Revista de Estudios Empresariales. Segunda época.
Número: 2 (2023). Páginas: 27-39 DOI: https://doi.org/10.17561/ree.n2.2023.7823

VALORACIÓN AUTOMÁTICA DE INMUEBLES RESIDENCIALES MEDIANTE MODELOS DE MACHINE LEARNING

AUTOMATIC VALUATION OF RESIDENTIAL PROPERTIES USING MACHINE LEARNING MODELS

Francisco Guijarro Martínez (Universitat Politècnica de València)[1]

Resumen

La literatura reciente sobre valoración inmobiliaria ha aportado diversas evidencias en el ámbito internacional sobre el buen desempeño de los modelos de *machine learning* en la predicción del comportamiento de los precios, sobre todo si se comparan con los obtenidos por los denominados métodos tradicionales valoración, muy extendidos sobre todo en la práctica profesional. Con todo, se siguen remarcando algunas limitaciones como el diseño de caja negra y la dificultad en la interpretación de los resultados proporcionados por estas técnicas. Este trabajo tiene por objeto comparar los resultados y el desempeño de diferentes modelos de *machine learning* aplicados en el ámbito de la valoración inmobiliaria residencial. Para ello se ha recopilado una amplia base de datos con ofertas de inmuebles en la ciudad de Madrid, que permite dividir la muestra en los grupos de entrenamiento y test. La comparación entre los modelos se ha llevado a cabo a través de diferentes métricas, entre las que destaca el MAPE (*Mean Absolute Percentage Error*) por ser uno de los preferidos por las sociedades de tasación. Las métricas utilizadas confirman un buen rendimiento generalizado para el conjunto de modelos entrenados, con variaciones relativamente pequeñas tras el proceso de validación.

Palabras clave: valoración inmobiliaria, machine learning, gradient boosting machine.

Clasificación JEL: R30, C55, C58

Abstract

Recent literature on real estate valuation has provided evidence on the good performance of machine learning models in predicting price behavior, especially when compared to those obtained by traditional valuation methods. The latter are widely used in professional practice. However, some limitations are still highlighted, such as the black box design and the difficulty in interpreting the results provided by these techniques. This work aims to compare the results and performance of different machine learning models applied in the field of residential real estate valuation. For this purpose, a large database of property listings in the city of Madrid has been compiled, which allows the sample to be divided into training and test groups. The comparison between the models has been carried out through different metrics, among which the MAPE (Mean Absolute Percentage Error) stands out as one of

[1] Email: fraguima@upvnet.upv.es
ORCID: https://orcid.org/0000-0002-8803-5165

Fecha de envío: 13/03/2023. Fecha de aceptación: 31/05/2023

the favorites of valuation companies. The metrics we have used confirm a good generalized performance for the set of trained models, with relatively small variations after the validation process.

Keywords: real estate valuation, machine learning, gradient boosting machine.
JEL Codes: R30, C55, C58

1. INTRODUCCIÓN

Los activos inmobiliarios representan un porcentaje significativo y creciente en la riqueza a nivel mundial, según los datos proporcionados por los diferentes informes sobre el tamaño de mercado de los activos residenciales, publicados por la consultora internacional MSCI (MSCI, 2022). La importancia de este mercado obliga a realizar valoraciones de sus activos de forma periódica y rigurosa, lo que hace que la industria de la valoración sostenga a un número importante de valoradores. Según Kok *et al.* (2017), sólo en el ámbito de Estados Unidos se alcanzaban los 74.000 valoradores. Se trata de una industria intensiva en mano de obra, característica que ha venido marcada fundamentalmente por el tipo de metodologías empleadas en la valoración de inmuebles. La utilización de los métodos tradicionales de valoración por homogeneización de comparables requiere de un elevado grado de especialización y de una dedicación horaria importante por parte de los valoradores. Sin embargo, eso no asegura que las valoraciones alcancen un alto grado de precisión en algunos casos. Según se señala en el estudio de Kok *et al.* (2017), los valoradores cometen un error relativo del 12% en sus valoraciones respecto del valor final de transacción. Esta cifra no es exclusiva de Estados Unidos, lugar donde se hizo el estudio, sino que es equiparable a la que se produce en otros países como Italia (7,7%) o Japón (13,9%), según el estudio de MSCI (MSCI, 2022). Ha de tenerse en cuenta que no todos los mercados alcanzan el mismo grado de madurez y transparencia, y estas variables pueden ser claves a la hora de alcanzar resultados fiables desde un punto de vista valorativo.

Como ha sucedido en otros sectores, el rápido avance en la capacidad de procesamiento computacional y la cada vez mayor disponibilidad y detalle de bases de datos en el ámbito inmobiliario, ha permitido la aplicación de diferentes técnicas vinculadas al ámbito de la inteligencia artificial. Esto ha posibilitado obtener valoraciones con un elevado grado de precisión, pero sobre todo reducir el tiempo y coste vinculados a la valoración masiva de inmuebles. Esta reducción de costes y el acotamiento de los errores en los procesos de valoración ha favorecido la aparición de nuevos modelos, vinculados al *big data* y al empleo de métodos propios de la inteligencia artificial y el *machine learning*. Entre las ventajas del uso de estas metodologías está la limitación de la intervención humana, con la consiguiente reducción de costes, la velocidad de cómputo, la posibilidad de actualizar valores de tasación casi en tiempo real, la acotación en los errores de valoración, y la eliminación de sesgos propios del análisis humano. Resulta complicado medir y analizar de forma sistemática los errores producidos por un único valorador, o por un pequeño grupo de ellos. Sin embargo, cuando la base de datos empleada por los modelos de *machine learning* incluye cientos de miles de inmuebles, resulta viable emplear diferentes métricas de valoración que permiten medir la bondad de los diferentes métodos empleados (Steurer *et al.*, 2021). Sin embargo, en la literatura se siguen señalando algunas limitaciones que acompañan a estos métodos desde sus orígenes, como la naturaleza de caja negra, que en ocasiones dificulta saber cómo los modelos han llegado a estimar el precio de los inmuebles; o qué importancia han tenido cada una de las variables en la formación de los precios (Valier, 2020), si bien ésta es una limitación que en muchos modelos ya ha sido superada.

Este trabajo realiza una comparativa entre diferentes métodos de *machine learning* para una amplia base de datos de inmuebles localizados en Madrid, permitiendo acotar los errores producidos por diferentes metodologías en el ámbito de la normativa española. El artículo se estructura como sigue: en la siguiente sección se presenta una revisión actualizada de la literatura; la sección 3 presenta los datos empleados en el análisis, realizando una breve descripción de los mismos; en la sección 4 se describen de forma sucinta los métodos de valoración de *machine learning* empleados, las métricas con que comparar sus resultados, y los propios resultados obtenidos sobre la ciudad de Madrid. Finalmente, el trabajo se cierra con un resumen de las principales conclusiones alcanzadas.

2. REVISIÓN DE LA LITERATURA

En la literatura podemos encontrar diferentes investigaciones que han tratado de sintetizar una taxonomía de los métodos de valoración, que podemos clasificar en dos grandes grupos: los métodos de valoración tradicionales y los métodos de valoración avanzados (Abidoye *et al.*, 2019).

Por un lado, los métodos de valoración tradicionales se fundamentan en la comparación directa entre el inmueble a valorar, al que denominamos inmueble problema, y un conjunto de inmuebles con características similares al inmueble problema, y que han sido objeto de transacción reciente. Precisamente el hecho de que estos inmuebles sean similares al inmueble problema en cuanto a ubicación, superficie, antigüedad, etc., hace que reciban la denominación de "comparables" (Ministerio de Economía, 2003). La principal diferencia entre este conjunto y el inmueble problema es que del segundo no se conoce su valor, que se pretende estimar, mientras que de los primeros se conoce el precio de transacción, con el requisito de que la compraventa se haya producido recientemente y, por lo tanto, los precios no se encuentren alejados de las cotizaciones actuales que marca el mercado. La comparación que se produce entre las características de los inmuebles puede ser objetiva (superficie, dormitorios, altura, etc.), pero en ocasiones nos encontramos con que la naturaleza de las variables puede ser intrínsecamente subjetiva (calidad de la edificación, calidad de la ubicación, estado de conservación, etc.), lo que puede introducir importantes sesgos en su determinación por parte de los valoradores.

Por otro lado, los métodos de valoración avanzados se basan en la aplicación de modelos de optimización matemática, estadísticos, o de inteligencia artificial (Ahn *et al.*, 2012; Guijarro, 2021; Kontrimas y Verikas, 2011). En general ofrecen una mayor precisión en las estimaciones, pero como contrapartida son intensivos en tiempo de computación y en la cantidad de información requerida para poder aplicarlos. Afortunadamente, el tiempo de cómputo se ha reducido de manera muy considerable en los últimos años, favorecido por el desarrollo tecnológico. Dentro del grupo de métodos de valoración avanzados se encuentran el modelo de valoración hedónica, los métodos basados en el análisis de series temporales -como el modelo ARIMA-, los modelos espaciales y de *kriging*, las redes neuronales artificiales, los árboles de decisión en sus diferentes variantes, las máquinas de vector soporte, etc. (Pagourtzi *et al.*, 2003).

Existe un consenso dentro de la literatura en que los métodos de valoración avanzados proporcionan estimaciones más precisas, robustas, fiables y eficientes que los métodos de valoración tradicionales (Baldominos *et al.*, 2018; Ho *et al.*, 2021; Kok *et al.*, 2017; Selim, 2009), aunque en sus comienzos algunos autores señalaron que estos métodos podían conducir a errores significativos en los procesos de valoración masiva de inmuebles (Lenk *et al.*, 1997). En una comparación más detallada y reciente, Valier (2020) concluye que los modelos de inteligencia artificial ofrecen una mayor precisión que los modelos hedónicos, aunque muchos autores se mantengan reticentes al empleo de los primeros por considerarlos una caja negra,

en la que es difícil establecer la relación exacta que se produce entre el precio y sus variables explicativas. En el ámbito específico de los modelos heredados de la inteligencia artificial, Tchuente y Nyawa (2022) concluyen sobre una muestra de diferentes ciudades francesas y para un periodo de 5 años que las redes neuronales artificiales y el modelo *random forest* superan de forma significativa a otros métodos cuando no se tienen en cuenta las características de geocodificación de los inmuebles, mientras que los modelos *adaboost*, *gradient boosting* y el propio *random forest* funcionan mejor cuando las características de geocodificación son incluidas en la muestra de datos. Simlai (2021) analiza el mercado inmobiliario de California, concluyendo que los métodos de inteligencia artificial proporcionan una descripción exhaustiva de los determinantes del valor de las viviendas en el conjunto de secciones censales de California. En comparación con los modelos hedónicos, las regresiones Ridge, LASSO y Elastic Net proporcionan predicciones fuera de muestra significativamente mejores.

Los métodos avanzados emplean mediciones objetivas de las características relevantes de los inmuebles, evitando el posible sesgo introducido por los tasadores a la hora de comparar características con un alto componente subjetivo. Además, el alto grado de informatización permite manejar un número de registros y variables inabarcables por los métodos tradicionales, con tiempos de cómputo cada vez más reducidos, y con la posibilidad de ser actualizados de manera periódica y a un coste mínimo (Arribas *et al.*, 2016; Grover, 2016). Además de las variables que definen las características de la vivienda o del edificio donde se ubica, se ha podido agregar otra serie de variables vinculadas, por ejemplo, a la calidad medioambiental (Guijarro, 2019). Y todo gracias a las facilidades actuales en la recopilación y manejo de datos, que han permitido superar las limitaciones de las primeras aplicaciones de estos métodos.

La aparición de los modelos automatizados de valoración o AVM (acrónimo del término en inglés, *Automated Valuation Model*) ha permitido poder avanzar en el uso de métricas para medir la *performance* o desempeño en los modelos de valoración inmobiliaria. En esta línea de investigación, en la investigación de Steurer *et al.* (2021) se analizan un total de 48 métricas diferentes, definiendo un total de 7 de ellas como las más apropiadas para la evaluación del desempeño de los modelos AVM. Por su parte, Sing *et al.* (2022) recopilan más de 300.000 transacciones de vivienda pública y privada en Singapur para el período comprendido entre 1995 y 2017. En sus conclusiones destacan que el modelo boosting es el mejor modelo predictivo que produce las estimaciones más sólidas y precisas para los precios de la vivienda en comparación con los modelos de árbol de decisión y de análisis de regresión múltiple; y todo ello bajo el análisis de diferentes métricas de validación.

El lector interesado en una revisión exhaustiva de la literatura académica sobre modelos AVM aplicados en el ámbito inmobiliario puede referirse a Wang y Li (2019).

3. DATOS

El ámbito geográfico de análisis de esta investigación se circunscribe Madrid capital, ciudad para la que se ha recopilado una muestra de inmuebles ofertados a la venta en un popular portal inmobiliario, y durante el periodo que va desde abril de 2022 hasta septiembre de ese mismo año. En total, la muestra cuenta con un número inicial de 28.948 registros. Puesto que algunos inmuebles aparecen publicitados por más de una agencia inmobiliaria, se ha procedido a eliminar los registros duplicados, de forma que del proceso han quedado un total de 18.935 inmuebles.

También resulta habitual encontrar registros con errores; por ejemplo, en la introducción de la superficie, o en el precio del inmueble (colocar el precio de alquiler cuando el inmueble está a la venta; o al revés). Otra situación que puede afectar gravemente a la consecución de un

modelo fiable de valoración de inmuebles es encontrarnos con registros cuyas características pueden considerarse claramente atípicas, si las comparamos con el resto de inmuebles de la base de datos. Este tipo de observaciones son denominadas *outliers*. Para depurar la muestra, se ha empleado la distancia de Mahalanobis, tomando como variables discriminatorias el precio y la superficie de los inmuebles. Este proceso de depuración se ha llevado a cabo a nivel de sección censal. De esta forma, un inmueble se ha considerado atípico y ha sido eliminado de la muestra cuando su distancia de Mahalanobis se situaba más allá del percentil 97,5% del obtenido para dicho estadístico sobre el conjunto de la muestra, lo que se ha correspondido con una distancia de 16,67. Esto significa que en el proceso de filtrado se ha depurado un 2,5% de la muestra. También se han excluido los inmuebles con un precio de oferta por debajo de los 30.000 euros o por encima de los 5 millones de euros. Se ha limitado la altura de las viviendas a 15, de forma que ese umbral se ha asignado a aquellas que se situaban por encima de esta altura.

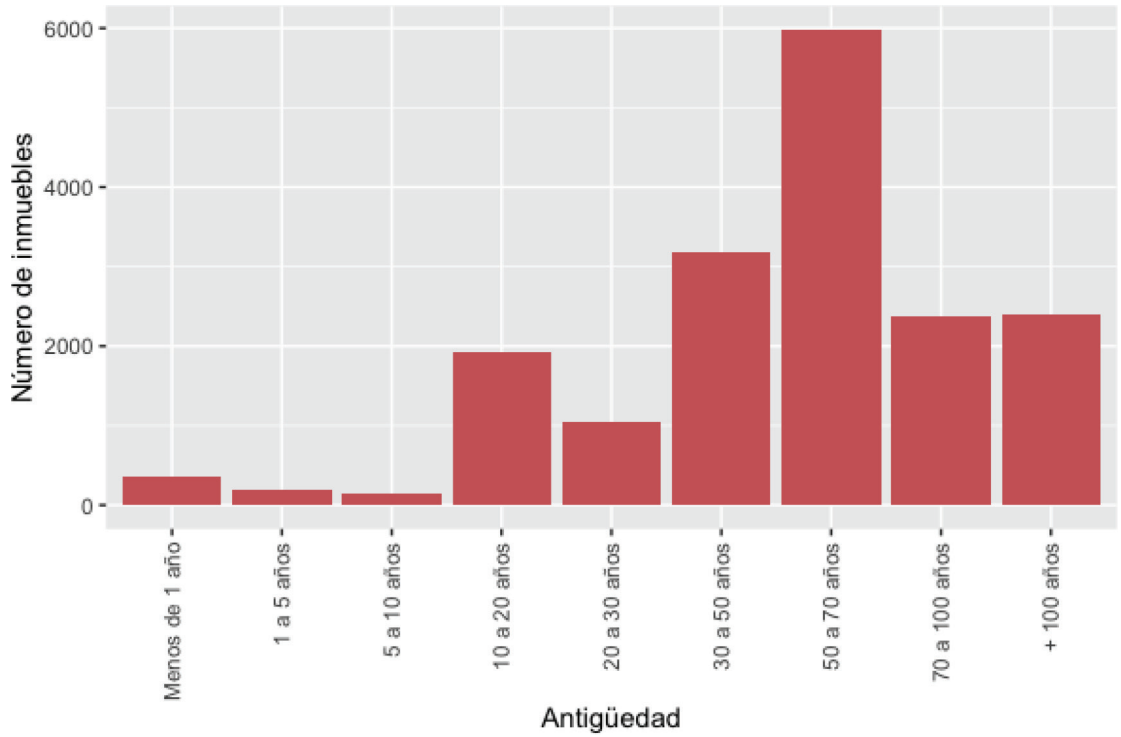
Con todo, la muestra ha quedado compuesta definitivamente por un total de 17.486 inmuebles: 8.811 correspondientes a los primeros 3 meses de análisis (trimestre T1) y 8.675 correspondientes a los 3 últimos (trimestre T2). La Tabla 1 muestra los principales estadísticos descriptivos de las variables numéricas consideradas en nuestro estudio. En la muestra se ha incluido un número importante de variables binarias, que se identifican fácilmente a través de sus valores mínimo (0) y máximo (1). Simplemente indican la presencia (1) o ausencia (0) de determinada característica. A través de la media podemos constatar, por ejemplo, que el 68% de los inmuebles tienen ascensor en su edificio, mientras que sólo un 14% cuenta con piscina.

TABLA 1. ESTADÍSTICOS DESCRIPTIVOS DE LAS VARIABLES ANALIZADAS

	Media	Desv. típica	Mediana	Mínimo	Máximo	Asimetría	Curtosis
Precio oferta	501.326,50	492.325,71	331.000,00	31.000,00	5.000.000,00	2,63	9,43
Superficie	106,07	61,96	88	1	529	1,85	4,51
Num. dormitorios	2,68	1,09	3	1	8	0,45	0,23
Num. baños	1,71	0,9	1	1	6	1,45	2,14
Ascensor	0,68	0,47	1	0	1	-0,77	-1,41
Num. planta	2,73	2,26	2	0	15	1,61	4,24
Terraza	0,45	0,5	0	0	1	0,2	-1,96
Aire acondicionado	0,52	0,5	1	0	1	-0,06	-2
Calefacción	0,75	0,43	1	0	1	-1,15	-0,67
Parking	0,18	0,39	0	0	1	1,64	0,68
Trastero	0,24	0,43	0	0	1	1,19	-0,58
Piscina	0,14	0,34	0	0	1	2,12	2,48
Zona ajardinada	0,05	0,22	0	0	1	4,03	14,27

Además de las variables numéricas reflejadas en la Tabla 1, aparece una variable categórica, la antigüedad. Esta variable viene clasificada en 9 niveles diferentes: desde las viviendas con menos de 1 año de antigüedad, a las que tienen más de 100 años. La Figura 1 permite constatar que el grupo más numeroso es el de viviendas entre 50 y 70 años, mientras que el menos representado en la muestra es el de viviendas entre 5 y 10 años.

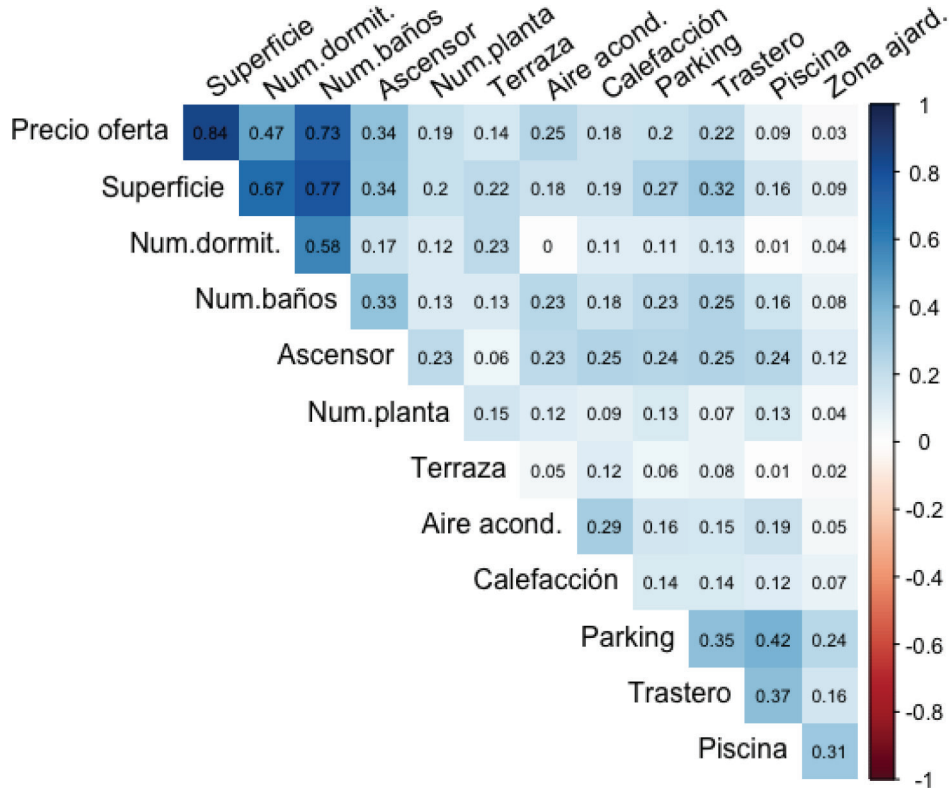
FIGURA 1. DIAGRAMA DE BARRAS DE LA VARIABLE ANTIGÜEDAD



Para completar este análisis univariante básico, se ha representado en la Figura 2 la matriz de correlaciones entre las variables numéricas. Esto permite estudiar cuáles están más correlacionadas con el precio, que es la variable que se intenta explicar. La figura muestra que la superficie tiene la mayor correlación con el precio (84%), algo previsible, seguida por el número de baños (73%) y el número de dormitorios (47%). A priori, variables como la zona ajardinada tienen una escasa vinculación con el precio (3%), algo que puede venir explicada inicialmente por la escasa representación en la muestra de las viviendas con esta característica. Debe señalarse que en esta matriz no aparece la variable antigüedad por ser categórica, pese a su previsible correlación significativa con el precio; ni tampoco ninguna de las variables ligadas a la ubicación, una de las características más referidas en la literatura (Pearson, 1991; Hamid, 2007; Kucklick y Müller, 2020).

En lo que respecta a la ubicación del inmueble, en el caso de los portales inmobiliarios españoles rara vez se informa sobre la localización exacta de los inmuebles. Los portales agrupan los inmuebles en zonas que ellos mismos definen y que, en muchos casos, se pueden asimilar a las secciones censales, pero son pocas las propiedades en las que se define la calle y el número exacto donde se ubican. Es precisamente la sección censal la variable que se ha tomado para informar sobre la localización aproximada del inmueble. Se han recopilado ofertas para 2.229 secciones censales de Madrid (sobre un total de 2.443 secciones registradas en el último censo), de forma que la sección censal con mayor número de inmuebles suma un total de 58 ofertas. Esto ha posibilitado incorporar la renta familiar disponible media de las secciones censales, como una variable explicativa de carácter socioeconómico de la zona. Para completar la información geográfica de los inmuebles, se ha creado una variable denominada "área de valor", que aglutina secciones censales vecinas en primer orden o superior, hasta completar un número mínimo de 50 inmuebles. El área de valor se puede considerar una sección censal de

FIGURA 2. MATRIZ DE CORRELACIONES ENTRE LAS PRINCIPALES VARIABLES NUMÉRICAS EMPLEADAS EN LA INVESTIGACIÓN



segundo nivel, creada a partir de una sección censal con escasa oferta de inmuebles a la que se ha añadido la oferta de otras secciones censales hasta completar un número suficiente de inmuebles, y que consideramos estadísticamente significativo para extraer conclusiones sobre la dinámica de precios. La configuración de la ubicación a partir de estas dos variables, permite poder incorporar al modelo el precio de los inmuebles en cada una de ellas, justamente en el trimestre T1 para no contaminar la muestra empleada en el entrenamiento (trimestre T2), así como el precio de los parking o aparcamientos de esas mismas zonas.

4. RESULTADOS DE LOS MODELOS DE MACHINE LEARNING

En esta sección se analiza el desempeño de diferentes modelos de *machine learning*. Para ello, se ofrecen las medidas de *performance* más habituales según se constata en la literatura, y que se relacionan a continuación:

- RMSE (*Root Mean Square Error*): desviación típica de los errores o residuos en la estimación. Si tomamos $\hat{y}_i$ como el precio estimado para el inmueble *i*-ésimo, y_i el precio observado para dicho inmueble, y tenemos *n* inmuebles en la muestra, el RMSE se construye a partir de la expresión (1):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(y_i - y_i\right)^2}{n}} \quad (1)$$

- MSE (*Mean Square Error*): varianza de los residuos de la estimación; se calcula como el cuadrado del RMSE.
- MAPE (*Mean Absolute Percentage Error*): el error porcentual absoluto medio, de definición similar al RMSE pero tomando errores en valor absoluto en lugar de su cuadrado y dividiendo por el precio observado del inmueble, según la expresión (2):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - y_i\right|}{y_i} \tag{2}$$

- RMSLE (*Root Mean Squared Log Error*): error logarítmico cuadrático medio calculado según la expresión (3):

$$RMSLE = \sqrt{\sum_{i=1}^{n}\frac{\left(log\left(y_i+1\right)-log\left(y_i+1\right)\right)^2}{n}} \tag{3}$$

Una vez configurada la muestra, el modelo de valoración se ha obtenido mediante su programación en el lenguaje estadístico R, a través de la función *automl* de la librería h20. Esta función permite el proceso de automatización en la selección de los algoritmos de *machine learning*, la generación de características, el ajuste de los hiperparámetros de cada uno de los algoritmos empleados, el modelado iterativo y la evaluación de modelos. Al tratarse de un modelo supervisado, sólo se han tenido en cuenta aquellos algoritmos que permiten trabajar con este tipo de modelos, de los que se da una breve explicación a continuación.

- Los modelos de aprendizaje profundo, o *deep learning* en inglés, se basan en una red neuronal artificial multicapa de alimentación directa que se entrena con descenso de gradiente estocástico mediante retropropagación (*backpropagation*). La red puede contener un gran número de capas ocultas compuestas por neuronas con funciones de activación tangencial, rectificadora y de tipo *maxout*. Además, la función *automl* incorpora funciones avanzadas como la velocidad de aprendizaje adaptativa, el recocido de velocidad, el entrenamiento por impulso, el abandono, la regularización L1 o L2, el punto de control y la búsqueda en cuadrícula.
- Los *distributed random forest*, también conocidos por sus siglas DRF, son una herramienta que se emplea tanto en problemas de clasificación como de regresión. Es precisamente este último el objeto de nuestra investigación. Cuando se proporciona un conjunto de datos, el algoritmo DRF genera un bosque de árboles de regresión, en lugar de un único árbol. Cada uno de ellos se comporta como un aprendiz débil construido sobre un subconjunto de filas y columnas. La incorporación de un mayor número de árboles reduce la varianza, de forma que en los problemas de regresión se toma la predicción media de todos sus árboles para realizar una predicción final.
- Los modelos lineales generalizados (GLM, por sus siglas en inglés *generalized linear models*) estiman modelos de regresión para resultados que siguen distribuciones exponenciales. Además de la distribución normal, la función *automl* considerada en nuestro trabajo también incluye otras conocidas distribuciones, como poisson, binomial y gamma.
- El modelo *Gradient Boosting Machine* (GBM) es un método que emplea un aprendizaje hacia delante. La heurística se basa en la obtención de buenos, que no óptimos, resultados predictivos mediante aproximaciones cada vez más refinadas, de forma que de forma iterativa se consigue ir alcanzando soluciones cada vez más próximas al óptimo. Se construyen árboles de regresión de forma secuencial sobre todas las características del conjunto de datos empleados.

- El modelo XGBoost es un algoritmo de aprendizaje supervisado que aplica un proceso denominado *boosting* para obtener modelos precisos. El término *boosting* hace referencia a la técnica de aprendizaje por conjuntos que consiste en construir varios modelos de forma secuencial en los que cada nuevo modelo intenta corregir las deficiencias del modelo anterior. En el proceso de refuerzo de árbol, cada modelo que se añade al conjunto es un nuevo árbol de decisión. XGBoost proporciona un *boosting* de árbol paralelo.
- El modelo XRT (*eXtremely Randomized Trees*). Se trata de un algoritmo de árboles extremadamente aleatorios, que emplean la aleatoriedad de un modelo de *Random Forest* pero tomando un subconjunto de las variables independientes en cada uno de los árboles entrenados. Los umbrales empleados para cada una de las variables se escogen de forma totalmente aleatorio, y no por optimización del criterio de impureza. Además, y a diferencia de *Random Forest*, las muestras de entrenamiento de cada uno de los árboles se escogen sin reemplazo; esto es, no siguen el tradicional modelo *bootstrap*.

La ventaja de emplear la función *automl* es que permite lanzar de forma multicore estos algoritmos en forma de grid, de forma que se cuenta con varias instancias de los mismos que se diferencian por los parámetros empleados en el entrenamiento. Precisamente para poder validar los resultados obtenidos, se ha dividido la muestra de forma aleatoria en un 75% para el entrenamiento y un 25% en cada uno de los trimestres para la validación de los modelos.

La Tabla 2 recoge el desempeño de los modelos entrenados para la muestra de viviendas en Madrid. Las diferentes métricas se han obtenido al aplicar dichos modelos al 25% de la muestra reservada para la validación, y los modelos se han ordenado de menor a mayor RMSE. Podemos comprobar cómo el modelo GBM ocupa las primeras posiciones, tanto en las versiones

TABLA 2. MÉTRICAS OBTENIDAS POR LOS MODELOS DE *MACHINE LEARNING*

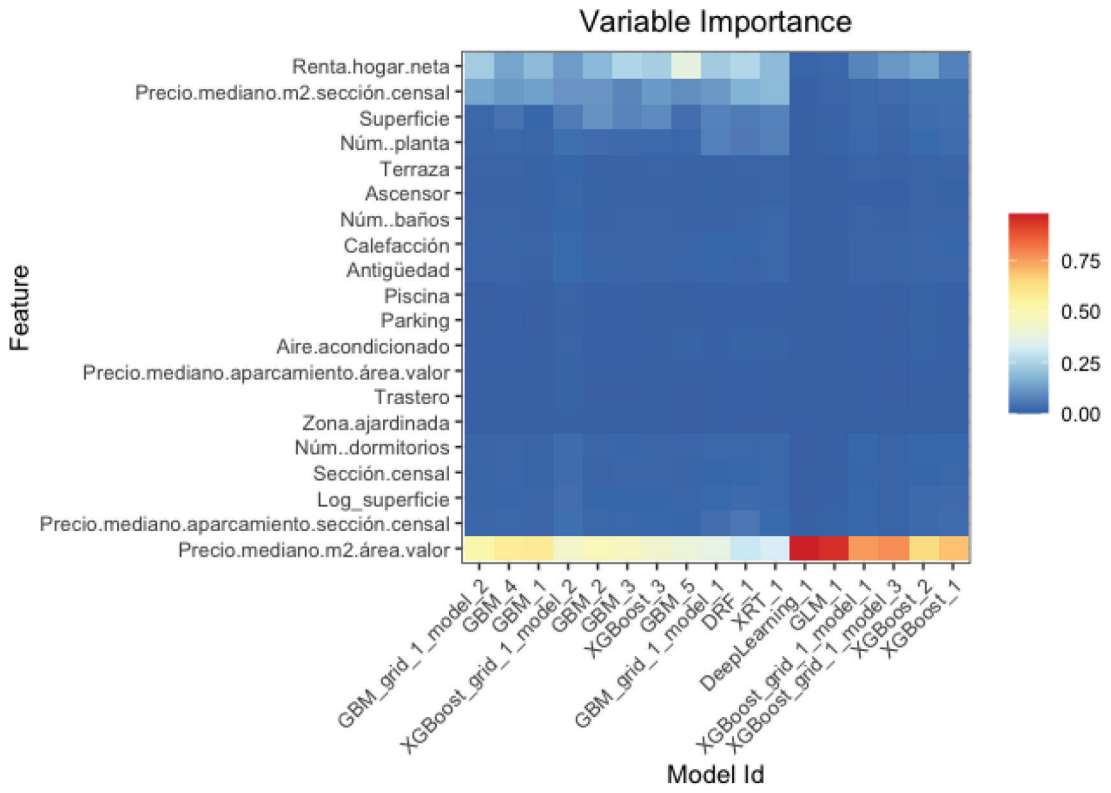
Modelo	RMSE	MSE	MAPE	RMSLE
GBM_grid_1_model_2	0,1710	0,0292	0,1217	0,0188
GBM_4	0,1714	0,0294	0,1227	0,0188
GBM_1	0,1719	0,0296	0,1225	0,0189
XGBoost_grid_1_model_2	0,1728	0,0298	0,1242	0,0190
GBM_2	0,1728	0,0298	0,1243	0,0189
GBM_3	0,1733	0,0300	0,1248	0,0190
XGBoost_3	0,1736	0,0302	0,1220	0,0191
GBM_5	0,1740	0,0303	0,1243	0,0191
GBM_grid_1_model_1	0,1743	0,0304	0,1221	0,0192
DRF_1	0,1751	0,0307	0,1270	0,0192
XRT_1	0,1753	0,0307	0,1257	0,0193
DeepLearning_1	0,1758	0,0309	0,1279	0,0193
GLM_1	0,1778	0,0316	0,1294	0,0195
XGBoost_grid_1_model_1	0,1792	0,0321	0,1291	0,0197
XGBoost_grid_1_model_3	0,1816	0,0330	0,1318	0,0199
XGBoost_2	0,1824	0,0333	0,1329	0,0201
XGBoost_1	0,1914	0,0366	0,1435	0,0210

individuales (GBM_4 y GBM_1) como en el formato de grid, donde se superponen diferentes modelos GBM. De las 5 métricas informadas en la tabla, queremos destacar los valores obtenidos en el MAPE. Podemos comprobar como la mayor parte de los modelos generan un MAPE por debajo del 13%, y que además la ordenación de los diferentes modelos es bastante similar entre las diferentes métricas utilizadas.

La figura 3 muestra la importancia relativa de las variables independientes utilizadas para explicar el precio de la vivienda en Madrid. La matriz se construye colocando por filas las variables independientes, y por columnas los diferentes modelos entrenados y ordenados según su precisión. Atendiendo a los tonos, que van desde el azul con una importancia mínima de 0 hasta el rojo con una importancia máxima de 1, podemos constatar cómo la variable más relevante es el precio mediano del metro cuadrado en el área de valor. Esto es, los precios observados en el trimestre T1 ejercen una gran influencia, como era previsible, en los precios de los inmuebles en el trimestre T2. Además, este hecho se reproduce para la práctica totalidad de los modelos. Las siguientes variables en importancia son la renta familiar disponible media de las secciones censales y el precio mediano del metro cuadrado en las secciones censales. En definitiva, el posicionamiento de estas 3 variables es un claro indicador de la relevancia que tiene la ubicación en la explicación de los precios inmobiliarios; tanto a nivel de sección censal como a nivel de la novedosa área de valor configurada en nuestra investigación.

Junto a estas variables, aparecen como especialmente relevantes la superficie del inmueble y el número de planta (altura) en la que se ubica dentro del edificio. Esto es, una variable que informa directamente de una característica intrínseca de la vivienda y otra que lo hace del edificio al que pertenece la vivienda.

FIGURA 3. IMPORTANCIA RELATIVA DE LAS VARIABLES PARA LOS DIFERENTES MODELOS DE *MACHINE LEARNING* ENTRENADOS



El resto de variables presenta una importancia muy moderada, próxima en muchos casos a 0. Como vimos en la matriz de correlaciones, ello no supone que no estén relacionadas con el precio. Pero su importancia se ve claramente minimizada cuando previamente se han incorporado las variables antes mencionadas, que son las que finalmente mayor porcentaje en la variabilidad de los precios explican.

5. CONCLUSIONES

La correcta valoración de inmuebles, de forma individual o en conjunto, resulta un proceso clave en muchas decisiones de tipo corporativo y financiero. Se trata de un sector con alta regulación, donde los agentes deben demostrar de forma periódica la precisión y calidad de sus procesos. Es por ello que las sociedades de tasación dedican un esfuerzo importante en el diseño e implementación de modelos de valoración focalizados en la acotación de los errores de valoración. En este sentido, la incorporación de la inteligencia artificial en estos procesos viene marcada por el uso intensivo de datos, la eficiencia en tiempo y recursos necesarios para su desarrollo, y una acotación significativa de los errores de predicción en comparación con métodos más tradicionales de valoración.

Este trabajo pretende servir de muestra sobre cómo algunas de las técnicas de machine learning ligadas a algoritmos de aprendizaje supervisado permiten implementar modelos de valoración sobre amplias bases de datos, limitando de forma muy significativa los errores de predicción. Varios de estos modelos se han aplicado sobre una muestra de viviendas en la ciudad de Madrid, destacando entre ellos el modelo Gradient Boosting Machine. Además, este modelo también permite identificar las variables más significativas desde un punto de vista de relevancia valorativa. Como destacan muchas de las referencias ligadas a este campo, la correcta modelización de algunas variables puede marcar una diferencia significativa en la capacidad de predicción de los modelos. Esto es, no es suficiente incorporar todas las variables que puedan explicar la variabilidad de los precios, sino que en ocasiones algunas de estas variables originales deben ser transformadas para facilitar el trabajo de los modelos de machine learning. En este caso, la variable área de valor vinculada a la ubicación de las viviendas se constituye como la más relevante, por encima de la sección censal o la renta neta del hogar ligada a la propia sección censal.

FINANCIACIÓN

Esta investigación no ha recibido financiación externa.

AGRADECIMIENTOS

Los autores quieren agradecer expresamente el apoyo recibido por parte de Euroval y el Instituto de Análisis Inmobiliario (INSTAI), que han facilitado los datos y el equipo informático necesarios para poder desarrollar los modelos de valoración inmobiliaria investigados en este trabajo.

REFERENCIAS

Abidoye, R. B., Junge, M., Lam, T. Y., Oyedokun, T. B., & Tipping, M. L. (2019). Property valuation methods in practice: evidence from Australia. *Property management*, 37(5), 701-718.

Ahn, J. J., Byun, H. W., Oh, K. J., & Kim, T. Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, 39(9), 8369-8379.

Arribas, I., García, F., Guijarro, F., Oliver, J., & Tamošiūnienė, R. (2016). Mass appraisal of residential real estate using multilevel modelling. *International Journal of Strategic Property Management*, 20(1), 77-87.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences*, 8(11), 2321.

Grover, R. (2016). Mass valuations. *Journal of Property Investment & Finance*, 34(2), 191-204.

Guijarro, F. (2019). Assessing the impact of road traffic externalities on residential price values: A case study in Madrid, Spain. *International Journal of Environmental Research and Public Health*, 16(24), 5149.

Guijarro, F. (2021). A mean-variance optimization approach for residential real estate valuation. *Real Estate Management and Valuation*, 29(3), 13-28.

Hamid, A. (2007). Combining geographic information systems and regression models to generate locational value residual surfaces in the assessment of residential property values. *Pacific Rim Property Research Journal*, *13*(1), 35-62.

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48-70.

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, 43(6), 202-211.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 11(1), 443-448.

Kucklick, J. P., & Müller, O. (2020). Location, location, location: Satellite image-based real-estate appraisal. arXiv preprint arXiv:2006.11406.

Lenk, M. M., Worzala, E. M., & Silva, A. (1997). High-tech valuation: should artificial neural networks bypass the human valuer?. *Journal of Property Valuation and Investment*, 15(1), 8-26.

Ministerio de Economía (2003). Orden ECO/805/2003, de 27 de marzo, sobre normas de valoración de bienes inmuebles y de determinados derechos para ciertas finalidades financieras. Madrid, España.

MSCI (2022). Real Estate Market Size 2021/22. Annual update on the size of the professionally managed global real estate investment market. https://www.msci.com/www/research-report/real-estate-market-size-2021-22/03296053034. Consultado el 02/01/2023.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21(4), 383-401.

Pearson, T. D. (1991). Location! Location! Location! What Is Location?. *The Appraisal Journal*, 59(1), 7.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, 36(2), 2843-2852.

Simlai, P.E. (2021). Predicting owner-occupied housing values using machine learning: An empirical investigation of California census tracts data. *Journal of Property Research*, 38(4), 305-336.

Sing, T. F., Yang, J. J., & Yu, S. M. (2022). Boosted tree ensembles for artificial intelligence based automated valuation models (AI-AVM). *The Journal of Real Estate Finance and Economics*, 65(4), 649-674.

Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2), 99-129.

Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, 308(1), 571-608.

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*, 38(3), 213-225.

Wang, D., & Li, V. J. (2019). Mass appraisal models of real estate in the 21st century: A systematic literature review. *Sustainability*, 11(24), 7006.