**Logotipo

Descripción generada automáticamente**

**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. |

**[](https://co.creativecommons.org/?page_id=13)** [](https://co.creativecommons.net/tipos-de-licencias/)

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA)

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

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

Universidad de Antioquia - [www.udea.edu.co](http://www.udea.edu.co)

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama.

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Tabla de contenido

[RESUMEN 7](#_Toc138110616)

[ABSTRACT 8](#_Toc138110617)

[I. INTRODUCCIÓN 9](#_Toc138110618)

[II. PLANTEAMIENTO DEL PROBLEMA 10](#_Toc138110619)

[III. JUSTIFICACIÓN 11](#_Toc138110620)

[IV. OBJETIVOS 12](#_Toc138110621)

[VII. METODOLOGÍA 14](#_Toc138110622)

[**Scraping:** 14](#_Toc138110623)

[**Modelo:** 15](#_Toc138110624)

[VIII RESULTADOS 16](#_Toc138110625)

[**Scraping:** 16](#_Toc138110626)

[**Modelado**: 23](#_Toc138110627)

[Tratamiento adicional de datos: 23](#_Toc138110628)

[Iteración 1: 25](#_Toc138110629)

[Iteración 2: 28](#_Toc138110630)

[Iteración 3: 36](#_Toc138110631)

[Iteración 4: 38](#_Toc138110632)

[Iteración 5: 39](#_Toc138110633)

[Iteración 7 48](#_Toc138110634)

[IX. DISCUSIÓN 52](#_Toc138110635)

[X. CONCLUSIONES 54](#_Toc138110636)

[XI. RECOMENDACIONES 55](#_Toc138110637)

[REFERENCIAS 56](#_Toc138110638)

LISTA DE TABLAS

[TABLA I RESULTADOS DE LA EXPLORACIÓN DE SITIOS WEB 17](#_Toc138110667)

[TABLA II BASE RESULTANTE POR WEB SCRAPING 24](#_Toc138110668)

[TABLA III MEDIDAS DE DISPERSIÓN Y TENDENCIA CENTRAL DE LA VARIABLE *PRECIO* 27](#_Toc138110669)

[TABLA IV CALCULO DE OUTLIERS VARIABLE *PRECIO* 27](#_Toc138110670)

[TABLA V 28](#_Toc138110671)

[TABLA VI ELIMNACIÓN DE ATÍPICOS SOBRE LAS VARIABLES CATEGÓRICAS NUMÉRICAS 30](#_Toc138110672)

[TABLA VII CORRELACIONES ENTRE VARIABLES PREDICTORA 35](#_Toc138110673)

[TABLA VIII MODELO ITERACION 2 36](#_Toc138110674)

[TABLA IX. BASE DE ENTRENAMIENTO RESULTADO ITERACIÓN 2 37](#_Toc138110675)

[TABLA X. COMPARACION RESULTADOS DE MÉTRICAS ITERACIÓN 3 38](#_Toc138110676)

[TABLA XI. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 4 39](#_Toc138110677)

[TABLA XII: DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5\_1 41](#_Toc138110678)

[TABLA XIII. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5\_1 41](#_Toc138110679)

[TABLA XIV. DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5\_2 42](#_Toc138110680)

[TABLA XV. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5\_2 42](#_Toc138110681)

[TABLA XVI. DISTRIBUCION DE PRECIOS SUB ITERACIÓN 5\_3 43](#_Toc138110682)

[TABLA XVII. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5\_4 45](#_Toc138110683)

[TABLA XVIII. EVALUACIÓN DE COEFICIENTE DE SILUETA 47](#_Toc138110684)

[TABLA XIX. RANGO DE PRECIOS POR CLUSTER 47](#_Toc138110685)

[TABLA XX: TABLA DE EJECUCIONES ITERACIÓN 6 (XGBOOST) 48](#_Toc138110686)

[TABLA XXI. RESULTDOS MÉTRICAS MODELO FINAL 50](#_Toc138110687)

[TABLA XXII. EVALUACIÓN DEL MODELO POR ZONAS 51](#_Toc138110688)

[TABLA XXIII. EVOLUCIÓN DEL MAPE A TRAVÉS DE LAS ITERACIONES 53](#_Toc138110689)

[TABLA XXIV. COMPARACIÓN DEL R2 Y RMSE ENTRE LA 1RA Y ÚLTIMA ITERACÓN 53](#_Toc138110690)

LISTA DE FIGURAS

[Fig. 1. DISTRIBUCIÓN DE INMUEBLES POR ZONAS 22](#_Toc138110691)

[Fig. 2. DISTRIBUCION DE LA VARIABLE *PRECIO* 27](#_Toc138110692)

[Fig. 3. DISTRIBUCIÓN DE VARIABLE PRECIO SIN OUTLIERS 29](#_Toc138110693)

[Fig. 4. DISTRIBUCIÓN VARIABLES NUMÉRICAS 30](#_Toc138110694)

[Fig. 5. DISTRIBUCIÓN VARIABLES CATEGÓRICAS NUMÉRICAS 31](#_Toc138110695)

[Fig. 6 DISTRIBUCION DE VARIABLES CATEGÓRICAS NUÉRICAS SIN OUTLIERS 32](#_Toc138110696)

[Fig. 7 DISTRIBUCIÓN VARIABLES CATEGÓRICAS 33](#_Toc138110697)

[Fig. 8 CORRELACIÓN VARIABLE OBJETIVO VS VARIABLES PREDICTORAS 34](#_Toc138110698)

[Fig. 9 CORRELACIÓN VARIABLES NUMÉRICAS 35](#_Toc138110699)

[Fig. 10 CORRELACIÓN VARIABLES CATEGÓRICAS NUMÉRICAS 35](#_Toc138110700)

[Fig. 11 CORRELACIÓN VARIABLES CATEGÓRICAS 36](#_Toc138110701)

[Fig. 12. DISTRIBUCIÓN DE PRECIOS DE ARRIENDO CON FILTRO 6000000 46](#_Toc138110702)

[Fig. 13. DISTRIBUCIÓN DE LOS CLUSTERS 48](#_Toc138110703)

[Fig. 14. DISTRIBUCION PRECIOS CLUSTER 0 49](#_Toc138110704)

[Fig. 15. DISTRIBUCION DE PRECIOS CLUSTER 1 49](#_Toc138110705)

[Fig. 16. DISTRIBUCION DE PRECIOS CLUSTER 2 49](#_Toc138110706)

[Fig. 17. DISTRIBUCION DE PRECIOS LUEGO DE SUBMUESTREO 51](#_Toc138110707)

[Fig. 18. IMPORTANCIA DE VARIABLES MODELO FINAL 52](#_Toc138110708)

[Fig. 19 FLUJO DE IMPLEMENTACIÓN DEL MODELO 55](#_Toc138110709)

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 7 iteraciones siendo la primera la línea base de la solución y la iteración 7 la implementación de una serie de tratamiento de datos donde se crea un modelo auxiliar de clustering que permite realizar un submuestreo de la base original y crear una nueva variable de la base de entrenamiento para así obtener los resultados deseados.

***Palabras clave*** — **regresión, scraping, machine learning, datos.**

# 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 iteration 7, the implementation of a series of data processing where an auxiliary clustering model is created that allows subsampling the original base and creating a new variable from the training base in order to obtain the desired results.

***Keywords*** —**regression, scraping, machine learning, data.**

# 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.

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. El modelo será construido mediante múltiples iteraciones donde al utilizar diferentes topologías y parámetros 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.

# 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 se selecciona porque permite abordar dos temáticas claves en los sistemas de machine learning: web scraping y modelos de regresión supervisados. Ese desarrollo permitirá construir un proceso automático que ofrezca al usuario una ayuda en la estimación de precios de sus inmuebles si este busca inversión o si busca una residencia.

Adicional a esto permite explorar el funcionamiento e implementación de los sistemas de recolección de datos, además de poder explorar soluciones de machine learnig que resuelven necesidades cotidianas de los usuarios y que se acoplan al comportamiento de un mercado específico.

# 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 de dos investigaciones previamente realizadas que se describen a continuación:

La predicción de precios de arriendo en la ciudad de Dhaka (Bangladés), dividido en dos principales iteraciones, la primera enfocada en el entrenamiento de modelos para 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. Las métricas sobre las cuales se enfocan los resultados son R2 y RMSE [1].

Cómo aplicar técnicas de machine learning en un ejercicio de predicción de precios de alquiler. La metodología consiste en mostrar la implementación de modelos como: Linear Regression, Regression Tree, Random Forest Regression y Gradient Boosting Regression Trees. La evaluación de lo modelos se centra en la métrica RMSE [2].

# 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.

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 reentrenar el algoritmo con mejor desempeño con estos grupos en busca del modelo final. Adicional, la etiqueta obtenida del clustering será una variable mas de la base de entrenamiento.
* **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, éste 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, 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 Medellin
  + **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:**

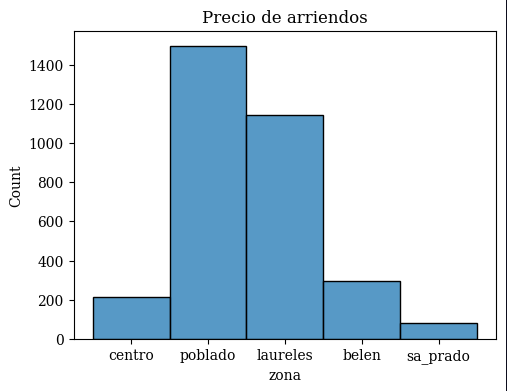
|  |
| --- |
| * + - Laureles |
| * + - La Castellana |
| * + - Robledo |
| * + - Simón Bolívar |
| * + - El Nogal |
| * + - Calasanz |
| * + - Pilarica |
| * + - Estadio |
| * + - San Javier |
| * + - San German |
| * + - La América |
| * + - Los Colores |
| * + - López de Mesa |
| * + - Córdoba |
| * + - 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 medellin
    - 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 vivienda*s. En la Fig 1 se puede apreciar la distribución del número de inmuebles por zona.

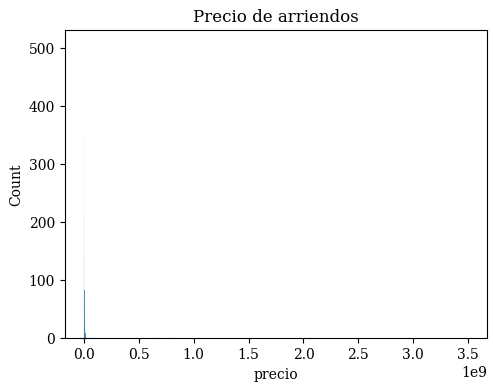
Fig. 1. DISTRIBUCIÓN DE INMUEBLES POR ZONAS

****

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.< DISTRIBUCION DE PRECIOS DE ARRIENDOS



En la figura 2 se observa la distribución de los precios. Este nos muestra que hay valores a una escala muy superior a la masa central de los datos, lo cual nos indica que 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
    - Prqueadero 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.
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 comodidaes 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

|  |  |
| --- | --- |
| Caracteritica | 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 categorias 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:**

- *ingral alacena* equivale a *integral*

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

**tipo\_pisos:**

- *madera laminada* equivale a *madera*

- *madera laminada marmol* 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

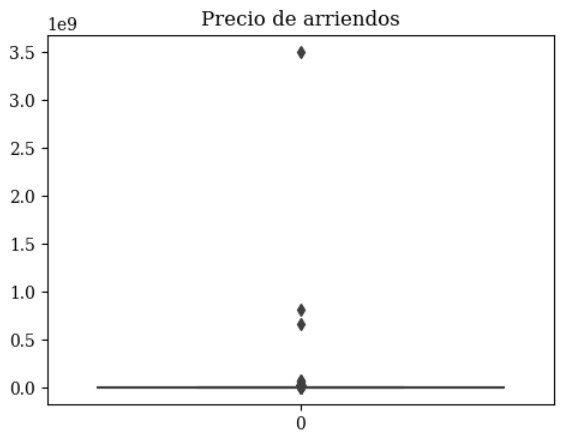
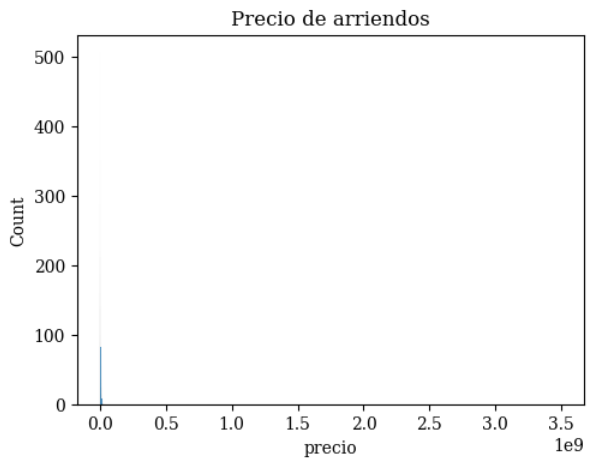
Se bien se ha logrado una reducción sustancial de valores aún se tienen demasiadas categorías de estas variables para hacer un proceso de encoding tradicional. Por tanto, la segunda intervención consiste en hacer un encoding por medio de frecuencias. Este proceso consiste en reemplazar el valor de cada 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 y nans de la base. 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. 2. DISTRIBUCION DE LA VARIABLE *PRECIO*



En la Fig. 2 se obtiene la distribución de la variable objetivo a través de un gráfico de dispersión y un gráfico de cajas, se puede observar la presencia de atípicos ya que marcan la existencia de precios a un valor muy superior al conjunto 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 MM 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*

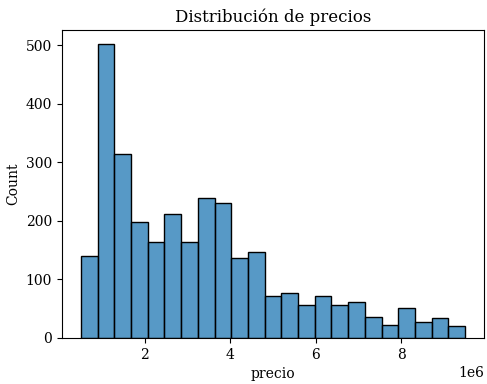
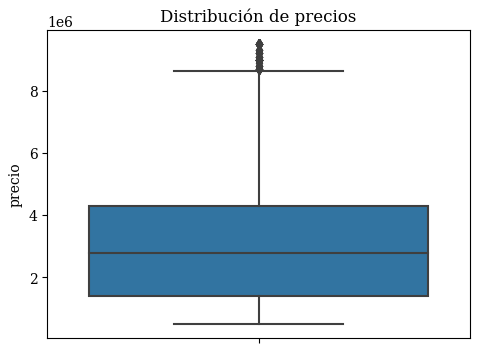
|  |  |
| --- | --- |
| Medida | Valor |
| Media | 447391999.773014 |
| Desviación estándar | 1233632087.282006 |
| Mínimo | 0.000000 |
| 25% | 1125805.250000 |
| 50%  75%  Máximo | 3900000.000000  20162403.136915  3500000000.000000 |
|  |  |

TABLA IV CALCULO 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. 3, donde se observa una distribución mas compacta de la variable y la eliminación de los datos atípicos

Fig. 3. 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  RMSE  MAPE | 0.72  1102245.72  0.28 |
|  |  |

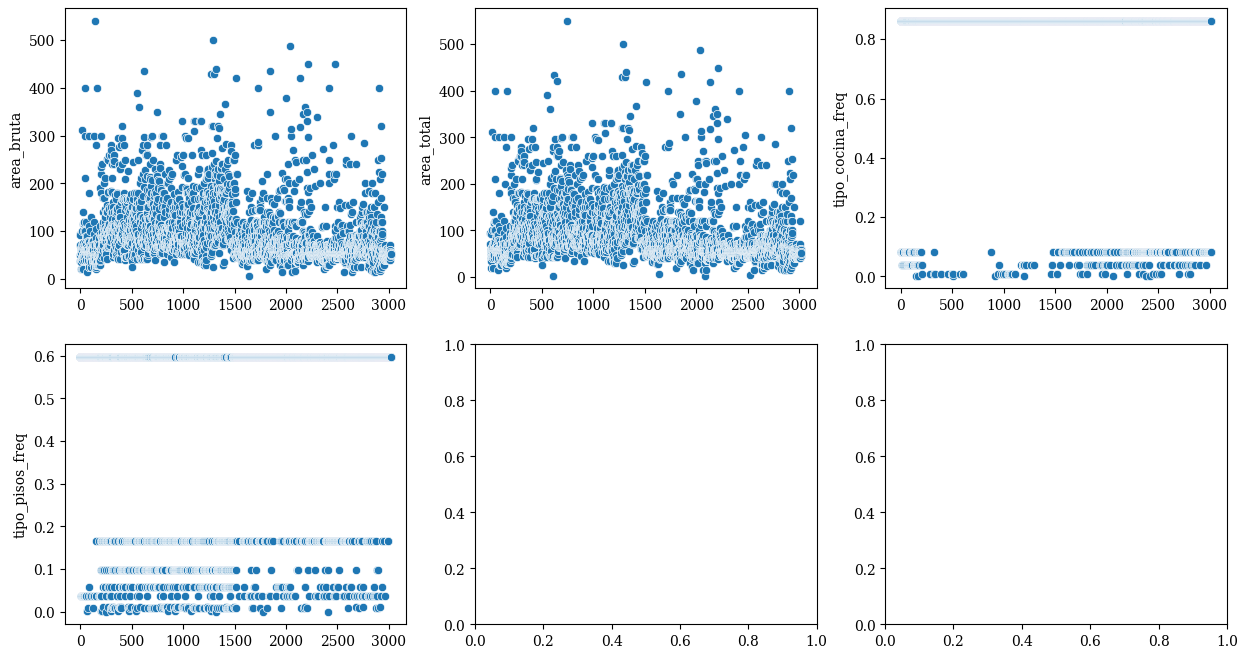
Dada la TABLA V se obtiene un modelo con métricas aceptables, con un MAPE 0.13 puntos por encima del deseado y un RMSE que nos indica un error de predicción de aproximadamente 1’102.245COP lo cual para un usuario es un margen de error muy amplio y que se buscará reducir en las próximas iteraciones.

### 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:

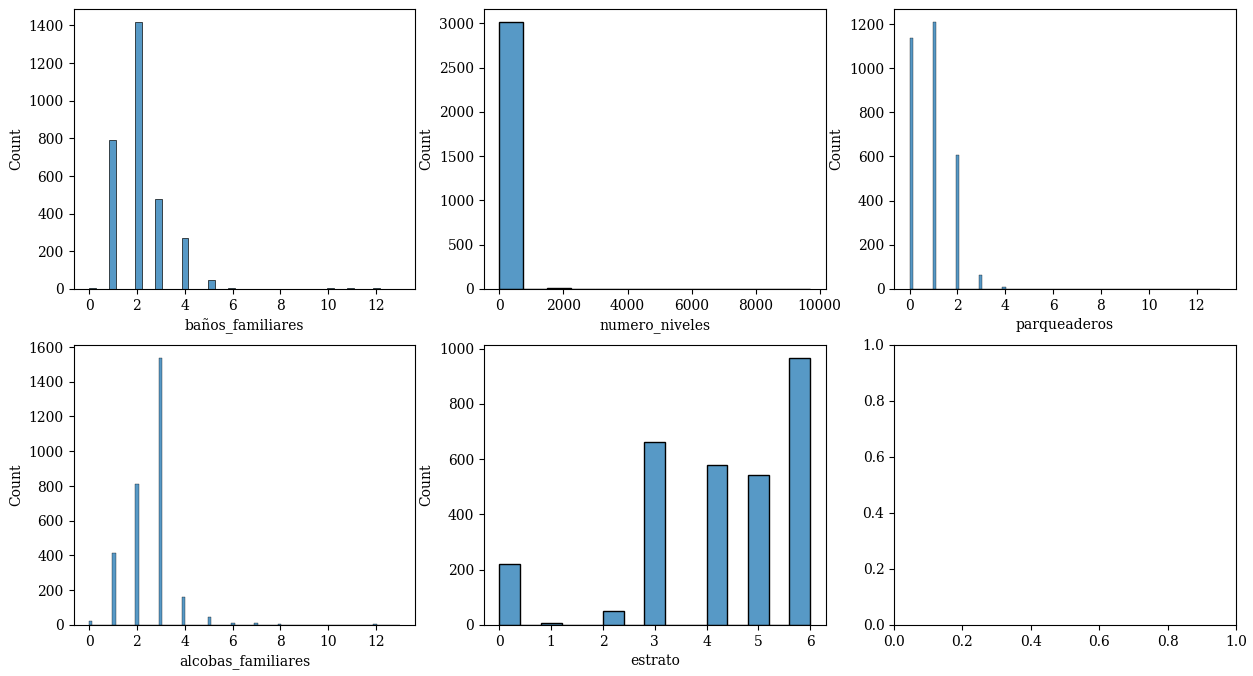
* Eliminación de outliers de la variable objetivo y codificación de variables categóricas: se preserva el tratamiento de datos realizado en la iteración 1.
* Observación de distribución de los datos: Se grafica la distribución de los datos de cada tipo de variables obtenidas

Fig. 4. DISTRIBUCIÓN VARIABLES NUMÉRICAS



En la figura 4 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 comporamiento observado de una gran cantidad de datos concentrados en una de las frecuencias.

Fig. 5. DISTRIBUCIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



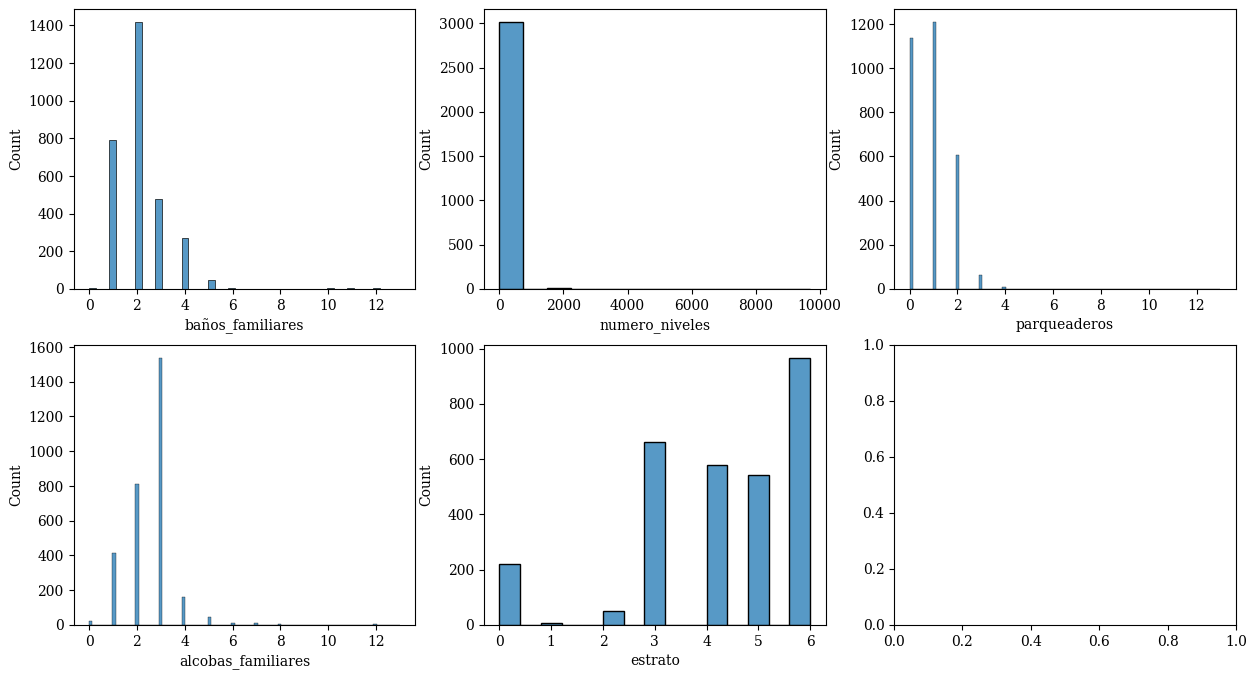
En la figura 5 se presentan las variables numéricas, pero con valores discretos, las cuales se les da un tratamiento de variables categóricas. Se puede observar la presencia de valores atípicos en las variables de *baños\_familiares, número\_niveles, parqueaderos y alcobas\_familiares*, pues valores 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 |

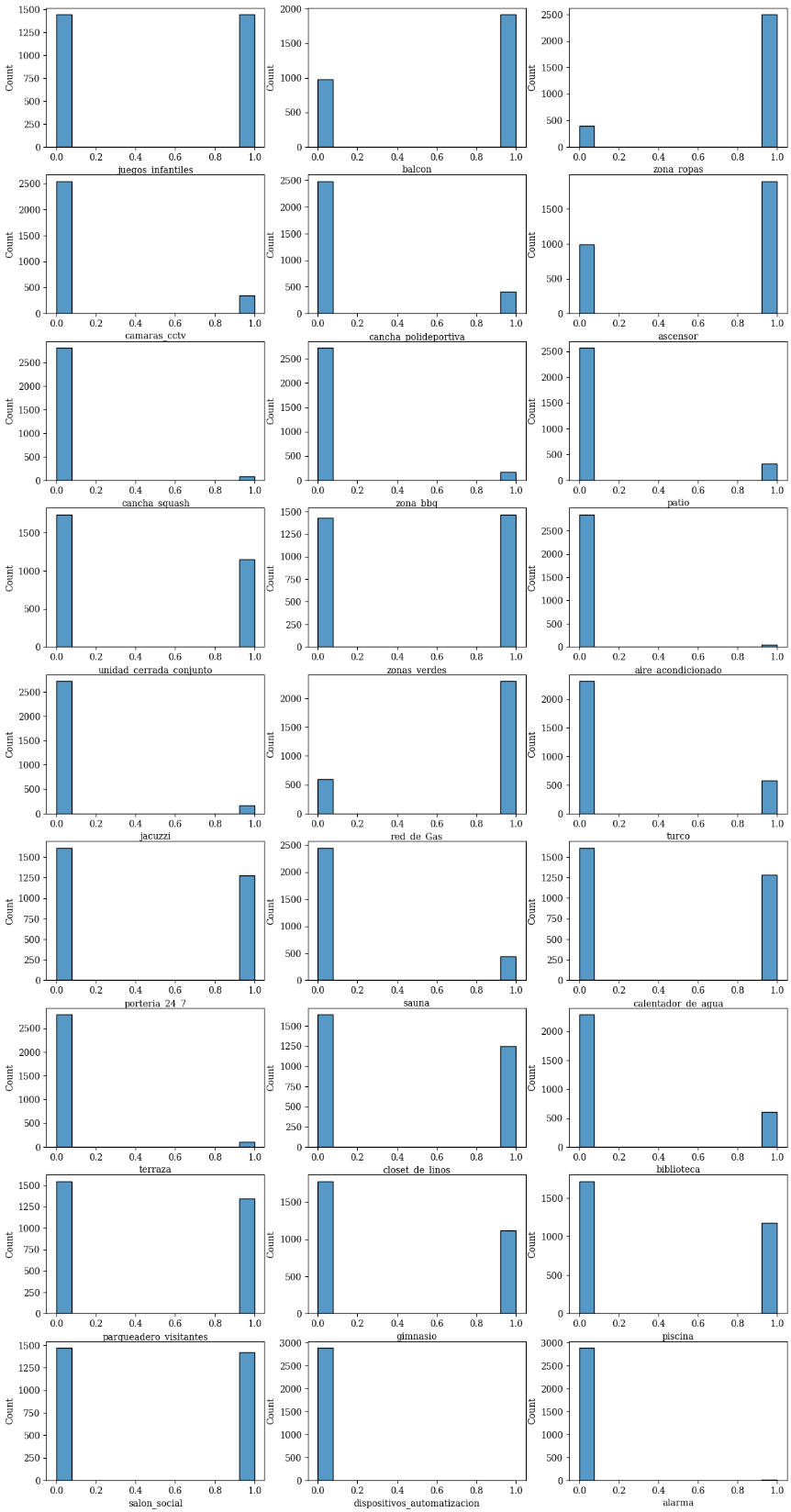
Con el tratamiento realizado en la TABLA IV se obtiene la distribución de la Fig 5. Con lo cual los valores para estas variables tienen una representación más lógica.

Fig. 6 DISTRIBUCION DE VARIABLES CATEGÓRICAS NUÉRICAS SIN OUTLIERS



En la Fig. 7 Se observa las distribuciones de las variables categóricas booleanas, a estas representar 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

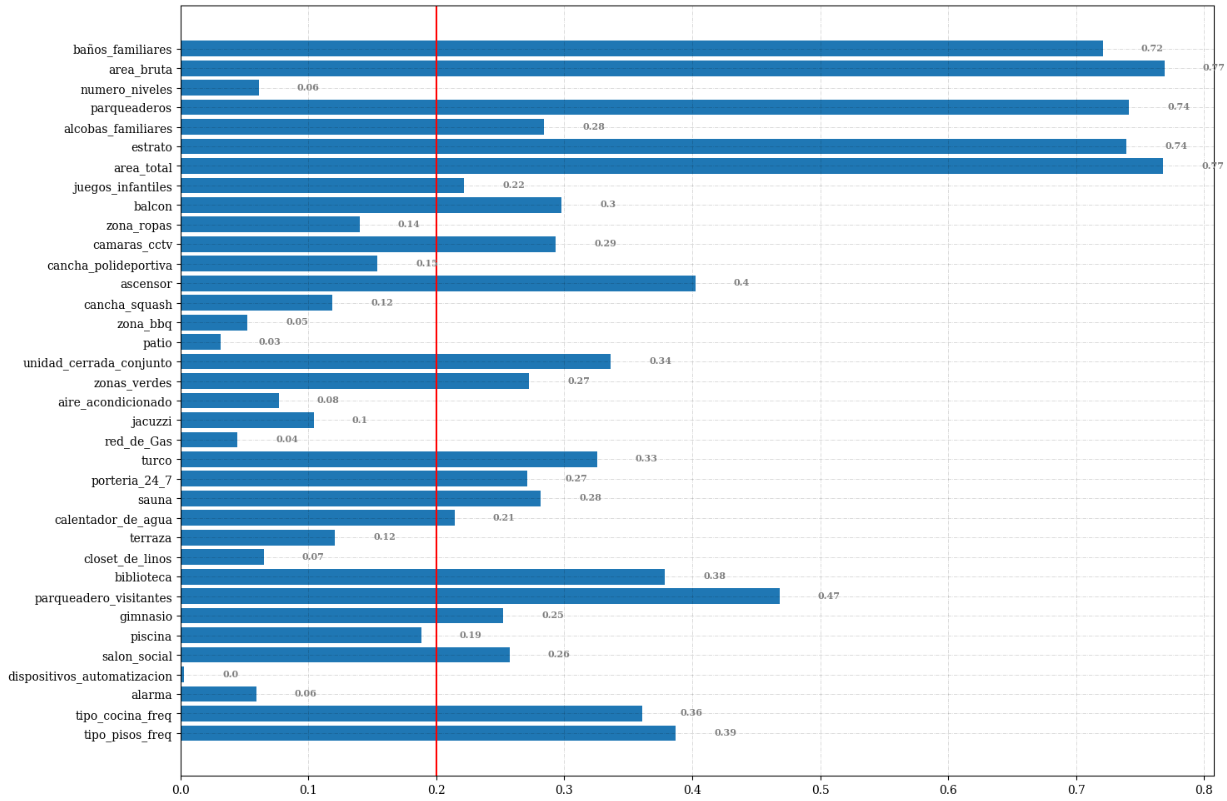
casi ilógico que una casa o apartamento no la posea, por ejemplo variables como *zona\_de\_ropas* ó *red\_de\_gas*, sin embargo, también sucede el caso contrario con comodidades que son escasas para la mayoría de las viviendas como lo son *cancha\_sqash* o *aire\_acondicionado*. Estas variables tendrán 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. 7 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. 8 CORRELACIÓN VARIABLE OBJETIVO VS VARIABLES PREDICTORAS



En la Fig. 8 se aprecia la correlación absoluta de cada una de las variables predictoras con respecto a la variable objetivo, adicional 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

Definiendo 0.7 como valor para determinar si dos variables están altamente correlacionadas se realizan las siguientes ejecuciones

Fig. 9 CORRELACIÓN VARIABLES NUMÉRICAS

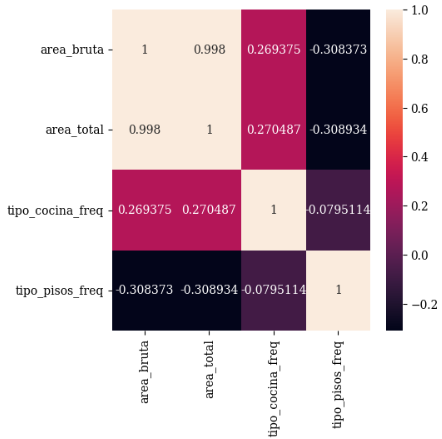
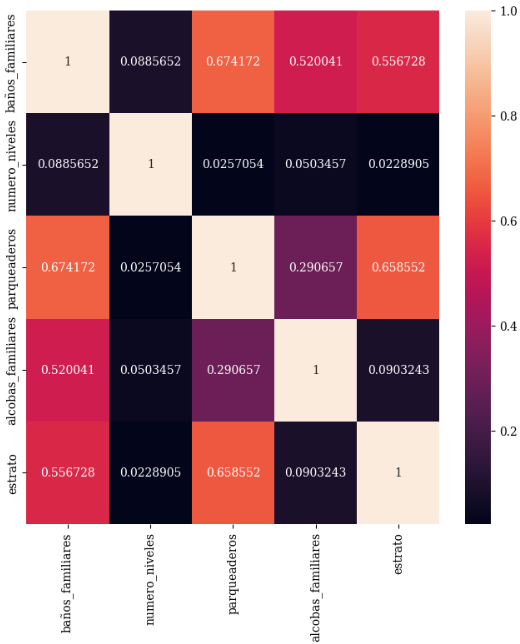


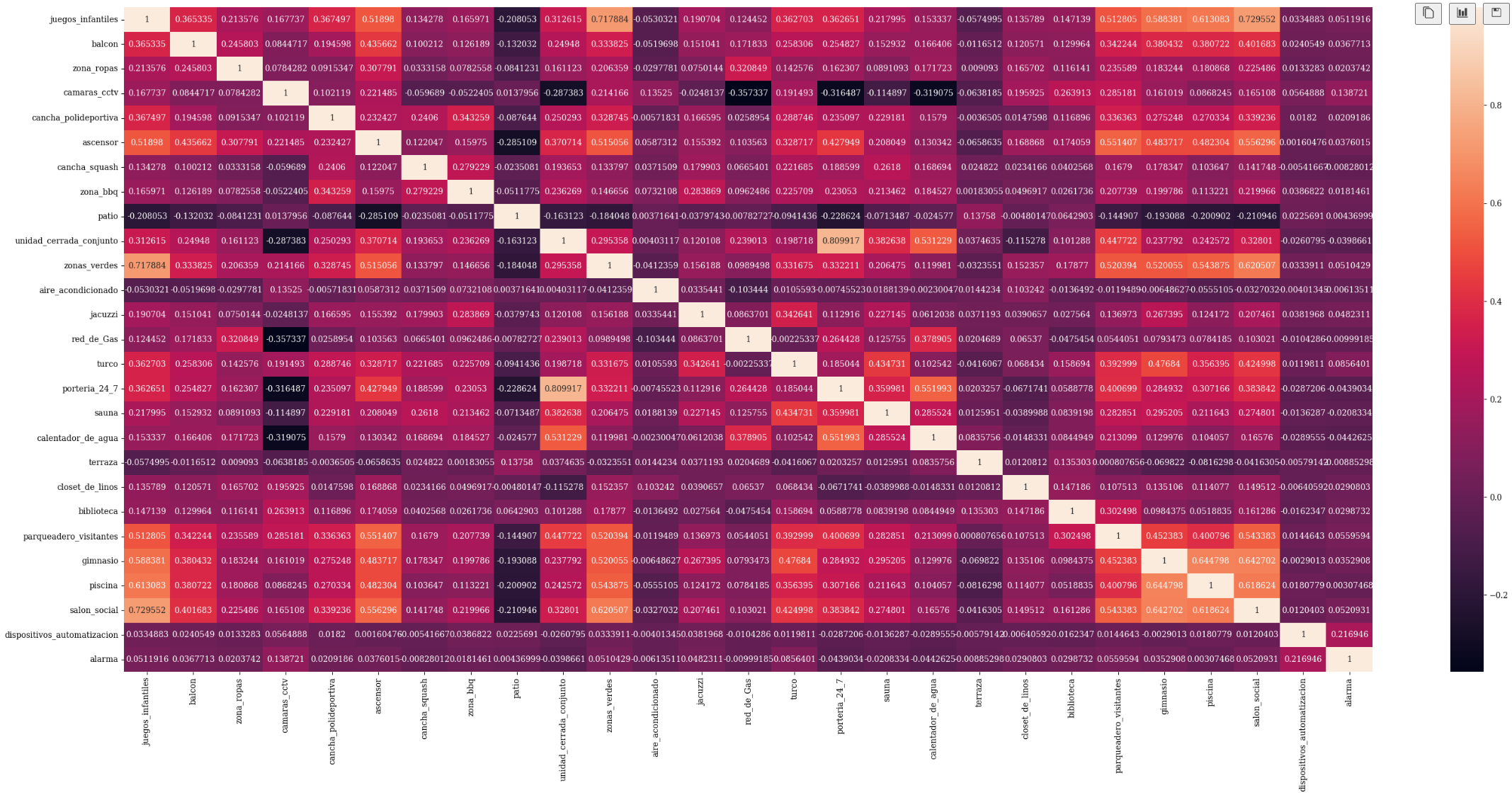
Fig. 10 CORRELACIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



En la Fig. 9 se observa una 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. 10 se tienen el análisis de correlaciones para las variables numéricas categóricas en el cual no se presencia valores de alta correlación entre ellas.

Fig. 11 CORRELACIÓN VARIABLES CATEGÓRICAS



En la Fig 11 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
  + Un conjunto de unidad cerrada es, en la mayoría de los casos, implica 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 CORRELACIONES ENTRE VARIABLES PREDICTORA

|  |  |  |
| --- | --- | --- |
| Variable1 | Variable 2 | Valor |
| baños\_familiares  baños\_familiares | area\_bruta  area\_total | 0.78  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 en los puntos previos se eliminan 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 en el escenario que se indica en la TABLA VIII

TABLA VIII MODELO ITERACION 2

|  |  |
| --- | --- |
| Medida | Valor |
| %Entrenamiento | 75% |
| % Prueba | 25% |
| Modelo | LinearRegression |
| R2  RMSE  MAPE | 0.70  1116690.71  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 ejutar 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 hiperparámtros 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 test
* 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 test
  + 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 |
|  |  |  |  |  |  |  |

* Para estas ejecuciones se obtiene una mejora sustancial en las métricas del modelo para los algoritmos de Gradient Boostin y XG Boost 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 evaluar 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

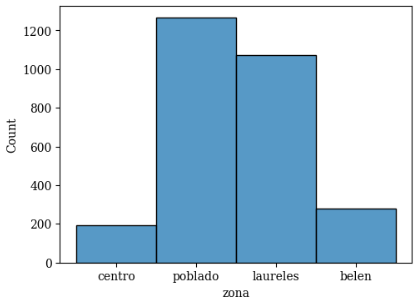
Dado que en cada sub iteración se está cambiando la base de entrenamiento se debe considerar repetir el tratamiento de datos de datos de distribución y correlación realizado en la iteración dos, así como una variación de los hiperparámetros si así lo necesita cada algoritmo.

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 iteracion 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, Belen y Centro

TABLA XII: 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

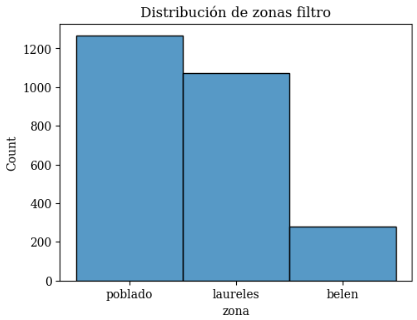
TABLA XIII. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5\_1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |
|  |  |  |  |  |  |  |

Se observan 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

TABLA XIV. 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 XV. COMPARACION RESULTADOS DE MODELOS ITERACIÓN 5\_2

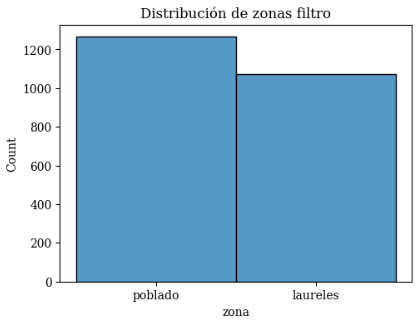
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modelo | R2 Train | R2 Test | RMSE Train | RMSE  Test | MAPE Train | MAPE test |
| RandomForestRegressor | 0.78 | 0.73 | 962915.73 | 1085599.87 | 0.22 | 0.25 |
| SVR | 0.74 | 0.71 | 1043737.84 | 1124187.99 | 0.2 | 0.24 |
| Gradient 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 |
|  |  |  |  |  |  |  |

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

TABLA XVI. 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*. 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:

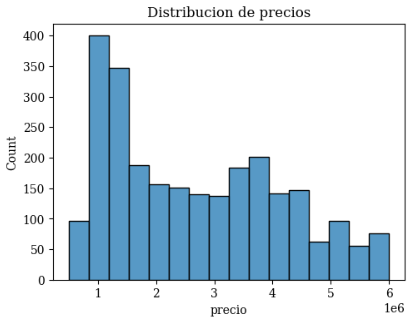
En el análisis por zonas se ha optado por eliminar aquellas con menos presencia en la base. Esta sub iteración se basa en hacer una exploración más a fondo sobre la variable *precio* analizando los rangos de valores que toma esta y así seleccionar un conjunto de datos para ejecutar el entrenamiento.

En la Fig 3 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 11) 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 iteracion 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. 12. DISTRIBUCIÓN DE PRECIOS DE ARRIENDO CON FILTRO 6000000



* Resultados modelo: Los modelos SVR y Gradient Boosting no obtuvieron resultados concluyentes para esta sub iteración

TABLA XVII. 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 |
|  |  |  |  |  |  |  |

* 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, se hará una experimentación adicional que consiste en implementar una estrategia de clustering para determinar de manera óptima el submuestreo de la base de entrenamiento. La etiqueta resultante del clustering será adicionada como nueva variable en la base de entrenamiento para la ejecución de los modelos.

*Iteración 6:*

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

* La distribución de la base precios tienen un sesgo hacia las casas de menor costo, dejando las zonas que tienen arriendos más caros con muy poca presencia ante el modelo
* El análisis por zonas no muestra una solución para encontrar la métrica deseada, incluso hubo deterioro en estas
* Al realizar una ejecución sobre un rango de precios específico hubo una mejora en las métricas, lo cual indica que el análisis se debe enfocar en la variable *precio*

Para seleccionar el tipo de submuestreo de manera óptima se implementará un algoritmo de *kmeans* considerando solo las variables *estrato* y *area\_bruta* (siendo estas numéricas y han mostrado mayor influencia en los análisis de correlación (Fig 8)). La cantidad de clusters a seleccionar será determinado calculando el coeficiente de siluetas.

Luego de entrenar el modelo *kmeans* las etiquetas resultantes se añaden como una variable adicional en el proceso de entrenamiento.

Clusteting.

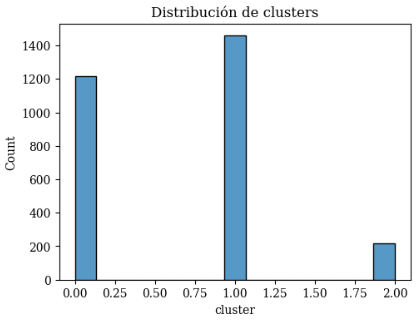
Se reutiliza la base de entrenamiento resultante de la iteración 2 y las variables *estrato* y *area\_bruta* para obtener los resultados de la TABLA XVIIII. En esta se puede observar como se obtiene un mejor valor de coeficiente de silueta para 3 clusters, por tanto, este será el valor para entrenar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

Al entrenar el modelo y calcular las etiquetas resultantes se obtiene la distribución de la Fig 13.

TABLA XVIII. EVALUACIÓN DE COEFICIENTE DE SILUETA

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

Fig. 13. DISTRIBUCIÓN DE LOS CLUSTERS



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

TABLA XIX. RANGO DE PRECIOS POR CLUSTER

|  |  |  |
| --- | --- | --- |
| Numero de Clusters | Mínimo | Máximo |
| 0 | 500000 | 9000000 |
| 1 | 700000 | 9500000 |
| 5 | 550000 | 8700000 |
|  |  |  |

Dada la tabla XIX y las distribuciones observadas en las figuras 14 -16, 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
* Dado que en la iteración anterior se mostró una mejoría en métricas al seleccionar un rango de precios, para este caso, se selecciona la base de entrenamiento de acuerdo a los cluster lo cual nos resulta en 3 ejecuciones:
  + ejecución 1: Base de entrenamiento solo con datos de cluster 0 (Tiene el rango más amplio de precios)
  + ejecución 2: Base de entrenamiento con datos de los clusters 0 y 1
  + ejecución 3: Base de entrenamiento con un porcentaje de datos de cada cluster

Fig. 14. DISTRIBUCION PRECIOS CLUSTER 0

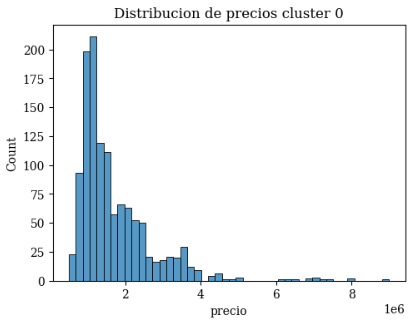


Fig. 15. DISTRIBUCION DE PRECIOS CLUSTER 1

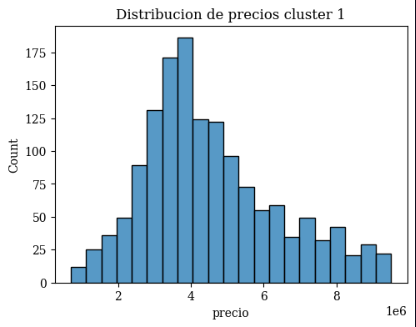


Fig. 16. DISTRIBUCION DE PRECIOS CLUSTER 2

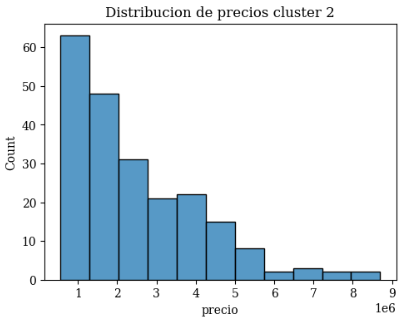


TABLA XX: TABLA DE EJECUCIONES ITERACIÓN 6 (XGBOOST)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ejecución | R2 Train | R2 Test | RMSE Train | RMSE  Test | MAPE Train | MAPE test |
| XGBOOST: Ejecución 1 | 0.75 | 0.72 | 511152 | 504298.55 | 0.15 | 0.15 |
| XGBOOST: Ejecución 2 | 0.79 | 0.75 | 950955.44 | 1073306 | 0.17 | 0.19 |
| XGBOOST: Ejecución 3 | 0.9 | 0.86 | 556334.30 | 699427.97 | 0.12 | 0.15 |
|  |  |  |  |  |  |  |

En la TABLA XX se tiene el consolidado de las diferentes ejecuciones donde en la ejecución 1, que corresponde a un entrenamiento con los datos del clúster 0, se obtiene la métrica buscada MAPE = 0.15, sin embargo, esta muestra de datos deja por fuera un alto porcentaje de información de la base inicial, por ello, se realiza la ejecución 2 donde se considera la información de los clústeres 0 y 1 obteniendo una métrica por encima del deseado.

Dado lo anterior y buscando añadir información a los datos del clúster 0 para ejecutar un entrenamiento con la mayor cobertura de datos de la base, se realiza una serie de ejecuciones donde se busca adherir información de los clústeres 1 y 2 considerando el sesgo en la Fig 14 y conservando el valor de la métrica MAPE <= 0.15. Esto se logra en la ejecución 3 donde la base de entrenamiento contiene:

* 100% de los datos del cluster 0.
* Datos del cluster 1 donde el precio esta en el rango [3800000, 6900000]
* Datos del cluster 2 donde el precio está en el rango [2000000, 6900000]

De esta forma se encuentra el modelo final con las métricas buscadas y una base de entrenamiento construida de acuerdo a los resultados de la ejecución del algortimo de kmeans y la distribución de los precios.

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 para el algoritmo de xgboost, el entrenamiento del algoritmo kmeans y la implementación 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 un análisis por zonas para verificar la efectividad del modelo y establecer una recomendación al negocio de cómo se debe implementar.

Fig. 17. DISTRIBUCION DE PRECIOS LUEGO DE SUBMUESTREO

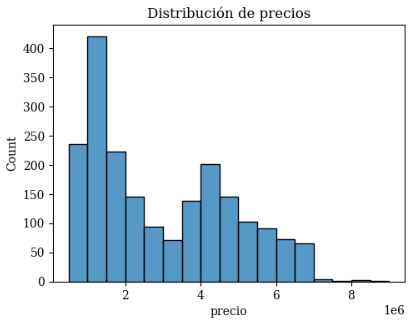


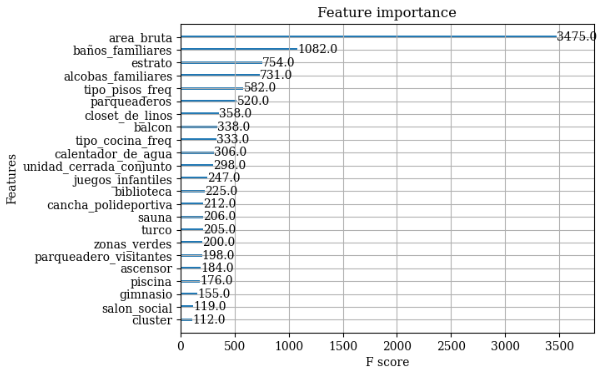
TABLA XXI. RESULTDOS MÉTRICAS MODELO FINAL

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modelo | R2 Train | R2 Test | RMSE Train | RMSE  Test | MAPE Train | MAPE test |
| XGBOOST | 0.92 | 0.86 | 488993.87 | 674427.58 | 0.11 | 0.15 |
|  |  |  |  |  |  |  |

En la figura 17 se tiene la distribución de precios obtenido del submuestreo realizado en la iteración 6 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 18. En esta se aprecia la gran influencia que tiene el area bruta de una vivienda para determinar su precio, seguido del número de baños familiares 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. 18. 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 las iteraciones se realiza un calculo de la métrica MAPE sobre cada conjunto determinado por las zonas obtenidas en el proceso de scraping, las cuales son: Centro, Poblado, Laureles, Belén y San Antonio de Prado

TABLA XXII. EVALUACIÓN DEL MODELO POR ZONAS

|  |  |
| --- | --- |
| Zona | MAPE |
| Centro | 0.25 |
| Poblado | 0.23 |
| Laureles | 0.33 |
| Belén | 0.35 |
| San Antonio de Prado | 0.44 |
|  |  |

En la tabla XXII se observa como el valor del MAPE no está dentro del rango deseado por el negocio en ninguna de las zonas, esto puede deberse a que dentro de cada una de ellas exista tipos de viviendas que no logra predecir de manera precisa el modelo, por tanto, las condiciones de implementación deben basarse no por zona sino por características internas de las viviendas.

Un ejemplo de esto es que, al seleccionar una muestra de viviendas donde el área bruta es mayor a 0.2 (valor escalado) y el estrato es igual a 0.5 (valor escalado) se obtiene un MAPE de 0.12. Debido a esto las condiciones de implementación del modelo quedaran sujetas a premisas como:

* Este modelo es funcional solo para inmuebles con un área bruta mayor a *N* mts
* Este modelo es funcional solo para inmuebles con un área bruta mayor a *N* mts y *M* baños
* Este modelo es funcional solo para inmuebles con un área bruta mayor a *N* mts y estrato *K*

El hallar estas condiciones se deja para un trabajo futuro del proyecto.

# IX. DISCUSIÓN

El proceso evolutivo de los resultados para la obtención del modelo final se observa en la TABLA XXIII, 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 |
|  |  |  |

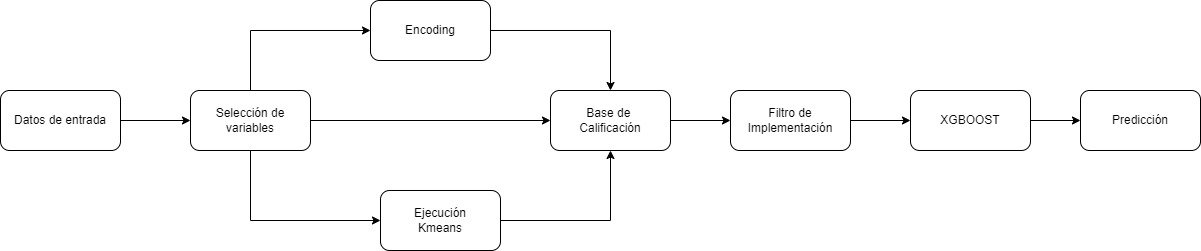
Es importante observar que la métrica objetivo debe ser acompañada de otros valores que evalúen la precisión del modelo para garantizar y reforzar los datos obtenidos. En la TABLA XXIV se observa como no solo se evoluciona el MAPE sino que se mejoran las métricas de R2 Y RMSE hasta el punto de lograr un error de predicción aproximadamente de 674.427COP, lo cual es mucho mas tolerable para el usuario que el obtenido en la iteración 1.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteración | Modelo | R2 | RMSE | MAPE |
| Iteracion 1 | LinearRegression | 0.72 | 1102245.72 | 0.28 |
| Iteración 7 | XGBOOST | 0.86 | 674427.58 | 0.15 |
|  |  |  |  |  |

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 y debe pasar el filtro de implementación del negocio como se observa en la Fig 19.

Fig. 19 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.

El filtro de implementación que contiene las reglas del negocio de dónde se decide utilizar el modelo debe construirse, según los resultados obtenidos en la iteración 7, de encontrar un patrón en las características de viviendas que conserve el MAPE inferior a 15%, de esta manera se asegura una alta precisión en la usabilidad del modelo y así otorgarle al negocio el activo que necesita.

# X. CONCLUSIONES

* 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.
* 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.
* Para este proyecto se obtiene un modelo que cumple con la métrica establecida por el negocio (MAPE <= 0.15), sin embargo, su implementación dependerá de la construcción de un filtro de negocio en base a una o mas variables que indique sobre qué tipo de inmuebles debe ser implementado.

# 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.

* Utilizar una buena herramienta de versionado de código y una herramienta de trazabilidad para los experimentos en la construcción del modelo de Machine Learning.
* 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
* Si se desea continuar con la implementación de este proyecto se podría explorar la construcción del modelo de clustering con más variables de la base de entrenamiento y evaluar efectividad de esta en las métricas del modelo de regresión.
* Este proyecto este 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 re entrenado.
* 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.

# 

# REFERENCIAS

* [1]. 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
* [2]. Y. Ma, Z. Zhang, A. Ihler, B. Pan. “Estimating Warehouse Rental Price using Machine Learning Techniques”. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL. 2018