



UNIVERSIDAD  
**COMPLUTENSE**  
MADRID

# Estimación del precio de la vivienda en Madrid mediante técnicas de *machine learning*

Comparativa de modelos y propuesta de  
implementación empresarial

Curso académico  
2024/2025

**Autor:** Diego Olalla Carrión  
**Tutores:** Carlos Ortega, Santiago Mota  
**Máster Universitario en Data Science,  
Big Data y Business Analytics**  
Universidad Complutense de Madrid

# ÍNDICE

1. Resumen ejecutivo .....	2
2. Introducción .....	3
2.1 Contexto y motivación del trabajo.....	3
2.2 Descripción del problema o reto de negocio.....	3
2.3 Relevancia en el ámbito del Data Science / Big Data / Business Analytics.....	5
3. Metodología.....	6
3.1 Fases del proyecto.....	6
3.2 Justificación de técnicas y herramientas utilizadas .....	8
4. Análisis de datos / Preparación de datos.....	8
4.1 Exploración y descripción del dataset.....	8
4.2 Transformaciones y limpieza de datos.....	10
4.3 Retos encontrados.....	11
5. Modelado / Desarrollo técnico .....	12
5.1 Modelos construidos .....	12
5.2 Comparativa de enfoques .....	13
5.3 Evaluación de resultados .....	13
6. Discusión de resultados .....	14
6.1 Interpretación en términos de negocio.....	14
6.2 Limitaciones del estudio .....	15
6.3 Posibles mejoras futuras.....	15
7. Conclusiones.....	16
7.1 Principales hallazgos.....	16
7.2 Impacto esperado en el negocio o la práctica profesional .....	17
7.3 Lecciones aprendidas.....	18
8. Bibliografía y referencias.....	19
9. Anexos .....	20

## 1. Resumen ejecutivo

Este trabajo aborda el análisis del mercado inmobiliario de Madrid mediante la construcción de un sistema predictivo de precios basado en técnicas avanzadas de machine learning. El estudio se apoya en un dataset de más de 5.200 propiedades con 36 variables que abarcan aspectos de localización, características estructurales, servicios y condiciones del inmueble. La complejidad del mercado madrileño marcado por fuertes disparidades territoriales, la concentración de zonas premium y la presencia de outliers significativos hace necesario un enfoque analítico que supere las metodologías tradicionales de tasación.

El proyecto se desarrolla en cuatro fases principales: análisis exploratorio de datos, preprocesamiento y limpieza, ingeniería de características y modelado comparativo con diversos algoritmos. Los resultados técnicos identifican a **XGBoost como el modelo con mayor capacidad predictiva**, alcanzando un  $R^2$  cercano al 0,95 y predicciones con error inferior al  $\pm 10\%$  en más del 70% de los casos. Sin embargo, desde una perspectiva operativa y de negocio, el análisis recomienda **RandomForest como modelo más adecuado para su aplicación práctica**, debido a su robustez, estabilidad y facilidad de implementación, con un rendimiento prácticamente equivalente ( $R^2 \approx 0,949$ ). (Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_2** para **preprocesado/ingeniería y comparativa de modelos con métricas completas**; y **Anexo I - Diego\_Olalla\_Carrión\_Informe\_Tecnico** para **criterios de selección y justificación**).

Los hallazgos confirman que el precio de la vivienda en Madrid está dominado por la **interacción entre ubicación premium y superficie construida**, reforzando la fórmula “ubicación + metros cuadrados” como núcleo del valor inmobiliario. Factores adicionales como la densidad de baños, la calidad del espacio interior y las amenidades de accesibilidad (ascensor, climatización) también ejercen un impacto significativo.

Desde un punto de vista estratégico, el sistema ofrece a promotoras, agencias y entidades financieras un marco de apoyo a la toma de decisiones con ventajas en tres áreas: mayor precisión en valoraciones, capacidad de segmentación de clientes e identificación de oportunidades de inversión. La propuesta final combina rigor analítico con viabilidad práctica, recomendando un **enfoque híbrido en el que RandomForest opere en producción y XGBoost actúe como benchmark de referencia y control de calidad**.

En conclusión, el proyecto demuestra que la aplicación de machine learning al mercado inmobiliario no solo mejora la precisión predictiva frente a las técnicas tradicionales, sino que también puede integrarse en procesos reales siempre que se equilibren precisión y estabilidad.

### Alcance del estudio y datos:

Elemento	Síntesis
Ámbito	Mercado residencial de Madrid
Muestra	> 5.200 propiedades

Elemento	Síntesis
Variables	36 (localización, estructura, servicios, condiciones)
Complejidad	Disparidades territoriales, zonas premium, outliers
Fases	EDA → Preprocesado/Limpieza → Ingeniería → Modelado comparativo

## 2. Introducción

### 2.1 Contexto y motivación del trabajo

El mercado inmobiliario madrileño constituye uno de los sectores más dinámicos y estratégicos de la economía española. Madrid concentra aproximadamente el 14% de la población nacional y se posiciona como núcleo financiero, administrativo y tecnológico del país. Esta relevancia se traduce en un mercado residencial caracterizado por su complejidad, con diferencias de precios que superan el 300% entre distritos, una fuerte polarización entre áreas premium (como Salamanca, Chamberí o Retiro) y zonas periféricas, y una creciente demanda por viviendas que integren criterios de sostenibilidad y eficiencia energética.

Según datos de Idealista (2019), el precio medio de la vivienda en Madrid alcanzaba los 3.770 €/m<sup>2</sup>, con distritos como Salamanca y Chamberí superando los 5.000 €/m<sup>2</sup>, frente a zonas como Villaverde o Usera, donde no se alcanzaban los 2.000 €/m<sup>2</sup>. Este diferencial refleja una segmentación marcada que condiciona tanto la oferta como la demanda. Asimismo, informes del Ayuntamiento de Madrid señalan que entre 2015 y 2019 el valor residencial en la ciudad aumentó en torno a un 37%, impulsado por la presión demográfica, la inversión extranjera y los procesos de regeneración urbana. Estos indicadores refuerzan la relevancia económica del problema y justifican la necesidad de modelos predictivos capaces de capturar dicha heterogeneidad.

En este contexto, la estimación precisa del precio de la vivienda adquiere un papel clave tanto para promotoras y agencias inmobiliarias como para entidades financieras, inversores y administraciones públicas. Los métodos tradicionales de valoración, basados en comparables simples o en modelos hedónicos clásicos, resultan insuficientes para capturar la heterogeneidad y no linealidad del mercado actual.

La **motivación** de este trabajo es ofrecer una solución predictiva más precisa y adaptable, basada en técnicas de ciencia de datos, que permita mejorar la toma de decisiones en un sector de alta relevancia económica y social.

### 2.2 Descripción del problema o reto de negocio

El problema central de este TFM es construir un sistema que estime el precio de la vivienda en Madrid con la fiabilidad y la consistencia que los enfoques tradicionales no siempre alcanzan, y que además pueda integrarse en procesos reales de tasación, fijación de precios y análisis de inversión. No se trata solo de "acertar" una cifra: el reto consiste en entregar un valor defendible, estable a lo largo del tiempo,

comprendible para equipos no técnicos y útil para tomar decisiones con menor incertidumbre.

La primera dimensión del reto es la **heterogeneidad geográfica y socioeconómica**. Madrid reúne mercados muy distintos dentro del mismo municipio: barrios premium consolidados, ejes residenciales con demanda estable y zonas emergentes donde conviven dinámicas de renovación y cambios de oferta. Estas diferencias se manifiestan en fuertes disparidades del precio por metro cuadrado, no solo entre distritos, sino incluso entre subzonas dentro de un mismo barrio. En la práctica, esto implica que el sistema debe capturar matices de localización y entorno sin perder legibilidad, y ofrecer resultados comparables entre áreas con realidades muy diferentes.

La segunda dimensión ataña a la **calidad y la forma de los datos**. El conjunto de partida contiene información incompleta en variables relevantes (como año de construcción u orientación) y variables de alta cardinalidad (por ejemplo, 171 barrios o 51 niveles de planta), además de descripciones textuales que encierran información útil pero no estructurada. Traducido al negocio, el desafío es doble: por un lado, evitar que los huecos introduzcan sesgos o decisiones arbitrarias; por otro, transformar un catálogo muy amplio de categorías en señales comprensibles que aporten valor real a la estimación, sin saturar a quien toma decisiones con un exceso de detalle difícil de utilizar. (Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1** para **limpieza, imputación y gestión de alta cardinalidad**).

La tercera dimensión es la **variabilidad por segmentos de mercado**. Mientras que la vivienda estándar y las zonas premium muestran pautas más estables y previsibles, el tramo intermedio y los activos de lujo presentan una dispersión mayor y una sensibilidad acusada a atributos muy particulares de cada inmueble. En consecuencia, el sistema debe identificar en qué contextos puede ofrecer estimaciones de alta precisión y en cuáles conviene presentarlas con mayor prudencia; por ello, se aporta una lectura segmentada por bandas de precio y áreas, acompañada de explicaciones que acotan el alcance de la cifra y evitan generar expectativas no realistas en entornos donde el comportamiento del mercado es más volátil. (Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_2** para **análisis por segmentos/bandas de precio**).

Con este contexto, el **objetivo de negocio** es claro: disponer de un modelo que combine precisión estadística con **estabilidad, interpretabilidad y aplicabilidad operativa**. En términos prácticos, ello implica que la herramienta proporcione una estimación central coherente con el mercado, acompañada de explicaciones comprensibles sobre los factores que más han pesado (ubicación, superficie, cualidades del espacio y amenidades), y que permita desagregar resultados por zonas y bandas de precio cuando sea necesario. El sistema debe facilitar, además, la identificación temprana de casos atípicos o dudosos y activar una revisión experta antes de utilizar la cifra en una negociación, una tasación interna o un análisis de inversión.

Para asegurar su **adopción**, el problema se define también con restricciones explícitas: transparencia en las reglas de decisión, reproducibilidad de resultados, tiempos de respuesta razonables para el uso diario y un plan de mantenimiento que contemple la evolución del mercado (actualizaciones de datos y revisiones periódicas). Igualmente, se incorporan consideraciones de **equidad y aceptación**:

vigilar posibles sesgos geográficos, documentar supuestos y límites, y mantener un circuito de revisión humana en los casos sensibles, de modo que la herramienta aporte legitimidad y confianza.

En síntesis, el reto de negocio no consiste únicamente en construir un predictor exacto, sino en diseñar un **sistema de apoyo a la decisión** que funcione en los contextos reales de promotoras, agencias y entidades financieras: que reduzca la incertidumbre, mejore la comparabilidad entre activos y zonas, y permita priorizar oportunidades con criterios homogéneos. El éxito se medirá tanto por la calidad técnica de las estimaciones como por su utilidad práctica: que las cifras sean defendibles, consistentes y accionables en los procesos cotidianos de valoración, fijación de precios y análisis de mercado.

### **Objetivo de negocio (operativo)**

<b>Componente</b>	<b>Exigencia</b>	<b>Anexo de detalle</b>
Precisión + estabilidad	Estimación central coherente y consistente	Notebook_2 (métricas/estabilidad)
Interpretabilidad	Explicar factores: ubicación, superficie, cualidades, amenidades	Notebook_2 (importancias), Informe_Tecnico (explicación de variables)
Desagregación	Resultados por zonas y bandas de precio	Notebook_2 (segmentación)
Gestión de atípicos	Detección temprana y revisión experta	Informe_Tecnico (protocolo de revisión)

**2.3 Relevancia en el ámbito del Data Science / Big Data / Business Analytics**  
El estudio se enmarca en el ámbito de la ciencia de datos aplicada a la economía urbana y responde a varios retos actuales en Data Science y Big Data:

- Aplicación de técnicas de *machine learning* a un mercado real de elevada complejidad, con variables estructurales, geoespaciales, categóricas y textuales.
- Gestión de datos incompletos y *outliers*, aplicando metodologías de imputación avanzada, codificación híbrida y estrategias de detección multivariante.
- Construcción y evaluación comparativa de múltiples algoritmos (lineales, basados en árboles y *boosting*) para identificar no solo el modelo con mayor precisión, sino también el más robusto para su uso práctico.
- Enfoque orientado a negocio, en el que los resultados no se limitan a métricas técnicas, sino que se traducen en implicaciones estratégicas para la toma de decisiones en el sector inmobiliario.

Este TFM no solo busca aportar un modelo predictivo de referencia, sino también un marco metodológico transferible a otros mercados urbanos, contribuyendo a la práctica profesional de la analítica avanzada en sectores donde la precisión y la estabilidad de los modelos tienen un impacto directo en decisiones económicas de alto valor.

### 3. Metodología

#### 3.1 Fases del proyecto

El desarrollo del sistema predictivo siguió un enfoque estructurado, alineado con las fases estándar de un proceso analítico en ciencia de datos:

##### 1. Definición del problema de negocio

- Objetivo: construir un modelo predictivo robusto para estimar precios de viviendas en Madrid con aplicabilidad práctica en el sector inmobiliario.
- Éxito esperado: superar un  $R^2$  de 0,90 y garantizar estabilidad en diferentes segmentos de mercado, evaluando también métricas como RMSE, MAE y MAPE por tramos de precio.

(Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_2** para **el cuadro detallado de métricas de validación**).

##### 2. Análisis exploratorio de datos (EDA)

- Se exploró un dataset de 5.200 propiedades y 36 variables.
- Se identificaron patrones de heterogeneidad, valores faltantes y *outliers*.
- Se analizaron correlaciones entre variables estructurales (habitaciones, baños, superficie) y ubicación.
- Se observaron diferencias extremas entre distritos premium y periféricos, confirmando la necesidad de un modelo capaz de capturar relaciones no lineales.

##### 3. Preparación y transformación de datos

- Imputación multienfoque de valores faltantes (mediana condicional, KNN, categorías “desconocido”).
- Codificación híbrida de variables categóricas de alta cardinalidad (barrios, orientaciones).
- Transformaciones de escala y normalización adaptativa (RobustScaler, logaritmos).
- Este proceso redujo la dispersión de variables críticas y permitió entrenar modelos más estables.

(Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1** para **limpieza, imputación y codificación**).

##### 4. Ingeniería de características

- Creación de variables derivadas como *premium\_district\_size* (superficie × indicador premium) y *bath\_density* (número de baños/superficie).
- Agrupación inteligente de barrios (171 → 15 grupos) y orientaciones arquitectónicas.
- Selección de características mediante filtrado estadístico, modelos y RFE-CV, reduciendo de 203 a 79 variables finales.
- La ingeniería de variables mejoró la capacidad predictiva en más de un 20% respecto a un modelo sin transformación.

## 5. Modelado predictivo

- Entrenamiento y validación de modelos lineales (Ridge, Lasso, ElasticNet), de árboles (RandomForest, GradientBoosting, DecisionTree, AdaBoost) y avanzados (XGBoost, CatBoost, LightGBM).
- Se descartaron KNN y SVR tras pruebas preliminares por su menor escalabilidad y precisión en este contexto.
- Validación cruzada estratificada por quintiles de precio y análisis de estabilidad con métricas múltiples ( $R^2$ , RMSE, MAE, error relativo por segmentos).

## 6. Evaluación y selección del modelo

- XGBoost alcanzó la mejor precisión ( $R^2 \approx 0,95$ ), pero RandomForest se destacó como modelo más robusto y práctico para su uso operativo.
- Se propuso un enfoque híbrido: XGBoost como *benchmark* de referencia y RandomForest como modelo productivo.
- La brecha entre entrenamiento y validación se mantuvo baja (<0,05), lo que refuerza la estabilidad del modelo en diferentes muestras.

(Véase **Anexo II** - Diego\_Olalla\_Carrión\_Notebook\_2 para **métricas completas y análisis de estabilidad**).

## 7. Interpretación y propuesta de negocio

- Análisis de importancia de variables (SHAP, *Feature Importance*).
- Identificación de los factores clave: ubicación premium, superficie, densidad de baños y amenidades de accesibilidad.
- Traducción de resultados en recomendaciones estratégicas para promotoras, agencias y entidades financieras.

## 8. Conclusiones

- Evaluación del impacto del sistema en términos de negocio: mayor precisión en valoraciones, segmentación de clientes e identificación de oportunidades de inversión.

(Véase **Anexo I** - Diego\_Olalla\_Carrión\_Informe\_Tecnico para la **discusión final, limitaciones y próximos pasos**).

### 3.2 Justificación de técnicas y herramientas utilizadas

El proyecto integró técnicas de ciencia de datos y *machine learning* con criterios de rigor metodológico y aplicabilidad:

- **Lenguaje de programación:** Python, por su ecosistema de librerías en análisis de datos (*pandas*, *NumPy*), *machine learning* (*scikit-learn*, *XGBoost*, *CatBoost*, *LightGBM*) y visualización (*matplotlib*, *seaborn*).
- **Entorno de desarrollo:** Jupyter Notebook, que facilita la trazabilidad del análisis y la integración de código, resultados y documentación.
- **Preprocesamiento de datos:** se emplearon técnicas de imputación condicional y KNN para datos faltantes, *target encoding* para categorías con alta cardinalidad y normalización robusta frente a *outliers*.
- **Ingeniería de características:** se generaron variables derivadas para capturar interacciones no lineales clave en el mercado inmobiliario, mejorando en más de un 20% la capacidad predictiva.
- **Modelos predictivos:** se compararon algoritmos de diferente naturaleza para equilibrar precisión, interpretabilidad y robustez. Los modelos lineales aportaron transparencia, los de árboles interpretabilidad parcial y los de *boosting* máxima precisión.
- **Validación:** se aplicó validación cruzada estratificada y análisis de errores por segmentos de precio, asegurando que el sistema fuera estable y aplicable en todos los rangos del mercado.
- **Interpretabilidad:** se utilizaron herramientas como SHAP para explicar el impacto de cada variable en la predicción, facilitando la comprensión de resultados a equipos no técnicos.

## 4. Análisis de datos / Preparación de datos

### 4.1 Exploración y descripción del dataset

El dataset utilizado contiene **5.255 propiedades residenciales** de la ciudad de Madrid, con un total de **36 variables** que abarcan aspectos de identificación, geolocalización, características estructurales, servicios, condiciones de construcción y metadatos textuales. Entre las variables clave destacan:

- **Precio de la vivienda:** rango entre 39.000 € y 9,48 M€, con media de 474.523 € y mediana de 247.000 €, lo que evidencia una alta dispersión y fuerte asimetría.
- **Superficie:** media de 119,8 m<sup>2</sup>, con una distribución heterogénea que distingue segmentos económicos, estándar, *premium* y de lujo. (Detalle en **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1, distribución de m<sup>2</sup> y segmentación por tramos**).
- **Número de habitaciones y baños:** dominancia de inmuebles de 2-3 habitaciones y 1-2 baños, siendo este último la variable estructural con mayor correlación con el precio ( $r \approx 0,64$ ). (Detalle en **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1, matriz de correlaciones y conteos por categoría.**)

- **Ubicación geográfica:** 11 distritos y 171 barrios, con diferencias extremas de valoración (hasta 8 veces entre barrios).
- **Servicios y amenidades:** presencia de ascensor, aire acondicionado, garaje, piscina o terraza, variables binarias con tasas de aparición muy heterogéneas (15%–85%).
- **Calidad y consistencia de datos:** verificación de tipos (conversión de *price* a numérico y fechas a formato estándar), control básico de duplicados por identificador y revisión de registros con campos críticos vacíos (detalle en **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1 – celdas de limpieza**).

El análisis exploratorio confirmó la **multimodalidad del mercado**, con tres concentraciones principales de precios en torno a 150.000 €, 250.000 € y 425.000 €. Además, el distrito de Salamanca y sus barrios *premium* destacaron como polos de valor, con multiplicadores superiores al 4× respecto a zonas periféricas como Villaverde (gráficos y tablas en **Anexo II -Diego\_Olalla\_Carrión\_Notebook\_1, densidades**).

#### **Resumen del dataset y variables clave**

<b>Aspecto</b>	<b>Valor / Descripción</b>
Observaciones y periodo	5.255 propiedades
Nº de variables	36 (identificación, geolocalización, estructura, servicios, construcción, metadatos textuales)
Precio	39.000–9,48 M€; media 474.523 €; mediana 247.000 €; alta dispersión y asimetría
Superficie	Media 119,8 m <sup>2</sup> ; segmentación: económico / estándar / premium / lujo
Habitaciones y baños	Dominancia 2–3 hab. y 1–2 baños; correlación (baños, precio) ≈ 0,64
Ubicación	11 distritos y 171 barrios; disparidades hasta ×8 entre barrios
Amenidades	Ascensor, garaje, piscina, terraza, prevalencia 15%–85%
Calidad de datos	Conversión de tipos, control de duplicados, revisión de vacíos críticos
Multimodalidad del precio	Modos aprox. en 150k, 250k y 425k; Salamanca >4× Villaverde

## 4.2 Transformaciones y limpieza de datos

El preprocesamiento del dataset fue una fase crítica para garantizar la calidad del modelo predictivo. Las principales acciones realizadas fueron:

- **Gestión de valores faltantes:**

- Variables con más del 20% de ausencia (*construct\_date, orientation*) se realizó imputación contextual y creación de **indicadores de ausencia**.
- Variables con 5%-20% de ausencia (*heating, loc\_street*) se realizó imputación mediante **KNN** y patrones geográficos.
- Variables con <5% de ausencia (*bath\_num, room\_num*) se realizó imputación por **mediana condicional**.

(Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1** para la implementación de **imputaciones y checks de ausencia**; y detalle en tabla en **Anexo - Diego\_Olalla\_Carrión\_Informe\_Tecnico** para **reglas y criterios de imputación**).

- **Codificación de variables categóricas:**

- Alta cardinalidad (*barrios, descripción textual*) a través de *target encoding* regularizado.
- Cardinalidad media (*distritos*) mediante combinación de **one-hot encoding** con agrupaciones jerárquicas en macrozonas.
- Cardinalidad baja (*tipo de vivienda*) mediante **one-hot encoding estándar**.

- **Normalización y escalado:**

- Transformación **logarítmica** aplicada a la variable *precio* para reducir asimetría.
- Uso de **RobustScaler** en variables afectadas por *outliers* (superficie, número de baños).
- **MinMaxScaler** en variables ordinales (ej. número de planta).

- **Tratamiento de outliers:**

- Precios extremos y superficies atípicas gestionados mediante **transformación logarítmica** y verificación de calidad de datos.
- *Outliers* multivariados identificados con **Isolation Forest**, conservando aquellos representativos de segmentos de lujo o singulares.

(Véase **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_1** para control de calidad y log-transform; **Anexo II - Diego\_Olalla\_Carrión\_Notebook\_2** para **detección multivariante con Isolation Forest y criterios de conservación**; y **Anexo I - Diego\_Olalla\_Carrión\_Informe\_Tecnico** para la **política de tratamiento de casos atípicos**).

- **Ingeniería de variables derivadas:**

- *premium\_district\_size* = superficie × indicador distrito *premium*.
- *bath\_density* = número de baños / superficie.
- *age\_size* = antigüedad × superficie.
- *size\_quality* = superficie × índice de calidad.

- **Estandarización adicional:**

- Control básico de incoherencias (superficies nulas/negativas, precios no numéricos) y documentación de reglas aplicadas.

(Detalle en **Anexo I** (Informe\_Tecnico) para el listado completo de **reglas y evidencias**; y **Anexo II** - Diego\_Olalla\_Carrión\_Notebook\_1 para las **celdas de validación y casting de tipos**).

#### 4.3 Retos encontrados

A lo largo de la preparación de los datos afrontamos varios retos que condicionaron desde el inicio la forma de trabajar y, sobre todo, la manera de presentar resultados útiles para la toma de decisiones. **El mercado residencial de Madrid es profundamente heterogéneo**: conviven realidades muy distintas, vivienda estándar, zonas premium consolidadas y activos singulares de lujo y esa mezcla se traducía en cifras muy dispares. Para que las conclusiones fueran sólidas y comparables, el primer paso fue poner orden en esa dispersión sin borrar la realidad del mercado. **Ajustamos la escala de las variables clave** para que los casos extremos no arrastrasen el análisis y las tipologías más comunes no quedaran en segundo plano; con ello, **los modelos dejaron de sobrerreaccionar ante valores atípicos y pasaron a captar con mayor fidelidad la tendencia general**.

La ubicación y la altura del inmueble exigieron un tratamiento específico por su enorme variedad. Partíamos de 171 barrios y hasta 51 niveles de planta; incorporar todo "tal cual" habría enturbiado la lectura y restado claridad a las conclusiones. **Optamos por agrupar los barrios en macrozonas con identidad reconocible** (centros premium, ejes residenciales consolidados, periferias dinámicas, entre otras) y por **simplificar la altura a las categorías que realmente marcan diferencias de precio** (bajos, plantas intermedias y áticos; viviendas exteriores o interiores). **Esta reorganización permitió conservar intacta la potencia explicativa de la localización** (factor decisivo en Madrid) sin diluir el análisis en un exceso de detalles poco informativos.

**La ausencia de datos en campos estratégicos constituyó un tercer frente.** En variables como el año de construcción o la orientación, ignorar los huecos habría introducido sesgos y, en el extremo opuesto, rellenarlos sin criterio habría maquillado la realidad. **Elegimos una vía intermedia y honesta**: señalar explícitamente cuándo una información no estaba disponible y completar los vacíos con referencias realistas según la zona y la tipología. **El resultado fue un conjunto continuo (sin huecos artificiales) que preserva los patrones originales del mercado** y añade contexto donde faltaban datos.

También **detectamos variables que, en el fondo, contaban la misma historia**. Para evitar que un mismo factor pesara varias veces (y con ello se distorsionara la

lectura) depuramos la información y nos quedamos con la combinación más clara y representativa. **Esta racionalización simplificó la interpretación** y nos permitió concentrarnos en **lo que de verdad mueve el precio: el tamaño y la calidad del espacio, la ubicación y ciertos servicios diferenciadores** que impactan la experiencia cotidiana de los residentes.

Por último, **la estructura del mercado nos llevó a trabajar por segmentos**. Observamos mayor estabilidad en vivienda estándar y en barrios premium bien definidos, mientras que el rango medio y el lujo presentaron más variación caso a caso. Para no comprometer precisión donde el mercado es más caprichoso, **evaluamos los resultados por bandas de precio y por clústeres de zonas**, y reforzamos la interpretación humana en los activos singulares. **Este enfoque segmentado elevó la utilidad práctica del análisis** para quienes fijan precios, negocian o valoran activos.

Como resultado de este recorrido (limpiar sin "borrar realidad", agrupar sin perder identidad local y completar información con sentido) **el conjunto de trabajo quedó configurado por 5.248 viviendas descritas mediante 79 variables de valor**. Consideramos que es un equilibrio sólido entre riqueza informativa y agilidad analítica (gráficos y tablas en **Anexo II- Diego\_Olalla\_Carrión\_Notebook\_1** y **Diego\_Olalla\_Carrión\_Notebook\_2**).

## 5. Modelado / Desarrollo técnico

### 5.1 Modelos construidos

Se entrenaron y evaluaron distintos algoritmos de *machine learning* con el objetivo de encontrar un equilibrio entre precisión predictiva, robustez y aplicabilidad práctica. Los modelos se organizaron en tres categorías:

1. **Modelos lineales:** Ridge, Lasso y ElasticNet, que ofrecieron interpretabilidad y simplicidad, pero con menor capacidad de capturar relaciones no lineales.
2. **Modelos basados en árboles:** RandomForest, DecisionTree, AdaBoost y GradientBoosting, que permitieron identificar interacciones complejas y ofrecieron mejor desempeño que los lineales.
3. **Modelos avanzados de boosting:** XGBoost, CatBoost y LightGBM, que mostraron el mejor rendimiento global en términos de precisión y generalización.

El *pipeline* incluyó división del dataset en entrenamiento (80%) y test (20%), con validación cruzada estratificada ( $k=5$ ) por quintiles de precio para garantizar estabilidad y control de sobreajuste. Las transformaciones (imputación, codificación y escalado) se encapsularon en *pipelines* para evitar fugas de información; la búsqueda de hiperparámetros se realizó con esquemas de validación interna y, cuando procedía, con parada temprana (*early stopping*) en los modelos de *boosting*. La reproducibilidad se aseguró mediante fijación de semillas y registro de configuraciones. Detalle en **Anexo II**: Diego\_Olalla\_Carrión\_Notebook\_1 (pipeline de **preprocesado**), Diego\_Olalla\_Carrión\_Notebook\_2 (**entrenamiento, validación y métricas**) y Diego\_Olalla\_Carrión\_Informe\_Tecnico - Sección de **Modelado**.

## 5.2 Comparativa de enfoques

Los resultados mostraron diferencias claras en el desempeño de cada familia de modelos:

### a) Modelos lineales

- Útiles para entender relaciones directas entre variables, pero insuficientes en un mercado tan heterogéneo.

### b) Modelos de árboles

- RandomForest destacando por su estabilidad y menor riesgo de sobreajuste.
- GradientBoosting mostró un equilibrio entre precisión e interpretabilidad parcial.
- DecisionTree y AdaBoost resultaron menos efectivos, con problemas de generalización.

### c) Modelos de *boosting* avanzados

- XGBoost fue el modelo con mejor rendimiento técnico.
- CatBoost obtuvo resultados muy similares, con la ventaja de manejar categorías de forma nativa y menor sobreajuste.
- LightGBM ofreció gran eficiencia computacional, con un rendimiento ligeramente inferior al de XGBoost.

Detalle en **Anexo II**: Diego\_Olalla\_Carrión\_Notebook\_2 (**resultados de modelos lineales, árboles y boosting**).

## 5.3 Evaluación de resultados

La comparación entre modelos permite extraer conclusiones relevantes tanto a nivel técnico como de negocio:

### a) Desempeño técnico

- XGBoost lidera en precisión predictiva ( $R^2 \approx 0,951$ ;  $RMSE \approx 0,171$ ), consolidándose como el modelo de referencia en términos métricos.
- RandomForest ofrece un rendimiento muy próximo ( $R^2 \approx 0,949$ ), con mayor estabilidad y menor sensibilidad a configuraciones de hiperparámetros.
- Los modelos lineales quedan rezagados en un mercado con fuertes no linealidades.

### b) Robustez y aplicabilidad

- RandomForest se perfila como el modelo más adecuado para su implementación en entornos empresariales, debido a su robustez frente a variaciones de datos y su facilidad de integración en sistemas de producción.
- XGBoost se recomienda como *benchmark* técnico, utilizado en paralelo para validar resultados y monitorizar la calidad predictiva.
- CatBoost y LightGBM son alternativas viables en escenarios específicos (p. ej., APIs en tiempo real o despliegues ligeros).

### c) Interpretabilidad de los modelos

- El análisis de importancia de variables y valores SHAP confirma que la interacción entre ubicación *premium* y superficie (*premium\_district\_size*) es el principal determinante del precio.
- Otros factores clave son la densidad de baños, la calidad del espacio y la ubicación geográfica en distritos como Salamanca, Chamberí y Retiro.

El modelo **XGBoost** es el campeón en términos de métricas técnicas, pero el modelo **RandomForest** se recomienda como opción empresarial para producción, dada su mayor robustez e interpretabilidad. La propuesta final es un **enfoque híbrido**, en el que RandomForest se despliega en producción y XGBoost actúa como modelo de control y validación.

Detalle en **Anexo II**: Diego\_Olalla\_Carrión\_Notebook\_2 (**métricas comparativas, CV estratificada, estabilidad por segmentos, SHAP/importancias**) y en el **Anexo I**: Diego\_Olalla\_Carrión\_Informe\_Tecnico – **Secciones de Evaluación de resultados e Interpretación**.

## 6. Discusión de resultados

### 6.1 Interpretación en términos de negocio

Los resultados del modelado confirman una jerarquía clara de determinantes del precio inmobiliario en Madrid:

- **Ubicación premium x superficie (*premium\_district\_size*)**: principal driver del precio, cuantificando multiplicadores de hasta 4x en distritos como Salamanca frente a zonas periféricas.
- **Densidad de baños (*bath\_density*)**: indicador de calidad interior que diferencia viviendas funcionales de alto valor frente a aquellas con distribución deficiente.
- **Calidad del espacio y amenidades**: variables como ascensor, terraza o aire acondicionado aportan valor diferencial, especialmente en zonas consolidadas con edificios antiguos.
- **Factores microgeográficos**: distritos emergentes como Tetuán muestran dinámicas particulares, con crecimientos que reflejan procesos de revalorización urbana.

Desde la perspectiva empresarial, el sistema predictivo aporta tres beneficios clave:

1. **Mayor precisión en valoraciones**: mejora frente a métodos tradicionales, reduciendo errores en más del 25%.
2. **Segmentación estratégica**: identificación de patrones de valor por rangos de precio y zonas, útil para agencias e inversores.
3. **Soporte a la toma de decisiones**: permite priorizar inversiones en metros cuadrados en ubicaciones *premium* y en mejoras interiores de alto impacto.

Detalle en **Anexo II**: Diego\_Olalla\_Carrión\_Notebook\_2 (**SHAP/importancias y segmentación por rangos**).

## 6.2 Limitaciones del estudio

A pesar de los resultados positivos, el trabajo presenta limitaciones que condicionan su aplicabilidad:

Ámbito	Resumen	Implicación operativa
Desempeño por segmentos	Alta precisión en viviendas estándar y premium; menor estabilidad en rango medio (600k-1M €) y lujo (>1M €) por mayor variabilidad intrínseca.	Lectura y reporte segmentados; prudencia en esos tramos y validación adicional antes de decisiones críticas.
Dependencia de variables dominantes	Fuerte influencia de ubicación y superficie puede hacer sensible el modelo a cambios estructurales del mercado.	Vigilancia de deriva y recalibraciones; seguimiento de métricas por zona y tamaño.
Falta de variables cualitativas internas	Estado de conservación, acabados, reformas o vistas no están plenamente reflejados en los datos.	Plan para enriquecimiento de variables cualitativas cuando estén disponibles (texto/fotos/inspecciones).
Ausencia de dimensión temporal	Estudio de un único periodo; limitada capacidad para proyectar tendencias medio-largo plazo.	Incorporar actualizaciones periódicas y variables temporales; comparar cortes por año/semestre.

## 6.3 Posibles mejoras futuras

Para reforzar la utilidad práctica del sistema, se propone:

Acción prioritaria	Qué se hace	Cómo se implementa	Impacto esperado
Incorporar dimensión temporal	Integrar histórico y modelos de forecasting para capturar ciclos.	Añadir sello temporal, agregaciones mensuales.	Mejor lectura de ciclos y estabilidad fuera de muestra.
Segmentar modelos por rango de precio	Submodelos para estándar, premium y lujo.	Estratificación por bandas; calibración y métricas por segmento; (opcional) stacking como metamodelo.	Mejor ajuste en colas y mensajes más precisos por segmento.
Validación geográfica cruzada	Asegurar consistencia entre barrios/distritos.	CV por bloques espaciales (p. ej., leave-one-district-out); evaluación por zona.	Generalización real entre áreas y detección de

Acción prioritaria	Qué se hace	Cómo se implementa	Impacto esperado
			sobreajuste espacial.
Despliegue híbrido con control de calidad	RF en producción y XGBoost como referencia/monitor.	RF como modelo estable; XGB en paralelo; monitorización de drift (datos y rendimiento), alertas y plan de recalibración.	Robustez operativa y control continuo de la calidad predictiva.

El sistema predictivo demuestra ser una herramienta poderosa para el mercado inmobiliario madrileño, pero debe entenderse como un **apoyo a la decisión y no como sustituto de la supervisión experta**. Su principal valor radica en la capacidad de cuantificar *drivers* de precio de forma granular y consistente, lo que aporta ventaja competitiva a promotoras, agencias y entidades financieras.

Detalle en **Anexo II: Diego\_Olalla\_Carrión\_Notebook\_2 (validación cruzada, segmentación por rangos de precio)**.

## 7. Conclusiones

### 7.1 Principales hallazgos

El trabajo confirma que la aplicación de técnicas de *machine learning* al mercado inmobiliario de Madrid permite alcanzar niveles de precisión **superiores al 94%** en la explicación de la variabilidad de precios, manteniendo coherencia entre rigor analítico y uso operativo. Los resultados se traducen en conclusiones accionables para fijación de precios, negociación y apoyo a valoraciones internas. Entre los hallazgos clave destacan:

- **Ubicación premium y superficie**

Constituyen el núcleo del valor inmobiliario, validando la fórmula “ubicación + metros cuadrados” como *driver* principal. La combinación de ambas variables explica buena parte de las diferencias de precio entre zonas y dentro de una misma área, y permite ordenar de forma consistente las estimaciones por rango de valor. Su peso se mantiene incluso cuando se introducen variables adicionales, lo que refuerza su carácter estructural en el mercado madrileño.

- **Densidad de baños y calidad del espacio**

Actúan como moduladores determinantes: a igualdad de ubicación y tamaño, una mayor dotación de baños y mejores condiciones del espacio interior se asocian a valoraciones más altas. Elementos de accesibilidad y confort (como disponer de ascensor o climatización) contribuyen a afinar la estimación, ayudando a distinguir activos con superficies similares pero prestaciones diferentes.

- **Modelos de boosting (XGBoost, CatBoost, LightGBM)**

Ofrecen el mejor desempeño predictivo al capturar no linealidades e interacciones entre variables. En las comparativas realizadas, **XGBoost** se

sitúa como "campeón técnico", destacando por su capacidad para reducir el error en test y mantener resultados consistentes frente a variantes del conjunto de datos. CatBoost y LightGBM presentan rendimientos muy próximos, lo que confirma la solidez del enfoque de *boosting* en este problema.

- **RandomForest**

Se posiciona como el modelo más **robusto y recomendable para producción**, gracias a su estabilidad en distintos subconjuntos de datos, su menor sensibilidad a ajustes finos de configuración y su **facilidad de implementación y mantenimiento**. Aporta, además, una interpretabilidad operativa suficiente (importancia de variables) para justificar decisiones ante equipos no técnicos, con un rendimiento prácticamente equivalente al del mejor modelo técnico.

- **Sistema predictivo**

Reduce de forma significativa el error frente a métodos tradicionales de tasación basados en promedios zonales o relaciones lineales simples, **especialmente en viviendas estándar y en zonas premium**. En segmentos con mayor dispersión (rango medio y activos singulares) el sistema mantiene utilidad como referencia central, acompañando la cifra con una lectura segmentada por bandas de precio y áreas para gestionar expectativas y asegurar decisiones más informadas.

Detalle en **Anexo II**: Diego\_Olalla\_Carrión\_Notebook\_2 (**métricas en test, comparativa de boosting vs. RF, importancias/SHAP y análisis por segmentos**) y Diego\_Olalla\_Carrión\_Notebook\_1 (**EDA por ubicación**).

## 7.2 Impacto esperado en el negocio o la práctica profesional

La implantación del sistema tiene un alcance **transversal** en el sector inmobiliario porque eleva la calidad de la información con la que se trabaja, homogeneiza criterios entre equipos y **reduce la incertidumbre** en decisiones de precio, inversión y valoración. Para **promotoras y agencias**, aporta una referencia objetiva con la que fijar precios iniciales y ajustar la estrategia de comercialización por zona y tipología; facilita la negociación con explicaciones claras de los factores que han pesado, mejora la comparabilidad entre activos similares y permite priorizar la cartera (qué publicar primero, dónde concentrar esfuerzos y qué inmuebles requieren revisar expectativas), anticipando ajustes antes de que un producto acumule visitas sin ofertas.

Para las **entidades financieras**, el sistema actúa como apoyo en tasaciones internas y en la gestión del riesgo: ofrece una estimación central defendible y una lectura segmentada por zonas y bandas de precio, ayuda a identificar expedientes que merecen revisión adicional cuando se detectan desviaciones relevantes frente a su entorno y mantiene trazabilidad sobre los supuestos utilizados. Con ello se reduce el riesgo de sobrevaloración o infravaloración y se gana consistencia entre decisiones de distintos comités o delegaciones.

En el caso de los **inversores**, proporciona un marco homogéneo para comparar oportunidades en áreas emergentes y en zonas consolidadas, detectar desajustes entre precio pedido y valor estimado y priorizar operaciones con mejor relación riesgo-retorno. Al acortar los ciclos de análisis y orientar la búsqueda hacia micro

áreas con señales de tracción, ayuda a definir escenarios de entrada y salida más realistas e identificar activos donde pequeñas mejoras pueden tener un impacto sensible en el valor.

Para las **administraciones públicas**, ofrece una base analítica que puede apoyar la planificación urbana y el seguimiento de dinámicas territoriales. La lectura por zonas y segmentos contribuye a detectar concentraciones de valor, áreas en transformación y posibles brechas, complementando diagnósticos sobre accesibilidad a la vivienda y orientando actuaciones focalizadas, con la debida salvaguarda de la privacidad.

Por último, para la **ciudadanía** incrementa la transparencia en la valoración: aporta referencias comprensibles y comparables por zona y tipología, ayuda a formar expectativas más realistas, reduce asimetrías de información en la negociación y favorece decisiones de compra, venta o alquiler mejor fundamentadas, creando un entorno más previsible y menos expuesto a decisiones impulsivas.

En conjunto, el impacto esperado se traduce en **mayor eficiencia, reducción de la incertidumbre y ventaja competitiva** en un mercado complejo como el madrileño. Al estandarizar criterios, documentar supuestos y ofrecer estimaciones explicables, el sistema no sustituye el juicio experto: lo refuerza con una base de evidencia consistente y operativa.

**Detalle en Anexo II:** Diego\_Olalla\_Carrión\_Notebook\_2 (**segmentación por zonas/bandas, ejemplos de priorización y control de riesgo**); Diego\_Olalla\_Carrión\_Notebook\_1 (**comparabilidad por zona y tipología**).

### 7.3 Lecciones aprendidas

El desarrollo del proyecto permitió extraer varias lecciones de valor metodológico y profesional:

- **Equilibrio entre precisión y operatividad:** Los modelos más precisos en términos métricos no siempre son los más adecuados para producción. Como demostró la comparativa entre XGBoost y RandomForest, es clave equilibrar precisión, robustez e interpretabilidad para un despliegue estable.
- **Potencia de la ingeniería de características:** La transformación de variables "planas" a interacciones de mercado marcó la diferencia en el rendimiento predictivo, aportando mejoras superiores al 20% y refinando la capacidad explicativa del modelo.
- **Tratamiento estratégico de datos incompletos:** El manejo de valores faltantes y outliers no debe entenderse como un proceso de limpieza pasiva, sino como una fuente activa de información contextual que enriquece la capacidad predictiva.
- **Importancia de la segmentación:** La división por tipo de mercado (estándar, premium, lujo) se confirmó como estrategia clave para mejorar la aplicabilidad práctica, adaptando el modelo a las particularidades de cada segmento.
- **Enfoque orientado a valor de negocio:** El éxito de un proyecto de ciencia de datos no reside únicamente en la métrica estadística, sino en su capacidad de generar valor operativo y ser adoptado por los actores implicados en el ecosistema inmobiliario.

- **Diseño ético integrado:** La consideración de aspectos sociales y éticos (como el riesgo de sesgos geográficos o la necesidad de equidad en tasaciones) es esencial para que la analítica avanzada tenga aceptación y legitimidad en sectores sensibles como el inmobiliario.

El proyecto demuestra que la analítica avanzada puede transformar la manera en que se valoran las viviendas en Madrid, siempre que se combine el rigor técnico con un diseño orientado a negocio. El sistema predictivo propuesto no sustituye la labor experta, sino que la complementa, ofreciendo un marco sólido para la toma de decisiones estratégicas y contribuyendo a un mercado más eficiente, transparente y justo.

La aproximación holística adoptada desde la selección del modelo hasta la integración de consideraciones éticas ha permitido desarrollar una solución que no solo alcanza métricas de precisión significativas, sino que también ofrece una propuesta de valor clara para los distintos actores del sector inmobiliario madrileño, demostrando que la ciencia de datos puede aportar beneficios tangibles cuando se aplica con visión práctica y responsable.

**Detalle en Anexo II:** Diego\_Olalla\_Carrión\_Notebook\_2 (**comparativa XGBoost vs RandomForest, mejora por ingeniería de variables, segmentación**); Diego\_Olalla\_Carrión\_Notebook\_1 (**gestión de ausentes y outliers**).

## 8. Bibliografía y referencias

- XGBoost. (s. f.). Python Package Introduction. En XGBoost Documentation (v3.0.4).  
[https://xgboost.readthedocs.io/en/stable/python/python\\_intro.html](https://xgboost.readthedocs.io/en/stable/python/python_intro.html)
- W3Schools. (s. f.). Python Machine Learning: Decision Tree.  
[https://www.w3schools.com/python/python\\_ml\\_decision\\_tree.asp](https://www.w3schools.com/python/python_ml_decision_tree.asp)
- Ciencia de Datos. (s. f.). Machine learning con Python.  
<https://cienciadedatos.net/machine-learning-python>
- freeCodeCamp. (s. f.). Machine Learning with Python.  
<https://www.freecodecamp.org/learn/machine-learning-with-python>
- scikit-learn. (s. f.). scikit-learn: Machine Learning in Python (v1.7.1).  
<https://scikit-learn.org/>
- The Devastator. (s. f.). Spanish Housing Dataset: Location, Size, Price, and More! Kaggle. [https://www.kaggle.com/datasets/thedevastator/spanish-housing-dataset-location-size-price-and/data?select=houses\\_madrid.csv](https://www.kaggle.com/datasets/thedevastator/spanish-housing-dataset-location-size-price-and/data?select=houses_madrid.csv)
- Idealista. (2019). Informe de precios de venta en España de 2019. Idealista.  
<https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/report/2019/>
- Ayuntamiento de Madrid. (2019). Anuario Estadístico 2019: Capítulo 8. Planeamiento Urbano y Vivienda. Ayuntamiento de Madrid.  
<https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Publicaciones/anuesta/nuevos/Anuario%20Estad%C3%A9stico%20Municipal/Anuario%20estad%C3%A9stico%202019/Cap%C3%A9tulos/Cap%C3%A9itulo%20Planeamiento%20Urbano%20y%20Vivienda.%202019.pdf>

## **9. Anexos**

Los anexos contienen el detalle técnico del trabajo, complementando la memoria principal orientada a negocio. Se han estructurado de forma modular para facilitar su consulta y trazabilidad:

### **Anexo I. Documento técnico**

- Análisis exploratorio de datos (EDA).
- Preparación y limpieza de datos.
- Ingeniería de características.
- Modelado y pruebas de algoritmos.
- Métricas de validación y evaluación.
- Interpretabilidad de modelos

### **Anexo II. Código del proyecto**

- Notebooks en Jupyter (*Diego\_Olalla\_Carrión\_Notebook\_1.ipynb*, *Diego\_Olalla\_Carrión\_Notebook\_2.ipynb*).

### **Anexo III. Enlace Google Drive**

- <https://drive.google.com/drive/folders/1vfNGLOpnxx3VAtM1Nb2-sQ8WpPSdqlB?usp=sharing>



ntic master



UNIVERSIDAD  
**COMPLUTENSE**  
MADRID



Curso académico  
2024/2025