Finance, Markets and Valuation Vol. 7, Num. 2 (July-December 2021), 100–123

Finance, Markets and Valuation

Comparación de técnicas de valoración inmobiliaria basadas en la inteligencia artificial

Comparison of real estate appraisal methodologies based on artificial intelligence

José Legido Casanoves[1]

[1]Universitat Politècnica de València, Valencia, España. joleca@ade.upv.es

JEL: R3

Resumen

El objetivo del presente trabajo es presentar y comparar modelos que estimen el valor de un inmueble en función de sus características, lo que puede aportar valor, por ejemplo, a potenciales inversores de este mercado. De esta manera, podrán comparar el precio de oferta de los inmuebles con el obtenido mediante el modelo, antes de tomar decisiones de inversión o desinversión. En este trabajo se aplican métodos de análisis de datos, supervisados (regression tree, random forerst, nearest neighbour, SVM) y no supervisado (clustering) con la finalidad de estimar el valor de viviendas ubicadas en la ciudad de Madrid, España. Se comprueba que la característica de la vivienda que más influyen en el precio es la superficie, y en menor medida el número de habitaciones, si posee ascensor, también el número de baños y si el edificio dispone de aparcamiento. En cuanto al mejor método de estimación, según el MAPE, ha sido el random forest.

Palabras clave: Mercado inmobiliario; Valoración vivienda; Inteligencia artificial

Abstract

The aim of this paper is to present and compare models that estimate the value of a property according to its characteristics. This information can be required by, for example, potential investors in the real estate markett. In this way, they will be able to compare the offer price of the property with the value obtained by the model, before making investment or disinvestment decisions. In this paper we apply supervised (regression tree, random forerst, nearest neighbour, SVM) and unsupervised (clustering) data analysis methods in order to estimate the value of dwellings located in the city of Madrid, Spain. It is found that the housing characteristic that most influences the price is the surface area, and to a lesser extent the number of rooms, whether the building has a lift, the number of bathrooms and whether the building has a car park. The best method of estimation, according to the MAPE, was the random forest method.

Keywords: Real estate market; Housing valuation; Artificial intelligence

DOI: 10.46503/RIVF7714

Corresponding autor
José Legido Casanoves

Received: 30 Sep 2021
Revised: 26 Nov 2021
Accepted: 29 Nov 2021

Finance, Markets and Valuation
ISSN 2530-3163.

Cómo citar: Legido Casanoves, J. (2021). Comparación de técnicas de valoración inmobiliaria basadas en la inteligencia artificial. *Finance, Markets and Valuation*, 7(2), 100–123. DOI: https:\\doi.org\10.46503/RIVF7714

1. Introducción

El sector inmobiliario juega un papel determinante en la economía. Su actividad repercute en aspectos tales como las políticas públicas, el sistema impositivo, la estabilidad del sistema financiero, el empleo, el gasto de los hogares, urbanismo, cambio climático etc. El impacto socioeconómico de las crisis financieras ocasionadas por la actividad del sector inmobiliario son buena prueba de esta situación (Astudillo, 2021). No es de extrañar, por tanto, que el sector inmobiliario haya sido objeto de numerosos estudios que lo han analizado desde diversas perspectivas, como la formación de precios (Aznar *et al.,* 2010), la toma de decisiones de inversión (Cervelló *et al.,* 2011, Çamlibel *et al.*, 2021), rentabilidad de los fondos de inversión inmobiliaria (Feng *et al.*, 2021; Highfield *et al.*, 2021) análisis del riesgo de impago hipotecario (Archer & Smith, 2013; Kim &Sim, 2021), evolución de precios, alquileres y tipos de interés (Lin &Tsai, 2021) o su impacto en la calidad de vida de los ciudadanos (Streimikiene, 2014), por citar solo algunos ejemplos. Uno de los campos de investigación más destacados es la valoración inmobiliaria. Y, dentro de este ámbito, la valoración masiva de inmuebles, especialmente de viviendas.

La valoración masiva es la valoración sistemática de un grupo de inmuebles en un momento determinado del tiempo aplicando métodos estandarizados y comprobaciones estadísticas (Gloudemans, 1999). En el caso de las valoraciones inmobiliarias, son diversos los interesados en estas valoraciones (Wang & Li, 2019). Por ejemplo, las Corporaciones Locales las emplean para calcular la base imposible de ciertos impuestos. Las entidades financieras las emplean para valorar los activos que se entregan como garantía en los préstamos hipotecarios. Los fondos de inversión inmobiliarios emplean la valoración masiva para estimar el valor de su cartera de inmuebles. Y las empresas de tasación, por ejemplo, emplean estos métodos para controlar la calidad de las valoraciones realizadas por sus empleados.

La utilidad de la valoración masiva de inmuebles ha supuesto que numerosos académicos hayan propuesto y aplicado diferentes metodologías para valorar carteras de inmuebles (Guijarro, 2019). Estas metodologías se pueden clasificar en dos grupos: las que se basan en un enfoque econométrico (Arribas *et al.,* 2016) y las que se basan en el uso de inteligencia artificial. Dentro de este último grupo encontramos enfoques basados en los árboles de decisión (Fan et al., 2006), rough sets (D'Amato, 2007) redes neuronales (Tay & Ho, 1992; Selim, 2009), máquinas de vector soporte (Kontrimas & Verikas, 2011) o random forest (Antipov & Pokryshevskaya, 2012).

En el presente trabajo se comparan varios métodos de valoración masiva de inmuebles basados en la inteligencia artificial (clustering, árbol de regresión, random forest, vecino más próximo y máquinas de vector soporte) a partir de una muestra de 1000 viviendas ofertadas en la ciudad de Madrid. La finalidad es conocer qué modelo de predicción del precio de los inmuebles es el más idóneo en base al promedio del error absoluto o MAPE.

El estudio se estructura de la siguiente manera. Tras esta introducción, se describe cómo se ha obtenido la base de datos objeto de análisis. A continuación, se presentan y aplican diferentes modelos de predicción basados en el uso de inteligencia artificial. En primer lugar, se presentan varios modelos no supervisados y a continuación, modelos supervisados. Finalmente, se presentan las conclusiones del trabajo así como posibles líneas de mejora a considerar en análisis futuros.

2. Base de datos

En este apartado se analiza la base de datos, que muestra las características de 1000 viviendas ofertadas en la ciudad de Madrid. Para realizar este análisis, en primer lugar se describirán las variables que contiene la base de datos, con el objetivo de conocer con mayor profundidad a qué se refiere cada variable.

La base de datos consta de las siguientes variables:

-rawPrice: Precio de la vivienda.

-surface: superficie de la vivienda, medida en metros cuadrados.

-rooms: habitaciones

-bathrooms: baños.

-elevator: ascensor (1 si posee ascensor, 0 en caso contrario)

-parking: zona de aparcamiento (1 si posee aparcamiento, 0 en caso contrario)

-terrace: terraza (1 si posee terraza la vivienda, 0 en caso contrario)

-buildingSubType: subtipo de vivienda.

La base de datos está formada por 1000 observaciones y 8 variables, de las cuales 7 son de tipo numérico y una de tipo categórica, que se trata de buildingSubType.

En las variables no se ha detectado ningún valor anómalo o inconsistente, en cambio sí que se han detectado valores faltantes, esto puede perjudicar a la hora de aplicar los métodos de análisis, ya que si una variable tiene un elevado porcentaje de valores faltantes el programa los sustituirá automáticamente por la media de la variable, y por lo tanto esto puede no ser adecuado. Para solucionar este problema se ha decidido imputar los valores por regresión, que consiste en predecir los valores faltantes de una variable a partir de sus relaciones con otras variables de la base de datos.

3. Método no supervisado

A continuación, se implementa un método de análisis no supervisado, conocido como clustering. El objetivo de utilizar este método es conocer cuál es la tendencia de agrupamiento de las viviendas de la base de datos y que características definen cada grupo. Esto permitirá obtener información de las viviendas que se están analizando y por lo tanto un mayor conocimiento del mercado inmobiliario.

Los métodos no supervisados, se caracterizan por qué no se conoce a priori el objetivo buscado, para este método se ha decidido aplicar la técnica clustering, a través de este modelo se reducirá la dimensión de los datos. Esta técnica clasifica un conjunto heterogéneo de elementos en grupos, en función de las similitudes o diferencias entre ellos, posteriormente se validará el modelo y se procederá a la interpretación.

En primer lugar, para empezar con el clustering, se necesitarán las variables numéricas y la variable categórica se utilizará posteriormente para facilitar la interpretación. Primero, habrá que observar en qué medida están representada las variables, con el objetivo de saber si es necesario escalar y centrar, en este caso, sí que haría falta escalar y centrar los datos, ya que todas las variables no están medidas en la misma escala y también hace falta que la media sea 0.

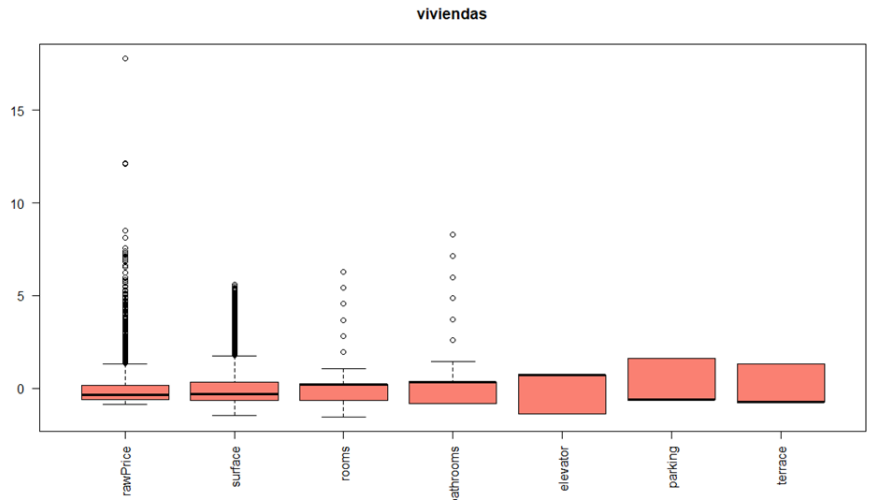


Figura 1. Gráfico boxplot de las variables escaladas y centradas de la base de datos

Como se observa en la Figura 1, las variables numéricas, ya han sido escaladas y centradas. También, se observan valores anómalos en algunas variables, pero no se han eliminado debido a que son valores que, a pesar de no ser habituales, se pueden dar en ciertas viviendas.

A continuación, se procederá a generar el mapa de colores, en el cual se ha utilizado la distancia euclídea (Figura 2).

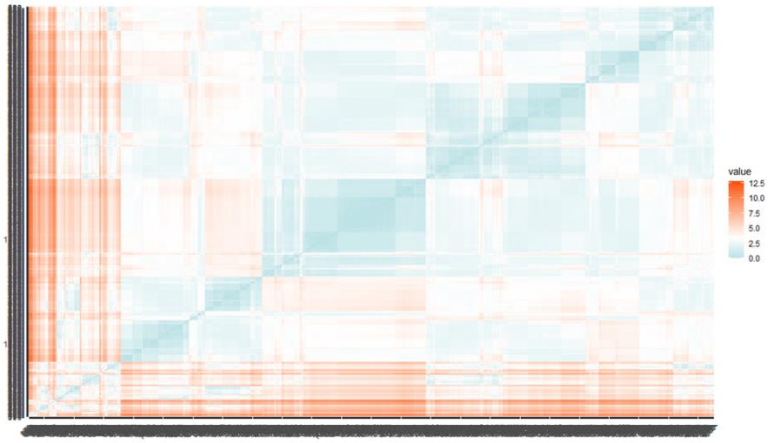


Figura 2. Mapa de color

Como se puede observar en el mapa, las viviendas se agrupan mayoritariamente en dos grupos y también hay varios grupos de menor tamaño. Este mapa da una idea del posible número de clusters.

El estadístico de Hopkins (Tabla 1) selecciona aleatoriamente observaciones, calculando la distancia entre el elemento y su elemento más cercano, como muestran los datos anteriores existe una elevada tendencia de agrupamiento ya que sus valores están entre 0.85 y 0.97.

Tabla 1. Estadístico de Hopkins

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.8587	0.9002	0.9231	0.9174	0.9360	0.9715

3.1. Modelos jerárquicos

En primer lugar, se aplicará el modelo jerárquico, el cual es útil cuando los individuos tienen una clara estructura jerárquica, además, puede ser utilizado como paso previo para la determinación del número de grupos a formar, con un método no jerárquico.

Para ello se aplicarán dos métodos: método ward y método de la media.

El método ward forma clusters maximizando la homogeneidad intra-clusters.

A través de un dendograma, se representará gráficamente en forma de árbol el proceso de agrupamiento con los 6 clusters que se han apreciado en el mapa de color, y se cortará el dendrograma a un determinado nivel, para obtener esa clasificación de los elementos en cada grupo.

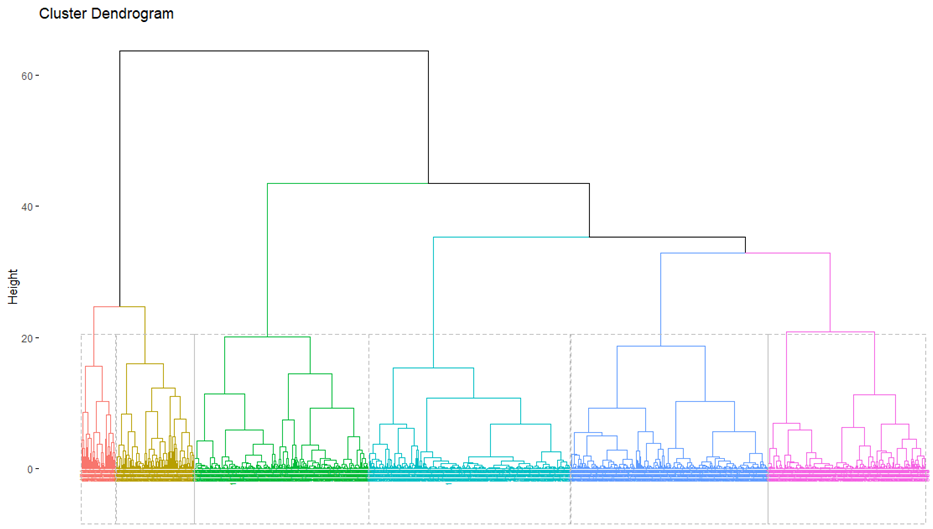


Figura 3. Dendograma con el método ward

En el dendograma de la Figura 3 se observa que el cluster 2 (239) es el que más viviendas agrupa, seguido del cluster 3 (234), el cluster 1 (207), el cluster 5 (187), el cluster 4 (93) y por último el cluster 6 (41).

El método de la media define la distancia entre clusters como la media de las distancias entre todas las parejas de elementos que la componen.

Como se observa en el dendograma de la Figura 4, el grupo 1 (954) agrupa a la gran mayoría de las viviendas, mientras que los otros grupos poseen un menor número de viviendas, en concreto el grupo 2 (31), el grupo 3 (9), el grupo 4 (5), el grupo 5 (1) y el grupo 6 (1).

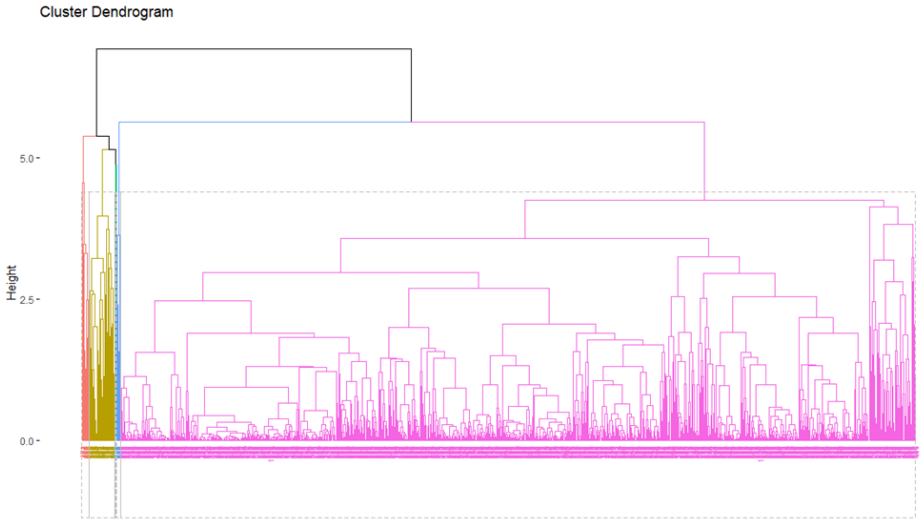


Figura 4. Dendograma con el método de la media

Tras haber visto los dos métodos del modelo jerárquico, el método Ward parece tener más sentido y ajustarse mejor a lo observado en el mapa de color, por lo que se utilizara el método ward.

Para obtener el número de clusters óptimo para este algoritmo, se pueden aplicar distintos criterios como se observa a continuación, donde se aplicará el coeficiente de silhoutte y el método de la suma de cuadrados intracluster.

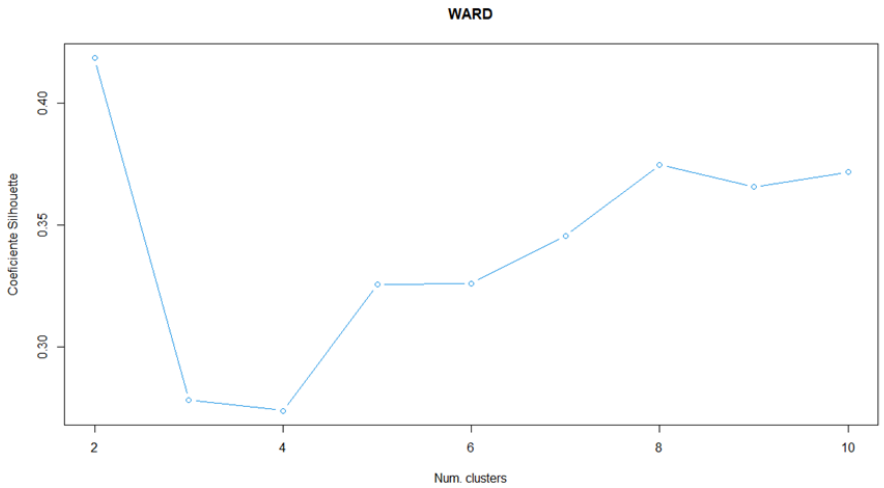


Figura 5. Número óptimo de clusters a través del coeficiente de silhouette

Como se observa en la Figura 5, el mayor coeficiente de Silhouette indica que el número óptimo de clusters es de 2, en cambio en el método de la suma de cuadrados intraclusters (Figura 6) no se ve claramente el codo bien diferenciado

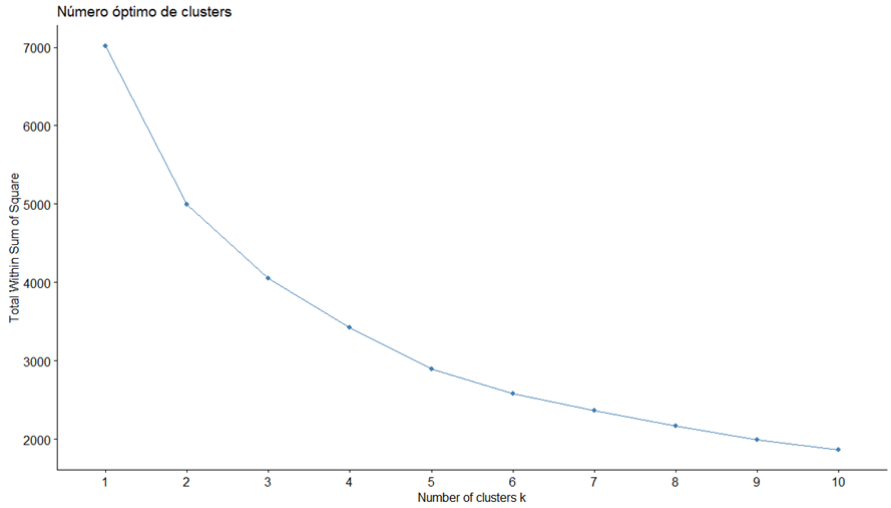


Figura 6. Número óptimo de clusters a través de la suma de cuadrados intracluster

A continuación, se proyecta el PCA aplicándolo a los datos y los cluster obtenidos a través del método Ward que es el que más se ajustaba a lo observado en el mapa de colores, se observa que la primera dimensión explica el 47,5% de toda la variabilidad del modelo y la segunda dimensión el 16,8%.

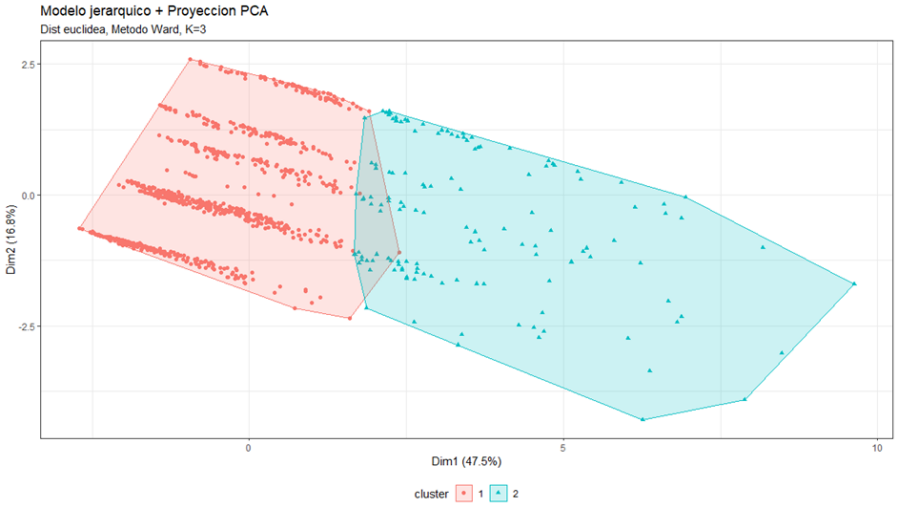


Figura 7. PCA scores

Como se observa en la Figura 7, tras utilizar 2 cluster tal y como indicaba el coefiente de silhoutte, se observa en el gráfico de scores del PCA, que ambos clusters se solapan entre si. Esto puede ser debido a que se diferencian en algunas variables que no están correlacionadas con las dos primeras componentes principales

3.2. Modelos de partición

El objetivo de los modelos de partición es obtener una partición de los individuos en clusters, de tal forma que todos los individuos pertenezcan a uno de los posibles clusters y que estos clusters sean diferentes. A continuación se emplean los métodos de k-medias y k-medoides.

En el método k-medias, en primer lugar, es necesario determinar el número de clusters necesarios con los que se trabajara. Para ello, se seleccionan como centroides los clusters iniciales y calculando la distancia euclídea de cada elemento a estos, se asignan al cluster que tenga su centroide más próximo

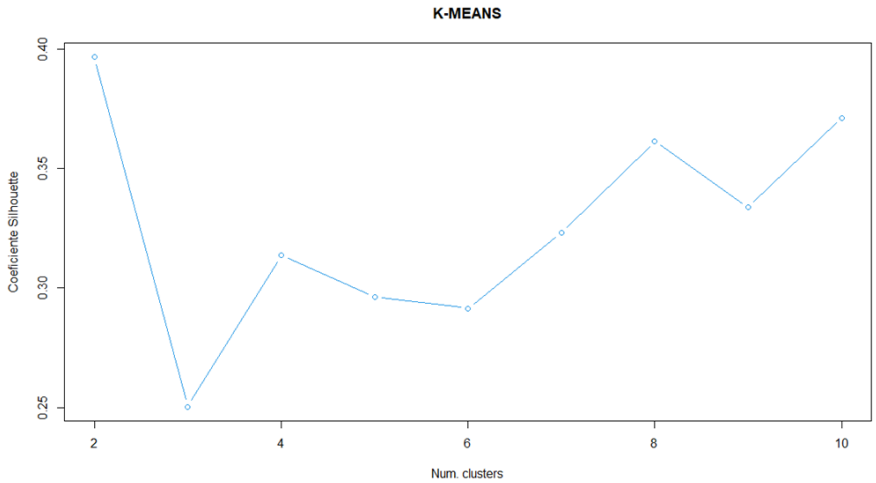


Figura 8. Número óptimo de clusters a través del coeficiente de silhouette

En la Figura 8 se ha aplicado el coeficiente de silhouette para establecer el numero óptimo de clusters. Este coeficiente indica la calidad de agrupamiento de los individuos, por lo tanto, contra más cerca este de 1, esto indicara que los individuos se han asignado al cluster correcto. Como se aprecia en el gráfico, el coeficiente de silhoutte indica que el número de clusters optimo son 2, con un coeficiente de 0.40.

En la Figura 9 se ha aplicado el método de la suma de cuadrados intracluster, para determinar el número óptimo de clusters, para ello se ha utilizado el método del codo, a pesar de esto no se puede ver un codo bien diferenciado.

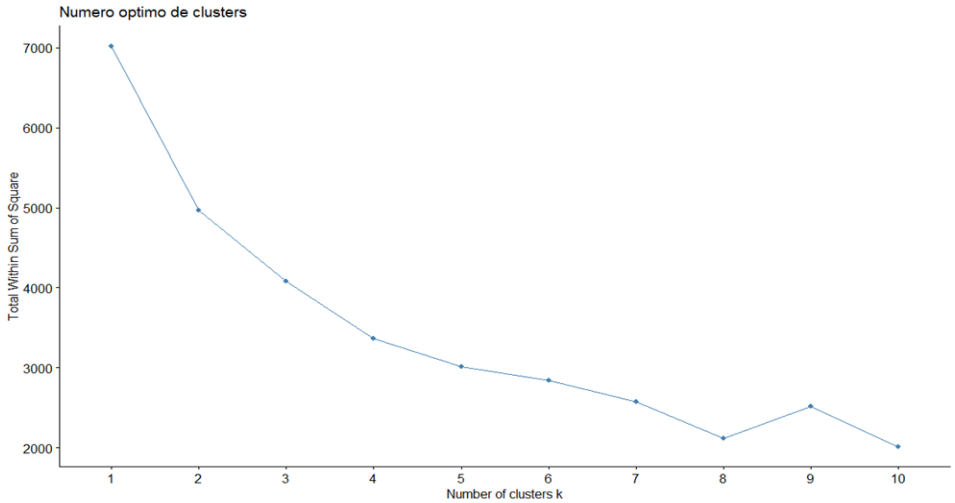


Figura 9. Número óptimo de clusters a través de la suma de cuadrados intracluster

En el gráfico de scores del PCA (Figura 10), tras utilizar 2 clusters tal y como indicaba el coeficiente de silhoutte, se observa en el gráfico que ambos clusters no se solapan entre si, a pesar de estar muy cerca el uno del otro, por lo que ambos clusters están completamente diferenciados.

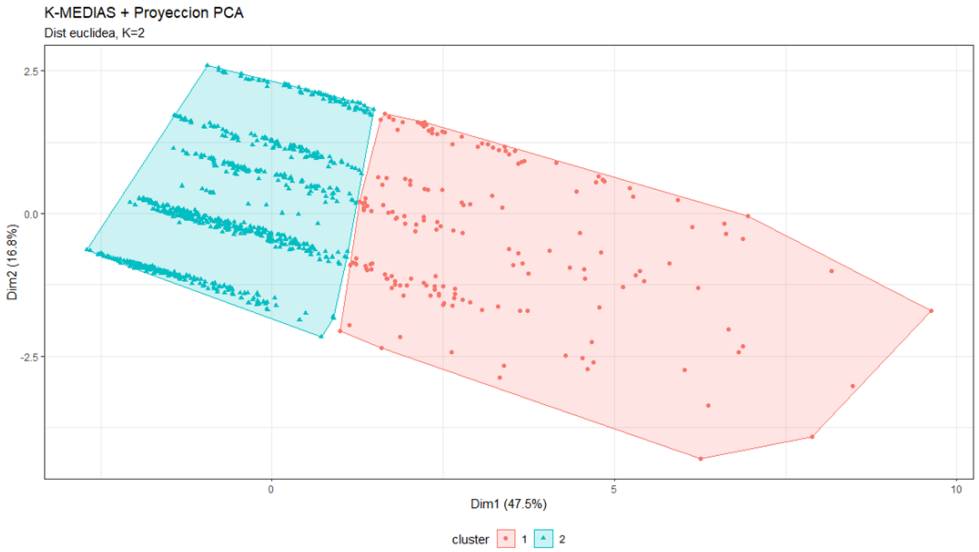


Figura 10. PCA scores

A continuación, se utilizará el método de los k-medoides que, en teoría, debería ser más robusto frente a los valores atípicos. En este algoritmo también es necesario determinar a priori el número de clusters.

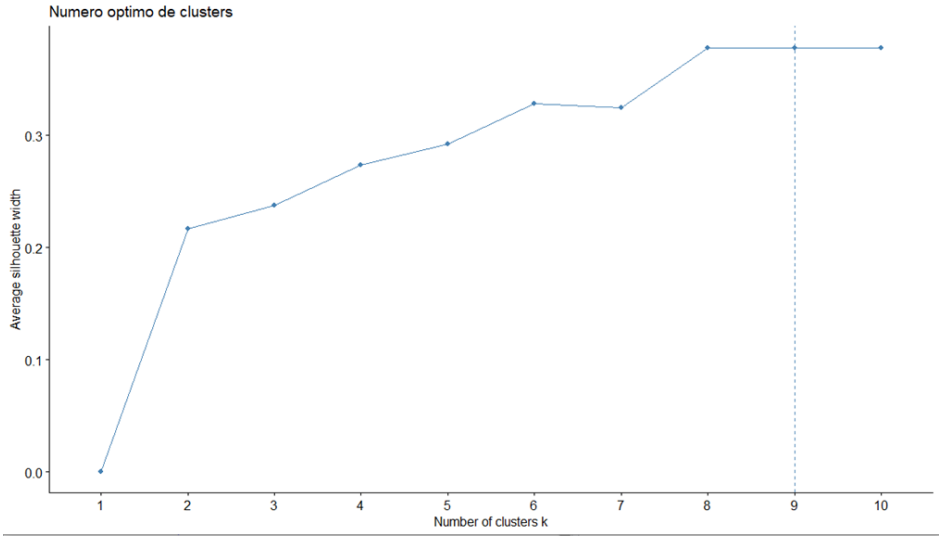


Figura 11. Número óptimo de clusters a través del coeficiente de silhouette

Como se observa en la Figura 11, el coeficiente de silhoutte indica que el número de cluster optimo son 9, en cambio, como se puede apreciar en la Figura 12, en el cual se ha utilizado el método de la suma de cuadrados intracluster, parece indicar que el número de clusters óptimo son 4. Por lo tanto, se optará a utilizar 4 clusters, ya que 9 clusters son demasiados y seguramente se solapen.

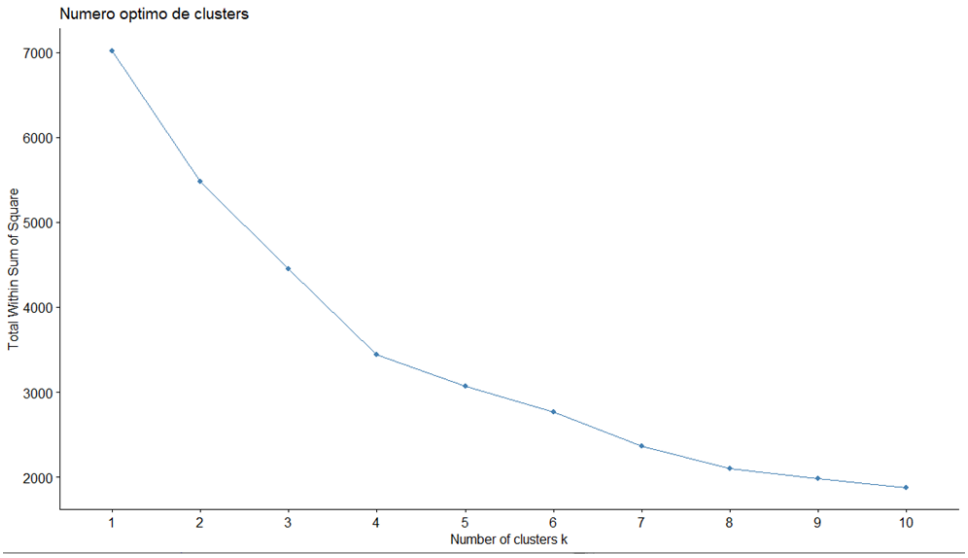


Figura 12. Número óptimo de clusters a través de la suma de cuadrados intracluster

Como se observa en el gráfico scores PCA (Figura 13), existen clusters que se solapan entre sí, como el cluster 4, que esta solapado por el cluster 1 y 2, como se ha explicado anteriormente, esto puede deberse a que se diferencian en algunas variables que no están correlacionadas con las dos primeras componentes principales. Por lo tanto, a partir del PCA se podría ver la posibilidad de reducir el número de clusters con el objetivo de que se reagrupan los clusters que se solapan.

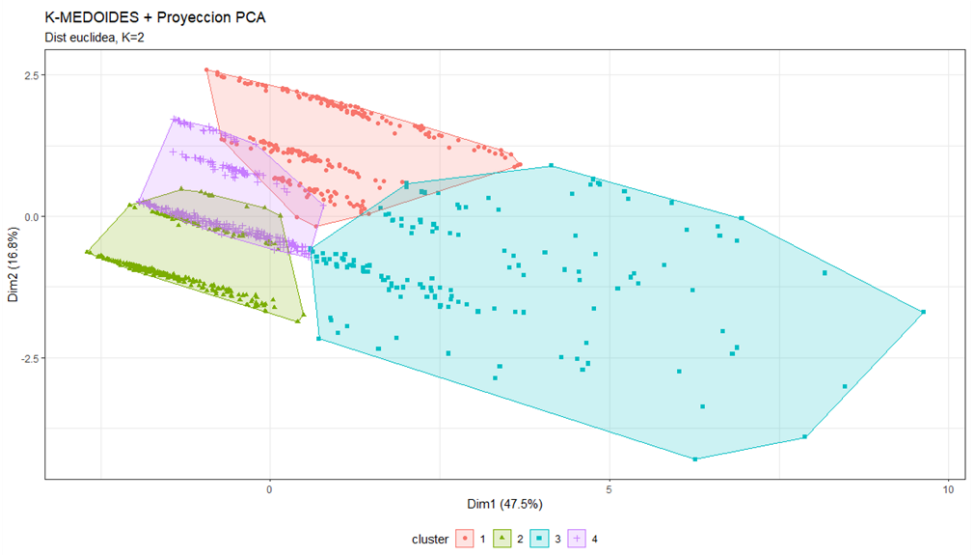


Figura 13. PCA scores

3.3. Selección y validación del modelo

Tras haber generado los diferentes modelos jerárquicos y de partición, y observado los resultados, es difícil decantarse por un modelo, pero en principio el criterio k-medias y el ward son los más fiables por los resultados, por lo tanto, se utilizará el coeficiente de silhoutte para la selección y validación del modelo.

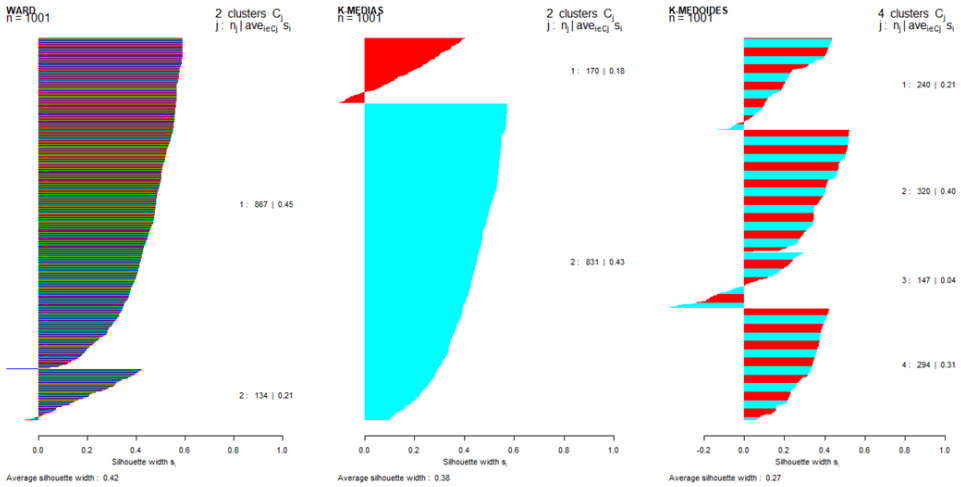


Figura 14. Comparación de los modelos de clustering

Tras ver los resultados de los diferentes modelos (Figura 14), se observa que el mejor modelo es el método ward, ya que es el modelo que presenta un mayor coeficiente de silhoutte (0.42) frente al algoritmo k-medias (0.38) y el método k-medoides (0.27), además el método ward también es el que tiene un menor número de viviendas mal clasificadas.

3.4. Interpretación de los resultados

En primer lugar, se va a utilizar el gráfico PCA, para observar cuales de las variables utilizadas en el análisis han tenido mayor contribución en la determinación de los clusters obtenidos a través del método Ward.

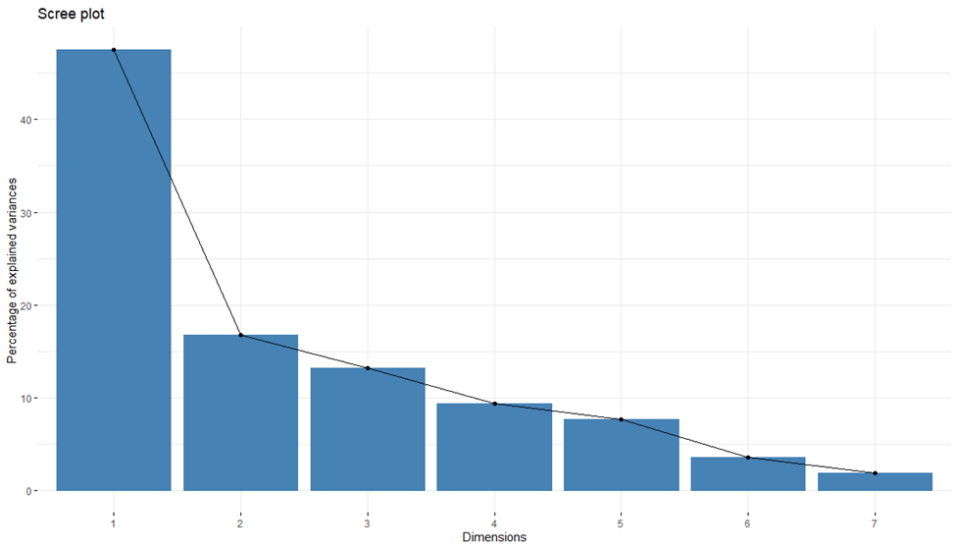


Figura 15. Dimensiones del modelo

Como se observa en el gráfico PCA (Figura 13), existen 7 dimensiones que explican el 100% de la variabilidad del modelo. La primera dimensión es la que más variabilidad explica con un 47,5%, seguido de la segunda dimensión que explica el 16,8% (Tabla 2).

Tabla 2. Aportación de cada dimensión al modelo

	Eigenvalue	Variance percent	Cumultaive variance pecent
Dimensión 1	3.3322	47.527	47.528
Dimensión 2	1.1786	16.811	64.338
Dimensión 3	0.9243	13.184	77.523
Dimensión 4	0.6569	9.370	86.893
Dimensión 5	0.5369	7.658	94.552
Dimensión 6	0.2509	3.578	98.130
Dimensión 7	0.1310	1.869	100.00

A continuación, se mostrará más detalladamente la contribución de las variables a las dos primeras dimensiones, que entre ambas explican el 64,3% de la variabilidad total del modelo.

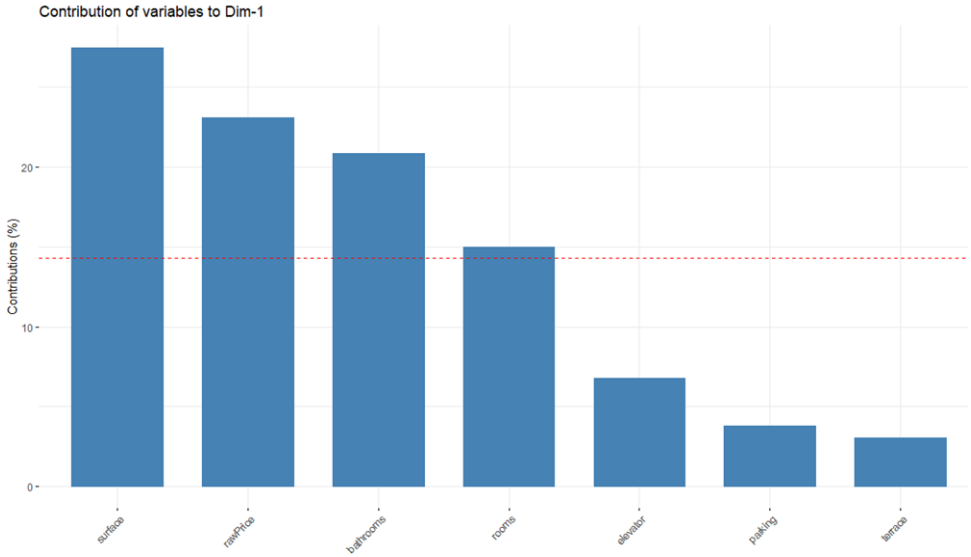


Figura 16. Contribución de las variables a la primera dimensión

La Figura 16 muestra la primera dimensión, la cual explica el 47,5% del total de la variabilidad del modelo, entre las variables, las que más influyen son la superficie, los baños, el precio y las habitaciones, parece ser que esta dimensión hace referencia al tamaño de las viviendas.

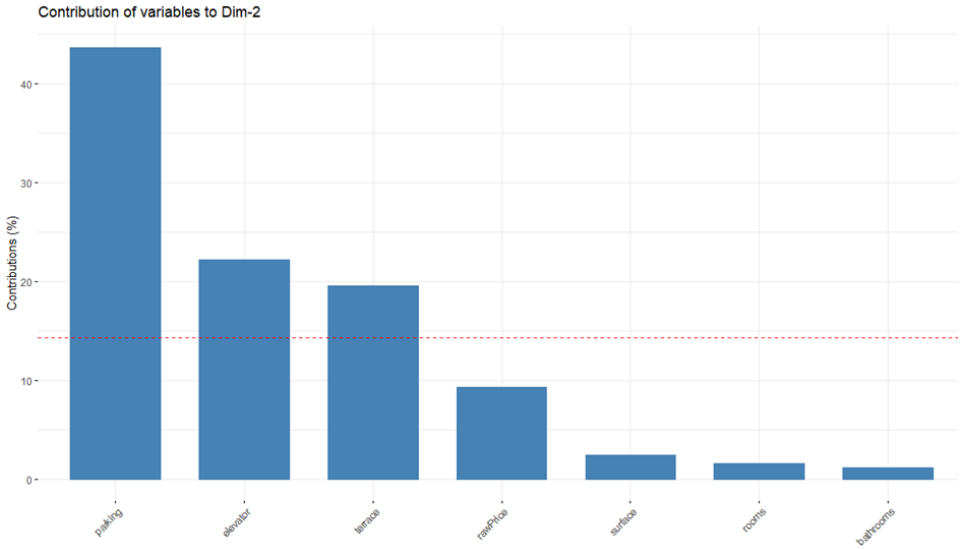


Figura 17. Contribución de las variables a la segunda dimensión

La Figura 16 muestra la segunda dimensión, la cual explica el 16,8% de la variabilidad total del modelo, entre las variables que más influyen se encuentran el parking, la terraza y el ascensor, esta dimensión hace referencia a los extra que trae consigo la vivienda.

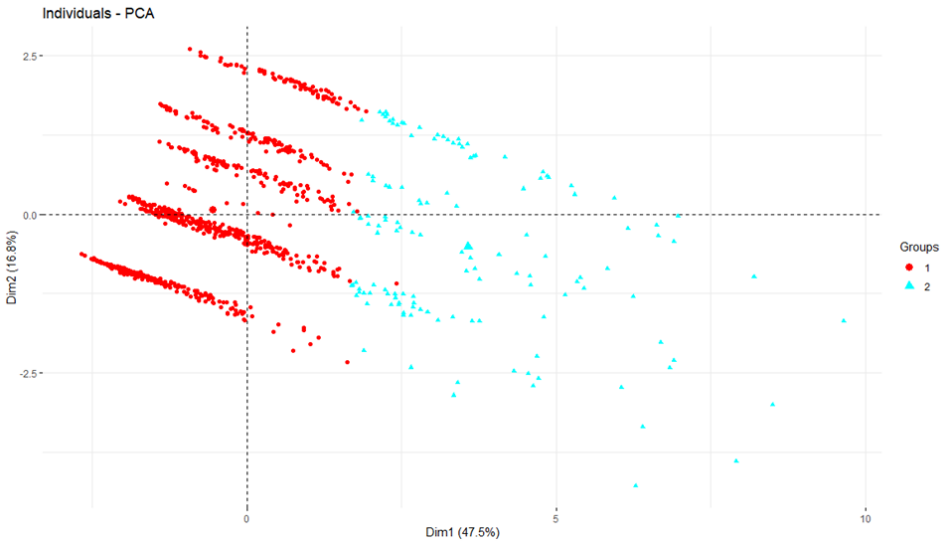


Figura 18. Distribución de los clusters

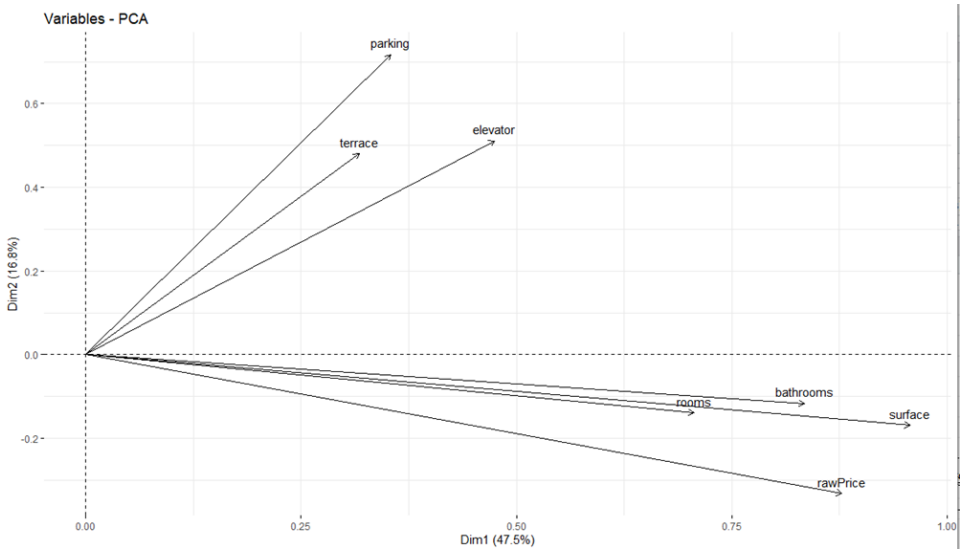


Figura 19. Distribución de las variables

Tras utilizar el método Ward, que es el que ofrecía un mayor coeficiente de silhoutte, las viviendas se han agrupado en 2 cluster, el primer cluster está formado por 867 viviendas y el segundo por 134 (Figura 18).

Al observar las Figuras 18 y 19 se puede apreciar que las viviendas del grupo 2 tienen mayor superficie, más habitaciones, más baños y por lo tanto tienen un mayor precio.

A simple vista, se puede apreciar que las viviendas del grupo 1 pertenecen a las personas con un menor poder adquisitivo en comparación a las personas del grupo 2, debido a que las viviendas que poseen tienen una menor superficie y precio.

En cuanto al aparcamiento, la terraza y el ascensor no son una característica distintiva de un grupo particular, sino que ambos grupos poseen viviendas con algunas o todas estas características.

A continuación, se va a realizar un gráfico descriptivo del perfil de ambos cluster con el objetivo de ver las diferencias entre ellos. Para ello, se calculará la media de cada variable para cada cluster.

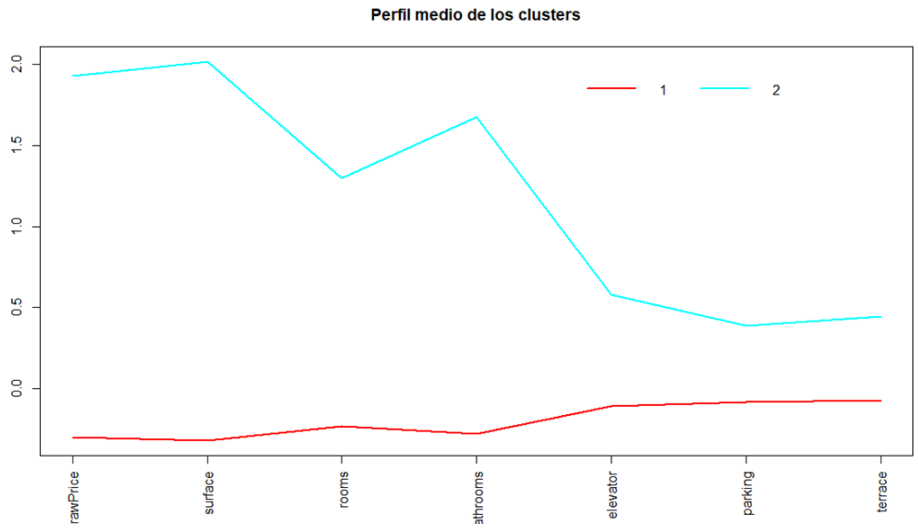


Figura 20. Perfil medio de los clusters

Entre el perfil medio de cada cluster (Figura 20) vemos una notable diferencia entre ambos. En el segundo cluster las viviendas tienen un mayor precio, una mayor superficie, más habitaciones y baños, y se diferencian mucho en estos aspectos en comparación a las viviendas del primer grupo. En cambio, no existe tanta diferencia entre las viviendas de ambos grupos cuando se trata de si la vivienda posee ascensor, terraza o parking, ya que es una característica que poseen determinadas viviendas de ambos grupos, aunque más las viviendas del grupo 2.

Uno de los mayores problemas en el mercado inmobiliario, es que no se conoce con exactitud el valor de los activos, por lo que se utilizan referencias próximas, aunque solo tienen carácter orientativo. Por lo tanto, en la valoración de un activo, existe la posibilidad de error, ya que al fijar el precio adecuado y no conocer el precio exacto, se puede cometer un error, ya que cuando se fija el precio se tiene en cuenta referencias próximos, por lo tanto, no son del todo acertadas ya que se realizan pocas operaciones y las características de cada inmueble varían.

A continuación, se realizará un método de análisis supervisado de regresión, con el objetivo de desarrollar modelos que estimen el valor de un inmueble en función de sus características, lo que podrá aportar valor a potenciales inversores de este mercado. Podrán comparar el precio de oferta de los inmuebles con el obtenido mediante el modelo, antes de tomar decisiones de inversión o desinversión.

Además, este modelo también podrá ser útil, para las personas que hayan decidido poner en venta su inmueble, y quieran tener un precio estimado de la vivienda en base a sus características.

4. Método supervisado

El método supervisado se caracteriza por que se conoce a priori el objetivo buscado, es decir, para los individuos conoces alguna variable respuesta asociada a ellos. Para realizar este método, se utilizarán diferentes modelos y se escogerá el modelo más valido, con el objetivo de predecir el comportamiento de un atributo.

En primer lugar, se debería explorar, limpiar y procesar los datos, en este caso no será necesario ya que los datos que se van a utilizar ya han sido preparados anteriormente. También, seria primordial reducir la dimensión de los datos, pero al solo tener 8 variables esto no será necesario.

En segundo lugar, se realizará la partición de los datos, para ello se dividirán los datos en dos particiones: entrenamiento y validación. Los datos de entrenamiento se utilizarán para construir el modelo y los de validación se usarán para comprobar cómo funciona el modelo cuando se aplica a nuevos datos. Para la partición se dividirá aleatoriamente los datos, correspondiendo el 75% del total a entrenamiento y el 25% restante a validación.

4.1. Árbol de regresión

A continuación, se realizará el árbol de regresión para predecir el precio de la vivienda (Rawprice), para ello se ajustará el árbol, teniendo en cuenta la variable “Rawprice”, como dependiente, y el resto de las variables serán independientes (Figura 21).

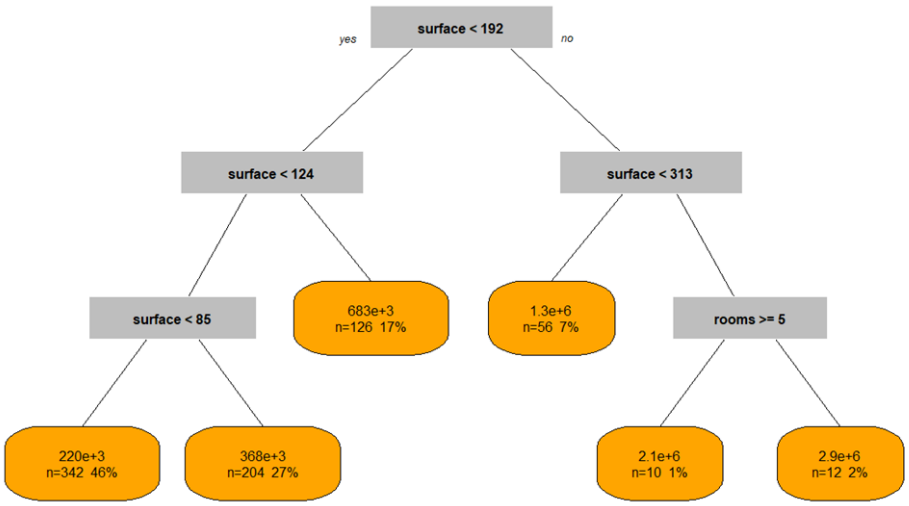


Figura 21. Árbol de regresión

Un árbol de regresión es una jerarquía de pruebas lógicas sobre algunas de las variables explicativas, por lo tanto, los modelos basados en árbol seleccionan automáticamente las variables más relevantes, y por esta razón, no todas las variables necesitan aparecer en el árbol final.

El árbol está formado por diferentes nodos, los cuales tienen dos ramas que están sujetas a si la vivienda cumple una condición de las variables predictoras. Cuando se llega al final de una rama y esta no tiene un nodo de decisión, es que se ha llegado a un nodo terminal. Se llega a este nodo cuando los subgrupos tienen un tamaño mínimo o ninguna mejora es posible.

Como el árbol anterior se ha formado para ajustarse perfectamente a lo datos introducidos, haría falta podarlo para evitar el sobreajuste, debido a que si se utiliza este árbol para nuevos datos dará un bajo desempeño. Por lo tanto, con la poda del árbol lo que se consigue es obtener el árbol óptimo, para facilitar su interpretabilidad y evitar problemas de sobreajuste.

Para la poda del árbol se utilizará la regla X-SE, para ello se generará un árbol con el menor xerror (Tabla 3):

Tabla 3. Error de los diferentes árboles

	CP	nsplit	rel error	xerror	xstd
1	0.5218	0	1.0000	1.0032	0.1545
2	0.0975	1	0.4781	0.5191	0.0737
3	0.0712	2	0.3805	0.4327	0.0636
4	0.0137	3	0.3055	0.3704	0.0608
5	0.0116	4	0.2953	0.3862	0.6463
6	0.0100	5	0.2839	0.3797	0.0645

Como se observa en la tabla, el árbol 4 es el que presenta un menor xerror (0.37024), con un cp=0.0137, por lo tanto, el árbol podado sería el siguiente (Figura 22).

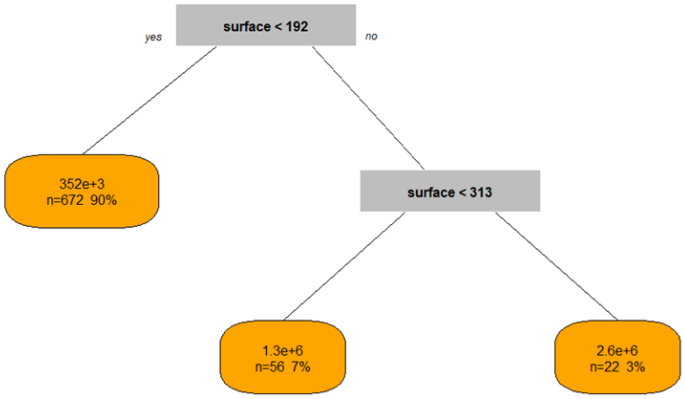


Figura 22. Árbol de regresión podado

Para construir los anteriores arboles se ha utilizado el árbol más grande (Figura 23).

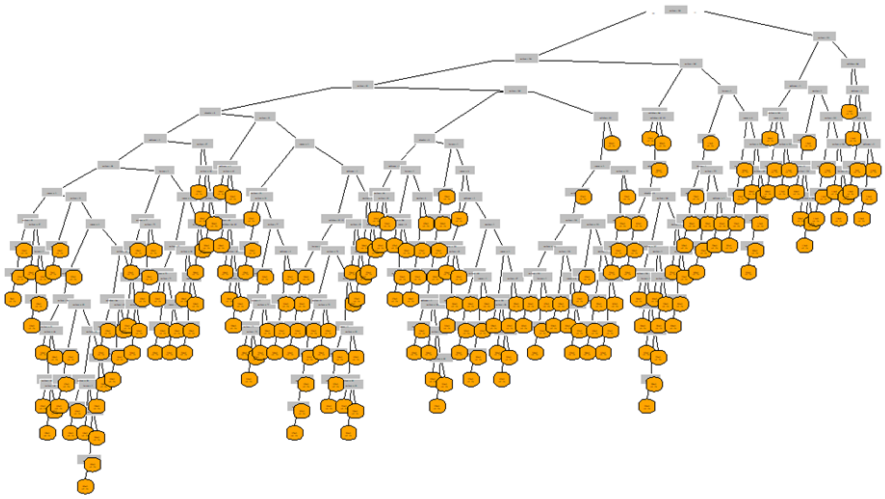


Figura 23. Árbol de regresión completo

4.2. Random Forest

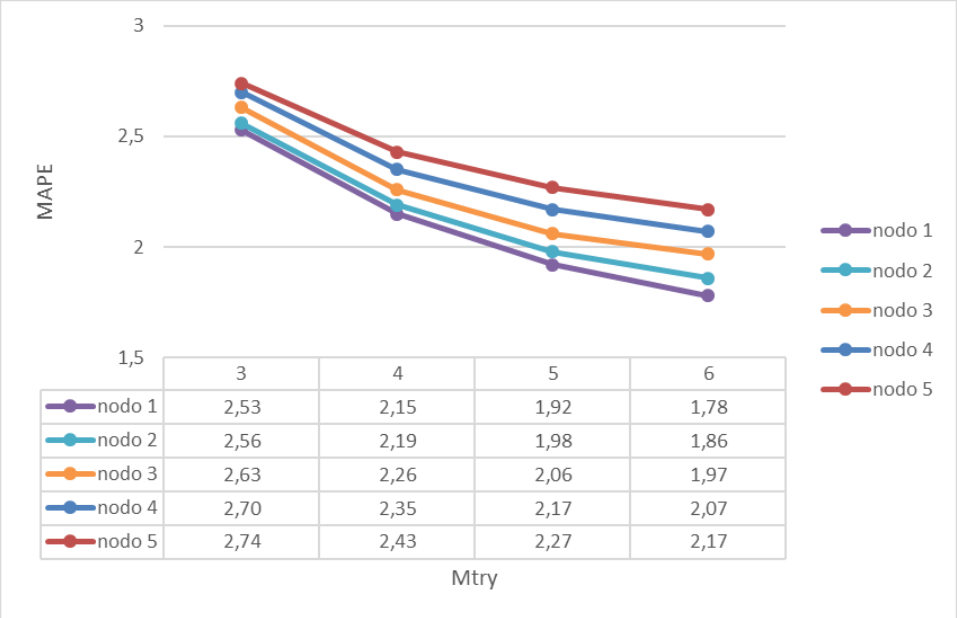
El bosque aleatorio (random forest) añade una nueva capa de aleatoriedad al proceso bagging. Además, en la construcción de cada árbol la división de las ramas de cada nodo se lleva a cabo, no a partir del conjunto total de predictores, sino a través de un subconjunto seleccionado al azar.

Con este algoritmo se realizará una combinación de diferentes árboles de decisión para obtener modelos más estables y con menos propensión al sobreajuste. Este modelo también proporcionara una medida de la importancia de las variables predictoras y la proximidad de unos datos a otros.

Para que funcione este algoritmo la base de datos no debe poseer ningún dato faltante.

En primer lugar, es necesario determinar la muestra aleatoria de variables que van a intervenir en la división de un nodo, en teoría el valor de mtry que ofrece mejores

predicciones en regresión es el número de variables dividido por tres, en este caso mtry tomaría el valor de 3. Pero como este valor es orientativo, también se realizará con otros valores y se seleccionará el que presente mejores resultados de predicción. Para ello se utilizará la medida de bondad de ajuste conocida como mape, que es el error porcentual absoluto medio. Además, también será útil determinar el parámetro nodesize donde se especifica el tamaño mínimo de los nodos terminales de los árboles. Cuanto mayor sea este parámetro, se cultivarán arboles más pequeños.



	3	4	5	6
nodo 1	2,53	2,15	1,92	1,78
nodo 2	2,56	2,19	1,98	1,86
nodo 3	2,63	2,26	2,06	1,97
nodo 4	2,70	2,35	2,17	2,07
nodo 5	2,74	2,43	2,27	2,17

Figura 24. MAPE de random forest

Como se puede observar en la Figura 24, el resultado que presenta un mejor Mape se obtiene a través de utilizar un mtry=6 y nodesize=1, consiguiendo un Mape del 17,8%, lo que indica que la predicción esta errada en un 17,8%.

A continuación, se realizará un gráfico para observar la mejoría del modelo al utilizar un mtry=6, en vez de un mtry=3 (Figura 25)

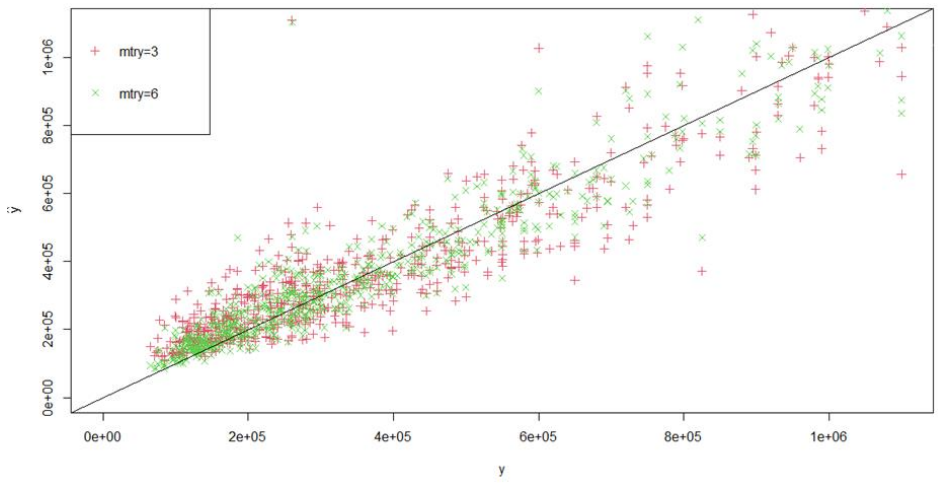


Figura 25. Comparación de la predicción entre mtry=3 y mtry=6

4.3. Vecino más próximo

cercanas, se trata de un modelo simple e intuitivo que se suele utilizar para la predicción. Para este modelo todas las variables deben ser numéricas, además no puede haber valores faltantes.

En primer lugar, se obtendrá el valor k, el cual indica el número de vecinos, este valor se puede obtener de dos formas, haciendo la raíz cuadrada del número de variables (este método indica que K=3) o a través del paquete kknn en el software R que obtiene el k optimo a través de un algoritmo. En este caso, el valor de k=10.

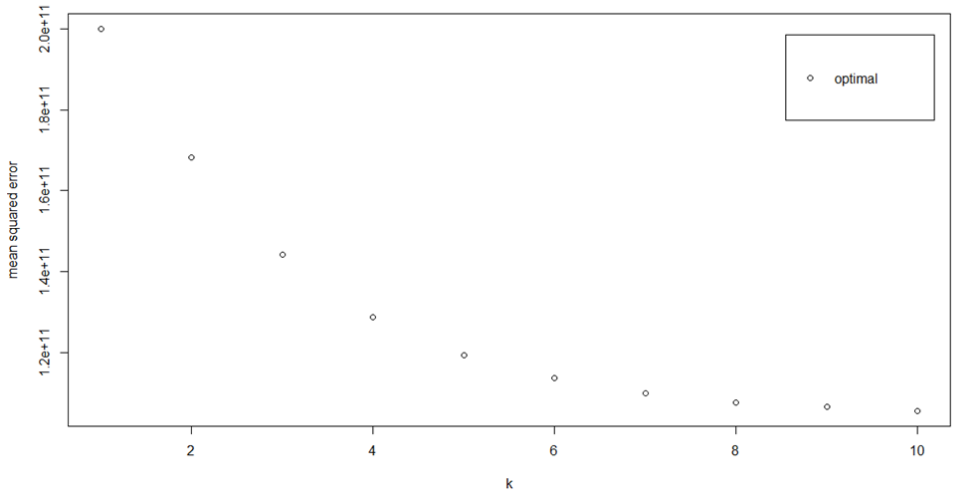


Figura 26. K óptimo

La Figura 26 muestra el error para cada k, por lo tanto, el valor óptimo de k son 10, ya que es el que presenta un menor error.

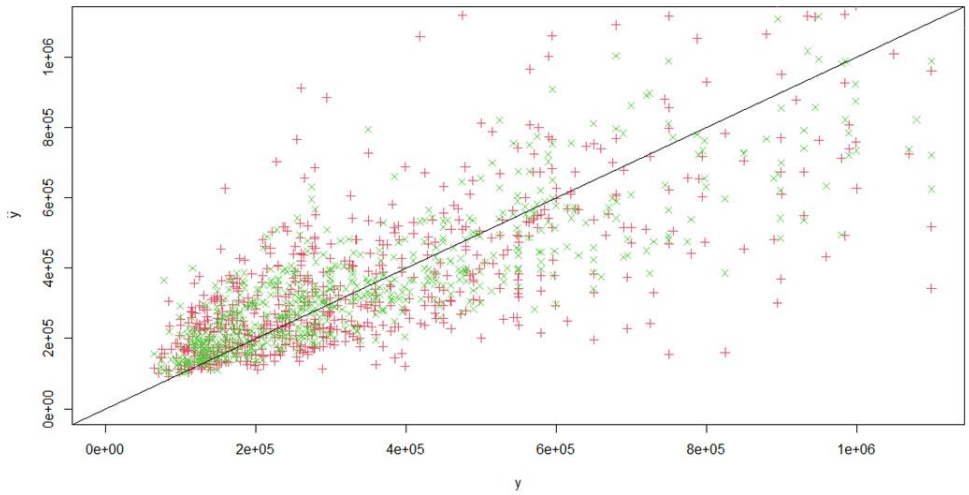


Figura 27. Comparación de la predicción entre k=3 y K=10

Como se observa en la Figura 27, al utilizar k=10 (puntos verdes) lo cual indicaba el paquete kknn que era el número de óptimo de k, ha mejorado el modelo respecto a utilizar k=3 (puntos rojos), que es la raíz cuadrada del número de variables, por lo que se utilizara k=10 para la predicción del vecino más próximo.

4.4. Máquinas de vector soporte vectorial

Las máquinas de soporte vectorial son un conjunto de algoritmos que buscan un hiperplano que separa de forma óptima los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior. Su nombre es debido al vector formado por los puntos más cercanos al hiperplano de separación que se denomina vector de soporte

Para la realización de este modelo, existen dos librerías en R: kernlab y e1071, por lo tanto, se realizará el modelo en las dos librerías y se seleccionará el modelo que presente mejores resultados (Figuras 28 y 29).

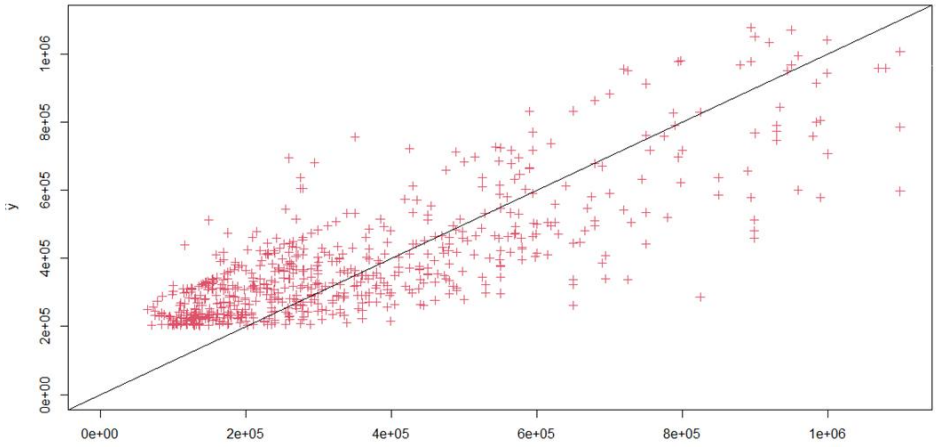


Figura 28. Predicción del modelo frente a las observaciones. Librería kernlab

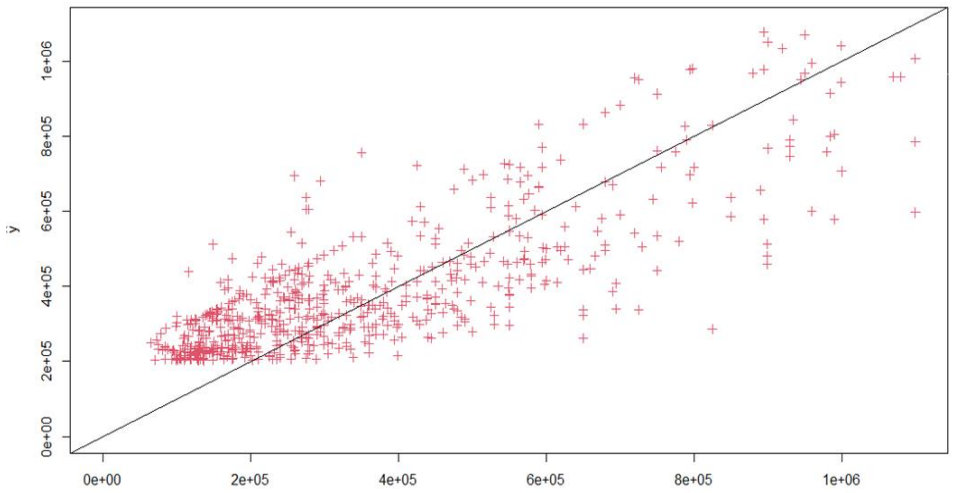


Figura 29. Predicción del modelo frente a las observaciones. Librería e1071

Al observar ambos gráficos y ver las predicciones de las dos librerías, se puede apreciar que la librería "e1071" obtiene una mejor predicción ya que los valores se acercan más a la diagonal. Por lo que se utilizara el modelo de esta librería para la predicción.

4.5. Comparación y selección del modelo

A continuación, se procederá a seleccionar el modelo de predicción que mejor resultado ha aportado, para ello se representará gráficamente las predicciones de los

modelos frente a las observaciones, por lo tanto, el modelo que más se acerque a la diagonal será el mejor modelo

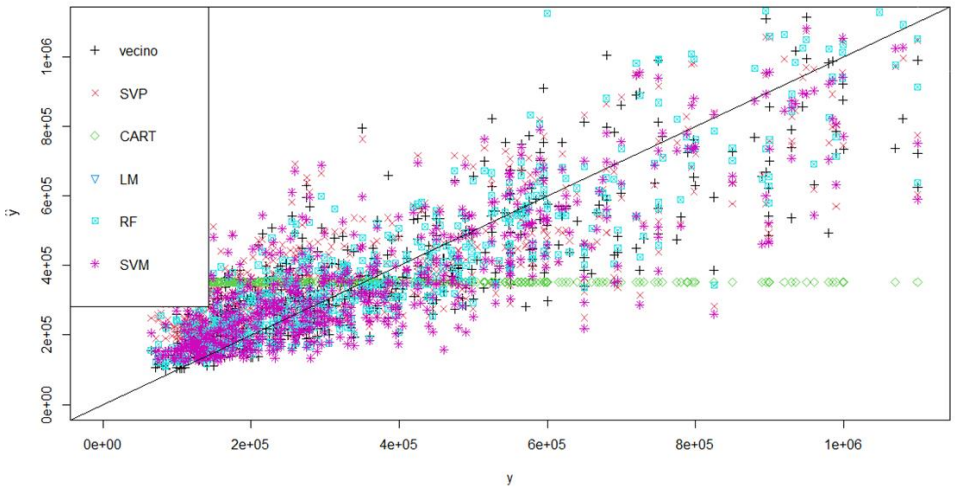


Figura 30. Predicción de todos los modelos frente a las observaciones

En el gráfico no se aprecia claramente el modelo que más se ajusta a la diagonal, por lo que se procederá a calcular las distancias de cada modelo a las observaciones a través de varias medidas de bondad de ajuste:

MAE: es la diferencia absoluta entre el valor verdadero y el valor predicho por el modelo.

MSE: es la suma de los cuadrados de la distancia entre el valor predicho y el valor verdadero

RMSE: es el error cuadrático medio, y representa a la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor predicho.

MAPE: es el promedio del error absoluto o diferencia entre el valor real y el predicho, expresado como un porcentaje de los valores reales

Los resultados se muestran en la Tabla 4.

Tabla 4. Medidas de bondad del ajuste

	MAE	MSE	RMSE	MAPE
Árbol de regresión	2.24 e+05	1.21 e+11	3.48 e+05	7.24 e-01
Random forest	7.00 e+04	1.51 e+10	1.23 e+05	1.78 e-01
Vecino más próximo	1.29 e+05	5.79 e+10	2.40 e+05	3.03 e-01
Máquina soporte vectorial kernlab	1.59 e+05	8.55 e+10	2.92 e+05	4.65 e-01
Máquina soporte vectorial e1071	1.26 e+05	5.77 e+10	2.40 e+05	3.34 e-01

La Tabla 4 muestra las diversas medidas de bondad de ajuste, tanto relativas como absolutas. Un medidor útil para saber si un modelo es bueno, es el MAPE, el cual determina el error porcentual absoluto medio, expresando la exactitud como un porcentaje del error, por lo tanto, contra menor sea el valor MAPE, menor será el error de la predicción y mejor será el resultado del modelo.

Teniendo como referencia el MAPE, el modelo que presenta mejores resultados es el random forest, ya que tiene un MAPE de 17,8%, lo que indica que la predicción esta errada en un 17,8%.

Además, como se ha podido observar en el análisis, las características de la vivienda que más influyen en el precio son sobre todo la superficie, y en menor medida el número de habitaciones, si posee ascensor, también el número de baños y si el edificio posee aparcamiento.

5. Conclusiones

El presente trabajo se ha centrado en el mercado inmobiliario, donde se ha tratado uno de los mayores problemas de este mercado, que es el precio.

Para abordar este problema, se han desarrollado varios modelos que estimen el valor de un inmueble en función de sus características.

El mejor resultado de predicción se ha obtenido a través del algoritmo Random forest, con un Mape del 17,8%, indicando que la predicción de esta técnica esta errada en un 17,8%.

Además, como se ha podido observar en el análisis, las características de la vivienda que más influyen en el precio son sobre todo la superficie, y en menor medida el número de habitaciones, si posee ascensor, también el número de baños y si el edificio posee aparcamiento.

Uno de los mayores problemas a la hora de intentar predecir el valor de las viviendas, ha sido la falta de la variable localización, ya que se trata de un factor incluso más influyente que la superficie. Este ha sido el mayor problema en los resultados obtenidos, ya que, de tener esta variable, los resultados habrían mejorado considerablemente. Debido a la falta de la localización, como se ha observado en la base de datos, existen bastantes viviendas con un elevado precio y poca superficie. Por lo tanto, como la variable que más influía, con gran diferencia, en la determinación del precio de la vivienda, era la superficie, se producía un mayor error en la predicción.

En un futuro trabajo sería interesante analizar una gran base de datos de una ciudad, que poseyera todas las características de la vivienda, no solo las relacionadas con el tamaño, como en este caso. A pesar de que, junto a la localización, son las variables que más influyen. Sería interesante disponer de variables como: año de construcción, reformada, tipo de vivienda (unifamiliar o plurifamiliar), número de pisos, servicios cercanos, jardín, piscina, etc... Debido a que cada vivienda tiene determinadas características que las diferencia de las demás y, por lo tanto, se obtendría una predicción del precio con un menor error.

Bibliografía

Antipov, E.A., Pokryshevsakaya, E.B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and CART-based approach for model diagnostics. *Expert Systems with Applications*, 39, 1772-1778. https://doi.org/10.1016/j.eswa.2011.08.077

Archer, W. R., & Smith, B. C. (2013). Residential mortgage default: the roles of house price volatility, euphoria and the borrower's put option. *The Journal of Real Estate Finance and Economics*, 46(2), 355–378. https://doi.org/10.1007/s11146-011-9335-y

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). https://doi.org/10.3846/1648715X.2015.1134702

Astudillo Cobos, & D. G. (2021). Análisis del mercado inmobiliario español como oportunidadde inversión. *Finance, Markets and Valuation* 7(1), 41–52 https://doi.org/10.46503/OERI9337

Aznar, J., Ferrís-Oñate, J., & Guijarro, F. (2010). An ANP framework for property pricing combining quantitative and qualitative attributes. *Journal of the Operational Research Society*, 61(5), 740-755. https://doi.org/10.1057/jors.2009.31

Çamlibel, M. E., Sümer, L., & Hepşen, A. (2021). Risk-return performances of real estate investment funds in Turkey including the Covid-19 period. *International Journal of Strategic Property Management*, 25(4), 267-277. https://doi.org/10.3846/ijspm.2021.14957

Cervelló, R., García, F., & Guijarro, F. (2011). Ranking residential properties by a multicriteria single price model. *Journal of the Operational Research Society,* 62, 1941–1950. https://doi.org/10.1057/jors.2010.170

D'Amato, M. (2007). Comparing rough set theory with multiple regression analysis as automated valuation methodologies. *International Real Estate Review*, 10(2), 42-65.

Fan, G.Z, Ong S.E., & Koh, H.C. (2006). Determinants of house price: A decision tree approach. *Urban Studies*, 43(12), 2301-2315. https://doi.org/10.1080/00420980600990928

Feng, Z., Miller, S. M., & Tirtiroglu, D. (2021). U.S. REIT industry profitability: a Bennet decomposition of industry dynamics. *International Journal of Strategic Property Management*, 25(4), 316–331. https://doi.org/10.3846/ijspm.2021.14958

Gloudemans, R.J. (1999). Mass appraisal of real property. International Association of Assessing Officers.

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. https://doi.org/10.3390/ijerph16245149

Highfield, M. J., Shen, L., & Springer, T. M. (2021). Economies of scale and the operating efficiency of REITs: a revisit. Journal of Real Estate Finance and Economics, 62(1), 108–138. doi: https://doi.org/10.1007/s11146-019-09741-9

Kim, D.S., & Shin, S. (2021). The economic explainability of machine learning and standard econometric models-an application to the U.S. mortgage default risk. *International Journal of Strategic Property Management*, 25(5), 396–412. https://doi.org/10.3846/ijspm.2021.15129

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of real estate by computational intelligence. *Applied Soft Computing*, 11, 443-448. https://doi.org/10.1016/j.asoc.2009.12.003

Lin, C.C., & Tsai, I.-C. (2021). House prices, rental costs, and mortgage interest rates. International *Journal of Strategic Property Management*, 25(5), 356–368. https://doi.org/10.3846/ijspm.2021.14966

Štreimikienė, D. (2014). Housing indicators for assessing quality of life in Lithuania. *Intellectual Economics*, 8(1), 25-41. https://doi.org/10.13165/IE-14-8-1-02

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36, 2843-2852. https://doi.org/10.1016/j.eswa.2008.01.044

Tay, D.; & Ho, D. (1991). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation & Investment*, 10, 252-541.

Wang, D., & Victor J. L. (2019). Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. *Sustainability* 11(24). https://doi.org/10.3390/su11247006