Predicción de precios de arriendos de viviendas en la ciudad de Medellín en base a información recolectada a través de Web Scraping

Walter Arboleda Castañeda

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Efraín Alberto Oviedo, Magíster (MSc) en Tecnologías de la Información y la Comunicación

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	Arboleda Walter [1]
Referencia Estilo IEEE (2020)	[1] C. Arboleda Walter, “Predicción de precios de arriendos de viviendas en la ciudad de Medellín en base a información recolectada a través de Web Scraping”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA)

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

MSE.	Mean Square Error
RMSE.	Root Mean Square Error
MAE	Mean Absolute Error
R2	R Squared
MAPE	Mean Absolute Percentaje Error
LR	Linear Regression
RFr.	Random Forest Regression
SVMr	Suport Vector Machine Regression
MLPr	Neural Network MLP Regressor
MM	Miles de Millones

RESUMEN

Con el fin de aumentar la oferta de sistemas predictores para las empresas se crea un modelo que permite estimar el precio de arriendo de viviendas (casas y apartamentos) en la ciudad de Medellín Colombia. El desarrollo es realizado en dos etapas, la primera consta de crear un script que permita recolectar la información a través de web scraping y que este sea diseñado de manera que facilite su ejecución automática y periódica. La segunda etapa consta de crear un modelo de regresión evaluado en diferentes zonas de la ciudad que evidencie dónde es más conveniente su implementación. Esta decisión se toma en base a la siguiente métrica propuesta por el negocio: *"El modelo es implementable en una zona si el MAPE es menor o igual a 15%"*. El modelo es el resultado de siete iteraciones siendo la primera la línea base de la solución y la séptima la implementación de un modelo auxiliar de clustering que permite dividir el conjunto de entrenamiento y así ejecutar un modelo de regresión por cada clúster resultante. Esto le otorga al negocio una estrategia de desarrollar múltiples modelos de regresión, cada uno con la métrica buscada cuyo filtro o zona de ejecución será determinada por un modelo de clasificación.

Palabras clave — **regresión, clasificación, clustering, scraping, machine learning, datos, xgboost, kmeans.**

ABSTRACT

In order to increase the supply of predictive systems for companies, a model is created to estimate the rental price of homes (houses and apartments) in the city of Medellín Col. The development is carried out in two stages, the first one consists of creating a script that allows the collection of information through web scraping and this is designed in a way that facilitates its automatic and periodic execution. The second stage consists of creating a regression model evaluated in different areas of the city that shows where its implementation is most convenient. This decision is made based on the following metric proposed by the business: *"The model is implementable in an area if the MAPE is less than or equal to 15%"*. The model is the result of 7 iterations, the first being the baseline of the solution and the seventh, the implementation of an auxiliary clustering model that allows dividing the training set and thus running a regression model for each resulting cluster. This gives to the business a strategy of developing multiple regression models, each one with the searched metric whose filter or execution zone will be determined by a classification model.

Keywords — **regression, classification, clustering, scraping, machine learning, data, xgboost, kmeans.**

I. INTRODUCCIÓN

Con el objetivo de aumentar los servicios de una empresa, los sistemas de predicción de precios otorgan una ventaja competitiva al ofrecerle al usuario sistemas que de manera fácil, oportuna y automática le ayudan a invertir su patrimonio o cubrir una necesidad básica. *"The uses of price prediction include increasing customer loyalty and engagement. By accurately reflecting the prices of different goods and services using accurate data, customers are more likely to trust the company and its business processes"* [1].

Al implementar estos sistemas se debe resolver dos grandes retos, el primero es garantizar una obtención de datos constante, confiable y de actualización automática en base a las dinámicas del mercado en el que este será implementado y segundo un modelo de predicción con métricas que cumpla con las expectativas del negocio a cargo de administrarlo.

En este documento se hará el desarrollo de un sistema de predicción de arriendos de viviendas en la ciudad de Medellín, cuyos datos serán obtenidos a través de web scraping donde el objetivo es construir un modelo de regresión que le permita al usuario estimar el valor de alquiler bien sea en la búsqueda de un lugar donde residir o en la estimación del precio de una inversión.

Este sistema debe contar con un script ejecutable de manera periódica y automática que permita actualizar y guardar un histórico de la información. Por otra parte el modelo será construido mediante múltiples iteraciones donde al utilizar diferentes tipologías e hiperparámetros se busca alcanzar las mejores métricas de desempeño.

Adicional a esto, este modelo será evaluado en diferentes zonas de la ciudad de Medellín, ya que el cambio de precios según los estratos socioeconómicos de las diferentes comunas, en las cuales se encuentra dividida la ciudad, y las características de las viviendas entre y al interior de estas puede sesgar la predicción y limitar su funcionalidad. De esta manera se le garantizará al negocio una implementación más confiable.

Todo el detalle desarrollado en este proyecto se encuentra almacenado en un repositorio público de GitHub con el siguiente enlace: https://github.com/wacGitHub07/udea-monografia.git

II. PLANTEAMIENTO DEL PROBLEMA

El planteamiento principal de este proyecto está basado en si es posible crear un sistema de predicción de precios cuyo comportamiento sea similar al mercado y que sea este susceptible a las dinámicas sociales y económicas que afectan estos valores, de manera que le permita al usuario hacer la predicción de precio de alquiler de manera confiable.

Para la construcción de este modelo y del proceso de recolección de datos debemos resolver los siguientes cuestionamientos:

Datos:

- ¿Cuáles son los métodos existentes de scraping que cumplan con las expectativas de usabilidad del negocio?
- ¿Cuáles con los sitios con información de arriendos en la ciudad de Medellín y cuáles de estos ofrecen una información completa y actualizada?
- ¿Cuál es el tratamiento adecuado para los datos obtenidos?

Modelo:

- ¿Cuáles son las arquitecturas disponibles para los modelos de regresión?
- ¿Cuáles son las métricas a evaluar?
- ¿Cuál es la estrategia para la evaluación por zonas?

III. JUSTIFICACIÓN

Este proyecto es seleccionado porque permite dar solución a una necesidad empresarial desde los frentes de ingeniería y ciencia de datos, ofreciéndole al usuario un proceso automático que lo asiste en:

- Estimar los precios de arriendo de vivienda según sus características para proyectar su próxima inversión.
- Encontrar su próxima residencia de acuerdo a sus expectativas en comodidades y presupuesto.

Adicional a esto, permite la construcción de un proyecto completo de machine learning, haciendo una recolección propia de datos desde la web, iterando distintos tipos de algoritmos y explorando diferentes alternativas que permita ajustar un modelo hasta cumplir con la necesidad planteada.

IV. OBJETIVOS

A. Objetivo general

Desarrollar e implementar un proceso de recolección de datos a través de web scraping sobre el mercado de precios de alquiler de inmuebles de la ciudad de Medellín. El resultado del scraping debe ser la base insumo para la construcción de un modelo de predicción de precios. Este modelo debe ser evaluado sobre diferentes zonas de la ciudad y ser implementado sobre aquellas donde la métrica MAPE sea menor o igual a 15%.

B. Objetivos específicos

- Explorar y seleccionar el sitio web que cumpla con la cobertura suficiente de precios de arriendos en la ciudad para satisfacer las expectativas del negocio.
- Seleccionar y desarrollar una estrategia de web scraping que permita una usabilidad intuitiva y buena mantenibilidad.
- Crear un modelo resultante de múltiples iteraciones entre topologías, tratamiento de datos y tuneo de hiperparámetros.
- Evaluar el desempeño del modelo en diferentes zonas de la ciudad para recomendar al negocio dónde es más prudente hacer su implementación.

VI. MARCO TEÓRICO

El desarrollo de este proyecto se apoya principalmente de dos trabajos que otorgan una guía de cómo abordar un problema de regresión y que metodología es más efectiva, también ayudan en la referencia de cómo abordar un proyecto a cerca del desarrollo de modelos de predicción de precios de viviendas:

- La predicción de precios de arriendo en la ciudad de Dhaka (Bangladés), es un proyecto dividido en dos principales fases, la primera enfocada en el entrenamiento de modelos para obtener una línea base con los algoritmos de regresión MLPr, SVMr, Lasso, Linear, Elastic Net, Ridge y Decision Tree. La segunda consiste en experimentar diferentes estrategias de ensamble para obtener una predicción más robusta, estos ensambles consideran Bagging, AdaBoosting, Gradient Boosting y Ensemble XGBoost donde las métricas sobre las cuales se enfocan los resultados son R2 y RMSE. Adicional a lo anterior se realiza un tratamiento de datos donde se trabajan distribuciones y correlaciones que apoyan en encontrar un mejor modelo que se ajuste a los datos. Este trabajo nos otorga una guía de como realizar el proceso evolutivo en la construcción del modelo mediante iteraciones hasta encontrar la métrica que se busca, pues en su línea base obtienen un RMSE promedio de 282473.462 hasta llegar a un 0.1864 en la iteración final [2].

- *"Estimating Warehouse Rental Price using Machine Learning Techniques"*. [3] es un proyecto cuya metodología consiste en mostrar la implementación de diferentes tipologías de modelos al desarrollar un proyecto de predicción de precios de alquiler donde se otorga una guía de cómo observar e interpretar los datos y cuáles son los algoritmos a considerar en un problema de regresión para hacer un comparativo de los resultados. En este proyecto se realiza un análisis de importancia de variables con el modelo final para otorgar más interpretabilidad a los resultados.

  Los algoritmos considerados son: Linear Regression, Regression Tree, Random Forest Regression y Gradient Boosting Regression Trees. La evaluación de los modelos se centra

en la métrica RMSE donde el modelo de Regression Tree es quien obtiene mejores resultados.

Adicional a estas referencias, las decisiones y camino elegido para abordar este proyecto se apoyan en dos trabajos adicionales:

- *"House Price Prediction using a Machine Learning Model: A Survey of Literature*". [4] Realiza una recopilación de artículos previos sobre la predicción de precios de arriendo donde se muestra cuáles son los algoritmos mas utilizados y las métricas más comúnmente utilizadas. Según el análisis realizado en trabajos previos algoritmos como Multiple Regression Analysis, SRV, XGBOOST y Artifial Neutral Network, son los algoritmos mas populares en la predicción de precios de arriendo, donde la métrica más evaluada es RMSE.

- "Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi". [5] Se aborda un problema de predicción de precios de alquiler con datos de 11 deferentes países, lo cual para este proyecto es un referente de "zonas". En este se resalta la metodología que considera el manejo de hiperparámetros para la regularización, penalización y control de sobre muestreo. En los resultados obtenidos los algoritmos de Boosting, Bagging, Forest, Ridge y Lasso son los que presentan mejores resultados enfocándose los en los valores obtenidos por la métrica MSE.

VII. METODOLOGÍA

La metodología de este trabajo se divide en dos grandes fases, scraping para la obtención de los datos y modelado para el desarrollo del modelo de predicción.

Scraping:

El desarrollo del sistema de recolección de información se realiza a través del lenguaje Python implementando una de las siguientes librerías:

- **Beautifull Soup:** Es una librería de Python para extraer datos de archivos HTML y XML.
- **Selenium:** Librería que admite lenguaje Python y que permite interactuar con sitios web que utilizar javascript para cargar su contenido.

La decisión del uso de estas será determinada por la complejidad del sitio sobre el cual se desee extraer la información. La complejidad será medida por las políticas de extracción de datos del sitio, la información contenida en el archivo *robots.txt* y la estructura de los datos de cada vivienda [6].

De los sitios explorados se seleccionará uno que cumpla con los criterios de:

- Cobertura de la ciudad
- Cantidad de información individual de las viviendas
- Cantidad de registros

Es importante resaltar que el modelo se construirá con la información de uno de los sitios web explorados ya que extraer datos de varios de estos se expone el sistema a información repetida no rastreable y la necesidad de hacer una homologación de datos que dilataría la solución al problema planteado.

Los sitios explorados son:

- https://www.espaciourbano.com/
- https://www.metrocuadrado.com/
- https://fincaraiz.com.co/

Modelo:

El desarrollo del modelo será mediante la implementación de diferentes iteraciones, donde cada una de ellas debe buscar mejor rendimiento del modelo que la anterior. Las iteraciones se definen acuerdo a la intervención sobre el proceso y el objetivo de esta:

- **Iteración 1**: Crear un modelo de línea base como punto de partida para un primer vistazo de datos y métricas.
- **Iteración 2**: Realizar tratamiento a los datos mediante atípicos, distribuciones, correlaciones y depuración. Con este tratamiento ejecutar de nuevo el modelo de la iteración anterior para observar mejoras.
- **Iteración 3**: Implementar nuevas tipologías de modelos con el fin de observar la arquitectura de predicción adecuada y una comparación entre métricas.
- **Iteración 4:** Implementar estrategias de ensamble de modelos, con el objetivo de robustecer el sistema de predicción y así tratar de lograr la métrica MAPE exigida por el negocio**.**
- **Iteración 5:** Realizar un entrenamiento de los algoritmos con mejores métricas e iterar el entrenamiento sobre subgrupos de la base de modelado basados en las zonas.
- **Iteración 6:** Implementar una estrategia de clustering que permita realizar la selección de subconjuntos de la base de entrenamiento y entrenar el algoritmo con mejor desempeño en cada uno de estos grupos. Con esto buscar uno o más modelos que se ajusten a un conjunto de datos con las métricas esperadas.
- **Iteración 7:** Ejecutar entrenamiento del modelo final mediante validación cruzada, ejecutar visualización de importancia de variables y análisis de ejecución del modelo por zonas para establecer las recomendaciones al negocio de cómo debe ser este implementado.

VIII RESULTADOS

Igual que en la sección de METOLOGÍA se exponen los resultados de acuerdo a las dos grandes fases de este proyecto:

Scraping:

Al explorar los sitios de arriendo se ha visualizado que no todos poseen la misma cobertura de la ciudad y que la estructura de la información es diferente, además, obtener los datos en cada uno de ellos depende de la cantidad de iteraciones y filtros que se deben realizar para obtener los datos deseados. En la TABLA I se muestran los sitios explorados y las observaciones que nos indican cual es el sitio ideal para usarlo como la fuente de información.

TABLA I. RESULTADOS DE LA EXPLORACIÓN DE SITIOS WEB

Sitio	Url	Librería	Observaciones
Espacio Urbano	https://www.espaciourbano.com/	Beautifull Soup	Presenta una buena cobertura sobre la ciudad y su configuración de políticas facilita la extracción de información
Metro Cuadrado	https://www.metrocuadrado.com/	Selenium	Se debe interactuar con la página para obtener los datos, sin embargo, se nota que solo tiene cobertura en zonas específicas de la ciudad
Finca Raiz	https://fincaraiz.com.co/robots.txt	Selenium	Presenta una buena cobertura sobre diferentes zonas de la ciudad, sin embargo, se debe interactuar con el sitio para obtener los datos

Debido a que el sitio de *Espacio Urbano* no contiene políticas de restricción en su archivo *robots.txt* que complejice la extracción de información, además de contar con una cobertura completa de los

barrios y zonas de la ciudad y el método de scraping puede ser utilizado a través de la librería Beautifull Soup, la cual tiene una facilidad de uso notable frente a los otros métodos, este sitio será la fuente de datos.

Al ejecutar el script de web scraping se obtienen los siguientes resultados:

- Se obtiene una cobertura de la ciudad dividida en 5 zonas: Centro, Poblado, Laureles, Belén y San Antonio de Prado, esta división es definida por el sitio web. Dentro de cada zona se encuentra los diferentes barrios y sectores de la ciudad, cada uno de estos que se listan a continuación contiene al menos una vivienda registrada para alquiler en el sitio.
    - **Centro:**
        - 12 de octubre
        - Alfonso López
        - Andalucía
        - Bombona
        - Prado Centro
        - Aranjuez
        - Avenida Oriental
        - Ayacucho
        - Caribe
        - Palmas
        - Boston
        - Buenos Aires
        - Campo Valdes
        - Castilla
        - Centro
        - El chagualo
        - El Salvador
        - Encizo
        - Florencia
        - Girardot
        - La Candelaria
        - La Milagrosa
        - Loreto
        - Los Angeles
        - Manrique
        - Moravia
        - Parque Bolivar
        - Pedregal
        - Popular
        - San Benito
        - San Pablo
        - Santa Cruz
        - Santander
        - Boyacá las Brisas
        - Sevilla
        - Villa Hermosa
        - Villanueva Medellín

- **Poblado:**
  - San Lucas
  - El Campestre
  - Las Santas
  - Los Parra
  - Loma del Indio
  - Los Balsos
  - La Florida
  - El Tesoro
  - Loma los González
  - Transversal Superior
  - Vizcaya
  - Alejandría
  - Ciudad Del Rio
  - Milla de Oro
  - Patio Bonito
  - Cola del Zorro
  - La Tomatera
  - Astorga
  - La Calera
  - Las Palmas
  - Oviedo
  - La Visitación
  - Aguacatala
  - Santa María de Los Ángeles
  - Las Santas
  - Las Loma
  - Loma de San Julián
  - Castropol
  - Provenza
  - Manila
  - Intercontinental
  - San Diego
  - La Concha
  - La Linde
  - Chuscalito
  - La Frontera
  - Transversal Inferior
  - Loma el Encierro
  - Provenza

- **Laureles:**
  - Laureles
  - La Castellana
  - Robledo
  - Simón Bolívar
  - El Nogal
  - Calasanz
  - Pilarica
  - Estadio
  - San Javier
  - San German
  - La América
  - Los Colores
  - López de Mesa
  - Córdoba

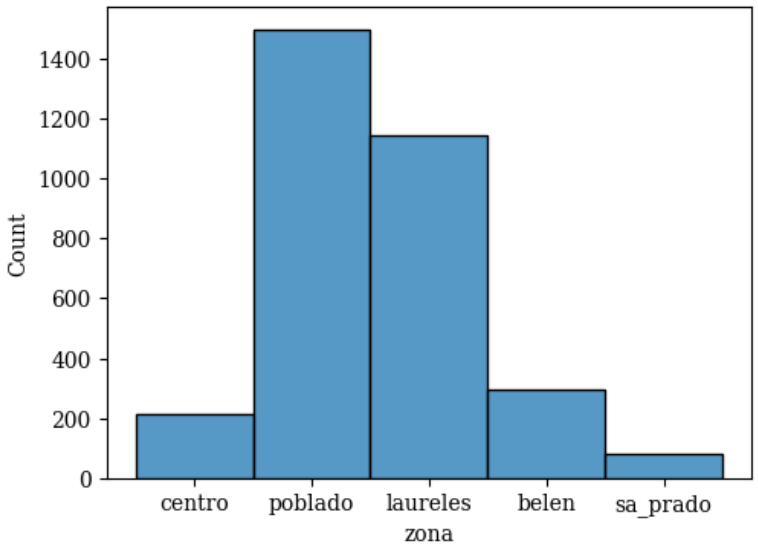
- Santa Mónica
- Suramericana
- Almería
- Santa Lucía
- Conquistadores
- Santa Gema
- Santa Rosa de Lima
- Belencito
- La Floresta
- Santa Teresita
- San Joaquín
- Florida Nueva
- Velódromo
- San Cristobal
- Carlos e Restrepo
- Avenida Nutibara
- La Pradera
- Estadio
- San Javier
- Nueva Pradera
- Santa Lucia

- **Belén:**
  - Loma de los Bernal
  - Rosales
  - La Mota
  - Santa Fé
  - Guayabal
  - Los Alpes
  - Belén La Palma
  - San Bernardo
  - Medellín
  - Altavista
  - Jardin
  - El Manzanillo
  - Granada
  - Las Playas
  - Rodeo Alto
  - Miravalle
  - Belén Rincón
  - Fátima
  - Trinidad
  - La Nubia
  - Malibú
  - La Castellana
  - Aliadas
  - La Gloria
  - Campo Amor
  - Porvenir
  - Buenavista
  - Las Mercedes
  - Alameda
  - Cristo Rey
  - Las Violetas

- **San Antonio de Prado:**
  - Barichara
  - San Antonio de Prado
  - Aragón
  - Pradito
  - Villas del Bosque
  - Villa Loma
  - Prados Del Sol
  - Ciudadela Villa del Bosque
  - El Vergel
  - Prado Verde
  - Prado Campestre
  - El Limonar
  - La Fabiola

- Para las zonas se obtiene un total de *3232 viviendas*. En la Fig 1 se puede apreciar la distribución del número de inmuebles por zona.

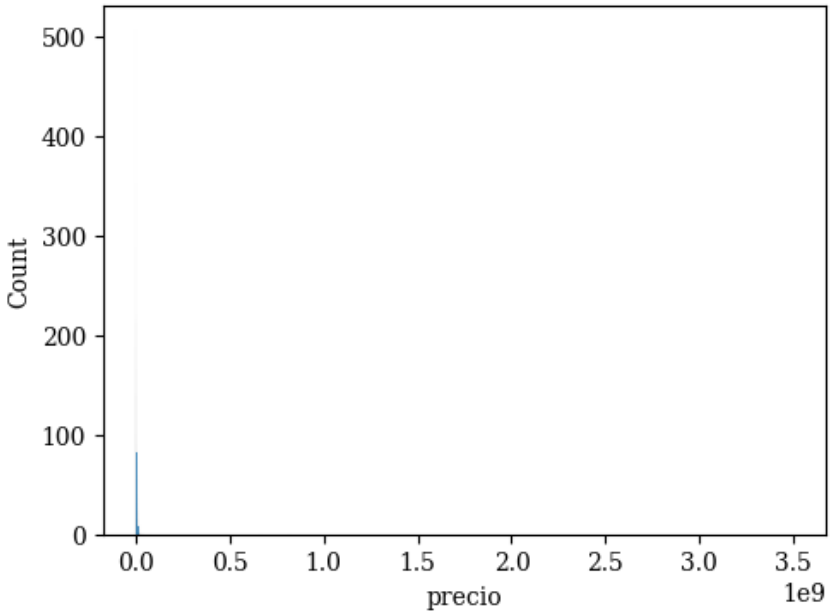
Fig. 1. DISTRIBUCIÓN DE INMUEBLES POR ZONAS



Se puede apreciar cómo las zonas del Poblado y Laureles tienen una importante ventaja sobre las demás, esto puede indicar un funcionamiento de la predicción del modelo con mejores métricas para estas zonas.

- Para los datos obtenidos se realiza una primera visualización de los precios de las viviendas.

Fig. 2. DISTRIBUCIÓN DE PRECIOS DE ARRIENDOS



En la Fig 2 se observa la distribución de los precios. Se tiene un histograma donde el eje vertical presenta la frecuencia de valores y el eje horizontal representa el precio en miles de millones, lo cual quiere decir, para un valor de 3.0, que hay presencia de inmuebles hasta por 3 mil millones COP lo cual es un claro valor atípico en un mercado de precios de arriendos que se evidencia causando un cambio de escala en el gráfico que dificulta apreciar el conjunto general de los datos. Por tanto, se debe realizar un tratamiento de datos atípicos para obtener una mejor distribución de esta variable.

- Las variables extraídas de cada inmueble se separan en 3 categorías: Numéricas, Categóricas Numéricas y Categóricas:

    - Numéricas: Variables continuas
        - Área Bruta
        - Área Total
        - Precio (Variable Objetivo)

    - Categóricas Numéricas: Variables discretas numéricas
        - Cantidad de baños familiares

- Número de Niveles
- Número de parqueaderos
- Número de alcobas
- Estrato

- Categóricas: Variables booleanas que indican si el inmueble cuenta con alguna de las siguientes comodidades:
  - Juegos Infantiles
  - Balcón
  - Zona de Ropas
  - Cámaras cctv
  - Cancha Polideportiva
  - Ascensor
  - Cancha Sqash
  - Zona BBQ
  - Patio
  - Unidad Cerrada Conjunto
  - Zonas Verdes
  - Aire Acondicionado
  - Jacuzzi
  - Red De Gas
  - Tipo de cocina
  - Tipo de Piso
  - Turco
  - Portería 24 7
  - Sauna
  - Calentador de Agua
  - Terraza
  - Closet de Linos
  - Biblioteca
  - Parqueadero Visitantes
  - Gimnasio
  - Piscina
  - Salón Social
  - Dispositivos de automatización
  - Alarma

En las variables se puede notar un alto número de categóricas, esto también es un punto a tener en cuenta en la selección del tipo de modelo, pues el algoritmo debe tener la condición de funcionar bien para este tipo de características.

El flujo de ejecución del sistema de scraping y la construcción de la base insumo sigue los siguientes pasos:

1. Se filtra la información por cada zona directamente en el sitio de *Espacio Urbano*.
2. Con la información filtrada se realiza la petición a la url.
3. Se aplica xpath al archivo html resultante mediante BeautifullSoup para la extracción de la información [7].
4. Se repite el proceso anterior y se crea un archivo por cada zona.
5. Se consolidada la información teniendo en cuenta que, dado que no todos los inmuebles tienen las mismas comodidades se debe hacer un proceso de construcción de columnas por cada una de estas y asignar un 1 al inmueble que la posea y un 0 al que no.
6. Se realiza el formato a los nombres y columnas eliminando espacios y caracteres especiales.

Con esto se obtiene una base de modelado conformada por:

TABLA II. BASE RESULTANTE POR WEB SCRAPING

Característica	Valor
Sitio Web	https://www.espaciourbano.com/
Cantidad de registros	3232
Variable Objetivo	Precio
Cantidad de variables numéricas	2
Cantidad de variables categóricas numéricas	5
Cantidad de variables categóricas	29

***Modelado**:*

A continuación, se presentan los diferentes resultados de las iteraciones, recordando que el objetivo principal es ir superando los resultados de las iteraciones anteriores

Tratamiento adicional de datos:

El resultado de la base construida a través de scraping contiene un gran número de variables categóricas a las cuales se les debe hacer un encoding para proceder con el entrenamiento, sin embargo, algunas de estas variables contienen un gran número de categorías que al aplicar este proceso resultará en problemas de dimensión para la base de modelado. Las variables con este comportamiento son:

- *Zona*: 5 categorías diferentes
- *Barrio sector*: 170 categorías diferentes
- *Tipo pisos*: 104 categorías diferentes
- *Ciudad*: 5 categorías diferentes
- *Tipo cocina*: 75 categorías diferentes

Las variables de *zona, ciudad* y *barrio_sector* no serán parte del entrenamiento sino del análisis de los resultados, por tanto, no se realiza intervención sobre estas.

Para las variables *tipo_pisos* y *tipo_cocina* se realiza dos intervenciones. La primera consiste en hacer una reducción de categorías agrupando valores similares, por ejemplo:

tipo _cocina:

- *integral alacena* equivale a *integral*
- *integral red de gas* equivale a *integral*

tipo_pisos:

- *madera laminada* equivale a *madera*
- *madera laminada mármol* equivale a *madera*

Al realizar este agrupamiento se obtiene las siguientes categorías para cada variable:

- Tipo pisos: 18 categorías diferentes
- Tipo cocina: 6 categorías diferentes

Si bien se ha logrado una reducción sustancial de valores aún se tienen demasiadas categorías de estas variables para hacer un proceso de encoding tradicional. Por tanto, la segunda intervención consiste en hacer un encoding por medio de frecuencias. Este proceso consiste en

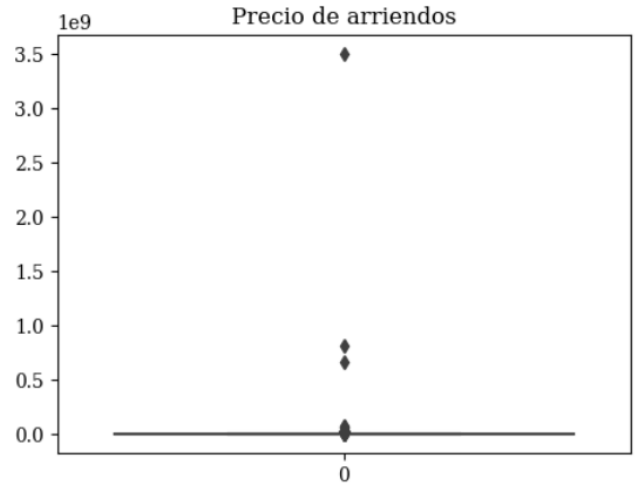
reemplazar el valor de cada categoría por la frecuencia que esta representa en la base general, convirtiendo así la variable original en numérica. Con esto se obtienen dos nuevas variables numéricas para la base de entrenamiento *tipo_cocina_freq* y *tipo_pisos_freq.*

Adicional a lo anterior se procede con la eliminación de datos nulos e indeterminados de la base donde para este caso no se requiere una intervención profunda dado que son valores que no representan ni un 1% de los datos, por tanto, se eliminan.

Iteración 1:

Al completar el tratamiento adicional a los datos se procede a crear un modelo línea base para obtener un primer vistazo del comportamiento de la información y que tan complejo debe ser el algoritmo a utilizar para alcanzar las métricas del negocio. También se hace un análisis básico de la distribución de la variable objetivo *"precio"*, para tener un punto de partida de la calidad de los datos.

Fig. 3. DISTRIBUCIÓN DE LA VARIABLE *PRECIO*



En la Fig. 3 (Gráfico de Cajas y Bigotes), se aprecia la distribución de la variable objetivo reforzando el comportamiento de datos atípicos obtenido en la Fig.2. En el eje vertical se tiene una escala de miles de millones donde se observa la presencia de precios tan grandes que evita observar el comportamiento general de los datos, esto se comprueba calculando los valores de tendencia central y dispersión. En la TABLA III se observa como los valores

de mínimo, percentil 75 y máximo están por fuera de lo que se podría considerar un valor regular de alquiler, pues un valor de 0 no tiene sentido para el mercado y valores de miles de millones para casas o apartamentos en arriendo se asemejan más a precios de venta o están por fuera del rango regular de precios en la ciudad para ese mercado.

Para regular la variable precio dentro de valores más "lógicos", se realiza una eliminación de datos atípicos por medio del cálculo del rango inter cuartil obteniendo los valores que se observan en la TABLA IV.

TABLA III. MEDIDAS DE DISPERSIÓN Y TENDENCIA CENTRAL DE LA VARIABLE *PRECIO*

Medida	Valor
Media	447391999.773014
Desviación estándar	1233632087.282006
Mínimo	0.000000
25%	1125805.250000
50%	3900000.000000
75%	20162403.136915
Máximo	3500000000.000000

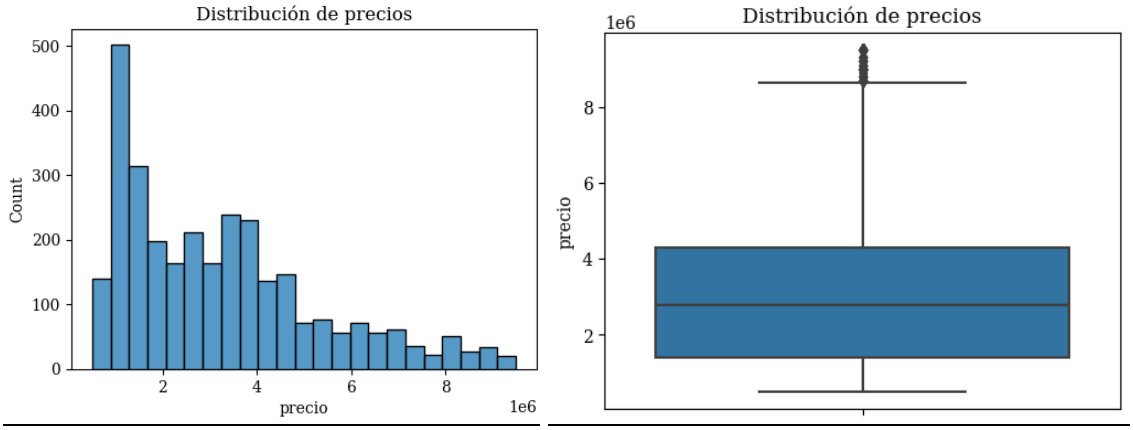
TABLA IV. CÁLCULO DE OUTLIERS VARIABLE *PRECIO*

Medida	Valor
IQR	3200000.0
Límite Inferior	-3300000.0
Limite Superior	9500000.0
Cantidad de Atípicos	194

Con este rango y límites se obtienen 194 valores atípicos, sin embargo, tiene que ser mas acido el corte, pues el límite inferior tiene un valor negativo el cual marca como permitidos valores superiores en ese signo o cercanos a 0 lo cual es ilógico para la variable. Por tanto,

como criterio de desarrollador la variable precio se toma entre los intervalos de [500000, 9500000], dando como resultado la distribución de la Fig. 4, donde se observa una distribución mas compacta de la variable y la eliminación de los datos atípicos.

Fig. 4. DISTRIBUCIÓN DE VARIABLE PRECIO SIN OUTLIERS



El tratamiento de los datos para la iteración I finaliza con un escalamiento de los datos a través de la clase *MinMaxScaler* de la librería *Scikit Learn.*

El modelo base será es una regresión lineal simple teniendo especial atención en las métricas de R2, RMSE, MAPE, con el objetivo de observar los valores objetivo del negocio.

TABLA V. MODELO LINEA BASE

Medida	Valor
%Entrenamiento	75%
% Prueba	25%
Modelo	LinearRegression
R2	0.72
RMSE	1102245.72
MAPE	0.28

Dada la TABLA V se obtiene un modelo con métricas aceptables, con un MAPE 0.13 puntos por encima del deseado y un RMSE que nos indica un error de predicción de aproximadamente 1’102.245COP lo cual para un usuario es un margen de error muy amplio y que se buscará reducir en las próximas iteraciones.

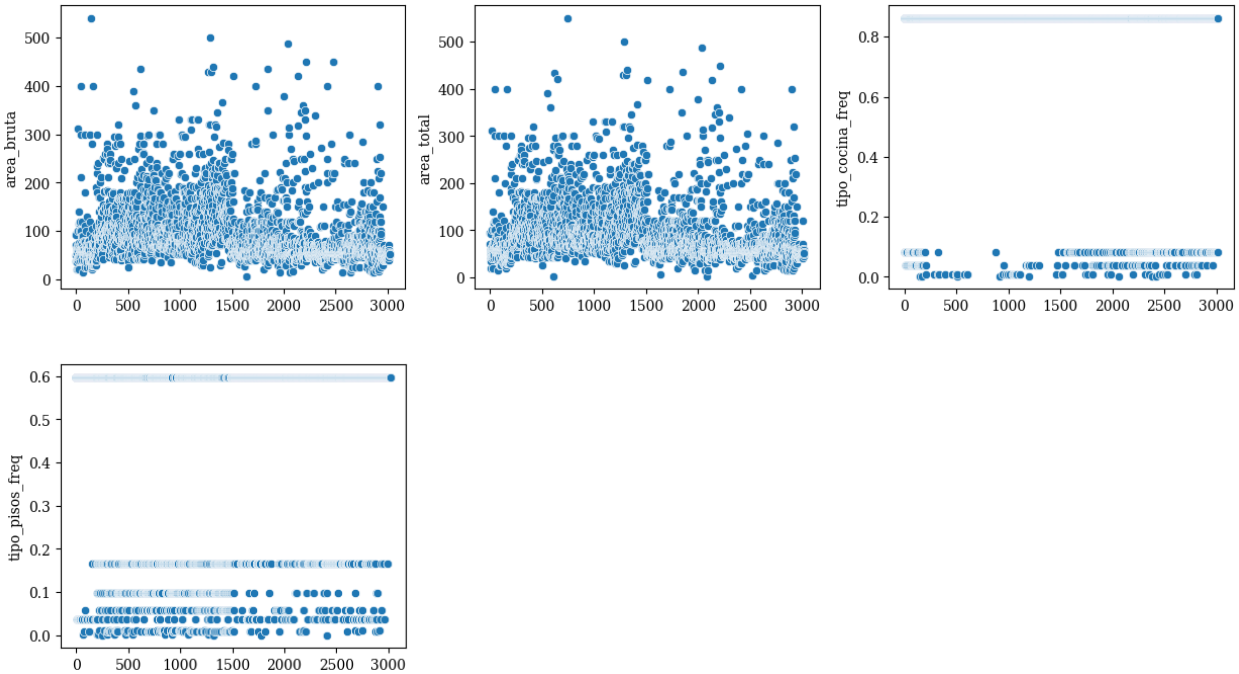
Iteración 2:

Habiendo logrado el modelo de línea base, el objetivo es mejorar las métricas obtenidas por medio del tratamiento y limpieza de los datos recolectados por el proceso de web scraping, lo cual resulta en los siguientes hallazgos y manipulación de los datos:

- Eliminación de outliers de la variable objetivo y codificación de variables categóricas: se preserva el tratamiento de datos realizado en la iteración 1.

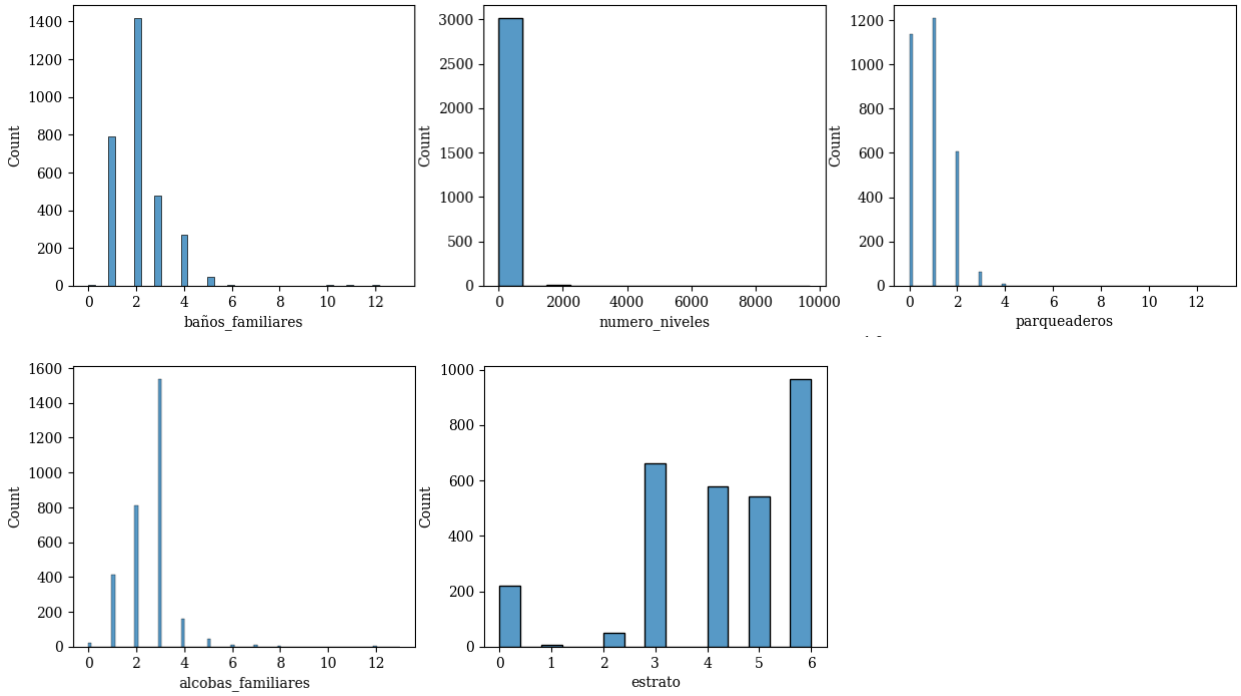
- Observación de distribución de los datos: Se grafica la distribución de los datos de cada tipo de variables obtenidas.

Fig. 5. DISTRIBUCIÓN VARIABLES NUMÉRICAS



En la Fig 5 se aprecia el comportamiento de las variables numéricas, estas no presentan novedades en sus distribuciones, recordando que las nuevas variables numéricas *tipo_pisos_freq* y *tipo_cocina_freq* son el resultado de aplicar encoding por frecuencias sobre las variables categóricas *tipo_pisos* y *tipo_cocina* respectivamente, donde allí se observaba una predominancia de una de las categorías sobre las demás, lo cual se traduce en el comportamiento observado de una gran cantidad de datos concentrados en una de las frecuencias.

Fig. 6. DISTRIBUCIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



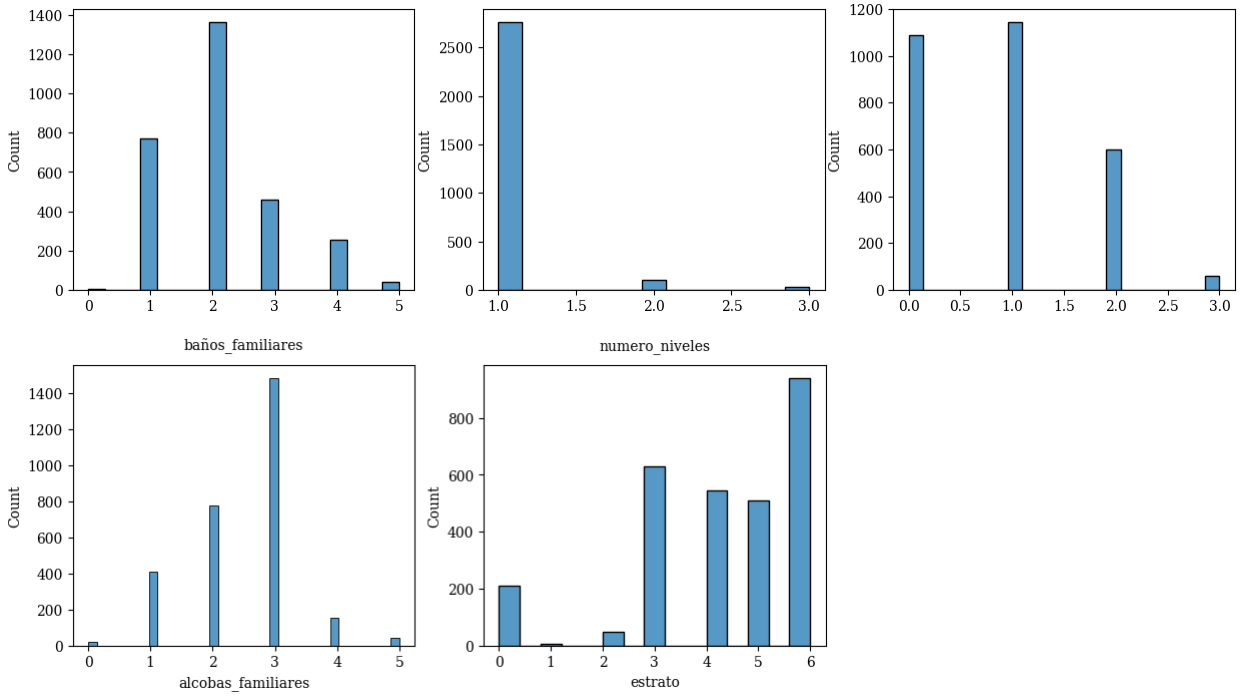
En la Fig 6 se presentan las variables numéricas, pero con valores discretos, las cuales se les da un tratamiento de variables categóricas. Se puede observar la presencia de valores atípicos en las variables de *baños_familiares*, *numero_niveles*, *parqueaderos y alcobas_familiares*, pues valores como 20, 800, 12 o 30 respectivamente no tienen un sentido lógico para una vivienda, por tanto, se realiza el siguiente tratamiento:

TABLA VI. ELIMNACIÓN DE ATÍPICOS SOBRE LAS VARIABLES CATEGÓRICAS NUMÉRICAS

Variables	Valor Permitido
numero_niveles	>0 y <= 3
baños_familiares	<= 5
parqueaderos	<= 3
alcobas_familiares	<= 5

Con el tratamiento realizado en la TABLA VI se obtiene la distribución de la Fig 7. Con lo cual los valores para estas variables tienen una representación más lógica.

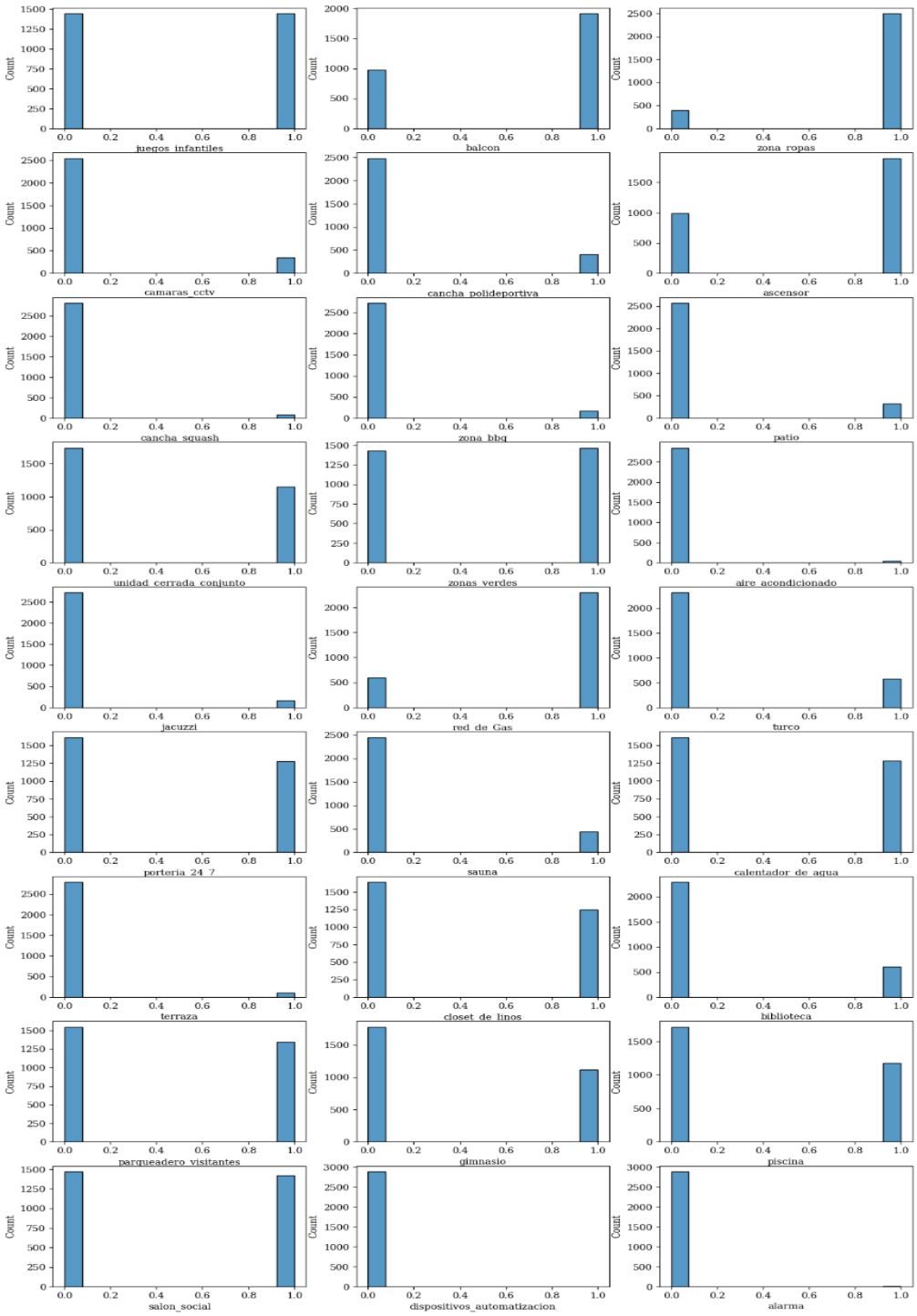
Fig. 7 DISTRIBUCION DE VARIABLES CATEGÓRICAS NUÉRICAS SIN OUTLIERS



En la Fig 8 Se observa las distribuciones de las variables categóricas booleanas, estas representan las comodidades con las que cuenta una vivienda. Es de resaltar que hay algunas de ellas que son tan comunes hoy en día que es casi ilógico que una casa o apartamento no la posea, por ejemplo, variables como *zona_de_ropas* o *red_de_gas*, sin embargo, también sucede el caso contrario con comodidades que son escasas para la mayoría de las viviendas como lo son *cancha_sqash* o *aire_acondicionado*. Estas variables

tendrán un valor predominante lo cual puede tener como consecuencia que no tengan representatividad en el entrenamiento del modelo. La eliminación de estas variables se considera en base a su distribución y su valor de correlación con respecto a la variable objetivo.

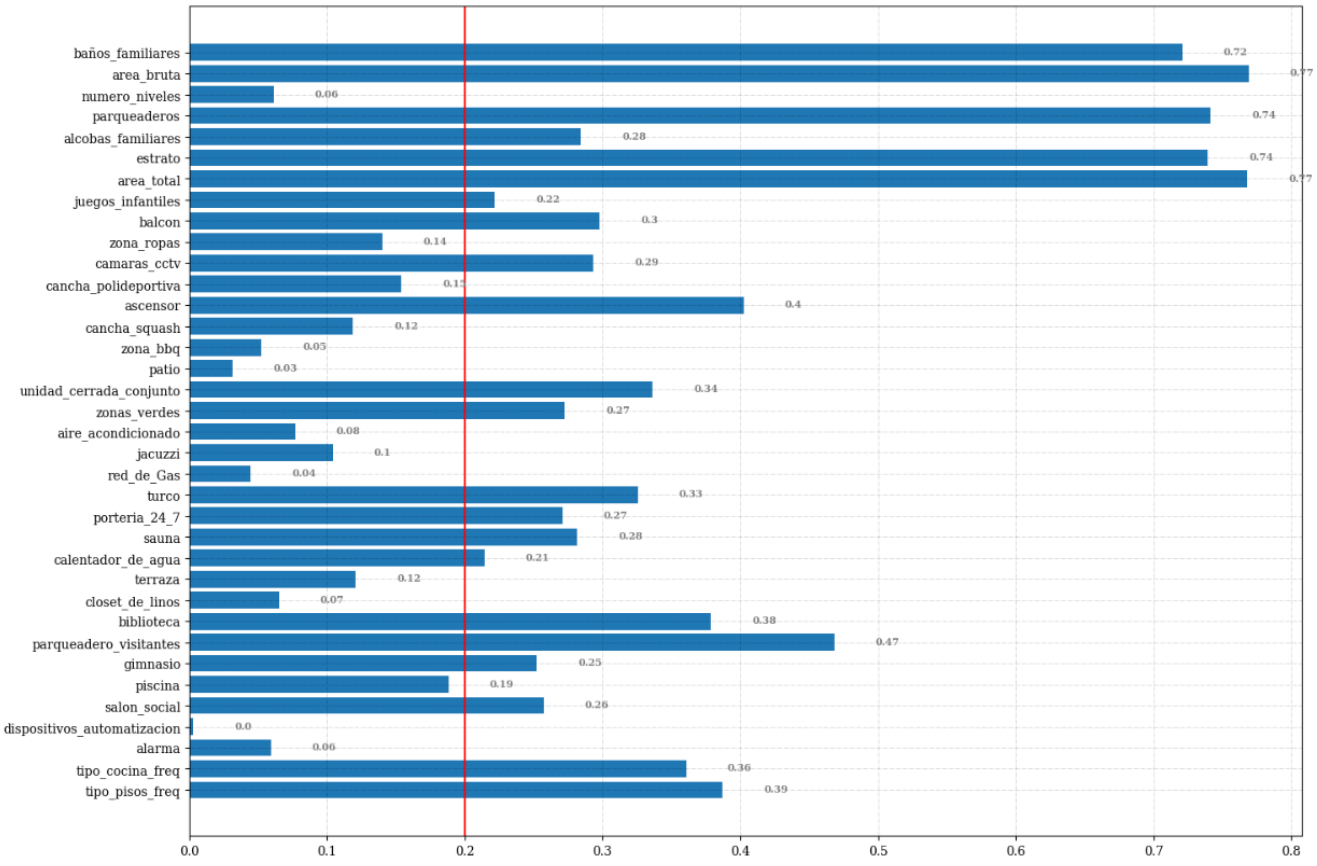
Fig. 8 DISTRIBUCIÓN VARIABLES CATEGÓRICAS



Se observa que variables como *cancha_squash*, *zona_bbq* o *patio*, cuentan con un sesgo muy grande entre valores, lo cual puede implicar que no sean influyentes en el modelo.

- Correlación: Para el calculo de correlaciones se realizan múltiples ejecuciones para determinar la relevancia y comparación de variables, estas son:
    - Variables predictoras vs variable precio
    - Variables numéricas
    - Variables categóricas numéricas
    - Variables categóricas

Fig. 9 CORRELACIÓN VARIABLE OBJETIVO VS VARIABLES PREDICTORAS



En la Fig. 9 se aprecia la correlación absoluta de cada una de las variables predictoras con respecto a la variable objetivo, en esta se establece un limite de 0.2 como valor frontera para definir cuáles son las variables poco explicativas para el precio. Según lo anterior

variables como *aire_acondicionado*, *jacuzzi* y *dispositivos_automatizacion* serán poco aportantes para el modelo. Variables por debajo de este valor de correlación y una distribución con un valor predominante serán eliminadas de la base de entrenamiento

De otro lado, definiendo 0.7 como valor para determinar si dos variables están altamente correlacionadas se realizan las siguientes ejecuciones:

Fig. 10 CORRELACIÓN VARIABLES NUMÉRICAS

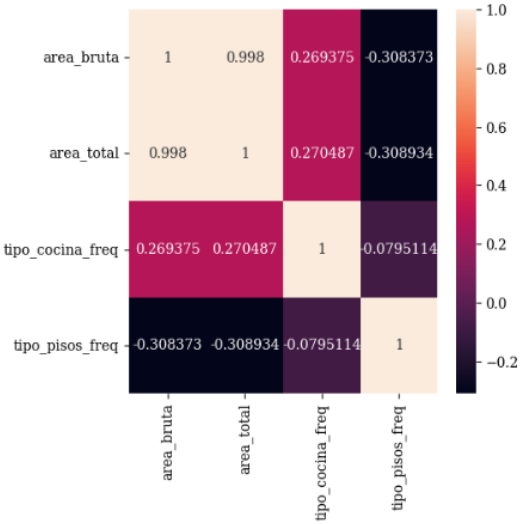
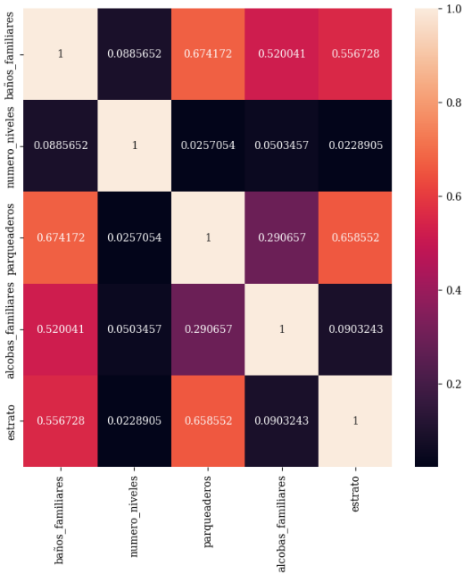


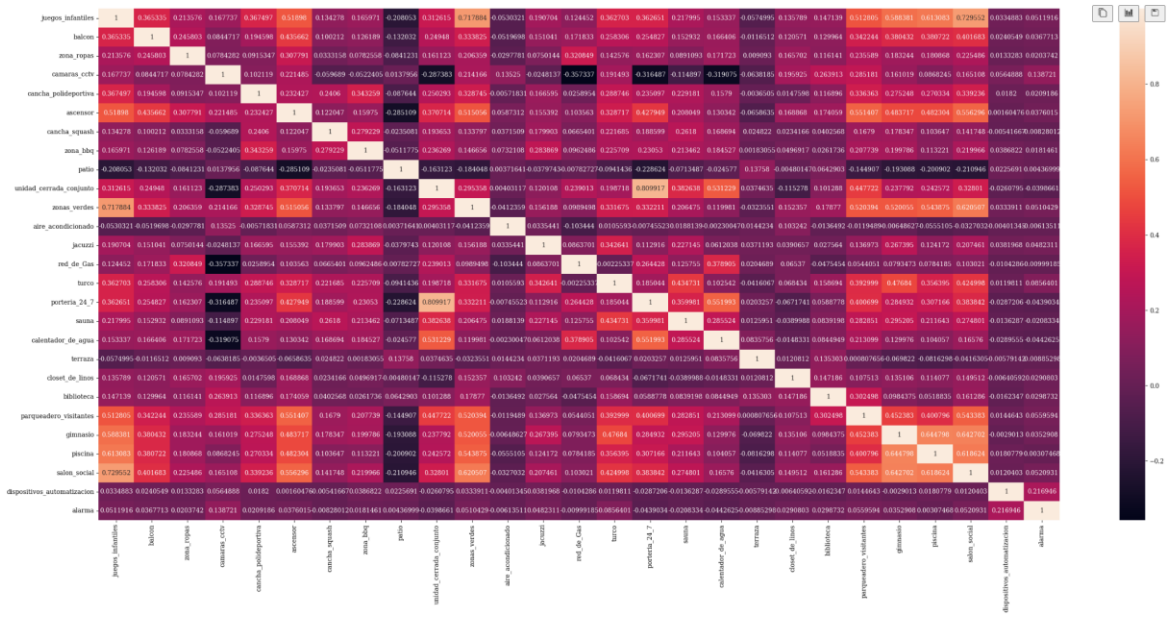
Fig. 11 CORRELACIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



En la Fig. 10 se observa que las variables *area_bruta* y *area_total* están altamente correlacionadas, este valor es esperado por la similaridad en el significado de ambas variables, por tanto, se puede prescindir de una de ellas para la fase de modelado.

En la Fig. 11 se tienen el análisis de correlaciones para las variables numéricas categóricas en el cual no se presencia valores de alta correlación entre ellas.

Fig. 12 CORRELACIÓN VARIABLES CATEGÓRICAS



En la Fig 12 se observa la correlación entre las variables categóricas, donde se resalta un alto valor entre:

- zonas_verdes - juegos_infantiles
- juegos_infantiles - salon social
- unidad_cerrada_conjunto - porteria_24_7

Para el descarte de una de estas variables se usará el siguiente criterio: “¿La existencia de una variable implica necesariamente la otra?":

- La existencia de zonas verdes no es condición suficiente para la existencia de juegos infantiles, por tanto, se conservan ambas variables.
- La existencia de juegos infantiles no es condición suficiente para la existencia de salón social, por tanto, se conservan ambas variables.
- Una vivienda en un conjunto de unidad cerrada implica, en la mayoría de los casos, la existencia de portería 24/7, por tanto, es una variable de la cual se puede prescindir.

Finalmente, al calcular la correlación entre todas las variables predictoras se obtienen los resultados de la TABLA VII.

TABLA VII CORRELACIÓN ENTRE VARIABLES PREDICTORAS

Variable1	Variable 2	Valor
baños_familiares	area_bruta	0.78
baños_familiares	area_total	0.78
area_bruta	area_total	0.99
juegos_infantiles	zonas_verdes	0.70
juegos_infantiles	salon_social	0.73
unidad_cerrada_conjunto	porteria_24_7	0.81

Dados estos resultados y aplicando la lógica de eliminación dictada en los puntos previos se descartan las siguientes variables:

- *numero_niveles*
- *zona_ropas*
- *camaras_cctv*
- *cancha_polideportiva*
- *zona_bbq*
- *patio*
- *aire_acondicionado*
- *jacuzzi*
- *red_de_Gas*
- *terraza*
- *cancha_squash*
- *area_total*
- *porteria_24_7*

- Modelado: Dado el tratamiento de datos realizado se procede con el entrenamiento del modelo obteniendo el escenario que se indica en la TABLA VIII.

TABLA VIII MODELO ITERACIÓN 2

Medida	Valor
%Entrenamiento	75%
% Prueba	25%
Modelo	LinearRegression
R2	0.70
RMSE	1116690.71
MAPE	0.28

Los resultados de este modelo no muestran mejoras en cuanto a lo obtenido en el modelo línea base, sin embargo, el tratamiento realizado a los datos nos provee una base de entrenamiento sin valores ilógicos y con variables significativas para la variable objetivo. Esta base será utilizada en las iteraciones posteriores donde se dará paso a ejecutar diferentes tipos de modelos.

TABLA IX. BASE DE ENTRENAMIENTO RESULTADO ITERACIÓN 2

Caracteritica	Valor
Cantidad de registros	2892
Variable Objetivo	Precio
Cantidad de variables numéricas	3
Cantidad de variables categóricas numéricas	4
Cantidad de variables categóricas	17

Iteración 3:

Habiendo realizado un tratamiento exhaustivo sobre los datos de entrenamiento y obtenido una base de modelado en la iteración 2 (TABLA IX), se procede a ejecutar diferentes tipologías de modelos de machine learning con el fin de encontrar aquel que nos otorgue la métrica objetivo.

Los tipos de algoritmos a considerar en esta ejecución corresponden a los modelos tradicionales de regresión, estos son:

- LinearRegression
- Regresión Lasso
- DecisionTreeRegressor
- RandomForestRegressor
- SVR

La ejecución de cada modelo consiste en hacer múltiples iteraciones desde un conjunto de hiperparámetros con ayuda de la librería *ParameterGrid* de *Scikit-Learn.* De cada ejecución se seleccionará el modelo con los mejores resultados en base a los siguientes criterios y configuración:

- Se evalúan las métricas R2, RMSE y MAPE.
- La diferencia entre R2 entrenamiento y prueba no debe ser superior a 0.05 (este criterio es global para todos los modelos en las iteraciones posteriores).
- Se selecciona el modelo con menor valor MAPE para el conjunto de prueba.
- La división del conjunto de entrenamiento y prueba es 75%, 25% respectivamente.

TABLA X. COMPARACION RESULTADOS DE MÉTRICAS ITERACIÓN 3

Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
LinearRegression	0.74	0.71	1051860.93	1090624.48	0.28	0.28
Regresión Lasso	0.73	0.69	1077563.62	1090480.11	0.28	0.29
DecisionTreeRegressor	0.73	0.69	1077563.62	1140152.11	0.28	0.29
RandomForestRegressor	0.79	0.75	952251.42	1011693.32	0.24	0.25
SVR	0.71	0.68	1121700.44	1153141.45	0.25	0.26

Dados los resultados de la TABLA X se obtienen mejoras en los algoritmos de RandomForestRegressor y SVR, siendo estos los que mejoran el alcance a la métrica del negocio (MAPE <= 0.15).

Aunque se obtiene una mejoría en las métricas es preciso encontrar un algoritmo con los resultados deseados, por tanto, se debe considerar la implementación de modelos mas robustos. En este caso se opta por la ejecución de modelos de boosting y evaluar los resultados que arrojan.

Iteración 4:

Reutilizando del nuevo la base de modelado de la iteración 2 se ejecuta el entrenamiento de los algoritmos de boosting como opción a la necesidad de ejecutar algoritmos más complejos en la búsqueda de la métrica MAPE deseada, esto debido a que los tipos de modelos tradicionales no han mostrado cercanía a esta.

Los modelos de tipología boosting a ejecutar son:

- Ada Boosting
- Gradient Boosting
- XGBOOST

El proceso de ejecución es similar a la iteración 3 donde cada tipo de modelo es sometido a diferentes entrenamientos en base a un conjunto de parámetros con las siguientes condiciones y configuraciones:

- Se evalúan las métricas R2, RMSE y MAPE.
- Se selecciona el modelo con menor valor MAPE para el conjunto de prueba.
- La división del conjunto de entrenamiento y prueba es 70%, 30% respectivamente (se amplía en esta iteración el conjunto de entrenamiento para buscar cambios en los resultados).

TABLA XI. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 4

Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
Ada Boosting	0.73	0.71	1080592.66	1108223.99	0.3	0.32
Gradient Boosting	0.8	0.75	930445.85	1016608.13	0.18	0.21
XGBOOST	0.72	0.7	1104313.47	11215224	0.21	0.21

Dado los resultados presentados en la TABLA XI, de esta iteración de obtienen las siguientes conclusiones:

- Para estas ejecuciones se obtiene una mejora sustancial en las métricas del modelo para los algoritmos de Gradient Boostin y XGBoost obteniendo un MAPE a solo 0.3 puntos por encima de la métrica del negocio.
- Uniendo los resultados de la iteración anterior y la actual, los algoritmos con mejores resultados son: Gradient Boosting, XGBoost, RandomForestRegressor y SVR, por tanto, en las iteraciones posteriores solo se ejecutarán estos algoritmos.
- Aunque la mejora en los algoritmos de boosting da un buen indicio del modelo a seleccionar, aún se debe obtener la métrica objetivo. Para esto, regresando a lo obtenido en los datos del scraping (Fig 1), se observa que hay un desbalance en la cantidad de registros por zonas, siendo *Poblado* la más predominante, por lo tanto, se procederá a entrenar los mejores algoritmos en diferentes combinaciones de zonas y verificar si los datos de algunas de estas afectan los resultados obtenidos hasta ahora.

Iteración 5:

Esta iteración se caracteriza por subdividirse en una serie de sub iteraciones donde en cada una se ejecuta los modelos que han mostrado mejores métricas hasta ahora, iterando por los diferentes valores que tiene la variable *zona* en la base de entrenamiento, los cuales son: Poblado, Laureles, Centro, Belén y San Antonio de prado. Las sub iteraciones consisten en ir eliminando las zonas con menor cantidad de registros y observar si los modelos obtienen mejores métricas, esto determinará si el modelo a buscar se define por un grupo específico de datos obtenido por el proceso de scraping.

Las sub iteraciones son:

- iteracion_5_1: Base con zonas Poblado, Laureles, Belén y Centro
- iteracion_5_2: Base con zonas Poblado, Laureles, Belén
- iteracion_5_3: Base con zonas Poblado, Laureles
- iteracion_5_4: Base con rango de precios

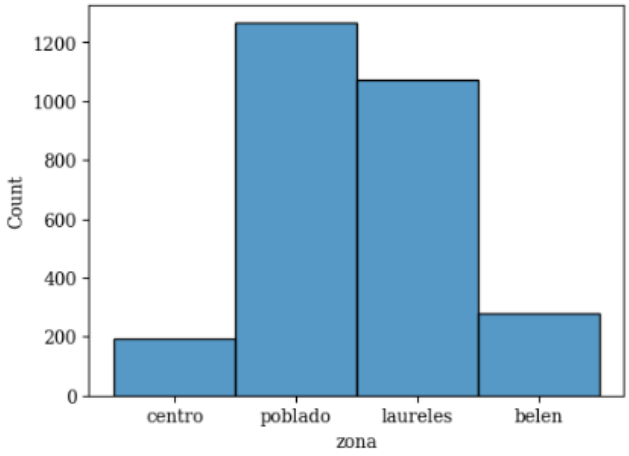
Dado que en cada sub iteración se está cambiando la base de entrenamiento se debe considerar repetir el tratamiento de datos de datos de distribución y correlación realizado en la iteración 2 además de una variación de los hiperparámetros si así lo necesita cada algoritmo.

En cada sub iteración se manejan los mismos criterios de elección y configuración de las iteraciones anteriores.

Nota: En los resultados obtenidos de repetir el tratamiento de datos de la iteración 2 solo se resaltarán cambios relevantes encontrados. Todos estos resultados se pueden encontrar en detalle en el repositorio adjunto.

Iteracion_5_1: Base con zonas Poblado, Laureles, Belén y Centro.

Fig. 13. DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5_1



- Tratamiento de datos: No se encuentran diferencias en cuanto a distribución, correlación o que lleve a hacer una nueva intervención en la base de entrenamiento
- Resultados de los modelos:

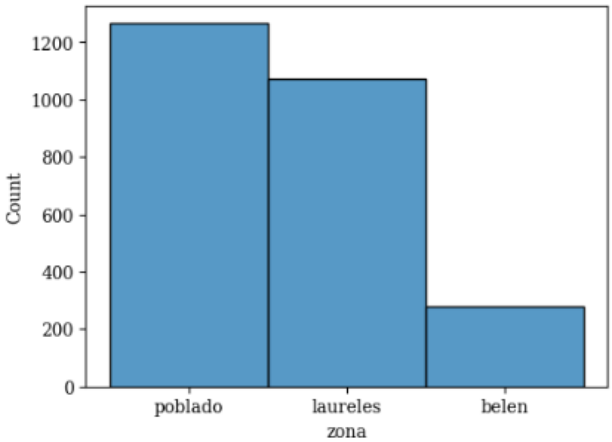
TABLA XII. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5_1

Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
RandomForestRegressor	0.73	0.68	1088509.69	1151104.27	0.28	0.31
SVR	0.73	0.68	1090509.75	1144851.89	0.24	0.26
Gradient Boosting	0.75	0.7	1037850.32	1101355.96	0.2	0.23
XGBOOST	0.71	0.66	1121918.81	1174164.96	0.21	0.22

En la TABLA XII Se observa un leve deterioro en las métricas con respecto a la iteración anterior a pesar de eliminar la zona con presencia minoritaria. Como aún no se obtiene la métrica del negocio se procede con la siguiente sub iteración.

Iteracion_5_2: Base con zonas Poblado, Laureles, Belén.

Fig. 14. DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5_2



- Tratamiento de datos: El tratamiento de datos realizado no muestra cambios que lleve a tomar nuevas decisiones sobre la base de entrenamiento, sin embargo, sí se

obtuvo la presencia de nuevas variables que según la correlación no afectan la variable precio.

- Resultados de los modelos:

TABLA XIII. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5_2

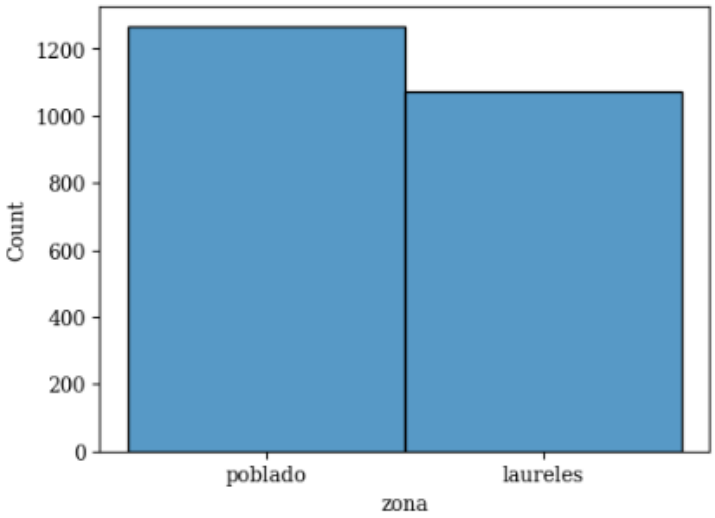
Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
RandomForestRegressor	0.78	0.73	962915.73	1085599.87	0.22	0.25
SVR	0.74	0.71	1043737.84	1124187.99	0.2	0.24
Gradient Boosting	0.77	0.73	982184.71	1087837.28	0.19	0.21
XGBOOST	0.73	0.69	1055116.34	1169434.38	0.18	0.19

En la TABLA XIII se observa como los algoritmos de XGBOOST y Gradient Boosting conservan buenos resultados aún bajo la eliminación de datos e incluso muestran una mejora en las métricas, esto marca un indicio de cuál puede ser el tipo de modelo final para el proyecto.

En busca de las métricas objetivo se procederá con la siguiente sub iteración, sin embargo, si en ella aún no se obtiene los resultados deseados se ejecutará una estrategia diferente de análisis por zonas.

Iteracion_5_3: Base con zonas Poblado, Laureles.

Fig. 15. DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5_3



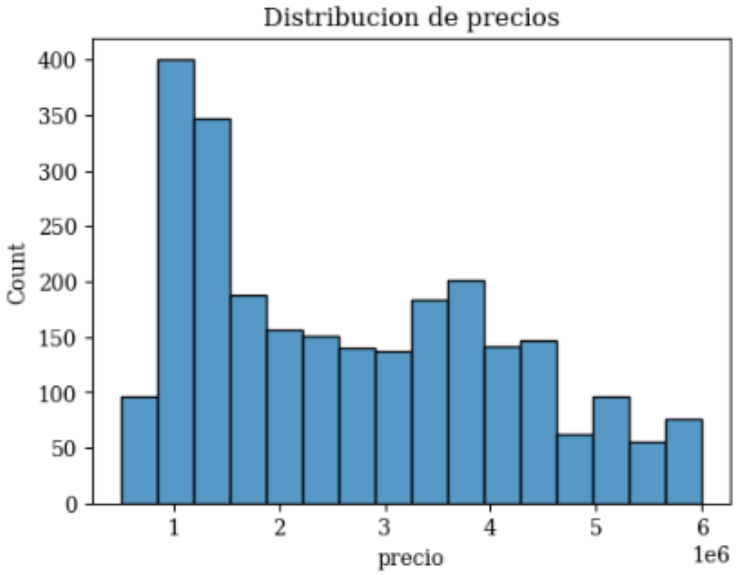
- Tratamiento de datos: Para este conjunto de entrenamiento surgen nuevas variables que por correlación muestran ser influyentes sobre la variable objetivo, por ejemplo, *camaras_cctv*, por el contrario, salen otras variables como *gimnasio* de tipo no influyentes. Esto es muestra que al estar en un conjunto de datos donde se encuentran las viviendas mas costosas algunas comodidades son muy frecuentes y son menos influyentes que en los demás grupos de inmuebles por su presencia como factor común.
- Resultado de los modelos: Para esta iteración no se tienen resultados dentro de los márgenes deseados, esto es, no hay ejecuciones con diferencias entre el R2 entrenamiento y prueba inferior a 0.05. Esto es un claro deterioro de las métricas obtenidas hasta ahora, por tanto, se debe ejecutar una nueva estrategia de submuestreo sobre la base.

Iteración_5_4: En el análisis por zonas se ha optado por eliminar aquellas con menos presencia en la base. Esta sub iteración se basa en hacer una exploración más a fondo sobre la variable *precio* analizando los rangos de valores que toma esta y así seleccionar un conjunto de datos para ejecutar el entrenamiento.

En la Fig 4 se observaba como la distribución de la variable precio contiene un sesgo hacia la izquierda dejando los inmuebles mas costosos con menos presencia en la base de entrenamiento. Lo que procede es realizar un submuestreo de la base con los precios inferiores a 6’000.000COP (Fig 16) y así plantear la hipótesis de si los inmuebles mas costosos son los que afectan encontrar las métricas deseadas.

- Tratamiento de datos: El tratamiento de datos realizado no muestra cambios que lleve a tomar nuevas decisiones sobre la base de entrenamiento.
  El proceso realizado en la iteración 2 ha mostrado resultados similares a lo largo de todas las iteraciones, por tanto, se concluye reutilizar la base resultante de esta para las próximas iteraciones.

Fig. 16. DISTRIBUCIÓN DE PRECIOS DE ARRIENDO CON FILTRO 6000000



- Resultados modelo: Los modelos SVR y Gradient Boosting no obtuvieron resultados concluyentes para esta sub iteración, esto es, no hay resultados cuyo R2 sea inferior a 0.05 entre los conjuntos de entrenamiento y prueba.

TABLA XIV. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5_4

Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
RandomForestRegressor	0.81	0.76	638433.45	717236.39	0.2	0.21
XGBOOST	0.8	0.75	661255.80	7311474	0.19	0.19

Dados los resultados presentados en la TABLA XIV, se dan las siguientes conclusiones:

- Al seleccionar un conjunto de datos de acuerdo a un rango de precios se nota una inmediata mejora en las métricas de los modelos, lo cual indica que la métrica objetivo podría hallarse sobre un rango de precios específico.
- En estas iteraciones el modelo de XGBOOST ha mostrado ser superior a los demás algoritmos en resultado y tiempo de ejecución, por tanto, para las próximas iteraciones solo se usará este algoritmo.
- Aunque se encontrase un modelo funcional para un rango de precios lo ideal es encontrar un modelo con las métricas adecuadas y con una cobertura general para todos los datos. Por ello, continuando con la estrategia de esta iteración se creará subconjuntos de la base de entrenamiento de acuerdo con la implementación de un algoritmo kmeans. Esto dará como resultado clústeres que serán las nuevas bases de entrenamiento para los modelos. Por tanto, en la próxima iteración las zonas de entrenamiento de los modelos serán determinadas por los clústeres resultantes.

Iteración 6:

Hasta ahora se ha identificado los siguientes comportamientos en la base de entrenamiento y los modelos:

- La distribución de la variable *precio* tiene un sesgo hacia las casas de menor costo, dejando las zonas que tienen arriendos más caros con muy poca presencia ante el modelo.
- Las múltiples ejecuciones en base a la variable *zona* no muestra una solución para encontrar las métricas deseadas, incluso hubo deterioro en estas.

- Al realizar una ejecución sobre un rango específico de precios hubo una mejora en las métricas, lo cual indica que se debe ejecutar una estrategia diferente en la creación de subconjuntos de la base de entrenamiento.

  Para seleccionar un subconjunto de manera óptima, se implementará un algoritmo de *kmeans* considerando solo las variables *estrato* y *area_bruta* (siendo estas numéricas y las que muestran mayor importancia en los análisis de correlación (Fig. 9 CORRELACIÓN VARIABLE OBJETIVO VS VARIABLES PREDICTORAS)). La cantidad de clústeres a seleccionar será determinada calculando el coeficiente de siluetas.

  Con lo anterior, el número de clústeres resultantes será la cantidad de subconjuntos de la base de entrenamiento donde se ejecutará el proceso de entrenamiento del modelo para cada una de ellas.

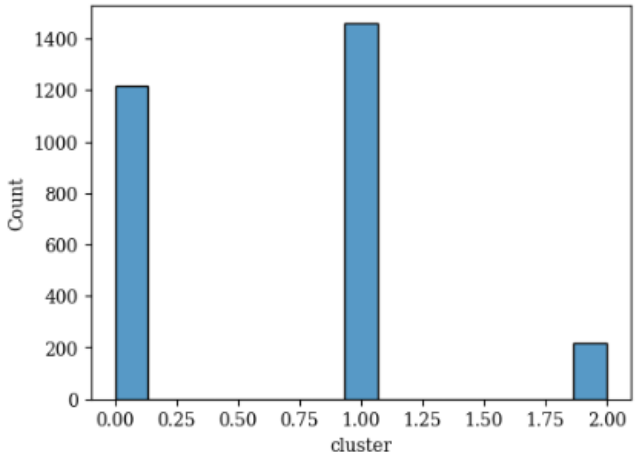
  Clusteting.
  Se reutiliza la base de entrenamiento resultante de la iteración 2 y las variables *estrato* y *area_bruta* para obtener los resultados de la TABLA XV. En esta se puede observar como se obtiene un mejor valor de coeficiente de silueta para 3 clústeres, por tanto, este será el valor para entrenar el algoritmo de *kmeans.*

  Al entrenar el modelo y calcular las etiquetas resultantes se obtiene la distribución de la Fig 17.

TABLA XV. EVALUACIÓN DE COEFICIENTE DE SILUETA

Numero de Clusters	Coeficiente de silueta
2	0.55
3	0.61
4	0.56
5	0.57

Fig. 17. DISTRIBUCIÓN DE LOS CLUSTERS



Obtenemos los rangos y distribución de la variable precio en cada cluster.

TABLA XVI. RANGO DE PRECIOS POR CLUSTER

Numero de Clusters	Mínimo	Máximo
0	500000	9000000
1	700000	9500000
2	550000	8700000

Dada la TABLA XVII y las distribuciones observadas en las figuras 18 -20, el entrenamiento del modelo se dividirá en múltiples ejecuciones siguiendo la estrategia a continuación:

- Solo se tendrá en cuenta el algoritmo xgboost, siendo este el de mejores resultados y rendimiento
- Se maneja el mismo criterio de selección del modelo de las iteraciones 3 y 4
- Se seleccionan las bases de entrenamiento de acuerdo a los clústeres lo cual nos resulta en 3 ejecuciones:
    - ejecución 1: Base de entrenamiento solo con datos de clúster 0 (Tiene el rango más amplio de precios)
    - ejecución 2: Base de entrenamiento con datos del clúster 1

- ejecución 3: Base de entrenamiento con datos del clúster 2

La efectividad de esta estrategia será medida si con al menos uno de los clústeres se obtiene un modelo con las métricas deseadas (MAPE <=0.15), pues este nos otorgará el modelo con el alcance de este proyecto y se podrán establecer las condiciones de implementación para recomendar al negocio.

Fig. 18. DISTRIBUCIÓN PRECIOS CLÚSTER 0

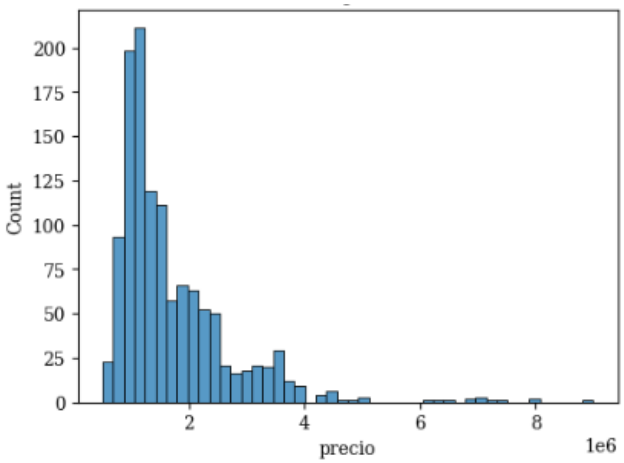


Fig. 19. DISTRIBUCIÓN DE PRECIOS CLÚSTER 1

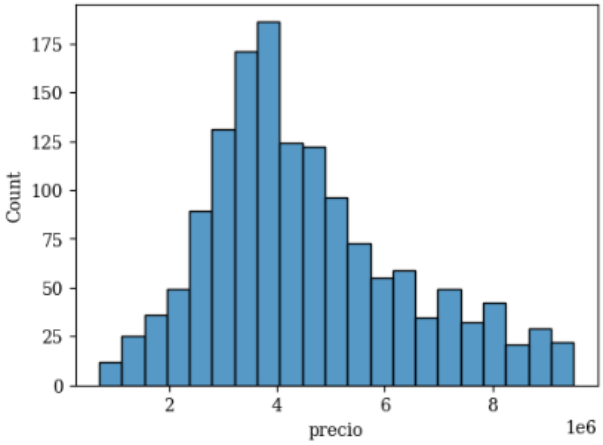


Fig. 20. DISTRIBUCIÓN DE PRECIOS CLÚSTER 2

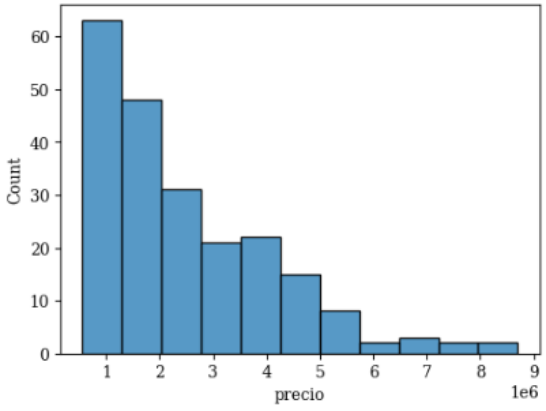


TABLA XVIII: TABLA DE EJECUCIONES ITERACIÓN 6 (XGBOOST)

Ejecución	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
Ejecución 1: XGBOOST clúster 0	0.75	0.72	511152.68	504298.55	0.14	0.15
Ejecución 2: XGBOOST clúster 1	0.50	0.42	1297740.17	1429225.26	0.23	0.25
Ejecución 3: XGBOOST clúster 2	0.50	0.47	1173213.30	1026328	0.33	0.37

En la TABLA XIX se obtiene el consolidado de los mejores resultados de entrenar un modelo en cada clúster. Esta muestra que en el clúster 0 se obtiene las métricas deseadas siendo este con mayor cobertura sobre los precios, en consecuencia, esta estrategia ha sido exitosa pues se ha logrado la métrica propuesta por el negocio (MAPE <= 0.15).

Para el modelo del clúster 1 no se logran las métricas deseadas, incluso están muy por debajo de lo observado a lo largo de las iteraciones, esto debido al cambio de distribución que se observa entre las Fig 19. Por consiguiente, no se puede implementar este modelo para este conjunto de datos y se debe ejecutar una búsqueda exhaustiva de técnicas e hiperparámetros para llegar a lo demandado por el negocio. Esto se deja como trabajo futuro de este proyecto.

Similar a los datos del clúster 1, para el clúster 2 no se obtienen las métricas deseadas. Este conjunto de datos presenta una minoría con respecto a los otros subconjuntos (Fig. 17. DISTRIBUCIÓN DE LOS CLUSTERS), lo cual dificulta ajustar un modelo e implementarlo con los resultados esperados. El encontrar un modelo para estos datos con los requisitos del negocio se deja como trabajo futuro de este proyecto.

Iteración 7:

Habiendo logrado un modelo con las métricas esperadas por el negocio se procede con la construcción del modelo final, este se hará mediante el entrenamiento de los mejores parámetros resultantes de la iteración 6 que consiste en: la ejecución del algoritmo kmeans, seleccionar los datos del cluster 0 y entrenar un modelo xgboost por medio de validación cruzada para garantizar un entrenamiento más robusto. Con los resultados obtenidos en el modelo final se hará una evaluación de la importancia de variables y establecer una recomendación al negocio de cómo se debe implementar.

Fig. 21. DISTRIBUCIÓN DE PRECIOS LUEGO DE SUBMUESTREO

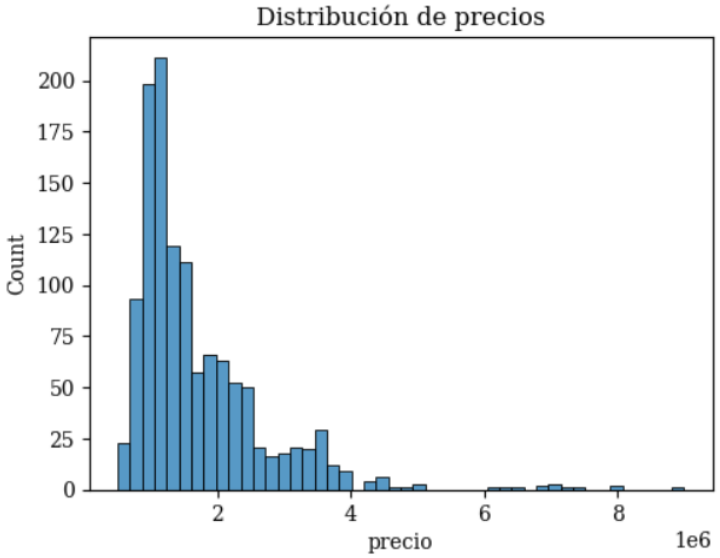


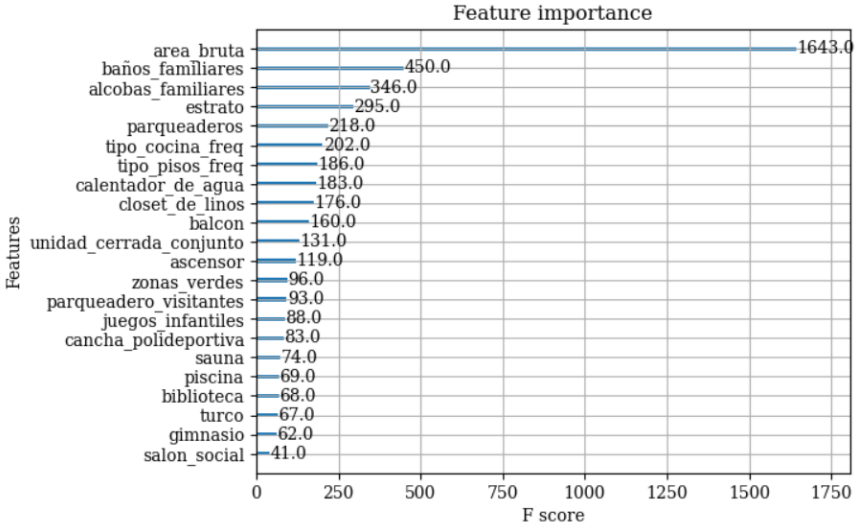
TABLA XX. RESULTDOS MÉTRICAS MODELO FINAL

Modelo	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE test
XGBOOST	0.78	0.73	476610.20	490008.01	0.14	0.15

En la figura 21 se tiene la distribución de precios obtenido de los datos del cluster 0 con el cual se obtiene la base de entrenamiento para la ejecución del proceso de creación del modelo final ejecutando validación cruzada que nos da como resultado lo obtenido en la TABLA XXI.

Aprovechando las ventajas de la librería *xgboost* se grafica de esta la importancia de variables sobre el modelo final en la Fig 22. En esta se aprecia la gran influencia que tiene el área bruta de una vivienda para determinar su precio, seguido del número de baños familiares, habitaciones y el estrato de esta. Estas variables están acordes al comportamiento del mercado pues el tamaño de una vivienda, el estrato y ciertas comodidades son influyentes para determinar el precio de arriendo. Cabe resaltar también la presencia de la variable *tipo_pisos_freq* en el ranking de variables mas importantes siendo este resultado del tratamiento de datos realizado.

Fig. 22. IMPORTANCIA DE VARIABLES MODELO FINAL



Dado que ninguna de las variables presenta una importancia cercana a 0 se conservan todas las utilizadas en el proceso de entrenamiento.

Para finalizar, inicialmente se establecía que la efectividad del modelo resultante sería determinada por los valores de la variable *zona* (centro, poblado, belén, laureles y sa_prado), sin embargo, la aplicación del algoritmo de clustering evita que el negocio deba filtrar por esta variable y solo tenga calcular el clúster al cual pertenece una vivienda, de esta forma, el modelo queda abierto a cualquier tipo de vivienda bajo una condición más óptima.

Con lo anterior se tiene un modelo funcional y con las métricas del negocio para los inmuebles pertenecientes al clúster 0. Por tanto, las recomendaciones para implementar el modelo siguen las siguientes premisas:

- Si un inmueble pertenece al clúster 0 se ejecuta la predicción.
- Si un inmueble pertenece al clúster 1 NO se ejecuta la predicción y se debe buscar un modelo independiente para este.
- Si un inmueble pertenece al clúster 2 NO se ejecuta la predicción y se debe buscar un modelo independiente para este.

IX. DISCUSIÓN

El proceso evolutivo de los resultados para la obtención del modelo final se observa en la TABLA XXII, donde se visualiza como el tratamiento de datos, variación de hiperparámetros e implementación de estrategias de generación de nuevas variables aumenta la precisión de los resultados hasta encontrar las métricas deseadas.

TABLA XXIII. EVOLUCIÓN DEL MAPE A TRAVÉS DE LAS ITERACIONES

Iteración	MAPE	Mejor Modelo
Iteración 1	0.28	LinearRegression
Iteración 2	0.28	LinearRegression
Iteración 3	0.25	RandomForest
Iteración 4	0.21	XGBOOST
Iteración 5	0.19	XGBOOST
Iteración 6	0.15	XGBOOST
Iteración 7	0.15	XGBOOST

Es importante observar que la métrica objetivo debe ser acompañada de otros valores que evalúen la precisión del modelo para garantizar y reforzar los datos obtenidos. En la TABLA XXIV se observa cómo no solo se evoluciona el MAPE sino que se mejoran las métricas de R2 y RMSE hasta el punto de lograr un error de predicción aproximadamente de 490.008COP, lo cual es mucho mas tolerable para el usuario que el obtenido en la iteración 1.

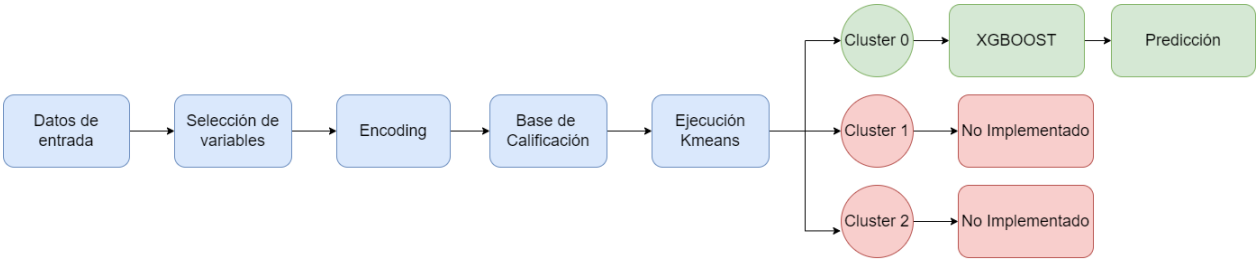
TABLA XXV. COMPARACIÓN DEL R2 Y RMSE ENTRE LA 1RA Y ÚLTIMA ITERACÓN

Iteración	Modelo	R2	RMSE	MAPE
Iteración 1	LinearRegression	0.72	1102245.72	0.28
Iteración 7	XGBOOST	0.74	490008.58	0.15

La estrategia de clusteting implementada para encontrar el modelo final le otorga un beneficio al negocio que le pueda dar más utilidad al modelo ya que al depender la predicción del resultado de un modelo de clasificación se disminuye el error de filtrado de los datos e incluso facilita automatizar el flujo de ejecución, pues si se aplicase el modelo dependiendo del valor de una variable como la zona se queda susceptible a los valores cambiantes y las variaciones en los tipos de vivienda dentro de una zona que puede acelerar la obsolescencia del modelo.

Al implementar este modelo se debe considerar todo el flujo de ejecución que este debe contener, pues este consiste en componentes de encoding de variables, ejecución de un algoritmo de clustering que determina la ejecución o no del proceso de predicción como se observa en la Fig 23.

Fig. 23 FLUJO DE IMPLEMENTACIÓN DEL MODELO



Es importante resaltar que los componentes de Encoding, Ejecución Kmeans y XGBOOST son objetos serializados obtenidos y almacenados desde el proceso de desarrollo.

Para completar el flujo se debe ejecutar un proceso de búsqueda de un modelo de machine learning para los clústeres 1 y 2 donde se debe realizar un proceso de iteraciones para encontrar un modelo que otorgue las métricas demandadas por el negocio para cada conjunto de datos. Por tanto, el proyecto ideal estaría compuesto de 4 modelos donde el primero es un modelo de clasificación que determinará cuál de los modelos de regresión se debe ejecutar.

El conjunto de datos del clúster 2 tiene muy pocos registros que puede ser razón de hasta ahora no lograr ajustar un modelo. Para este caso se debería completar estos datos con ejecuciones periódicas del proceso de scraping para completar una muestra que permita encontrar un buen modelo de regresión para las viviendas en este clúster.

X. CONCLUSIONES

- La necesidad inicial del negocio era encontrar un modelo aplicable a diferentes zonas de la ciudad de Medellín, sin embargo, se propone una solución mas sofisticada, pues la implementación del modelo no dependerá del valor de una variable categórica sino de las características internas de la vivienda, dando esto cobertura sobre múltiples zonas incluso si solo se tiene uno de los tres modelos que se deben lograr.
- La solución a este proyecto se concreta de la implementación de dos modelos de tipologías diferentes, clasificación y regresión, lo que muestra las múltiples opciones y recursos que se tienen para intervenir los datos y llegar a los resultados deseados
- Al extraer información de sitios web es importante estructurar una estrategia que permita revisar que los datos están siendo recolectados de manera correcta, pues esto puede causar problemas futuros en la implementación.
- Al querer extraer información de sitios web siempre es importante verificar sus políticas de seguridad y tratamiento de datos, pues es un punto importante en la ética del manejo de información
- Al realizar un proceso de scraping es importante tomar un tiempo prudente en elaborar una estructura ordenada y que garantice buenas características de mantenibilidad, seguridad y escalabilidad para hacer una ejecución efectiva y facilitar las iteraciones e implementación de la solución.
- El consumir este tipo de información o datos de mercados similares puede resultar en bases con una gran cantidad de variables categóricas, por ello, es importante definir una buena estrategia de encoding que eviten problemas de dimensionalidad en los datos.
- En la construcción de un modelo de Machine Learning se resalta la importancia de la estrategia de iteraciones y múltiples experimentos, pues esto lleva a tener una base de construcción y toda la trazabilidad de los resultados hasta llegar a las métricas deseadas.
- En la construcción de un proyecto de Machine Learnig no solo se debe garantizar el ejercicio estadístico sino también construir una solución intuitiva, con fácil mantenibilidad (pocas variables si el caso lo permite) y teniendo siempre presente las métricas del negocio pues son las que le darán vida al modelo final.

- En la construcción de este proyecto se resalta la importancia del tratamiento de datos, pues una base de entrenamiento con datos lógicos, donde se realice un buen análisis de distribuciones y correlación garantiza resultados más precisos e interpretables.
- Se resalta la importancia de utilizar una buena herramienta de versionado de código y una herramienta de trazabilidad para los experimentos en la construcción del modelo de Machine Learning.

XI. RECOMENDACIONES

Según el proceso ejecutado y lo aprendido en el desarrollo de este proceso se entregan las siguientes recomendaciones para trabajos futuros y de creación de modelos en general.

- Se puede considerar hacer de nuevo una implementación de este modelo con la información de sitios más robustos como lo es https://fincaraiz.com.co/, Pues este podría traer mucha mas cobertura de la ciudad y más características para maniobrar en el proceso de entrenamiento.
- Para la continuación de este proyecto se pueden considerar tipologías de modelos, como las observadas en [4] y [5] y no considerados en las iteraciones ejecutadas, estos pueden ser: Regresión Lineal Múltiple, Regresión Robusta, Proceso de Regresión Gaussiano, Redes Neuronales, entre otros. Estos pueden ser considerados para la búsqueda de los modelos para los datos en los clústeres 1 y 2.
- Este proyecto está construido con datos altamente cambiantes y un mercado que puede ser muy volátil, por tanto, es recomendable la implementación de alertas que indiquen cuando el modelo debe ser calibrado o reentrenado.
- Al crear un proceso de scraping se recomienda implementar su ejecución periódica pues el contenido de estos datos puede ser valioso para este y otros proyectos.
- En una posible recalibración o reentrenamiento de los modelos se debe considerar incluir mas variables al modelo de clasificación kmeans, pues al solo estar entrenado con 2 variables queda susceptible a la variación de muy pocas características y, al contar este tipo de bases de entrenamiento con gran cantidad de variables categóricas, se puede considerar cambiar a algoritmos como kmodes.

REFERENCIAS

- [1]. P. Herman. (2023 March 7). The importance of price prediction [Online]. https://www.future-processing.com/blog/the-importance-of-price-prediction/#what-is-price-prediction

- [2]. Asif Ahmed Neloy, H M Sadman Haque, Md. Mahmud Ul Islam. “Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring”. North South University. Dhaka 1229, Bangladesh

- [3]. Y. Ma, Z. Zhang, A. Ihler, B. Pan. “Estimating Warehouse Rental Price using Machine Learning Techniques”. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL. 2018

- [4]. Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, Ismail Ibrahim, " House Price Prediction using a Machine Learning Model: A Survey of Literature", International Journal of Modern Education and Computer Science (IJMECS), Vol.12, No.6, pp. 46-54, 2020.DOI: 10.5815/ijmecs.2020.06.04

- [5]. Embaye WT, Zereyesus YA, Chen B (2021) Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. PLoS ONE 16(2): e0244953. https://doi.org/10.1371/journal.pone.0244953

- [6]. Platzi. Curso de Fundamentos de Web Scraping con Python y Xpath. [Online]. https://platzi.com/cursos/webscraping/

- [7]. Platzi. Curso de Web Scraping: Extracción de Datos en la Web. [Online]. https://platzi.com/cursos/web-scraping/