Facultad de Ciencias Económicas y Empresariales

ANÁLISIS PREDICTIVO DEL PRECIO DE LA VIVIENDA EN LOS DISTRITOS DE CIUDAD LINEAL Y LA LATINA CON MODELOS DE MACHINE LEARNING

Autora: Ana Bruno Cueto
Director: Eduardo César Garrido Merchán

MADRID | Abril 2022

RESUMEN

El mercado inmobiliario es uno de los más importantes de la economía de una país, y también uno de los más cambiantes, pues el precio medio de los inmuebles residenciales fluctúa constantemente. Gracias al desarrollo de las tecnologías y del Machine Learning se han creado modelos y algoritmos que permiten predecir eficazmente los precios de las viviendas y detectar cuales son las variables que más influyen en determinar su valor. En el presente trabajo, tras extraer datos de la web de la plataforma Idealista mediante técnicas de Web Scraping, se analizarán las viviendas de los distritos madrileños de Ciudad Lineal y La Latina, y se construirán modelos de Machine Learning con la herramienta RStudio para obtener predicciones lo más precisas posibles sobre su precio de mercado. Como medida del error cometido al tratar de predecir, se empleará el RMSE (Raíz del error cuadrático medio).

Palabras Clave:

Machine Learning, mercado inmobiliario, ensembles, raíz del error cuadrático medio, modelos predictivos, RStudio.

ABSTRACT

The real estate market is one of the most important in the economy of a country, and also one of the most changing, as the average price of residential real estate is constantly fluctuating. Thanks to the development of technologies and Machine Learning, models and algorithms have been created to effectively predict housing prices and detect which are the most influential variables in determining their value. In this dissertation we will analyze the houses in the Madrid districts of Ciudad Lineal and La Latina, and we will build Machine Learning models with RStudio to obtain the most accurate predictions possible about their market price. As a measure of the error made when trying to predict, the RMSE (Root Mean Squared Error) will be used.

Key Words:

Machine Learning, real estate market, ensembles, root mean square error, predictive models, RStudio.

ÍNDICE

ÍNDICE DE GRÁFICOS, TABLAS, ECUACIONES E ILUSTRACIONES

No table of figures entries found.

1. INTRODUCCIÓN

El mercado inmobiliario es uno de los sectores más importantes de todas las economías del mundo. La compraventa de viviendas es una de las principales fuentes de inversión y una de las decisiones más importantes que toman las personas en sus vidas. De hecho, el último Informe Juventud elaborado en España en el año 2021 por el Instituto Juventud reveló como la vivienda, junto con el trabajo, la educación o la seguridad es uno de los temas que más preocupa a los jóvenes. Además, la actividad inmobiliaria es un componente importante en el Producto Interior Bruto español, reflejo de su actividad económica y de sus perspectivas de crecimiento.

No obstante, el mercado inmobiliario es muy complejo, y los valores de las viviendas dependen de muchas características. ¿Qué factores son los que más influyen en el precio? ¿Qué tendencias pueden detectarse? ¿Qué precio puede tener una vivienda con X características? Como bien arranca el arranca el artículo *"The world's most valuable resource is no longer oil, but data"* de *"The Economist"*, el recurso más valioso del mundo ya no es el petróleo, sino los datos, pues las empresas más potentes del mundo, como Amazon o Google cuentan con ellos como su principal activo. Así, la información, los datos, son críticos y esenciales tanto para las inmobiliarias como para los compradores, para así conocer a la perfección el mercado y poder tomar las mejores decisiones de compra o inversión.

Dicho esto, en este trabajo de fin de grado, se analizará el precio de la vivienda en los distritos madrileños de Ciudad Lineal y La Latina utilizando una de las herramientas más revolucionarias de los últimos años, los modelos de Machine Learning. El objetivo es tanto estudiar la relación entre diversas variables, como la superficie, el número de habitaciones o el número de baños, como llegar a elaborar a modelos que puedan predecir con el mínimo error posible el precio de las viviendas analizadas.

Para ello, se han recopilado datos sobre el mercado inmobiliario en los distritos seleccionados, utilizando fuentes públicas como el Instituto Nacional de Estadística (INE) y portales de anuncios de compraventa de viviendas como Idealista. A

continuación estos datos, tras explorarlos estadísticamente, han sido modificados y procesados para poder emplear técnicas como la regresión lineal, los árboles de decisión o las redes neuronales, comparándose los resultados obtenidos para determinar cuál es el modelo que menos error genera al predecir el precio de la vivienda en los distritos de Ciudad Lineal y La Latina.

Los resultados obtenidos en este trabajo de fin de grado pretenden proporcionar una visión general del mercado inmobiliario de dos distritos madrileños, siendo interesantes para profesionales del sector inmobiliario, inversores y particulares interesados en la compraventa de viviendas en estas zonas de la ciudad. Además, el estudio de los modelos de Machine Learning permitirá obtener predicciones del precio de la vivienda en el futuro y, por lo tanto, tomar decisiones más informadas en cuanto a la compra o inversión en propiedades en estas áreas.

2. ESTADO DEL ARTE

En los últimos años, la aplicación de técnicas de Machine Learning en el ámbito del sector inmobiliario ha experimentado importantes avances, convirtiéndose las técnicas de aprendizaje automático en una herramienta valiosa en el análisis de datos, la predicción de precios de la vivienda, la detección de oportunidades inmobiliarias e incluso como mecanismo para detectar riesgos y burbujas inmobiliarias.

Esto queda demostrado en el estudio de Choy y Ho (2023), dónde se ilustra cómo el aprendizaje automático en el mercado inmobiliario puede proporcionar predicciones de precios más precisas que las técnicas de estadística tradicional. Se comprueba que modelos como *Extra Trees, k-Nearest Neighbors y Random Forest* superan al modelo de precios hedónicos en cuanto a poder explicativo y minimización de costes.

Los modelos de Machine Learning también han sido utilizado para detectar oportunidades de inversión. Este es el caso de Baldominos et al. (2018), dónde se desarrolló una aplicación de aprendizaje automático que identificó en tiempo real oportunidades en el mercado inmobiliario, es decir, viviendas que aparecían listadas con un precio sustancialmente inferior al de mercado. Todo esto debido a que se consiguió detectar que en ocasiones las personas interesadas en vender una vivienda no actualizaban los precios o bien por diversas razones, fijaban deliberadamente un precio inferior al de mercado para vender la vivienda más rápidamente.

Los estudios en la materia han ido desde el empleo de métodos simples como la regresión múltiple, los cuales puede predecir con eficacia y analizar el precio de la vivienda en cierta medida [Zhang, 2021], hasta métodos de aprendizaje automático más avanzados como los ensembles [Alfaro-Navarro et al., 2020] o el *deep learning* [Mohamed et al., 2023].

Por otro lado, el Machine Learning también se ha empleado como mecanismo de detección de riesgos, como por ejemplo Park y Ryu (2021), que analizan la relación entre los mercados de la vivienda y de valores, centrándose en las burbujas del mercado inmobiliario y detectando eficazmente los cambios y la volatilidad futura en los precios del mercado. Otro ejemplo de esto, fue llevado a cabo en el mercado suizo, dónde

analizando la evolución de los precios de venta de los inmuebles residenciales en todos los distritos suizos se consiguió detectar once distritos críticos que mostraban señales de burbujas, y siete distritos en los que las burbujas ya habían estallado [Ardila et al., 2013].

En definitiva, se han llevado a cabo numerosos análisis del mercado inmobiliario con modelos de Machine Learning y en diversos lugares: Florida [Bhushan Jha et al.], Virginia [Park & Kwon Bae, 2015], Boston [Sanyal et al., 2022] y Suiza [Moosavi, 2017] entre muchos otros. En España también existen análisis que desarrollan modelos para el mercado español, como por ejemplo Alfaro-Navarro et al. (2020) que consiguen proporcionar de forma totalmente automática el mejor modelo para cada municipio español. No obstante, el presente trabajo pretenderá esclarecer y realizar predicciones en los distritos madrileños de Ciudad Lineal y La Latina.

3. DEFINICIÓN DEL PROYECTO

En esta sección se proporcionará una definición clara y precisa de algunos aspectos del trabajo. En primer lugar, se explicará las hipótesis del trabajo, es decir, aquello que se intentará confirmar o refutar. A continuación, se detallarán los objetivos, así como las asunciones y restricciones que condicionan el diseño y desarrollo de los modelos predictivos.

3.1. Hipótesis

La investigación y análisis del mercado inmobiliario de los distritos de Ciudad Lineal y La Latina pretende dar respuestas a las siguientes preguntas de investigación:

Pregunta I: ¿Son los modelos propuestos capaces de obtener una raíz del error cuadrático medio (RMSE) inferior a 60.000 euros en la predicción de un piso?

Pregunta II: ¿Es el trabajo capaz de cuantificar monetariamente cada una de las variables de los pisos?

Pregunta III: ¿Pueden los modelos planteados obtener una raíz del error cuadrático medio (RMSE) inferior a la media?

3.2. Objetivos

En el presente trabajo se van a emplear diversos modelos de Machine Learning para analizar las características que componen el precio de la vivienda. Los objetivos marcados son los siguientes:

- Cuantificar monetariamente el efecto de variables como: el número de habitaciones, el tipo de calefacción de la vivienda, la presencia de garaje y el estado del inmueble entre otras.
- Elaborar varios modelos de Machine Learning que nos permitan ayudar a estimar y predecir los precios de la vivienda en distritos de Madrid.
- Construir ensembles para lograr la mejor predicción posible para el conjunto de datos.

3.3. Asunciones

A efectos de poder entender mejor tanto los resultados como las preguntas planteadas, es necesario considerar que para la elaboración del proyecto se ha asumido lo siguiente:

- El precio de los pisos queda mayoritariamente explicado por las variables del data set elaborado (Distrito, Precio, M2, Habitaciones, Ascensor, Exterior/Interior, Planta, Baños, Garaje, Estado, Terraza/Balcón y Calefacción).
- En los distritos de Ciudad Lineal y La Latina no hay viviendas más caras o baratas de las mostradas.
- Las viviendas analizadas son todas las viviendas que existen en esos distritos y por lo tanto la muestra es representativa.

3.4. Restricciones

Además también será necesario tener en cuenta que dada la naturaleza del trabajo, se cuenta con una serie de restricciones.

- El tiempo empleado para realizar el trabajo es de 8 meses y el valor de la asignatura es de 6 créditos, por lo que el análisis será acorde a ese tiempo de estudio.
- La capacidad de computación está restringida a la que ofrece un ordenador portátil de la marca Apple, modelo MacBook Air (chip Apple M1).

4. MARCO TEÓRICO

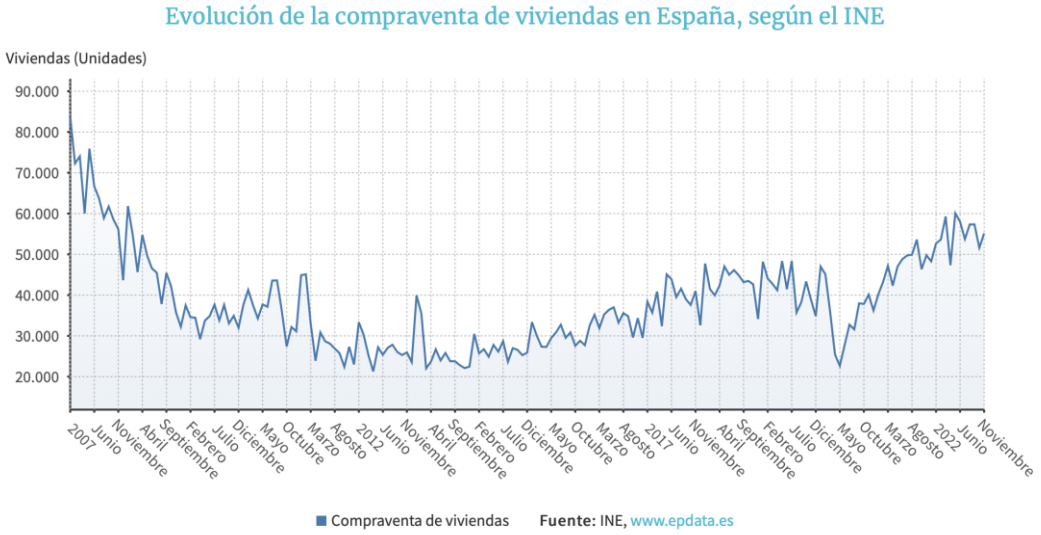
En esta sección del trabajo se presentará el marco teórico que fundamentará el análisis predictivo de los precios de las viviendas. En primer lugar, se analizará el mercado inmobiliario en España, prestando especial atención a la evolución del precio de la vivienda en los últimos años y como han afectado las diversas crisis a su determinación. En segundo lugar se realizará una aproximación al concepto de Machine Learning, sus tipos y la metodología empleada en el presente trabajo. Este marco teórico será fundamental para comprender los conceptos y herramientas que se utilizarán y para situar el trabajo en el contexto adecuado.

4.1. El mercado inmobiliario en España

El mercado inmobiliario es aquel relacionado con todas las operaciones de compraventa, alquiler y otro conjunto de acciones de oferta y demanda de bienes inmuebles. Estos inmuebles pueden ser de naturaleza residencial, comercial, industrial o urbana. Este mercado es esencial para el desarrollo de la economía de una país, contribuyendo a mejorar el bienestar de la sociedad. No sólo es una fuente de empleo, sino que genera grandes ingresos públicos y proporciona un servicio de primera necesidad como es la vivienda. [Otero Moreno y Blanco García Lomas, 2014. p.14].

Pese a su gran importancia, en España, este mercado ha experimentado grandes éxitos, pero también fuertes recesiones, como bien puede observarse en el Gráfico 1:Evolución de la compraventa en España (2007-2022) y que han afectado de forma considerable a la economía española, debido a su estrecha relación con el sector.

Gráfico 1:Evolución de la compraventa en España (2007-2022)

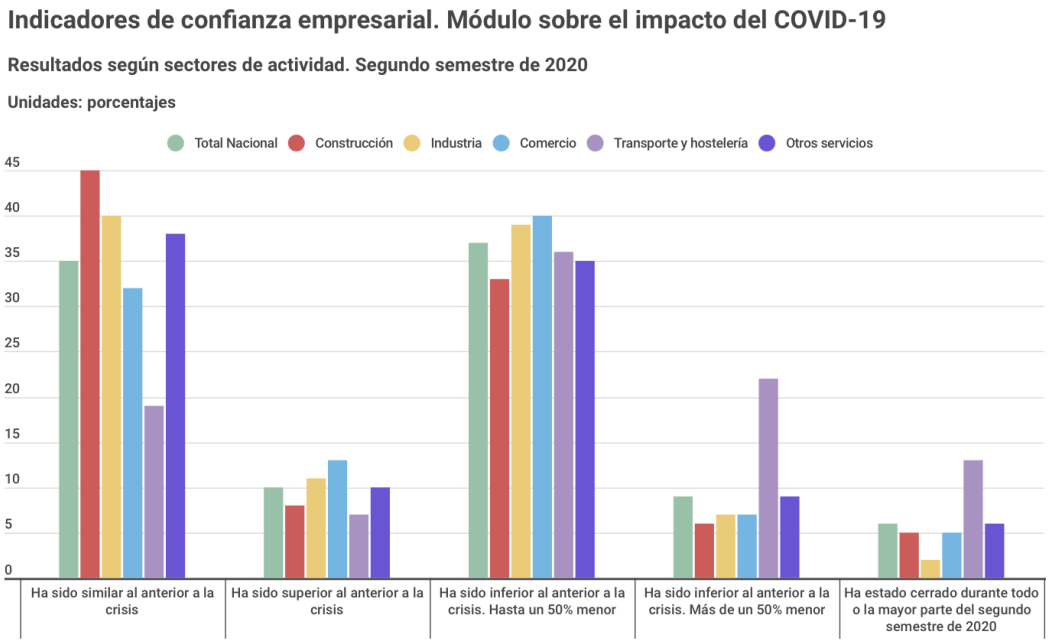


Fuente: Elaborado por Epdata con datos del Instituto Nacional de Estadística

De hecho, el mercado inmobiliario fue el gran protagonista en la crisis del año 2008, que surgió como "consecuencia del impacto de la crisis de los mercados crediticios y financieros internacionales, cuyo origen fue el estallido del boom de las hipotecas *subprime* en Estados Unidos. A estos factores externos se unieron otros de carácter internos, propios de la economía española, como el exuberante comportamiento expansivo de la actividad inmobiliaria, del crédito hipotecario, el elevado endeudamiento de las familias y el elevado ritmo de gasto público realizado en la etapa del boom económico, entre otros factores" [Martínez Álvarez, y García Martos, 2014. p.5]. Todo esto provocó la explosión de la burbuja inmobiliaria, causando el desplome de las compraventas, lo cual arrastró a las promotoras, constructoras, bancos y cajas de ahorro, originando una crisis económica no sólo en España, si no a nivel global.

También el mercado inmobiliario, como casi todos los aspectos de la economía (Ver Gráfico 2: Impacto del Covid-19 en los distintos sectores de actividad) sufrió las consecuencias de la pandemia de la Covid-19. Si bien el más afectado fue el sector del transporte y la hostelería debido a las restrictivas medidas de movilidad implantadas por el Gobierno, en el sector inmobiliario se debatía sobre si nos encontrábamos en la "antesala del estallido de una nueva burbuja" [Ocaña P. de Tudela, y Torres, 2019. p.1].

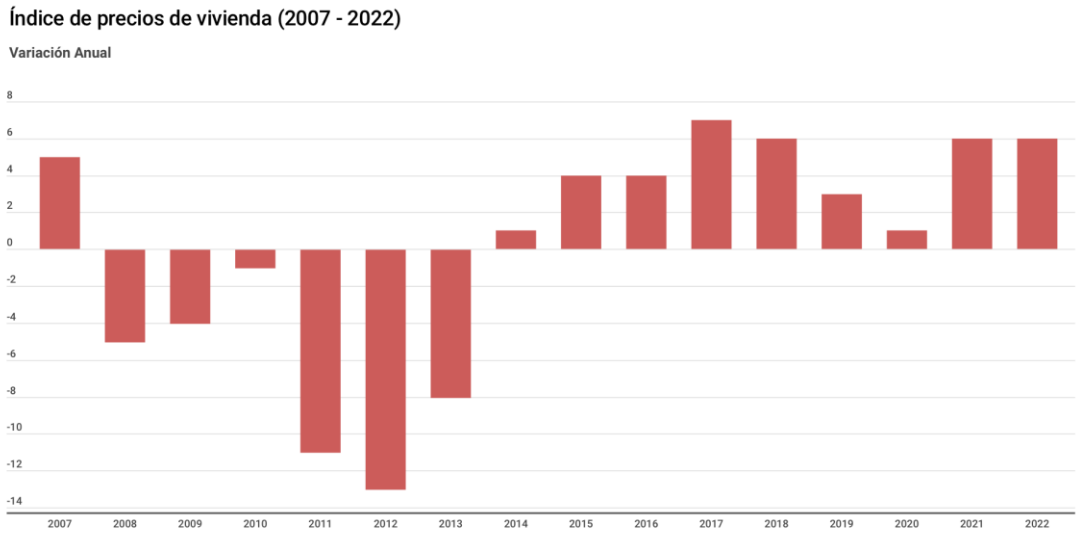
Gráfico 2: Impacto del Covid-19 en los distintos sectores de actividad



Fuente: Elaboración propia a partir de los datos del Instituto Nacional de Estadística (2021)

Pese a que en los primeros meses, la actividad del mercado inmobiliario se redujo intensamente por las restricciones adoptadas, a partir del verano de 2020, con el levantamiento parcial de las restricciones, la actividad comenzó a recuperarse, si bien los desarrollos de la pandemia condicionaron su evolución. La crisis sanitaria supuso un cambio en las preferencias sobre el tipo de vivienda demandada hacia viviendas de mayor tamaño y con espacios exteriores. A pesar de la intensa contracción de la actividad económica, los precios de la vivienda no mostraron caídas generalizadas (Ver Gráfico 3: Evolución IPV (2007-2022)), en un contexto de ausencia previa de grandes desequilibrios en este mercado y de una posición financiera de los hogares más sólida respecto a la que tenían en los años anteriores a la crisis de 2008 [Alves y San Juan, 2021. p.2-3].

Gráfico 3: Evolución IPV[1] (2007-2022)

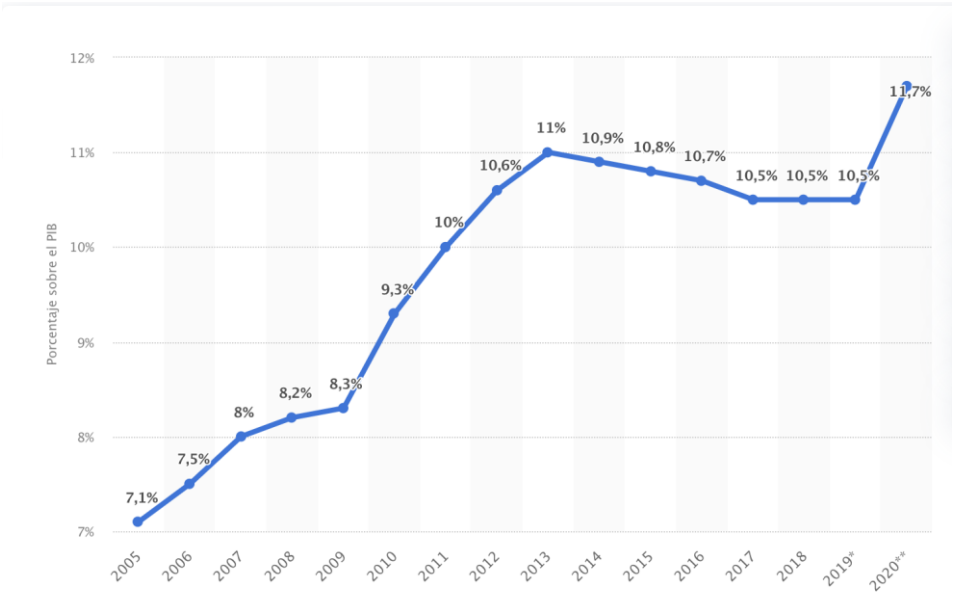


Fuente: Elaboración propia a partir de los datos del Instituto Nacional de Estadística (2022)

No obstante, pese a los reveses que ha sufrido este sector en nuestro país, cuenta con un gran peso en el Producto Interior Bruto (en adelante PIB), que además ha ido aumentando paulatinamente desde la crisis de 2008 (ver Gráfico 4: Evolución anual del peso de las actividades inmobiliarias sobre el PIB en España desde 2005 hasta 2020). "De esta forma, mientras que en 2005 representaba un 7,1% sobre el total del PIB, a partir de 2012 suponía ya más del 10%, alcanzando su máximo en 2020, con 11,7%" [Statista, 2022].

[1] El Índice de Precios de Vivienda (IPV) tiene como objetivo la medición de la evolución de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano, a lo largo del tiempo. [INE]

Gráfico 4: Evolución anual del peso de las actividades inmobiliarias sobre el PIB en España desde 2005 hasta 2020



Fuente: Statista (2022)

Estos datos, junto con el 5% del PIB que aportó el sector de la construcción "dimensionan la importancia del sector inmobiliario y de la construcción en nuestra economía y reflejan cómo la actividad inmobiliaria se ha comportado razonablemente bien teniendo en cuenta el shock al que nos hemos enfrentado en el último año" [Tinsa, 2021].

4.2. Machine Learning

En esta sección, se abordará el Machine Learning, una disciplina de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos. Se explicarán los tipos más relevantes de Machine Learning, así como la metodología empleada para la construcción de los modelos a los que se refiere el trabajo..

4.2.1. Definición del concepto

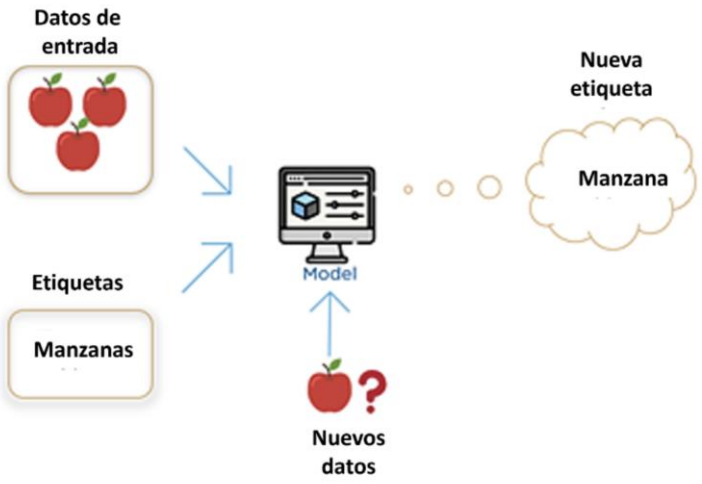
El Machine Learning es una rama de la Inteligencia Artificial que se enfoca en el desarrollo de algoritmos y modelos matemáticos que permiten a una máquina aprender a partir de datos sin ser programada explícitamente para realizar una tarea específica. En otras palabras, el Machine Learning permite a las máquinas aprender de forma autónoma y mejorar su rendimiento en una tarea a medida que reciben más datos, con el

objetivo de extraer conocimiento a partir de un conjunto de datos aparentemente desestructurado. Actualmente está en presente en recomendaciones de películas en plataformas digitales, coches autónomos e inteligentes, reconocimiento del habla de asistentes virtuales, robótica, motores de búsqueda, diagnóstico médico, detección de fraude en pagos con tarjeta, ...siendo empleado por empresas como Apple, Google, Amazon o Spotify.

Existen principalmente tres tipos de Machine Learning, aunque también hay variantes y combinaciones de estos:

1. Aprendizaje supervisado: es una técnica de aprendizaje automático en la que se le proporcionan al modelo datos previamente etiquetados, con el objetivo de que aprenda la relación entre las variables de entrada (inputs) y las de salida (outputs). El objetivo de estos modelos es que el sistema sea capaz de hacer predicciones precisas sobre nuevos datos. Se utiliza en una amplia variedad de aplicaciones, como la clasificación de correos electrónicos como spam o no spam, la predicción de precios de acciones o la identificación de objetos en imágenes.

Ilustración 1: Ejemplo de Aprendizaje supervisado

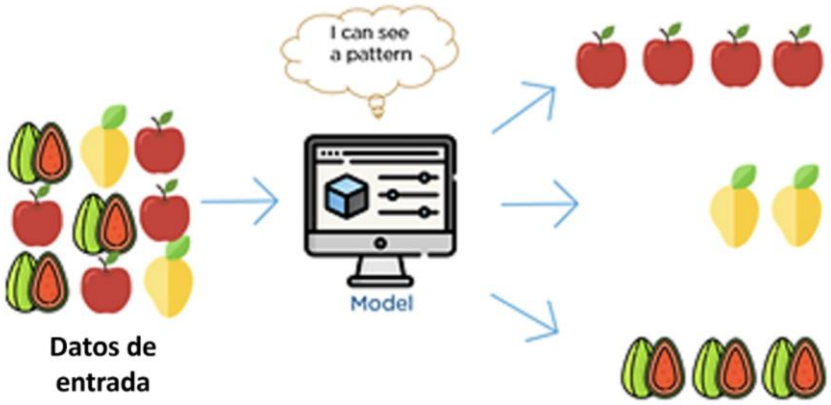


Fuente: KeepCoding (2022)

2. Aprendizaje no supervisado: en el caso de técnicas de aprendizaje no supervisado se proporcionan al modelo datos sin etiquetar, sin un resultado conocido, con el objetivo de encontrar patrones o estructuras en los datos. Sus

principales usos incluyen la agrupación de consumidores con gustos similares o la identificación de estructuras en imágenes.

Ilustración 2: Ejemplo de Aprendizaje no supervisado



Fuente: KeepCoding (2022)

3. Aprendizaje por refuerzo: en estos algoritmos el modelo toma decisiones recibiendo recompensas o castigos sobre sus acciones. El objetivo es que el modelo encuentre la secuencia óptima para maximizar la recompensa y mejorar su comportamiento futuro. Se utiliza en una amplia variedad de aplicaciones, como el control de robots, los juegos de mesa y la optimización de sistemas de recomendación. No obstante es preciso remarcar que este tipo de aprendizaje puede ser más puede ser más complicado y costoso que los anteriores, pues requiere una gran cantidad de datos para ser efectivo.

Ilustración 3: Ejemplo de Aprendizaje por refuerzo

Fuente: KeepCoding (2022)

Los modelos de Machine Learning se basan en algoritmos. Según el matemático, astrónomo y geógrafo persa Muhammad ibn Musa al-Khwarizmi (siglo IX) un algoritmo es un conjunto de operaciones autocontenidas para ser realizadas paso a paso. En el caso del Machine Learning, los algoritmos son utilizados para aprender patrones en los datos y luego utilizar estos patrones para hacer predicciones sobre datos nuevos. Existen tanto algoritmos de aprendizaje supervisado, como pueden ser clasificación (target categórico) o regresión (target numérico); como algoritmos de aprendizaje no supervisado como puede ser el clustering.

Por otro lado, un modelo es una representación simplificada de la realidad, y la clave de los mismo es encontrar el equilibrio entre su sesgo y su varianza. El sesgo (*bias*) hace referencia a cuánto las predicciones de un modelo se desvían en promedio de los valores reales, es decir la capacidad del modelo comprender la verdadera relación entre las variables predictoras y la variable de respuesta. Por otro lado, la varianza hace referencia a cuánto cambia un modelo en función de los datos que se utilizan para entrenarlo, permitiendo ver si el modelo está memorizando los datos en lugar de aprender la verdadera relación entre las variables. Así, "a medida que aumenta la complejidad de un modelo, este dispone de mayor flexibilidad para adaptarse a las observaciones, reduciendo así el bias (sesgo) y mejorando su capacidad predictiva. Sin embargo, alcanzado un determinado grado de flexibilidad, aparece el problema de *overfitting* (sobreajuste), el modelo se ajusta tanto a los datos de entrenamiento que es incapaz de predecir correctamente nuevas observaciones. El mejor modelo es aquel que consigue un equilibrio óptimo entre bias y varianza" [Amat Rodrigo, 2020]. Una de las formas de lograr este equilibro es a través de los métodos de ensemble, pues combinan múltiples modelos consiguiendo mejores predicciones que los modelos originales (Ver 6.5 Ensembles).

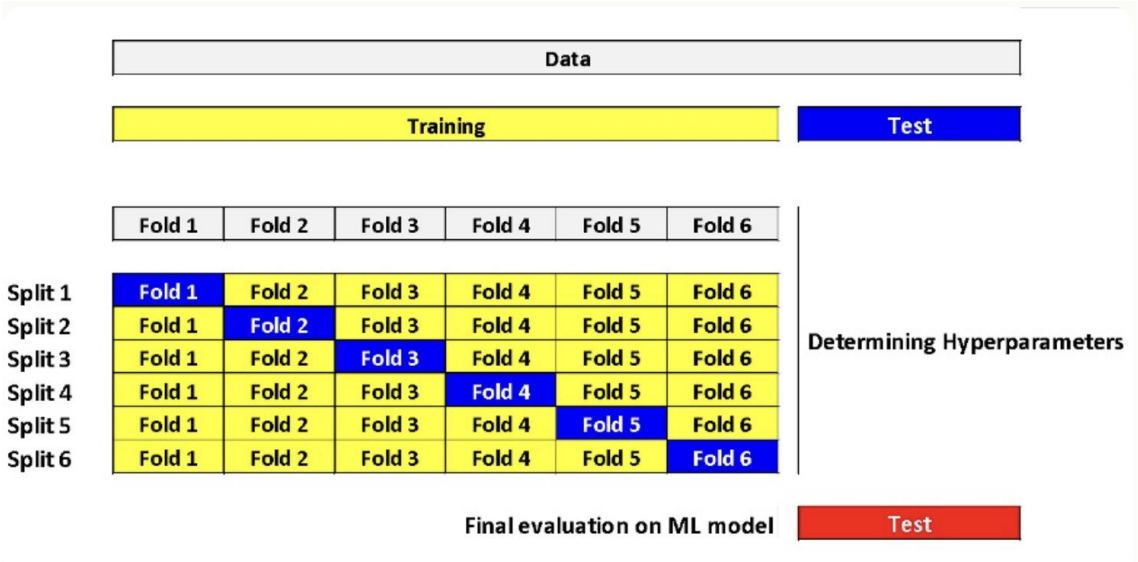
Los modelos podemos clasificarlos según su finalidad (explicativos o predictivos) o según su interpretabilidad (interpretables o de caja negra). Para evaluar el rendimiento de los modelos predictivos existen diversas métricas destinadas a calcular el error, esto es, la diferencia entre los valores reales y los predichos. En este trabajo se empleará la raíz de error cuadrático medio (RMSE), que mide el nivel de dispersión de los errores.

4.2.2. Metodología empleada para la construcción de los modelos

Antes de adentrarse en la elaboración de los modelos hay que precisar la metodología que va a ser empleada para construir los mismos. Para probar la eficacia de un modelo de Machine Learning, lo más común es realizar una partición aleatoria en los datos (Train, 80% y Test, 20%), utilizando una parte para el entrenamiento y ajuste del modelo y otra para realizar predicciones y evaluar los errores de predicción. Pese a que este enfoque es generalmente aceptado, no deja de ser una única estimación, basada en una única partición aleatoria, la cual no podemos saber si extensible a todo el conjunto de datos y a todas las posibles particiones.

Es por ello, que existen técnicas más avanzadas para evaluar el rendimiento y la generalización de los modelos, como la validación cruzada en k pliegues (*k-fold Cross-Validation*). En este método, en lugar de dividir los datos en Training y Test, se hacen k particiones[2] iguales. Posteriormente el modelo se entrena k veces, utilizando los folds k-1 para entrenar el modelo y el k-fold restante para evaluar su rendimiento. Así cada fold habrá sido usado una vez como como conjunto de validación, habiendo hecho predicciones sobre todo el conjunto de datos. La media de los errores registrados será la métrica que evalué el rendimiento del modelo. [Cross-Validation : definición e importancia en Machine Learning, 2022].

Ilustración 4: Funcionamiento de K-fold Cross-Validation



Fuente: d'Archimbaud (s. f.)

[2] Aunque k puede establecerse en cualquier número, lo más común es utilizar una validación cruzada de 10 pliegues (10-fold CV), ya que la evidencia empírica sugiere que el uso de un número mayor tiene pocas ventajas añadidas. [Lantz, 2013. p.319]

Por otro lado, el nivel de complejidad de muchos modelos puede controlarse a través de los llamados hiperparámetros. La mayoría de los algoritmos de aprendizaje automático permiten ajustar al menos un parámetro, y los modelos más sofisticados ofrecen un gran número de formas de ajustar el modelo. Algunos ejemplos pueden incluir la cantidad de capas en una red neuronal, la profundidad de un árbol de decisión, o el tipo de kernel utilizado en una máquina de vectores de soporte (SVM) entre otros. Así, para realizar este ajuste en los modelos habrá que atender: (i) al modelo de Machine Learning que se quiera implementar, (ii) a los hiperparámetros del modelo que pueden ajustarse (Ver Ilustración 5: Hiperparámetros de cada uno de los modelos de Machine Learning con el paquete Caret), y (iii) a los criterios quieren utilizarse para evaluar los modelos, en el presente caso el RMSE. [Lantz, 2013].

Ilustración 5: Hiperparámetros de cada uno de los modelos de Machine Learning con el paquete Caret

Model	Learning task	Method name	Parameters
k-Nearest Neighbors	Classification	`knn`	`k`
Naïve Bayes	Classification	`nb`	`fL`, `usekernel`
Decision Trees	Classification	`C5.0`	`model`, `trials`, `winnow`
OneR Rule Learner	Classification	`OneR`	None
RIPPER Rule Learner	Classification	`JRip`	`NumOpt`
Linear Regression	Regression	`lm`	None
Regression Trees	Regression	`rpart`	`cp`
Model Trees	Regression	`M5`	`pruned`, `smoothed`, `rules`
Neural Networks	Dual use	`nnet`	`size`, `decay`
Support Vector Machines (Linear Kernel)	Dual use	`svmLinear`	`C`
Support Vector Machines (Radial Basis Kernel)	Dual use	`svmRadial`	`C`, `sigma`
Random Forests	Dual use	`rf`	`mtry`

Fuente: Lantz (2013)

Esta técnica será implementada a través de una búsqueda de rejilla (*Grid Search*), en la cual se especificará un rango de posibles combinaciones de parámetros para luego entrenar el modelo con cada combinación y después seleccionar la combinación de valores de hiperparámetros que produce el mejor rendimiento (el menor RMSE).

Ahora bien, es preciso señalar que estas técnicas más avanzadas (*cross-validation* y búsqueda de hiperparámetros) sólo serán empleadas en la elaboración de algunos de los modelos (redes neuronales, Random Forest y ensembles), ya que R cuenta con la

librería Caret[3], que hace muy sencilla su implementación. Por lo que cabría esperar que los resultados de estos modelos sean mejores que en los que se emplean métodos menos complejos.

Una vez realizada una aproximación a estos conceptos, los modelos empleados serán los explicados en el apartado 6. MODELOS DE MACHINE LEARNING.

[3] Para más información sobre la librería Caret véase por ejemplo: Kuhn, M. (2019). *The caret Package*. https://topepo.github.io/caret/index.html

5. ANÁLISIS EXPLORATORIO DE LOS DATOS

"El análisis exploratorio de los datos se refiere al conjunto de técnicas estadísticas cuyo objetivo es explorar, describir y resumir la naturaleza de los datos y comprender las relaciones existentes entre las variables de interés, maximizando la comprensión del conjunto de datos. Un análisis exploratorio de datos posee importantes ventajas:

- permite identificar posibles errores (datos incorrectamente introducidos, detectar la ausencia de valores o una mala codificación de las variables),
- revela la presencia de valores atípicos (outliers),
- permite comprobar la relación entre variables (correlaciones) y su posible redundancia, y
- realizar un análisis descriptivo de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos."[Guía práctica de introducción al Análisis Exploratorio de Datos, 2021. p.3].

Por ello, previo a la elaboración de los modelos de Machine Learning, realizaremos un análisis exploratorio de los datos objeto de estudio, con el objetivo de poder comprenderlos y detectar patrones que puedan ser útiles para entender las diferencias de precio entre distritos o viviendas.

En este caso, para realizar este análisis se utilizará el lenguaje de programación R, además de la herramienta Excel e Infogram. R es un lenguaje de programación open source (de código abierto) que permite llevar a cabo análisis estadísticos, modelado y gráficos de manera profunda gracias a la amplia gama de herramientas disponibles. Su flexibilidad permite a los usuarios escribir y compartir sus propios paquetes y funciones, lo que significa que puede ser personalizado para satisfacer las necesidades específicas de cada usuario. [An Introduction to R, 2022]. Por otro lado Infogram es una plataforma en línea, que cuenta con una versión gratuita y que permite a los distintos usuarios crear y compartir gráficos, mapas, presentaciones, y otros tipos de visualizaciones de datos. Además permite importar datos de fuentes diversas como Excel, MySQL, Oracle o Google Drive.

5.1. Preparación y limpieza de los datos

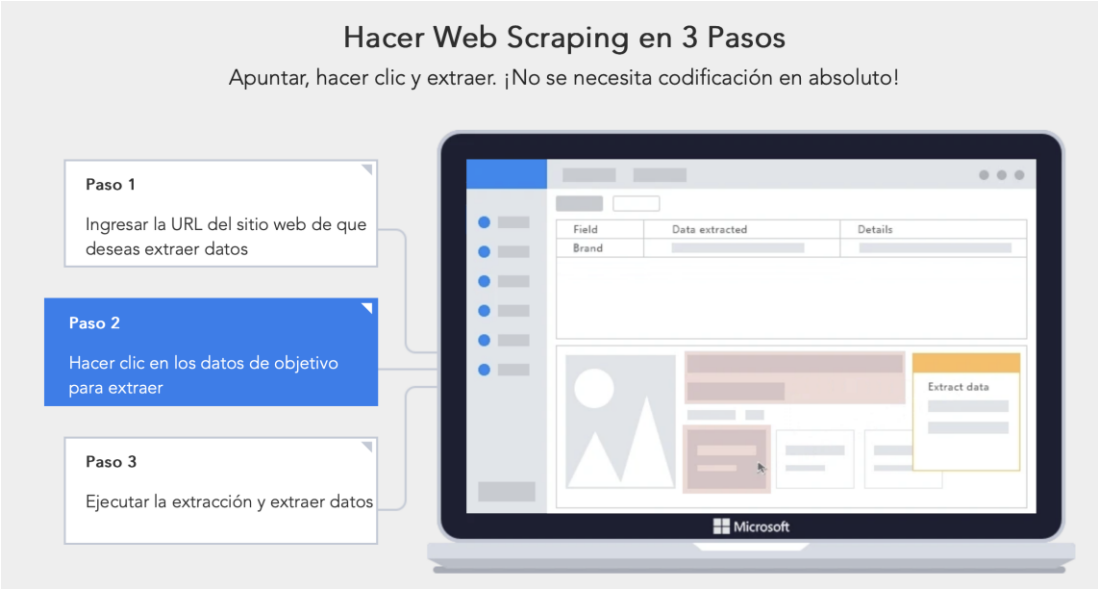
La base de datos empleada en el proyecto es de elaboración propia mediante el Web Scrapping de la página de Idealista, uno de los portales inmobiliarios más importantes de España.

El Web Scrapping es una técnica que permite extraer información de diversas páginas webs de forma automatizada, para después recoger todos esos datos en un fichero Excel o .csv. Es importante remarcar que es esencial que las páginas web de las que se pretenden obtener los datos deben guardar la misma estructura, pues lo que se extraen son unos campos determinados, por lo que si la disposición cambia, se obtendrán datos aleatorios.

El Web Scrapping tiene funcionalidades muy diversas: detectar cambios en sitios web, obtener precios para elaborar comparadores o labores de investigación entre otros. En este caso la finalidad será extraer todas las características posibles de las viviendas en los distritos de Ciudad Lineal y La Latina (número de baños, metros cuadrados, la existencia de ascensor o terraza…) para así elaborar una base de datos que nos permita construir modelos de Machine Learning.

En el presente proyecto se empleó la aplicación de Octoparse, que permite hacer Web Scrapping sin la necesidad de programar y con los pasos que se muestran a continuación.

Ilustración 6: Web Scrapping en Octoparse

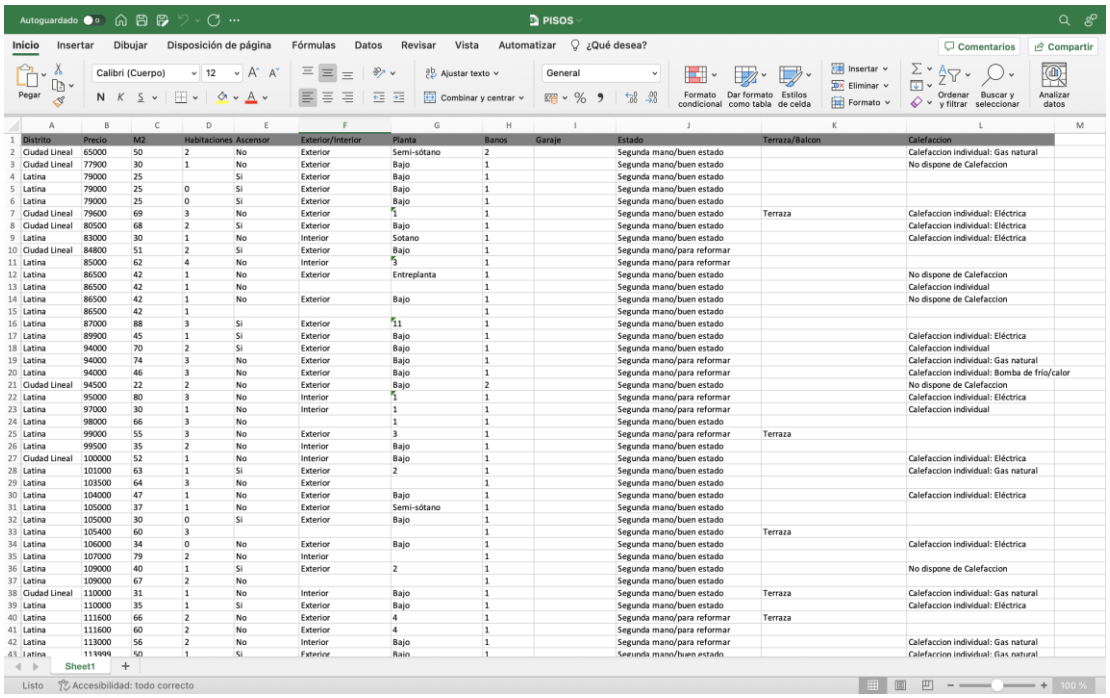


Fuente: Octoparse.es

Gracias a esta aplicación (que cuenta con una versión gratuita), con fecha de noviembre del año 2022 se extrajeron todas las viviendas disponibles en la página de Idealista en los distritos objeto de estudio, junto con sus características, de manera muy eficiente y óptima.

Una vez extraídos los datos a un fichero Excel se procedió a hacer una limpieza y orden de los datos para que a la hora de importarlos a RStudio no hubiera problemas y pudieran utilizarse para elaborar los modelos. También se dicotomizaron las variables Ascensor y Garaje con la función buscar y reemplazar de Excel. El resultado fue un fichero de 1920 observaciones, las cuales contaban con 12 variables. Se muestra a continuación:

Ilustración 7: Excel con los datos de las viviendas a analizar



Distrito	Precio	M2	Habitaciones	Ascensor	Exterior/Interior	Planta	Banos	Garaje	Estado	Terraza/Balcon	Calefaccion
Ciudad Lineal	65000	50	2	No	Exterior	Semi-sótano	2		Segunda mano/buen estado		Calefaccion individual: Gas natural
Ciudad Lineal	77900	30	1	No	Exterior	Bajo	1		Segunda mano/buen estado		No dispone de Calefaccion
Latina	79000	25		Si	Exterior	Bajo	1		Segunda mano/buen estado		
Latina	79000	25	0	Si	Exterior	Bajo	1		Segunda mano/buen estado		
Latina	79000	25	0	Si	Exterior	Bajo	1		Segunda mano/buen estado		
Ciudad Lineal	79600	69	3	No	Exterior	1	1		Segunda mano/buen estado	Terraza	Calefaccion individual: Eléctrica
Ciudad Lineal	80500	68	2	Si	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	83000	30	1	No	Interior	Sotano	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Ciudad Lineal	84800	51	2	Si	Exterior	Bajo	1		Segunda mano/para reformar		
Latina	85000	62	4	No	Interior	3	1		Segunda mano/para reformar		
Latina	86500	42	1	No	Exterior	Entreplanta	1		Segunda mano/buen estado		No dispone de Calefaccion
Latina	86500	42	1	No			1		Segunda mano/buen estado		Calefaccion individual
Latina	86500	42	1	No	Exterior	Bajo	1		Segunda mano/buen estado		No dispone de Calefaccion
Latina	86500	42	1				1		Segunda mano/buen estado		
Latina	87000	88	3	Si	Exterior	11	1		Segunda mano/buen estado		
Latina	89900	45	1	Si	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	94000	70	2	Si	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual
Latina	94000	74	3	No	Exterior	Bajo	1		Segunda mano/para reformar		Calefaccion individual: Gas natural
Latina	94000	46	3	No	Exterior	Bajo	1		Segunda mano/para reformar		Calefaccion individual: Bomba de frio/calor
Ciudad Lineal	94500	22	2	No	Exterior	Bajo	2		Segunda mano/buen estado		No dispone de Calefaccion
Latina	95000	80	3	No	Interior	1	1		Segunda mano/para reformar		Calefaccion individual: Eléctrica
Latina	97000	30	1	No	Interior	1	1		Segunda mano/para reformar		Calefaccion individual
Latina	98000	66	3	No		1	1		Segunda mano/buen estado		
Latina	99000	55	3	No	Exterior	3	1		Segunda mano/para reformar	Terraza	
Latina	99500	35	2	No	Interior	Bajo	1		Segunda mano/buen estado		
Ciudad Lineal	100000	52	1	No	Interior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	101000	63	1	Si	Exterior	2	1		Segunda mano/buen estado		Calefaccion individual: Gas natural
Latina	103500	64	3	No	Exterior		1		Segunda mano/buen estado		
Latina	104000	47	1	No	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	105000	37	1	No	Exterior	Semi-sótano	1		Segunda mano/buen estado		
Latina	105000	30	0	Si	Exterior	Bajo	1		Segunda mano/buen estado		
Latina	105400	60	3				1		Segunda mano/buen estado	Terraza	
Latina	106000	34	0	No	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	107000	79	2	No	Interior		1		Segunda mano/buen estado		
Latina	109000	40	1	Si	Exterior	2	1		Segunda mano/buen estado		No dispone de Calefaccion
Latina	109000	67	2	No			1		Segunda mano/buen estado		
Ciudad Lineal	110000	31	1	No	Interior	Bajo	1		Segunda mano/buen estado	Terraza	Calefaccion individual: Gas natural
Latina	110000	35	1	Si	Exterior	Bajo	1		Segunda mano/buen estado		Calefaccion individual: Eléctrica
Latina	111600	66	2	No	Exterior	4	1		Segunda mano/para reformar	Terraza	
Latina	111600	60	2	No	Exterior	4	1		Segunda mano/para reformar		
Latina	113000	56	2	No	Interior	Bajo	1		Segunda mano/para reformar		Calefaccion individual: Gas natural

Fuente: Elaboración propia a partir de los datos de Idealista

5.2. Categorización de las variables y tratamiento de valores perdidos

Antes de proceder a realizar cualquier análisis de los datos es preciso asegurar que no existen valores perdidos (NAs) y que el programa interpreta bien las variables, como numéricas (enteras o decimales) o categóricas. Con la función str() comprobamos que las variables numéricas son el precios, los m2, las habitaciones y los baños; mientras que las categóricas son “Distrito”, “Ascensor”, “Exterior/Interior”, “Planta”, “Garaje”, “Estado”, “Terraza/Balcón” y “Calefacción”. El único cambio que realizar es transformar las variables Baños y Habitaciones en números enteros, puesto que no admiten decimales.

Por otro lado, el fichero cuenta con valores perdidos[4], en concreto 3.417 (Ver Ilustración 8: Número de valores perdidos en cada una de las variables). Si analizamos los NAs, en cada una de las variables, podemos observar valores cercanos y superiores al 50% en las variables “Terraza/Balcón” y “Garaje” respectivamente

[4] Podemos comprobarlo con la función *sum(is.na(PISOS))* en RStudio.

Ilustración 8: Número de valores perdidos en cada una de las variables

```
> colSums(is.na(PISOS))
        Distrito           Precio               M2     Habitaciones         Ascensor 
               0                0                0               23               68 
Exterior/Interior           Planta            Banos           Garaje           Estado 
             113              131                2             1415                3 
   Terraza/Balcon      Calefaccion 
             917              739 
```

Fuente: Elaboración propia con RStudio

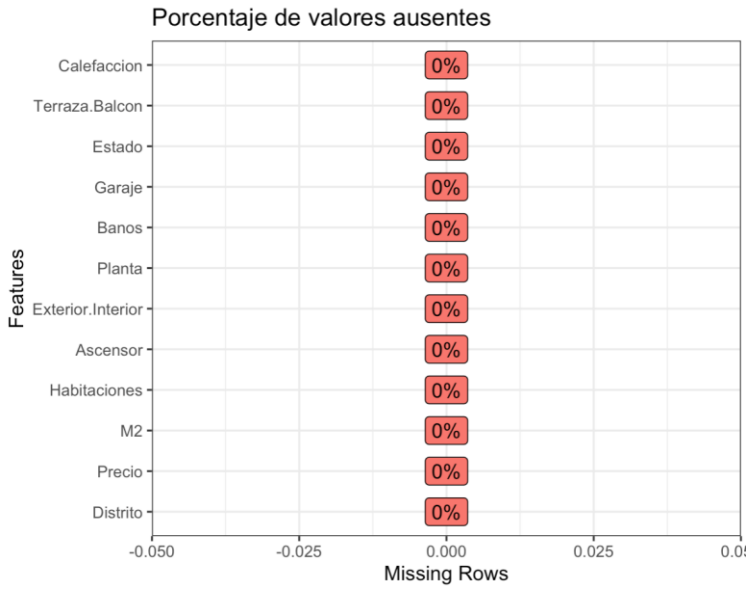
"Existen varias maneras de tratar con valores ausentes:

- Rellenar los valores con la media, mediana o el valor más frecuente de la variable.
- Completar los valores que faltan por el valor que esté directamente antes o después en la fila o columna.
- Completar todos los datos faltantes con 0, si se trata de valores numéricos. Esta opción es poco aconsejable ya que puedes modificar de manera significativa los resultados.
- Eliminar las filas que presenten valores ausentes, siempre y cuando el conjunto de datos sea lo suficientemente grande y no se pierde información relevante al eliminar esas filas.
- Y una forma abrupta de tratamiento que depende del contexto de análisis, es eliminar las variables que presentan un porcentaje mayor del 50% de datos ausentes" [Guía práctica de introducción al Análisis Exploratorio de Datos, 2021. p.13]

En el presente caso, lo más recomendable, en vista a futuros análisis como la correlación entre variables, parece sustituir los valores perdidos en las variables numéricas (habitaciones y baños) por su valor medio. Para ello, una vez calculado dicho valor medio (2 para baños y 3 para habitaciones), se reemplazan los perdidos por los valores medios. En cuanto a la variable Garaje, la cual cuenta con el mayor números de valores perdidos, se sustituirán estos por el carácter "No", ya que parece razonable que si en el anuncio de los pisos en Idealista no se habla de esta variable es porque las viviendas no disponen de garaje. Lo mismo haremos con la variable Terraza/Balcón y Ascensor, sustituyéndolos por "Ninguno" y "No" respectivamente. En la variable Estado, Exterior/Interior, Planta y Calefacción, los reemplazaremos por "Desconocido".

Una vez realizados los cambios, comprobamos que ya no existen valores ausentes:

Ilustración 9: Porcentaje de valores ausentes tras los cambios implementados



Fuente: Elaboración propia en RStudio

5.3. Análisis descriptivo

En este paso se aplicarán funciones de estadística descriptiva, para así elaborar un resumen de la información que se puede obtener a partir de los datos de una muestra. Todo ello acompañado de gráficos que puedan arrojar claridad.

La función summary nos permite obtener un resumen de las variables de las viviendas, su mínimo, máximo, media, mediana…(Ver Ilustración 10: Función summary en).

Ilustración 10: Función summary en RStudio

```
> summary(PISOS)
          Distrito       Precio             M2          Habitaciones  Ascensor  Exterior/Interior
 Ciudad Lineal:992   Min.   :  65000   Min.   : 22.00   Min.   :0.00   No: 696   Length:1920
 Latina       :928   1st Qu.: 165900   1st Qu.: 65.00   1st Qu.:2.00   Si:1224   Class :character
                     Median : 238976   Median : 80.00   Median :3.00             Mode  :character
                     Mean   : 320215   Mean   : 95.49   Mean   :2.67
                     3rd Qu.: 355000   3rd Qu.:105.25   3rd Qu.:3.00
                     Max.   :2800000   Max.   :609.00   Max.   :8.00

    Planta        Banos                        Garaje                              Estado
 1      :371   Min.   :1.000   No                  :1415   Desconocido                :   3
 2      :309   1st Qu.:1.000   Si                  : 474   Promoción de obra nueva    : 101
 Bajo   :276   Median :1.000   Garaje opc. 20.000 €:   8   Segunda mano/buen estado   :1455
 3      :266   Mean   :1.504   Garaje opc. 30.000 €:   7   Segunda mano/para reformar: 361
 4      :192   3rd Qu.:2.000   Garaje opc. 15.000 €:   3
 5      :135   Max.   :7.000   Garaje opc. 25.000 €:   3
 (Other):371                    (Other)             :  10
         Terraza/Balcon                          Calefaccion
 Balcon          :104    Desconocido                        :739
 Ninguno         :917    Calefaccion individual: Gas natural:479
 Terraza         :740    Calefaccion individual             :196
 Terraza y Balcon:159    Calefaccion individual: Eléctrica  :139
                         Calefaccion central                :116
                         No dispone de Calefaccion          :112
                         (Other)                            :139
```

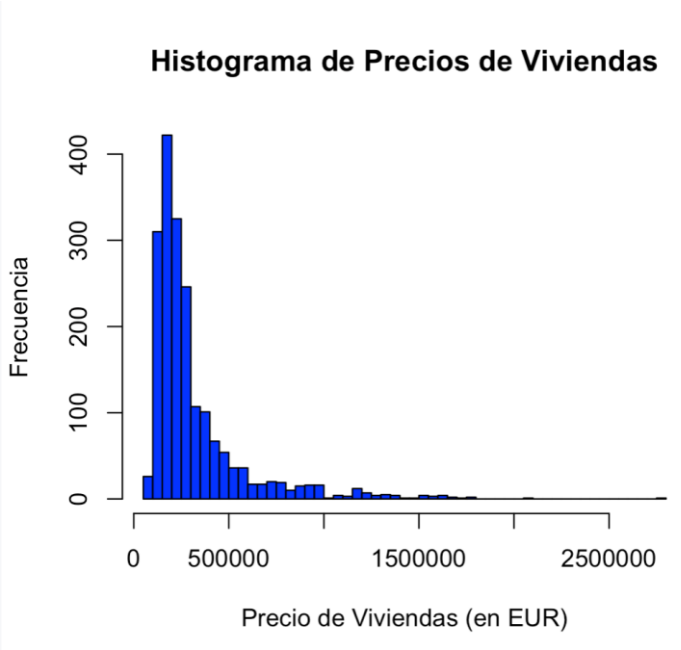
Fuente: Elaboración propia con RStudio

Con esta función podemos saber que el precio mínimo de las viviendas es 65.000, mientras que el precio máximo es 2.800.000 euros. También podemos conocer los m2 del piso más pequeño (22) y los del más grande (609). Otros datos interesantes los encontramos en la media, teniendo el piso medio un precio de 320.215 euros, 95,49 m2, 2,6 habitaciones y 1,5 baños.

Ahora, para examinar la distribución de las variables numéricas (Precio, M2, Habitaciones y Baños), y poder examinar gráficamente la información obtenida usaremos histogramas[5] (Ver Gráficos 5, 6, 7, y 8). En estos podemos observar asimetrías a la derecha. En el caso del precio, el de la mayoría de las viviendas ronda los 300.000 – 500.000 euros, observándose algunos outliers[6] con algunos precios muy bajos y precio superiores a los 2.500.000 euros. Los metros cuadrados de las viviendas se encuentran en torno a los 80-100 m2; el número de habitaciones parece ser mayoritariamente 3; y el número de baños también 3.

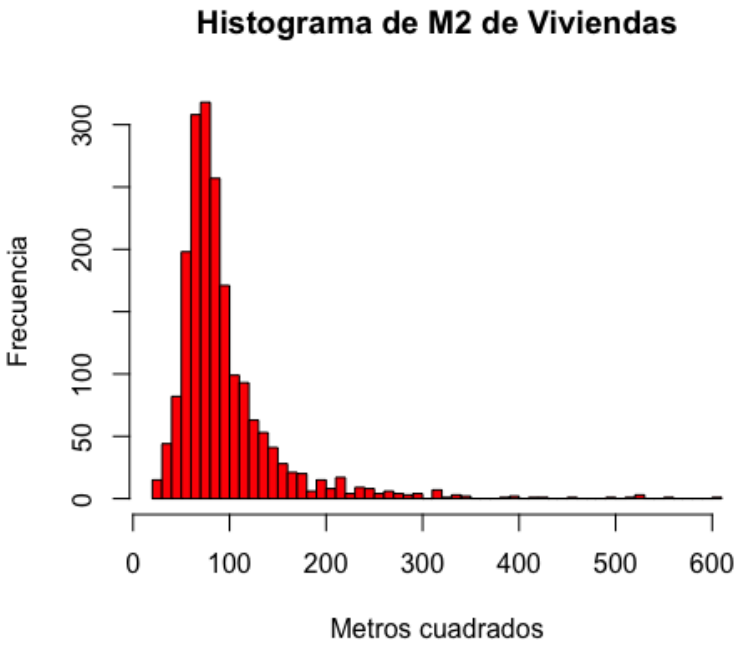
[5] Un histograma es un tipo de gráfico que representa la distribución de un conjunto de datos continuos. El objetivo de un histograma es mostrar la frecuencia con la que los datos caen dentro de ciertos intervalos.
[6] Valores atípicos, anormales dentro del conjunto de datos.

Gráfico 5: Histograma de Precios de Viviendas



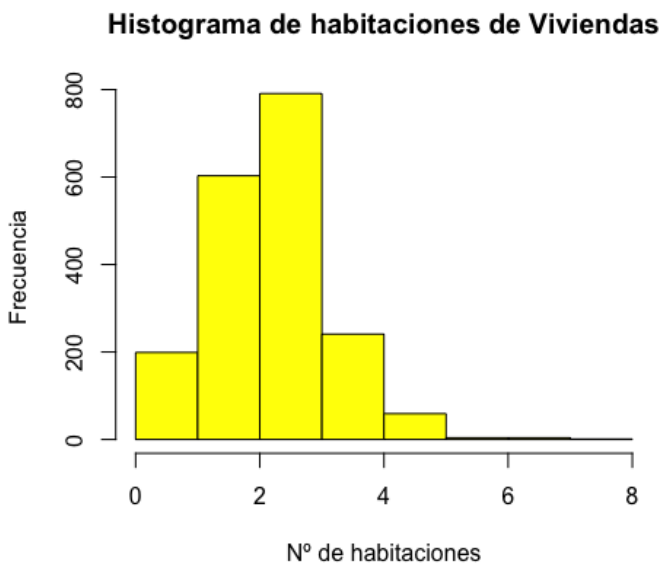
Fuente: Elaboración propia en RStudio

Gráfico 6: Histograma de los M2 de las viviendas



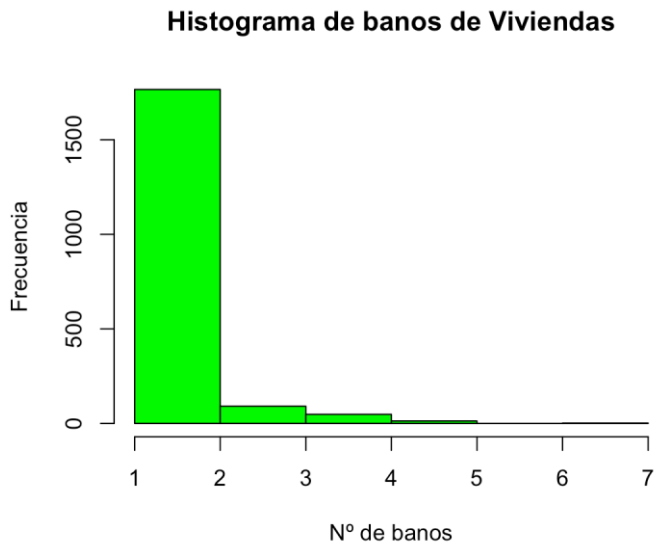
Fuente: Elaboración propia en RStudio

Gráfico 7: Histograma del número de habitaciones en las viviendas



Fuente: Elaboración propia en RStudio

Gráfico 8: Histograma del número de baños de las viviendas



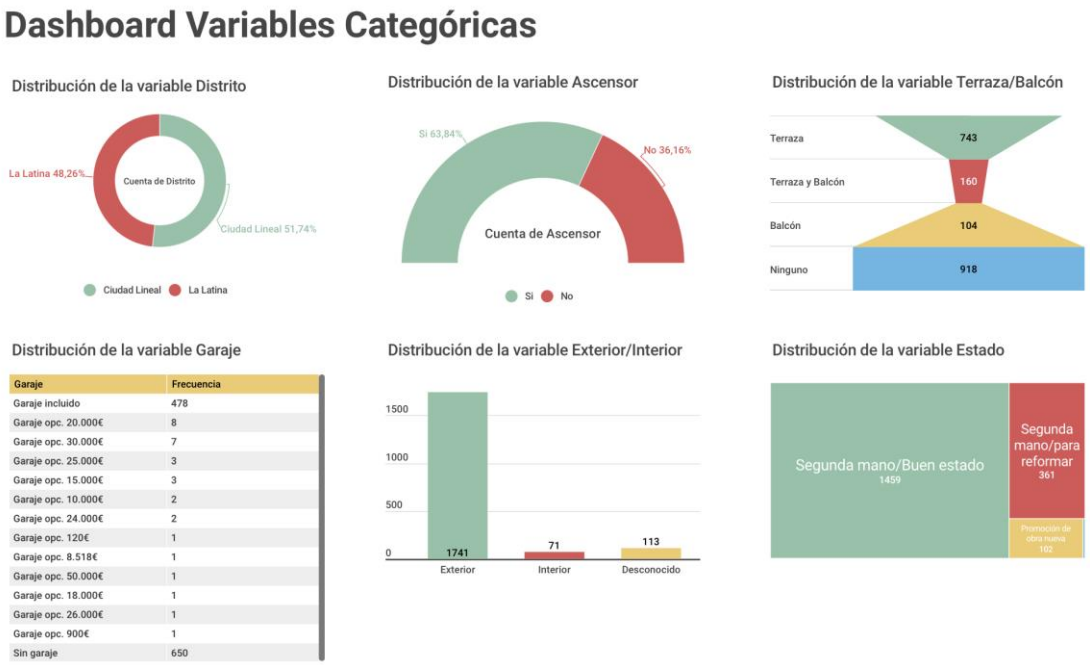
Fuente: Elaboración propia

Para el caso de las variables categóricas, esto es aquellas que hacen referencia a una cualidad o a una categoría (Distrito, Ascensor, Exterior/Interior, Planta, Garaje, Estado, Terraza/Balcón y Calefacción) se emplearán gráficos de barra o de tarta para examinar

la distribución a priori de dichas variables[7]. (Ver Ilustración 11: Dashboard de la distribución de las variables categóricas).

De esta manera, gracias a esta análisis podemos conocer que existe paridad (48% frente a 52%) en la distribución de las viviendas en los distritos objeto de estudio, lo cual hará que las conclusiones extraídas puedan ser aplicables a todas las viviendas, ya que no existe dominio de un distrito sobre otro. Por otro lado, observamos que la mayoría de las viviendas cuentan con ascensor, son exteriores y están en buen estado. En la variable garaje observamos gran variabilidad. En la variable Terraza/Balcón, se observa una predominancia de las viviendas con terraza.

Ilustración 11: Dashboard de la distribución de las variables categóricas



Fuente: Elaboración propia

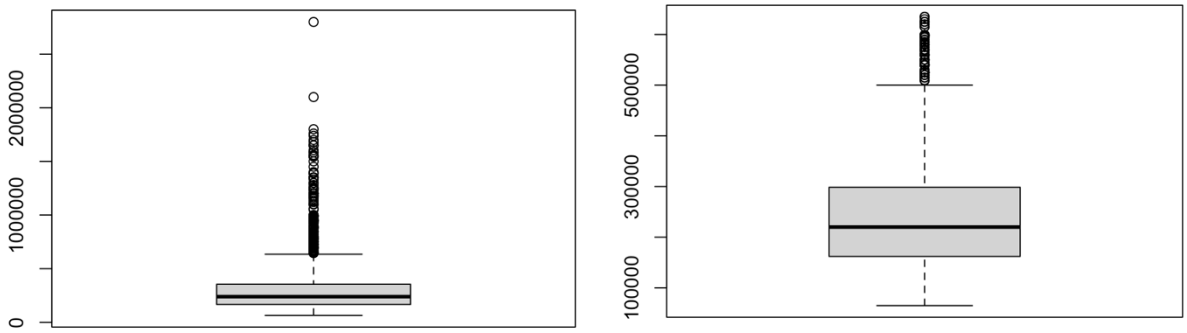
5.4. Identificación de valores atípicos

En el presente caso hemos podido observar, gracias a los histogramas, que existen outliers, los cuales pueden llegar a modificar los resultados, por lo que es preciso detectarlos para evaluar su influencia en el conjunto de datos y en su caso, eliminarlos. En este caso, usaremos el rango intercuartílico (IQR) basándonos en el precio para eliminar los valores atípicos. El rango intercuartílico es la diferencia entre el percentil

[7] Las variables Planta y Calefacción no han sido representadas visualmente puesto que no ofrecían claridad.

75 y el percentil 25 del conjunto de datos. Los valores que están a más de 1,5 veces el IQR por encima del percentil 75 o por debajo del percentil 25 se consideran valores atípicos y por lo tanto serán eliminados. [Guía práctica de introducción al Análisis Exploratorio de Datos, 2021. p.16]. Los resultados de eliminar los valores atípicos pueden observarse en el siguiente gráfico de bigotes:

Gráfico 9: Gráfico de Bigotes antes y después de eliminar los valores atípicos



Fuente: Elaboración propia

5.5. Correlación de variables

La correlación es una medida estadística que mide la relación lineal entre dos variables, es decir, la tasa constante a la que cambian juntas. Al interpretar este coeficiente, es importante tener en cuenta los siguientes puntos:

- A medida que el coeficiente de correlación se acerca a 0, la relación lineal se debilita.
- Los coeficientes positivos indican una correlación positiva, donde los valores de ambas variables tienden a aumentar juntos.
- Los coeficientes negativos indican una correlación negativa, donde los valores de una variable aumentan mientras que los valores de la otra disminuyen.

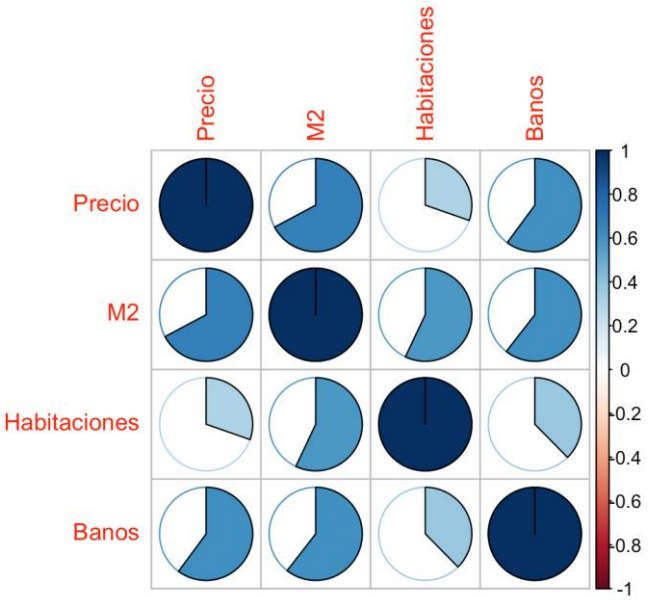
Sin embargo, es importante tener en cuenta que solo porque dos variables estén correlacionadas, no significa que haya una relación causal. Puede haber correlaciones aleatorias o ficticias. Por ejemplo, el gasto estadounidense en ciencia, espacio y tecnología se correlaciona al 99.79% con los suicidios por ahorcamiento, estrangulamiento y asfixia, lo cual no obedece a un relación de causa-efecto. [Vigen, s.f.].

En el caso de que existiera alta correlación entre varias variables podría eliminarse una de ellas, para bien reducir la redundancia o bien mejorar la interpretación y rendimiento del modelo. En el presente caso (Ver Gráfico 10: Mapa de correlaciones entre las variables numéricas), donde todas las correlaciones son positivas, podemos extraer las siguientes conclusiones:

- Existe una alta correlación (0.67) entre las variables precio y m2, lo cual puede considerarse lógico, puesto que a más m2 mayor será el precio de la vivienda.
- La correlación también es significativa (0.60) entre los m2 y el número de baños. Sin embargo, el número de habitaciones guarda una correlación menor (0.56) con los m2 de las viviendas, lo cual puede indicarnos que a medida que las viviendas aumentan de tamaño lo que tiende a incrementarse es el número de baños y no tanto de habitaciones.
- Parece ser que el número de baños de una vivienda influye de manera notable en el precio de las viviendas, pues la correlación es de (0.60).

Analizados los resultados, no se procederá a eliminar ninguna variable, ya que a pesar de tener correlaciones altas, podría tener consecuencias impredecibles en los resultados, además de que la base de datos no cuenta con excesivas variables y por lo tanto no es preciso simplificarla.

Gráfico 10: Mapa de correlaciones entre las variables numéricas



Fuente: Elaboración propia en RStudio

6. MODELOS DE MACHINE LEARNING

En esta sección del trabajo se presentan los modelos de Machine Learning utilizados para el análisis del precio de la vivienda en los distritos de Ciudad Lineal y La Latina. Se explicará tanto el funcionamiento de los modelos, como la metodología y los resultados obtenidos, comparando su precisión y eficacia en la predicción del precio de la vivienda.

Tanto el código como el data set empleado para la elaboración de los modelos de Machine Learning está disponible en GitHub: https://github.com/AnaBrunoCueto/Analisis-precio-vivienda-Ciudad-Lineal-y-La-Latina-con-Machine-Learning

6.1. Regresión lineal múltiple

La regresión lineal consiste en especificar la relación entre una única variable dependiente numérica (Y, valor que se desea predecir, *target*) y una o más variables independientes numéricas o cualitativas (X, predictores, *features*). Este modelo se basa en la suposición de que la relación entre las variables es lineal, sigue una línea recta. Si sólo hay una variable independiente, se habla de regresión lineal simple; en caso contrario, se habla de regresión múltiple. Ambos modelos suponen que la variable dependiente es continua. [Lantz, 2013. p.160]

"Los modelos lineales múltiples siguen la siguiente ecuación:

Ecuación 1: Regresión lineal múltiple

$$Y_1 = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_n X_n + \epsilon_i$$

- β_0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- β_1: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose constantes el resto de las variables. Se conocen como coeficientes parciales de regresión.
- ϵ_i: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo". [Amat Rodrigo, 2016)]

Un modelo explicativo de regresión lineal se utiliza para entender la relación entre la variable dependiente y las variables independientes. En este tipo de modelo, el objetivo es identificar qué variables independientes tienen un impacto significativo en la variable dependiente y cómo se relacionan entre sí. En nuestro caso, nos ayudará a comprender cuanto afecta cada variable al precio de la vivienda.

Por otro lado, un modelo predictivo de regresión lineal se utiliza para hacer predicciones sobre el valor de la variable dependiente a partir de los valores de las variables independientes. En este tipo de modelo, el objetivo es obtener la mejor predicción posible para un conjunto de datos nuevos y desconocidos. Los resultados del modelo se utilizan para hacer predicciones y tomar decisiones basadas en esas predicciones. En el presente caso nos ayudará a estimar el precio de las viviendas.

En definitiva, mientras que un modelo explicativo de regresión lineal se enfoca en entender la relación entre las variables, un modelo predictivo está dirigido a hacer predicciones lo más precisas posibles.

6.1.1. Regresión lineal múltiple explicativa

El objetivo de este apartado, será elaborar un modelo de regresión explicativo[8], para entender la relación entre las variables independientes y la variable dependiente, el precio. Además nos va a permitir identificar qué variables independientes tienen un impacto significativo en el precio y como se relacionan entre sí. El foco va a estar en los betas, que van a medir como afecta la variación de las variables independientes en el precio.

Es preciso remarcar que los modelos de regresión lineal asumen una distribución normal para la variable dependiente, lo cual en el presente caso no se da, como ya hemos podido observar con el histograma de la distribución de la variable precio (Ver Gráfico 5: Histograma de Precios de Viviendas). Dado que la media es mayor que la mediana, la distribución es asimétrica a la derecha, por lo que a pesar de que estas asunciones de

[8] Elaborado e interpretado con el código y la información disponible en: Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

normalidad en la mayoría de los casos se incumplen, es preciso tenerlo en cuenta para correcciones futuras.

Otro problema es que los modelos de regresión lineal requieren en puridad, que las variables sean numéricas, sin embargo tenemos varias variables factor en nuestro data set (Ver Ilustración 12: Resumen de nuestro data set). Se explicará a continuación como deben interpretarse las mismas.

Ilustración 12: Resumen de nuestro data set

```
> str(PISOS_filtrado)
tibble [1,740 × 12] (S3: tbl_df/tbl/data.frame)
 $ Distrito         : Factor w/ 2 levels "Ciudad Lineal",..: 1 1 2 2 2 1 1 2 1 2 ...
 $ Precio           : num [1:1740] 65000 77900 79000 79000 79000 79600 80500 83000 84800 85000 ...
 $ M2               : num [1:1740] 50 30 25 25 25 69 68 30 51 62 ...
 $ Habitaciones     : int [1:1740] 2 1 3 0 0 3 2 1 2 4 ...
 $ Ascensor         : Factor w/ 2 levels "No","Si": 1 1 2 2 2 1 2 1 2 1 ...
 $ Exterior/Interior: chr [1:1740] "Exterior" "Exterior" "Exterior" "Exterior" ...
 $ Planta           : Factor w/ 23 levels "-1","-2","1",..: 22 19 19 19 19 3 19 23 19 12 ...
 $ Banos            : int [1:1740] 2 1 1 1 1 1 1 1 1 1 ...
 $ Garaje           : Factor w/ 14 levels "Garaje opc. 10.000€",..: 13 13 13 13 13 13 13 13 13 13 ...
 $ Estado           : Factor w/ 4 levels "Desconocido",..: 3 3 3 3 3 3 3 3 4 4 ...
 $ Terraza/Balcon   : Factor w/ 4 levels "Balcon","Ninguno",..: 2 2 2 2 2 3 2 2 2 2 ...
 $ Calefaccion      : Factor w/ 11 levels "Calefaccion central",..: 8 11 10 10 10 7 7 7 10 10 ...
```

Fuente: Elaboración propia en RStudio

Además, antes de construir el modelo va a ser preciso tener en cuenta que hay variables muy correlacionadas (Ver Gráfico 10: Mapa de correlaciones entre las variables numéricas), como son el precio y los metros cuadrados, o el precio y los baños. La multicolinealidad es la correlación alta entre más de dos variables explicativas y puede llegar a ser un problema, ya que puede resultar difícil captar la influencia de los factores si están muy correlacionados. Es idílico pensar en casos donde las variables no estén correlacionadas, por lo que para ser un verdadero problema, la correlación debe ser fuerte, mayor de 0.90, lo cual no ocurre.

Una vez elaboramos el modelo de regresión lineal con todas las variables (*fullmodel*), obtenemos los valores de beta, que nos van a permitir cuantificar monetariamente las variables. Son los siguientes:

Tabla 1: Betas Modelo Explicativo

BETAS MODELO EXPLICATIVO

VARIABLE	BETAS
(Intercept)	33694,63
DistritoLatina	-51404,76
M2	1828,90
Habitaciones	-7029,20
AscensorSi	32543,66
`Exterior/Interior`Exterior	-2336,13
`Exterior/Interior`Interior	-9451,76
Planta-2	-17912,47
Planta1	36961,61
Planta10	46198,94
Planta11	12846,70
Planta12	47789,61
Planta13	78235,35
Planta14	78596,94
Planta2	40645,55
Planta3	42238,83
Planta4	44969,94
Planta5	49171,50
Planta6	62497,27
Planta7	75115,56
Planta8	36434,12

Planta9	35055,37
PlantaBajo	10430,89
PlantaDesconocido	45066,05
PlantaEntreplanta	13629,17
PlantaSemi-sótano	-11423,61
PlantaSotano	-149734,48
Banos	45936,55
GarajeGaraje opc. 12.000€	42378,86
GarajeGaraje opc. 15.000 €	105945,48
GarajeGaraje opc. 18.000 €	71638,05
GarajeGaraje opc. 19.000 €	78895,74
GarajeGaraje opc. 20.000 €	95421,91
GarajeGaraje opc. 24.000 €	159246,60
GarajeGaraje opc. 25.000 €	138423,12
GarajeGaraje opc. 26.000 €	41735,02
GarajeGaraje opc. 30.000 €	160476,75
GarajeGaraje opc. 8.518 €	23065,99
GarajeNo	44354,16
GarajeSi	103285,32
EstadoPromoción de obra nueva	-20432,09
EstadoSegunda mano/buen estado	-37928,34
EstadoSegunda mano/para reformar	-57428,84
`Terraza/Balcon`Ninguno	-6949,34
`Terraza/Balcon`Terraza	-308,88
`Terraza/Balcon`Terraza y Balcon	3604,96
CalefaccionCalefaccion central: Gas	1354,59
CalefaccionCalefaccion central: Gas natural	-54578,30

CalefaccionCalefaccion central: Gasoil	-972,90
CalefaccionCalefaccion individual	-26314,63
CalefaccionCalefaccion individual: Bomba de frío/calor	-19290,78
CalefaccionCalefaccion individual: Eléctrica	-31586,41
CalefaccionCalefaccion individual: Gas natural	-19006,38
CalefaccionCalefaccion individual: Gas propano/butano	53362,60
CalefaccionDesconocido	-19583,26
CalefaccionNo dispone de Calefaccion	-25185,75

Fuente: Elaboración propia en Infogram

Los coeficientes de beta reflejan el aumento del precio de la vivienda por un incremento de "uno" en cada una de las variables, manteniéndose las demás constantes. Así por ejemplo, cada M2 adicional, cabe esperar un aumento medio del precio de 1.828,90 euros, suponiendo que todo lo demás se mantiene igual. De la misma manera, cada baño adicional produce un incremento medio de 45.936,55 euros. En el caso de las variables factor, R toma uno de los niveles/categorías como referencia, y por ello no aparecen en la estimación. Las betas se interpretarán en relación a la variable referencia. Dicho esto, puede decirse, que según el modelo, los pisos del Distrito de la Latina cuestan 51.404,76 euros menos que los de Ciudad Lineal, de la misma manera que los pisos con ascensor cuestan 32.543,66 euros más que los que no lo tienen.

Por otro lado, y una vez conocidas las interpretaciones de las betas, es preciso comprobar como de bien el modelo está evaluando los datos (función *summary*). El resultado, proporciona tres formas de evaluar el rendimiento o ajuste de nuestro modelo:

- Residuals (errores). Proporciona el número de errores de nuestras predicciones. En nuestro caso, el error máximo es de 338.877, lo que indica que el modelo subestimó el precio en casi 340.000 euros al menos en una observación. Por otro lado, el 50% de los errores se sitúan dentro de los valores 1Q y 3Q (el primer y tercer cuartil), por lo que la mayoría de las predicciones se situaron entre 34.493 euros por encima del valor real y 25.375 euros por debajo del valor real. Pese a que el tamaño de los errores es bastante alto, no es sorprendente debido a la naturaleza del precio de la vivienda.

- El p-valor. En R el p-valor se encuentra representado con asteriscos (*), que indican niveles de significación[9]. El p-valor debe que ser muy pequeño para que haya evidencias de que los resultados no se deben al azar y las estimaciones son reales. Es común utilizar un nivel de significación de 0,05 para indicar cuando una variable es estadísticamente significativa. En el presente caso, las variables más significativas, aquellas con tres asteriscos son: Distrito, M2, AscensorSi, Baños y Calefacción Individual Eléctrica.
- R cuadrado. El R cuadrado hace referencia a la proporción de varianza de la Y (precios) explicada por el modelo. Lo ideal es que es sea cercana 1, esto es cuanto mayor, mejor bondad de ajuste. En nuestro caso es 0,7141, lo que significa que casi el 72% de la variabilidad de los precios viene explicada por el modelo. No obstante, es preciso tener en cuenta que el R cuadrado será mayor cuantas más variables tengamos, pues los modelos con más características siempre explican más variación. Por ello, para una interpretación más precisa, hay que fijarse en el R-cuadrado ajustado, que corrige el R-cuadrado penalizando los modelos con un gran número de variables independientes, como el del presente caso. Este valor es 0,7046 por lo que aún con tantas variables, la predicción del modelo sigue siendo bastante buena.

De esta manera, gracias a este modelo, hemos podido cuantificar el efecto de cada una de las variables sobre el precio de la vivienda y estimar cuales son las más significativas.

6.1.1. Regresión lineal múltiple predictiva

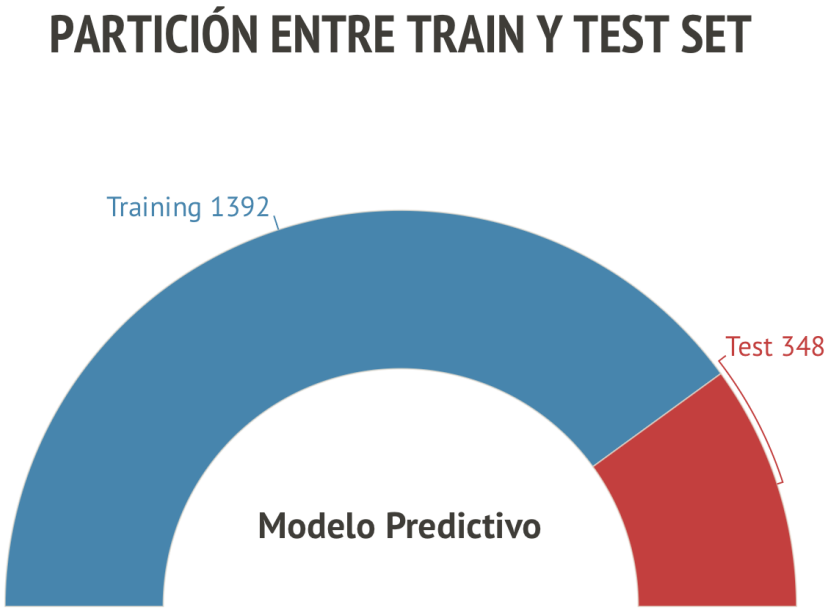
El objetivo de este apartado será construir un modelo de regresión predictivo[10], que nos permita predecir el valor de la variable dependiente (precio) a partir de las variables independientes (m2, habitaciones, baños, distrito, ascensor, etc.). En este caso, el foco no estará en los betas (estimaciones), si no en las predicciones, en obtener las mejores posibles, esto es las que tengan el mínimo error posible.

[9] "El nivel de significación proporciona una medida de la probabilidad de que el coeficiente verdadero sea cero dado el valor de la estimación". [Lantz, 2013, p. 183]

[10] Elaborado e interpretado con el código y la información disponible en: Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

Para lograr estos objetivos es preciso realizar, a diferencia de en los modelos explicativos, una partición en los datos. El conjunto de entrenamiento (*Training Set*) se usará para estimar el modelo, mientras que el conjunto de test (*Test Set*) para poder comparar los valores predichos con los reales y así cuantificar el error del modelo. En nuestro modelo la partición será de 80% y 20% respectivamente, quedando las observaciones distribuidas de la siguiente manera:

Ilustración 13: Ejemplo de la partición entre el Training y el Test Set

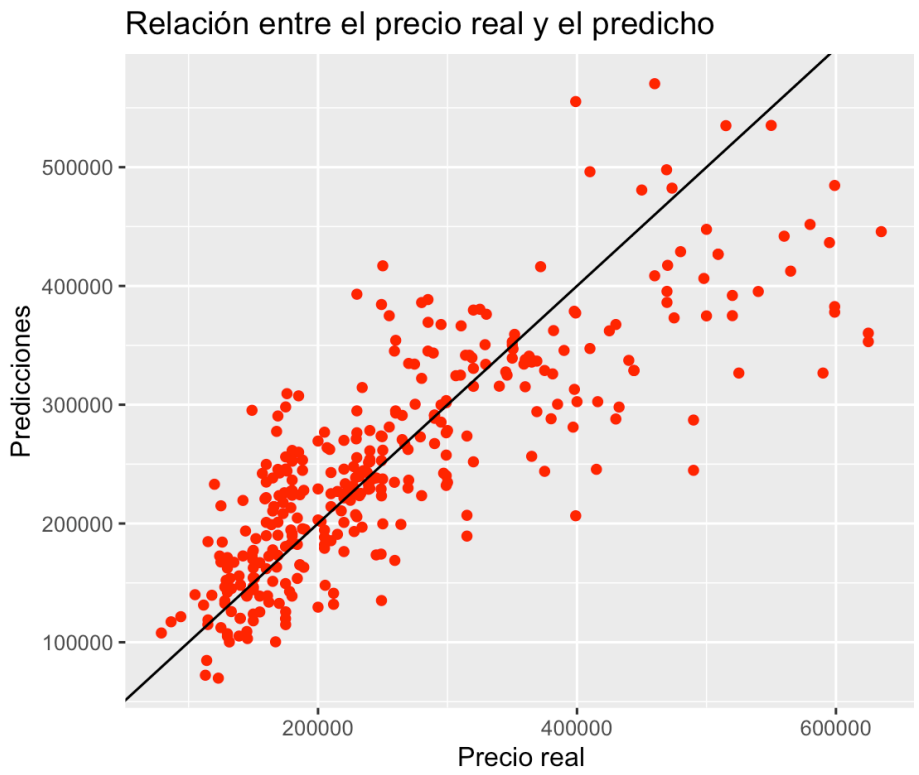


Fuente: Elaboración propia con Infogram

Antes de elaborar el modelo, es precio mencionar que en este caso, no podremos usar todas las variables (*fullmodel*). Al realizar la partición, la variables categóricas que tienen muchos niveles dejan observaciones fuera del modelo de entrenamiento y dentro del test o viceversa. Así, al obtener las predicciones en el Test Set se obtienen errores en RStudio, puesto que se incorporan nuevos niveles que no habían sido tenidos en cuenta previamente. Por ello, en este caso elaboraremos el modelo dejando fuera las variables Planta, Garaje y Calefacción.

Una vez elaborado el modelo, realizadas las predicciones en el Test Set y cuantificados los errores, obtenemos el siguiente gráfico de dispersión:

Gráfico 11: Gráfico de dispersión entre el precio real y el predicho



Fuente: Elaboración propia en RStudio

Además de otras métricas que comentaremos más adelante, este gráfico de dispersión puede ser muy útil para evaluar la precisión de un modelo de regresión en la predicción de los precios de las viviendas. Dado que los puntos del gráfico están situados cerca de la línea diagonal, esto sugiere que el modelo es preciso y se ajusta bien a los datos. Ahora bien, también hay que tener en cuenta los errores, que son todos aquellos puntos que se alejan de la diagonal. En concreto, el error cuadrático medio de nuestro modelo (RSME) es de 69.527,03 euros, lo que significa que las predicciones están en media a 70.000 euros de sus valores reales. A pesar de ser alto, teniendo en cuenta que la desviación típica de los precios es 116.319,9, el modelo está haciendo predicciones notablemente precisas.

Una vez analizado el modelo, existen varios métodos para mejorar su rendimiento:

- El primero de ellos es añadir relaciones no-lineales, ya que los modelos de regresión lineal asumen relaciones lineales entre las variables dependientes y la independiente. En el presente caso, el efecto de los M2 puede no ser constante, ya que los compradores pueden estar dispuestos a pagar más por una vivienda más grande, pero luego el aumento en el precio puede ser menor o incluso

disminuir. Para tener en cuenta esta relación no lineal se elevará la variable M2 al cuadrado en el nuevo modelo. Lo mismo se hará con los baños.

- La segunda de las opciones es convertir variables numéricas en binarias cuando pensemos que el efecto de una variable numérica no es acumulativo, sino que sólo tiene efecto una vez alcanzado un umbral específico. En el presente caso no puede extraerse esta conclusión de ninguna de las variables numéricas.
- La última opción es añadir el interacciones entre variables. Por ejemplo, los M2 y el número de baños tienden separadamente a encarecer las viviendas, pero es lógico pensar que unidas tendrán un efecto aún mayor sobre el precio. En el caso de las viviendas, podrían añadirse interacciones entre muchas variables, por lo que optaremos por crear un interacción entre los metros cuadrados y el distrito, y los baños y los metros cuadrados[11].

Así, teniendo en cuenta estos cambios elaboramos nuevamente el modelo, y comparándolo con el modelo inicial, obtenemos los siguientes resultados:

Tabla 2: Resultados Modelo 1 vs. Modelo 2

	Modelo 1	Modelo 2
R2	0,646	0,687
R2 Ajustado	0,642	0,684
RMSE Test Set	69.527,03	66.390,03

Fuente: Elaboración propia

En relación con nuestro primer modelo, el valor R-cuadrado ha mejorado de 0,64 a 0,68 y el RMSE se ha reducido considerablemente. El modelo explica ahora casi el 69% de la variación del precio de la vivienda. También, puede comprobarse como algunos de los ajustes que hemos implementado se confirman. En concreto las relaciones no lineales implementadas en las variables M2 y Baños, son estadísticamente significativas, al igual que las interacciones. Por ejemplo, la interacción entre los M2 y el Distrito tiene importantes efectos. Puede afirmarse pues que además del aumento de

[11] También se añadió esta relación no lineal en la variable Habitaciones, pero se comprobó que no mejoraba las predicciones del modelo, por lo que se excluyó.

3.753,662 euros por metro adicional, las viviendas que se encuentran el distrito de Ciudad Lineal son 795,924 euros más caras.

6.2. Árboles de regresión y Árboles modelo

Los árboles de regresión y los árboles modelos son técnicas de aprendizaje automático que se utilizan para modelar relaciones no lineales entre las variables de entrada y salida. Ambos métodos se basan en la construcción de un árbol de decisiones, que representa un conjunto de reglas para la predicción de la variable de salida. Estos árboles de decisión pueden ser de regresión, en los cuales la variables Y es cuantitativa, y de clasificación, en los que la variable Y es cualitativa.

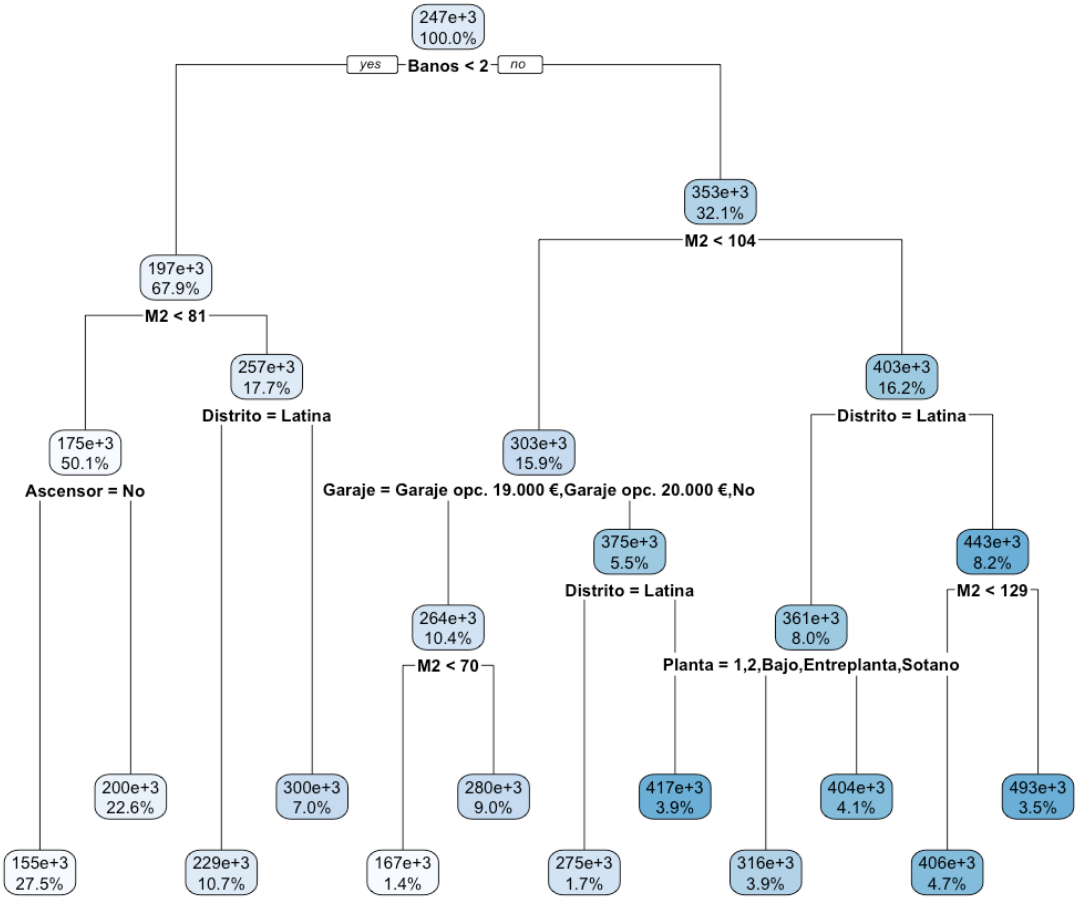
Así, el objetivo de este apartado será elaborar construir un árbol de regresión[12], esto es un modelo muy parecido a un diagrama de flujo en el que los nodos de decisión, los nodos hoja y las ramas definen una serie de decisiones que pueden utilizarse para clasificar ejemplos. A pesar de llamarse árboles de regresión, no se utilizan métodos de regresión lineal, sino que se van a realizar predicciones basadas en el valor medio de las observaciones que llegan a cada hoja. Además, como fue mencionado, los modelos de regresión lineal, asumen una distribución normal de los datos, lo cual en la vida real no siempre se da. Así, los árboles de regresión, son más adecuados para modelos con muchas características o muchas relaciones complejas y no lineales entre las variables y el resultado. [Lantz, 2013. pp.187-188]

Este tipo de modelo será útil tanto para predecir el precio de una vivienda en función de las preferencias de los compradores o inversores y así determinar un valor de mercado para la misma, como para identificar las características que más influyen en el precio de las viviendas y conocer qué tipos de propiedades son más valoradas por los compradores.

Dicho esto, una vez dividido el conjunto de datos en training y test set (80% - 20%) , elaboramos el modelo, obteniendo el siguiente resultado:

[12] Elaborado e interpretado con el código y la información disponible en: Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

Ilustración 14: Árbol de Regresión



Fuente: Elaboración propia en RStudio

Es preciso aclarar el funcionamiento de este árbol de regresión. Empezando por el nodo principal, los datos se van dividiendo según la variable o característica que produzca el mayor aumento de homogeneidad en el resultado después de la división. Para los árboles de decisión numéricos, la homogeneidad puede medirse a través de estadísticos como la varianza, la desviación típica o la desviación absoluta de la media. Un criterio común para realizar estas divisiones por ramas, es la reducción de la desviación típica (SDR), que mide la reducción de la desviación típica del valor original a la desviación típica ponderada tras la división. [Lantz, 2013. p.189]. Por otro lado, el Yes/No que aparece al inicio del árbol hace referencia a la dirección en la que debemos movernos para avanzar en la ramas. Si la vivienda tiene menos de dos baños, avanzaremos por las ramas de la izquierda; si no tiene menos de dos baños, por las de la derecha, y así sucesivamente. En cuanto a los cuadrados, el primer número que aparece es la

predicción del precio de la vivienda, mientras que el segundo hace referencia al porcentaje del conjunto de datos que representa esa predicción.

Con esto, podemos extraer alguna conclusión inicial. Dado que el número de baños se utilizó en primer lugar en el árbol, es el factor de predicción más importante del precio de la vivienda, seguido por los M2 y el Distrito. Ahora bien, una vez comparados los precios predichos con los reales, podemos observar que las predicciones no son del todo precisas (Ver Ilustración 15: Comparación ente los precios predichos y los reales). Pese a que entre el primer y tercer cuartil las predicciones se acercan a la realidad, el modelo no está identificando correctamente los precios más baratos y los más caros. Otra forma de evaluar el rendimiento del modelo es considerar, como hemos hecho anteriormente, el error cuadrático medio de la raíz (RSME), medida comúnmente utilizada en los modelos de regresión. En este caso, este valor es de 72.354,11, un valor bastante superior a los modelos de regresión múltiple predicativos (Ver Tabla 2: Resultados Modelo 1 vs. Modelo 2).

Ilustración 15: Comparación ente los precios predichos y los reales

```
> summary(p.rpart)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 154621  154621  200047  246708  299806  493032
> summary(PISOS_filtrado$Precio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65000  162000  220000  249176  298250  635000
```

Fuente: Elaboración propia

Por ello, y de la misma manera que en la regresión predictiva, existen varias formas de mejorar el modelo, siendo una de ellas la construcción de árboles modelo. Estos funcionan de forma similar a los árboles de decisión, pero en cada hoja, se construye un modelo de regresión lineal múltiple. Esto suele dar lugar a resultados más precisos que los árboles de regresión, que utilizan un único valor para la predicción en los nodos hoja.

Una vez elaborado el modelo, observamos que en este caso el factor de predicción más importante es el metro cuadrado, seguido del número de baños. Ahora bien, los nodos no terminan en este caso en una predicción del precio de la vivienda, sino en un modelo de regresión múltiple. Por ejemplo, el modelo de regresión lineal 1 se muestra a

continuación, y tiene la misma interpretación que los modelos de regresión múltiple construidos anteriormente. Así, por ejemplo, cada metro cuadrado adicional, se espera que aumente el precio de la vivienda en 1.561,13 euros.

Ilustración 16: Modelo de regresión lineal 1 del árbol

```
LM num: 1
Precio =
        35103.6778 * Distrito=Ciudad Lineal
        + 1561.1339 * M2
        - 7026.6018 * Habitaciones
        + 22025.8891 * Ascensor=Si
        + 324.6457 * Exterior/Interior=Desconocido,Exterior
        + 8775.7211 * Exterior/Interior=Exterior
        + 247.8719 * Planta=Entreplanta,12,3,2,Desconocido,16,15,-2,4,1,5,8,10,11,9,6,7,14,13
        + 17739.247 * Planta=12,3,2,Desconocido,16,15,-2,4,1,5,8,10,11,9,6,7,14,13
        - 12838.4258 * Planta=Desconocido,16,15,-2,4,1,5,8,10,11,9,6,7,14,13
        + 9945.2197 * Planta=16,15,-2,4,1,5,8,10,11,9,6,7,14,13
        - 204.3844 * Planta=1,5,8,10,11,9,6,7,14,13
        + 9470.2307 * Planta=5,8,10,11,9,6,7,14,13
        + 36401.7906 * Planta=10,11,9,6,7,14,13
        - 60292.753 * Planta=11,9,6,7,14,13
        + 26170.5591 * Planta=6,7,14,13
        + 33201.8619 * Banos
        + 36249.371 * Garaje=Garaje opc. 50.000 €,Garaje opc. 24.000 €,Garaje opc. 18.000 €,Garaje opc. 26.000 €,Garaje opc. 19.000 €,Garaje opc. 15.000 €,Garaje opc. 20.000 €,G
araje opc. 30.000 €,Si,Garaje opc. 25.000 €
        + 1291.0339 * Garaje=Garaje opc. 19.000 €,Garaje opc. 15.000 €,Garaje opc. 20.000 €,Garaje opc. 30.000 €,Si,Garaje opc. 25.000 €
        + 13529.6438 * Estado=Segunda mano/buen estado,Promoción de obra nueva,Desconocido
        + 32851.2526 * Estado=Promoción de obra nueva,Desconocido
        + 211.366 * Terraza/Balcon=Balcon,Terraza,Terraza y Balcon
        + 5362.9549 * Calefaccion=Calefaccion individual: Bomba de frío/calor,Calefaccion individual,Calefaccion individual: Gas natural,Calefaccion central: Gas natural,Descono
cido,Calefaccion central: Gas,Calefaccion central: Gasoil,Calefaccion central,Calefaccion individual: Gas propano/butano
        - 249.6347 * Calefaccion=Calefaccion individual,Calefaccion individual: Gas natural,Calefaccion central: Gas natural,Desconocido,Calefaccion central: Gas,Calefaccion cen
tral: Gasoil,Calefaccion central,Calefaccion individual: Gas propano/butano
        + 186.9633 * Calefaccion=Calefaccion individual: Gas natural,Calefaccion central: Gas natural,Desconocido,Calefaccion central: Gas,Calefaccion central: Gasoil,Calefaccio
n central,Calefaccion individual: Gas propano/butano
        + 5028.0203 * Calefaccion=Calefaccion central: Gas natural,Desconocido,Calefaccion central: Gas,Calefaccion central: Gasoil,Calefaccion central,Calefaccion individual: G
as propano/butano
        + 493.5747 * Calefaccion=Calefaccion central: Gas,Calefaccion central: Gasoil,Calefaccion central,Calefaccion individual: Gas propano/butano
        + 35525.5028 * Calefaccion=Calefaccion central,Calefaccion individual: Gas propano/butano
        - 9392.78
```

Fuente: Elaboración propia en RStudio

Ahora bien, en este caso el modelo parece estar evaluando mejor los valores extremos, las viviendas más baratas y caras. De la misma manera, el error cuadrático medio se ha reducido considerablemente hasta 67.142,65. Por todo ello, comprobamos como el árbol modelo mejora la predicción del árbol de regresión.

Ilustración 17: Comparación entre los precios predichos y los reales del modelo mejorado

```
> summary(p.m5p)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  96742  169595  223019  247818  301076  582910
> summary(PISOS_filtrado$Precio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65000  162000  220000  249176  298250  635000
```

Fuente: Elaboración propia en RStudio

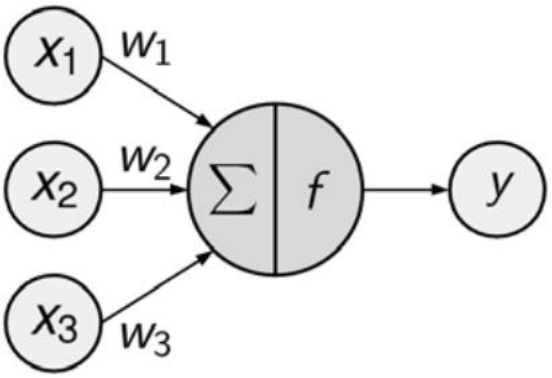
6.3. Redes neuronales

Las redes neuronales son un tipo de modelo de Machine Learning, subconjunto del aprendizaje automático y núcleo de los algoritmos de *deep learning*, inspirado en el funcionamiento del cerebro humano. Estas redes se basan en la interconexión de nodos

artificiales, llamados neuronas, que procesan y transmiten información a través de un conjunto de capas.

Las redes neuronales artificiales (RNA) se componen de una capa de nodos, que contiene una capa de entrada (x), una o más capas ocultas y una capa de salida (y). Cada nodo, o neurona artificial, se conecta a otro y tiene un peso (w), los cuales ayudan a determinar la importancia de las variables. A continuación, todas las entradas se multiplican por sus respectivos pesos y se suman. Si la salida de cualquier nodo individual está por encima del valor umbral especificado, ese nodo se activa, enviando datos a la siguiente capa de la red (según una función de activación, f). En caso contrario, no pasa ningún dato a la siguiente capa de la red. Este proceso de paso de datos de una capa a la siguiente define a la red neuronal como una red *feedforward*, que serán las empleadas en el trabajo [What are Neural Networks? | IBM, s. f.].

Ilustración 18: Funcionamiento de una red neuronal



Fuente: Lantz, 2013. p.208

La función de activación es el mecanismo por el que la neurona artificial procesa la información y la transmite a toda la red. Esta información puede transmitirse de diversas maneras: en función de si supera un umbral o no (función de activación umbral, *threshold*); usando regresión logística con salidas en el rango de 0 a 1 (función sigmoide); o produciendo una salida cero para entradas negativas y una salida lineal para entradas positivas (función rectificadora) entre otros métodos [Villanueva García, 2020].

Así, el objetivo de este apartado será elaborar construir una red neuronal[13], que nos ayude a predecir el precio de la vivienda teniendo en cuenta las diversas variables. Para ello debemos tener en cuenta que las redes neuronales funcionan mejor cuando los datos de entrada se escalan con una función de normalización o estandarización. Dado que los datos siguen una distribución uniforme, lo más recomendable es normalizarlas en intervalos de 0 a 1.[Lantz, 2013. pp.218-219]

Una vez realizadas esta labores de preprocesamiento, particionados los datos en Training y Test Set (80% y 20% respectivamente) y elaboramos el modelo usando la libreria *nnet,* que implementa redes neuronales *feed fordward* con una única capa oculta y utiliza una función de activación sigmoide por defecto.

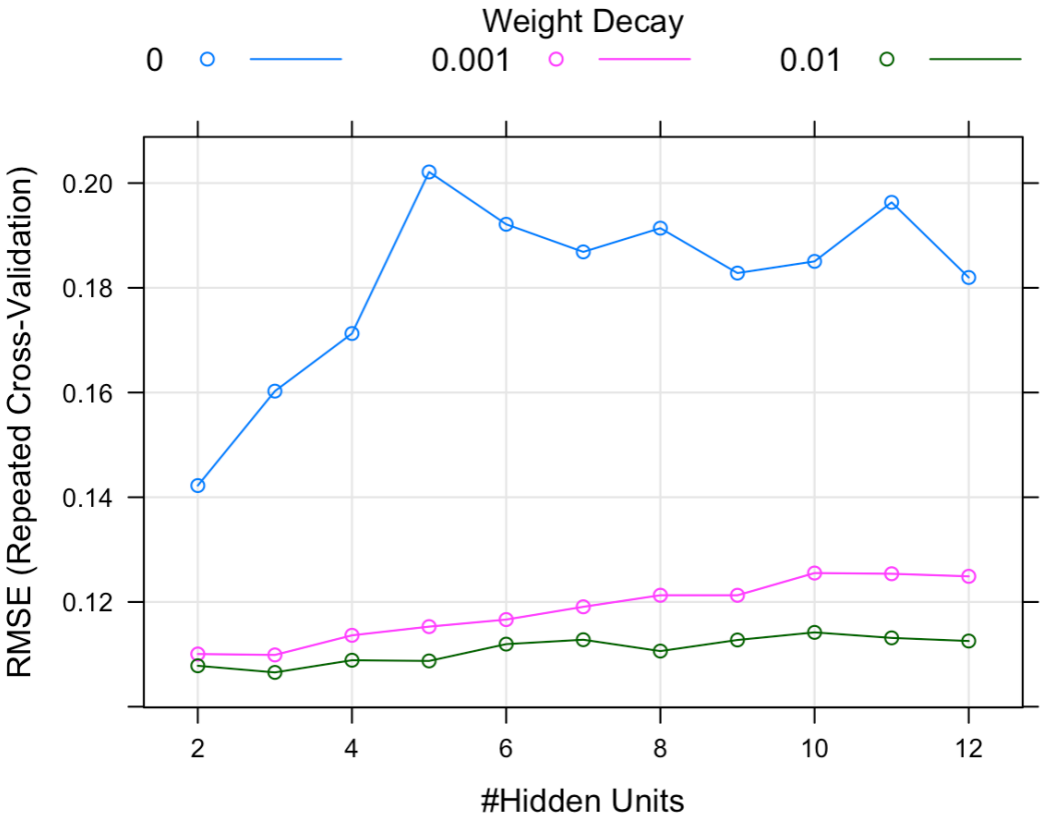
Para construir este modelo, en primer lugar crearemos la función *trainControl*, que generará los parámetros que controlan cómo se crea el modelo. En el presente caso el modelo será entrenado utilizando la validación cruzada repetida[14] con 10 pliegues (*folds*) y 3 repeticiones. Por otro lado, los hiperparámetros a optimizar en este caso son *size* (el número de neuronas en la capa oculta) y *decay* (parámetro de regularización empleado para mayor penalizar a los pesos más grandes y así prevenir el sobreajuste y lograr la generalización del modelo). Para fijar estos valores recurriremos a la búsqueda de rejilla o *Grid Search*, empleando varias combinaciones de posibles valores de *size* y *decay*, donde *size* estará entre 2 y 12, y *decay* entre 0, 0.001 y 0.01.

En el Gráfico 12: Combinaciones de hiperparámetros size y decay y RMSE observamos que los hiperparámetros óptimos, esto es, aquellos que tienen un menor RMSE son *size*= 3 y *decay*=0.01 , que producen un error de 62.729,04 en las estimaciones del test set.

[13] Elaborado e interpretado con el código y la información disponible en: Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (2021). *Aprendizaje Estadístico* (Septiembre de 2021).

[14] En la validación cruzada repetida, se realiza el mismo proceso que en la validación cruzada, pero se repite varias veces con diferentes divisiones aleatorias de los datos. En cada repetición, se divide el conjunto de datos en conjuntos de entrenamiento y prueba de manera diferente y se entrena y evalúa el modelo en cada uno de ellos. Al final, se promedia el rendimiento del modelo en todas las repeticiones. Este método ayuda a obtener estimaciones más precisas.

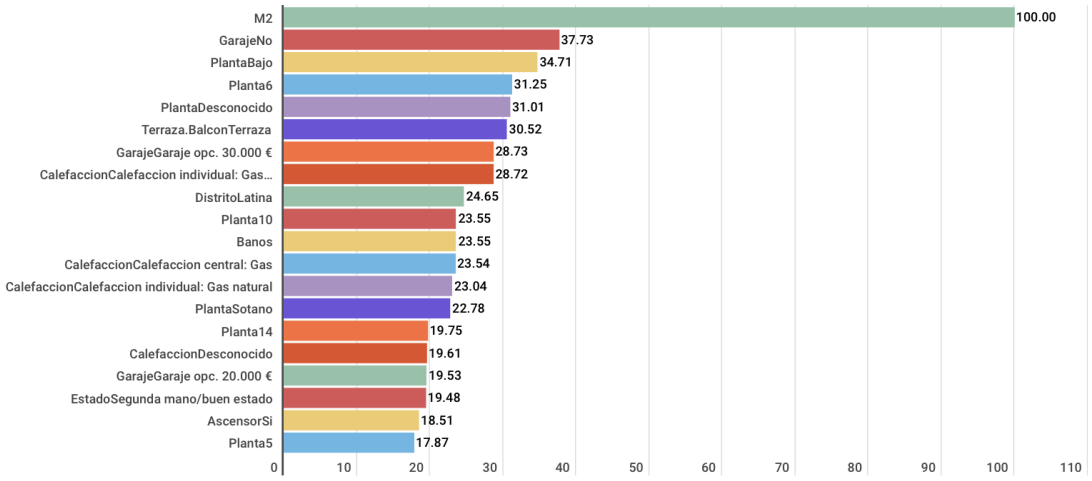
Gráfico 12: Combinaciones de hiperparámetros size y decay y RMSE



Fuente: Elaboración propia en RStudio

Además, también podemos conocer la importancia de las variables en el modelo, lo cual puede proporcionar información valiosa para mejorar la eficiencia, la interpretación y la calidad del modelo en un futuro.

Gráfico 13: Importancia de las variables[15] en la red neuronal



Fuente: Elaboración propia en Infogram

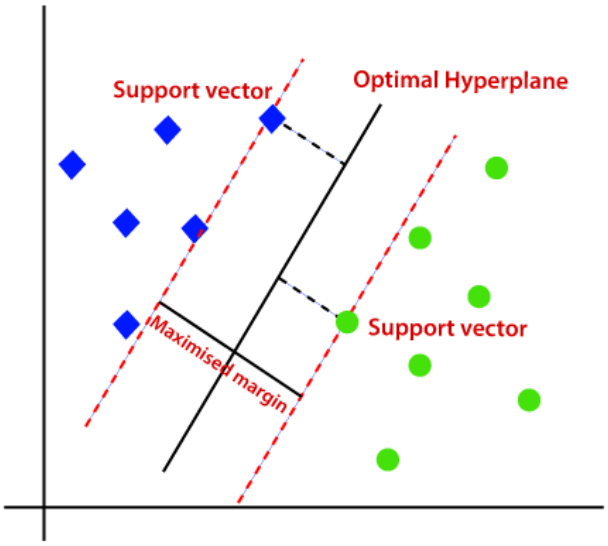
6.4. Support Vector Machine

El modelo de Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado utilizado en problemas de clasificación y regresión. El objetivo del SVM es encontrar un hiperplano en un espacio N-dimensional, siendo N el número de variables[16], que separe las diferentes clases de datos de manera óptima. Para separar las clases de datos, existen muchos hiperplanos posibles. Ahora bien, el objetivo será encontrar uno que tenga el máximo margen, es decir que maximice la distancia entre los puntos de datos más cercanos a cada lado del hiperplano. Los vectores de soporte (*Support Vectors*) son los puntos de datos que están más cerca al hiperplano de separación entre dos clases. Estos puntos son importantes porque son los que determinan la ubicación y orientación del hiperplano, y por lo tanto, son críticos para la capacidad del SVM para separar las clases de datos. Todos los demás puntos de datos se consideran redundantes y no se utilizan para construir el modelo. [Gandhi, 2018].

[15] Nota: por simplicidad solo se muestran algunas variables

[16] Cuando el número de características es 2, el hiperplano será una línea. Si es 3, será un plano bidimensional y así hasta llegar a un espacio N-dimensional.

Ilustración 19: Funcionamiento del Hiperplano



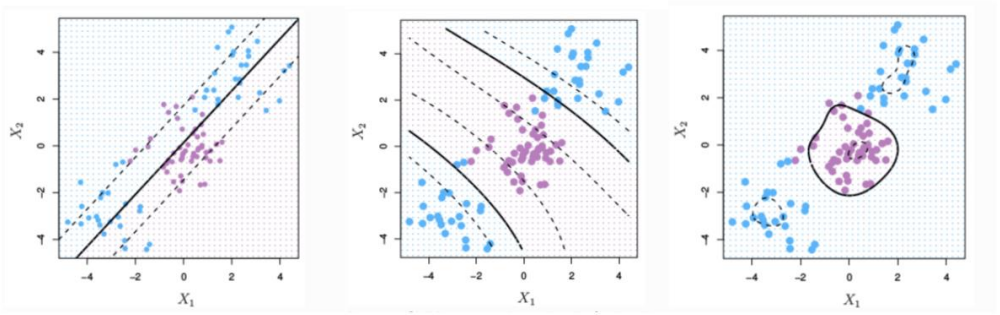
Fuente: Saini, (2021)

Ahora bien, puede ocurrir que los datos los datos no sean linealmente separables. En estos casos, SVM utiliza una función matemática llamada kernel. "Los kernels son funciones que transforman un espacio de pocas dimensiones en un espacio de dimensiones mayores mediante transformaciones complejas de los datos. También puede definirse como una función que cuantifica la similitud entre dos observaciones en un nuevo espacio dimensional". [Gil Martínez, 2020. p.6]. El uso del kernel es fundamental en el SVM, ya que permite al modelo encontrar un hiperplano de separación óptimo incluso cuando los datos no son linealmente separables en el espacio original.

En este trabajo se mostrará la ejecución de los más populares, que son:

- "Kernel lineal: equivalente a un support vector classifier, segmentación mediante una línea recta.
- Kernel radial (RBF kernel, radial basis function kernel): cuyos límites se establecen de forma radial.
- Kernel polinomial: con límites más flexibles". [Sánchez Burón, 2020]

Ilustración 20: Tipos de Kernel: Lineal, Polinomial y Radial



Fuente: Amat Rodrigo (2017)

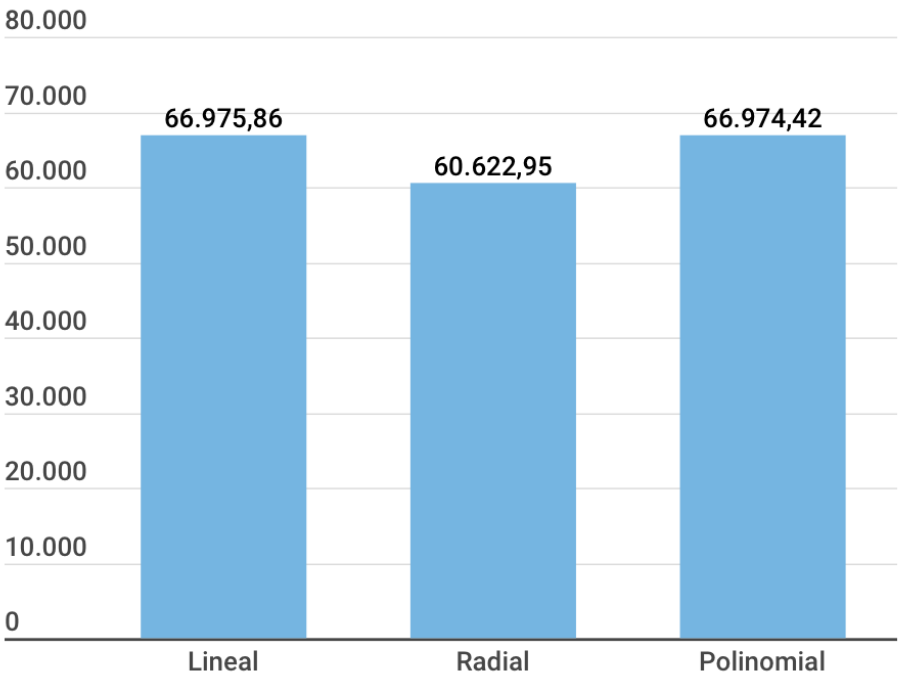
Con todo, el objetivo de este apartado será elaborar construir un modelo[17] Support Vector Machine, el cual se entrenará para encontrar la relación no lineal entre las características de las viviendas y su precio. El modelo SVM ajustará una función para separar las viviendas con diferentes precios, y luego se utilizará para predecir el precio de una nueva vivienda en función de sus características.

En este caso no va a ser preciso escalar los datos, pues el paquete de R que emplearemos (*kernlab*), lo hará automáticamente. Así, una vez elaborados[18] y comparados todos los modelos con los distintos kernels (Ver Gráfico 14: Comparación de los modelos de SVM con los distintos kernels) podemos observar que el kernel radial es el que estima el precio de la vivienda con el menor RSME.

[17] Elaborado e interpretado con el código y la información disponible en: Sánchez Burón, A. (2020). *Support Vector Machine con kernlab*. RPubs.

[18] En este caso, los hiperparámetros serán asignados por defecto siendo C (que especifica el coste de la violación de las restricciones) 1 y épsilon (empleado en la función de pérdidas de los métodos de regresión) 0.1. [Fernández Casal et al., 2021]

Gráfico 14: Comparación de los modelos de SVM con los distintos kernels



Fuente: Elaboración propia en Infogram

6.5. Ensembles

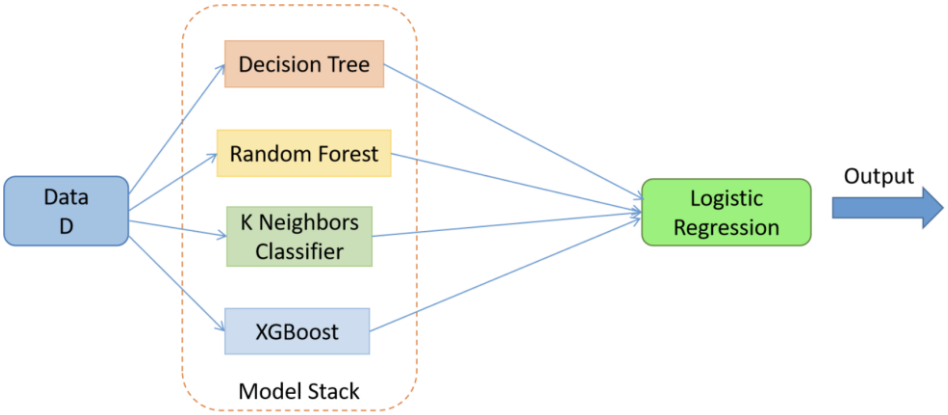
En ocasiones un único algoritmo puede no ser suficiente, especialmente con conjuntos de datos que tienen muchas variables y muchas observaciones, ya que puede ocurrir que realice buenas predicciones de las viviendas más caras, pero no de las más baratas. En estos casos, es recomendable emplear ensembles, que son conjuntos de modelos de Machine Learning, los cuales se combinan para obtener una única predicción, que a priori debería ser mejor que las obtenidas por los algoritmos por separado. La ventaja de combinar modelos radica en que los errores tienden a compensarse, resultando en un mejor error de generalización, aunque a costa de más dificultad y tiempo de computación. [Martínez Heras, 2019]

Las técnicas de ensemble pueden clasificarse principalmente en:

- Bagging. Se aplica principalmente a problemas de aprendizaje supervisado. Consta de dos pasos, *bootstrapping* y agregación. El *bootstrapping* es un método de muestreo aleatorio en el que las muestras se obtienen mediante reemplazamiento, siendo por lo tanto cada modelo diferente. En la agregación, se combinan los resultados de los modelos de base. Los modelos de bagging más comunes son los Random Forests (Ver 6.6 Random Forest).

- Boosting. Se trata de un método de ensemble en el que cada modelo aprende de los errores del modelo precedente para hacer mejores predicciones en el futuro.
- Stacking. En este caso, se entrenan varios modelos, normalmente de diferentes tipos, y a su vez un modelo supervisor que aprende a combinar mejor las predicciones de los modelos primarios. En definitiva, un algoritmo toma las salidas de los submodelos como entrada e intenta aprender cómo combinar mejor las predicciones de entrada para obtener una mejor predicción de salida. [Khandelwal, 2021].

Ilustración 21: Funcionamiento de un Ensemble con Stacking



Fuente: Khandelwal (2021)

Una de las principales ventajas del stacking es que puede aprovechar las fortalezas de diferentes modelos de aprendizaje automático y reducir las debilidades de cada modelo individual, produciendo una predicción final más precisa. Pese a que también tiene algunas desventajas, como un mayor costo computacional debido a la necesidad de entrenar múltiples modelos y un mayor riesgo de sobreajuste si no se realiza una validación cruzada adecuada, será el método que emplearemos en el presente caso.

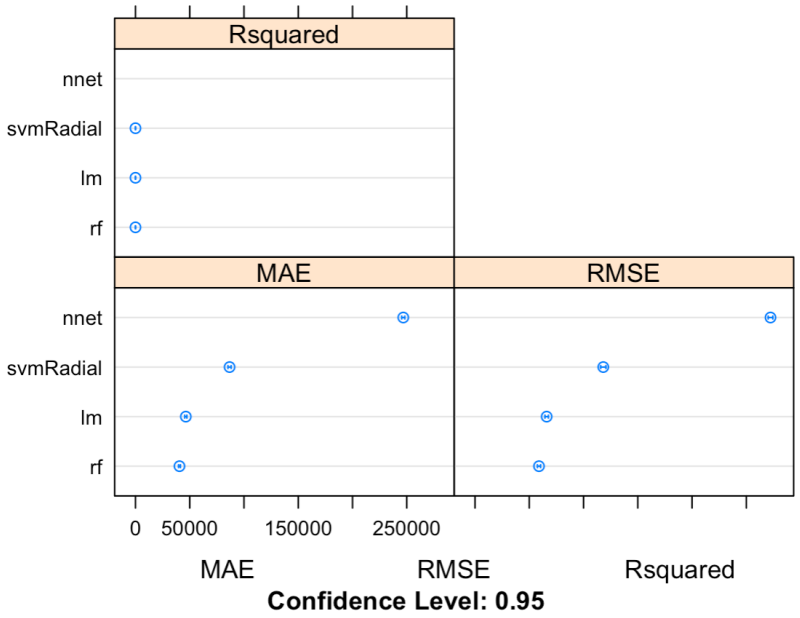
En primer lugar, y tras realizar la partición de los datos estableceremos la función *trainControl* para controlar y personalizar el proceso de entrenamiento de modelos, que como en los casos anteriores será la validación cruzada repetida con 10 pliegues (*folds*) y 3 repeticiones. Ahora bien, en este caso se hará un búsqueda aleatoria (*random*[19]) de hiperparámetros. Posteriormente se fijará la lista de algoritmos, los cuales serán

[19] El método por defecto para optimizar los parámetros de ajuste en el Training Set es utilizar una búsqueda de rejilla (*Grid Search*), lo cual suele ser eficaz pero, en los casos en que hay muchos parámetros de sintonización, puede resultar ineficaz. Una alternativa es utilizar una selección aleatoria de combinaciones de parámetros de sintonización para cubrir en menor medida el espacio de parámetros. [Kuhn, 2019]

Random Forest, regresión lineal, red neuronal, y Support Vector Machine con kernel radial.

La clave para los ensembles consigan mejores resultados que cualquiera de los algoritmos individuales es que los modelos que los forman sean diversos y que sus errores no estén correlacionados, ya que así podrán compensar los errores de los demás y la precisión general del conjunto será mayor. [Amat Rodrigo, 2020]. Así, una vez entrenado el conjunto de modelos o ensemble, observamos que los modelos menos correlacionados (0,06) son la red neuronal y la regresión lineal. Sin embargo la red neuronal (Ver Ilustración 22: Resultados Ensemble con los modelos de primer nivel) cuenta con un RMSE mucho mayor al resto de modelos. Por ello, dado que la regresión lineal y el Random Forest cuentan con el menor RMSE y además tienen una correlación relativamente baja (0,43), se emplearán como modelos de segundo nivel.

Ilustración 22: Resultados Ensemble con los modelos de primer nivel



Fuente: Elaboración propia en RStudio

Al crear el modelo stacked con Random Forest como algoritmo de segundo nivel, empleando la validación cruzada repetida con 10 pliegues (*folds*) y 3 repeticiones y una búsqueda aleatoria (*random*) de hiperparámetros, puede concluirse que el hiperparámetro mtry[20] óptimo es 2, y el RMSE en el Tes Sett es de 61.229,07.

[20] Para más información sobre el significado de este hiperparámetro véase 6.6. Random Forest.

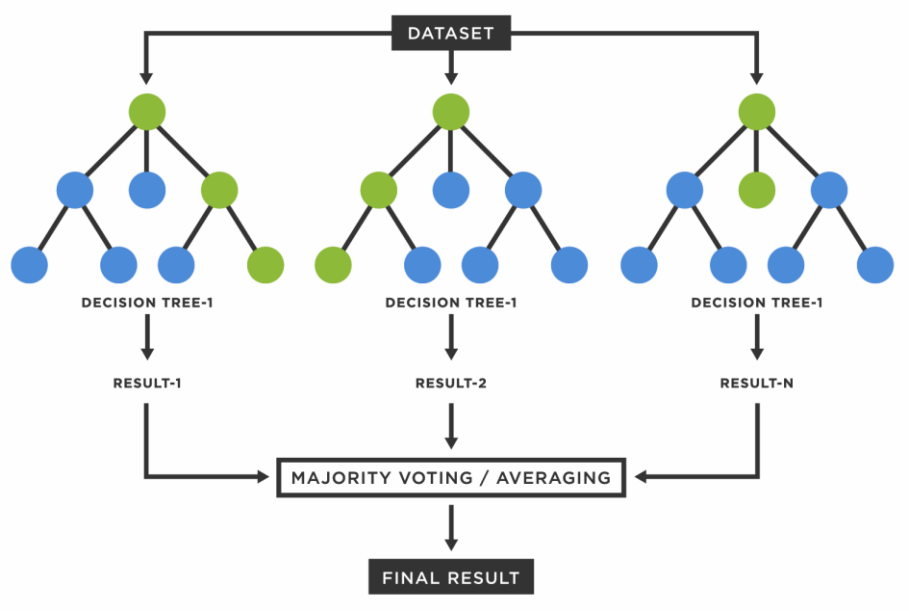
Por otro lado, al emplear la regresión lineal como algoritmo de segundo nivel, los resultados mejoran, pues el RMSE en el Test Set desciende a 58.531,83

6.6. Random Forest

"Un modelo Random Forest está formado por un conjunto (ensemble) de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping. Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo" [Amat Rodrigo, 2020].

Por lo tanto, un Random Forest es en definitiva un conjunto de árboles de decisión. Sin embargo, existen algunas diferencias entre ambos. Un árbol de decisión tiende a crear reglas que utiliza para tomar decisiones. Sin embargo, un Random Forest elige características y hace observaciones al azar, construye un bosque de árboles de decisión y luego calcula la media de los resultados. La teoría es que un número grande de árboles no correlacionados creará predicciones más precisas que un árbol de decisión individual. Esto se debe a que el conjunto de árboles trabaja en conjunto para protegerse de los errores individuales y del sobreajuste. [What is a Random Forest?, s. f.]

Ilustración 23: Funcionamiento del Random Forest



Fuente: (What is a Random Forest?, s. f.)

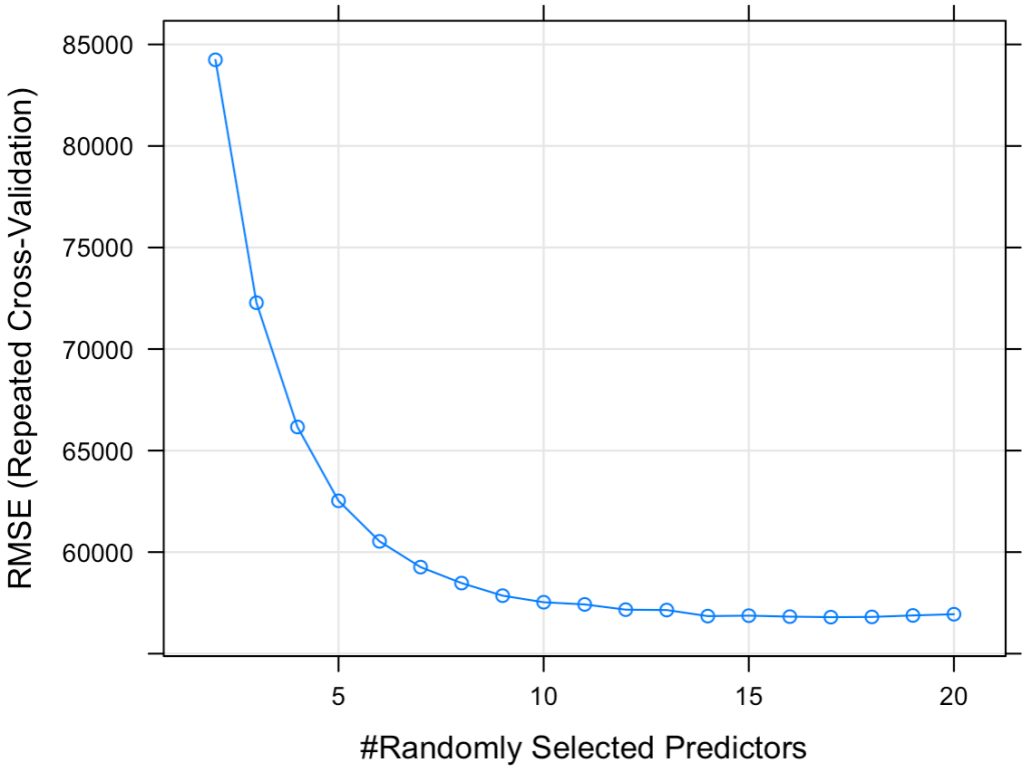
Así, el objetivo de este apartado será elaborar construir un modelo[21] Random Forest. Este método, que es un caso particular de ensembles, puede llegar a mejorar las predicciones pues combina la versatilidad y potencia en un único enfoque de aprendizaje automático y es menos propenso a sobreajustar (*overfitting*) los datos de entrenamiento en comparación con otros algoritmos de aprendizaje automático, como los árboles de decisión. Además, se utiliza para tareas de predicción, como es el presente caso, por lo que nos va a permitir estimar el precio de las viviendas objeto de estudio, y determinar cuáles de estas variables son más importantes para hacer una predicción precisa del precio.

El punto clave de este modelo será la optimización del hiperparámetro *mtry*, que representa el número de variables que se consideran en cada división de los árboles de decisión y permite controlar cuánto se descorrelacionan los árboles entre sí. Al igual que en la red neuronal, primeramente crearemos la función *trainControl*, utilizando la validación cruzada repetida con 10 pliegues (*folds*) y 3 repeticiones. Para la búsqueda del hiperparámetro *mtry*, recurriremos una vez más a la búsqueda de rejilla o *Grid Search*, empleando varias combinaciones de *mtry* entre 2 y 20.

[21] Elaborado e interpretado con el código y la información disponible en: Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

En el Gráfico 15: Combinaciones de hiperparámetros mtry y RMSE observamos que el hiperparámetro *mtry* óptimo es 17, el cual genera un RMSE de 59.980,77 en las estimaciones del Test Set.

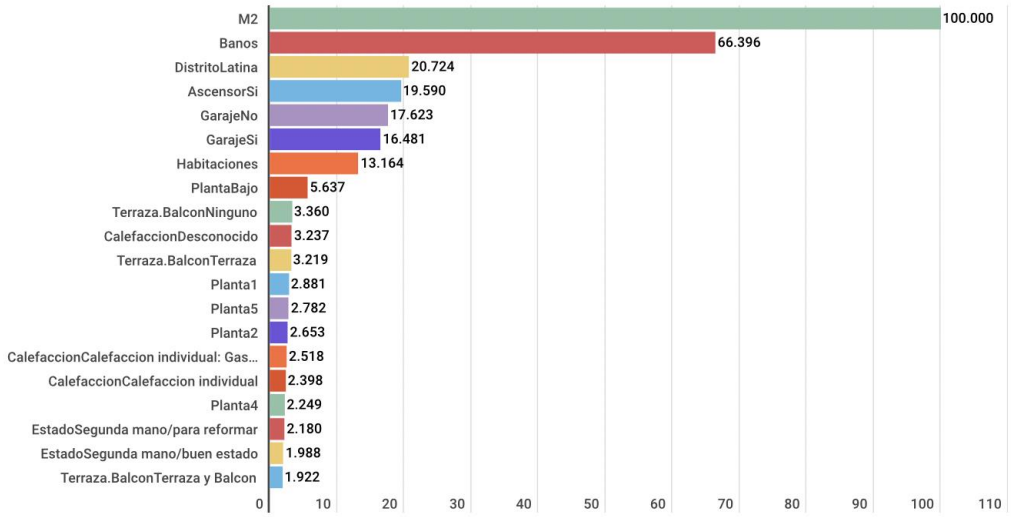
Gráfico 15: Combinaciones de hiperparámetros mtry y RMSE



Fuente: Elaboración propia en RStudio

Pese a que estos árboles no pueden dibujarse como los árboles modelo o los árboles de regresión, podemos conocer la importancia de las variables en el modelo y comprobar que las variables más destacadas difieren de las que consideradas como tal en la red neuronal.

Gráfico 16: Importancia de las variables en Random Forest



Fuente: Elaboración propia en Infogram

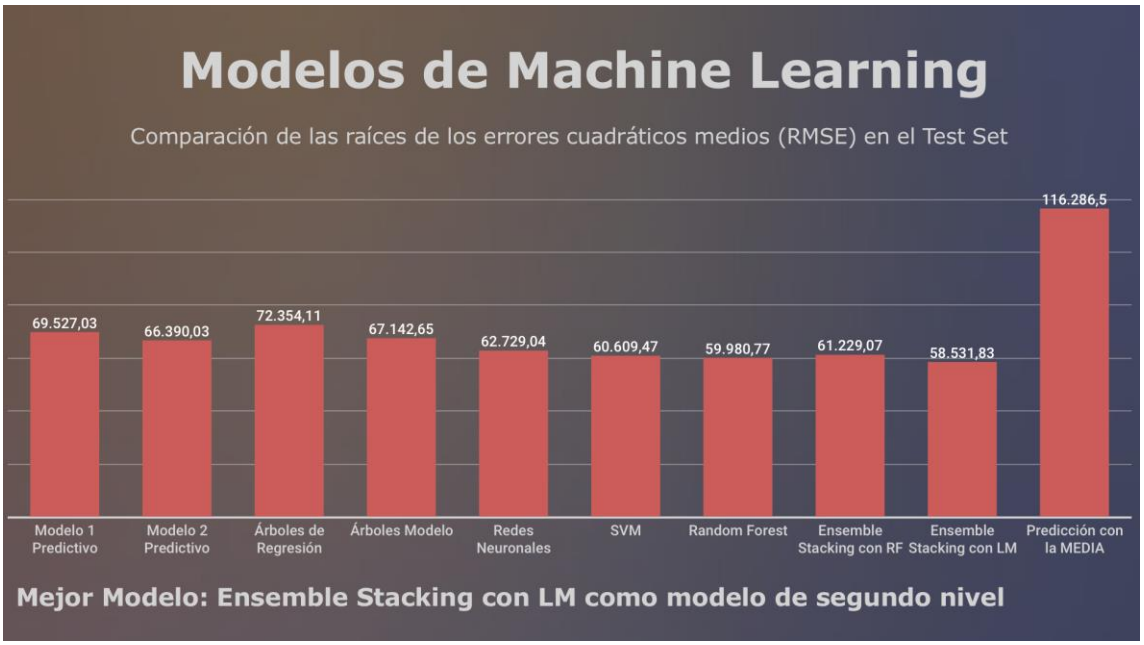
7. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se ha explorado la aplicación de diversos modelos de Machine Learning con el objetivo realizar predicciones sobre el precio de la vivienda en los distritos de Ciudad Lineal y La Latina. Tras una limpieza y análisis exploratorio de los datos de entrada, se han construido modelos de aprendizaje automático utilizando seis técnicas diferentes: regresión lineal, tanto explicativa como predictiva, árboles de regresión y árboles modelos, redes neuronales, máquinas de vectores soporte, ensembles y un tipo especial de ellos, el Random Forest.

Tras la elaboración de todos los modelos y la comparación de su RMSE en el Test Set (Ver Gráfico 17: RMSE de los distintos modelos de Machine Learning elaborados), el Ensemble, en concreto el que emplea la regresión lineal como algoritmo de segundo nivel ha sido el mejor modelo, con un RMSE de 58.531,83. Esto era de esperar por la naturaleza tabular de los datos, pues los datos tabulares a menudo tienen muchas características o columnas, lo que puede hacer que sea difícil para un solo modelo capturar toda la complejidad de la relación entre las variables. El ensemble aborda estos problemas al combinar múltiples modelos, aprovechar las fortalezas de cada uno y compensar errores, lo que cómo ha quedado demostrado, resulta en mejores predicciones.

Por otro lado, también queda demostrado cómo se obtienen mejores predicciones cuando se utilizan métodos más sofisticados como la búsqueda de hiperparámetros o la validación cruzada, pues aquellos modelos en los que se ha empleado (Red neuronal, Random Forest y Ensemble) cuentan con resultados mejores.

Gráfico 17: RMSE de los distintos modelos de Machine Learning elaborados



Fuente: Elaboración propia en Infogram

Ahora bien, una vez obtenidas estas predicciones es necesario comprobar si son buenas, si superan el *benchmark*. Para ello lo más sencillo es compararlas con el modelo nulo, la media. Es decir, el objetivo sería que los modelos construidos tuvieran un error de predicción menor que el error resultante de emplear la media como predictor. Así, el RMSE prediciendo con la media es de 116.286,5, resultando nuestros modelos bastante inferiores, por lo que podemos afirmar que las predicciones son buenas. El trabajo es por lo tanto capaz de confirmar la tercera hipótesis planteada, pues los modelos elaborados han obtenido un RMSE inferior a la media.

La primera hipótesis también se confirma, pues varios de los modelos cuentan con un RMSE inferior a 60.000 euros, en concreto el Random Forest y el ensemble empleando regresión lineal como algoritmo de segundo nivel. En cuanto a la segunda hipótesis, gracias especialmente al modelos de regresión lineal explicativo, se ha podido cuantificar monetariamente las variables. Son los valores de Beta los que nos han permitido conocer que, por ejemplo, el aumento del metro cuadrado equivale a 1828,90 euros, o que la presencia de un baño más incrementa el precio de la vivienda en 45.936,65 euros.

Si bien se han conseguido confirmar las hipótesis y cumplir los objetivos planteados inicialmente, existen varios ajustes que podrían ayudar a mejorar las estimaciones. Algunos de los modelos han utilizado técnicas más avanzadas, como la validación cruzada y la búsqueda de hiperparámetros, porque el paquete *Caret* ofrecía una sencilla implementación. Ahora bien, para poder hacer una comparación más justa, en un trabajo futuro, se recomienda extender estos métodos a todos los modelos empleados.

Por otro lado, en este trabajo únicamente se han empleado algunas de las librerías de las que dispone R para elaborar modelos de aprendizaje automático. Sin embargo existen muchas más. Por ejemplo, en el caso de las redes neuronales ha sido empleado el paquete *nnet*, que permite elaborar redes neuronales de arquitectura sencilla, con una única capa oculta. Ahora bien, otros paquetes como *neuralnet* o *keras* permiten elaborar redes más complejas y que quizá hubieran obtenido resultados mejores. Además también existen otros modelos y algoritmos predictivos de Machine Learning que podrían haberse implementado, como *Naive Bayes*, *Gradient Boosting*, *XGBoost* o *k-nearest neighbors* y que posiblemente pudieran mejorar las predicciones.

En definitiva, los resultados obtenidos se encuentran limitados por las restricciones planteadas, pero son buenos y cumplen con los objetivos enunciados. No obstante, dentro del mundo del aprendizaje automático existen infinidad de posibilidades para ayudar a mejorar las predicciones y que en un trabajo futuro podrían implementarse.

8. BIBLIOGRAFÍA

¿Qué es el Web Scraping? (2020, 10 septiembre). IONOS Digital Guide. Recuperado 4 de febrero de 2023, de https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/que-es-el-web-scraping/

Alfaro-Navarro, J., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). *A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems*. Complexity. vol. 2020, Article ID 5287263, 12 pages,. https://doi.org/10.1155/2020/5287263

Alves, P., y San Juan, L. (2021). El impacto de la crisis sanitaria del Covid-19 sobre el mercado de la vivienda en España. En Banco de España. Recuperado 7 de febrero de 2023, de https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/ArticulosAnaliticos/21/T2/Fich/be2102-art16.pdf

Amat Rodrigo, J. (2016, julio). *Introducción a la Regresión Lineal Múltiple*. Recuperado 21 de marzo de 2023, de https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple.html

Amat Rodrigo, J. (2017, abril). *Máquinas de Vector Soporte (Support Vector Machines, SVMs)*. Recuperado 21 de marzo de 2023, de https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Amat Rodrigo, J. (2020). *Arboles de decision, Random Forest, Gradient Boosting y C5.0*. Ciencia de datos. Recuperado 3 de abril de 2023, de https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

An Introduction to R. (2022, 31 octubre). Recuperado 10 de febrero de 2023, de https://cran.r-project.org/doc/manuals/r-release/R-intro.html

Ardila, D., Cauwels, P., Sanadgol, D., & Sornette, D. (2013). Is There A Real Estate Bubble in Switzerland? ((Diagnostic as of 2012-Q4)). *Swiss Finance Institute Research, Paper No. 13-07*. https://doi.org/10.2139/ssrn.2237561

Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernárdez, S., & Afonso, C. (2018a). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences*, *8*(11), 2321. https://doi.org/10.3390/app8112321

Bhushan Jha, S., Babiceanu, R. F., Pandey, V. & Kumar Jha, R. (s. f.). Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. https://arxiv.org/pdf/2006.10092.pdf

Bhushan Jha, S., Babiceanu, R. F., Pandey, V. & Kumar Jha, R. (s. f.). Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. https://arxiv.org/pdf/2006.10092.pdf

Caughlin, D. E. (2022). Chapter 48 Applying k-Fold Cross-Validation to Logistic Regression. En *R for HR: An Introduction to Human Resource Analytics Using R*. https://rforhr.com

Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. *Land*, *12*(4), 740. https://doi.org/10.3390/land12040740

Cross-Validation: definición e importancia en Machine Learning. (2022). Formation Data Science | DataScientest.com. Recuperado 2 de abril de 2023, de https://datascientest.com/es/cross-validation-definicion-e-importancia

d'Archimbaud, E. (s. f.). *Types of Cross Validation Techniques used in Machine Learning*. Kili. Recuperado 2 de abril de 2023, de https://kili-technology.com/data-labeling/machine-learning/cross-validation-in-machine-learning

Evolución de la compraventa de viviendas en España, según el INE. (s. f.). EpData. Recuperado 5 de febrero de 2023, de https://www.epdata.es/evolucion-compraventa-viviendas-ine/1ccf1579-f9ba-4518-8a30-ccf991fb301c/espana/106

Extraer datos de inmuebles de Idealista: Tutoriales de casos. (2022). Octoparse. Recuperado 15 de noviembre de 2022, de https://helpcenter.octoparse.es/hc/es/articles/5012405988377-Extraer-datos-de-inmuebles-de-Idealista

Fernández Casal, R., Costa Bouzas, J., y Oviedo de la Fuente, M. (2021). *Aprendizaje Estadístico* (Septiembre de 2021). https://rubenfcasal.github.io/aprendizaje_estadistico/aprendizaje_estadistico.pdf

Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. Recuperado 2 de abril de 2023, de https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gil Martínez, C. (2018, junio). *Árboles de decisión y métodos de ensemble*. GitHub. Recuperado 21 de marzo de 2023, de https://github.com/CristinaGil/Ciencia-de-Datos-R/blob/fd22e722f751f0fba2e6e4ed357abbdb4123f197/PDF/Arboles_de_decision_y_metodos_ensemble_(bagging,_random%20forests,_boosting).pdf

Gil Martínez, C. (2020, 16 febrero). *Máquina de Vector Soporte*. GitHub. Recuperado 21 de marzo de 2023, de https://github.com/CristinaGil/Ciencia-de-Datos-R/blob/master/PDF/Maquinas_de_Vector_Soporte_(SVM).pdf

Guía práctica de introducción al Análisis Exploratorio de Datos. (2021). En *Ministerio de Asuntos Económicos y Transformación Digital*. Recuperado 7 de febrero de 2023, de https://datos.gob.es/sites/default/files/doc/file/analisis_exploratorio_de_datos_2021_v6.pdf

Hernández, F. (2021). *Modelos Predictivos*. https://fhernanb.github.io/libro_mod_pred/

Herramienta de Web Scraping Gratis | Octoparse. (s. f.). https://www.octoparse.es/

Irizarry, R. A. (2023). *Introducción a la ciencia de datos: Análisis de datos y algoritmos de predicción con R*. http://rafalab.dfci.harvard.edu/dslibro/

Khandelwal, Y. (2021). *Ensemble Stacking for Machine Learning and Deep Learning*. Analytics Vidhya. Recuperado 3 de abril de 2023, de https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/

Kuhn, M. (2019). *The caret Package*. https://topepo.github.io/caret/index.html

Kumar, A. (2023, 15 enero). *Overfitting & Underfitting in Machine Learning*. Data Analytics. Recuperado 16 de marzo de 2023, de https://vitalflux.com/overfitting-underfitting-concepts-interview-questions/

Laboa, H. (2019, 3 agosto). Compra tu casa de forma inteligente -III. Exploratory data analysis. *Haritz Laboa*. Recuperado 7 de febrero de 2023, de https://www.hlaboa.com/post/Compra_tu_casa_de_forma_inteligente-III_exploratory_data_analysis/

Lantz, B. (2013). *Machine Learning with R*. Packt Publishing. https://supermariogiacomazzo.github.io/STOR538_WEBSITE/Textbooks%20in%20R/Machine%20Learning%20with%20R.pdf

Las actividades constructoras e inmobiliarias representaron un 17% del PIB: Susana de la Riva en Capital Radio. (2021, 26 febrero). Tinsa Tasaciones Inmobiliarias. Recuperado 7 de febrero de 2023, de https://www.tinsa.es/blog/tinsa/las-actividades-constructoras-e-inmobiliarias-representaron-un-17-del-pib/

Martínez Álvarez, J.A. y García Martos, D. (2014). Auge y crisis del sector inmobiliario reciente: Interrelación con la política económica, Instituto de Estudios Fiscales, DOC no 5/2014. Recuperado 5 de febrero de 2023, de

https://www.ief.es/docs/destacados/publicaciones/documentos_trabajo/2014_05.pdf

Martinez Heras, J. (2019, 31 mayo). *Ensembles: voting, bagging, boosting, stacking*. IArtificial.net. Recuperado 10 de febrero de 2023, de https://www.iartificial.net/ensembles-voting-bagging-boosting-stacking/

Martínez Pagés, J., y Ángel Maza, L. (2003). Análisis del precio de la vivienda en España: Documento de Trabajo no 0307. En Banco de España, Servicio de Estudios. Recuperado 4 de febrero de 2023, de https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/03/Fic/dt0307.pdf

Méndez González, J. (2019, 13 octubre). *Stepwise Regresión*. RPubs. Recuperado 15 de marzo de 2023, de https://rpubs.com/jorge_mendez/609253

Mendoza Vega, J. B. M. (2021, 21 abril). *Variables dummy (one-hot encoding) con R*. Medium. Recuperado 20 de marzo de 2023, de https://medium.com/@jboscomendoza/variables-dummy-one-hot-encoding-con-r-1f62b4ec8242

Ministerio de Derechos Sociales y Agenda 2030. (2021). Resumen ejecutivo Informe Juventud en España 2020. En *Instituto de la Juventud*. Recuperado 4 de febrero de 2023, de https://www.injuve.es/sites/default/files/adjuntos/2021/03/informe-juventud-en-espana-2020-resumen-ejecutivo.pdf

Mohamed, H. H., Ibrahim, A. H., & A. Hagras, O. (2023). Forecasting the Real Estate Housing Prices Using a Novel Deep Learning Machine Model. *Civil Engineering Journal*, *9*, 46-64. https://doi.org/10.28991/cej-sp2023-09-04

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Moosavi, V. (2017, 30 marzo). *Urban Data Streams and Machine Learning: A Case of Swiss Real Estate Market*. arXiv.org. https://arxiv.org/pdf/1704.04979.pdf

Ocaña P. de Tudela, C., y Torres, R. (2019). El mercado de la vivienda: situación y perspectivas a corto plazo. En Funcas. Recuperado 6 de febrero de 2023, de https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_CIE/273art02.pdf

Ocaña P. de Tudela, C., y Torres, R. (2020). Impacto de la pandemia sobre el sector inmobiliario. En Funcas. Recuperado 6 de febrero de 2023, de https://www.funcas.es/wp-content/uploads/2020/09/Cie278art03.pdf

Otero Moreno, M., y Blanco García Lomas., J. (2014). El sector inmobiliario en España. En Instituto de estudios económicos. Recuperado 5 de febrero de 2023, de https://www.ieemadrid.es/wp-content/uploads/El-sector-inmobiliario-en-España.pdf

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928-2934. https://doi.org/10.1016/j.eswa.2014.11.040

Park, D., & Ryu, D. (2021). A Machine Learning-Based Early Warning System for the Housing and Stock Markets. *IEEE Access*, *9*, 85566-85572. https://doi.org/10.1109/access.2021.3077962

Russell, S. J., Norvig, P., y Rodríguez, J. M. C. (2004). *Inteligencia artificial: un enfoque moderno*. Pearson Educación. https://luismejias21.files.wordpress.com/2017/09/inteligencia-artificial-un-enfoque-moderno-stuart-j-russell.pdf

Saini, A. (2021). *Support Vector Machine (SVM): A Complete guide for beginners*. Analytics Vidhya. Recuperado 2 de abril de 2023, de https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Sánchez Burón, A. (2020). *Support Vector Machine con kernlab*. RPubs. Recuperado 2 de abril de 2023, de https://rpubs.com/AdSan-R/SVMkernlab

Sanyal, S., Biswas, S. Kr., Das, D., & Chakraborty, M. (2022). *Boston House Price Prediction Using Regression Models*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9848309/

Statista. (2022, 20 septiembre). *Peso de las actividades inmobiliarias sobre el PIB España 2005-2020*. Recuperado 7 de febrero de 2023, de https://es.statista.com/estadisticas/549634/aportacion-de-las-actividades-inmobiliarias-al-pib-en-espana/

The world's most valuable resource is no longer oil, but data. (2017, 6 mayo). The Economist. Recuperado 20 de marzo de 2023, de https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Tipos de aprendizaje del Deep Learning. (2022, 8 agosto). KeepCoding Tech School. Recuperado 7 de febrero de 2023, de https://keepcoding.io/blog/tipos-aprendizaje-deep-learning/

Vigen, T. (s. f.). *Spurious correlations*. Recuperado 10 de febrero de 2023, de https://www.tylervigen.com/spurious-correlations

Villanueva García, J. D. (2020, 23 octubre). *Redes neuronales desde cero (I) - Introducción*. IArtificial.net. Recuperado 1 de abril de 2023, de https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion/

What are Neural Networks? | IBM. (s. f.). https://www.ibm.com/topics/neural-networks

What is a Random Forest? (s. f.). TIBCO Software. Recuperado 3 de abril de 2023, de https://www.tibco.com/reference-center/what-is-a-random-forest

Wickman, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. https://r4ds.had.co.nz

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