en la vida y las finanzas de las personas, siendo una actividad de interés para individuos y hogares con deseo de comprar o vender una vivienda, y también para instituciones financieras, empresas inmobiliarias, bancos, inversores o autoridades fiscales, entre otros.

El precio de los inmuebles está afectado por un gran número de variables, lo que hace de esta actividad un problema complejo para los interesados. Tradicionalmente, para la estimación de los precios de vivienda se han empleado métodos como la comparación del valor con otros inmuebles o la estimación de los costos de construcción, sin embargo, estos precios también se ven afectados por otros parámetros como la oferta, la demanda y los procesos inflacionarios, lo cual dificulta conocer de forma inmediata y con precisión el valor del precio de las viviendas (Martínez Sanchez & Téllez Buitrago, 2021).

Actualmente las organizaciones generan millones de datos por segundo, y los avances tecnológicos han originado nuevas técnicas para llevar a cabo el procesamiento y análisis de esta información, tales como los métodos de aprendizaje automático. El uso de estos métodos representa una ventaja competitiva para diferentes sectores de la economía, entre ellos el sector inmobiliario, permitiendo realizar la toma de decisiones a partir de los datos y predecir los precios de venta de acuerdo con diferentes características de los inmuebles. Por esta razón, en el presente trabajo se presenta una solución tecnológica para esta problemática del sector inmobiliario, a través del desarrollo de un modelo predictivo para estimar los precios de venta de vivienda en la ciudad de Medellín y el Área Metropolitana del Valle de Aburrá.

En este documento se presenta inicialmente la introducción, en la cual se establece el planteamiento del problema, la justificación y los objetivos del proyecto. A continuación, se expone el contexto y estado del arte, en el cual se presentan los conceptos básicos y los métodos de aprendizaje automático para la estimación del precio de venta de las viviendas. Por último, se presenta el proceso detallado de la contribución, desde la obtención de los datos y su tratamiento, hasta el desarrollo e implementación de un modelo para la predicción de los precios de viviendas en la ciudad de Medellín y el Área Metropolitana.

## 1.1 Justificación

Estimar con precisión el valor de venta de un inmueble, es un problema complejo al que se enfrentan constructores, vendedores, agentes inmobiliarios, compradores, inversionistas, entidades de créditos, entre otros. Su dificultad radica en el amplio número de factores que afectan el valor de un inmueble. Atributos como el tamaño, el número de habitaciones o la localización, son algunas de las características que afectan el precio de las viviendas. Adicionalmente, este valor es altamente sensible a los cambios de precios del mercado, impidiendo a los interesados conocer con certeza el valor real de un inmueble.

Disponer de una herramienta para la estimación precisa y objetiva del valor del precio de una vivienda, es de gran importancia en el sector inmobiliario. Un análisis preciso de estos precios permite establecer cuando un proyecto de vivienda presenta un precio razonable en función de sus características, o cuando un precio es atribuible a la especulación o una alta demanda en la ciudad.

En la actualidad existen diferentes técnicas para determinar los precios de una vivienda, estos valores son estimados con base en técnicas tradicionales, como la comparación con las viviendas localizadas en el mismo sector, la estimación a partir de los costos de construcción, proyecciones de costos, modelos estadísticos u otros estimadores lineales; en este último se estima el precio por cada metro cuadrado para un tipo de proyecto específico, y se multiplica por la cantidad de metros cuadrados de la vivienda.

La gran cantidad de variables involucradas en el proceso para determinar los precios de las viviendas, sumado a la falta de disponibilidad de datos precisos e históricos de las ofertas inmobiliarias, han favorecido al uso de métodos tradicionales para esta actividad, con una lenta adopción de las nuevas tecnologías por parte del sector inmobiliario.

Debido a la importancia del sector inmobiliario en la economía Colombiana, y la creciente demanda en unidades de vivienda en la ciudad de Medellín, es relevante el uso de herramientas tecnológicas para resolver la problemática de predicción de los precios de la

vivienda. El uso de estas herramientas permitirá contar con estimaciones más precisas de los precios de los inmuebles, además, brindará mayor comprensión de la dinámica del mercado inmobiliario en la ciudad, soportando la toma de decisiones por parte de los involucrados en los procesos de compra y venta de los inmuebles.

## 1.2 Planteamiento del trabajo

Los avances tecnológicos en el almacenamiento y procesado de grandes volúmenes de datos, permiten a las diferentes industrias beneficiarse y usar el análisis de los datos en la toma de decisiones. El uso de herramientas tecnológicas como los métodos de aprendizaje automático, permite utilizar los datos de ofertas de inmuebles e información histórica para predecir el valor de una vivienda a partir de sus diferentes atributos.

En el presente trabajo se presenta el desarrollo de un modelo predictivo para estimar los precios de venta de vivienda en la ciudad de Medellín y el Área Metropolitana, a partir de la implementación de métodos de aprendizaje automático. Para el desarrollo del modelo se propone el uso de los datos de ofertas inmobiliarias en diferentes ubicaciones alrededor de la zona de estudio, y la aplicación de diferentes técnicas de aprendizaje automático para la predicción del valor de venta de los inmuebles. El resultado del presente documento busca apoyar a la toma de decisiones relacionadas con la compra y venta de inmuebles en la ciudad y alrededores.

## 1.3 Estructura de la memoria

A continuación se presenta la estructura de capítulos del presente documento:

**Capítulo 1: “Introducción”.** En este capítulo se realiza el planteamiento del trabajo de fin de máster, y una introducción al contenido, incluyendo la descripción del problema a tratar, su importancia y posibles causas. Adicionalmente, se presenta la propuesta para solucionar la problemática planteada.

**Capítulo 2: “Contexto y estado del arte”.** En este capítulo se expone el contexto y estado del arte relacionado con el mercado inmobiliario en la ciudad. Se presentan las metodologías y procedimientos utilizados para determinar el precio de venta de las viviendas, la revisión de los estudios previos, y el estado actual del uso de herramientas tecnológicas en la solución de la problemática.

**Capítulo 3: “Objetivos concretos y metodología de trabajo”.** En este capítulo se presentan los objetivos que se pretenden alcanzar con el trabajo propuesto y la metodología de trabajo a aplicar en el desarrollo de la contribución.

**Capítulo 4: “Desarrollo específico de la contribución”.** En este capítulo se expone el proceso detallado de cada una de las fases de la metodología. Se presentan las actividades orientadas a generar una modelo para la predicción de los precios de vivienda en la ciudad de Medellín y el Área Metropolitana, desde la descripción del conjunto de datos, el tratamiento, el análisis y los algoritmos empleados en la fase de modelado.

**Capítulo 5: “Conclusiones y trabajo futuro”.** En este capítulo se presentan las conclusiones y los resultados obtenidos a partir del desarrollo de las actividades planeadas, además, se expone el trabajo futuro y posibles usos de la aportación del presente trabajo.

**Bibliografía.** Se presentan las fuentes de las cuales se ha extraído la información relevante para este Trabajo de Fin de Máster.

## 2. Contexto y estado del arte

En este capítulo se presentan las definiciones y los conceptos básicos para el análisis y la resolución de la problemática expuesta. Además, se presenta el contexto sobre el mercado inmobiliario en la ciudad, y el conjunto de estudios y antecedentes a nivel mundial relacionados con el uso de métodos de aprendizaje automático para la predicción de precios de los inmuebles.

### 2.1. Mercado inmobiliario

El mercado inmobiliario está conformado por el conjunto de todas las propiedades disponibles para la venta en un determinado lugar, sumado a todas las personas dispuestas a adquirir estos inmuebles. Lo anterior, conforma la oferta y la demanda de bienes inmuebles, a partir de las cuales se generan las diferentes transacciones de compra y venta.

Los bienes inmuebles pueden ser de diferente naturaleza, entre ellos se destacan los inmuebles de uso residencial, comercial, industrial, las oficinas, los inmuebles de usos especiales, entre otros. De los anteriores, el tipo de inmuebles principal para el presente estudio, corresponde aquellos destinados a la vivienda, definiendo vivienda como el espacio independiente, cerrado y cubierto que ofrece refugio y protección a las personas (Real Academia Española, 2022).

#### 2.1.1. Mercado Inmobiliario en la ciudad de Medellín

Para identificar el comportamiento del mercado inmobiliario en la ciudad de Medellín, es necesario analizar los elementos que lo conforman: la localización, el producto inmobiliario, el precio, la oferta y la demanda.

##### **Localización**

La ciudad de Medellín es la capital del departamento de Antioquia, ubicada al noroccidente de Colombia, sobre la cordillera central de los Andes. La ciudad está localizada en el centro geográfico del Valle de Aburrá, con una altitud media de 1.495 metros sobre el nivel del mar y con 376,4 kilómetros cuadrados de superficie (Alcaldía de Medellín, 2022).

El Valle de Aburrá, es una subregión del departamento de Antioquia bajo la figura de Área Metropolitana, la cual está compuesta por la ciudad de Medellín y 9 municipios más, estos son: Caldas, Sabaneta, Itagüí, Envigado, Bello, Copacabana, Girardota, La Estrella y Barbosa, como se observa en la Figura 1. La subregión se extiende en una superficie aproximada de

1.158 km<sup>2</sup>, con una longitud aproximada de 75 kilómetros y una amplitud variable (Área Metropolitana del Valle de Aburrá, 2022).

Debido a su proximidad geográfica, la ciudad de Medellín y los demás municipios que conforman el Área Metropolitana del Valle de Aburrá, presentan una interacción constante a nivel económico, demográfico, social o ecológico. Por lo tanto, el mercado inmobiliario fluctúa entre los diferentes municipios que conforman la región, y las interacciones entre la oferta y la demanda se presentan entre los diferentes municipios.



Figura 1. Mapa del Área Metropolitana del Valle de Aburrá (Área Metropolitana del Valle de Aburrá, 2020)

### Producto inmobiliario

El producto inmobiliario de la ciudad de Medellín y el Área Metropolitana está compuesto principalmente por inmuebles destinados a la vivienda. Para el periodo comprendido entre octubre de 2021 y octubre de 2022, en la ciudad de Medellín se licenciaron un total de 1.119.096 m<sup>2</sup>, de los cuales el 76%, es decir 485.830 m<sup>2</sup> correspondió a proyectos de vivienda. Para el mismo periodo, en el Área Metropolitana esta cifra ascendió a un total de 4.858.030 m<sup>2</sup> licenciados, con un 82% destinado a vivienda, equivalentes a 3.960.678 m<sup>2</sup> (DANE, Licencias de construcción, 2022).

## **Precio de venta de las viviendas**

El precio de las viviendas varía a partir de las diferentes características del inmueble. Las viviendas son un bien altamente heterogéneo, y el precio de venta está asociado a características como: la localización, el tamaño, el número de habitaciones, la estructura, la antigüedad, entre otros factores.

La Constitución Política de Colombia (1991), consagra en el artículo 51 el derecho de todos los colombianos a una vivienda digna, y la responsabilidad del Estado en su cumplimiento. “El Estado fijará las condiciones necesarias para hacer efectivo este derecho y promoverá planes de vivienda de interés social, sistemas adecuados de financiación a largo plazo y formas asociativas de ejecución de estos programas de vivienda”.

Para garantizar el cumplimiento del derecho a la vivienda, el Estado, a través del Ministerio de vivienda define los valores máximos de venta para la Vivienda de Interés Social (VIS) y la Vivienda de Interés Prioritario (VIP), siendo estos dos tipos, aquellas viviendas que se desarrollan para garantizar el derecho a la vivienda de los hogares de menores ingresos, o que viven bajo condiciones de pobreza extrema, respectivamente.

El Decreto 1467 (2019) y la Ley 1955 (2019), establecen un límite superior para los inmuebles VIS de ciento cincuenta (150) salarios mínimos mensuales legales vigentes (SMMLV), en aglomeraciones urbanas cuya población supere un millón (1.000.000) de habitantes, como es el caso de las ciudades y municipios del presente documento; y clasifican en el segmento VIP los inmuebles con precio menor o igual a 90 SMMLV.

A partir de lo anterior, los precios del mercado inmobiliario en la ciudad de Medellín y el Área Metropolitana, oscilan aproximadamente entre 90 salarios mínimos legales mensuales vigentes, correspondiente a las viviendas tipo VIP, hasta viviendas con valores superiores a los mil millones de pesos.

## **Oferta en el mercado inmobiliario de la ciudad**

La oferta de vivienda comprende el número de inmuebles disponibles a la venta. De acuerdo con los datos del censo de viviendas de Coordinada Urbana, suministrado por la Cámara Colombiana de la Construcción (CAMACOL), para noviembre 2022 la oferta de viviendas en el departamento de Antioquia estaba conformada por 23.384 unidades (CAMACOL, 2022).

La Tabla 1 presenta la distribución de las viviendas ofertadas en el departamento de Antioquia según el rango de precios, para la fecha de noviembre de 2022. Se observa que el 38% de las unidades de vivienda corresponde a viviendas VIS (Incluyendo VIP) frente a 62% de viviendas NO VIS, o lo que es equivalente viviendas con precios superiores a los 150 SMMLV.

Adicionalmente, el 45% de las viviendas se encuentra en un rango de precio comprendido entre las viviendas VIS (150 SMMLV), hasta 500 SMMLV.

Tabla 1. Distribución de unidades de vivienda en el departamento de Antioquia

Distribución de viviendas en Antioquia según el rango de precios				Total de Viviendas VIS	Total de Viviendas NO VIS	Total Viviendas
VIP	VIS (sin VIP)	> VIS y hasta 500 SMMLV	Mayor a 500 SMMLV			
227	8.703	10.478	3.976	8.930	14.454	23.384

Fuente: CAMACOL, 2022

De acuerdo con el Informe de Calidad de vida del área metropolitana, en el año 2020 la oferta de viviendas para el Valle de Aburrá fue de 17.825 unidades (Medellín cómo vamos, 2020). En la Figura 2 se presenta la evolución de la oferta de viviendas para el Valle de Aburrá, entre los años 2014 a 2020.

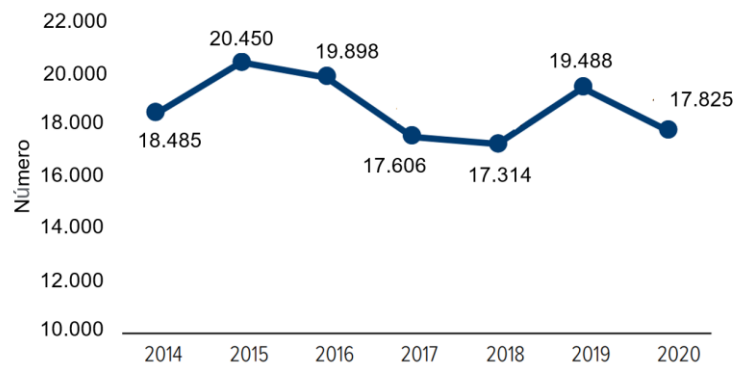


Figura 2. Evolución de la oferta de viviendas en el Valle de Aburrá (Medellín cómo vamos, 2020)

A continuación, en la Figura 3 se presenta la distribución de la oferta VIS y NO VIS de viviendas para el Valle de Aburrá, en el periodo comprendido entre los años 2014 a 2020. Para el año 2020, el 70% de la oferta correspondió a viviendas NO VIS, frente al 30% de Viviendas de Interés Social.

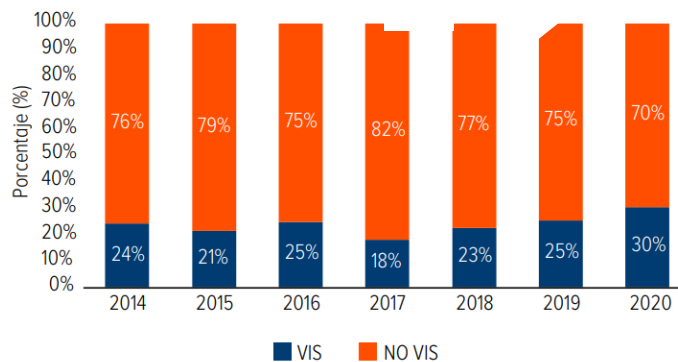


Figura 3. Distribución de la oferta de viviendas VIS y NO VIS en el Valle de Aburrá (Medellín cómo vamos, 2020)



## Demanda del mercado inmobiliaria en la ciudad

La demanda de vivienda se refiere a la población y hogares que potencialmente comprarán una vivienda en la ciudad y alrededores. De acuerdo con las cifras del censo nacional de población y vivienda del 2018, el Área Metropolitana del Valle de Aburrá contaba para este año con una población de 3.931.447 habitantes y 1.393.745 viviendas (DANE, Censo Nacional de Población y Vivienda, 2018).

El DANE además define el déficit cuantitativo de vivienda como: “los hogares que habitan en viviendas con deficiencias estructurales y de espacio”, es decir, los hogares para los que se necesitan viviendas adicionales en el Área Metropolitana (DANE, Déficit habitacional, 2020).

Según la Encuesta de Calidad de Vida del año 2019, el déficit cuantitativo de vivienda en el Área Metropolitana del Valle de Aburrá fue del 3,49%, esto es un total de 48.641 viviendas. Por su parte, la ciudad de Medellín al cierre del año 2019 contaba con 892.151 viviendas y un déficit cualitativo de vivienda de 3,94%, es decir 35.150 viviendas.

## 2.2. Avalúo inmobiliario

El avalúo inmobiliario es el proceso por el cual se determina el valor de un inmueble. La ley 1363 de 2013 define la valuación como: “la actividad, por medio de la cual se determina el valor de un bien, de conformidad con los métodos, técnicas, actuaciones, criterios y herramientas que se consideren necesarios y pertinentes para el dictamen” (Congreso de la República de Colombia, 2013). El dictamen de la valuación se denomina avalúo y, a partir del propósito del avalúo inmobiliario se pueden presentar las diferentes tipologías indicadas en la Tabla 2.

Tabla 2. Tipos de avalúo Inmobiliario

Tipo de avalúo	Descripción
<b>Avalúo comercial</b>	Realizado para determinar el valor de un inmueble a una determinada fecha, considerando la oferta y la demanda. Este avalúo parte de la premisa de que las condiciones que afectan el precio de un inmueble cambian constantemente, lo que puede generar un aumento o disminución en el valor comercial
<b>Avalúo catastral</b>	En el cual se calcula el valor de los predios mediante investigación y análisis estadístico del mercado inmobiliario. Generalmente, corresponde a un porcentaje entre el 0 y el 100 % del avalúo comercial
<b>Avalúo administrativo</b>	Son los avalúos ordenados por las entidades oficiales o administrativas. Pueden ser llevados a cabo por el Instituto Geográfico Agustín Codazzi o por cualquier persona natural o jurídica.

<b>Avalúo judicial</b>	Aquellos que se realizan durante un proceso judicial. Por lo general, se presentan cuando un juez ordena la venta de un inmueble para saldar una deuda, cuando este se presentó como garantía de pago
------------------------	---

Fuente: Elaboración propia a partir de ley 1673 del 2013

El Instituto Geográfico Agustín Codazzi (2018) en la Resolución 620, enuncia las técnicas utilizadas para determinar el valor comercial de los inmuebles, las cuales se denominan métodos valuatorios. Se presentan entonces cuatro diferentes métodos: método de comparación o de mercado, método de capitalización de rentas o ingresos, método de costo de reposición y método residual, como se observa en la Figura 4.

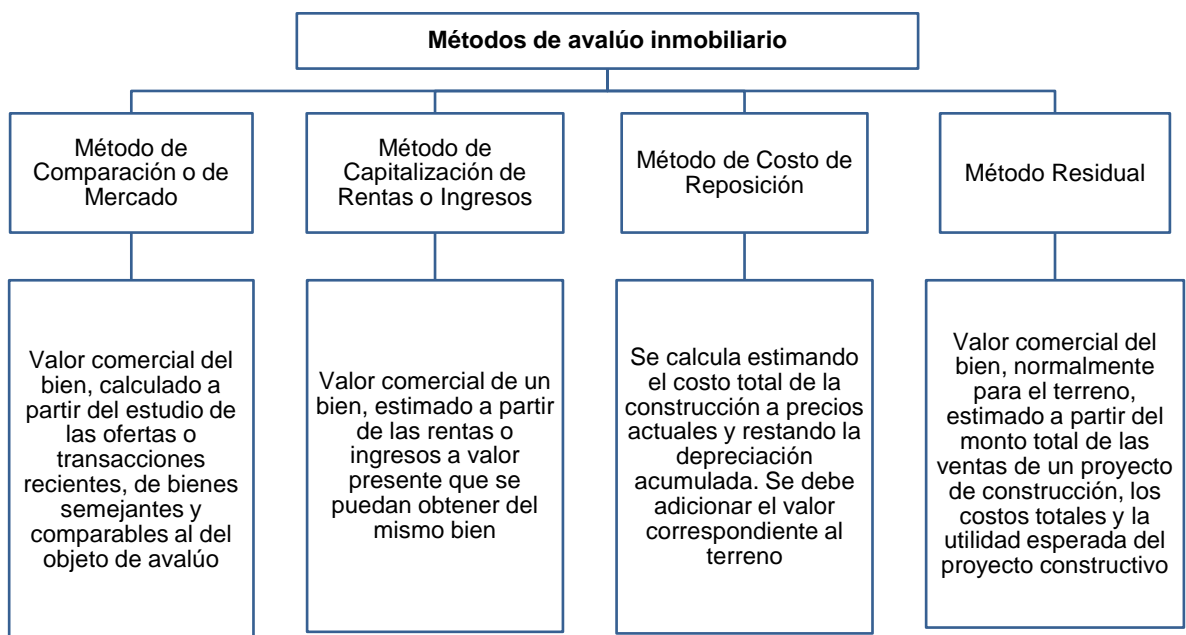


Figura 4. *Métodos de avalúo inmobiliario* (Elaboración propia a partir de Instituto Geográfico Agustín Codazzi, 2018)

## 2.3. Machine Learning

A continuación se presentan algunos conceptos relevantes de los métodos de aprendizaje automático o machine learning, con el fin de comprender su funcionamiento, y cómo las técnicas de aprendizaje automático pueden ser aplicadas en la solución de la problemática de la predicción de precios de las viviendas.

El aprendizaje automático o Machine learning pertenece al campo de la inteligencia artificial, (AI), la cual, es un subcampo de la informática que busca simular la inteligencia humana en una máquina. El aprendizaje automático brinda la posibilidad a las máquinas de aprender de las experiencias pasadas, sin ser programado explícitamente, y se basa en el desarrollo de algoritmos que pueden acceder a los datos y aprender de ellos (Iberdrola, 2022).

Los procesos de aprendizaje automático inician con datos u observaciones de ejemplo, y buscan identificar patrones en los mismos para realizar mejores decisiones en el futuro. De acuerdo con los objetivos específicos de los modelos de machine learning, estos algoritmos generalmente se dividen en cuatro categorías diferentes: aprendizaje automático supervisado, aprendizaje automático no supervisado, aprendizaje automático semi-supervisado y aprendizaje automático por refuerzo.

### Aprendizaje automático supervisado

Los algoritmos de aprendizaje automático supervisados utilizan ejemplos etiquetados, estos son, datos donde se conoce la variable objetivo o lo que es equivalente, la variable que se desea predecir. El propósito de estos modelos es predecir eventos futuros, estimando la variable objetivo para cualquier nuevo conjunto de datos de entrada en el modelo, utilizando un entrenamiento suficiente del modelo y la minimización del error.

De acuerdo con el tipo de la variable objetivo, los problemas de aprendizaje supervisado se pueden agrupar en clasificación y regresión:

- **Clasificación:** Un problema de clasificación se presenta cuando la variable objetivo es de tipo categórica o cualitativa, como por ejemplo las etiquetas “hombre” y “mujer” (Clasificación binaria) o “verde”, “azul” y “rojo” (Clasificación Multiclase).
- **Regresión:** Se trata de un problema de regresión cuando la variable objetivo es de tipo cuantitativo o numérico, como lo son el precio o el peso. Por ejemplo, la predicción de los precios de vivienda es un problema de regresión, ya que la variable objetivo precio es de tipo numérica.

## **Aprendizaje automático no supervisado**

Los algoritmos de aprendizaje automático no supervisados, se utilizan cuando los datos para entrenar el modelo no están clasificados o etiquetados, y el objetivo es encontrar características en los datos que puedan ser utilizadas para la clasificación. Algunos ejemplos de aprendizaje no supervisado son:

- **Clustering:** Un problema de clustering es aquel donde se desea descubrir las agrupaciones en los datos, como por ejemplo agrupar clientes por comportamiento de compra.
- **Reglas de Asociación:** Un problema de aprendizaje de reglas de asociación es donde se desea determinar las reglas que describen grandes porciones de los datos, como por ejemplo personas que compran X producto también tienden a comprar Y producto.

## **Aprendizaje automático semi-supervisado**

Los métodos de aprendizaje automático semi-supervisados, utilizan un conjunto de datos etiquetados y no etiquetados para clasificar nuevas instancias del conjunto de datos. Generalmente utilizan una pequeña cantidad de datos etiquetados y una gran cantidad de datos no etiquetados, a partir de esto, los métodos pueden mejorar considerablemente la precisión.

## **Aprendizaje automático por refuerzo.**

Este método permite que las máquinas determinen automáticamente el comportamiento ideal dentro de un contexto específico con el fin de maximizar su rendimiento.

### **2.3.1. Técnicas de Machine Learning para la predicción de precios de vivienda**

A continuación se presenta la descripción de algoritmos de machine learning más utilizados para desarrollar los modelos de regresión, en los cuales la variable objetivo es de tipo numérico, como en el caso específico de la predicción de precios de vivienda.

La regresión es un procedimiento estadístico utilizado para estimar la relación entre variables; específicamente, la regresión busca determinar la influencia de una o más variables en una variable numérica. La variable cuyo valor se desea estimar es llamada dependiente o respuesta, mientras las variables usadas para la estimación son llamadas variables dependientes o exploratorias. El objetivo de estos modelos es la predicción de la variable

objetivo con la mayor precisión posible, asumiendo que esta se comporta con un tipo de función lineal, logística u otro.

### **Regresión Lineal simple**

Es el modelo de regresión más simple, y describe la variable objetivo como una combinación de las variables independientes, asumiendo una relación lineal entre las variables. El objetivo final es determinar los coeficientes que conforman la función lineal, para obtener la predicción de la variable objetivo (Beyeler, 2017).

### **Árboles de decisión**

Los árboles de decisión son un método de aprendizaje supervisado empleado para la clasificación y la regresión; en estos modelos, los datos son divididos continuamente en subconjuntos más pequeños a partir de ciertos parámetros, utilizando reglas simples que son organizadas en una estructura de árbol. En estos modelos, se establece el mínimo número de reglas o preguntas necesarias para realizar una correcta predicción

### **Random Forest**

Random Forest es un algoritmo de aprendizaje supervisado que consiste en la combinación de varios árboles de decisión, los cuales son elegidos de forma aleatoria. La idea principal detrás de este algoritmo es construir una gran cantidad de árboles, cada uno de ellos para ser entrenado en un subconjunto aleatorio de datos y características (Suthaharan, 2016). Random Forest utiliza una técnica llamada bagging, en la cual se utilizan árboles de decisión paralelos, como resultado se obtienen muchos árboles aleatorios entrenados y cada árbol produce una predicción diferente. La predicción final es el promedio de las predicciones de los árboles de decisión.

### **Gradient Boosting (GBT)**

El modelo de Gradient Boosting (GBT), en español, algoritmo de aumento del gradiente, hace parte de los modelos de tipo ensamble (igual que Random Forest), en los cuales se combinan varios modelos de árboles de decisión. A diferencia de Random Forest, esta técnica utiliza la técnica de boosting, que consiste en mejorar un modelo débil, combinándolo con otros modelos débiles, hasta obtener un mejor modelo. En este caso, se utiliza el entrenamiento secuencial de un conjunto de árboles de decisión de poca profundidad. La finalidad es que el árbol tenga información del error del árbol anterior para mejorarlo.

**Extreme Gradient Boosting (XGBoost)**

El Extreme Gradient Boosting o XGBoost, es una versión mejorada del Gradient Boosting GBT, en la cual se utiliza una combinación de modelos con menores predicciones. En este modelo, los árboles de decisión son construidos en paralelo, en lugar de secuencia, además, este modelo está diseñado específicamente para mejorar la velocidad y el desempeño.

**Redes Neuronales artificiales (NN)**

Este algoritmo tiene su nombre debido a la similitud que existe (en su funcionamiento) con el cerebro humano. Una red neuronal artificial (ANN) es un modelo computacional que imita el comportamiento de las redes biológicas neuronales y se basa en una colección de unidades conectadas llamadas neuronas artificiales. Cada conexión transmite una señal de una neurona artificial a otra. La neurona artificial que recibe la señal puede procesarla y luego transmitirla a otra neurona que está conectada a ella.

La forma en que las neuronas se conectan en una red neuronal artificial se llama arquitectura de la red. Una red neuronal consta de varias neuronas que se agrupan en varias capas. En general, todas las neuronas que están en la misma capa tienden a tener un comportamiento similar. Cada arquitectura de la red neuronal tiene dos capas básicas, la capa de entrada que consta de todas las entradas de la red y la capa de salida que proporciona los resultados finales, además puede constar de más, incluyendo las capas ocultas.

**Vecinos más cercanos (KNN)**

Los métodos basados en vecindades se basan en el principio de encontrar un número predefinido de ejemplos del conjunto de entrenamiento, más cercanos en distancia a un nuevo ejemplo y en función de ellos realizar la predicción.

Los métodos basados en vecindades son conocidos como métodos de aprendizaje automático perezosos, ya que no constituyen ningún modelo, simplemente recuerdan todos los datos del conjunto de entrenamiento.

**Máquinas de soporte vectorial (SVM)**

Por sus siglas en inglés Support Vector Machine, es un modelo de aprendizaje supervisado que contiene algoritmos orientados a resolver problemas de clasificación y predicción basado en regresiones.

### 2.3.2. Métricas para evaluación de los modelos

En las secciones anteriores se definieron los conceptos principales del aprendizaje automático. Esta sección se centra en las diferentes métricas de desempeño utilizadas para evaluar la calidad de las predicciones de los diferentes modelos de machine learning.

#### Error cuadrático medio o Mean Squared Error (MSE)

El error cuadrático medio consiste en calcular la media de todos los errores cuadráticos, es decir, la diferencia entre el valor predicho por el modelo y el valor real, todo ello elevado al cuadrado, según el número total de datos (Martinez Heras, 2020). El MSE siempre es positivo ya que el error está al cuadrado y cuanto más se acerque esta función a cero, más precisa será la predicción del modelo de aprendizaje automático. Sin embargo, esta métrica es menos adecuada cuando hay muchos valores atípicos, ya que un valor atípico crea una gran diferencia entre el valor predicho y el valor real. La función para el MSE se muestra en la Ecuación 1.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Ecuación 1 MSE

Donde:

$y_i$ : Valores predichos por el modelo

$\hat{y}_i$ : Valores reales

$n$ : Número de predicciones

#### Raíz del error cuadrático medio o Root Mean Square Error (RMSE)

Esta métrica surge de una modificación del MSE, consiste en calcular la raíz cuadrada de forma que se obtiene el “Root mean square error” o RMSE. Esta fórmula escala el error y se puede denotar como se muestra en la Ecuación 2.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Ecuación 2 RMSE

### **Error absoluto medio o Mean Absolute Error (MAE)**

El error absoluto medio muestra el promedio de los errores absolutos entre los valores de la predicción y los valores reales. Comparando el MAE con el MSE, el MAE es más indulgente con los valores atípicos, ya que no eleva al cuadrado el error, sino que lo convierte en un valor absoluto. Su fórmula se puede ver en la Ecuación 3.

$$MAE = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{y}_i]$$

Ecuación 3 MAE

### **Error porcentual absoluto medio o Mean Absolute Percentage Error (MAPE)**

Es una medida de la precisión de la predicción que compara el error con el valor real, lo que puede convertirlo en una función útil cuando los datos están más desviados. Además, el MAPE se denota en porcentajes, lo que lo hace mejor interpretable. El MAPE se puede notar como se puede ver en la Ecuación 4.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Ecuación 1 MAPE

### **Error logarítmico cuadrático medio o Mean Squared Logarithmic Error (MSLE)**

Es el valor logarítmico del promedio de la diferencia entre los valores predichos y los valores reales. Este tipo de predicción es especialmente útil en áreas con grandes valores atípicos. La ecuación del MSLE se muestra en la Ecuación 5.

$$MSLE = \frac{1}{n} \sum_{i=1}^n (\log(\hat{y}_i + 1) - \log(y_o + 1))^2$$

Ecuación 2 MSLE



## 2.4. Estudios previos

En este apartado se presentan los antecedentes y casos de estudios relacionados con la aplicación de técnicas de Machine Learning para la predicción de precios de vivienda en diferentes ciudades alrededor del mundo y en Colombia

### 2.4.1. Estudios previos alrededor del mundo

Antipov y Pokryshevskaya (2012) utilizaron el método de Random Forest para la valoración de viviendas a partir de un conjunto de datos de 2.695 apartamentos residenciales de la ciudad de San Petersburgo, Rusia. Como resultado de su investigación, concluyeron que este modelo presentaba mejor desempeño que otras técnicas, tales como las redes neuronales artificiales o el análisis de regresión múltiple.

En el año 2014, los ingenieros Xibin Wanga, Junhao Wena, Yihao Zhanga y Yubiao Wang de la universidad de Chongqing en China, utilizaron el modelo de máquinas de vectores de soporte (SVM) y la selección de parámetros para la predicción del precio de viviendas en la ciudad de china Chongqing (Wang et al., 2014).

En el año 2015, se desarrolló un modelo lineal para predecir los precios de las casas unifamiliares en el condado de Los Ángeles, Estados Unidos, empleando datos entre los años 2003 y 2009, siendo uno de los estudios con mayor cobertura de área y tiempo. Para este estudio se analizó un conjunto de datos de más de 15 millones de registros. Se empleó el modelo de Random Forest contra un modelo lineal, obteniendo mejores resultados con el primero.

Con el fin de apoyar la toma de decisiones de los agentes inmobiliarios e instituciones financieras, Park y Bae (2015) desarrollaron un modelo para la predicción de precios de vivienda en el Condado de Fairfax, Virginia, utilizando técnicas de machine learning a partir de datos de 5.359 viviendas. Para su modelo utilizaron algoritmos como C4.5, el cual es utilizado para generar un árbol de decisión, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), Naive Bayes y AdaBoost (Adaptive Boosting). En su modelo utilizaron dos técnicas; la primera fue resolver un problema de clasificación mediante la técnica de Random Forest, y el segundo fue utilizar una regresión utilizando el algoritmo Naive Bayes. Los resultados mostraron que el algoritmo RIPPER mejoró significativamente la predicción de precios.

En respuesta al problema de clasificación para determinar si los precios de la vivienda aumentarían o disminuirían, Banerjee y Dutta (2017), analizaron diferentes algoritmos de

aprendizaje automático. Para ello, utilizaron un conjunto de datos publicado en el sitio web Kaggle.com y utilizaron diferentes técnicas de aprendizaje automático, como máquinas de soporte vectorial (SVM), redes neuronales (NN) y la técnica de Random Forest. Los resultados mostraron que la técnica de Random Forest presentó mejores resultados, con una precisión del 79%, sin embargo, también presentó el mayor sobreajuste. Por el contrario, la técnica SVM fue la más consistente y, por tanto, la más fiable.

Con respecto al avalúo de los bienes inmuebles, Kok et al. (2017) analizaron el rendimiento de varias técnicas de aprendizaje automático. Examinaron 84.305 observaciones de los estados de California, Florida y Texas, durante el período de 2011 a 2016, y compararon diferentes técnicas de aprendizaje, como las técnicas de regresión de mínimos cuadrados ordinarios (OLS), Random Forest, GBR y XGBoost. Los resultados mostraron que, en términos generales, XGBoost fue el algoritmo que mejor funcionó.

Ceh et al. (2018) compararon el algoritmo Random Forest contra el modelo de precios hedónicos con el fin de analizar qué técnica obtendría mejores predicciones. Los autores utilizaron una muestra de 7.407 propiedades durante el período comprendido entre 2008 y 2013 en Ljubljana (Eslovenia). Los resultados mostraron que el modelo Random Forest tuvo un mejor desempeño predictivo, utilizando como métricas de medición el coeficiente de determinación y el MAPE.

En un concurso organizado por Kaggle.com, en el que los participantes debían proponer un algoritmo para la predicción del precio de la vivienda a partir del conjunto de datos Ammen, Fan et al. (2018) utilizaron algoritmos predictivos basados en regresiones como Random Forest, máquinas de soporte vectorial (SVM), XGBM, ridge y regresión lineal LASSO. Los datos fueron proporcionados por Ames Housing en Iowa, con registros de 2006 a 2010. Los resultados mostraron que ridge, LASSO y XGBM tenían un error de predicción más bajo.

Motivado por los aumentos de precios de vivienda en Estocolmo en los últimos 20 años, Nilsson (2019) utiliza en su tesis de estudio las redes neuronales para predecir los precios de vivienda en esta ciudad a partir de diferentes parámetros de los apartamentos.

Para predecir los precios de alquiler de apartamentos en Dhaka (Bangladesh), Neloy et al. (2019) compararon varios algoritmos, entre ellos: Regresión lineal. Redes Neuronales, Support Vector Machine, Random Forest, árbol de decisión (DT), Ensemble AdaBoosting Regressor, Ensemble Gradient Boosting Regressor, Ensemble XGBoost y Ridge Regression, Lasso Regression, and Elastic Net Regression. Los resultados mostraron que los algoritmos de Random Forest tenían un error cuadrático medio más bajo.

Voutas Chatzidis (2019) utilizó diferentes algoritmos de máquina basados en regresión para predecir los precios de la vivienda en los Países Bajos a partir de 546.110 registros de transacciones comprendidas entre los años 2005 hasta 2018. El autor utilizó algoritmos LGBM, XGBM, CatBoost y Random Forest. Como resultado CatBoost obtuvo los mejores resultados con una tasa de precisión del 90%.

En un estudio que analizó toda España, Alfaro-Navarro et al. (2020) propusieron una nueva metodología para llevar a cabo la predicción automatizada de los precios de la vivienda. Se generó un modelo diferente para cada municipio y se logró una muestra de 790.631 inmuebles para los 433 municipios analizados y 48 provincias. Los modelos se realizaron utilizando algoritmos de bagging, boosting y Random Forest. Los resultados mostraron que los algoritmos de boosting y Random Forest fueron ligeramente mejores a partir de la métrica de MAPE.

Hong (2020) comparó el comportamiento predictivo de métodos tradicionales de regresión lineal con el aprendizaje automático utilizando tres algoritmos (XGBM, LGBM, CatBoost) para predecir el precio de transacción de los apartamentos en Seúl. Para ello, el autor utilizó una muestra de 620.617 observaciones para el período comprendido entre 2009 y 2019. Los resultados mostraron que los algoritmos de machine learning tenían más poder predictivo que el modelo Ordinary Least Squares (OLS). Además, se observó que el algoritmo CatBoost era superior en términos de predicción de precios, incluso cuando se trataba de valores atípicos. Además, se encontró que el modelo con un conjunto de los tres algoritmos presentaba una mayor precisión que los algoritmos individuales.

Para predecir el precio de transacción de los apartamentos en Gangnam (Corea del Sur), Hong et al. (2020) compararon el comportamiento predictivo de modelos de precios hedónicos frente al aprendizaje automático mediante el uso de la técnica de Random Forest. Para ello, los autores utilizaron una muestra compuesta por 16.601 apartamentos para el período comprendido entre 2006 y 2017. Los resultados mostraron que la técnica Random Forest fue superior en cuanto a la predicción del precio con una desviación de error de 5,5%

Por último, Hu et al. (2021) analizaron el rendimiento predictivo a través de algoritmos de aprendizaje supervisado para precios de alquiler de viviendas en Shenzhen (China) para el periodo comprendido entre 2017 y 2018. Los autores utilizaron los algoritmos Random Forest (RF), Extra-Trees Regression (ETR), Gradient Boosting Regression (GBR), Support Vector Regression (SVR), redes neuronales multicapa (MLP-NN) y K vecinos más cercanos (KNN). Los resultados mostraron que los algoritmos Random Forest y ETR tuvieron un mejor desempeño predictivo. Encontrando como factores determinantes la cercanía de las viviendas a oportunidades de empleo y hospitales.

La Tabla 3 presenta el resumen de los estudios sobre la aplicación de técnicas de machine learning para la predicción de precios de vivienda en diferentes ciudades alrededor del mundo.

Tabla 3. Resumen de estudios relacionados en el mundo

Autor	Ubicación	Modelos aplicados	Mejor Resultado
Wang et al., 2014	Chongqing, China	Máquinas de vectores de soporte (SVM),	SVM
Park y Bae, 2015	Condado de Fairfax, Virginia, USA	C4.5, RIPPER, Naive Bayes y AdaBoost.	RIPPER
Antipov y Pokryshevskaya, 2012	San Petersburgo, Rusia	CHAID, CART, KNN, Regresión múltiple, Redes neuronales artificiales	Random Forest
Banerjee y Dutta, 2017	Kaggle	Máquinas de vectores de soporte (SVM), redes neuronales (NN) y la técnica de Random Forest.	SVM
Hong, 2020	Seúl, Corea del sur	XGBM, LGBM, CatBoost	CatBoost
Hong et al., 2020	Gangnam, Corea del Sur	Random Forest	Random Forest
Alfaro-Navarro et al., 2020	España	Bagging Boosting Random Forest	Boosting Random Forest
Nilsson, 2019	Estocolmo, Suecia	Redes neuronales artificiales	Redes neuronales artificiales
Voutas Chatzidis, 2019	Países bajos	LGBM, XGBM, CatBoost y Random Forest	CatBoost
Neloy et al., 2019	Dhaka, Bangladesh	Regresión lineal. Redes Neuronales, Máquinas de vectores de soporte (SVM), Random Forest, árbol de decisión (DT), Ensemble AdaBoosting Regressor, Ensemble Gradient Boosting Regressor, Ensemble XGBoost y Ridge Regression, Lasso Regression, and Elastic Net Regression.	Random Forest
Hu et al., 2021	Shenzhen, China	Random Forest regression (RFR), extra-trees regression (ETR), gradient-boosting regression (GBR), support vector regression (SVR), multi-layer perceptron neural network (MLP-NN) and <i>k</i> -nearest neighbor algorithm ( <i>k</i> -NN).	Random Forest regression (RFR) Extra-trees regression (ETR)
Fan et al., 2018	Ammen Kaggle	RF, SVM, XGBM, ridge y regresión lineal LASSO.	Ridge, LASSO y XGBM
Ceh et al., 2018	Ljubljana, Eslovenia	Random Forest	Random Forest
Kok et al., 2017	California, Florida y Texas, USA	Regresión de mínimos cuadrados ordinarios (OLS), RF, GBR y XGBM	XGBM
Lowrance et al. 2015	Los Ángeles, USA	Random Forest Regresión lineal	Random Forest

Fuente: Elaboración propia

## 2.4.2. Estudios previos en Colombia

Respecto a estudios en Colombia para resolver la problemática de estimar los precios de los inmuebles, en el año 2019 se desarrolló como trabajo de grado un modelo para la predicción de los precios de vivienda en el municipio de Rionegro, para apoyar la toma de decisiones de compra y venta de propiedad raíz (Grajales Alzate, 2019). En este estudio se utilizó técnicas de web scraping sobre los portales Finca raíz y Mercado libre, para la construcción de un conjunto de datos de 3.033 registros y 18 columnas, de inmuebles del municipio de Rionegro. Para los modelos se utilizó cinco técnicas de machine learning: regresión lineal, árboles de decisión, Random Forest, gradient boosting machine y por último máquinas de soporte vectorial. Como resultados encontraron que, a partir del coeficiente de determinación, el modelo con mejor desempeño para predecir el precio de las viviendas en el municipio de Rionegro fue un modelo de gradient boosting, seguido por Random Forest. Además, entre los resultados se obtuvo que las variables que más influyen en la predicción del precio de vivienda en Rionegro son el área de la vivienda, el área construida, el tipo de vivienda, el estrato y el número de baños.

En su trabajo de Investigación Martínez Sánchez y Téllez Buitrago (2021) desarrollaron un método automático para la predicción de del avalúo comercial de un inmueble en la ciudad de Bogotá, implementando cuatro modelos de machine learning, utilizando las técnicas de árboles de decisión, regresión lineal, Random Forest y redes neuronales profundas, empleando el estrato socioeconómico como una variable dentro del análisis. Para el desarrollo de los modelos utilizaron un conjunto de datos construido utilizando la herramienta de web Scraping Dexi.io para la extracción de información de la página web de Finca Raíz con un dataset final de 1.549 registros y 10 columnas. Finalmente, utilizando como medidas de desempeño el coeficiente de determinación, el RMSE (Raíz del error cuadrático medio) y el coeficiente de variación, obtuvieron mejores resultados con el modelo de Random Forest para el avalúo de un inmueble en la ciudad de Bogotá.

Este mismo año, Soto Hincapié y David Rodríguez (2021) en su estudio exploraron las mejores técnicas de aprendizaje supervisado para la predicción de precios de vivienda en el Valle de San Nicolás, comprendido por los municipios de Rionegro, Guarne, El Carmen de Viboral, El Retiro, El Santuario y Marinilla, en el departamento de Antioquia. Para su análisis utilizaron un conjunto de datos de 2.481 registros extraídos del portal Fincaraiz.com a través de la técnica de web scraping, y emplearon modelos estadísticos como la regresión lineal múltiple, la regresión por mínimos cuadrados ordinarios y la regresión Ridge, técnicas de ensamble con modelos como Random Forest, Gradient Boosting y XGBoost y redes

neuronales. Como resultado para el conjunto de datos de viviendas en el Valle de San Nicolás, los mejores modelos para la predicción de precios fueron el Random Forest y el XGBoost.

En la Tabla 4 se presenta el resumen de los estudios sobre la aplicación de técnicas de machine learning para la predicción de precios de vivienda en Colombia.

Tabla 4. Resumen de estudios relacionados en Colombia

Autor	Ubicación	Modelos	Mejores Resultados
Grajales Alzate, 2019	Rionegro	Regresión Lineal. Árboles de decisión de regresión. Máquinas de soporte vectorial (SVM). Random Forest Gradient Boosting Machine (GBM).	Gradient boosting R2: 0.75, MAPE:0.16 Random Forest R2: 0.77, MAPE:0.15
Martínez Sánchez y Téllez Buitrago, 2021	Bogotá	Árboles de decisión Regresión lineal Random Forest Redes neuronales profundas	Random Forest R2 0.91 CV:0.17
Soto Hincapié & David Rodríguez, 2021	Valle de San Nicolás	Regresión lineal múltiple Regresión por mínimos cuadrados ordinarios Regresión Ridge Random Forest Gradient Boosting XGBoost Redes neuronales	Random Forest R2 0.93 CV 0.17  XGBoost. MAPE: 0.24

Fuente: Elaboración propia

## 2.5. Conclusiones de los estudios previos

Los métodos de aprendizaje automático han sido ampliamente utilizados en la última década para la predicción de precios de vivienda en diferentes ciudades alrededor del mundo. En general, los resultados del aprendizaje automático en el sector inmobiliario son prometedores en cuanto a la precisión de los modelos, sin embargo, el contexto de estos estudios revela la importancia de utilizar múltiples criterios para seleccionar el tipo de modelo de aprendizaje automático a aplicar.

Como se evidencia en los casos de estudio previos, se ha dedicado un gran número de investigaciones para determinar el algoritmo más adecuado para realizar la predicción de precios. En este contexto, los algoritmos varían desde técnicas simples, como varios tipos de regresión y árboles de decisión, hasta otras más complejas, como métodos de ensamble o la implementación de redes neuronales.

El presente trabajo pretende ampliar la literatura con la propuesta de un método de aprendizaje automático a partir de datos de la ciudad de Medellín y el Área Metropolitana, considerando las particularidades del comportamiento de los precios de vivienda en esta ciudad, en el desarrollo de un código fuente que puede ser utilizado para trabajar con otros conjuntos de datos.

### **3. Objetivos concretos y metodología de trabajo**

En este capítulo se presenta el objetivo general, los objetivos específicos y la metodología de trabajo.

#### **3.1. Objetivo general**

El objetivo principal de este proyecto es desarrollar un modelo predictivo de los precios de venta de viviendas en la ciudad de Medellín y el Área Metropolitana, para apoyar la toma de decisiones en los procesos de compra y venta de los inmuebles.

#### **3.2. Objetivos específicos**

Los objetivos específicos de este trabajo son:

- Seleccionar el conjunto de datos a utilizar, a partir de información de ofertas inmobiliarias en la ciudad de Medellín y el Área Metropolitana.
- Realizar el análisis descriptivo y la exploración de las variables del conjunto de datos
- Determinar las técnicas de aprendizaje automático que serán utilizadas para la predicción de precios de venta de vivienda.
- Construir los modelos predictivos para pronosticar el precio de viviendas en la ciudad de Medellín y el Área Metropolitana.
- Evaluar el desempeño de los modelos y determinar el modelo que más se ajusta al comportamiento de los datos analizados.

### 3.3. Metodología del trabajo

En este capítulo se presenta la descripción de la metodología para el desarrollo del presente trabajo. Se seleccionó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es una de las metodologías más utilizadas para el desarrollo de proyectos de minería de datos. A continuación, se presenta en la Figura 5, las etapas de la metodología.

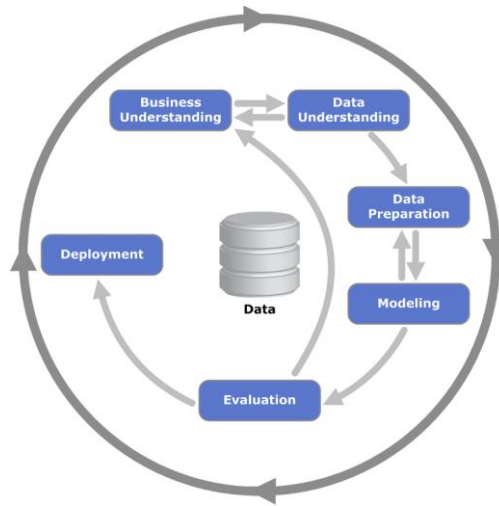


Figura 5. Metodología CRISP-DM (Rodríguez, 2010)

En la Tabla 5 se expone la descripción de las etapas de la metodología CRISP-DM.

Tabla 5. Descripción de las etapas de la metodología CRISP-DM

Etapa	Descripción
<b>Comprensión del negocio</b>	En esta etapa se debe realizar la identificación del problema, los beneficios del proyecto, los objetivos, y la evaluación de la situación actual.
<b>Comprensión de los datos</b>	Esta etapa incluye la recolección de los datos a utilizar en el proyecto, incluyendo las fuentes y técnicas de recolección, la descripción de los datos y su exploración
<b>Preparación de los datos</b>	En esta etapa se seleccionan los datos y se realizan las actividades de limpieza y transformación
<b>Modelado</b>	Es la etapa en la cual se realiza la selección de la técnica de modelado y la obtención del modelo.
<b>Evaluación</b>	En esta etapa se establecen las métricas para evaluar la calidad del modelo, de acuerdo con los resultados en esta etapa, se decide si continuar con la última fase o regresar a las etapas anteriores.
<b>implementación</b>	En esta etapa se utiliza el conocimiento adquirido mediante la implementación del modelo

Fuente: Elaboración propia



Teniendo en cuenta que las actividades asociadas a la comprensión del negocio están incluidas en los capítulos anteriores, para el desarrollo de la contribución se iniciará con las actividades asociadas a la comprensión de los datos.

En la Figura 6 se presenta el diagrama de procesos para describir las fases y actividades a ejecutar en el desarrollo de la contribución, las cuales están basadas en la metodología seleccionada.

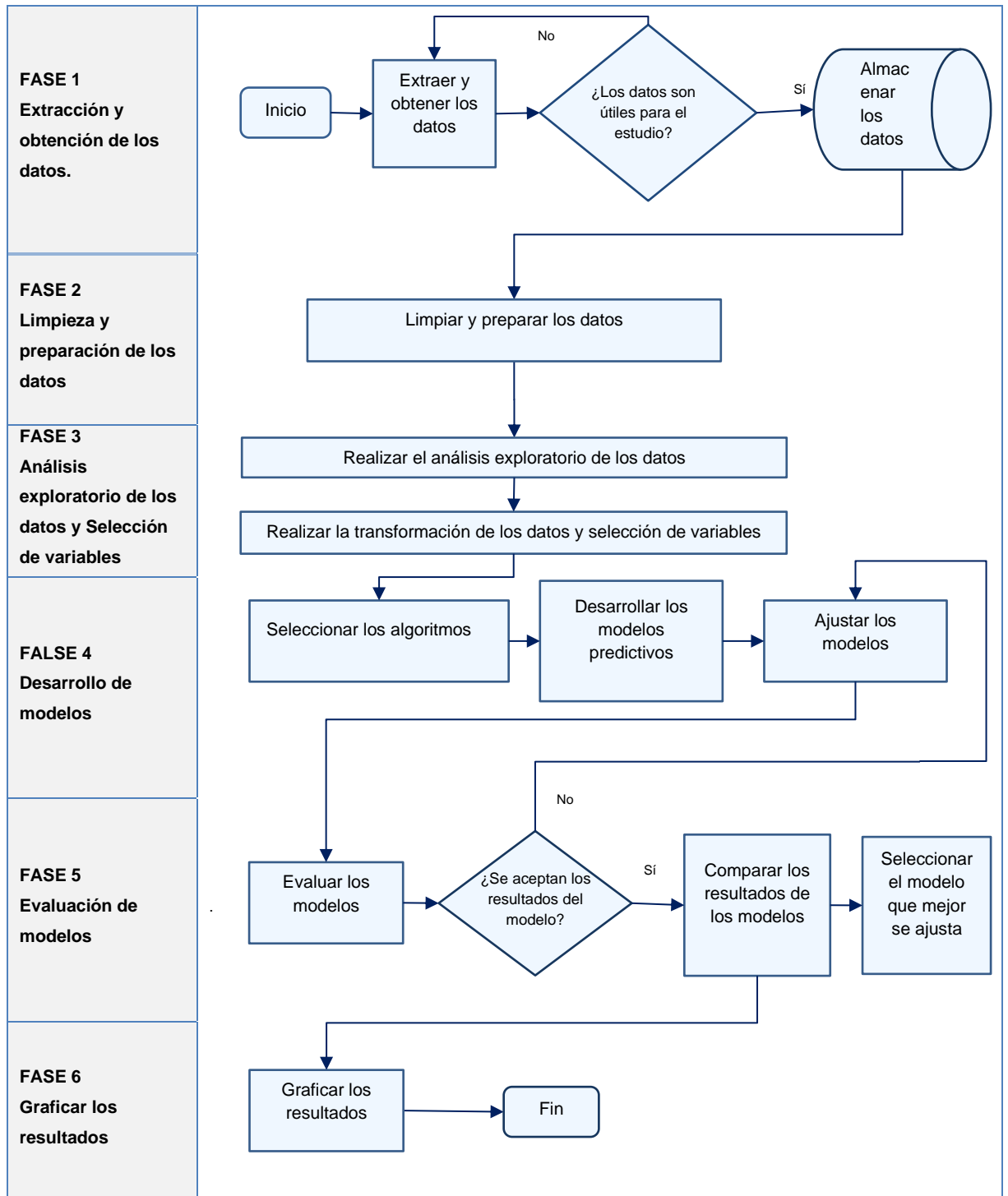


Figura 6. Diagrama de procesos de la metodología de trabajo (Elaboración propia)

A continuación, en la Tabla 6 se presentan los objetivos que se pretenden alcanzar en cada fase para el desarrollo de la contribución.

Tabla 6. Objetivos de las fases de la metodología de trabajo

<b>Fase</b>	<b>Nombre de Fase</b>	<b>Objetivos</b>
<b>1</b>	<b>Extracción y obtención de los datos.</b>	Obtener los datos de las publicaciones de venta de viviendas en la ciudad de Medellín y el Área Metropolitana, a partir de conjuntos de datos públicos de los portales inmobiliarios.
<b>2</b>	<b>Limpieza y preparación de los datos</b>	Realizar los procesos necesarios para la limpieza y preparación del conjunto de datos
<b>3</b>	<b>Análisis exploratorio de los datos y Selección de variables</b>	Analizar los datos de entrada y el comportamiento de las variables relacionadas con el precio de venta de las viviendas Transformar los datos y realizar la selección de variables de acuerdo con los requerimientos de los modelos a aplicar
<b>4</b>	<b>Desarrollo de modelos</b>	Seleccionar las técnicas de aprendizaje automático a aplicar para el desarrollo de los modelos a partir de la bibliografía existente Desarrollar los modelos para la predicción de los precios de venta de vivienda a partir de las técnicas de aprendizaje automático seleccionadas Ajustar los modelos en sus hiperparámetros y su aplicación en datos de prueba y entrenamiento
<b>5</b>	<b>Evaluación de modelos</b>	Determinar y aplicar las medidas de desempeño para la evaluación de los modelos y realizar el análisis comparativo de las mismas Seleccionar el modelo que mejor se ajusta a los datos
<b>6</b>	<b>Graficar los resultados</b>	Graficar los resultados obtenidos con los modelos para la predicción de precios de venta de las viviendas

Fuente: Elaboración propia

## 4. Desarrollo específico de la contribución

En este capítulo se exponen las actividades realizadas en el desarrollo del presente trabajo, iniciando con la descripción de la fuente de datos, seguida del análisis, los algoritmos y las métricas utilizados en el presente documento.

### 4.1. Extracción y obtención de los datos

A continuación se presenta la fuente de los datos empleada, se describe el proceso de extracción y el conjunto de datos original.

#### 4.1.1. Fuente de los datos

El conjunto de datos empleado en el presente documento, corresponde a los datos públicos de la empresa Properati, el cual contiene información de propiedades en Latinoamérica, listadas en su portal web para la realización de procesos de venta o alquiler.

Properati es un portal inmobiliario con origen en Argentina y presencia en toda Latinoamérica, el cual facilita los conjuntos de datos de los inmuebles anunciados en su portal web, con el objetivo de que todos los interesados puedan acceder a información abierta y actualizada para realizar análisis sobre las viviendas y el sector inmobiliario.

Se utiliza el conjunto de datos de propiedades listadas en el portal web de Properati (Properati, 2022), almacenadas en la herramienta Bigquery de Google. A partir de consultas SQL se obtienen los datos correspondientes a propiedades en Colombia, localizadas en el departamento de Antioquia y, con operaciones de venta. Finalmente se obtuvo un conjunto de datos inicial que presenta las siguientes características resumidas en la Tabla 7.

Tabla 7. Fuente de los datos

Fuente	Descripción	Detalles	Acceso
Properati	Conjunto de datos con información de inmuebles localizados en Antioquia, Colombia, publicados para la venta en el portal inmobiliario Properati, en el periodo comprendido entre mayo de 2015 y mayo del 2021	578.490 filas 28 columnas Tamaño: ~900 MB	Consultas SQL en Bigquery

Fuente: Elaboración propia

### 4.1.2. Protección de los datos

En el presente apartado se explica el cumplimiento del Reglamento general de protección de datos (RGPD), el cual establece las pautas a seguir en lo relativo al tratamiento de los datos personales.

La fuente de datos empleada para el presente análisis es de acceso público, y está disponible para ser utilizada libremente, de acuerdo con la iniciativa de datos abiertos liderada por el portal inmobiliario Properati, por esta razón no se requiere tratamiento específico para el uso de datos personales.

El conjunto de datos ofrecidos por el portal inmobiliario Properati, se publican bajo los términos de la licencia CC BY 3.0 de Creative Commons y permite copiar, distribuir, exhibir y modificar los datos, siempre que se cite debidamente la fuente y los cambios ejecutados sobre estos (Properati Data, 2022).

### 4.1.3. Herramientas

A continuación, se describen las herramientas principales utilizadas para la extracción y el tratamiento de los datos.

**BigQuery:** Los datos de las propiedades y desarrollos inmobiliarios publicados en el portal web de la empresa Properati, se encuentran disponibles para los usuarios de BigQuery, el cual es un producto de Google Cloud para consultar grandes volúmenes de datos. Se realiza la extracción de los datos utilizando consultas SQL en este almacén de datos de Google, las consultas se realizan directamente desde el notebook, estableciendo una conexión directa entre el código propuesto y la fuente de los datos.

**Jupyterlab:** Es una aplicación web que permite la creación y edición de notebooks (cuadernos) y otros tipos de archivos. Se utiliza como entorno para el desarrollo del código del presente proyecto. Jupyterlab incluye celdas que a su vez pueden contener código, texto, imágenes, ecuaciones, gráficos, entre otros. Para el presente proyecto se utiliza Jupyterlab desde la plataforma de Google Cloud.

**Python:** Se utiliza el lenguaje de programación Python y algunas de sus librerías. Las librerías de Python son un conjunto de funciones disponibles para ser utilizadas en este lenguaje, algunas de las librerías utilizadas en el presente documento son: Pandas, Numpy, Matplotlib, Seaborn, Scikit-Learn, Scipy, math.

#### 4.1.4. Descripción del conjunto de datos

El conjunto de datos original contiene información sobre la venta de inmuebles localizados en Antioquia, Colombia, con un total de 578.490 registros publicados en el portal inmobiliario Properati, en el periodo comprendido entre mayo de 2015 y mayo de 2021. El conjunto de datos presenta un total de 28 columnas correspondiente a las variables explicativas que se presentan en la Tabla 8, además, se incluye la descripción de la variable y el tipo de datos.

Tabla 8. Variables del conjunto de datos original

Variable	Descripción	Tipo de dato
<b>type</b>	Tipo de aviso (Propiedad, Desarrollo/Proyecto)	object
<b>type_i18n</b>	Tipo de aviso (Propiedad, Desarrollo/Proyecto)	object
<b>country</b>	País en el que está publicado el aviso (Argentina, Uruguay, Colombia, Ecuador, Perú)	object
<b>id</b>	Identificador del aviso. No es único: si el aviso es actualizado por la inmobiliaria (nueva versión del aviso), se crea un nuevo registro con la misma id, pero distintas fechas de alta y de baja	object
<b>start_date</b>	Fecha de creación del aviso	object
<b>end_date</b>	Fecha de baja del aviso	object
<b>created_on</b>	Fecha de creación de la primera versión del aviso	object
<b>place</b>	Campos referidos a la ubicación de la propiedad o del desarrollo. Diccionario con coordenadas latitud y longitud del inmueble	object
<b>property</b>	Campos relativos a la propiedad (vacío si el aviso es de un desarrollo/proyecto).	object
<b>development</b>	Campos relativos al desarrollo inmobiliario (vacío si el aviso es de una propiedad).	object
<b>lat</b>	Latitud de la localización del inmueble	float64
<b>lon</b>	Longitud de la localización del inmueble	float64
<b>I1</b>	País de localización de la propiedad: Colombia	object
<b>I2</b>	Provincia o departamento de localización de la propiedad	object
<b>I3</b>	Ciudad o municipio de localización de la propiedad	object
<b>I4</b>	Barrio de localización de la propiedad	object
<b>operation</b>	Tipo de operación (Venta, Alquiler)	object
<b>property_type</b>	Tipo de propiedad (Casa, Departamento, Lote, Oficina, Locales, Depósitos)	object
<b>rooms</b>	Cantidad de ambientes (útil en Argentina)	float64
<b>bedrooms</b>	Cantidad de dormitorios (útil en el resto de los países)	float64
<b>bathrooms</b>	Cantidad de baños	float64

<b>surface_total</b>	Área total del inmueble en m <sup>2</sup>	float64
<b>surface_covered</b>	Área cubierta del inmueble en m <sup>2</sup>	float64
<b>price</b>	Precio del inmueble publicado en el anuncio	float64
<b>currency</b>	Moneda del precio publicado del inmueble	object
<b>price_period</b>	Periodo del precio (Diario, Semanal, Mensual)	object
<b>title</b>	Título del anuncio	object
<b>description</b>	Descripción del anuncio	object

Fuente: Elaboración propia

## 4.2. Limpieza y preparación de los datos

Se realiza el proceso de limpieza y preparación del conjunto de datos utilizando la herramienta de Python, el cual incluye las siguientes actividades:

### Selección de los datos

Se filtra el conjunto de datos original para obtener únicamente los registros de interés correspondientes a venta de inmuebles en la ciudad de Medellín y el Área Metropolitana del Valle de Aburrá. Para esto se filtra por la variable 'l4', correspondiente a municipio y se selecciona los registros correspondientes a: Medellín, Caldas, Sabaneta, Itagüí, Envigado, Bello, Copacabana, Girardota, La Estrella y Barbosa.

### Consistencia en los datos

Se realiza el proceso de verificación de los valores correspondientes en cada variable para identificar los datos inconsistentes. Para esto se exploran los datos con la herramienta `profile_report` de la librería `pandas`. Se corrigen y reemplazan los datos inconsistentes, por ejemplo, el dato con fecha 9999-12-31 en la variable `end_date` se cambia por un valor consistente.

### Tipo de datos

Se identifica el tipo de dato de cada variable a analizar y se cambia el tipo de datos, por una tipología de datos consistentes para las posteriores etapas. Se cambia el tipo de datos en algunas variables a tipo categórico, `datetime` o numérico

### Normalización de los datos

Se realiza la normalización de las variables categóricas, se identifica la necesidad de convertir los datos en minúsculas y sin tildes para facilidad en el manejo del conjunto de datos.

### Registros duplicados

Se identifican los registros correspondientes al mismo inmueble y se eliminan los valores duplicados.

### Tratamiento de datos perdidos

Se identifica un alto porcentaje de valores perdidos en algunas variables del conjunto de datos. Sobre estos datos es necesario realizar un tratamiento, ya sea eliminarlos o complementarlos con un valor estimado. Se definió el tratamiento de valores perdidos en el conjunto de datos, de acuerdo con el tipo de dato y cada variable. A continuación, se presenta en la Figura 7, el gráfico de porcentaje de valores nulos de cada variable del conjunto de datos.

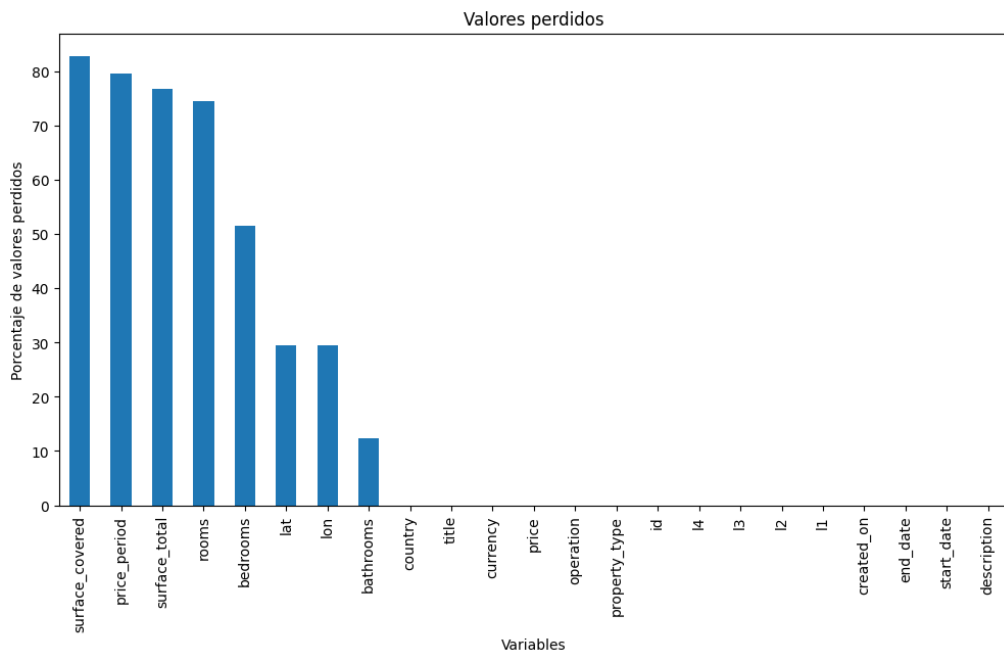


Figura 7. Valores nulos en el conjunto de datos (Elaboración propia)

Las variables `surface_covered` y `surface_total` presentan altos porcentajes de valores perdidos ya que es una información omitida en el 82% y 76% de los anuncios respectivamente. El 79% de los anuncios no presenta registro de `price_period`, adicionalmente, el 51% de los datos no presenta información del número de habitaciones (`bedrooms`), el 29% de los datos no presenta datos de coordenadas de latitud y longitud y, el 12% no incluye datos del número de baños (`bathrooms`). Todos los registros presentan datos en la variable precio, el cual corresponde a la variable objetivo del estudio.

Se eliminan los datos sin registro de superficie total, ya que es una variable explicativa importante en la determinación de los precios de un inmueble y, no se encuentra implícita en otras columnas como descripción o título.

### **Selección de variables relevantes**

A partir del porcentaje de valores perdidos, se omiten las variables no relevantes para el análisis, como el caso de `price_period` y `rooms`. La variable `price_period` se refiere a la frecuencia de pagos del precio, la cual es relevante para los datos con operaciones de alquiler y tiene poca importancia para transacciones de venta de los inmuebles.

Por su parte, `rooms` es una propiedad de número de ambientes del inmueble, y de acuerdo con la documentación del conjunto de datos, esta variable es utilizada en otros países como Argentina, mientras en Colombia se utiliza la variable `bedrooms`.

Otras variables no consideradas son `start_date` y `end_date`. Una vez identificados los valores de anuncios duplicados, la columna `start_date` no aporta información adicional a la columna `created_on`, siendo el 100% de los registros iguales, por esta razón se define el uso de una de las dos para posteriores análisis. Por su parte, la columna `end_date` se refiere a la fecha de baja del anuncio, el cual se omitirá para los análisis, al tratarse en su gran mayoría de fechas en el futuro.

Respecto a la variable `currency` (moneda), se filtran los datos únicamente con precio de venta en moneda pesos colombianos COP, y por así es posible omitir esta variable en los análisis posteriores. Igualmente, con la variable `operation`, se consideran únicamente los valores de venta de inmuebles y, se eliminan esta columna. La variable `type` considera un único tipo de datos, correspondiente a Propiedades, y por tanto no se requiere ese tipo de anuncios en posteriores análisis.

En la Figura 8 se presenta la distribución de tipos de propiedades en el conjunto de datos original. Inicialmente se cuenta con registros de tipo de propiedades correspondiente a casas, departamentos, lotes, oficinas, locales y depósitos. Para el presente estudio se pretende analizar la vivienda y por tanto se seleccionaron únicamente los datos correspondientes a tipo de propiedad vivienda y casas.



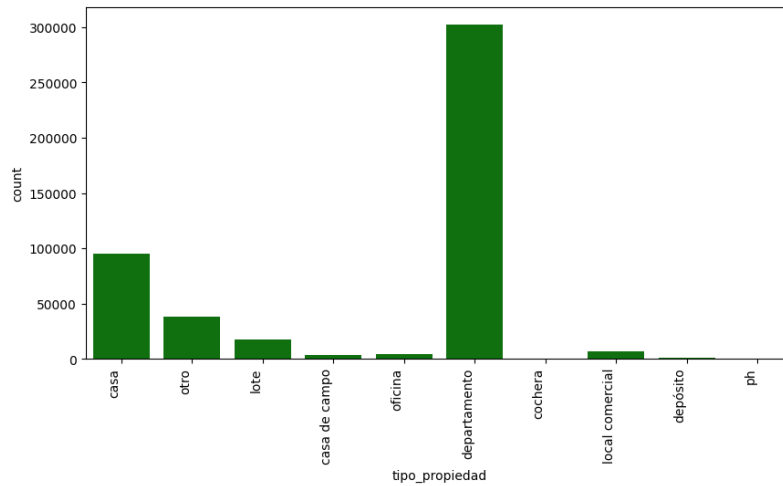


Figura 8. Distribución de tipo de propiedad en el conjunto de datos original (Elaboración propia)

### Agregar columnas Adicionales

Se crea una nueva variable para estimar el precio por m<sup>2</sup> de superficie de cada vivienda, denominada de 'price\_m2', la cual consiste en la división entre el precio (price) y la superficie total (surface\_total).

### Renombre de las Columnas

Se renombraron las variables del conjunto de datos por nombres en español y, se dividieron las variables en los siguientes grupos para su tratamiento y análisis exploratorio: Precio de venta, Localización, Características y, por último, Información Adicional, como se observa en la Tabla 9.

Tabla 9. Variables del conjunto de datos por grupos

Precio de venta	Localización	Características	Información Adicional
precio	pais	tipo_propiedad	descripción
moneda	departamento	habitaciones	título
precio_m2	municipio	baños	id
fecha_publicacion	barrio	superficie total	
	latitud	superficie_cubierta	
	longitud		

Fuente: Elaboración propia

A continuación, en la Tabla 10 se presentan las variables del conjunto de datos al finalizar el proceso de limpieza y preparación, se incluye el grupo, el nombre de la variable, la descripción y el tipo de dato.

Tabla 10. Variables del conjunto de datos modificado

Grupo	Variable	Descripción	Tipo de dato
<b>Precio de Venta</b>	precio	Precio de venta del inmueble en pesos Colombianos COP	float64
	precio_por_m2	Precio del inmueble sobre la superficie total (precio/supecficie_total)	float64
	moneda	Moneda del pecio: Peso Colombiano	category
	fecha_publicacion	Fecha de creación de la publicación.	datetime
<b>Localización</b>	pais	País de ubicación del inmueble: Colombia	category
	departamento	Departamento de ubicación del inmueble: Antioquia.	category
	municipio	Municipio de ubicación del inmueble	category
	barrio	Barrio de ubicación del inmueble	category
	latitud	Latitud de la localización del inmueble	float64
	longitud	longitud de la localización del inmueble	float64
<b>Características</b>	tipo_propiedad	Tipo de inmueble: casa, apartamento	category
	habitaciones	Cantidad de habitaciones del inmueble	Int64
	baños	Cantidad de baños del inmueble	Int64
	supecficie_total	Superficie total del inmueble en m <sup>2</sup>	float64
	supecficie_cubierta	Superficie cubierta del inmueble en m <sup>2</sup>	float64
<b>Información Adicional</b>	descripcion	Descripción del inmueble	category
	titulo	Título de la publicación de venta del inmueble	category
	id	Código único de cada inmueble publicado	object

Fuente: Elaboración propia

Al finalizar el proceso de limpieza y preparación de los datos, el conjunto de datos modificado esta conformado por 108.823 registros y 18 columnas explicativas. Esto se debe a que gran parte de ellos datos originales, exactamente el 76%, no presenta registros de la variable superficie total, y son omitidos en el análisis debido a su importancia en la determinación del precio del inmueble. Adicionalmente el conjunto de datos original presentaba registros de otros municipios del departamento de Antioquia, y otros tipos de propiedades.

## 4.3. Análisis exploratorio de los datos

En esta fase se realizará la exploración de los datos a través de tablas, gráficos y herramientas de visualización. El objetivo de esta fase es comprender la distribución de las diferentes variables del conjunto de datos, identificar valores atípicos y analizar las relaciones de los diferentes parámetros con el precio de los inmuebles. Este análisis se realizó a partir de los grupos de variables definidos previamente: Precio de venta, características, localización, información adicional.

### 4.3.1. Precio de Venta

En primer lugar, es fundamental analizar el comportamiento de la variable objetivo a predecir, en este caso la variable precio. En el análisis exploratorio de datos para la variable precio, se pretende responder las siguientes preguntas:

- ¿Cómo es la distribución de la variable precio y del precio por m<sup>2</sup>?
- ¿Cuál es el precio total y precio por m<sup>2</sup> de las casas y departamentos ofrecidos en Medellín y el Área Metropolitana?
- ¿Cómo ha sido la evolución en el tiempo de los precios de venta de las propiedades y el precio por m<sup>2</sup>?

A continuación, se realiza la exploración de la variable precio a través de diferentes gráficos, para dar respuesta a las preguntas planteadas.

#### Distribución de las variables precio y precio por m<sup>2</sup>

La Figura 9, muestra a través de un gráfico de histograma, la distribución de la variable objetivo precio, donde se adiciona una línea de color azul para indicar la media y otra línea de color rojo que indica la mediana. En este gráfico se observa una forma de distribución asimétrica para la variable precio, con un sesgo positivo o hacia la derecha, el cual se puede confirmar al observar que la media presenta un valor mayor que la mediana. Este comportamiento se debe a que los datos de precios inmobiliarios generalmente tienden a presentar observaciones con precios extremos.

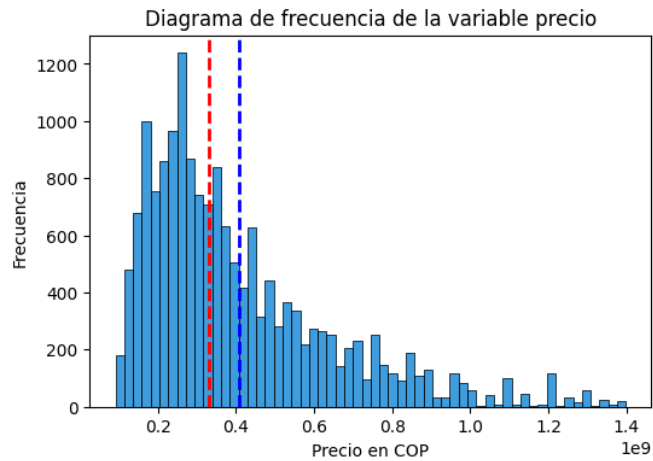


Figura 9. *Distribución de la variable precio* (Elaboración propia)

La Figura 10, presenta el histograma de la variable precio por m<sup>2</sup>, el cual presenta una distribución asimétrica menos pronunciada que la variable precio.

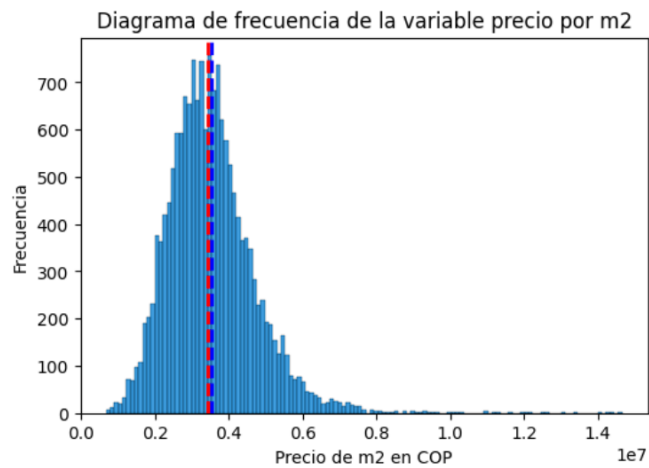


Figura 10. *Distribución de la variable precio m<sup>2</sup>* (Elaboración propia)

La Tabla 11 presenta el resumen con las estadísticas descriptivas de la variable objetivo precio y precio por m<sup>2</sup>. A partir de estos resultados se observa que el precio de los inmuebles oscila entre 91 y 1.399 millones de pesos colombianos, con una media de 407 millones de pesos y una mediana de 330 millones. Debido a la asimetría de la distribución del precio, generalmente se emplea la mediana como medida de tendencia central en lugar de la media.

Respecto al precio por m<sup>2</sup> de los inmuebles, en la Tabla 11 se observa que presenta un rango entre 673 mil hasta 15 millones de pesos, con una media de 3.5 millones de pesos por m<sup>2</sup> de superficie y una mediana de 3.4 millones de pesos por m<sup>2</sup> de superficie.

Tabla 11. Estadísticas descriptivas de las variables precio y precio por m<sup>2</sup>

Parámetro	Precio	Precio por m <sup>2</sup>
N válido	16.759	16.759
Perdidos	\$ -	\$ -
Media	\$ 407.219.300	\$ 3.521.997
Mediana	\$ 330.000.000	\$ 3.413.654
Moda	\$ 350.000.000	\$ 3.000.000
Rango	\$ 1.308.000.000	\$ 13.977.401
Mínimo	\$ 91.000.000	\$ 673.758
Máximo	\$ 1.399.000.000	\$ 14.651.160
Percentil 25	\$ 230.000.000	\$ 2.727.273
Percentil 75	\$ 520.000.000	\$ 4.139.535
Percentil 90	\$ 750.000.000	\$ 4.980.016
Percentil 95	\$ 900.000.000	\$ 5.530.464
Percentil 99	\$ 1.211.400.000	\$ 7.090.122

Fuente: Elaboración propia

### Evolución de las variables precio y precio por m<sup>2</sup>

Con el fin de identificar posibles patrones en el comportamiento de los precios de venta a lo largo del periodo de estudio, se realiza una gráfica de la evolución en el tiempo de la variable precio. Se utiliza la variable fecha\_publicacion con la cual se agrupa la información por año y mes en el eje x. En el eje vertical, se grafica los promedios mensuales de los precios, de esta forma la Figura 11 presenta la evolución mensual de los promedios del precio de venta de los inmuebles, para el periodo comprendido entre mayo del 2015 y mayo del 2021.

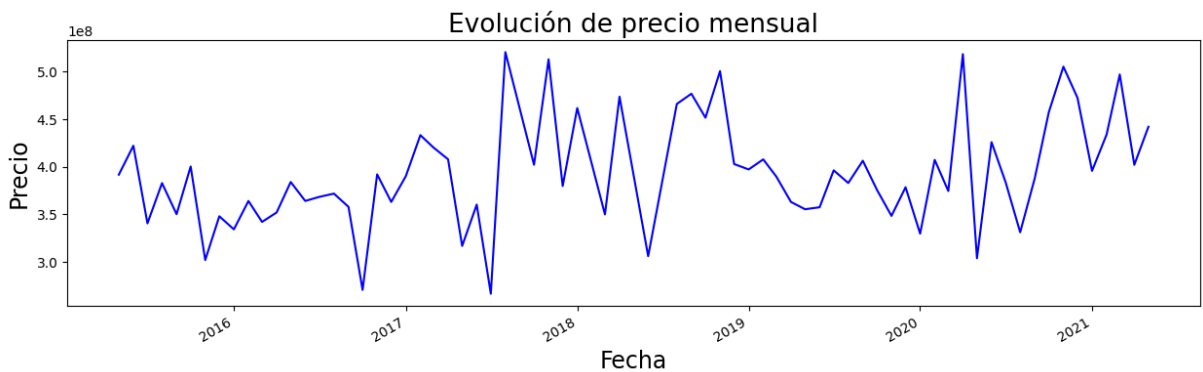


Figura 11. Evolución de los precios de venta (Elaboración propia)

Los menores registros mensuales se han presentado en octubre de 2016 y julio del 2017, mientras hay diferentes registros con valores altos a lo largo del tiempo, demostrando la heterogeneidad de los precios en el periodo de estudio. Se observa una posible tendencia de los precios a disminuir en los meses de julio a octubre, sin embargo, no es un patrón bien definido.

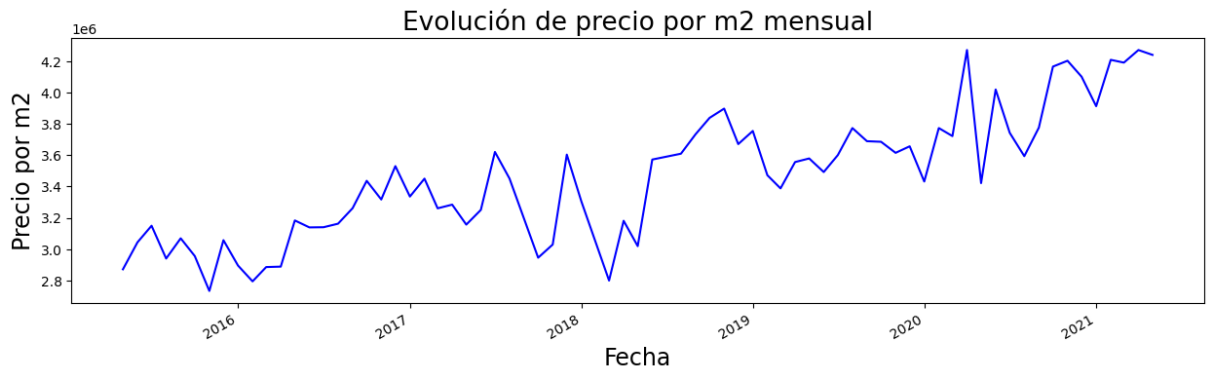


Figura 12. *Evolución de los precios por m<sup>2</sup>* (Elaboración propia)

En la Figura 12 se observa la evolución mensual del precio por m<sup>2</sup> promedio de los inmuebles, evidenciando una tendencia positiva representada en el aumento de estos precios. El menor valor mensual fue registrado en noviembre de 2015 y el mayor registro mensual es de abril de 2020, seguido de abril de 2021.

### 4.3.2. Localización

La localización geográfica de un inmueble es generalmente uno de los factores más influyentes en el precio. En este análisis se pretende responder las siguientes preguntas.

- ¿Cómo están distribuidos los inmuebles del conjunto de datos en términos de municipios y barrios?
- ¿Cómo es la distribución geográfica de precios de las propiedades ofertadas en la ciudad de Medellín y el Área Metropolitana?
- ¿Cuál es el precio total y precio por m<sup>2</sup> de los inmuebles según su ubicación en términos de municipio o barrio?

#### Distribución de la localización de los inmuebles

Para comprender la localización del inmueble en el conjunto de datos, se presentan los siguientes gráficos de barras. La Figura 13, presenta la distribución de datos por su correspondiente municipio, siendo el municipio con mayores datos Medellín seguido de Envigado y, por último, el municipio con menos registros de inmuebles es Barbosa.

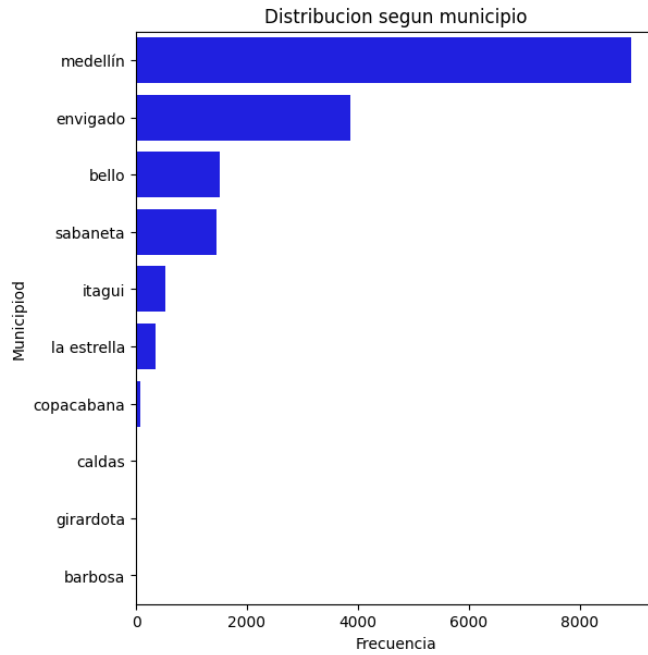


Figura 13. Distribución de los datos según municipio (Elaboración propia)

### Distribución geográfica de precios

A continuación se presenta la distribución de precios de vivienda en la ciudad de Medellín y el Área Metropolitana, con el fin de visualizar cómo los precios son afectados por la localización del inmueble e identificar patrones en los mismos. En la Figura 14 se presenta un mapa de calor con la distribución de precios de vivienda en la ciudad de Medellín y el Área Metropolitana a partir del conjunto de datos a analizar. El mapa de calor fue elaborado sobre el mapa de la herramienta Open Street Map y a partir de este, se evidencia un incremento de precios en el sector sur del Área Metropolitana, específicamente en los municipios de Sabaneta y Envigado y el centro de la ciudad de Medellín.

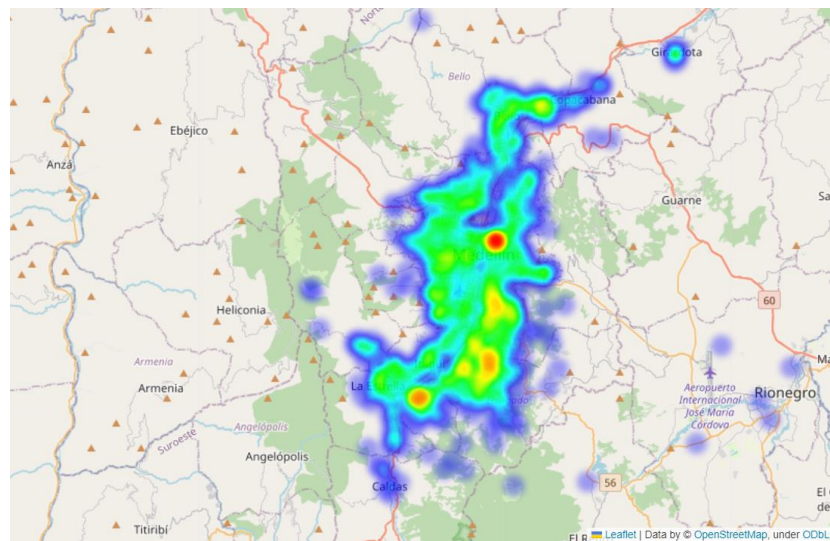


Figura 14. Mapa de Calor con Distribución de Precios (Elaboración propia)

### Distribución de precios por municipios

A continuación, se presenta en la Figura 15 el diagrama de caja o boxplot con las medidas de tendencia central con condiciones de dispersión para la variable precio según el municipio de ubicación. Respecto al precio, el municipio de Medellín presenta los mayores precios promedios con una media de aproximadamente 457 millones de pesos colombianos, seguido de Envigado y Sabaneta. Los menores valores se presentan en el municipio de Barbosa, esto también se debe a que es el municipio con menos registros, presentando además menor variación en los valores.

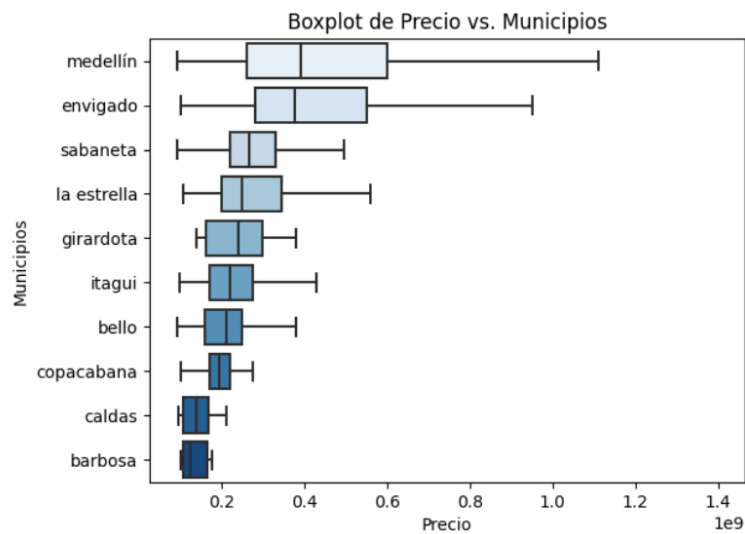


Figura 15. *Boxplot distribución de precios por municipio* (Elaboración propia)

La Figura 16 presenta el diagrama de caja para la variable precio según el municipio de ubicación y el tipo de propiedad, en el cual se observa como las casas presentan mayor rango entre los valores, y mayor o menor precio promedio según el municipio de ubicación. Medellín continúa siendo el municipio con mayores precios promedios de casas y departamentos.

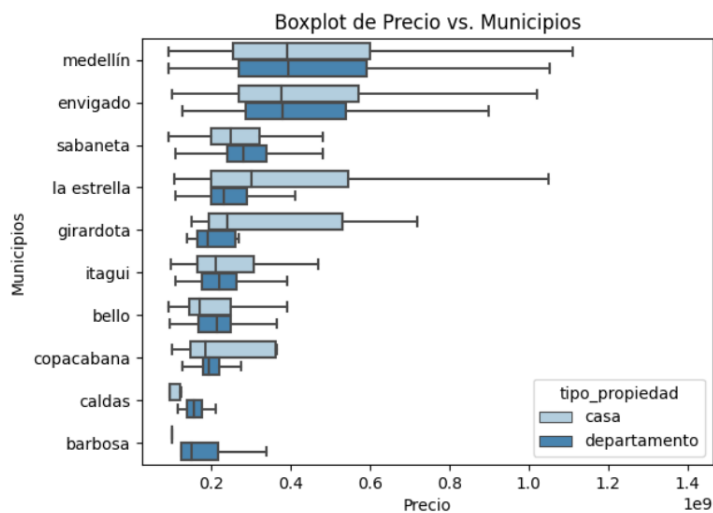


Figura 16. *Boxplot distribución de precios por municipio según tipo de propiedad* (Elaboración propia)



En la Figura 17 se presenta el diagrama de caja para la variable precio por m<sup>2</sup> según el municipio de ubicación. Respecto al precio por m<sup>2</sup>, el municipio de Envigado presenta los mayores precios promedio con una media de aproximadamente 3.7 millones de pesos colombianos por m<sup>2</sup> de superficie.

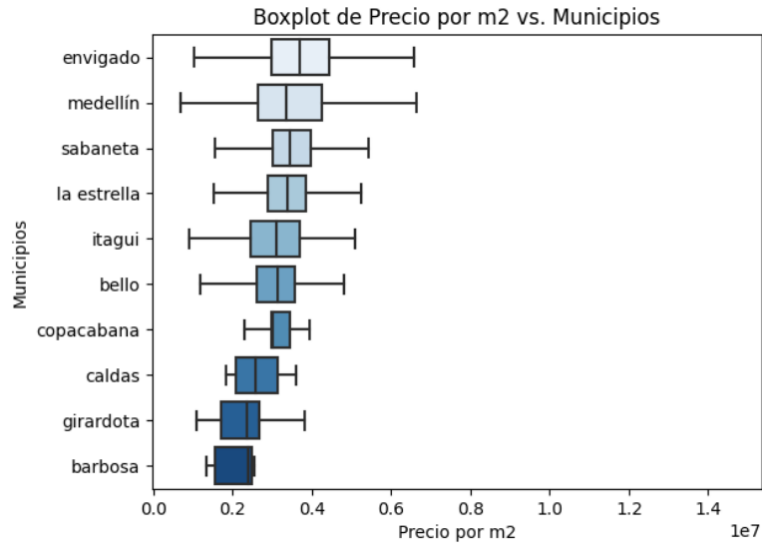


Figura 17. *Boxplot distribución de precios m<sup>2</sup> por municipio* (Elaboración propia)

La Figura 18 presenta el diagrama de caja para la variable precio por m<sup>2</sup> de superficie según el municipio de ubicación y el tipo de propiedad. Los mayores precios por m<sup>2</sup> para las casas y departamentos se presentan en el municipio de Envigado, adicionalmente, los departamentos presentan mayor precio por m<sup>2</sup> en comparación con las casas, en la totalidad de municipios analizados, esto generalmente ocurre ya que los apartamentos presentan menores áreas de superficie, incrementando el valor por m<sup>2</sup>.

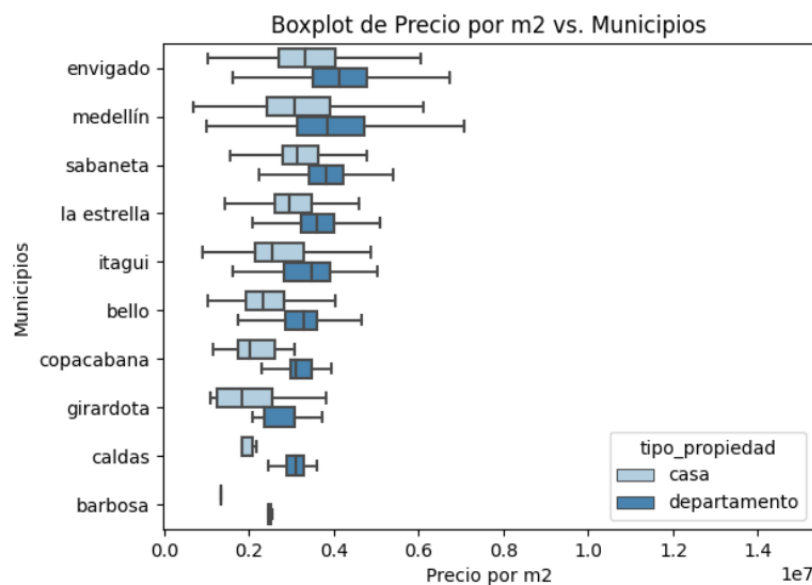


Figura 18. *Boxplot distribución de precios m<sup>2</sup> por municipio según tipo de propiedad* (Elaboración propia)

### 4.3.3. Características

Los bienes inmuebles son un producto altamente heterogéneo, y su precio está determinado por un alto número de variables, asociadas a las características de las propiedades. En el presente análisis exploratorio se pretende responder:

- ¿Cuál es la distribución de las principales características de las propiedades ofrecidas en Medellín y el Área Metropolitana?
- ¿Qué factores del inmueble ayudan a explicar el precio de las propiedades ofrecidas para la venta en Medellín y el Área Metropolitana?

#### Distribución de las variables del grupo características

A continuación, se presenta la distribución de las siguientes variables incluidas en el grupo de características del conjunto de datos: habitaciones, baños, superficie total y tipo de propiedad.

- Número de habitaciones y número de baños

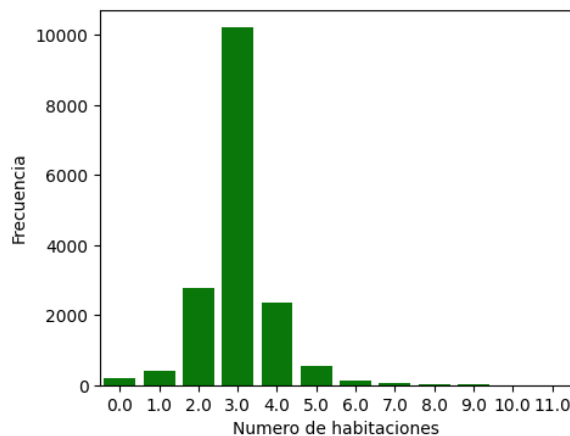


Figura 19. Distribución de la variable número de habitaciones (Elaboración propia)

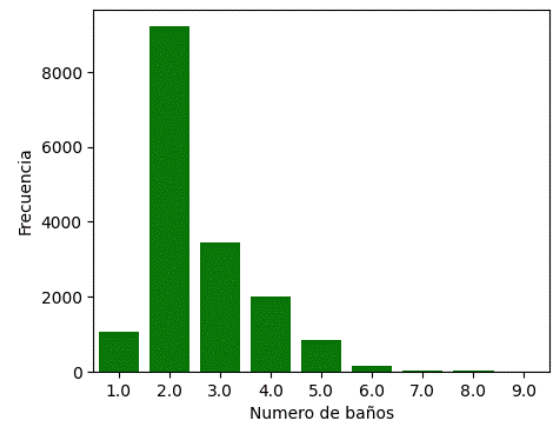


Figura 20. Distribución de la variable número de baños (Elaboración propia)

A partir de las gráficas anteriores, se observa en la Figura 19 que la mayoría de las viviendas presenta 3 habitaciones, además el conjunto de datos incluye viviendas con cero habitaciones, correspondiente a estudios y, viviendas con 12 habitaciones.

A partir de la Figura 20, se observa que la mayoría de los inmuebles presenta 2 baños, siendo el valor mínimo 1 y el valor máximo 9 baños.

- Superficie total y Tipo de propiedad

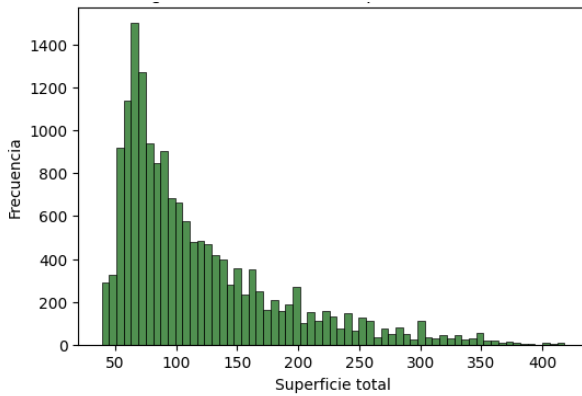


Figura 21. Distribución de la variable superficie total (Elaboración propia)

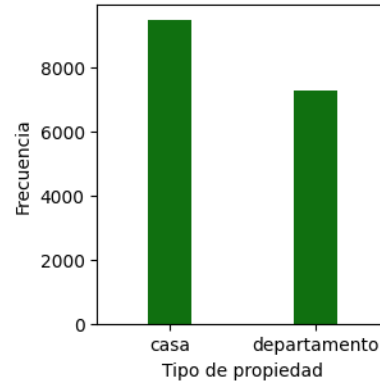


Figura 22. Distribución de la variable tipo de propiedad (Elaboración propia)

La superficie total, es una variable numérica que presenta distribución asimétrica como se observa en la Figura 21. En esta distribución se observa un sesgo positivo o hacia la derecha, con un valor promedio de 118 m<sup>2</sup> de superficie total, un valor mínimo de 39 m<sup>2</sup> y un valor máximo de 418 m<sup>2</sup>.

Por último, para la variable categórica tipo de propiedad, se evidencia en la Figura 22 que el presente dataset contiene valores correspondientes a casas y departamentos. Después del proceso de limpieza y preparación, el mayor número de registros corresponde a casas, esto se debe a que la mayoría de registros corresponden a la ciudad de Medellín, donde las casas son predominantes.

**Relación de las variables del grupo características con el precio**

Para analizar el impacto de las características anteriores en la variable objetivo precio, se presenta los gráficos de caja para las variables de tipo categóricas, en los cuales es posible identificar el valor medio del precio, los cuartiles, valor mínimo y máximo, así como los outliers o datos atípicos, los cuales se encuentran alejados por más de 3 puntos de desviación estándar

En la Figura 23 se presenta el diagrama de cajas para los precios según el número de habitaciones. En general, se observa como la media de los precios es menor al contar con menos habitaciones, a excepción de los estudios que presentan un promedio de precios superior pese a tener cero habitaciones

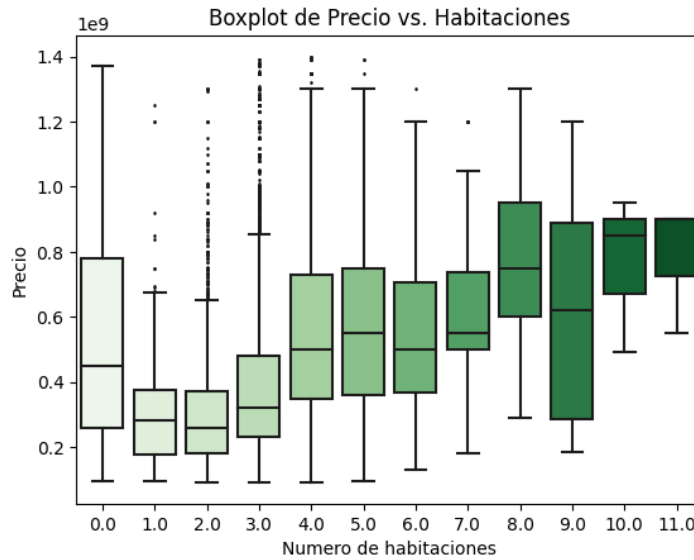


Figura 23. *Boxplot de precio y número de Habitaciones* (Elaboración propia)

En la Figura 24 se presenta el diagrama de cajas para los precios según el número de baños. Se observa que la media de precio de los inmuebles se incrementa cuando el número de baños es mayor.

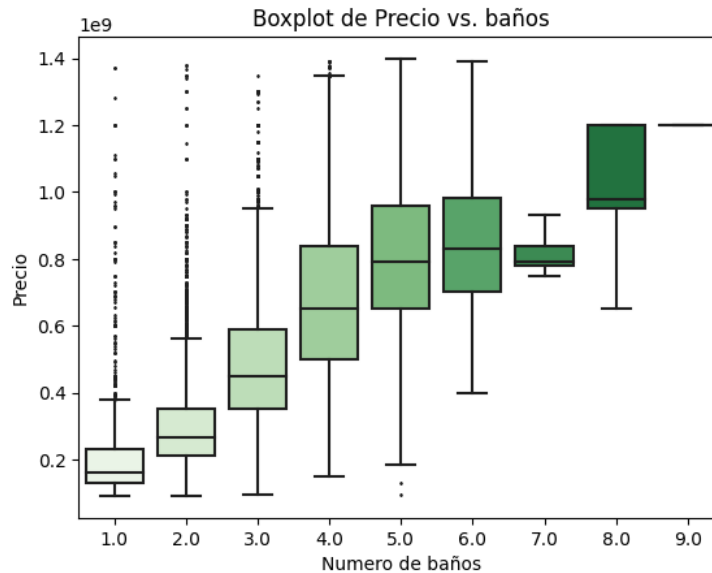


Figura 24. *Boxplot de precio y número de baños* (Elaboración propia)

En la Figura 25 se muestra la distribución de los precios según el tipo de propiedad, utilizando un diagrama de caja. En este gráfico se observa que los departamentos tienden a ser en promedio menos costosos que las casas, sin embargo las casas presentan precios máximos mayores. En la Figura 26 se muestra la distribución de los precios por m<sup>2</sup> según el tipo de

propiedad, en el cual se evidencia que los departamentos presentan mayor valor por m<sup>2</sup> de superficie promedio que las casas.

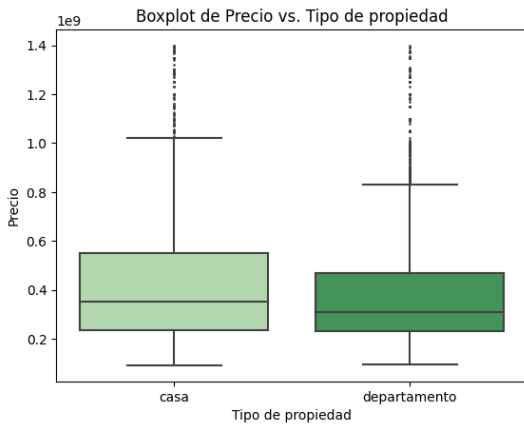


Figura 25. *Boxplot de precio y Tipo de propiedad* (Elaboración propia)

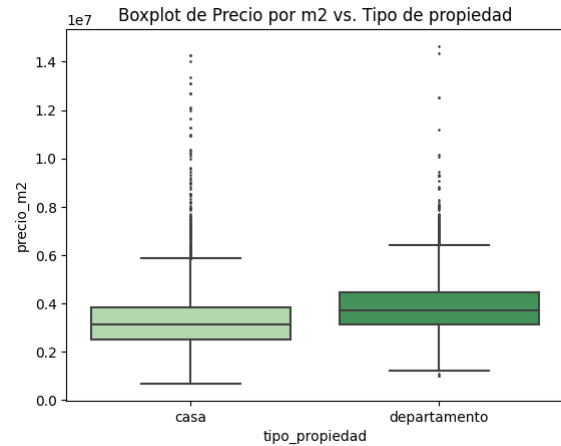


Figura 26. *Boxplot de precio por m<sup>2</sup> y Tipo de propiedad* (Elaboración propia)

### 4.3.4. Matriz de correlaciones

A continuación se pretende analizar cuáles atributos están más correlacionados con el precio, para ello se incluye la matriz de correlaciones entre las diferentes variables.

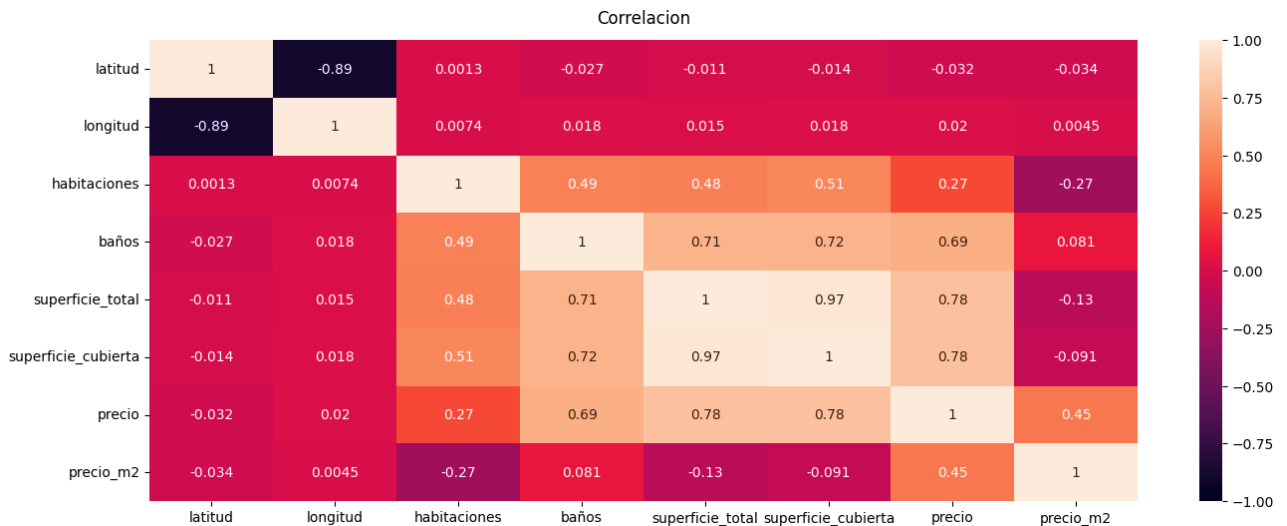


Figura 27. *Matriz de coeficiente de Correlación entre las variables* (Elaboración propia)

De la matriz de correlaciones de la Figura 27, se hace énfasis en el coeficiente de correlación de las diferentes variables frente al precio. A partir de esto, se obtiene que las variables con mayor correlación positiva son la superficie total y la superficie cubierta, ambas con una correlación de 0,78; seguido de la variable baños con una correlación de 0,69 y el número de

habitaciones con un coeficiente de 0,27. La única variable con correlación negativa frente al precio es la latitud, con una correlación de -0,034.

Para analizar con mayor profundidad la relación entre las variables en el conjunto de datos, a continuación, se presenta en la Figura 28 un análisis visual con gráficos de dispersión con la relación entre las variables con mayor correlación frente al precio, diferenciando entre el tipo de propiedad casa o departamento.

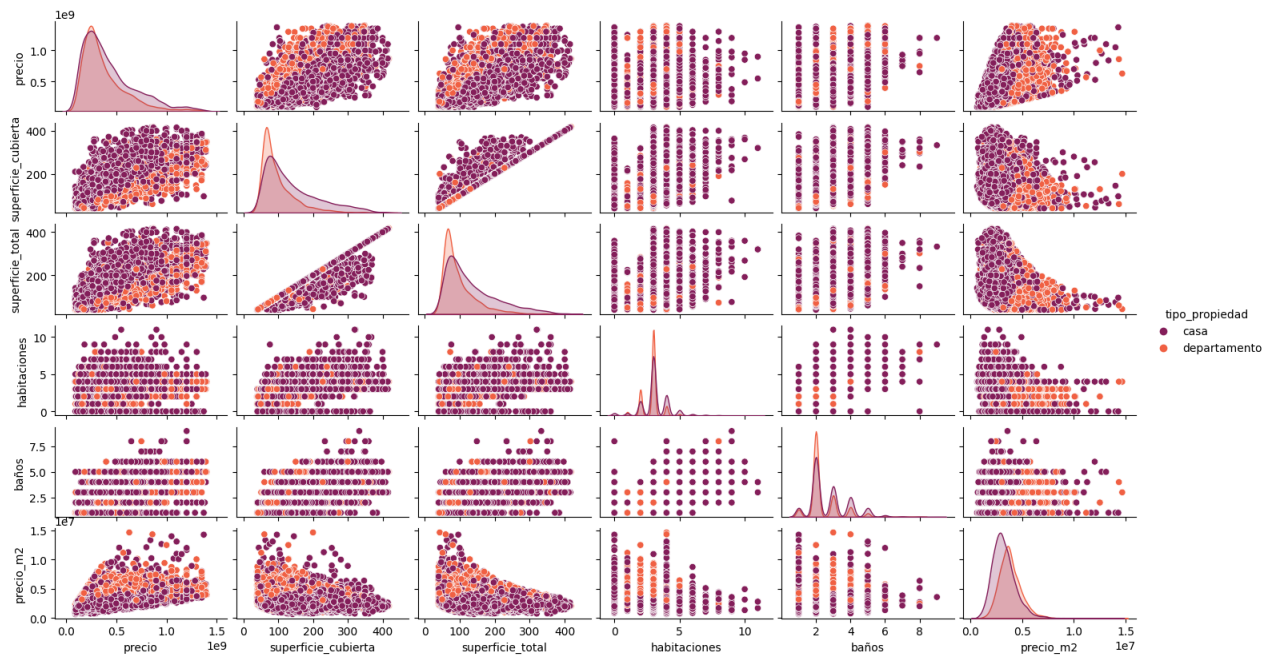


Figura 28. *Relación entre las variables* (Elaboración propia)

Del gráfico anterior, se destaca una posible linealidad entre las variables precio y superficie total, sin embargo se presenta significativa aleatoriedad en la distribución de las variables. Se evidencia también que las viviendas con mayores baños tienen a presentar además mayores habitaciones. En el gráfico también es posible visualizar una tendencia negativa entre la relación del precio por m<sup>2</sup> y la superficie total, en el cual, el precio por m<sup>2</sup> disminuye al contar con áreas más grandes de propiedades.

## 4.4. Transformaciones y selección de variables

En esta fase se realizan las transformaciones correspondientes en el conjunto de datos con el objetivo de preparar y estructurar los mismos para la ejecución de los modelos. A continuación, se enuncian algunas de las tareas de la fase de transformación de los datos.

### Selección de las variables para el modelo

- **Subconjunto de datos 1:** Se selecciona un conjunto de datos inicial con las variables numéricas: latitud, longitud, superficie total, número de habitaciones y baños, como variables de entrada, y la variable precio como variable objetivo.
- **Subconjunto de datos 2:** Se selecciona un segundo conjunto de datos para modelar con algunas variables categóricas adicionales, correspondientes a municipio, barrio y tipo de propiedad, adicionales a las variables anteriores, latitud, longitud, superficie total, número de habitaciones y baños, como variables de entrada, y la variable precio como variable objetivo.
- **Subconjunto de datos 3:** Se presenta un tercer conjunto de datos similar al segundo, sin las variables municipio, tipo de propiedad, latitud, longitud, superficie total, número de habitaciones y número de baños, como variables de entrada, y la variable precio como variable objetivo.

A continuación, se describen las razones por las cuales otras variables no fueron consideradas para la fase de modelado. Variables como título y descripción no fueron consideradas, debido a la necesidad de aplicar técnicas de procesamiento de lenguaje natural para estructurar la información contenida en estas variables.

La variable superficie cubierta no fue considerada debido al alto número de datos perdidos de esta variable y su relación directa con la superficie total, siendo redundante para los modelos. Otras variables categóricas como id, fecha\_publicacion, fueron eliminadas al no aportar información adicional sobre los inmuebles. Respecto a la localización, las variables país y departamento contenían datos únicos del país Colombia y departamento Antioquia, por lo tanto se pudieron omitir en la siguiente fase.

### Transformaciones previas al modelado

Se transforman las variables categóricas tipo de propiedad, municipio y barrio a variables de tipo numéricas, utilizando la opción `get_dummies` de la librería Pandas.

Se estandarizan los datos mediante la función `StandardScaler`.

## 4.5. Desarrollo de los Modelos

En la fase de modelado se realizan una serie de actividades orientadas a seleccionar, configurar, entrenar y probar los modelos de machine learning a implementar para dar solución a la problemática de predicción de precios de vivienda en la ciudad de Medellín y el Área Metropolitana.

### 4.5.1. Selección de los algoritmos

Para la selección de los algoritmos de aprendizaje automático, se tuvo en consideración los antecedentes y estudios previos analizados en el presente documento, en los cuales gran número de investigadores han dedicado sus esfuerzos a determinar el algoritmo más adecuado para el problema de la predicción de precios de vivienda en diferentes ciudades alrededor del mundo y en Colombia.

Debido a la naturaleza del problema, en el cual la variable objetivo precio es de tipo numérica, las técnicas más apropiadas para abordar la problemática, son los modelos de regresión desde el paradigma de aprendizaje supervisado. De acuerdo con los estudios previos, entre los algoritmos más populares para abordar esta problemática, se encuentran, la Regresión lineal, árboles de decisión, Random Forest, Gradient Boosting, Extreme Gradient Boosting, K vecinos más cercanos y las redes neuronales artificiales.

El algoritmo con los resultados más prometedores es el Random Forest, el cual ha presentado mejores predicciones en los estudios previos analizados. Además, los investigadores anteriores también recomiendan considerar este modelo en futuras investigaciones. También se puede observar algunos estudios, donde algoritmos como k-vecinos más cercanos XGBoost han presentado mejores resultados.

De acuerdo con los resultados presentados en los estudios previos, se seleccionan cinco algoritmos para ser implementados en el presente informe, estos son: Regresión Lineal, Árboles de decisión, Nearest Neighbor o k vecinos más cercanos, Random Forest, y XGBoost

Posteriormente a la implementación de estos modelos, se realizará una comparación entre los diferentes métodos para finalmente proponer el modelo con mejores resultados a la problemática.



### 4.5.2. Implementación de los algoritmos

Se realiza la implementación de los modelos utilizando el lenguaje de programación Python, incluyendo diferentes librerías para la construcción de los modelos de machine learning.

Inicialmente se elimina la variable objetivo precio en los conjuntos de datos de entrenamiento y pruebas. Se definen los porcentajes sobre el conjunto de datos para las fases entrenamiento y prueba del modelo: 80% los datos entrenamiento y 20% los datos de prueba. Posteriormente, se implementan los modelos de machine learning de acuerdo con su complejidad, iniciando con un modelo de regresión lineal, seguido del modelo de árboles de decisión, k vecinos más cercanos, Random Forest y por último extreme gradient boosting machine.

Inicialmente se implementan los modelos con una configuración inicial utilizando los hiperparámetros. Posteriormente, se evalúan los resultados de los modelos respecto a los datos del conjunto de pruebas. En general el subconjunto de pruebas 2 presenta mejores resultados iniciales y por tanto se presenta la configuración de hiperparámetros en este subconjunto de datos. Los resultados de todas las iteraciones para los tres subconjuntos de datos se presentan en el Anexo II.

### 4.5.3. Configuración de hiperparámetros

A partir de los modelos iniciales, se requiere realizar diferentes configuraciones de los modelos candidatos, para lo cual se modifican los hiperparámetros, que corresponden a los valores de las configuraciones que se utilizan en cada modelo durante la fase de entrenamiento.

Adicionalmente, en esta fase de configuración de hiperparámetros se analizan los resultados con el fin de obtener una predicción óptima, evitando el overfitting o sobreajuste y el underfitting de los modelos. El primero se presenta cuando el modelo se encuentra muy ajustado a los datos, y es incapaz de generalizar, se identifica con una alta precisión en los datos de entrenamiento y bajo rendimiento en los datos de prueba. El segundo, se presenta cuando el modelo es muy simple y no es posible predecir correctamente a partir de los datos, en este caso se trata de bajo rendimiento en los datos de entrenamiento.

A continuación, se presentan los hiperparámetros más importantes utilizados en los diferentes modelos, para la selección y configuración de los hiperparámetros se realizaron diferentes iteraciones considerando el  $R^2$  y RMSE como métricas de evaluación.

### Árboles de decisión

El hiperparámetro más importante en este modelo corresponde a `max_depth` o la profundidad máxima del árbol de decisión, el cual se utiliza para controlar el sobreajuste. Aumentar este valor hace que el modelo sea más complejo, lo que hace más posible el sobreajuste.

En la Figura 29 y la Figura 30 se pueden observar los resultados al cambiar `max_depth`. Los gráficos muestran una correlación entre el aumento en la profundidad del árbol y un mejor resultado para los datos de entrenamiento, sin embargo árboles con profundidad mayor a 7, tienden a él sobreajuste. Dado que utilizar más árboles también sugiere mayores recursos computacionales, el número de `max_depth` 7

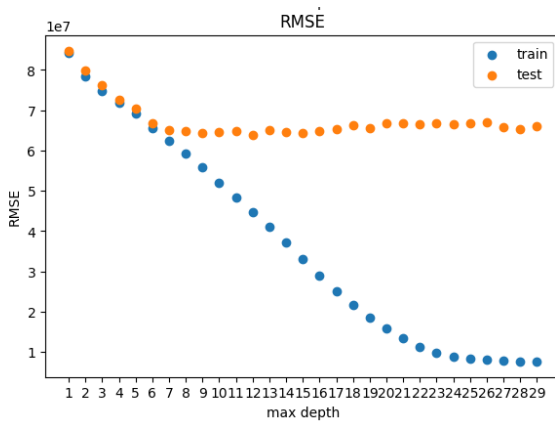


Figura 29. Métricas de error RMSE y `max_depth` para Árbol de decisión (Elaboración propia)

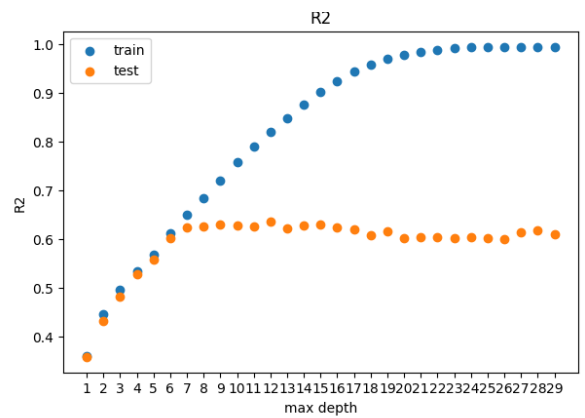


Figura 30. Métricas de error  $R^2$  y `max_depth` para Árbol de decisión (Elaboración propia)

### Random Forest

Los hiperparámetros considerados para este modelo, corresponden a `n_estimators` y `max_features`; `n_estimators` es el número de árboles de decisión en el Random Forest. Con un aumento en la cantidad de árboles, la precisión del modelo aumenta y se reduce el sobreajuste. Sin embargo, esto hará que el modelo sea más lento; por lo tanto, elegir un valor de `n_estimators` que el procesador pueda manejar permite que el modelo sea más estable y funcione bien. El valor por defecto es 100.

En las Figuras 31 y 32 se pueden observar los resultados al cambiar `n_estimators`. El gráfico muestra una correlación entre un aumento en el número de árboles y un mejor resultado. Dado que utilizar más árboles también sugiere mayores recursos computacionales, el número de árboles que se decide utilizar es de 300

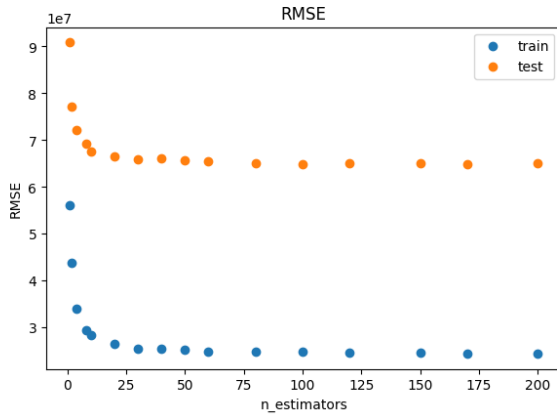


Figura 31. Métricas de error RMSE y  $n\_estimators$  para Random Forest (Elaboración propia)

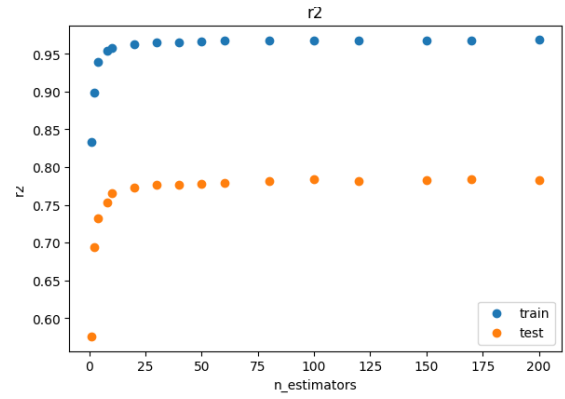


Figura 32. Métricas de error  $R^2$  y  $n\_estimators$  para Random Forest (Elaboración propia)

Un modelo de bosque aleatorio solo puede tener un número máximo de características o  $max\_features$  en un árbol individual. Por lo tanto, encontrar un  $max\_features$  óptimo es importante para el rendimiento del modelo. La Figuras 33 y la Figura 34 muestran los resultados de las métricas cuando la variable  $max\_features$  cambia. No se observa un incremento significativo en los resultados del modelo con este hiperparámetro y se decide utilizar un  $max\_features = 5$ .

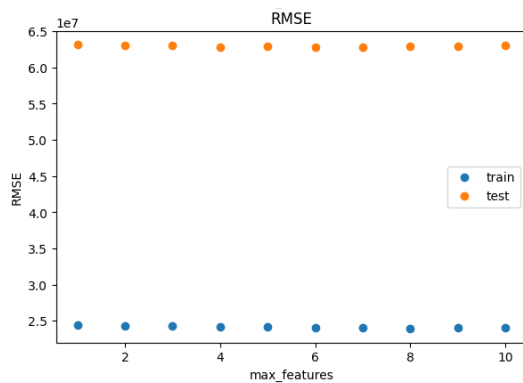


Figura 33. Métricas de error RMSE y  $max\_features$  para Random Forest (Elaboración propia)

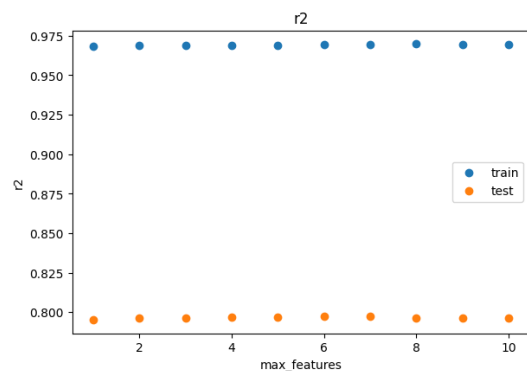


Figura 34. Métricas de error  $R^2$  y  $max\_features$  para Random Forest (Elaboración propia)

### K vecino más cercano para la regresión

A continuación, se presentan los parámetros utilizados para el algoritmo K Neighbors, a partir de cómo afectan el modelo en términos de sobreajuste y desajuste.  $k\_neighbors$  representa el número de vecinos que se usarán para las consultas de  $kneighbors$ , de esta forma, utilizar

un  $k\_neighbors=1$  significa que cada muestra se usa a sí misma como referencia, ese es un caso de sobreajuste.

La Figura 35 y la Figura 36 muestran el comportamiento de las métricas de error con diferentes valores del número de  $k$ , para este caso aumentar el número de vecinos mejora los puntajes de las pruebas y se seleccionó un  $k\_neighbors = 10$ .

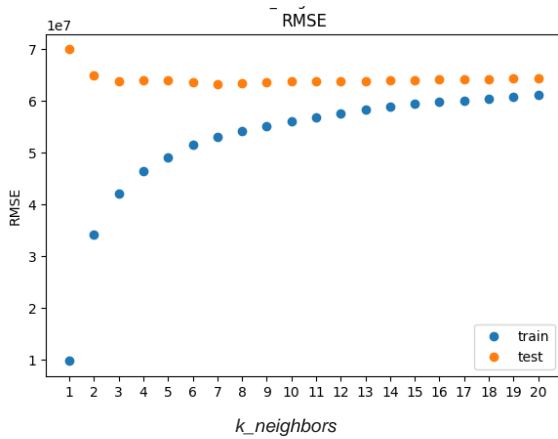


Figura 35. Métricas de error RMSE y  $k\_neighbors$  para  $k$  vecinos más cercanos (Elaboración propia)

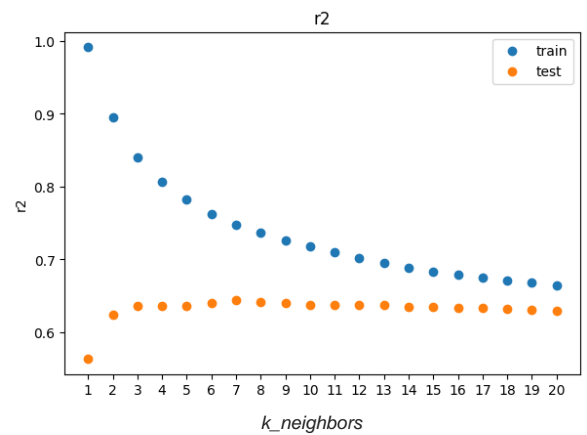


Figura 36. Métricas de error para y  $k\_neighbors$  para  $k$  vecinos más cercanos (Elaboración propia)

### XGBoost para la regresión

Para este modelo se analizan los hiperparámetros  $n\_estimators$  o número de árboles, la tasa de aprendizaje  $eta$  y la profundidad máxima del árbol  $max\_deph$

$n\_estimators$ , determina cuántos árboles de decisión se construirán. Si  $n\_estimator$  se establece igual a 1, solo se creará un único árbol de decisión. El valor predeterminado para  $n\_estimators$  es 100 y debe ser un número entero mayor que 0. Cuanto mayor sea el valor que tome para  $n\_estimators$ , más preciso será el rendimiento del modelo, pero se necesita más tiempo para entrenar el modelo y es posible que se ajuste demasiado a los datos de entrenamiento.

En las Figuras 37 y 38 se pueden observar los resultados al cambiar  $n\_estimators$ . A partir de lo anterior para el presente modelo se utilizó un  $n\_estimators$  de 1000.

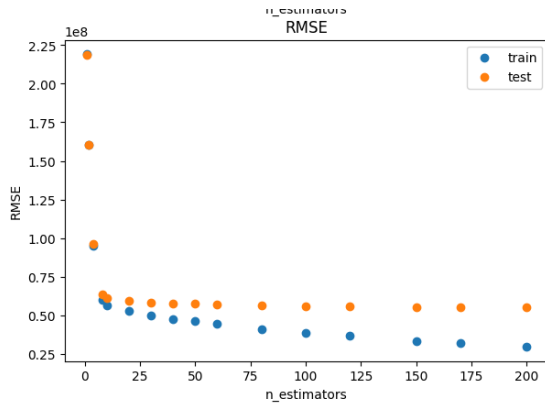


Figura 37. Métricas de error RMSE y  $n\_estimators$  para XGBoost (Elaboración propia)

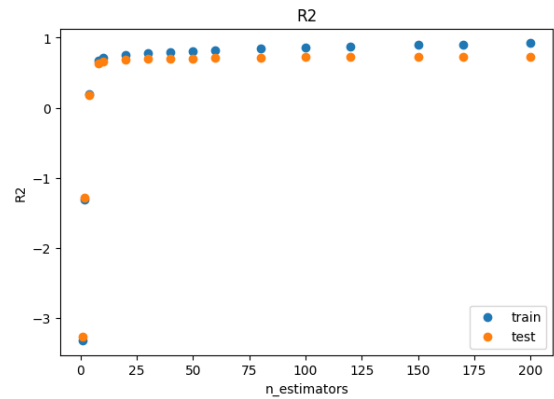


Figura 38. Métricas de error  $R^2$  y  $n\_estimators$  para XGBoost (Elaboración propia)

La tasa de aprendizaje (eta), determina qué tan rápido aprende el modelo XGBoost. Una tasa de aprendizaje baja hace que el cálculo sea más lento y requiere más árboles para lograr la misma reducción en el error residual, que un modelo con una tasa de aprendizaje alta; sin embargo, optimiza las posibilidades de alcanzar el mejor modelo. La tasa de aprendizaje presenta un valor predeterminado de 0,3; para el presente modelo se utilizó una tasa de aprendizaje del 0,05. En las Figuras 39 y 40 se pueden observar los resultados al cambiar la tasa de aprendizaje.

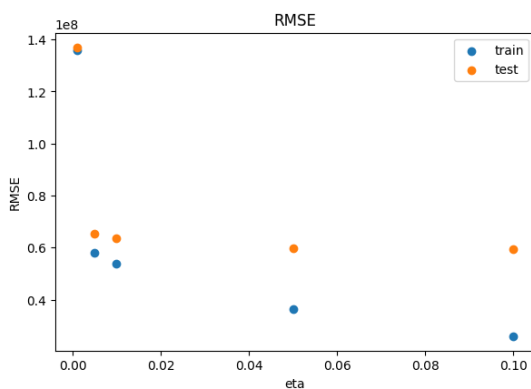


Figura 39. Métricas de error RMSE y  $eta$  para XGBoost (Elaboración propia)

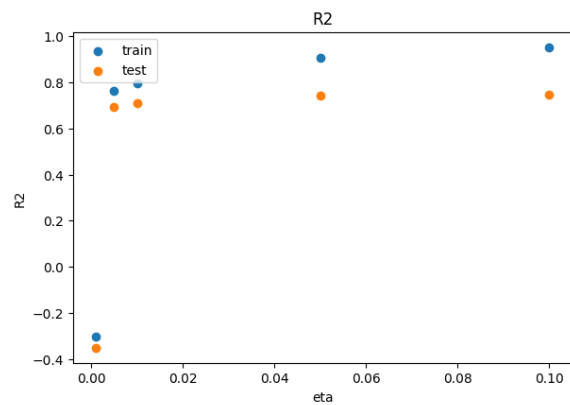


Figura 40. Métricas de error  $R^2$  y  $eta$  para XGBoost (Elaboración propia)

La profundidad máxima del árbol de decisión,  $max\_depth$ , se utiliza para controlar el sobreajuste. Aumentar este valor hará que el modelo sea más complejo, lo que hace más posible el sobreajuste. Para el presente modelo se utiliza el valor predeterminado de 6.

## 4.6. Evaluación de los modelos

En esta etapa se presentan las métricas de desempeño utilizadas para la evaluación de los diferentes modelos, con el fin de determinar el modelo que mejor se comporta a partir del conjunto de datos. A continuación se presentan las métricas de evaluación empleadas:

**Error Absoluto Medio (MAE)**, indica el promedio de los errores absolutos entre los valores de la predicción y los valores reales, dando una aproximación de cuán precisos son los modelos.

**Error porcentual absoluto medio (MAPE)**, es una medida de la precisión de la predicción que compara el error con el valor real.

**Raíz cuadrada del error absoluto medio (RMSE)**, es una medida de la precisión de la predicción que compara el error con el valor real, cuanto más se acerque esta función a cero, más precisa será la predicción del modelo de aprendizaje automático.

**Coefficiente de determinación  $R^2$** , el cual mide la relación lineal entre dos variables y da información sobre la variación del modelo. Un valor de  $R^2$  más cercano a 1 indica que las predicciones se ajustan mejor a los datos, lo que significa que tiene una baja variabilidad y, por lo tanto, se espera que prediga precios futuros con la misma precisión.

A continuación, se presenta en la Tabla 12 la comparación de las métricas de evaluación para los algoritmos implementados. A partir de los resultados de los cinco métodos diferentes se puede observar que el modelo Random Forest presenta mejores resultados.

Tabla 12. Métricas de evaluación para los algoritmos

Algoritmo	$R^2$	RSME	MAE	MAPE
Regresión Lineal	0,63	83.770.503	64.515.463	22,5%
Árbol de decisión	0,67	79.444.601	59.407.054	20,0%
Vecinos más cercanos	0,70	75.760.177	52.767.784	17,9%
Random Forest	0,81	61.620.160.	41.914.886	14,3%
XGBoost	0,79	62.126.693	43.157.639	14,6%

Fuente: Elaboración propia

En la Tabla 13 se presenta la comparación de los resultados utilizando como unidad de medida el dólar americano USD para las métricas de RSME y MAE, la conversión de unidades se realiza para mayor facilidad en el análisis y divulgación de los resultados, utilizando la Tasa de cambio de \$4.648.

Tabla 13. Métricas de evaluación para los algoritmos, unidades dólares americanos USD

Algoritmo	R <sup>2</sup>	RSME	MAE	MAPE
Regresión Lineal	0,63	\$18.022,91	\$1.3880,26	22,5%
Árbol de decisión	0,67	\$17.092,21	\$1.2781,21	20,0%
Vecinos más cercanos	0,70	\$16.299,52	\$1.1352,79	17,9%
Random Forest	0,81	\$13.257,35	\$9.017,833	14,3%
XGBoost	0,79	\$13.366,33	\$9.285,206	14,6%

Fuente: Elaboración propia

## 4.7. Predicciones

A continuación se presentan los resultados de las predicciones para cada uno de los modelos implementados, utilizando el conjunto de datos de prueba. En las Figuras 41 a 45, se presentan las gráficas de la predicción de los precios de un inmueble en el conjunto de datos de prueba en comparación con el precio real.

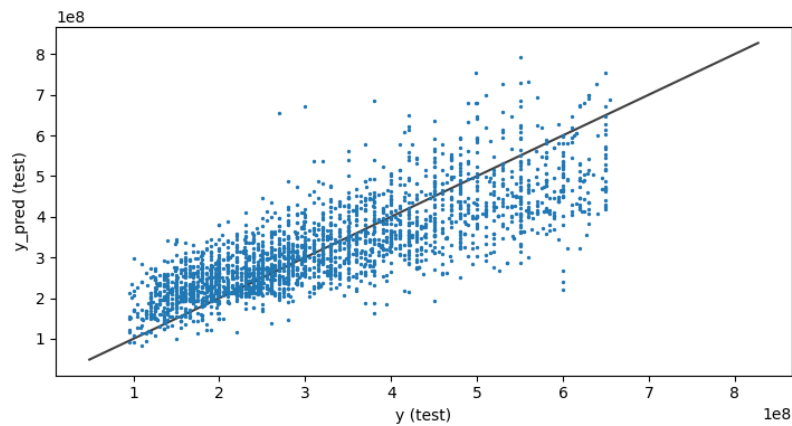


Figura 41. Predicción del precio en relación con el precio real del modelo Regresión Lineal (Elaboración propia)

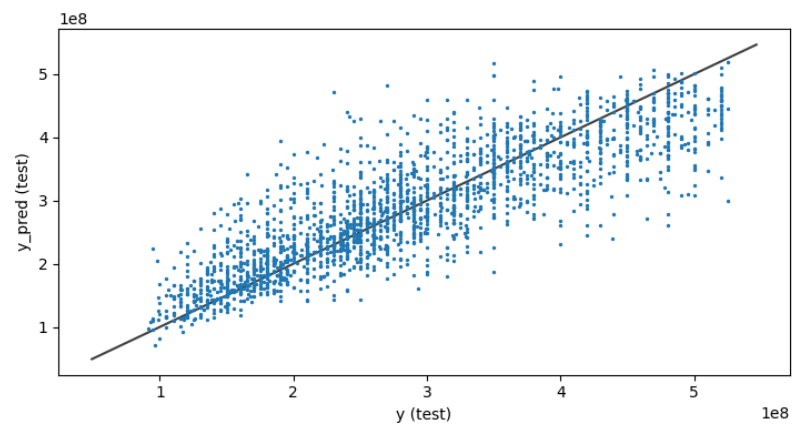


Figura 42. Predicción del precio en relación con el precio real del modelo Árboles de decisión (Elaboración propia)

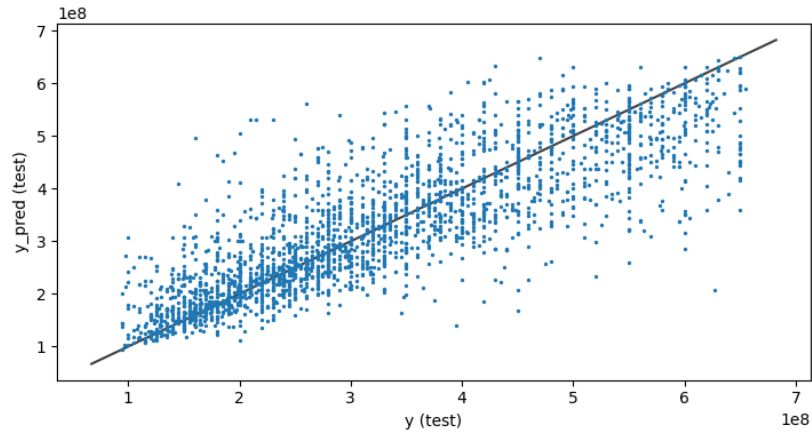


Figura 43. Predicción del precio en relación con el precio real de del modelo Random Forest (Elaboración propia)

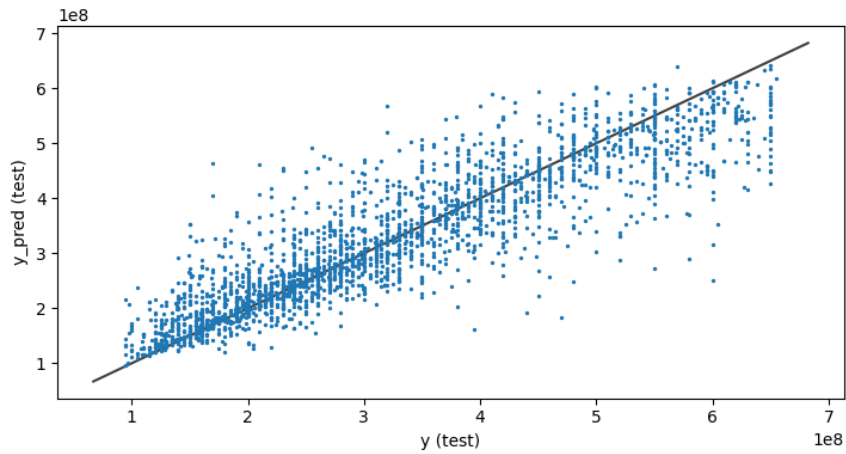


Figura 44. Predicción del precio en relación con el precio real del modelo K Vecino más cercano (Elaboración propia)

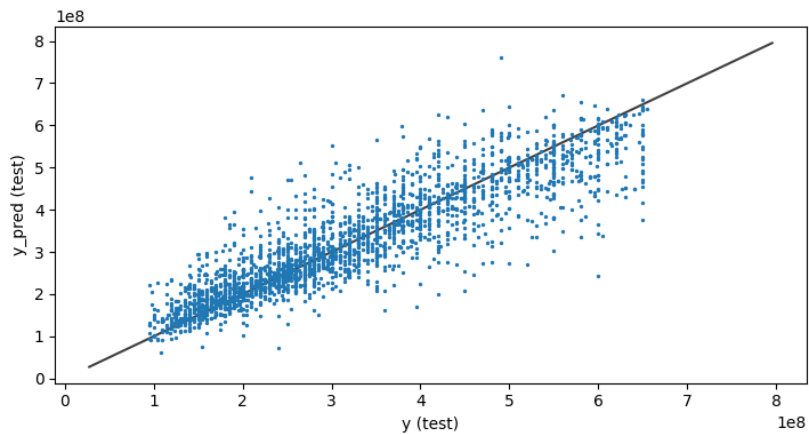


Figura 45. Predicción del precio en relación con el precio real de del modelo XGBoost regresor (Elaboración propia)



## 4.8. Análisis de los resultados

A continuación, se presenta un análisis de los resultados obtenidos de acuerdo con las métricas de error implementadas para la evaluación de los modelos.

De acuerdo con los resultados de la evaluación realizada con las métricas seleccionadas, el modelo que presentó mejores resultados es el Random Forest, con todas las métricas con un  $R^2$  de 0,81, MAPE de 14,3% y MAE de \$9.017 USD, en comparación con el modelo de XGBoost que presentó un  $R^2$  de 0,79, un MAPE de 14,6% y MAE de \$9.285 USD. Ambos modelos con resultados muy similares y comparables. Mientras modelos como la regresión lineal o los árboles de decisión presentaron menor desempeño.

Los resultados presentados corresponden a las métricas de evaluación obtenidas con subconjunto de datos 2, con el cual se obtuvieron mejores resultados, después del proceso de configuración de hiperparámetros. A pesar de que los resultados son prometedores, se debe mencionar que se evidencia algún sobreajuste aun en los modelos configurados, ya que en todos los casos evaluados los resultados en el conjunto de entrenamiento son superiores

Los resultados obtenidos se consideran aceptables, teniendo en cuenta las limitaciones del conjunto de datos, sin embargo los modelos pueden ser mejorados a futuro por ejemplo, por ejemplo utilizando características en el modelo Random Forest, o combinando los datos con otras fuentes relevantes.

Como métricas de evaluación, si bien se calculó tanto el RMSE como el MAE, esta última es más relevante para el presente análisis, puesto que tanto la variable objetivo precio, como otras variables del conjunto de datos presentan distribuciones asimétricas y valores extremos, y si bien se optó por convertir estas variables utilizando transformaciones log (logaritmos) o normalizando los datos, estos procesos no presentaron mejores resultados que los suministrados.

## 5. Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones del presente documento y las perspectivas futuras a partir del trabajo desarrollado.

### 5.1. Conclusiones

En el presente documento se utilizaron los datos públicos del portal inmobiliario Properati para desarrollar un modelo que permitiera la predicción de los precios de venta de inmuebles en la ciudad de Medellín y el Área metropolitana. Para esto se consideraron atributos asociados a los inmuebles, como la superficie total, el número de habitaciones, el número de baños, el tipo de propiedad y la localización, esta última a través de las variables municipio, latitud y longitud. Se realizó el tratamiento y análisis de los datos, para posteriormente realizar la aplicación de técnicas de machine learning para la predicción.

A partir del análisis de datos realizado, se obtiene un panorama general del comportamiento del mercado inmobiliario en la ciudad de Medellín y alrededores, en cuanto al tipo de producto y precios. El conjunto de datos analizados incluye propiedades tipo casas y departamentos, con precios de venta entre los 91 y los 1.300 millones de pesos colombianos. Adicionalmente, la mayoría de los inmuebles contiene una distribución de 3 habitaciones y 2 baños.

En el presente estudio se utilizaron los modelos de aprendizaje automático de Regresión lineal, Árboles de decisión, Random Forest, K vecinos más cercanos y XGBoost para la regresión, para predecir los precios de vivienda en la ciudad de Medellín y el Área Metropolitana. Los mejores resultados se obtuvieron con el modelo Random forest con un coeficiente de determinación  $R^2$  de 0,81, MAPE de 14,3% y MAE de \$9.017 USD.

A partir de los resultados de los modelos, las variables que más influyen en la predicción de los precios de vivienda en la ciudad de Medellín y el Área Metropolitana, son la superficie total, seguida del número de baños. Por otra parte, las variables menos significativas para los modelos corresponden a la latitud y la localización en términos de municipio.

Los modelos utilizados en este estudio demostraron el potencial de las valoraciones de los inmuebles a partir de las técnicas de aprendizaje automático, y su potencial en el futuro para reforzar o reemplazar los avalúos tradicionales de las propiedades en la ciudad de Medellín y el Área Metropolitana. Los modelos propuestos permiten predecir los precios de vivienda, sin embargo se destaca la importancia de la selección de hiperparámetros de los modelos a aplicar, para mejorar los resultados y reducir el sobreajuste.

## 5.2. Líneas de trabajo

A continuación se presentan las líneas de trabajo futuro a partir del estudio desarrollado y algunos aspectos a considerar en posteriores análisis sobre la temática.

Para el trabajo desarrollado se utilizó el conjunto de datos del portal inmobiliario Properati; en trabajos futuros se recomienda adicionar datos de otros portales inmobiliarios, para un mayor cubrimiento de las viviendas publicadas para la venta en la ciudad de Medellín y el Área Metropolitana. Adicionalmente, completar el conjunto de datos con ofertas inmobiliarias de proyectos nuevos en etapas de construcción y ventas, obteniendo los datos directamente de los portales inmobiliarios de las empresas constructoras.

Una etapa futura a partir del presente trabajo consiste en el despliegue del modelo a través de una API (Application Programming Interface) o interfaz de programación de aplicaciones, en la cual se realice la predicción de los precios de vivienda a partir de determinados atributos de los inmuebles. Otra posibilidad es agregar los modelos en un portal inmobiliario permitiendo a los usuarios la estimación de los inmuebles de forma rápida y precisa.

El presente trabajo sirve de base para futuros análisis de los precios en la ciudad y alrededores. A partir del conocimiento generado en el presente documento, es posible aplicar en el futuro técnicas más avanzadas que presenten mayor precisión en los resultados y menor sobreajuste en la fase de entrenamiento. Así mismo, es posible emplear el desarrollo como un punto de partida para estimar los precios en otras ciudades con características o condiciones similares.

## 6. Bibliografía

- Alcaldía de Medellín. (2022). *Datos Generales de Medellín*. Obtenido de <https://www.medellin.gov.co/es/conoce-algunos-datos-generales-de-la-ciudad/>
- Alfaro-Navarro, J.-L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. *Complexity*, 2020. doi:<https://doi.org/10.1155/2020/5287263>
- Antipov, E., & Pokryshevskaya, E. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39, 1772-1778. doi:10.1016/j.eswa.2011.08.077
- Área Metropolitana del Valle de Aburrá. (2020). *Estado Global de la metrópolis*. Obtenido de [https://www.metropolis.org/sites/default/files/resources/Global-state-metropolis-Valle-de-Aburra\\_AMVA\\_Oct2019.pdf](https://www.metropolis.org/sites/default/files/resources/Global-state-metropolis-Valle-de-Aburra_AMVA_Oct2019.pdf)
- Área Metropolitana del Valle de Aburrá. (2022). *El Valle de Aburrá*. Obtenido de <https://www.metropol.gov.co/Paginas/Noticias/area-silvestre/programas-emitidos/el-valle-de-aburra.aspx>
- Banerjee, D., & Dutta, S. (2017). Predicting the housing price direction using machine learning techniques. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, (págs. 2998-3000). doi:10.1109/ICPCSI.2017.8392275
- Beyeler, M. (2017). *Machine Learning for OpenCV*. Packt Publishing Ltd.
- CAMACOL. (2022). *Coordenada Urbana*. Obtenido de Tablas de coyuntura: <https://camacol.co/descargable/tablas-de-coyuntura-nivel-municipal-septiembre-2021>
- Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *SPRS international journal of geo-information*, 7, 168.
- Congreso de la República de Colombia. (25 de Mayo de 2019). Ley 1955 de 2019. *Por el cual se expide el Plan Nacional de Desarrollo 2018-2022*. Obtenido de [http://www.secretariassenado.gov.co/senado/basedoc/ley\\_1955\\_2019.html](http://www.secretariassenado.gov.co/senado/basedoc/ley_1955_2019.html)
- Constitución Política de Colombia. (1991). *Artículo 51*. Obtenido de Gaceta Asamblea Constituyente de 1991 N° 85: <http://www.secretariassenado.gov.co/index.php/constitucion-politica>
- DANE. (2018). *Censo Nacional de Población y Vivienda*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivienda-2018>

- DANE. (2020). *Déficit habitacional*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/deficit-habitacional>
- DANE. (1 de 10 de 2021). *Boletín Técnico - Indicadores económicos alrededor de la construcción*. Obtenido de DANE Información para todos: [https://www.dane.gov.co/files/investigaciones/boletines/pib\\_const/Bol\\_ieac\\_Iltrim22.pdf](https://www.dane.gov.co/files/investigaciones/boletines/pib_const/Bol_ieac_Iltrim22.pdf)
- DANE. (2022). *Licencias de construcción*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/construccion/licencias-de-construccion>
- DANE. (1 de 10 de 2022). *Producto Interno Bruto -PIB- nacional trimestral*. Obtenido de DANE Información Para Todos: [https://www.dane.gov.co/files/investigaciones/boletines/pib/bol\\_PIB\\_Iltrim22\\_produccion\\_y\\_gasto.pdf](https://www.dane.gov.co/files/investigaciones/boletines/pib/bol_PIB_Iltrim22_produccion_y_gasto.pdf)
- Fan, C., Cui, Z., & Zhong, X. (2018). House Prices Prediction with Machine Learning Algorithms. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (págs. 6–10). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3195106.3195133
- Gobernación de Antioquia. (2019). *Encuesta Calidad de Vida 2019*. Obtenido de <https://antioquia.gov.co/index.php/encuesta-calidad-de-vida-2019>
- Grajales Alzate, Y. V. (2019). Modelo de predicción de precios de viviendas en el Municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz. *Modelo de predicción de precios de viviendas en el Municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz*. Obtenido de <http://hdl.handle.net/20.500.11912/5285>
- Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48-70. doi:10.1080/09599916.2020.1832558
- Hong, J. (2020). An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System. *Housing Finance Research*, 4. doi:10.52344/hfr.2021.4.0.33
- Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in south korea. *International Journal of Strategic Property Management*, 24(3), 140-152. doi:10.3846/ijspm.2020.11544
- Iberdrola. (2022). *Qué es el 'Machine Learning*. Obtenido de <https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico>
- Instituto Geográfico Agustín Codazzi. (23 de Noviembre de 2008). Resolución 620 de 2008. *Por la cual se establecen los procedimientos para los avalúos ordenados dentro del marco de la Ley 388 de 1997*. Obtenido de <https://www.igac.gov.co/es/contenido/resolucion-620-de-2008>

- Kang, J., Lee, H. J., Jeong, S. H., Lee, H. S., & Oh, K. J. (2020). Developing a forecasting model for real estate auction prices using artificial intelligence. *Sustainability (Switzerland)*, 12(7). doi:10.3390/su12072899
- Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management Special Real Estate*, 43, 202-211. doi:<https://doi.org/10.3905/jpm.2017.43.6.202>
- Liyanaarachchi, L., Wijethunga, I. A., & Madushanka, M. (s.f.). Housing Price Prediction using Machine Learning. *Housing Price Prediction using Machine Learning*.
- Lowrance, R. E., Lecun, Y., & Shasha, D. (2015). *Predicting the Market Value of Single-Family Residential Real Estate*. Master's thesis, New York University. Obtenido de <http://pqdtopen.proquest.com/#viewpdf?dispub=3685886>
- Manu Shahi, A. G., & Sengar, N. (2020). Machine Learning House Price Prediction. *International Journal for Modern Trends in Science and Technology*, 6(12), 186-189. doi:10.46501/ijmtst061236
- Marjan, Č., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, 7, 168. doi:10.3390/ijgi7050168
- Martinez Heras, J. (2020). *Error Cuadrático Medio para Regresión*. Obtenido de <https://www.iartificial.net/error-cuadratico-medio-para-regresion/>
- Martínez Sanchez, D. N., & Téllez Buitrago, J. V. (2021). *Método automático para la predicción del avalúo comercial de un inmueble en la ciudad de Bogotá*. Master's thesis, Universidad Católica de Colombia.
- Medellín cómo vamos. (2020). *Informe de Calidad de Vida de Medellín 2020*. Obtenido de <https://www.medellincomovamos.org/system/files/2021-12/docuprivados/MCV%20ILB%20Metro%202020%20final.pdf>
- Medellín como vamos. (2022). *Área metropolitana del Valle de Aburrá*. Obtenido de <https://www.medellincomovamos.org/territorio/area-metropolitana-del-valle-de-aburra>
- Mora-Garcia, R.-T., Céspedes-Lopez, M.-F., & Perez-Sanchez, V. R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, 11(11), 2100. doi:10.3390/land11112100
- Neloy, A. A., Haque, H. M., & Ul Islam, M. M. (2019). Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring. *Proceedings of the 2019 11th International Conference on Machine Learning and Computing* (págs. 350–356). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3318299.3318377
- Nilsson, P. (2019). *Prediction of residential real estate selling prices using neural networks*. Master's thesis, KTH Royal Institute of Technology. Obtenido de <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-249637>

- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934. doi:10.1016/j.eswa.2014.11.040
- Phan, T. D. (2019). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018*, 35–42. doi:10.1109/iCMLDE.2018.00017
- Presidencia de la República de Colombia. (13 de Agosto de 2019). Decreto 1467 de 2019. *Por el cual se adiciona el Decreto 1077 de 2015 en relación con el precio máximo de la Vivienda de Interés Social*. Obtenido de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=99353>
- Properati. (2022). *Properati Data*. Obtenido de [https://bigquery.cloud.google.com/dataset/properati-data-public:properties\\_co](https://bigquery.cloud.google.com/dataset/properati-data-public:properties_co)
- Quang, T., Minh, N., Hy, D., & Bo, M. (2020). Housing Price Prediction via Improved Machine Learning Techniques. 174, págs. 433-442. Elsevier B.V. doi:10.1016/j.procs.2020.06.111
- Real Academia Española. (2022). *Diccionario de la lengua española, 23.ª ed.* Obtenido de <https://dle.rae.es/>
- Rodríguez Rojas, O. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. Obtenido de [http://www.oldemarrodriguez.com/yahoo\\_site\\_admin/assets/docs/Documento\\_CRISP-DM.2385037](http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037)
- Romero, M. (2017). La reconfiguración territorial: macroproyecto de la centralidad sur, Valle de Aburrá, Colombia. 2018. *Revista de la Facultad de Trabajo Social*, 33(33), 8-28. doi:10.18566/rfts.v33n33.a01
- Soto Hincapié, R. A., & David Rodríguez, E. (2021). *Modelo de Predicción del Precio de la Vivienda en el Valle de San Nicolás*. Master's thesis, Universidad de Antioquia.
- Suthaharan, S. (2016). *Decision Tree Learning. In: Machine Learning Models and Algorithms for Big Data Classification* (Vol. 36). Boston: Integrated Series in Information Systems. doi:[https://doi.org/10.1007/978-1-4899-7641-3\\_10](https://doi.org/10.1007/978-1-4899-7641-3_10)
- Tissnesh Betancur, A., & Posada Cuartas, E. (2014). Determinantes De La Oferta De Vivienda Nueva En Medellín.
- Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174, 433–442. doi:10.1016/j.procs.2020.06.111
- Voutas Chatzidis, I. (2019). *Prediction of housing prices based on spatial and social parameters using regression and deep learning methods*. Master's thesis, Aristotle University of Thessaloniki.

# Anexos

A continuación se presentan los anexos del presente trabajo.

## Anexo I. Notebook

En este Notebook se encuentra el código en Python con el análisis de datos realizado para obtener los resultados presentados en este trabajo. Es posible observar el código en el siguiente enlace del repositorio de github ViviendasMedellin:

<https://github.com/atmedig/ViviendasMedellin/blob/main/ViviendasMedellin.ipynb>



The screenshot shows a GitHub notebook viewer interface. At the top, it displays the repository name 'ViviendasMedellin' and the file name 'ViviendasMedellin.ipynb'. Below this, there is a search bar with '1 contributor' and a 'Download' button. The main content area shows the notebook's title: 'Proyecto: Predicción de Precios de vivienda en la ciudad de Medellin y el Área Metropolitana', followed by the author 'Elaborado por: Angie Medina'. The 'Importar las librerías' section is visible, containing the following Python code:

```
In [1]: from google.cloud import bigquery
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from pandas import Series
%matplotlib inline
import unicodedata
import plotly.express as px
from _plotly_future_ import v4_subplots
from plotly.offline import init_notebook_mode, iplot
import plotly.graph_objs as go
import matplotlib.pyplot as plt
```

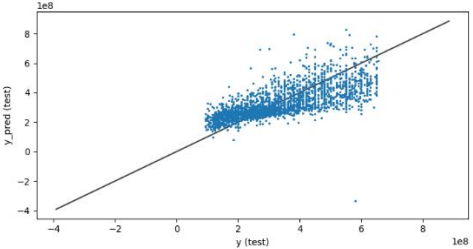
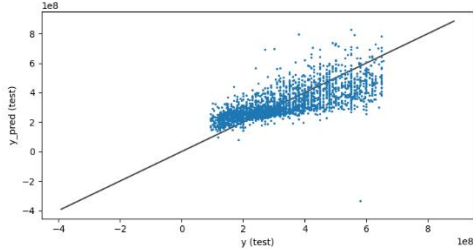


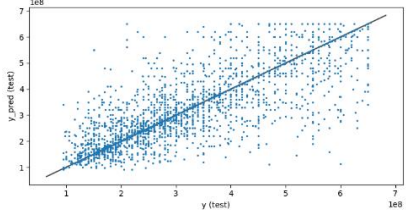
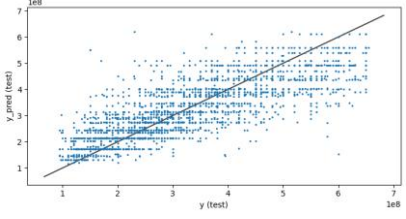
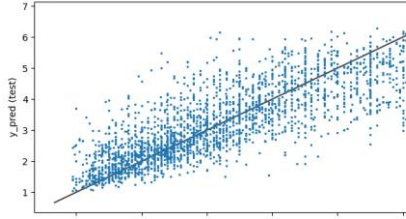
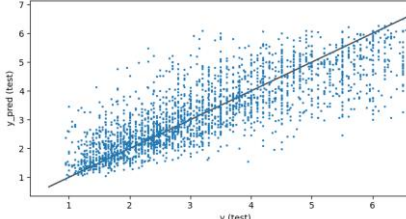
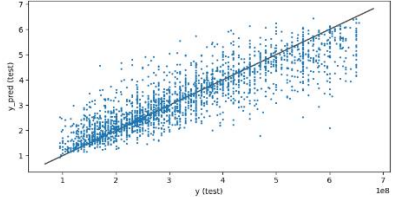
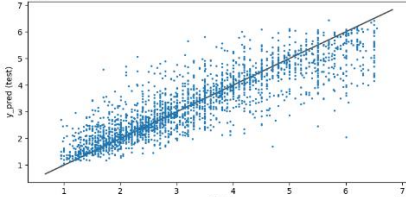
## Anexo II. Resultados de los modelos para diferentes conjuntos de datos

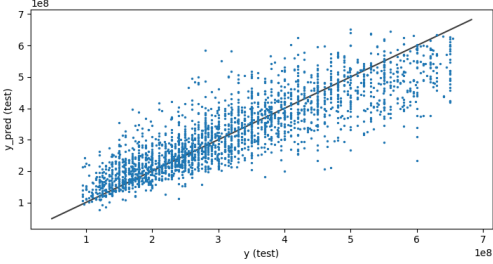
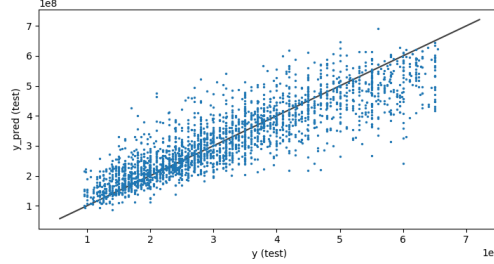
En este anexo se incluyen los resultados de métricas de evaluación para los tres subconjuntos de datos considerados durante las iteraciones para el proceso de modelado.

### Resultados de métricas de evaluación para los modelos, utilizando el subconjunto de datos 1:

El subconjunto de datos 1 Incluye las variables numéricas: latitud, longitud, superficie total, número de habitaciones y baños, como variables de entrada, y la variable precio como variable objetivo.

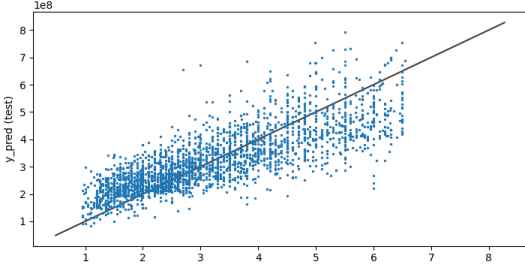
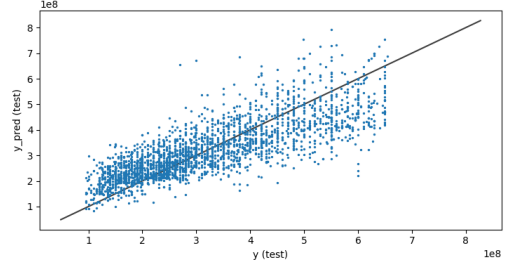
Modelo Inicial	Modelo Optimizado
<b>Modelo: Regresión lineal</b>	<b>Modelo: Regresión lineal</b>
R2 en Train: 0.5278017066715841 R2 ajustado en train: 0.5275779796303028 ----- R2 en test: 0.5359929466226653 R2 ajustado en Test: 0.5351121435600659 ----- RMSE en train: 94488344.491277 RMSE en test: 94746860.92959999 ----- MAE en train: 73281532.29378462 MAE en test: 72830906.2534759 ----- MAPE en train: 0.26444507173107606 MAPE en test: 0.26166034160505963	R2 en Train: 0.5278017066715841 R2 ajustado en train: 0.5275779796303028 ----- R2 en test: 0.5359929466226653 R2 ajustado en Test: 0.5351121435600659 ----- RMSE en train: 94488344.491277 RMSE en test: 94746860.92959999 ----- MAE en train: 73281532.29378462 MAE en test: 72830906.2534759 ----- MAPE en train: 0.26444507173107606 MAPE en test: 0.26166034160505963
	
<b>Modelo: Árbol de Decisión</b>	<b>Modelo: Árbol de Decisión óptimo</b>
R2 en Train: 0.993647476412116 R2 ajustado en train: 0.9936444665933025 ----- R2 en test: 0.6094076602351934 R2 ajustado en Test: 0.608662169174926 ----- RMSE en train: 10959460.520083074 RMSE en test: 86928943.19938555 ----- MAE en train: 2700241.6978929667 MAE en test: 52155953.2148413 ----- MAPE en train: 0.01011959031170725 MAPE en test: 0.17916885098148858	R2 en Train: 0.6760928305801259 R2 ajustado en train: 0.6759393637131592 ----- R2 en test: 0.6517164908129178 R2 ajustado en Test: 0.6510553603854556 ----- RMSE en train: 78257561.96679892 RMSE en test: 82085981.39737895 ----- MAE en train: 59427055.63567791 MAE en test: 61342475.27402256 ----- MAPE en train: 0.2072134182879804 MAPE en test: 0.21197874266531502

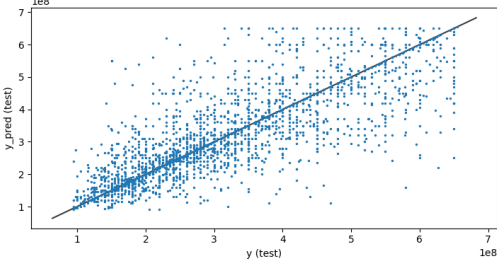
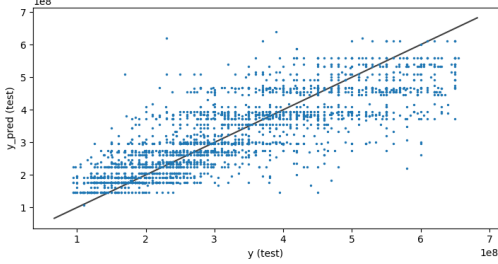
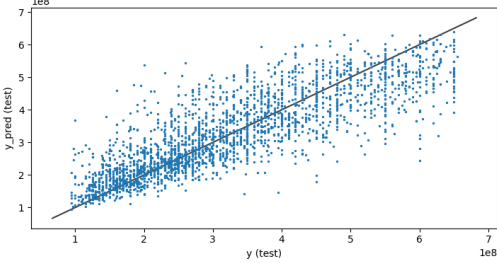
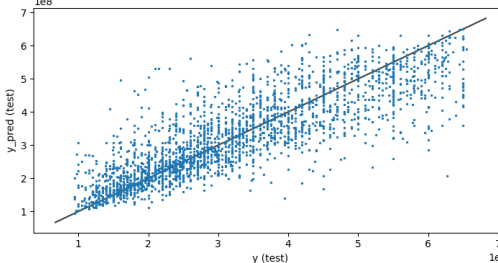
	
<p><b>Modelo: k Vecinos más cercanos</b></p> <p>R2 en Train: 0.7581422321233116 R2 ajustado en train: 0.7580276401741612</p> <p>-----</p> <p>R2 en test: 0.6619630119975107 R2 ajustado en Test: 0.6613213320658431</p> <p>-----</p> <p>RMSE en train: 67623252.22057417 RMSE en test: 80869480.03391927</p> <p>-----</p> <p>MAE en train: 48318471.87828392 MAE en test: 58700077.631212115</p> <p>-----</p> <p>MAPE en train: 0.16886572770639796 MAPE en test: 0.20459031303714084</p> 	<p><b>Modelo: k Vecinos más cercanos óptimo</b></p> <p>R2 en Train: 0.7862777688978363 R2 ajustado en train: 0.786176507535616</p> <p>-----</p> <p>R2 en test: 0.6648054863551814 R2 ajustado en Test: 0.6641692021607153</p> <p>-----</p> <p>RMSE en train: 63568341.81451925 RMSE en test: 80528755.93635903</p> <p>-----</p> <p>MAE en train: 44690882.31558386 MAE en test: 57827437.357670456</p> <p>-----</p> <p>MAPE en train: 0.15565027128295109 MAPE en test: 0.20057461759213113</p> 
<p><b>Modelo: Random Forest</b></p> <p>R2 en Train: 0.9608898064243883 R2 ajustado en train: 0.9608712760569214</p> <p>-----</p> <p>R2 en test: 0.7791996537327044 R2 ajustado en Test: 0.7787805186790459</p> <p>-----</p> <p>RMSE en train: 27193243.783594638 RMSE en test: 65358550.81980068</p> <p>-----</p> <p>MAE en train: 18349602.047780983 MAE en test: 44415215.269857734</p> <p>-----</p> <p>MAPE en train: 0.0641829785003399 MAPE en test: 0.15433247633526193</p> 	<p><b>Modelo: Random Forest óptimo</b></p> <p>R2 en Train: 0.9608898064243883 R2 ajustado en train: 0.9611323188734593</p> <p>-----</p> <p>R2 en test: 0.779182077489831 R2 ajustado en Test: 0.7787629090720061</p> <p>-----</p> <p>RMSE en train: 27102383.677947957 RMSE en test: 65361152.11758489</p> <p>-----</p> <p>MAE en train: 18277306.58256161 MAE en test: 44385234.768827185</p> <p>-----</p> <p>MAPE en train: 0.06405493603617671 MAPE en test: 0.15453836516234015</p> 
<p><b>Modelo: Regresión XGBoost</b></p> <p>R2 en Train: 0.8633244001422687 R2 ajustado en train: 0.8632596433907016</p> <p>-----</p> <p>R2 en test: 0.7492848953059801</p>	<p><b>Modelo: XGBoost óptimo</b></p> <p>R2 en Train: 0.8985599399207638 R2 ajustado en train: 0.8985118777298801</p> <p>-----</p> <p>R2 en test: 0.7682993543188299</p>

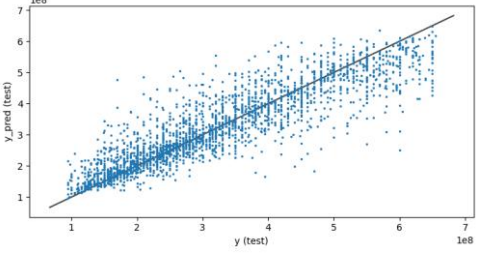
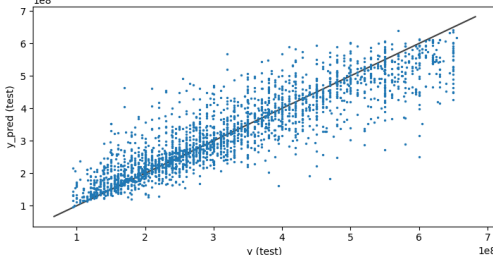
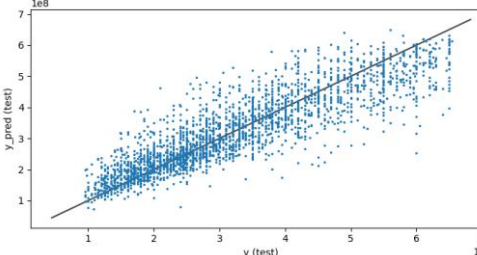
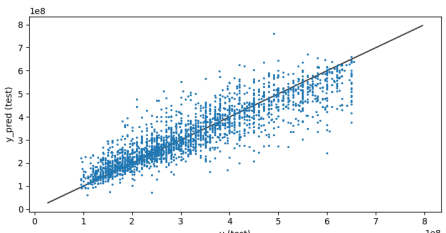
<p>R2 ajustado en Test: 0.748808974454245</p> <p>-----</p> <p>RMSE en train: 50834847.80679945 RMSE en test: 69645456.40455653</p> <p>-----</p> <p>MAE en train: 37234934.3508855 MAE en test: 50905661.63181818</p> <p>-----</p> <p>MAPE en train: 0.1302557953261137 MAPE en test: 0.17337137829110408</p> 	<p>R2 ajustado en Test: 0.7678595277324951</p> <p>-----</p> <p>RMSE en train: 43794622.00103184 RMSE en test: 66952401.78339006</p> <p>-----</p> <p>MAE en train: 31509114.16071598 MAE en test: 48091807.413636364</p> <p>-----</p> <p>MAPE en train: 0.110712871140984 MAPE en test: 0.163844968710924</p> 
--	---

**Resultados de métricas de evaluación de los modelos, utilizando el subconjunto de datos 2**

El subconjunto de datos 2 incluye las variables categóricas municipio, barrio y tipo de propiedad, y las variables numéricas latitud, longitud, superficie total, número de habitaciones y baños, como variables de entrada, y la variable precio como variable objetivo.

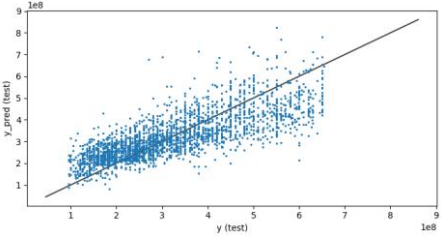
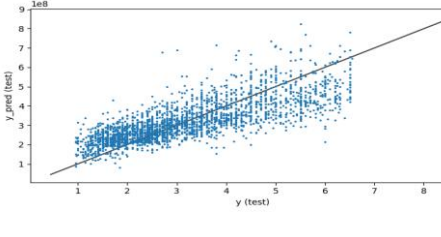
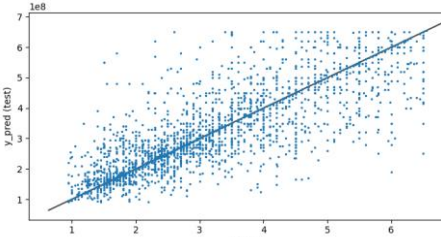
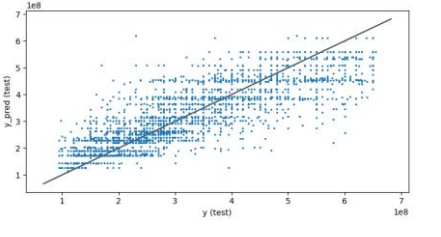
Modelo Inicial	Modelo Optimizado
<b>Modelo: Regresión lineal</b>	<b>Modelo: Regresión lineal</b>
<p>R2 en Train: 0.6200834708178123 R2 ajustado en train: 0.6186749011212532</p> <p>-----</p> <p>R2 en test: 0.6372752673064543 R2 ajustado en Test: 0.6318343963160511</p> <p>-----</p> <p>RMSE en train: 84753985.12468371 RMSE en test: 83770503.10493058</p> <p>-----</p> <p>MAE en train: 65341981.8904667 MAE en test: 64515463.73923532</p> <p>-----</p> <p>MAPE en train: 0.22958255555710014 MAPE en test: 0.225507712313094</p> 	<p>R2 en Train: 0.6200834708178123 R2 ajustado en train: 0.6186749011212532</p> <p>-----</p> <p>R2 en test: 0.6372752673064543 R2 ajustado en Test: 0.6318343963160511</p> <p>-----</p> <p>RMSE en train: 84753985.12468371 RMSE en test: 83770503.10493058</p> <p>-----</p> <p>MAE en train: 65341981.8904667 MAE en test: 64515463.73923532</p> <p>-----</p> <p>MAPE en train: 0.22958255555710014 MAPE en test: 0.225507712313094</p> 

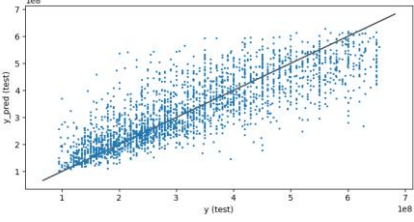
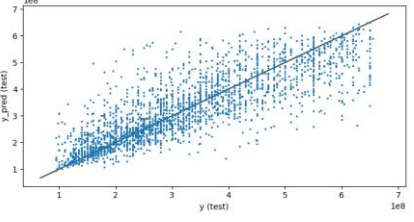
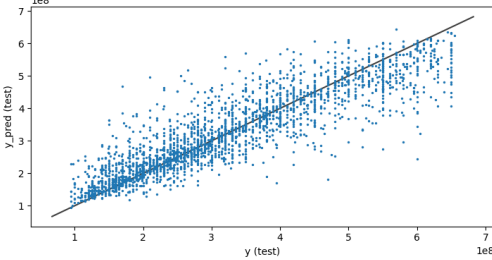
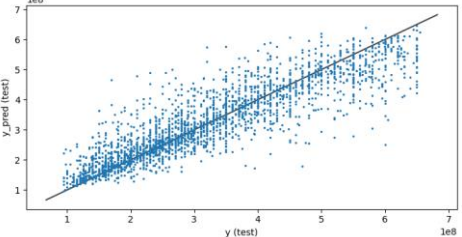
Modelo: <b>Árbol de Decisión</b>	Modelo: <b>Árbol de Decisión óptimo</b>
<p>R2 en Train: 0.9960498155977235                      R2 ajustado en train: 0.9960351699858129                      -----                      R2 en test: 0.6477807451909381                      R2 ajustado en Test: 0.6424974563688022                      -----                      RMSE en train: 8642208.071316047                      RMSE en test: 82548480.98444405                      -----                      MAE en train: 1918961.6627782884                      MAE en test: 50284466.163982205                      -----                      MAPE en train: 0.007235543535883827                      MAPE en test: 0.16950526826541212</p> 	<p>R2 en Train: 0.6930476948363262                      R2 ajustado en train: 0.6919096456014766                      -----                      R2 en test: 0.6737701461855816                      R2 ajustado en Test: 0.6688766983783654                      -----                      RMSE en train: 76181844.40296194                      RMSE en test: 79444601.29436885                      -----                      MAE en train: 57556892.491115585                      MAE en test: 59407054.5710741                      -----                      MAPE en train: 0.20107225109070587                      MAPE en test: 0.20343734054471735</p> 
Modelo: <b>k Vecinos más cercanos</b>	Modelo: <b>k Vecinos más cercanos óptimo</b>
<p>R2 en Train: 0.7858256030376491                      R2 ajustado en train: 0.7850315350196311                      -----                      R2 en test: 0.7007944289043829                      R2 ajustado en Test: 0.6963063453379487                      -----                      RMSE en train: 63635551.12391078                      RMSE en test: 76082955.87329896                      -----                      MAE en train: 45102261.74171797                      MAE en test: 55244167.96712121                      -----                      MAPE en train: 0.15582443316973527                      MAPE en test: 0.19061738405724077</p> 	<p>R2 en Train: 0.846183352996281                      R2 ajustado en train: 0.8456130682815356                      -----                      R2 en test: 0.7033277763917565                      R2 ajustado en Test: 0.6988776930376328                      -----                      RMSE en train: 53928420.455027714                      RMSE en test: 75760177.30898745                      -----                      MAE en train: 36067448.444612816                      MAE en test: 52767784.42525252                      -----                      MAPE en train: 0.12406258398095027                      MAPE en test: 0.1794554653933299</p> 
Modelo: <b>Random Forest</b>	Modelo: <b>Random Forest óptimo</b>
<p>R2 en Train: 0.9669502096526114                      R2 ajustado en train: 0.9668276750178031                      -----                      R2 en test: 0.8030548308780223                      R2 ajustado en Test: 0.8001006533411926                      -----</p>	<p>R2 en Train: 0.9676553484494561                      R2 ajustado en train: 0.9675354281708677                      -----                      R2 en test: 0.8037360349183487                      R2 ajustado en Test: 0.8007920754421239                      -----</p>

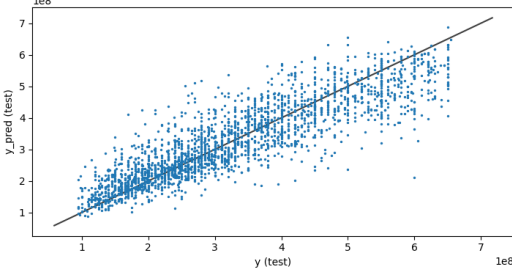
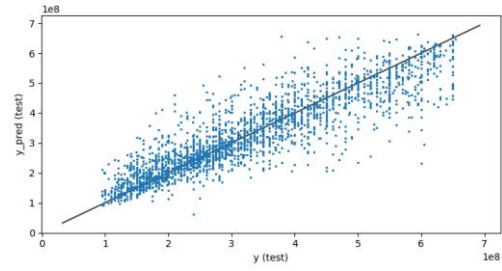
<p>RMSE en train: 24997719.342144046                  RMSE en test: 61727004.97215823</p> <p>-----</p> <p>MAE en train: 16831094.85692706                  MAE en test: 41963152.76657325</p> <p>-----</p> <p>MAPE en train: 0.058284988855716424                  MAPE en test: 0.14383876803934023</p> 	<p>RMSE en train: 24729610.251801573                  RMSE en test: 61620160.23860431</p> <p>-----</p> <p>MAE en train: 16756478.809442624                  MAE en test: 41914886.854855575</p> <p>-----</p> <p>MAPE en train: 0.05818883756239156                  MAPE en test: 0.14365493884058372</p> 
<p><b>Modelo: Regresión XGBoost</b></p>	<p><b>Modelo: XGBoost óptimo</b></p>
<p>R2 en Train: 0.8743601399647666                  R2 ajustado en train: 0.873894320538835</p> <p>-----</p> <p>R2 en test: 0.7823775673832335                  R2 ajustado en Test: 0.779113230893982</p> <p>-----</p> <p>RMSE en train: 48739352.499907844                  RMSE en test: 64886503.07099967</p> <p>-----</p> <p>MAE en train: 35746190.968557626                  MAE en test: 47193490.31363636</p> <p>-----</p> <p>MAPE en train: 0.12294362767744328                  MAPE en test: 0.16009262198662264</p> 	<p>R2 en Train: 0.9497091026162727                  R2 ajustado en train: 0.949522645253599</p> <p>-----</p> <p>R2 en test: 0.8004960962281832                  R2 ajustado en Test: 0.797503537671606</p> <p>-----</p> <p>RMSE en train: 30836190.921910528                  RMSE en test: 62126693.19318917</p> <p>-----</p> <p>MAE en train: 21410298.69277394                  MAE en test: 43157639.96363636</p> <p>-----</p> <p>MAPE en train: 0.07474755280085553                  MAPE en test: 0.14604677439607805</p> 

**Resultados de métricas de evaluación para los modelos, utilizando el subconjunto de datos 3**

El Subconjunto de datos 3 incluye las variables categóricas tipo de propiedad y municipio y las variables numéricas latitud, longitud, superficie total, número de habitaciones y número de baños, como variables de entrada, y la variable precio como variable objetivo.

<b>Modelo Inicial</b>	<b>Modelo Optimizado</b>
<p><b>Modelo: Regresión lineal</b></p> <p>R2 en Train: 0.5832046846015941 R2 ajustado en train: 0.5825324978677195 ----- R2 en test: 0.6066358161267305 R2 ajustado en Test: 0.6040853999841502 ----- RMSE en train: 88772295.03483161 RMSE en test: 87236844.15166783 ----- MAE en train: 68446466.98522495 MAE en test: 67851359.75813195 ----- MAPE en train: 0.24450462227200717 MAPE en test: 0.24195290087172786</p> 	<p><b>Modelo: Regresión lineal</b></p> <p>R2 en Train: 0.5832046846015941 R2 ajustado en train: 0.5825324978677195 ----- R2 en test: 0.6066358161267305 R2 ajustado en Test: 0.6040853999841502 ----- RMSE en train: 88772295.03483161 RMSE en test: 87236844.15166783 ----- MAE en train: 68446466.98522495 MAE en test: 67851359.75813195 ----- MAPE en train: 0.24450462227200717 MAPE en test: 0.24195290087172786</p> 
<p><b>Modelo: Árbol de Decisión</b></p> <p>R2 en Train: 0.9953200970713455 R2 ajustado en train: 0.9953125495568984 ----- R2 en test: 0.6607139388800582 R2 ajustado en Test: 0.6585141436706612 ----- RMSE en train: 9406638.533562845 RMSE en test: 81018751.79618506 ----- MAE en train: 2104110.8697288022 MAE en test: 49574538.41587915 ----- MAPE en train: 0.008019572291274419 MAPE en test: 0.17033808178369125</p> 	<p><b>Modelo: Árbol de Decisión óptimo</b></p> <p>R2 en Train: 0.694866509278327 R2 ajustado en train: 0.6943744051760342 ----- R2 en test: 0.6780573726619431 R2 ajustado en Test: 0.675970025345106 ----- RMSE en train: 75955805.20472613 RMSE en test: 78920854.8977507 ----- MAE en train: 57422284.201056935 MAE en test: 59478427.81293636 ----- MAPE en train: 0.19907614010460967 MAPE en test: 0.20516703728867502</p> 
<p><b>Modelo: k Vecinos más cercanos</b></p> <p>R2 en Train: 0.7764299589372556 R2 ajustado en train: 0.7760693963058102 ----- R2 en test: 0.6847786641068787 R2 ajustado en Test: 0.6827348949573047 ----- RMSE en train: 65016387.72063993 RMSE en test: 78092682.58482215</p>	<p><b>Modelo: k Vecinos más cercanos óptimo</b></p> <p>R2 en Train: 0.8385725738148446 R2 ajustado en train: 0.8383122316988074 ----- R2 en test: 0.6875060322869994 R2 ajustado en Test: 0.6854799463025901 ----- RMSE en train: 55246490.449355766 RMSE en test: 77754110.63314848</p>

<p>-----</p> <p>MAE en train: 46230263.21943366 MAE en test: 56509611.76803031</p> <p>-----</p> <p>MAPE en train: 0.15944534097747126 MAPE en test: 0.1947548946619036</p> 	<p>-----</p> <p>MAE en train: 36968431.133851066 MAE en test: 53869933.31414141</p> <p>-----</p> <p>MAPE en train: 0.12716890048178398 MAPE en test: 0.1841610944049799</p> 
<p><b>Modelo: Random Forest</b></p>	<p><b>Modelo: Random Forest óptimo</b></p>
<p>R2 en Train: 0.9648906131357258 R2 ajustado en train: 0.9648339904645662</p> <p>-----</p> <p>R2 en test: 0.7986801973933007 R2 ajustado en Test: 0.797374920259695</p> <p>-----</p> <p>RMSE en train: 25764852.227767475 RMSE en test: 62408793.53061836</p> <p>-----</p> <p>MAE en train: 17243944.02559258 MAE en test: 42273348.644897945</p> <p>-----</p> <p>MAPE en train: 0.059864635616444145 MAPE en test: 0.14596678868101076</p> 	<p>R2 en Train: 0.9661966037669014 R2 ajustado en train: 0.9661420873324111</p> <p>-----</p> <p>R2 en test: 0.8010412680196564 R2 ajustado en Test: 0.7997512991242841</p> <p>-----</p> <p>RMSE en train: 25281113.673736162 RMSE en test: 62041750.26708862</p> <p>-----</p> <p>MAE en train: 17034447.239428744 MAE en test: 42077534.42889633</p> <p>-----</p> <p>MAPE en train: 0.05917996907096696 MAPE en test: 0.14476928088554464</p> 
<p><b>Modelo: XGBoost</b></p>	<p><b>Modelo: XGBoost óptimo</b></p>

<p>R2 en Train: 0.875614236633731  R2 ajustado en train: 0.8754136334673116  -----  R2 en test: 0.7823114432567138  R2 ajustado en Test: 0.7809000376637939  -----  RMSE en train: 48495492.17042971  RMSE en test: 64896360.137242116  -----  MAE en train: 35392308.84108344  MAE en test: 47477123.72575758  -----  MAPE en train: 0.12244693767867079  MAPE en test: 0.16128956875333217</p> 	<p>R2 en Train: 0.9534881889347178  R2 ajustado en train: 0.9534131770014942  -----  R2 en test: 0.7996993904394594  R2 ajustado en Test: 0.7984007213461988  -----  RMSE en train: 29654981.571923617  RMSE en test: 62250619.03658134  -----  MAE en train: 20499343.899138175  MAE en test: 42802525.807575755  -----  MAPE en train: 0.07206898314236707  MAPE en test: 0.14522232820970163</p> 
--	--