**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.

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

Biblioteca Carlos Gaviria Díaz

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

Universidad de Antioquia - www.udea.edu.co

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](#_Toc134133591)

[ABSTRACT 8](#_Toc134133592)

[I. INTRODUCCIÓN 9](#_Toc134133593)

[II. PLANTEAMIENTO DEL PROBLEMA 10](#_Toc134133594)

[III. JUSTIFICACIÓN 11](#_Toc134133595)

[IV. OBJETIVOS 12](#_Toc134133596)

[V. HIPÓTESIS 13](#_Toc134133597)

[VII. METODOLOGÍA 15](#_Toc134133598)

[Scraping**:** 15](#_Toc134133599)

[Modelo**:** 16](#_Toc134133600)

[VIII RESULTADOS 17](#_Toc134133601)

[Scraping**:** 17](#_Toc134133602)

[Modelado: 24](#_Toc134133603)

[REFERENCIAS 38](#_Toc134133604)

LISTA DE TABLAS

[TABLA I RESULTADOS DE LA EXPLORACIÓN DE SITIOS WEB 17](#_Toc134133611)

[TABLA II BASE RESULTANTE POR WEB SCRAPING 23](#_Toc134133612)

[TABLA III MEDIDAS DE DISPERSIÓN Y TENDENCIA CENTRAL DE LA VARIABLE *PRECIO* 25](#_Toc134133613)

[TABLA IV CALCULO DE OUTLIERS VARIABLE *PRECIO* 25](#_Toc134133614)

[TABLA V 26](#_Toc134133615)

[TABLA VI CORRELACIONES ENTRE VARIABLES PREDICTORAS 28](#_Toc134133616)

[TABLA VII CORRELACIONES ENTRE VARIABLES PREDICTORA 33](#_Toc134133617)

[TABLA VIII MODELO ITERACION 2 34](#_Toc134133618)

LISTA DE FIGURAS

[Fig. 1. DISTRIBUCIÓN DE INMUEBLES POR ZONAS 21](#_Toc134133636)

[Fig. 2. DISTRIBUCION DE LA VARIABLE *PRECIO* 24](#_Toc134133637)

[Fig. 3. DISTRIBUCIÓN DE VARIABLE PRECIO SIN OUTLIERS 26](#_Toc134133638)

[Fig. 4. DISTRIBUCIÓN VARIABLES NUMÉRICAS 27](#_Toc134133639)

[Fig. 5. DISTRIBUCIÓN VARIABLES CATEGÓRICAS NUMÉRICAS 28](#_Toc134133640)

[Fig. 6 DISTRIBUCION DE VARIABLES CATEGÓRICAS NUÉRICAS SIN OUTLIERS 29](#_Toc134133641)

[Fig. 7 DISTRIBUCIÓN VARIABLES CATEGÓRICAS 30](#_Toc134133642)

[Fig. 8 CORRELACIÓN VARIABLE OBJETIVO VS VARIABLES PREDICTORAS 31](#_Toc134133643)

[Fig. 9 CORRELACIÓN VARIABLES NUMÉRICAS 32](#_Toc134133644)

[Fig. 10 CORRELACIÓN VARIABLES CATEGÓRICAS NUMÉRICAS 32](#_Toc134133645)

[Fig. 11 CORRELACIÓN VARIABLES CATEGÓRICAS 32](#_Toc134133646)

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 N iteraciones siendo la primera la línea base de la solución y la iteración N la implementación de una serie de tratamiento de datos, algoritmos y parámetros para acercar la predicción a los resultados esperados.

***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 N iterations, the first being the baseline of the solution and the iteration N the implementation of a series of data processing, algorithms and parameters to bring the prediction closer to the expected 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.

# V. HIPÓTESIS

*A. Hipótesis de trabajo*

Es posible crear un sistema de predicción de precios de arriendo con un comportamiento similar al mercado y que en base a unas características cualquier usuario puede obtener un valor estimado y así evitar invertir tiempo en asesorías y búsquedas presenciales

*B. Hipótesis estadística*

Se desea conocer si la ciudad debe ser dividida por zonas y aplicar el modelo solo donde el MAPE sea igual o inferior al 15%, por tanto:

1. *Hipótesis nula: El modelo de predicción de precios de arriendo es generalizable para todas las zonas de la ciudad*
2. *Hipótesis alternativa: El modelo de predicción de precios de arriendo NO es generalizable para todas las zonas de la ciudad.*

VI. MARCO TEÓRICO

* [1] **Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring:** Trabajo previo cuyo objetivo es 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.
* [2] **Estimating Warehouse Rental Price using Machine Learning Techniques:** Trabajo previo cuyo objetivo es explicar 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.

# 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 topologías 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**.**

# 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 complejicen la extracción de información, además cuenta 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:**

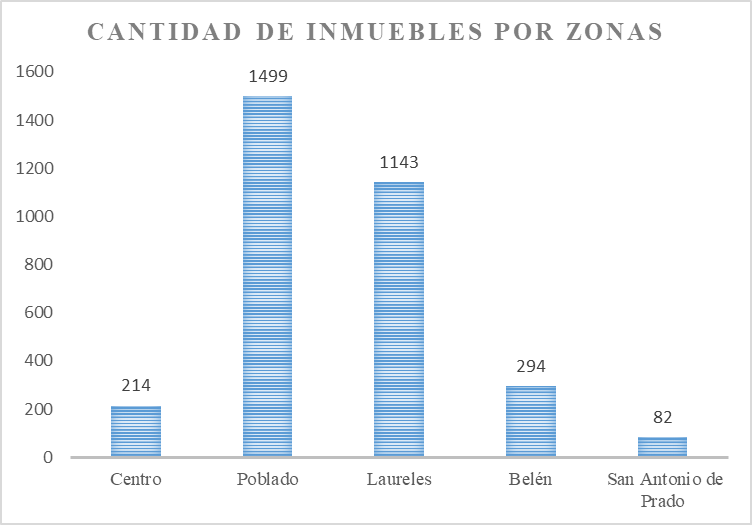
|  |
| --- |
| * + - Laureles |
| * + - La Castellana |
| * + - Robledo |
| * + - Simón Bolívar |
| * + - El Nogal |
| * + - Calasanz |
| * + - Pilarica |
| * + - Estadio |
| * + - San Javier |
| * + - San German |
| * + - La América |
| * + - Los Colores |
| * + - López de Mesa |
| * + - Córdoba |
| * + - 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 generar un funcionamiento de la predicción del modelo con mejores métricas para estas zonas.

* 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 características:
    - 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 de la topología del 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 características 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. Existen variables categóricas no booleanas que contienen más de 10 categorías, estas se formatean agrupando los datos en categorías más generales para eliminar ambigüedades en su contenido, ejemplo de estas: tipo de pisos y tipo de cocina.
7. 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 |

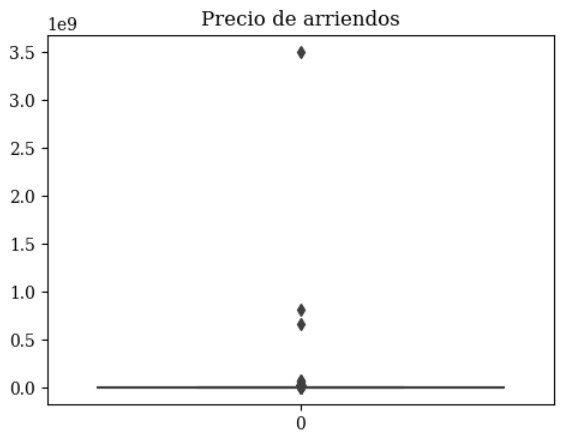
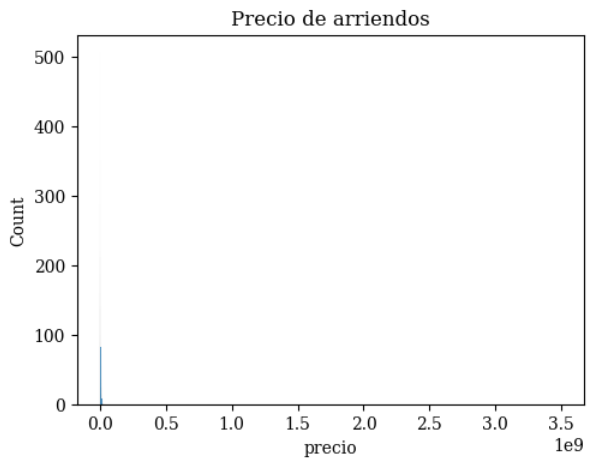
## Modelado:

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

Iteración 1:

Al completar la base insumo por parte del proceso de scraping se procede a crear un modelo línea base para obtener un primer vistazo del comportamiento de los datos 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*

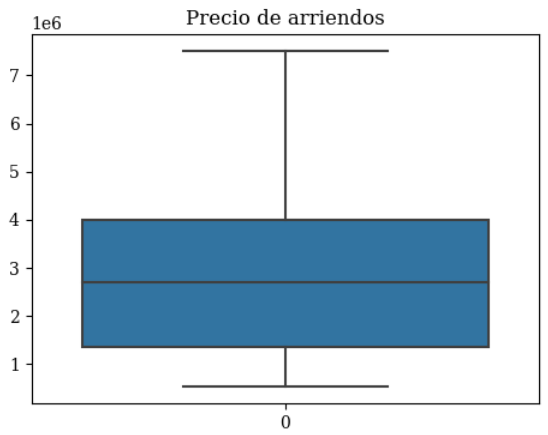
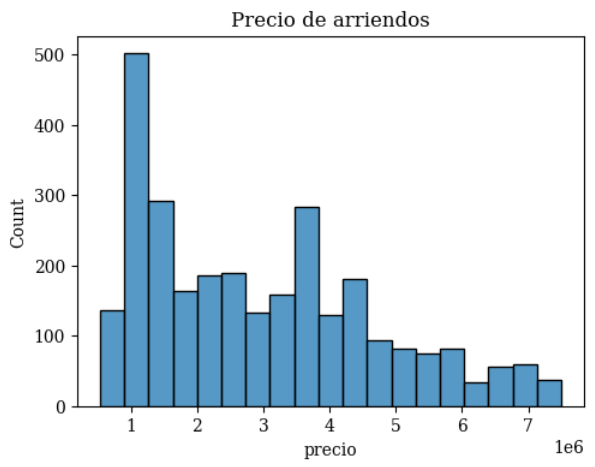
|  |  |
| --- | --- |
| Medida | Valor |
| Media | 447391999.773014 |
| Desviación estándar | 1233632087.282006 |
| Mínimo | 0.000000 |
| 25% | 1125805.250000 |
| 50%  75%  Máximo | 3900000.000000  20162403.136915  3500000000.000000 |
|  |  |

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, 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* y una codificación por frecuencias de las variables *tipo\_pisos* y *tipo\_cocina*, pues al estas tener más de 10 categorías, incluso luego de la generalización de sus valores, resulta más conveniente hacer la transformación a las variables numéricas *tipo\_piso\_freq* y *tipo\_cocina\_freq*, cambiando así dos variables categóricas por dos numéricas.

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 | 0.730  1085319.29  0.273 |
|  |  |

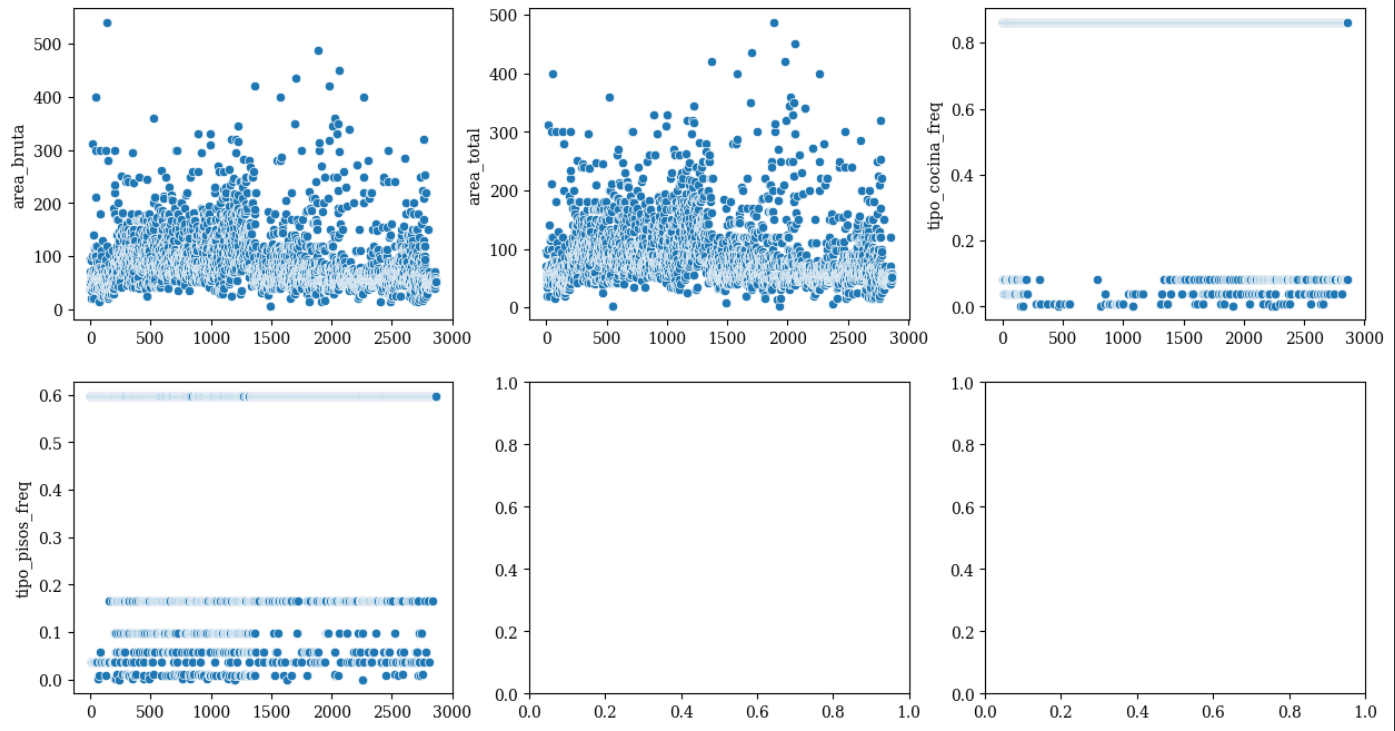
Dada la TABLA V se obtiene un modelo con métricas aceptables, con un MAPE 0.12 unidades por encima del deseado y un RMSE que nos indica un error de predicción de aproximadamente 1’085,319 COP lo cual para un usuario es un margen de error muy amplio y que se buscará reducir en las próximas iteraciones.

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:

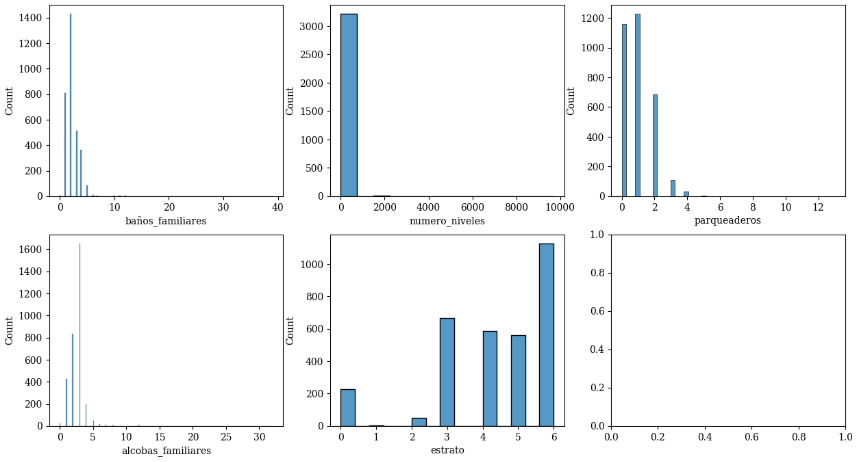
* Eliminación de outliers de la variable objetivo y codificación de variables categóricas: se preserva el tratamiento de datos realizado en la iteración 1.
* Observación de distribución de los datos: Se grafica la distribución de los datos de cada tipo de variables obtenidas

Fig. 4. DISTRIBUCIÓN VARIABLES NUMÉRICAS



En la figura 4 se aprecia el comportamiento de las variables numéricas, estas no presentan novedades en sus distribuciones, por tanto, no se realiza ninguna intervención.

Fig. 5. DISTRIBUCIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



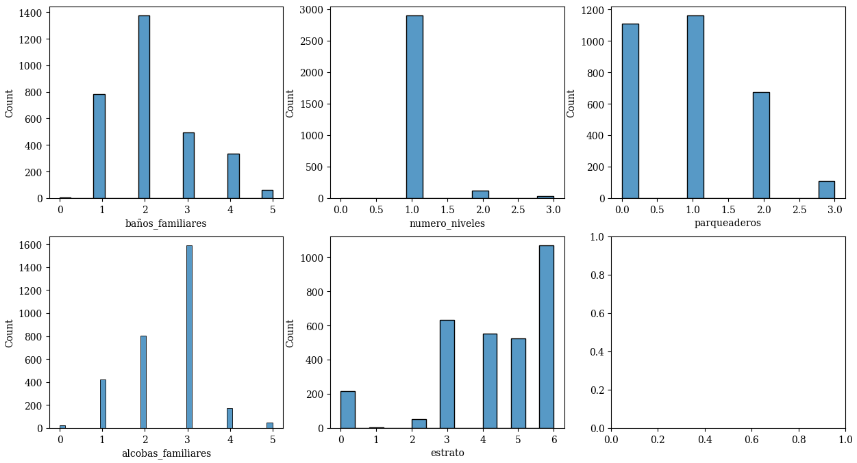
En la figura 5 se presentan las variables numéricas, pero con valores discretos, las cuales se les da un tratamiento de variables categóricas. Se puede observar la presencia de valores atípicos en las variables de *baños\_familiares, número\_niveles\_parqueaderos y alcobas\_familiares*, pues valores 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 | <= 3 |
| baños\_familiares | <= 5 |
| parqueaderos | <= 3 |
| alcobas\_familiares | <= 5 |

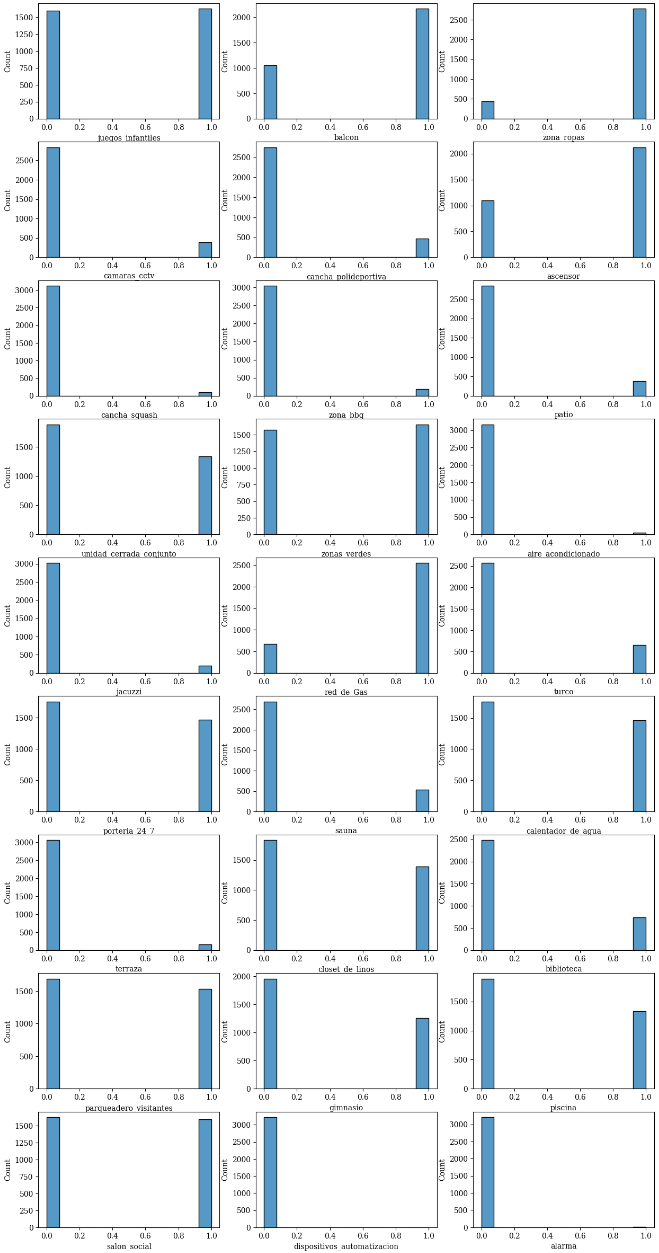
Con el tratamiento realizado en la TABLA IV se obtiene la distribución de la Fig 5. Con lo cual los valores para estas variables tienen una representación mas lógica.

Fig. 6 DISTRIBUCION DE VARIABLES CATEGÓRICAS NUÉRICAS SIN OUTLIERS



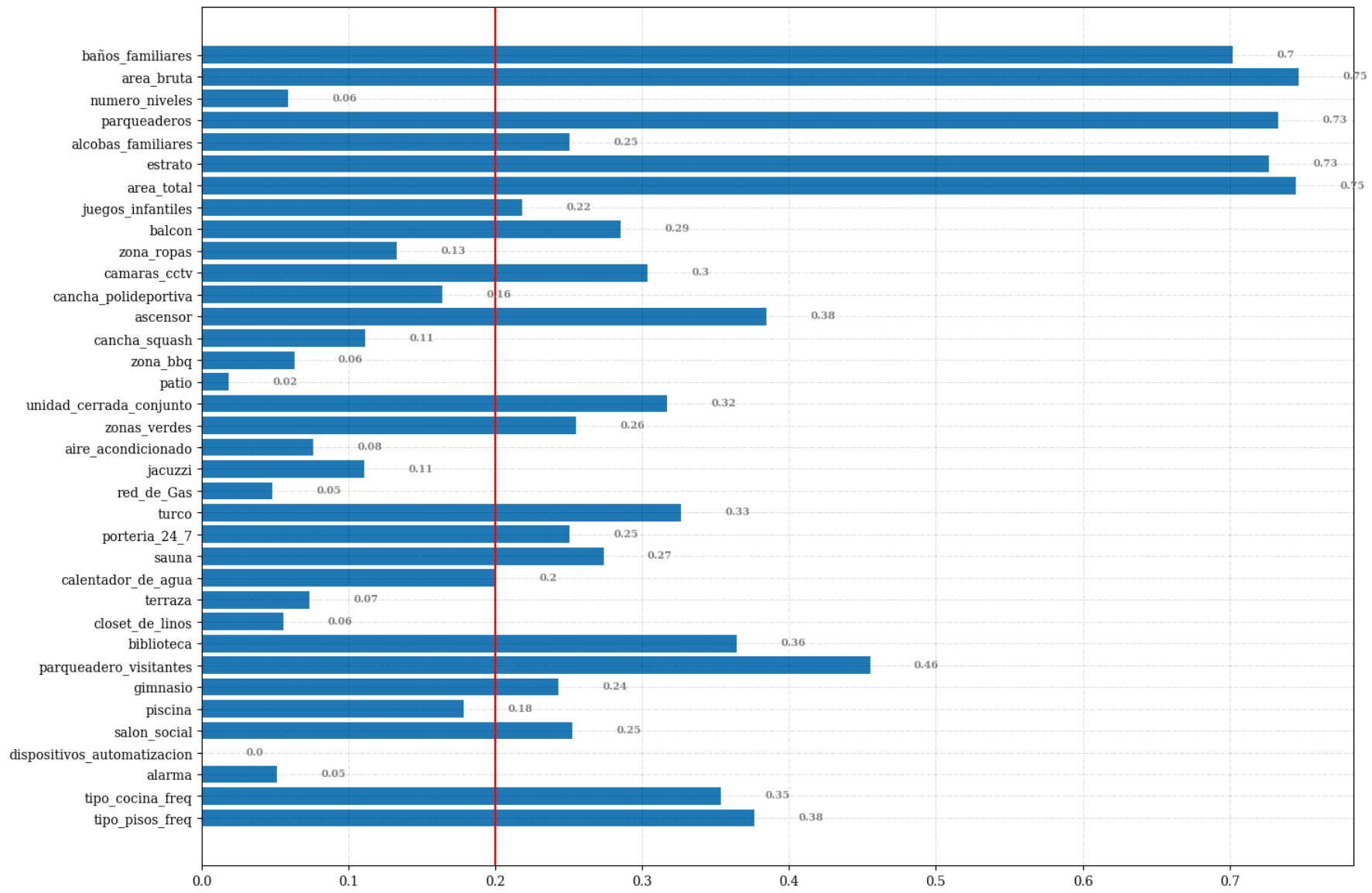
En la Fig. 7 Se observa las distribuciones de las variables categóricas, al estas representar las comodidades con las que cuenta una vivienda es de resaltas que hay algunas de ellas que son tan comunes hoy en día que es casi ilógico que una casa o apartamento no la posea, por ejemplo variables como *zona\_de\_ropas* 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 son candidatas para ser eliminadas pues este comportamiento indica una posible muy poca influencia en la predicción del modelo.

Fig. 7 DISTRIBUCIÓN VARIABLES CATEGÓRICAS



* Correlación: Para el cálculo de la correlación se realizan varias comparaciones, variables predictoras vs variable objetivo y en cada uno de los tipos de variables que se ha considerado

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 cuales 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 tal como indico su distribución, por tanto, serán eliminadas en el proceso de modelado.

Fig. 9 CORRELACIÓN VARIABLES NUMÉRICAS

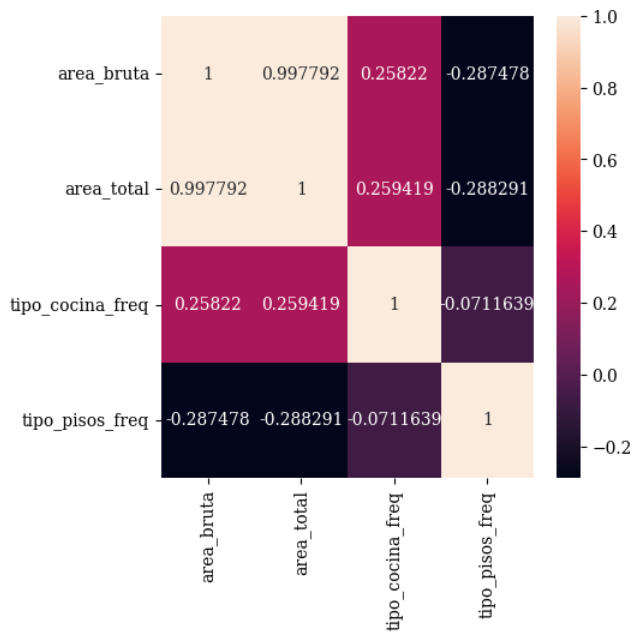
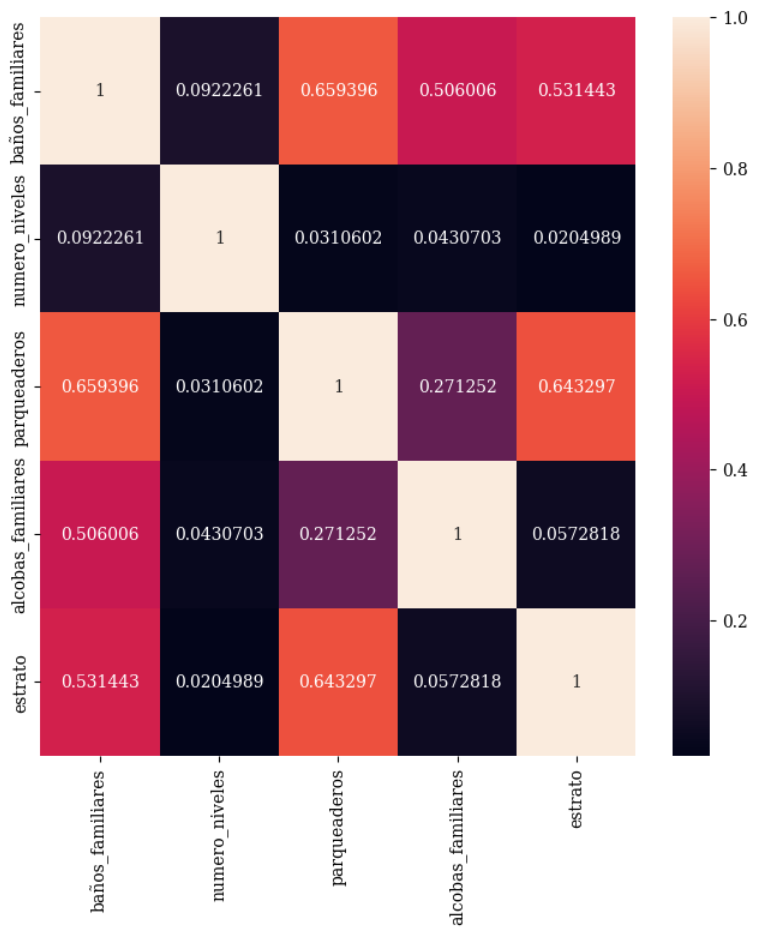
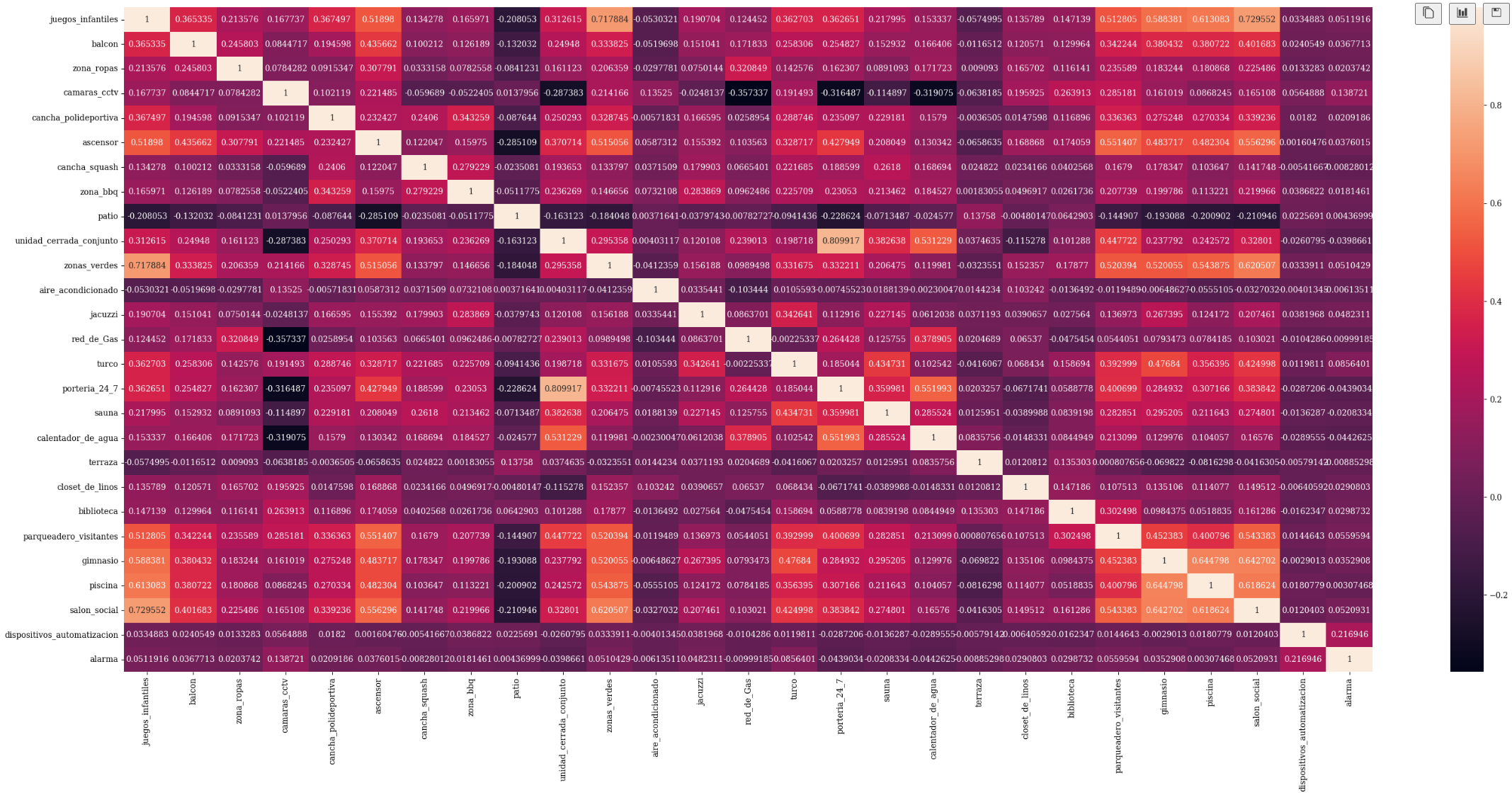


Fig. 10 CORRELACIÓN VARIABLES CATEGÓRICAS NUMÉRICAS



Definiendo 0.7 como un alto valor de correlación para marcar dos variables como igualmente influyentes sobre la variable objetivo. 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.

Fig. 11 CORRELACIÓN VARIABLES CATEGÓRICAS



En la Fig 10 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 estas la eliminación de una de estas variables esta atada a la siguiente premisa “¿Si una vivienda tiene la característica x necesariamente tiene la característica y?”. Siguiendo esto se concluye que la variable a eliminar es *porteria\_24\_7.*

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 | 0.77  0.77 |
| area\_bruta | area\_total | 0.99 |
| juegos\_infantiles | zonas\_verdes | 0.71 |
| 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*
  + *cancha\_squash*
  + *aire\_acondicionado*
  + *jacuzzi*
  + *terraza*
  + *dispositivos\_automatizacion*
  + *alarma*
  + *red\_de\_Gas*
  + *zona\_ropas*
  + *cancha\_polideportiva*
  + *zona\_bbq*
  + *patio*
  + *'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 | 0.73  907309.29  0.275 |
|  |  |

Se obtiene una mejora en la métrica RMSE con un margen de error aproximado de 907,310 COP, y un MAPE aún 0.12 unidades por encima del deseado. Dado estos resultados se considera como insuficiente la regresión lineal para obtener las métricas buscadas con lo cual la siguiente iteración se enfoca en iterar sobre diferentes tipos de algoritmos

IX. DISCUSIÓN

Discusion interna de los resultados

A que aporta el proceso realizado

Analizar el proceso evolutivo de la solución

X. CONCLUSIONES

XI. RECOMENDACIONES

# 

# 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
* Añadir referencia del encoder por frecuencia