Universidades de Burgos, León y Valladolid

Máster universitario en Inteligencia de Negocio y Big Data en Entornos Seguros

Trabajo Final de Máster:

Predicción del precio de la vivienda mediante aprendizaje automático

Presentado por Bungisa Beto Bibeyi
en agosto 2024
Tutor Juan José Rodríguez Díez

Índice de contenido

Índice de ilustraciones

Índice de tablas

Resumen

El propósito de este trabajo de fin de máster es analizar los datos del mercado inmobiliario de la Comunidad de Madrid mediante técnicas de aprendizaje automático. Para ello, se ha llevado a cabo la recopilación y el análisis exhaustivo de los datos con el objetivo de identificar las características comunes que determinan el valor de un inmueble.

Los datos analizados han sido utilizados para entrenar diversos modelos de aprendizaje automático. Posteriormente, se ha realizado una evaluación de estos modelos para seleccionar el que mejor se adapta al problema planteado. Finalmente, el modelo seleccionado ha sido optimizado para ofrecer el mejor rendimiento posible.

Este proyecto tiene como finalidad automatizar tareas repetitivas en el sector inmobiliario y reducir el tiempo necesario para la tasación del precio de las viviendas.

Abstract

The purpose of this master's thesis is to analyze real estate market data from the Community of Madrid using machine learning techniques. To achieve this, a thorough collection and analysis of the data has been carried out with the aim of identifying common characteristics that determine the value of a property.

The analyzed data have been used to train various machine learning models. Subsequently, an evaluation of these models has been conducted to select the one that best fits the problem at hand. Finally, the selected model has been optimized to offer the best possible performance.

This project aims to automate repetitive tasks in the real estate sector and reduce the time required for property valuation.

I - *INTRODUCCIÓN*

El mercado inmobiliario es uno de los sectores más importantes del mundo. Este sector contribuye profundamente a la economía de un país. Desde una perspectiva económica, la intensidad del mercado promueve la producción general que beneficia al sector y las actividades que están relacionadas como la construcción y los servicios financieros. Esta industria es una base crucial de empleo y constituye un servicio de primera necesidad.

Sin embargo, el sector financiero tiene una serie de complejidades. Esta complejidad no solo involucra la compra y venta de propiedades, sino también en temas legales, económicos y sociales, etc. Este dinámico sector no solo responde a los cambios de la economía sino también tiene una cierta influencia en ella. Su compresión es muy importante para predecir y adaptarse a los cambios en el mercado.

Para abordar esta complejidad surge la necesidad de utilizar herramientas que faciliten el análisis de los datos. A través de modelos de aprendizaje automático, pueden ayudar a anticipar valores futuros de las viviendas, tomando en cuenta el histórico y las características principales. Con ello, se pueden identificar oportunidades de inversión o riesgos potenciales. Otro aspecto que puede beneficiar la integración de esta herramienta es la automatización de tareas repetidas.

En este trabajo de fin de máster, se analizará el precio de la vivienda en la Comunidad de Madrid utilizando un modelo de aprendizaje automático. El objetivo principal es descubrir patrones en los datos que permitan una estimación precisa de los precios de las viviendas.

Para ello, ha sido necesario recopilar el conjunto de datos sobre el mercado inmobiliario de la Comunidad de Madrid a través de Idealista. Estos datos han sido analizados y preprocesados para ser entrenados utilizando técnicas de Regresión Lineal, Bosque Aleatorio y Refuerzo de Gradiente Extremo.

Finalmente, se evaluarán los modelos para seleccionar el que ofrezca el mejor rendimiento en la predicción de precios. El modelo elegido será optimizado para garantizar la máxima precisión en las estimaciones de precios de viviendas.

II - OBJETIVOS DEL PROYECTO

1. Objetivo general

El objetivo del proyecto es desarrollar un modelo de predicción del precio de la vivienda a partir de un conjunto de datos de inmuebles en venta de la Comunidad de Madrid.

2. Objetivos específicos

- Analizar la importancia del mercado inmobiliario.
- Reunir fuentes de datos necesarias para el desarrollo del modelo.
- Determinar las diferentes características de la vivienda que influyen en el precio.
- Crear diferentes modelos de aprendizaje automático que permitan predecir el precio de la vivienda.
- Elegir el modelo que ofrece mejor rendimiento respecto del problema planteado.

III - Conceptos teóricos

1. El mercado inmobiliario

El mercado inmobiliario se centra en la compra, venta, alquiler y gestión de propiedades. Afecta directamente al Producto Interno Bruto (PIB) de los países, no solo en inversión privada, sino también en el ámbito público, siendo uno de los puntos clave en los programas políticos. En España, el mercado inmobiliario es dinámico e importante debido a su alto turismo. Abarca diversas áreas económicas, como el empleo, creando puestos de trabajo en construcción y gestión de propiedades MOM01.

El mercado inmobiliario puede dividirse en residencial, comercial e industrial. Cada sector cubre necesidades específicas y afecta la evolución del mercado de manera distinta.

En España, el mercado inmobiliario evolucionó desde el auge de los 2000, impulsado por bajas tasas de interés y crédito accesible, hasta la crisis de 2008. La recuperación se dio gracias a políticas financieras e inversión extranjera, estabilizando el sector desde 2014. En 2021 y 2022, hubo un aumento en las ventas de viviendas, pero en 2023 hubo una caída del 9,7% por el aumento de tipos de interés y la desaceleración económica, a pesar de un aumento del 3,5% en los precios. Las previsiones para 2024 indican un aumento del 2,5% en precios y una reducción del 5% en ventas, según informes de BBVA Research, PlanRadar y Bankinter. Sin embargo, la firma S&P prevé un aumento de los precios del 4% en 2024, 3% en 2025, 2,4% en 2026 y 2% en 2027 GA01.

El mercado inmobiliario en España puede verse afectado por factores demográficos, económicos, legislativos y tecnológicos. El crecimiento de la población en grandes ciudades, tasas de interés, inflación, crecimiento del PIB, políticas urbanísticas y avances en Inteligencia Artificial (IA) influyen en el sector.

La IA está transformando el mercado inmobiliario, optimizando la gestión de propiedades y el análisis de datos. Facilita decisiones fundamentadas mediante la revisión de grandes volúmenes de datos. El análisis predictivo utiliza datos históricos para prever la oferta y demanda a corto y medio plazo. También mejora la tasación y valoración de inmuebles sin necesidad de tasadores, mediante el uso de planos y características. Además, influye en marketing y creación de marca personal CM01.

2. Inteligencia artificial

Hoy en día, la inteligencia artificial (IA) es una de las palabras más escuchadas debido a su impacto en el sector empresarial y en la vida social. En el ámbito empresarial, la IA se utiliza en áreas como los servicios financieros, la sanidad y el transporte, incorporando herramientas como la automatización de tareas, el análisis de datos y la toma de decisiones. En el ámbito social, aplicaciones como los asistentes de voz y tecnologías relacionadas con la salud están provocando cambios positivos en la vida cotidiana.

La influencia de la IA en el mundo es considerable. Pero, ¿qué es la inteligencia artificial? una definición posible es "*la habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana*" LR01.

2.1. Aprendizaje automático

Dentro de la inteligencia artificial existen diversas ramas. Este proyecto se centrará en el aprendizaje automático. Esta rama se dedica al desarrollo de algoritmos y modelos matemáticos con el objetivo de aprender a partir de un conjunto de datos para realizar tareas específicas. Las técnicas de aprendizaje automático permiten que un robot o una máquina aprendan de manera independiente para realizar tareas y extraer información de los datos. Dentro del aprendizaje automático, se destacan tres modelos principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

2.1.A. Aprendizaje supervisado

El aprendizaje supervisado es una técnica de aprendizaje automático que utiliza un conjunto de datos etiquetados para que el modelo aprenda la relación entre las características de los datos. Tras el entrenamiento, el modelo recibe datos no etiquetados para determinar el valor correcto IAAR.

Esta técnica se emplea en problemas de clasificación y regresión. En cuanto a los algoritmos, se describen los siguientes en este trabajo:

- **Árbol de decisión**: es un algoritmo que se estructura como un árbol, con una raíz, nodos internos y hojas. El punto de partida es la raíz, que contiene la primera condición. Este nodo raíz genera la primera partición binaria, representada por dos flechas que indican si se cumple o no la condición. Los nodos internos tienen condiciones adicionales para seguir dividiendo, mientras que las hojas representan las regiones donde no se realizarán más particiones. La profundidad del árbol es la trayectoria más larga desde la raíz hasta una de las hojas FM01.
- **Regresión lineal**: es un algoritmo que estudia estadísticamente la relación entre una variable dependiente y un conjunto de variables independientes. Su objetivo es encontrar la línea recta que mejor se ajuste al conjunto de datos AWS01.
- **Bosque aleatorio**: es un algoritmo que se compone de un conjunto de Árbol de decisión.Este entrena varios árboles utilizando subconjuntos aleatorios de datos y combina los resultados del entrenamiento para identificar el resultado más frecuente IBM01.
- **Extreme gradient boosting (Refuerzo gradiente extremo)**: este algoritmo genera varios modelos de predicción débil. Cada modelo se basa en los resultados del modelo anterior para crear un modelo más fuerte. Se utiliza un algoritmo de optimización para elegir el modelo más robusto. Durante el entrenamiento, se ajustan iterativamente los parámetros del modelo débil para encontrar el mínimo de una función objetivo. Esta técnica es una variante de XGBoost, que está diseñada para tareas de regresión FM01.

A.1. Técnicas de entrenamiento

Para garantizar que los modelos se adapten a nuevos datos, es necesario evaluar su rendimiento. Una de las técnicas más utilizadas es la validación cruzada. Esta técnica consiste en entrenar el modelo con diferentes subconjuntos de datos AWS02. Para ello, se realizan las siguientes acciones:

1. **División del conjunto de datos**: se divide el conjunto de datos en K subconjuntos del mismo tamaño.

2. **Entrenamiento y evaluación**: el modelo se entrena K veces. En cada iteración, se seleccionan todos los subconjuntos menos uno, que se utiliza para la evaluación.
3. **Resultados**: después de todas las iteraciones, se promedian las mediciones de desempeño obtenidas.

A.2. Evaluación del modelo

La evaluación del modelo de aprendizaje supervisado es esencial para conocer el rendimiento de los modelos entrenados. Existen diferentes métricas de evaluación, pero se mencionarán medidas de regresión porque el problema considerado es de predicción numérica. Estas métricas son las siguientes:

- **Error absoluto medio (MAE)**: permite conocer la diferencia promedio entre el valor real y el valor predicho.
- **Raíz del error cuadrático medio (RMSE)**: mide la diferencia cuadrática promedio entre el valor predicho y el valor real.
- **R cuadrada (R2)**: es el coeficiente de determinación y mide la capacidad del modelo para predecir valores futuros.

A.3. Ajuste de hiperparámetros

En la fase de entrenamiento del modelo de aprendizaje automático, es necesario identificar los hiperparámetros que se ajustan mejor al modelo y al conjunto de datos. Esta técnica consiste en definir diferentes conjuntos de hiperparámetros para que sean ejecutados en el modelo con la finalidad de determinar cuáles hiperparámetros se ajustan mejor a los datos. Estos hiperparámetros son parámetros ajustables que se definen manualmente para controlar el proceso de aprendizaje de un modelo AWS03.

2.1.B. Aprendizaje no supervisado

La técnica de aprendizaje no supervisado consiste en aprender de datos sin etiquetar para encontrar patrones o agrupaciones ocultos sin necesidad de ninguna intervención humana. La idea del algoritmo es dividir los datos en grupos, de manera que los valores dentro de cada grupo sean lo más similares posible entre sí y se distinguen claramente de los valores de otros grupos AWS04.

2.1.C. Aprendizaje por refuerzo

El aprendizaje por refuerzo es una técnica que consiste en entrenar al modelo con un conjunto de datos sin etiquetas, utilizando un enfoque de ensayo y error. Estos modelos emplean un sistema de recompensas y castigos para gestionar la información. El modelo aprende de cada acción que realiza y determina los mejores caminos para alcanzar los objetivos establecidos IA01.

2.1.D. Análisis exploratorio de datos

Uno de los pasos fundamentales en cualquier proyecto de aprendizaje automático es el análisis de los datos. Esta técnica permite no solo comenzar a entender los datos que se están procesando, sino también tener un esquema mental de ellos. Este esquema será clave para las fases posteriores del proyecto.

El análisis exploratorio fue desarrollado por el matemático John Tukey en la década de 1970.

Tukey propuso una nueva disciplina científica que incluía los métodos y técnicas estadísticas como un componente más del análisis de datos. El rápido desarrollo de las tecnologías ha impulsado esta disciplina debido a la capacidad de acceder a una mayor cantidad de datos y la utilización de análisis cuantitativo en diferentes disciplinas BBG01.

Dentro del análisis exploratorio se pueden distinguir diferentes tipos:

- **Univariante no gráfica**: consiste en analizar los datos de una sola variable con el objetivo de describir y encontrar patrones entre ellos.
- **Gráfico univariante:** los diagramas de tallo y hojas, histogramas y diagramas de caja ayudan a observar los datos desde diferentes perspectivas.
- **Multivariante no gráfica**: esta técnica descubre la relación entre dos o más variables.
- **Gráfico multivariante**: se emplean visualizaciones para mostrar relaciones entre dos variables. Los gráficos más utilizados son el gráfico de barras, diagrama de dispersión, mapa de calor, entre otros.

Existen diversas metodologías para llevar a cabo un análisis exploratorio, cada una con sus propias técnicas y enfoques. Sin embargo, en este trabajo de fin de máster se detallarán específicamente los siguientes pasos para el análisis exploratorio:

- **Recopilación de datos**: consiste en la obtención de los datos que se quieren analizar.
- **Exploración inicial:** se trata de entender el conjunto de datos. Esta parte incluye la revisión y la limpieza de los datos.
- **Visualización de datos**: se utilizan diferentes herramientas para mostrar la distribución y explorar la relación entre las variables.
- **Estadísticas descriptivas**: en esta parte se describe las características de los datos mediante estadísticas descriptivas.
- **Identificación de valores atípicos**: se trata de detectar y analizar los valores extremos del conjunto de datos.
- **Iteración y refinamiento**: es un proceso iterativo que permite que, cada vez que se encuentran nuevas tendencias en los datos, exista la posibilidad de realizar ajustes o nuevas hipótesis.

IV - Técnicas y herramientas

Este apartado detalla los procedimientos utilizados para desarrollar un modelo de predicción del precio de las viviendas. Inicialmente, se recopilaron datos mediante la API de Idealista y se procesaron para el análisis exploratorio y la creación del modelo. Para realizar estos pasos, se emplearon herramientas como Jupyter Notebook y VSCode.

1. Metodología usada

Para el desarrollo del modelo de predicción del precio de la vivienda, se ha basado en la metodología KDD (Knowledge Discovery in Databases). Esta metodología consiste en detectar patrones dentro de grandes conjuntos de datos con el objetivo de convertirlos en conocimiento IBM03. La técnica se compone de las siguientes etapas:

- **Elección de la técnica**: esta etapa implica adquirir conocimiento sobre los problemas planteados. Para ello, se debe buscar información en libros, trabajos y artículos relacionados con el tema.
- **Recogida de datos**: consiste en buscar fuentes de datos que proporcionen la información necesaria para el desarrollo del proyecto. Puede ser necesario desarrollar o utilizar herramientas que faciliten la obtención del conjunto de datos.
- **Análisis exploratorio de datos**: esta etapa tiene como objetivo estudiar y comprender el conjunto de datos. Se aplican técnicas estadísticas para explorar, describir y resumir la naturaleza de los datos, buscando obtener un conjunto de datos de la máxima calidad posible.
- **Entrenamiento**: en esta parte se trata de entrena al modelo para que aprenda de los datos.
- **Evaluaciones**: consiste en evaluar el modelo entrenado para determinar su rendimiento.

2. Alternativas estudiadas

En este apartado se comentan las alternativas planteadas durante el desarrollo del proyecto que fueron descartadas debido a diversas circunstancias. Después de haber sido estudiados los modelos de Regresión lineal, Árbol de decisión, Bosque aleatorio y Extreme gradient boosting (XGBoost), así como los enfoques de procesamiento de datos basados en BERT, se decidió descartar el Árbol de decisión y BERT.

2.1. Árbol de decisión

Durante la elección del modelo, se planteó el uso de Árbol de decisión como un posible algoritmo para la predicción del precio de la vivienda. Sin embargo, se descartó debido a que, en comparación con otros modelos, presentaba diversas problemáticas. Entre estos inconvenientes se encontraban su capacidad predictiva inferior en comparación con otros modelos y su tendencia al sobreajuste, lo que lo hace menos adecuado si el conjunto de datos puede cambiar con el tiempo XI01.

2.2. BERT

Para la parte del procesamiento de texto, se planteó el uso de un transformador basado en redes neuronales que pudiera realizar esta tarea. Sin embargo, se descartó debido al tiempo de desarrollo del proyecto y a los recursos necesarios para el procesamiento IBM04.

2.3. Herramientas

En el desarrollo del modelo de predicción del precio de la vivienda, se utilizaron diversas herramientas para la recopilación, procesamiento y análisis de datos, así como para la gestión del código. A continuación se detallan las herramientas principales utilizadas:

- **Java:** Lenguaje de programación.
- **Visual Studio Code (VSCode):** Entorno de desarrollo que se utilizó para programar en Java para la obtención del conjunto de datos de la API de Idealista.
- **Maven:** Herramienta de gestión de proyectos y construcción para Java que se utilizó para la construcción de la aplicación.
- **API Idealista:** Medio a través del cual se obtuvo el conjunto de datos de Idealista.
- **Jupyter Notebook:** Aplicación web que se utilizó tanto en la fase de preprocesamiento y análisis de datos como en la creación y evaluación de modelos en Python, permitiendo la ejecución interactiva de código y la visualización en tiempo real.
- **Python:** Lenguaje de programación orientado a la ciencia de datos y al aprendizaje automático que se utilizó como lenguaje de programación en Jupyter Notebook.
- **GitHub:** Plataforma para el control de versiones que se utilizó para guardar el código y los datos en un repositorio remoto, facilitando el seguimiento de cambios en el proyecto.
- **Pandas (pd):** Librería para la manipulación y análisis de datos, que facilitó la carga, limpieza y transformación de datos de manera eficiente.
- **NumPy (np):** Librería fundamental para operaciones matemáticas y estadísticas avanzadas.
- **Seaborn (sns):** Librería para la creación de gráficos estadísticos detallados.
- **Matplotlib.pyplot (plt):** Librería para la generación de gráficos personalizados en 2D.
- **NLTK (Natural Language Toolkit):** Conjunto de herramientas para el procesamiento del lenguaje natural, incluyendo tokenización y lematización, útil para el análisis de textos descriptivos.
- **Folium:** Librería para la visualización de datos geoespaciales mediante la creación de mapas interactivos.
- **Scikit-Learn (sklearn):** Librería para el aprendizaje automático que proporciona herramientas para el preprocesamiento y modelado de datos.

V - ASPECTOS RELEVANTES DEL DESARROLLO DEL PROYECTO

En este apartado se describe el proceso de creación del modelo de predicción del precio de la vivienda, abarcando desde la recopilación del conjunto de datos hasta el análisis exploratorio de los mismos.

1. Elección de la técnica

Dado que el problema planteado consiste en predecir el precio de la vivienda, se seleccionó la técnica de aprendizaje supervisado. Esta técnica es adecuada para la predicción de valores continuos a partir de datos etiquetados. Está diseñada para enfrentar problemas de regresión, donde el objetivo es estimar un valor numérico basándose en una serie de características. En este caso, se cuenta con un conjunto de datos que incluye el precio de las viviendas y sus características (como el tamaño, la ubicación, el número de habitaciones, etc.). Esto permite entrenar modelos que pueden aprender cómo cada una de estas características afecta al precio de la vivienda.

La técnica de aprendizaje no supervisado es más adecuada para conjuntos de datos que no tienen etiquetas y se centra en identificar patrones y estructuras subyacentes en los datos. Su propósito principal es descubrir agrupaciones naturales, subgrupos y relaciones entre variables sin necesidad de una variable objetivo definida. Aunque el aprendizaje no supervisado puede ser útil en la fase de análisis exploratorio para obtener una visión general de los datos, en este caso, se decidió no emplear esta técnica.

2. Recogida de datos

Para obtener el conjunto de datos, fue necesario realizar una solicitud de acceso a la API de Idealista. Se accedió al apartado de contacto del sitio web de Idealista para explicar los motivos de la solicitud. Tras varios días, se recibió un correo electrónico de Idealista con la clave de acceso y el manual de instrucciones de su API.

La API de Idealista utiliza un sistema de autenticación por token, que consiste en información codificada que contiene la identidad del usuario. Para obtener el token, es necesario usar la API Key y el secret proporcionados por Idealista. La API Key es un código alfanumérico único y privado utilizado para autenticar y controlar el acceso a un servidor. Por lo tanto, se realiza una petición de autenticación por API Key al servidor de autenticación de Idealista, el cual responde con un token. Este token se utiliza para realizar peticiones a la API y acceder al conjunto de datos de la vivienda de Idealista.

En la primera petición a la API, se observó que el mensaje devuelto era un conjunto de datos en formato JSON. Con el usuario proporcionado por Idealista, solo se pueden realizar 100 peticiones al mes, con un máximo de 50 resultados por petición.

Con estas limitaciones, se procedió a automatizar el proceso de obtención del conjunto de datos mediante la creación de una aplicación en Java. En la aplicación se implementaron las funcionalidades necesarias para que, en cada solicitud a la API, la respuesta se guardara en un archivo en formato JSON como en la ilustración 1.

```
{
    "elementList": [
            {
            "propertyCode": "99999999",
            "thumbnail": "https://img4.idealista.com/",
            "externalReference": "99999999",
            "numPhotos": 42,
            "price": 1.04E7,
            "priceInfo": {
                "price": {
                    "amount": 1.04E7
                }
            },
            "propertyType": "chalet",
            "operation": "sale",
            "size": 1410.0,
            "rooms": 7,
            "bathrooms": 9,
            "address": "XXXXXXXX",
            "province": "Madrid",
            "municipality": "La Moraleja",
            "district": "La Moraleja urbanización",
            "country": "es",
            "latitude": XXXXXXXX,
            "longitude": XXXXXXXX,
            "showAddress": false,
            "url": "https://www.idealista.com/inmueble/XXXXXXXX/",
            "description": "Se trata de un chalet de más de 1.400 m2.
                            El chalet es sencillamente impresionante,
                            en él se combina la elegancia,
                            el estilo señorial y el buen gusto.",
            "hasVideo": true,
            "status": "good",
            "newDevelopment": false,
            "parkingSpace": {
                "hasParkingSpace": true,
                "isParkingSpaceIncludedInPrice": true
            },
            "priceByArea": 7376.0,
            "detailedType": {
                "typology": "chalet",
                "subTypology": "independantHouse"
            },
            "suggestedTexts": {
                "subtitle": "La Moraleja urbanización, La Moraleja",
                "title": "XXXXXXXX"
            },
            "hasPlan": true,
            "has3DTour": true,
            "has360": false,
            "hasStaging": true,
            "highlight": {
                "groupDescription": "Top"
            },
            "topNewDevelopment": false,
            "topPlus": false
        }
            ]

}
```


Ilustración 1: Formato de dato de la API de Idealista

2.1. Descripción de atributos

Tras ejecutar la aplicación Java, se obtuvo un total de 4,768 registros y 48 columnas de viviendas en venta en la Comunidad de Madrid. Para comprender el conjunto de datos, a continuación se realiza una descripción detallada de cada variable:

- **propertyCode**: identificador único de la vivienda en Idealista.
- **thumbnail**: URL de la imagen en miniatura.
- **externalReference**: referencia externa de la vivienda.
- **numPhotos**: número total de fotos disponibles para la vivienda.
- **price**: precio de la propiedad.
- **propertyType**: tipo de propiedad, como apartamento, casa, oficina.
- **operation**: tipo de operación, como venta o alquiler.
- **size**: tamaño de la vivienda en metros cuadrados.
- **rooms**: número de habitaciones en la vivienda.
- **bathrooms**: número de baños en la vivienda.
- **address**: dirección completa de la vivienda.
- **province**: provincia donde se encuentra la vivienda.
- **municipality**: municipio donde se encuentra la vivienda.
- **country**: país donde se encuentra la vivienda.
- **latitude**: coordenada de latitud de la vivienda.
- **longitude**: coordenada de longitud de la vivienda.
- **showAddress**: indica si la dirección completa está visible en el anuncio.
- **url**: URL del anuncio de la vivienda en Idealista.
- **description**: descripción textual de la vivienda.
- **hasVideo**: indica si hay un video disponible para la vivienda.
- **status**: estado de la vivienda, como disponible o reservado.
- **newDevelopment**: indica si la propiedad es una nueva construcción.
- **priceByArea**: precio por metro cuadrado de la vivienda.
- **hasPlan**: indica si hay planos disponibles para la vivienda.
- **has3DTour**: indica si hay un recorrido en 3D disponible.
- **has360**: indica si hay fotos panorámicas de 360 grados.
- **hasStaging**: indica si se han realizado decoración en la vivienda.
- **topNewDevelopment**: indica si la propiedad está destacada como uno de los principales nuevos desarrollos.
- **topPlus**: indica si la propiedad está destacada como una de las principales.
- **priceInfo**.price.amount: monto del precio de la vivienda.
- **parkingSpace.hasParkingSpace**: indica si la propiedad incluye espacio de estacionamiento.

- **parkingSpace.isParkingSpaceIncludedInPrice**: indica si el espacio de estacionamiento está incluido en el precio de la propiedad.
- **detailedType.typology**: tipología detallada de la vivienda, como apartamento o casa.
- **detailedType.subTypology**: subtipología de la vivienda, como dúplex o ático.
- **suggestedTexts.subtitle**: subtítulo sugerido para el anuncio.
- **suggestedTexts.title**: título sugerido para el anuncio.
- **highlight.groupDescription**: descripción del grupo de características destacadas.
- **district**: distrito en el que se encuentra la vivienda.
- **neighborhood**: barrio en el que se encuentra la vivienda.
- **floor**: número del piso en el que se encuentra la vivienda.
- **exterior**: indica si la propiedad es exterior (con vistas al exterior) o interior.
- **hasLift**: indica si el edificio cuenta con ascensor.
- **newDevelopmentFinished**: indica si el nuevo desarrollo está terminado.
- **parkingSpace.parkingSpacePrice**: precio del espacio de estacionamiento.
- **priceInfo.price.priceDropInfo.formerPrice**: precio anterior de la propiedad antes de cualquier reducción.
- **priceInfo.price.priceDropInfo.priceDropValue**: valor de la reducción del precio.
- **priceInfo.price.priceDropInfo.priceDropPercentage**: valor de la reducción en porcentaje.

3. *Análisis exploratorio de datos*

Dentro de un proyecto de aprendizaje automático, el análisis exploratorio de los datos es una parte crucial. En esta fase, se estudia y prepara el conjunto de datos para que pueda ser entrenado por un modelo de aprendizaje automático. El objetivo es localizar y solucionar problemas en los datos, así como comprender como se relacionan entre ellos.

3.1. Limpieza de los datos

Esta parte, consistio en realizar una mejora de la calidad de los datos extraidos. Para ello, se procedió a la carga del conjunto de datos con el uso de Jupyter notebook y realizar las diferentes etapas que implica la limpieza de datos.

3.1.A. Columnas irrelevantes

Durante la fase de recogida de datos, se identificaron varias columnas que no aportaban información relevantes al análisis del precio de la vivienda. Entre ellas se encuentran:

- **thumbnail**
- **externalReference**
- **numPhotos**
- **propertyCode**

- **showAddress**
- **hasVideo**
- **hasPlan**
- **url**
- **groupDescription**
- **has3DTour**
- **has360**
- **hasStaging**
- **topNewDevelopment**
- **newDevelopmentFinished**
- **topPlus**
- **newDevelopment**
- **amount**
- **operation**
- **province**
- **country**

3.1.B. Valores atípicos

Dentro del conjunto de datos, se observó que varias columnas presentaban valores que diferían significativamente del resto de los valores del mismo grupo. Estos valores atípicos fueron identificados en cada columna, pero se decidió conservarlos para su análisis posterior. Las columnas afectadas son:

- **price**
- **Size**
- **rooms**
- **bathrooms**
- **latitude**
- **longitude**
- **priceByArea**
- **parkinSpacePrice**
- **formerPrice**
- **priceDropValue**
- **priceDropPercentage**

3.1.C. Datos faltantes

En esta etapa se identificaron y resolvieron los registros ausentes en cada columna. Se creó una función que mostraba en porcentaje la cantidad de registros faltantes en cada columna, como se puede apreciar en la ilustración 2. Se procedió a eliminar las columnas con más del 70% de valores nulos. Las columnas afectadas fueron parkingSpacePrice, formerPrice, priceDropValue y recompense. Para la columna subTypology, se decidió eliminarla debido a su 54% de registros nulos, ya que la columna propertyType cubría esa información.

Para la columna floor, se sustituyeron los valores nulos por categorías basadas en la columna typology, clasificados en bajo, mediano, alto, elevado y muy alto. La columna HasLift, que indicaba la presencia de un ascensor, tenía varios registros con valores ausentes. Dado que esta columna influye en el precio de la vivienda, se decidió reemplazar los valores nulos por falso, indicando la ausencia de ascensor. Lo mismo se aplicó a las columnas hasParkingSpace e isParkingSpaceIncludeInPrice, transformando los valores booleanos a numéricos.

Para la columna exterior, se reemplazaron los valores nulos por falso para los registros con tipología piso y por verdadero para los registros con tipología countryhouse o chalet. Las columnas neighborhood y district se eliminaron, ya que la información estaba cubierta por las columnas de latitud, longitud y municipio.

Tras el proceso de limpieza, el tamaño del conjunto de datos se redujo a 4.517 registros y 20 columnas. Se mantuvieron las columnas hasParkingSpace e isParkingSpaceIncludeInPrice, que aún tienen valores nulos, para el análisis exploratorio de datos.

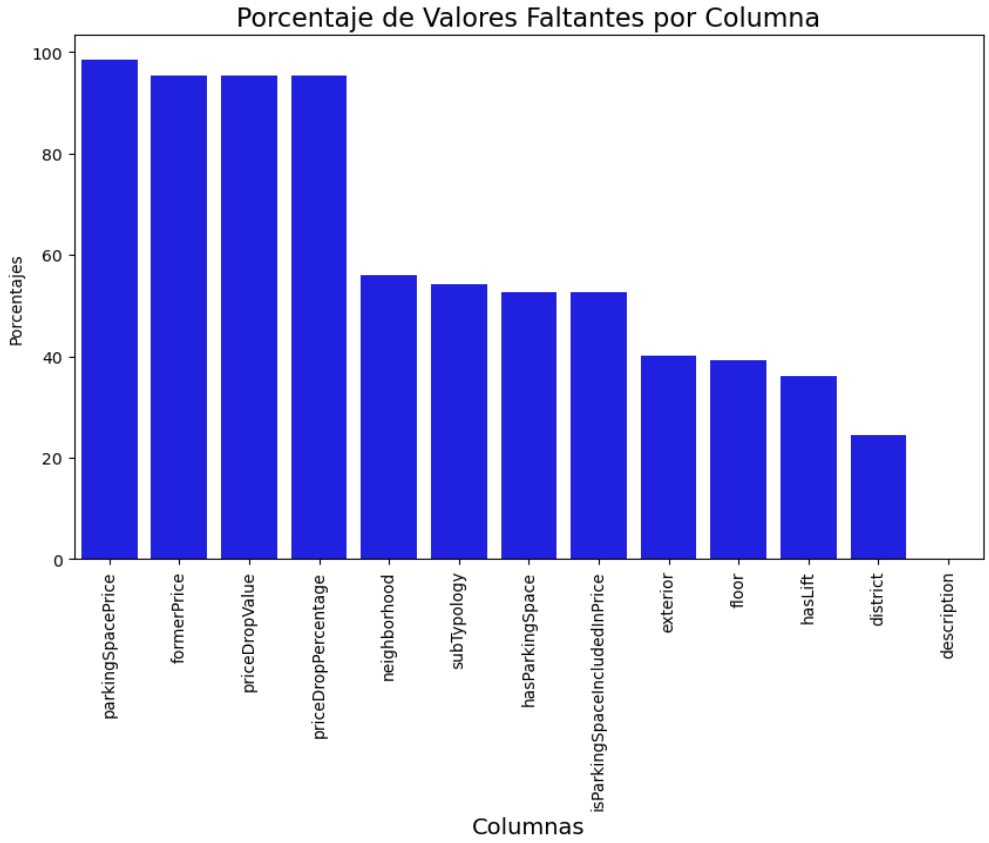

Ilustración 2: Porcentaje de valores faltantes por columna

3.1.D. Análisis univariable

El análisis univariable consistió en estudiar individualmente cada variable del conjunto de datos. Este análisis permitió comprender la distribución y las características de cada variable.

Al analizar la columna precio, se observó que el valor mínimo es de 19.800 euros y el valor máximo es de 1.500.000 euros, indicando una alta variabilidad en los precios de la vivienda. Los cuartiles muestran una distribución amplia de los precios, como se puede apreciar en la ilustración 3. La distribución del precio está sesgada a la derecha. El histograma de la ilustración 4 revela que, a medida que aumenta el precio, disminuye la cantidad de viviendas disponibles.

Respecto al valor máximo de la vivienda, se consideró la posibilidad de que fuera un valor atípico, lo que podría sesgar el conjunto de datos y afectar el rendimiento del modelo. Por lo tanto, fue necesario determinar si este valor era atípico para decidir su eliminación en el conjunto de datos.

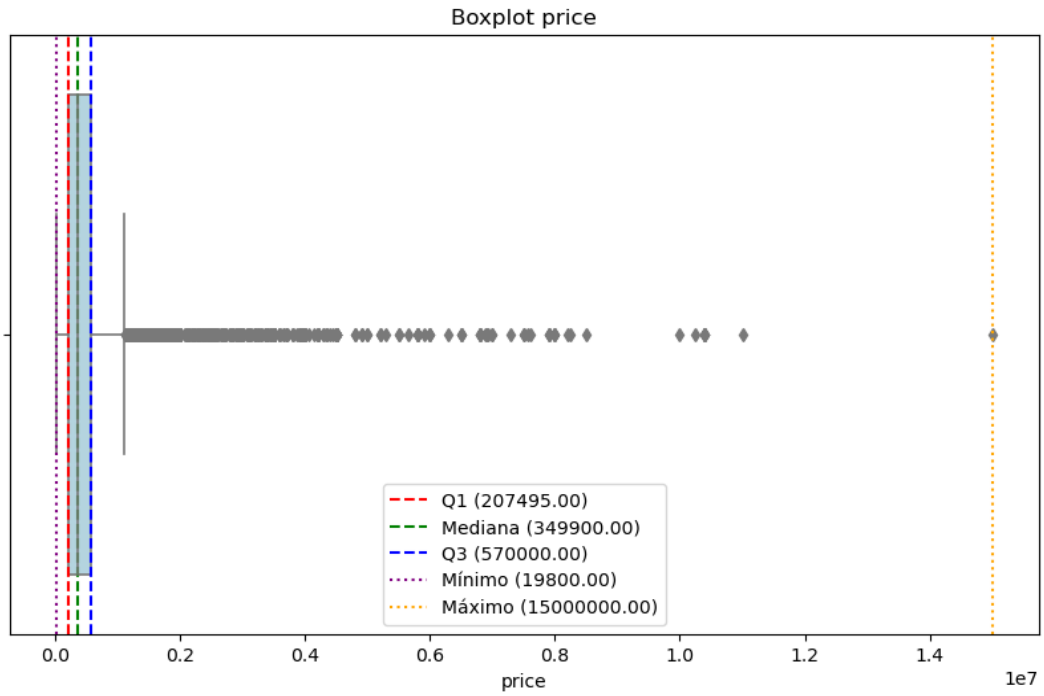

Ilustración 3: Diagrama de caja y bigote de la columna precio

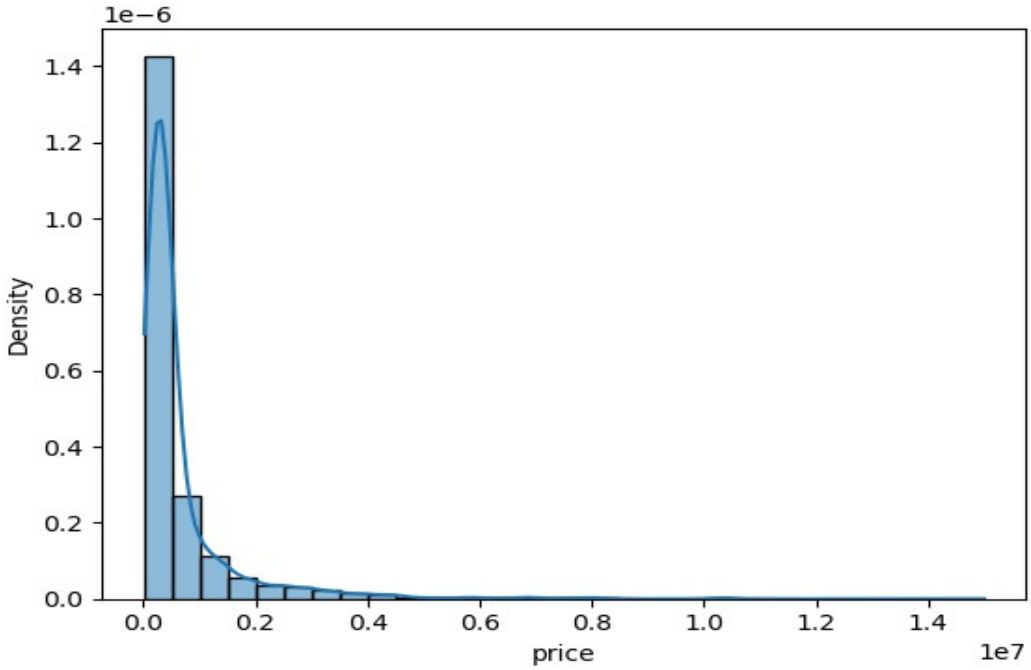

Ilustración 4: Histograma de la columna precio

La columna size presentaba un rango de tamaño de vivienda muy amplio, que abarcaba desde viviendas de dimensiones muy reducidas, con un mínimo de 14 metros cuadrados, hasta viviendas extremadamente grandes, con un tamaño máximo de 110.000 metros cuadrados. Esta amplia variabilidad en el tamaño puede observarse en la imagen . El tamaño máximo de 110.000 metros cuadrados es inusualmente alto y podría ser un valor atípico. Estos valores extremos pueden generar sesgos en el modelo, afectando su rendimiento y precisión.

La distribución del tamaño de las viviendas muestra que el 50 % de las viviendas tienen un tamaño inferior a 120 metros cuadrados, mientras que el 25 % de las viviendas tienen un tamaño entre 120 y 235 metros cuadrados. Esta distribución se puede apreciar en la ilustración 5. La alta concentración de viviendas con tamaños menores a 120 metros cuadrados sugiere que la mayoría de las propiedades en el conjunto de datos son relativamente pequeñas en comparación con el rango completo de tamaños.

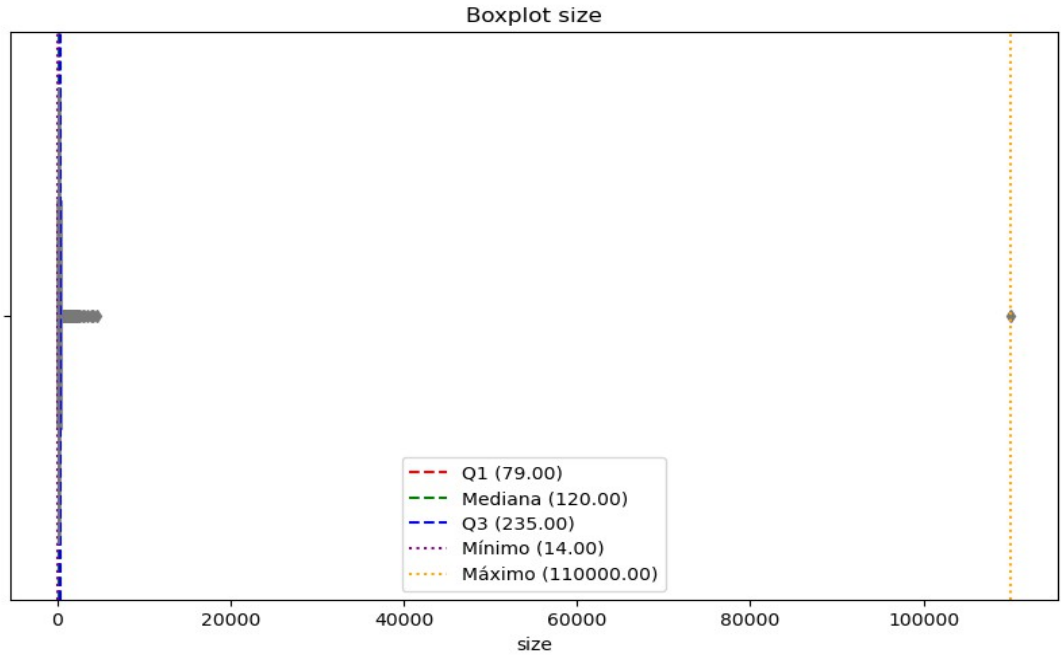

Ilustración 5: Diagrama de caja y bigotes de la columna size

En cuanto a la distribución de las habitaciones, el 50% de las viviendas tiene una cantidad de habitaciones inferior a 3, y solo el 25% de las viviendas tiene entre 3 y 4 habitaciones. La distribución de los valores en la columna de habitaciones está sesgada a la derecha, como se puede observar en el histograma de la ilustración 6, donde a medida que aumenta el número de habitaciones, disminuye la cantidad de viviendas.

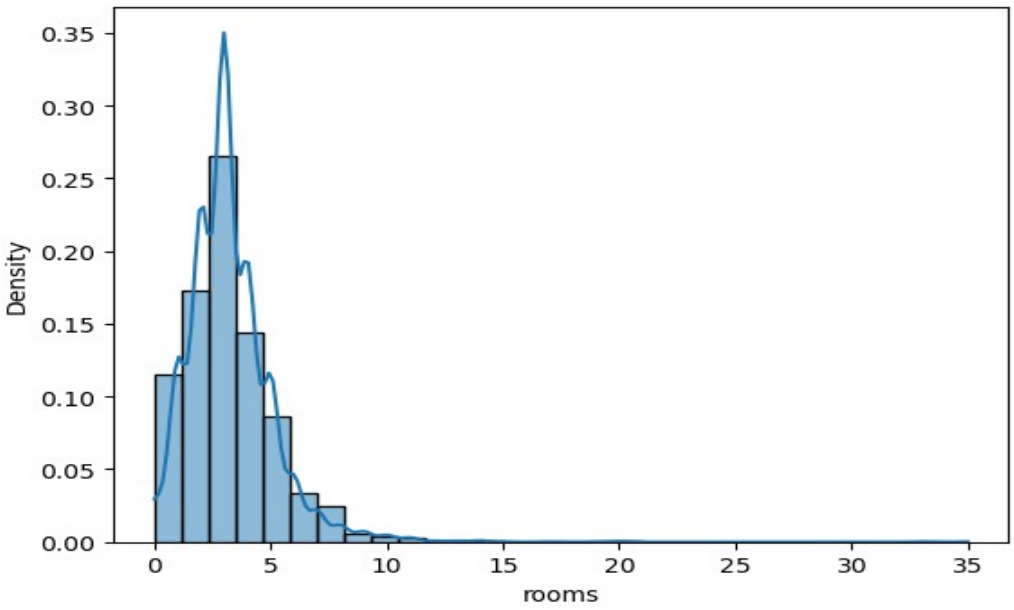

Ilustración 6: Histograma de la columna rooms

Por otro lado, la distribución de la columna baños está sesgada a la derecha, de forma que a medida que aumenta el número de baños, disminuye la cantidad de viviendas, como se puede apreciar en la ilustración 7. En cuanto a la cantidad de valores, se observa que el 75% de los registros tienen tres baños o menos, y el valor máximo es de 38. Esta cantidad máxima sugiere que se trata de un valor atípico dentro del conjunto de datos como se puede observar en la ilustración 8.

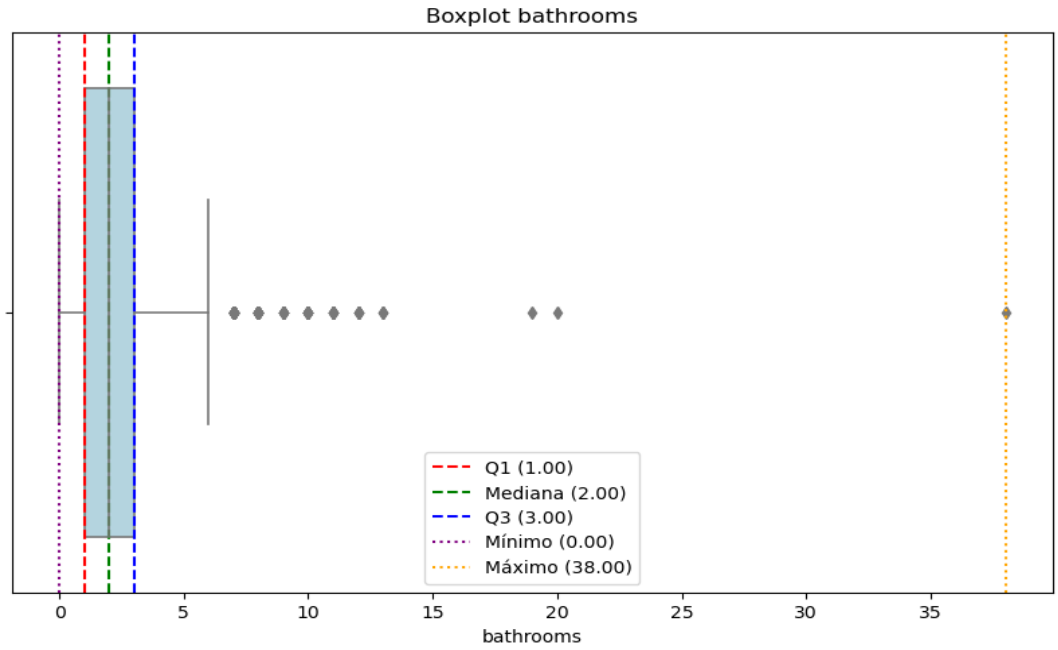

Ilustración 7: Diagrama de cajas y bigotes de la columna bathrooms

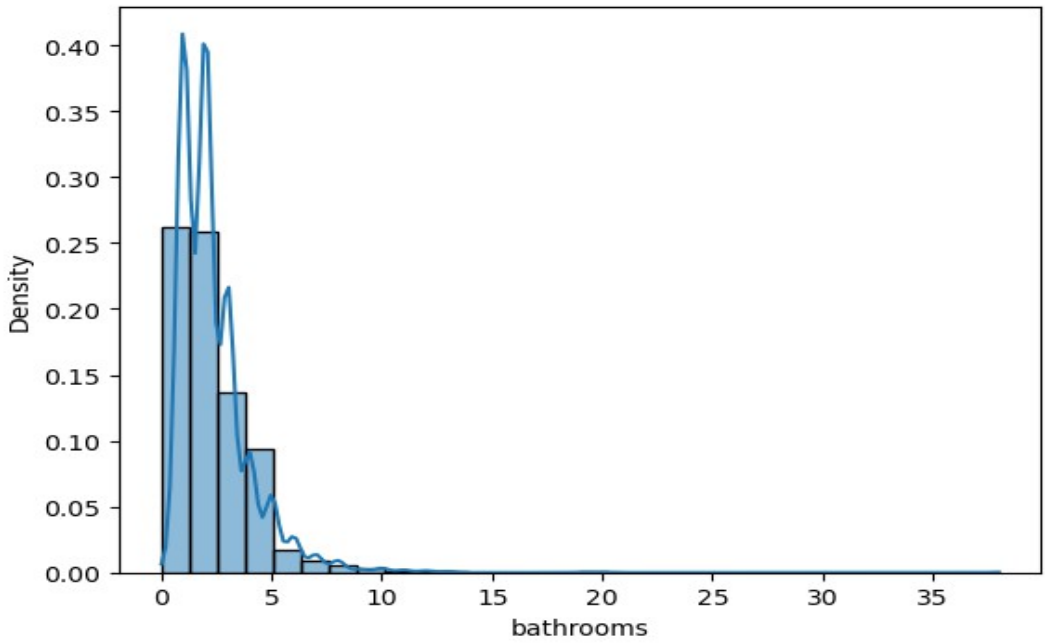

Ilustración 8: Histograma de la columna bathrooms

A continuación, en el análisis de la distribución del precio por metro cuadrado, se observa que, a medida que aumenta el valor de la columna, se reduce la cantidad de viviendas, lo que sugiere que la distribución de los datos está sesgada hacia la derecha como se puede ver en la ilustración 9. La columna presenta un valor mínimo de 41 euros y un valor máximo de 20.087 euros, lo que lleva a pensar que podrían existir valores atípicos en el conjunto de datos como se puede apreciar en la ilustración 10.

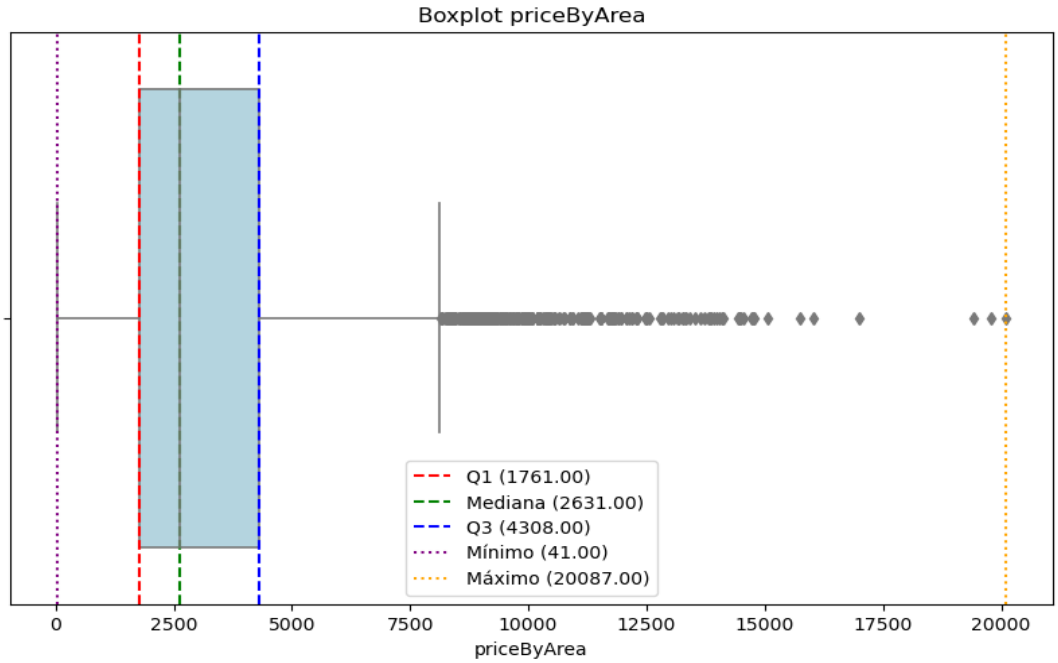

Ilustración 9: Diagrama de cajas y bigotes de la columna priceByArea

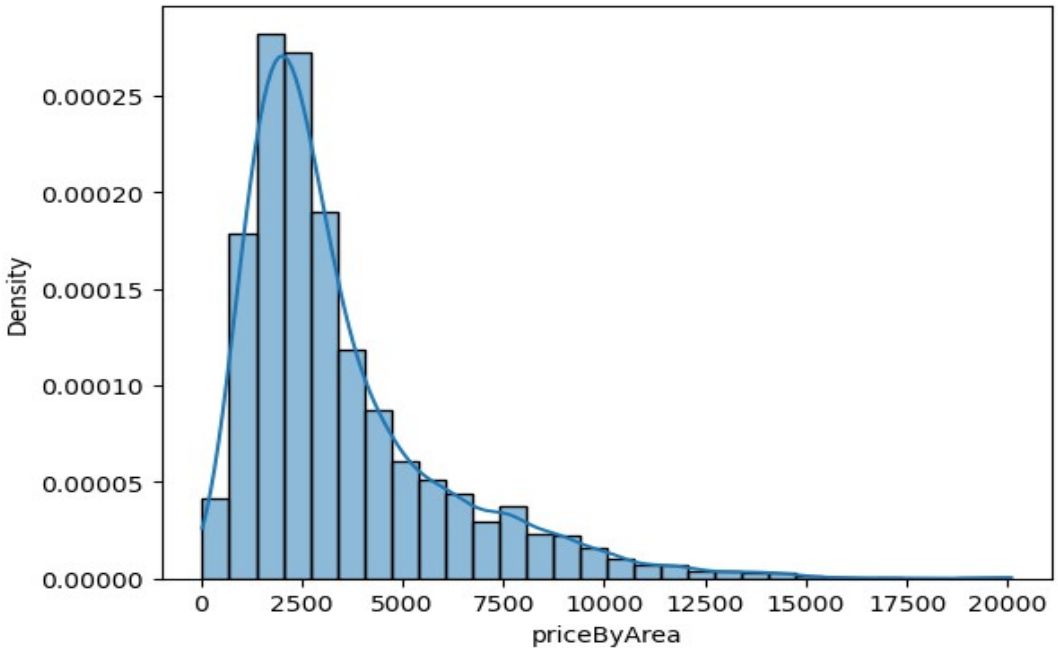

Ilustración 10: Histogramas de la columna priceByArea

Con el análisis de las columnas de longitud y latitud se confirma el análisis previo de cada una de las columnas, al observar que las viviendas están distribuidas en diferentes regiones de la Comunidad de Madrid. También se puede evidenciar la existencia de una región específica, concretamente la ciudad de Madrid, que concentra una alta cantidad de viviendas como se puede ver en la ilustración 11.

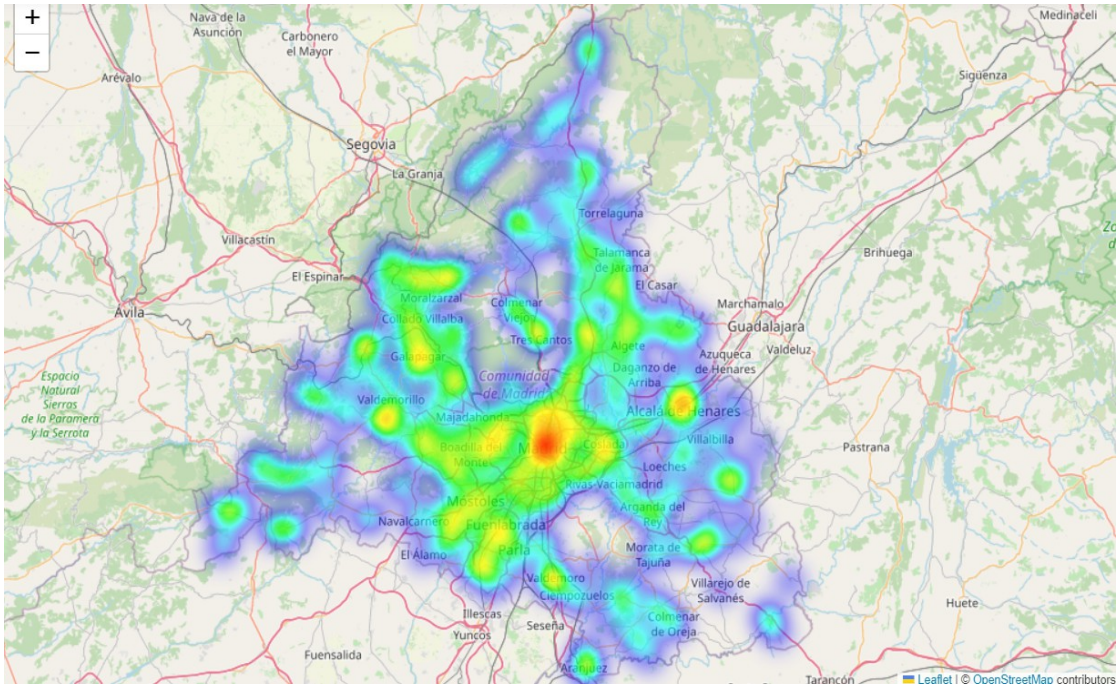

Ilustración 11: Mapa de calor de las columnas latitud y longitud

Finalmente, se procedió a crear una función que determinara si una columna tenía valores atípicos y a eliminarlos. Esta función se utilizó en las columnas en las que era posible que hubiera valores atípicos.

En el caso del análisis de variables categóricas, se realizaron gráficos de conteo por cada columna para mostrar la distribución de los valores. Dentro de ese conjunto de visualizaciones se pudo observar que la columna que indicaba la presencia de un aparcamiento solo tenía un valor verdadero para todos los valores no nulos. Esta columna tiene una fuerte relación con la columna que indica si el precio del aparcamiento se incluye en el precio de la vivienda.

Para solucionar ese problema, se determinó que la mayoría de viviendas que no indicaba la presencia de un aparcamiento es porque generalmente no tiene un aparcamiento. Esta hipótesis surge de que la presencia de un aparcamiento en una vivienda puede influir significativamente en el precio de la vivienda. En caso de que una vivienda tuviera un aparcamiento, sería señalizado. Por ello, se procedió a poner falso a todos los valores nulos de la columna hasParkingSpace (tiene aparcamiento).

En cuanto a la columna que indica si el aparcamiento se incluye en el precio de la vivienda, se procedió a poner falso a todas las columnas nulas que tienen el valor de la columna aparcamiento a falso.

Las columnas título, subtítulo y dirección se eliminan del conjunto de datos. Debido a que las columnas como tipología, tipo de propiedad, municipio, latitud y longitud ya cubren toda esa información. Por lo tanto, la reducción en el número de columnas simplifica el conjunto de datos sin perder información relevante.

3.1.E. Análisis multivariable

En la fase de análisis multivariable se procedió a realizar un análisis para explorar la relación entre el precio de la vivienda y otras variables relevantes en el conjunto de datos. Este análisis tuvo como objetivo identificar como las variables independientes influyen en el precio de la vivienda y determinar las correlaciones más significativas.

La visualización de la ilustración 12, es un mapa de calor que sirve para mostrar la relación de correlación entre las diferentes columnas. Para facilitar el entendimiento del gráfico, se muestra el valor de la correlación entre cada par de variables, donde los valores cercanos al 1 o -1 son correlaciones fuertes, ya sean positivas como negativas. Este valor también se refleja mediante el uso de colores, como el rojo para las correlaciones positivas y el azul para las correlaciones negativas. El tono de cada color determina el nivel de intensidad de la relación, donde los tonos oscuros muestran una intensidad fuerte y los tonos claros indican una intensidad débil.

Dentro de esta visualización de la correlación entre los pares de variables, se puede destacar las siguientes correlaciones:

- **Precio y tamaño de la vivienda:** la correlación entre estas variables es una correlación positiva moderada al tener un valor de 0.49. Esto sugiere que, a medida que aumenta el tamaño de la vivienda, el precio tiende a incrementarse.
- **Precio y número de habitaciones:** la correlación entre esta variable es de 0.28. Este valor indica que existe una correlación positiva modera por lo que sugiere que el número de habitaciones influye en el precio de la vivienda.
- **Precio y número de baños:** la correlación entre el precio y el número de de baños: se puede apreciar que existe una correlación moderada al tener un valor de 0.50. Por lo que un número mayor de baños influye en el precio de la vivienda.
- **Precio y precio por metros cuadrados:** la correlación entre estas variables es de 0.56. Este valor indica que existe una correlación fuerte por lo que sugiere que a medida que aumenta el precio por metros cuadrado de una vivienda, también aumenta el precio de la vivienda.
- **Precio y aparcamiento:** el valor de la correlación es de 0.34. Este valor indica que la presencia de un aparcamiento influye en el precio de la vivienda.
- **Precio y ascensor:** se puede apreciar que el valor de la correlación es de 0.25. Con este valor se puede deducir que la presencia de un ascensor influye en el precio de la vivienda.
- **Tamaño de vivienda y número de habitaciones:** con un valor de correlación de 0.59 se puede deducir que el tamaño de la vivienda esta fuertemente relacionada con el número de habitaciones.
- **Tamaño de la vivienda y número de baños:** el valor de la correlación es de 0.66. Este valor indica que existe una correlación fuerte por lo que sugiere que el tamaño de la vivienda puede influir en el números de baños.

Para finalizar, se han observado más correlaciones entre los pares de variables dentro del conjunto de datos. En esta sección, solo se detallan aquellas correlaciones más significativas que tienen una fuerte influencia en el precio de la vivienda.

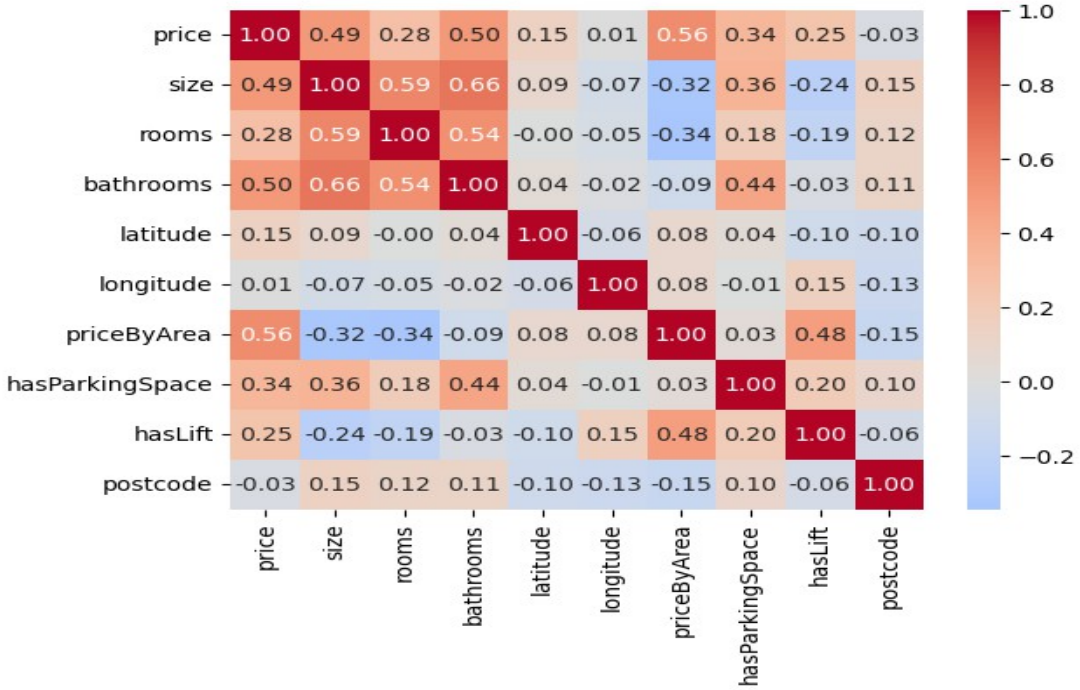

Ilustración 12: Mapa de correlación entre las variables numéricas

4. Modelado

Una vez que se ha realizado la fase del análisis exploratorio y se tiene un conjunto de datos adecuado, se procedió al desarrollo de los diferentes modelos. Los modelos utilizados para los objetivos planteados son modelos de Regresión lineal, bosque aleatorio y extreme gradient boosting.

La mayoría de los modelos de aprendizaje automático solo entiende de valores numérico. Por ello, fue necesario realizar una transformación de las variables categórica a numéricos. El siguiente paso, fue una evaluación de los distintos modelos aplicando una métrica con la finalidad de determinar el modelo que ofreció mejor rendimiento. El mejor modelo entrenado, se le realizó un proceso de optimización para tratar de mejorar el rendimiento del modelo.

4.1. Transformación

Los modelos de aprendizaje automático usualmente suelen procesar los datos en formato numérico. Debido a ello, si se quiere construir un modelo de aprendizaje automático hay que codificar todas las variables categóricas en numéricas. En este caso, se ha decantado por utilizar el método de codificación ONE-HOT.

Las columnas que han sido necesarias realizar una transformación son las siguientes:

- **propertyType** (Tipo de propiedad)
- **municipality** (Municipio)
- **status** (Estado)
- **typology** (Tipologia)
- **floor** (Planta)

- **hasLift** (Tiene Ascensor)
- **title** (Titulo)

En el caso de las variables categóricas se ha utilizado la función StandardScaler que consiste en escalar los datos de forma que tenga una media de 0 y una desviación estándar de 1 con el objetivo de evitar sesgos en la optimización. Con esta función se intenta mejorar el rendimiento del modelo, ya que los datos se ajustan mejor al modelo.

Para simplificar el proceso, se decidió crear un pipeline donde se realizó la transformación de forma automática para cada tipo de variables.

4.2. Entrenamiento

A continuación, se llevó a cabo la separación del conjunto de datos en dos partes. El primer conjunto de datos estaba compuesto de variables predictoras llamadas X. Mientras el segundo conjunto solo tendría la variable objetivo nombrada como y.

Estos dos conjuntos serán utilizados para entrenar el modelo y el método de entrenamiento del modelo que se ha elegido es la validación cruzada. Este método, consiste en dividir el conjunto de datos (X,y) en k particiones y se ejecuta k iteraciones de entrenamiento y validación. En este caso, se ha divido el conjunto de datos en 5 particiones para realizar 4 iteraciones para el entrenamiento y una iteración para la validación de cada modelo.

En la fase de entrenamiento, se presenta al modelos un conjunto de datos y el modelo ajusta automáticamente los parámetros propios del modelo para realizar las predicciones. Mientras que en la fase de entrenamiento se realiza la validación con la partición restantes y se verifica el rendimiento del modelo.

Para automatizar el proceso, se ha creado una función donde se le pasa por parámetros el clasificador y transformador de las columnas categóricas y numéricas. En esta función se ha creado un pipeline en el que se utiliza para la función de validación cruzada. Una vez que se ha ejecutado la validación cruzada, los resultados de la evaluación y el modelos se guarda en la carpeta resultados y modelos.

4.3. Evaluación

Una vez que se ha completado la fase de entrenamiento, se procede a evaluar el rendimiento de cada uno de los modelos. Para la evaluación se han utilizado las métricas RMSE, MAE y R2. Cada una de estas métricas proporciona una información diferente sobre el desempeño del modelo.

A continuación se describe el análisis de los modelos:

- **Regresión lineal**: en la tabla 1 se puede observar que el valor promedio del MAE es de 4.562,67 euros, lo que indica la diferencia absoluta entre el valor real y el valor predicho relativamente grande. El RMSE es relativamente alto y señala que el modelo se ajusta peor a nuevos datos. Este modelo tiene una tasa de acierto del 84%, como se muestra en la tabla 1.
- **Bosque aleatorio**: en este modelo, el valor medio del MAE es de 8.301,27 euros, lo que indica el margen de error. El valor medio del RMSE es de 15.783,15 euros, lo que señala que el modelo se ajusta peor ante cambios en el conjunto de datos. En cuanto a la métrica R2, el modelo tiene una tasa de acierto del 99.16% ,como se puede observar en la tabla 1.

- **Extreme gradient boosting**: este modelo presenta un valor medio de MAE de 8.429,89 euros, un valor medio de RMSE de 14.026,51 euros y un R2 de 0.9934. Esto indica que el modelo tiene un pequeño margen de error, aunque este aumenta con nuevos conjuntos de datos. La tasa de acierto del modelo es del 99.34%, como se muestra en la tabla 1.

En la comparación de modelos, se ha elegido el Extreme gradient boosting como el mejor para los objetivos planteados. Aunque el modelo de Bosque aleatorio ofrece un buen rendimiento, el Extreme gradient boosting presenta mejores resultados en las métricas RMSE y R2

El modelo de Extreme gradient boosting tiene un margen de error mayor que el modelo de Bosque aleatorio, pero su rendimiento en nuevos conjuntos de datos fue un factor crucial para su elección.

Modelo	Mean RMSE	Mean MAE	Mean R2
Regresión lineal	69114,72	4562,67	0.842709
Bosque aleatorio	15783,15	8301,27	0.993421
Extreme gradient boosting	14026,51	8429,89	0.993421

Tabla 1: Métricas de evaluación de los modelos

4.4. Optimización del modelo

En este apartado consiste en optimizar el modelo elegido con el parámetro que mejor se ajuste al conjunto de datos. Para ello, se ha definido una lista de parámetros para el modelo Extreme gradient boosting y se ha entrenado el modelo utilizando la técnica de evaluación cruzada para determinar el parámetro que ofrece el mejor rendimiento.

Una vez entrenado el modelo, se utiliza los parámetros que mejor resultado han generado para entrenar el modelo con el conjunto de datos de entrenamiento. Este modelo entrenado se realiza una prueba con el conjunto de dato de prueba.

En la tabla 2, se observa que las métricas generadas por el modelo optimizado tienen un MAE de 6.100,1 euros, un MSE de 10.937,82 euros y un R2 de 0.9960. Estos resultados indican que el modelo optimizado ofrece mejores valores de MAE, MSE y R2 en comparación con el modelo no optimizado como se muestra en la tabla 2.

Modelo	Mean RMSE	Mean MAE	Mean R2
Extreme gradient boosting	10937,82	6100,1	0.995981

Tabla 2: Métricas de evaluación del modelo Extreme gradient boosting

5. Repositorio del proyecto

En este apartado contiene el enlace del repositorio donde se ha guardado el proyecto realizado.

URL: https://github.com/BungisaBeto/tfm-bungisa-beto

VI - Trabajos relacionados

En este apartado contamos con varios trabajos, investigaciones y proyectos relacionados con el uso de la inteligencia artificial en el mercado inmobiliario. Así como con la creación de un modelo de predicción del precio de vivienda.

Autores como Andoni Fernández Gariano en su trabajo de fin de grado titulado de 2023 llamado “Análisis del Impacto de las Smart Technologies en el mercado inmobiliario español: un enfoque a la digitalización del sector”. Hace una introducción al mercado inmobiliario en España, así como el impacto de las Smart Techologies o las tecnologías inteligentes definidas por en el proyecto como “aquellas tecnologías que implementan análisis de Big Data, Inteligencia Artificial, conectividad a Internet y otras tecnologías avanzadas para mejorar y automatizar determinadas actividades”. Resume los beneficios del uso del Big Data en el ámbito inmobiliario, por ejemplo, la capacidad de desarrollar pronósticos a corto y medio plazo en el ámbito de la gestión, fijación de un mejor precio en el mercado, la ciberseguridad y la tecnología Blockchain, la labor comercial, el uso de tours virtuales, el uso de Chatbots, etc FGM01.

María Benítez Cullerés en el proyecto de fin de Máster “Las nuevas PropTech y su transformación digital aplicada al sector inmobiliario”. Habla de la importancia de apostar por Proptech (conceptos “property” y “technology” y se refiere a empresas que utilizan tecnología para mejorar o reinventar servicios dentro del sector inmobiliario. Utilizando el aprendizaje automático, Big Data y la geolocalización. en las empresas dedicadas al sector inmobiliario. Así como el “boom” que se produjo en 2017 en España dando especial interés al primer mapa de Proptech diseñado por Aguirre Newman y Finnovating Proptech, sobre las primeras 58 empresas y su clasificación en 5 categorías.

El uso de esas nuevas tecnologías ayuda a conseguir una mejor rentabilidad en el patrimonio inmobiliario minimizando los riesgos a la hora de las inversiones, mayor transparencia en el mercado, mayor alcance y promoción, ayudan a invertir el capital de manera mas inteligente y diversificada BCM01.

Núñez Tabales, Rey Carmona, y Caridad y Ocerin en el 2026 presentaron un artículo de investigación denominado “Artificial intelligence (AI) techniques to analyze the determinants attributes in housing prices” en el que describen el uso de nuevas metodologías alternativas utilizadas para la valoración de los inmuebles en la ciudad de Sevilla. Dentro del estudio, se compararon dos métodos: la metodología hedónica clásica y las Redes Neuronales Artificiales (RNA) para determinar el precio de mercado de las viviendas. En ese estudio se tomaron en cuenta las características del inmueble como la ubicación, la superficie, el garaje, el trastero, la piscina, etc. Dentro de esas características, la superficie construida es la variable que más influye en la asignación de precios de una vivienda.

Dentro del ámbito de valoración en el mercado inmobiliario, hicieron varios estudios entre la regresión múltiple y los sistemas de inteligencia artificial, mostrándose la superioridad de las RNA, en la mayoría de los casos por su precisión y capacidad para estimar valores atípicos (outliers). No obstante, no se puede afirmar categóricamente que las RNA sean siempre más eficientes que los métodos tradicionales, sugiriendo que la elección del método debe basarse en el caso específico NRC01.

VII - Conclusiones y líneas de trabajo futuras

En este trabajo se han estudiado diferentes modelos de aprendizaje automático. El objetivo fue encontrar un modelo con las características necesarias para predecir el precio de las viviendas en la Comunidad de Madrid. Para ello, se recopilaron datos utilizando la API de Idealista hasta conseguir una muestra adecuada. Luego, se realizó un análisis exploratorio para obtener un conjunto de datos de alta calidad y adecuado para el entrenamiento del modelo.

A continuación, se llevó a cabo el entrenamiento de los diferentes modelos utilizando la validación cruzada. Los resultados de los modelos entrenados fueron evaluados para seleccionar el mejor modelo. Dentro de los modelos entrenados, el modelo de Extreme gradient boosting y el de Bosque aleatorio ofrecieron buenos resultados, con un RMSE de 14.026,51 euros para el Extreme gradient boosting y un RMSE de 15.783,15 euros para el Bosque aleatorio. El modelo Extreme gradient boosting fue mejor debido a su menor tasa de error en comparación con el de Bosque aleatorio cuando se enfrentó a nuevos datos. Estos resultados eran de esperar, ya que ambos modelos combinan múltiples algoritmos, lo que puede mejorar la detección de relaciones entre las variables del conjunto de datos.

Tras la elección del modelo, se procedió a realizar un proceso de optimización con diferentes parámetros para determinar cuáles se ajustaban mejor a la naturaleza de los datos. La versión optimizada del modelo Extreme gradient boosting ofreció un mejor rendimiento que la versión anterior, reduciendo el RMSE a 10.937,82 euros. Una vez obtenido el modelo optimizado, se procedió a probarlo con nuevos datos para comparar las predicciones del modelo con los valores reales. Los resultados confirmaron la evaluación previa del rendimiento del modelo.

Con las pruebas realizadas, se puede confirmar que se han logrado los objetivos marcados inicialmente. Se ha desarrollado un modelo capaz de predecir el precio de la vivienda con un margen de error absoluto de 6.100 euros. Este margen de error es bastante aceptable, ya que actualmente la tasación de viviendas suele tener un margen del 10%. Por lo tanto, estas técnicas pueden ser utilizadas para desarrollar modelos que se integren en sistemas de tasación de precios de viviendas, ayudando al sector inmobiliario a realizar tasaciones automatizadas y reducir la carga de trabajo.

Por otro lado, en este trabajo se han utilizado los modelos de Regresión lineal, Bosque aleatorio y Extreme gradient boosting para la predicción del precio de las viviendas. Sin embargo, existen una gran variedad de modelos, como las redes neuronales o la Regresión por vecinos más cercanos, que también podrían utilizarse y ofrecer resultados interesantes.

Para finalizar, se ha demostrado que existen diferentes técnicas de aprendizaje automático, y la elección de unas u otras depende en cierta medida de los objetivos del proyecto. Por ello, es importante realizar una buena planificación y estudiar el problema en detalle para encontrar la alternativa que se adapte mejor al objetivo. Además, la fuente de datos es crucial por lo que es necesario llevar a cabo un análisis exhaustivo de los datos mediante un análisis exploratorio, ya que la calidad del modelo depende en gran medida de esto.

Este proyecto se puede ampliar entrenando grandes conjuntos de datos para observar el comportamiento del modelo en diferentes escenarios. Además, se podría desarrollar una aplicación con una interfaz de usuario que permita seleccionar los modelos entrenados y cargar un conjunto de datos, ya sea mediante un archivo o una base de datos, para realizar predicciones. En esta aplicación, se podría planificar el entrenamiento o ajuste de los modelos para que se ejecuten en un periodo determinado.

VIII - REFERENCIAS

Bibliografía

- Amazon. (2024). *¿Qué es el aprendizaje no supervisado?*. Obtenido de:https://www.ibm.com/es-es/topics/unsupervised-learning
- Amazon. (2024). *¿En qué consiste el ajuste de hiperparámetros?*. Obtenido de:https://aws.amazon.com/es/what-is/hyperparameter-tuning/
- Amazon. (2023). *Guía para desarrolladoresAmazon Machine Learning*. Obtenido de:https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.-pdf#cross-validation
- Amazon. (024). *¿Qué es la regresión lineal?*. Obtenido de:https://aws.amazon.com/es/what-is/linear-regression/#:~:text=La%20regresi%C3%B3n%20lineal%20es%20una,independiente%20como%20una%20ecuaci%C3%B3n%20lineal
- Benítez Cullerés, M.. (2018). *Las nuevas PropTech y su transformación digital aplicada al sector inmobiliario*. Obtenido de:https://upcommons.upc.edu/bitstream/handle/2117/130390/Mem%C3%B2ria_BenitezMaria.pdf
- Bruce P., Bruce A., Gedeck P.. (2022). *Estadística práctica para ciencia de datos con R y Python» 2022*.:
- Culmia. (2023). *Inteligencia Artificial en el sector inmobiliario: ¿Cómo nos afecta?*. Obtenido de: https://www.culmia.com/blog/inteligencia-artificial-sector-inmobiliario
- Fernadez Aviles Gema, Montero Jose Maria,. (2024). *Fundamentos de ciencia de datos con R*. Obtenido de:https://cdr-book.github.io/index.html
- Fernández Garitano, A.. (2023). *Análisis del Impacto de las Smart Technologies en el Mercado Inmobiliario Español: Un Enfoque sobre la Digitalización del Sector.* Obtenido de:https://digibuo.uniovi.es/dspace/bitstream/handle/10651/69300/tfg_AndoniFern%C3%A1ndezGaritano.pdf?sequence=4&isAllowed=y
- Gavilán A.. (2023). *Informe anual 2023. Capítulo 4: El mercado de la vivienda en España: evoluciónreciente, riesgos y problemas de accesibilidad*. Obtenido de:https://www.bde.es/f/webbe/GAP/Secciones/SalaPrensa/IntervencionesPublicas/DirectoresGenerales/economia/Arc/Fic/IIPP-2024-04-23-gavilan2-es-or.pdf
- IAAR. (2024). *La Era de las máquinas inteligentes*. Obtenido de:https://iaarbook.github.io/machine-learning/
- Iaarbook. (2024). *La Era de las máquinas inteligentes*. Obtenido de:https://iaarbook.github.io/
- Ibm. (2024). *¿Qué son los grandes modelos de lenguaje (LLM)?*. Obtenido de:https://www.ibm.com/es-es/topics/large-language-models
- Ibm. (2024). *¿Qué es la minería de datos?*. Obtenido de:https://www.ibm.com/es-es/topics/data-mining
- Ibm. (2024). *¿Qué es un bosque aleatorio?*. Obtenido de:https://www.ibm.com/es-es/topics/random-forest
- Lasse Rouhiainen. (2018). *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro futuro*. Obtenido de:
- Moreno Otero M., Garcia Lomas J.B. (2024). *El sector inmobiliarioen España*. Obtenido de:https://www.ieemadrid.es/wp-content/uploads/El-sector-inmobiliario-en-Espa%C3%B1a.pdf

- Núñez Tabales, J. M., Rey Carmona, F. J., & Caridad y Ocerin, J. M.. (2016). *Artificial intelligence (AI) techniques to analyze the determinants attributes in housing prices*. Obtenido de:https://www.redalyc.org/articulo.oa?id=92549096001
- Xiang LI. (2024). *Comparing Linear Regression and Decision Trees forHousing Price Prediction*. Obtenido de:https://www.atlantis-press.com/article/125998094.pdf