uc3m | Universidad Carlos III de Madrid

Grado en Estadística y Empresa
2025-2026

Trabajo Fin de Grado

"Análisis estadístico y modelización predictiva de los precios de viviendas en la Comunidad de Madrid"

Eloy Celaya López

Tutora

Sandra Benítez Peña

Madrid, 2026

RESUMEN

El mercado inmobiliario constituye uno de los pilares de la economía española y presenta una especial relevancia en la Comunidad de Madrid, donde la evolución de los precios de compraventa refleja tanto dinámicas macroeconómicas como factores micro vinculados a las características de los inmuebles.

El presente trabajo, titulado *"Análisis estadístico y modelización predictiva de los precios de compraventa de viviendas en la Comunidad de Madrid"*, persigue un doble objetivo. En primer lugar, se realiza un análisis descriptivo y multivariante a partir de fuentes oficiales con el fin de estudiar la evolución histórica de los precios y su relación con variables socioeconómicas y territoriales.

En segundo lugar, se emplean datos micro procedentes de plataformas inmobiliarias que recogen información detallada de las viviendas (superficie, número de habitaciones, ubicación, entre otras características), sobre la que se desarrollan modelos predictivos del precio de compraventa.

Con los resultados obtenidos, se logra alcanzar una visión complementaria: por un lado, el comportamiento agregado de los precios en la región y, por otro, la identificación de los determinantes específicos que influyen en el valor de un inmueble. Asimismo, se discute las implicaciones de los hallazgos tanto para el análisis académico como para la toma de decisiones en el sector inmobiliario. De forma complementaria, se ha desarrollado una aplicación web para mostrar todos los resultados mediante visualizaciones y predicciones interactivas.

Palabras clave: precios de la vivienda; análisis estadístico; análisis multivariante; modelización predictiva; aplicación web

ABSTRACT

The real estate market constitutes one of the pillars of the Spanish economy and is of particular relevance in the Community of Madrid, where the evolution of housing prices reflects both macroeconomic dynamics and micro-level factors related to property characteristics.

This Bachelor's Thesis, entitled *"Statistical Analysis and Predictive Modeling of Housing Sale Prices in the Community of Madrid"*, pursues a twofold objective. First, a descriptive and multivariate analysis is conducted using official data sources in order to examine the historical evolution of housing prices and their relationship with socioeconomic and territorial variables.

Second, micro-level data from real estate platforms are employed, providing detailed information on property characteristics such as floor area, number of rooms, location, among others. Based on these data, predictive models for housing sale prices are developed.

The results provide a complementary perspective: on the one hand, an understanding of the aggregate behaviour of housing prices in the region, and on the other, the identification of specific determinants that influence property values. Additionally, the implications of the findings are discussed from both an academic perspective and a decision-making standpoint within the real estate sector. Finally, a web-based application is developed to present the results through interactive visualizations and predictions.

Keywords: housing prices; statistical analysis; multivariate analysis; predictive modeling; web application

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Motivación

El acceso a la vivienda se ha convertido en uno de los principales retos económicos y sociales de nuestro país, especialmente en grandes áreas urbanas como la Comunidad de Madrid[1]. El incremento de los precios en los últimos años, junto con la elevada demanda residencial y la limitada oferta disponible, ha generado una situación de notable tensión en el mercado inmobiliario.

En el último trimestre de 2025, el valor tasado medio de la vivienda en la Comunidad de Madrid alcanzó los 3732.5 €/m², muy por encima de la media nacional (2153.4 €/m²) (Ministerio de Transportes, Movilidad y Agenda Urbana, 2025). Este diferencial refleja las fuertes desigualdades territoriales existentes entre municipios y distritos, donde factores como la renta media, la accesibilidad o la cercanía al centro urbano condicionan significativamente el precio de la vivienda.

Ante este escenario, resulta esencial comprender cómo han evolucionado los precios a lo largo del tiempo y qué variables explican sus diferencias actuales, tanto a nivel regional como dentro del propio municipio de Madrid. El análisis estadístico y predictivo ofrece una herramienta idónea para identificar patrones, modelizar tendencias y evaluar el peso de factores estructurales como la superficie, el número de habitaciones o la localización.

Este trabajo se propone, por un lado, realizar un análisis descriptivo de datos para estudiar la evolución del mercado de la vivienda, apoyándose principalmente en técnicas de visualización y en la comparación descriptiva entre distintos ámbitos territoriales y por otro, el desarrollo de modelos predictivos para los determinantes del precio de compraventa de las viviendas en la Comunidad de Madrid. El enfoque combina el uso de fuentes oficiales, como el valor tasado medio del Ministerio de Vivienda y Agenda Urbana o el Colegio de Registradores, con datos recientes del mercado inmobiliario, los cuales se explican con mayor detalle en la Sección 3. De esta forma, se busca aportar una visión empírica y cuantitativa que contribuya a comprender las dinámicas del mercado residencial madrileño y su relevancia en el contexto socioeconómico actual.

1.2. Estado del arte

El análisis del precio de la vivienda ha sido objeto de un amplio interés en la literatura económica y estadística, especialmente en contextos urbanos donde la heterogeneidad territorial desempeña un papel fundamental.

En el ámbito español, trabajos como el de Martínez Pagés y Maza (2003) analizan la evolución del precio de la vivienda desde una perspectiva macroeconómica, poniendo de manifiesto la influencia de factores estructurales y de demanda. En el caso específico de la Comunidad de Madrid, diversos estudios han destacado la existencia de una marcada desigualdad espacial en los precios inmobiliarios. Robles Quiñonero (2016) y Álvarez (2012) muestran que el valor de la

[1] https://www.elconfidencial.com/espana/madrid/2025-09-14/el-gran-reto-de-madrid-absorber-en-los-desarrollos-un-millon-de-personas-mas-con-la-vivienda-estancada_4205233/

vivienda presenta patrones espaciales claros a nivel intraurbano, asociados tanto a la localización como a características socioeconómicas del entorno. Estos resultados refuerzan la necesidad de incorporar herramientas de análisis espacial en el estudio del mercado inmobiliario. En esta línea, la literatura reciente ha subrayado la presencia de autocorrelación espacial en los precios inmobiliarios, evidenciada mediante diversos indicadores. La consideración de estos efectos resulta clave para evitar inferencias sesgadas y para comprender la dinámica territorial del mercado.

Paralelamente, numerosos trabajos han abordado la predicción del precio de la vivienda mediante modelos estadísticos y de aprendizaje automático. Estudios como Amri y Tularam (2012) o Zhang (2021) comparan modelos de regresión lineal con enfoques más flexibles, destacando la mejora en capacidad predictiva cuando se incorporan relaciones no lineales y múltiples características del inmueble.

En conjunto, la literatura pone de manifiesto la relevancia de combinar análisis descriptivo, espacial y predictivo. El presente trabajo se sitúa en esta línea, integrando fuentes oficiales y datos microeconómicos de plataformas inmobiliarias, y aportando como elemento diferencial el desarrollo de una aplicación web interactiva que permite explorar los resultados y realizar predicciones personalizadas.

1.3. Objetivos

El objetivo principal del presente trabajo es analizar estadísticamente la evolución y los determinantes del precio de compraventa de las viviendas en la Comunidad de Madrid, integrando una perspectiva temporal, espacial y predictiva, facilitando la interpretación de los resultados mediante herramientas interactivas de visualización.

El estudio se centra en identificar las tendencias del mercado inmobiliario de la Comunidad de Madrid, analizar la estructura actual de precios entre municipios y distritos y estimar modelos que permitan predecir el valor de una vivienda en función de sus características. Todo ello se apoya en una aplicación web[2] interactiva que actúa como soporte para la exploración, el análisis y la comunicación de los resultados obtenidos.

Para llevar esto a cabo, se proponen los siguientes objetivos específicos:

1. Analizar la evolución temporal del valor tasado medio por metro cuadrado de las viviendas de la Comunidad de Madrid entre 2005 y 2025, utilizando datos oficiales.
2. Estudiar la distribución espacial de precios de la vivienda dentro de la región, identificando diferencias entre municipios.
3. Analizar el actual mercado de la vivienda en el municipio de Madrid a partir de una base de datos reciente de anuncios de compraventa, analizando las características estructurales que más influyen en el precio, tales como la superficie, el número de habitaciones o la ubicación.
4. Elaborar visualizaciones y mapas dinámicos que faciliten la interpretación de los resultados y la comprensión de las desigualdades territoriales dentro de la región.

[2] https://tfg-eloy-celaya.streamlit.app/

5. Desarrollar modelos predictivos de precios de vivienda mediante técnicas de regresión y aprendizaje automático, evaluando su precisión y capacidad explicativa.
6. Integrar los resultados del análisis descriptivo y predictivo en una aplicación web interactiva que permita al usuario explorar distintas visualizaciones, filtrar y comparar información espacial y temporal y obtener estimaciones del precio de una vivienda a partir de sus características gracias a los modelos desarrollados.

Con ello, el trabajo pretende aportar una visión cuantitativa y empírica del mercado inmobiliario madrileño, combinando el análisis histórico con herramientas modernas de predicción y visualización y ofreciendo resultados de utilidad tanto académica como práctica.

1.4. Estructura del documento

El presente Trabajo de Fin de Grado se organiza en varias Secciones que permiten desarrollar de manera ordenada los objetivos planteados y facilitar la comprensión de la metodología y los resultados obtenidos.

En primer lugar, en la Sección 1 se introduce el contexto del estudio, exponiendo la motivación del trabajo, los objetivos perseguidos y el alcance del análisis, así como una breve revisión del estado del arte y la estructura general del documento. La Sección 2 recoge el marco teórico y metodológico en el que se fundamenta el trabajo. En él se describen los principales conceptos estadísticos y de análisis multivariante empleados, con especial atención a las técnicas de tratamiento de datos mixtos, distancias robustas, imputación multivariante y modelos predictivos basados en proximidad. En la Sección 3 se presenta el conjunto de datos utilizado, detallando su procedencia, las variables consideradas y el proceso de preprocesamiento aplicado, incluyendo la limpieza de datos, la imputación de valores faltantes y el análisis exploratorio inicial. Posteriormente, la Sección 4 se dedica al desarrollo de los modelos estadísticos y de aprendizaje automático. En este capítulo se describen los modelos implementados, los criterios de ajuste y validación, así como la comparación de resultados entre los distintos enfoques considerados. En la Sección 5 se presenta la aplicación web desarrollada. Se comenta el método utilizado para su desarrollo, así como los resultados que se muestran y su uso. Finalmente, la Sección 6 recoge las conclusiones del trabajo, destacando las aportaciones principales, las limitaciones del estudio y posibles líneas de investigación futura.

El documento se completa con un apartado de referencias bibliográficas y anexos que contienen información técnica complementaria.

2. METODOLOGÍA

2.1. Técnicas exploratorias

En esta sección se describe el conjunto de técnicas empleadas para llevar a cabo el análisis exploratorio de los datos (*Exploratory Data Analysis*, EDA) (Tukey, 1977). El objetivo principal de esta fase es comprender la estructura del conjunto de datos, identificar patrones relevantes y detectar posibles problemas antes de proceder a la modelización predictiva.

Dado que en este trabajo se utilizan tanto datos históricos agregados como datos microeconómicos a nivel de vivienda, las técnicas exploratorias aplicadas varían en función del tipo de información analizada. En particular, el análisis exploratorio cumple un papel fundamental en la preparación de los datos para la modelización, ya que permite justificar decisiones clave relativas a la selección de variables, la aplicación de transformaciones y el tratamiento de observaciones atípicas (Dhummad, 2023).

A continuación, se describen de forma estructurada las principales técnicas exploratorias empleadas.

2.1.1.Análisis univariante

El análisis exploratorio univariante tiene como objetivo estudiar la distribución individual de cada variable, identificar posibles asimetrías, valores atípicos y errores de registro, y evaluar la necesidad de aplicar transformaciones.

Para las variables cuantitativas continuas se emplean tanto representaciones gráficas, como histogramas y diagramas de caja, como medidas descriptivas básicas, incluyendo la media, la mediana, la desviación típica y los cuartiles, que permiten caracterizar numéricamente la forma y dispersión de las distribuciones. En el caso de las variables discretas y binarias, el análisis se basa en el estudio de frecuencias absolutas y relativas, así como en el cálculo de proporciones, lo que facilita la identificación de desequilibrios entre categorías.

Este análisis resulta clave para detectar distribuciones altamente asimétricas, como las correspondientes al precio o la superficie de la vivienda, que presentan una elevada dispersión y la presencia de valores extremos. La combinación de herramientas gráficas y medidas estadísticas permite justificar de forma objetiva la aplicación de transformaciones, como la transformación logarítmica, con el fin de mejorar el comportamiento estadístico de los modelos posteriores (Cleff, 2019).

2.1.2. Análisis bivariante y multivariante

El análisis bivariante y multivariante se emplea para estudiar la relación entre la variable objetivo y las variables explicativas, así como para detectar posibles dependencias entre predictores. En esta fase se utilizan gráficos de dispersión, diagramas de caja condicionados y matrices de correlación (Denis, 2021).

Este análisis permite identificar relaciones funcionales relevantes, evaluar la intensidad de la asociación entre variables y detectar posibles problemas de multicolinealidad. Asimismo, proporciona una base empírica para la selección de variables y la posterior interpretación de los modelos estimados.

2.1.3. Preprocesamiento de los datos

El preprocesamiento de los datos constituye una etapa esencial en cualquier análisis estadístico, ya que permite garantizar la coherencia interna del conjunto de datos y su adecuación a las técnicas analíticas posteriores (Alberto & Di Lecce, 2023). Esta fase engloba un conjunto de procedimientos orientados a mejorar la calidad de la información disponible y a reducir posibles sesgos derivados de errores de registro o distribuciones problemáticas.

Entre las prácticas habituales de preprocesamiento se encuentra la detección y gestión de observaciones inconsistentes o no plausibles, como valores extremos que no responden a comportamientos reales del fenómeno estudiado y que pueden interpretarse como errores de medición o registro (Joshi & Patel, 2020). Es también común abordar el tratamiento de valores atípicos mediante distintas estrategias (Aggarwal, 2017) con el fin de limitar su influencia sobre los resultados del análisis.

Por otro lado, el preprocesamiento puede incluir la aplicación de transformaciones sobre determinadas variables, especialmente cuando presentan distribuciones altamente asimétricas. En este contexto, transformaciones como la logarítmica se utilizan frecuentemente para estabilizar la varianza, aproximar la normalidad y facilitar la modelización estadística posterior (West, 2021).

En general, las técnicas de preprocesamiento se apoyan en la información obtenida a través del análisis exploratorio de los datos y se aplican de manera sistemática para asegurar la consistencia y la robustez de los resultados obtenidos en las etapas posteriores del análisis.

2.2. Técnicas predictivas

El objetivo de esta sección es presentar los modelos predictivos empleados para estimar los precios inmobiliarios a partir de variables socioeconómicas, territoriales y temporales disponibles. La predicción de precios de viviendas puede abordarse mediante diferentes enfoques, desde modelos lineales tradicionales hasta técnicas de *machine learning* más complejas capaces de capturar relaciones no lineales.

2.2.1. Marco común de modelización

Con el objetivo de garantizar una comparación coherente entre los distintos modelos predictivos considerados, se establece un marco metodológico común que se aplica de forma sistemática a todos ellos. Este enfoque permite aislar el efecto propio de cada algoritmo, asegurando que las diferencias observadas en el rendimiento se deban a la capacidad del modelo y no a decisiones metodológicas inconsistentes.

Preparación de los datos y métodos de validación

Una vez realizado el preprocesamiento explicado en la Sección 2.1.4, los datos deben prepararse adecuadamente para su utilización en procesos de modelización y evaluación. En este contexto, resulta fundamental establecer procedimientos de validación (Michelucci, 2024) que permitan evaluar el rendimiento de los modelos y su capacidad de generalización a nuevas observaciones.

Una estrategia habitual, la validación simple, consiste en dividir el conjunto de datos en subconjuntos con finalidades distintas, como el ajuste del modelo y su evaluación posterior. Esta técnica de validación permite obtener una primera estimación del comportamiento del modelo

fuera de la muestra utilizada para su entrenamiento y constituye un enfoque ampliamente utilizado por su sencillez e interpretabilidad.

No obstante, para obtener estimaciones más robustas del rendimiento del modelo, es frecuente recurrir a métodos de validación basados en múltiples particiones de los datos. Entre ellos, la validación cruzada *k*-fold es una de las técnicas más extendidas. Este procedimiento divide el conjunto de datos en *k* subconjuntos o pliegues, utilizando de forma iterativa cada uno de ellos como conjunto de validación y los restantes como conjunto de entrenamiento. De este modo, se obtienen varias estimaciones del rendimiento del modelo, cuya agregación permite reducir la dependencia de una partición concreta.

En general, la elección del método de validación depende de factores como el tamaño del conjunto de datos, la complejidad del modelo y los objetivos del análisis. El uso de diferentes estrategias de validación permite obtener una visión más completa del comportamiento de los modelos y contribuye a una evaluación más fiable de su capacidad predictiva.

Métricas de evaluación

La evaluación del rendimiento de los modelos depende del tipo de problema abordado. En términos generales, los métodos de aprendizaje supervisado pueden orientarse a problemas de regresión, cuando la variable respuesta es cuantitativa, o a problemas de clasificación, cuando la respuesta es categórica. Cada uno de estos enfoques requiere métricas de evaluación específicas, adaptadas a la naturaleza de la variable objetivo y a los objetivos del análisis.

En el contexto de los modelos de regresión, las métricas de evaluación se centran en cuantificar la capacidad explicativa del modelo y la precisión de sus predicciones sobre variables continuas. Entre las métricas más utilizadas se encuentran el coeficiente de determinación R^2 y el error cuadrático medio (RMSE), que permiten evaluar de forma complementaria el ajuste del modelo y la magnitud de los errores de predicción.

La evaluación y comparación de distintos modelos de regresión se realiza habitualmente mediante estas métricas, calculadas tanto sobre el conjunto de entrenamiento como sobre conjuntos de validación o prueba, lo que permite analizar simultáneamente el comportamiento del modelo dentro y fuera de la muestra de ajuste.

El coeficiente de determinación $R^2 \in [0,1]$ mide la proporción de la variabilidad total de la variable respuesta que es explicada por el modelo. Se define como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde y_i representa el valor observado, $\hat{y}_i$ el valor predicho por el modelo e $\bar{y}$ la media muestral de la variable respuesta. Valores de R^2 cercanos a 1 indican un alto poder explicativo, mientras que valores bajos sugieren que el modelo apenas mejora respecto a una predicción basada en la media.

Por su parte, el error cuadrático medio (RMSE) cuantifica el error medio de predicción penalizando de forma más severa los errores grandes. Se calcula como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Finalmente, la comparación de las métricas entre los subconjuntos de entrenamiento y *test* permite detectar posibles problemas de sobreajuste del modelo entrenado: diferencias pequeñas entre ambos conjuntos indican un modelo estable y con buena capacidad de generalización, mientras que discrepancias elevadas sugieren que el modelo se ajusta excesivamente a los datos de entrenamiento.

Preprocesamiento y prevención del *data leakage*

En el desarrollo de modelos predictivos, resulta fundamental evitar situaciones de *data leakage*, entendidas como la incorporación involuntaria de información procedente de datos que no deberían estar disponibles durante la fase de ajuste del modelo. Este fenómeno puede conducir a estimaciones excesivamente optimistas del rendimiento y a conclusiones erróneas sobre la capacidad predictiva real de los modelos.

Para prevenir este tipo de problemas, es habitual estructurar el preprocesamiento de los datos de forma que las transformaciones aplicadas a las variables explicativas se aprendan únicamente a partir de la información disponible en el conjunto de entrenamiento. Entre estas transformaciones se incluyen, por ejemplo, la estandarización de variables numéricas, la codificación de variables categóricas o la imputación de valores perdidos.

En la práctica, este enfoque se implementa mediante procedimientos sistemáticos que integran las etapas de preprocesamiento y modelización, garantizando que cualquier transformación se aplique posteriormente de forma coherente a los conjuntos de validación o prueba. De este modo, se preserva la independencia entre las fases de ajuste y evaluación, contribuyendo a una estimación más fiable del rendimiento predictivo de los modelos.

2.2.2. Regresión Lineal Múltiple

La Regresión Lineal Múltiple (C. Montgomery, A. Peck, & Geoffrey Vining, 2012) es un modelo estadístico paramétrico que asume una relación lineal entre la variable respuesta y y un conjunto de predictores $X_1, X_2, \dots, X_p$. El modelo puede expresarse como:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

donde $\varepsilon \sim N(0, \sigma^2)$ representa el término de error.

Los parámetros β se estiman mediante Mínimos Cuadrados Ordinarios (OLS), mediante el problema de optimización:

$$\min_{\beta} (y - X\beta)'(y - X\beta).$$

La solución analítica del vector de coeficientes del modelo viene dada por:

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Para que las estimaciones obtenidas mediante OLS sean insesgadas y eficientes, el modelo se apoya en los siguientes supuestos:

- Linealidad: la relación entre la variable dependiente y los predictores es lineal.
- Independencia: las observaciones son independientes entre sí.
- Homocedasticidad: la varianza del término de error es constante.
- Normalidad de los errores: los residuos siguen una distribución aproximadamente normal.
- Ausencia de multicolinealidad severa entre las variables explicativas.

Cada coeficiente β_j representa el efecto marginal de la variable explicativa X_j sobre la variable dependiente, manteniendo constantes el resto de predictores. En el contexto de modelos con variable dependiente especificada en escala logarítmica, los coeficientes pueden interpretarse aproximadamente en términos de variaciones porcentuales del precio.

La regresión lineal múltiple se utiliza en este trabajo como modelo de referencia (*baseline*)**.** Su principal fortaleza reside en su elevada interpretabilidad (Dwivedi, y otros, 2023), que permite evaluar la coherencia económica de los coeficientes estimados y analizar el efecto marginal de cada característica del inmueble sobre el precio.

No obstante, este modelo presenta limitaciones relevantes en el análisis del mercado inmobiliario, donde las relaciones entre variables pueden ser no lineales y existir interacciones complejas (Wang, y otros, 2025), así como heterogeneidad espacial y valores atípicos. Estas limitaciones justifican la incorporación posterior de modelos más flexibles basados en árboles y métodos de ensemble, que permiten capturar estructuras más complejas en los datos.

2.2.3. Modelos de regresión regularizada

Los modelos de regresión regularizada constituyen una extensión natural de la regresión lineal clásica, diseñada para mejorar la estabilidad de las estimaciones en presencia de multicolinealidad o cuando el número de variables explicativas es elevado (Zou & Hastie, 2005). Estos métodos incorporan un término de penalización sobre la magnitud de los coeficientes, lo que permite controlar la complejidad del modelo, reducir la varianza de las estimaciones y, en determinados casos, realizar selección de variables mediante *shrinkage*.

Dentro de este marco general, el modelo *Elastic Net* ofrece una formulación unificada que combina dos tipos de penalización: la penalización L_1, asociada a la selección de variables, y la penalización L_2, orientada a estabilizar los coeficientes en contextos de alta correlación entre predictores.

Elastic Net

El modelo *Elastic Net* introduce simultáneamente una penalización basada en la norma absoluta (L_1) y otra basada en la norma cuadrática (L_2) de los coeficientes,

$$\min_{\beta} \ (y - X\beta)'(y - X\beta) + \lambda(\alpha \, \|\beta\|^2 + (1 - \alpha)\|\beta\|_1).$$

Esta combinación permite aprovechar las ventajas de ambos enfoques, resultando especialmente adecuado cuando existen grupos de variables altamente correlacionadas, ya que

favorece una selección más estable que la obtenida con *Lasso* y evita la dispersión excesiva de coeficientes característica de algunos escenarios de *Ridge*.

Desde un punto de vista conceptual, *Elastic Net* puede entenderse como un compromiso entre selección de variables y estabilidad del modelo, ofreciendo un marco flexible que engloba a otros modelos regularizados como casos particulares.

***Ridge Regression* (regularización L_2)**

La regresión *Ridge* se obtiene como un caso particular del modelo *Elastic Net* cuando la penalización L_1 es nula (es decir, cuando $\alpha = 1$), quedando únicamente la penalización L_2. En este contexto, los coeficientes se reducen progresivamente hacia cero sin llegar a anularse exactamente, lo que permite mitigar los efectos de la multicolinealidad y mejorar la estabilidad numérica del modelo.

Ridge resulta especialmente útil cuando todas las variables explicativas aportan información relevante, pero presentan relaciones de dependencia entre sí.

***Lasso Regression* (regularización L_1)**

Por su parte, la regresión *Lasso* surge cuando la penalización L_2 es nula (es decir, cuando $\alpha = 0$), manteniéndose únicamente la penalización L_1. Esta formulación induce que algunos coeficientes sean exactamente cero, lo que implica una selección automática de variables y da lugar a modelos más parsimoniosos.

No obstante, en situaciones con fuerte correlación entre predictores, *Lasso* tiende a seleccionar una única variable representativa dentro de cada grupo correlacionado, lo que puede generar cierta inestabilidad en la selección.

2.2.4. Árbol de Decisión

Los árboles de decisión (Song & Lu, 2015) son modelos predictivos no paramétricos que permiten capturar relaciones no lineales e interacciones complejas entre las variables explicativas sin imponer una forma funcional predefinida. En problemas de regresión, los árboles aproximan la relación entre predictores y variable respuesta mediante particiones recursivas del espacio de variables.

Un árbol de decisión divide el espacio de características $\mathcal{X} \subset \mathbb{R}^p$ en un conjunto finito de regiones disjuntas $R_1, R_2, \dots, R_M$. En cada región R_m, la predicción se define como una constante:

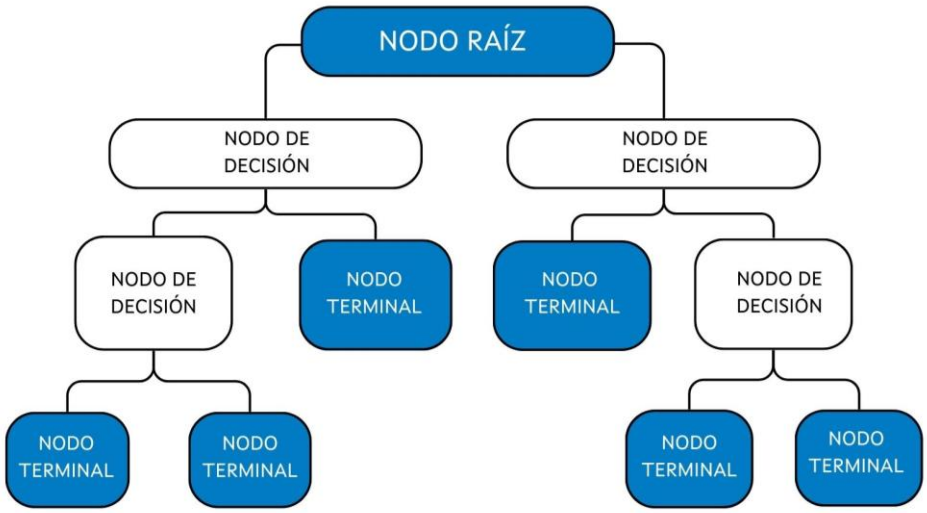
$$\hat{y}(x) = \sum_{m=1}^{M} c_m \, \mathbb{I}(x \in R_m),$$

donde c_m representa el valor predicho en la región R_m, que en regresión suele corresponder a la media de la variable respuesta de las observaciones que caen en dicha región e $\mathbb{I}(\cdot)$ denota a la función indicador.

La construcción del árbol se realiza mediante un procedimiento recursivo y jerárquico que, partiendo del nodo raíz y continuando a través de los sucesivos nodos de decisión (véase la

Figura 1), selecciona en cada nodo la variable explicativa y el punto de partición que optimizan la función de pérdida escogida.

Figura 1. Esquema general de un árbol de decisión.



Esquema general de un árbol de decisión. El proceso comienza con la partición del espacio de características en el nodo raíz, a partir del cual se realizan divisiones recursivas mediante nodos de decisión basadas en reglas de partición. Este procedimiento conduce a la definición de regiones disjuntas del espacio de características, asociadas a los nodos terminales, donde se obtiene la predicción final del modelo.

En el caso de regresión, el criterio habitual es la minimización del error cuadrático medio (MSE) dentro de los nodos:

$$\text{MSE} = \sum_{i \in R_m} (y_i - \bar{y}_{R_m})^2,$$

donde $\bar{y}_{R_m}$ es la media de la variable respuesta en la región R_m.

En cada paso, el algoritmo evalúa todas las posibles divisiones de todas las variables explicativas y selecciona aquella que maximiza la reducción de la suma de errores cuadrados total:

$$\Delta\text{MSE} = \text{MSE}_{\text{parent}} - (\text{MSE}_{\text{left}} + \text{MSE}_{\text{right}}).$$

Este proceso se repite de manera recursiva hasta alcanzar un criterio de parada. Así, el crecimiento del árbol puede detenerse en función de distintos parámetros, como:

- Profundidad máxima del árbol.
- Número mínimo de observaciones en un nodo.
- Reducción mínima del error requerida para realizar una partición.

Estos criterios controlan la complejidad del modelo y determinan el número de regiones M. Un árbol muy profundo presenta una gran capacidad de ajuste a los datos de entrenamiento, pero puede generalizar mal a nuevas observaciones, conduciendo a problemas de sobreajuste.

Desde el punto de vista estadístico, los árboles de decisión presentan bajo sesgo y alta varianza. Su flexibilidad les permite adaptarse con precisión a la estructura de los datos, pero pequeñas perturbaciones en el conjunto de entrenamiento pueden producir árboles sustancialmente distintos. Este comportamiento los hace especialmente propensos al sobreajuste, especialmente cuando se permite un crecimiento excesivo del árbol.

Una de las principales ventajas de los árboles de decisión es su interpretabilidad, ya que el proceso de predicción puede representarse mediante reglas del tipo "si–entonces", fácilmente comprensibles desde un punto de vista conceptual. Además, los árboles son invariantes a transformaciones monótonas de las variables y no requieren supuestos de linealidad ni normalidad de los errores.

No obstante, esta interpretabilidad se ve comprometida cuando el árbol alcanza gran profundidad, y su elevada varianza limita su capacidad de generalización. Estas limitaciones justifican el uso de los árboles de decisión como modelos exploratorios y como componentes básicos de métodos de *ensemble* más robustos, como *Random Forest* y *Gradient Boosting*, que se introducen en secciones posteriores.

2.2.5. Random Forest

El modelo *Random Forest* (Biau & Scornet, 2016) es un método de *ensemble* basado en árboles de decisión que tiene como objetivo principal mejorar la capacidad de generalización de los árboles individuales mediante la reducción de su varianza. Este enfoque fue propuesto como una extensión del método de *bagging* (*Bootstrap Aggregating*) y se ha consolidado como una de las técnicas más robustas para problemas de regresión con relaciones no lineales y alta complejidad estructural.

Un *Random Forest* consiste en un conjunto de B árboles de decisión independientes, entrenados sobre distintas submuestras del conjunto de datos original. Cada árbol se construye siguiendo dos principios fundamentales:

- Remuestreo con reemplazo (*bootstrap*): Para cada árbol $b = 1, \dots, B$, se genera una muestra bootstrap $\mathcal{D}_b$ del conjunto de entrenamiento original.
- Selección aleatoria de predictores: En cada nodo del árbol, el conjunto de variables candidatas para realizar la partición se restringe a un subconjunto aleatorio de tamaño $m \ll p$, donde p es el número total de predictores.

Estas dos fuentes de aleatoriedad introducen diversidad entre los árboles, reduciendo la correlación entre ellos.

En problemas de regresión, la predicción final del *Random Forest* se obtiene como la media de las predicciones de los árboles individuales:

$$\hat{y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b\,(x),$$

donde $\hat{y}_b(x)$ representa la predicción del árbol b para una observación con características x.

Desde el punto de vista estadístico, el *Random Forest* puede interpretarse como un estimador que reduce la varianza de los árboles individuales mediante promediado. Si los árboles tienen varianza σ^2 y correlación ρ, la varianza aproximada del estimador *ensemble* viene dada por:

$$\mathrm{Var}(\hat{y}_{RF}) \approx \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

Esta expresión refleja que, incluso con un número elevado de árboles, la reducción de varianza depende críticamente de la correlación entre ellos. La selección aleatoria de predictores en cada nodo resulta clave para disminuir dicha correlación y mejorar la estabilidad del modelo.

Una ventaja adicional del *Random Forest* es la posibilidad de evaluar la importancia relativa de las variables explicativas (Genuer, Poggi, & Taleau-Malot, 2010). Esta se mide habitualmente mediante la reducción media del error (o de la impureza) asociada a cada predictor a lo largo de todos los árboles del *ensemble*. Esta medida proporciona una aproximación global al peso de cada variable en el proceso de predicción, aunque debe interpretarse con cautela en presencia de variables correlacionadas.

El *Random Forest* combina una elevada capacidad predictiva con una notable robustez frente al sobreajuste, especialmente en comparación con los árboles de decisión individuales. Además, no requiere supuestos de linealidad ni normalidad y es capaz de capturar interacciones complejas entre variables.

No obstante, su principal limitación reside en la pérdida de interpretabilidad respecto a modelos más simples (Love, y otros, 2023), como la regresión lineal o los árboles individuales. Asimismo, aunque el sesgo del modelo suele ser bajo, el coste computacional puede ser elevado cuando se emplea un número grande de árboles o conjuntos de datos de gran tamaño (Oshiro, Perez, & Baranauskas, 2012).

En conjunto, el *Random Forest* ofrece un equilibrio favorable entre flexibilidad y estabilidad, lo que lo convierte en una herramienta especialmente adecuada para la modelización del precio de la vivienda en contextos caracterizados por relaciones no lineales y heterogeneidad estructural.

2.2.6. XGBoost

El modelo *XGBoost* (Chen & Guestrin, 2016) es una implementación avanzada de los métodos de *Gradient Boosting* basada en árboles de decisión. A diferencia de los enfoques basados en *bagging*, como *Random Forest*, *XGBoost* construye los modelos de forma secuencial, de modo que cada nuevo árbol se entrena para corregir los errores cometidos por el conjunto de árboles previamente ajustados. Este enfoque permite mejorar progresivamente el rendimiento predictivo, manteniendo un control explícito de la complejidad del modelo.

Sea $\hat{y}_i^{(t)}$ la predicción del modelo tras t iteraciones. En *XGBoost*, la predicción final se obtiene como la suma de las contribuciones de T árboles:

$$\hat{y}_i = \sum_{t=1}^{T} f_t\,(x_i), f_t \in \mathcal{F},$$

donde cada f_t representa un árbol de regresión perteneciente al espacio de funciones $\mathcal{F}$.

El modelo *XGBoost* se basa en la minimización de una función objetivo que combina una función de pérdida y un término de regularización que penaliza la complejidad del modelo:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t),$$

donde $\ell(\cdot)$ es una función de pérdida diferenciable (en este caso, el error cuadrático medio) y el término de regularización $\Omega(f_t)$ se define como:

$$\Omega(f_t) = \gamma\, K_t + \frac{1}{2}\lambda \sum_{j=1}^{K_t} w_{tj}^2\,,$$

siendo K_t el número de hojas del árbol t, w_{tj} los valores predichos en cada hoja, y γ y λ hiperparámetros de regularización. Este término penaliza tanto el número de hojas como la magnitud de las predicciones, reduciendo el riesgo de sobreajuste.

El entrenamiento del modelo se realiza de forma iterativa. En cada iteración t, se ajusta un nuevo árbol que aproxima el gradiente negativo de la función de pérdida evaluado en las predicciones actuales. Para ello, se utiliza una expansión de segundo orden de la función objetivo mediante un desarrollo de Taylor, lo que permite una optimización eficiente y estable. Este procedimiento hace que cada nuevo árbol se centre en las observaciones peor predichas en iteraciones anteriores, mejorando progresivamente el ajuste del modelo.

Desde el punto de vista estadístico, *XGBoost* reduce el sesgo del modelo mediante la combinación secuencial de árboles poco profundos, al tiempo que controla la varianza gracias a la regularización explícita incluida en la función objetivo. Este equilibrio entre flexibilidad y control de la complejidad permite a *XGBoost* capturar relaciones no lineales complejas sin incurrir en un sobreajuste excesivo.

Al igual que *Random Forest*, *XGBoost* proporciona medidas de importancia de las variables explicativas, basadas en la contribución de cada predictor a la reducción de la función de pérdida a lo largo del proceso de entrenamiento. Estas medidas reflejan la relevancia de las variables en un contexto no lineal y permiten comparar su papel relativo en la predicción del precio de la vivienda.

XGBoost destaca por su elevada capacidad predictiva y su buen comportamiento fuera de muestra, especialmente en problemas con alta heterogeneidad y relaciones no lineales complejas. La inclusión de regularización explícita y el entrenamiento secuencial permiten mejorar la estabilidad del modelo respecto a otros métodos basados en árboles. Como limitación principal, *XGBoost* presenta una menor interpretabilidad directa que modelos más simples y requiere una adecuada selección de hiperparámetros para obtener un rendimiento óptimo. Asimismo, su mayor complejidad computacional implica un mayor coste de entrenamiento.

En conjunto, *XGBoost* representa el modelo más flexible y potente considerado en este estudio, proporcionando un punto de comparación final frente a los modelos anteriores y permitiendo evaluar hasta qué punto la incorporación de técnicas avanzadas de *boosting* mejora la predicción del precio de la vivienda.

2.2.7. Support Vector Regression

El modelo *Support Vector Regression* (SVR) (Smola & Schölkopf, 2004) es una extensión de las Máquinas de Vectores Soporte al problema de regresión, cuyo objetivo es estimar una función $f(\mathbf{x})$ que aproxime la relación entre un conjunto de variables explicativas $\mathrm{x} \in \mathbb{R}^p$ y una variable respuesta continua y, controlando explícitamente la complejidad del modelo para favorecer la generalización.

En su formulación básica, el SVR busca una función de la forma

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b,$$

donde w es el vector de coeficientes y b el término independiente, que minimiza la función objetivo:

$$\min_{\mathbf{w},b} \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} L_\varepsilon \left(y_i - f(\mathbf{x}_i)\right),$$

siendo $C > 0$ un hiperparámetro de regularización y $L_\varepsilon(\cdot)$ la función de pérdida ε-insensible, definida como

$$L_\varepsilon(u) = \begin{cases} 0, & \text{si } |u| \leq \varepsilon, \\ |u| - \varepsilon, & \text{si } |u| > \varepsilon. \end{cases}$$

Esta función de pérdida ignora errores pequeños, penalizando únicamente aquellos que superan el umbral ε, lo que permite obtener soluciones más robustas frente al ruido y evita un ajuste excesivamente preciso a todas las observaciones.

Para modelizar relaciones no lineales entre las variables explicativas y el precio de la vivienda, se emplea el truco del *kernel*, que permite expresar la función de predicción como

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b,$$

donde α_i, α_i^* son los multiplicadores de Lagrange asociados a cada observación y $K(\mathbf{x}_i, \mathbf{x})$ es una función *kernel*. En este trabajo se utiliza el *kernel* radial (RBF), definido como

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2\right),$$

donde el hiperparámetro $\gamma > 0$ controla la influencia de cada observación: valores elevados de γ producen funciones más locales y flexibles, mientras que valores pequeños generan superficies de regresión más suaves.

El comportamiento del SVR viene, por tanto, determinado principalmente por tres hiperparámetros:

- C: controla el compromiso entre el ajuste a los datos y la regularización del modelo. Valores elevados penalizan fuertemente los errores, aumentando la flexibilidad del modelo.
- ε: define la anchura del tubo de insensibilidad alrededor de la función de regresión.
- γ: regula la complejidad del modelo no lineal inducido por el *kernel* RBF.

Dado que el SVR es sensible a la escala de las variables, se debe aplicar un preprocesamiento previo de estandarización.

2.3. Entorno de trabajo

El desarrollo del presente Trabajo Fin de Grado se ha realizado utilizando el lenguaje de programación *Python*[3], debido a su amplia adopción en el ámbito del análisis estadístico, el aprendizaje automático y la visualización de datos, así como a la disponibilidad de librerías especializadas de código abierto.

El análisis de datos y la modelización predictiva se han llevado a cabo principalmente en el entorno *Google Colab*[4], que permite la ejecución de código en la nube, facilita la gestión de dependencias y proporciona recursos computacionales suficientes para el entrenamiento de modelos de aprendizaje automático. Este entorno ha permitido trabajar de forma reproducible y eficiente con conjuntos de datos de tamaño medio. Todos los códigos y datos utilizados se encuentran en un repositorio público de *GitHub*[5].

Para el desarrollo de la aplicación web interactiva[6] se ha utilizado *Visual Studio Code*[7] como entorno de desarrollo integrado, junto con la librería *Streamlit*[8], orientada a la creación de aplicaciones web para análisis de datos a partir de *Python*. Para poder publicar la web, se subió todo a *GitHub* y se conectó con *Streamlit*. Esta combinación ha permitido integrar de forma directa los modelos entrenados y las visualizaciones interactivas en una única aplicación funcional.

Entre las principales librerías empleadas a lo largo del trabajo destacan *pandas* y *numpy* para la manipulación y transformación de datos, *matplotlib*, *seaborn* y *plotly* para la visualización gráfica, y *scikit-learn* para la implementación y evaluación de los modelos predictivos. Para la representación cartográfica y el análisis espacial se han utilizado librerías específicas como *GeoPandas*, compatibles con los formatos geográficos oficiales empleados.

El uso de herramientas basadas en *software* libre y ampliamente documentadas garantiza la reproducibilidad del análisis y facilita la extensión futura del trabajo a nuevos conjuntos de datos o metodologías.

[3] http://www.python.org
[4] https://colab.research.google.com/
[5] https://github.com/ecelaya/TFG-Eloy-Celaya-Lopez
[6] https://tfg-eloy-celaya.streamlit.app/
[7] https://code.visualstudio.com/
[8] https://streamlit.io/

3. Casos de estudio

El presente estudio se apoya en tres fuentes principales de información que permiten analizar el mercado inmobiliario de la Comunidad de Madrid desde una perspectiva tanto histórica como microeconómica. Estas bases de datos se complementan entre sí, proporcionando una visión integral de la evolución temporal de los precios y de los factores estructurales que los determinan.

3.1. Ministerio de Vivienda y Agenda Urbana

La primera fuente corresponde a la Estadística del valor tasado de la vivienda libre y protegida, publicada trimestralmente por el Ministerio de Vivienda y Agenda Urbana[9] (MIVAU). Esta base de datos recoge el valor medio de tasación por metro cuadrado (€/m²) de las viviendas, tanto nuevas como usadas, en municipios de más de 25.000 habitantes. Las cifras se obtienen a partir de las tasaciones hipotecarias realizadas por sociedades de tasación homologadas por el Banco de España, lo que garantiza su fiabilidad y representatividad a nivel nacional.

En este trabajo se utiliza la serie correspondiente a la Comunidad de Madrid para el periodo 2005–2024, con el objetivo de analizar la evolución del precio de la vivienda y las diferencias entre municipios. Esta fuente ofrece un indicador de referencia para medir el comportamiento del mercado residencial a largo plazo. Se utilizará también, de la misma fuente, el conjunto de datos de valor tasado de vivienda libre, que contiene las medias trimestrales a nivel nacional y de provincia, de donde se utilizarán las medias de Madrid y España desde el año 1995 hasta la actualidad.

Ambas bases de datos de esta fuente tienen como única opción de descarga una hoja de cálculo enfocada en lo visual más que en la usabilidad, como podemos ver en la Figura 2.

Figura 2. Ejemplo de los datos del MIVAU en bruto en Excel.

Tabla 4.
Valor tasado medio de vivienda libre de los municipios mayores de 25.000 habitantes.

Tercer trimestre de 2025 (*)
(*) Para acceder a la información de años anteriores consulte las pestañas que se encuentran en la parte inferior de este excel
Unidad: euros/m²

Provincia	Municipio	Valor tasado de vivienda			Número de tasaciones		
		Hasta cinco años de antigüedad	Con más de cinco años de antigüedad	Total	Hasta cinco años de antigüedad	Con más de cinco años de antigüedad	Total
Almería	**Almería**	**2,109.5**	**1,563.2**	**1,581.1**	**105**	**552**	**657**
	Níjar	n.r	1,184.0	1,191.8	10	85	95
	Roquetas de Mar	1,635.3	1,569.1	1,569.8	40	420	460
	Ejido (El)	n.r	1,216.4	1,224.0	16	289	305
Cádiz	Algeciras	1,601.8	1,346.5	1,348.7	42	362	404
	Arcos de la frontera	n.r	1,108.8	1,111.0	4	95	99
	Cádiz	**3,087.9**	**2,968.2**	**2,971.0**	**39**	**288**	**327**
	Chiclana de la Frontera	2,109.9	2,019.0	2,020.7	26	314	340
	Jerez de la Frontera	2,086.3	1,608.9	1,619.7	80	646	726
	Línea de la Concepción (La)	n.r	1,445.8	1,446.6	6	200	206
	Puerto de Santa María	n.r	2,173.9	2,176.4	18	276	294
	Puerto Real	n.r	1,670.6	1,678.0	17	112	129
	Rota	n.r	2,856.7	2,846.9	5	105	110
	San Fernando	1,972.8	1,766.2	1,771.3	27	242	269
	Sanlúcar de Barrameda	n.r	1,768.4	1,772.2	12	190	202
	San Roque	2,985.3	2,264.0	2,295.8	24	142	166
Córdoba	**Córdoba**	**2,056.0**	**1,607.0**	**1,633.9**	**112**	**975**	**1,087**
	Lucena	n.r	1,039.0	1,044.5	8	85	93
	Puente Génil	n.r	914.9	915.9	5	61	66
Granada	Almuñecar	n.r	2,503.9	2,505.2	8	126	134
	Granada	**2,570.6**	**2,186.2**	**2,194.2**	**69**	**558**	**627**
	Motril	1,639.8	1,350.3	1,363.6	23	138	161
Huelva	Huelva	1,777.4	1,392.6	1,400.9	48	398	446

T1A2023 | T2A2023 | T3A2023 | T4A2023 | T1A2024 | T2A2024 | T3A2024 | T4A2024 | T1A2025

[9] https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=35000000

Para el *dataset* a nivel municipio, encontramos una hoja por cada trimestre, con una fila para cada municipio y una columna con el valor tasado y otra del número de tasaciones. Tras dejar solo los datos con el mismo formato en cada hoja, los datos se importaron a *Python*, donde con un bucle se unieron todas las hojas añadiendo la columna de "Periodo", con el año y trimestre correspondiente. Del periodo, con formato T1A2025, se creó una columna de "Fecha" y se eliminó la de "Provincia" ya que todos los municipios son de Madrid. Tras todo este proceso, este *dataframe* final contiene 2236 entradas de datos y 8 columnas. Se puede ver una muestra con una entrada en la Tabla 1.

Tabla 1. Ejemplo del dataset del MIVAU ya tratado.

Municipio	Valor_Tasado	Num_Tasaciones	Periodo	Trimestre	Año	Mes	Fecha
Alcorcón	2645	652	T1A2005	1	2005	3	2005-03-31
...	...	...	...	...	...	...	...

La Tabla 2 contiene datos a nivel Provincia, además de las medias nacionales para cada trimestre. En este caso, cada hoja contenía información correspondiente a tres anualidades, por lo que se procedió a su consolidación mediante Excel, seleccionando exclusivamente las observaciones que posteriormente serían objeto de análisis, correspondientes a España y a la Comunidad de Madrid. Posteriormente, del mismo modo que con el *dataframe* anterior, tras dar el formato correcto a cada columna se creó la columna "Fecha", lo que nos deja con un *dataframe* final de 244 entradas y 5 columnas, como se muestra en la Tabla 2.

Tabla 2. Ejemplo del dataset del MIVAU a nivel provincial.

Región	Año	Trimestre	Valor_Tasado	Fecha
España	1995	1	670.8	1995-03-31
...	...	...	...	...

3.2. Colegio de Registradores

Se utilizará también una base de datos con el precio medio declarado de vivienda (€/m^2) por distrito y barrio, con datos anuales desde el 2007 hasta el 2024, de viviendas nuevas y usadas en el Municipio de Madrid, a nivel de distrito y de barrio, y son publicados por el Ayuntamiento de Madrid en su Banco de Datos[10]. Estos precios son los precios reales de viviendas que constan en el Registro de la Propiedad.

El Banco de Datos del Ayuntamiento de Madrid ofrece una interfaz sencilla, donde una vez seleccionada la base de datos que se desea utilizar, permite seleccionar qué variables queremos incluir. Para este estudio, se seleccionaron los 21 distritos junto a los 131 barrios del municipio de Madrid, además de los 18 años disponibles (2007-2024) y el total de viviendas, sin diferenciar entre nuevas y usadas. Como métrica, se seleccionó únicamente el precio medio en euros por metro cuadrado, que es la variable que se estudiará posteriormente.

[10] https://servpub.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=0504020100060

Una vez descargado el conjunto de datos, se realizó una limpieza inicial mínima desde la propia plataforma para eliminar títulos y elementos no informativos, conservando únicamente los datos numéricos. El archivo resultante contiene 153 filas, correspondientes a la Ciudad de Madrid (*agregado*), los distritos y los barrios, y 20 columnas, incluyendo las variables identificativas (*distrito* y *barrio*) y una columna para cada año del periodo analizado. Cabe señalar que a nivel barrio solo encontramos datos a partir del 2016 y, para alguno de los barrios, ningún dato.

3.3. Centro Nacional de Información Geográfica

Para la representación cartográfica de los municipios de la Comunidad de Madrid se emplearon los límites administrativos oficiales proporcionados por el Centro Nacional de Información Geográfica[11] (CNIG), perteneciente al Instituto Geográfico Nacional (IGN). Estos datos, disponibles en formato vectorial y actualizados periódicamente, incluyen la delimitación precisa de los términos municipales, provinciales y autonómicos. Su utilización permitió generar un mapa georreferenciado fiable y coherente con la realidad administrativa española, garantizando así la correcta integración del análisis territorial y la visualización interactiva desarrollada en este Trabajo Fin de Grado.

3.4. Portal de viviendas

Para el desarrollo de los modelos predictivos, se ha utilizado una base de datos obtenida a través de la plataforma *Kaggle*[12] con información detallada de miles de inmuebles ofertados a través de plataformas de compraventa en Madrid en el año 2023. Se incluyen 17 variables incluyendo el precio de oferta además de diversas variables con características físicas, económicas o de ubicación. La Tabla 3 contiene una muestra del dataset.

Tabla 3. Ejemplo del dataset de la plataforma Kaggle para la modelización.

Address	ZipCode	Latitude	Longitude	Price	Date	Rooms	Bathrooms	Surface
Retiro	28007	40.40258	-3.672911	445000	07/06/2023	2	2	102
...	...	...	...	...	...	...	...	...

Floor	Elevator	Air_Conditioner	Heater	Parking	Balcony	Terrace	Swimming_Pool
0	1	1	1	0	0	0	1
...	...	...	...	...	...	...	...

[11] https://centrodedescargas.cnig.es/CentroDescargas/limites-municipales-provinciales-autonomicos
[12] https://www.kaggle.com/datasets/alefernandezarmas/madrid-real-state-prices?resource=download

4. Análisis descriptivos y resultados predictivos

En esta sección se presentan los resultados empíricos del estudio, abarcando desde el preprocesamiento de los datos y el análisis exploratorio, tanto histórico como del mercado actual, hasta la evaluación de los distintos modelos de predicción considerados. El análisis se centra en examinar el comportamiento de los datos, identificar patrones relevantes y comparar el rendimiento de los enfoques de modelización bajo diferentes decisiones de especificación, con el objetivo de obtener una visión integral del problema de predicción del precio de la vivienda.

4.1. Preprocesamiento de datos

El preprocesamiento constituye una etapa fundamental en cualquier análisis estadístico, ya que garantiza que los datos utilizados son fiables y los adecua para su posterior uso en el análisis. Esta fase incluye diversos procedimientos que tienen como objetivo corregir inconsistencias, homogeneizar formatos y construir nuevas variables que faciliten el estudio posterior. A través del preprocesamiento se integran distintas fuentes de datos, se identifican valores atípicos o faltantes (nulos) y se transforma la información. El objetivo es obtener un conjunto de datos limpio y estructurado, estableciendo una base sólida sobre la que desarrollar la exploración inicial y posteriores técnicas que se apliquen.

4.1.1. Ministerio de Vivienda y Agenda Urbana

En el *dataset* de los datos de los municipios encontramos 2236 entradas y 8 columnas, y ningún valor nulo. Cabe destacar que para el municipio de Navalcarnero solo existen datos desde el 2020, probablemente porque antes no superaba el umbral de 25000 habitantes. Debido a uniones con datos oficiales para distintas visualizaciones, se procedió a comprobar que todos los municipios tuvieran nombres oficiales y hubo que cambiar 'Rozas de Madrid (Las)' por 'Las Rozas de Madrid'. A partir de este *dataset*, se creó uno nuevo agrupando por año y calculando la media del valor tasado. Esto será utilizado para la realización de mapas interactivos, donde tener 4 mapas distintos para cada año aumentaría exponencialmente los recursos y tiempo de carga necesarios, cuando los cambios trimestrales no son diferenciales. De esta forma los mapas mostrarán para cada año la media del valor de sus 4 trimestres, minimizando los recursos, pero manteniendo un resultado completo.

La tabla a nivel provincia no presentaba valores faltantes ni incongruencias tras ser importada a *Python*. Esta tabla dispone de 244 entradas y 5 columnas.

Aunque parte del análisis será realizado con estas dos tablas por separado, para un análisis más profundo y visualizaciones más completas, se procedió también a unir ambas tablas. Para ello, se llevó a cabo una unión mediante las claves "*Año"* y "*Trimestre",* obteniendo para cada municipio los valores medios de España y Madrid, que servirán como referencia para la elaboración de visualizaciones. Con esto, se obtuvieron las diferencias absolutas y porcentuales de cada municipio con ambas regiones de referencia, que servirán para profundizar más en el análisis.

4.1.2. Colegio de Registradores

Tras la importación del conjunto de datos a *Python* y ver que está formado por 153 filas (una por distrito y barrio) y 20 columnas, se realizó una primera inspección exploratoria para detectar posibles inconsistencias. Se observó que los valores numéricos se encontraban codificados como texto, utilizando el carácter "." como separador de miles y la "," como separador decimal, lo que impedía su correcta interpretación numérica. Asimismo, los valores faltantes aparecían representados mediante el carácter "-". Estos valores se reemplazaron por valores ausentes (*NaN*) y las columnas correspondientes a los años se convirtieron a formato numérico.

Dado que el análisis posterior se realiza a nivel de distrito y que para los barrios únicamente se dispone de información a partir de 2016, se filtró el conjunto de datos para eliminar todas las observaciones a nivel de barrio. No obstante, se mantuvo la observación correspondiente a la Ciudad de Madrid como referencia agregada, con el objetivo de disponer de un punto de comparación global frente a la evolución de los distintos distritos.

Finalmente, con el objetivo de facilitar el análisis temporal y la posterior representación cartográfica, el conjunto de datos se transformó de formato ancho (*wide format*), con una columna por año, a formato largo (*long format*), generando una estructura con una columna de año y una columna de precio. Esta transformación permite un tratamiento más flexible de la información y resulta especialmente adecuada para la visualización de series temporales y mapas por periodo. La Tabla 4 contiene una muestra del *dataset* tratado.

Tabla 4. Ejemplo del dataset del Colegio de Registradores.

Distrito	**Año**	**€/m²**
Centro	2007	4798.92
...	...	...

4.1.3. Centro Nacional de Información Geográfica

Para poder representar adecuadamente la información territorial relativa a los precios de la vivienda en la Comunidad de Madrid resulta imprescindible disponer de una limitación geográfica precisa al nivel deseado, en este caso a nivel municipio.

Los límites municipales utilizados en este Trabajo Fin de Grado proceden del Centro de Descargas del Centro Nacional de Información Geográfica (CNIG), donde se distribuyen en formato *shapefile* distintos ficheros que conforman la delimitación oficial de los municipios españoles. Sin embargo, este formato presenta un nivel de detalle elevado y contiene múltiples archivos asociados (.shp, .shx, .dbf, .prj, entre otros), por lo que fue necesario llevar a cabo un preprocesamiento para adaptar los datos a los requisitos de la visualización interactiva en *Streamlit* y *Plotly*.

En primer lugar, se cargó el *shapefile* mediante la librería *GeoPandas*, lo que permitió unificar todos los componentes del fichero y manipular tanto los atributos como la geometría de los polígonos. Posteriormente, se filtraron los registros para conservar únicamente los municipios pertenecientes a la Comunidad de Madrid, reduciendo así el tamaño del conjunto de datos y focalizando el análisis. A continuación, las geometrías se simplificaron empleando algoritmos de

reducción de complejidad geométrica (*topology-preserving simplification*), con el objetivo de disminuir el peso del archivo sin comprometer la forma general de los municipios. Esta simplificación es fundamental para asegurar un rendimiento óptimo en aplicaciones web interactivas, donde la carga de polígonos extremadamente detallados puede generar tiempos de renderizado elevados.

Finalmente, el conjunto de datos se exportó a formato *GeoJSON*[13], un estándar ampliamente utilizado en visualización web y totalmente compatible con librerías como *Plotly*. Este formato permitió integrar el mapa de la Comunidad de Madrid en la aplicación de *Streamlit*, vinculando cada municipio con las series temporales de precios de vivienda mediante un identificador común. Gracias a este preprocesamiento, los datos geográficos del CNIG pudieron incorporarse de manera eficiente y precisa para su utilización en mapas avanzados para visualizar los datos.

4.1.4. Portal de viviendas

El preprocesamiento constituye una de las etapas más determinantes en cualquier proyecto de modelado predictivo, especialmente cuando se trabaja con datos reales del mercado inmobiliario. Los *datasets* procedentes de portales de vivienda, administraciones públicas o fuentes estadísticas presentan habitualmente problemas como valores ausentes, incoherencias, diferencias de formato, escalas heterogéneas o variables redundantes. Si estas cuestiones no se abordan adecuadamente, los modelos pueden aprender patrones espurios, generar sesgos o perder capacidad predictiva. Cabe destacar que, actualmente, muchos de los *datasets* que encontramos y podemos utilizar tienen la mayor parte del preprocesado de datos ya realizada, encontrando incluso una completa ausencia de valores nulos.

En este proyecto, el preprocesamiento cumple una doble función. Por un lado, garantiza la calidad y coherencia interna del conjunto de datos, transformando la información original en un formato estructurado y consistente para el modelado. Esto incluye la identificación y tratamiento de valores faltantes, la estandarización de variables numéricas, la codificación de variables categóricas y la detección de posibles *outliers* o registros anómalos. Por otro lado, permite adaptar el *dataset* a las exigencias de los algoritmos de *machine learning*, optimizando la representación de la información para maximizar el rendimiento de los modelos.

Además, el proceso de preprocesamiento posibilita realizar una primera valoración del valor predictivo de las características disponibles, facilitando decisiones clave como la selección de variables, la reducción de dimensionalidad o la creación de nuevas variables derivadas (*feature engineering*). Todo ello es especialmente relevante en el mercado inmobiliario, donde conceptos como la ubicación, el estado del inmueble o la antigüedad muestran relaciones no lineales con el precio.

En definitiva, un preprocesamiento riguroso no solo mejora el desempeño del modelo, sino que asegura que las conclusiones derivadas del análisis sean sólidas, interpretables y representativas de la dinámica real del mercado de vivienda en Madrid.

El primer paso del preprocesamiento es analizar detenidamente la base de datos con la que nos encontramos. En este caso, el *dataset* de viviendas tiene 14130 entradas de datos y 16 columnas, incluyendo numéricas, categóricas y booleanas. Una de estas variables contiene la

[13] https://geojson.org/

fecha de publicación del anuncio. Todos los registros corresponden al año 2023, por lo que esta variable no aporta información adicional al análisis y no se espera que la fecha tenga influencia en el rendimiento de los modelos. Por tanto, se procede a su eliminación. La variable *"Address"* contiene información textual no estructurada relativa a la dirección de la vivienda. Dado su elevado nivel de detalle y su alta cardinalidad, así como la disponibilidad de variables geográficas más adecuadas como la latitud, la longitud y el código postal, se opta por eliminar esta variable del análisis con el fin de evitar complejidad innecesaria en la modelización.

Aunque el código postal constituye una variable de localización potencialmente relevante, el análisis exploratorio reveló la presencia de numerosos valores inconsistentes que no corresponden al municipio de Madrid o nulos. No obstante, las coordenadas geográficas asociadas a dichas observaciones resultaron ser coherentes y válidas. Dado que la información espacial queda adecuadamente capturada mediante la latitud y la longitud, y con el objetivo de evitar introducir ruido y redundancia en la modelización, se opta por excluir la variable código postal del conjunto de datos.

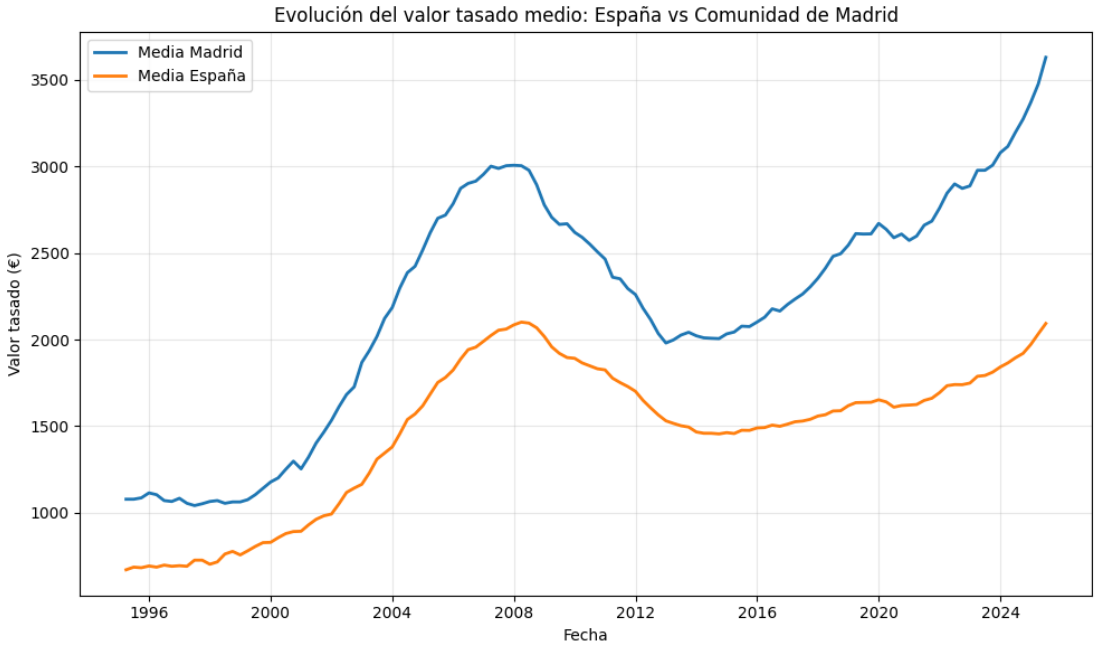
Tras esto, se procedió al análisis de valores nulos, uno de los pasos más importantes en el preprocesamiento de datos. Para facilitar este análisis, se creó una tabla representando el porcentaje de valores nulos de cada variable. Como se ha indicado anteriormente, este *dataset* no contiene valores nulos, por lo que no fue necesario ningún tipo de imputación.

4.2. ANÁLISIS HISTÓRICO

4.2.1. Análisis histórico de la Comunidad de Madrid

El análisis histórico permite comprender la evolución del mercado inmobiliario a lo largo del tiempo y contextualizar las diferencias actuales entre municipios. Mediante el estudio de las series temporales del valor tasado y su comparación con las medias de referencia de la Comunidad de Madrid y del conjunto nacional, es posible identificar patrones de crecimiento, ciclos económicos y episodios de divergencia entre zonas. Este análisis constituye un punto de partida fundamental, ya que ofrece una visión dinámica del comportamiento del mercado, permitiendo detectar qué municipios han experimentado una trayectoria sistemáticamente superior o inferior a la media, así como periodos en los que se han producido cambios estructurales relevantes.

Figura 3. Evolución del valor tasado medio en España y la Comunidad de Madrid (1995-2025).

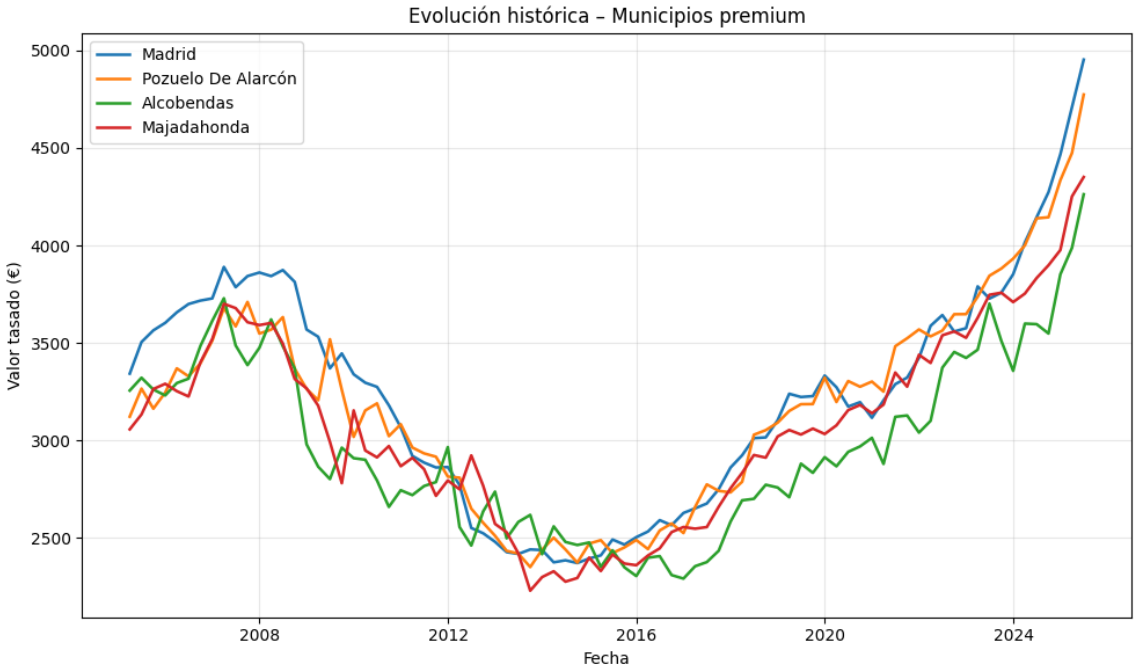


La evolución presentada en la Figura 3 revela que el valor tasado en la Comunidad de Madrid ha sido sistemáticamente superior al promedio nacional, manteniendo una brecha constante en todas las fases del ciclo inmobiliario. Ambas series muestran un fuerte crecimiento hasta 2008, seguido de una caída prolongada tras la crisis financiera, con un punto mínimo en torno a 2013–2014. Posteriormente, se observa una recuperación sostenida, especialmente marcada en Madrid desde 2020, donde el incremento de precios es notablemente más intenso que en el conjunto de España. Esta divergencia reciente refuerza el carácter especialmente dinámico y tensionado del mercado madrileño.

A continuación, nos centraremos en la evolución de algunos municipios de la Comunidad de Madrid agrupados por precio de la vivienda y renta.

En primer lugar, encontramos los municipios *premium* con los precios más elevados, como pueden ser Madrid, Pozuelo de Alarcón, Alcobendas y Majadahonda.

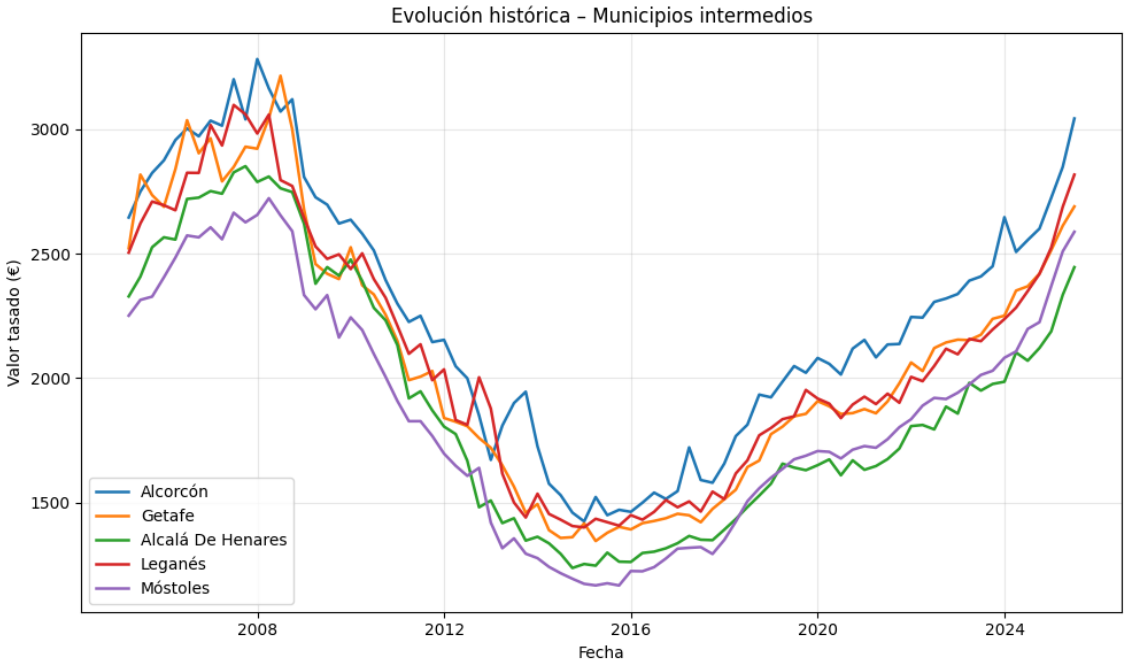
Figura 4. Evolución del Valor Tasado Medio en Municipios Premium (2005-2025).



Como podemos ver en la Figura 4, estos municipios presentan una evolución y niveles de precios muy similares a lo largo del tiempo. En 2008, justo *antes d*e la crisis financiera, el precio de la vivienda alcanzó valores cercanos a los 3800 €/m^2. Posteriormente, durante el periodo de crisis, se produjo un descenso significativo, situándose por debajo de los 2500 €/m^2 en torno a 2014. En la actualidad, los precios experimentan el mayor incremento de los últimos años, superando ampliamente los máximos alcanzados en 2008 y aproximándose a los 5000 €/m^2 en municipios como Madrid y Pozuelo.

La gran mayoría de municipios de la región los podemos situar como intermedios o de renta media. Ejemplos de este tipo de municipios son Alcorcón, Getafe, Leganés, Móstoles o Alcalá de Henares, situados hacia el sur y este de la Comunidad de Madrid.

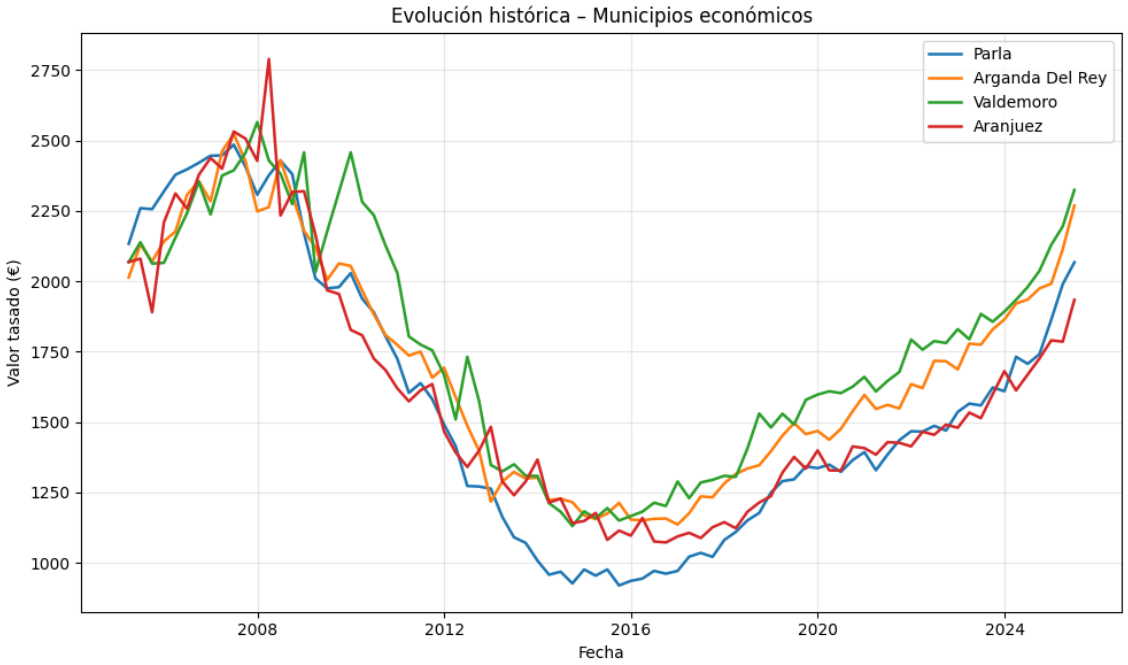
Figura 5. Evolución del Valor Tasado Medio en los Municipios de renta intermedia (2005-2025).



Del mismo modo que con los municipios *premium*, en la Figura 5 podemos ver una tendencia muy similar entre los municipios del mismo grupo. Observamos el mismo patrón que en la Figura 4, con un pico hacia el año 2008 de unos 3000 €/m^2, una drástica bajada con la crisis financiera hasta los 1000-1500 €/m^2 y por último un cambio de tendencia hacia el año 2016, donde los precios comienzan a subir de nuevo. La diferencia más significativa con respecto a los municipios anteriores es que, aunque los precios siguen subiendo, lo hacen de forma menos drástica. Podemos ver como en algunos de estos municipios el precio actual no supera al de 2008 todavía, rondando los 2500 €/m^2.

Por último, encontramos los municipios de renta más baja, con el precio de vivienda más económico y accesible. Estos municipios generalmente se encuentran más alejados del centro y tienen características que los hacen menos atractivos y populares. Encontramos municipios como Parla, Arganda del Rey, Valdemoro o Aranjuez.

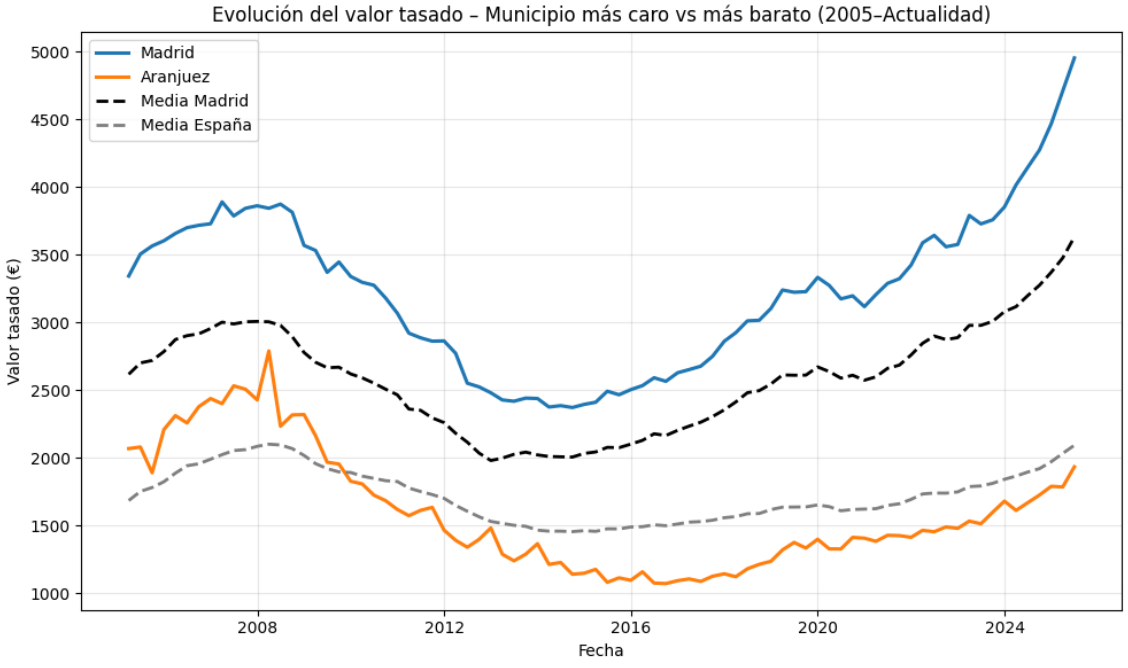
Figura 6. Evolución del Valor Tasado Medio en los Municipios económicos (2005-2025).



En la Figura 6 apreciamos una vez más un patrón parecido, aunque con precios más bajos, llegando hasta los 1000 €/m^2 en el punto más bajo de algunos de los municipios. Como ocurría con los municipios de renta intermedia, los precios actuales no superan los máximos de antes de la crisis, que se sitúan en torno a los 2500 €/m^2.

Por último, procedemos a comparar el municipio con el valor tasado más elevado, Madrid, con el que tiene el valor más bajo, Aranjuez, mostrando también las medias de la Comunidad de Madrid y de España.

Figura 7. Evolución del Valor Tasado Medio. Municipio más caro (Madrid) vs más barato (Aranjuez) (2005-2025).

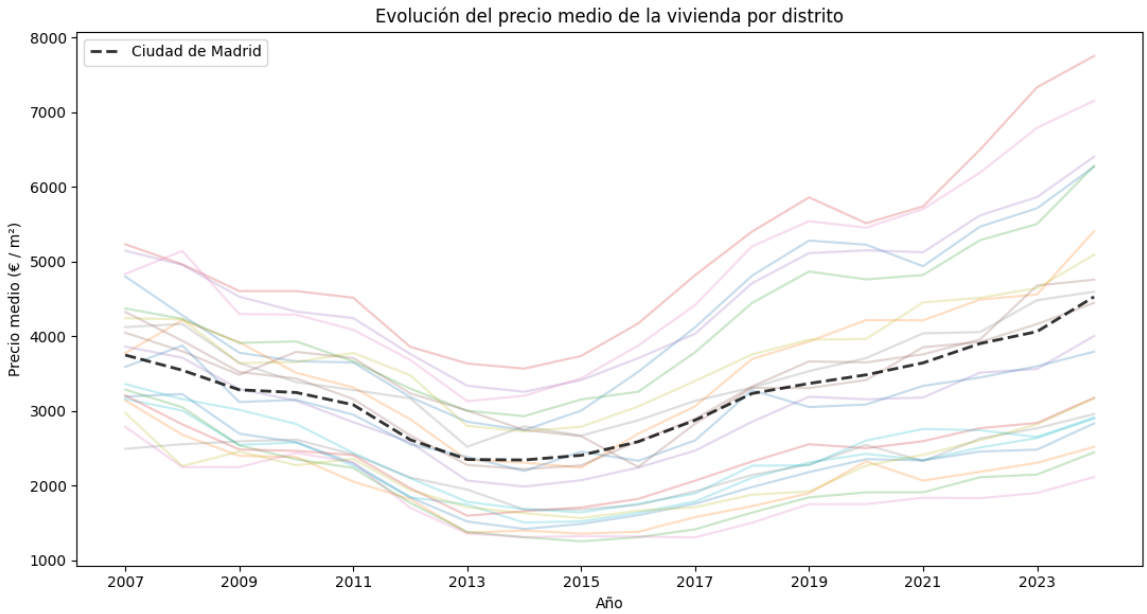


En la Figura 7 se observa claramente que el municipio de Madrid presenta, de forma consistente a lo largo del tiempo, los precios de la vivienda más elevados, seguido por la media de la Comunidad de Madrid. En el extremo inferior se sitúan Aranjuez y la media nacional, cuyas series se alternan a lo largo del periodo analizado. Aquí podemos apreciar como el precio actual de Madrid y la media de la Comunidad son mucho más altos que el máximo de antes de la crisis, y han sufrido un grave incremento en los últimos años. La media de España ha sufrido menos variaciones a lo largo del tiempo, aunque de nuevo vemos el efecto de la crisis financiera. Estas variaciones son más obvias en Aranjuez, que sufrió un mayor incremento y caída, aunque actualmente su valor es menor que el de 2008. Se aprecia una marcada disparidad en los precios de la vivienda en España, ya que el municipio de Madrid mantiene, de forma sostenida a lo largo del tiempo, niveles de precios superiores al doble de la media nacional.

4.2.2. Análisis histórico del municipio de Madrid

Como se ha mostrado anteriormente, el municipio de Madrid presenta, de forma consistente, los valores tasados más elevados de la Comunidad de Madrid. Este municipio, además de ser la capital de la Comunidad de Madrid, concentra cerca del 50% de la población de la región, lo que pone de manifiesto su relevancia. Es por ello que este municipio se va a analizar en profundidad, estudiando la evolución del precio medio de las viviendas de cada uno de sus distritos. Para ello, se va a utilizar el *dataset* obtenido anteriormente del portal del Ayuntamiento de Madrid junto con los datos del Colegio de Registradores.

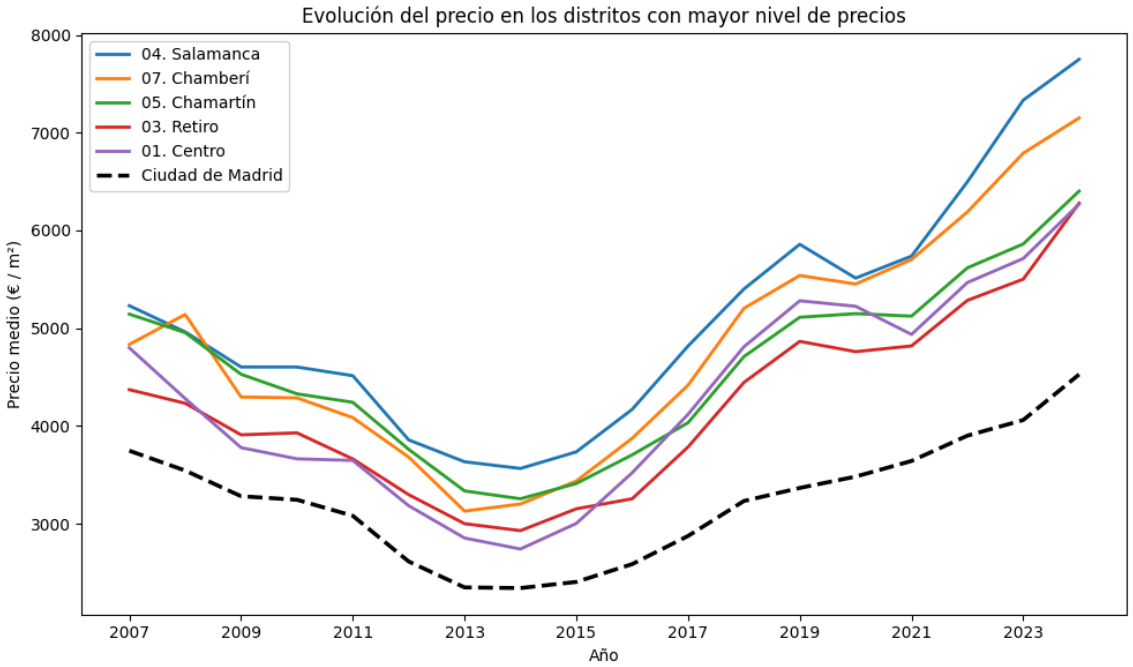
Figura 8. Evolución del precio medio de los distritos de la ciudad de Madrid vs la media (2007-2024).



La Figura 8 muestra la evolución del precio medio de la vivienda por metro cuadrado para los distintos distritos del municipio de Madrid durante el periodo 2007–2024. Como referencia agregada, se incluye la serie correspondiente a la Ciudad de Madrid, representada mediante una línea discontinua.

En la Figura 8 se observa un patrón temporal común a la mayoría de distritos, caracterizado por una fase de ajuste pronunciado entre 2008 y 2014, coincidente con la crisis financiera, seguida de una recuperación sostenida a partir de 2015. No obstante, la intensidad de estas dinámicas difiere notablemente entre distritos, lo que confirma una elevada heterogeneidad espacial dentro del municipio.

Figura 9. Evolución del precio medio en los distritos más caros y en la Ciudad de Madrid (2007-2024).

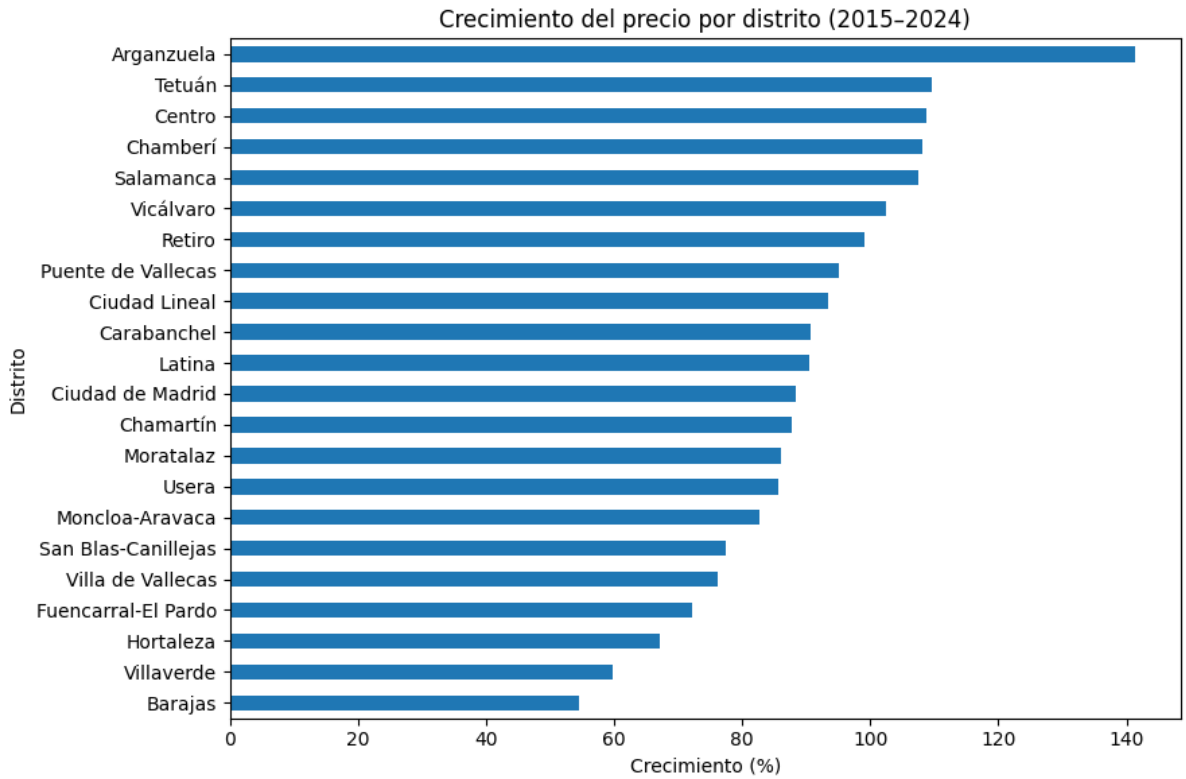


Con el objetivo de profundizar en la dinámica de los distritos con mayor nivel de precios, se analiza de forma específica la evolución temporal de los cinco distritos más caros en 2024, incluyendo la serie agregada de la Ciudad de Madrid como referencia.

La Figura 9 muestra que estos distritos presentan niveles de precios persistentemente superiores a la media municipal a lo largo de todo el periodo analizado. A pesar de las diferencias en magnitud, las trayectorias temporales son notablemente similares, con una caída pronunciada durante la crisis financiera y una recuperación intensa a partir de 2015.

Asimismo, se observa que la brecha entre estos distritos y la media de la ciudad se amplía en los últimos años, lo que sugiere una creciente segmentación del mercado inmobiliario urbano, especialmente en el segmento de mayor valor.

Figura 10. Crecimiento del precio por distrito, en porcentaje (2015-2024).



La Figura 10 muestra el crecimiento porcentual acumulado del precio medio de la vivienda entre 2015 y 2024 por distrito. Este análisis permite complementar la información sobre niveles de precios con una perspectiva dinámica centrada en la revalorización relativa.

Los resultados indican que los distritos con mayores niveles de precios no son necesariamente los que han experimentado los mayores crecimientos. Destacan distritos como Arganzuela, Tetuán o Centro, con incrementos superiores al 100%, lo que sugiere procesos intensos de revalorización durante la fase expansiva reciente del mercado.

Por el contrario, algunos distritos periféricos, como Barajas o Villaverde, presentan crecimientos más moderados, a pesar de partir de niveles de precio más bajos. La Ciudad de Madrid, considerada como conjunto, muestra un crecimiento intermedio, reflejando nuevamente el efecto de agregación de dinámicas heterogéneas.

Este resultado muestra que el análisis conjunto de niveles y crecimientos es fundamental para entender la evolución del mercado inmobiliario y anticipar posibles procesos de convergencia o divergencia espacial dentro del municipio.

4.3. ESTUDIO MICROECONÓMICO ACTUAL

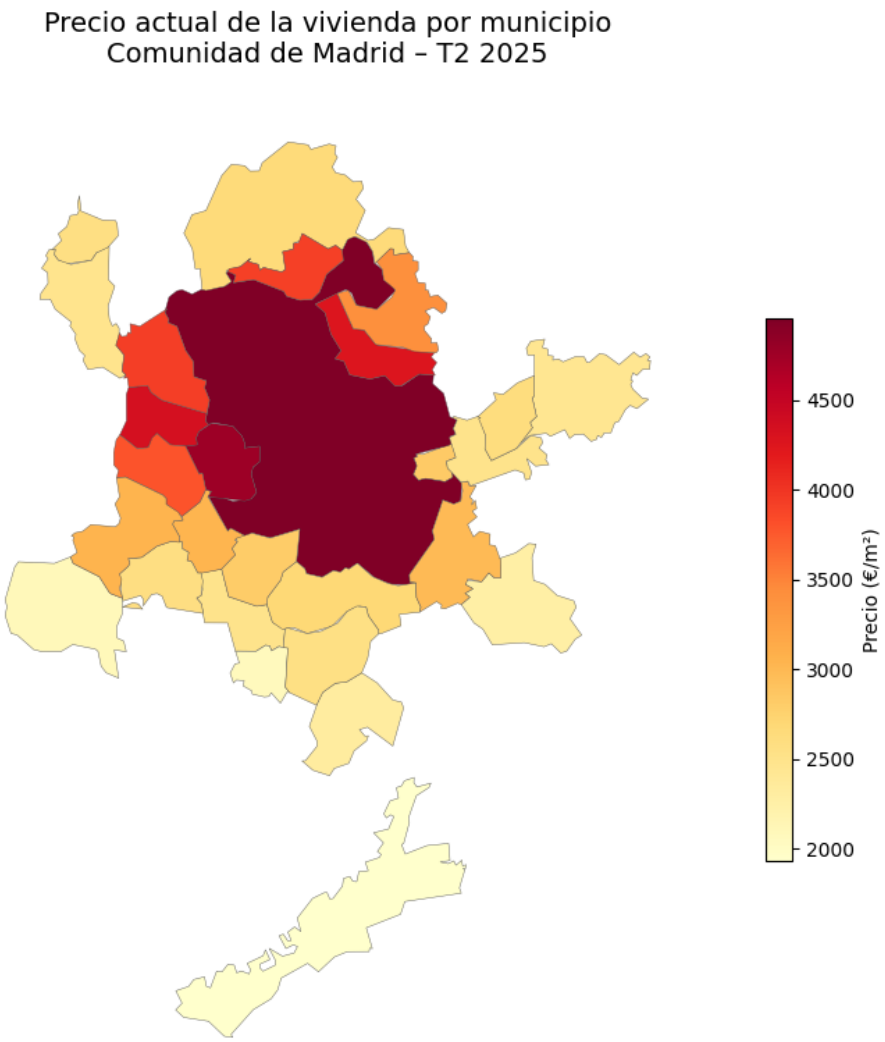
El análisis histórico desarrollado en la Sección 4.2 ha permitido contextualizar la evolución temporal del precio de la vivienda en la Comunidad y el municipio de Madrid. No obstante, para comprender adecuadamente la situación del mercado en el momento actual, resulta necesario realizar un estudio descriptivo centrado exclusivamente en el nivel y la distribución espacial de los precios en el presente.

Este apartado tiene como objetivo ofrecer una fotografía estática del mercado inmobiliario actual, analizando cómo se distribuyen los precios de la vivienda dentro de la Comunidad de Madrid y, de forma más detallada, dentro del propio municipio de Madrid. El enfoque adoptado es estrictamente descriptivo y territorial, permitiendo identificar patrones de segmentación y desigualdad espacial que caracterizan el mercado residencial madrileño en la actualidad.

4.3.1. Distribución actual del precio de la vivienda en la Comunidad de Madrid

La Figura 11 muestra la distribución espacial del precio medio de la vivienda por metro cuadrado en los municipios de la Comunidad de Madrid de más de 25000 habitantes en el segundo trimestre del año 2025. Para cada municipio se ha representado el valor medio anual del precio de la vivienda, con el fin de eliminar posibles fluctuaciones estacionales y facilitar la comparación territorial.

Figura 11. Mapa del precio actual de la vivienda por municipio en la Comunidad de Madrid



El mapa revela una marcada heterogeneidad espacial en los precios de la vivienda. Se observa una clara concentración de valores elevados en el municipio de Madrid y en varios municipios del eje noroeste, tradicionalmente asociados a niveles de renta más altos. Por el contrario, los

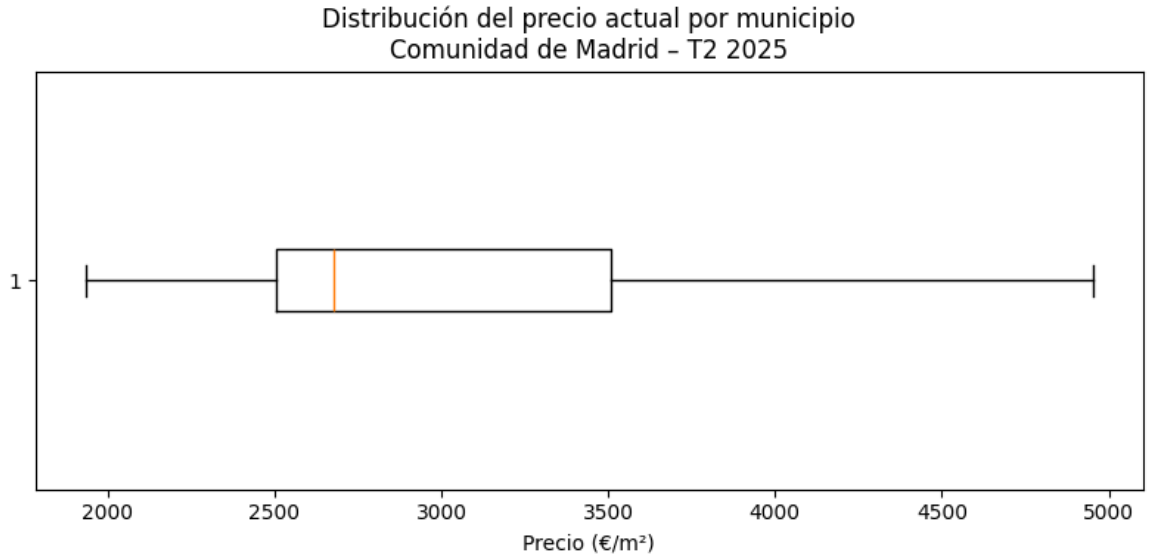
municipios situados en las zonas sur y sureste de la región presentan, en general, precios notablemente inferiores.

Este patrón evidencia la existencia de un mercado fuertemente segmentado a nivel territorial, donde la localización constituye un elemento clave en la determinación del nivel de precios, incluso cuando se analizan valores medios agregados.

4.3.2. Desigualdad territorial y dispersión de precios

La Figura 12 muestra la distribución del precio actual de la vivienda por metro cuadrado considerando el conjunto de municipios de la Comunidad de Madrid en el segundo trimestre de 2025. El gráfico evidencia una dispersión significativa de los precios, con municipios que se sitúan claramente por encima del valor central de la distribución.

Figura 12. Boxplot del precio actual por distrito en la Comunidad de Madrid.

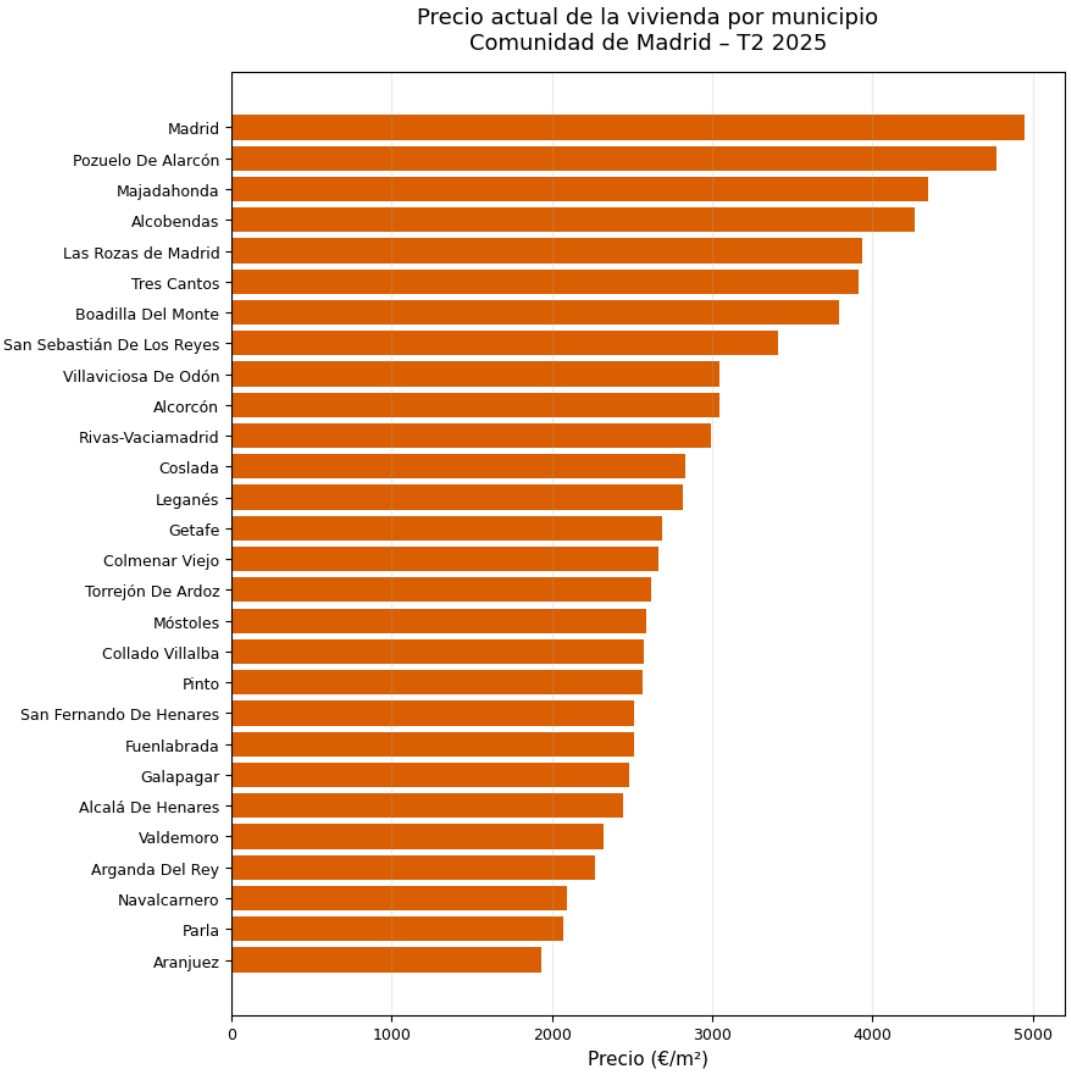


La amplitud observada recalca la existencia de fuertes desigualdades territoriales en el mercado inmobiliario regional, incluso dentro de una misma área metropolitana.

La Figura 13 muestra el *ranking* de los municipios por precio medio de la vivienda en el momento actual. Este *ranking* permite identificar de forma clara los extremos de la distribución y facilita la comparación directa entre municipios.

Los municipios con precios más elevados se concentran mayoritariamente en el entorno de la capital y en zonas tradicionalmente asociadas a un mayor poder adquisitivo. En contraste, los municipios con precios más bajos se localizan principalmente en áreas periféricas, lo que refuerza el carácter policéntrico y segmentado del mercado inmobiliario madrileño.

Figura 13. Precio actual de la vivienda por distrito en la Comunidad de Madrid.

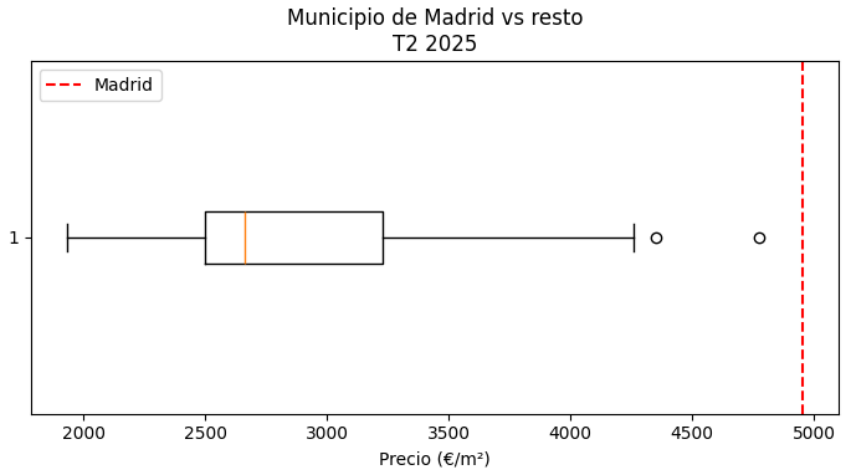


4.3.3. El municipio de Madrid en el contexto original

La Figura 14 compara el precio actual de la vivienda por metro cuadrado del municipio de Madrid con la distribución de precios del resto de municipios de la Comunidad en el segundo trimestre de 2025. El gráfico muestra que el precio de Madrid se sitúa claramente por encima del rango central de la distribución regional, superando ampliamente la mediana y el intervalo intercuartílico del resto de municipios.

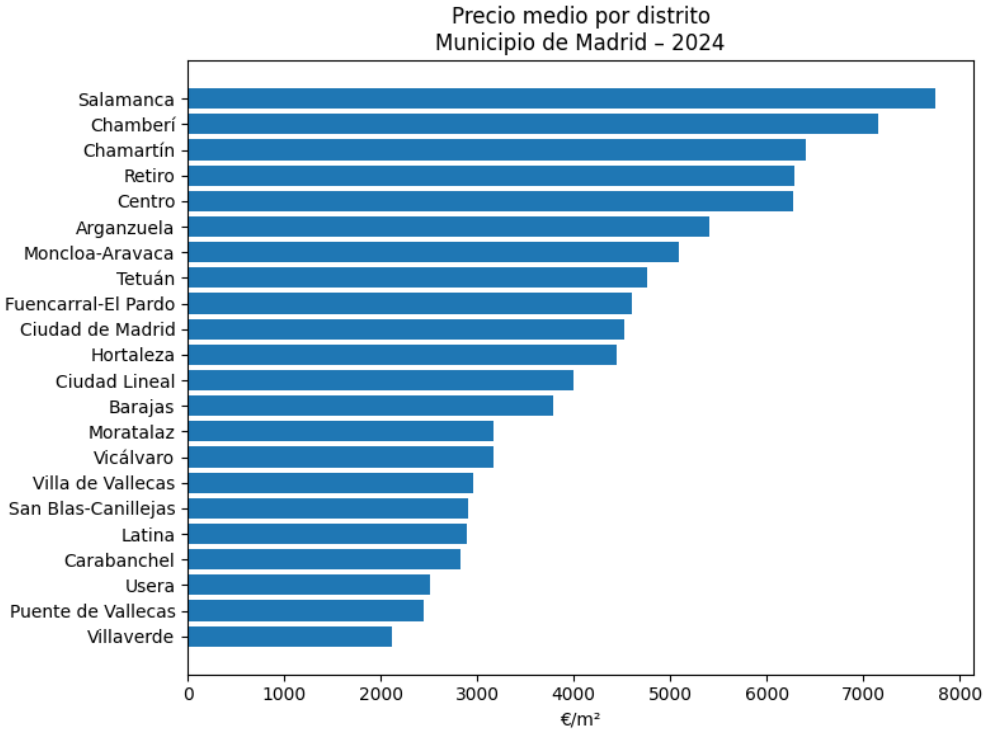
Este resultado indica que el municipio de Madrid constituye un submercado diferenciado dentro del conjunto regional, con un nivel de precios que no puede interpretarse como una simple prolongación del resto de municipios de la Comunidad.

Figura 14. Boxplot del Municipio de Madrid vs el resto de Municipios.



Para comprender mejor la distribución espacial de los precios de la vivienda, la Figura 15 muestra el precio medio por metro cuadrado en los distintos distritos del municipio de Madrid en el último año disponible. Los resultados evidencian una marcada heterogeneidad interna dentro del propio municipio, con diferencias de precio muy significativas entre distritos.

Figura 15. Precio medio por distrito en el Municipio de Madrid.



Los distritos de Salamanca, Chamberí y Chamartín se sitúan claramente en los niveles de precio más elevados, superando ampliamente la media municipal, lo que refleja su consolidación como áreas de alta demanda residencial. A continuación, distritos centrales como Retiro y Centro mantienen también precios elevados, reforzando la existencia de un núcleo urbano con valores inmobiliarios significativamente superiores al resto de la ciudad.

En contraste, los distritos periféricos, como Villaverde, Puente de Vallecas o Usera, presentan niveles de precio considerablemente más bajos, lo que manifiesta la persistencia de un gradiente centro–periferia en el mercado inmobiliario madrileño. Esta estructura confirma la coexistencia de submercados claramente diferenciados dentro del municipio, incluso cuando se analizan valores medios agregados.

4.3.4. Síntesis del estudio actual

El análisis descriptivo del mercado inmobiliario actual revela un alto grado de segmentación territorial, tanto a nivel regional como dentro del propio municipio de Madrid. Los precios de la vivienda muestran una elevada dispersión y una clara dependencia de la localización, reflejando la coexistencia de submercados con dinámicas y niveles de precio muy diferenciados.

Esta fotografía del mercado actual permite contextualizar los resultados históricos presentados previamente y establece una base sólida para los análisis posteriores, centrados en el estudio detallado de los determinantes del precio de la vivienda a nivel micro.

4.4. MODELIZACIÓN PREDICTIVA

4.4.1. Análisis exploratorio de los datos

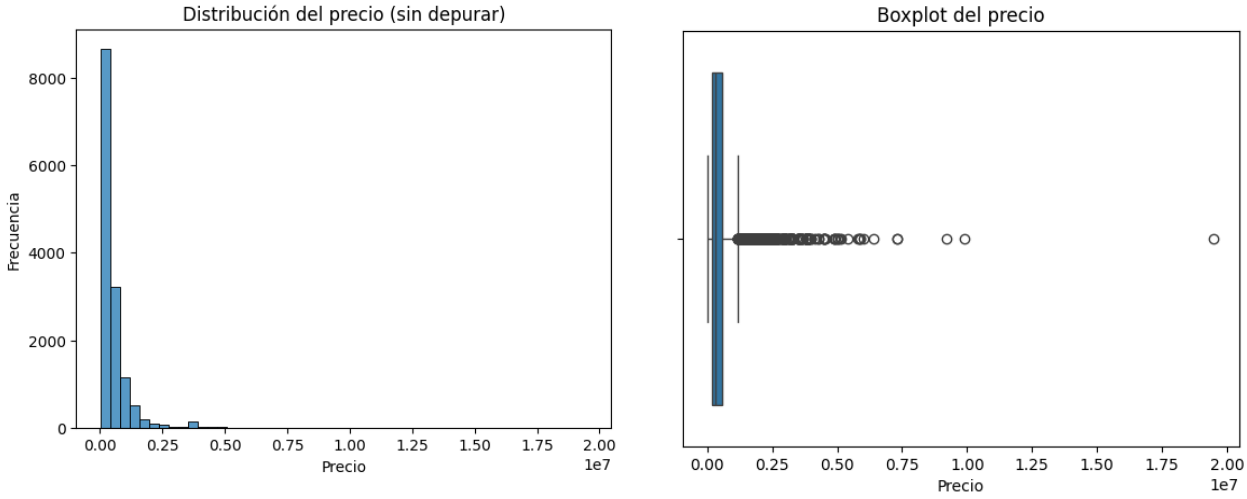
Análisis univariante

En este apartado se realiza un análisis exploratorio de los datos con el objetivo de comprender la estructura del conjunto de datos, evaluar la calidad de la información disponible y detectar patrones relevantes en las variables. Este análisis permite estudiar de forma descriptiva la distribución de las variables, identificar posibles valores atípicos y analizar las relaciones existentes entre las características de los inmuebles y su precio. Asimismo, el análisis exploratorio sirve como base para justificar las decisiones metodológicas adoptadas posteriormente, tales como la selección de variables, la aplicación de transformaciones y la definición del enfoque de modelización empleado en el estudio.

En primer lugar, se lleva a cabo un análisis exploratorio univariante con el objetivo de estudiar de forma individual la distribución de las variables numéricas incluidas en el conjunto de datos. En particular, se presta especial atención a la variable objetivo, correspondiente al precio de la vivienda, ya que constituye el principal objeto de interés del estudio. El análisis univariante permite identificar la forma de las distribuciones, evaluar la presencia de asimetrías y valores atípicos, y obtener una primera aproximación descriptiva que servirá de base para las transformaciones y decisiones metodológicas adoptadas en fases posteriores del trabajo.

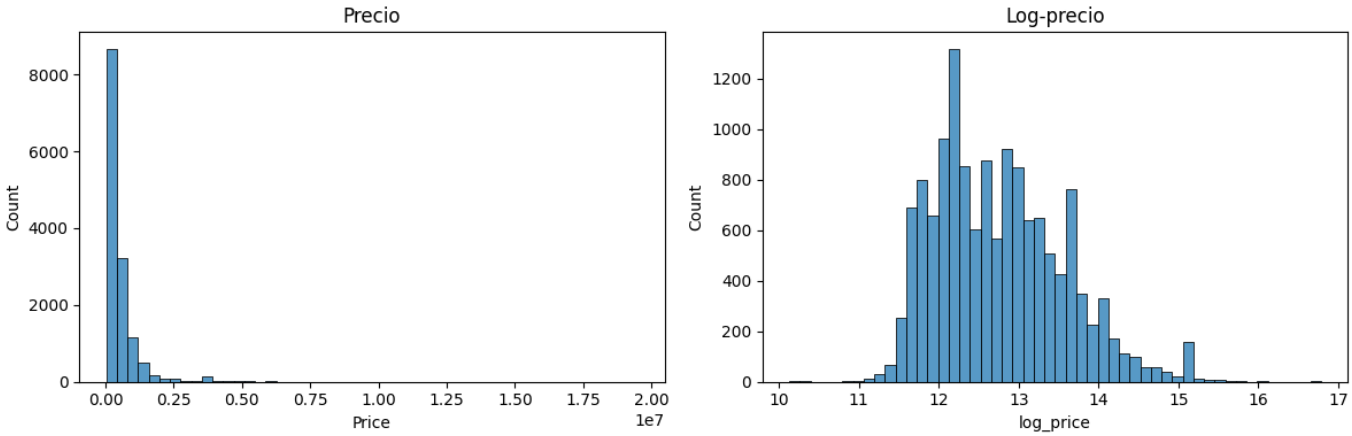
El análisis univariante inicial de la variable precio revela la existencia de un reducido número de observaciones con valor cero. Dichos valores se interpretan como errores de registro y no como transacciones válidas de mercado. En consecuencia, se procede a la eliminación de estas observaciones antes de continuar con el análisis.

Figura 16. Distribución de la variable precio.



El histograma y el boxplot de la variable precio (Figura 16) muestran una distribución claramente asimétrica a la derecha, con una elevada concentración de observaciones en los rangos de precios más bajos y una cola larga asociada a viviendas de alto valor. Asimismo, el boxplot identifica un número considerable de observaciones extremas, que no corresponden a errores de registro, sino a propiedades de elevado precio. Estas características justifican la necesidad de aplicar una transformación sobre la variable precio con el objetivo de mejorar su comportamiento estadístico antes de la fase de modelización.

Figura 17. Distribución de la variable precio antes y después de la transformación logarítmica.

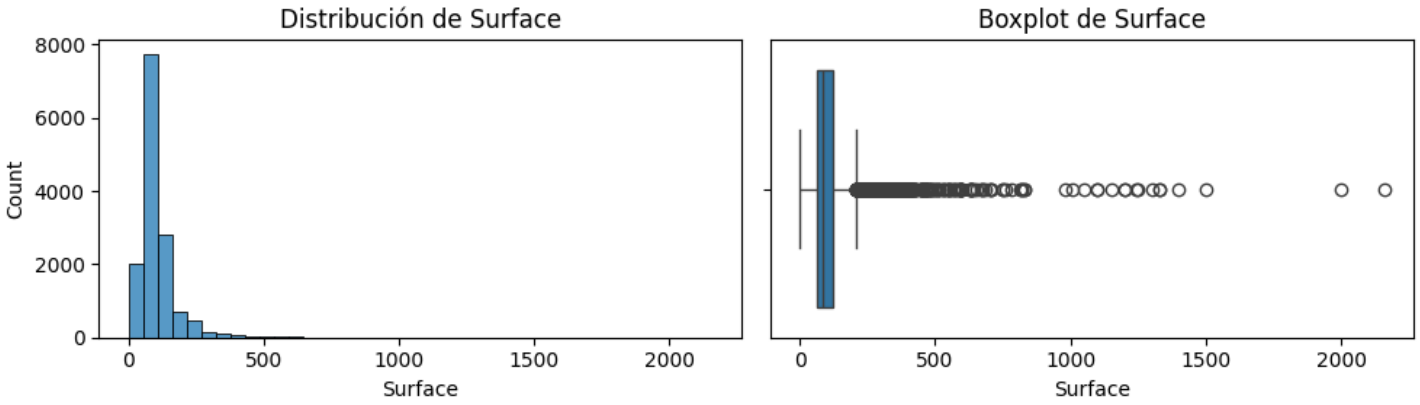


La Figura 17 compara la distribución del precio de la vivienda con la distribución de su transformación logarítmica. Mientras que el precio original (izquierda) presenta una marcada asimetría positiva y una elevada concentración de observaciones en los valores más bajos, la variable log-precio (derecha) muestra una distribución considerablemente más simétrica. Esta transformación reduce la influencia de las viviendas de precio muy elevado y permite aproximar mejor los supuestos de normalidad, por lo que se adopta el log-precio como variable objetivo en los modelos desarrollados en las secciones posteriores.

A continuación, se estudiarán las variables predictivas, comenzando por las numéricas.

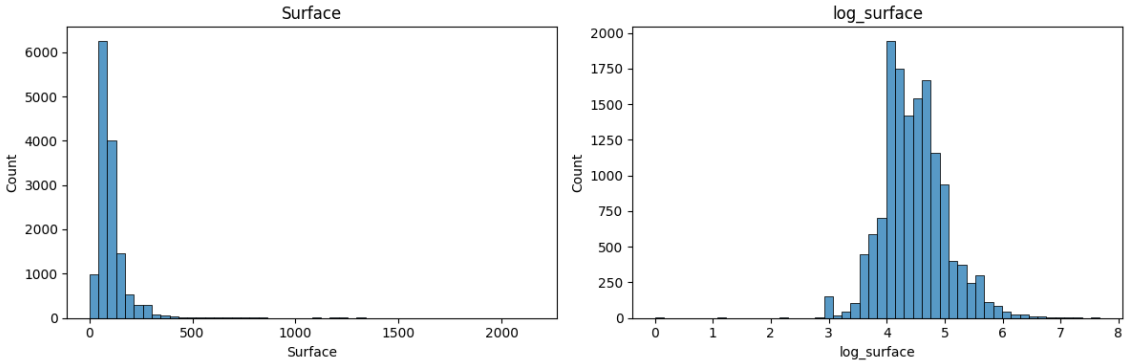
Del mismo modo que pasa con la variable a predecir, la variable “Surface” (metros cuadrados construidos de la vivienda) contiene algunos valores demasiado pequeños o con valor nulo, por lo que antes de analizarla se procede a eliminar estas entradas.

Figura 18. Distribución de la variable superficie.



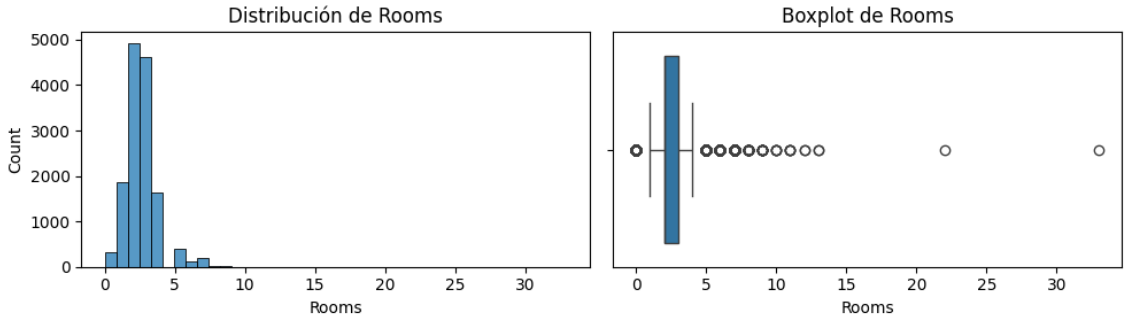
El análisis univariante de la superficie construida (Figura 18) muestra una distribución claramente asimétrica a la derecha, con una elevada concentración de viviendas de tamaño reducido y una cola larga asociada a propiedades de gran superficie. El *boxplot* identifica un número considerable de observaciones extremas, que no corresponden a errores de registro, sino a viviendas singulares. Con el objetivo de reducir la asimetría y limitar la influencia de estos valores extremos en la modelización, se propone la aplicación de una transformación logarítmica sobre la variable superficie (“Surface”).

Figura 19. Distribución de la variable superficie antes y después de la transformación logarítmica.



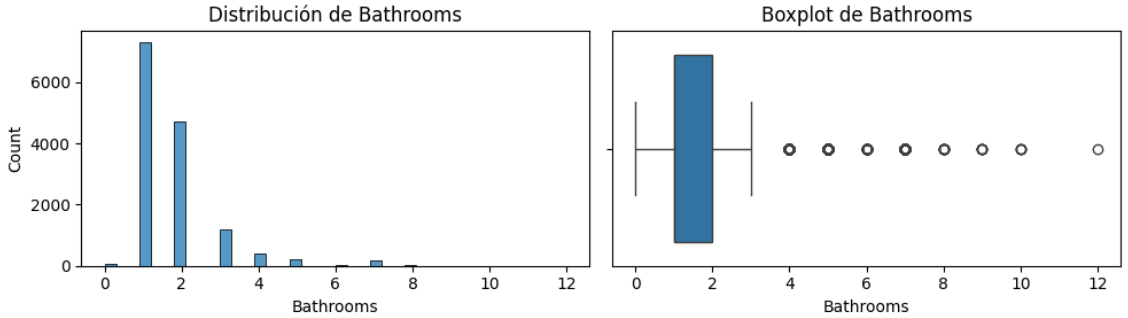
La Figura 19 compara la distribución de la superficie construida con la distribución de su transformación logarítmica. Mientras que la superficie original presenta una marcada asimetría positiva y una cola larga asociada a viviendas de gran tamaño, la variable *log-surface* muestra una distribución considerablemente más simétrica. Esta transformación permite reducir la influencia de valores extremos y facilita el cumplimiento de los supuestos del modelo, por lo que se adopta “*log-surface”* como variable explicativa en el análisis posterior.

Figura 20. Distribución de la variable número de habitaciones.



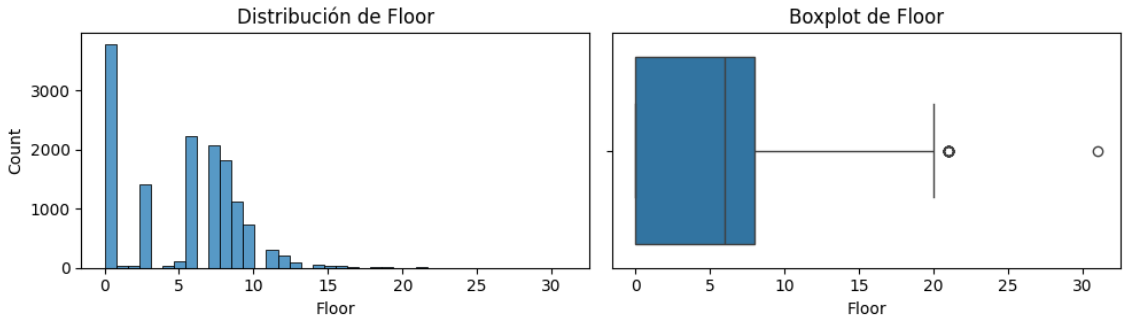
Por su parte, el análisis univariante del número de habitaciones (Figura 20) muestra una distribución discreta, con una clara concentración entre dos y cuatro habitaciones, siendo tres el valor más frecuente. El *boxplot* identifica un reducido número de observaciones extremas correspondientes a viviendas con un número elevado de habitaciones. Dado que estos valores no constituyen errores de registro, sino propiedades singulares, la variable se mantiene sin aplicar transformaciones adicionales.

Figura 21. Distribución de la variable número de baños.



Continuando con el análisis univariante del número de baños en la Figura 21, esta muestra una distribución discreta, con una elevada concentración de viviendas con uno o dos baños. El *boxplot* identifica un reducido número de observaciones extremas correspondientes a viviendas con un mayor número de baños, que no constituyen errores de registro. En consecuencia, la variable se mantiene sin aplicar transformaciones adicionales.

Figura 22. Distribución de la variable planta.



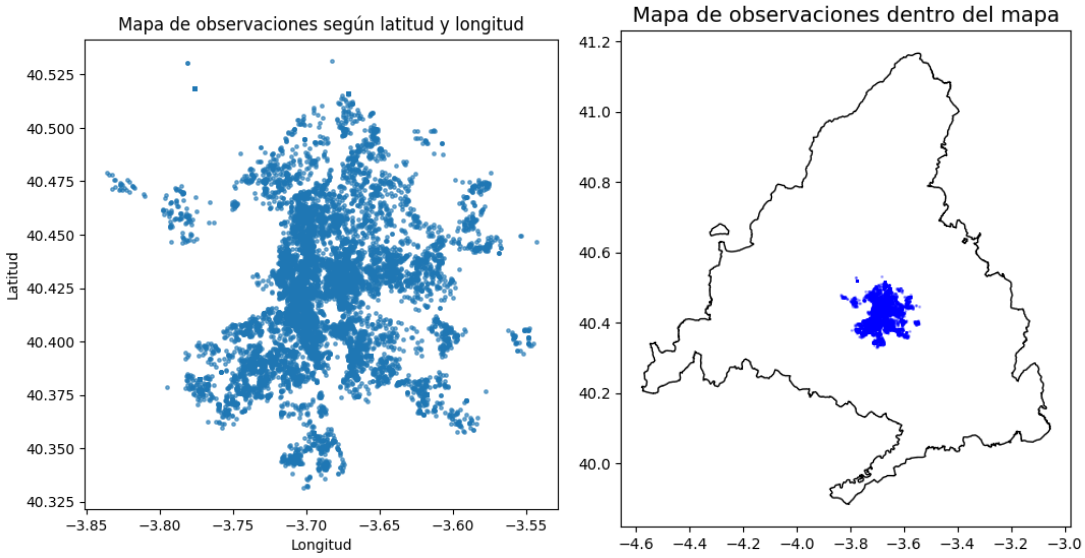
El análisis univariante de la variable del número de planta donde se sitúa la vivienda, "Floor" (Figura 22) muestra una distribución discreta, con una elevada concentración de viviendas en

plantas bajas y medias, y una presencia reducida de viviendas situadas en plantas muy altas. El boxplot identifica algunos valores extremos asociados a edificios singulares, que no se consideran errores de registro. Por ello, la variable se mantiene sin aplicar transformaciones adicionales.

Encontramos siete variables categóricas binarias, que incluyen si hay o no ascensor ("Elevator"), aire acondicionado ("Air_Conditioner"), calefacción ("Heater"), *parking* ("Parking"), balcón ("Balcony"), terraza ("Terrace") y piscina ("Swimming_Pool"). Todas son correctas y toman valores 0 o 1, y serán estudiadas más a fondo en el análisis bivariante para ver su relación con la variable objetivo.

Para terminar el análisis univariante vamos a observar las variables de localización: latitud y longitud. El análisis exploratorio reveló la existencia de tres observaciones con latitud y longitud iguales a cero, valor que no corresponde a ninguna ubicación real dentro del área de estudio. Dado que estas coordenadas no aportan información válida y pueden distorsionar el análisis espacial, dichas observaciones fueron eliminadas del conjunto de datos. Como parte del análisis exploratorio, se realizó una validación de las coordenadas geográficas de las viviendas mediante la representación conjunta de la latitud y la longitud. El mapa resultante (Figura 23) muestra que las observaciones se concentran en el área correspondiente a la ciudad de Madrid, sin detectarse agrupaciones anómalas fuera del ámbito de estudio.

Figura 23. Mapas con las variables latitud y longitud.

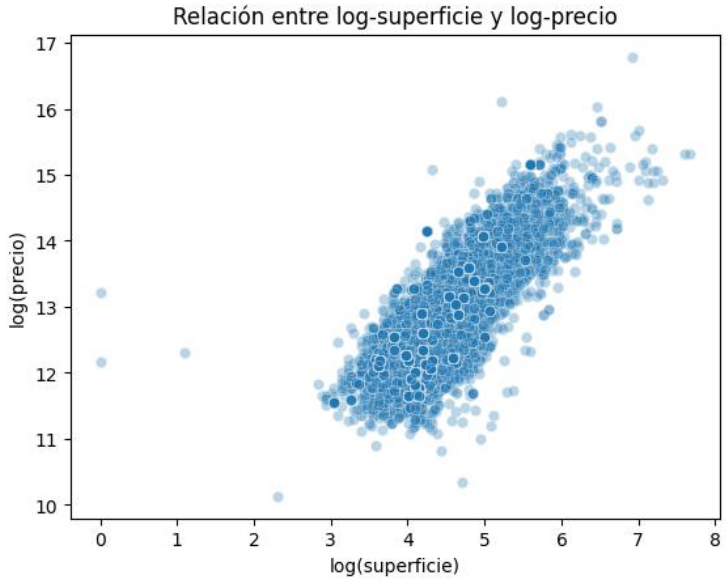


Análisis bivariante

Tras este análisis y pertinentes transformaciones de las variables numéricas, se procedió a un análisis bivariante, cuya finalidad es analizar la relación de cada variable predictiva con la variable objetivo precio de la vivienda. El objetivo es identificar patrones, relaciones funcionales y posibles dependencias entre variables, así como evaluar la relevancia de las características estructurales, los equipamientos y la localización en la determinación del precio.

En primer lugar, se analiza la relación entre el logaritmo del precio y el logaritmo de la superficie.

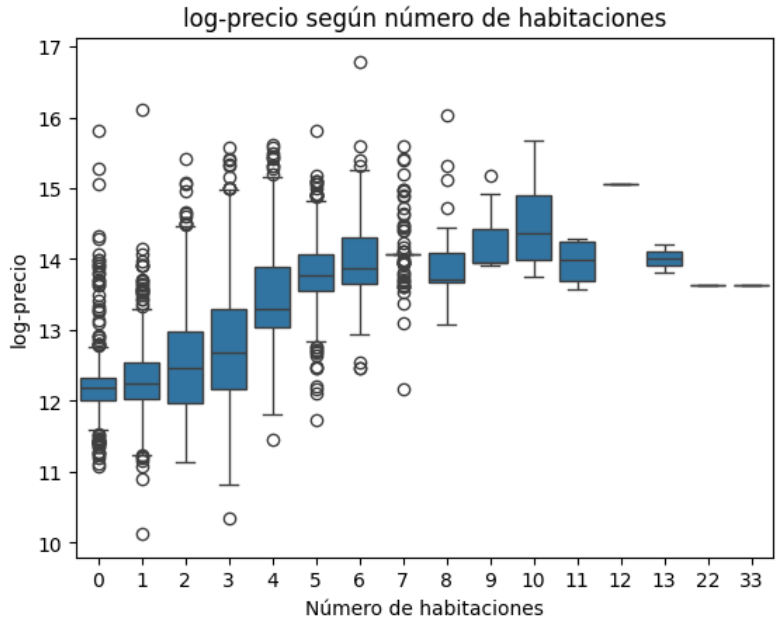
Figura 24. Relación entre la variable respuesta log-precio y log-superficie.



La Figura 24 muestra la relación entre dichas variables transformadas. Se observa una relación positiva clara y aproximadamente lineal entre ambas variables, lo que indica que la superficie es uno de los principales determinantes del precio. La transformación logarítmica permite linealizar la relación y reducir la influencia de valores extremos.

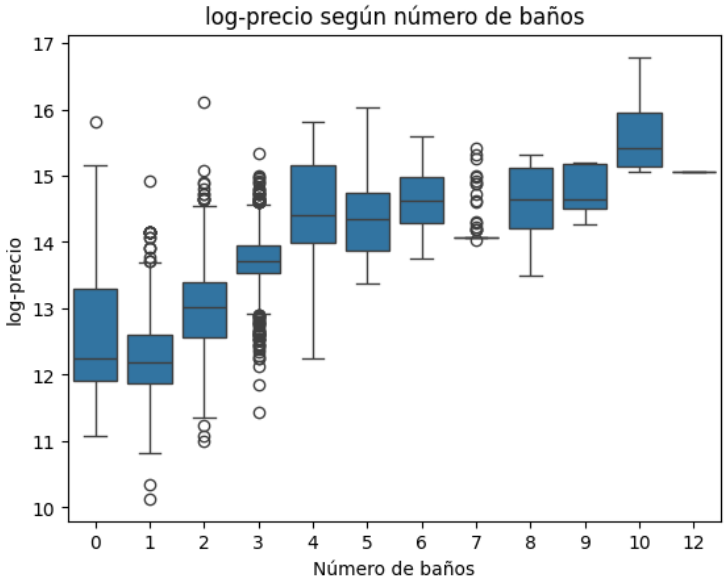
A continuación, se examina la relación entre el logaritmo del precio y el número de habitaciones de la vivienda, que se representa gráficamente en la Figura 25. Se observa una relación positiva clara, con incrementos significativos del precio medio al aumentar el número de habitaciones, especialmente en los tramos más bajos; a partir de un cierto número de habitaciones, el logaritmo del precio tiende a estabilizarse y su variabilidad se reduce, reflejando una menor presencia de viviendas con un número extremo de habitaciones. En conjunto, la variable presenta un comportamiento coherente y se mantiene como variable explicativa en el modelo.

Figura 25. Relación entre la variable respuesta log-precio y el número de habitaciones.



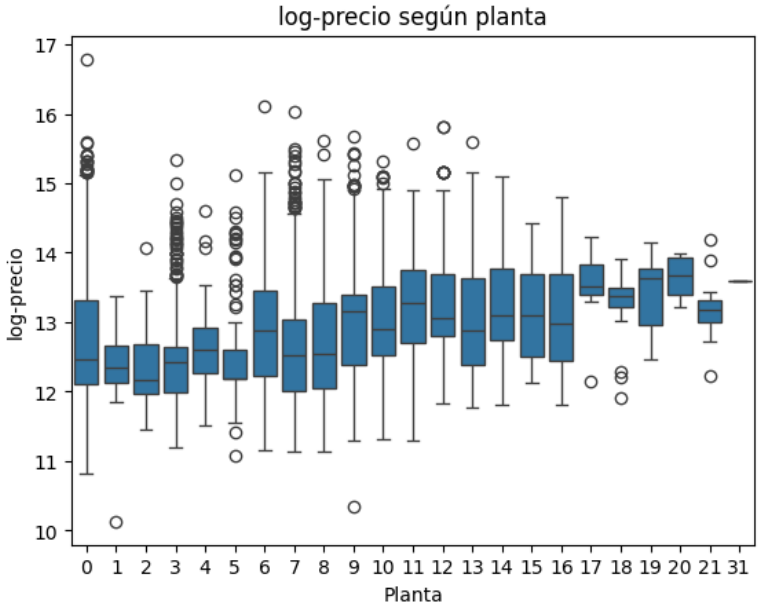
Por su parte, la Figura 26 muestra la relación entre el logaritmo del precio de la vivienda y el número de baños. Se observa una relación positiva clara, con incrementos del precio medio especialmente significativos en los valores más bajos de la variable independiente. En conjunto, la variable presenta un comportamiento coherente y se mantiene como variable explicativa en el modelo.

Figura 26. Relación entre la variable respuesta log-precio y el número de baños.



Finalmente, la Figura 27 muestra la relación entre el logaritmo del precio de la vivienda y la planta en la que se sitúa el inmueble. Se aprecia una relación positiva moderada, de modo que las viviendas ubicadas en plantas más altas tienden a presentar precios ligeramente superiores. No obstante, la dispersión es elevada, especialmente en plantas bajas e intermedias, lo que sugiere que el efecto de la planta sobre el precio depende de otros factores adicionales. En consecuencia, la variable se mantiene como explicativa sin aplicar transformaciones adicionales.

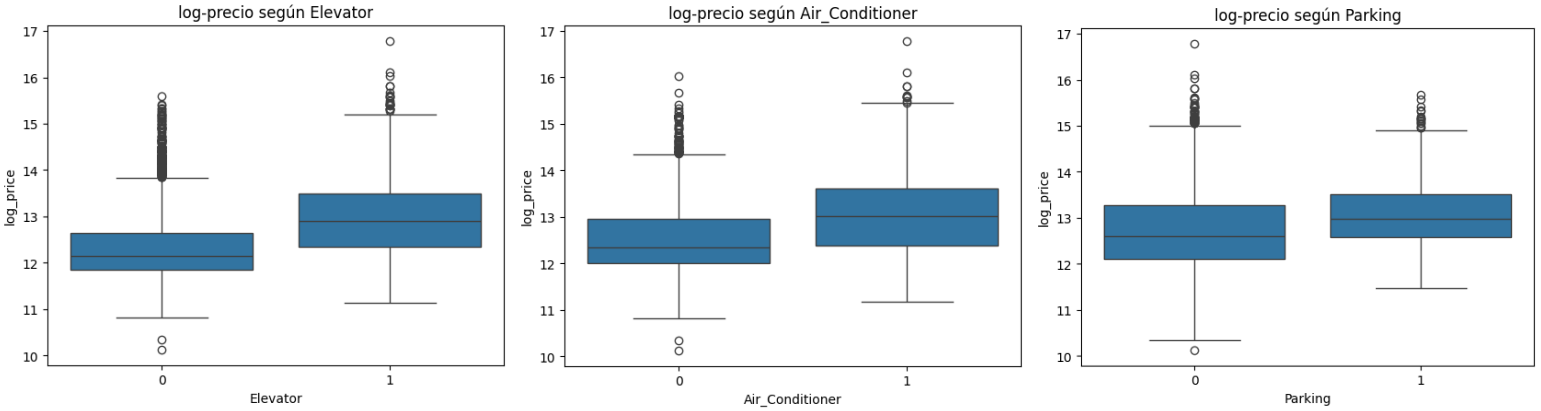
Figura 27. Relación entre la variable respuesta log-precio y la planta.



Continuamos el análisis estudiando ahora las variables categóricas binarias. Se presenta la relación de cada una de estas variables con el logaritmo del precio mediante *boxplots*.

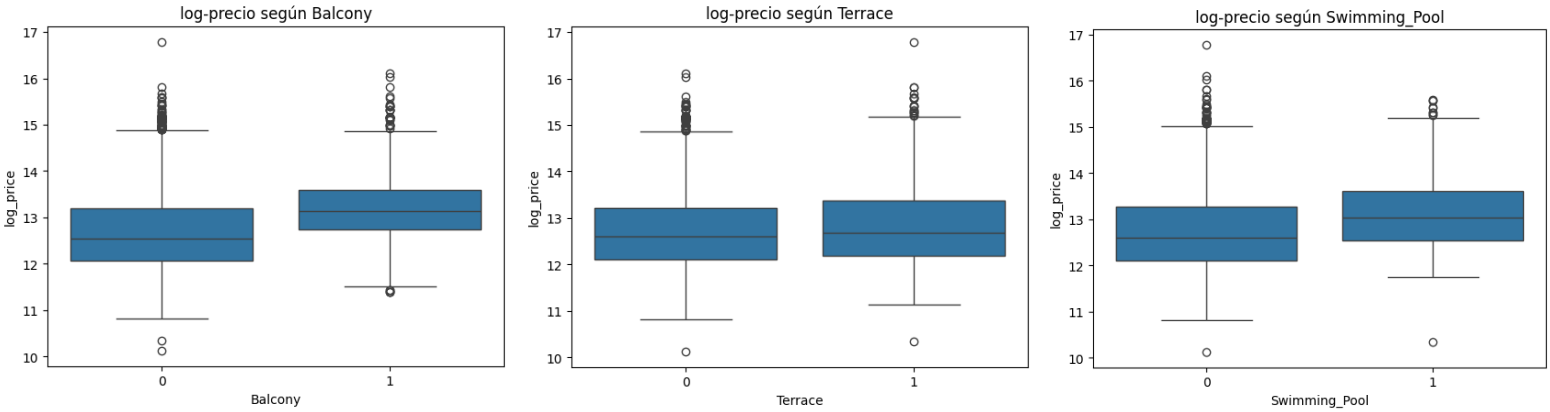
El análisis exploratorio bivariante de las variables categóricas muestra diferencias sistemáticas en el precio de la vivienda en función de la presencia de determinados equipamientos. En general, las viviendas que disponen de ascensor, aire acondicionado o plaza de aparcamiento presentan valores medianos del log-precio claramente superiores, como se observa en la Figura 28.

Figura 28. Boxplots de la distribución de la variable respuesta log-precio según si hay o no ascensor, aire acondicionado y aparcamiento.



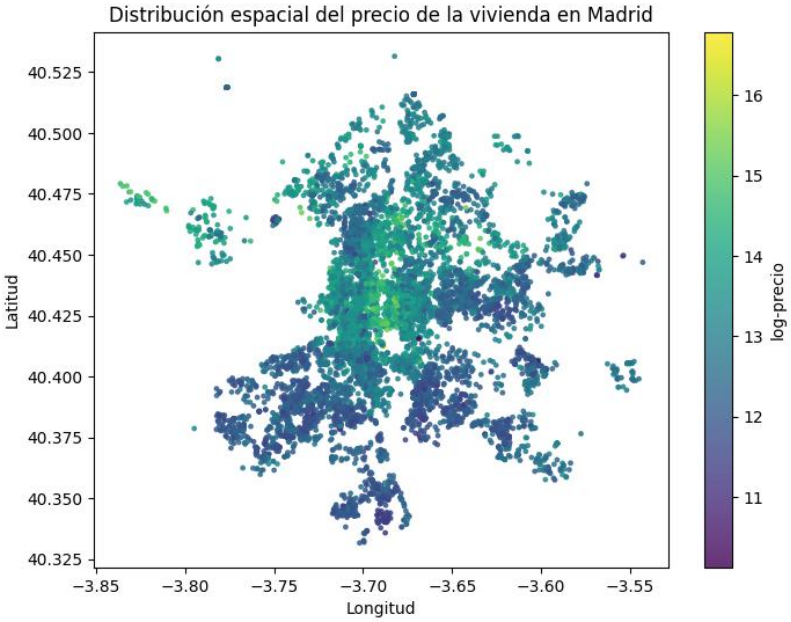
Asimismo, la presencia de características como balcón, terraza o piscina se asocian a niveles de precio más elevados (Figura 29), si bien en estos casos la dispersión es mayor y el número de observaciones es más reducido. Estos resultados indican que los equipamientos influyen positivamente en el precio de la vivienda, aunque su efecto debe interpretarse conjuntamente con otras características estructurales y de localización.

Figura 29. Boxplots de la distribución de la variable respuesta log-precio según si hay o no balcón, terraza y piscina.



Por último, pasamos a pasar a analizar las variables de localización conjuntamente con la variable objetivo. Para la latitud y longitud se realiza un mapa con una escala de colores para cada ubicación dependiendo del precio.

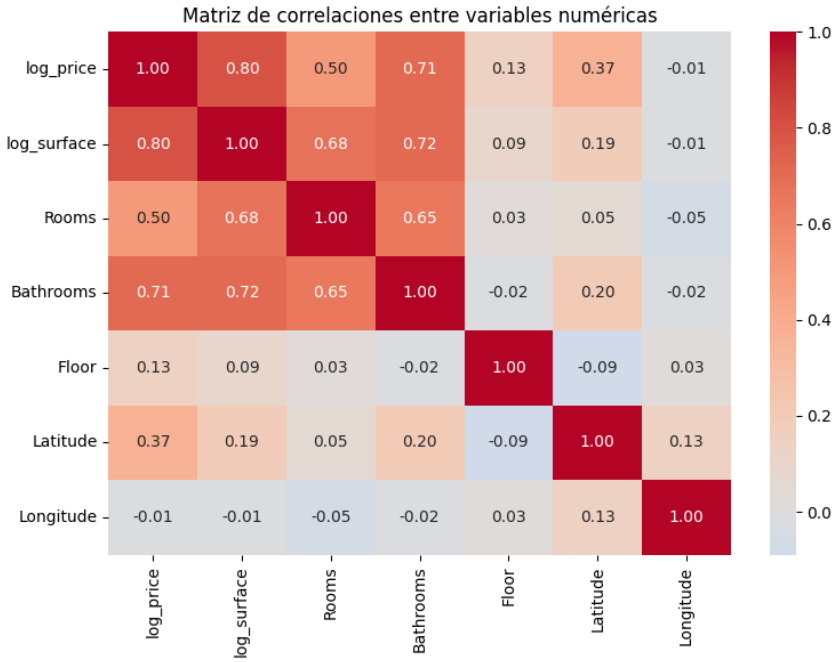
Figura 30. Mapa con la relación entre la variable respuesta log-precio con las variables latitud y longitud.



La Figura 30 muestra la distribución espacial del logaritmo del precio de la vivienda en el municipio de Madrid. Se observa una marcada heterogeneidad espacial, con concentraciones de precios más elevados en determinadas zonas y una disminución progresiva hacia áreas periféricas. Este patrón destaca la importancia de la localización como factor explicativo del precio de la vivienda y justifica la inclusión de variables geográficas en los modelos.

Por su lado, la matriz de correlaciones entre las variables numéricas de la Figura 31 confirma los resultados obtenidos en el análisis exploratorio previo.

Figura 31. Matriz de correlaciones para las variables numéricas.



En particular, se observa una fuerte correlación positiva entre el logaritmo del precio y el logaritmo de la superficie, lo que refuerza su papel como principal determinante del precio de la vivienda. El número de baños y de habitaciones presentan correlaciones positivas moderadas con el precio, mientras que la planta muestra una relación más débil. Las variables de localización presentan correlaciones individuales reducidas, lo que sugiere que su efecto no es puramente lineal y debe interpretarse de forma conjunta. En términos generales, no se aprecian problemas severos de colinealidad entre las variables consideradas; no obstante, en la Sección 4.4.3 se analizan los factores de inflación de la varianza (VIF) con el fin de evaluar la posible presencia de multicolinealidad.

4.4.2. Preparación de los datos para los modelos

Una vez realizada la limpieza inicial del conjunto de datos y el análisis exploratorio, en este apartado se describen las transformaciones finales y decisiones metodológicas necesarias para preparar los datos de cara a la aplicación de distintos modelos predictivos.

A partir del análisis exploratorio, se selecciona un conjunto final de variables explicativas que incluyen características estructurales de la vivienda, equipamientos y variables de localización. Se excluyen aquellas variables que no aportan información relevante para la modelización o que introducen ruido, como la dirección textual o el código postal.

Las transformaciones logarítmicas previamente realizadas permiten linealizar relaciones, reducir la asimetría de las distribuciones y atenuar la influencia de valores extremos, facilitando la estimación de modelos lineales y la comparación entre algoritmos.

Antes de la utilización de los modelos, y siguiendo las técnicas de validación explicadas en la Sección 2.1.1, el conjunto de datos se divide en un conjunto de entrenamiento y un conjunto de prueba (*test*), con el objetivo de evaluar el rendimiento predictivo de los distintos algoritmos sobre observaciones no utilizadas en el ajuste. Concretamente, se emplea una partición del 60% de las observaciones para entrenamiento y del 40% para prueba, fijando una semilla aleatoria para garantizar la reproducibilidad de los resultados. Adicionalmente, para el ajuste de hiperparámetros de los modelos que lo requieren, se utiliza una búsqueda en rejilla (*grid search*) (Alibrahim & Ludwig, 2021) a través de una validación cruzada de 5 o 10 pliegues (*10-fold CV*), según el modelo utilizado y su complejidad computacional. Asimismo, el preprocesamiento de las variables se diseña de forma específica para cada tipo de modelo considerado. Mientras que los algoritmos sensibles a la escala de las variables requieren la normalización de las variables numéricas, los modelos basados en árboles no precisan dicho escalado. Esta estrategia permite adaptar el conjunto de datos a las características de cada algoritmo y evitar posibles problemas de fuga de información (*data leakage*), asegurando una comparación coherente entre modelos.

4.4.3. Resultados predictivos

Regresión Lineal Múltiple.

En primer lugar, se estimó un modelo de regresión lineal múltiple como aproximación base al problema de predicción del precio de la vivienda. Los coeficientes de este modelo se estimaron mediante el método de Mínimos Cuadrados Ordinarios (MCO u OLS), cuyo objetivo es minimizar

la suma de los residuos al cuadrado, como se estudió en la Sección 2.2.2. La elección de este modelo responde a su elevada interpretabilidad y a su amplio uso como referencia en estudios empíricos del mercado inmobiliario. Este modelo servirá como punto de referencia (*baseline*) para el resto de modelos predictivos.

Con el fin de evaluar adecuadamente la capacidad de generalización del modelo, el conjunto de datos se dividió en un subconjunto de entrenamiento y otro de *test*. Todas las transformaciones de las variables, incluida la estandarización de los predictores numéricos, se implementaron mediante un *pipeline*, garantizando que dichas transformaciones se ajustaran exclusivamente sobre los datos de entrenamiento y evitando así problemas de *data leakage*. Aunque para los modelos predictivos se emplea estandarización de las variables, la inferencia estadística se realiza sobre el modelo estimado en unidades originales.

El uso de estandarización permite comparar la importancia de cada variable dentro del modelo. En este contexto, los coeficientes estimados no se interpretan como efectos marginales en unidades reales, sino como medidas comparables del impacto relativo de cada predictor sobre el logaritmo del precio. La Tabla 5 presenta los coeficientes estandarizados ordenados por su valor absoluto, con el objetivo de identificar las variables con mayor influencia relativa en el modelo.

Tabla 5. Coeficientes del modelo de Regresión Lineal.

VARIABLE	**COEFICIENTE**
log_surface	0.472530
Bathrooms	0.243886
Elevator	0.197918
Balcony	0.188716
Terrace	- 0.172562
Latitude	0.163610
Swimming_Pool	- 0.117210
Rooms	- 0.092631
Air_Conditioner	0.079113
Floor	0.066421
Parking	- 0.047951
Longitude	- 0.023426
Heater	0.017046

El modelo de regresión lineal estimado se especifica sobre el logaritmo del precio de la vivienda, con el objetivo de reducir la asimetría de la distribución y facilitar la interpretación de los coeficientes en términos porcentuales.

La especificación adoptada viene dada por la expresión:

$$\begin{aligned}\log(\text{Precio}_i) = {} & -188.236 + 0.869 \log(\text{Surface}_i) + 0.224\ \text{Bathrooms}_i - 0.073\ \text{Rooms}_i \\ & + 0.017\ \text{Floor}_i + 4.808\ \text{Latitude}_i - 0.632\ \text{Longitude}_i + 0.2\ \text{Elevator}_i \\ & + 0.192\ \text{Balcony}_i - 0.17\ \text{Terrace}_i - 0.117\ \text{SwimmingPool}_i \\ & + 0.081\ \text{AirConditioner}_i - 0.047\ \text{Parking}_i + 0.013\ \text{Heater}_i + \varepsilon_i\end{aligned}$$

donde $\varepsilon_i$representa el término de error aleatorio asociado a la observación i.

La ecuación anterior recoge la especificación completa del modelo de regresión lineal estimado. En ella se combinan variables continuas y binarias que capturan tanto las características estructurales del inmueble como sus equipamientos y su localización. Esta especificación permite analizar el efecto marginal de cada predictor sobre el precio de la vivienda, manteniendo constantes el resto de variables del modelo.

Dado que la variable dependiente se ha especificado mediante logaritmos, las predicciones obtenidas por el modelo corresponden inicialmente al logaritmo del precio. Para recuperar el precio estimado en niveles, se aplica la transformación inversa mediante la función exponencial. En concreto, si $\hat{y}_i$ denota la predicción del logaritmo del precio para la vivienda i, el precio estimado se obtiene como $\widehat{\text{Precio}}_i = \exp(\hat{y}_i)$.

El contraste de significación global del modelo arroja un p-valor prácticamente nulo, lo que permite rechazar la hipótesis nula de ausencia de efecto conjunto de las variables explicativas. En consecuencia, se concluye que el modelo es globalmente significativo desde el punto de vista estadístico.

Tabla 6. Coeficientes, errores, p-valores e intervalos de confianza del modelo de Regresión Lineal.

VARIABLE	COEFICIENTE	STD. ERROR	P-VALOR	IC 95%
const	-188.236	4.12	0.000	[-196.3, -180.18]
log_surface	0.869	0.01	0.000	[0.850, 0.889]
Bathrooms	0.224	0.005	0.000	[0.215, 0.233]
Elevator	0.2	0.007	0.000	[0.185, 0.214]
Balcony	0.192	0.009	0.000	[0.175, 0.209]
Terrace	-0.17	0.007	0.000	[-0.184, -0.155]
Latitude	4.808	0.1	0.000	[4.610, 5.006]
Swimming_Pool	-0.117	0.013	0.000	[-0.142, -0.091]
Rooms	-0.073	0.004	0.000	[-0.081, -0.066]
Air_Conditioner	0.081	0.007	0.000	[0.067, 0.094]
Floor	0.017	0.001	0.000	[0.015, 0.019]
Parking	-0.047	0.013	0.000	[-0.072, -0.021]
Longitude	-0.632	0.081	0.000	[-0.790, -0.474]
Heater	0.013	0.007	0.067	[-0.01, 0.027]

La Tabla 6 recoge los coeficientes estimados por el modelo especificado sobre el logaritmo del precio, junto con sus errores estándar, p-valores e intervalos de confianza al 95%. Los coeficientes representan efectos marginales condicionales: el impacto de cada predictor sobre $\log(\text{Precio})$ manteniendo constantes el resto de variables del modelo. El error estándar ("std. Error") cuantifica la precisión de cada estimación, mientras que el p-valor permite contrastar la significación individual de cada coeficiente, con hipótesis nula $H_0: \beta_j = 0$ frente a la hipótesis alternativa $H_1: \beta_j \neq 0$. Los intervalos de confianza al 95% complementan esta evidencia indicando el rango de valores plausibles para cada parámetro.

Dado que tanto el precio como la superficie se han especificado en logaritmos, el coeficiente asociado a $\log(\textit{Surface})$ puede interpretarse como una elasticidad. En concreto, manteniendo constantes el resto de características del inmueble, un aumento del 1% en la superficie se asocia con un incremento aproximado del 0.87% en el precio de la vivienda. Por ejemplo, un aumento

del 10% en la superficie implicaría un incremento cercano al 8.7% en el precio de la vivienda. Añadir un baño adicional se asocia con un incremento aproximado del 22.4% en el precio de la vivienda, manteniendo constantes la superficie, la localización y el resto de características. Dado que la latitud y la longitud se miden en grados geográficos, su interpretación es algo compleja. Un aumento de 0.01 grados de latitud (≈ 1.1 km hacia el norte) se asocia con un incremento aproximado del 4.8% en el precio de la vivienda, manteniendo constantes el resto de variables. Este resultado indica que las zonas más al norte del municipio de Madrid tienden a presentar precios más elevados. La presencia de ascensor se asocia con un incremento aproximado del 20% en el precio de la vivienda, en comparación con viviendas similares sin ascensor. Esto refleja una prima de accesibilidad claramente valorada por el mercado. Otras variables relacionadas con la calidad y las comodidades del inmueble, como la presencia de balcón, presentan coeficientes positivos, reflejando su contribución al valor de mercado.

Algunos coeficientes muestran signos que podrían resultar contraintuitivos si se interpretan de forma aislada. Sin embargo, en un modelo de regresión múltiple los coeficientes representan efectos marginales condicionados, es decir, el impacto de cada variable manteniendo constantes el resto de predictores. En este sentido, el coeficiente negativo asociado al número de habitaciones refleja el efecto de añadir habitaciones dado un nivel fijo de superficie total, lo que puede implicar una menor eficiencia en la distribución del espacio.

En conjunto, los coeficientes estimados por el modelo de regresión lineal son coherentes desde un punto de vista económico e inmobiliario. Las variables relacionadas con el tamaño, la localización y la calidad del inmueble concentran los efectos más relevantes, mientras que los signos aparentemente contraintuitivos se explican por la naturaleza condicional de los coeficientes en un modelo multivariante. Estos resultados refuerzan el papel del modelo lineal como herramienta interpretativa, aunque también dejan ver sus limitaciones para capturar relaciones más complejas.

Los intervalos de confianza al 95% proporcionan información adicional sobre la precisión y la estabilidad de los coeficientes estimados. En particular, el hecho de que la mayoría de los intervalos no incluya el valor cero indica que los efectos asociados a estas variables son estadísticamente significativos al nivel del 5%. Además, la amplitud relativamente reducida de los intervalos para las variables más relevantes sugiere que sus estimaciones son precisas y poco sensibles a la variabilidad muestral. En conjunto, estos resultados refuerzan la evidencia de que las variables relacionadas con el tamaño del inmueble, su localización y la presencia de determinadas comodidades ejercen un impacto robusto y consistente sobre el precio de la vivienda.

Todas las variables, excepto “Heater” (presencia o no de calefacción), resultan estadísticamente significativas al nivel del 5%. La variable “Heater” presenta un p-valor ligeramente superior ($p \approx 0.07$), por lo que su efecto se considera marginal.

Para analizar la posible existencia de multicolinealidad entre las variables explicativas, se calcularon los factores de inflación de la varianza (VIF), mostrados en la Tabla 8, y se calculó la matriz de correlaciones, mostrada anteriormente en la Sección 4.2.1. Todos los valores de VIF se sitúan por debajo de los umbrales comúnmente aceptados, lo que indica la ausencia de multicolinealidad.

En consecuencia, los signos observados en los coeficientes no se deben a inestabilidad numérica del modelo, sino a la información parcialmente solapada capturada por variables correlacionadas que describen características similares de la vivienda.

En lo que respecta al poder predictivo de este modelo, el rendimiento del modelo de regresión lineal se evaluó mediante el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE) en los conjuntos de entrenamiento y *test*.

Tabla 7. Métricas de rendimiento del modelo de Regresión Lineal en el subconjunto de entrenamiento y de test.

CONJUNTO	**R^2**	**RMSE**
ENTRENAMIENTO	0.7811	0.3775
TEST	0.7761	0.3755

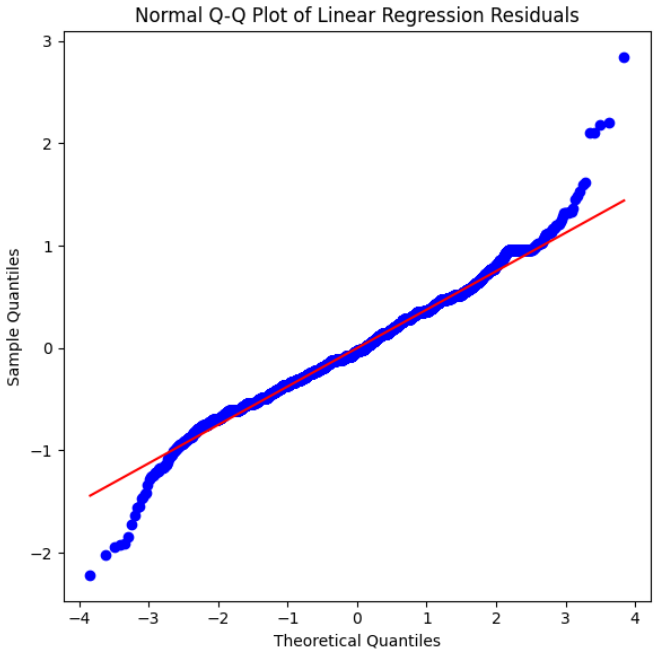
Tal y como se observa en la Tabla 7, en el conjunto de entrenamiento, el modelo alcanza un valor de R^2 de 0.7811, lo que indica que el modelo es capaz de explicar aproximadamente el 78% de la variabilidad del logaritmo del precio. El RMSE asociado, con un valor de 0.3775, refleja el error medio de predicción en dicha escala. En el conjunto de *test*, el valor de R^2 se sitúa en 0.7761, lo que implica que el modelo mantiene una capacidad explicativa similar fuera de muestra, explicando en torno al 78% de la variabilidad observada. Asimismo, el RMSE en test se mantiene en 0.3755, prácticamente idéntico al obtenido en entrenamiento, lo que sugiere que el modelo no presenta sobreajuste severo y generaliza de forma razonablemente estable a datos no observados. Dado que el modelo se estima sobre el logaritmo del precio, las métricas de error se expresan en dicha escala, facilitando la comparación entre modelos y reduciendo la influencia de valores extremos.

A continuación, se muestran las figuras para el diagnóstico de residuos.

Tabla 8. VIF de las variables numéricas del modelo.

VARIABLE	**VIF**
log_surface	2.569011
Bathrooms	2.370073
Rooms	2.088464
Latitude	1.098403
Floor	1.034405
Longitude	1.022560

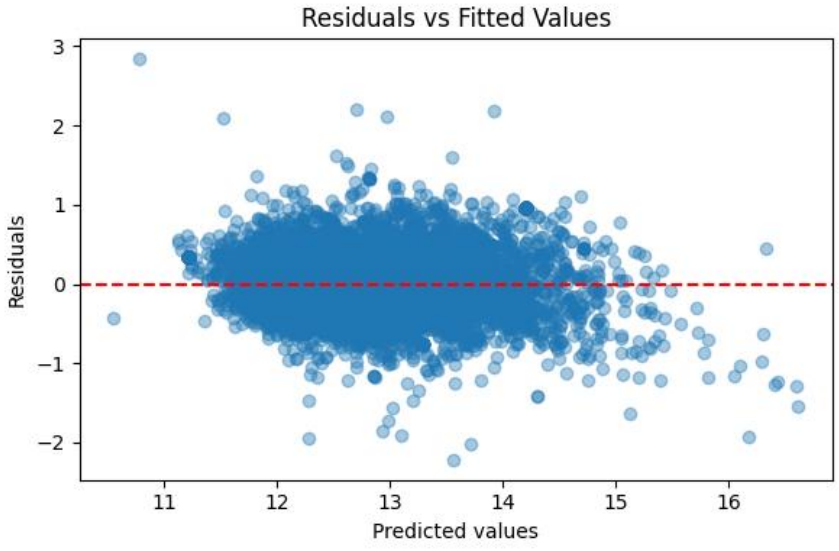
Figura 32. Gráfico Q-Q normal de los residuos del modelo de Regresión Lineal.



La Figura 32 muestra el gráfico Q–Q normal de los residuos del modelo de regresión lineal. En la región central de la distribución, los puntos se alinean de forma razonablemente cercana a la recta teórica, lo que indica que los residuos presentan una normalidad aproximada en torno a su media. No obstante, se observan desviaciones claras en ambas colas, especialmente en los cuantiles extremos, lo que sugiere la presencia de colas más pesadas que las de una distribución normal.

Este comportamiento es habitual en datos de precios inmobiliarios, donde las observaciones correspondientes a viviendas muy baratas o muy caras resultan más difíciles de capturar mediante una especificación lineal. En cualquier caso, la desviación observada no es lo suficientemente severa como para invalidar el modelo, si bien expone sus limitaciones en la modelización de valores extremos.

Figura 33. Gráfico de residuos frente a valores ajustados por el modelo de Regresión Lineal.



La Figura 33 representa los residuos frente a los valores ajustados por el modelo. Los residuos se encuentran centrados alrededor de cero, sin mostrar patrones sistemáticos claros ni estructuras no lineales evidentes, lo que sugiere que la especificación funcional lineal resulta razonable como aproximación media.

Sin embargo, se aprecia un ligero incremento de la dispersión de los residuos para valores predichos más elevados, indicando la presencia de heterocedasticidad moderada. En particular, el error de predicción tiende a ser mayor en viviendas de mayor precio, un fenómeno coherente con la mayor heterogeneidad existente en los segmentos altos del mercado inmobiliario.

Para analizar mejor el rendimiento, se aplicó un procedimiento de validación cruzada *k-fold* con diez pliegues (*10-fold CV*) sobre el conjunto de entrenamiento con el objetivo de evaluar la estabilidad del modelo de regresión lineal. Los resultados obtenidos muestran un valor medio de R^2 de 0.7794, con una desviación estándar de 0.0105, lo que indica que el modelo explica de forma consistente aproximadamente el 78% de la variabilidad del logaritmo del precio a lo largo de las distintas particiones.

De manera complementaria, el error cuadrático medio presenta un valor medio de RMSE de 0.3786, con una desviación estándar reducida de 0.0072, lo que sugiere que el rendimiento del modelo es estable y poco sensible a la división concreta de los datos en los subconjuntos de entrenamiento y validación.

Tratando de mejorar el ajuste del modelo de regresión lineal y aproximar de forma más adecuada la relación entre las variables explicativas y la variable respuesta, se analiza el efecto de distintas transformaciones sobre determinadas variables discretas. En particular, se consideran transformaciones del número de habitaciones y del número de baños, variables que presentan relaciones no estrictamente lineales con el logaritmo del precio de la vivienda.

En el caso del número de habitaciones, se compara su inclusión en forma original frente a una transformación mediante raíz cuadrada, mientras que para el número de baños se evalúa el uso de su transformación logarítmica. Estas transformaciones permiten reducir la asimetría de las distribuciones y suavizar el efecto de valores elevados, favoreciendo una relación más próxima a la linealidad.

En el modelo base, el coeficiente de determinación en el conjunto de *test* alcanza un valor de R^2= 0.7811, que sube hasta 0.7828 utilizando el modelo con las variables transformadas. La importancia de variables es muy similar en ambos modelos, con los coeficientes muy similares. Esta mejora, aunque de magnitud moderada, resulta consistente tanto en el conjunto de entrenamiento como en validación cruzada, donde también mejoran ligeramente el RMSE y R^2 medios, con desviaciones típicas ligeramente inferiores, lo que indica un comportamiento más estable del modelo.

Los gráficos Q–Q de los residuos muestran, en ambos modelos, desviaciones respecto a la normalidad en las colas, un comportamiento habitual en datos inmobiliarios. No obstante, el modelo con variables transformadas presenta una mejor alineación con la recta teórica en la región central de la distribución, así como una reducción de la asimetría en los extremos, lo que sugiere una aproximación más adecuada a la normalidad de los errores, como se puede apreciar en la Figura A.1 del Anexo.

De forma consistente, los gráficos de residuos frente a valores ajustados evidencian una dispersión más homogénea de los residuos en el modelo con transformaciones (Figura A.2 del Anexo), con una menor presencia de patrones sistemáticos y una ligera reducción de la heterocedasticidad observada en el modelo base, especialmente para valores elevados del precio estimado.

Finalmente, los gráficos de regresión parcial permiten analizar el efecto marginal de cada variable controlando por el resto de predictores. En este contexto, las transformaciones aplicadas al número de habitaciones y al número de baños dan lugar a relaciones más próximas a la linealidad, con pendientes mejor definidas y menor dispersión, lo que puede interpretarse como una captación más realista de rendimientos decrecientes en su impacto sobre el precio de la vivienda. Se puede ver la comparación en las Figuras A.3 y A.4 del Anexo. Estos resultados apoyan el uso de las variables transformadas, al contribuir a una mejora en el comportamiento estadístico del modelo sin alterar su estructura ni su interpretabilidad.

Con el objetivo de evaluar la parsimonia del modelo, se estimó una versión reducida excluyendo la variable "Heater", que no resultó estadísticamente significativa al nivel del 5%. Para evaluar si la reducción del número de variables supone una pérdida significativa de capacidad explicativa, se aplica el contraste de razón de verosimilitudes entre el modelo completo y el modelo reducido. En el caso del modelo sin transformaciones, el estadístico del contraste toma un valor de $LR = 4.21$, con un p-valor de 0.040, lo que conduce al rechazo de la hipótesis nula al nivel de significación del 5%. Este resultado indica que el modelo reducido presenta una pérdida significativa de ajuste respecto al modelo completo. Por el contrario, al considerar el modelo con variables transformadas, el contraste de razón de verosimilitudes arroja un estadístico de $LR = 2.94$, con un p-valor de 0.086. En este caso, no se rechaza la hipótesis nula al nivel de significación habitual, lo que sugiere que el modelo reducido mantiene una capacidad explicativa comparable a la del modelo completo.

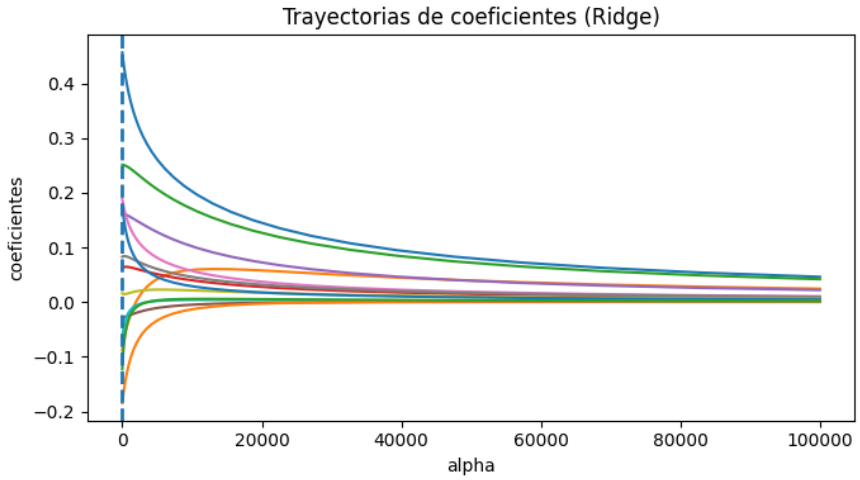
En conjunto, el modelo de regresión lineal proporciona una aproximación interpretable y estable al problema de predicción del precio de la vivienda. No obstante, las desviaciones observadas en los gráficos y la presencia de heterocedasticidad moderada sugieren que la relación entre las variables puede presentar componentes no lineales, lo que motiva la consideración de modelos más flexibles en secciones posteriores.

Regresión lineal penalizada.

A pesar de la buena capacidad explicativa del modelo de regresión lineal múltiple, la estimación por mínimos cuadrados ordinarios puede verse afectada por problemas de multicolinealidad o por una elevada varianza de las estimaciones cuando se incluyen múltiples predictores. Con el objetivo de evaluar la robustez del modelo lineal base y analizar si la introducción de regularización mejora su estabilidad y capacidad de generalización, se consideran modelos de regresión penalizada. En particular, se analizan las regresiones *Ridge* y *Lasso*, explicados en la Sección 2.2.2., que incorporan penalizaciones de norma L_2 y L_1, respectivamente, permitiendo reducir la complejidad del modelo y, en el caso de Lasso, realizar selección automática de variables.

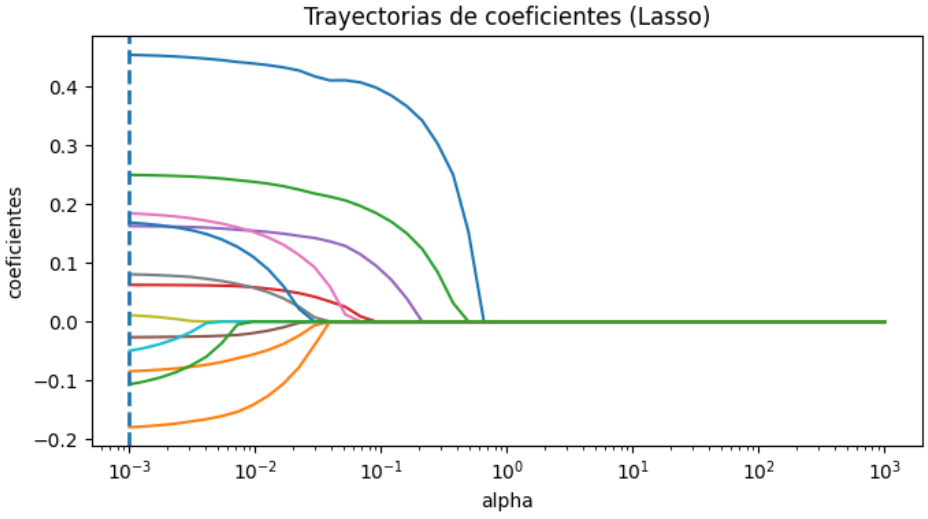
Para el modelo *Ridge*, cuya trayectoria de coeficientes puede observarse en la Figura 34, se realiza la selección del parámetro de regularización mediante validación cruzada, arrojando un valor óptimo de $\lambda \approx 5.7$. No obstante, las métricas de ajuste y los coeficientes estimados resultan prácticamente idénticos a los del modelo OLS. Este resultado indica que el modelo lineal base no presenta problemas relevantes de multicolinealidad ni sobreajuste y que la regularización no aporta mejoras adicionales en términos de capacidad predictiva. En consecuencia, *Ridge* actúa como una validación de la estabilidad del modelo de regresión lineal base. Este resultado es coherente con los bajos valores de VIF observados previamente y con la estabilidad mostrada por el modelo de regresión lineal múltiple en validación cruzada.

Figura 34. Trayectorias de los coeficientes para el modelo Ridge.



Por su parte, el modelo *Lasso*, cuya trayectoria de coeficientes puede observarse en la Figura 35, también realiza la selección del parámetro de regularización mediante validación cruzada, arrojando un valor óptimo muy reducido ($\lambda = 0.001$). La regularización *Lasso* no elimina ninguna de las variables explicativas, manteniendo una especificación prácticamente idéntica a la del modelo básico de regresión lineal. Este resultado refuerza la estabilidad del modelo lineal obtenido y sugiere que la información relevante para la predicción del precio se encuentra distribuida entre múltiples predictores.

Figura 35. Trayectorias de los coeficientes para el modelo Lasso.



Árbol de Decisión.

Con el objetivo de capturar posibles relaciones no lineales entre las variables explicativas y el precio de la vivienda, se estimó un árbol de decisión como primer modelo no paramétrico. A diferencia de la regresión lineal, los árboles de decisión permiten modelizar interacciones complejas y umbrales en los predictores, sin imponer una forma funcional predefinida.

Dado que los árboles de decisión pueden presentar una elevada varianza cuando su estructura es excesivamente compleja, se prestó especial atención a la selección de sus hiperparámetros de regularización. En una primera fase, se analizó de forma exploratoria el efecto de la profundidad máxima del árbol sobre el rendimiento mediante validación cruzada, representando gráficamente el coeficiente de determinación medio en función de dicho parámetro. Este análisis mostró una mejora rápida del ajuste para profundidades bajas, seguida de una clara estabilización del rendimiento a partir de valores moderados, lo que sugiere la existencia de un punto de rendimientos decrecientes. Este comportamiento es análogo al denominado *método del codo* y proporciona una primera aproximación heurística a la complejidad adecuada del modelo.

A continuación, con el fin de seleccionar una configuración óptima de manera más sistemática, se llevó a cabo una optimización conjunta de los principales hiperparámetros del árbol mediante validación cruzada de 5 pliegues. En particular, se consideraron la profundidad máxima del árbol y el tamaño mínimo de los nodos hoja, permitiendo capturar la interacción entre ambos parámetros. Se pueden observar los diferentes valores que fueron probados en la Tabla 36. Los resultados indican que la configuración óptima corresponde a una profundidad máxima de 12 y un mínimo de 5 observaciones en los nodos hoja.

Figura 36. Valores de rejilla utilizados para la obtención de los hiperparámetros óptimos en el Árbol de Decisión.

HIPERPARÁMETRO	VALORES DE REJILLA (GRID SEARCH)
PROFUNDIDAD	{4, 6, 8, 10, 12, 20}
OBS. MÍNIMAS POR HOJA	{1, 5, 10, 25, 50, 100}

El rendimiento del árbol de decisión se evaluó mediante el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE) en los conjuntos de entrenamiento y *test*.

Figura 37. Métricas de rendimiento para la evaluación del Árbol de Decisión.

CONJUNTO	R^2	RMSE
ENTRENAMIENTO	0.9678	0.1447
TEST	0.9130	0.2341

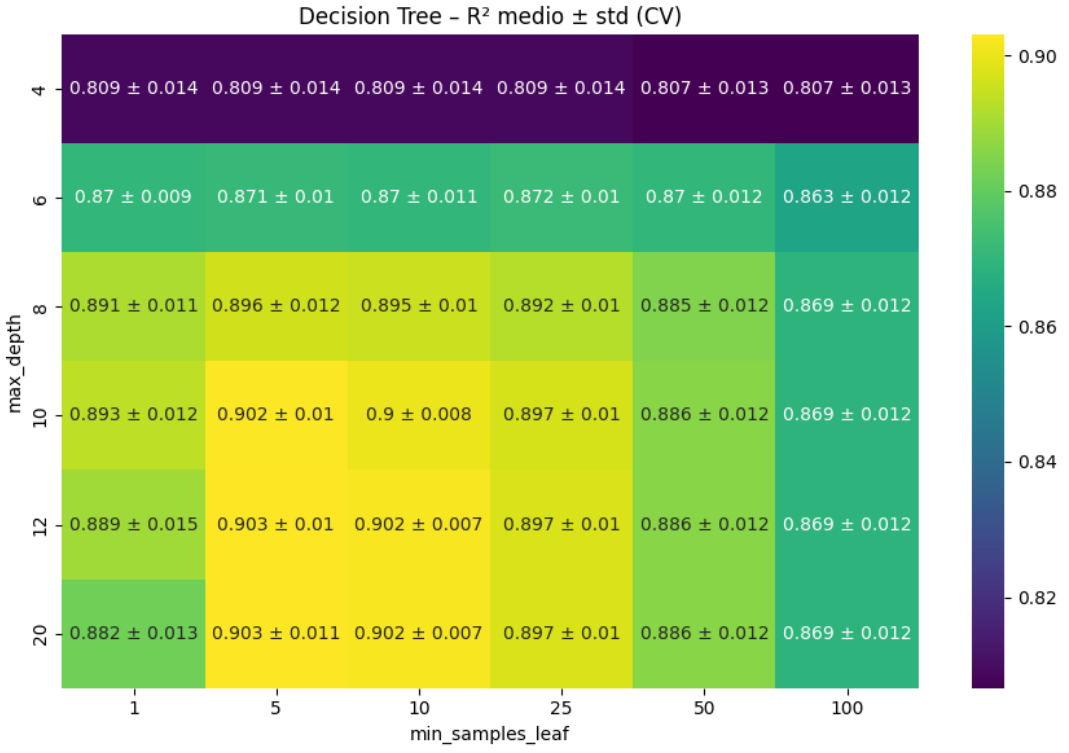
Los resultados muestran un valor elevado de R^2 en el conjunto de entrenamiento, lo que indica que el modelo es capaz de explicar una gran proporción de la variabilidad del precio cuando se ajusta sobre los datos observados. En concreto, el valor de R^2 en entrenamiento indica que el árbol explica aproximadamente el 96,78% de la variabilidad del logaritmo del precio. Sin

embargo, este porcentaje se reduce en el conjunto de *test*, donde el modelo explica en torno al 91% de la variabilidad, lo que refleja una pérdida de capacidad explicativa fuera de muestra. Esta diferencia entre el rendimiento en entrenamiento y en *test* sugiere la presencia de alta varianza, un comportamiento característico de los árboles de decisión individuales, que tienden a adaptarse en exceso a los datos de entrenamiento cuando presentan una estructura compleja.

Para evaluar la estabilidad del modelo, se aplicó un procedimiento de validación cruzada de 5 pliegues sobre el conjunto de entrenamiento. Los valores medios de R^2 obtenidos confirman que el árbol de decisión mantiene una capacidad explicativa elevada de forma consistente entre particiones, explicando aproximadamente el 90% de la variabilidad del precio en promedio. No obstante, el error medio obtenido mediante validación cruzada es superior al observado en entrenamiento, lo que refuerza la evidencia de que el modelo presenta una capacidad de generalización limitada en comparación con su ajuste *in-sample*.

A modo de contraste con la configuración óptima obtenida mediante validación cruzada, se consideró adicionalmente una selección heurística de los hiperparámetros del árbol de decisión, basada en la inspección visual de los resultados de validación cruzada y en criterios de simplicidad y estabilidad del modelo. En particular, se fijaron los valores de la profundidad máxima del árbol en 8 y del tamaño mínimo de los nodos hoja en 10, obteniendo un árbol con una estructura más simple que el anterior. La Figura 38 muestra el valor de R^2 medio para diferentes combinaciones de estos dos parámetros. Las Figuras A.5 y A.6 en el Anexo representan, de forma separada, la evolución del R^2 en función de cada parámetro, manteniendo fijo el otro, lo que permite identificar regiones en las que incrementos adicionales en la complejidad del modelo no se traducen en mejoras apreciables del rendimiento predictivo tras la validación cruzada.

Figura 38. Valor del R^2 medio para diferentes combinaciones de profundidad y tamaño mínimo de nodo hoja en el Árbol de Decisión.



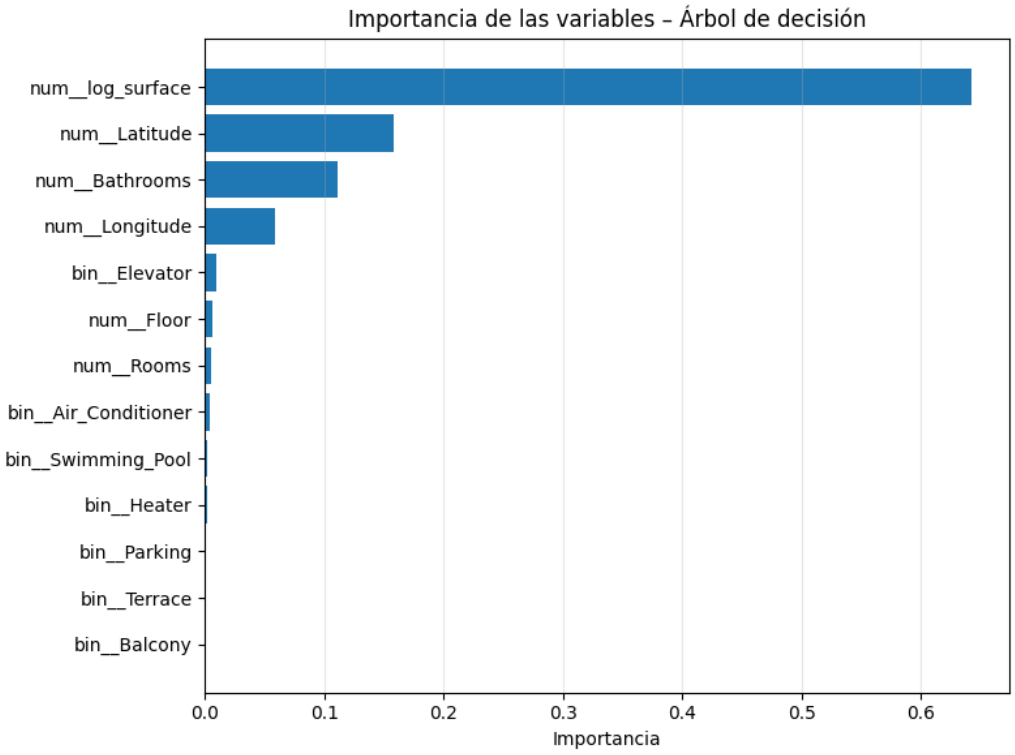
La elección de una profundidad máxima inferior a la óptima (10 frente a la profundidad de 12 anterior) responde al objetivo de reducir la complejidad estructural del árbol, limitando el número de particiones sucesivas y, por tanto, la capacidad del modelo para ajustarse de forma excesivamente específica a los datos de entrenamiento. Asimismo, el establecimiento de un tamaño mínimo de nodo hoja de 10 observaciones en contraste a los 5 del modelo óptimo, introduce un grado adicional de regularización, evitando la creación de hojas con un número reducido de observaciones y contribuyendo a una mayor estabilidad de las predicciones.

Desde el punto de vista empírico, esta combinación de hiperparámetros se sitúa en una región del espacio de búsqueda donde el coeficiente de determinación medio obtenido mediante validación cruzada es elevado y presenta una variabilidad reducida entre particiones, lo que sugiere un compromiso adecuado entre capacidad explicativa y robustez. De este modo, la configuración heurística seleccionada permite evaluar el impacto de una elección más conservadora de los hiperparámetros frente a la solución óptima obtenida de manera sistemática.

Con esta configuración heurística, el árbol de decisión alcanza un valor de R^2 de 0.921 en el conjunto de entrenamiento y de 0.896 en el conjunto de *test*, con valores de RMSE de 0.227 y 0.258, respectivamente. Estos resultados reflejan una reducción moderada de la capacidad explicativa respecto al modelo optimizado, acompañada de una disminución del riesgo de sobreajuste. Los resultados de la validación cruzada de 5 pliegues confirman esta tendencia, con un R^2 medio de 0.890 y una desviación estándar reducida, lo que indica un comportamiento más estable entre particiones en comparación con el ajuste óptimo.

Por otro lado, se llevó a cabo un análisis de la importancia de las variables, que se puede visualizar en la Figura 39.

Figura 39. Importancia de variables, en porcentaje, en el Árbol de Decisión.



Este análisis de la importancia de las variables indica que la superficie del inmueble, medida a través de su logaritmo ("num_log_surface"), es el predictor más relevante en la determinación del precio. Asimismo, las variables de localización, como la latitud ("num_Latitude") y la longitud ("num_Longitude"), y ciertas características estructurales, como el número de baños ("num_Bathrooms"), presentan una contribución significativa al proceso de predicción. En comparación con el modelo lineal, la jerarquía de importancias obtenida mediante el árbol de decisión refleja una mayor sensibilidad a la muestra de entrenamiento, lo que es coherente con la naturaleza no paramétrica del modelo. Se puede observar en la Figura 39 como muchas de las variables tienen una importancia muy reducida. Por ello, se estimó un modelo alternativo empleando únicamente las variables más relevantes, con el fin de evaluar la robustez del modelo y su capacidad de generalización bajo una estructura más parsimoniosa. En concreto, se seleccionan las 4 variables con mayor importancia para el nuevo modelo.

Tabla 9. Rendimiento en validación cruzada del Árbol de Decisión del modelo base vs el modelo reducido.

ÁRBOL	R^2 CV	RMSE CV
COMPLETO	0.900	0.254
REDUCIDO	0.901	0.253

La comparación de rendimiento entre el árbol completo y el árbol reducido, resumida en la Tabla 9, muestra que ambos modelos presentan un rendimiento muy similar tras la validación cruzada, siendo incluso ligeramente superior en el modelo reducido con cuatro variables. Las diferencias observadas en R^2 y RMSE son numéricamente muy pequeñas, lo que indica que la eliminación de variables con baja importancia no solo no conlleva una pérdida apreciable de capacidad predictiva, sino que puede contribuir a una ligera mejora de la generalización. Este resultado confirma que la mayor parte de la señal explicativa se concentra en un subconjunto reducido de predictores y refuerza la estabilidad de la jerarquía de importancias obtenida.

En conjunto, el árbol de decisión proporciona una mejora notable en la capacidad explicativa respecto al modelo lineal, al ser capaz de capturar relaciones no lineales entre las variables. Sin embargo, la diferencia observada entre el rendimiento en entrenamiento y en *test*, junto con la sensibilidad del modelo a la complejidad de su estructura, sugiere la presencia de varianza elevada, una característica habitual de los árboles de decisión individuales.

Estas características indican que, si bien el árbol de decisión resulta útil para explorar estructuras complejas en los datos, puede beneficiarse de enfoques adicionales orientados a reducir la varianza del modelo y mejorar su estabilidad.

Random Forest.

Con el objetivo de mejorar la capacidad de generalización observada en el árbol de decisión individual, se empleó un modelo de *Random Forest*, basado en la agregación de múltiples árboles entrenados sobre submuestras aleatorias del conjunto de datos y subconjuntos aleatorios de variables explicativas. Este enfoque permite reducir la varianza del modelo a costa de un ligero incremento del sesgo, mejorando habitualmente el rendimiento fuera de muestra.

La selección de los hiperparámetros se llevó a cabo mediante un procedimiento de *Grid Search* con validación cruzada de 5 pliegues**,** optimizando el coeficiente de determinación R^2. En particular, se consideraron como hiperparámetros clave el número de árboles del bosque ($n\_estimators$), la profundidad máxima de cada árbol (max_depth), el tamaño mínimo de los nodos hoja ($min\_samples\_leaf$) y la proporción de variables consideradas en cada partición ($max\_features$). Los diferentes hiperparámetros probados figuran en la Tabla 10.

Tabla 10. Valores de rejilla utilizados para la obtención de los hiperparámetros óptimos en el Random Forest.

HIPERPARÁMETRO	VALORES DE REJILLA (GRID SEARCH)
Nº DE ÁRBOLES	{250, 400, 500}
PROFUNDIDAD	{10, 15, 25}
MÍNIMO POR HOJA	{1, 5, 10}
PROP. DE VARIABLES	{sqrt, 1/3, 0.5, 1}

Los resultados del proceso de optimización indican que la configuración óptima del modelo corresponde a un *Random Forest* compuesto por 500 árboles, con una profundidad máxima de 25, un tamaño mínimo de nodo hoja igual a 1 y una proporción de variables consideradas en cada división del 50% del total de predictores. Esta combinación maximiza el valor medio de R^2 en validación cruzada, permitiendo capturar relaciones complejas entre las variables explicativas sin incurrir en un sobreajuste severo.

Tabla 11. Métricas de rendimiento para la evaluación del Random Forest.

CONJUNTO	R^2	RMSE
ENTRENAMIENTO	0.992	0.070
TEST	0.946	0.187

El rendimiento del modelo *Random Forest* se evaluó mediante las métricas R^2 y RMSE en los conjuntos de entrenamiento y *test*. En el conjunto de entrenamiento, el modelo alcanza un valor de R^2 de 0.992, lo que indica que es capaz de explicar aproximadamente el 99% de la variabilidad del logaritmo del precio, junto con un RMSE reducido de 0.07, reflejando un ajuste muy preciso sobre los datos observados. En el conjunto de *test*, el valor de R^2 se sitúa en 0.946, lo que implica que el modelo explica en torno al 95% de la variabilidad fuera de muestra. Aunque se observa una disminución respecto al rendimiento *in-sample*, este comportamiento es consistente con la elevada flexibilidad del modelo y no resulta indicativo de un sobreajuste severo. De hecho, el RMSE en *test*, con un valor de 0.187, se mantiene en niveles moderados, lo que confirma una mejora sustancial en la capacidad predictiva respecto al modelo lineal.

Para evaluar la estabilidad del modelo, se aplicó un procedimiento de validación cruzada con 5 pliegues sobre el conjunto de entrenamiento. Los resultados obtenidos muestran un valor medio de R^2 de 0.941, con una desviación estándar de 0.004, lo que indica que el modelo explica de forma consistente aproximadamente el 94% de la variabilidad del logaritmo del precio a lo

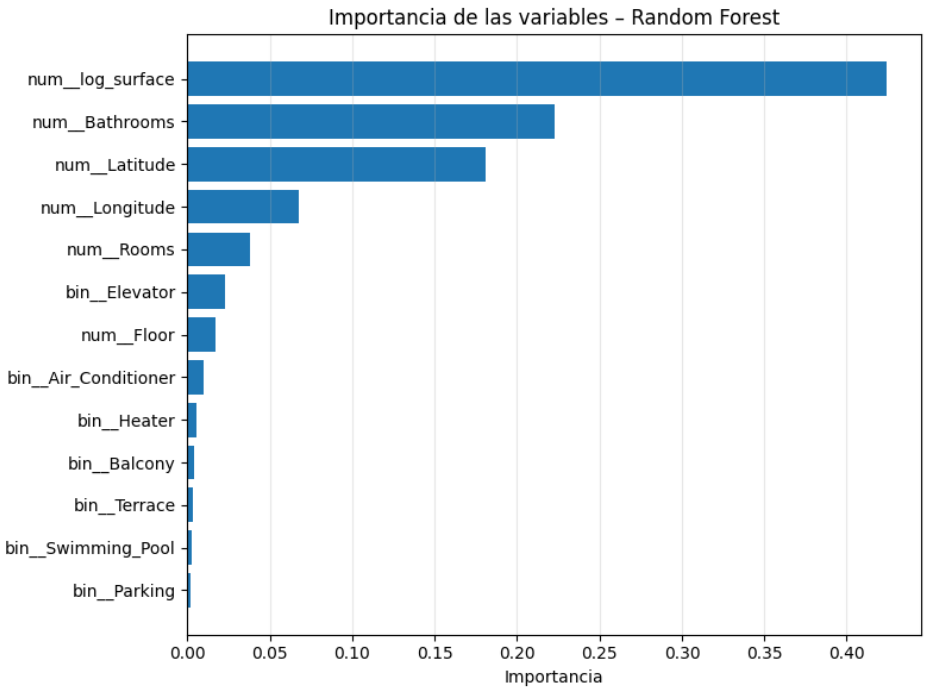
largo de las distintas particiones. De forma complementaria, el error cuadrático medio presenta un valor medio de RMSE de 0.195, con una desviación estándar reducida de 0.008, lo que confirma que el rendimiento del modelo es estable y poco sensible a la división concreta de los datos en entrenamiento y validación.

A partir del análisis visual del comportamiento de los principales hiperparámetros (Figuras A.7, A.8, A.9 y A.10 del Anexo) y con el objetivo de disponer de una alternativa menos compleja al modelo optimizado, se definió una configuración heurística del *Random Forest*. En concreto, se seleccionaron 300 árboles, una profundidad máxima de 20, una proporción de variables consideradas en cada división del 50%, y un tamaño mínimo de nodo hoja de 5 observaciones. Esta elección se sitúa en regiones del espacio de búsqueda con rendimiento elevado y baja variabilidad, introduciendo un mayor grado de regularización respecto a la solución óptima.

El *Random Forest* con selección heurística de hiperparámetros alcanza un valor de R^2 de 0.962 en el conjunto de entrenamiento y de 0.938 en el conjunto de *test* (el modelo es capaz de explicar aproximadamente un 96% de la variabilidad del logaritmo del precio), con valores de RMSE de 0.156 y 0.200, respectivamente. Estos resultados confirman una elevada capacidad predictiva fuera de muestra, aunque ligeramente inferior a la obtenida por el modelo optimizado.

En comparación con la configuración óptima, el modelo heurístico presenta una reducción moderada del rendimiento, a cambio de una estructura más parsimoniosa y un menor riesgo de sobreajuste, lo que refuerza la utilidad de este enfoque como alternativa estable y bien regularizada.

Figura 40. Importancia de variables, en porcentaje, en el Random Forest.



La Figura 40 muestra la importancia de las variables obtenida a partir del modelo *Random Forest*, calculada mediante la reducción media de la impureza. Los resultados indican que el logaritmo de la superficie ("num_log_surface") es, con diferencia, el predictor más relevante del precio de

la vivienda. Le siguen variables estructurales como el número de baños (“num_Bathrooms”) y las coordenadas geográficas (latitud - ”num_Latitude” - y longitud - ”num_Longitude” -), lo que pone de manifiesto la importancia conjunta de las características físicas y de la localización.

En comparación con el árbol de decisión individual, el *Random Forest* presenta una distribución más equilibrada de la importancia entre los predictores, reduciendo la dependencia excesiva de una única variable dominante y ofreciendo una interpretación más estable y robusta, coherente con su mejor capacidad de generalización.

Con el objetivo de analizar el impacto de una reducción del conjunto de variables explicativas, se estimó un modelo *Random Forest* utilizando únicamente las cuatro variables más relevantes identificadas en el análisis de importancia. El rendimiento de este modelo se comparó con el *Random Forest* completo, considerado como referencia.

El modelo reducido alcanza un valor de R^2 de 0.991 en el conjunto de entrenamiento y de 0.937 en el conjunto de *test*, con valores de RMSE de 0.076 y 0.201, respectivamente. Estos resultados reflejan una muy ligera pérdida de capacidad predictiva en la muestra *test* respecto al modelo completo manteniendo, no obstante, un nivel de ajuste elevado fuera de muestra a pesar de la reducción de complejidad del modelo.

La validación cruzada de 5 pliegues confirma esta tendencia, con un R^2 medio de 0.933 y una desviación estándar reducida (0.006), junto con un RMSE medio de 0.209, lo que indica un comportamiento estable entre particiones. En conjunto, estos resultados sugieren que un subconjunto reducido de variables es capaz de capturar gran parte de la estructura del precio de la vivienda, a costa de una pérdida moderada de precisión, ofreciendo una alternativa más parsimoniosa y fácilmente interpretable.

En conjunto, el modelo *Random Forest* logra una mejora sustancial en la capacidad predictiva respecto al árbol de decisión individual. Mientras que el árbol presenta una elevada capacidad de ajuste *in-sample* pero una pérdida apreciable de rendimiento fuera de muestra, el *Random Forest* reduce de forma notable esta varianza al combinar múltiples árboles entrenados sobre distintas submuestras de los datos. Como resultado, el modelo ofrece un mejor equilibrio entre capacidad explicativa y generalización, manteniendo una elevada precisión predictiva en contextos caracterizados por relaciones complejas y no lineales entre las variables.

XGBoost.

Una vez más, con el objetivo de superar las limitaciones presentadas por el árbol de decisión, se empleó un modelo de *XGBoost***,** una implementación avanzada de los métodos de *Gradient Boosting* que construye árboles de forma secuencial, incorporando regularización explícita para controlar la complejidad del modelo.

Con el objetivo de determinar una configuración adecuada de hiperparámetros, se llevó a cabo un proceso de optimización mediante validación cruzada de 5 pliegues, considerando como hiperparámetros principales el número de árboles ($n\_estimators$), la profundidad máxima (max_depth), la tasa de aprendizaje ($learning\_rate$) y los parámetros de muestreo de observaciones y variables ($subsample$ y $colsample\_bytree$). La rejilla de valores probados figura en la Tabla 12. El procedimiento de búsqueda sistemática condujo a una configuración óptima compuesta por 500 árboles, una profundidad máxima de 7, una tasa de aprendizaje con

un valor de 0.05, una proporción de observaciones de 0.8 para entrenar cada árbol ($subsample$) y una proporción de variables de 0.6 seleccionadas aleatoriamente para cada árbol ($colsample\_bytree$).

Tabla 12. Valores de rejilla utilizados para la obtención de los hiperparámetros óptimos en el XGBoost.

HIPERPARÁMETRO	VALORES DE REJILLA (GRID SEARCH)
Nº DE ÁRBOLES	{200, 300, 500}
PROFUNDIDAD	{5, 7, 10}
TASA DE APRENDIZAJE	{0.05, 0.1}
PPOP. DE OBSERVACIONES	{0.6, 0.8, 1}
PROP. DE VARIABLES	{0.6, 0.8, 1}

De forma complementaria, se analizó el comportamiento del número de árboles mediante representaciones gráficas del rendimiento medio en validación cruzada (Figura A.11 del Anexo). Dicho análisis mostró que una reducción del número de árboles de 500 a 300 conlleva un ahorro computacional moderado, pero también una ligera disminución del rendimiento predictivo, observable tanto en el coeficiente de determinación como en el error cuadrático medio. Dado que el coste adicional asociado al uso de 500 árboles no resulta prohibitivo en el contexto del presente estudio, se adoptó esta configuración como solución final, priorizando el máximo rendimiento predictivo.

Tabla 13. Métricas de rendimiento para la evaluación del XGBoost.

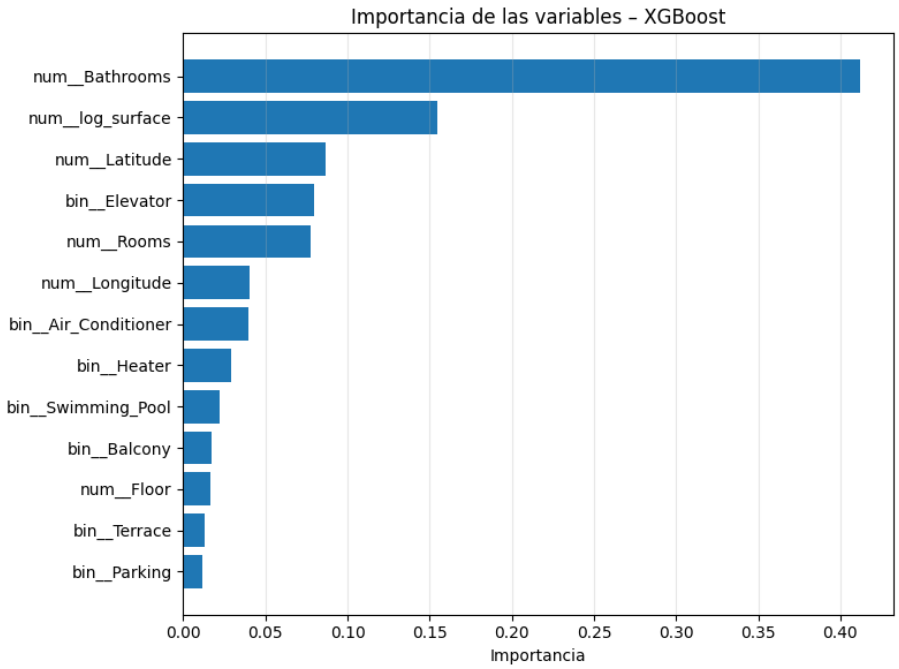
CONJUNTO	R^2	RMSE
ENTRENAMIENTO	0.990	0.079
TEST	0.950	0.179

El rendimiento del modelo *XGBoost* se evaluó mediante las métricas R^2 y RMSE en los conjuntos de entrenamiento y *test*. Los resultados muestran un ajuste elevado en el conjunto de entrenamiento, junto con un rendimiento igualmente alto fuera de muestra. En concreto, el valor de R^2 en el conjunto de test indica que el modelo es capaz de explicar aproximadamente el 95% de la variabilidad del logaritmo del precio, lo que supone una mejora respecto al árbol de decisión individual y un rendimiento comparable, e incluso ligeramente superior, al obtenido con el modelo *Random Forest*. Asimismo, el error de predicción medido mediante el RMSE se mantiene en niveles reducidos, lo que refleja una elevada precisión en la estimación del precio de la vivienda.

La estabilidad del modelo se evaluó mediante un procedimiento de validación cruzada *k-fold* con cinco particiones. Los resultados obtenidos muestran un valor medio de R^2 de 0.947, con una desviación estándar de 0.005, lo que indica que el modelo es capaz de explicar de forma consistente aproximadamente el 94.7% de la variabilidad del logaritmo del precio a lo largo de las distintas particiones. El error cuadrático medio presenta un valor medio de RMSE de 0.186,

con una desviación estándar reducida de 0.0095, lo que confirma la estabilidad del rendimiento predictivo del modelo y su capacidad de generalización frente a diferentes subconjuntos de entrenamiento y validación.

Figura 41. Importancia de variables, en porcentaje, en el XGBoost.



La Figura 41 muestra la importancia de las variables estimada a partir del modelo XGBoost. Los resultados indican que el número de baños y la superficie de la vivienda constituyen los principales determinantes del precio, seguidos de las coordenadas geográficas y de determinadas características estructurales como la presencia de ascensor. En comparación con el árbol de decisión y el modelo *Random Forest*, el *XGBoost* presenta una distribución más equilibrada de la importancia entre los predictores, reflejando su capacidad para capturar interacciones no lineales complejas entre las variables explicativas.

Con el objetivo de analizar el efecto de la reducción del número de predictores, se estimaron versiones simplificadas del modelo *XGBoost* utilizando únicamente las variables con mayor importancia identificadas en el modelo completo, manteniendo constantes el resto de hiperparámetros. En particular, se consideraron configuraciones con seis y cinco variables explicativas.

El modelo reducido con seis variables mantiene un rendimiento elevado, alcanzando un valor de R^2 en el conjunto de test cercano a 0.94, frente al 0.95 del modelo completo, lo que supone una pérdida leve de capacidad explicativa. El error cuadrático medio en *test* se sitúa en torno a 0.195, y los resultados de validación cruzada confirman una buena estabilidad, con valores medios de R^2 próximos a 0.94 y baja variabilidad entre particiones.

Por el contrario, la reducción adicional a cinco variables implica un deterioro más apreciable del rendimiento, con un incremento del RMSE hasta aproximadamente 0.20 y una ligera disminución adicional del R^2. Este resultado sugiere que la variable excluida en esta configuración aporta información relevante, por lo que la simplificación adicional no resulta recomendable.

En conjunto, estos resultados indican que es posible reducir la complejidad del modelo sin comprometer de forma sustancial su rendimiento, aunque existe un umbral a partir del cual la pérdida de información se vuelve significativa.

En conjunto, el modelo *XGBoost* ofrece muy buen rendimiento predictivo, combinando una elevada capacidad explicativa con una buena estabilidad fuera de muestra. La incorporación de regularización y el entrenamiento secuencial de árboles permiten reducir la varianza observada en el árbol de decisión individual, manteniendo al mismo tiempo la flexibilidad necesaria para capturar relaciones no lineales complejas.

Estos resultados ponen de manifiesto el potencial de los métodos de *boosting* como herramientas avanzadas para la predicción del precio de la vivienda en contextos caracterizados por una elevada heterogeneidad y complejidad estructural.

Support Vector Regression.

Con el objetivo de explorar enfoques alternativos a los modelos basados en árboles y evaluar el rendimiento de métodos fundamentados en principios de regularización explícita, se introdujo el modelo *Support Vector Regression* (*SVR*). A diferencia de los árboles de decisión y de los modelos ensemble, el *SVR* aborda el problema de regresión desde una perspectiva geométrica y de optimización convexa, buscando estimar una función que maximice la capacidad de generalización mediante el control directo de la complejidad del modelo.

Este enfoque resulta especialmente adecuado en contextos donde se sospecha la existencia de relaciones no lineales entre las variables explicativas y el precio de la vivienda, y donde el control del sobreajuste es prioritario. Asimismo, el *SVR* constituye un referente clásico en la literatura estadística y de aprendizaje automático, lo que lo convierte en un punto de comparación natural frente a modelos más complejos como *Random Forest* y *XGBoost*.

La selección de hiperparámetros del modelo *SVR* se llevó a cabo mediante un procedimiento de optimización con validación cruzada, considerando los parámetros C, γ y ε. Los valores probados se encuentran en la Tabla 14. Los valores óptimos obtenidos coinciden con aquellos que mostraban un mejor comportamiento en el análisis heurístico previo de cada hiperparámetro de forma individual, lo que refuerza la consistencia del proceso de selección. En particular, el modelo final se estimó con $C = 1$, $\gamma =$ scale y $\varepsilon = 0.05$, incorporando un *kernel* radial y un preprocesamiento basado en estandarización de las variables numéricas.

Tabla 14. Valores de rejilla utilizados para la obtención de los hiperparámetros óptimos en el modelo SVR.

PARÁMETRO	VALORES DE REJILLA (GRID SEARCH)
C	{0.1, 1, 10, 100}
Γ	{scale, 0.01, 0.1, 1, 10, 100}
E	{0.001, 0.01, 0.05, 0.1}

El rendimiento del modelo se evaluó mediante las métricas R^2 y RMSE en los conjuntos de entrenamiento y *test*, cuyos resultados se resumen en la Tabla 15.

Tabla 15. Métricas de rendimiento para la evaluación del modelo SVR.

CONJUNTO	R^2	RMSE
ENTRENAMIENTO	0.939	0.2
TEST	0.92	0.23

Los resultados muestran un ajuste elevado en el conjunto de entrenamiento, con un valor de R^2 cercano a 0.94, lo que indica que el modelo es capaz de capturar una proporción significativa de la variabilidad del logaritmo del precio. En el conjunto de *test*, el valor de R^2 se sitúa en torno a 0.92, reflejando una pérdida moderada de capacidad explicativa fuera de muestra. Este comportamiento es coherente con la naturaleza del modelo *SVR* con *kernel RBF*, que ofrece una elevada flexibilidad, pero presenta una capacidad de generalización más limitada que los modelos *ensemble* empleados posteriormente.

Con el objetivo de evaluar la estabilidad del modelo, se aplicó un procedimiento de validación cruzada *k*-fold con cinco particiones sobre el conjunto de entrenamiento. Los resultados obtenidos muestran un valor medio de R^2 de 0.920, con una desviación estándar de 0.006, lo que indica una variabilidad moderada entre particiones. De forma complementaria, el error cuadrático medio presenta un valor medio de RMSE de 0.228, con una desviación estándar reducida de 0.012, lo que sugiere que el rendimiento del modelo es relativamente estable frente a distintas divisiones de los datos.

En conjunto, estos resultados exponen que el *SVR* constituye una alternativa sólida para la modelización no lineal del precio de la vivienda, aunque su rendimiento predictivo es inferior al alcanzado por los modelos basados en árboles, tanto en términos de capacidad explicativa como de error de predicción.

4.4.4. Comparación de resultados

Tabla 16. Comparación de resultados en validación cruzada de todos los modelos desarrollados.

CONJUNTO	R^2 CV	RMSE CV
REGRESIÓN LINEAL	0.7785	0.3787
SVR	0.9187	0.2297
ÁRBOL DE DECISIÓN	0.9005	0.2539
RANDOM FOREST	0.9414	0.1949
XGBOOST	0.9482	0.1833

La Tabla 16 recoge de forma conjunta las principales métricas de rendimiento obtenidas por los distintos modelos analizados, permitiendo una comparación directa de su capacidad predictiva y su comportamiento en los conjuntos de entrenamiento y prueba. A partir de estos resultados

se observan diferencias claras asociadas al grado de complejidad y a la capacidad de cada modelo para capturar relaciones no lineales en los datos.

Los modelos lineales, incluida la regresión lineal múltiple y sus variantes regularizadas (*Ridge* y *Lasso*), presentan un rendimiento correcto pero limitado. Si bien la regularización contribuye a estabilizar las estimaciones y a reducir el riesgo de sobreajuste, las mejoras respecto al modelo lineal base son prácticamente nulas, lo que sugiere que la relación entre las variables explicativas y el precio de la vivienda no es estrictamente lineal.

El modelo de *Support Vector Regression* (SVR) ofrece una mejora respecto a los enfoques puramente lineales al introducir no linealidades mediante el uso de funciones *kernel*. No obstante, sus resultados se sitúan en una posición intermedia en la comparación global, mostrando un rendimiento inferior al de los modelos *ensemble*.

Por su parte, los modelos basados en árboles de decisión capturan de forma más efectiva las interacciones y no linealidades presentes en los datos, lo que se traduce en una mejora apreciable de las métricas de error. Sin embargo, el árbol de decisión individual presenta una mayor diferencia entre el rendimiento en entrenamiento y prueba, evidenciando una mayor propensión al sobreajuste.

En este contexto, los modelos *Random Forest* y *XGBoost*, ofrecen los mejores resultados globales. Ambos alcanzan los menores errores de predicción y muestran una capacidad de generalización claramente superior. En especial, *XGBoost* destaca por combinar un alto rendimiento predictivo con una mayor estabilidad, consolidándose como el modelo más competitivo entre los analizados.

En conjunto, la comparación de resultados pone de manifiesto que el incremento de complejidad del modelo, cuando se controla adecuadamente, permite capturar patrones más complejos del mercado inmobiliario y mejorar significativamente la precisión de las predicciones. Estos resultados justifican la elección de modelos *ensemble* como herramienta principal para la predicción del precio de la vivienda en el marco de este estudio. En concreto, para la aplicación web se utilizará el modelo *XGBoost*.

5. APLICACIÓN WEB Y RESULTADOS INTERACTIVOS

Con el objetivo de trasladar los resultados del trabajo a un entorno interactivo, se desarrolló una aplicación web[14] utilizando el *framework Streamlit*[15], una herramienta de código abierto basada en *Python* que permite crear aplicaciones web orientadas a la visualización y análisis de datos de forma sencilla y eficiente.

La aplicación fue desarrollada inicialmente en un entorno local utilizando *Visual Studio Code* como editor principal y *Python* como lenguaje de programación. Esta fase permitió implementar de forma iterativa las distintas funcionalidades de la aplicación, incluyendo la carga de datos, la generación de visualizaciones interactivas y la integración del modelo predictivo previamente entrenado. Durante el desarrollo local, se realizaron pruebas continuas para verificar el correcto funcionamiento de la aplicación y asegurar la coherencia entre los resultados obtenidos en el entorno de análisis y los mostrados en la interfaz web.

Una vez finalizado el desarrollo local, la aplicación fue publicada utilizando la plataforma *Streamlit Cloud*, que permite desplegar aplicaciones directamente desde un repositorio de *GitHub*. Para ello, el código de la aplicación y los archivos necesarios (modelo entrenado, *datasets* y dependencias) fueron organizados en un repositorio estructurado, facilitando su correcta ejecución en un entorno remoto. El uso de *GitHub* como intermediario permitió gestionar versiones del código y asegurar la reproducibilidad de la aplicación. Tras el despliegue, la aplicación quedó accesible públicamente a través de un enlace web, permitiendo su uso sin necesidad de instalar *software* adicional ni disponer de conocimientos técnicos específicos.

Este proceso de despliegue evidencia la viabilidad de integrar modelos estadísticos y análisis de datos en aplicaciones accesibles, acercando los resultados del trabajo a un entorno de uso real.

La aplicación web desarrollada constituye el producto final del trabajo y permite al usuario interactuar con los datos y resultados obtenidos a lo largo del análisis. A través de una interfaz sencilla e intuitiva, la aplicación ofrece distintas secciones que facilitan tanto la exploración del comportamiento histórico del mercado inmobiliario como la obtención de predicciones personalizadas. La navegación por la aplicación se organiza en apartados claramente diferenciados, lo que permite acceder de forma independiente a las visualizaciones descriptivas y al módulo de predicción del precio de la vivienda.

La aplicación web incorpora las siguientes funcionalidades principales, que se pueden observar en la Figura 42:

- Menú de Inicio con una explicación sobre la utilización de la app y la memoria del trabajo en formato pdf. [**Inicio**]
- Visualización interactiva del análisis histórico del mercado inmobiliario mediante gráficos dinámicos. [**Análisis Histórico**]
- Exploración de los datos utilizados a lo largo del proyecto, que figuran en la web ya limpios para que el usuario pueda buscar u ordenarlos a su gusto. [**Datos**]

[14] https://tfg-eloy-celaya.streamlit.app/
[15] https://streamlit.io/

- Módulo de predicción del precio de la vivienda, donde el usuario introduce las características de una vivienda concreta y obtiene una estimación basada en el modelo entrenado. [**Modelización**]
- Presentación clara de los resultados, facilitando su interpretación sin necesidad de conocimientos técnicos avanzados. [**Predicción de Precio**]

Figura 42. Menú principal de la aplicación web desarrollada.

Uno de los principales valores añadidos de la aplicación es la posibilidad de interactuar directamente con los resultados. A diferencia de los gráficos estáticos incluidos en la memoria, la aplicación permite modificar entradas, explorar distintos escenarios y observar de forma inmediata cómo cambian los resultados mostrados. Este enfoque interactivo mejora la comprensión del comportamiento de los datos y del modelo, y refuerza la utilidad práctica del trabajo al permitir una exploración dinámica de los resultados.

Aunque la aplicación cumple con los objetivos planteados, presenta algunas limitaciones inherentes al enfoque adoptado. En particular, las predicciones generadas están condicionadas por el ámbito geográfico y temporal del conjunto de datos utilizado para el entrenamiento del modelo, por lo que su uso fuera de este contexto debe interpretarse con cautela. Asimismo, la aplicación está diseñada como una herramienta de apoyo al análisis y no como un sistema de valoración oficial, por lo que los resultados obtenidos deben entenderse como estimaciones aproximadas. Otra limitación es la limitada capacidad de procesamiento de *Streamlit*, que en ocasiones afecta al rendimiento al cargar visualizaciones más complejas, como los mapas interactivos. Al ser esta aplicación gratuita, tras un día de inactividad la web se desactiva para minimizar recursos. El usuario al entrar encontrará una pestaña indicando que la app está inactiva con un botón para volver a activarla. Tras seleccionar el botón, la app se reiniciará y funcionará adecuadamente.

6. CONCLUSIONES

El objetivo de este Trabajo de Fin de Grado ha sido analizar y modelizar el precio de la vivienda a partir de un conjunto amplio de variables estructurales y espaciales, combinando análisis exploratorio, modelización predictiva y desarrollo de una herramienta interactiva. A lo largo del trabajo se ha seguido un enfoque progresivo que abarca desde el análisis histórico del mercado inmobiliario hasta la aplicación de modelos avanzados de aprendizaje automático.

El análisis histórico ha permitido contextualizar los datos y comprender la evolución espacial y temporal de los precios, identificando patrones relevantes que sirven de base para la modelización posterior. Este enfoque resulta fundamental para interpretar adecuadamente los resultados predictivos y evitar conclusiones aisladas del contexto del mercado.

En cuanto a la modelización, los resultados muestran que los enfoques capaces de capturar relaciones no lineales ofrecen un rendimiento superior frente a modelos más restrictivos. En este sentido, los métodos *ensemble* destacan por su capacidad de generalización y precisión, consolidándose como la opción más adecuada para el problema considerado. Estos resultados confirman que la complejidad del mercado inmobiliario requiere modelos flexibles, siempre que su uso esté debidamente controlado.

Como resultado aplicado del trabajo, se ha desarrollado una aplicación web interactiva que integra tanto el análisis histórico como el modelo predictivo final. Esta herramienta permite visualizar la información de forma intuitiva y generar estimaciones personalizadas del precio de la vivienda, reforzando la utilidad práctica del estudio y facilitando la transferencia de los resultados a un entorno real.

Más allá del rendimiento predictivo, este trabajo pone de relieve consideraciones metodológicas y éticas relevantes. Variables espaciales como la latitud y la longitud pueden actuar como *proxies* de características socioeconómicas del entorno, lo que exige cautela desde la perspectiva del *fairness* en *machine learning*. Como líneas de trabajo futuro, se propone explorar representaciones espaciales alternativas, como la agregación por zonas, el uso de clústeres espaciales o la aplicación de métodos específicos de modelización espacial, como la *Geographically Weighted Regression* (GWR) (Brunsdon, Fotheringham, & Charlton, 2002).

En conjunto, este Trabajo de Fin de Grado evidencia el valor de combinar análisis histórico, modelización avanzada y herramientas interactivas para el estudio del mercado inmobiliario. La integración de rigor metodológico, capacidad predictiva y aplicabilidad práctica refuerza la relevancia de los resultados obtenidos y sienta las bases para desarrollos futuros en contextos reales.

REFERENCIAS

Abbott, S. (2001). Bivariate Data Analysis: A Practical Guide. *Mathematical Gazette, 85*(502), 184-185.

Aggarwal, C. C. (2017). *Outlier Analysis* (2 ed.). Springer Cham. doi:https://doi.org/10.1007/978-3-319-47578-3

Alberto, A., & Di Lecce, V. (2023). Data preprocessing impact on machine learning algorithm performance. *Open Computer Science, 13*(1). doi:https://doi.org/10.1515/comp-2022-0278

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. *2021 IEEE Congress on Evolutionary Computation (CEC)* (págs. 1551-1559). IEEE. doi:10.1109/CEC45853.2021.9504761

Álvarez, N. (2011). *Modelos geoestadísticos del precio de la vivienda: aproximación al conocimiento intraurbano de la ciudad de Madrid.* Tesis doctoral, Universidad Autónoma de Madrid.

Amri, S., & Tularam, G. (2012). Performance of Multiple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices. *Journal of Mathematics & Statistics, 8*(4), 419-434. doi:https://doi.org/10.3844/jmssp.2012.419.434

Ayuntamiento de Madrid. (2024). *Precio medio declarado de la vivienda (euros/m2) por Distrito y Barrio según Tipo de vivienda.* Obtenido de https://servpub.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=0504020100060

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197-227. doi:https://doi.org/10.1007/s11749-016-0481-7

Brunsdon, C., Fotheringham, S., & Charlton, M. (2002). Geographically Weighted Regression. *Journal of the Royal Statistical Society, 47*(3), 431-443. doi:https://doi.org/10.1111/1467-9884.00145

C. Montgomery, D., A. Peck, E., & Geoffrey Vining, G. (2012). *Introduction to Linear Regression Analysis.* John Wiley & Sons.

Centro Nacional de Información Geográfica. (2025). *Límites municipales, provinciales y autonómicos*. Obtenido de Centro de Descargas CNIG: https://centrodedescargas.cnig.es/CentroDescargas/limites-municipales-provinciales-autonomicos

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Cornell University. doi:https://doi.org/10.48550/arXiv.1603.02754

Cleff, T. (2019). *Applied Statistics and Multivariate Data Analysis for Business and Economics: A Modern Approach Using SPSS, Stata, and Excel* (1st ed. 2019 ed.). Springer Nature. doi:https://doi.org/10.1007/978-3-030-17767-6

Denis, D. J. (2021). *Applied Univariate, Bivariate, and Multivariate Statistics Using Python.* John Wiley & Sons.

Dhummad, S. (2023). The Imperative of Exploratory Data Analysis in Machine Learning. *Scholars Journal of Engineering and Technology, 13*(1), 30-44. doi:https://doi.org/10.36347/sjet.2025.v13i01.005

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., . . . Ranjan. (2023). Explainable AI (XAI): Core Ideas, Techniques and Solutions. *ACM Computing, 55*(9), 1-33. doi:http://dx.doi.org/10.114 5/3561048

Fernández Armas, A. (2023). *Madrid real state prices*. Obtenido de Kaggle: https://www.kaggle.com/datasets/alefernandezarmas/madrid-real-state-prices?resource=download

Genuer, R., Poggi, J.-M., & Taleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225-2236. doi:https://doi.org/10.1016/j.patrec.2010.03.014

James, G., Daniela Written, Hastie, T., & Tibhsirani, R. (2021). *An Introduction to Statistical Learning.* Springer.

Joshi, A. P., & Patel, D. B. (2020). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Computer Science and Technology, 13*(0203), 78-81. doi:http://dx.doi.org/10.13005/ojcst13.0203.03

Kahane, L. H. (2014). *Regression Basics.* SAGE Publications, Inc. doi:https://doi.org/10.4135/9781483385662

Love, P. E., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2023). Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics, 59*. doi:https://doi.org/10.1016/j.aei.2023.102024

Martínez Pagés , J., & Maza, L. Á. (2003). *Análisis del precio de la vivienda en España.* Banco de España. Obtenido de https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/03/Fic/dt0307.pdf

Matplotlib development team. (2025). *Matplotlib*. Obtenido de https://matplotlib.org/stable/index.html

Michelucci, U. (2024). Model Validation and Selection. In: Fundamental Mathematical Concepts for Machine Learning in Science. Springer, Cham. doi:https://doi.org/10.1007/978-3-031-56431-4_7

Ministerio de Transportes, Movilidad y Agenda Urbana. (2025). *Valor tasado de vivienda libre de los municipios mayores de 25000 habitantes.* Obtenido de https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=35000000

Myles, A., Feudale, R., Liu, Y., Woody, N., & Brown, S. (2004). An introduction to decision tree modeling. *Journal of Chemometrics, 18*(6), 275-285. doi:https://doi.org/10.1002/cem.873

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *8th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2012). 7376*, págs. 154-168. Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-642-31537-4_13

Plotly Technologies. (2025). *Plotly Open Source Graphing Library for Python*. Obtenido de https://plotly.com/python/

Quiñonero, D. R. (2016). *Análisis de la estructura espacial del precio de la vivienda en la ciudad de Madrid.* TFM, Universidad Politécnica de Cartagena. Obtenido de http://hdl.handle.net/10317/5651

scikit-learn developers. (2025). *scikit learn*. Obtenido de https://scikit-learn.org/stable/

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222. doi:https://doi.org/10.1023/B%3ASTCO.0000035301.49549.88

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghao archives of psychiatry, 27*(2), 130-135. doi:https://doi.org/10.11919/j.issn.1002-0829.215044

Streamlit, Inc. (2025). *Get started with Streamlit*. Obtenido de Streamlit: https://docs.streamlit.io/get-started

The pandas development team. (2025). *User guide*. Obtenido de pandas: https://pandas.pydata.org/docs/user_guide/index.html

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley Publishing Company.

Vies, J., Monllor, P., & Carrasco, F. (2023). Análisis espacial del precio de la vivienda en España: evidencia a nivel municipal. *Revista de Estudios Empresariales. Segunda Época*, 109-126. doi:https://doi.org/10.17561/ree.n2.2023.7854

Wang, H., K. Roy Chowdhury, P., Yoon, J., Bhaduri, P., Srikrishnan, V., Judi, D., & Daniel, B. (2025). Explaining drivers of housing prices with nonlinear hedonic regressions. *Machine Learning with Applications, 13*, 100707. doi:https://doi.org/10.1016/j.mlwa.2025.100707

West, R. M. (2021). Best practice in statistics: The use of log transformation. *Ann Clin Biochem, 59*(3), 162-165. doi:https://doi.org/10.1177/00045632211050531

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming*, 1-9. doi:https://doi.org/10.1155/2021/7678931

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*(2), 301-320. doi:https://doi.org/10.1111/j.1467-9868.2005.00503.x

ANEXO

Con el fin de no sobrecargar el cuerpo principal del trabajo y mantener una extensión adecuada del manuscrito, en este anexo se incluyen algunas gráficas y resultados complementarios que no se presentan en el texto principal. Estas representaciones sirven como apoyo al análisis realizado y permiten ampliar o matizar determinados aspectos metodológicos y empíricos, sin que su omisión en el manuscrito principal afecte a la comprensión ni a las conclusiones fundamentales del estudio.

Figura A. 1. Gráfico Q-Q normal de los residuos del modelo de Regresión Lineal al realizar transformaciones en las variables "Bathrooms" y "Rooms".

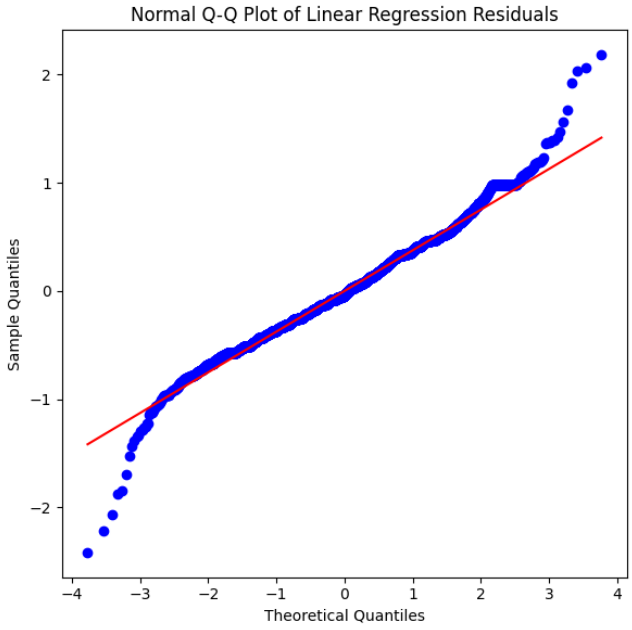


Figura A. 2. Gráfico de residuos frente a valores ajustados por el modelo de Regresión Lineal al realizar transformaciones en las variables "Bathrooms" y "Rooms".

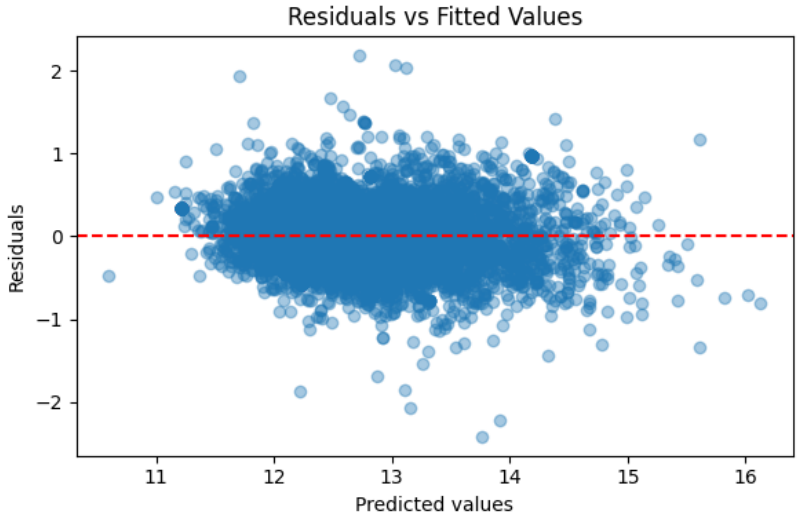


Figura A. 3. Efecto marginal de cada variable predictora sin transformar las variables "Bathrooms" y "Rooms".

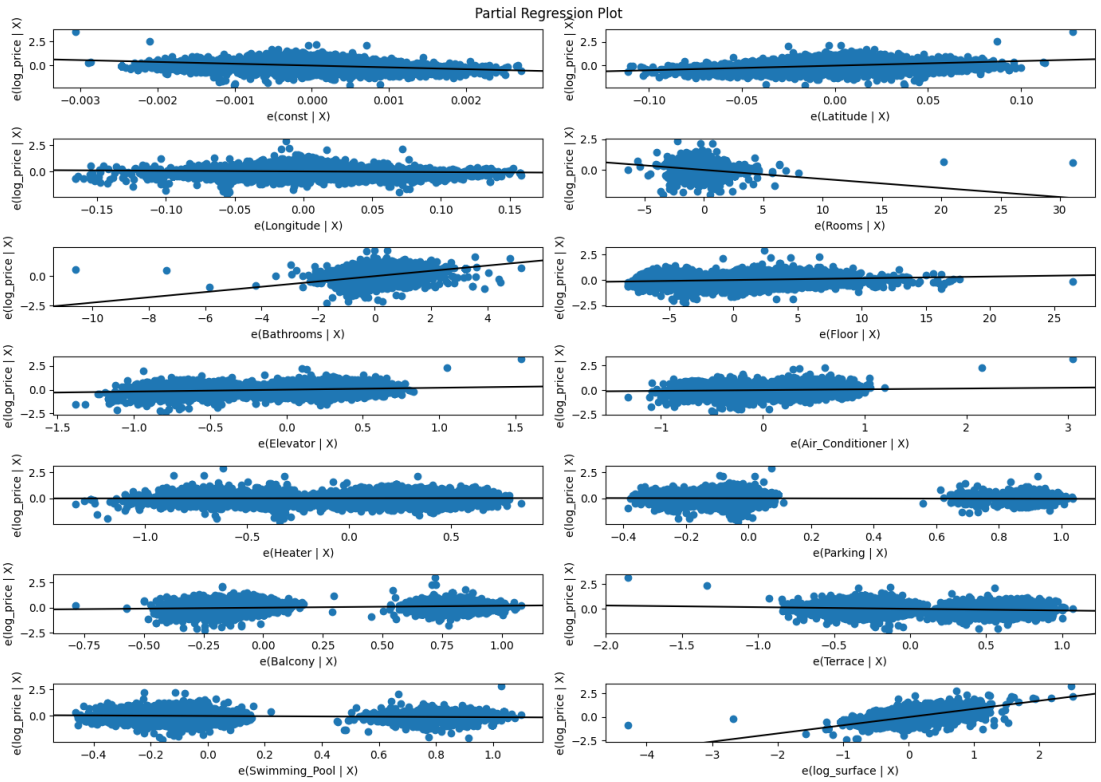


Figura A. 4. Efecto marginal de cada variable predictora al transformar las variables "Bathrooms" y "Rooms".

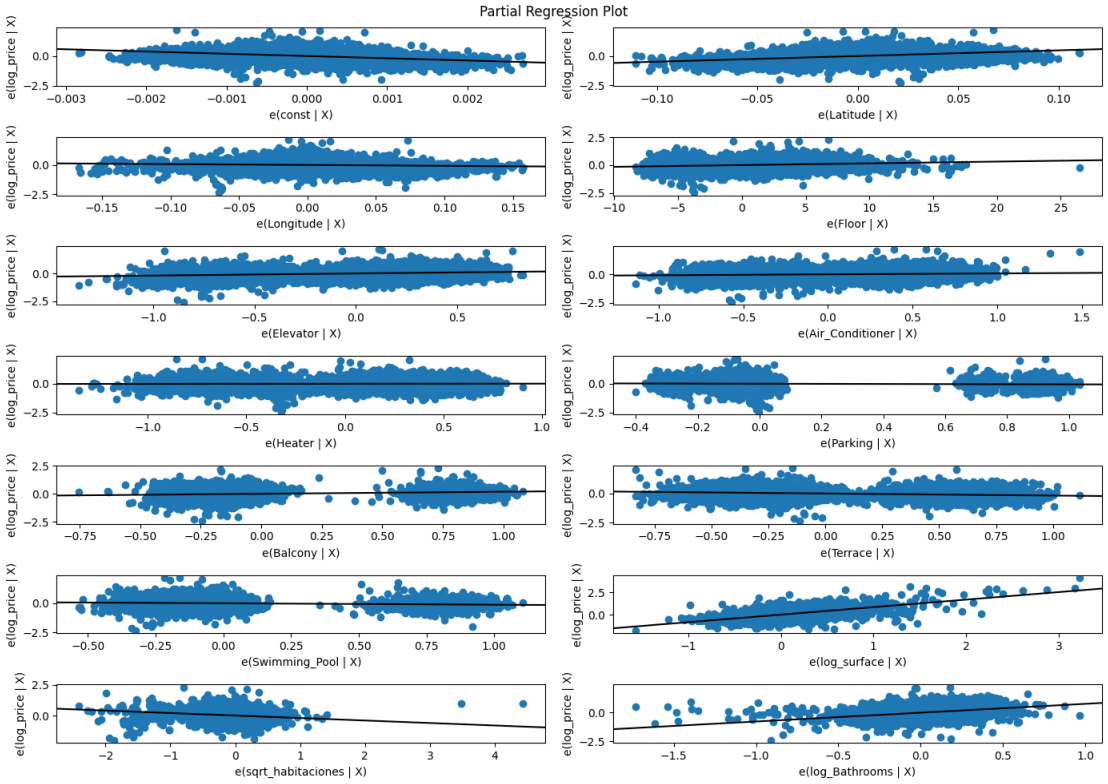


Figura A. 5. Valor de R^2 medio para diferentes combinaciones de profundidad máxima en árbol de decisión.

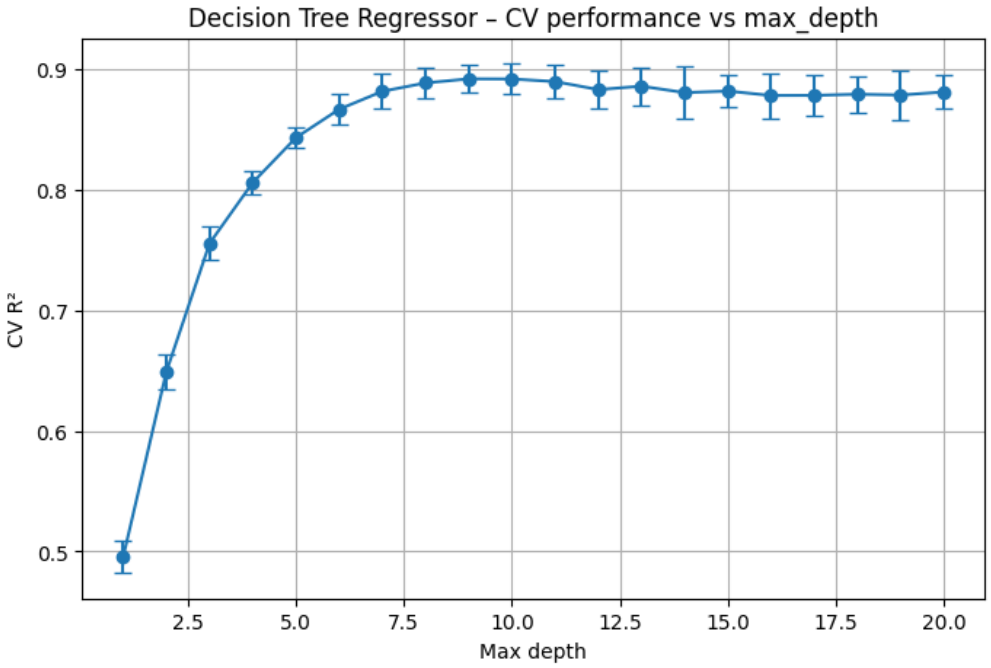


Figura A. 6. Valor de R^2 medio para diferentes combinaciones del mínimo número de observaciones por hoja en árbol de decisión.

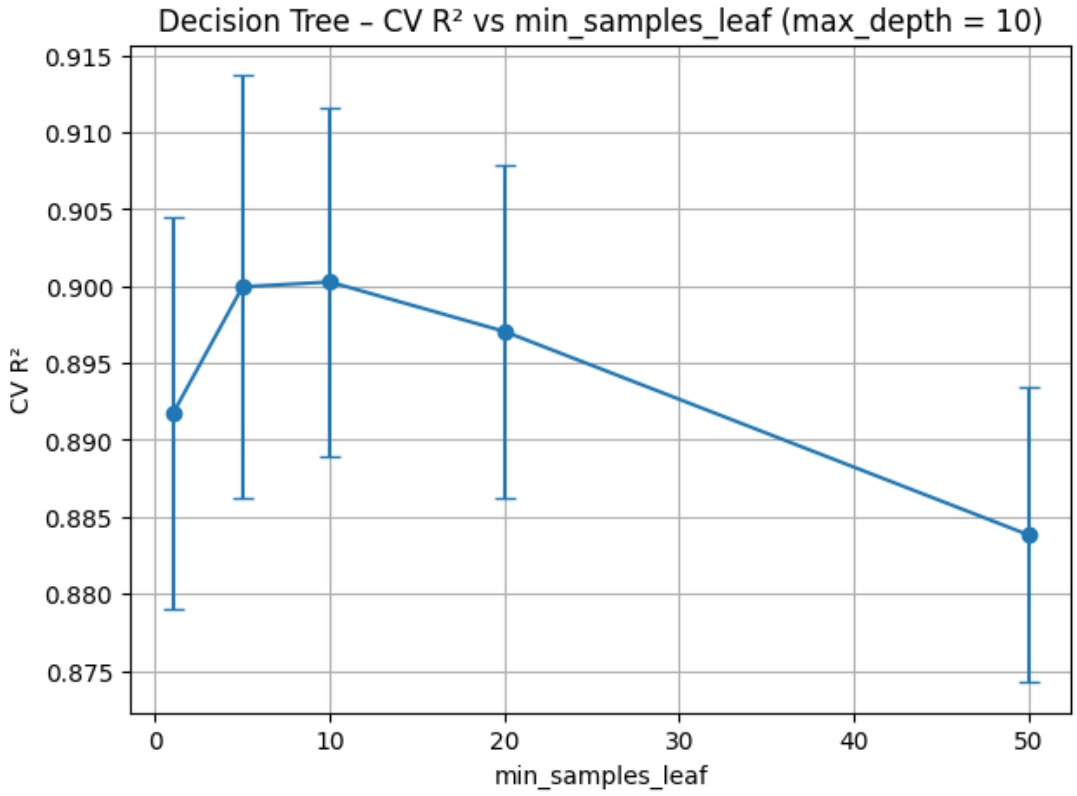


Figura A. 7. Valor de R^2 medio para diferentes combinaciones del número de árboles en Random Forest.

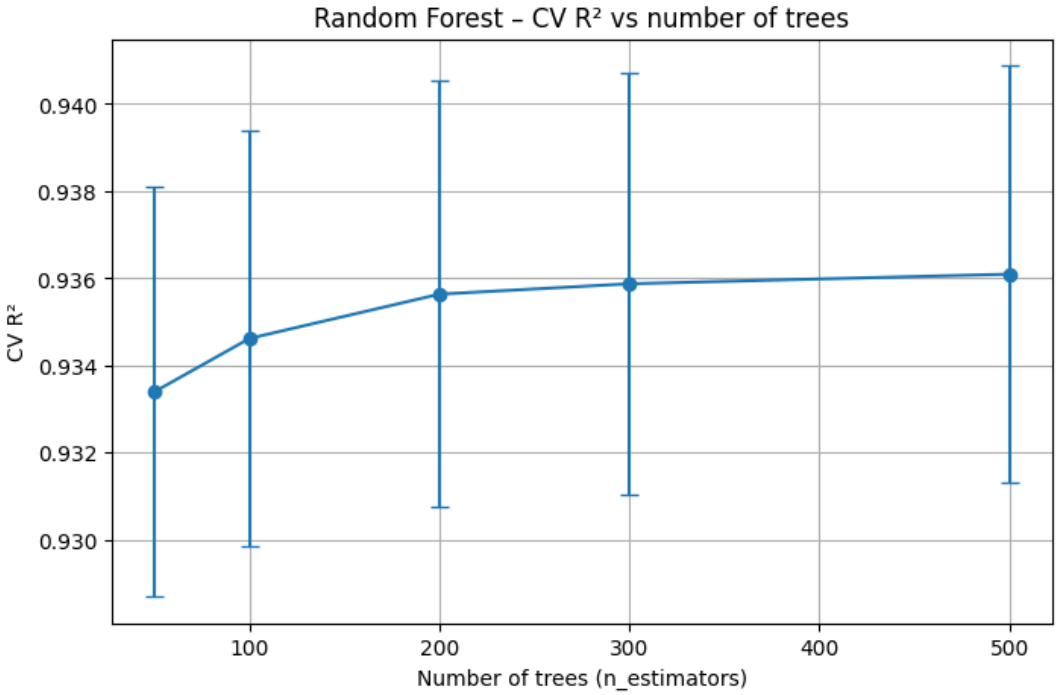


Figura A. 8. Valor de R^2 medio para diferentes combinaciones de la profundidad máxima en Random Forest.

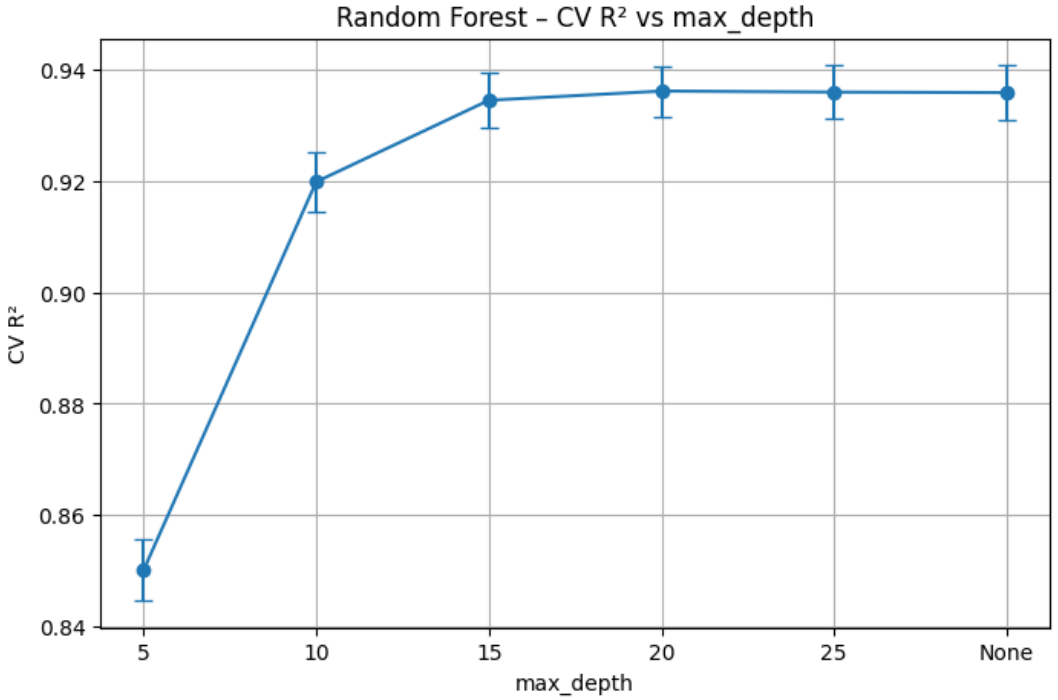


Figura A. 9. Valor de R^2 medio para diferentes combinaciones del número máximo de variables a probar en cada nodo en Random Forest.

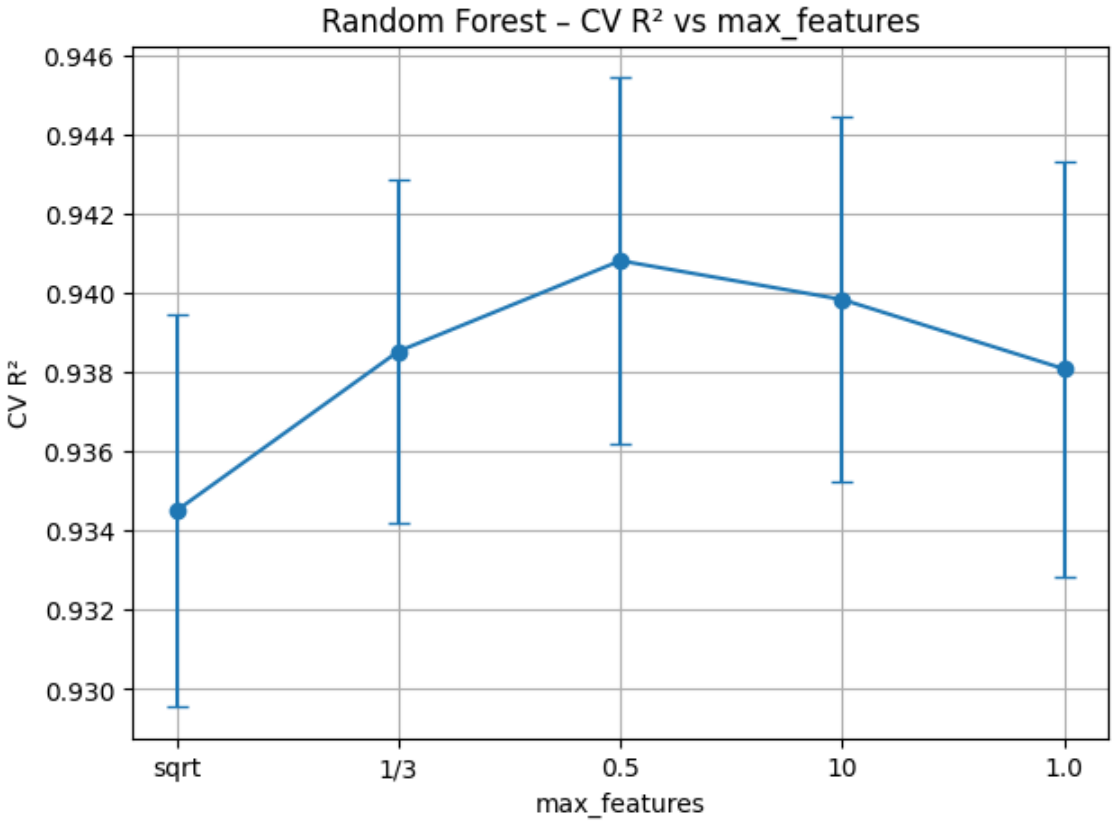


Figura A. 10. Valor de R^2 medio para diferentes combinaciones del mínimo número de observaciones por hoja en Random Forest.

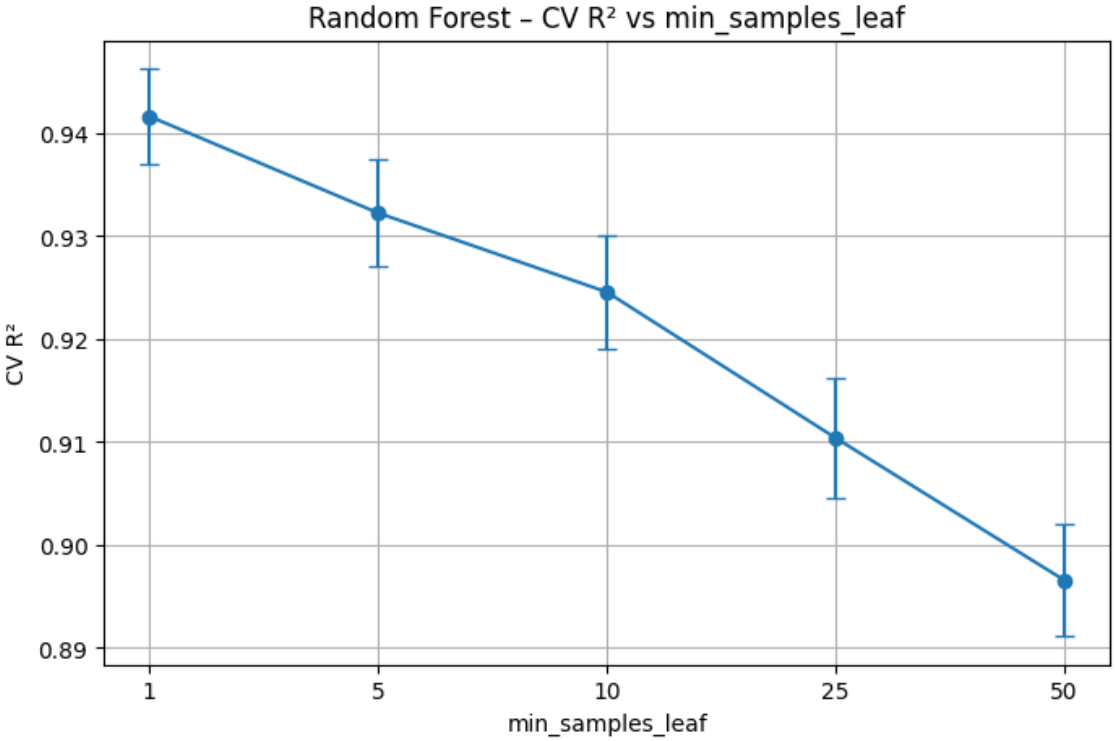
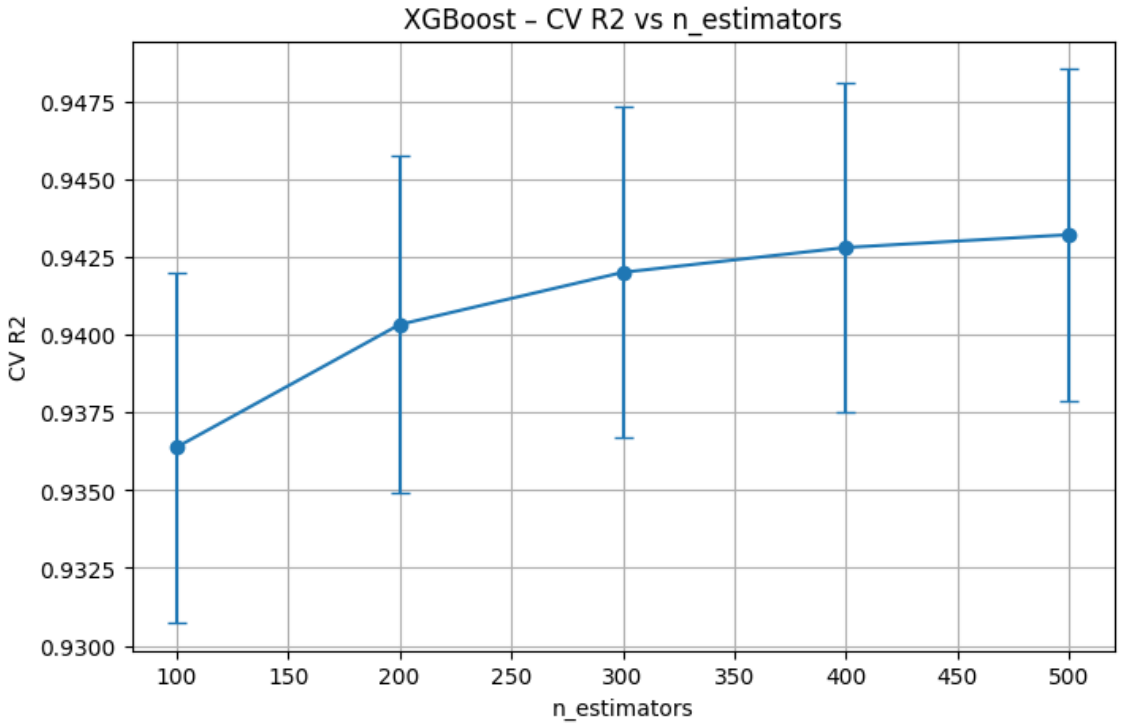


Figura A. 11. Valor de R^2 medio para diferentes combinaciones del número de árboles en XGBoost.



DECLARACIÓN DE USO DE INTELIGENCIA ARTIFICIAL GENERATIVA (IAG) EN EL TRABAJO DE FIN DE GRADO (TFG)

He usado IAG en mi TFG

Marca lo que corresponda:

SI	**NO**

Si has marcado SI, completa las siguientes 3 partes de este documento:

Parte 1: declaración sobre comportamiento legal, ético y responsable

Ten presente que el uso de IAG conlleva unos riesgos y puede generar una serie de consecuencias académicas graves: la evaluación de tu TFG puede verse comprometida si el uso de la IAG comporta la utilización de datos de carácter confidencial, materiales protegidos por derechos de autoría, o datos de carácter personal, y se hace sin cumplir las condiciones exigidas en cada caso (autorización de los interesados, autorización de los titulares, seguimiento de las instrucciones de la Universidad).

Pregunta	
1. En mi interacción con herramientas de IAG he facilitado **datos de carácter confidencial** contando siempre con la debida autorización de los interesados. La confidencialidad abarca toda información que una persona u organización desea proteger por razones legales, comerciales, de privacidad o estratégicas (como patentes o secretos comerciales).	
SÍ, he usado estos datos con la autorización de los interesados	NO, no he usado datos de carácter confidencial
2. En mi interacción con herramientas de IAG he facilitado **materiales protegidos por derechos de autoría** contando siempre con la autorización de los respectivos titulares.	
SÍ, he usado estos materiales con autorización de los titulares de derechos de autor; o bien sin ella porque se ajustan a una de las excepciones o límites que permite la ley: ● obra en dominio público ● obra licenciada (licencias Creative Commons) ● uso de fragmentos con fines de investigación (derecho de cita)	NO, no he usado materiales protegidos por derechos de autoría

3. En mi interacción con herramientas de IAG he facilitado **datos de carácter personal** con la debida autorización de los interesados.	
SÍ, he usado estos datos con autorización de los interesados y conforme a las instrucciones contenidas en la guía aprobada por la Universidad	NO, no he usado datos de carácter personal
4. Mi utilización de la herramienta de IAG ha **respetado sus términos de uso**, así como los principios éticos esenciales, no orientándola de manera maliciosa a obtener un resultado inapropiado para el trabajo presentado, es decir, que produzca una impresión o conocimiento contrario a la realidad de los resultados obtenidos, que suplante mi propio trabajo o que pueda resultar en un perjuicio para las personas.	
SI	NO

Parte 2: declaración de uso técnico

Utiliza el siguiente modelo de declaración tantas veces como sea necesario, a fin de reflejar todos los tipos de iteración que has tenido con herramientas de IAG. Incluye un ejemplo por cada tipo de uso realizado donde se indique: *[Añade un ejemplo].*

Declaro haber hecho uso del sistema de IAG: ChatGPT, en su versión 5.2 **para:**

Documentación y redacción:

- *Soporte a la reflexión en relación con el desarrollo del trabajo: proceso iterativo de análisis de alternativas y enfoques utilizando la IAG*

He utilizado la IA para contrastar distintos enfoques y técnicas de modelado para ayudarme a decidir cuáles utilizar finalmente.

- *Revisión o reescritura de párrafos redactados previamente*

He utilizado la IAG para la reescritura de algunos párrafos ya redactados mejorando su cohesión y claridad, o en ocasiones para sintetizarlos y acortarlos, siempre manteniendo el contenido e ideas originales.

- *Búsqueda de información o respuesta a preguntas concretas*

Algunas consultas realizadas están orientadas a obtener más detalle o explicaciones sobre conceptos teóricos como algunos hiperparámetros o funcionamiento de algunos modelos.

- *Búsqueda de bibliografía*
- *Resumen de bibliografía consultada*

Desarrollar contenido específico

Se ha hecho uso de IAG como herramienta de soporte para el desarrollo del contenido específico del TFG, incluyendo:

- *Asistencia en el desarrollo de líneas de código (programación)*

He utilizado la IA para obtener ayuda en algunas librerías específicas con funciones u otras opciones, aunque el groso del código fue realizado por mí.

- *Generación de esquemas, imágenes, audios o vídeos*

No procede – Todas las figuras son de elaboración propia a través de Python.

- *Procesos de optimización*

He utilizado la IA para depurar y corregir errores de código.

- *Tratamiento de datos: recogida, análisis, cruce de datos…*

No procede.

- *Inspiración de ideas en el proceso creativo*

Una vez definido el objetivo del trabajo, he solicitado ideas para abordar sobre todo la creación de modelos de la forma más correcta posible, sin delegar en la IA la definición de objetivos o estructura del trabajo.

Parte 3: reflexión sobre utilidad

Aporta una valoración personal (formato libre) sobre las fortalezas y debilidades que has identificado en el uso de herramientas de IAG en el desarrollo de tu trabajo. Menciona si te ha servido en el proceso de aprendizaje, o en el desarrollo o en la extracción de conclusiones de tu trabajo.

El uso de estas herramientas ha sido muy útil como apoyo para el desarrollo del trabajo, pero también como aprendizaje. Me ha facilitado enormemente la capacidad de entender conceptos metodológicos complejos y poder aplicarlos correctamente, así como las tareas de programación, sobre todo al encontrarme con errores en el código o cosas que no funcionaban correctamente.
Por otro lado, también me he encontrado con limitaciones importantes, como la veracidad de la información, que he tenido que contrastar siempre con fuentes académicas.
Creo que, si se usa solo como herramienta de apoyo y nunca como sustituto del razonamiento ni nuestra propia toma de decisiones y fuentes oficiales, puede ayudar enormemente a mejorar la calidad y coherencia del trabajo final y facilitar muchas tareas, sobre todo en trabajos técnicos en los que se usen lenguajes de programación.
Otro tema a tener en cuenta es que la IA nunca debe sustituir nuestro pensamiento para dar una estructura coherente y redactar las ideas de nuestro trabajo.