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

YURI VANESA GRAJALES ALZATE

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MEDELLÍN

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

YURI VANESA GRAJALES ALZATE

Trabajo de grado para optar al título de magister en tecnologías de la información y la comunicación.

Asesor

JOSÉ RICARDO ZAPATA GONZÁLEZ

PhD. TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

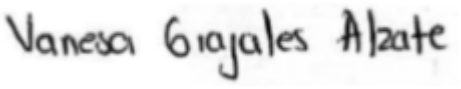
MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MEDELLÍN

2019

DECLARACIÓN ORIGINALIDAD

“Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en esta o cualquier otra universidad”. Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.

FIRMA AUTOR (ES) 

Ciudad y fecha: Medellín 24 de Febrero de 2020

CONTENIDO

LISTA DE GRÁFICAS

LISTA DE TABLAS

RESUMEN

El presente proyecto de grado tiene como objetivo realizar un modelo para la predicción de precios de compra y venta de vivienda (casas y apartamentos) nueva y usada en el municipio de Rionegro - Antioquia. Los datos se obtuvieron por medio de la técnica de *web scraping* y la predicción de precios se realizó mediante técnicas de aprendizaje automático. El propósito de este trabajo es apoyar la toma de decisiones informada de las personas que desean invertir o vender su inmueble en este municipio. Los resultados mostraron que variables como el área de la casa, el área construida, el tipo de vivienda y el estrato son factores importantes a la hora de determinar el precio.

PALABRAS CLAVE: Web scraping, precios de vivienda, machine learning.

El código utilizado para el desarrollo del presente trabajo se puede encontrar en: https://github.com/VanesaGrajales/preciosviviendas

1. INTRODUCCIÓN

En un contexto industrializado y cada vez más globalizado, donde los precios de los bienes y servicios están en constante cambio, entre ellos el de las viviendas, el cual ha venido aumentando en las principales ciudades de Colombia y no solo por la volatilidad de los precios, sino también debido a la escasez de tierras que ha venido experimentado el país (La República, 2017); el tener una idea de cuánto puede incrementar un precio de una vivienda en una determinada ciudad o municipio puede ser de gran ayuda, tanto para las personas que desean adquirirla para habitar en ella como para quienes la quieran como opción de inversión y así se pueda apoyar el proceso de toma de decisiones en el sector de la vivienda.

Por otro lado, diversas investigaciones han trabajado con los precios de vivienda en las grandes urbes alrededor del mundo: (Phan, 2019), (Li, Ye, Lee, et.al, 2017); (Park y Bae, 2015), sin embargo, el monitoreo de dichos precios a escalas pequeñas, es decir, municipios que no son grandes ciudades, sigue siendo un gran desafío, ya que las deficiencias y falta de disponibilidad y análisis de los precios de venta de las viviendas en estos lugares no se encuentran abiertas al público o no se han realizado, lo que dificulta la predicción de los retornos de las inversiones en el sector, así como sus posibles efectos en otros sectores de la economía (Caplin, Chopra, Leahy, Lecun, & Thampy, 2008).

Es por esto que el presente trabajo se enfoca en determinar los precios de vivienda del municipio de Rionegro - Antioquia, una región con alto potencial de crecimiento y a la cual se le están realizando importantes inversiones y vías de acceso como el túnel de oriente las cuáles aumentan la valorización de las viviendas del sector y por ende cada vez es más complicado determinar cuánto es justo o conveniente invertir en casas o apartamentos o por el contrario en cuánto es considerable vender en este municipio.

En este estudio, se presentan un total de 9 apartados en los cuáles se abordará con más profundidad el problema y su solución. En el segundo capítulo se presenta cuál fue el problema encontrado y la justificación de porque es necesario abordarlo, en el tercer capítulo se abordará los objetivos del trabajo. En el cuarto capítulo, se muestra todo lo relacionado con el marco conceptual y los trabajos realizados anteriormente en el campo del presente estudio, posteriormente, se presenta la metodología que se utilizó para el desarrollo de este trabajo. En el capítulo 6 se presentan los resultados encontrados: extracción de los datos de las páginas web, análisis estadístico realizado y los modelos de machine learning implementados, luego se encuentran las conclusiones de los resultados encontrados y las recomendaciones para trabajos futuros.

2. PLANTEAMIENTO DEL PROBLEMA

2.1. Problema

Durante los últimos años la construcción de viviendas en el municipio de Rionegro Antioquia ha incrementado exponencialmente (ver figura 1) gracias a las condiciones climáticas, ambientales, de conexión al Valle de Aburrá y otras regiones, cercanía al aeropuerto internacional e importantes proyectos viales como el túnel de oriente (Alcaldía de Rionegro, 2016), lo cual ha permitido que la valorización en dicho municipio haya incrementado por lo menos un 16% (Propiedades Oriente Raíz, 2019). Esto hace que la rentabilidad de quienes quieran invertir o vender en este municipio aumente significativamente.

Figura 1. Incremento en el número de viviendas en el municipio de Rionegro
Copyright 2016 por Alcaldía de Rionegro

En concordancia con lo anterior, el hecho de que la valorización esté incrementando podría generar un fenómeno especulativo, en donde una inversionista no tenga como conocer el precio real de una vivienda de acuerdo con las condiciones del mercado del sector en el cuál desea invertir/vender, y si

además se suma el hecho que el actual entorno competitivo hace crítico para los desarrolladores, propietarios o inversionistas en activos inmobiliarios, permanecer atentos a las tendencias y la evolución del mercado para responder de manera adecuada a la aparición de nuevas oportunidades de compra y venta (Deloitte, 2016), podría generar pérdidas significativas.

El avaluó catastral (el cual se obtiene mediante la investigación y análisis estadístico del mercado inmobiliario) podría tomarse como referencia para determinar el precio real de una vivienda, que de acuerdo el Instituto Geográfico Agustín Codazi (2017) es aproximadamente un 70% del avaluó comercial (precio del mercado), sin embargo, en muchas propiedades este porcentaje puede ser mucho menor, ya que en la mayoría de las veces el precio de la vivienda depende de cómo se esté comportando el mercado en un determinado sector y en un determinado tiempo, por tanto el valor del catastro no siempre refleja el verdadero precio de la vivienda. Es por esto que quien desee comprar o vender una vivienda debe empezar a buscar el precio desde cero.

Metro Cuadrado (2019), presenta una forma en que puede determinarse el precio de una vivienda, la cual puede llegar a ser muy tedioso para una persona, además de tomar mucha parte de su tiempo y no se asegura que efectivamente ese sea el precio. La metodología se presenta a continuación:

- ✓ Realizar una plantilla con las siguientes variables: tipo de inmueble, años de construido, área, estrato, piso en el que queda, número de baños, número de parqueaderos, edificio inteligente, altura, número de líneas telefónicas, precio de arriendo, observaciones.

- ✓ Llamar a averiguar por los inmuebles que están el sector de la vivienda y anotarlos en la plantilla. El total de inmuebles averiguados no debe ser inferior a 20 inmuebles.

- ✓ Dividir el precio de cada una para saber el precio por metro cuadrado.

- ✓ Luego sumar los precios por metro cuadrado obtenido y dividirlos por el total de viviendas. Esto dará una idea de en cuanto comprar o vender una vivienda.

Otra opción podría ser consultar a un asesor, que en muchos casos son inmobiliarias quienes a su vez utilizan en su equipo de trabajo, personas que se encargan del avaluó de las viviendas, y la mayoría de las veces dichas entidades cobran por comisión. Generalmente la tarifa mínima para un avaluó es de 300 mil pesos por un inmueble cuyo valor comercial es de 100 millones de pesos. Si el costo del inmueble es mayor, se les suma a los 300 mil pesos el 1 por mil y si la vivienda queda fuera del área metropolitana, como es este caso, tiene un recargo que oscila entre el 1.5 y 2 por ciento.

Cualquiera de estos procesos conlleva gran cantidad tiempo, esfuerzo y dinero para los posibles inversores y/o vendedores (Boeing & Waddell, 2017). Esto sumado a que actualmente en el municipio de Rionegro no existe una plataforma o modelo que permita estimar el precio de un inmueble de forma rápida, eficiente y a un bajo costo, hace que la mayoría de las veces se tomen decisiones basadas más en la especulación que en la información.

2.2. Justificación

Invertir es una decisión que debe ser fruto de un proceso de meditación por lo que antes de realizar una inversión es importante hacer un estudio del mercado sobre donde se quiere destinar los fondos (Villa, 2018). La propiedad raíz es una buena opción para invertir ya que por lo general genera rentabilidad a largo plazo y además se está adquiriendo un patrimonio a largo plazo.

Por otro lado, la primera recomendación de las inmobiliarias al momento de comprar una vivienda en cualquier lugar es determinar el presupuesto con el que se cuenta para ello, ya que una vez se tenga esto, se buscarán condiciones de las viviendas que se ajusten a ello (Alisenda Inmobiliaria, 2016), la segunda recomendación es tener claro qué condiciones de la vivienda se quiere para empezar a hacer relaciones entre vivienda y precio.

Es por esto que tener claridad sobre cuál es la tendencia de los precios de vivienda en el sector que se desea adquirir (ya que los precios del catastro no están acordes con los valores comerciales de los inmuebles), además de las condiciones de dichas viviendas ayuda a mejorar la toma de decisiones de inversión, ya que se cuenta con un precio inicial de las viviendas de manera que se pueda negociar de forma justa y equitativa la compra o venta de dichos bienes raíces y además permite que las personas decidan si se ajusta a su presupuesto y sus necesidades o es necesario buscar en otros lugares. Debido a esto, el tener un modelo o plataforma que permita determinar el precio de la vivienda sería altamente eficaz y beneficioso para el inversionista.

A pesar de la demanda creciente de compra y venta de viviendas en el municipio de Rionegro, este no cuenta con un modelo o sistema que permita determinar dichos precios, lo cual hace que la búsqueda de información con respecto a este

tema siga siendo muy tedioso para las personas que deseen invertir en este municipio.

En línea con lo anterior, el principal beneficio de este proyecto radica en el ahorro de tiempo y el esfuerzo empleado a la hora de buscar los precios óptimos para invertir o vender propiedades en el municipio de Rionegro - Antioquia, así como un mejor conocimiento de los precios de una vivienda dada sus condiciones, lo cual conllevará a mejor toma de decisiones de manera rápida y eficaz.

3. OBJETIVOS

3.1. Objetivo General.

Crear un modelo de datos para la predicción de precios de las viviendas en el municipio de Rionegro usando aprendizaje de maquina (machine learning), mediante la extracción de información disponible en páginas web para la toma de decisiones informada de compra y venta de propiedad raíz.

3.2. Objetivos Específicos

- Consolidar la información relacionada con los precios de vivienda (casas y apartamentos) de páginas web previamente establecidas en el alcance del municipio de Rionegro.

- Realizar un análisis descriptivo de los datos de las viviendas que están en venta, de acuerdo con sus condiciones habitacionales y el precio, utilizando herramientas estadísticas y de análisis de datos.

- Generar modelos de aprendizaje de maquina (machine learning) que permitan predecir la relación entre el perfil de las viviendas y el precio de estas.

- Determinar el modelo que más se ajusta al comportamiento de los datos analizados.

4. MARCO REFERENCIAL

4.1. Marco contextual

El oriente antioqueño está viviendo un momento de alta valorización debido al desarrollo de importantes obras viales y a que es una zona con potencial de explotación turística. Los municipios con más auge son Rionegro y Guarne, ya que en estos se desarrollarán importantes obras como equipamiento con presencia de todos los servicios (bancos, salud, colegios, etc), inversión vial por parte de los municipios, proyecto de tren ligero, conectividad con el túnel de oriente, entre otros (López de Mesa, 2019).

Los analistas estiman que para la próxima década serán muchas las familias que decidan asentarse en el Oriente de Antioquia, por ser una región con mejor calidad del aire y menos problemas de circulación y seguridad, comparada con Medellín (Argos, 2017). Entre los municipios para elegir dónde comprar vivienda en el oriente antioqueño se encuentran:

- **Rionegro:** es considerado monumento nacional de Colombia, reconocido como centro empresarial con gran actividad y desarrollo económico, cercano al aeropuerto internacional José María Córdova y con un clima templado – frío que es uno de sus mayores atractivos.

- **El Retiro:** Su cercanía a la ciudad, la amplia oferta de servicios, centros comerciales, sitios de interés cultural y recreativos lo hacen uno de los más apetecidos por los inversionistas. Un lugar donde se pueden disfrutar el aire limpio y la vida tranquila de un pueblo con todas las comodidades de la ciudad.

- **Guarne:** Su cercanía a la autopista Medellín- Bogotá, así como su clima hacen de este un gran atractivo tanto para vivir como para invertir**.**

Al realizar un análisis sobre las licencias de construcción que se están adjudicando en los municipios del oriente antioqueño, se puede evidenciar que el municipio de La Ceja es el que más ha dado licencias para construcción de viviendas con un total de 903 adjudicadas, seguido de Rionegro con un número importante de 680 licencias, lo que confirma un importante crecimiento urbanístico y de expansión de estos territorios (Camara de Comercio del Oriente Antioqueño, 2018).

En la Figura 2, se presenta el panorama de la construcción en el Oriente Antioqueño. Aquí se puede ver como los proyectos de vivienda han crecido en un 200% durante los últimos 6 años (Cámara de Comercio del Oriente Antioqueño (CCOA), 2018)

Figura 2. La construcción en el Oriente Antioqueño. Copyright 2017 por Oriente Comercial Digital

Esto muestra que gran cantidad de personas que está prefiriendo estos municipios, ya sea para invertir o para vivir allí. Así mismo las beneficiosas posibilidades de vivir o invertir en el oriente antioqueño y el auge de la alta valorización hacen que las personas que tienen propiedades allí también las puedan vender para generar grandes rendimientos.

Variables como la ubicación de la casa o apartamento, cercanía a colegios, centros de salud, número de habitaciones, numero de baños, variables del POT como zona de reserva o no, entre otros pueden influir en el precio de estas viviendas, pero con una búsqueda de información solo en páginas web no es posible determinarlo, por esto, este trabajo pretende determinar esas variables que más influyen en el precio de dichas viviendas, con el fin de que las personas puedan tomar mejores decisiones acerca de cómo y dónde quieren su casa dependiendo de su presupuesto.

Otro punto importante para tener en cuenta en este proyecto es que, si se cuenta con información de ubicación de las viviendas, existe el potencial, por ejemplo, de facilitar la determinación de las ubicaciones óptimas de servicios como educación y salud (Rubio, Guzmán, & Otero, 2019) para construir políticas públicas en estos sectores que contribuyan a mejor calidad de vida de las personas allí residentes.

4.2. Marco conceptual

Uno de los métodos de recolección de datos en la actualidad es por medio del web scraping, el cual es una técnica que permite acceder y extraer automáticamente grandes cantidades de información de un sitio web, lo cual puede ahorrar gran cantidad de tiempo y esfuerzo (Kho, 2018). Uno de los lenguajes de programación más utilizados para desarrollar este tipo de técnica es Python (Van Rossum, 1989) por medio del cual se puede extraer información útil de las páginas web mediante librerías como: beautifulsoap, selenium, scrapy, etc.

Por otro lado, la metodología más utilizada actualmente para proyectos de minería de datos tanto en la academia como en la industria es la Cross Industry Standard Process for Data Mining más conocida como CRISP-DM (DaimlerChrysler - SPSS - NCR, 1996), la cual es completamente libre, se puede ajustar a las necesidades del proyecto y permite determinar las actividades específicas a realizar en cada etapa del proceso. Consta de 6 etapas que pueden ser ajustadas de acuerdo con las necesidades y/o el problema a resolver. Cabe resaltar que para el presente proyecto solo se tendrán en cuenta las primeras 5 etapas de dicha metodología, ya que la última: despliegue, no hace parte del alcance. En la figura 1 se puede apreciar cómo funciona esta metodología.

Aquí se observa que primero empieza con la parte de entendimiento del negocio para luego continuar con la parte de entendimiento de los datos, seguido por la preparación de estos. Cabe destacar que estas tres etapas son fundamentales para el desarrollo de un buen modelo de datos y que aporte verdadero conocimiento y por ende la mayor parte de tiempo del proyecto debería estar concentrada en estas etapas. La siguiente etapa es modelado y por ultimo evalaución

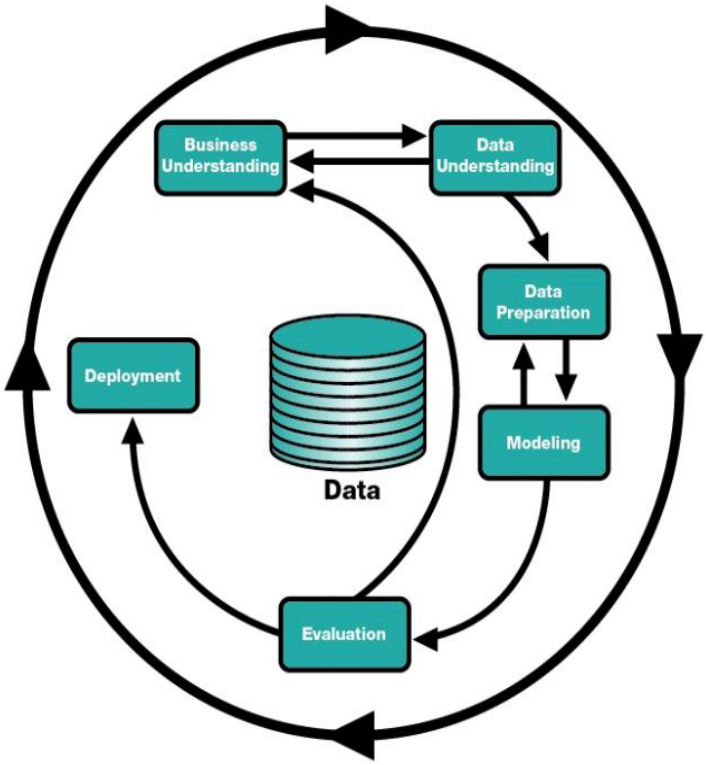


Figura 3. Metodología Crisp - DM para desarrollo de proyectos de minería de datos. Copyright 2012 por Smart Vision Europe

Por otro lado, los algoritmos de machine learning y estadísticos más utilizados para desarrollar modelos de regresión son:

- **Regresión Lineal:** Procedimiento estadístico que se utiliza para determinar la relación entre dos variables y la fuerza de dicha relación, además de predecir una variable determinada en función de otras (Cardona, González, Rivera, & Cárdenas, 2013).

- **Árboles de decisión:** Son una organización jerárquica del espacio de representación, donde cada nodo interior contiene una pregunta sobre un atributo concreto y cada nodo hoja se refiere a una decisión (clase).

- **KNN (k Vecinos Cercanos):** Almacenan el conjunto de entrenamiento y al realizar la clasificación se busca en los ejemplos almacenados casos similares y se asigna la clase más probable en éstos.

- **RandomForest:** La idea básica es generar múltiples modelos en un conjunto de datos de entrenamiento y luego combinar (en promedio) sus reglas de salida. Es una técnica popular de Ensembling que se usa para mejorar el rendimiento predictivo de árboles de Decisión.

- **Redes Neuronales:** Son inspiradas en los sistemas biológicos, adaptados y simulados en computadoras. Las redes neuronales intentan aprender mediante ensayos repetidos como organizarse mejor a sí mismas para conseguir maximizar la predicción.

- **Máquinas de soporte vectorial para regresión (SVMR):** Es una mejora a la máquina de soporte vectorial (SVM), la cual utiliza los mismos principios que la SVM para la clasificación, aunque con algunas diferencias debido a que la salida es un número real,

entonces se hace complicado predecir la información disponible, que tiene infinitas posibilidades. Por esto, es necesario establecer un margen de tolerancia (epsilon) en aproximación al SVM. A pesar de esto, la idea principal es siempre la misma: minimizar el error, individualizando el hiperplano que maximiza el margen (Sayad, 2020).

Es de resaltar que estos algoritmos ya se encuentran implementados en las librerías scikit learn (https://scikit-learn.org/stable/) del lenguaje de programación Python.

4.3. Marco legal

Ley 527 de 1999: Por medio de la cual se define y reglamenta el acceso y uso de los mensajes de datos, del comercio electrónico y de las firmas digitales, y se establecen las entidades de certificación y se dictan otras disposiciones.

Ley 1581 de 2012: Protección de datos personales. Por la cual se dictan disposiciones generales para la protección de datos personales.

La ley 1712 de 2014 por medio de la cual se crea la ley de transparencia y del derecho a la información pública nacional y se dictan otras disposiciones; tiene como objeto regular el derecho a la información pública, los procedimientos para el ejercicio y la garantía del derecho y las excepciones a la publicidad de la información.

4.4. Estado del arte

En este apartado se presentan los casos de estudio más recientes relacionados con la utilización de técnicas de machine learning para la predicción de precios de vivienda en diferentes lugares del mundo.

Park & Bae (2015) utilizaron algoritmos de machine como como C4.5, RIPPER, Naive Bayes y AdaBoost para desarrollar un modelo de predicción de precios de la vivienda, para ayudar a los agentes inmobiliarios a tomar decisiones con mejor información. Ellos utilizaron alrededor de 5359 datos de casas del Condado de Fairfax, Virginia. Los análisis dieron como resultado que el algoritmo RIPPER es mejor que los otros modelos en cuanto a desempeño de la predicción del precio de la vivienda. En aplicaciones prácticas, este modelo también puede ayudar a las instituciones financieras para realizar una mejor evaluación de las propiedades inmobiliarias, análisis de riesgo y préstamos Este trabajo presenta los diferentes algoritmos que pueden llevarse a cabo para la predicción del precio de viviendas y que pueden ser de utilidad en el presente proyecto de grado, además se resalta la importancia de hacer predicciones de precios de vivienda.

Por otro lado, Lim, Wang, & Chang (2016) compararon el desempeño de las Redes Neuronales Artificiales (RNA), series de tiempo ARIMA (Media Movil Integrada Autoregresiva) y Regresión lineal Múltiple para predecir el índice de precios de condominios en Singapur. En este caso las variables tenidas en cuenta para construir los modelos fueron macroeconómicas: gasto del consumidor, salario promedio mensual, producto interno bruto (PIB), índice de precios al consumidor, tasa de interés preferencial, tasa de interés real, población, índice de precios de reventa, cambio del índice de precios de

reventa de HDB, Straits Times Index, el número de condominios disponibles y variables de características de las casas: precios de los condominios, area, numero de baños y cuartos, edad, distancia, colegios, centros comerciales y centros de salud. Los resultados mostraron que las RNA muestran más alta precisión en la predicción que los demás modelos utilizados. Este análisis da una señal de como las RNA pueden ser un buen modelo para determinar los precios de vivienda. Sin embargo, el mejor modelo depende de la preparación de los datos y de las variables tenidas en cuenta, por esto en este proyecto se estudiarán varios modelos y de acuerdo con esto se determinará el mejor.

Otro caso donde las RNA han demostrado tener un buen desempeño en la predicción de precios de vivienda fue en una investigación realizada en la ciudad de Boston – Estados Unidos, en la cual, se diseñó una Red Neuronal Artificial de 2 capas para entrenar un modelo de regresión multivariable con un algoritmo de propagación hacia atrás. Los resultados obtenidos de la predicción están muy cerca de los datos que se utilizaron en el test set. Se encontró que las variables más relevantes son la tasa de criminalidad, la tasa de impuestos a la propiedad y la proporción de alumnos por maestro (Wang, y otros, 2018).

Otras técnicas utilizadas para la predicción de precios de vivienda y que han sido efectivas son SVM y Stepwise, el cual es un método de uso común para la selección de subconjuntos. Esto lo demuestra un estudio realizado en la ciudad de Melbourne en Australia, en el cuál, partiendo de datos históricos y utilizando diferentes técnicas de machine learning se encontró que estos dos eran los mejores, además se concluyó que las variables más importantes para predecir el precio de las viviendas fueron el tipo de casa, la latitud y la distancia, seguidos por el número de cuartos. El análisis se realizó en el lenguaje de programación R (Phan, 2019). Este

documento da cuenta de los métodos de machine learning más utilizados para predecir precios de vivienda, lo cual será útil para tenerlo en cuenta dentro de este proyecto.

Como se mencionaba anteriormente, existen varios factores que afectan el precio de la vivienda, y pueden ser variados dependiendo de la localización de la vivienda y sus características. Por tanto, Febrita, Alfiyatin, Taufiq, & Mahmudy (2018) diseñaron un sistema para extraer reglas difusas (funciones de membresía y reglas de inferencia), que se pudiesen usar para predecir los precios de la vivienda según la ubicación de los objetos cercanos. Se utilizó el método de agrupamiento de K-Means para extraer valores iniciales y para formar funciones de membresía difusas y reglas de inferencia de varios grupos de residenciales. Esta investigación produce un sistema difuso de buena interpretabilidad que muestra un resultado satisfactorio de las predicciones. Este método puede ser útil a utilizar en este proyecto para hacer agrupaciones de viviendas según su ubicación dentro de un municipio.

Por el lado de la obtención de datos para la predicción de precios de vivienda se ha encontrado que estos pueden ser extraídos de diferentes fuentes y formas entre las cuáles se encuentra el webscraping. Por esto cabe mencionar un trabajo, donde se coleccionaron 11 millones de datos de renta de casas en Estados Unidos de la página craiglist (página de avisos clasificados). Los datos fueron extraídos entre mayo y julio de 2014. Se desarrollaron herramientas para limpiar, extraer, organizar y analizar los datos. El lenguaje de programación utilizado fue Python, tanto para extraer los datos como para analizarlos. El procedimiento de webscraping consistió en visitar la página web que tenía datos de renta de casas, posteriormente se llevó a formato HTML en el servidor web. Luego el scraper extrajo todos los

elementos de HTML usando Xpath Query Language (Kay, 2008), finalmente el scraper guardó los elementos como estructurados. Los datos coleccionados fueron: región, precio de renta, área, renta por área, número de cuartos, entre otros (Boeing & Waddell, 2017). Este análisis es bastante similar a como se tiene planteado el presente proyecto, además de que brinda una idea de cuánto tiempo puede tardar en promedio la recolección de los datos que sería en promedio entre 2 y 3 meses, sin embargo, es de considerar que el número de renta de casas puede ser mayor al de venta.

El webscraping no es algo nuevo y se viene utilizando desde hace tiempo, ejemplo de ello es que mediante técnicas de webscraping buscaron comparar los precios de varios sitios de comercio electrónico y llevarlas a una herramienta en línea para una búsqueda más efectiva. El framework de la herramienta se diseña utilizando HTML y CSS como front-end y PHP como back-end. Los scripts de scraping utilizan el lenguaje Python y el scraping funciona en etiquetas HTML. La novedad de esta herramienta es que los resultados se obtienen dinámicamente y se muestran cada vez que el usuario ingresa la consulta, lo cual ayuda a mejorar la capacidad de almacenamiento y procesamiento. Además, la precisión del proceso de recuperación de datos es del 93% con cálculos mínimos y menos tiempo (Mehak, Zafar, Aslam, & Bhatti, 2019). Este es de gran utilidad para mirar el proceso que se llevó a cabo para extraer los datos desde diferentes páginas web. Aunque en este caso fue para varios productos que no eran necesariamente casas, se podría realizar solo para viviendas.

Así mismo, el webscraping se ha utilizado para recolectar imágenes de vehículos. Para este caso, uno de los procesos de scraping llevado a cabo fue: Primero, buscaron la imagen en Google y extrajeron el código fuente HTML. Posteriormente extrajeron la URL y se hizo la validación respectiva. Solo se almacenaron las URL validadas. Finalmente descargaron

las imágenes de las URL validadas (Lee & Kang, 2019). Este documento permite evidenciar un proceso de web scraping, pero aplicado a imágenes específicas buscadas en Google, mientras que lo que se quiere en este trabajo es extraer los datos directamente desde las URL de las páginas web que contienen los datos de precios y características de viviendas.

En línea con lo anterior, en China utilizaron datos de la página web de la empresa de bienes raíces líder en el país para investigar tendencias espaciotemporales de las variaciones de los precios de vivienda. Se realizaron tres estudios, primero se identificó la distribución espacial de los precios de la vivienda a nivel micro. Luego, se observó la dinámica espacio - temporal de las propiedades residenciales en el mercado y tercero, se investigó si existe disparidad geográfica en términos de precios de la vivienda. Los resultados mostraron que los cambios en los precios de vivienda eran altamente relevantes para la ubicación de las viviendas nuevas, lo cual permitió una mejor comprensión de la estructura y dinámica de los mercados inmobiliarios en las ciudades chinas. (Li, Ye, Lee, Gong, & Qin, 2017).

En Colombia, también se construyó una base de datos con la información de precios y características de venta o arriendo de casas o apartamentos nuevos o usados en 5 ciudades de Colombia (Bogotá, Medellín, Cali, Barranquilla y Bucaramanga), mediante webscraping. Las características tenidas en cuenta fueron precios, ubicación área del inmueble, estrato, número de habitaciones y baños, estado, antigüedad, tipo de inmueble, entre otras. Debido a que ciertos anuncios estaban disponibles solo en un periodo de tiempo corto (por ejemplo, menor a un mes), los datos fueron descargados cada 15 días para recolectar el mayor número de información posible de las páginas de Internet. Una vez se descargó la información, se compilaron los datos en una base de datos para cada una de

las ciudades para cada tipo de inmueble (apartamentos y casas) y para cada tipo de negocio (venta y arriendo). Consecuentemente, en total se construyeron veinte bases de datos, con las cuales se realizó un análisis descriptivo de dichos datos (Rubio, Guzmán, & Otero, 2019).

Anteriormente se han realizado trabajos de extracción de datos de vivienda de páginas web en el país, sin embargo, solo se extrajo la información y se realizó un análisis descriptivo mas no la realización de un modelo predictivo que es lo que se pretende en este proyecto. Además, es importante aclarar que el mercado inmobiliario está en constante movimiento, ya que depende mucho de la oferta y la demanda, por tanto, la información también cambia constantemente.

En general estos trabajos muestran que el tema de predicción de precios de vivienda es un tema actual y de interés para muchas ciudades y países, además los estudios muestran diferentes modelos de machine learning como RNA, NaiveBayes, regresión múltiple, máquinas de soporte vectorial, entre otros para predecir los precios de vivienda en las diferentes ciudades estudiadas. Por esto, en este proyecto también se trabajará con diferentes modelos para determinar cuáles tienen mejor desempeño para lograr el objetivo.

Otro punto para resaltar es que como se evidenció en el estado del arte, aún no se ha trabajado con datos de precios de vivienda relacionados específicamente con el municipio de Rionegro que por su desarrollo exponencial y por la alta demanda de compra y venta de viviendas necesita un modelo específico para que las personas puedan tomar mejores decisiones de inversión y venta de viviendas.

5. METODOLOGÍA

La metodología cuantitativa tiene su foco en los aspectos observables que son susceptibles de cuantificación, y utiliza la estadística para el análisis de los datos (Centro Virtual Cervantes, 2019).

Este proyecto se realizó en base a una metodología cuantitativa ya que se presentaron eventos como la predicción de precios y además se utilizaron técnicas estadísticas y de machine learning para el análisis de los datos.

Además, se siguieron las etapas de la metodología CRISP-DM, ampliamente utilizada tanto en la academia como en la industria para este tipo de proyectos y con la cual se buscó alcanzar los objetivos del presente trabajo.

Las etapas llevadas a cabo fueron:

9.1. Extracción y entendimiento de la información relacionada con los precios de vivienda (casas y apartamentos) de las páginas web del municipio de Rionegro

Esta parte del proyecto abarcó las etapas de entendimiento del negocio o del problema y entendimiento de los datos.

Debido a que no se contaba con los datos para realizar la predicción de los precios de vivienda, se recurrió a la técnica de web scraping, para lo cual es necesario analizar aspectos cómo: accesibilidad de los datos de origen y análisis de patrones de los datos. Los pasos que se siguieron para la extracción de los datos fueron:

a. Visitar el sitio web y examinar la estructura de los datos en los sitios web a estudiar. Un ejemplo de una de las páginas web, se puede observar en la figura 4: Básicamente está compuesta por la imagen de la casa, el valor, el área privada y construida, precio por metro cuadrado, el estrato, estado, antigüedad, sector, mapa de ubicación, servicios públicos y una descripción de esta donde se muestra: número de pisos, número de baños, si tiene patio o no, parqueadero, patio y zona de ropas, algunas tienen a los lugares que están cercanas, etc. Las páginas que se utilizaron en este trabajo fueron:

   - Finca raíz (https://www.fincaraiz.com.co/)
   - Mercado libre(https://www.mercadolibre.com.co/?matt_tool=26606016&matt_word=mercado-libre&gclid=EAIaIQobChMIkoGHmLSs5gIVZf7jBx38BgMqEAAYASAAEgKQQPD_BwE)

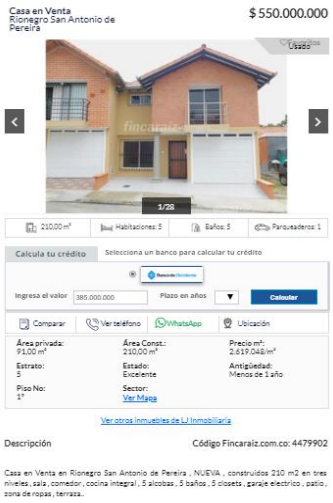


Figura 4. Muestra de datos de una vivienda en la página de Finca Raíz. Copyright 2019 Finca Raíz

b. Se ubicaron los datos que se querían extraer dentro de los múltiples niveles de etiquetas HTML. En pocas palabras, había una gran cantidad de código en una página web y se quería encontrar las piezas de código relevantes que contienen los datos necesarios.

c. Por último, se generó el código de Python necesario para extraer los datos.

Debido a que eran varias páginas web, fue necesario extraer varias bases de datos ya que las páginas eran ligeramente diferentes en su estructura. Por esto se construyó una base de datos por página web y posteriormente se realizará una unión de los datos para obtener una sola base de datos.

Una vez se extrajeron los datos de las páginas web, el siguiente paso fue convertirlos en datos estructurados para que pueda ser evaluado la calidad y confiabilidad, mediante los siguientes criterios:

- Significado de cada columna de la base de datos.
- De qué tipo son las variables
- Cuáles son los posibles valores de las variables.
- Variables de interés
- Porcentaje de datos faltantes en cada columna relevante.
- Porcentaje de duplicados en cada columna y en la base de datos general.
- Detección inicial de outliers

5.1. Análisis descriptivo de los datos

Es parte es la base para obtener los conjuntos de datos que serán la entrada para cada uno de los modelos de machine learning que se analizarán posteriormente. Esta etapa contiene las siguientes sub - etapas que se realizaron con el fin de obtener datos de calidad para realizar el análisis predictivo y serán explicadas más adelante en el documento

Subetapa	Descripción	Actividades
1	Eliminar variables irrelevantes y redundantes para el análisis	Se descartaron aquellas variables que no fueron relevantes para lograr los objetivos y el alcance del proyecto.
2	Limpieza y calidad de los datos.	Se eliminaron datos duplicados exactos para disminuir el tamaño de la data.
3	Descripción estadística de los datos. Análisis univariable.	Se realizaron descripciones estadísticas para variables numéricas y tablas de contingencia para categóricas. Se construyeron Histogramas y diagramas de dispersión para variables numéricas. Se realizó la interpretación de los resultados. Estos análisis se realizaron con el fin de encontrar

		patrones, tendencias o relaciones entre los datos.
4	Análisis de correlaciones. Análisis bivariable	Correlaciones entre variables. Correlación con la variable objetivo. Interpretación de los resultados.
5	Tratamiento de outliers	Para el tratamiento de outliers, se eliminaron las observaciones que tuvieran valores por encima o por debajo de los rangos normales.
6	Tratamiento de datos nulos	Se realizó una imputación de las variables que contenían menos del 50% de datos nulos con la mediana. Las variables con más del 50% de datos nulos se eliminaron.
7	Reducción o creación de nuevas de variables	Se creó la variable tipo de vivienda a partir de la descripción inicial de los resultados.

Tabla 1. Etapas para la preparación de los datos

Durante el desarrollo de esta etapa, se pueden encontrar diferentes estructuras de datos que conlleven a la generación de diferentes estrategias de modelamiento.

5.2. Generar modelos de machine learning para predecir precios de las viviendas

En esta etapa se construyeron los modelos analíticos que dan respuesta al objetivo propuesto. También se evaluó la calidad de los modelos construidos en función de las predicciones aportadas por el modelo vs los datos reales y la calidad de los agrupamientos. Es importante resaltar que existen gran diversidad de técnicas de aprendizaje automático para predecir y agrupar datos, por lo cual, en esta etapa, fue muy importante plantear estrategias de modelamiento que mejor se adapten a los objetivos y a las características de los datos.

Una buena estrategia en este caso fue iniciar con modelos simples e interpretables que permitan ir ampliando el conocimiento del efecto de las variables y su relación con la variable objetivo. Es por esto por lo que se utilizaron los siguientes modelos en su orden:

a. Regresión Lineal.

b. Arboles de decisión de regresión.

c. Máquinas de soporte vectorial (SVM).

d. Random Forest

e. Gradient Boosting Machine (GBM).

5.3. Definición del modelo que más se ajuste a los datos analizados y conclusiones del estudio.

En esta etapa se hizo una comparación entre los resultados de los diferentes modelos para encontrar el mejor de acuerdo con los resultados obtenidos. Dado que este es un problema de regresión, los métodos de evaluación fueron los siguientes:

a. MAPE: Error Porcentual Absoluto Medio. Es una medida que es sensible a los ceros de los datos.

b. Coeficiente de determinación (R2): Es una medida de la relación lineal entre dos variables. Entre mayor sea su valor indica un mejor ajuste del modelo encontrado a los datos.

Los métodos de cálculos de cada una de estas medidas se encuentran en la tabla 2.

Medidas de error	
MAPE = $\frac{\sum_{i=1}^{n} \frac{\lvert y_i - \hat{y}_i \rvert}{y_i}}{n}$ (1)	R2 = $R^2 = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$ (2)

Tabla 2. Medidas de error

Donde:

y_i: Dato real, $\hat{y}_i$: Dato estimado, n: total de observaciones

6. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

6.1. Consolidación de información.

6.1.1. Extracción de información

En esta etapa se indagó en las páginas web de Finca Raíz, metro cuadrado y mercado libre para obtener datos de casas, apartamentos y fincas que están en venta en Rionegro. Se emplearon las librerías Beautiful Soup (Richardson, 2004) y Selenium (Huggins, 2004) de python entre otras, para obtener los datos. La primera se utilizó para extraer los enlaces de cada una de las viviendas y la segunda para extraer los datos específicos de cada vivienda. La información fue llevada a archivos csv para su posterior procesamiento.

0	https://www	4617440	Apartament	$ 355.000.00	73.71	3	2	1	apriv68,00 a	DescripciÃƒÂ	1	a Rionegro San Antonio d	68,00
1	https://www	4630318	Casa en Ven	$ 345.000.00	154	4	4	1	apriv154,00	DescripciÃƒÂ	2	gro El Porvenir	154,00
2	https://www	2076556	FOREST APA	Desde $ 333	65	2		2		DescripciÃƒÂ	1	Sector RÃƒÂ-o Golf - Rionegro Calle 40E	
3	https://www	4640678	Finca en Ver	$ 2.300.000.(	14720	3	2	NaN	area 14,72 H	DescripciÃƒÂ	3	egro Vereda	14,72
4	https://www	4624799	Apartament	$ 173.000.00	58.3	2	2	1	apriv58,30 a	DescripciÃƒÂ	1	a Rionegro Horizontes de	58,30
6	https://www	4918112	Casa en Ven	$ 780.000.00	220	3	4	NaN	apriv310,00	DescripciÃƒÂ	2	gro porvenir	310,00
7	https://www	3698025	Ventus Apar	Desde $ 330	66.29	2		1		DescripciÃƒÂ	1	Balcones II - Rionegro Calle 41 48AA-15	
8	https://www	4727145	Apartament	$ 190.000.00	57	3	2	1	aconst57.00	DescripciÃƒÂ	1	a Rionegro fontibon	

Tabla 3. Datos obtenidos de vivienda

Entre los datos extraídos se encuentra: Código de la propiedad en el sitio web, barrio donde se oferta la vivienda, número de habitaciones, número de baños, número de garajes, la descripción de la vivienda que aparece en el sitio web, así como un “otros” que contemplan aquellos datos que están en general en ese sitio web como por ejemplo si la casa tiene Jacuzzi o no. Estos datos dependen de la página web de la cuál son extraídos.

Los datos que se obtuvieron de la página de finca raíz son presentados en la tabla 4.

Variable	Descripción	Codificación
URL	url del sitio donde se encuentra ubicada la vivienda específica	url
Código	Código de propiedad asignado por la página	Cod
Barrio	Barrio de residencia	barrio
Precio	Precio de la vivienda	price
Área	Área de la vivienda	area
Número de cuartos	Número de cuartos	room
Número de baños	Número de baños	bath
Número de garajes/parqueaderos	Variable numérica entera	garaje
Tipo de vivienda	Casa finca/campestre, apartamento, finca, casa urbana	tipo_vivienda
Área privada	Variable numérica de tipo float	apriva
Área construida	Variable numérica de tipo float	aconst
Precio del m2	Variable numérica de tipo entero	preciom2

Administración	Variable numérica de tipo entero	admon
Estrato	2, 3, 4, 5, 6	estrato
Estado	excelente, buena y remodelar	esta
Antigüedad	Menos de 1 año, entre 1 y 8 años, entre 9 y 16 años, entre 16 y 30 años y más de 30 años.	anti
número de pisos	Cantidad de pisos de la vivienda	piso

Tabla 4. Variables extraídas de finca raíz.

En la tabla 5 se presentan los datos extraídos de la página web de mercado libre.

Variable	Descripción	Código
URL	url del sitio donde se encuentra ubicada la vivienda específica	url
Barrio	Variable de tipo string	barrio
Precio	Variable numérica de tipo entero	price
Área	Variable numérica de tipo float	area
Número de cuartos	Variable numérica entera	room
Número de baños	Variable numérica entera	bath
Número de garajes	Variable numérica entera	garage
Tipo de vivienda	Casa finca/campestre/urbana, apartamento, finca	tipo_vivienda

Área construida	Variable numérica de tipo float	aconst
Administración	Variable numérica de tipo entero	admon
Antigüedad	Menos de 1 año, entre 1 y 8 años, entre 9 y 16 años, entre 16 y 30 años y más de 30 años.	antg
número de pisos	Cantidad de pisos de la vivienda	Piso

Tabla 5. Variables extraídas de mercado libre

Es importante destacar que, dentro de los datos extraídos de las viviendas, se encuentran tres tipos de vivienda: apartamento, casa, finca. Estas podrían ser analizadas por separado de manera que no haya sesgos en los precios, ya que en la mayoría de los casos el precio de una finca puede llegar a doblegar el precio de un apartamento, sin embargo, debido a la cantidad de datos, estos tipos de vivienda serán analizados por igual. El total se extrajo 2315 datos de la página web de Finca Raíz, de los cuáles, el 50% corresponde a los apartamentos. Y de mercado libre se extrajeron en total 728 datos, de los cuáles el 56% corresponde a los apartamentos.

Se obtuvieron bases de datos por cada página y luego se unieron de manera que se pudiera trabajar con solo una base de datos. Donde no había datos disponibles tanto de una página como de la otra, se tomaron como datos nulos, de manera que las bases de datos pudiesen unirse. En total quedaron 18 columnas de variables.

Para más información de este apartado ver ANEXO 1.

6.1.2. Preparación de las bases de datos

Después de la extracción de los datos se observaron los siguientes puntos que hacían que fuera imposible hacer un análisis adecuado de los datos:

- Existían datos que a pesar que los datos demandaban que fuera una columna de enteros, esto no sucedía de esta manera, por ejemplo, en la columna parqueaderos (garage), en vez de tener un registro con 1 o 5, se tenían registros: *Entre 1* y *5.* Ya que se tenían pocos registros similares, y además no se tenía certeza de cuál de los dos valores era "*verdadero"* entonces se optó por eliminar dichos registros.

- Los tipos de vivienda fue una variable que se tuvo que crear, a partir de la variable barrio, mediante la extracción de los primeros caracteres de esta que estaban de nominados como *Apartamento*, *Casa campestre*, *Casa*, *Finca.*

- Dentro de la variable área existían algunos datos que no estaban en m2 si no en hectáreas, por lo cual se tuvo que hacer la conversión respectiva.

- Dentro de las variables *room* y *bath,* también existían algunos datos que no eran coherentes con la lógica del campo, por ejemplo, *Desde 3* o *Desde 2,* por tanto, también se realizaron ajustes a esto campos.

- Algunos datos estaban contenidos en una sola columna por lo que hubo que dividirlos en varias para poder realizar los respectivos

análisis. Este fue el caso del área construida, el área privada, el estrato, el estado, entre otros.

garage	boxcube
1	apriv68,00 aconst73,71 preciom2: 4.816.171/admon$180,000 Estrato: 5 Estado: Excelente anti 1 a 8 aÃ±os Piso No: 10Âº Sector: Ver Ma
1	apriv154,00 aconst154,00 preciom2: 2.240.260/Estrato: 3 Estado: Bueno anti 16 a 30 aÃ±os Piso No: 3Âº Sector: Ver Mapa
2	

Tabla 6. Datos contenidos en una misma columna.

- Debido a las divisiones que se realizaron en el paso anterior, muchos datos quedaron en columnas erróneas, por ejemplo, en área privada se podrían encontrar datos de estrato o estado, por esta razón se tuvo que realizar un tratamiento para que todos los datos quedaran en sus respectivas columnas.

	vacio	apriva	aconst	preciom2	admon	estrato	esta	antg	otro1
0		68,00	aconst73,71	preciom2:	4.816.171/admon$180,000	Estrato:	5	Estado:	Excelente
1		154,00	aconst154,00	preciom2:	2.240.260/Estrato:	3	Estado:	Bueno	anti
2			None	None	None	None	None	None	None

Tabla 7. Datos en la columna equivocada

	apriva	aconst	preciom2	admon	estrato	est
0	68.00000	73.71000	4816171.00000	180000.00000	5.00000	Excelen
1	154.00000	154.00000	2240260.00000	nan	3.00000	Bueno
3	nan	nan	15625.00000	nan	3.00000	Excelen

Tabla 8. Datos organizados.

Una vez se tuvo una base de datos consolidada y con los datos en los campos correctos, se realizaron las siguientes actividades para mejorar la calidad de los datos encontrados:

- Había una gran cantidad de datos duplicados, es decir, eran exactamente las mismas viviendas. Investigando en las páginas web se logró detectar que muchas de las viviendas están en varias páginas de esta debido a que son avisos “destacados”, es decir, quieren que se vean lo más posible. Estos datos fueron eliminados

- Existían campos que no eran necesarios para los análisis posteriores como url, código, etc. porque no aportaban información al modelo de machine learning.

- Cuando se realizó la división de los datos que estaban en una misma columna se encontró que había columnas que eran insignificantes para el análisis. Es el caso de columnas que tenían solo strings, por ejemplo, el *“m2”* que quedaba de extraer el área.

- Aunque existían buena parte de datos nulos especialmente en la variable *admon,* los cuáles corresponde al 77% de los datos y otro1 (número de pisos) con el 57% de los datos, no son eliminados en esta etapa ya que se estudiará más adelante cuál es el mejor tratamiento para estos.

barrio	0	barrio	0.000000
price	0	price	0.000000
area	0	area	0.000000
room	21	room	1.397206
bath	35	bath	2.328676
garage	334	garage	22.222222
description	0	description	0.000000
tipo_vivienda	0	tipo_vivienda	0.000000
barrio_or	0	barrio_or	0.000000
apriva	482	apriva	32.069195
aconst	332	aconst	22.089155
preciom2	57	preciom2	3.792415
admon	1160	admon	77.178975
estrato	276	estrato	18.363273
esta	770	esta	51.230872
antg	298	antg	19.827013
otro1	863	otro1	57.418496

Tabla 9. Cantidad y porcentaje de datos nulos

- Se realizaron conversiones de categóricas a numéricas con el fin de a facilitar los análisis posteriores, por ejemplo, la variable antigüedad, estado y barrio.

Antiguedad

```
#Antiguedad
#print(df['antg'].unique())
df['antg']= df['antg'].replace({"1Âº": "1a8", "Mapa":"", "2Âº":"1a8", "MÃ¡sde30":"mas30","Excelente":""
                                ,"Sector:":"","Piso":"", "Estado:":"", "5Âº":"1a8", "Ver":"","1-ago":"1a
8","13Âº":"9a15",
                                "Menos-1":"Menosde1","3Âº":"1a8","sep-15":"9a15", "Remodelar":"","N
o:":"","4Âº":"1a8",
                                "Bueno":"", "16-30":"16a30", "8Âº":"1a8", "Estrato:":"", "de":"", "3":"",
"1-8":"1a8", "9-15":"9a15"})
df['antg']=df['antg'].astype('category')
```

6.2. Análisis descriptivo de los datos de las viviendas

Esta etapa es fundamental para poder desarrollar un buen modelo de predicción. .

6.2.1. Análisis univariable de los datos

En esta sección se estudió cada una de las variables para observar su comportamiento particular y con esto poder establecer relaciones posteriores entre ellas. En el anexo 2, se puede ver con más detalle los análisis realizados en esta etapa, así como los códigos construidos.

	price	area	room	bath	garage	tipo_vivienda	apriva	aconst	preciom2	admon
count	1496	1496	1475	1461	1163	1496	1017	1167	1439	341
mean	1259886018	4522.62431	3.29491525	3.07255305	1.78675838	1.957887701	2060.98653	4249.1171	3680026.59	283295.666
std	1.2977E+10	116556.027	1.1760592	1.40846496	1.47879914	0.926873742	6747.86161	131741.304	3930970.8	352983.484
min	110000000	41	1	1	1	1	2	41	133	120
25%	285000000	74	3	2	1	1	72	67	2500000	130000
50%	450000000	150	3	3	1	2	126	118	3488372	200000
75%	1200000000	350	4	4	2	3	580	210	4519969.5	356000
max	5E+11	4500000	14	12	10	4	80000	4500000	106382979	5750000

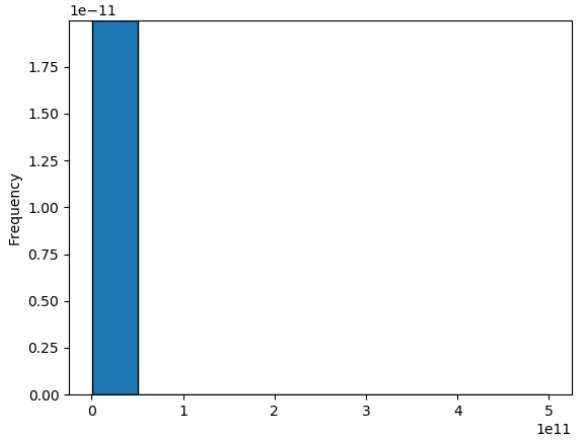
Tabla 10. Análisis Descriptivo general de los datos

Tanto en la gráfica 1 como en la tabla 10, se puede observar que la variable precio está completa, no tiene valores nulos, sin embargo, la variabilidad es alta ya que el valor mínimo es de 110 millones y el máximo de 500.000 millones. La media se mantiene en 1200 millones, mientras que la mediana es de 450 millones. En el gráfico 1 se puede observar que la mayor cantidad de precios de viviendas está concentrada en menos de 1000 millones.

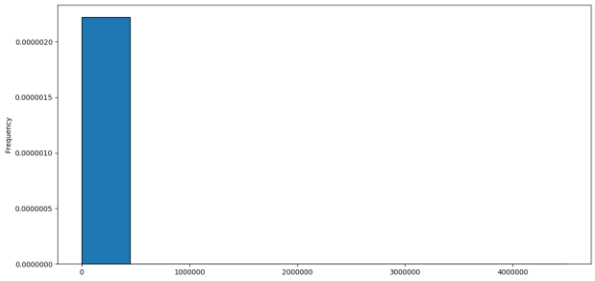
Para el área también existe una alta variabilidad ya que la media y la mediana están distanciadas en gran medida como se observa en la gráfica 2. En el caso de los cuartos y los baños (gráficas 3 y 4), la distribución es más

homogénea y tanto la media como la mediana se encuentran en 3 para cada uno.

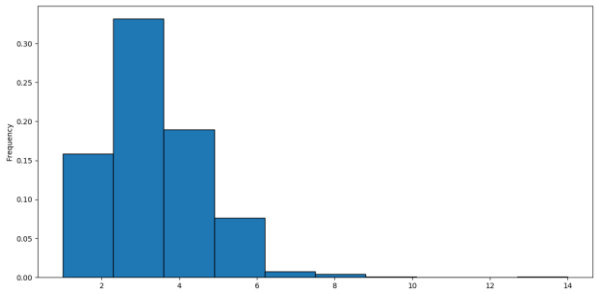
En cuanto a la distribución del número de parqueaderos tampoco se observa una variabilidad significativa (gráfica 5). Por otro lado, Apriva (gráfica 6), aconst (gráfica 7) y preciom2 (gráfica8) también se observa gran variabilidad en los datos y un importante sesgo hacia la izquierda, lo cual podría influir en los resultados del modelo.



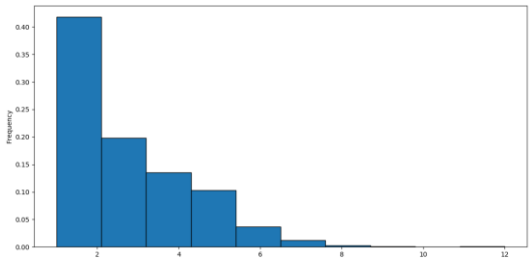
Gráfica 1. Distribución de la variable precio



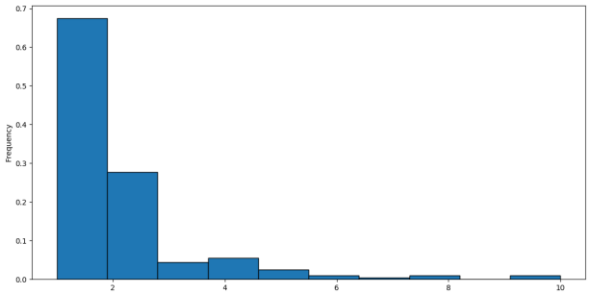
Gráfica 2.Distribución de la variable área



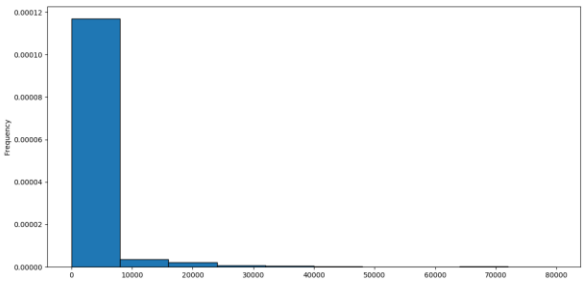
Gráfica 3. Distribución de la variable room (cuartos)



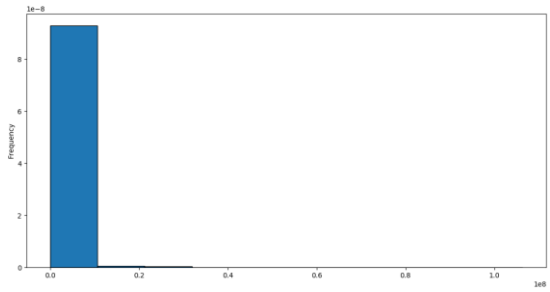
Gráfica 4. Distribución de la variable bath (baños)



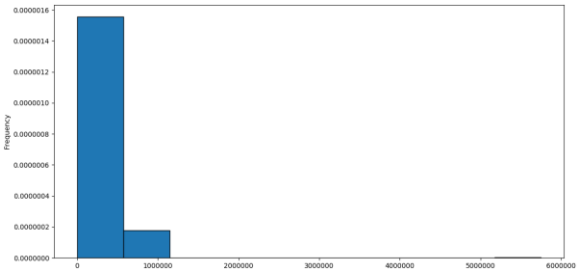
Gráfica 5. Distribución de la variable garaje



Gráfica 6. Distribución de la variable apriva (área privada)



Gráfica 7. Distribución de la variable aconst (área construida)



Gráfica 8. Distribución de la variable preciom2 (precio m2)

Luego de analizar las variables cuantitativas se procede a analizar las variables cualitativas. Para el caso del estrato, se observa que la mayor cantidad de viviendas se encuentra en estrato 4, seguido por el estrato 5. Esta variable se convierte en un punto importante para definir la efectividad y eficiencia del modelo, ya que el sistema no sería confiable, por ejemplo, para para estrato 1, que podría ser por ejemplo una finca incluso muy grande lo cual podría afectar su estimación de precio.

Así mismo se puede identificar que a la mayor cantidad de viviendas son relativamente nuevas, ya que, la antigüedad está entre 1 a 8 años, seguido por menos de 1 año. El estado de la mayoría de las viviendas es excelente, y el piso más común es 1.

	estrato	esta	antg	piso
count	1220.0	727	1199	637.0
unique	4.0	3	5	16.0
top	4.0	Excelente	1a8	1.0
freq	548.0	537	587	289.0

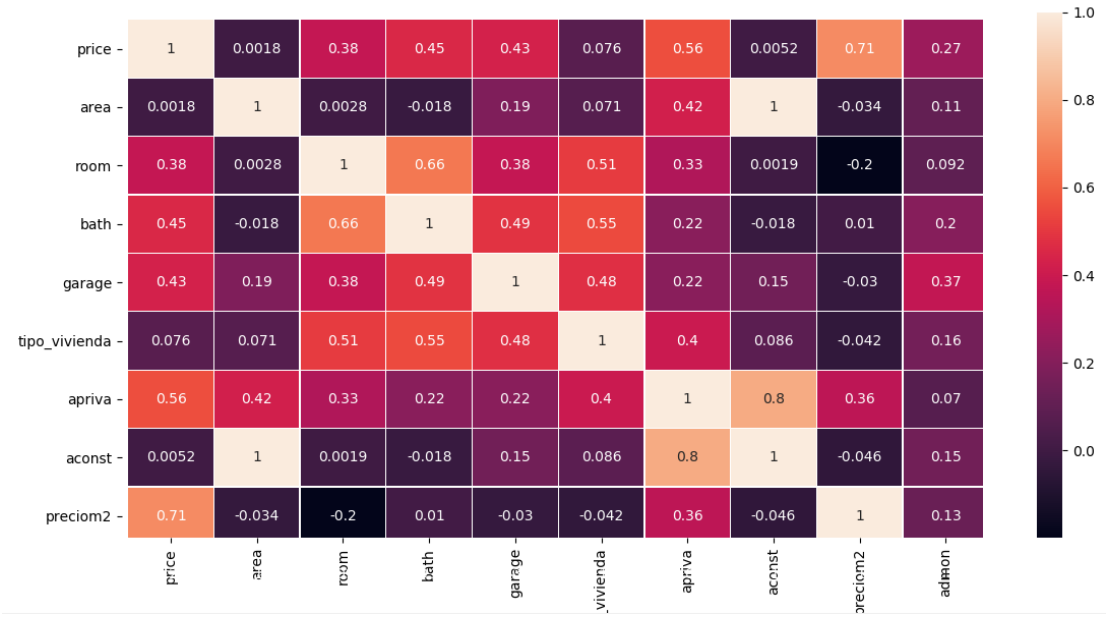
Gráfica 9. Análisis variables categóricos

En cuanto a los análisis de las variables categóricas, se puede observar que la variable más completa es el estrato, seguido por la antigüedad, por tanto, estas variables podrían ser claves para determinar el precio de una vivienda. Dado que las variables estado y piso tienen aproximadamente el 50% menos de datos que la variable objetivo, es decir, el precio, estas podrían no tener una influencia importante, aunque estas validaciones se darán posteriormente.

6.2.2. Análisis bi-variables

En esta parte se busca establecer relaciones entre las variables y poder analizar posibles combinaciones de los datos.

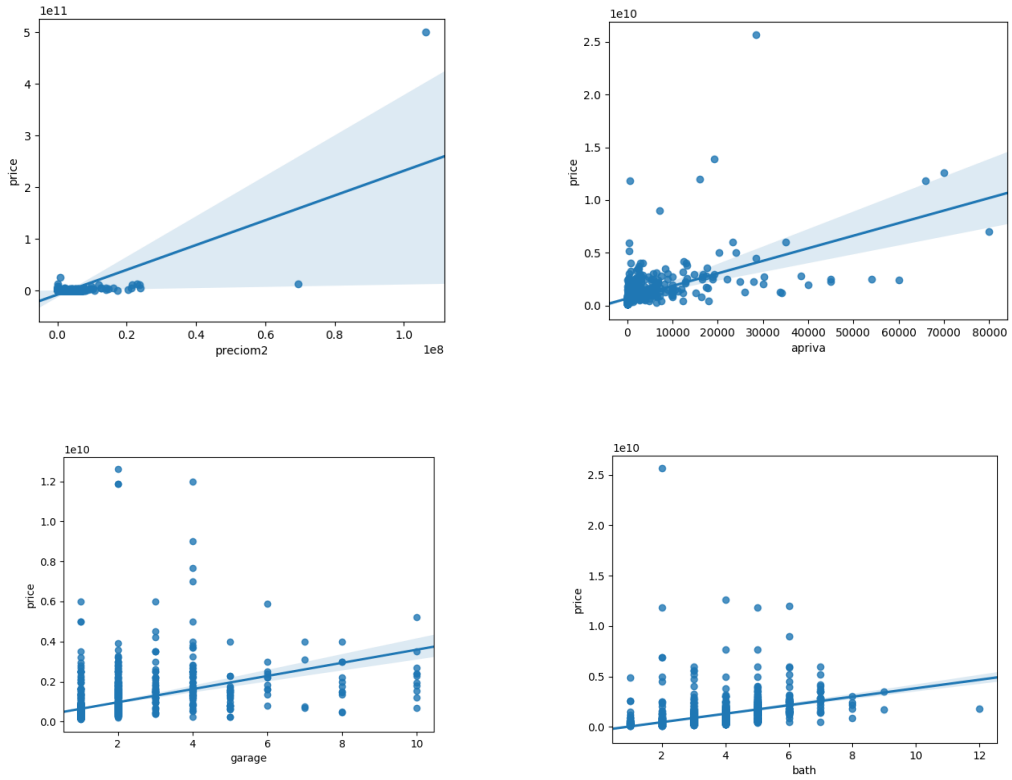
De acuerdo con el gráfico 10, se observa que las variables que más tienen relación con la variable precio son preciom2 con una correlación de 0.71, seguido por área privada, número de baños y número de garajes. Es por esto que se realizan algunos gráficos de correlaciones (ver gráfica 10) donde se observa una relación lineal entre precio, baños y garajes, sin embargo, no se ve claramente para el precio de m2.



Gráfica 10. Mapa de correlación de variables

Por otro lado, las variables con mayor correlación entre sí fueron el área construida con el área privada que para algunas viviendas podrían ser iguales, el número de baños con el número de cuartos, lo cual sería razonable dado que entre más cuartos tenga una vivienda más baños debería tener en la práctica, el tipo de vivienda con el número de cuartos también fueron otras de las variables más correlacionadas, sin embargo debieran considerarse dentro del modelo ya que además de que las correlaciones no son altas, pueden ser un factor determinante al momento de tomar la decisión del precio de una vivienda.

De acuerdo con el gráfico 11, se observa gran cantidad de datos atípicos en las diferentes variables que deberían ser eliminados para evitar la afectación al modelo, especialmente en la variable preciom2



Gráfica 11. Diagramas de dispersión

Para las variables categóricas se analizan las tablas de contingencia para detectar relaciones entre ellas (ver tabla 11)

esta	Bueno	Excelente	Remodelar	All
estrato				
3.0	46	95	2	143
4.0	70	193	2	265
5.0	35	111	0	146
6.0	9	53	1	63
All	160	452	5	617

Tabla 11. Tabla de contingencia estado y estrato

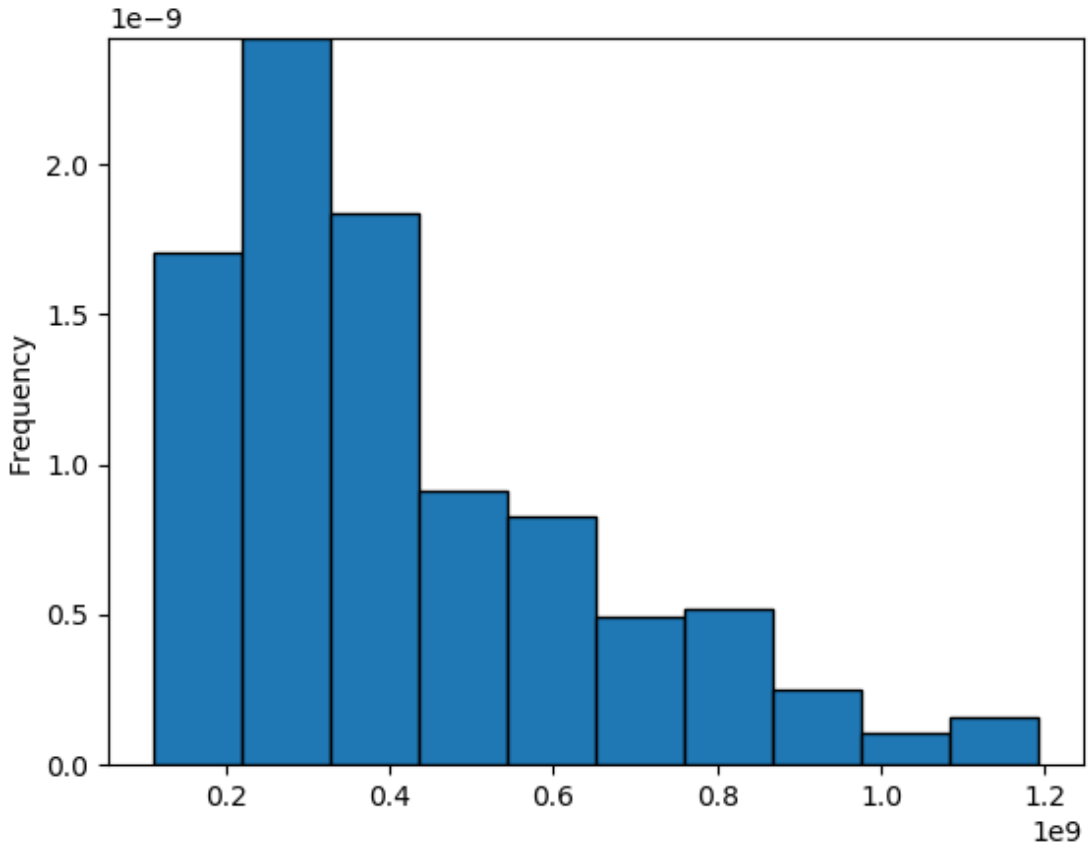
La mayoría de las viviendas en estado excelente se encuentra en el estrato 5, en remodelar hay muy pocas viviendas por lo que convendría descartar esta variable como se ve en la tabla 12.

antg	16a30	1a8	9a15	Menosde1	mas30	All
estrato						
3.0	29	90	46	33	8	206
4.0	45	272	60	113	10	500
5.0	36	138	43	67	2	286
6.0	6	36	21	21	2	86
All	116	536	170	234	22	1078

Tabla 12. Tabla de contingencia antigüedad y estrato.

6.2.3. Tratamiento de datos atípicos

Luego de analizar los datos, se eliminaron algunos atípicos relacionados con el precio de las viviendas, como se puede evidenciar en la gráfica 12.



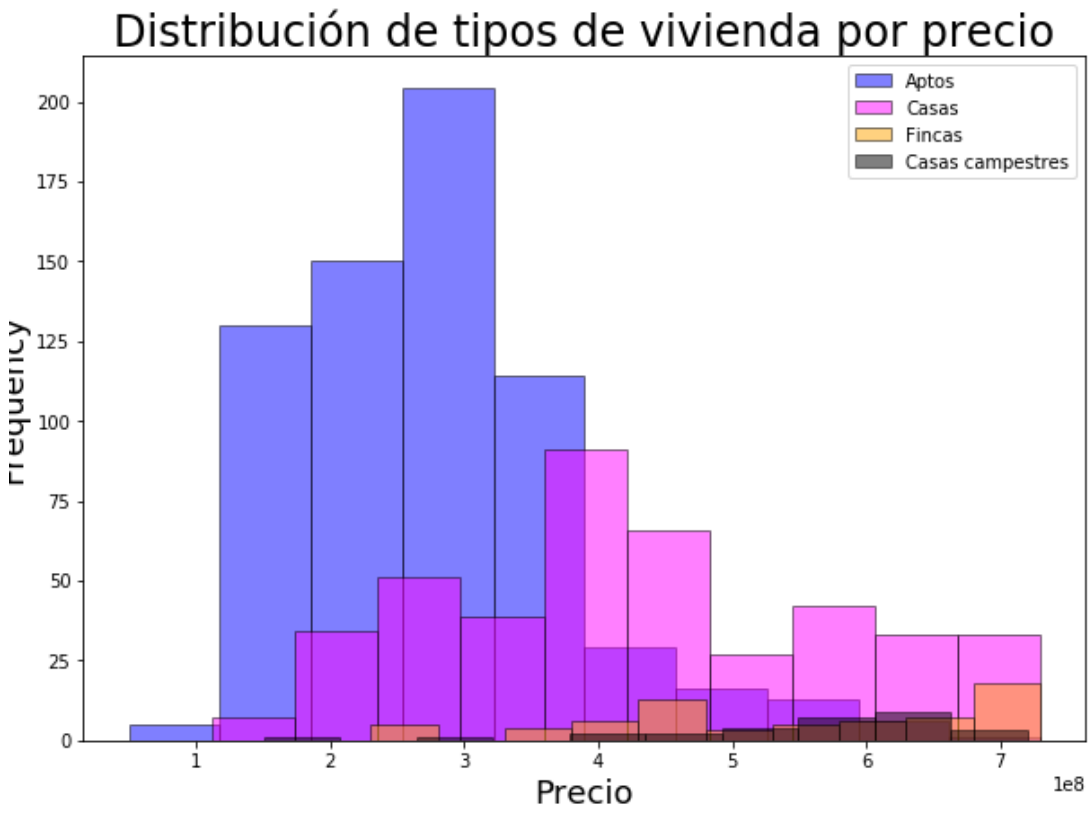
Gráfica 12. Distribución de precios de vivienda después de limpieza de datos

Al dejar solo los valores menores de 1200 millones, se encuentra que las variabilidades en los rangos intercuartiles se hacen más cortas, así mismo la mediana de baños disminuye a 2 mientras que los cuartos permanecen en 3 y garage en 1.

	price	area	room	bath	garage	ipo_vivienda	apriva	aconst	preciom2	admon
count	1116	1116	1103	1094	862	1116	764	976	1081	250
mean	420522629	4486.67922	2.99909338	2.54753199	1.35150812	1.67114695	423.183639	4795.7753	3337149.2	224973.864
std	230854418	134699.857	0.99499614	0.95124916	0.8864007	0.80598152	1492.745	144036.6	1497785.37	381942.424
min	110000000	41	1	1	1	1	13	41	133	120
25%	251500000	66	2	2	1	1	64	64	2571429	120000
50%	347500000	104.5	3	2	1	1	87	87.5	3307087	166420
75%	550000000	175	3	3	1	2	160	151	4135802	222250
max	1193465184	4500000	8	8	10	4	18000	4500000	17400000	5750000

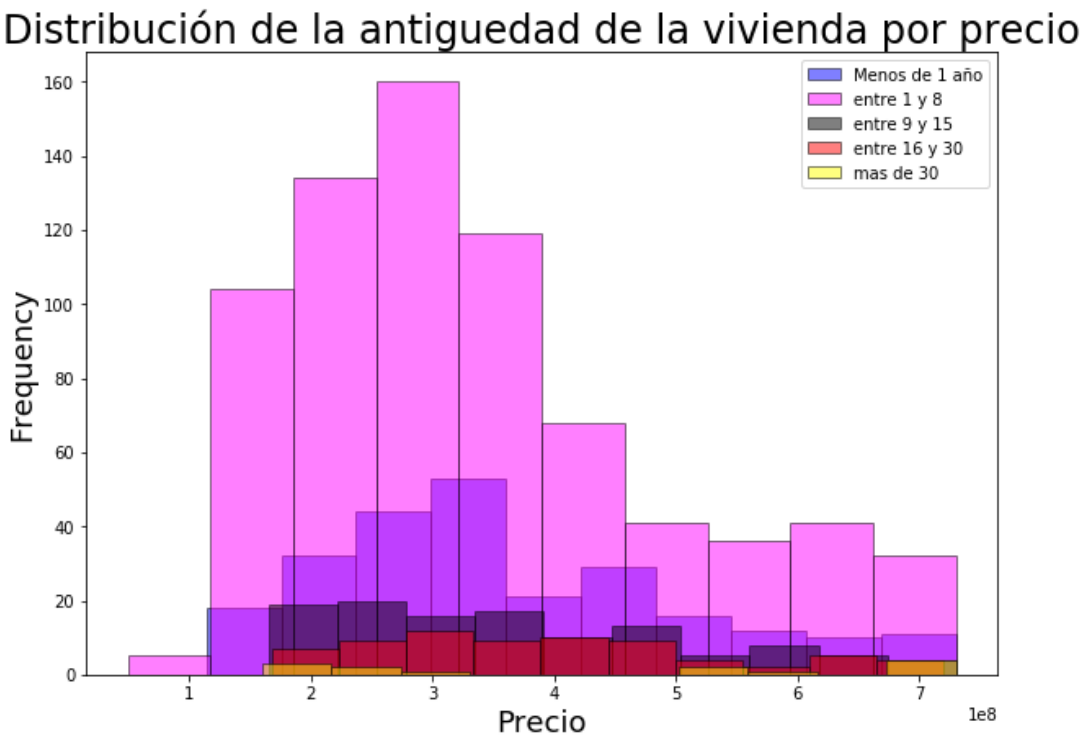
Tabla 13. Descripción estadística para precios menores de 1200 millones

Una vez se hizo el ajuste de los precios a menos de 1200 millones, se puede apreciar la distribución de los precios de vivienda por precio, donde se observa que en general los apartamentos son los de menor precio, las casas tienen una distribución más uniforme, pero son menos cantidad.



Gráfica 13. Distribución de tipos de vivienda por precio.

En el caso de la antigüedad, las viviendas entre 1 y 8 años son las de mayor cantidad y las que menos precio tienen. Las viviendas menores a un año están mejor distribuidas, pero son más pocas en comparación con las anteriores, caso particular son las viviendas con antigüedad mayor a 30 años que como hay una cantidad aislada que tienen precios menores a 300 millones hay otras aisladas que el precio llega a los 700 millones.



Gráfica 14. Distribución de la antigüedad de la vivienda por precio

6.2.4. Tratamiento de datos nulos

De acuerdo con la tabla 14 y como se había mencionado en apartados anteriores y de acuerdo con las distribuciones encontrados y los rangos intercuartiles, en los datos, existen varios datos nulos, entre las variables más destacables en este tema se encuentran: *admon* con cerca del 70% de los datos, *apriva* con cerca del 30% y *garage* con cerca del 22%. La primera variable fue eliminada completamente ya que eran demasiados nulos y si se imputaban los datos quedaría un sesgo importante, mientras que las ultimas fueron imputadas porque si se eliminaban estas variables, el modelo quedaría sin dos variables que pudieran influir significativamente en el precio. Se llenan con la mediana porque sigue existiendo una diferencia importante

entre la media y la mediana, a pesar de haber eliminado los datos atípicos presentes en los datos.

	price	area	room	bath	garage	ipo_vivienda	apriva	aconst	preciom2	admon
count	1116	1116	1116	1116	1116	1116	1116	1116	1116	1116
mean	420522629	4486.67922	2.99910394	2.53673835	1.27150538	1.67114695	317.147222	4205.13144	3336206.39	179536.905
std	230854418	134699.857	0.98917872	0.94489126	0.79275524	0.80598152	1244.69036	134699.775	1474099.42	182137.905
min	110000000	41	1	1	1	1	13	41	133	120
25%	251500000	66	2	2	1	1	74	66	2600000	166420
50%	347500000	104.5	3	2	1	1	87	87.5	3307087	166420
75%	550000000	175	3	3	1	2	120.25	141.25	4102564	166420
max	1193465184	4500000	8	8	10	4	18000	4500000	17400000	5750000

Tabla 14. Tratamiento de datos nulos

6.3. Modelamiento de los datos.

Antes de proceder con el modelamiento de los datos, se realizó una partición de datos con el 80% de los datos para entrenamiento y 20% de los datos para prueba.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
print(X_train.shape)
print(y_test.shape)
```

Posteriormente, fueron implementando los modelos de machine learning de acuerdo con su complejidad, inicialmente se inició con un modelo de regresión lineal, luego se realizó la predicción con árboles de decisión, posteriormente con random forest, gradient bosting machine y por último se utilizó una máquina de soporte vectorial.

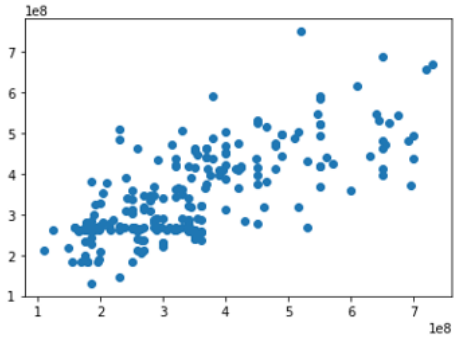
De acuerdo con los resultados obtenidos en cada uno de los modelos, se encontró que el modelo con el mejor desempeño fue GB, seguido por Random Forest ya que presentan los mayores R2 con respecto a los demás. El modelo con el menor desempeño fue *Suport Vector Machine,* ya que su R2 es el menor seguido por la regresión lineal.

Se elige medida para determinar el desempeño de los modelos el R2, ya que es un indicador que permite determinar que tanto se ajusta el modelo al precio de las viviendas y por tanto es de fácil interpretabilidad.

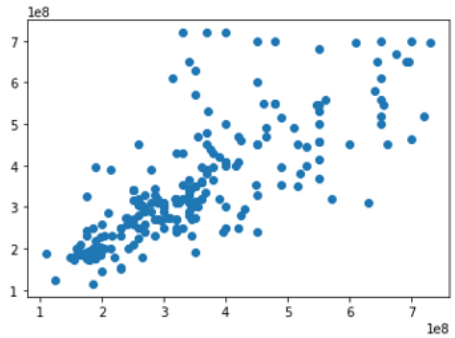
	R2	RMSE	MAE	MAPE	SMAPE
Regresión Lineal	0.530519	9.658456e+07	7.552920e+07	0.249191	0.109879
Decision Tree	0.551575	9.439382e+07	6.075658e+07	0.176933	0.088676
Random Forest	0.702875	7.683654e+07	5.148748e+07	0.152312	0.074897
Gradient Bosting Machine	0.710440	7.585210e+07	5.510707e+07	0.167122	0.080208
SVM	0.034413	1.385142e+08	1.064484e+08	0.339682	0.154847

Tabla 15. Resultados modelos de machine learning

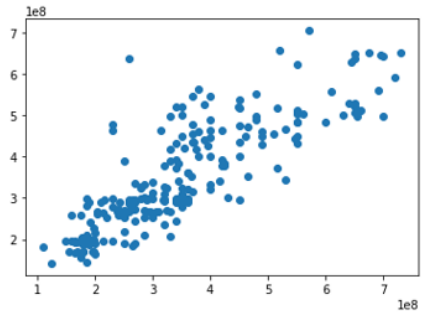
Sin embargo, cabe destacar que las medidas de MAPE son relativamente bajas, lo que confirma que los modelos con mejores desempeños son Random Forest y Gradient Boosting.

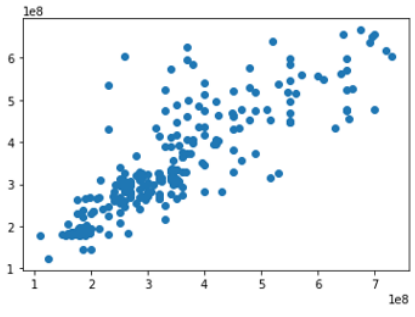


Predicciones Regresión Lineal



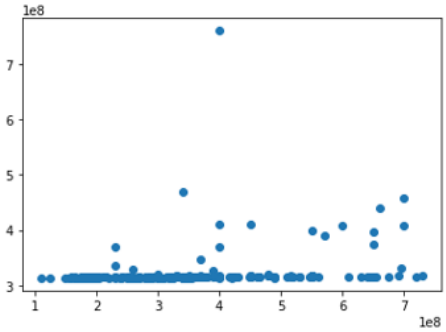
Predicciones árboles de decisión





Predicciones Gradient Boosting

Predicciones Random Forest



Predicciones Máquina de soporte vectorial

Gráfica 15. Resultados modelos de predicción

Luego de obtener los resultados anteriores, se procedió a realizar una validación cruzada con 10 conjuntos de prueba para los modelos con menor R2 encontrado, de manera que se pueda garantizar que los resultados son independientes del conjunto de prueba.

```
Cross-validation scores: [0.82007193 0.78296132 0.70891265 0.59836476 0.80252586 0.67771331
 0.80998721 0.65028433 0.75746135 0.79569438]
Average cross-validation score: 0.74
```

Gráfica 16. Resultados Cross Validation modelo random forest

```
#Validación cruzada para 10 grupos aleatorios de datos gradient boosting
from sklearn.model_selection import KFold,cross_val_score
scores = cross_val_score(gradient_boosted, X, y, cv=10)
print("Cross-validation scores: {}".format(scores))
print("Average cross-validation score: {:.2f}".format(scores.mean()))
```

```
Cross-validation scores: [0.80227522 0.77194485 0.73888581 0.63177276 0.80076383 0.72844551
 0.82625837 0.67745147 0.72407652 0.75014975]
Average cross-validation score: 0.75
```

Gráfica 17. Resultados Cross Validation modelo gradient boosted

Comparando los resultados de los diferentes conjuntos de entrenamiento, obtenidos a través de validación cruzada, tanto para el random forest como para el gradient boosting, se muestra que en general, los modelos obtenidos son estables, ya que la varianza entre los diferentes resultados es de 0.003 para el modelo obtenido a través de gradient bosting y de 0.005 el obtenido para random forest. De esta manera cualquiera de los dos modelos podría ser aplicable para la predicción de precios de viviendas.

Adicionalmente, se realizó una hiperparametrización del modelo para determinar los parámetros del gradient boosting que mejor se ajustan a los datos; estos pueden evidenciarse en el gráfico 18.

```
0.7785428520517582
{'learning_rate': 0.1, 'max_depth': 6, 'max_features': 3, 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 100}
```

Gráfica 18. Resultados Hiperparametrización

Luego de correr el modelo con estos resultados, se encontró que el modelo mejora en un 2%.

Posteriormente se procedió a verificar cuáles son las variables que el modelo de gradient boosting considera más importantes para determinar el precio de una vivienda. Los resultados se muestran en la tabla 16.

```
#Selección de caarcteristicas para el mejor modelo de random forest encontrado
for name, importance in zip(b, gradient_boosted.feature_importances_):
    print(name, "=", importance)
```

```
area = 0.28351808959154357
room = 0.0754250420162363
bath = 0.053274324785183784
garage = 0.05759291486796008
apriva = 0.14443244916229012
aconst = 0.09995553351300661
barrio = 0.04744573620167542
tipo_vivienda = 0.05011764758119632
estrato = 0.08796116295238988
esta = 0.01963514287996115
antg = 0.045821752809432154
piso = 0.034820203639124525
```

Tabla 166. Importancia de variables para gradient boosting

Las variables más relevantes para determinar el precio de una vivienda son el área de la casa, el área privada, el área construida y el estrato de esta.

```
#Selección de caarcteristicas para el mejor modelo de random forest encontrado
for name, importance in zip(b, regressorRF2.feature_importances_):
    print(name, "=", importance)
```

```
area = 0.31507286932359135
room = 0.024382082377121084
bath = 0.09345710633235858
garage = 0.05203410820979993
apriva = 0.07871204181558535
aconst = 0.205174445677024
barrio = 0.016361500687163695
tipo_vivienda = 0.10188214040403817
estrato = 0.08404468923920619
esta = 0.002771008932015859
antg = 0.015703808566301918
piso = 0.010404198435794484
```

Tabla 17. Importancia de variables para random forest

Así mismo se observa la importancia de variables para el segundo mejor modelo encontrado, random forest. Aquí se observa que para dicho modelo la variable área es más significante que para gradient boosting al igual que el área construida, el tipo de vivienda y el número de baños parecen ser más importantes que el estrato para este modelo (ver tabla 17).

Lo anterior entra un poco en discrepancia con el análisis bi-variable realizado, donde se evidenciaba que el área al igual que el área construida tenía muy poca correlación con la variable objetivo, es decir, el precio de la vivienda. El área privada (apriva) si se encuentra tanto correlacionada con la variable precio como de importancia para ambos modelos.

7. CONCLUSIONES

- Mediante la extracción de información de páginas web, se logró evaluar un modelo de gradient boosting que pudiese predecir el precio de las viviendas en el municipio de Rionegro, para que de esta manera se pueda apoyar a las personas en la decisión de invertir o comprar vivienda en este municipio. Además, este modelo es un primer acercamiento a la dinámica de los bienes raíces en este municipio ya que permite determinar en qué rango se mueven y se podrán mover los precios de vivienda en este municipio de rápido crecimiento. Es de anotar que este modelo aún es susceptible de mejora de manera que cada vez haya más precisión sobre la inversión en propiedad raíz en dicho municipio.

- Se logró consolidar una base de datos estructurada para el posterior análisis y construcción de los modelos de machine learnig a través de la información extraída de las páginas web de mercado libre y finca raíz, sin embargo, se hizo necesario realizar un importante trabajo de limpieza de los datos ya que no todos los datos podían obtenerse de manera estructurada. La base de datos obtenida contó con 3033 registros y 18 columnas, incluyendo la url de la vivienda y el código de esta. Es importante destacar que a pesar de que las bases de datos de ambas páginas web no contaban con las mismas columnas, se logró una unión de estas, a pesar de que quedó aproximadamente un 50% de los datos como nulos o duplicados.

- Los modelos que mejor desempeño presentaron fueron gradient bosting machine y random forest, ya que arrojaron R2 de 0.75 y 0.77 respectivamente. Las variables que más influyen en la predicción de

los precios de vivienda en el municipio de Rionegro, de acuerdo con los modelos son el área de la vivienda, el área construida, el área privada, el tipo de vivienda, el estrato y el número de baños, sin embargo, estos resultados no siempre concordaron con el análisis bi variable realizado ya que algunas variables como el área que no tenía una correlación lineal con el precio, terminó siendo la variable más importante al momento de tomar la decisión del precio de una vivienda y otras como el estrato que también tenía correlaciones bajas con el precio de la vivienda, para los modelos si fue relativamente importante. En el caso del área privada, por ejemplo, además de encontrarse correlacionada con la variable objetivo, fue importante para determinar el precio de una vivienda, con esto se ratifica la importancia de esta variable al momento de tomar una decisión de compra o venta de vivienda en el municipio de Rionegro.

- Crear un modelo de datos para la predicción de precios de las viviendas en el municipio de Rionegro usando aprendizaje de maquina (machine learning), mediante la extracción de información disponible en páginas web para apoyar la toma de decisiones informada de compra y venta de propiedad raíz.

8. TRABAJOS FUTUROS

Se sugiere para trabajos futuros en el tema, tener en cuenta las siguientes temáticas:

- ✓ Abarcar más páginas web de manera que se pueda tener un cubrimiento más amplio de las viviendas que están a la venta en el municipio de Rionegro.

- ✓ Tener un periodo más prolongado de recolección de los datos, de manera que se pueda evaluar cómo van cambiando los precios a través del tiempo y si es que verdaderamente el municipio se está valorizando o es solo una *"burbuja"* o comportamiento atípico

- ✓ Tener en cuenta otros municipios aledaños para observar los cambios en el comportamiento entre municipios y así tener más información para tomar decisiones de inversión.

9. REFERENCIAS

Bibliografía

Alcaldía de Rionegro. (2016). *Plan de Desarrollo 2016 - 2019.* Recuperado el 10 de 10 de 2019, de https://www.rionegro.gov.co/Documents/Plan%20de%20Desarrollo%202016-2019.pdf

Centro Virtual Cervantes. (2019). *Metodología cuantitativa.* Recuperado el 9 de Marzo de 2020, de https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/metodologiacuantitativa.htm

Oriente Comercial Digital. (2017). *El presente y el futuro de la construcción se edifica.* Recuperado el 10 de Octubre de 2019, de http://www.orientecomercialdigital.com/sitio/noticias_detalle.php?id=588

Sayad, S. (2020). *Máquina de vectores de soporte - Regresión (SVR).* Recuperado el 10 de Octubre de 2019, de https://www.saedsayad.com/support_vector_machine_reg.htm

Alisenda Inmobiliaria. (01 de enero de 2016). *El proceso de compra de una vivienda explicado paso a paso.* Recuperado el 20 de Septiembre de 2019 de https://www.alisedainmobiliaria.com/blog/el-proceso-de-compra-de-una-vivienda-explicado-paso-a-paso/

Argos. (20 de 04 de 2017). *El Oriente Antioqueño es el futuro Medellín.* Recuperado el 20 de Septiembre de 2019 de http://grandesrealidades.argos.co/oriente-antioqueno-futuro-medellin/

Boeing, G., & Waddell, P. (2017). New Insights into rental housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listing. *Journal of planning education and research*.

Cámara de Comercio del Oriente Antioqueño (CCOA). (2018). *El presente y futuro de la construcción se edifica en el oriente.* Recuperado el 05 de mayo de 2019, de http://www.orientecomercialdigital.com/sitio/noticias_detalle.php?id=588

Cámara de Comercio del Oriente Antioqueño. (05 de 04 de 2018). *Concepto económico del oriente antioqueño.* Recuperado el 05 de mayo de 2019 de https://www.ccoa.org.co/Portals/0/Biblioteca%20virtual/Publicaciones%20regionales/2018/Concepto%20econ%C3%B3mico%202018.pdf?ver=2019-02-01-105326-537

Cardona, D. F., González, J. L., Rivera, M. L., & Cárdenas, E. H. (2013). Aplicación de la regresión lineal en un problema de pobreza. *Revista interaccion, 12*, 73-84. Recuperado el 12 de junio de 2019 de http://www.unilibre.edu.co/revistainteraccion/volumen12/art4.pdf

DaimlerChrysler - SPSS - NCR. (1996). *CRISP-DM.* Recuperado el 10 de mayo de 2019, de http://crisp-dm.eu/home/about-crisp-dm/

Deloitte. (2016). *Bienes Raíces y Transacciones de Inversión Inmobiliaria.* Recuperado el 05 de mayo de 2019, de https://www2.deloitte.com/content/dam/Deloitte/mx/Documents/bienes-raices/Bienes-Raices-Folleto-2016.pdf

Kay, M. (2008). *XSLT 2.0 and XPath 2.0 Programmer's Reference.* Indianapolis.

Kho, J. (26 de septiembre de 2018). *How to Web Scrape with Python in 4 Minutes.* Recuperado el 03 de mayo de 2019, de https://towardsdatascience.com/how-to-web-scrape-with-python-in-4-minutes-bc49186a8460

Lee, Y., & Kang, S.-J. (2019). Web Scraping Crawling-based Automatic Data Augmentation for Deep Neural Networks-based Vehicle Classifications. *2019 IEEE International Conference on Consumer Electronics, ICCE 2019.* Las Vegas.

Li, S., Ye, X., Lee, J., Gong, J., & Qin, C. (2017). Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective. *Applied Spatial Analysis and Policy*, 421-433.

Lim, W., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.* Changsha; China.

López de Mesa, A. M. (29 de abril de 2019). Así se pinta el desarrollo urbanístico en Antioquia. *El colombiano*.

Mehak, S., Zafar, R., Aslam, S., & Bhatti, S. (2019). Exploiting filtering approach with web scrapping for smart online shopping: PPenny Wise: A wise tool for online shopping. *2nd International Conference on Computing, Mathematics and Engineering Technologies.* Sukkur, Pakistan: iCoMET 2019.

Murillo, D., & Saavedra, D. (2017). Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R. *4to Congreso Internacional AmITIC 2017.* Popayan, COlombia.

Phan, T. (2019). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. *Proceedings - International Conference on Machine Learning and Data Engineering.* Sydney, Australia: iCMLDE 2018.

Propiedades Oriente Raíz. (14 de enero de 2019). *Valor del metro cuadrado en el Oriente Antioqueño 2019.* Recuperado el 1 de junio de 2019, de https://www.orienteraiz.co/blog/valor-metro-cuadrado-oriente-antioqueno-2019/

Rubio, J., Guzmán, F., & Otero, J. (2019). Una base de datos de precios y características de vivienda en Colombia con información de Internet. *Revista de Economía del Rosario*, 75-100.

Van Rossum, G. (1989). Python. Países Bajos.

Villa, M. (28 de abril de 2018). *¿Por qué es importante invertir tu dinero?* Recuperado el 01 de mayo de 2019, de http://forbes.es/business/42248/por-que-es-importante-invertir-tu-dinero/

Wang, J., Hu, S., Zhan, X., Luo, Q., Yu, Q., Liu, Z. (2018), Predicting House Price with a Memristor-Based Artificial Neural Network. *IEEE Access*, 16523-16528.

. ANEXO 1

EJEMPLO PYCHARM PROJECT CON LA EXTRACCIÓN DE LOS DATOS

Para la extracción de los datos se trabajó con el programa Pycharm y código Python, con este fue construido el código utilizado para la extracción de las páginas web empleadas: finca raíz y mercado libre.

El código se encuentra en el repositorio de GitHub señalado en el enlace: https://github.com/VanesaGrajales/preciosviviendas

ANEXO 2

JUPYTER NOTEBOOK CON EL ANÁLISIS DE LOS DATOS Y LOS MODELOS DE PREDICCIÓN.

En este Jupyter se encuentran los análisis realizados a los datos para obtener los resultados presentados en este trabajo. En la primera parte se presentan toda la limpieza realizada a los datos extraídos de la web, en la segunda parte se encuentra el análisis estadístico realizado para la compresión de los datos, en la tercera parte se encuentran los modelos de machine learning empleados para realizar la predicción de los precios de vivienda, así como su respectiva evaluación, en la penúltima parte se encuentra la validación cruzada y la hiperparametrización realizada al mejor modelo obtenido y en la última parte se puede evidenciar un resumen de los resultados obtenidos así como la conclusión sobre el mejor modelo.

El código se encuentra en el repositorio de GitHub señalado en el enlace: https://github.com/VanesaGrajales/preciosviviendas