ntic master

UNIVERSIDAD COMPLUTENSE MADRID

Anexo I. Documento técnico:

Análisis exploratorio, preparación de datos y modelado

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. INTRODUCCIÓN Y OBJETIVOS

1.1 Contexto del mercado inmobiliario madrileño

El mercado residencial de Madrid constituye uno de los sectores más relevantes y complejos de España, tanto por su peso económico como por la diversidad de factores que influyen en los precios. En la última década ha estado marcado por la presión demográfica (concentrando el 14% de la población nacional), la diversificación económica derivada de su rol como hub financiero y tecnológico, la elevada heterogeneidad territorial con diferencias de más del 300% entre distritos y la creciente sofisticación del mercado, donde cobran importancia elementos como la eficiencia energética o la sostenibilidad. Estas características hacen necesaria la aplicación de modelos predictivos avanzados que vayan más allá de las valoraciones tradicionales basadas en comparables simples.

Características distintivas del mercado madrileño:

Característica	**Descripción**	**Impacto en valoración**
Concentración geográfica	Distrito de Salamanca representa el 6,2% de la oferta pero concentra el 70% de propiedades premium	Alta variabilidad de precios por m²
Prima por antigüedad	Propiedades históricas con sobreprecio del 15-25%	Relación inversa entre antigüedad y depreciación
Servicios críticos	Importancia de ascensores en edificios históricos y sistemas de climatización	Variables con alto poder predictivo

1.2 Objetivos del estudio

El objetivo principal del proyecto es desarrollar un sistema predictivo de alta precisión para la estimación de precios inmobiliarios en Madrid mediante técnicas de machine learning. Para alcanzarlo se plantean los siguientes objetivos específicos: realizar un análisis exploratorio sobre 5.255 propiedades, cuantificar el impacto de las variables categorizadas (ubicación, características y servicios), diseñar una arquitectura predictiva basada en múltiples algoritmos con validación cruzada, implementar ingeniería de características para capturar sinergias, aplicar una validación robusta con el fin de superar un R^2 del 90% y finalmente generar recomendaciones estratégicas basadas en los resultados obtenidos.

1.3 Descripción del conjunto de datos

El dataset empleado contiene 5.255 registros y 36 variables correspondientes a propiedades residenciales en 11 distritos y 171 barrios de Madrid durante 2019, con precios que oscilan entre 39.000€ y 9.488,250€. La información se organiza en distintas categorías de variables con diferentes niveles de completitud:

Categoría	Variables	Ejemplos representativos	Completitud
Identificación	2	house_id, obtention_date	100%
Geolocalización	6	loc_district, loc_neigh, loc_street	95-100%
Características estructurales	6	m2_real, m2_useful, room_num, bath_num	85-100%
Condición y construcción	3	condition, construct_date, energetic_certif	43-100%
Servicios y amenidades	12	air_conditioner, garage, lift, swimming_pool	35-100%
Variable objetivo	1	price (€)	100%
Metadatos	6	ad_description, orientation, heating	47-100%

En cuanto a las variables clave, destacan las numéricas (precio medio de 474.523€ con elevada dispersión, superficie real de 119,8 m², media de 2,8 habitaciones y 1,8 baños), las categóricas (distrito, barrio, estado de la vivienda y tipo de calefacción) y doce binarias que reflejan servicios con tasas de presencia del 15% al 85%. Los principales desafíos metodológicos incluyen valores faltantes en variables estratégicas (como ground_size o kitchen), distribuciones altamente asimétricas en precio y superficie, elevada cardinalidad en categorías (171 barrios, 51 plantas, 16 orientaciones) y la presencia de outliers multivariantes, especialmente en el distrito de Salamanca.

2. ANÁLISIS EXPLORATORIO DE DATOS: MERCADO INMOBILIARIO DE MADRID

2.1 Estructura y Composición del Conjunto de Datos

La estructura integra variables numéricas continuas y discretas, binarias y categóricas, con aplicaciones que van desde la segmentación hasta el modelado predictivo. La completitud es excelente en variables críticas como precio y localización, aceptable en variables de configuración (habitaciones y baños) y limitada en aspectos como fecha de construcción y orientación.

Composición Estructural del Dataset:

Categoría de Variable	Cantidad	Ejemplos Representativos	% Dataset	Aplicación Analítica
Variables Numéricas Continuas	4	price, m2_real, price_per_m2, construct_date	23,5%	Modelado predictivo, análisis de tendencias

Categoría de Variable	Cantidad	Ejemplos Representativos	% Dataset	Aplicación Analítica
Variables Numéricas Discretas	2	room_num, bath_num	11,8%	Segmentación, análisis de configuración
Variables Binarias/Booleanas	11	air_conditioner, lift, garage, terrace	64,7%	Evaluación de amenidades, scoring de propiedades
Variables Categóricas	14	loc_district, house_type, condition	-	Análisis geoespacial, segmentación cualitativa

Estadísticas de Completitud:

- Excelente (>95%): precio, house_type, loc_district, amenidades básicas
- Buena (70-95%): room_num, bath_num, condition
- Mejorable (<70%): construct_date (42.8%), orientation (46,3%)

2.2 Análisis de la Variable Objetivo: Precio

El precio de las propiedades refleja la segmentación natural del mercado madrileño, desde vivienda social hasta propiedades ultra-premium. La media (474.523 €) está muy por encima de la mediana (247.000 €), lo que evidencia fuerte asimetría y dispersión. La desviación estándar confirma la heterogeneidad del mercado y la amplitud del rango revela oportunidades en distintos niveles de especialización.

Estadísticas Descriptivas del Precio:

Métrica	Valor	Interpretación Empresarial	Implicación Estratégica
Media	474.523 €	Precio medio global del mercado	Punto de referencia para segmentación
Mediana	247.000 €	Valor típico de transacción	Indicador de mercado mid-tier dominante
Desviación Estándar	670.215 €	Alta heterogeneidad del mercado	Necesidad de estrategias por segmento
Rango	39.000 € - 9.488.250 €	Espectro desde social hasta ultra-lujo	Especialización segmentada

Métrica	Valor	Interpretación Empresarial	Implicación Estratégica
Coeficiente de Variación	1,41	Dispersión superior a la media	Mercado altamente variable
Ratio Media/Mediana	1,92	Asimetría distributiva significativa	Polarización del mercado

La distribución presenta asimetría positiva elevada (4,28), curtosis leptocúrtica (26,73) y multimodalidad con tres concentraciones principales en 150K€, 250K€ y 425K€.

2.3 Exploración de Variables Numéricas

El análisis de variables numéricas muestra patrones relevantes: la superficie construida presenta segmentos diferenciados (económico, estándar, premium y lujo), la configuración habitacional está dominada por viviendas de 2-3 habitaciones y 1-2 baños, y la antigüedad evidencia un patrón interesante donde las propiedades históricas céntricas retienen un valor premium.

Análisis Estadístico de Variables Numéricas Principales:

Variable	Media	Mediana	Desv. Std	Skewness	Kurtosis	Completitud	Correlación con Precio
m2_real	119,80	90,00	174,94	29,90	1,243,71	98.2%	0,534***
room_num	2,80	3,00	1,18	1,06	4,78	87,4%	0,430***
bath_num	1,75	2,00	0,95	1,82	6,95	85,9%	0,637***
construct_date	1975,75	1974,00	26,52	-0,76	0,97	42,8%	- 0,310***
price_per_m2	4.632,17	3.023,81	5.891,23	8,44	157,32	98,2%	0,352***

2.4 Exploración de Variables Categóricas

Las variables categóricas aportan información fundamental sobre la ubicación, tipología y descripción de los inmuebles. Se aplicaron distintas estrategias de codificación, desde la eliminación de identificadores hasta técnicas de target encoding, one-hot encoding y procesamiento de lenguaje natural en descripciones textuales.

Análisis de Variables Categóricas y Estrategias de Codificación:

Nivel	Variable	Categorías	Valor Más Común	%	Ratio Precio	Estrategia Recomendada	Justificación
Alta	house_id	5.248	-	0,02 %	-	Eliminación	Identificador sin valor predictivo
Alta	ad_description	4.817	-	0,60 %	-	Procesamiento NLP + TF-IDF	Extracción de características latentes
Media	loc_neigh	171	Recoletos	10,2 %	3,24x	Target encoding + clustering	Reducción dimensional preservando información
Media	floor	51	"planta 1ª exterior"	18,9 %	1,86x	Extracción estructurada	Separación en nivel y posición
Baja	house_type	9	"Piso"	85,2 %	0,94x	One-hot encoding	Preservación categórica completa
Baja	loc_district	11	"Tetuán"	26,3 %	0,87x	One-hot encoding + clustering	Preservación con agrupación selectiva

El análisis geoespacial jerárquico confirma la importancia de la ubicación: Madrid capital duplica los precios de la periferia, con variabilidad extrema entre distritos (3,74x) y barrios (8,27x).

2.5 Análisis de Correlaciones y Multicolinealidad

Las correlaciones muestran que el número de baños es el predictor estructural más fuerte, seguido de superficie, habitaciones y localización en Salamanca. También destacan como relevantes las amenidades de accesibilidad y confort. El análisis VIF confirma la existencia de clusters de variables correlacionadas que deben tratarse para evitar multicolinealidad.

Principales Correlaciones con Variable Precio:

Variable	Correlación con Precio	Significancia	VIF	Interpretación Empresarial
bath_num	0,637	***	3,82	Predictor premium más relevante
m2_real	0,534	***	5,14	Driver fundamental con efecto no lineal
room_num	0,430	***	2,96	Configuración espacial con valor incremental
loc_district_Salamanca	0,387	***	2,14	Ubicación premium de mayor impacto
lift	0,250	***	1,67	Amenidad de accesibilidad
air_conditioner	0,242	***	1,58	Confort climático significativo
balcony	0,200	***	1,32	Espacio exterior como diferenciador
construct_date	-0,310	***	2,21	Valoración premium en ubicaciones históricas
loc_district_Villaverde	-0,287	***	1,83	Ubicación con mayor descuento relativo

Clusters identificados:

- **Espacial:** m2_real ↔ room_num ↔ bath_num
- **Calidad:** lift ↔ air_conditioner ↔ condition
- **Ubicación:** loc_district ↔ price_per_m2 ↔ construct_date

2.6 Identificación y Tratamiento de Valores Atípicos

El análisis de outliers se realizó con métodos univariados, multivariados y de influencia, confirmando la presencia de propiedades ultra-premium y superficies extremas. Se adoptó una estrategia diferenciada: retención con transformación logarítmica en precios altos, corrección de errores de medición en superficies y conservación de outliers multivariados mediante flags.

Resumen de Detección de Outliers por Método:

Método	Variable	Outliers	% Dataset	Umbrales Aplicados	Caracterización	Estrategia
IQR Univariado	price	681	12,98%	<-825K, >2,175M €	Segmento ultra-premium	Transformación logarítmica
IQR Univariado	m2_real	445	8,48%	<-47,5, >292,5 m²	Propiedades atípicas	Verificación + transformación
Z-Score (z>3)	price	234	4,46%	>2,48M €	Segmento lujo extremo	Retención selectiva
Isolation Forest	Multivariado	53	1,01%	Contamination=0,01	Combinaciones inusuales	Análisis caso por caso

2.7 Conclusiones del Análisis Exploratorio

El análisis exploratorio revela cinco hallazgos principales: la existencia de una estructura multimodal con tres segmentos diferenciados, la predominancia de la ubicación como factor que explica hasta un 74% de la varianza, el número de baños como predictor más potente, un patrón bimodal en la antigüedad con primas tanto en propiedades históricas como nuevas, y la interacción contextual en el valor de amenidades según la localización.

Las recomendaciones para el modelado incluyen un enfoque híbrido basado en transformaciones logarítmicas y segmentación, la gestión de multicolinealidad mediante clusters y la incorporación de términos de interacción para capturar efectos contextuales.

3. PREPROCESAMIENTO DE DATOS INMOBILIARIOS

3.1 Gestión Estratégica de Valores Faltantes

En el mercado inmobiliario, la ausencia de datos no siempre implica simple omisión, sino que puede contener información relevante sobre las características de los inmuebles o las prácticas del mercado. Por ello, se aplicó una estrategia de enfoque múltiple adaptada a los patrones de cada variable. El diagnóstico identificó ausencias tanto sistemáticas como aleatorias o estructurales, lo que derivó en metodologías específicas de imputación.

Diagnóstico y Estrategias para Valores Faltantes:

Variable	Valores Faltantes	% Ausencia	Patrón Identificado	Estrategia Aplicada
construct_date	3.003	57,22%	Sistemático	Indicador binario + Imputación por mediana condicional
orientation	2.817	53,68%	Aleatorio	Creación de categoría "Desconocido" + Análisis de texto
heating	2.398	45,69%	Estructural	Imputación condicional por distrito/zona
loc_street	1.365	26,01%	Geográfico	Agrupación jerárquica por distrito
energetic_certif	1.265	24,10%	Regulatorio	Imputación por año construcción + normativa aplicable
planta_num	522	9,95%	Técnico	KNN (k=5) con variables espacialmente correlacionadas
room_num	135	2,57%	Estructural	Mediana condicionada por superficie y tipología
floor	120	2,29%	Descriptivo	Extracción desde descripción + moda por tipología
bath_num	6	0,11%	Puntual	Mediana condicionada por precio y superficie

La metodología se adaptó según el nivel de ausencia: en variables con más del 20% se aplicaron indicadores de ausencia combinados con imputación contextual, en el rango de 5-20% se utilizó KNN con validación cruzada y en casos inferiores al 5% se recurrió a medidas de tendencia central. El resultado fue un dataset sin valores faltantes, con preservación estadística (±5%) y mantenimiento de la integridad informativa.

3.2 Transformación Inteligente de Variables Categóricas

El tratamiento de variables categóricas buscó equilibrar la riqueza informativa con la necesidad de mantener una dimensionalidad manejable, considerando además la jerarquía geográfica propia del mercado madrileño. Se aplicaron estrategias diferenciadas según el nivel de cardinalidad y la naturaleza de las variables.

Estrategias de Codificación Optimizadas:

Tipo de Variable	Cantidad	Técnica Implementada	Variables Ejemplo	Justificación Metodológica
Alta Cardinalidad	4	Eliminación selectiva / NLP	house_id, ad_description	Evita curse of dimensionality y sobreajuste
Media Cardinalidad	3	Target Encoding regularizado	loc_neigh	Maximiza señal predictiva con control de varianza
Baja Cardinalidad	7	One-Hot Encoding optimizado	house_type, condition	Preserva naturaleza nominal sin asumir orden
Ordinal	1	Codificación Ordinal jerárquica	floor	Mantiene relaciones verticales intrínsecas
Multidireccional	1	Encoding Direccional	orientation	Captura características espaciales complejas

En su implementación técnica, se utilizó Target Encoding con regularización bayesiana, compresión de categorías infrecuentes en one-hot encoding y codificación vectorial para orientación espacial. Como resultado, el dataset pasó de 43 a 65 variables, manteniendo controlada la dispersión (<25%) y reduciendo la entropía informativa a menos del 3%.

3.3 Normalización Adaptativa de Variables Numéricas

Dada la alta dispersión y asimetría de muchas variables, se aplicó un esquema de normalización adaptativa en función de sus características distribucionales. Esto permitió optimizar el desempeño de algoritmos sensibles a la escala sin distorsionar la información subyacente.

Análisis Distribucional y Estrategia de Normalización:

Variable	Media	Mediana	Desv. Std	Skewness	Kurtosis	Técnica Aplicada	Justificación
price	474.523 €	247.000 €	670,215	4,28	26,73	RobustScaler	Distribución con cola larga y outliers legítimos

Variable	Media	Mediana	Desv. Std	Skewness	Kurtosis	Técnica Aplicada	Justificación
m2_real	119,80 m²	90,00 m²	174.94	29.90	1.243,71	Log + MinMaxScaler	Extrema asimetría positiva
price_per_m2*	3.962 €/m²	2.744 €/m²	3.215	2,17	9,24	RobustScaler	Mejor comportamiento estadístico
room_num	2,80	3,00	1,18	1,06	4,78	RobustScaler	Distribución discreta con sesgo moderado
bath_num	1,75	2,00	0,95	1,82	6,95	RobustScaler	Valores atípicos significativos
floor_encoded	1,45	1,00	2,48	5,73	133,21	MinMaxScaler	Preservación de proporcionalidad ordinal
planta_num	2,15	2,00	3,42	4,54	93,04	RobustScaler	Valores extremos con significado inmobiliario

*Variable derivada que sustituye a m2_real por mejor comportamiento estadístico.

El framework incluyó el uso de RobustScaler en variables con outliers, transformación logarítmica en aquellas con skewness extremo y MinMaxScaler para escalas ordinales. Tras su aplicación, se preservó el 98,7% de la correlación entre variables, se redujo el sesgo (skewness promedio de 7,89 a 1,62) y el tiempo de entrenamiento de modelos cayó un 78%.

El dataset resultante cuenta con 5.248 observaciones y 0,92 completamente limpias, con un balance de 37% numéricas y 63% categóricas, y una eficiencia computacional mejorada en un 65% respecto al estado inicial.

4. INGENIERÍA DE CARACTERÍSTICAS: OPTIMIZACIÓN PREDICTIVA DEL MODELO INMOBILIARIO

4.1 Transformación Logarítmica de la Variable Objetivo

La variable precio presentaba una distribución altamente asimétrica que podía distorsionar los modelos predictivos. Para mitigar este efecto se aplicó una

transformación logarítmica, lo que redujo significativamente la asimetría y la curtosis, acercando la distribución a la normalidad. Esto generó impactos estadísticos relevantes, como la homogeneización de la varianza y la mejora en la estabilidad frente a outliers, además de favorecer la convergencia en algoritmos como Ridge, reduciendo un 42% las iteraciones requeridas. En términos prácticos, la transformación estableció una base robusta para la modelización, permitiendo capturar mejor la dinámica de todos los segmentos de mercado.

4.2 Ingeniería Avanzada de Características Derivadas

Se diseñaron variables derivadas para capturar interacciones complejas del mercado madrileño, mejorando la capacidad predictiva del modelo. Estas variables de interacción, como la densidad de baños, el tamaño premium o la combinación de edad y superficie, contribuyeron de forma significativa al incremento en R^2, logrando mejoras de hasta un 23,4% frente al uso exclusivo de variables originales.

Variables de interacción de alto impacto:

Variable derivada	**Metodología de construcción**	**Importancia relativa**	**Incremento predictivo**
premium_district_size	m² × indicador_distrito_premium	11,4%	+18,7%
bath_density	n_baños / m²	9,0%	+15,2%
age_size	(año_actual - año_construcción) × m²	6,2%	+9,8%
size_quality	m² × índice_calidad_compuesto	4,3%	+7,5%
rooms_bathrooms	Ratio habitaciones/baños normalizado	4,7%	+6,9%

Se incorporaron también características refinadas de ubicación (índice de centralidad, factor de prestigio zonal, gradiente norte-sur), además de descomposiciones optimizadas de orientaciones, niveles de planta y tipologías mediante técnicas vectoriales y de embeddings supervisados.

4.3 Estratificación Inteligente de Variables Categóricas

Para reducir la dimensionalidad manteniendo la capacidad explicativa, se aplicaron técnicas avanzadas de agrupación y codificación en variables categóricas de alta cardinalidad. En particular, la optimización de barrios (171 categorías) permitió consolidarlos en 15 agrupaciones sin pérdida significativa de precisión, mientras que las orientaciones arquitectónicas se reorganizaron en grupos de impacto diferencial en precio.

Optimización barrios (171 → 15 categorías):

Metodología	Implementación	Rendimiento
Target encoding regularizado	Precio medio con penalización bayesiana (λ=10)	R^2=0,891
Clustering jerárquico	Clustering jerárquico de barrios según precios	R^2=0,889
Quintiles con corrección	5 grupos + categoría "Otros" para <10 propiedades	R^2=0,886
One-hot encoding original	171 variables dummy (línea base)	R^2=0,878

Refinamiento de orientaciones arquitectónicas:

Grupo	Orientaciones incluidas	Efecto precio	Frecuencia
Premium	Sur, Sureste, Suroeste	+15-20%	32,7%
Intermedio	Este, Oeste, Norte-Sur	+5-10%	41,2%
Básico	Norte, Noroeste, Interior	Referencia	26,1%

Asimismo, los estados de conservación se transformaron en una escala ordinal 1-10 derivada de descripciones textuales, ajustada con factores de antigüedad. Esta estrategia redujo la dimensionalidad en un 40% y aportó un incremento adicional en el R^2 ajustado (+0,013).

4.4 Selección Multidimensional de Características

El proceso de selección se desarrolló en cuatro fases complementarias: reducción de redundancia, filtrado estadístico, selección basada en modelos y optimización con RFE-CV. Esto permitió pasar de 203 a 79 características finales, balanceando relevancia predictiva y eficiencia.

Top-10 características por importancia global:

Posición	Característica	Importancia	Estabilidad CV	Categoría
1	premium_district_size	11,4%	±0,8%	Derivada
2	bath_density	9,0%	±1,1%	Derivada
3	m2	7,8%	±0,4%	Original
4	age_size	6,2%	±0,9%	Derivada

Posición	Característica	Importancia	Estabilidad CV	Categoría
5	loc_district_Salamanca	5,1%	±0,6%	Categórica
6	rooms_bathrooms	4,7%	±0,7%	Derivada
7	size_quality	4,3%	±0,5%	Derivada
8	lift	3,9%	±0,3%	Original
9	loc_district_Tetuán	3,5%	±0,4%	Categórica
10	balcony	3,2%	±0,5%	Original

4.5 Reducción Dimensional para Optimización Computacional

Para garantizar la escalabilidad y eficiencia, se evaluaron diferentes configuraciones de PCA, equilibrando precisión, coste computacional e interpretabilidad. La versión con 30 componentes mantuvo un R^2 cercano al original con un tiempo de inferencia un 62% más rápido, lo que la convierte en la opción óptima para aplicaciones en tiempo real.

Análisis comparativo de configuraciones PCA:

Configuración	Componentes	Varianza explicada	R^2 Test	Tiempo inferencia	Uso memoria
Original_79	79	-	0.914	1,00x	1,00x
PCA_30	30	100,0%	0.884	0,38x	0,42x
PCA_20	20	100,0%	0.857	0,29x	0,31x
PCA_90%	2	94,1%	0.641	0,11x	0,14x

La interpretación semántica de los componentes principales confirmó tres ejes fundamentales: un factor tamaño-calidad, un factor antigüedad-estado y un factor singularidad.

4.6 Conclusiones y Valor de Negocio

La estrategia de ingeniería de características transformó sustancialmente el potencial predictivo del modelo inmobiliario. Entre los resultados destacan un incremento del 23,4% en la capacidad predictiva respecto a variables originales, una reducción dimensional del 40% manteniendo un 96,7% del rendimiento, una mejora del 89% en la velocidad de inferencia y una reducción del 72% en el error de valoración del segmento premium.

Desde el punto de vista de negocio, los beneficios incluyen mayor precisión segmentada en valoraciones, escalabilidad para procesar millones de propiedades en

tiempo real, adaptabilidad a la evolución del mercado y una interpretabilidad alineada con conceptos estratégicos como tipología, ubicación y calidad constructiva.

5. DESARROLLO DE MODELOS PREDICTIVOS

5.1 División de Datos y Estrategia de Validación

Para garantizar la robustez y generalización de los resultados, se aplicó una estrategia de validación multinivel. La muestra se dividió en un 80% para entrenamiento y un 20% para test, manteniendo la estratificación por distritos y rangos de precio. Además, se utilizó validación cruzada k-fold con k=5, estratificada por quintiles de precio y con semilla fija para asegurar reproducibilidad. Las desviaciones estándar se mantuvieron por debajo de 0,015 en todos los modelos, confirmando estabilidad. Como medidas anti-overfitting, se monitorizó la brecha entre train y test (con umbral 0,05), se aplicó regularización adaptativa y se incorporó early stopping en los algoritmos de boosting.

5.2 Modelos Lineales

Los modelos lineales ofrecieron una base interpretable y robusta para el análisis inmobiliario. Ridge, Lasso y ElasticNet mostraron un rendimiento muy similar, con R^2 alrededor de 0,888, RMSE en torno a 0,236 y diferencias train-test reducidas (<0,011). Ridge destacó por la estabilidad de sus coeficientes y su capacidad para identificar variables influyentes como la superficie y la pertenencia a distritos premium. Lasso eliminó 23 variables irrelevantes, mostrando utilidad en selección automática, mientras que ElasticNet resultó más adecuado para propiedades de gama media al balancear regularizaciones L1 y L2.

Rendimiento de Modelos Lineales:

Modelo	**R^2 Test**	**RMSE**	**MAE**	**Diferencia Train-Test**	**Tiempo Entrenamiento (s)**
Ridge	0,8881	0,2357	0,1847	0,0102	0,84
Lasso	0,8875	0,2364	0,1851	0,0098	0,97
ElasticNet	0,8879	0,2361	0,1849	0,0100	1,22

5.3 Modelos Basados en Árboles

Los modelos de árboles fueron capaces de capturar relaciones no lineales del mercado, ofreciendo mayor precisión que los lineales. RandomForest alcanzó el mejor rendimiento (R^2=0,9490), siendo dominado por variables derivadas como premium_district_size y bath_density. GradientBoosting mostró un excelente equilibrio entre precisión y generalización, con la menor brecha train-test (0,0132). En cambio, DecisionTree evidenció sobreajuste extremo y AdaBoost fue limitado por su sensibilidad a outliers.

Rendimiento de Modelos de Árboles:

Modelo	R^2 Test	RMSE	MAE	Diferencia Train-Test	R^2 CV (media ± std)
RandomForest	0,9490	0,1735	0,1258	0,0439	0,9462 ± 0,0081
GradientBoosting	0,9410	0,1866	0,1432	0,0132	0,9399 ± 0,0071
DecisionTree	0,8860	0,2594	0,1742	0,1140	0,8791 ± 0,0215
AdaBoost	0,8816	0,2644	0,2058	0,0532	0,8803 ± 0,0124

5.4 Modelos Avanzados

Los algoritmos de boosting confirmaron su liderazgo en predicción inmobiliaria. XGBoost fue el modelo con mejor rendimiento global (R^2=0,9506), destacando por alcanzar un 72,3% de predicciones con error menor a ±10%. CatBoost logró resultados casi idénticos con un sobreajuste mínimo gracias a su tratamiento nativo de variables categóricas. LightGBM ofreció gran eficiencia computacional, reduciendo en un 60% el tiempo de entrenamiento. En contraste, SVR y KNN obtuvieron rendimientos inferiores, aunque mostraron utilidad potencial en ensambles.

Rendimiento de Modelos Avanzados:

Modelo	R^2 Test	RMSE	MAE	Diferencia Train-Test	Tiempo Inferencia (ms/pred)
XGBoost	0,9506	0,1709	0,1195	0,0222	0,87
CatBoost	0,9494	0,1728	0,1203	0,0096	1,12
LightGBM	0,9471	0,1767	0,1221	0,0154	0,43
SVR (RBF)	0,9229	0,2101	0,1447	0,0641	3,56
KNN_5	0,8553	0,3027	0,2112	0,1442	2,34

5.5 Análisis Comparativo y Recomendaciones

El ranking global sitúa a XGBoost como líder, seguido de cerca por CatBoost y RandomForest. La correlación entre modelos de boosting fue superior al 0.99, lo que permitió construir un metamodelo de ensamble que alcanzó un R^2 de 0,9521. El análisis de errores por segmentos reveló un mayor margen de mejora en propiedades de gama media-alta (600K-1M€) y un reto especial en el segmento de lujo (>1M€), donde la variabilidad intrínseca del mercado complica la precisión.

Análisis de errores por segmento:

Rango de Precio	Error Medio	Característica Principal
< 300K€	7,3%	Subestimación
300K-600K€	9,1%	Distribución equilibrada
600K-1M€	11,4%	Sobrestimación
> 1M€	13,8%	Alta variabilidad

Recomendaciones de implementación:

- Máxima precisión: XGBoost como referencia técnica y control de calidad.
- Interpretabilidad: RandomForest en informes ejecutivos.

En términos de negocio, el modelo logró un 27,7% de predicciones con error inferior al ±5% y un 72,3% dentro del ±10%, consolidando su aplicabilidad práctica en un mercado altamente complejo como el madrileño.

6. EVALUACIÓN Y OPTIMIZACIÓN

6.1 Métricas de Evaluación y Comparación de Modelos

La evaluación de los modelos se realizó mediante un marco multi-métrico que integró precisión estadística, estabilidad y aplicabilidad al negocio. Los modelos de boosting (XGBoost, CatBoost y LightGBM) mostraron los mejores resultados en términos de R^2 y error absoluto, mientras que RandomForest se destacó por su estabilidad.

Métricas Estadísticas Tradicionales:

Modelo	R^2 Test	RMSE	MAE	R^2 Train-Test Gap	CV R^2 (Mean ± Std)
XGBoost	0,9506	0,1709	0,1195	0,0222	0,9284 ± 0,0081
CatBoost	0,9494	0,1728	0,1203	0,0096	0,9478 ± 0,0071
RandomForest	0,9490	0,1735	0,1258	0,0439	0,9462 ± 0,0081
LightGBM	0,9471	0,1767	0,1221	0,0185	0,9456 ± 0,0089
GradientBoosting	0,9410	0,1866	0,1432	0,0132	0,9399 ± 0,0071

En métricas de negocio, el porcentaje de predicciones con error inferior al ±5% fue bajo (12,8-18,7%), pero se alcanzó hasta un 27.7% dentro del ±10% y un 64,3% dentro del ±20% en configuraciones ensemble. La correlación entre modelos de boosting (>0,99) justificó la construcción de un metamodelo con R^2=0,9521.

6.2 Optimización de Hiperparámetros

La optimización se realizó con RandomizedSearchCV (5-fold), buscando balance entre cobertura y eficiencia. En XGBoost, la mejor configuración incluyó 148 estimadores, profundidad máxima de 10 y subsample de 0,8543, reduciendo sobreajuste y mejorando estabilidad.

Impacto de la Optimización por Modelo:

Modelo	R^2 Base	R^2 Optimizado	Mejora (%)	Tiempo (seg)
XGBoost	0,9455	0,9506	+0,51%	847
CatBoost	0,9448	0,9494	+0,49%	1.127
LightGBM	0,9421	0,9471	+0,53%	654
RandomForest	0,9435	0,9490	+0,58%	423

La brecha media Train-Test se redujo del 0,045 inicial a 0,022 tras la optimización.

6.3 Análisis de Importancia de Características

El análisis confirmó que la variable derivada premium_district_size domina el poder predictivo (51,5%), seguida por bath_density y otras interacciones complejas. En conjunto, las características geográficas aportaron cerca del 20% del poder explicativo, y las variables derivadas superaron consistentemente a las originales.

Top 10 Características Más Influyentes (XGBoost):

Ranking	Característica	Importancia	Interpretación de Negocio
1	premium_district_size	0,5147	Interacción tamaño × ubicación premium
2	bath_density	0,1236	Proporción baños/superficie
3	loc_district_Tetuán	0,1036	Ubicación específica valorada
4	size_quality	0,0921	Calidad del espacio habitacional
5	poly_m2_real × bath_num	0,0897	Interacción metros² × baños
6	lift	0,0142	Presencia de ascensor
7	loc_district_Villaverde	0,0134	Efecto distrito con descuento relativo
8	room_num	0,0129	Número de habitaciones
9	bath_num	0,0125	Número de baños
10	loc_district_San Blas	0,0119	Ubicación geográfica

6.4 Análisis de Residuos y Diagnóstico del Modelo

Los residuos confirmaron la existencia de heteroscedasticidad en rangos extremos y ligera desviación respecto a normalidad, aunque sin autocorrelación espacial relevante. El análisis segmentado mostró un rendimiento sólido en el rango medio del mercado, con debilidades en los segmentos de lujo y bajo coste.

Segmentación de Performance por Quintiles de Precio:

Quintil	Rango Precio (€)	R^2 Segmento	RMSE Segmento	Error % Medio	% Dentro ±10%
Q1 (Bajo)	600-900	0,8234	0,1456	8,9%	31,2%
Q2	900-1,200	0,9012	0,1389	11,2%	28,6%
Q3 (Medio)	1.200-1.600	0,9456	0,1234	9,7%	33,9%
Q4	1.600-2.200	0,9201	0,1567	13,4%	24,2%
Q5 (Alto)	2.200+	0,7891	0,2134	18,6%	15,8%

Los principales casos problemáticos fueron propiedades de lujo, inmuebles atípicos (127 casos con error >30%) y zonas en transformación urbanística.

6.5 Validación del Modelo Final

Se construyó un índice compuesto (0-100) que combinó precisión, robustez y consistencia segmentada. Ningún modelo alcanzó un umbral de confianza "alto", lo que refuerza la necesidad de supervisión humana en determinados casos. RandomForest obtuvo la mayor puntuación (60,1/100), seguido de GradientBoosting (58,2/100).

Ranking Final de Aplicabilidad:

Modelo	Score Negocio	Nivel Confianza	Recomendación
RandomForest	60,1/100	Medio	Supervisión humana
GradientBoosting	58,2/100	Bajo	Uso complementario
XGBoost	44,7/100	Bajo	Uso complementario
CatBoost	45,1/100	Bajo	Uso complementario

La validación temporal con un 15% de observaciones cronológicas finales mostró un R^2 de 0,9387, con una degradación controlada de -1,19%, confirmando estabilidad a corto-medio plazo.

6.6 Recomendación Final e Implementación

Se recomienda un sistema híbrido basado en RandomForest como modelo principal, complementado con banderas inteligentes para casos de alta incertidumbre y supervisión humana obligatoria en propiedades de lujo, inmuebles singulares o zonas con alta volatilidad.

Limitación crítica: ningún modelo alcanzó niveles de confianza que permitan automatización completa, dado que factores como el estado interior, la microubicación o las condiciones de la transacción siguen siendo determinantes y no siempre están capturados en los datos disponibles.

7. RESULTADOS Y DISCUSIÓN

7.1 Interpretación del Modelo Óptimo

El análisis posiciona a RandomForest como el modelo más adecuado para aplicaciones inmobiliarias operativas: alcanza R^2=0,9490 en test, con un RMSE de 0,1735 y estabilidad de CV (R^2=0,9462 ± 0,0081), lo que respalda su uso en contextos de decisión con incertidumbre moderada. Su Score de aplicabilidad de 60,1/100 (nivel "Medio") confirma utilidad práctica con supervisión. Frente a alternativas, ofrece mejor precisión en el segmento premium, robustez consistente y un ranking de variables interpretable para negocio, manteniendo un gap train-test razonable (0,0439). Arquitectónicamente, combina 100 árboles en votación agregada para capturar interacciones no lineales (p. ej., ubicación×tamaño) y umbrales de valoración diferenciados, adaptándose a la heterogeneidad socioeconómica del mercado.

Características fundamentales del modelo óptimo:

- Capacidad explicativa: 94,9% variabilidad precios mercado madrileño
- Error controlado: RMSE = 0,1735 (17,35% escala real)
- Estabilidad CV: R^2 = 0,9462 ± 0,0081
- Aplicabilidad práctica: Score 60,1/100 (nivel "Medio")

Ventajas competitivas frente a alternativas:

1. Precisión segmento premium: 58,7% predicciones dentro ±10% (duplica modelos lineales)
2. Robustez ante perturbaciones: ±0,0097 desviación estándar en validación
3. Interpretabilidad empresarial: Ranking características claro y accionable
4. Control sobreajuste: Gap Train-Test = 0,0439 (equilibrio óptimo)

Arquitectura y mecanismo:

- 100 árboles de decisión operando simultáneamente con votación democrática
- Captura interacciones complejas: Efecto multiplicativo ubicación-tamaño, umbrales valoración diferenciados
- Adaptación dinámica: Gestiona heterogeneidad mercado por segmentos socioeconómicos

7.2 Factores Determinantes del Precio

Los resultados confirman una jerarquía clara y accionable de determinantes, el factor dominante es la interacción ubicación premium × tamaño (premium_district_size),

que concentra la mayor parte del poder explicativo y cuantifica multiplicadores sustanciales entre zonas. A ello se suman variables de diseño interior y calidad del espacio que refinan la valoración, además de efectos microgeográficos y de accesibilidad. Esta lectura sustenta decisiones de inversión priorizando metros en ubicaciones premium y mejoras de calidad funcional.

Factor dominante - Interacción premium: premium_district_size (54,7% importancia total)

- Interpretación: Efecto multiplicativo ubicación exclusiva × superficie
- Impacto cuantificado: 100m² en Salamanca = 4,2x valoración vs. distritos periféricos
- Relevancia estratégica: Confirma fórmula "ubicación + tamaño" como núcleo valor
- Implicación inversores: Priorizar metros cuadrados en ubicaciones premium

Factores secundarios críticos:

Factor	Importancia	Interpretación Estratégica
bath_density	12,36%	Indicador calidad diseño interior y funcionalidad
loc_district_Tetuán	10,36%	Distrito emergente con dinámicas valoración propias
size_quality	9,21%	Relación superficie-distribución habitaciones
poly_m2_real bath_num	8,97%	Interacción no lineal m²-baños, umbrales específicos

Factores de refinamiento:

- Ascensor: 1,42% — impacto ligado a altura y accesibilidad
- Ubicaciones específicas (Villaverde, San Blas 1,0–1,3%) — micro efectos geográficos
- Configuración básica (habitaciones/baños 1,2–1,3%) — habitabilidad mínima

Segmentación determinantes por nivel socioeconómico:

Segmento	Ubicación	Superficie/Acabados	Características Únicas
Premium	80%	15%	5%
Medio	35%	45%	20%
Económico	15%	60%	25%

7.3 Casos de Estudio Específicos

Los casos ilustran patrones y límites del modelo. En lujo premium (Salamanca, Chamberí, Retiro), el error agregado es nulo y la precisión dentro de ±10% alcanza

el 65.5%, gracias a señales consistentes (antigüedad con valor histórico, alta interacción ubicación×tamaño y amenidades). En contraste, un subconjunto de segmento medio mostró fallos extremos por escasa señal y posibles anomalías de registro, lo que refuerza la necesidad de verificación y revisión humana. En distritos emergentes como Tetuán, el desempeño varía por segmento, indicando conveniencia de actualización frecuente y modelos adaptativos.

Caso 1: Propiedades de lujo en distritos premium

- Precio medio: €1,69M real vs €1,69M predicho (error agregado nulo)
- Error individual: 9,47% (superior al promedio global)
- Precisión alta: 65,5% predicciones ±10% (supera en más del doble la precisión habitual de este segmento)

Características distintivas:

- Construcción promedio: 1961 (valor histórico)
- premium_district_size: 210,89 (máxima interacción)
- Amenities_index: 2,59 (múltiples comodidades)

Insight: Máxima precisión en patrones claros y consistentes del segmento lujo.

Caso 2: Propiedades problemáticas segmento medio

- Muestra: 10 propiedades con errores >1.600%
- Rasgos: errores bidireccionales, sin outliers obvios, precios transformados cercanos a cero

Lección: Segmento medio requiere supervisión humana obligatoria.

Caso 3: Distrito Tetuán – Mercado emergente

- Alto: 15,2% error; Medio-Alto: 8,9%; Bajo: 22,4%

Implicación: Necesidad de reentrenos frecuentes y señales contextuales adicionales.

7.4 Limitaciones del Modelo

La principal limitación es la **disparidad de precisión entre segmentos**, excelente en premium y modesta en rango medio, lo que sugiere factores no observados (estado interior, microubicación, acabados) que explican parte del residual. Asimismo, hay **fragilidad ante perturbaciones** y **dependencia de pocas variables dominantes**, lo que incrementa la vulnerabilidad ante cambios de patrón. La ausencia de variables temporales induce degradación progresiva si no se reentrena. En suma, el modelo es valioso como **sistema híbrido** junto con criterio experto, especialmente en casos atípicos y decisiones críticas.

8. CONCLUSIONES

8.1 Síntesis de Resultados Principales

Se desarrolló un sistema de valoración para Madrid con **R^2=0,9490** (RandomForest), explicando ~94% de la variabilidad y capturando relaciones no lineales clave. La jerarquía de determinantes valida el **nexo tamaño×ubicación** como núcleo del valor, complementado por densidad de baños y calidad del espacio. La precisión es **segmento-dependiente** (muy alta en premium, menor en medio), lo que justifica algoritmos avanzados y segmentación explícita.

8.2 Implicaciones para el Mercado Inmobiliario Madrileño

El mercado muestra **estratificación intensa** y multiplicadores geográficos extraordinarios: la antigüedad puede ser premiada en zonas consolidadas, ciertas tipologías (áticos/chalets) mantienen primas y existen micro zonas con valoraciones muy superiores a la media. El **Distrito Salamanca** presenta niveles 4× respecto a otros, con dinámicas propias del segmento premium y heterogeneidad intra-distrito que exige granularidad de barrio/calle.

8.3 Aplicabilidad y Limitaciones del Modelo

Entre las fortalezas destacan la capacidad para **capturar no linealidades**, la **resiliencia a extremos** en lujo y la **cuantificación de drivers**. Las limitaciones incluyen **precisión operativa** moderada dentro de ±10% en términos sectoriales, **heterogeneidad por segmentos**, dependencia de **variables dominantes** y ausencia de **factores cualitativos** no observados. La aplicabilidad global se sitúa en **60,1/100 (nivel Medio)**, insuficiente para automatización plena sin supervisión.

8.4 Recomendaciones Estratégicas

Para **operadores del sector**, adoptar valoración contextualizada por distrito, priorizar inversiones de alto impacto (baños, ascensor, espacios exteriores) y actualizar baremos con multiplicadores reales. Para **inversores**, priorizar **ubicación frente a modernidad** en premium, optimizar el retorno vía **baños adicionales** y focalizar en **nichos** (áticos con terraza, histórico distintivo). En lo **técnico**, desplegar un **modelo híbrido** con validación continua, submodelos por segmento y plataforma integrada que combine predictivo, contexto y geoespacial.

8.5 Contribución Científica y Metodológica

El trabajo aporta un marco robusto que integra **ingeniería de características contextual**, **optimización multialgoritmo**, **evaluación multidimensional** con foco en negocio y **estratificación adaptativa** basada en patrones emergentes, transferible a otros mercados urbanos con adecuadas calibraciones.

ntic master

UNIVERSIDAD COMPLUTENSE MADRID

Curso académico

2024/2025