Facultad de Administración y Dirección de Empresas
Universidad Politécnica de Valencia

Análisis y predicción del precio en alquiler de vivienda en la ciudad de Valencia

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

Autora: Laura Pellicer Martínez

Tutor: Francisco Guijarro Martínez

2021-2022

Resumen

El presente trabajo fin de grado se centra en el análisis y predicción del precio de una vivienda en la ciudad de Valencia. Para ello, en primer lugar, se construirá una base de datos con todos aquellos anuncios ofertados en el portal web inmobiliario seleccionado. Así pues, la recogida de toda esta información se realizará mediante el proceso de *web scraping*. Una vez obtenida se procederá al preprocesamiento y transformación de todos esos datos, dejándolos en un formato legible para la herramienta R Studio. Acto seguido se procederá al estudio de cada una de las variables, analizando además la dependencia entre cada una de ellas y la variable precio. Una vez entendido el panorama se pasará a formular un modelo de regresión capaz de predecir el precio de la vivienda a partir de una serie de variables. En un primer momento se probará con un modelo sencillo de regresión múltiple y se utilizarán algoritmos más complejos para tratar de reducir el margen de error.

Palabras clave: regresión lineal, mercado del alquiler, web scraping, machine learning

Abstract

This thesis focuses on analyzing and predicting the price of a house in the city of Valencia. To do this, first of all, a database will be built with all the advertisements offered on the selected real estate website. Thus, the collection of all this information will be carried out through the process of web scraping. Once obtained, the data will be preprocessed and transformed into a readable format for the R Studio tool. We will then study each of the variables, analyzing the dependence between each of them and the price variable. Once the overview is understood, we will formulate a regression model capable of predicting house prices from a series of variables. Initially, a simple multiple regression model will be tested and more complex algorithms will be used to reduce the margin of error.

Keywords: linear regression, rental market, web scraping, machine learning.

Tabla de contenidos

Índice de ilustraciones

Índice de tablas

Índice de gráficos

CAPÍTULO 1

Introducción

La cita del filósofo Thomas Hobbes “la información es poder” cobra especial relevancia en la época que estamos viviendo. Vivimos en un momento donde la información se sitúa como el eje central de nuestras vidas, estamos a un click de distancia de obtener prácticamente cualquier dato que se necesite. Es por esto que ya se habla de la revolución de la información, cuyo fin último no deja de ser otro más que el de proporcionar las herramientas necesarias para tomar decisiones adecuadas.

Precisamente un ámbito en el que tomar una decisión fundamentada y correcta no es tarea fácil es en el mercado inmobiliario. Comprar una casa es sin duda una de las decisiones más importantes que uno toma a lo largo de su vida. El precio de una casa puede depender de una amplia variedad de factores que van desde la ubicación o sus características hasta la demanda y oferta del mercado. No obstante, la compleja situación que atraviesa el mercado del alquiler en España provoca que tampoco sea una labor sencilla decidir en esta otra modalidad.

En este sentido se han utilizado durante muchos años diferentes modelos econométricos para tratar de predecir el precio de una vivienda. Sin embargo, precisamente gracias a esta nueva era que estamos viviendo, han aparecido nuevas técnicas que parecen mucho más adecuadas y potentes para resolver este tipo de problema.

Relacionado con este tema hay diferentes investigaciones que resultan interesantes, en concreto, en 2018 se realiza un estudio muy relevante sobre el mercado de la vivienda en Melbourne (Phan 2018). En dicha publicación se parte de un conjunto de variables como la longitud y latitud, el número de habitaciones y baños, el tipo de propiedad o la distancia al centro de la ciudad. Tras analizarlas y seleccionar cuales son más relevantes se procede a la formulación de diferentes modelos utilizando redes neuronales, árboles de regresión, entre otros. De esta manera se consiguen predicciones del precio del inmueble muy cercanas a la realidad, sin apenas margen de error. Otro caso que arrojan resultados llamativos es el desarrollado para un condado de Virginia (Park y Bae 2015). De este caso resulta muy interesante la cantidad y sobre todo la concreción de las variables que se recogen. Son detalles tan específicos como el número de chimeneas, el tipo de exterior (vallado parcial, cubierto…) hasta variables que cuantifican la calidad de las escuelas existentes en ese barrio.

Adicionalmente, este tipo de técnicas también se está utilizando para predecir y analizar el precio de los alquileres vacacionales. Una de las plataformas más famosas que se dedican a este tipo de actividad es AirBnB. En estos estudios cobra especial importancia las diferentes opiniones que dejan los usuarios al finalizar su estancia en cada una de las propiedades, ya que muchos de los futuros clientes las toman como referencia (Rezazadeh Kalehbasti, Nikolenko, y Rezaei 2021).

1.1 Motivación

Durante toda mi etapa universitaria he vivido en diferentes pisos de estudiantes, unos más baratos otros más caros, unos mejor ubicados otros menos, unos más espaciosos y otros menos. Como para gran parte de la población la pandemia supuso un cambio y una nueva forma de hacer las cosas. En ese momento las clases se tuvieron que impartir en modalidad online por lo que decidí volver a mi casa con mi familia. Sin embargo, ahora que ya parece que todo vuelve a la normalidad me surge otra vez la oportunidad de volverme a vivir a Valencia y, por tanto, la necesidad de buscar un piso en el que alojarme. No obstante, ahora esta tarea de búsqueda no es como las anteriores veces, las necesidades y requerimientos han cambiado. En estos momentos ya no necesito un piso que este cerca de la universidad por lo que tengo la oportunidad de explorar nuevas zonas.

Paralelamente a este proceso de mudanza en el que me estaba empezando a embarcar, llegaba el momento de realizar el trabajo fin de grado correspondiente a la parte de ADE. Entonces, tras pasar muchas horas viendo diferentes portales web en los que se anuncian muchas propiedades para ser arrendadas me surge la idea. Es en ese momento en el que decido que puede ser muy interesante analizar de forma exhaustiva la oferta de viviendas que hay actualmente en la ciudad de Valencia. Así, una vez realizado y con todo ese bagaje poder crear un modelo que sea capaz de predecir el precio de una propiedad según las características de la misma.

No obstante, aparte de satisfacer esta inquietud personal también se quiere utilizar este trabajo como una herramienta más de aprendizaje.

1.2 Objetivos

El objetivo general de este trabajo es analizar y predecir el precio de las viviendas en alquiler en la ciudad de Valencia. De la misma forma se establecen una serie de objetivos a nivel especifico que se pretenden alcanzar al finalizar este proyecto. Se procede a enumerarlos a continuación:

- ✓ Diseñar un proceso para crear una base de datos con diferentes propiedades en alquiler, anotando para cada una de ellas sus características y funcionalidades que posee.
- ✓ Comprender y conocer el panorama general de la situación del mercado del alquiler inmobiliario en Valencia. Siendo capaz de aportar diversas métricas o gráficos atendiendo a los diversos criterios de clasificación de los alojamientos (barrio, número de habitaciones, metros cuadrados…).
- ✓ Averiguar qué factores son los más influyentes a la hora de establecer el precio de un piso en alquiler.
- ✓ Elaborar un modelo capacitado para predecir el precio de una vivienda tras facilitarle una serie de características sobre esta.

1.3 Relación con las asignaturas

Para la realización de este trabajo se han requerido de conocimientos aprendidos a lo largo de toda la carrera. En concreto, se pueden agrupar en tres grandes bloques o áreas de conocimiento.

El primero de ellos hace referencia a la estadística, a largo del grado se ha cursado tres asignaturas relacionadas con este ámbito. Desde la más básica como es introducción a la estadística, para continuar con métodos estadísticos en Economía y acabar con econometría.

Gracias a todas estas asignaturas se ha conseguido desarrollar una fuerte base de conocimientos que ha sido primordial para poder desarrollar este trabajo. En la asignatura más elemental se aprendieron temas tan básicos, pero imprescindibles, como la estadística descriptiva o las diferencias entre las principales distribuciones. A continuación, en la siguiente materia, se profundizó en la inferencia estadística. En esta se aprendieron a utilizar tanto las pruebas paramétricas como las no paramétricas con las que poder establecer diversas pruebas que permitan validar o rechazar distintas hipótesis. Hecho que sin duda ha sido muy importante en el desarrollo del trabajo. No obstante, también se estudiaron temas como el análisis de la varianza y una primera introducción a los modelos de regresión, aspectos que igualmente han sido importantes para el desarrollo del trabajo. Por último, en econometría se profundizó en los modelos de regresión aprendiendo a hacer un análisis y diagnóstico de los mismos con el que poder detectar problemas como la multicolinealidad o la heterocedasticidad.

Otro de los bloques que recoge aspectos que han sido necesarios para el desarrollo de este trabajo son aquellos relacionados con la investigación ya sea de mercados o más relacionada con la empresa. En particular, las asignaturas que recogen toda esta información son: investigación operativa e investigación comercial. En la primera se aprendieron aspectos fundamentales como modelizar un problema, así como a saber definir las variables necesarias para este. En la segunda, aunque estaba fundamentada en la investigación y metodología comercial se realizó un trabajo de investigación de mercados con el software R que ha servido de base e inspiración para desarrollar este.

Por último, otro de los módulos que se ha necesitado para el correcto desarrollo de este trabajo ha sido el de derecho. En la asignatura de derecho de la empresa se aprendió a leer y comprender textos jurídicos, así como a localizar la información necesaria en el código legal. Además, se trataron temas relacionados con la competencia desleal que junto con asignaturas de la rama de informática han sido claves para entender el marco legal en el que se enmarca el *web scrapping*.

1.4 Orden documental

En este apartado de la introducción se procede a explicar cuáles van a ser los temas a tratar en cada uno de los apartados del presente trabajo. De esta manera se pretender facilitar y guiar al lector para una mayor comprensión de la memoria.

En este primer capítulo titulado introducción se ha realizado una pequeña presentación de lo que se puede esperar del trabajo. Asimismo, se ha explicado cuales han sido las inquietudes que han motivado la elaboración de este proyecto. A continuación, se ha enunciado a nivel general cual es el objetivo del mismo, así como los objetivos específicos que se pretenden alcanzar al finalizarlo. Por último, se ha hecho un repaso sobre aquellas materias cursadas durante la carrera que guardan una estrecha relación con el presente trabajo fin de grado.

El siguiente apartado recibe por nombre marco conceptual y en este se revisarán dos conceptos clave para el desarrollo de este proyecto. Primeramente, se explicará qué es el *web scraping* y para qué se utiliza además de revisar la legalidad de esta práctica. A continuación, se profundizará en el mundo del aprendizaje automático, enfatizando en su inmenso auge debido a su creciente uso. Asimismo, se explicarán cuáles son los principales algoritmos que se utilizan a día de hoy en este tipo de prácticas.

A continuación, se encuentra el que constituye el grueso del trabajo, la metodología. En este se repasarán cronológicamente todos los pasos y análisis necesarios hasta llegar a construir el modelo final. Primero se articulará todo un proceso capaz de recolectar una muestra lo suficientemente significativa y representativa de las propiedades en alquiler de la ciudad de

Valencia. A continuación, se hace necesario el preprocesamiento y transformación oportuno de todos esos datos previamente recogidos. Posteriormente se realizará tanto un análisis individualizado como bivariante para conocer y entender cada una de las variables, así como la dependencia de cada una de ellas con la variable precio. Para acabar se confeccionará un modelo de regresión múltiple junto con otros modelos basados en otras técnicas y algoritmos.

En el siguiente apartado, que recibe por nombre resultados y discusión, se analizarán las métricas obtenidas previamente enmarcándolas y dándoles sentido dentro del contexto del trabajo. De la misma forma se elegirá el mejor modelo atendiendo a los diferentes criterios estipulados.

El último capítulo, como cualquier otro trabajo, reúne las conclusiones obtenidas tras la elaboración de este. Así pues, se hará un repaso por los diferentes objetivos planteados reflexionado sobre si han sido alcanzados o no. Además, se recapitulará cuáles han sido las lecciones aprendidas y los errores que se han cometido y se finalizará dejando enumeradas cuales podrían ser algunas líneas futuras de próximos trabajos.

CAPÍTULO 2

Marco teórico

En este segundo apartado del presente trabajo se va a realizar una revisión de la situación actual referente a las tecnologías empleadas, así como definirse todos aquellos conceptos necesarios para el desarrollo del mismo.

2.1 Web Scraping

La recopilación automatizada de datos en Internet es casi tan antigua como el propio Internet. Si bien es cierto que el *web scraping* no es un término nuevo, en los últimos años esta metodología ha ido cogiendo fuerza. Esta práctica recibe numerosos nombres entre los que destacan *screen scraping*, *data mining* o *web harvesting*.

Teóricamente el *web scraping* se puede definir como un programa o código diseñado para automatizar la descarga y el análisis de los datos de la web (Patel 2020). La forma más habitual de hacerlo es crear un programa que sea capaz de consultar un servidor web, solicitar los datos (generalmente en forma de HTML) y luego analizar esos datos para extraer la información necesaria.

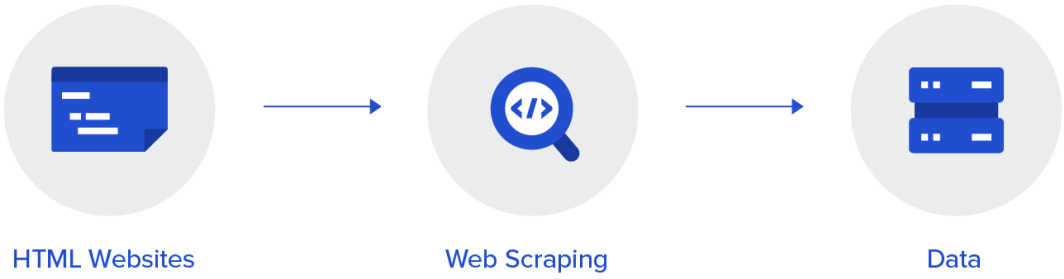

Ilustración 1 *Proceso del web scraping*
Fuente: Barnett 2020

Quizás la mejor forma de entender de que se trata es ver algunos ejemplos para los que se utiliza. En los *e-commerce* el precio es uno de los factores más importantes para los clientes por eso cualquier cambio en los precios por parte de la compañía o de la competencia puede tener un impacto directo en el devenir del negocio. Para poder optimizar al máximo su estrategia de fijación de precios, es posible que las empresas necesiten realizar un seguimiento de los precios en el mercado y compararlos con los suyos varias veces al día. Afortunadamente, recopilar toda esa información ya no es una tarea laboriosa ni tediosa, sino que ahora gracias al *web scraping* se automatiza. De esta forma las empresas son capaces de tomar mejores decisiones acerca de su estrategia y generar más rentabilidad para su negocio (Uses of Web Scraping s. f.).

No obstante, aparte de monitorizar el precio fijado por la competencia, las empresas deben ser capaces de responder a los diferentes estados de ánimo de sus clientes. Se requiere analizar y medir los sentimientos para poder descubrir que campañas tienen éxito entre los clientes, para proteger y mejorar la percepción de la marca o para interactuar con la audiencia de una forma mucho más efectiva. Gracias al *web scraping,* las empresas pueden recopilar información de redes sociales como Twitter, Facebook o Instagram, así como reseñas en diferentes plataformas especializadas en ello (Suganya y Vijayarani 2020)

Otro de los usos principales del *web scraping*, y en el que se enmarca este trabajo, es el del sector inmobiliario. Esta nueva herramienta les permite crear su propia base de datos con un largo listado de diferentes propiedades. Además, estos datos se pueden actualizar diariamente ya que los agentes y propietarios publican nuevas propiedades todos los días. Todos estos datos ayudan a los agentes inmobiliarios a encontrar propiedades que coincidan mejor con los requisitos y el presupuesto del cliente. Por otro lado, pueden controlar las tendencias de precios o los rendimientos de los alquileres según la ubicación y las comodidades que estos presenten.

Ahora que ya se tiene una idea más aproximada de lo qué es y para que se utiliza, se procede a explicar los cuatro tipos más comunes de raspadores (Diouf et al. 2019):

- ✓ Extensiones para el navegador: se caracterizan por su sencillez ya que solo hay que agregar una extensión al navegador que se utilice de forma habitual. Se utiliza especialmente para ejecutar pequeños proyectos debido a que solo pueden raspar una página web a la vez. Además, tampoco se pueden agregar funciones avanzadas que requieran de su ejecución fuera del navegador.
- ✓ Software: mediante esta opción de raspador el usuario ha de descargarse un software o programa descargable e instalarlo en su ordenador. Con este tipo de raspador sí que se pueden desempeñar funciones más avanzadas como extraer información de varias páginas a la vez o la integración de consejos de ayuda y sugerencias.
- ✓ Autoconstruido o prediseñado: es posible construir tu propio raspador, pero esto requerirá de conocimientos avanzados de programación por lo que a menos que se trate de un usuario experto programador, los raspadores autoconstruidos tendrán una funcionalidad limitada. No obstante, otro de los inconvenientes que tienen es que requieren de un mantenimiento continuo. Este será el tipo de raspador elegido para extraer toda la información relacionada con el mercado de la vivienda en alquiler de la ciudad de Valencia.
- ✓ Raspadores alojados en la nube: este tipo de herramientas operan desde un servidor externo proporcionado por su proveedor. De esta forma la capacidad del computador del usuario no limitará el proyecto ni tampoco impedirá que pueda seguir ejecutando otras tareas en segundo plano. Cuando se haya extraído toda la información el programa envía una notificación anunciando que ya están listos todos los datos para ser exportados. Suelen ser la opción más cara, pero a cambio permite el raspado masivo de datos.

Llegados a este punto es momento de analizar la legalidad y ética de esta práctica. Al fin y al cabo, el *web scraping* no deja de ser un proceso que consiste en copiar información de una página web para el uso propio con lo que a priori no parece muy clara la legitimidad de esta herramienta.

Esta problemática se plantea por primera vez en España con el caso de la compañía aérea Ryanair contra la empresa organizadora de viajes y diferentes experiencias Atrápalo. Esta última a lo que se dedica es a realizar búsquedas por las diferentes aerolíneas ofreciéndole al usuario un listado con las mejores opciones. Así pues, el gigante aéreo demandó a Atrápalo por los siguientes motivos (STS 572-2012 2012):

- ✓ Incumplimiento de las condiciones y términos de uso de la página web.
- ✓ Violación del derecho de propiedad intelectual al vulnerar su derecho al software desarrollado para generar información sobre precios y vuelos.
- ✓ Vulneración de su derecho sui géneris sobre sus bases de datos.
- ✓ Realización de actos de competencia desleal.

El 9 de octubre de 2012 la Sala Primera del Tribunal Supremo desestimó el recurso de casación presentado por la empresa Ryanair Limited. De esta manera se confirmaba la sentencia dictada en recurso de apelación pronunciada por la Audiencia de Barcelona de 15 de diciembre de 2009, interpuesta contra la sentencia del juzgado de lo Mercantil de Barcelona de 21 de enero de 2009, (autos 214/2008).

La sentencia de apelación confirmaba que no se habían infringido las condiciones de uso de la página web de Ryanair, ya que no había ningún tipo de relación contractual y el uso de dichas páginas era público para todos los usuarios. El uso de la web no supone ningún tipo de contrato de hecho, como argumentaba Ryanair. Según la sentencia, no se pueden imponer límites al uso que los usuarios hacen de la información de la web, ya que esta se diseña precisamente para ofrecer al público la información de Ryanair sin ningún tipo de restricciones. Solo en casos puntuales en que existieran derechos de exclusividad podría entenderse que hubiera límites.

En primer lugar, según Ryanair, la utilización de la Web supone de hecho un contrato en el que se aceptarían las condiciones impuestas por la titular de la página. En cambio las sentencias niegan la existencia de esta relación contractual entre Ryanair y Atrápalo, por lo que no hay una infracción de las condiciones de uso de la página Web. Según las sentencias, Ryanair no puede imponer límites al uso que los usuarios hagan de la información, ya que de hecho existe un ofrecimiento público de información de facto sin restricciones. Esos límites solo existirían en los casos fundados en derechos de exclusividad.

Por lo que se refiere al derecho de Propiedad Intelectual, la sentencia rechaza que se hubiese infringido dicho derecho. Basándose en la premisa que establece la Directiva 96/9 de que en que para que los fabricantes de bases de datos tengan un “derecho sui generis” no se establece en función de la “originalidad” sino el de la “inversión”. Si bien reconoce que puede existir una inversión importante, esta se ha realizado en crear un software que genera información, no en recopilar información y crear una base de datos. Por lo tanto lo que es original es el software que se ha utilizado, no la página web que solo genera información con los datos que quiere difundir la empresa para vender sus productos. No existe una base de datos recopilados, sino que son datos proporcionados por la empresa a los posibles clientes (vuelos, precios, etc). Por lo tanto en este caso no existe originalidad.

Para que una base de datos esté protegida por el derecho de autor tiene que tener una estructura donde la selección y disposición de sus contenidos constituya una creación intelectual. La protección del derecho de autor afectaría a la estructura de la base de datos, donde residiría la originalidad, no a su contenido, ni por lo tanto a los elementos constitutivos de esta.

En tercer lugar, se rechaza que se haya vulnerado el derecho sui generis. La normativa europea, con la finalidad de fomentar y proteger las inversiones en los sistemas de almacenamiento y tratamiento de datos que contribuyan al desarrollo del mercado de la información, reconoció a los fabricantes de bases de datos un derecho "sui generis" no a la originalidad, sino a la "inversión". Por lo tanto, el derecho "sui generis" sobre una base de datos protege la inversión económica, evaluada cualitativa o cuantitativamente, que realiza su fabricante ya sea de medios financieros, tiempo, esfuerzo y energía utilizados, o inversiones de la misma naturaleza, con la finalidad de obtener, verificar o presentar su contenido.

De acuerdo con ello, la Ley 5/1998, de 6 de marzo, considera como bases de datos "las colecciones de obras, de datos o de otros elementos independientes dispuestos de manera sistemática o metódica, y accesibles individualmente por medios electrónicos o de otra forma". Por otra parte, la sentencia de 9 de noviembre de 2004 del Tribunal de Justicia Europeo confirma los requisitos básicos que debe reunir una bases de datos: 1) que exista una recopilación de elementos independientes, es decir, separables unos de otros, sin que resulte afectado el valor de su contenido informativo; 2) que los elementos estén dispuestos de forma sistemática o metódica; 3) que esté dotada de algún instrumento técnico que permita la localización de cualquier elemento independiente; 4) que resulten accesibles individualmente de una u otra manera.

Por lo tanto, para la tutela del derecho "sui generis" no es necesario que la base de datos sea original, o que los datos contenidos en la base estén tutelados por la protección dispensada por el derecho de autor, sino que es suficiente que suponga una inversión económica que sea necesario tutelar por el esfuerzo de recopilación realizado.

Por último, en referencia al cuarto fundamento del recurso, "-existencia de ilícito concurrencial por desleal aprovechamiento del esfuerzo ajeno"- la sentencia rechaza que haya competencia desleal. Confirma, apoyándose como precedentes en la STJ de 5 de marzo de 2009, Apis-Hristovich EOOD contra LAKORD AD, C_545/07, que no existe una base de datos basada en recopilación de información ya existente, por lo que no existe extracción de la misma. En este caso, la técnica del *screen scraping* sería perfectamente legal, ya que no supone ningún perjuicio económico por la inversión realizada en la base de datos.

En el caso de la web de Ryanair se trata de una información basada en la creación de datos. Esta información se gestiona mediante un software donde Ryanair incorpora a su web los datos de sus productos y los ofrece a los clientes.

La agencia de viajes lo que hace es facilitar, con ánimo de lucro, el acceso de sus propios clientes a las ofertas que Ryanair que ofrece desde su web.

Son los clientes, y no la agencia de viajes Atrápalo, los que utilizan los datos de Ryanair y los que contratan sus productos. La empresa Atrápalo se limita a cobrar un sobreprecio como intermediaria (que los califica el fallo como lícitos, ya que son sus servicios).

En conclusión, la técnica del web scraping es una práctica totalmente legal, pero como muchas otras herramientas puede ser ilegal dependiendo según el uso que se haga de ella. Hay que presentar especial atención a todos los temas relacionados con la propiedad intelectual, así como no violar ningún derecho de autor. Además, es clave no realizar ninguna acción que pueda incurrir en competencia desleal.

2.2 Aprendizaje automatizado

En las últimas dos décadas, el aprendizaje automático se ha convertido en uno de los pilares de la tecnología de la información y, con ello, en una parte bastante central, aunque normalmente oculta, de nuestra vida. Con la cantidad cada vez mayor de datos disponibles, hay razones suficientes para creer que el análisis inteligente de datos se volverá aún más omnipresente como un ingrediente necesario para el progreso tecnológico.

Sin embargo, el aprendizaje de las máquinas no es algo nuevo, ha existido por lo menos desde la década de 1970. Fue en ese momento cuando aparecieron los primeros algoritmos que ya permitían resolver algunos problemas sencillos. No obstante, ha sido la inmensa potencia de cálculo que tienen hoy en día los computadores junto con la gran cantidad de datos disponible lo que ha marcado la diferencia. Gracias a esa potencia junto con la multitud de información utilizable es lo que permite que se puedan abordar problemas mucho más complejos y en una gama cada vez más amplia de dominios.

La idea general que subyace a la mayor parte del aprendizaje automático o *machine learning* es que un ordenador aprende a realizar una tarea estudiando un conjunto de ejemplos de entrenamiento. A continuación, el ordenador o el sistema de ordenadores realiza la misma tarea con datos que son totalmente desconocidos (Louridas y Ebert 2016).

Son muchos los tipos de algoritmos de aprendizaje automatizado, entre todos ellos destacan especialmente dos: aprendizaje supervisado y no supervisado. Sin lugar a duda, el aprendizaje supervisado es el uso más común en el ámbito del *machine learning*. Este tipo de aprendizaje es un enfoque de aprendizaje automático que se define por el uso de conjuntos de datos etiquetados. Estos conjuntos de datos están diseñados para poder ser entrenados o supervisados mediante algoritmos que clasifiquen todos estos datos o predigan resultados con precisión. Por otro lado, el aprendizaje no supervisado utiliza algoritmos de aprendizaje automático para analizar y agrupar conjuntos de datos no etiquetados. Este tipo de algoritmos descubren patrones ocultos en los datos sin necesidad de intervención humana, de ahí su nombre (Supervised vs. Unsupervised Learning 2021). Por tanto, la principal diferencia entre ambos enfoques es el uso o no de conjuntos de datos etiquetados. Sintetizando, el aprendizaje supervisado utiliza datos de entrada y salida etiquetados, mientras que un algoritmo de aprendizaje no supervisado no lo hace.

Asimismo, cada uno de estos tipos de aprendizaje aún puede desagregarse en otros grupos más pequeños. Para el contexto de este trabajo solo son importantes los algoritmos supervisados por lo que serán estos los que se van a estudiar en profundidad. En concreto, este tipo de aprendizaje aún puede desagregarse en otros dos subgrupos. Por un lado, se encuentran los problemas de clasificación que utilizan un algoritmo para asignar los distintos datos en categorías específicas. Un ejemplo de aplicación de este tipo de problema es la intención de querer clasificar una multitud de clientes entre aquellas que prefieren comprar online de los que prefieren hacerlo en una tienda física. Del otro lado, se hallan los modelos de regresión los cuales se utilizan para comprender la relación entre las variables dependientes e independientes. Estos son útiles para predecir valores numéricos basados en diferentes puntos de datos como pueden ser las ventas de una determinada empresa en los próximos meses.

Probablemente el algoritmo de regresión más famoso sea el de regresión lineal, el cual no necesita mayor explicación. Se trata de un modelo que predice el valor de una variable en función del valor de otra variable y donde dicha regresión se ajusta a una línea recta o superficie que minimiza la diferencia entre los valores de salida reales y los predichos.

No obstante, con el avance de la tecnología se han desarrollado otros modelos que a veces pueden ser computacionalmente muy complejos pero muy precisos. El primero de ellos es el *decision tree* o árbol de decisión. Un árbol de regresión puede considerarse como una variante de este tipo de algoritmo ya que está diseñado para aproximar funciones de valor real, en lugar de usarse para métodos de clasificación. Este tipo de árbol se construye a través de un proceso conocido como partición recursiva binaria, que es un proceso iterativo que divide los datos en particiones o ramas, y luego continúa dividiendo cada partición en grupos más pequeños a medida que el método va avanzando en cada rama. Inicialmente, todos los registros del conjunto de entrenamiento se agrupan en la misma partición. A continuación, el algoritmo comienza a asignar los datos en las dos primeras particiones o ramas, utilizando todas las divisiones binarias posibles en cada campo. El algoritmo selecciona la división que minimiza la suma de las desviaciones al cuadrado de la medida en las particiones separadas. Esta regla de división se aplica luego a cada una de las nuevas ramas. Este proceso continúa hasta que cada nodo alcanza un tamaño de nodo mínimo especificado por el usuario y se convierte en un nodo terminal (Myles et al. 2004).

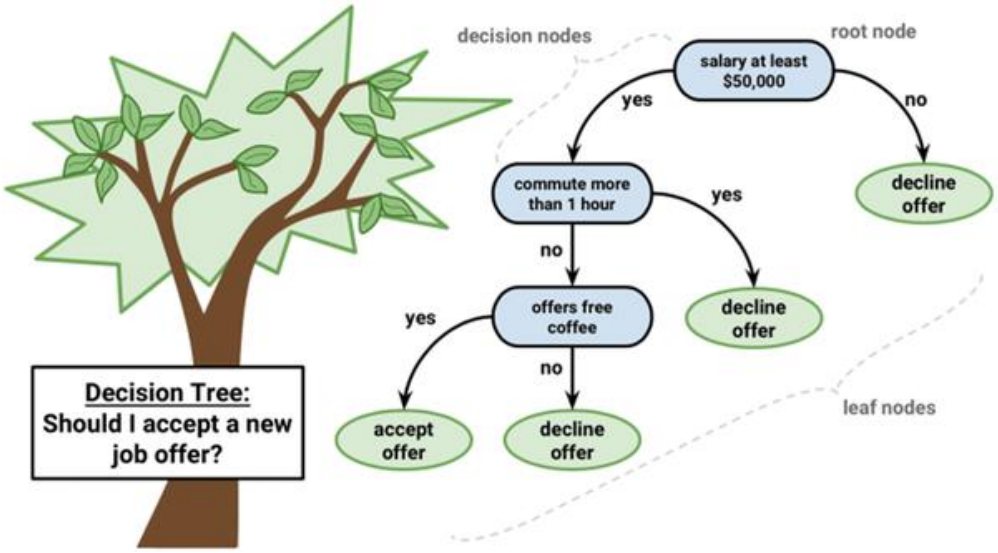

Ilustración 2 *Ejemplo de árbol de decisión*
Fuente: Varghese 2019

Dado que el árbol crece a partir del conjunto de entrenamiento, un árbol completamente desarrollado sufre de sobreajuste (explicar elementos aleatorios del conjunto de entrenamiento que probablemente no sean característicos de la población más grande). Para corregir esta tendencia de los árboles de decisión se han creado nuevos algoritmos que, manteniendo su esencia, funcionan mejor según determinadas circunstancias.

El primero de ellos es el bosque aleatorio o *random forest* el cual se basa en desarrollar múltiples arboles de decisión que se fusionan para obtener una predicción más precisa. La lógica que hay detrás de este es que múltiples modelos no correlacionados (arboles de decisión individuales) funcionan mucho mejor como grupo en vez de solos. Por tanto, en el caso de los modelos de regresión este algoritmo elige el promedio de las salidas de todos los árboles. El hecho diferenciador de esta forma de proceder es que, si bien los árboles de decisión individuales pueden producir errores, el resto de los otros árboles seguirá estando en lo correcto por lo que ese error quedará aislado y no afectará al resultado final.

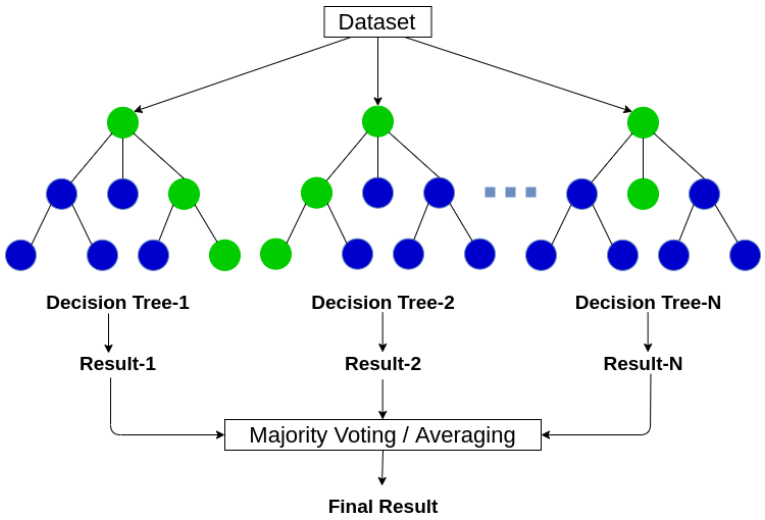

Ilustración 3 *Esquema del algoritmo random forest*
Fuente: Abhishek 2020

El segundo de ellos y, además el de más reciente creación, es el *xgBoost.* Su principal diferencia frente al bosque aleatorio reside en el hecho de que se trata de un algoritmo secuencial. De esta manera unos árboles se van agregando secuencialmente a los anteriores para poder aprender del resultado obtenido de los árboles previos. Esto permite ir corrigiendo los errores de cada uno de los árboles previos hasta que ese error sea mínimo (Espinosa-Zúñiga 2020).

Por último, otros de los algoritmos más famosos es el *k-nearest Neighbors* (KNN) este se usa tanto en problemas de regresión como en problemas de clasificación. Es un modelo de aprendizaje automático supervisado no paramétrico que almacena todos los datos disponibles y predice nuevos basándose en una métrica de similitud elegida. La idea es predecir el valor del nuevo caso basándose en los k valores más cercanos disponibles con respecto a la métrica de similitud.

CAPÍTULO 3

Metodología

En este capítulo se recoge todos aquellos pasos que han sido necesarios hasta obtener el modelo de regresión final. Primero se realizará el respectivo proceso de *web scraping*, a continuación, se preprocesarán y transformarán todos los datos previamente recolectados. El siguiente paso consistirá en análisis profundo de todas las variables que componen la base de datos. Para acabar formulando los distintos modelos de regresión que sean capaces de explicar el precio en alquiler de una vivienda en la ciudad de Valencia.

3.1 Obtención de la base de datos

Para poder realizar un análisis exhaustivo del precio del alquiler en la ciudad de Valencia, primero se requiere de la recogida y confección de una base de datos que sea representativa de la situación actual en lo que se refiere al arrendamiento de viviendas. En la actualidad son muchas las plataformas que permiten la búsqueda de pisos, entre ellas destacan: idealista, fotocasa y enalquiler. Finalmente se ha decido avanzar en el proceso con la web de idealista ya que es la que mayor oferta presenta, así como la que mayores facilidades ofrece a la hora de realizar el proceso de *web scraping*.

Cada uno de los anuncios de esta página tienen una apariencia similar a la que se presenta en la ilustración 4. Como se puede apreciar, a la izquierda del anuncio hay un carrusel fotográfico mediante el cual el usuario se puede hacer una idea aproximada de la vivienda en cuestión. A continuación, en el lado derecho se encuentra toda la información descriptiva del piso. Lo primero que se lee es la dirección de este y a continuación, separado por una coma el barrio en el que se encuentra ubicado. En este momento es importante mencionar que una ciudad normalmente es dividida en distritos y dentro de cada uno de esos distritos encontramos diferentes barrios. Por lo que a priori del anuncio solo podríamos extraer el barrio, más adelante se explicará cómo se ha solucionado este pequeño inconveniente. En la segunda línea, se muestra el precio que tiene ese arrendamiento mensualmente. En todos los anuncios el importe viene fijado en términos mensuales. Más abajo, en el siguiente reglón aparece detallado el número de habitaciones, los metros cuadrados, la planta, si es un piso exterior o interior y si tiene ascensor o no. Seguidamente se puede leer una pequeña descripción en la que el anunciante enfatiza en las características más atrayentes o relevantes del domicilio. Por último, también se incluye en el caso de que la hubiera, una foto de la inmobiliaria encargada de la gestión y un teléfono de contacto.

Ilustración 4 *Anuncio de la plataforma Idealista*
Fuente: Idealista (2021)

Ahora ya sí, conociendo la estructura y disposición de los anuncios en el portal de idealista se puede empezar con el proceso de *web scraping*. Para dicha tarea se ha desarrollado el código que se adjunta en el anexo II, el cual básicamente tiene tres partes o funcionalidades clave. La primera de ellas tiene como objetivo ser capaz de recorrer las diversas páginas en las que hay anuncios y no solo la que corresponde con el url facilitado. Trata de simular algo parecido a la acción que realiza el usuario cuando quiere seguir viendo ofertas y pulsa la tecla que se suele encontrar al final de la página denominada siguiente. Por otro lado, la segunda parte hace referencia a la recogida propiamente dicha de los datos. El proceso se ha diseñado para que almacene la información de las cuatro primeras líneas, es decir, dirección, precio, número de habitaciones, metros, planta y descripción.

En este momento es importante hablar de cómo se ha solucionado la pequeña problemática del distrito comentada previamente. La solución por la que en última instancia se ha optado es por filtrar los anuncios por distritos y por tanto repetir el proceso tantas veces como distritos tiene la ciudad de Valencia. De esta forma, aunque el proceso es un poco más laborioso, se asegura poder contar sin margen de error de este dato. Para acabar, la última parte corresponde al proceso de almacenamiento de toda esa información que se ha recogido en la etapa previa. En definitiva, toda esa información es almacenada en un archivo tipo excel donde columna a columna se guardan cada una de las características del alojamiento anunciado en el portal de idealista.

Al finalizar el proceso de *web scraping* se obtiene un archivo excel con apariencia similar a la que se muestra en la ilustración 5.

	A	B	C	D	E	F
1						
2		Columna1	Columna3	Columna4	Columna5	Columna6
3		Piso en Museo, El Carme, València	950€/mes	2 hab. 65 m² Planta 3ª exterior sin ascensor 4 horas	SC inmobiliaria oficina Juan Llorens 47, ofrece en alquiler vivienda en pleno centro de El Carmen: C/ Museu. A 100 m del cauce del río. 3...	Ciutat Vella
4		Piso en calle del Palau, 7, La Seu, València	900€/mes	2 hab. 95 m² Planta 4ª exterior sin ascensor Publicado ayer	Se alquila Piso con encanto. Luminoso. Espacios amplios. En edificio emblematico del Barrio de la Seu (En pleno centro de Valencia). Tot...	Ciutat Vella
5		Piso en calle de l'Àngel Custodi, 10, El Carme, València	725€/mes	2 hab. 60 m² Planta 1ª exterior con ascensor Publicado ayer	Alquilo piso en el barrio del Carmen, muy bien situado, cerca de las Torres de Serrano, a gente responsable y cuidadosa. El piso está com...	Ciutat Vella
6		Piso en calle de Salvador Giner, El Carme, València	1.100€/mes	3 hab. 103 m² Planta 3ª exterior con ascensor Publicado ayer	Piso en alquiler luminoso. En el Carme, Ciutat Vella, València. Piso Ubicado al antiguo cauce del rio Turia. Cerca de todos los sitios de...	Ciutat Vella
7		Piso en La Xerea, València	750€/mes	2 hab. 85 m² Planta 3ª exterior sin ascensor	Se alquila fantastica vivienda reformada y decorada con mucho gusto combinando moderno con clásico respetando el encanto de sus ladrillos...	Ciutat Vella

Ilustración 5 *Base de datos obtenida tras el proceso del web scraping*
Fuente: Elaboración propia

El siguiente paso es transformar todos esos datos de manera que puedan ser soportados y correctamente analizados por un software estadístico. Para ello se procede a explicar columna a columna como se ha ido segmentado la información.

En la primera de ellas aparece toda la información que está relacionada con la ubicación de la vivienda. La primera palabra se ha utilizado para determinar el tipo de vivienda en cuestión, obteniendo de esta manera la primera variable de la base de datos. En la ilustración 5 todas las filas corresponden con pisos, pero se han encontrado áticos, dúplex, casas, entre otras. Seguida de esta primera palabra y a continuación de la preposición en, se detalla la calle donde está situada la vivienda. Por lo que todo el texto comprendido entre la preposición y la siguiente coma equivaldría a la segunda de las variables: la calle. Por último, falta extraer la variable barrio para ello se ha empezado la segmentación de los datos por la parte izquierda. Se ha separado en primer lugar la ciudad que es este caso no tiene mayor interés ya que la base de datos solo incluye viviendas de la ciudad de València. Con lo que, descartando esta información, entre comas solo queda el número del piso (no siempre se facilita) y el nombre del barrio en cuestión. Por tanto, ahora ya sí se puede obtener la tercera de las variables, simplemente extrayendo el texto que queda entre comas empezando lo más a la izquierda posible. En conclusión, de esta primera columna se ha podido obtener el tipo, la calle y el barrio de la vivienda.

A continuación, se encuentra el precio del domicilio mes a mes. En este caso no hay mayor dificultad lo único que hay que hacer es aislar los caracteres alfanuméricos dado que la moneda siempre es la misma y la periodicidad del pago también.

En la tercera columna se enumeran las características más relevantes de la vivienda anunciada. La primera variable que se puede extraer es el número de habitaciones y para ello simplemente hay que coger el primer número que aparece. A continuación, se quiere obtener el número de metros cuadrados que tiene el piso, en este caso esta información son aquellos números que se encuentran comprendidos entre el punto y la m de m^2. Seguidamente se encuentra una palabra que hace referencia a donde se encuentra la vivienda ya sea en una planta, en un bajo o en el entresuelo. Por tanto, esta será la siguiente variable que extraer, el tipo de planta donde se aloja el piso. Inmediatamente seguido se detalla el número de planta donde se sitúa la vivienda, otro detalle que es interesante incluir en la base de datos. Así pues, en esta columna también aparece detallado si se trata de una vivienda exterior o por el contrario es interior. El hecho de que sea un piso interior y, por tanto, que todas las habitaciones estén orientadas hacia un patio interior puede ser un detalle importante a la hora de establecer el precio por lo que será la siguiente variable a incluir en el dataset. Por último, se especifica si la vivienda cuenta con ascensor o no, es importante fijarse que el hecho que marca si hay o no es la preposición (con o sin ascensor). Esta será la última variable que se va a obtener de dicha columna dado que el momento en el que se haya publicado el anuncio no es relevante para el análisis del precio de la vivienda. En resumidas cuentas, se han obtenido un total de seis variables: número de habitaciones, metros cuadrados, tipo de planta, número de planta, tipo de luz y la existencia o no de ascensor.

La penúltima información que se ha conseguido obtener del proceso de *web scraping* ha sido la descripción que ha puesto el anunciante sobre la vivienda. En este caso analizar dicha variable es un proceso más largo y laborioso. No se trata de únicamente ir segmentado la información, sino que hay que analizar el contenido y tratar de buscar los patrones más repetidos. Por ese motivo se abordará en el siguiente paso de preprocesado de los datos.

Para finalizar solo queda hablar del distrito del piso, aquí no hay mayor dificultad ya que no necesita ninguna modificación. Así pues, tras estas primeras transformaciones la base de datos pasaría de tener un aspecto como se mostraba en la ilustración 6 a el que se muestra a continuación:

TIPO	CALLE	BARRIO	PRECIO	HABITACIONES	METROS	ASCENSOR	LUZ	DESCRIPCIÓN	ZONA	TIPO PLANTA	N_PLANTA
Piso	Museo	El Carme	950	2	65	Sin ascensor	Exterior	SC inmobiliaria oficina Juan Llorens 47, ofrece en alquiler vivienda en pleno centro de El Carmen: C/ Museu. A 100 m del cauce del río. 3...	Ciutat Vella	Planta	3
Piso	calle del Palau	La Seu	900	2	95	Sin ascensor	Exterior	Se alquila Piso con encanto. Luminoso. Espacios amplios. En edificio emblematico del Barrio de la Seu (En pleno centro de Valencia). Tot...	Ciutat Vella	Planta	4
Piso	calle de l'Àngel Custodi	El Carme	725	2	60	Con ascensor	Exterior	Alquilo piso en el barrio del Carmen, muy bien situado, cerca de las Torres de Serrano, a gente responsable y cuidadosa. El piso está com...	Ciutat Vella	Planta	1
Piso	calle de Salvador Giner	El Carme	1100	3	103	Con ascensor	Exterior	Piso en alquiler luminoso. En el Carme, Ciutat Vella, València. Piso Ubicado al antiguo cauce del rio Turia. Cerca de todos los sitios de...	Ciutat Vella	Planta	3
Piso		La Xerea	750	2	85	Sin ascensor	Exterior	decorada con mucho gusto combinando moderno con clásico respetando el encanto de sus ladrillos...	Ciutat Vella	Planta	3

Ilustración 6 *Base de datos con todas las variables transformadas*
Fuente: Elaboración propia

3.2 Preprocesamiento

3.2.1 Variable descripción

A la hora de analizar textos el primer paso siempre es la *tokenización*, este el proceso por el cual se dividen cadenas de texto más largas en piezas más pequeñas o tokens. En nuestro caso vamos a segmentar la variable descripción en una lista de palabras. Para realizarlo se requieren de una serie de librerías específicas tal como se puede apreciar en el anexo II.

Para poder crear una lista con todas las palabras que aparecen en las distintas descripciones de los pisos se necesitan hacer varias tareas previas:

- Quitar las mayúsculas: para alcanzar una mayor homogenización se transforman todas las palabras a minúsculas. De esta manera se evita que haya palabras con el mismo significado pero que en vez de aparecer una sola vez aparezcan duplicadas por estar escritas de formas diferentes. Por ejemplo, la palabra piso aparece escrita de diversas formas: piso, Piso y PISO.
- Quitar los signos de puntuación: se ha procedido a eliminar todas las comas, puntos, puntos y comas… ya que no aportan nada al estudio de esta variable.
- Quitar las palabras vacías: se han prescindido de todas las *stop words*, estas son aquellas palabras que carecen de sentido cuando se escriben solas. Básicamente se trata de las conjunciones, artículos, preposiciones y adverbios.

Una vez ya se ha generado toda una lista con todas las palabras se ha procedido a analizar cuáles son las más repetidas. Como era esperable las más reiteradas han sido: piso, vivienda, alquiler y Valencia. Además, se ha generado una nube de palabras como se puede ver en la ilustración 7 en la que se empiezan a ver otras palabras que pueden resultar interesantes ya que podrían añadir más especificaciones sobre el piso anunciado.

Ilustración 7 *Nube de palabras más utilizadas con la variable descripción*
Fuente: Elaboración propia

La primera característica que se ha decido obtener a partir de la variable descripción ha sido el hecho de si el piso ha sido reformado o es de nueva construcción. Para ello se ha creado una lista de palabras entre las que se incluye todas las palabras derivadas de reforma, así como palabras como a estrenar, de nueva obra o de nueva construcción. Una vez creada se ha buscado la aparición de alguna de ellas en la variable descripción y en el caso de que existiera para un determinado anuncio se ha puesto un sí en la nueva variable reformado que se ha creado.

Otra particularidad de la vivienda que se ha creído interesante estudiar es si el alquiler incluye garaje o no. Se ha procedido de forma similar, aunque en este caso la lista de palabras es mucho más reducida ya que solo se ha buscado garaje y aparcamiento. Por último, se ha estudiado también si la vivienda posee o no piscina.

Finalmente, mencionar que si al buscar alguna de las características en la descripción no aparecía se ha puesto la variable a no ya que se han considerado funcionalidades de la casa lo suficientemente importantes como para aparecer en la descripción en el caso de poseerlas. En definitiva, después de finalizar este análisis se ha conseguido extraer tres aspectos añadidos a los ya especificados en el anuncio del piso.

3.2.2 Datos anómalos

Una de las etapas fundamentales en el preprocesamiento de los datos es el análisis y tratamiento de *outliers* o valores anómalos. Estos son aquellas observaciones anormales y extremas que se encuentran en una muestra estadística o serie temporal de datos que pueden afectar potencialmente a la estimación de los parámetros de este (Marco Sanjuán 2018).

Para la detección de estos valores se ha procedido a utilizar el test de Tukey, el cual toma como referencia la diferencia entre el tercer y primer cuartil, también conocido como rango intercuartílico. Por tanto, siendo q_1 y q_3 el primer y tercer cuartil y IQR el rango intercuartílico, será un valor atípico leve aquel que:

$$l_{min} < q_1 - 1{,}5IQR$$

$$l_{max} > q_3 + 1{,}5IQR$$

Mientras que se calificara como valor atípico extremo aquel que:

$$l_{min} < q_1 - 3IQR$$

$$l_{max} > q_3 + 3IQR$$

l_{min} y l_{max} determinan respectivamente los límites inferiores y superiores a partir de los cuales se considera un outlier. Dado que se ha partido de un tamaño de muestra bastante grande solo se tendrán en cuenta los límites extremos para imputar un dato como outlier.

Tras calcularlos con ayuda del software R se obtiene que el límite superior para el precio es 2800€ y para los metros cuadrados 287m^2. Otra forma gráfica de localizarlos es mediante el diagrama de caja, bigotes o *box whisker*.

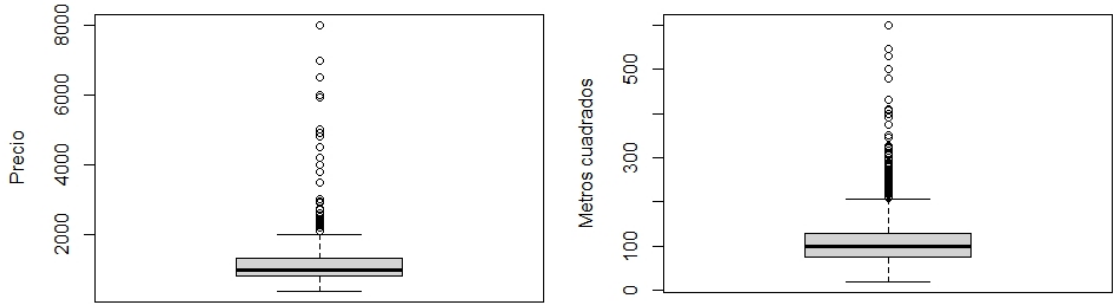

Gráfico 1 *Diagrama de bigotes de la variable precio y metros cuadrados*
Fuente: Elaboración propia

Así pues, se ha procedido a eliminar aquellas viviendas que tenían un precio superior a 2800€ y/o una superficie construida superior a 287 m^2. En total se han eliminado 67 filas de la base de datos lo que supone un 3%.

3.2.3 Valores duplicados y valores faltantes

Otro de los pasos fundamentales del preprocesamiento es eliminar todos aquellos valores duplicados para evitar así información redundante. En nuestro caso se considerará que dos filas son duplicadas cuando tengan el mismo valor en todos los campos o variables, es decir, cuando sean completamente iguales en todos los aspectos. En un primer momento se había pensado en buscar filas duplicadas únicamente mediante el atributo descripción ya que parecía bastante complicado que dos pisos tuvieran la misma descripción. Sin embargo, se ha observado que en ocasiones aquellos anuncios que son gestionados por una inmobiliaria muchas veces la descripción es la misma independientemente de la vivienda.

Tras buscar las distintas filas duplicadas con el código que se puede ver en el anexo II, se concluye que hay un total de 22 anuncios duplicados lo que supone un 1% de la muestra.

Por último, para acabar con la fase de preprocesamiento queda analizar los diferentes valores faltantes que hay en la base de datos. Tal como se puede observar en la tabla resumen 1 la variable calle tiene un 27,12% de valores faltantes siendo esta la que mayor porcentaje presenta. Al ser una variable que es muy difícil aplicarle algún tipo de estrategia para el tratamiento de valores faltantes se ha decidido descártala y no tenerla en cuenta en el posterior análisis. A continuación, las variables con mayor cantidad de valores faltantes son las relacionadas con la planta de la vivienda. Seguidamente las otras variables con mayor cantidad son las que tienen que ver con el tipo de luz y la posesión de ascensor. Así pues, en este momento es importante destacar que en la base de datos como se explicará en el siguiente apartado hay alrededor de un 1,5% de casas, chalets y cortijos. Para este tipo de viviendas nunca se especifica ninguna de las cuatro características (ascensor, luz, tipo de planta y planta).

Una vez conocedores de esta información y después de prescindir de la variable calle, se ha decido seguir adelante con el resto de las variables asumiendo el pequeño porcentaje de valores faltantes que tienen algunas de ellas. De esta forma al acabar el proceso de preprocesamiento la base de datos está formada por un total de 2135 filas y 14 variables.

VARIABLE	DESCRIPCIÓN	TIPO VARIABLE	% NA
tipo	Tipo de vivienda (piso, casa, estudio, ático…)	Variable cualitativa	0,00%
calle	Calle en la que se encuentra la vivienda	Variable cualitativa	27,12%
barrio	Barrio al cual pertence (aiora, gran vía, el carme…)	Variable cualitativa	0,00%
precio	Precio del coste mensual en € del alquiler de la vivienda	Variable cuantitativa	0,00%
habitaciones	Número de habitaciones de la vivienda (1,2,3..)	Variable cuantitativa	0,37%
metros	Metros cuadrados que tiene la vivienda (100, 200, 150…)	Variable cuantitativa	0,00%
ascensor	Indica si la vivienda tiene o no ascensor (con ascensor o sin ascensor)	Variable cualitativa	2,06%
luz	Tipo de luz que tiene la vivienda (exterior o interior)	Variable cualitativa	2,30%
descripción	Descripción de la vivienda	Variable cualitativa	0,42%
distrito	Distrito al que pertence la vivienda (ciutat Vella, eixample, campanar…)	Variable cualitativa	0,00%
tipo planta	Tipo de planta de la vivienda (bajo, entresuelo, planta…)	Variable cuantitativa	4,82%
planta	Número de planta en la que se encuentra la vivieda (1,2,5…)	Variable cualitativa	4,82%
piscina	Indica si la vivienda tiene o no piscina (sí o no)	Variable cualitativa	0,00%
garaje	Indica si la vivienda tiene o no garaje (sí o no)	Variable cualitativa	0,00%
reformado	Indica si la vivienda está reformada o no (sí o no)	Variable cualitativa	0,00%

Tabla 1 *Resumen de la base de datos*
Fuente: Elaboración propia

3.3 Análisis individualizado de las variables

Tras acabar con el preprocesamiento de los datos es momento de analizar en profundidad el contenido de la base de datos creada. Para ello en este primer apartado se va a estudiar cada variable de forma aislada con el objetivo de hacerse una idea del panorama general del mercado de alquiler en Valencia.

3.3.1 Variable tipo de vivienda

Con esta primera variable se pretende enmarcar a la vivienda anunciada en el portal web en un formato especifico. En total, se han obtenido un conjunto de siete opciones posibles: ático, casa, chalet, cortijo, dúplex, estudio y piso. Tal como se puede observar en el gráfico 2 el que más abunda con una notable diferencia es la vivienda tipo piso, representa el 85,58% de la muestra. El segundo más numeroso es el ático, el cual supone un 7,35% de la muestra y por tanto, se puede comprobar como el resto apenas representan el 7%.

En este punto es importante mencionar que tras repasar los anuncios de las viviendas que se han catalogado como chalet, no corresponden con la propia definición de este tipo de alojamiento. Comúnmente se ha definido a estos como aquella vivienda que es, de al menos una planta, y que posee jardín. Sin embargo, se han encontrado anuncios catalogados como chalets

que se encuentren en pleno barrio del Carmen y que por la tanto la posibilidad de tener jardín en ellos es prácticamente nula. Por ende, muchos de los domicilios que se han definido como chalets en realidad son casa. Siendo conocedores de este pequeño fallo se ha decido continuar con el análisis sin hacer una reagrupación ya que hay un total de 23 chalets y tan solo 6 casas con lo que la distorsión de la muestra es prácticamente nula.

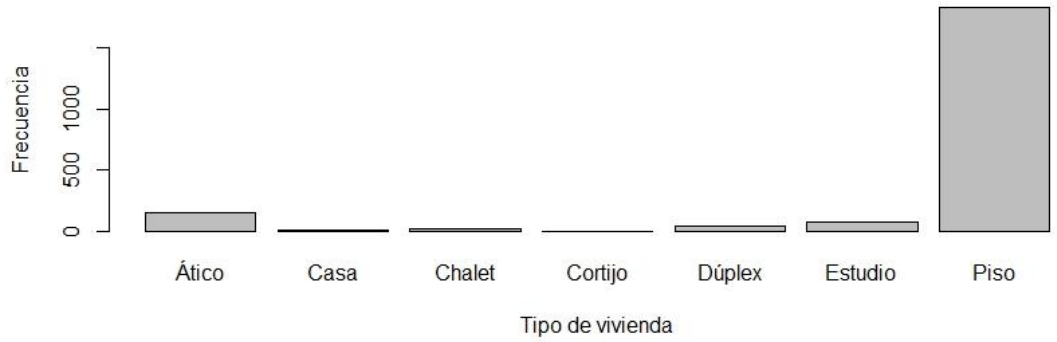

Gráfico 2 *Diagrama de barras de la variable tipo de vivienda*
Fuente: Elaboración propia

3.3.2 Variable barrio y distrito

Se ha decidido analizar estas dos variables juntas ya que una es la agregación de la otra por lo que en este caso tiene más sentido analizarlas en conjunto. La ciudad de Valencia está formada por un total de 19 distritos sin embargo, en nuestro dataset solo hay 16. Han quedado fuera los siguientes distritos: pobles del Nord, pobles de l'oest y pobles del sud. El único motivo por el que no aparecen es porque o bien no hay propiedades en alquiler en esos distritos o bien porque, aunque sí que las hay en el momento de realizar el *web scraping* no había disponibles. Por otro lado, en lo que a barrios se refiere para el total de distritos se pueden distinguir un total de 73 barrios. Además, se ha podido comprobar como por normal general cada distrito está formado por un mínimo de dos barrios y un máximo de siete.

Los distritos que mayor cantidad de viviendas en alquiler poseen son: Ciutat Vella, Eixample, Poblats Marítims y Extramurs. Tres de los cuales se encuentran ubicados en pleno centro de la ciudad. Los barrios con mayor oferta son: Sant Francesc, Russafa y el Cabanyal, todos ellos ubicados dentro de los distritos ya enunciados con mayor mercado de alquiler. Por otro lado, tal como muestra la tabla 2, en la parte baja de la lista se encuentra los distritos de Benicalap y Jesús donde apenas se agrupa un 2,5% de la oferta. Asimismo, los barrios de la Fontenta de Sant Lluis y Ciutat Fallera son lo que menos inmuebles en alquiler poseen. Por último, el distrito de Olivereta es el que presentan una distribución más desigual entre todos sus barrios, reuniendo casi toda la oferta uno de ellos (Nou Moles).

Ciutat Vella	23,89%
El Carme	3,61%
El Mercat	4,12%
El Pilar	1,73%
La Seu	3,28%
La Xerea	2,76%
Sant Francesc	8,38%
Eixample	15,78%
Russafa	7,31%
El Pla del Remei	4,45%
Gran Vía	4,03%
Poblats Maritims	10,82%
El cabanyal-el canyamelar	6,28%
Playa de la Malvarrosa	1,83%
El Grau	2,15%
Beteró	0,33%
Natzaret	0,23%
Extramurs	9,60%
Arrancapins	3,47%
La roqueta	2,95%
El botànic	1,78%
La Petxina	1,41%
Camins al grau	6,93%
Penya-roja	2,76%
Aiora	1,78%
La creu del Grau	1,17%
Albors	0,98%
Camí Fondo	0,23%

Quatre Carreres	5,39%
Mont-Olivet	1,26%
Ciutat de les Arts	1,45%
Malilla	1,45%
La Punta	0,37%
En Corts	0,42%
Na Rovella	0,37%
Fonteta de Sant Lluis	0,05%
Pla del Real	4,73%
Mestalla	2,30%
Exposició	1,50%
Jaume Roig	0,66%
Ciutat universitaria	0,28%
Saïda	4,22%
Sant Antoni	1,45%
Morvedre	0,89%
Trinitat	0,56%
Tormos	0,56%
Marxalenes	0,75%
Campanar	3,70%
Sant Pau	1,78%
Nou campanar	0,70%
Campanar	0,56%
El Calvari	0,33%
Les Tendetes	0,23%
Beniferri	0,09%
Patraix	2,76%
Patraix	1,12%
Barrio de Favara	0,61%
Safranar	0,14%
Vara de Quart	0,66%
Sant Isidre	0,23%

Olivereta	2,76%
Nou Moles	2,25%
La Llum	0,09%
Soternes	0,05%
Tres Forques	0,23%
La Fontsanta	0,14%
Algirós	2,76%
Ciutat Jardi	0,61%
Amistat	0,80%
Illa Perduda	0,33%
La carrasca	0,66%
Vega baixa	0,37%
Rascanya	2,72%
Sant Llorenç	1,59%
Torrefiel	0,42%
Els Orriols	0,70%
Benimaclet	1,41%
Benimaclet	1,26%
Camí de Vera	0,14%
Benicalap	1,26%
Benicalap	1,08%
Nou benicalap	0,14%
Ciutat fallera	0,05%
Jesus	1,26%
La Raiosa	0,37%
L'Hort de Senabre	0,37%
Sant Marcellí	0,19%
La Creu Coberta	0,14%
Camí Reial	0,19%

Tabla 2 *Reparto de la oferta de viviendas en alquiler según el distrito y barrio*
Fuente: Elaboración propia

3.3.3 Variable precio

La siguiente variable a analizar va a ser el dato a predecir y, por tanto, la variable dependiente del modelo. Es por eso por lo que es fundamental entender el comportamiento de la misma, así como la distribución que presenta. Esta variable recoge el precio que mensualmente el inquilino tiene que pagar por el alquiler de la propiedad.

Para comprender su funcionamiento se calculan primero algunas métricas básicas que aparecen resumidas en la siguiente tabla:

Mínimo	375€
Máximo	2750€
Moda	900€
Media	1094€
Mediana	970€
Desviación típica	409,34€

Tabla 3 *Resumen variable precio*
Fuente: Elaboración propia

Se puede comprobar como el máximo precio para una vivienda es de 2750€ lo cual concuerda con lo calculado anteriormente ya que en el apartado de estudio de valores anómalos se cifró como límite máximo 2800€. A continuación, lo siguiente a estudiar son las medidas de posición central. A través de estas se puede concluir que el precio de alquiler más repetido es 900€/mes pero que sin embargo la media es ligeramente superior, un 1,22%. Si se orden de menor a mayor en primer lugar, estaría la moda (900€), a continuación, la mediana (970€) y por

último la media (1094€). Esto ya es un primer síntoma de que la variable precio tiene una distribución sesgada. Para acabar con estas primeras medidas generales se ha obtenido una desviación típica de algo más de 400€ lo que supone un coeficiente de variación del 37,41%.

Retomando el tema de la distribución de la variable ya se ha anticipado que no parece que vaya a seguir una distribución normal. No obstante, se ha procedido a realizar el test de *Shapiro-Wilk* para lograr un veredicto claro. Al realizarlo se ha obtenido un p valor < 2.2e-16 por lo que se puede rechazar la hipótesis nula teniendo así evidencias suficientes para afirmar que los datos no se distribuyen de forma normal. Todo esto aparte de ser averiguado de forma numérica también se puede corroborar de forma gráfica. En el histograma se puede ver de una forma muy clara que se parte de una distribución sesgada a la izquierda.

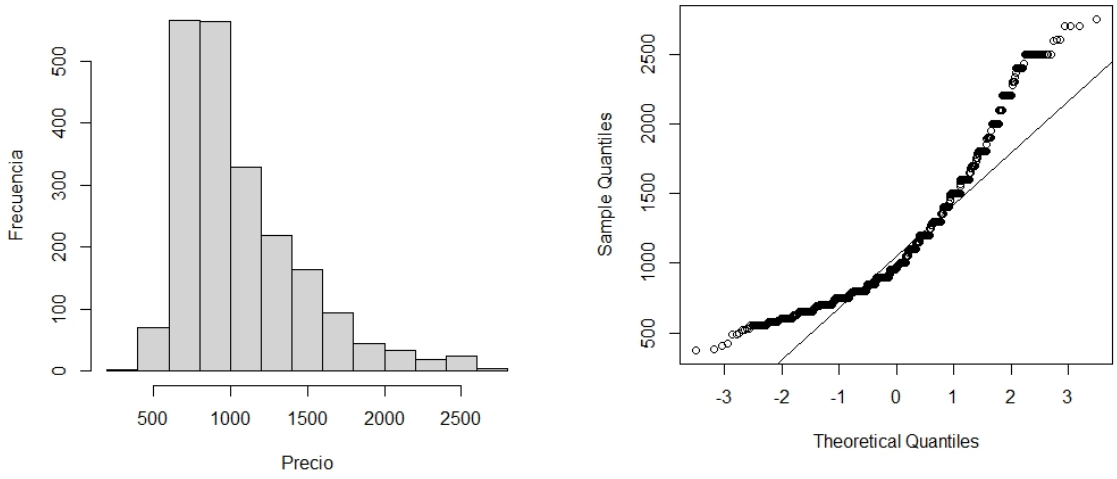

Gráfico 3 *Histograma y qqplot de la variable precio*
Fuente: Elaboración propia

Para tratar de corregir la no normalidad de esta variable se ha aplicado el logaritmo en base 10. De esta manera se ha conseguido normalizar la distribución de la variable precio, pasando de un coeficiente de asimetría de 1,32 a 0,50.

3.3.4 Variable habitaciones

La siguiente variable a analizar son el número de habitaciones que tiene una propiedad, siendo el mínimo ninguna habitación y el máximo siete habitaciones. Las viviendas que no tienen ninguna habitación deberían ser los estudios ya que se caracterizan por presentar un concepto abierto donde no existe segmentación entre las diferentes estancias de la casa. Se puede ver en la gráfica 4 como la mayoría de los inmuebles tienen entre una y cuatro habitaciones, representando casi el 94% de la muestra. Así pues, la que más abunda es aquella que tiene tres habitaciones con un 40%. Por último, se ha comprobado que el número de viviendas que no tienen habitaciones corresponden con el número de anuncios catalogados como estudios.

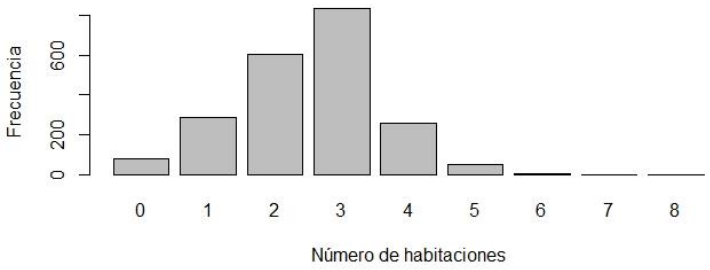

Gráfico 4 *Diagrama de barras de la variable número de habitaciones*
Fuente: Elaboración propia

3.3.5 Variable metros

Una de las variables que a priori parece fundamental para establecer el precio de una vivienda son los metros que posea esta. Así pues, tal como se ha realizado con la variable precio se procede a calcular, en primer lugar, una serie de métricas básicas.

Mínimo	19m^2
Máximo	286m^2
Moda	100m^2
Media	103,2m^2
Mediana	97m^2
Desviación típica	40,49m^2

Tabla 4 *Resumen variable metros cuadrados*
Fuente: Elaboración propia

Al igual que pasaba con el precio máximo, los metros cuadrados máximos están por debajo del límite superior para considerar un dato como anómalo. Además, se ha calculado que el 76,39% de las viviendas tienen menos de 125 metros cuadrados con lo que apenas el 24% se encuentra entre los 125 y 286 metros cuadrados. La propiedad más habitual es aquella que tiene 100 metros cuadrados de superficie, en total se han encontrado 102 viviendas (un 4,77%). Así pues, la media está algo por encima mientras que la mediana es ligeramente inferior. Por último, se ha obtenido que los datos se desvían en promedio 40,49 metros cuadrados sobre la media aritmética, con lo que se obtiene un coeficiente de variación del 39,23%. Este es ligeramente superior al de la variable precio con lo que se puede concluir que la variable metros posee una mayor dispersión.

Por otro lado, en lo que se refiere a la distribución de la variable metros, esta tampoco presenta una distribución normal. Al realizar la prueba de Shapiro se ha obtenido un p valor < 2.2e-16 con lo que se puede rechazar la hipótesis nula de que la variable siga una distribución normal.

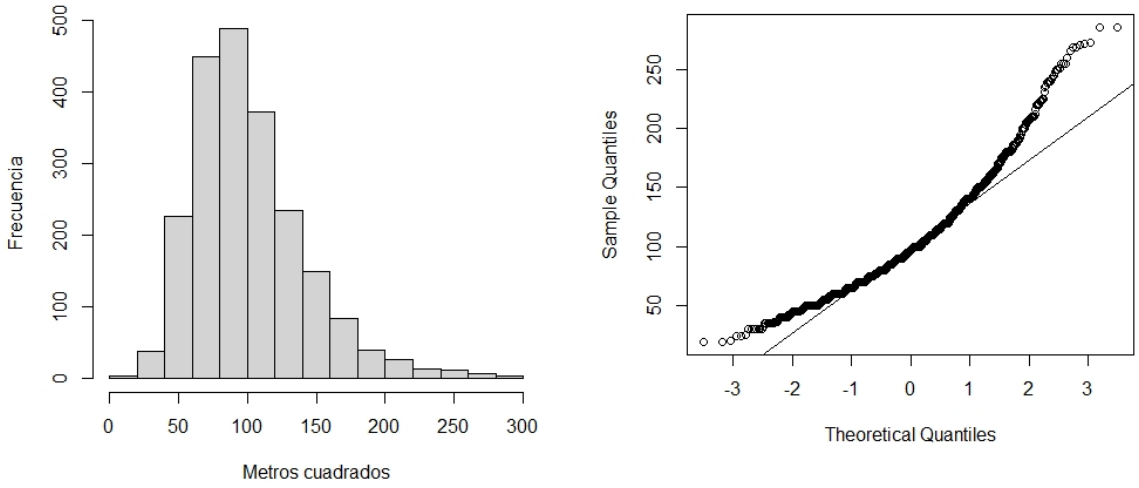

Gráfico 5 *Histograma y qqplot de la variable metros cuadrados*
Fuente: Elaboración propia

Al igual que se ha realizado con el precio, se ha aplicado el logaritmo en base 10 a la variable metros cuadrados para reducir así el sesgo que presenta. Tras aplicarlo se ha obtenido un coeficiente de asimetría de -0,17, inicialmente se partía de un coeficiente de 1,07.

3.3.6 Variable ascensor

A continuación, se va a estudiar la primera variable dicotómica presente en la base de datos construida con todas aquellas propiedades en alquiler. Este tipo de variables se definen por tener solo dos valores posibles, en este caso que la propiedad sí que tenga ascensor o que no lo tenga. Tal como se puede observar en el gráfico de tartas la gran mayoría de las viviendas sí que tiene ascensor, el 84% de ellas.

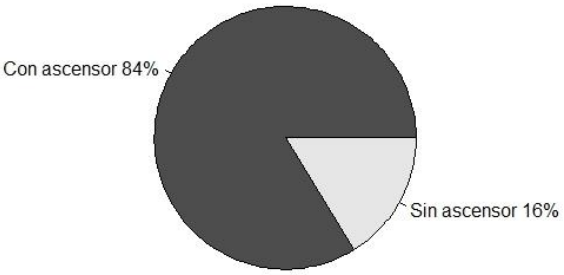

Gráfico 6 *Diagrama de tarta variable ascensor*
Fuente: Elaboración propia

3.3.7 Variable luz

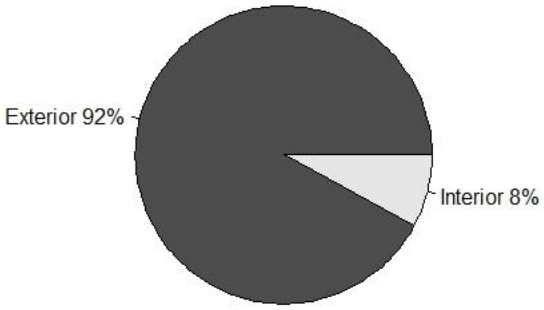

Gráfico 7 *Diagrama de tarta variable luz*
Fuente: Elaboración propia

La siguiente variable en la que se va a profundizar es en el tipo de luz que posee el inmueble. Se recuerda que se ha considerado que una propiedad tiene luz exterior si las ventanas dan a una calle o similar. Por el contrario, si la luz proviene del patio interior de la finca se considera que ese piso tiene luz de tipo interior. Así pues, tal como se podía esperar la inmensa mayoría de los alquileres tienen luz exterior y solo un 8% presentan luz interior.

3.3.8 Variable tipo de planta y planta

En este caso al igual que se hizo con la variable barrio y distrito, estas también se van a estudiar de forma conjunta ya que están íntimamente relacionadas la una con la otra. Las propiedades aparte de poder estar situadas en una planta de un edificio también puede ser que se sitúen en el bajo o en la entreplanta. La diferencia principal entre ellas es que un bajo son aquellos que se encuentran a pie de calle mientras que un entresuelo es el que se sitúa entre el bajo y el primer piso. La gran mayoría de las viviendas, el 96% de ellas, se enmarcan en la categoría de planta. En cuanto a la variable planta anotar que los bajos se han codificado como un 0 y las entreplantas como 0,5, evitando de esta forma que hubiera una mayor cantidad de valores nulos. Tal como aparece representado en el gráfico el mínimo es el 0 y la planta máxima es la 23ª. Además, se puede comprobar como la mayoría de las propiedades son lo que comúnmente llamamos un primero. Por último, destacar que el 75% de los inmuebles están situadas entre un primero y un quinto.

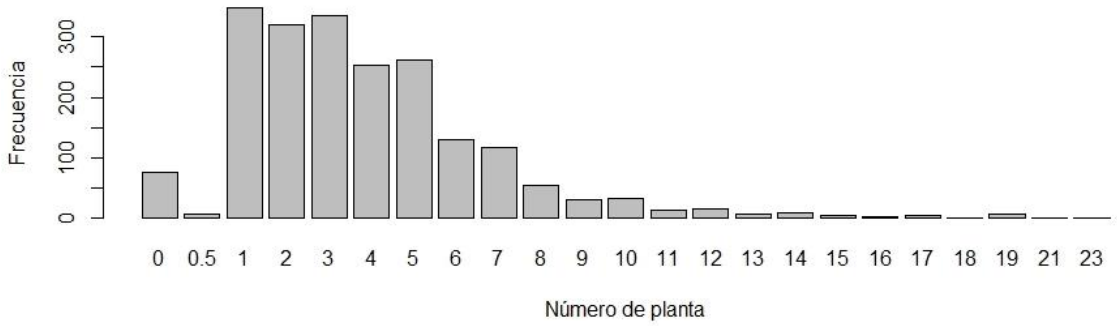

Gráfico 8 *Diagrama de barras variable número de planta*
Fuente: Elaboración propia

3.3.9 Variable piscina

Para acabar queda por analizar las últimas tres variables creadas a partir de la descripción que el anunciante añade a las ya estudiadas características de la vivienda. Vuelven a ser todas ellas variables dicotómicas ya que todas indican la posesión o no de la característica en cuestión. En esta primera variable se quiere averiguar el porcentaje de inmuebles que poseen piscina frente a los que no. En total se ha obtenido que solo el 2% de las viviendas tienen, es decir, solo 59 de las 2135 propiedades.

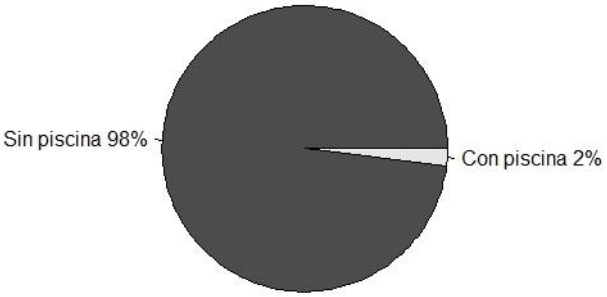

Gráfico 9 *Diagrama de tarta variable piscina*
Fuente: Elaboración propia

3.3.10 Variable garaje

Siguiendo con las variables obtenidas a partir de la descripción de la propiedad se encuentra la variable garaje. Se ha obtenido que apenas el 3% de las viviendas que están en alquiler en la ciudad de Valencia tienen garaje.

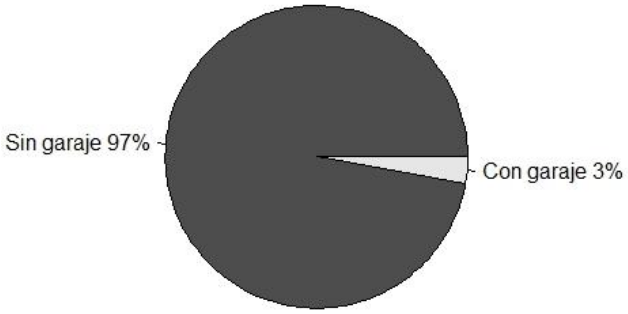

Gráfico 10 *Diagrama de tarta variable garaje*
Fuente: Elaboración propia

3.3.11 Variable reformado

Para acabar con este primer análisis individualizado queda por estudiar la característica que le puede aportar un valor añadido a la vivienda en el caso de que esté reformada. De esta manera se concluye que el 18% de los alojamientos están reformados.

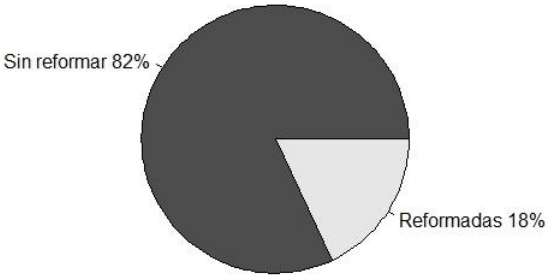

Gráfico 11 *Diagrama de tarta variable reformado*
Fuente: Elaboración propia

En conclusión, con este primer análisis se ha podido dibujar y caracterizar como son las viviendas que se encuentran actualmente en la ciudad de Valencia. El 50% de la oferta se encuentra concentrada en los distritos de Ciutat Vella, Eixample y Poblats Marítims. Además, la gran mayoría suelen tener un precio alrededor de los 1000€ mensuales y unos 100 metros cuadrados. Por otra parte, lo más común es que se trate de pisos con ascensor, luz exterior y dos o tres habitaciones. Así pues, lo que más predominan son pisos ubicados en la primera planta, aglomerando al 75% entre un primero y un quinto. Por último, remarcar que alrededor del 20% son pisos reformados o de nueva construcción.

3.4 Análisis bivariante

En el apartado anterior se realizó un análisis individual de cada una de las variables, sin embargo, ahora lo que se propone es un estudio donde se examine la dependencia de unas variables con otras. En concreto, lo que es de especial interés para el caso de estudio que se plantea en este trabajo es calcular la dependencia entre la variable precio (variable dependiente) y el resto de las variables (variables independientes).

Para realizar este tipo de estudio lo más adecuado es realizar diversas pruebas paramétricas según la naturaleza de las variables. No obstante, previamente se ha determinado que la variable precio no sigue una distribución normal por lo que es motivo suficiente para desestimar el uso de este tipo de pruebas. En su defecto lo correcto es aplicar pruebas no paramétricas o pruebas de distribución libre.

El primer tipo de prueba que se va a realizar es Wilcoxon de suma de rangos, esta se utiliza como sustituta a los procedimientos t en problemas con dos muestras independientes. Por tanto, esta prueba se va a aplicar entre la variable precio y aquellas variables de tipo dicotómico. Para esta se establece de forma genérica las siguientes hipótesis:

H_o: las distribuciones son idénticas en las dos poblaciones

H_1: las distribuciones no son iguales en las dos poblaciones

Además, para poder aplicar este tipo de prueba se han de cumplir las siguientes condiciones (Crespo Abril 2017):

- Independencia: todas las muestras han de ser escogidas siguiendo un muestreo de tipo aleatorio simple ya que de esta forma se garantiza que cada dato de la muestra ha sido seleccionado de forma independiente. No obstante, este requisito también implica que los elementos de una muestra son independientes de la otra. Se exige pues la independencia entre las muestras y dentro de ellas.
- Muestras grandes: las muestras han de ser de al menos de 10 observaciones ya que cuanto mayor es el tamaño de la muestra mejor es el resultado.
- No existen empates: se entiende por empates valores iguales. Así pues, si el número de datos repetidos no es muy elevado y las muestras son de gran tamaño se sigue pudiendo aplicar esta prueba. Los resultados serán aproximados, pero de forma general podemos asumirlos como válidos.

Tras comprobar que nuestro caso de estudio cumple con todos los requerimientos para aplicar este tipo de prueba se procede a calcular los diversos valores de p para poder concluir si se acepta o rechaza la hipótesis nula. Se ha escogido como nivel de significancia un α=0,05, por lo que para valores de p superiores a α se aceptará la hipótesis nula. Admitiendo de esta manera que la distribución de las dos poblaciones es similar o idéntica.

El segundo tipo de prueba que se requiere es la de Kruskal-Wallis esta sería equivalente a la popular prueba de ANOVA. En este caso en vez de tener dos muestras independientes como la anterior prueba se tienen k muestras, siendo k superior a dos. Por esta razón las hipótesis genéricas de esta prueba son las siguientes:

H_o: las distribuciones son idénticas en las k poblaciones estudiadas

H_1: al menos una de las distribuciones toma valores diferentes al resto

Por lo que se refiere a los prerrequisitos previos a comprobar para poder aplicar esta prueba son los mismos que para la prueba de Wilcoxon. De la misma manera el nivel de significación también se mantiene.

Para acabar la última prueba que se va a aplicar es el coeficiente de correlación de rangos de Spearman. Este es una medida estadística que mide el grado de asociación entre ambas variables, determinando de esta manera la dependencia o independencia de dos variables aleatorias (Tejada 2014). Este coeficiente puede oscilar desde el valor -1 hasta +1 cuanto más cercano sea el resultado a 1 mayor es la asociación entre las dos variables. Dicho de otra manera, cuando aumenta el rango de una variable el de la otra variable también se incrementa. Por el contrario, cuanto más cercanos están los valores a -1 esta asociación que hay entre ambas variables es negativa, es decir que mientras una variable aumenta la otra disminuye. Así pues, cuando el resultado de este coeficiente sea 0 significa que no hay correlación entre ambas variables.

Evidentemente, hay distintos rangos dentro de todos los posibles valores comprendidos entre el -1,0 y el +1,0. Se establece que hay una correlación perfecta cuando el coeficiente se encuentra entre -0,91 y -1,00 o +0,91 y 1,00. Cuando tiene un valor comprendido entre -0.76 y -0,90 o +0,76 y +0,90 se dice que tiene una relación fuerte. Por último, cuando este coeficiente se sitúa entre los valores -0,51 y -0,75 o +0,51 y +0,75 tienen una correlación considerable (Hernández Sampieri 2014).

3.4.1 Relación precio y ascensor

Empecemos por estudiar si hay alguna relación entre la variable precio y el hecho de que una vivienda en alquiler posea o no ascensor. Tras realizar la prueba, se ha obtenido un p valor < 2,2e-16 por lo que se rechaza la hipótesis nula, aceptado por tanto que el precio medio de las viviendas difiere según si tienen o no ascensor.

Así pues, si calculamos el precio medio de las viviendas con ascensor es de 1117€ y su mediana de 1000€ mientras que las propiedades sin ascensor tienen un precio medio de 937€ y una mediana de 850€.

3.4.2 Relación precio y luz

Repetimos el proceso anterior pero ahora con la variable tipo de luz en vez de con la variable ascensor. En este caso se ha logrado un p valor = 0,002384 igual que antes, aunque es este caso con un valor de significancia menor se rechaza la hipótesis nula.

En este caso si calculamos el precio medio de los inmuebles con luz exterior es de 1093€ y una mediana de 975€. En el caso de los alojamientos con luz de tipo interior su precio medio mensual es de 1017€ y la mediana es de 900€.

3.4.3 Relación precio y garaje

Se procede ahora a estudiar la variable garaje, se recuerda que esta característica se obtuvo gracias a la descripción que se adjuntaba en los anuncios. En este caso a diferencia de los dos valores de p anteriores es superior a 0,05. Se ha obtenido un p valor = 0,4078 por lo que no se tienen evidencias suficientes para rechazar la hipótesis nula, admitiendo así que la distribución de los precios es similar en las viviendas con garaje y en las que no lo tienen.

De forma complementaria se ha calculado el precio medio de las propiedades sin garaje el cual es de 1096€ mensuales con una mediana de 970€. Por otro lado, el precio medio de los inmuebles con garaje es de 1029€ al mes y una mediana de 935€.

3.4.4 Relación precio y piscina

Siguiendo con otra de las variables que se ha extraído de la descripción adjuntada en el anuncio es momento de analizar el factor piscina. Al realizar la prueba se ha calculado un p valor = 0,1239 por lo que al igual que antes se acepta la hipótesis nula.

Así pues, aquí a penas se observan diferencias entre el precio medio de las viviendas con piscinas y de aquellas que no poseen. Las que sí que tienen su precio medio es de 1094€ y su mediana de 950€ mientras que aquellas propiedades que no la tienen su precio medio se sitúa en los 1096€ y su mediana en los 1100€.

3.4.5 Relación precio y reforma

Por último, queda analizar la relación entre el precio de una vivienda en alquiler con el hecho de que haya sido o no reformada. Asimismo, se ha obtenido un p valor = 0,5651, igual que ha pasado con las dos variables anteriores se ha de admitir que no hay diferencias entre las distribuciones de las dos poblaciones.

Una vez más, se ha calculado el precio medio de aquellas viviendas que han sido reformadas y se ha obtenido que es 1087€ y su mediana de 970€ mientras que los inmuebles en alquiler sin haber sido reformados tienen un precio medio de 1095€ y una mediana de 968€. Cuanto menos llama la atención que las viviendas sin reformar tengan un precio medio más alto y una mediana prácticamente igual que las que no han sido reformadas.

3.4.6 Relación precio y tipo de vivienda

Una vez finalizados todos los análisis con aquellas variables dicotómicas se procede a analizar aquellas variables que tienen más de dos valores posibles. Como se ha explicado al inicio de esta sección para este tipo de variables se va a aplicar la prueba de Kruskal. Al realizarla se ha obtenido un p valor < 2.2e-16 con lo que se puede rechazar H_o. De esta manera se concluye que el precio medio de las viviendas en alquiler no es el mismo en los siete tipos de inmuebles identificados, y que por tanto, este depende del tipo de vivienda que se alquila.

El tipo de viviendas con un precio más elevado son las casas, este se sitúa en los 1825€ mensuales. A continuación, con un precio más o menos similar encontramos los chalets (1370€), el cortijo (1400€) y los dúplex (1343€). Seguidamente se encuentran los áticos (1260€) y los pisos (1076€). Para acabar los estudios son las viviendas con un precio medio más bajo (906€). En el siguiente gráfico se puede ver la mediana del precio de cada uno de los tipos de viviendas.

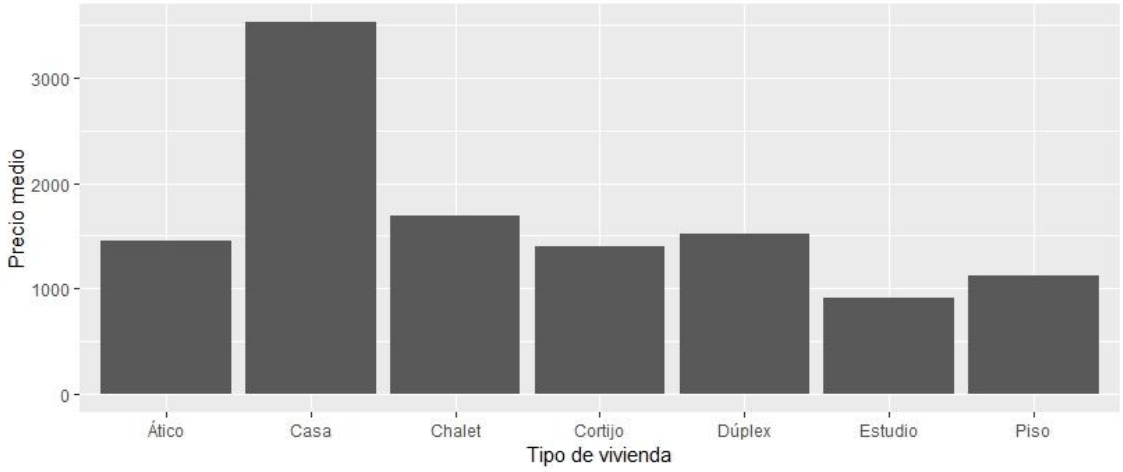

Gráfico 12 *Diagrama de barras entre el tipo de la vivienda y su precio medio*
Fuente: Elaboración propia

3.4.6 Relación precio y habitaciones

Se procede a examinar la relación entre el precio de una vivienda con el número de habitaciones que tiene. Tras efectuar la prueba se obtiene que su p valor < 2.2e-16 por lo que no se tienen las evidencias suficientes para aceptar H_o. De esta manera se deduce que el precio medio de los inmuebles es diferente según el número de habitaciones que posee la vivienda.

Las propiedades con ninguna o una habitación tienen un precio medio alrededor de los 900€ mensuales. Los inmuebles que tienen entre dos y cinco habitaciones tienen un precio medio entre los 1000€ y los 1800€. A partir de las seis habitaciones el precio medio de los pisos sube muchísimo pasando a valer entre los 2600€ mensuales y los 3160€.

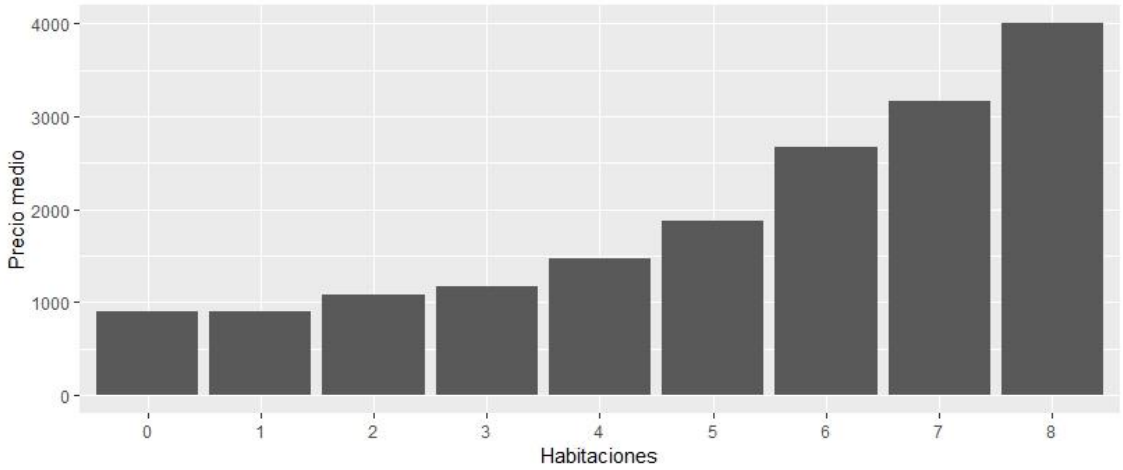

Gráfico 13 *Diagrama de barras entre el número de habitaciones y su precio medio*
Fuente: Elaboración propia

3.4.6 Relación precio y distrito/barrio

Pasemos ahora a analizar las variables que tienen que ver con la ubicación de los inmuebles en alquiler. Al realizar la prueba en ambas variables, en los dos casos obtenemos un p valor < 2.2e-16 por lo que estamos en disposición de afirmar que el precio medio de las propiedades es significativamente distinto según la ubicación en la que se sitúen.

Entre los distritos con un precio promedio más caro se encuentra el Eixample (1352,52€) y Ciutat Vella (1178,88€), sin embargo, entre los más baratos destacan Olivereta (815,68€) y Jesús (772,30€). Por lo que se refiere a los barrios con un precio promedio más caro son Ciutat Fallera (2000€) y el Pla del Remei (1498,31€). Es importante destacar que del barrio de Ciutat Fallera solo se ha encontrado un chalet en alquiler por lo que esta medida no es representativa. Así pues, los barrios con un alquiler medio más bajo son: la Llum (642,50€) y la Fontsanta (556,67€). Se pueden ampliar más detalles sobre distintos distritos y barrios en la tabla 5

3.4.6 Relación precio y planta/número de planta

Se procede ahora a estudiar la variable relacionada con la ubicación de la propiedad dentro de un edificio. Primero se ha analizado la dependencia entre la variable planta y el precio, obteniéndose un p valor = 0,332. De esta manera se puede concluir que el precio medio de las viviendas no cambio según si se encuentra en el bajo, entresuelo o en una planta. Asimismo, en lo que se refiere al número de planta sí que se han encontrado diferencias significativas al obtener un p valor = 5.203e-07. Por tanto, se puede afirmar que el precio medio de los pisos es distinto según la planta en la que se encuentren.

Todos los pisos situados entre la primera y séptima planta tienen un precio medio que ronda los 1000€. Sin embargo, a partir del octavo se ha observado un ligero incremento en el precio del alquiler, situándose este en los 1250€. Así pues, el precio medio más elevado se encuentra en las plantas undécima y vigésimo tercera.

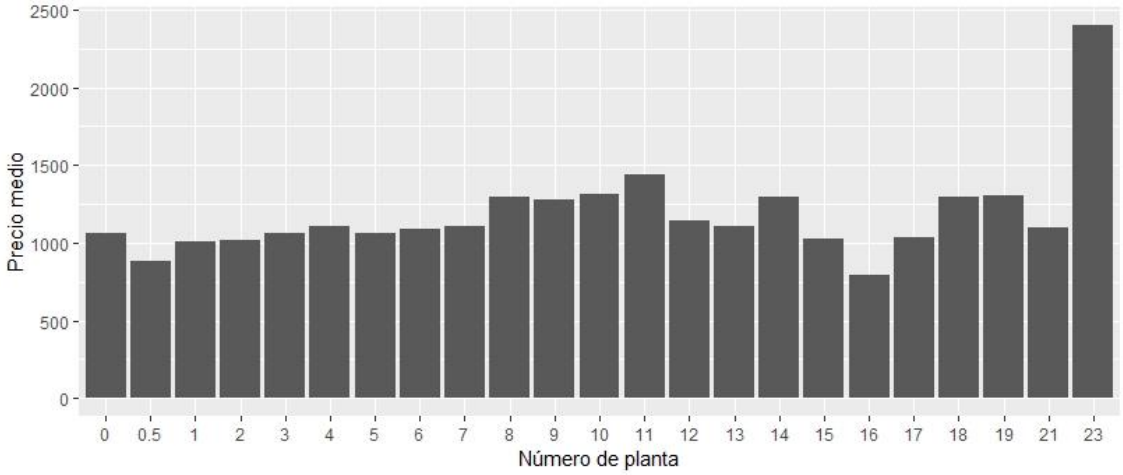

Gráfico 14 *Diagrama de barras entre el número de planta y su precio medio*
Fuente: Elaboración propia

3.4.7 Relación precio y metros cuadrados

Por último, queda hablar de la dependencia entre la variable precio y los metros cuadrados que tiene una propiedad. Al ser dos variables continuas se va a utilizar el coeficiente de Spearman para cuantificar el nivel de asociación entre estas dos variables. Con ayuda del software R se ha obtenido un valor de 0,596378, por lo que se puede decir que ambas variables tienen una correlación considerable.

En la gráfica 15, se puede observar cómo hay cierto comportamiento o patrón que refleja como a media que aumenta la cantidad de metros cuadrados también lo hace el precio de la vivienda. Sin embargo, esta tendencia está lejos de ser perfecta ya que si por ejemplo nos centramos en las propiedades de 200 metros cuadrados se observa como estas tienen un precio que va desde los 1000€ mensuales hasta los 2500€ mensuales.

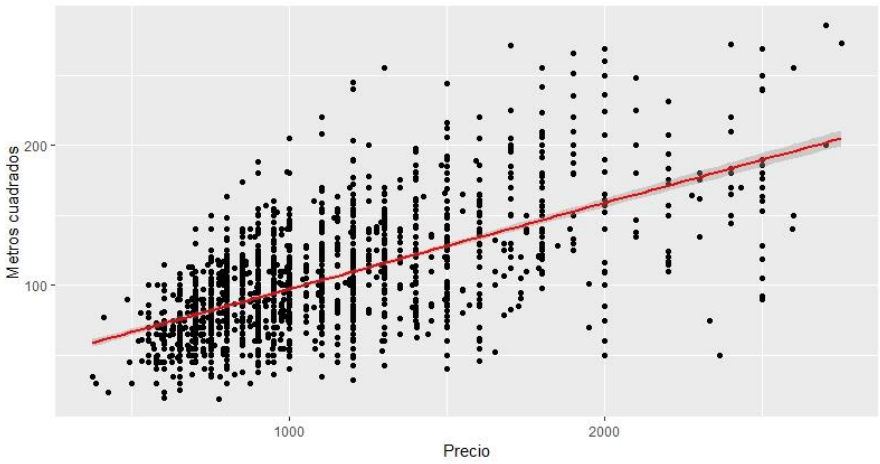

Gráfico 15 *Diagrama de dispersión entre la variable precio y metros*
Fuente: Elaboración propia

3.4.8 Otros análisis

Una medida que es muy utilizada en el sector inmobiliario para poder determinar el precio de compra de una vivienda o en su defecto el precio en alquiler es el precio del metro cuadrado. Por eso, se ha procedido a calcular el precio medio del metro cuadrado en cada uno de los distritos y barrios de la ciudad de Valencia. En la tabla que se muestra a continuación se puede ver tanto el precio promedio, los metros cuadrados como el precio del metro cuadrado.

	Precio medio	m2 medios	precio/m2
Ciutat Vella	1.178,88 €	96,87	12,17 €
El Carme	1.136,88 €	86,51	13,14 €
El Mercat	1.175,77 €	99,00	11,88 €
El Pilar	939,86 €	102,68	9,15 €
La Seu	1.172,64 €	86,10	13,62 €
La Xerea	1.148,98 €	98,39	11,68 €
Sant Francesc	1.260,17 €	102,80	12,26 €
Poblats Maritims	974,87 €	85,35	11,42 €
El cabanyal-el canyamelar	1.005,75 €	82,86	12,14 €
Playa de la Malvarrosa	880,74 €	82,56	10,67 €
El Grau	1.011,74 €	95,20	10,63 €
Beteró	855,71 €	91,86	9,32 €
Natzaret	709,00 €	74,20	9,56 €
Eixample	1.352,52 €	124,14	10,90 €
Russafa	1.244,24 €	111,61	11,15 €
El Pla del Remei	1.498,31 €	141,19	10,61 €
Gran Vía	1.387,91 €	128,02	10,84 €
Benicalap	1.087,41 €	102,30	10,63 €
Benicalap	1.077,83 €	98,48	10,94 €
Nou benicalap	856,67 €	82,33	10,40 €
Ciutat fallera	2.000,00 €	250,00	8,00 €
Extramurs	1.086,93 €	108,25	10,04 €
Arrancapins	992,50 €	115,55	8,59 €
La roqueta	1.152,22 €	98,78	11,66 €
El botànic	1.217,24 €	117,71	10,34 €
La Petxina	1.017,67 €	98,13	10,37 €
Camins al grau	1.115,20 €	111,32	10,02 €
Penya-roja	1.470,92 €	140,49	10,47 €
Aiora	860,68 €	91,11	9,45 €
La creu del Grau	929,80 €	93,60	9,93 €
Albors	791,19 €	86,57	9,14 €
Camí Fondo	1.140,00 €	113,40	10,05 €
Pla del Real	1.064,89 €	108,80	9,79 €
Mestalla	1.018,96 €	105,14	9,69 €
Exposició	1.057,81 €	99,00	10,68 €
Jaume Roig	1.335,71 €	151,14	8,84 €
Ciutat universitaria	845,83 €	92,17	9,18 €
Quatre Carreres	992,08 €	101,68	9,76 €
Mont-Olivet	787,22 €	94,67	8,32 €
Ciutat de les Arts	1.092,87 €	103,77	10,53 €
Malilla	998,55 €	100,61	9,92 €
La Punta	1.120,00 €	120,75	9,28 €
En Corts	1.138,89 €	103,22	11,03 €
Na Rovella	992,50 €	100,75	9,85 €
Fonteta de Sant Lluis	850,00 €	100,00	8,50 €

	Precio medio	m2 medio	precio/m2
Olivereta	815,68 €	85,19	9,58 €
Nou Moles	851,46 €	86,77	9,81 €
La Llum	642,50 €	77,50	8,29 €
Soternes	750,00 €	80,00	9,38 €
Tres Forques	710,00 €	86,80	8,18 €
La Fontsanta	556,67 €	64,00	8,70 €
Algirós	951,36 €	99,81	9,53 €
Ciutat Jardi	888,46 €	101,38	8,76 €
Amistat	942,65 €	95,29	9,89 €
Illa Perduda	982,86 €	103,57	9,49 €
La carrasca	901,79 €	92,57	9,74 €
Vega baixa	1.131,25 €	116,25	9,73 €
Campanar	1.044,68 €	111,86	9,34 €
Sant Pau	1.145,13 €	117,08	9,78 €
Nou campanar	1.060,33 €	106,07	10,00 €
Campanar	959,58 €	112,33	8,54 €
El Calvari	699,29 €	70,71	9,89 €
Les Tendetes	920,00 €	121,60	7,57 €
Beniferri	1.050,00 €	173,00	6,07 €
Rascanya	903,62 €	97,84	9,24 €
Sant Llorenç	1.063,24 €	107,62	9,88 €
Torrefiel	718,33 €	87,78	8,18 €
Els Orriols	653,00 €	81,73	7,99 €
Benimaclet	827,33 €	92,40	8,95 €
Benimaclet	831,30 €	91,48	9,09 €
Camí de Vera	791,67 €	100,67	7,86 €
Patraix	922,53 €	103,25	8,93 €
Patraix	821,88 €	98,58	8,34 €
Barrio de Favara	1.396,15 €	135,46	10,31 €
Safranar	698,33 €	60,00	11,64 €
Vara de Quart	780,64 €	92,29	8,46 €
Sant Isidre	706,00 €	98,60	7,16 €
Saïda	828,98 €	93,28	8,89 €
Sant Antoni	810,00 €	106,68	7,59 €
Morvedre	878,68 €	90,00	9,76 €
Trinitat	867,00 €	87,83	9,87 €
Tormos	681,67 €	84,58	8,06 €
Marxalenes	888,69 €	81,81	10,86 €
Jesus	772,30 €	94,04	8,21 €
La Raiosa	756,25 €	103,25	7,32 €
L'Hort de Senabre	783,75 €	81,13	9,66 €
Sant Marcellí	945,00 €	115,25	8,20 €
La Creu Coberta	714,67 €	87,00	8,21 €
Camí Reial	652,00 €	85,50	7,63 €

Tabla 5 *Precio del metro cuadrado según el distrito y el barrio*
Fuente: Elaboración propia

El distrito con el precio del metro cuadrado más alto corresponde con el centro histórico de la ciudad, Ciutat Vella. En este se encuentra monumentos tan emblemáticos como el mercado central, la plaza del ayuntamiento, las torres de Serrano o la plaza de la Virgen. Tiene un precio el metro cuadrado de 12,17€/m^2, sin embargo, hay bastante variedad de precios en los seis barrios que lo componen. Destaca con un precio significativamente menor el barrio del Pilar y con un precio ligeramente mayor el barrio de la Seu.

Dejando a un lado el centro histórico y situándose en la zona cercana a la playa se encuentra el segundo distrito con el precio del metro cuadrado más caro, Poblats Marítims. Su precio se sitúa en los 11,42€/m^2, siendo el barrio del Cabanyal el único que se encuentra por encima de ese precio. El resto de los barrios tiene un precio inferior llegando a estar el de Beteró un 18,44% por debajo del precio del metro cuadrado del distrito.

Volviendo a la parte más céntrica de la ciudad se encuentra el tercer distrito más caro de la ciudad, l'Eixample. Este contiene uno de los barrios más de moda de la ciudad (Russafa) cuyo precio del metro cuadrado es de 11,15€/m^2, el cual está un 2,32% por encima del precio del distrito (10,90€/m^2).

En la parte baja de la lista se sitúan los distritos de la Saïda y Jesús como los distritos con el precio del metro cuadrado más bajo. Este último es casi un 50% más barato que el distrito más caro de la ciudad, Ciutat Vella. Con lo que se vislumbra así la diferencia de precios entre vivir en una parte de Valencia o en otra. Además, en el distrito de Jesús se encuentra uno de los barrios más baratos, La Raiosa.

También se ha identificado que los barrios con unas viviendas más grandes son Beniferri, Jaume Roig y Penya-Roja. Por el contrario, los más pequeños se encuentran en El Calvari, Safranar y La Fontsanta.

Además, también se ha querido analizar la relación entre los metros y precios y cada tipo de inmueble. De esta manera en el gráfico 16 se pueden apreciar unas primeras conclusiones. Los estudios son aquellas viviendas con menor superficie construida y las casas las que mayor superficie construida tienen. Además, otra idea que se puede extraer al ver el gráfico es que los estudios apenas presentan variaciones en el precio o metros.

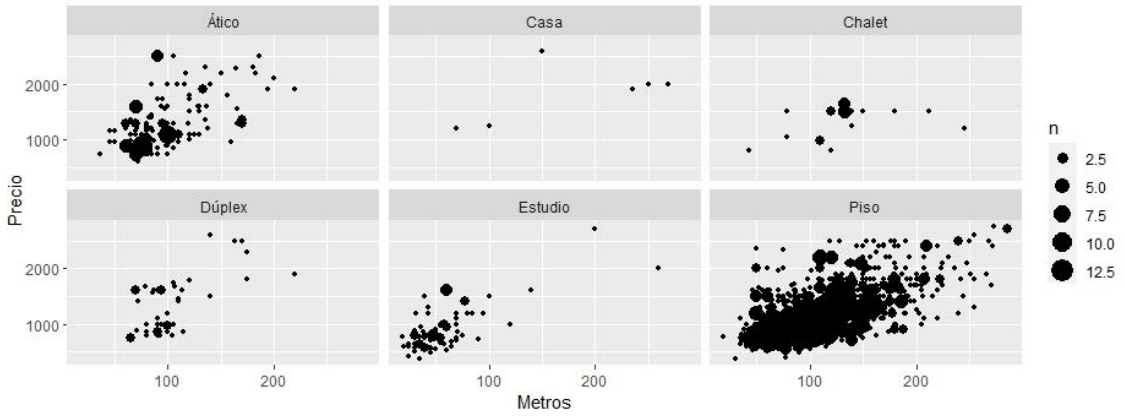

Gráfico 16 *Comparativa de metros y precio según el tipo de vivienda*
Fuente: Elaboración propia

No obstante, se ha decido calcular el precio del metro cuadrado en cada tipo de vivienda, tal como se muestra en la tabla 6. Se concluye así que las propiedades con un precio del metro cuadrado más alto son los estudios. A continuación, se encontrarían los áticos y dúplex, siendo estos un 21% más baratos que los estudios atendiendo al precio del metro cuadrado. Tal como se anticipaba en la gráfica 16 se comprueba que las casas son las propiedades que de media mayor número de metros cuadrados poseen, siendo estas también las que menor precio tiene su metro cuadrado.

Tipo de propiedad	Precio promedio	Metros promedio	Precio / m2
Casa	1825,00€	179,00	10,20€/m^2
Piso	1075,63€	105,18	10,23€/m^2
Chalet	1369,56€	131,30	10,43€/m^2
Dúplex	1343,32€	103,67	12,96€/m^2
Ático	1260,29€	96,22	13,10€/m^2
Estudio	906,11€	57,35	15,80€/m^2

Tabla 6 *Comparativa del precio/m2 según tipo de vivienda*
Fuente: Elaboración propia

También se ha querido averiguar en qué barrio predomina cada tipo de vivienda. Así, se ha observado que en el barrio de Favara en Patrix es donde más casas en alquiler hay. Por lo que a los áticos se refiere donde más abundan son en el Cabanyal, en la Xerea y Sant Francesc. Sin embargo, los dúplex se encuentran principalmente en el barrio del Carmen y Gran Vía. Por último, en lo que a los estudios se refieren vuelven a sonar el Cabanyal y Sant Francesc. Se concluye así que tanto los áticos, dúplex como los estudios predominan en los barrios con precio por metro cuadrado más elevado.

Además, también se tenía la curiosidad de indagar en que distritos se encuentran los pisos ubicados en plantas más altas. A la vista del gráfico 17 se puede ver claramente como casi ningún distrito posee viviendas en alquiler en una planta alta de un edificio. En los únicos que se ha encontrado alguna vivienda situada en la parte alta son Campanar, Quatre Carreres y Rascanya. Destaca también la cantidad de bajos que hay en Poblats Marítims y Ciutat Vella.

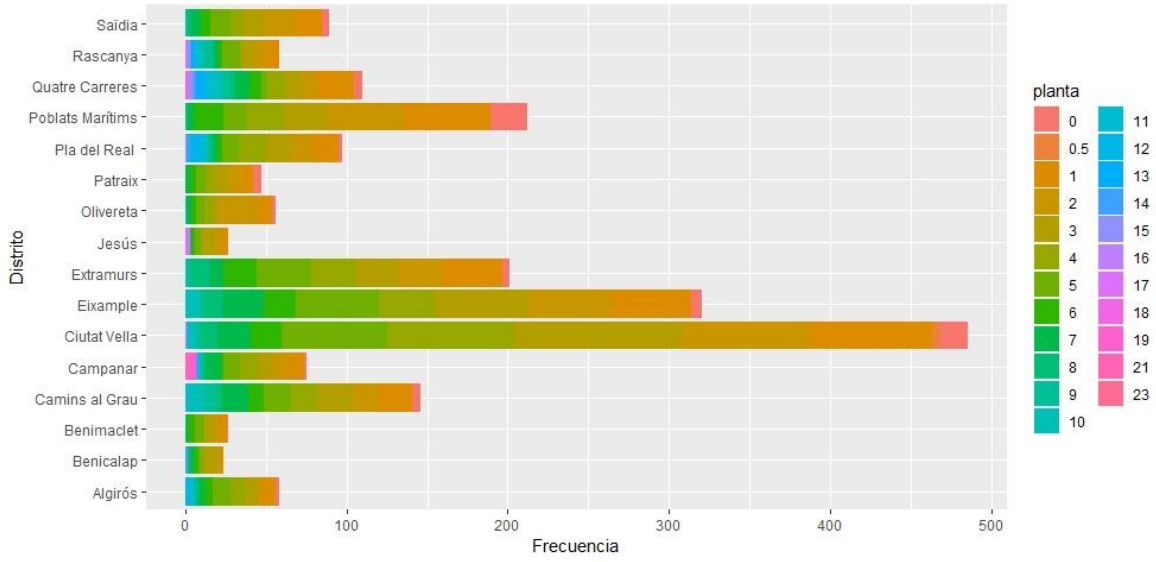

Gráfico 17 *Distritos con pisos más altos*
Fuente: Elaboración propia

3.5 Construcción modelo

En este último apartado del bloque de la metodología se va a tratar de crear un modelo que sea capaz de predecir el precio a establecer a una propiedad que está en alquiler. Para ello, se va a crear un modelo de regresión lineal múltiple cuya utilidad principal es tratar de estudiar las posibles relaciones existentes entre la variable dependiente (precio) y las diferentes variables independientes, explicativas o exógenas (Romero Villafranca y Zúnica Ramajo 2020). Así pues, la ecuación básica o recta de regresión de este tipo de modelo es:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_i X_{it} + \cup_t$$

Y_t: variable dependiente.

X_{it}: variables explicativas o regresores.

β_0: valor medio de Y cuando $X_1 \ldots X_i$ es 0, también denominado término constante o independiente del modelo.

β_1: incremento del valor medio de Y cuando X_i aumenta en una unidad manteniéndose constantes el resto de las variables independientes.

$\cup_t$: término de error del modelo o residuos que se distribuyen de forma normal y son independientes.

Para poder evaluar los parámetros del modelo y la varianza del error es necesario establecer algunas hipótesis respecto a la perturbación, variables y parámetros β. En primer lugar, se establecen las relativas a la perturbación

1. Las perturbaciones $\cup_t$ son variables aleatorias de media nula.
2. Todas las perturbaciones tienen la misma varianza.
3. Todas las perturbaciones están incorrelacionadas entre sí.
4. Las perturbaciones tienen una distribución conjunta normal.
5. La perturbación no depende de las variables explicativas X_i.

A continuación, se enumeran las hipótesis relativas a las variables explicativas y la explicada:

1. Las variables explicativas, X_i, y la explicada, Y, son obtenidas sin error de observación.
2. Las variables explicativas X_i son no aleatorias.
3. El valor de Y_j condicionado a las X_i es el observado de una variable aleatoria, cuyo valor medio es una combinación lineal de los valores de X_i, y cuya varianza es constante.
4. Entre las variables explicativas X_i no deben existir relaciones lineales exactas.

Por último, se encuentran las hipótesis relativas a los parámetros del modelo:

1. Los parámetros β_i del modelo son constantes y forman parte del modelo de forma lineal.

Una vez establecidas las hipótesis generales del modelo y su recta de regresión es momento de adecuarlas al contexto del problema. Así pues, estará formado por un total de ocho variables explicativas y la variable independiente que es el precio. De esta forma, la recta de regresión quedaría así:

$$precio = \beta_0 + \beta_1 metros + \beta_2 tipo + \beta_3 habitaciones + \beta_4 barrio + \beta_5 luz + \beta_6 ascensor + \beta_7 planta \cup_t$$

No obstante, tal como se ha comentado previamente algunas de ellas son cuantitativas y otras, sin embargo, son cualitativas. Una práctica muy común cuando hay este segundo tipo de variables en los modelos de regresión consiste en crear variables ficticias. Estas básicamente son variables binarias que toman valores de 0 o 1 según la presencia o ausencia de algún efecto de la variable categórica en cuestión. En total, en el modelo hay seis variables cualitativas y que por tanto se han de transformar a variables ficticias. Teniendo en cuenta los niveles de cada una de estas variables categóricas se han creado un total de 116 variables ficticias.

Por otro lado, es importante recordar que por lo que respecta a las variables continuas (metros y precio) a ambas se ha aplicado logaritmos para normalizarlas. De esta forma, se ha obtenido un primer modelo con las siguientes métricas:

R^2	0,5841
R^2 ajustado	0,5537
Raíz del error cuadrático medio	265,9593
Error absoluto medio	191,0862
AIC	-2657,423
P valor	< 2,2e-16

Tabla 7 *Métricas modelo inicial de regresión lineal*
Fuente: Elaboración propia

Es importante anotar que tanto en la raíz del error cuadrático medio como en el error absoluto medio se presenta ya el valor en euros. Es decir, al resultado obtenido que está en términos logaritmos se ha hecho el antilogaritmo para volverlo a mostrar en euros y poder apreciar bien la magnitud del error.

A continuación, aplicando el principio de economía o principio de parsimonia del filósofo y lógico Guillermo de Ockham según el cual: "en igualdad de condiciones, la explicación más sencilla suele ser la más probable" se procede a buscar el modelo más simple. Para ello, hay que verificar que todas las variables explicativas sean realmente necesarias o significativas para explicar la variabilidad del precio de una vivienda en alquiler. Para ello se obtiene la tabla anova y se observan los p valores correspondientes a cada uno de los parámetros que acompañan a las variables explicativas. En aquellos casos en que su p valor sea superior al nivel de significancia ($\alpha = 0,05$) se puede concluir que esa variable no es significativa.

Tras estudiar que parámetros eran significativos para explicar el precio de una vivienda en alquiler en Valencia se han eliminado las variables luz y ascensor. Así pues, para este modelo se han obtenido como se puede apreciar en la tabla 8 unas métricas prácticamente similares a las del anterior modelo.

R^2	0,556
R^2 ajustado	0,5243
Raíz del error cuadrático medio	267,2795
Error absoluto medio	190,8787
AIC	-2592,521
P valor	< 2,2e-16

Tabla 8 *Métricas del modelo de regresión tras eliminar variables no significativas*
Fuente: Elaboración propia

Una vez obtenido el mejor modelo es momento de realizar una validación del modelo donde el estudio de los residuales en los modelos de regresión juega un papel crucial. Por ello, se va a proceder a comprobar que cumpla con todas las hipótesis establecidas relativas con el error o perturbación:

- ✓ Normalidad de los residuos: para poder verificar la normalidad del error se utiliza el gráfico o papel probabilístico normal de los residuos. De esta manera podemos concluir que los residuos del modelo presentan una distribución normal, ya que tal como se puede observar en la gráfica 18 los residuos se ajustan de una forma casi perfecta a la línea. Asimismo, con el histograma se puede corroborar esta distribución normal de los residuos ya que tienen la típica forma de campana.

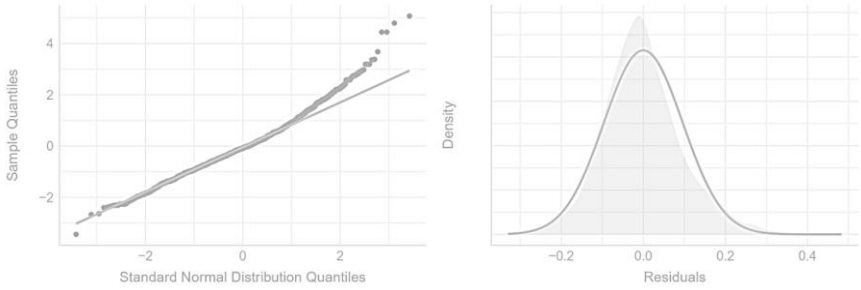

Gráfico 18 *Papel probabilístico normal de los residuos*
Fuente: Elaboración propia

- ✓ Linealidad de los residuos: con el gráfico de los residuos frente a los valores predichos es posible determinar si la perturbación depende de alguna variable explicativa. En el caso de que esto ocurriera, los residuos no estarían distribuidos al azar, sino más bien se obtendría una figura geométrica reconocible. En este caso, se puede concluir que los residuos sí que tiene una relación lineal tal como se puede apreciar en la gráfica 19.

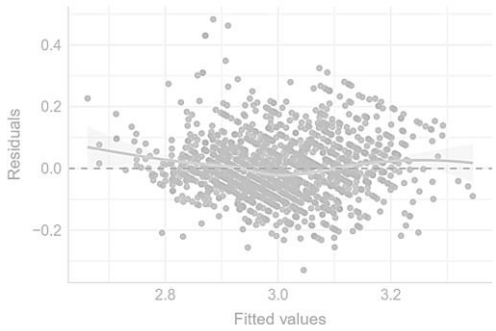

Gráfico 19 *Residuos frente a los valores predichos*
Fuente: Elaboración propia

- ✓ Homogeneidad en la varianza de los residuos: en este caso se representa la varianza de los residuos frente a los valores predichos. Lo que se busca en este gráfico es que la disposición de los puntos ocupe una franja cuyos límites inferior y superior sean paralelos al eje de abscisas. Tal como se puede ver en el gráfico pese a haber ciertos valores con una varianza más dispar, en términos generales se puede concluir que hay homogeneidad en los residuos.

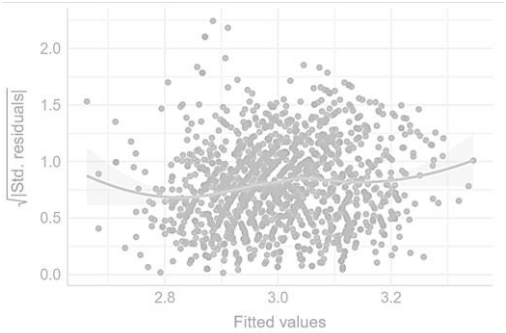

Gráfico 20 *Varianza de los residuos frente a los valores predichos*
Fuente: Elaboración propia

- ✓ Problemas de multicolinealidad: una de las hipótesis del modelo de regresión decía que las variables explicativas no debían estar relacionadas linealmente entre sí. Para verificar que no haya problemas de multicolinealidad se va a utilizar el factor de inflación de la varianza de la variable explicativa (VIF). Se establece que aquellas variables que tienen un VIF superior a cinco pueden tener problemas potenciales de multicolinealidad, en el caso de que sean mayores de diez se certifica entonces que existe multicolinealidad. Tal como se puede observar en la tabla 9 las variables ficticias relacionadas con el número de habitaciones presentan una fuerte multicolinealidad entre ellas. Para solucionar este problema se han eliminado las variables habitaciones.1, habitaciones.3 y habitaciones.5. De esta manera se ha conseguido erradicar el problema.

Moderate Correlation

	Term	VIF	Increased SE	Tolerance
habitaciones.6		7.08	2.66	0.14

High Correlation

	Term	VIF	Increased SE	Tolerance
tipo.Estudio		64.11	8.01	0.02
habitaciones.1		206.52	14.37	0.00
habitaciones.2		327.16	18.09	0.00
habitaciones.3		392.08	19.80	0.00
habitaciones.4		173.28	13.16	0.01
habitaciones.5		38.41	6.20	0.03

Tabla 9 *Factor de inflación de la varianza*
Fuente: Elaboración propia

- ✓ Valores atípicos: por último, falta comprobar que no haya valores atípicos para eso se ha utilizado la medida de *Leverage*. A la vista del gráfico 21 no se observa que ninguno sea un *outlier* o dato anómalo.

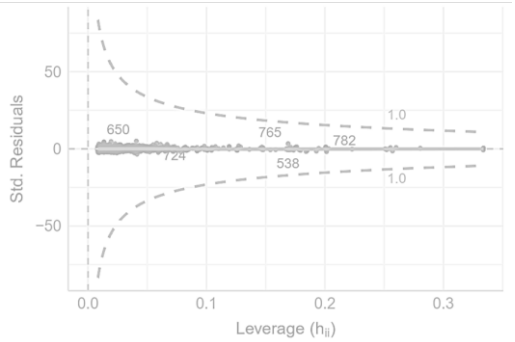

Gráfico 21 *Comprobación de valores atípicos*
Fuente: Elaboración propia

Tras realizar todas las comprobaciones y correcciones oportunas, el modelo mejorado presenta las métricas que se muestran en la tabla 10.

R^2	0,5718
R^2 ajustado	0,5545
Raíz del error cuadrático medio	264,0536
Error absoluto medio	190,5182
AIC	-2592,521
P valor	< 2,2e-16

Tabla 10 *Medidas finales del modelo de regresión lineal*
Fuente: Elaboración propia

Hasta el momento solo se ha utilizado el algoritmo de regresión lineal, sin embargo, tal como se ha introducido en el marco teórico existen múltiples algoritmos para resolver problemas de regresión. Tras probar varios, se ha hecho una selección de los cuatro con los que se ha obtenido unos errores menores, sus métricas se presentan en la siguiente tabla:

	Raíz del error cuadrático medio	Error absoluto medio
Random Forest	243, 1742	167,5934
XgbLinear	245,9988	173,9190
Cubist	250,4988	180,9246
Knn	293,2877	203,7011

Tabla 11 *Comparativa de los modelos*
Fuente: Elaboración propia

CAPÍTULO 4

Resultados y discusión

Una vez construidos los diferentes modelos es momento de analizar y reflexionar las distintas métricas que se han ido obteniendo para cada uno de ellos.

En primer lugar, se procede a estudiar el modelo de regresión lineal múltiple cuyos resultados se muestran en la tabla 10. Se ha obtenido un coeficiente de determinación (R^2) de 0,5718, lo cual significa que el 57,18% de la variabilidad de la variable precio se explica con el modelo creado. Siendo precisos y ajustándonos a la especificación del modelo explicaría la variabilidad en el logaritmo del precio. Por tanto, las variables tipo de vivienda junto con el barrio, el logaritmo de los metros cuadrados de la propiedad y el número de planta y habitaciones son capaces de explicar el 57,18% de la variabilidad del logaritmo del precio de un alojamiento en alquiler. De esta forma, habría algo más de un 40% de variabilidad en el precio que no puede ser explicada con estas variables y que por lo tanto habría que buscar nuevas variables que fueran capaces de captar dicha diversidad en el precio.

Otra de las métricas que resulta clave a la hora de analizar un modelo es el error absoluto medio (MAE). Para el de regresión lineal múltiple se ha obtenido un valor de 190,52€ lo que supone que la diferencia absoluta media entre los valores predichos y los valores objetivos es de esa cantidad. Por otro lado, se ha obtenido que la raíz del error cuadrático medio (RMSE) es de 264,05€. Esta métrica lo que calcula es la raíz cuadrada de la diferencia cuadrática promedio entre el valor real y el predicho por el modelo. No es ninguna sorpresa que el RMSE sea mayor al MAE ya que esta primera métrica por su naturaleza penaliza a los errores grandes.

Estas métricas anteriores se pueden comprender mejor si se representan gráficamente los valores predichos frente a los valores reales tal como se muestran en la gráfica 22. A simple vista la primera conclusión que se extrae es que el modelo predice mucho mejor en aquellas viviendas con un precio más bajo y que, sin embargo, tiene mayores dificultades a la hora de predecir viviendas con precios más altos.

Una primera característica que se puede observar tras analizar el gráfico es la tendencia del modelo a predecir un precio inferior al real. Esto es fácilmente percibidle gracias a la recta de regresión que separa ambas zonas, en la zona superior se sitúan todos los alquileres que se han predicho con un valor inferior al real y en la parte de abajo tendríamos aquellos cuyo valor predicho es inferior al real. Esta tendencia es especialmente notable en aquellos alquileres cuyo precio mensual es más alto. Por ejemplo, si se estudian los alquileres que tienen un precio superior a los 2000€ mensuales (parte central superior) para todos ellos se ha predicho un valor inferior a los 2000€.

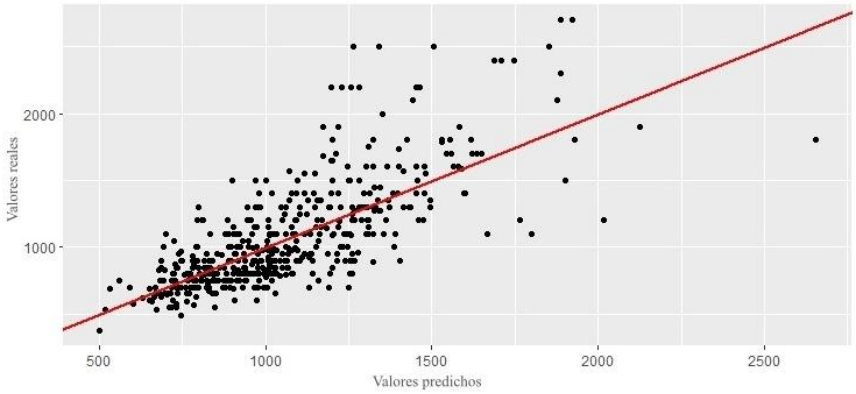

Gráfico 22 *Valores predichos frente a los valores reales*
Fuente: Elaboración propia

Además, a la vista del gráfico se observan tres comportamientos distintos según el precio real de la vivienda por eso se ha decidido calcular la raíz del error cuadrático medio y el error absoluto medio diferenciándolos para cada precio.

FRANJA DE PRECIOS	RMSE	MAE
0 – 1000€	174,1966	138,1430
0 – 1500€	187,6195	146,2243
0 – 2000€	213,9355	163,5073
0 – 2700€	264,0535	190,5182

Tabla 12 *Medidas de error según la franja de precios*
Fuente: Elaboración propia

Tal y como se venía anticipando el error aumenta conforme lo hace el precio de la vivienda y ha quedado demostrado numéricamente en la tabla 12. Se puede apreciar como el RSME disminuye un 51,58% y un 37,91% el MAE, si solo se tiene en cuenta las propiedades con un precio inferior a los 1000€ mensuales. Veamos si esto también ocurre en los otros cuatro modelos creados:

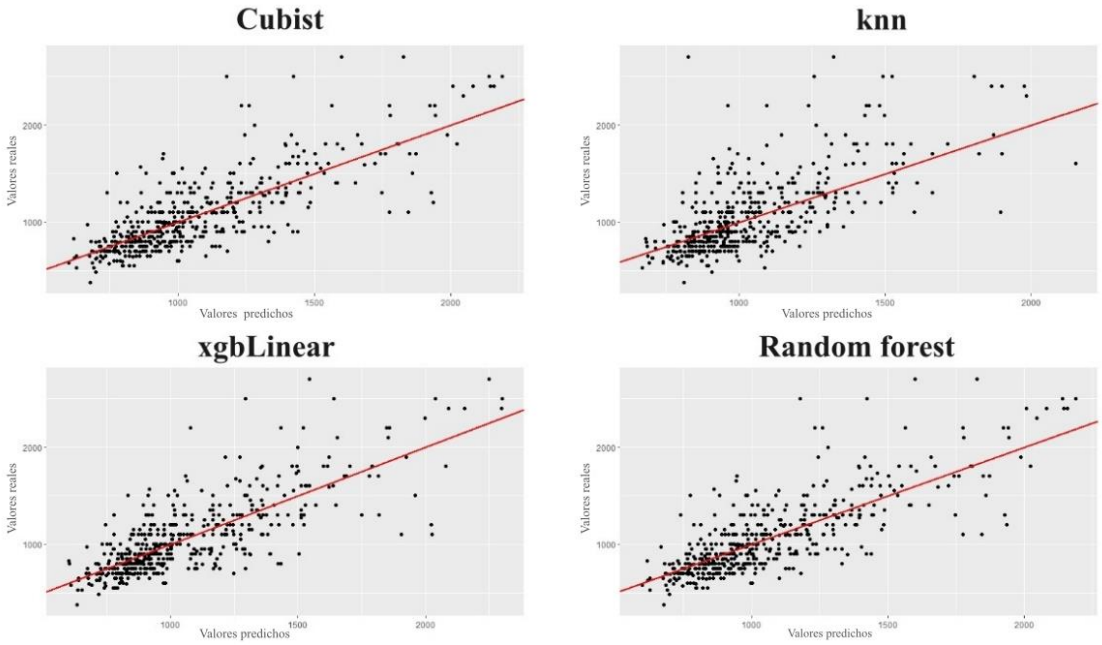

Gráfico 23 *Comparativa entre los distintos modelos*
Fuente: Elaboración propia

	CUBIST		KNN		XGBLINEAR		RANDOM FOREST	
FRANJA DE PRECIOS	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
0 – 1000€	187,66	140,77	179,85	144,50	175,86	129,16	170,05	125,97
0 – 1500€	204,42	152,35	191,06	152,30	196,78	144,66	185,00	135,79
0 – 2000€	227,46	168,05	228,51	175,31	217,67	159,30	208,56	151,16
0 – 2700€	250,50	180,92	293,29	203,70	246,00	173,92	243,17	167,59

Tabla 13 *Comparativa de las medidas de error según la franja de precios*
Fuente: Elaboración propia

Efectivamente, tal como infieren tantos los gráficos como la tabla creada, todos los modelos predicen mejor el precio del alquiler de las viviendas cuanto menor es su precio. Parte de la explicación de este comportamiento se explica por la propia naturaleza de la composición de la muestra. Si recuperamos el histograma de la variable precio (grafica 3) en él fácilmente se observaba que la gran mayoría de las viviendas tenían un precio inferior a los 1000€. De hecho, en la tabla 14 se puede comprobar como más de la mitad de las observaciones de la muestra corresponden con este tipo de viviendas. De ahí que al tener mayor cantidad de observaciones el modelo pueda entrenarse mejor por la cantidad mayor de observaciones que posee.

	Número de viviendas	Porcentaje de viviendas
< 1000 €	1204	56,39%
1000€ - 1500€	645	30,21%
1500€ - 2000€	207	9,70%
> 2000€	79	3,70%

Tabla 14 *Proporción de observaciones según la franja de precios*
Fuente: Elaboración propia

Asimismo, tras detectar estas cuatro particiones o subgrupos se ha decidido crear un modelo de clasificación que ayude a entender mejor el comportamiento y las características de cada uno de los subgrupos. Así pues, los grupos que se han establecido han sido los mismos cuatro que previamente, es decir, viviendas con un precio inferior a los 1000€, viviendas con un precio entre los 1000€ y 1500€, viviendas con un precio entre los 1500€ y 2000€ y viviendas con un precio superior a los 2000€.

Tras ejecutar el algoritmo del *random forest* se ha obtenido un modelo cuya *accuracy* (exactitud del modelo, porcentaje de aciertos) es de 72,22%. Además, se ha decidido obtener la matriz de confusión (tabla 15) para obtener un mayor nivel de detalle.

Predichos / Real	< 1000 €	1000€ - 1500€	1500€ - 2000€	> 2000€
< 1000 €	270	72	9	2
1000€ - 1500€	23	71	19	4
1500€ - 2000€	0	6	18	5
> 2000€	0	0	0	5

Tabla 15 *Matriz de confusión para el modelo de clasificación*
Fuente: Elaboración propia

Si se analiza detalladamente la matriz de confusión, una vez más se puede comprobar como la franja que mayor tasa de acierto tiene es la de viviendas con un precio inferior a los 1000€ mensuales. Tal como se puede muestra en la segunda columna de la tabla 15, el modelo de clasificación creado ha predicho bien 270 viviendas de un total de 293, lo que sitúa la tasa de acierto o *accuracy* de ese grupo en un 92,15%. No obstante, si se sigue analizando la matriz de confusión, se corrobora también la conclusión de que cuando falla el modelo este tiende a predecir un precio inferior al real. Por ejemplo, en el caso de la franja de alojamientos con un precio entre 1500€ y 2000€, solo 18 viviendas han sido clasificadas correctamente, 19 han sido clasificadas con un precio entre los 1000€ y los 1500€ y 9 de ellas han sido etiquetadas con un precio inferior a los 1000€.

Para acabar se ha impreso el modelo de clasificación construido con la finalidad de poder detectar algunas reglas que sirvan para establecer el precio de una propiedad.

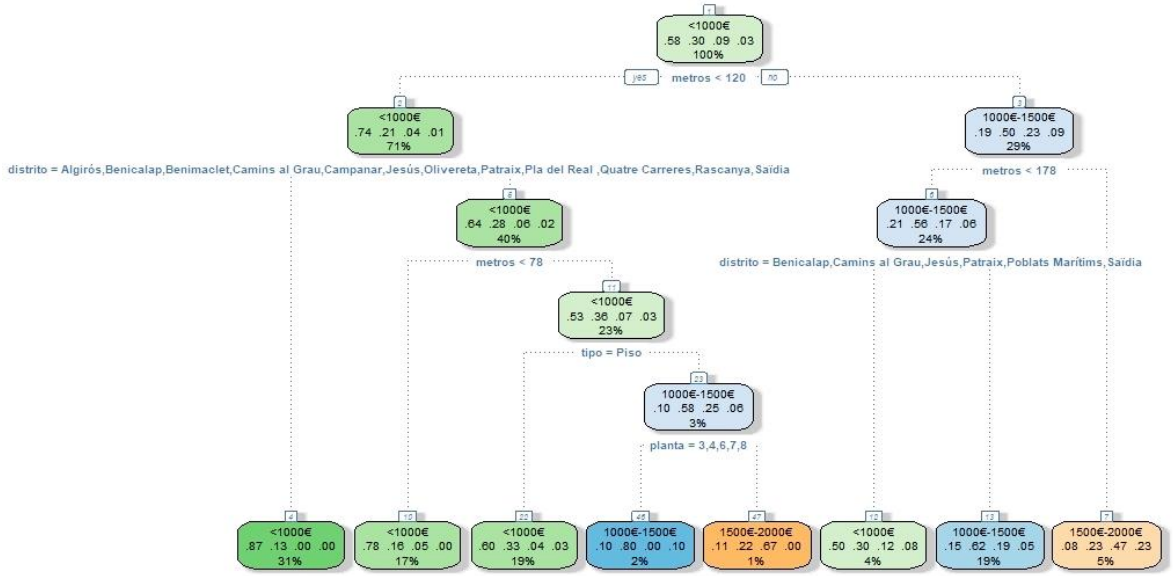

Gráfico 24 *Clasificación viviendas según franja de precios*
Fuente: Elaboración propia

Se procede ahora a enunciar las ocho reglas de decisión que se pueden extraer del árbol de decisión del gráfico 24:

1. Aquellas viviendas con una superficie inferior a los 120m^2 y que no pertenecen a los distritos de Ciutat Vella, Eixample, Extramurs o Poblats Marítims tienen por lo general un precio inferior a los 1000€ mensuales.
2. Las propiedades que tienen una superficie inferior a los 78m^2 y que pertenecen a los distritos de Ciutat Vella, Eixample, Extramurs o Poblats Marítims tienen por lo general un precio inferior a los 1000€ mensuales.
3. Los pisos que tiene una superficie comprendida entre los 78m^2 y los 120m^2 y que están ubicadas en los distritos de Ciutat Vella, Eixample, Extramurs o Poblats Marítims su precio suele estar por debajo de los 1000€ mensuales.
4. Aquellos estudios, dúplex o áticos que se encuentran en Ciutat Vella, Eixample, Extramurs o Poblats Marítims, tienen una superficie entre los 78m^2 y los 120m^2 y están situados entre un tercero y un octavo su precio suele oscilar entre los 1000 y 1500€.

5. Los estudios, dúplex o áticos que se encuentran en Ciutat Vella, Eixample, Extramurs o Poblats Marítims, tienen una superficie entre los 78m^2 y los 120m^2 y están situados a partir de una novena planta suelen presentar un precio comprendido entre los 1500€ y los 200€.
6. Aquellas viviendas con una superficie entre los 120m^2 y los 178m^2 y que pertenecen a los distritos de Benicalap, Camins al Grau, Jesús, Patraix, Poblats Marítims o Saïda su precio suelen ser inferior a los 1000€.
7. Aquellas viviendas con una superficie entre los 120m^2 y los 178m^2 y que no pertenecen a los distritos de Benicalap, Camins al Grau, Jesús, Patraix, Poblats Marítims o Saïda su precio suelen estar comprendido entre los 1000 y 1500€
8. Las viviendas con una superficie superior a los 178m^2 tienen por normal general un precio que oscila entre los 1500€ y 2000€.

CAPÍTULO 5

Conclusiones

En el presente proyecto ha sido abordada la tarea de analizar y predecir el precio en alquiler de vivienda en la ciudad de Valencia. Para ello, ha sido necesario llevar a cabo distintas actividades como la recogida de toda la información a través de un proceso de *web scraping*. A su vez, todos esos datos recolectados han tenido que ser transformados para dotarlos de un formato apto para la herramienta de análisis. Además, también se tuvo que realizar un preprocesamiento de los datos para asegurar la consistencia de los mismos.

Con todas estas tareas previas realizadas, se dio comienzo a los respectivos análisis. Primero estudiando cada variable de forma aislada y acto seguido de forma conjunta. Todo ellos sirvieron para entender mejor el panorama y poder desarrollar los diferentes modelos de regresión. Finalmente se ha verificado la validez de esos modelos y se estudió los resultados obtenidos con cada uno de ellos.

Tras haber repasado todo el trayecto recorrido a lo largo de este proyecto se procede a enumerar las conclusiones obtenidas:

- ✓ El precio promedio de las viviendas de alquiler ronda los 1000€ mensuales, siendo la cifra más repetida los 900€/mes. Es por eso que, algo más de la mitad de la oferta inmobiliaria tiene un precio inferior a los 1000€ al mes.
- ✓ La vivienda media tiene alrededor de unos 100 metros cuadrados, aunque es cierto que esta variable presenta una desviación típica frente a la media de 40 metros cuadrados.
- ✓ Hay notables diferencias de precio entre unos distritos y otros, entre los más caros destacan Ciutat Vella y Poblats Marítims, entre los más baratos: Saïda y Jesús. Además, afinando la búsqueda se desmarcan como los barrios más caros tanto el céntrico barrio del Carme como la Seu.
- ✓ En cuanto a los distritos con propiedades más grande, es decir, mayor superficie construida despunta el Eixample. Al lado contrario se encuentra el distrito de Oliverta con los inmuebles más pequeños, muy seguido del distrito de Poblats Marítims.
- ✓ Atendiendo al criterio del precio del metro cuadrado los estudios son la opción más cara mientras que las casas es la más barata.
- ✓ Con tan solo las variables tipo de inmueble, barrio, metros y número de habitaciones y planta se puede explicar el 57,18% de la variabilidad del precio de una propiedad en alquiler en la ciudad de Valencia.
- ✓ El mejor algoritmo (según las características y variables del problema) para predecir el precio de un alojamiento es el *random forest*.
- ✓ Tanto la variable metros cuadrados como el distrito son fundamentales para establecer el precio. Tanto es así que con solo estas dos ya que se pueden definir una horquilla de precios entre las que se debería fijar el precio de la vivienda.

5.1 Objetivos conseguidos

Al principio de este trabajo se establecieron cuatro objetivos muy concretos que se pretendían alcanzar al finalizarlo. Es momento de reflexionar si han sido alcanzados, así como de hacer balance de lo que se ha conseguido y lo que no.

El primero de los objetivos que se planteó fue diseñar un proceso que fuera capaz de crear una base de datos que recogiera la oferta actual que hay en el mercado inmobiliario del alquiler en la ciudad de Valencia. Este era sin duda un objetivo clave para poder desarrollar el resto del proyecto ya que era la materia prima de la que se iba a nutrir el resto del trabajo. Tras investigar la multitud de posibilidades que ofrece el *web scraping* se desarrolló un código que fuera capaz de recolectar la información básica que ofrece el portal de Idealista sobre un inmueble. No obstante, y pese haber sido capaces de conseguir reunir más de diez características sobre la propiedad, no se puede afirmar que este objetivo haya sido plenamente alcanzado. Tal como se comentó en el marco teórico, esta técnica está teniendo cada vez más popularidad y por lo tanto un auge enorme. Es por esto que, páginas como Idealista o enAlquiler están dedicando números esfuerzos para evitar este tipo de comportamientos ya sea con bloqueos a la IP o con estructuras de páginas web complejas y cambiantes. Esto ha dificultado, y mucho, el proceso de extracción de la información no pudiendo obtener toda la información que se hubiera querido.

Con el segundo de los objetivos se pretendía lograr un alto nivel de conocimiento sobre el panorama general del alquiler en Valencia. No hay duda de que este objetivo ha sido alcanzado ya que se ha adquirido un bagaje enorme relacionado con ese tema. Después de realizar este trabajo, se conoce cuáles son los barrios que mayor oferta presentan, así como cuales son los que tienen un precio por metro cuadrado mayor, entre otros muchos detalles. Además, para conseguir un mayor nivel de entendimiento se han ido elaborando a lo largo de toda la memoria distintas gráficas y tablas que facilitan la comprensión de toda esta información.

El penúltimo de los objetivos pretendía resolver la duda de qué factores son más influyentes a la hora de establecer el precio de una vivienda. Para aclarar esta cuestión primero se realizaron diversos test que permitieron cuantificar la relación de todas las variables con el precio. No obstante, al construir los modelos se siguió investigando sobre este tema para poder esclarecer aún más las conclusiones obtenidas. Todos los modelos coincidían que las variables más importantes eran en primer lugar, los metros cuadrados de la vivienda y, en segundo lugar, el barrio (o en su defecto el distrito) al que pertenecía esta. Así pues, para poder visualizar aún más claro esto, se decidió construir gráficamente el árbol de clasificación. Con él se puede ver rápidamente que tanto los metros cuadrados como el distrito son las primeras características a tener en cuenta a la hora de situar un inmueble entre una franja de precios u otra.

Por último, el cuarto objetivo abarcaba todo el tema de la construcción de un modelo que fuera capaz de predecir el precio de una propiedad en alquiler en Valencia. No hay duda de que junto al primero objetivo estos eran los dos más ambiciosos. Se elaboró un primer modelo de regresión que después de refinarlo y validarlo seguía teniendo unos resultados, a mi juicio, bajos. Por lo que se decidió explorar con métodos más potentes para ver si con alguno de ellos se podía reducir el error. En efecto, se consiguió mejorar las primeras métricas, sin embargo siguen siendo unos resultados que tienen mucho margen de mejora. Es por esto por lo que, pese a todos los esfuerzos de crear ya no solo un modelo de regresión, sino también de clasificación no se ha logrado obtener un modelo que realmente pueda predecir sin demasiado margen de error el precio.

5.2 Lecciones aprendidas

Otra de las etapas fundamentales al finalizar un proyecto es hacer un repaso analizando aquellos errores que se han cometido, así como realzar las lecciones que se han podido aprender con la realización de este.

El primero de los errores en el que se ha incurrido está relacionado con la verificación de la validez de la muestra. En un primer momento tras finalizar el proceso de *web scraping* todo parecía correcto ya que se había obtenido una muestra de más de 2000 observaciones con a penas valores faltantes. Sin embargo, no ha sido hasta la mitad de la realización del trabajo cuando se ha visto que esta muestra tenía ciertos problemas ya que había muy pocas observaciones atendiendo a unas características determinadas del inmueble. Es cierto que este problema es consecuencia de la propia naturaleza de la oferta de viviendas en alquiler en la ciudad de Valencia. Es decir, no habría motivos para decir que esta muestra no es representativa, pero habría sido interesante aumentar el tamaño de la muestra para suplir esos estratos con pocas observaciones. De esta manera quizás los modelos habrían tenido menos dificultades con aquellas viviendas que tienen un precio más alto dado que hubieran dispuesto de una gama de entrenamiento mucho más amplia.

Otro de los errores que se ha cometido está ligado con la detección y tratamiento de datos anómalos. Haciendo memoria se recordará que se estudiaron tanto la variable precio como los metros cuadrados calculando los límites superiores a partir de cuales considerarlos como *outliers.* Sin embargo, durante el desarrollo del proyecto al buscar otros estudios y trabajos se ha encontrado una forma que parece más adecuada para este caso. Con la forma con la que se procedió solo se obtuvieron valores extremos por la parte superior, es decir no había límite inferior. Este tipo de variables es muy común que tengan una cola muy larga por la derecha debido al contexto que representan. Por eso, la solución más optima habría sido crear una nueva variable con el precio del metro cuadrado y a esta variable aplicarle el logaritmo. De esta manera se analizarían ambas variables de forma conjunta y seguramente ya aparecerían anómalos tanto en la parte inferior como en la superior. A posteriori, se ha probado este método en el contexto del problema y apenas se han notado diferencias con la forma original de proceder. Tan solo han aparecido dos datos anómalos por la parte derecha de la cola y prácticamente los mismo por la parte izquierda. Con lo que en este caso no tiene mayor importancia, pero se anota para futuros trabajos.

Detectados los errores que se han hecho durante el transcurso de este trabajo y comentadas cuales hubieran sido una mejor propuesta, se procede ahora a revisar todas aquellas lecciones que se han aprendido gracias a la realización de este trabajo.

Probablemente uno de los aspectos en los que se ha experimentado un mayor cambio ha sido con el software R. Durante estos meses, se ha tenido una curva de aprendizaje muy positiva con esta herramienta, llegando a alcanzar un buen dominio de ella. Se ha aprendido a utilizar muchas librerías nuevas como todas las relacionadas con el procesamiento de textos o aquellas más centradas en la validación de modelos. Además, se ha conseguido obtener una gran agilidad a la hora de realizar distintos gráficos y tablas, adaptándolas siempre a las características de los datos a mostrar.

Además, indudablemente también se ha obtenido muchos conocimientos relacionados con toda el área del *machine learning*. Se han descubierto nuevos algoritmos para resolver distintos problemas de regresión, así como de clasificación. Aprendiendo a evaluar cada uno de ellos con diferentes métricas que cuantifican el error que tiene el modelo.

5.3 Líneas futuras

Una de las primeras líneas sobre las que se puede trabajar en el futuro es en la introducción de más características en el modelo. Así pues, se pueden agrupar en dos grandes grupos según si están o no relacionadas puramente con las características de la propiedad. En un primer grupo se encontrarían aquellas variables no conectadas con las funcionalidades de la vivienda, en este se puede incluir los tipos de interés vigentes en el momento o distintas variables que recojan la demanda potencial de viviendas. Para esta última se tendrían que incluir distintas variables socioeconómicas y demográficas. En cuanto a las variables íntimamente relacionadas con el inmueble se puede completar la información con el año de construcción de la vivienda o el tipo de calefacción que posee.

Otra de las líneas en las que hay que trabajar en un futuro es en la generalización de este modelo para poder aplicarlo en distintas ciudades españolas. De esta manera resultaría muy positivo poder aplicar un mismo modelo para predecir el precio de una vivienda independiente de si se trata de Barcelona, Madrid, Valencia o Sevilla.

CAPÍTULO 6

Bibliografía

Abhishek, Sharma. 2020. «Decision Tree vs. Random Forest - Which Algorithm Should You Use?» *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (2 de marzo de 2022).

Barnett, Neal. 2020. «Web Scraping Using Python Selenium». *Toptal Engineering Blog*. https://www.toptal.com/python/web-scraping-with-python (2 de marzo de 2022).

Crespo Abril, Fortunato. 2017. *Métodos estadísticos: ejercicios resueltos y teoría*. Valencia: Editorial de la Universidad Politécnica de Valencia.

Diouf, Rabiyatou et al. 2019. «Web Scraping: State-of-the-Art and Areas of Application». En *2019 IEEE International Conference on Big Data (Big Data)*, , 6040-42.

Espinosa-Zúñiga, Javier Jesús. 2020. «Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito». *Ingeniería, investigación y tecnología* 21(3). http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1405-77432020000300002&lng=es&nrm=iso&tlng=es (25 de febrero de 2022).

Hernández Sampieri, Roberto. 2014. *Metodología de la investigación*. 6ª ed. México [etc: McGraw-Hill/Interamericana.

Louridas, Panos, y Christof Ebert. 2016. «Machine Learning». *IEEE software* 33(5): 110-15.

Marco Sanjuán, Francisco Javier. 2018. «Outlier - Definición, qué es y concepto». *Economipedia*. https://economipedia.com/definiciones/outlier.html (7 de febrero de 2022).

Myles, Anthony J. et al. 2004. «An Introduction to Decision Tree Modeling». *Journal of Chemometrics* 18(6): 275-85.

Park, Byeonghwa, y Jae Kwon Bae. 2015. «Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data». *Expert Systems with Applications* 42(6): 2928-34.

Patel, Jay M. 2020. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*. Berkeley, CA: Apress. http://link.springer.com/10.1007/978-1-4842-6576-5 (9 de febrero de 2022).

Phan, The Danh. 2018. «Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia». En *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, , 35-42.

Rezazadeh Kalehbasti, Pouya, Liubov Nikolenko, y Hoormazd Rezaei. 2021. «Airbnb Price Prediction Using Machine Learning and Sentiment Analysis». En *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, eds. Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, y Edgar Weippl. Cham: Springer International Publishing, 173-84.

Romero Villafranca, Rafael, y Luisa Rosa Zúnica Ramajo. 2020. *Métodos estadísticos para ingenieros*. Valencia: Editorial de la Universidad Politécnica de Valencia.

STS 572-2012. 2012. https://vlex.es/vid/desleal-ryanair-atrapalo-as-411388894 (9 de febrero de 2022).

Suganya, E., y S. Vijayarani. 2020. «Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms». En *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing, eds. Ajith Abraham, Aswani Kumar Cherukuri, Patricia Melin, y Niketa Gandhi. Cham: Springer International Publishing, 677-85.

«Supervised vs. Unsupervised Learning: What's the Difference?» 2021. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (23 de febrero de 2022).

Tejada, Haroldo Elorza Pérez. 2014. *Estadistica para las ciencias sociales, del comportamiento y de la salud*. Universidad de Alicante.

«Uses of Web Scraping». https://www.webharvy.com/articles/web-scraper-use-cases.html (9 de febrero de 2022).

Varghese, Danny. 2019. «Comparative Study on Classic Machine Learning Algorithms». *Medium*. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 (2 de marzo de 2022).

CAPÍTULO 7

Anexos I: El TFG y los ODS

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Reflexión sobre la relación del TFG con los ODS en general y con el/los ODS más relacionados.

En este anexo se va a analizar la relación existente entre el ya desarrollado presente trabajo fin de grado y los Objetivos de Desarrollo Sostenible (ODS) estipulados por la Organización de las Naciones Unidas (ONU). Para ello, en un primer momento se va a explicar en qué consisten estos objetivos, además de explicar cómo surgen. Finalmente, se establecerá la relación que hay entre alguno de ellos y este proyecto.

En el año 2015 todos los Estados Miembros de las Naciones Unidas firmaron los 17 ODS que formarían parte de la llamada Agenda 2030. En esta se establecía un plan para alcanzar un futuro sostenible en un plazo de 15 años. Con ellos se pretende tal como leer en su página web no dejar a nadie atrás, para ellos se centra en los desafíos globales como la pobreza, justicia, , el clima, la degradación ambiental, la paz, la prosperidad y la desigualdad.

Con uno de los objetivos que se puede relacionar es aquel que abarca toma el tema de la industria, innovación e infraestructuras, en concreto, se trata del objetivo número 9. Este tipo de técnicas de *machine learning* que se han utilizado suponen toda una innovación y avance frente a las tradicionales técnicas. De esta manera aplicando este tipo de métodos se puede lograr obtener predicciones muy certeras sobre el precio de una vivienda. Hecho que sin la aplicación de este tipo de técnicas es un proceso mucho más complejo porque son muchas las variables que hay que tener en cuenta a la hora de cifrar una vivienda.

Por otro lado, se encuentra el objetivo número 11, el cual está íntimamente relacionado con las ciudades y las comunidades sostenibles. Tal como se ha comentado las técnicas de web scraping son muy potentes y son una fuente inmensa para recolectar gran cantidad de información. Hoy en día el mercado del alquiler, especialmente en España, supone todo un reto debido a la poca oferta que hay fruto del estallido de la pandemia. Por eso, poder rastrear diferentes portales inmobiliarios es vital para poder entender como se encuentra el mercado y a partir de ahí definir nuevas medidas que ayuden a solucionar este enorme problema.

CAPÍTULO 7

Anexos II: Códigos

7.1 Método web scraping

```
// contador de paginas
var page = 1;
// propiedades
var properties = [];
// metodo paginacion
function crawlNextPage() {
  // abrir URL en una ventana
  var response = window.open('https://www.idealista.com/alquiler-viviendas/valencia/poblats-
maritims/', 'new', true);
  // iniciar una ventana en funcion onload
  response.onload = function crawlNextPage() {
    // esperar a que el contenido se obtenga en una ventana recien creada
    setTimeout(function () {
      if ($(response.document).find('li[class="next"] a')[0]) {
        // imprimir información de depuración
        console.log('crawling next page...');
        // extracición lógica de datos
        $(response.document).find('article[class="item item-multimedia
container"]').each(function() {
          // extración características
          var features = {
            'title': $(this).find('a[class="item-link"]')
                        .text(),
            'details_url': 'https://www.idealista.com/' + $(this).find('a[class="item-link "]')
                           .attr('href'),
            'price': $(this).find('span[class="item-price h2-simulated"]')
                        .text(),
            'details': '',
            'description': $(this).find('div[class="item-description description"]')
                           .text()
                           .trim(),
          };
          // extracción de los detalles
          $(this).find('span[class="item-detail"]').each(function() {
            features.details += ' ' + $(this).text();
          });
          features.details = features.details.trim()
          console.log(features);
          properties.push(features);
        });
        //extracción y click boton de siguiente
        $(response.document).find('li[class="next"] a')[0].click();
```

```
      //incrementar contador de página
      page++;
      // rastrear la paginación recursivamente
      if (page < 25)
        crawlNextPage();
      else
        console.log('Reached page limit number!');
    } else {
      console.log('All done!');
    }
  },3000)
};
```

7.2 Script R studio

```
#########################
#LIBRERIAS#
#########################
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(regexPipes)
library(ggplot2)
library(dplyr)
library(openxlsx)
library(ggplot2)
library(descr)
library(tidyverse)
library(tokenizers)
library(dataMaid)
library(dummy)
library(EDAWR)
library(psych)
library(Hmisc)
library(reshape2)
library(caret)
library(tibble)
library(nortest)
library(MASS)
library(faraway)
library(ggpubr)
library(car)
library(EnvStats)
library(moments)
library(performance)
library(see)
library(patchwork)
library(rstanarm)
library(Metrics)
library(rpart)
library(rattle)
library(rpart.plot)
```

```
########################
#CARGA DE  LOS DATOS#
########################
datos <- read.xlsx("C:\\Users\\34673\\Documents\\UNIVERSIDAD\\TFG_ADE\\DATOS\\base
de datos_1.xlsx")
glimpse(datos)
#Renombrando las variables
colnames(datos) <- c('tipo',
                     'calle',
                     'barrio',
                     'precio',
                     'habitaciones',
                     'metros',
                     'ascensor',
                     'luz',
                     'descripcion',
                     'distrito',
                     'tipo_planta',
                     'planta',
                     'piscina',
                     'garaje',
                     'reformado',
                     'renta')
#Factorizacion de las variables
datos$tipo <- as.factor(datos$tipo)
datos$ascensor <- as.factor(datos$ascensor)
datos$luz <- as.factor(datos$luz)
datos$distrito <- as.factor(datos$distrito)
datos$tipo_planta <- as.factor(datos$tipo_planta)
datos$planta <- as.factor(datos$planta)
datos$barrio <- as.factor(datos$barrio)
datos$habitaciones <- as.factor(datos$habitaciones)
datos$piscina <- as.factor(datos$piscina)
datos$garaje <- as.factor(datos$garaje)
datos$reformado <- as.factor(datos$reformado)
########################
#ANALISIS VARIABLE DESCRIPCION#
########################
corpus <- Corpus(VectorSource(datos$descripcion))
#Cambiar a minusculas
d  <- tm_map(corpus, tolower)
#Quitar espacios en blanco
d  <- tm_map(d, stripWhitespace)
#Quitar signos de puntuaci?n
d <- tm_map(d, removePunctuation)
#Quitar palabras vacias genericas
stopwords("spanish")
d <- tm_map(d, removeWords, stopwords("spanish"))
#Crear la matriz de terminos
tdm <- TermDocumentMatrix(d)
findFreqTerms(tdm, lowfreq=20)
frecuentes<-findFreqTerms(tdm, lowfreq=20)
asociaciones <- findAssocs(tdm, frecuentes, rep(0.45, rep=5) )
m <- as.matrix(tdm)
#Se ordena y se suma
v <- sort(rowSums(m),decreasing=TRUE)
#Dar formato de data.frame
df <- data.frame(word = names(v),freq=v)
```

```
#Crear grafica palabras mas frecuentes
barplot(df[1:20,]$freq, las = 2, names.arg = df[1:20,]$word,
        col ="lightblue", main ="PALABRAS M?S FRECUENTES", ylab = "Frecuencia de
palabras")
masfrecuentes <- df[1:20,]$word
#Crear nube de palabras mas frecuentes
wordcloud(words = df$word, freq = df$freq, min.freq = 6,
          max.words=80, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
#Pisos con piscina
sum(str_count(d, "piscina"))
# Pisos con garaje
sum(str_count(d, "garaje"))
#Pisos reformados
sum(str_count(d, "reformado"))
#########################
#ANALISIS DE VALORES ANOMALOS#
#########################
boxplot(datos$precio, ylab = "Precio")
summary(datos$precio)
IQR(datos$precio)
#Calculamos el bigote maximo --> 3 cuadrante + (3*IQR) = 2800
Tmax = 1300+(3*500)
Tmin = 375 -(3*500)
#Encontramos outlier
datos$precio[which(datos$precio > Tmax)]
boxplot(datos$metros, ylab = "Metros cuadrados")
summary(datos$metros)
IQR(datos$metros)
#Calculamos el bigote maximo --> 3 cuadrante + (3*IQR) = 287
Tmax = 128+(3*53)
Tmin = 19-(3*53)
#Encontramos outlier
datos$metros[which(datos$metros > Tmax)]
#Elimimanos estos datos anomalos
datos_sin <- subset(datos, precio < 2800 & metros < 287)
#Comprobamos datos anomalos con variable precio/m2
datos$preciom2 <- datos$precio / datos$metros
datos$logpreciom2 <- log10(datos$preciom2)
boxplot(datos$logpreciom2)
summary(datos$logpreciom2)
IQR(datos$logpreciom2)
#Calculamos el bigote maximo --> 3 cuadrante + (3*IQR) = 1.6532
Tmax = 1.1162+(3*0.179006)
Tmin = 0.9372-(3*0.179006)
#########################
#NA VALUES Y DUPLICADOS#
#########################
#Deteccion de duplicados
datos_sin[duplicated(datos_sin),]
#NUmero total de filas duplicadas --> 21
nrow (datos_sin) - nrow (distinct (datos_sin))
#Elimando los duplicados
datos_sin <- datos_sin %>% distinct()
#DetecciOn de missing values(NA)
sum(is.na(datos_sin))
#Missing values por variable
apply(is.na(datos_sin),2, sum)
```

```
########################
#ANALISIS INDIVIDUALIZADO DE CADA VARIABLE#
########################
freq(datos_sin$tipo, xlab = "Tipo de vivienda", ylab = "Frecuencia")
ggplot(datos_sin, aes(tipo)) + geom_bar()
freq(datos_sin$barrio)
hist(datos_sin$precio, main = "Histograma variable precio", xlab = "Precio", ylab =
"Frecuencia")
summary(datos_sin$precio)
sd(datos_sin$precio)
mode <- function(x) {
  return(as.numeric(names(which.max(table(x)))))
}
mode(datos_sin$precio)
freq(datos_sin$habitaciones, xlab = "Número de habitaciones", ylab = "Frecuencia")
summary(datos_sin$habitaciones)
hist(datos_sin$metros, main = "Histograma variable metros", xlab = "Metros cuadrados", ylab =
"Frecuencia")
summary(datos_sin$metros)
sd(datos_sin$metros)
mode(datos_sin$metros)
freq(datos_sin$metros, main = "Número de viviendas según m2")
freq(datos_sin$ascensor, main = "Número de viviendas con o sin ascensor")
proporciones <- c(1751, 340)
etiquetas <- c("Con ascensor", "Sin ascensor")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas,"%",sep="")
#Dibujamos el diagrama de tarta con la variable luz
pie(proporciones,labels = etiquetas,
    col = gray.colors(2))
freq(datos_sin$luz, main = "Número de viviendas según el tipo de luz")
proporciones <- c(1921, 165)
etiquetas <- c("Exterior", "Interior")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas,"%",sep="")
#Dibujamos el diagrama de tarta con la variable piscina
pie(proporciones,labels = etiquetas,
    col = gray.colors(2))
freq(datos_sin$distrito, main = "Número de viviendas según el distrito")
freq(datos_sin$tipo_planta, main = "Número de viviendas según el tipo de planta", col =
palette("Tableau 10"))
freq(datos_sin$planta, xlab = "Número de planta", ylab = "Frecuencia")
freq(datos_sin$piscina, main = "Número de viviendas según si tienen piscina o no")
proporciones <- c(2093, 42)
etiquetas <- c("Sin piscina", "Con piscina")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas,"%",sep="")
#Dibujamos el diagrama de tarta con la variable garaje
pie(proporciones,labels = etiquetas,
    col = gray.colors(2))
freq(datos_sin$garaje, main = "Número de viviendas según si tienen garaje o no")
proporciones <- c(2076, 59)
etiquetas <- c("Sin garaje", "Con garaje")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
```

```
etiquetas <- paste(etiquetas,"%",sep="")
#Dibujamos el diagrama de tarta
pie(proporciones,labels = etiquetas,
   col = gray.colors(2))
freq(datos_sin$reformado, main = "Número de viviendas según si estan reformados o no")
proporciones <- c(1750, 385)
etiquetas <- c("Sin reformar", "Reformadas")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas,"%",sep="")
#Dibujamos el diagrama de tarta
pie(proporciones,labels = etiquetas,
   col = gray.colors(2))
#########################
#ESTUDIO DE LA NORMALIDAD DE LAS VARIABLES#
#########################
#Distribucion de la variable precio
shapiro.test(datos_sin$precio)
skewness(datos_sin$precio, na.rm = TRUE)
qqnorm(datos_sin$precio)
qqline(datos_sin$precio)
ggdensity(datos_sin, x = "precio", fill = "lightgray", title = "Distribución de la variable precio")
+
  stat_overlay_normal_density(color = "red", linetype = "dashed")
ggplot(datos_sin$precio)
#Distribucion de la variable metros
shapiro.test(datos_sin$metros)
skewness(datos_sin$metros, na.rm = TRUE)
qqnorm(datos_sin$metros)
qqline(datos_sin$metros)
ggdensity(datos_sin, x = "metros", fill = "lightgray", title = "Distribución de la variable metros")
+
  stat_overlay_normal_density(color = "red", linetype = "dashed")
#Distribucion de la variable renta
shapiro.test(datos_sin$renta)
skewness(datos_sin$renta, na.rm = TRUE)
qqnorm(datos_sin$renta)
qqline(datos_sin$renta)
#Transformación de la variable precio y metros
datos_sin$logprecio <- log10(datos_sin$precio)
skewness(datos_sin$logprecio, na.rm = TRUE)
ggdensity(datos_sin, x = "inversaprecio", fill = "lightgray", title = "Inversa del precio") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
datos_sin$logmetros <- log10(datos_sin$metros)
skewness(datos_sin$logmetros, na.rm = TRUE)
#########################
#ANALISIS BIVARIANTE ENTRE LAS VARIABLES#
#########################
#Relación entre la variable precio y tipo
datos_sin %>% group_by(tipo) %>% summarise(media = mean(precio),
                                           mediana = median(precio),
                                           sd = sd(precio),
                                           n = n())
data1 <- na.omit(datos[, c('precio', 'tipo')])
data1 <- data1 %>% group_by(tipo)
data1 <- mutate(na.omit(data1),
              Precio_medio = mean (precio))
ggplot(data1)+
```

```
  geom_bar(aes(x=tipo, y=Precio_medio),
           stat = "summary")+
  labs(x= "Tipo de vivienda", y="Precio medio")
  #geom_jitter(width = 0.25)
kruskal.test(precio ~ tipo, data = datos_sin)
pairwise.wilcox.test(x = datos_sin$precio, g = datos_sin$tipo, p.adjust.method = "holm")
#Relación entre la variable precio y barrio
kruskal.test(precio ~ barrio, data = datos_sin)
#Relación entre la variable precio y habitaciones
data1 <- na.omit(datos[, c('precio', 'habitaciones')])
data1 %>% group_by(habitaciones) %>% mutate (media = mean(precio),
                                    mediana = median(precio),
                                    sd = sd(precio),
                                    n = n())
kruskal.test(precio ~ habitaciones, data = data1)
data1 <- na.omit(datos[, c('precio', 'habitaciones')])
data1 <- data1 %>% group_by(habitaciones)
data1 <- mutate(na.omit(data1),
                Precio_medio = mean (precio))
ggplot(data1)+
  geom_bar(aes(x=habitaciones, y=Precio_medio),
           stat = "summary")+
  labs(x= "Habitaciones", y="Precio medio")
#Relación entre la variable precio y planta
data1 <- na.omit(datos_sin[, c('precio', 'tipo_planta')])
kruskal.test(precio ~ tipo_planta, data = data1)
#Relación entre la variable precio y numero de planta
data1 <- na.omit(datos_sin[, c('precio', 'planta')])
data1 %>% group_by(planta) %>% summarise(media = mean(precio),
                                      mediana = median(precio),
                                      sd = sd(precio),
                                      n = n())
kruskal.test(precio ~ planta, data = data1)
data1 <- data1 %>% group_by(planta)
data1 <- mutate(na.omit(data1),
                Precio_medio = mean (precio))
ggplot(data1)+
  geom_bar(aes(x=planta, y=Precio_medio),
           stat = "summary")+
  labs(x= "Número de planta", y="Precio medio")
#Relación entre la variable precio y distrito
kruskal.test(precio ~ distrito, data = datos_sin)
pairwise.wilcox.test(x = datos_sin$precio, g = datos_sin$distrito, p.adjust.method = "holm" )
boxplot(precio  ~ distrito, data = datos_sin )
#Relación entre la variable precio y ascensor
data1 <- na.omit(datos_sin[, c('precio', 'ascensor')])
wilcox.test(precio ~ ascensor, data=datos_sin, alt="two.sided", paired = F)
data1 %>% group_by(ascensor) %>% summarise(media = mean(precio),
                                    mediana = median(precio),
                                    sd = sd(precio),
                                    n = n())
boxplot(data1$precio  ~ data1$ascensor, xlab = "Ascensor", ylab = "Precio")
#Relación entre la variable precio y luz
data1 <- na.omit(datos_sin[, c('precio', 'luz')])
wilcox.test(precio ~ luz, data=data1, alt="two.sided", paired = F)
data1 %>% group_by(luz) %>% summarise(media = mean(precio),
                                    mediana = median(precio),
                                    sd = sd(precio),
```

```
                                  n = n())
#Relación entre la variable precio y garaje
wilcox.test(precio ~ garaje, data = datos_sin)
datos_sin %>% group_by(garaje) %>% summarise(media = mean(precio),
                                mediana = median(precio),
                                sd = sd(precio),
                                n = n())
#Relación entre la variable precio y piscina
wilcox.test(precio ~ piscina, data = datos_sin)
datos_sin %>% group_by(piscina) %>% summarise(media = mean(precio),
                                    mediana = median(precio),
                                    sd = sd(precio),
                                    n = n())
#Relación entre la variable precio y reforma
wilcox.test(precio ~ reformado, data = datos_sin)
datos_sin %>% group_by(reformado) %>% summarise(media = mean(precio),
                                     mediana = median(precio),
                                     sd = sd(precio),
                                     n = n())
#Relación entre la variable precio y metros
cor(datos_sin$precio, datos_sin$metros, method = "spearman")

ggplot(datos_sin, aes(x=precio, y=metros)) + geom_point() +
  xlab("Precio") + ylab("Metros cuadrados") +
  geom_smooth(formula = y ~ x, method="lm", col = "red")
#Tabla para saber el precio del m2 en cada distrito
df_grafica1 = datos_sin %>% group_by(distrito)  %>%
  summarise(precio_promedio = mean(precio),
            metros_promedio = mean(metros),
            precio_m2 = precio_promedio / metros_promedio,
            .groups = 'drop')
#Comparando tipo vivienda con otras variables
data1 <- subset(datos_sin, tipo != "Cortijo")
ggplot(data1, aes(metros, precio))+
  geom_count() +
  labs(x = "Metros", y="Precio")+
  facet_wrap(~ tipo)
df_grafica2 = data1 %>% group_by(tipo)  %>%
  summarise(precio_promedio = mean(precio),
            metros_promedio = mean(metros),
            precio_m2 = precio_promedio / metros_promedio,
            .groups = 'drop')
tabla <- table(data1$barrio, data1$tipo)
#Tabla para saber el precio del m2 en cada barrio
df_grafica3 = datos_sin %>% group_by(barrio)  %>%
  summarise(precio_promedio = mean(precio),
            metros_promedio = mean(metros),
            precio_m2 = precio_promedio / metros_promedio,
            .groups = 'drop')
#Numero de planta segun el distrito
data1 <- na.omit(datos_sin[, c('planta', 'distrito')])
ggplot(data1) + geom_bar(aes(y = distrito, fill = planta))+
                      labs(x = "Frecuencia", y = "Distrito", col = "Número planta")
```

```
########################
#MODELO DE REGRESION#
########################
#MODELO 1
data_r <- datos_sin[, c('logprecio','logmetros', 'tipo', 'barrio', 'planta', 'habitaciones', 'ascensor',
'luz')]
data_r <- datos_sin[, c('logprecio','logmetros', 'tipo', 'barrio', 'planta', 'habitaciones')]
data_r <- na.omit(data_r)
set.seed(1234)
data_r12 <- dummyVars(" ~ .", data = data_r)
data_r13 <- data.frame(predict(data_r12, newdata = data_r))
inTraining <- createDataPartition(data_r13$logprecio, p = 0.75, list = FALSE)
training <- data_r13[inTraining,]
testing <- data_r13[-inTraining,]

lm_4 <- lm(formula = logprecio ~., data = training)
summary(lm_4)
anova(lm_4)
AIC(lm_4)
prediction4 <-predict(lm_4,testing)
lm4_rmse <- RMSE(prediction4, testing$logprecio)
data_mod_sin <- data.frame(Predicted = prediction4, Observed = testing$logprecio, pred_price
= 10^(prediction4), real_price = 10^(testing$logprecio))
rmse_linear_sin <- RMSE(data_mod_sin$pred_price, data_mod_sin$real_price)
mae_linear_sin <- MAE(data_mod_sin$pred_price, data_mod_sin$real_price)

lm_4_op <- MASS::stepAIC(lm_4, direction = "both", trace = FALSE)
lm_4_op <- lm(formula = logprecio ~ logmetros + tipo.Ático + tipo.Dúplex +
        tipo.Estudio + barrio.Aiora + barrio.Albors + barrio.Arrancapins +
        barrio.Benicalap + barrio.Benimaclet + barrio.Camí.Fondo +
        barrio.Campanar + barrio.Ciutat.de.les.Arts.i.de.les.Ciencies +
        barrio.Ciutat.Jardí + barrio.Ciutat.Universitària + barrio.El.Botànic +
        barrio.El.Cabanyal.El.Canyamelar + barrio.El.Carme + barrio.El.Grau +
        barrio.El.Mercat + barrio.El.Pilar + barrio.El.Pla.del.Remei +
        barrio.En.Corts + barrio.Exposició + barrio.Gran.Vía + barrio.Jaume.Roig +
        barrio.L.Amistat + barrio.L.Illa.Perduda + barrio.La.Carrasca +
        barrio.La.Creu.del.Grau + barrio.La.Petxina + barrio.La.Punta +
        barrio.La.Roqueta + barrio.La.Seu + barrio.La.Vega.Baixa +
        barrio.La.Xerea + barrio.Malilla + barrio.Marxalenes + barrio.Mestalla +
        barrio.Mont.Olivet + barrio.Morvedre + barrio.Na.Rovella +
        barrio.Nou.Campanar + barrio.Nou.Moles + barrio.Patraix +
        barrio.Penya.Roja + barrio.Playa.de.la.Malvarrosa + barrio.Russafa +
        barrio.Sant.Francesc + barrio.Sant.Llorenç + barrio.Sant.Pau +
        barrio.Trinitat + planta.5 + planta.7 + planta.19 +
        habitaciones.2  + habitaciones.4 + habitaciones.6 +
        planta.0 + planta.9, data = training)
summary(lm_4_op)

check_model(lm_4_op)
check_collinearity(lm_4_op)
plot(lm_4_op)

prediction4<-predict(lm_4_op,testing)
lm4_op_rmse <- RMSE(prediction4, testing$logprecio)

data_mod <- data.frame(Predicted = prediction4, Observed = testing$logprecio, precio_predicho
= 10^(prediction4), precio_real = 10^(testing$logprecio))
rmse_linear <- RMSE(data_mod$precio_predicho, data_mod$precio_real)
```

```
mae_linear <- MAE(data_mod$precio_predicho, data_mod$precio_real)
data_mod1 <- subset(data_mod, precio_real < 1000)
data_mod2 <- subset(data_mod, precio_real < 1500)
data_mod3 <- subset(data_mod, precio_real < 2000)
rmse_linear1 <- RMSE(data_mod1$precio_predicho, data_mod1$precio_real)
rmse_linear2 <- RMSE(data_mod2$precio_predicho, data_mod2$precio_real)
rmse_linear3 <- RMSE(data_mod3$precio_predicho, data_mod3$precio_real)
mae_linear1 <- MAE(data_mod1$precio_predicho, data_mod1$precio_real)
mae_linear2 <- MAE(data_mod2$precio_predicho, data_mod2$precio_real)
mae_linear3 <- MAE(data_mod3$precio_predicho, data_mod3$precio_real)

ggplot(data_mod,
      aes(x = precio_predicho,
          y = precio_real)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 0.95)
########################
#OTROS MODELOS CON OTROS METODOS#
########################
data_r <- anomalos[, c('logprecio','logmetros', 'tipo', 'barrio', 'planta', 'habitaciones')]
data_r <- na.omit(data_r)
set.seed(1234)
data_r12 <- dummyVars(" ~ .", data = data_r)
data_r13 <- data.frame(predict(data_r12, newdata = data_r))
inTraining <- createDataPartition(data_r13$logprecio, p = .75, list = FALSE)
training <- data_r13[inTraining,]
testing <- data_r13[-inTraining,]

fitControl <- trainControl(method = "repeatedcv", search = 'random', number = 5, repeats = 2)
cubist <- train(logprecio~., data = training, method = "cubist", trControl=fitControl, verbose =
FALSE, tuneLength = 10)
summary(cubist)
prediction5<-predict(cubist,testing)
data_mod4 <- data.frame(Predicted = prediction5, Observed = testing$logprecio, pred_price =
10^(prediction5), real_price = 10^(testing$logprecio))
rmse_cubist <- RMSE(data_mod4$pred_price, data_mod4$real_price)
mae_cubist <- mae(data_mod4$pred_price, data_mod4$real_price)

knn <- train(logprecio~., data = training, method = "knn", trControl=fitControl, tuneLength =
10)
summary(knn)
prediction5<-predict(knn,testing)
data_mod4 <- data.frame(Predicted = prediction5, Observed = testing$logprecio, pred_price =
10^(prediction5), real_price = 10^(testing$logprecio))
rmse_knn <- RMSE(data_mod4$pred_price, data_mod4$real_price)
mae_knn <- mae(data_mod4$pred_price, data_mod4$real_price)

xgbLinerar <- train(logprecio~., data = training, method = "xgbLinear", trControl=fitControl,
verbose = FALSE, tuneLength = 10)
summary(xgbLinerar)
prediction5<-predict(xgbLinerar,testing)
data_mod4 <- data.frame(Predicted = prediction5, Observed = testing$logprecio, pred_price =
10^(prediction5), real_price = 10^(testing$logprecio))
rmse_xgbLinerar <- RMSE(data_mod4$pred_price, data_mod4$real_price)
mae_xgbLinerar <- mae(data_mod4$pred_price, data_mod4$real_price)
```

```
rf <- train(logprecio~., data = training, method = "rf", trControl=fitControl, verbose = FALSE,
tuneLength = 10)
summary(rf)
prediction5<-predict(rf,testing)
data_mod4 <- data.frame(Predicted = prediction5, Observed = testing$logprecio, pred_price =
10^(prediction5), real_price = 10^(testing$logprecio))
rmse_rf <- RMSE(data_mod4$pred_price, data_mod4$real_price)
mae_rf <- mae(data_mod4$pred_price, data_mod4$real_price)

data_mod1 <- subset(data_mod4, real_price < 1000)
data_mod2 <- subset(data_mod4, real_price < 1500)
data_mod3 <- subset(data_mod4, real_price < 2000)
rmse_rf <- RMSE(data_mod1$pred_price, data_mod1$real_price)
rmse_rf <- RMSE(data_mod2$pred_price, data_mod2$real_price)
rmse_rf <- RMSE(data_mod3$pred_price, data_mod3$real_price)
mae_rf <- MAE(data_mod1$pred_price, data_mod1$real_price)
mae_rf <- MAE(data_mod2$pred_price, data_mod2$real_price)
mae_rf <- MAE(data_mod3$pred_price, data_mod3$real_price)

ggplot(data_mod4,
      aes(x = pred_price,
          y = real_price)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 0.95)
########################
#MODELOS CLASIFICACION#
########################
datos_sin$f_precio <- cut(datos_sin$precio, breaks = c(370, 1001, 1501, 2001, 8000),
                          labels = c("<1000€", "1000€-1500€", "1500€-2000€", ">2000€"))
                          #labels = c('baratas', 'medio', 'caras', 'muy caras'))

data_r <- datos_sin[, c('f_precio', 'metros', 'tipo', 'distrito', 'planta', 'habitaciones')]
data_r <- na.omit(data_r)
set.seed(1234)
inTraining <- createDataPartition(data_r$f_precio, p = .75, list = FALSE)
training <- data_r[inTraining,]
testing <- data_r[-inTraining,]

fitControl <- trainControl(method = "repeatedcv", search = 'random',number =10)

rf_clas <- train(f_precio~., data = training, method = "rf", verbose = FALSE,
trControl=fitControl)
summary(rf_clas)
prediction5<-predict(rf_clas,testing)
table(prediction5, testing$f_precio)

C5.0_clas <- train(f_precio~., data = training, method = "C5.0", trControl=fitControl, verbose =
FALSE, tuneLength = 10)
summary(C5.0_clas)
prediction5<-predict(C5.0_clas,testing)
table(prediction5, testing$f_precio)

xgbTree_clas <- train(f_precio~., data = training, method = "xgbTree", trControl=fitControl,
verbose = FALSE, tuneLength = 10)
```

```
summary(xgbTree_clas)
prediction5<-predict(xgbTree_clas,testing)
table(prediction5, testing$f_precio)

rpart_clas <- train(f_precio~., data = training, method = "rpart", trControl=fitControl, verbose =
FALSE, tuneLength = 10)
summary(rpart_clas)
prediction5<-predict(rpart_clas,testing)
table(prediction5, testing$f_precio)

fit <- rpart(f_precio~., data = training, method = 'class')
fancyRpartPlot(fit)
```