COMILLAS
UNIVERSIDAD PONTIFICIA
ICADE

Facultad de Ciencias Económicas y Empresariales ICADE

El Big Data en el Sector Inmobiliario

Autor: Jose Luis Moreno Rodríguez

Director: María de las Mercedes Barrachina Fernández

MADRID | JUNIO 2023

Resumen:

El Big Data ha revolucionado el sector inmobiliario en el contexto de la revolución Proptech. En la era de la revolución 4.0, caracterizada por la digitalización y la conectividad, se han generado una gran cantidad de datos en el sector inmobiliario, como transacciones de propiedades, características de las viviendas, precios y ubicaciones geográficas. El Big Data permite aprovechar esta abundancia de datos para obtener información valiosa y tomar decisiones informadas. Además, la revolución Proptech ha impulsado aún más el uso del Big Data en el sector inmobiliario al combinar tecnología e innovación en soluciones como plataformas de búsqueda y alquiler de viviendas, análisis de precios en tiempo real y servicios personalizados para compradores y vendedores. En conjunto, el Big Data en el sector inmobiliario proporciona ventajas significativas al analizar tendencias, predecir precios y demanda, optimizar la gestión de propiedades y mejorar la experiencia del cliente.

Abstract:

Big Data has revolutionized the real estate sector in the context of the Proptech revolution. In the era of the 4.0 revolution, characterized by digitization and connectivity, a wealth of data has been generated in the real estate sector, such as property transactions, housing characteristics, prices and geographic locations. Big Data makes it possible to leverage this abundance of data to gain valuable insights and make informed decisions. In addition, the Proptech revolution has further boosted the use of Big Data in real estate by combining technology and innovation in solutions such as home search and rental platforms, real-time price analysis, and personalized services for buyers and sellers. Overall, Big Data in real estate provides significant advantages by analyzing trends, predicting prices and demand, optimizing property management and improving the customer experience.

Palabras clave:

Big data, machine learning, smart cities, Internet o Things, Proptech, real estate, avance tecnológico.

El Big Data en las Inversiones Inmobiliarias.

ÍNDICE

Índice de figuras

Índice de tablas

1. Introducción

Vivimos en un mundo que se encuentra en constante cambio y donde la innovación y adaptación al medio juegan un papel fundamental para la supervivencia de las empresas. La tecnología avanza cada vez más rápido, los inventos se suceden y las empresas buscan como introducirlos en su modelo de negocio para conseguir ser más competitivas y eficientes. El sector inmobiliario, a pesar de su marcado carácter tradicional, no iba a ser de otro modo. La revolución que está cambiando en estos últimos años uno de los sectores más antiguos del mundo como es el inmobiliario se le conoce como PropTech y se inspira principalmente en las técnicas de análisis inteligente de datos y *machine learning*. La innovación y disrupción tecnológica obligan a las empresas a adaptarse a los cambios si no quieren quedar obsoletas. El sector inmobiliario está directamente vinculado con la forma de vida de las personas por lo que afecta a todos los individuos.

1.1 Metodología

La metodología seguida en este trabajo se ha estructurado en dos pasos principales. En primer lugar, se ha llevado a cabo una exhaustiva revisión de la literatura, comenzando por la exploración de conceptos generales como la revolución proptech y el impacto del big data en diversos sectores. Posteriormente, se ha centrado en la aplicación específica del big data en el sector inmobiliario, analizando estudios previos, frameworks y casos de éxito que demuestran su relevancia y beneficios en este ámbito.

En segundo lugar, se ha realizado un caso práctico utilizando una base de datos proporcionada por Kaggle. Se ha llevado a cabo un análisis exploratorio de los datos, examinando variables claves relacionadas con el sector inmobiliario, como el precio, la ubicación, las características de las propiedades, entre otros. Este análisis ha permitido obtener información valiosa sobre patrones, tendencias y relaciones entre las variables.

Además, se ha desarrollado un modelo supervisado, utilizando técnicas de aprendizaje automático, para predecir los precios de las viviendas en función de las características proporcionadas. También se ha implementado un modelo no supervisado, específicamente un algoritmo de clustering, para identificar grupos o segmentos de propiedades con características similares.

2. Introducción: La Revolución PropTech.

Vivimos en un mundo que se encuentra en constante cambio y donde la innovación y adaptación al medio juegan un papel fundamental para la supervivencia de las empresas. La tecnología avanza cada vez más rápido, los inventos se suceden y las empresas buscan como introducirlos en su modelo de negocio para conseguir ser más competitivas y eficientes. El sector inmobiliario, a pesar de su marcado carácter tradicional, no iba a ser de otro modo. A la revolución que está cambiando en estos últimos años uno de los sectores más antiguos del mundo como es el inmobiliario se le conoce como PropTech. El sector inmobiliario está directamente vinculado con la forma de vida de las personas por lo que afecta a todos los individuos.

2.1 Definición

El término PropTech surge de la combinación de "propiedad" y "tecnología" y se refiere a la transformación digital que ha sufrido el sector inmobiliario en los últimos años. Esta revolución surge a raíz del gran avance tecnológico y la creciente demanda por parte de los consumidores de un mercado que ofrezca soluciones más eficientes y accesibles. Podríamos definirlo de la siguiente manera: "Conjunto de tecnologías innovadoras destinadas a cubrir necesidades dentro del ámbito del sector inmobiliario mediante la optimización, la mejora o la reinvención de cualquier servicio relacionado con el sector" (Hernández, Puigdevall y López, 2021, p. 216).

La Revolución PropTech ha transformado radicalmente la forma en que se busca y se accede a la información sobre propiedades. Recordemos que, antiguamente, para el proceso de compraventa de un inmueble, los usuarios debían acudir a anuncios o contactos para encontrar una vivienda. Después, se contactaba con el vendedor y finalmente se cerraba la operación. Comparar inmuebles era casi imposible, llevaba mucho tiempo y solo podías ceñirte a un pequeño grupo de ellos, los que se publicitaban en los periódicos o eran de personas conocidas. El proceso de utilización de plataformas online y aplicaciones móviles ha permitido a los usuarios acceder a una amplia gama de inmuebles, a parte de la posibilidad de analizar un gran número de datos sobre propiedades, desde detalles de precios y características, hasta imágenes y recorridos virtuales. Esto ha agilizado el proceso de búsqueda de propiedades, brindando a los

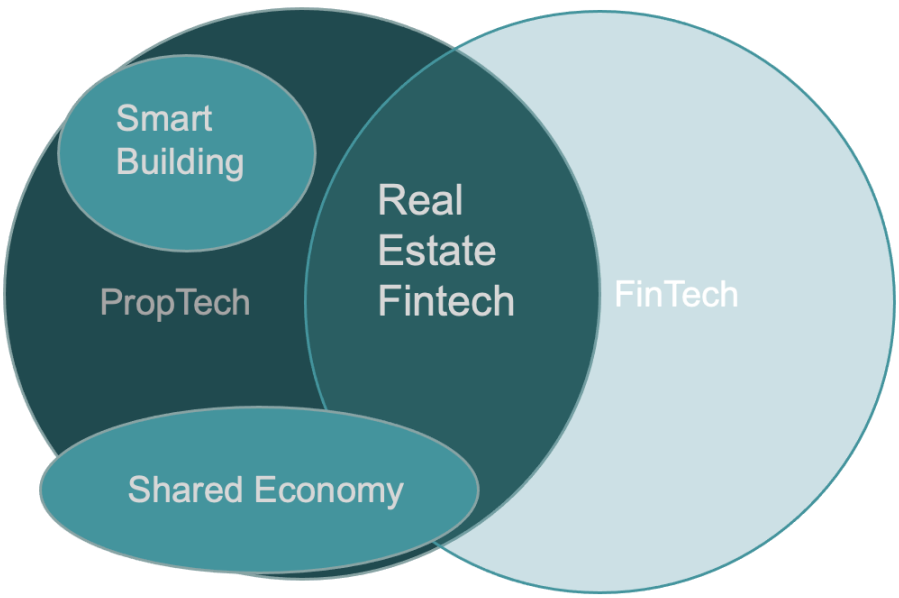
compradores una mayor transparencia y facilitando la toma de decisiones informadas (Hernández, Puigdevall, López, 2021, p. 218-219).

En este sentido, podemos catalogar a las empresas responsables de este cambio como disruptivas. El término de disrupción hace referencia a los modelos de negocio que cambian la forma en la que vivimos, que rompen con lo establecido, por ejemplo, Airbnb. La disrupción digital presenta principalmente dos características: la velocidad con la que se producen los datos y la actitud frente a la asunción de riesgos. (Capiello, 2020) Las empresas que no se adapten al cambio corren el riesgo de quedar obsoletas y, por consecuencia, tener que abandonar el mercado. (Ma, 2021)

2.2 Clasificación y contenido de las PropTech:

Como hemos visto, el término PropTech es un término muy amplio. Las raíces de este cambio, según muchos expertos, se encuentran en la motivación para la innovación y el avance de las Fintech, Smart Building Technologies y Shared Economy (Baum, 2020, p.

Figura 1. Raíces del Proptech



(Elaboración propia a partir de Baum, 2020)

Se va a realizar una clasificación de las PropTech según el servicio que ofrecen. Primero, se diseña y construye un bien inmueble, posteriormente, se valora y comercializa y, una vez se ha conseguido vender, se llevan a cabo postransacciones. Así, podríamos

clasificar las empresas PropTech de la siguiente manera: (Hernández, Puigdevall, López, p. 221)

1. Construcción:
    1.1 Contech: Tecnología innovadora existente en todas las fases de un proceso de una construcción, desde el diseño del inmueble y la elaboración de planos hasta la edificación e instalación de componentes especiales. (Rousseau, 2021)
    1.2 Smart homes: Construcción de casas inteligentes, ocupadas con nueva tecnología que facilita el día a día de las personas.
2. Preparación del bien:
    2.1 Imagen: Empresas que mejoran la experiencia visual del inmueble por parte del cliente.
    2.2 Big data: Aportan softwares mediante los cuales los profesionales pueden ver datos sobre inmuebles en tiempo real. Como veremos más adelante, sus principales usos son valoración de inmuebles, generación de informes y análisis de tendencias. Destacan las empresas de valoración de bienes a partir de datos y Machine Learning. Realmente y como veremos en el apartado "El big data en el sector inmobiliario", el big data puede ser utilizado en todas las fases de manera efectiva, obteniendo resultados óptimos a partir de su utilización.
3. Comercialización del bien:
    3.1 Marketplaces: Plataformas digitales donde los ofertantes publicitan los activos a modo de escaparate para los consumidores. Principales actores en el proceso de compraventa. Empresas como Idealista o Fotocasa son Marketplaces.
    3.2 Softwares CRM: Permiten a las empresas estar en contacto continuo con los clientes, mejorando la rentabilidad, optimizando los recursos, aumentando la satisfacción del cliente y potenciando el crecimiento del negocio. Para ello, dentro del software se integran ventas, marketing, atención al cliente y puntos de contacto o *touchpoints* (Pons, 2018).

4. Postransacción:
    4.1 Property Management: Softwares que permiten la gestión de activos inmobiliarios una vez realizada la transacción para su mantenimiento y puesta en rentabilidad.
    4.2 Internet de las cosas: Se refiere a la incorporación de dispositivos y sensores conectados a la red en los edificios y propiedades, con el objetivo de recopilar datos en tiempo real, automatizar procesos y mejorar la eficiencia operativa.

Como vemos y se apreciará posteriormente, la transición y el ordenamiento de datos son un elemento clave en este proceso de disrupción. En 2006, Clive Humby pronunció la siguiente frase "Los datos en el nuevo petróleo". Para esa época, esta pronunciación parecía descabellada, pero el tiempo le ha dado la razón. Existe una gran diferencia entre los datos y el petróleo, y es que, el petróleo es escaso, pero los datos están en todas partes. Michael Palmer completó dicha frase terminándola de la siguiente manera "Es cierto que los datos son el nuevo petróleo, pero hay que refinarlos". Parece que el sector inmobiliario, como hemos dicho, de carácter tradicional, ha aprendido a refinar estos datos y a sacar un partido inmenso de ellos en su actuación para la mejora de procesos en las fases de implementación de un inmueble.

2.3 Crecimiento, inversores y expectativas.

Podemos tomar como origen del mercado PropTech el año 2000, con la aparición de portales inmobiliarios como Idealista o Rightmove. Pronto les seguiría las empresas *peer to peer* mundialmente conocidas actualmente como Airbnb o Homeaway. Éstas no tardaron en ganar cuota de mercado y romper con los cánones básicos del tradicional sector inmobiliario. En los últimos años, el crecimiento del proptech está siendo exponencial. En la actualidad, estas empresas están marcando la evolución del sector inmobiliario, obligando a las empresas más tradicionales a tomar decisiones estratégicas sobre innovación y desarrollo si quieren permanecer en el mercado (Hernández, Puigdevall, López, p. 216).

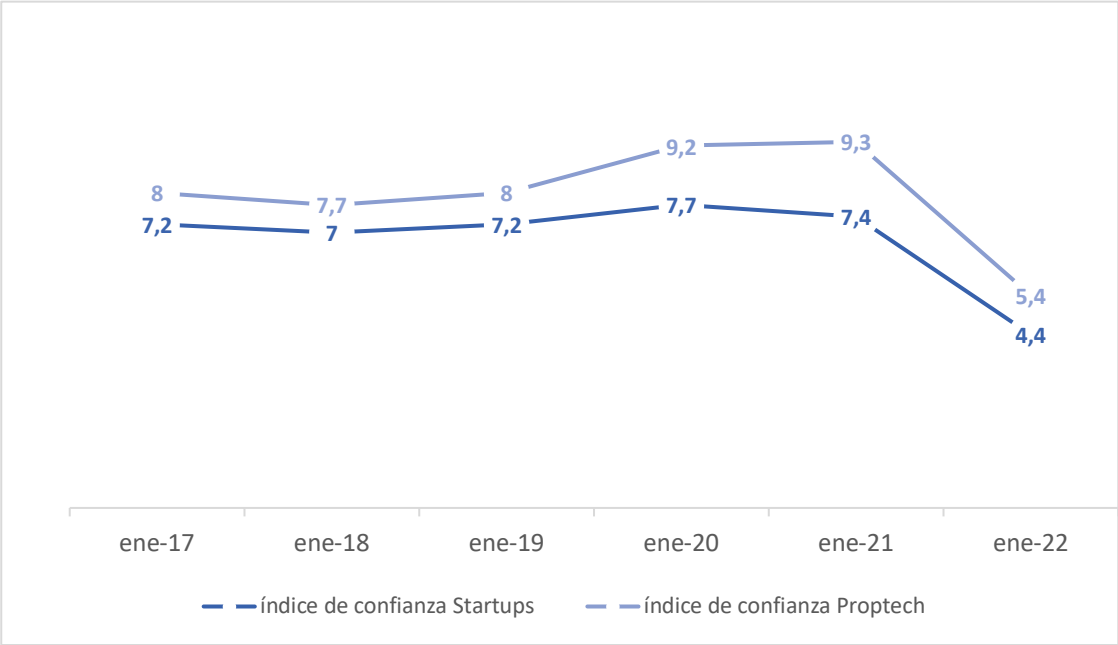
Los inversores se han sentido atraídos por las oportunidades de disrupción y mejora que ofrecen las empresas Proptech. Grandes fondos de inversión, empresas de

capital de riesgo y corporaciones han canalizado capital hacia startups Proptech para impulsar la innovación y promover la adopción de soluciones tecnológicas en el mercado inmobiliario. Uno de los principales atractivos del mercado Proptech para los inversores es su potencial de crecimiento y rentabilidad. Según el informe "Emerging Trends in Real Estate 2021" de PwC y el Instituto Urban Land, las inversiones en Proptech han generado rendimientos significativos en comparación con otros sectores. Además, el informe destaca que se espera un aumento continuo en la inversión en tecnología inmobiliaria en los próximos años.

A finales de 2022, PwC elaboró un estudio llamado “Global PropTech Confidence Index”, el cual trataba sobre la confianza que los inversores y las propias empresas y startups seguían manteniendo sobre el sector. Los resultados o *keypoints* obtenidos fueron los siguientes:

- El 71% de los inversores esperan realizar más o el mismo número de inversiones PropTech, frente al 88% de mediados de 2022.
- El 74% de los inversores esperan ver más fusiones y adquisiciones en los próximos 12 meses.
- El 33% de los inversores declararon que las empresas de su cartera estaban rindiendo por debajo de las expectativas en términos de crecimiento de clientes, un aumento del 14% a mediados de 2022.
- El 57% de los fundadores de startups creen que será más difícil conseguir capital en los próximos 12 meses, por debajo del 71% de hace seis meses, pero por encima del 20% de finales de año de 2021.
- El 54% de los startups han declarado que, si no consiguen capital adicional, tienen menos de 12 meses de margen.

Figura 2. Índice de confianza en Proptech



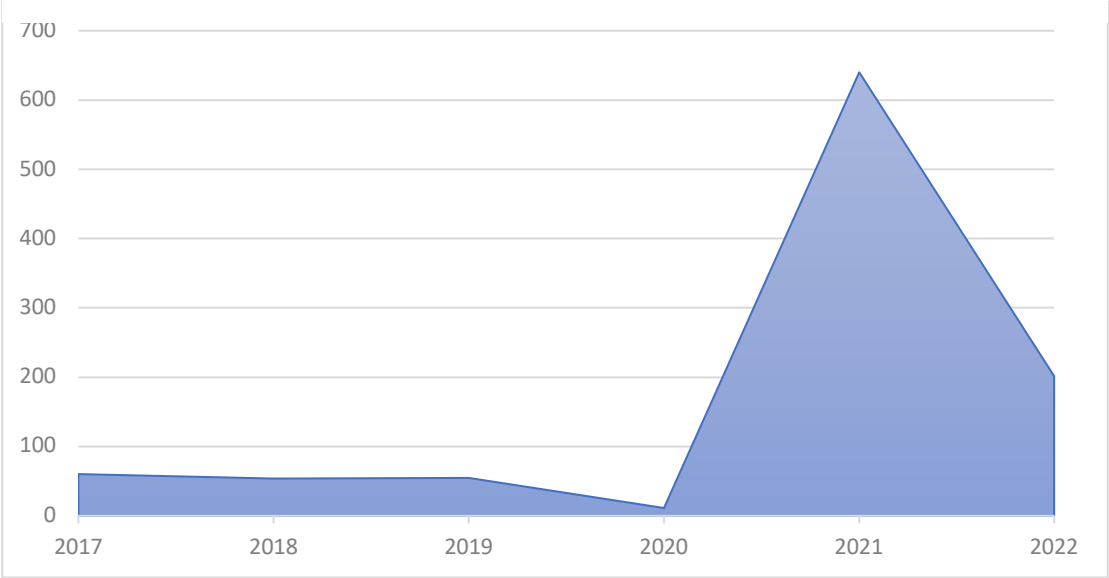
(Elaboración propia a partir de estudio de PwC)

Como vemos en la **figura 2**, el índice de confianza en proptech ha bajado significativamente, aunque se encuentra por encima del índice de confianza en los startups. La pérdida de confianza de los inversores en estos últimos dos años no es sectorial, sino más bien por motivos macroeconómicos. La incertidumbre macroeconómica, pronunciada por los inversores, se basa en la desestabilización de los tipos, indicadores de recesión y los niveles de inflación que comienza a moderarse ligeramente. Siguiendo con el gráfico, vemos como los inversores prefieren invertir en general en empresas proptech que en startups dedicadas a otro sector, esto se debe a que invirtiendo en proptech estás apostando por innovación y futuro en un mercado estable y duradero en el tiempo.

Cabe destacar el significativo impacto que tuvo el Covid en la revolución y disrupción de este sector. Si atendemos a la **figura 2**, la confianza en proptech se incrementa notablemente a raíz de la pandemia. La crisis puso de manifiesto la urgente necesidad de acelerar el proceso de digitalización de las empresas, lo que dio lugar al crecimiento de la economía digital. El uso de nuevas tecnologías digitales logró el objetivo de mantener la operatividad de las empresas a pesar de la orden de "trabajar desde casa". En el sector inmobiliario no iba a ser diferente, la apuesta por parte de los

inversores hacia este sector gracias a la pandemia creció con mucha fuerza en España, llegando a obtener las firmas españolas un total de 640 millones de euros en financiación en el ejercicio de 2021 (CBRE, 2021).

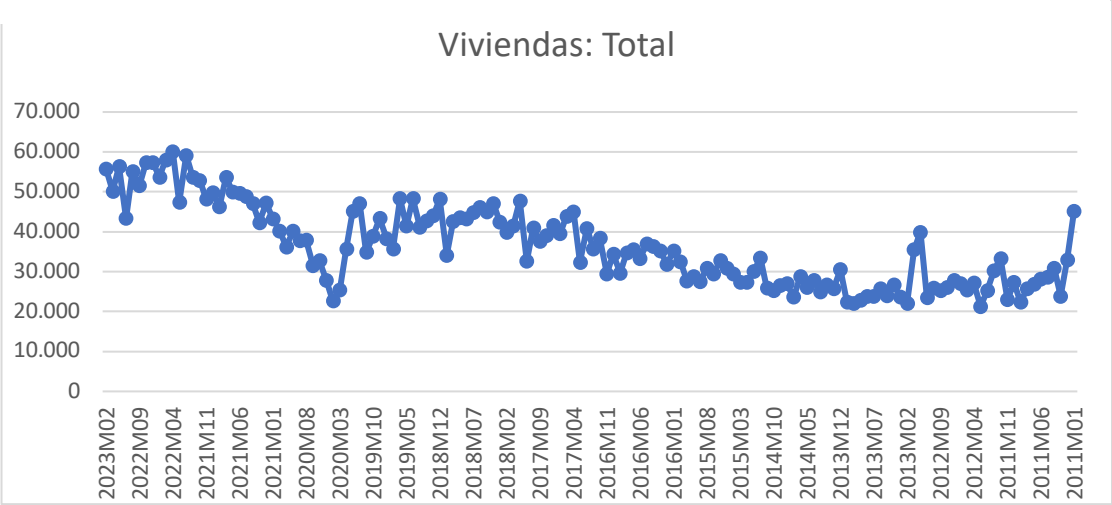
Figura 3. Inversión en empresas Proptech españolas (Mill de euros)



(Elaboración propia a partir de estudio de CBRE, 2023).

La compraventa de casas también creció considerablemente en España **(Ver figura 4),** situándose en máximos desde 2007, año del boom inmobiliario y anterior a la gran crisis financiera. En términos globales, el comercio de real estate alcanzó la cifra de US$1.3 trillones. El sector multifamiliar atrajo mucho más capital en Estados Unidos y Europa, mientras que la inversión industrial se mantuvo fuerte en todas las regiones. El sector minorista y hotelero mostraron signos de recuperación, aunque el sector de oficinas continua en decadencia (Barkham, Mellott, Raaum, 2022).

Figura 4. Compra de viviendas en España



(Elaboración propia a partir de los datos del INE (Instituto Nacional de Estadística)))

En cuanto a la situación geográfica, la mayoría de las empresas PropTech se encuentran en China, América del Norte y Europa, existiendo *unicorns companies* solo en China y Estados Unidos (Baum, 2020. P. 20). La financiación de estas empresas es mayor en China y Estados Unidos, donde la cifra de aportaciones de capital por parte de los inversores locales es mayor que en Europa. En Europa destacan, Alemania y Francia. En España, la tecnología más utilizada en 2022 ha sido el Big Data, alcanzando el 37%. Sin embargo, en el resto de los países europeos, la tecnología más usada ha sido la Inteligencia Artificial, el big data en estos países representa un 31% (Barkham, Mellott, Raaum, 2022).

La revolución proptech ha reducido los tiempos de transacción en la compraventa de viviendas. A parte, el avance en los servicios de imagen permite la perfecta visualización del inmueble, pudiendo acceder a su compra desde cualquier parte del mundo, lo que incentiva su intercambio y la inversión extranjera. Por otro lado, la digitalización y procesamiento de datos permite analizar terrenos y viviendas de todo el mundo, consiguiendo analizar tendencias en distintas zonas geográficas y permitiendo la búsqueda de oportunidades de inversión más allá de las fronteras. Todo esto fomenta la inversión en activos inmobiliarios (Siniak, et al, 2020).

3. Big Data

2.1 Definición

El término "big data" ha ganado terreno en los últimos años en un amplio abanico de disciplinas, como la empresa, la informática, los estudios de la información, los sistemas de información, la estadística y muchas más. A medida que la tecnología se desarrolla, producimos constantemente un volumen cada vez mayor de datos. Así, el autor Kenneth Cukier, define el big data de manera sencilla de la siguiente manera: “se trata de hacer cosas a partir del análisis de inmensas cantidades de información, que simplemente no son posibles con volúmenes más pequeños”.

Un estudio de text mining a partir del big data analizó las palabras claves más frecuentes incluidas en los resúmenes y definiciones relacionados con Big Data, teniendo en cuenta también sus relaciones mutuas. Las cuatro palabras que más aparecían en estos textos eran las siguientes: Información, método, tecnología e impacto (De mauro, Greco y Grimaldi, 2015). La mayoría de los expertos que se han posicionado y elaborado artículos sobre el tema relacionan el término con, al menos, una de las cuatro palabras claves. La gasolina del big data es la información. Una de las razones por la cual el big data se ha convertido en un fenómeno a nivel mundial, ha sido por el grado en que la **información** puede ser recopilada y almacenada. Mediante la digitalización, que es el proceso de convertir información continua y analógica en formato discreto, digital y legible por los humanos o por una máquina, la información es almacenada correctamente. Los **avances tecnológicos**, como los centros de datos, más conocidos como “data centers” permiten almacenar y procesar esa información. No tienen por qué ser megaproyectos, la ampliación en la capacidad de los procesadores de información como discos duros o la mejora en los procesadores de los ordenadores también forman parte de la revolución big data. A través de los “**métodos**” transformamos el big data en valor. Manyika y Chen proponen, por separado, aunque de manera similar, una lista de métodos analíticos de Big Data usados por las empresas para transformar la información en valor. En esta lista se incluyen: aprendizaje de reglas de asociación, fusión e integración de datos, algoritmos genéticos, aprendizaje automático, procesamiento del lenguaje natural, procesamiento neuronal, análisis de redes, … La última palabra en la que los expertos hacen énfasis es

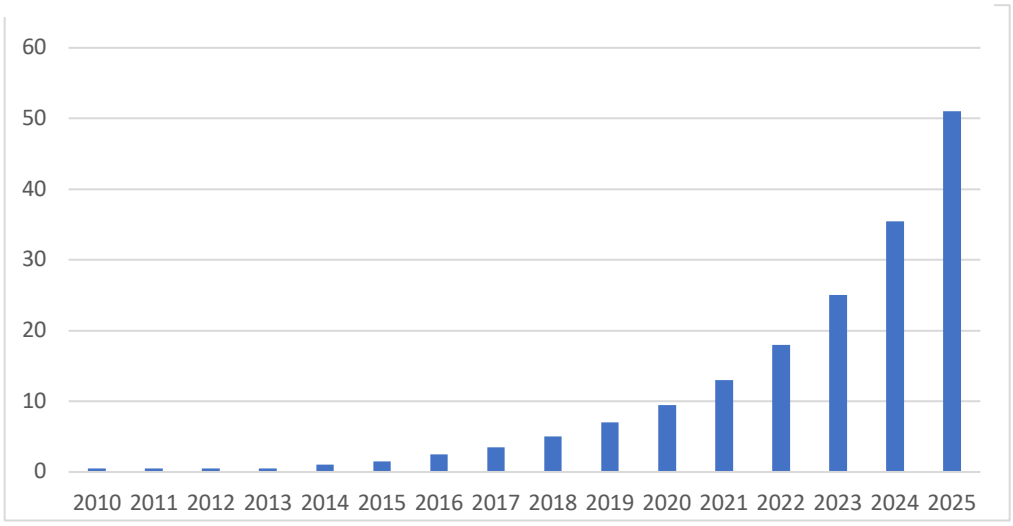
impacto, pues los macrodatos están cambiando la forma en que vivimos y, sobre todo, la forma en la que las organizaciones compiten (De Mauro, Greco y Grimaldi, 2015).

2.2 Las 3 V´s o drivers del big data.

Tradicionalmente, el big data ha estado relacionado con las 3 V´s, éstas son: Variabilidad, volumen y velocidad (Maté, 2014).

- **El volumen** en el Big Data se refiere a la enorme cantidad de datos que se generan y acumulan constantemente. Los conjuntos de datos masivos exceden las capacidades de los sistemas tradicionales de almacenamiento y procesamiento. El volumen del Big Data plantea desafíos y oportunidades para las organizaciones en la gestión y análisis de grandes volúmenes de datos. (Mayer Schonberger, Cukier, 2013). El volumen de datos no parará de crecer a lo largo del tiempo, pues cada vez son más. Los rastreadores que las empresas elaboran para hacer un seguimiento de las personas permiten recopilar un gran número de información acerca del usuario. Los datos se multiplican con el paso del tiempo, por lo que, si comparamos el volumen de datos de hoy con respecto a hace cinco años, se han incrementado en exceso, pero los expertos prevén que en los siguientes años seguirán creciendo de forma exponencial.

Figura 5. Evolución del volumen de datos

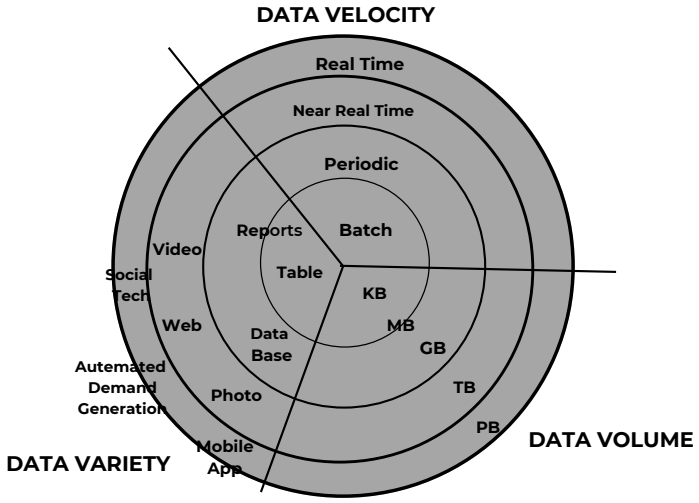


(Elaboración propia a partir de Statista)

- **La velocidad** es una de las características clave del Big Data, se refiere a la rapidez con la que se generan, procesan y analizan los datos en tiempo real. El crecimiento exponencial de los dispositivos conectados y las tecnologías de sensores ha acelerado la velocidad de generación de datos del Big Data. La velocidad de generación de datos ha aumentado de manera significativa, lo que plantea desafíos y oportunidades para las organizaciones en la gestión y aprovechamiento de los datos a alta velocidad. Muchas veces, la velocidad de generación de datos supera al volumen en importancia, ya que permite generar una respuesta rápida para satisfacer las necesidades del consumidor (McAfee, Brynjolfsson, 2012).

- **La Variedad** en el Big Data se refiere a la diversidad de tipos y formatos de datos presentes en grandes conjuntos de información. Los datos pueden presentarse en diferentes formatos y estructuras, como datos estructurados (database), no estructurados (imagen, vídeos, textos, …) y semiestructurados (mezcla de ambas). Los datos estructurados han sido creados mediante unas reglas o patrones establecidos con anterioridad, mientras que los no estructurados no siguen una estructura fija. La variedad de datos en el Big Data también ofrece oportunidades para descubrir patrones y tendencias ocultas que pueden proporcionar una ventaja competitiva a las organizaciones. Al combinar y analizar datos de diversas fuentes y formatos, se pueden obtener conocimientos más profundos y completos (Wadhwani, et al, 2017).

Figura 6. Las 3 primeras V´s del big data



(Elaboración propia basado en Maté, 2014)

2.3 Big data analytics:

Sin entrar primeramente en definiciones complejas, podríamos definir el concepto de big data analytics como la aplicación práctica de los datos obtenidos mediante técnicas de captación de información. El big data analytics se encarga de transformar los datos masivos en conocimiento para poder usarlos e incrementar la eficiencia en las organizaciones. Una organización que realmente no entiende sus datos extraídos no conseguirá crear valor a partir de ellos, pues no sabrá lo que está buscando exactamente. Muchas empresas, como Netflix, han basado su ventaja competitiva en la utilización de los datos. Detrás de las recomendaciones de las compañías hay un algoritmo que trabaja en tiempo real y que envía al consumidor un producto según sus preferencias, acertando en un gran número de veces. Utilizar los datos para mejorar la experiencia del consumidor es un reto por afrontar en todas las empresas, pues como hemos dicho anteriormente, los datos son el nuevo petróleo, pero hay que saber usarlos. De una manera más compleja, el análisis de datos es el proceso de aplicar algoritmos para analizar conjuntos de datos y extraer patrones, relaciones e información útil y desconocida. Esto permite a las organizaciones comprender el significado y la importancia de los datos y utilizarlos para la toma de decisiones (Elgendy, Elragal, 2014). En este sentido, se ve como el volumen de datos no adquiere importancia sin una posterior estructuración y manejabilidad.

El análisis de big data se ha ampliado para incluir técnicas avanzadas como reglas de asociación, agrupamiento, clasificación y árboles de decisión, así como análisis de redes sociales y análisis de textos o text minning. El análisis de las redes sociales se basa en el desarrollo y la evaluación de marcos y herramientas informáticas para recopilar, monitorear, resumir, analizar y visualizar datos de las redes sociales. Por otro lado, el análisis de la red social se centra en las relaciones entre entidades sociales y sus implicaciones. (Batrinca, Treleaven, 2015).

Las compañías crean datos a partir solo de existir, es intrínseco a ellas, por lo que en el volumen de datos no va a estar la ventaja competitiva que la empresa quiere adquirir. Las empresas toman ventaja a partir del manejo y utilización de la información. En este sentido, vemos como la transición de los datos puede ser la siguiente: del big data

(recopilación de datos masivos) a la inteligencia de negocios (datos que realmente aportan valor a la compañía.).

Figura 7. Del big data al big data analytics

(*Elaboración propia)*

Se refiere a las 3 V´s cuando se habla de términos vinculados tradicionalmente con el big data. Sin embargo, debido al auge en conocimiento y uso del big data las empresas han ido implementando otras V´s en sus proyectos (Tabares, Hernández, 2014). Por ejemplo, según Normandeau, se deben implementar 3 nuevas Vs: Veracidad, volatilidad y validez.

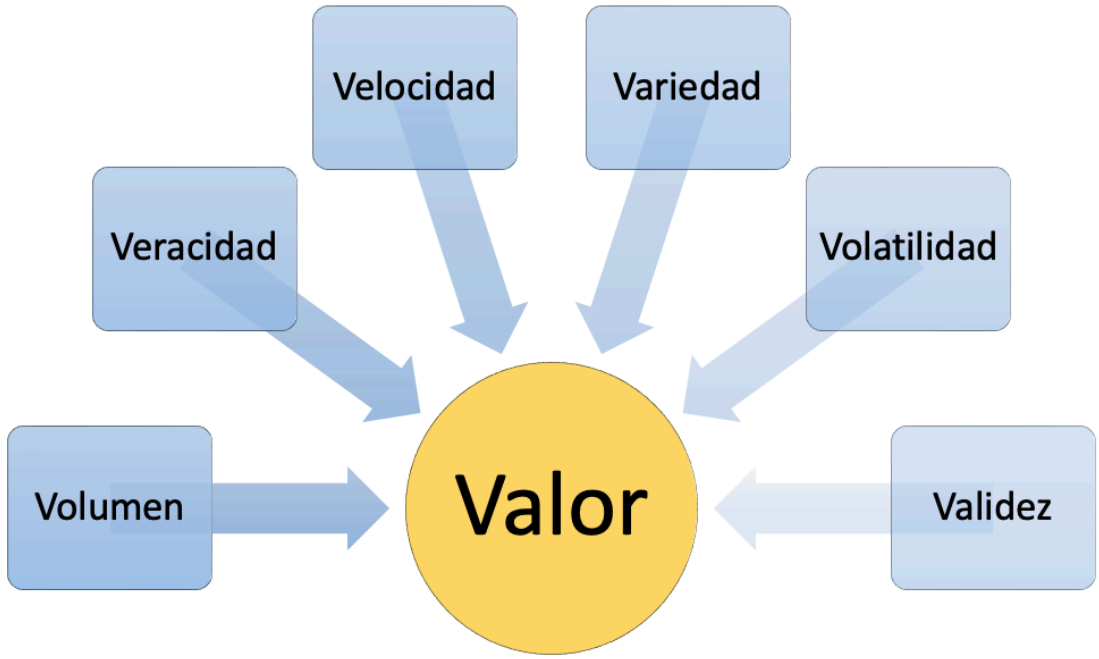
Por **veracidad** se refiere a la fiabilidad y confiabilidad de los datos. Debido al gran volumen de datos percibidos por las organizaciones, estos se presentan en formas distintas y a veces de forma complicada. Esto puede llevar al desaprovechamiento de los datos o a una disminución en su calidad (Urrea, 2022).

La **volatilidad** de los datos es, principalmente, el tiempo que los datos pueden estar almacenados y durante cuánto tiempo estos son válidos, es decir, si se deterioran con el paso del tiempo (Tabares, Hernández, 2014). Cuanto mayor sean las 3 V´s del big data, más tiempo serán útiles dichos datos.

La **validez** de los datos no debe confundirse con la veracidad de estos. A diferencia de la veracidad, la validez de los datos se refiere a la legalidad y transparencia de los datos fundamentalmente. Es necesario conocer la disposición legal acerca de los datos, así como su uso para procesamiento y retención de ellos (Moro, Visconti, 2020).

Todas las V´s deben actuar conjuntamente para crear **valor.** El valor se obtiene al final de la cadena, a partir del procesamiento de los datos analíticos. Esta V se califica como la más importante y permite a la empresa obtener la ventaja competitiva mediante la creación de valor a partir de la utilización de datos. En big data analytics, los datos estructurados y no estructurados se usan para adquirir mayor conocimiento de mercado y así mejorar la toma de decisiones empresariales. Para adquirir valor en el proceso se deben tener en cuenta las demás V´s.: "La naturaleza y procedencia de Big Data hace que los datos contengan ruido, el cual no garantiza la obtención de valor en dichos datos. La eliminación de este ruido proporciona Veracidad y, por ende, se facilita la Validez de estos, de acuerdo con el propósito específico para el que son utilizados. Al mismo tiempo, cuando se genera valor es necesario conocer la Volatilidad de los datos para efectos de conocer los diferentes métodos de análisis a aplicar" (Tabares y Hernández, 2014).

Figura 8. Creación de valor



(Elaboración propia a partir de Tabares y Hernández, 2014)

2.3.1 Aplicaciones del Big Data Analytics.

El análisis masivo de datos se ha convertido en un elemento fundamental en las empresas debido a su gran capacidad para: **predecir** comportamientos y necesidades futuras, **explica**r patrones que se consideran ocultos, y **simular** posibles distintos escenarios futuros. La inteligencia de negocios ha demostrado la gran versatilidad y

adaptabilidad a los diferentes sectores. Hoy en día, esta tecnología se usa para casi todos los sectores o ámbitos de la economía. A continuación, se describen algunos de ellos:

- **Data mining:** Es un proceso por el cual se intenta, a partir del análisis de un gran volumen de datos, encontrar patrones de comportamiento. Permiten llegar a una conclusión sobre qué mezcla de datos es el resultado ideal, de donde obtendríamos una ventaja. Se suele llegar a la conclusión a través de árboles de decisión o redes neuronales (Oluwunmi et al, 2019).
- **Marketing:** El uso del big data en este sector se basa en la predicción y análisis del comportamiento del consumidor. Este permite a las empresas recopilar, analizar y comprender grandes cantidades de datos sobre el comportamiento del consumidor, ayudando a predecir patrones de compra, preferencias y tendencias futuras (Lies, 2019). Asimismo, proporciona información valiosa para tomar decisiones estratégicas en el ámbito del comercio electrónico.
- **Medicina:** El big data analytics tiene un papel crucial en la mejora de la atención médica y la salud pública. Mediante el análisis de grandes conjuntos de datos de registros médicos electrónicos, datos de sensores y otros datos de salud, es posible identificar patrones, predecir enfermedades y mejorar los resultados clínicos (Sánchez, Verspoor, 2014). Además, el uso del big data en el sector de la salud puede mejorar la toma de decisiones clínicas y optimizar los recursos sanitarios. (Mayer-Schönberger, V, Ingelsson,)
- **Cadena de valor:** Permite la optimización de la cadena de suministros. El big data analytics puede ayudar a mejorar la eficiencia y la gestión de la cadena de suministro al analizar grandes volúmenes de datos sobre inventario, demanda, proveedores y transporte. El análisis de big data ayuda a la detección de anomalías, la optimización de rutas de envío y la gestión de la demanda, lo que conduce a una cadena de suministro más ágil y rentable. (Sanders, p.30)
- **Finanzas:** El big data analytics es ampliamente utilizado en el sector financiero para el análisis de datos de mercado, predicción de tendencias, detección de fraudes y gestión de riesgos. También permite a las instituciones financieras tomar decisiones más informadas y mejorar la eficiencia operativa.

Uno de los principales usos se basa en el [1]*credit scoring,* evitando gracias a esta herramienta el *default* en las deudas. (Goldstein et al, 2021)

- **Sector inmobiliario:** Debido a la importancia de este sector en la elaboración del trabajo, aunque se profundizará con un mayor énfasis en los siguientes apartados, es de gran importancia destacar por adelantado la implementación de esta tecnología en el sector para la valoración de inmuebles, gestión de propiedades y reducción de costes.

Estos son algunos de los principales usos del big data, aunque se puede afirmar que el big data es aplicado, o se aplicará, a la mayoría de sectores existentes.

[1] Sistema de calificación de créditos que intenta automatizar la concesión o no de préstamos.

4. Big data en el sector inmobiliario:

El sector inmobiliario es uno de los más antiguos del mundo. Se puede definir como el conjunto de todas las transacciones de compraventa de inmuebles de una determinada zona geográfica. Igualmente se incluyen en él los estudios sobre la evolución del sector y la predicción sobre su actividad, precios futuros, tamaño de mercado, y cualquier variable que pueda intervenir en su evolución (Hernández, Puigdevall, López, p. 180). Es uno de los mercados más grandes del mundo, interviniendo en él agentes con funcionalidades distintas (Ver siguiente tabla). Como vemos, el ecosistema del sector inmobiliario cuenta con múltiples actores, desde la creación del proyecto, hasta su puesta a disposición del consumidor y su gestión posterior (Hernández, Puigdevall, López, p. 184). Para el entendimiento del sector inmobiliario, también es importante recalcar la vinculación entre el sector inmobiliario y la vida emocional de las personas: "La vivienda no es solo un bien inmobiliario, es también una forma de consolidación espiritual" (Mario Benedetti).

Tabla 1. Agentes del sector inmobiliario

Generadores de producto	**Prestadores de servicio**	**Consumidor final**
Arquitectos, promotores y constructores	Agencias inmobiliarias, servicios de gestión, consultoras.	Pequeños tenedores: Particulares, Pymes, autónomos
Viviendas, edificios, oficinas, …	Servicios de financiación, asesoramiento legal y fiscal	Grandes tenedores: Fondos de inversión, socimis, family offices.

(Elaboración propia a partir de Hernández, Puigdevall, López)

El uso del big data en el sector inmobiliario es cada vez más usual, pues permite a las empresas trabajar de una manera más eficiente. Esta enorme cantidad de datos, con su variedad y complejidad, aporta un nuevo modelo de ingresos y un amplio espacio para el desarrollo. En las aplicaciones actuales de los macrodatos, las empresas inmobiliarias, incluyendo las promotoras, las agencias y las empresas de gestión de la propiedad, lo utilizan constantemente para alcanzar los objetivos del ámbito empresarial (Danyang et al, 2014).

4.1 Aplicaciones del big data en el sector inmobiliario

El big data y big data analytics se usan en el sector inmobiliario, para lo siguiente:

- Conocer las **preferencias de los usuarios**: A través de información recopilada de plataformas que usan los consumidores, por ejemplo, Airbnb, Zillow e Idealista, podemos obtener información sobre sus gustos y preferencias. La disponibilidad de este tipo de información permite a las empresas basar sus construcciones en lo que de verdad demandan los consumidores y, conforme a ello, elaborar los proyectos basándose en lo que el consumidor realmente quiere. Además, permite segmentar la población en cuanto a los gustos según su geografía (Oluwunmi et al, 2019).
- **Marketing:**
- **Urbanismo:** El uso de los macrodatos en el urbanismo se centra en el aprovechamiento, planificación y diseño del suelo urbano. Por ejemplo, los macrodatos se han utilizado para analizar las relaciones entre distintos tipos de patrones de crecimiento urbano y la vitalidad de las ciudades. Por ejemplo, utilizando datos de redes sociales y [2]POI de 363 ciudades de China, He et al. (2018) descubrieron que los diferentes patrones de expansión urbana tenían diferentes impactos en la vitalidad urbana, y la expansión de los bordes era la más efectiva para atraer residentes. También, los macrodatos han sido utilizados eficazmente para establecer los límites naturales de las ciudades e identificar los centros urbanos. En esta línea, Jiang y Miao (2015), utilizaron datos extraídos de redes sociales para desglosar los límites administrativos originales y establecer los límites naturales de Chicago, Nueva York y San Francisco (Kong, et al, 2020)**.**
- **Construcción:** El procesamiento de datos en la construcción se usa con muchos fines. Éste permite, entre otras cosas: valorar la predicción de fallos y elaborar un análisis de los residuos generados por la construcción. Actualmente unos de sus principales usos son: mejorar el diseño de edificios (evitando fallos comunes anteriores), supervisar eficazmente el rendimiento del inmueble , el tiempo de respuesta de un edificio monitorizado por big data

[2] POI: Punto de interés, por sus siglas en inglés: Point Of Interests. Por ejemplo: Agencia, hospital, parques, …

es de un 40% menos (Munawar et al, 2022); la energía, ahorro de energía a través de contadores inteligentes, se calcula que hoy en día los contadores realizan un total de 24 millones de lecturas al año, frente a 220 millones de lecturas al día que harán los contadores inteligentes en los próximos años (Zhou et al, 2016); y para la identificación de riesgos (los macrodatos pueden emplearse para evaluar el peligro de fallos estructurales en los edificios, por ejemplo debido a los daños de un seísmo). También, permite elaborar una predicción del retraso en la entrega del inmueble (Munawar et al, 2022). Muchas promotoras se enfrentan a multas y sanciones por incumplimientos de fechas de entrega. Con la utilización de big data se puede fijar una fecha más exacta para la finalización del proyecto y su consecuente entrega del bien inmueble a su comprador. Sannie Anibire, en un estudio, elaboró un modelo basado en redes neuronales y machine learning, que conseguía predecir el retraso en las entregas en un 93,75% (Munawar et al, 2022).

La **minería de datos** también se emplea como soporte en la construcción. Ésta detecta regularidades útiles e información necesaria para la toma de decisiones en proyectos de gestión de la construcción. Un método de minería de datos como el análisis de conglomerados puede resultar muy útil, ya que permite la combinación de diferentes objetos de construcción en grupos homogéneos e investigarlos y sacar conclusiones.

- **Valoración de inmuebles:** La valoración de inmuebles se considera una de las grandes aportaciones del big data a la industria del ladrillo, pues permite valorar de una manera más rápida, ágil, cómoda y, sobre todo, de manera exacta cualquier inmueble. La valoración de inmueble mediante el big data se basa en, a través del machine learning, predecir cuanto debe valer un inmueble (Kok, Leena, Martinez, 2017). El big data hace esto posible debido a su poder para el análisis de datos en tiempo real. Este permite el seguimiento de las tendencias del valor de la propiedad, la demanda de alquiler y compra en tiempo real. También posibilita analizar otras variables demográficas como el auge de una población o aumento de calidad de vida y variables macroeconómicas como el seguimiento de intereses. Toda esta información se recopila durante intervalos de tiempo para elaborar el modelo. Posteriormente, se analiza más a fondo utilizando el análisis predictivo para determinar los valores inmobiliarios (Oluwunmi, 2019).

La tasación de inmuebles ha presentado desajustes históricamente. En 2011, Cannon y Cole realizaron un estudio sobre la exactitud de las tasaciones del sector inmobiliario de Estados Unidos, creando una base de datos que comprendía datos procedentes del National Council of Real Estate Investment Fiduciaries (NCREIF) entre los años 1984-2010, y compararon las tasaciones inmobiliarias con las transacciones de compraventa reales. Los autores documentaron que, de media, las tasaciones están más de un 12% por encima o por debajo del precio de la transacción posterior. (Kok, Leena, Martinez, 2017). Estos resultados coinciden con los de Fisher, Miles y Webb en su estudio realizado en el 1999 para el periodo 1978-1998. Una vez finalizado su estudio, documentaron una desviación típica media del 9% a 12,5% entre las tasaciones y los precios de transacción (Kok, Leena, Martinez, 2017). Mientras siguen existiendo la falta de exactitud e ineficacia en las tasaciones de inmuebles, el sector inmobiliario ha sido testigo de un aumento significativo de la disponibilidad de datos y la llegada de técnicas de aprendizaje automático, permitiendo elaborar a través de modelos cuantitativos una tasación de la propiedad de una manera más fiable y precisa (Kok, Leena, Martinez, 2017).

4.2 Obtención de datos:

Para la extracción de datos, los agentes cuentan con múltiples herramientas que permiten crear bases de datos masivas para poder ser analizadas posteriormente. El mundo está inmerso en una era de enormes cantidades de información. Tan solo este año, se crearán más de mil billones de gigabytes de datos nuevos a nivel global. El fenómeno del Big Data marca una nueva etapa de oportunidades y desafíos en términos de innovación, competitividad y productividad para las empresas que deseen destacar frente a sus competidores. Cada día, se generan en el mundo más de 2.5 exabytes de datos, lo cual equivale a un millón de terabytes. El crecimiento del volumen de datos no se limita a una simple expansión, sino que se está produciendo de manera exponencial. De hecho, este crecimiento exponencial es tan grande que el 90% de los datos generados no se utilizan ni se analizan. Cada segundo, cualquier dispositivo electrónico genera una gran cantidad de datos. De la obtención de éstos, se puede obtener una gran ventaja competitiva frente a los competidores. Aunque, como hemos dicho, no vale con almacenarlos, hay que saber usarlos para conseguir beneficios (Puyol, 2014).

A continuación, se presenta las principales fuentes donde las empresas de nuestro sector analizado recopilan datos para su uso en la toma de decisiones y búsqueda de información relevante sobre el mercado.

4.2.1 Web Scrapping

El valor de los datos hoy en día está claro, de ahí a que se le llame el nuevo petróleo. Debido a esto, se encuentran muy reconocidos en internet y generalmente pertenecen a las grandes tecnológicas como Google, Meta, … Dicha cuestión termina en un pobre intercambio de datos de estas compañías de cara al público (Wei, et al, 2022). Gracias al web scraping, los usuarios pueden acceder a datos de internet y descargarlos si tienen permiso para ello.

El web scraping es una técnica usada por la mayoría de empresas que, mediante software, extraen información o contenido de un sitio web. Los usos del web scraping son infinitos, cualquier web contiene datos y cualquiera puede ser descargado, el límite de descarga lo ponen la creatividad y la legalidad (Krotov, Silva, 2018).

Normalmente, los datos de la web se desechan utilizando el protocolo de transferencia de hipertexto (HTTP) o a través de un navegador web. Esto se realiza manualmente por un usuario o automáticamente por un bot o rastreador web. Debido a la enorme cantidad de datos heterogéneos generados de manera constante en la WWW (World Wide Web), el web scraping está ampliamente reconocido como una técnica eficaz y potente para recopilar big data de una variedad de escenarios. Los métodos actuales de web scraping se han personalizado desde procedimientos ad hoc más pequeños, asistidos por humanos, hasta la utilización de sistemas totalmente automatizados capaces de convertir sitios web enteros en conjunto de datos (Zhao, 2017). Estos conjuntos de datos se obtienen de una manera no estructurada, así que, para poder sacar partido de ellos, hay que estructurarlos.

En definitiva, hacer web scraping se refiere a descargar, de forma automatizada o paso a paso, información relevante de una página web (Glez-Peña, et al 2014).

Las empresas inmobiliarias pueden sacar partido a esta técnica y, por ejemplo, obtener información relevante de marketplaces, creando bases de datos de inmuebles en venta y analizando las variables más significativas.

4.2.2 Datos procedentes de Remote Sensing.

El avance en la tecnología de sensores ha permitido que la ciencia de teledetección consiga proporcionar datos de imágenes con muchas resoluciones, y múltiples espectros y escalas temporales. Dado que ofrece una imagen aérea de arriba abajo del inmueble que se va a tasar y del entorno comunitario que lo rodea, la tecnología de teledetección se ha revelado como una técnica avanzada y fundamental para medir las condiciones ambientales externas de una vivienda (Wei, et al, 2022). En concreto, tiene tres direcciones de aplicación:

- **Imágenes multiespectrales:** La información del entorno superficial que aportan las imágenes multiespectrales es mucho más especializada y concreta que la de las imágenes en color verdadero. Esto puede ayudar a evaluar con mayor eficacia las condiciones ambientales que rodean la propiedad, factor crítico para la tasación inmobiliaria. En general, la aplicación de imágenes multiespectrales de teledetección se centra predominantemente en el cálculo del índice medioambiental (Wei, et al, 2022).
- **Night-time light:** Dado que la valoración inmobiliaria tiene fuertes atributos sociales y económicos, es necesario centrarse en las características sociales y económicas que rodean y caracterizan a los bienes inmuebles. La luz nocturna se ha de tener en consideración, puesto que se ha descubierto que está muy correlacionada con el producto interior bruto (PIB) y la densidad de población (Wei, et al, 2022). En comparación con los datos económicos y sociales totales tradicionales proporcionados por el gobierno, la luz nocturna muestra el crecimiento económico de las ciudades y, por tanto, ha sido utilizada por algunos investigadores para modelizar los precios de la vivienda con algunos resultados prometedores. (Zhao, et al, 2017)
- **Radar láser**: La principal ventaja de los radares láseres frente a las imágenes de teledetección es que, a diferencia de éstas, que sólo pueden presentar las características bidimensionales de los objetos terrestres, el radar láser puede proporcionar algunos nuevos parámetros tridimensionales basados en el espacio para el modelo de evaluación (Wei, et al, 2022). Actualmente, entre sus principales usos, destacan la medición del volumen y superficie de las

casas, y la mensuración del ángulo de visión que tiene el inmueble sobre el mar (Wei, et al, 2022).

4.2.3 Datos procedentes de IoT (Internet of Things)

Las cosas físicas están creando continuamente grandes volúmenes de big data a través de sensores en todo el entorno urbano: en edificios, carreteras, farolas, infraestructuras y otros lugares.

Teléfonos inteligentes, tabletas, aplicaciones de consumo, ordenadores, plataformas de medios sociales y vehículos conectados a la red forman parte de lo que se conoce como Internet de las cosas (IoT), que permite recopilar y aplicar conocimientos sobre el entorno y sus habitantes (Yovanof y Hazapis, 2009). La respuesta por parte de las empresas ha sido la creciente utilización del análisis de macrodatos para elaborar fiables conclusiones sobre la evolución del mercado y el conocimiento de clientes y productos (Donner, Eriksson, Steep, 2018). Las fuentes más valiosas de estudio para las empresas del sector inmobiliario son las siguientes:

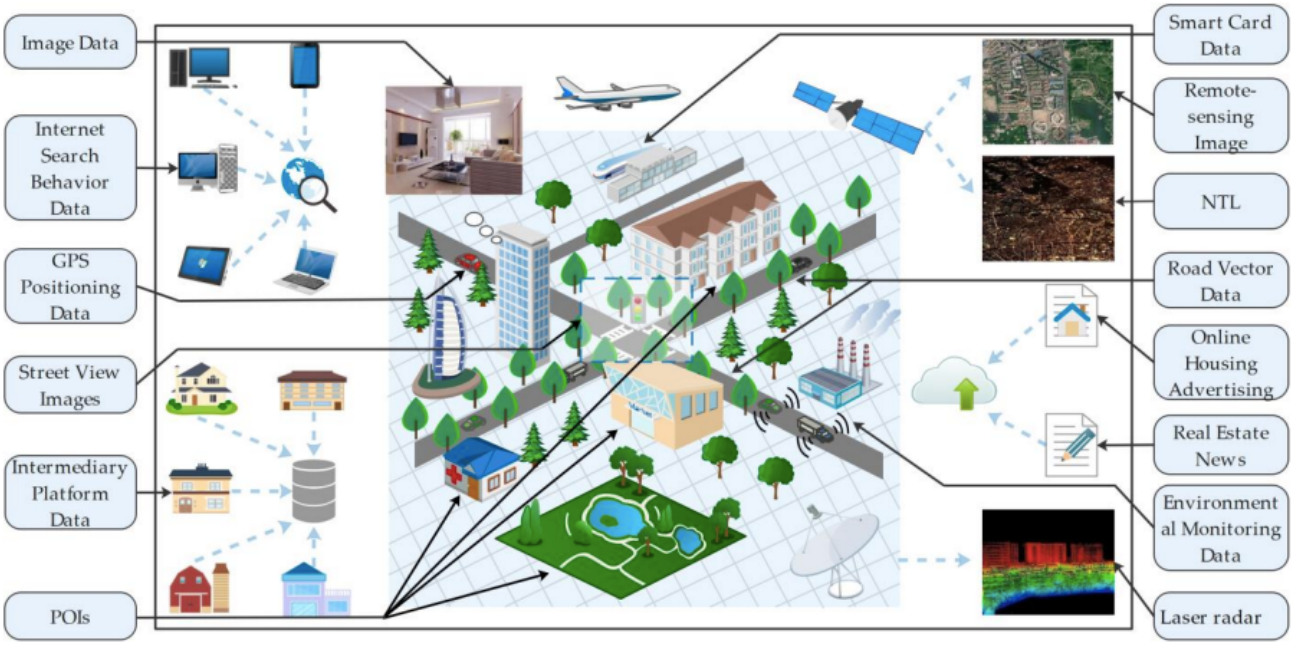
- **Sensores de monitoreo medioambiental**: Información proporcionada por sensores capaces de detectar, por ejemplo, el número de personas en un lugar, el ruido, la contaminación, etc. Estos datos los facilitan sobre todo los departamentos gubernamentales mediante el uso de sensores medioambientales, lo cual puede aportar información valiosa en relación con el valor inmobiliario (Wei, et al, 2022). La identificación de movimientos, aglomeraciones y reuniones revelará qué tipo de actividades se desarrollan en un barrio, por ejemplo, si la gente va de compras, come y bebe en restaurantes al aire libre o camina en una dirección determinada. Vincular la contaminación y el ruido al sector inmobiliario es un uso sencillo de este tipo de datos que puede aportar información valiosa a los promotores inmobiliarios. Además de los sensores en edificios e infraestructuras, los teléfonos inteligentes permiten hacer un seguimiento de entornos, como el ruido y el tiempo (Donner, Eriksson, Steep, 2018).

  Muchos estudios aclaman la vinculación que tienen estas variables con el sector inmobiliario, en especial, con los precios de la vivienda. Por ejemplo, el índice de contaminación del aire guarda una relación inversa con el precio de la vivienda en Pekín, China (Mei, et al, 2020). Con respecto a la

contaminación acústica, un estudio realizado por Zambrano-Monserrate y Ruano (2019) en la ciudad de Machala, Ecuador, muestra que cada aumento de 1 decibelio en el ruido ambiental reduce el precio de la vivienda en un 1,97% (Wei, et al, 2022).

- **Sensores de movimiento:** Este tipo de información se puede recopilar gracias a los servicios que ofrecen los coches conectados a la red, el transporte público y los teléfonos inteligentes. Aunque se plantean algunos problemas de privacidad, los sensores de los teléfonos inteligentes permiten seguir a las personas dentro de una ciudad (Han et al., 2015). El gobierno utilizó este tipo de rastreo para combatir la pandemia de la Covid-19 controlando los movimientos de la ciudadanía (Masdeau, 2020). El conocimiento del movimiento proporciona información valiosa, de modo que los patrones de consumo y actividad social pueden desglosarse por demografía o situación geográfica. La identificación de patrones de movimiento en una ciudad, como saber que un determinado grupo demográfico tiende a trabajar en el lugar X y se desplaza a comercios y restaurantes del lugar Y a una determinada hora del día, es útil para saber dónde ubicar los nuevos desarrollos y analizar el impacto de éstos en el transporte. Este tipo de datos también puede utilizarse para identificar tendencias de crecimiento de localidades o de desarrollo económico (Donner, Eriksson, Steep, 2018).
- **Tarjetas inteligentes**: Tarjetas de viaje como las de metro han sido objeto de estudio para encontrar patrones sobre el precio de los inmuebles en las ciudades. Un estudio realizado por Zhu et al., (2018) consiguió descubrir que la distancia de desplazamiento en metro y la frecuencia de viaje de los pasajeros están significativamente correlacionadas negativamente con los ingresos de los individuos. Por lo general, un individuo con un alto status social no tiende a usar transporte público. Otros sugieren que la tendencia de movilidad y el tiempo dedicado a viajar por las personas en las ciudades están relacionados con los precios regionales de la vivienda (Wei, et al, 2022).

Figura 9. Smart real estate collection data



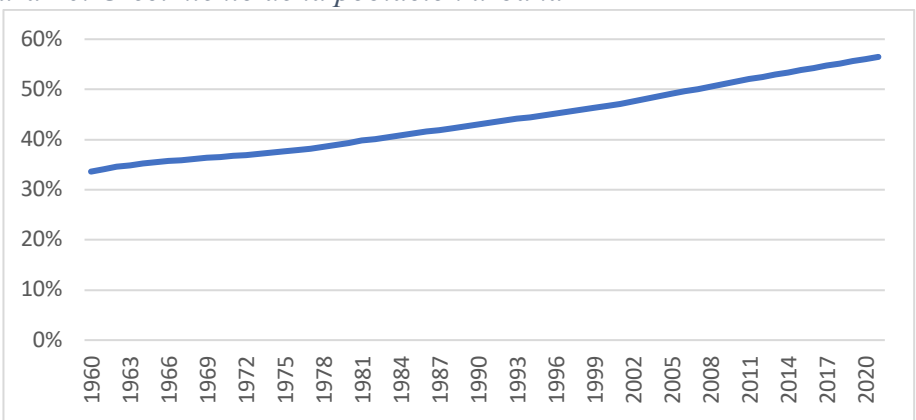
(Wei et al, 2022).

4.3 Smart cities

4.3.1 Definición y ámbitos de actuación

Actualmente, según el banco mundial, el 81% de la población española vive en ciudades (56% de la población mundial). Además, de los que no viven en zonas urbanas, el 70% accede a los servicios urbanos al menos una vez al día. Este porcentaje se encuentra en continuo crecimiento (**Ver figura 10**). En el año 2000 solo el 46% de la población global vivía en ciudades, nada que ver con la actual cifra de 56% o la previsión de las Naciones Unidas para 2050, la cual se sitúa en 66%.

Figura 10. Crecimiento de la población urbana



(*Elaboración propia a partir del Banco Mundial)*

El proceso de urbanización ha mejorado considerablemente la vida de las personas, permitiendo el acceso a agua, energía, luz; así como posibilitado tener un lugar donde residir, método para transportarse, etc.… Las ciudades están llamadas a seguir siendo el principal lugar donde habiten las personas. Éstas se encuentran inmersas en un proceso de transformación digital, pasando de un tipo gestión tradicional a convertirse en Smart Cities o ciudades inteligentes. Ahora bien, ¿Qué es una ciudad inteligente? Harrison y otros definieron las **Smart cities** como: una ciudad capacitada, interconectada e inteligente. Más tarde, Giffinder y Gurum añadían 6 características que debía considerar una ciudad inteligente: economía, gobernanza, medioambiente y sostenibilidad, personas, movilidad y calidad de vida. Por lo tanto, podríamos concluir que una ciudad inteligente es aquella que hace uso de las TIC para hacer del conjunto de la ciudad (personas, administración, educación, …) un lugar más eficiente y entendido (Yin, et al, 2015).

Muchos autores basaron su definición en el uso de los datos. Harrison et al. describieron una ciudad inteligente en torno a 3 factores: **Recopilación o instrumentación, interconexión e inteligencia (Ver siguiente figura)**. La instrumentación permite la captura e integración de datos en vivo del mundo real a través de sensores, para que seguidamente se produzca una interconexión que permita que los datos obtenidos de la instrumentación se integren a través de múltiples procesos, sistemas, organizaciones, industrias o cadenas de valor. Por último, inteligencia se refiere a que el procesamiento de datos debe producir nuevas percepciones que impulsen decisiones y acciones que puedan demostrar un valor añadido tangible (Yin, et al, 2015).

Figura 11. Procesamiento de datos smart city

(Elaboración propia basada en Yin, et al, 2015)

Tabla 2. Ámbitos de actuación de las smart cities

Ámbito	Especialización	Descripción
Gobernanza (Más eficiente)	E-Gobernanza Transparencia Servicios públicos Seguridad Emergencia Respuestas eficientes	Mejorar la eficiencia interna y externa del gobierno; permitir a los ciudadanos y a otras acceder a documentos y políticas oficiales; garantizar el funcionamiento eficaz de los servicios públicos y gestionar la seguridad pública; responder rápida y responder rápida y eficazmente en situaciones de emergencia.
Ciudadanía (Mejor estilo de vida)	Transporte público Tráfico inteligente Turismo Entretenimiento Sanidad Educación Consumo Cohesión social	Viajar y desplazarse de forma más eficiente; acceder a información contextualizada, precisa y en tiempo real en la vida cotidiana; Mejoras en servicios públicos, como la educación la sanidad y el deporte; actividades enriquecedoras para el tiempo libre, comunicarse y compartir más con los demás; vías de escape en caso de desastres naturales.
Empresarial (Más prosperidad)	Gestión empresarial Logística Cadena de suministro Transacciones Publicidad Innovación Sector inmobiliario Agricultura	Mejorar la eficiencia y la calidad de la gestión; utilizar plataformas y métodos más eficientes de logística; hacer una publicidad más amplia, transparente y precisa; ampliar los socios comerciales y los clientes; fomentar el espíritu empresarial y la inversión; mejorar la actividad la actividad empresarial en una ciudad, como la producción, el comercio, la agricultura y la consultoría.
Medioambiente (Más sostenible)	Gestión del agua Gestión de residuos Control de la contaminación Construcción Vivienda Comunidad Espacio público	Suministro de energía y agua más sostenibles, teniendo en cuenta el comportamiento de los ciudadanos; utilizando más energía verde o renovable; reciclando, tratando los residuos de forma eficiente y segura; control y prevención de la contaminación en la ciudad; ofreciendo movilidad y zonas verdes.

(Elaboración propia a partir de Yin et al)

4.3.2 Tecnología capaz de implementar el cambio

La tecnología que va a conducir el cambio entre la gestión tradicional de ciudades a la gestión inteligente es variada. La recopilación de datos del entorno es posible mediante la tecnología descrita anteriormente en el apartado de obtención de datos:

sensores, vehículos, POIs, móviles, … **(Ver figura 8).** También ayudarán aplicaciones como Zillow (Marketplace inmobiliario) y Yelp (reseñas de restaurantes y reservas) que permiten conocer las preferencias de las poblaciones (Donner, Eriksson, Steep, 2018). Para el procesamiento de los datos, existen numerosos avances que permiten obtener conclusiones y elaborar un análisis rentable del conjunto de datos: (Donner, Eriksson, Steep, 2018).

- **Computing power**: proporciona la fuerza tecnológica para analizar los datos y permite una visualización compleja.
- **Posibilidad de gran almacenamiento de datos**: El almacenamiento ha sido uno de los principales retos de los macrodatos, ya que el rápido aumento de las cantidades de información requiere nuevas tecnologías de almacenamiento.
- **Desarrollo de algoritmos:** identifica patrones de interés en los datos y posibilita el análisis predictivo basado en big data.
- **Inteligencia artificial:** Permite a las aplicaciones pensar, aprender y evaluar basándose en los datos recogidos, lo que permite a las máquinas realizar tareas que requieran de inteligencia de una manera eficiente.
- **Machine learning**: es el resultado de la práctica de utilizar algoritmos para analizar datos, aprender, y luego tomar decisiones o hacer predicciones.
- **Metamateriales:** Consiste en materiales que no se encuentran en el medioambiente, es decir, se han elaborado a partir de otras materias. Engloba desde la mejora del rendimiento de las antenas, las superlentes, los materiales para tejados que disminuye la temperatura en los edificios, hasta la tecnología láser utilizada para la fibra óptica.

4.3.3 ¿Cómo afectará al sector inmobiliario?

La revolución hacia smart cities no debe ser vista como un proyecto a futuro, pues ya está sucediendo. Se aprecian los primeros pasos en la utilización de tecnologías tales como conectividad a través de fibra óptica, recolección de datos por medio de sensores inteligentes, creación y uso de programa para grandes análisis de datos, teléfonos móviles, etc. (Bouskela, et al, 2016). La disrupción tecnológica por la ruptura de la gestión tradicional de las ciudades hacia una gestión tecnológica e inteligente va a abarcar cuatro grandes ámbitos en el sector inmobiliario: (Donner, Eriksson, Steep, 2018).

En primer lugar, contribuirá al **desarrollo de predicciones de las preferencias** de los usuarios y del uso de los inmuebles. La capacidad de recopilar y analizar preferencias de los usuarios tendrá un profundo impacto en el desarrollo y la gestión inmobiliaria. El campo de actuación será similar al que usan las agencias y particulares para prever las tendencias del mercado y evaluación de ingresos y rentabilidades futuras a partir de demanda (Hashem et al., 2016). Por ejemplo, gracias a los sensores de localización a través de GPS, podremos saber la transitividad de cierta zona y si constituye o no un punto estratégico para el desarrollo de un proyecto. (Wei et al, 2022)

En segundo lugar, va a **revolucionar la construcción y gestión del bien inmueble**. Las smart cities facilitarán la transformación de los bienes inmuebles y promoverán cambios en el uso de los inmuebles existentes por parte de los usuarios acorde a sus demandas. Mediante el escaneo láser durante las etapas de construcción, se pueden crear modelos digitales de los inmuebles, capturando todos los aspectos de la estructura (Donner, Eriksson, Steep, 2018). Una vez que el edificio está ocupado, es posible estudiar su utilización a través de una combinación de sensores, datos telefónicos, GPS y transacciones comerciales. El análisis del uso de los inmuebles puede llevar a la toma de decisiones de modificación. Obtener información detallada, precisa, completa y a tiempo real sobre los edificios permite que sea más económico y rápido adaptarlos a las necesidades de los usuarios. Esta discusión puede resultar en cambios en los bienes inmuebles, un aumento en los alquileres y/o modificaciones en la forma en que el inquilino utiliza el edificio (Donner, Eriksson, Steep, 2018).

En tercer lugar, la **identificación del poder adquisitivo de los usuarios** a través de, por ejemplo, tarjetas inteligentes, permitirán fijar precios y adaptar los proyectos a las zonas geográficas. Los gestores y promotores inmobiliarios podrán dirigirse mejor a los distintos tipos de clientes, por ejemplo, a precios de gama alta o de gama baja. (Wei, et al, 2022) El conocimiento de los patrones de compra también pueden significar un elemento a tener en cuenta para fijaciones de precio (Donner, Eriksson, Steep, 2018).

Por último, el avance hacia ciudades inteligentes permitirá la **identificación de riesgos** con mayor antelación. Los datos digitales de las ciudades pueden utilizarse para evaluar el riesgo en la estructura de los edificios, como los debido a daños sísmicos (Yu, et al, 2018). También permitirá mitigar el riesgo de sobre impagos o mal uso del bien inmueble de clientes personales. La información digital recopilada sobre diferentes grupos de usuarios, como arrendatarios y clientes finales, en establecimientos

comerciales, restaurantes, oficinas y propiedades residenciales, brinda datos acerca de las características de riesgo individuales. Estos datos a su vez pueden ser utilizados para identificar los posibles riesgos asociados a los activos comerciales en una ciudad. (Donner, Eriksson, Steep, 2018)

4.3.4 Retos de las Smart cities

Las ciudades inteligentes se van a tener que enfrentar a 2 grandes retos:

1. La privacidad de los datos.
2. Ciberseguridad.

La **privacidad de los datos** es todo un reto a gran escala para el big data. Actualmente, cada búsqueda que realizamos en Google queda registrada por la empresa. Con esa información, se elabora un perfil informático que las empresas usan para enviarnos publicidad personalizada. Saben nuestra edad, género, e incluso si estamos en una relación, sin nunca habérselo comunicado, solo analizando nuestro patrón de búsquedas. No solo pasa con Google, las redes sociales también elaboran perfiles a través de patrones de comportamiento en cuanto a "me gustas", comentarios, número de veces que repetimos un vídeo o tiempo que estamos parados delante de una publicación. Además, conocen nuestro ambiente, ubicación, e infinidad de datos a los que les dejamos acceder. (Yebra del Puerto, 2018)

Como hemos comentado, las ciudades inteligentes se van a basar en el análisis de datos para crear valor. Por consecuencia, los datos recopilados sobre personas individuales se van a multiplicar en los próximos años. Debido a la naturaleza de la conectividad en una ciudad inteligente, los datos se transferirán y emplearán en diversos procesos, involucrando a múltiples actores que se comunicarán y accederán a la información, desde los fabricantes de sensores inteligentes, hasta las autoridades de transporte de la ciudad, e incluso los individuos que utilizan la ciudad inteligente a través de sus dispositivos móviles. Cada organización que contribuye al desarrollo de la ciudad inteligente utilizará y manejará los datos de manera única, lo que puede comprometer la privacidad personal (Braun, et al, 2018). Un ejemplo claro para conocer la vulnerabilidad de las personas en un entorno Smart puede ser el siguiente: Mediante la matrícula de un vehículo, podemos conocer quién es su propietario. En el futuro, los vehículos, aparte de poder ser rastreados por GPS como en la actualidad, van a estar conectados a internet y, seguramente, estará continuamente compartiendo información con el teléfono móvil. Por tanto, a partir de la

matrícula, se podrá acceder a los datos más íntimos que una persona guarda en su smarthphone. Por otro lado, la trayectoria de un vehículo puede rastrearse fácilmente, aunque todas las comunicaciones entre el vehículo y la infraestructura estén encriptadas y cada dispositivo esté autenticado por otros (Khatoun y Zeadally, 2017).

En cuanto a la **ciberseguridad**, esta se centra específicamente en proteger la información de ataques a través de la web. La ciberseguridad es de vital importancia debido al creciente riesgo de ciberataques e incidentes que afectan a sectores críticos en las ciudades inteligentes. Al considerar las amenazas cibernéticas en las ciudades inteligentes, existen peligros que afectan tanto a los datos como a los sistemas de procesamiento. Las amenazas a los datos incluyen el riesgo de exposición de información personal identificable (IPI), riesgo de actividades maliciosas (acoso, chantaje, sobornos, …) y ataques distribuidos de denegación de servicios (DDoS) (Rawat y Ghafoor, 2018). Muchos autores ponen ahora el foco en hacer frente a estos retos. Para solucionarlos se discute cual será la mejor opción tecnología que consiga la tranquilidad de la ciudadanía en un entorno urbano inteligente (Ver siguiente tabla). Un problema grande para la contribución hacia un entorno smart seguro es que, las empresas buscan obtener beneficios de su actividad, por lo que ofrecerán la protección justa y necesaria para garantizar que su compañía no sufre daños reputacionales por vulnerar la privacidad de los usuarios (Braun, et al, 2018).

Tabla 3. Tecnologías contra vulneraciones en smart cities.

Tecnología	Explicación.
Blockchain	La tecnología Blockchain proporciona servicios transparentes, descentralizados, democráticos y seguros sin necesidad de terceros. Va a ser una herramienta crucial para el futuro, ya que permite la implementación de métodos de identidad digital.
Enfoque en datos	La ciberseguridad basada en datos es un conjunto de principios y técnicas que funcionan conjuntamente para tomar decisiones de ciberseguridad basadas en datos analizados de sistemas o aplicaciones, en lugar de intuición o presentimiento. Puede hacerse mediante el análisis de datos, análisis de riesgos y visualización de datos.
HSCCA	Acrónimo de *Hybrid Smart City Cybersecurity Architecture*. Analiza las amenazas además de crear un sistema de datos seguros. Para la creación de una ciudad inteligente, esta tecnología permite la recopilación, recuperación y almacenamiento de datos valiosos, al igual que el suministro de una fuente de red bien organizada y de alto nivel.
Enfoque en la probabilidad	El modelo representa el flujo de datos entre diversos agentes mediante Bigraph, averiguando así el agente culpable de la fuga de datos
ICADS ontologies	Es un sistema de estructuración integrado por capas para la seguridad de los usuarios.

(Elaboración propia basada en Almaner y Almaiah, 2021).

5. Aplicación práctica

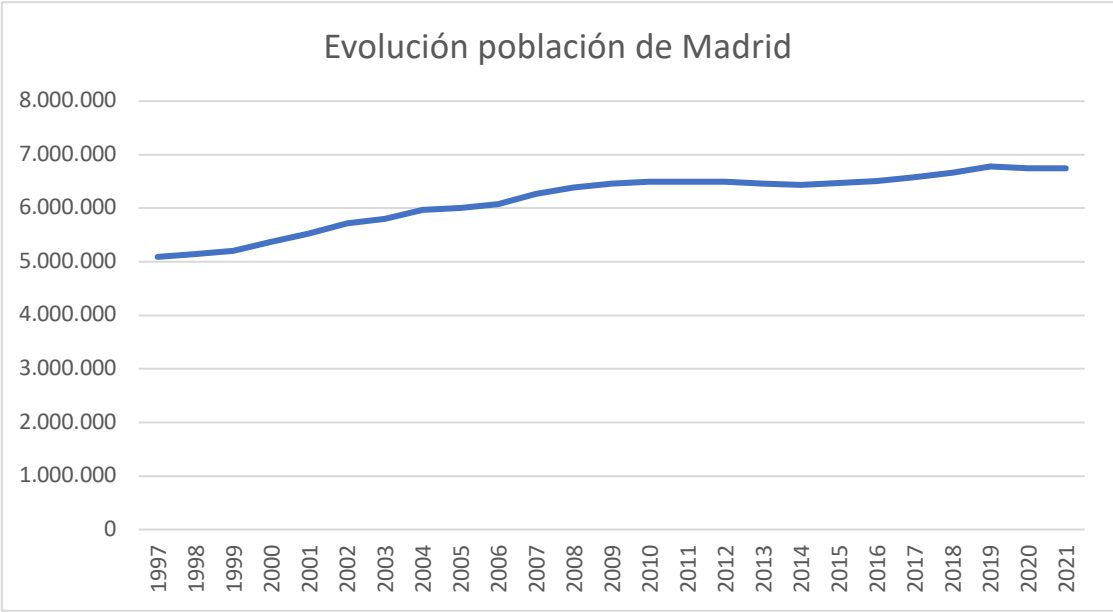
En el presente apartado se pondrá en práctica todo lo aprendido sobre la transformación del sector inmobiliario hacia un sector digitalizado, que usa datos masivos para tomar decisiones debidamente respaldadas. Se tratará de otorgar una visión panorámica de lo que son, hoy en día, la mayoría de las empresas del sector inmobiliario o empresas Proptech.

No se entrará en el análisis más científico posible, aunque sí será de lo más preciso, para hacer entender al lector la utilización de estas nuevas técnicas. Se desarrollará desde un punto de vista cercano a las empresas que no cuentan con recursos infinitos como los gigantes de la industria, los cuales pueden acceder a un volumen inmenso de datos con una variabilidad muy significativa. Se verá como la aplicación de estas nuevas técnicas es también posible para inversores particulares o pequeñas agencias a partir de datos extraídos de la web, sin un gasto excesivo en la elaboración del database.

5.1 Objeto de estudio y objetivos

El objeto de estudio del proyecto será realizar un estudio sobre la oferta de viviendas unifamiliares de Madrid. La capital de España ha experimentado un aumento en el número de ofertas inmobiliarias debido a, principalmente, el auge de la población. En concreto, entre 1996 y 2021, (25 años), la población de Madrid ha ascendido en un 32%, pasando de 5,91 mill de ciudadanos a 6,75 mill. (**Ver siguiente figura**).

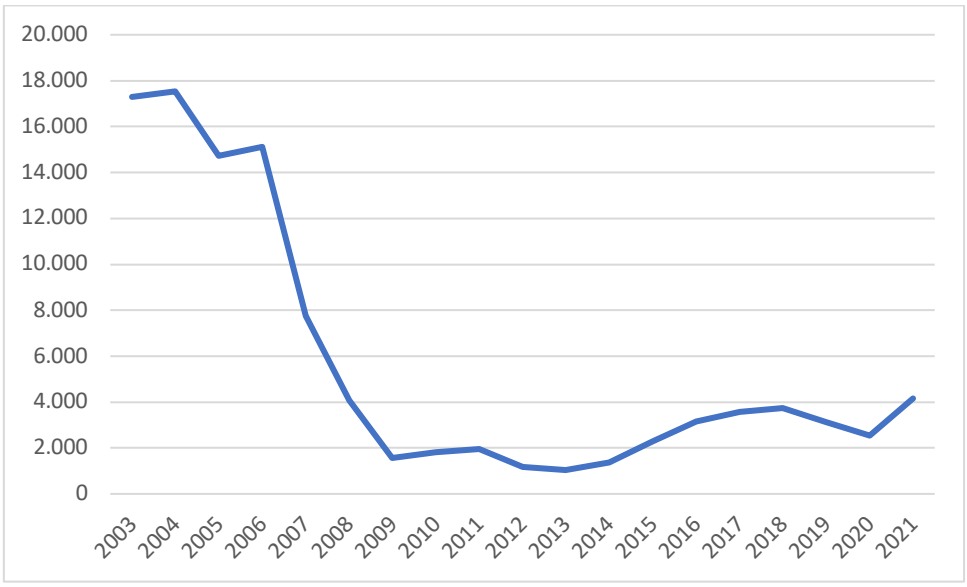
Figura 12. Evolución poblacional de Madrid



(Elaboración propia a partir de los datos del INE)

La oferta de viviendas de obra nueva en la comunidad también ha crecido en los últimos años **(Ver figura 13).** No obstante, la situación se encuentra lejos del incremento que sufrió dicha variable en los años previos a la crisis financiera, años también conocidos como los años del "Boom Inmobiliario". Parece que la construcción de casas en la actualidad se elabora de una manera más consciente y asumiendo menos riesgos.

Figura 13. Construcción de viviendas en Madrid



(Elaboración propia a partir del IDE)

Estas variables y relaciones permiten entender de mejor manera a la demografía de una zona territorial, pues para una empresa inmobiliaria, no es el mismo mercado una zona con prosperidad poblacional y económica como Madrid, que una zona geográfica donde el número de habitantes sea decreciente.

Partiendo de estas premisas, se pretende hacer entender al lector como una empresa puede beneficiarse de la revolución 4.0 en la que está sumergida el mundo actual, sin la necesidad de grandes inversiones ni patrimonios para la obtención de información relevante del sector y el uso de la inteligencia de datos para poder sacar partido de ella.

Los **objetivos** del análisis serán los siguientes:

- Realizar un **estudio de las variables**, así como su relación entre ellas. De esta manera, se pretenderá obtener una imagen global del mercado, conociendo su situación, factores que influyen, y como varían unas variables con respecto a otras. Se atenderá a técnicas de análisis de macrodatos, tanto analíticas como estadísticas, para poder llevar a cabo el proyecto. Se aplicarán técnicas de *Big Data Analytics*, para así poder convertir los datos en valor empresarial.
- Realizar un modelo que consiga predecir los precios de los inmuebles de Madrid de la manera más exacta posible. En esta sección, se atenderá a la información extraída para realizar un modelo cuantitativo que consiga proporcionar una respuesta a nuestra cuestión. Dicho modelo se va a basar en el *machine learning* o aprendizaje automático para alcanzar sabias conclusiones sobre lo que debe valer un piso. Con esto, pretendemos conocer la principal aportación de la tecnología del big data a la industria del ladrillo: valoración y tasación de inmuebles de una manera precisa, evitando las diferencias explicadas en el apartado del big data en el sector inmobiliario.

A partir de estos dos objetivos, se analizarán los datos extraídos de internet para analizar patrones de tendencia, conocer el mercado inmobiliario de Madrid y poner en

prácticas las técnicas de modelización. También se quiere hacer ver la gran ventaja que las empresas, sean del tamaño que sean, pueden obtener a partir del uso de estas técnicas.

5.2 Metodología

En este apartado, se presenta la metodología empleada para alcanzar las conclusiones de este trabajo, centrándose en el uso de RStudio como la plataforma principal de análisis de datos. La elección de RStudio se basa en el conocimiento y experiencia previa que se tiene en esta herramienta, considerándola como la opción más adecuada para abordar el análisis de una extensa base de datos de inmuebles.

RStudio, un entorno de desarrollo integrado ampliamente utilizado en la comunidad científica y empresarial ofrece numerosas ventajas y funcionalidades para el análisis y visualización de datos. Aunque, en su mayoría y debido al gran volumen de datos, para la visualización de datos hemos recurrido a una plataforma especializada en la visualización de datos, Tableau.

La metodología seguida en este trabajo ha consistido en una serie de pasos rigurosos y sistemáticos, aprovechando las capacidades de la plataforma para realizar diversas tareas relacionadas con la manipulación, limpieza y análisis de la base de datos de inmuebles. La familiaridad con el software ha permitido un flujo de trabajo más eficiente y una mayor confianza en la calidad de los resultados obtenidos.

A continuación, se presentan los pasos seguidos para la elaboración del proyecto:

1. **Definición de los objetivos** del estudio: Antes de comenzar el análisis, es importante establecer los objetivos específicos que se desean lograr. Dichos objetivos han sido enumerados en el apartado anterior.
2. **Recopilación de datos:** Se ha extraído una base de datos de inmuebles de Madrid. Para su recopilación, se ha asegurado que los datos sean relevantes, como el precio de venta, ubicación, tamaño del inmueble, características y cualquier otra variable que sea de interés.
3. **Limpieza y preparación de los datos:** En esta fase, se realiza una limpieza exhaustiva de los datos para eliminar valores atípicos, duplicados o faltantes. Asimismo, se ha procedido a la eliminación de variables irrelevantes o repletas de valores nulos.

4. **Análisis exploratorio de datos (Análisis EDA):** Se ha realizado un análisis exploratorio de los datos para comprender las características de los inmuebles en Madrid. Se han incluido también técnicas de visualización de datos a través de gráficos y tablas, identificación de relaciones entre variables y cálculo de estadísticas descriptivas.
5. **Machine Learning:** Para esta sección, se han utilizado técnicas de los dos tipos de modelos explicados en el apartado de Modelos Predictivos en el Sector Inmobiliario: supervisados y no supervisados. Se ha requerido de su ayuda para analizar los datos y obtener información más profunda. Se incluyen modelos cuantitativos para predecir los precios de los inmuebles en función de variables explicativas, análisis de clústeres para identificar grupos de inmuebles similares. Con más detalle se explicará, en esta sección, los pasos para elaborar el modelo de valoración de inmuebles.
6. **Interpretación de los resultados**: Se han analizado los resultados obtenidos del modelado estadístico y de otro análisis realizado y se han interpretado. También, se ha evaluado la importancia de las variables en relación con los precios de los inmuebles para extraer conclusiones sobre nuestra variable objetivo.
7. **Conclusiones**: Basándose en los análisis realizados, se han extraído conclusiones finales, así como los hallazgos claves sobre los datos de inmuebles en Madrid.

5.3 Elaboración del caso práctico

En este apartado se verá con mayor detalle los procedimientos a seguir en cada paso visto en el apartado anterior. Se explicará desde la recopilación de la información hasta las conclusiones finales el desarrollo, haciendo hincapié en aquellos puntos que resultan de mayor relevancia.

5.3.1 Recopilación de la información

En este proceso, es muy importante asegurarse de que los datos recogidos para la elaboración del modelo cumplen con las expectativas que debe tener una buena base de datos: **Variabilidad** (Datos de formalidad variada), **volumen** (gran número de datos) y **velocidad** (rapidez con la que se generan y analizan datos en tiempo real) (Aguilar, 2016).

Para realizar nuestro estudio se ha accedido a una base de datos pública. La base de datos utilizada en este estudio fue extraída de Kaggle, una reconocida plataforma que alberga una amplia variedad de conjuntos de datos de diversas temáticas. La elección de esta base de datos en particular se basó en su relevancia y pertinencia para el objetivo de investigación planteado.

La recopilación de información representa el primer paso en la construcción de cualquier estudio o investigación, por lo tanto, es fundamental comprender el proceso de recopilación de datos y garantizar su fiabilidad y validez, lo que permitirá obtener resultados sólidos y confiables en nuestro estudio.

Dicho esto, comprobamos que nuestra base de datos es adecuada:

1. Presenta un gran número de datos. Si el database contiene un gran número de elementos, cumple la condición de volumen. Cuenta con 21.742 elementos, cumple la condición.
2. La variedad de los datos se pone de manifiesto debido al gran número de variables que el database presenta. Concretamente la base de datos está compuesta con 58 columnas o variables, cada cual diferente entre sí, proporcionando un gran número de datos diferentes al modelo. **(Ver anexo)**
3. Al tratarse de una base de datos estática (descargada de internet), en vez de una generada por diversas fuentes de procesos explicados anteriormente como sensores, dispositivos móviles o redes sociales, es difícil verificar su velocidad. Sin embargo, dicha obtención de datos según las variables podría automatizarse para lograr dicho objetivo.

Figura 14. Visión de R Studio sobre database

```
# A tibble: 6 × 58
   ...1    id title        subti…¹ sq_mt…² sq_mt…³ n_rooms n_bat…⁴ n_flo…⁵ sq_mt…⁶
  <dbl> <dbl> <chr>        <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1     0 21742 Piso en ven… San Cr…      64      60       2       1      NA      NA
2     1 21741 Piso en ven… Los Án…      70      NA       3       1      NA      NA
3     2 21740 Piso en ven… San An…      94      54       2       2      NA      NA
4     3 21739 Piso en ven… San An…      64      NA       2       1      NA      NA
5     4 21738 Piso en ven… Los Ro…     108      90       2       2      NA      NA
6     5 21737 Piso en ven… San An…     126     114       4       2      NA      NA
# … with 48 more variables: latitude <lgl>, longitude <lgl>, raw_address <chr>,
```

(*Elaboración propia a partir de Kaggle)*

5.3.2 Limpieza y preparación de los datos.

Como hemos visto, la base de datos presenta un gran volumen y variabilidad, pero no todos estos datos son útiles para nuestro proyecto**.** La limpieza de los datos adquiere gran relevancia en este contexto, ya que es fundamental asegurar que los datos sean consistentes, precisos y confiables antes de proceder al análisis. En este proceso, se tenderá a reducir la presencia de datos nulos, datos repetidos, y se intentará mantener una uniformidad en la escritura de cada variable y así generar una base de datos general (Villamar, et al, 2022).

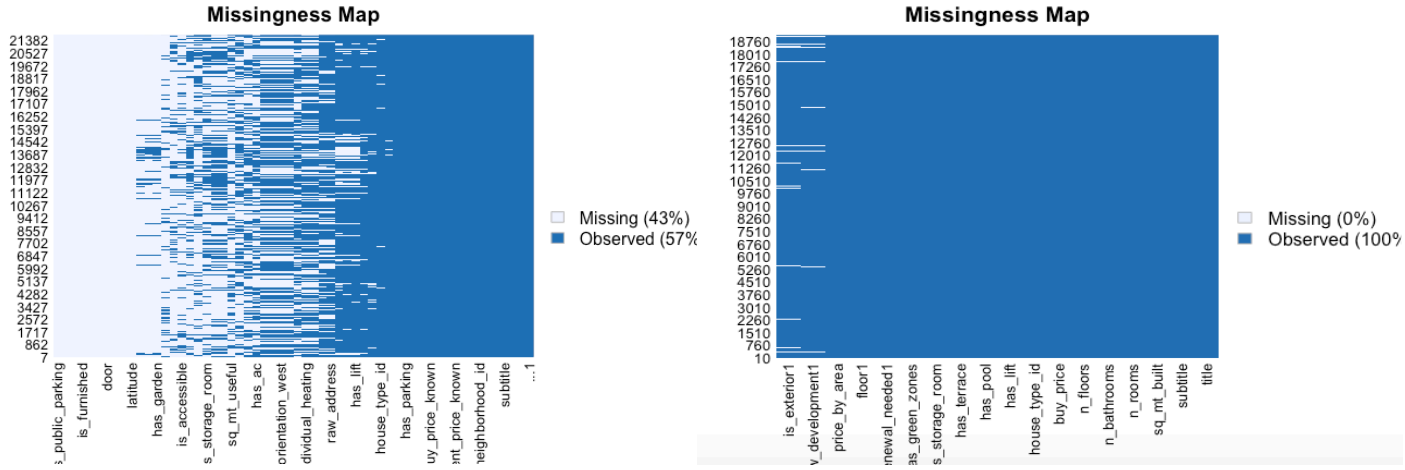
En primer lugar, hay ciertas **variables que tratan de explicar el mismo suceso**. Por ejemplo, existe una variable "is_floor_under" que se refiere a si un piso es un "bajo". En la variable "floor", se contempla en que planta está situado el piso, por lo que ya está dando la información sobre si es un bajo o no.

Por otro lado, para el **tratamiento de datos nulos**, dependiendo de la forma en la que éstos se presenten, se actuará de una forma u otra. Las acciones que se han llevado a cabo son las siguientes:

- Para el preprocesamiento de datos nulos o valores NA, en primer lugar, notamos que existen varias variables que presentan en su **totalidad valores nulos**, es decir, están vacías, por lo que habría que proceder a su eliminación ya que no aportan nada. En la siguiente futura vemos en azul claro los valores nulos del modelo.
- Otras variables que presentan valores nulos como has green zones, todos los valores que sí contempla están calificados como "TRUE" por lo que los valores nulos se calificaron como "FALSE". Esto tiene sentido ya que en los anuncios publicitados el vendedor hace manifiesto de que el piso cuenta con zonas verdes cuando la tiene, pero lo oculta cuando no lo tiene. Se incluyen aquí varias variables como "has pool", "has terrace", "is new development" y "has storage room".
- Para la variable "house type id", que determina el tipo de casa, se ha presupuesto que los valores nulos corresponden con pisos, ya que, para la venta de un piso, se especificaría su tipo de inmueble si este presentara características no comunes.

- Las variables que tratan de identificar a un piso con un número, normalmente seguidas de **"id"** pueden distorsionar el modelo, por lo que son también eliminadas.
- En las variables "n bathroom" y "sq mt built", debido a su gran correlación con el precio de la vivienda y al irrelevante número de valores nulos, se ha procedido a la **eliminación de estos elementos** o filas.

Figura 15. Visualización valores nulos



(Elaboración propia a partir de Kaggle)

De esta manera, hemos conseguido dejar la cifra de material no observado hasta prácticamente el 0% y seguimos teniendo una representación clara y sólida de la población total.

Por último, las **variables que comprendían valores booleanos**, con los valores "TRUE" o "FALSE", se han convertido a numéricos donde, 1=TRUE y 0=FALSE. De esta manera podremos realizar análisis matemáticos de estas variables. Algunas de estas variables son "has parking" o "has pool".

5.3.3 Análisis exploratorio de datos

En el presente apartado, se abordará el análisis exploratorio de datos realizado como parte de este trabajo. El análisis exploratorio de datos desempeña un papel fundamental en la comprensión inicial de un conjunto de datos y permite obtener una visión general de su estructura, patrones y relaciones subyacentes. Esta etapa es crucial

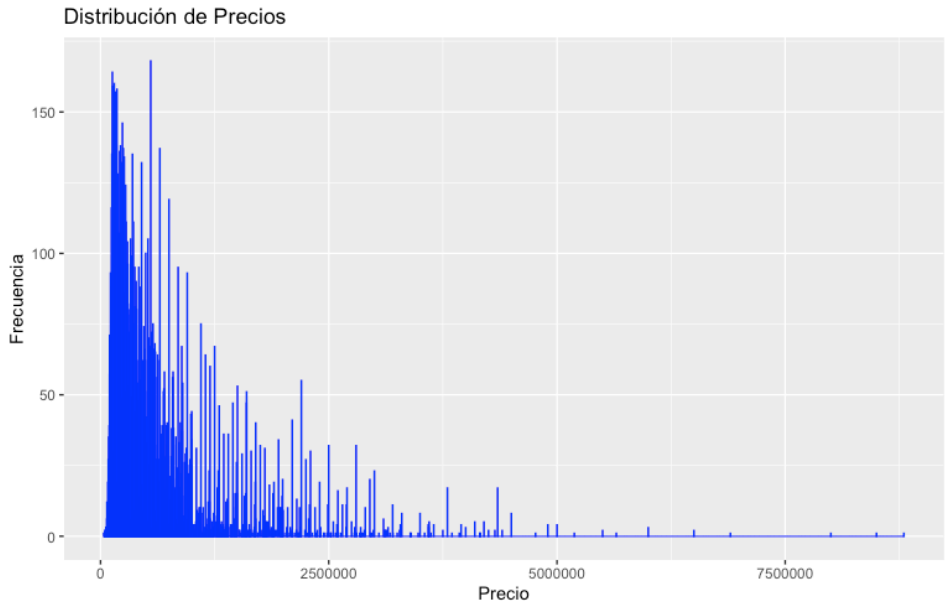
para identificar tendencias, detectar valores atípicos, explorar la distribución de las variables y establecer las bases para análisis posteriores más avanzados.

El objetivo principal del análisis exploratorio de datos es obtener una comprensión profunda de los datos y extraer información relevante que permita tomar decisiones fundamentadas.

a) Análisis de la variable precio (Buy price)

En primer lugar, se abordará el análisis de la variable precio (buy price), que desempeña un papel central en este trabajo. El objetivo principal de este análisis es obtener una comprensión detallada de los factores que influyen en el precio y su relación con otras variables relevantes. Los precios se encuentran muy concentrados, la mayoría en torno a la media, que se sitúa en 544.192 euros (**Ver figura 16**). El precio mínimo que podemos encontrar es de 36.000, que se corresponde a un piso en el barrio de Usera. Mientras que el piso más caro es un ático que cuesta 8.800.000 euros y se encuentra en la zona de el Retiro. El tercer percentil corresponde a un precio de 649.000 euros, por lo que por encima de este precio podemos catalogar las casas como caras si hablamos de Madrid en su conjunto.

Figura 16. Distribución variable precios



(Elaboración propia a partir de Kaggle)

Para hacer más preciso el análisis, es conveniente analizar las diferentes zonas de la comunidad. Visualizar los precios en su conjunto puede llevar a equivocaciones sobre

pensamientos de que casas pueden ser más o menos caras de lo normal. Ver los precios del metro cuadrado en las distintas zonas puede ser una medida más precisa para contrastar pensamientos sobre un inmueble. En la siguiente **tabla**, a la izquierda se pueden ver los barrios más caros de la capital según el database, es recoletos, con un precio de 8.771 euros por metro cuadrado, al contrario, a la derecha la zona más barata de Madrid. El retiro contiene el piso más caro de la base de datos, lo cual es un outlier, con lo cual su precio por metro cuadrado se puede haber visto afectado por dicho inmueble y realmente ser menor por lo general. Por otro lado, el barrio más barato de Madrid es San Cristóbal, con un precio por metro cuadrado de 1.517,8 euros. Dicho indicador es también una medida de status del barrio, de esta manera, los habitantes de los barrios más caros suelen tener un mayor status social que los habitantes de los barrios más baratos.

Por otro lado, se aprecia como la mayoría de los barrios más caros de Madrid se encuentran cercanos al centro: Recoletos, Chamberí, etc., por lo que existe una tendencia de la gente de mayores ingresos a demandar casas en el centro de la ciudad, aumentando así su precio, en vez de a los alrededores como puede ser Orcasitas o Usera. Recordemos la vinculación emocional entre las viviendas y las familias, vivir en el centro de la ciudad también hace la vida más cómoda: trayectos más cortos, mayor oferta de restauración, mayor seguridad, ….

Tabla 4. Precio/m2 barrios Madrid

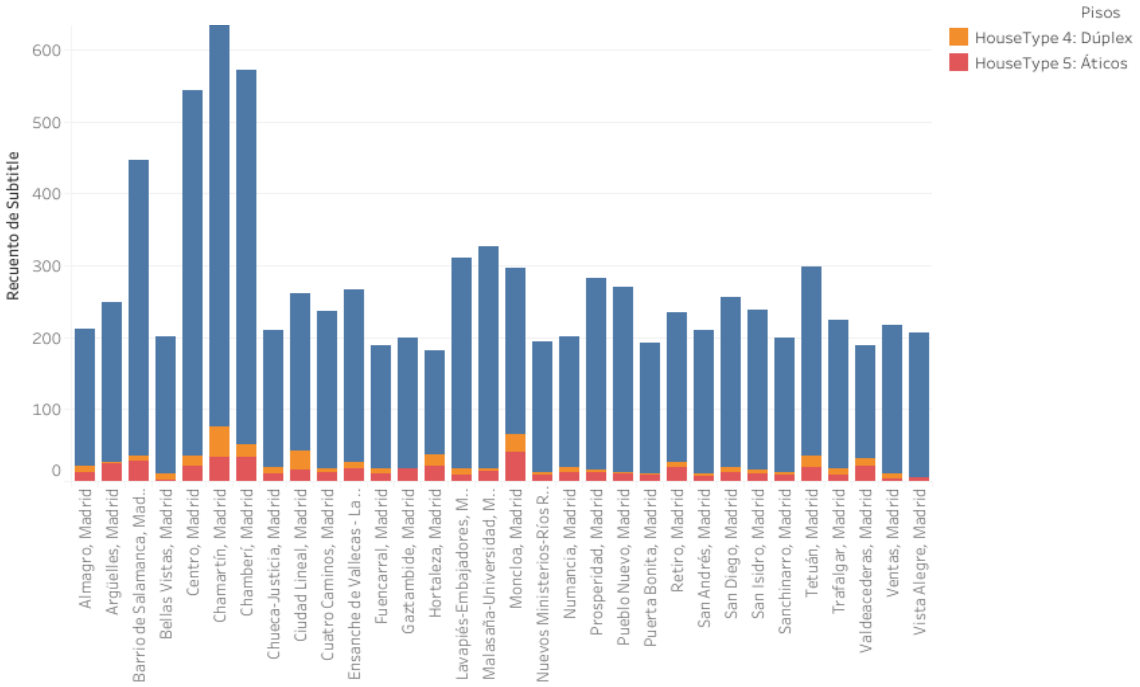
Barrios más caros	Precio/m2	Barrios más baratos	Precio/m23
Recoletos, Madrid	8770,6	San Cristóbal, Madrid	1517,8
Castellana, Madrid	7214,1	Entrevías, Madrid	1565,5
Barrio de Salamanca, Madrid	6980,3	Villaverde, Madrid	1726,1
Almagro, Madrid	6637,5	Los Ángeles, Madrid	1880,0
Jerónimos, Madrid	6488,9	San Andrés, Madrid	1888,5
Trafalgar, Madrid	6455	Los Ángeles, Madrid	1880,0
El Viso, Madrid	6332,3	Orcasitas, Madrid	1933,1
Goya, Madrid	6313,4	Portazgo, Madrid	1943,6
Chueca-Justicia, Madrid	6110,8	Puente de Vallecas, Madrid	1968,5
Lista, Madrid	6019,5	Usera, Madrid	1977,8
El retiro	5917,7	12 de Octubre-Orcasur, Madrid	1992,0

(Elaboración propia a partir de datos de Kaggle)

Se puede realizar todavía un análisis más profundo sobre la actividad económica de un barrio y elaborar **un recuento de la oferta inmobiliaria** de cada zona. En la siguiente **figura** se muestran los barrios de Madrid con mayor número de pisos en venta,

acompañado del tipo de piso que se vende en dichos barrios. Como vemos, las zonas donde existe mayor movimiento de mercado son las zonas céntricas, zonas que como se ha podido ver, suelen ser más caras.

Figura 17. Oferta innmobiliaria por barrio



(Elaboración propia a partir de datos de Kaggle)

b) Análisis de variables

Una vez analizado el precio en conjunto y según los barrios, de la cual se obtiene una visión general de la ciudad de Madrid, se procede a realizar análisis un poco más sofisticado, para los que se requiere herramientas analíticas más avanzadas.

En primer lugar, se realiza un análisis estadístico de las variables numéricas de nuestro modelo. En cuanto a las **variables más simples** (Media, mediana, moda, desv. típica, máx y mín) parece no existir ninguna anomalía, salvo la elevada cifra de la desviación típica del precio, superior incluso a la media de la variable. Recordemos que el precio por área en Madrid oscila mucho dependiendo del barrio, por lo que, al analizar Madrid en su conjunto, tiene sentido que la desviación típica de la ciudad presente altos valores. El que haya pisos sin habitaciones también llama la atención, aunque puede tratarse de estudios con salón, cocina y dormitorio unidos. Igualmente, saltan a la luz varios *outliers,* como

residencias de 14 baños, 15 habitaciones, o un precio/m2 de 18.888 euros. También se observa el piso más caro ya nombrado anteriormente, con un precio de 8.800.000 euros.

Se analizan ahora datos más complejos, vemos que éstas **(curtosis y coeficiente de asimetría**) en varias variables si presentan valores altos y más llamativos. La curtosis describe la forma de la distribución de una variable en relación con una distribución normal. Una curtosis tan alta, como es en el caso de las variables: "sq mt built" y "buy Price", indican que existe una excesiva concentración de los valores con respecto a la media, aparte de colas muy pesadas. Este tipo de distribución se le conoce como leptocúrtica (Aguilar, 2019). Por su parte, un coeficiente de asimetría de tres indica una asimetría o sesgo positivo en la distribución de los datos. La asimetría se refiere a la medida de la falta de simetría en una distribución estadística. Cuando el coeficiente de asimetría es positivo, significa que la cola derecha de la distribución es más larga o pesada que la cola izquierda. Esto indica que hay valores extremadamente altos que se alejan de la media, lo que resulta en una distribución desplazada hacia la derecha (Pérez, 2010, p. 32). Recordemos la figura de la distribución del precio (**Figura 16),** en ésta vemos como se cumple lo descrito, alta concentración de elementos agrupados alrededor de la media y cola más larga hacia valores más grandes (alta curtosis), hacia la derecha (asimetría positiva).

Tabla 5. Datos estadísticos variables numéricas

Variable estadística	sq_mt_built	n_rooms	n_bathrooms	buy_price	floor1	Price_by_area
Media	120,89	2,79	1,83	544.192,04	2,61	4.060,88
Mediana	94	3	2	340.000	2	3.800
Desv. Típica	84,53	1,27	1,02	600.427,75	2,05	1.926,38
Moda	70	3	1	550.000	1	5.000
Máx	894	15	14	8.800.000	9	18.889
Min	15	0	1	36.000	0	447
Curtosis	9	3	4	17	0	2
Asimetria	2,47	0,75	1,70	3,32	0,85	1,13

(Elaboración propia a partir de datos de Kaggle).

Seguidamente, a través de la **matriz de correlación** (**Tabla 6),** se puede observar las correlaciones que mantienen las variables numéricas entre sí. Las correlaciones explican como varía una variable con respecto a otra, es decir, sirve para medir la relación entre dos variables. Una correlación cercana a uno expresa que las variables están muy directamente correlacionadas, pues, cuando una aumenta, la otra también, prácticamente

en la misma medida. Por el contrario, una correlación negativa implica una relación inversa entre los objetos estudiados (Lahura, 2003).

Como se puede apreciar en la tabla, la variable que tiene una mayor relación con nuestro objetivo *precio* es el número m2 construidos, seguido del número de baños. Resulta curioso que el número de baños explica mejor que el número de cuartos el precio de una casa, aunque el número de habitaciones también guarda una correlación media con nuestra variable objetivo. El precio por metro cuadrado (Price by area) también mantiene una correlación alta con el precio. Si se analizan las correlaciones de las viviendas de obra nueva (is new development), se puede comprobar como esta mantiene una relación inversa con los metros construidos y el número de habitaciones, lo que se traduce en una tendencia por parte de los constructores a elaborar casas de menor tamaño y número de habitaciones. La correlación de ésta con el precio es prácticamente cero, la gente no está dispuesta a pagar más porque la casa esté sin estrenar. Las otras variables dicotómicas del modelo guardan una relación directa con el precio, es decir, si el inmueble tiene alguna de estas características, se comercializará a un mayor precio.

Tabla 6. Matriz de correlaciones

MATRIZ DE CORRELACIÓN														
	sq_mt_built	n_rooms	n_bathrooms	buy_price	has_lift	has_pool	has_terrace	has_storage_room	has_green_zones	is_exterior1	is_new_development1	is_renewal_needed1	floor1	price_b
sq_mt_built	1	0,6849971	0,810739706	0,865034594	0,31025134	0,1471003	0,152708926	0,314770525	0,131357769	0,193375851	-0,009227015	0,120433042	0,147291	0,32719
n_rooms		1	0,637911061	0,511269414	0,18155271	0,0332734	0,187963831	0,175288538	0,077351639	0,183335889	-0,053801289	0,226496662	0,1459884	0,05979
n_bathrooms			1	0,746030764	0,35159959	0,2062872	0,113683963	0,314883156	0,170354798	0,173543722	0,040053516	0,037407029	0,1254623	0,39699
buy_price				1	0,31877488	0,0859971	0,061801007	0,254648571	0,043822228	0,137513392	0,002603409	0,070540084	0,1578639	0,65842
has_lift					1	0,2601576	0,039601017	0,259831041	0,21126278	0,11204658	0,129807539	-0,014571339	0,2182775	0,37777
has_pool						1	0,115436852	0,395446527	0,616482194	0,138479131	0,27736463	-0,157947717	0,0663118	0,06151
has_terrace							1	0,112460213	0,142044505	0,148428346	0,079629425	0,047662956	0,1665632	-0,10606
has_storage_room								1	0,318689195	0,161650709	0,142727404	-0,063103606	0,063596	0,10615
has_green_zones									1	0,13692585	0,027725404	-0,103614634	0,0459115	-0,02657
is_exterior1										1	0,066990911	-0,009095977	0,0621587	-0,02527
is_new_development1											1	-0,115841018	-0,011221	0,06401
is_renewal_needed1												1	0,0349724	-0,04895
floor1													1	0,14897
														1

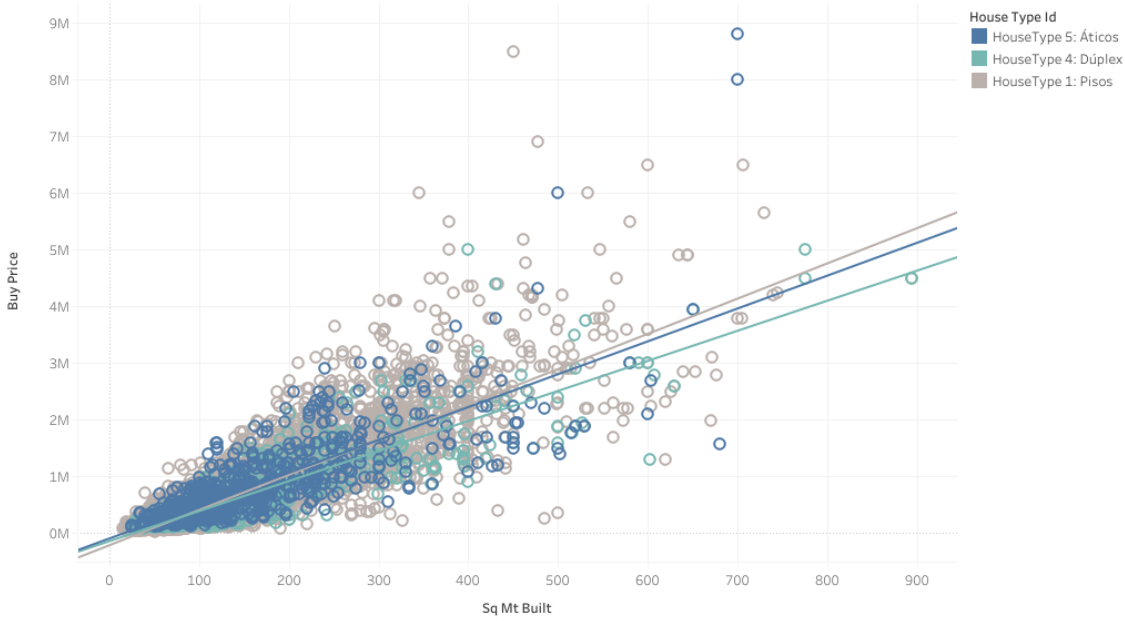
(*Elaboración propia a partir de datos de Kaggle)*

En este apartado, se explorarán detalladamente las relaciones existentes entre el precio de los inmuebles y otras variables relevantes que requieren una mayor atención en nuestro estudio. Comprender la influencia de estas variables en el precio de los inmuebles es fundamental para obtener una visión más completa y precisa de los factores que impactan en el mercado inmobiliario de Madrid.

La primera relación para analizar es la relación entre el **precio de la casa, los metros cuadrados construidos y el tipo de casa** (**Ver figura 18**). En esta figura, se contempla en una primera instancia la relación positiva entre el precio y los metros cuadrados que

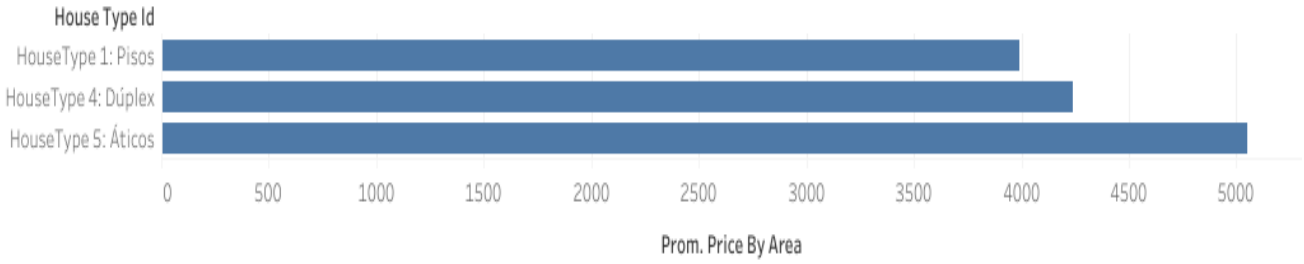
esta tiene. A más grande, más cara, es por eso que nuestro modelo de regresión tiene pendiente positiva en los tres modelos de casas. Asimismo, en cuanto al tamaño no hay sorpresa, los dúplex son las viviendas con mayor metro cuadrado. En dicha figura también parece que el precio no guarda relación con el tipo de inmueble, pues en la línea de tendencia tiene incluso menos pendiente que los pisos, pero no es así. Si analizamos la **figura 19** observamos como es más cara la superficie en caso de áticos, seguidos de pisos y por último dúplex. Los dúplex suelen encontrarse en las afueras de la ciudad, por consiguiente, es normal que el precio relativo sea menor. Es común en la sociedad actual la elección entre una casa de grandes dimensiones a las afueras, o una casa más menuda cercana a la poli.

Figura 18. Relación precio, metros cuadrados y tipo de casa



(*Elaboración propia a partir de datos de Kaggle)*

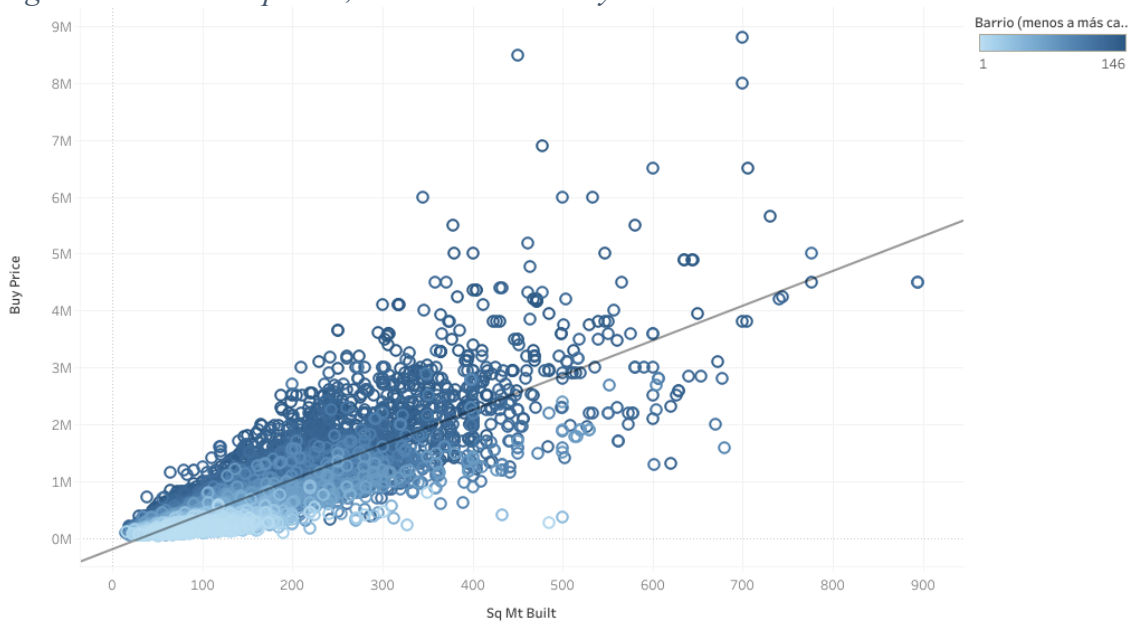
Figura 19. Precio/m2 de los tipos de casa



(Elaboración propia a partir de datos de Kaggle)

La siguiente relación para estudiar es la existente entre el **precio, los metros construidos y el barrio. (Ver figura 20)** El color más oscuro hace referencia a los barrios que tienen un precio del metro cuadrado más caro, en el extremo opuesto se sitúa el azul claro. Se pueden apreciar dos grandes conclusiones en el análisis conjunto de las tres variables. La primera conclusión es que los colores oscuros quedan encima de los claros, lo cual parece obvio ya que la variable precio se encuentra en el eje "y". Además, ningún inmueble de un barrio más económico, aunque tenga mayor área, tiene un precio parecido a los pisos de un barrio más costoso. Un detalle más curioso es que, a medida que aumenta los metros construidos, van desapareciendo los puntos claros, permaneciendo solo los oscuros en el gráfico. Esto lleva a la conclusión de que, por líneas generales, en los barrios más baratos, el tamaño de las viviendas es menor.

Figura 20. Relación precio, metros cuadrados y barrio



(Elaboración propia a partir de datos de Kaggle)

5.3.4 Machine learning.

En la sección de análisis exploratorio, se lleva a cabo un estudio exhaustivo de los datos, mediante la identificación de patrones, tendencias y posibles relaciones entre las variables. Este análisis preliminar permite comprender mejor la naturaleza de los datos y tomar decisiones informadas sobre el enfoque de modelado.

Una vez completada la etapa de análisis exploratorio, se pasa a la sección de machine learning, en la cual se aplican técnicas y algoritmos para construir modelos predictivos o descriptivos. Estos modelos se basan en los patrones y relaciones descubiertos durante el análisis exploratorio de datos. Utilizando métodos de entrenamiento y validación, se ajustan los modelos a los datos disponibles y se evalúa su rendimiento en la predicción o clasificación de nuevos datos (Janiest, et al, 2021).

A) Modelos no supervisados.

Se elaborará en este apartado una clusterización de los datos para obtener una visión pragmática de la clasificación de los inmuebles en Madrid.

La clusterización es una técnica de aprendizaje no supervisado en la que se agrupan objetos o casos similares en conjuntos llamados clústeres. El objetivo principal de la clusterización es encontrar patrones o estructuras inherentes en los datos sin tener una variable objetivo o etiquetas predefinidas (Fung, 2001). El algoritmo de clusterización busca maximizar la similitud intra-clúster y minimizar la similitud inter-clúster, es decir, busca formar grupos compactos y bien diferenciados, permitiendo la división de un gran conjunto de datos en grupos heterogéneos (Jankowska, et al, 2019).

Una vez comprendida la técnica de clusterización, se procede a realizarlo con los datos recopilados para el proyecto. El primer paso consiste en preparar los datos, pues se necesita que todas las variables sean numéricas. Recordemos que, en la limpieza de datos, las variables dicotómicas pasaron de figurar como “TRUE” y “FALSE” a “1” y “0” respectivamente. El problema principal se encuentra en las variables “House Type Id” y “Subtitle” (barrios). Dentro de la variable House Type Id se encontraban los tipos de inmuebles que componen el dataset. En su mayoría son pisos, aunque también hay dúplex y áticos. Éstos se han recogido con los valores “1”, “2” y “3” respectivamente. Más complicada es la reclasificación de los barrios, pues aparte de ser una cifra mucha más elevada (146 barrios) no es clara la forma o patrón que esta debe seguir. Finalmente, para convertir dicha variable en numérica, se ha optado por ordenar los barrios según cueste el metro cuadrado en la zona, siendo el número uno San Cristóbal y el número ciento cuarenta y seis Recoletos.

Tabla 7. Conversión de house type y subtitle en numéricas

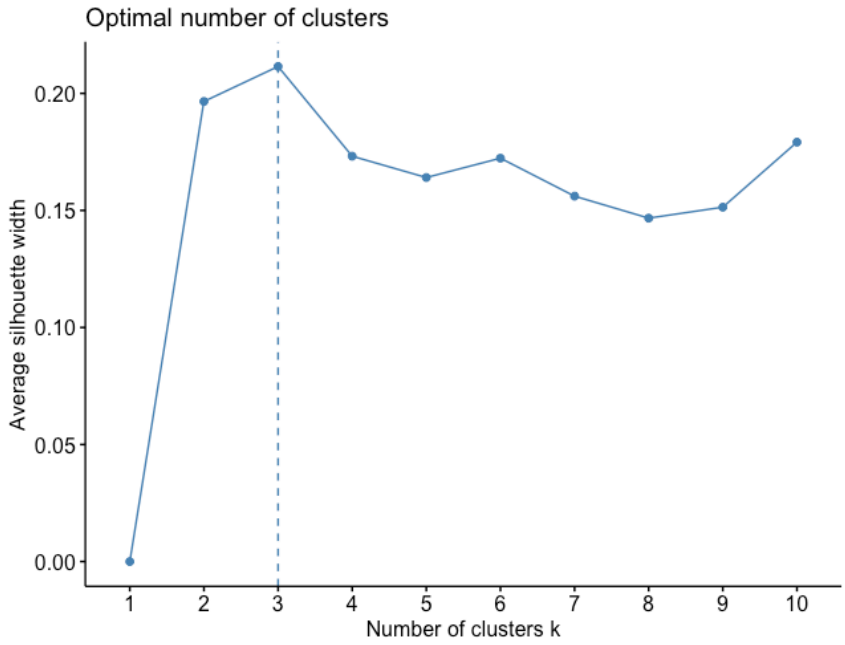
house_type_id	subtitle_numerico	m2_price_subtitle
1	28	2.229.011
1	107	4.500.592
1	11	2.017.210
1	126	5.440.485
1	31	2.281.014
1	110	4.567.754
1	100	4.301.417
1	18	2.133.443

(Elaboración propia a partir de datos de Kaggle)

Dicho esto, y una vez preparada la base de datos, éstos se escalan para verificar que todas las variables tengan un impacto significativo en el proceso de clusterización. Al escalar los datos, se ajusta la varianza y la magnitud de las variables para que tengan un rango similar, lo que evita que una variable con valores grandes o dominantes influya de manera desproporcionada en el resultado del clustering.

Tras finalizar este paso, se averigua el número "k" (agrupamientos) óptimo para nuestra base de datos. Para determinarlo existen principalmente 2 métodos: Método Elbow y Coeficiente de Silhouette. El método del codo es a veces difícil de analizar y se basa en la interpretación para definir el valor de "K" en K-medias en base a su gráfico (Saputra y Oswari, 2020), lo que da lugar a la subjetividad. Es por eso que se utilizará el método de Silhouette para la determinación del número de clusters, ya que otorga objetividad. Tras realizar la función del coeficiente de Silhouette obtenemos que el número de "K" óptimo para nuestro proyecto es de 3 **(Ver figura21).**

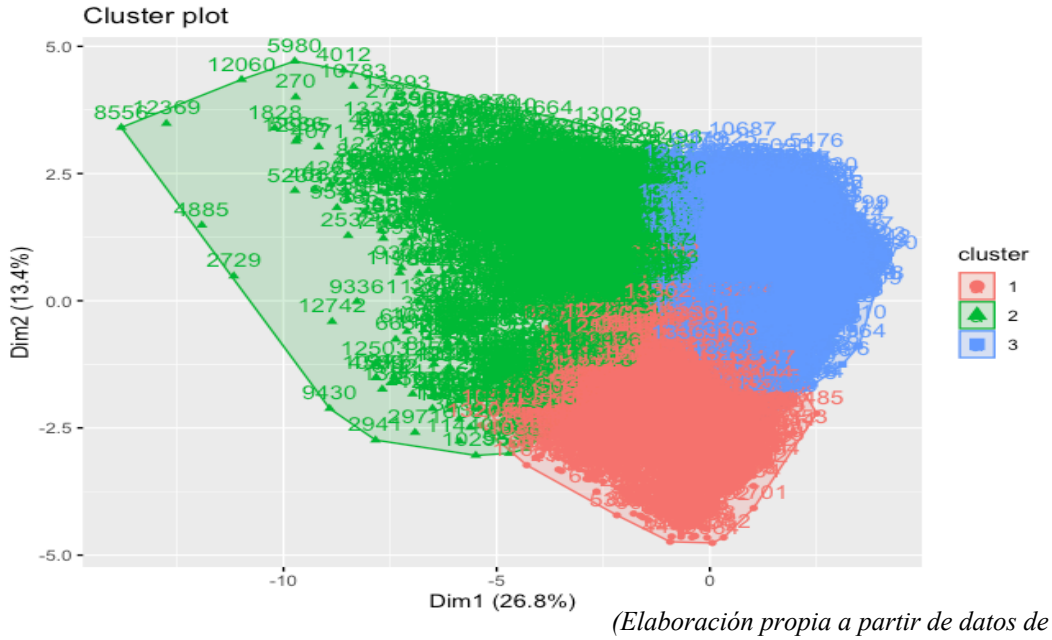
Figura 21. Número óptimo de K



(Elaboración propia a partir de datos de Kaggle)

Una vez el número de clústeres a obtener es conocido, se realiza la función K-means, un algoritmo para agrupar datos en "K" clústeres y basar su distancia a su centroide (Saputra y Oswari, 2020). Tras obtener los resultados, obtenemos tres grandes clústeres: el grupo 1, formado por 8.195 elementos; grupo 2 por 2.458; y cluster 3 por 2.731 elementos (Ver figura X).

Figura 22.Cluster plot



(Elaboración propia a partir de datos de Kaggle

Una vez obtenidos los tres subgrupos, se examinan las características de cada uno de ellos. Los clústeres resultantes deben ser interpretados como grupos de objetos que comparten características similares o pertenecen a la misma categoría. De esta manera, permite que sea utilizada para descubrir patrones ocultos en los datos, identificar grupos homogéneos o heterogéneos y segmentar el mercado.

Entre los resultados obtenidos, se analizan ahora las características de cada conjunto (**Ver tabla 8).**

El Clúster 1, con el mayor número de elementos, contiene la menor media en la variable subtitle numérico (barrios), los cuales estaban agrupados de más barato a más caro. Por consecuencia y como habíamos visto anteriormente (relación barrios más baratos - casas más pequeñas), también ocupa el último puesto en la variable sq mt built. La mayor diferencia con los demás clústeres se encuentra en el precio, pues su media es de 291.450,2 euros frente a 1.482.504,2 euros en el clúster 2 y 496.838 euros en el clúster

El número 2, como hemos visto, cuenta con la mayor media de precios. En él se agrupan los domicilios más grandes y con mayor número de baños y habitaciones. También presenta, con gran diferencia, las zonas más caras de la ciudad. Cabe destacar también la alta media en los barrios de este clúster (126, 9).

El **último clúster,** parece a primera vista más difícil de analizar. El primer punto para destacar es que recoge una lista de vecindarios parecida al primer grupo, aunque el precio y precio/m2, aparte de la superficie construida es mayor en éste. Por otro lado, si algo también tiene significativo es la alta media en las variables dicotómicas y tipo de casa, donde llega prácticamente a igualar al anterior grupo. Por su parte, en las variables dicotómicas revela que agrupa propiedades que otros inmuebles prácticamente no tienen: piscina, zonas verdes, terraza y trastero. Además, el número de pisos de obra nueva es significativamente alto en este clúster.

Tabla 8. Media de clústeres por variable

Cluster	sq_mt_built	n_rooms	n_bathrooms	buy_price	house_type_id	has_lift	has_pool	has_terrace
1	82,70	2,38	1,34	291.450,18	1,07	0,62	0,01	0,35
2	252,71	4,31	3,32	1.482.504,50	1,24	0,98	0,12	0,50
3	122,64	2,69	2,00	496.837,89	1,24	0,99	0,82	0,54

Cluster	has_storage_room	has_green_zones	is_exterior1	is_new_development1	is_renewal_needed1	floor1	price_by_area	subtitle_numerico
1	0,14	0,05	0,80	0,01	0,19	2,30	3.548,72	73,21
2	0,54	0,14	0,96	0,01	0,33	3,34	5.895,10	126,87
3	0,74	0,74	0,96	0,24	0,05	2,92	4.066,27	76,33

(Elaboración propia a partir de datos de Kaggle)

Visto esto, se concluye que gracias a la clusterización hemos conseguido agrupar elementos de características similares en tres grandes grupos:

- **Clúster 1**: Reúne viviendas baratas, de menores superficies y que se sitúan en los barrios más baratos. En resumen, recoge las viviendas (en su mayoría pisos) más baratas de los barrios más baratos.
- **Clúster 2**: Contiene principalmente los inmuebles de los barrios más costosos. Dichas viviendas son las más grandes y caras. Se podría conocer este grupo de casas como el sector más exclusivo de la ciudad, con una media de precio de más de un millón de euros, al alcance de pocos.
- **Clúster 3**: Agrupa viviendas de barrios más modestos, parecidos a los del clúster 1, pero que cuentan con mayores superficies, presentan mejores condiciones y calidad de vida (trastero, piscina, …) y por consecuencia tienen un precio más alto. Recoge también multitud de chalets y áticos de estas localidades, poco representativos en el clúster 1.

B) Modelos supervisados

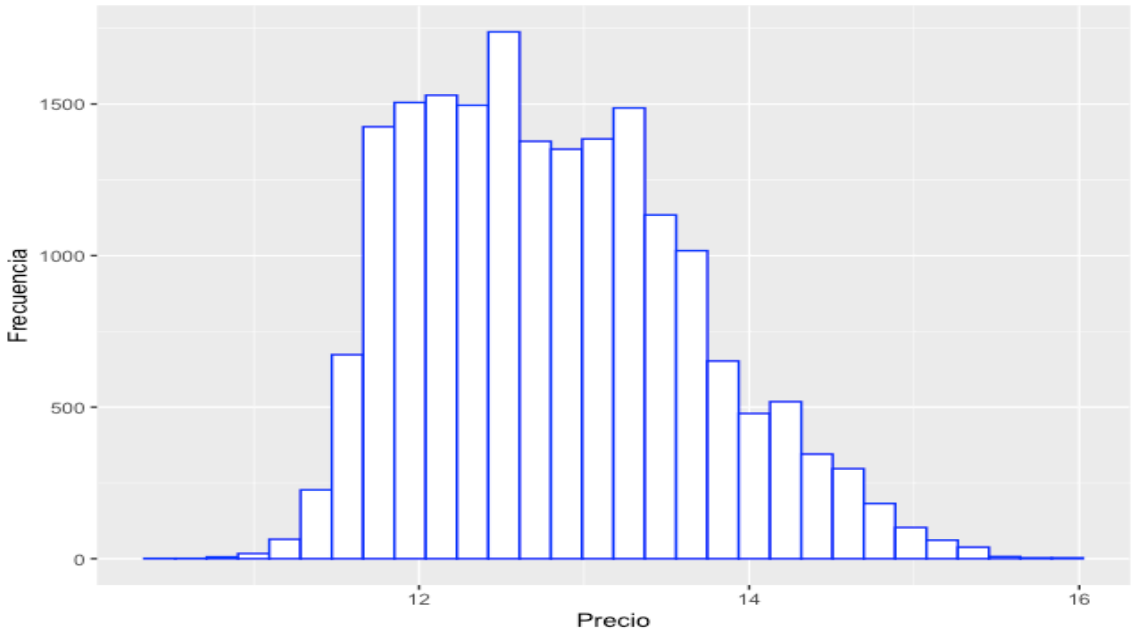
Una vez aplicada la técnica de modelo no supervisado, se procede a desarrollar varios modelos supervisados para realizar predicciones sobre nuestra variable objetivo, el precio de los inmuebles.

i) Preparación de los datos

El primer paso es comprobar que los datos están bien preparados. A lo largo de este apartado, se han modificado las variables de tal manera que ya todas están exentas de valores nulos y todas son numéricas. Sin el cumplimiento de estos requisitos, no se podría haber llevado a cabo el análisis clúster.

Sin embargo, todavía falta un pequeño matiz anterior a la elaboración del modelo. Si recordamos la distribución del precio (**Figura 16**), este presenta una distribución logarítmica. Es común convertir la variable objetivo en una distribución normal antes de elaborar un modelo. Esto se hace para cumplir con el supuesto de normalidad en muchos modelos estadísticos y para mejorar la interpretación de los resultados (Espinoza, 2018).

Figura 23. Nueva distribución de precio



(Elaboración propia a partir de datos de Kaggle)

Para la realización de los modelos supervisados, que entrenan con datos históricos para realizar sus predicciones, se utilizará la herramienta analítica R Studio. Ésta es una potente herramienta de minería de datos y aprendizaje automático, que sirve para elaborar modelos predictivos. R Studio proporciona una interfaz gráfica intuitiva que permite cargar datos, realizar análisis exploratorio, seleccionar variables relevantes y construir modelos predictivos de manera eficiente y efectiva.

Antes de empezar con la elaboración del modelo y aunque ya se han comentado la mayoría, se observa cuáles son las variables predictivas y cual la variable objetivo que componen nuestro modelo.

Tabla 9. Variables predictoras y objetivo

Variable	Función	Tipo
sq_mt_built	Predictor	Numérica Continua
n_rooms	Predictor	Numerica discreta
n_bathrooms	Predictor	Numérica discreta
house_type_id	Predictor	Categórica
has_lift	Predictor	Categórica
has_pool	Predictor	Categórica
has_terrace	Predictor	Categórica
has_storage_room	Predictor	Categórica
has_green_zones	Predictor	Categórica
is_exterior1	Predictor	Categórica
is_new_developmer	Predictor	Categórica
is_renewal_needed:	Predictor	Categórica
floor1	Predictor	Numérica discreta
price_by_area	Predictor	Numérica Continua
subtitle_numerico	Predictor	Categórica
buy_price	Objetivo	Numérica Continua

(Elaboración propia a partir de datos de Kaggle)

Siguiendo con la preparación de los datos y para no caer en errores de *overfitting* se llevarán a cabo diversas acciones de control. El *overfitting* consiste en la sobre explicación del modelo, esto es, que el modelo se ajusta tanto a los datos de entrenamiento que pierde la capacidad de generar predicciones en los datos de prueba, lo cual es nuestro objetivo (Cohen y Jensen, 1997). Para evitar el problema de sobreajuste se llevará a cabo la **eliminación de variables** que realmente no guardan relación con nuestra variable objetivo. Consiste en determinar las variables cuya supresión obtendría la menor (o ninguna) diferencia en el resultado del modelo y, a continuación, eliminar esa característica respectivamente. En la siguiente figura se presenta el ranking de las variables de mayor y menor relación con nuestra variable objetivo. Esto corrobora las hipótesis planteadas en el análisis exploratorio de datos. Vemos que las variables que ocupan las últimas posiciones prácticamente no explican nada del modelo, por lo que se llevará a su eliminación, quedándonos con las 10 primeras para ello (desde house type id hasta sq mt built).

Tabla 10. Ranking de predictoras según su importancia

		#	Uni...eg.	RReliefF
1	sq_mt_built		35017.858	0.071
2	n_bathrooms		27591.956	0.047
3	subtitle_numerico		23544.566	0.100
4	price_by_area		23058.577	0.056
5	n_rooms		7777.205	0.057
6	has_lift		6863.819	0.000
7	has_storage_room		1974.215	0.000
8	floor1		869.455	0.070
9	has_pool		723.446	0.014
10	house_type_id		544.645	0.007
11	is_exterior1		489.842	-0.000
12	has_green_zones		333.663	0.020
13	has_terrace		87.108	0.000
14	is_new_development1		58.762	0.000
15	is_renewal_needed1		46.283	0.001

(Elaboración propia a partir de datos de Kaggle)

ii) Elaboración de modelos supervisados

La preparación de datos es una etapa fundamental en el proceso de elaboración de un modelo. Antes de poder construir y entrenar un modelo, es necesario asegurarse de que los datos estén limpios, estructurados y en el formato adecuado. Una vez preparados, se procede a su elaboración en la plataforma R Studio. En este caso, se elaborarán dos modelos basados en árboles de decisión: Modelo *gradient boosting* y modelo *random forest.*

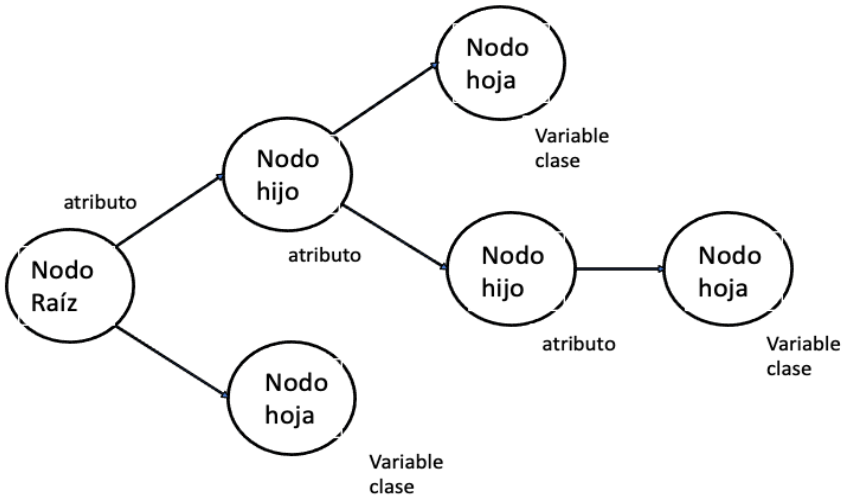
Árbol de decisión: consiste en un algoritmo de aprendizaje automático que utiliza un enfoque basado en reglas para tomar decisiones (Quinlan, 1993). Este árbol consta de nodos, hojas y ramas, y se utiliza para clasificar los datos **(Ver figura 24).** El nodo principal o raíz del árbol es el atributo que se utiliza como punto de partida para clasificar los datos. A medida que se avanza en el árbol, cada nodo interno representa una pregunta

sobre un atributo específico del problema. Estas preguntas se utilizan para tomar decisiones y seguir recorriendo el árbol hasta llegar a una hoja que representa una clase o categoría de clasificación. (Martínez, et al, 2009). Este tipo de modelos se puede aplicar al problema de predicción de precios de viviendas porque permite identificar las características más relevantes para determinar el precio de una vivienda, como el tamaño, la ubicación, características, etc.

Gradient boosting: es un algoritmo de aprendizaje automático que combina múltiples modelos débiles para construir un modelo más fuerte y preciso. Funciona de manera iterativa, donde cada nuevo modelo se enfoca en corregir los errores del modelo anterior (Cheng, Guestrin, 2016). Puede aplicarse al problema de predicción de precios de viviendas debido a su capacidad para capturar relaciones no lineales y complejas, así como su capacidad de aprender patrones sutiles y realizar predicciones precisas utilizando una combinación de múltiples modelos.

Random forest: Un modelo de Random Forest es un algoritmo de aprendizaje automático que combina múltiples árboles de decisión para obtener predicciones más precisas y estables. Cada árbol en el conjunto se entrena con una muestra aleatoria del conjunto de datos y una selección aleatoria de características. Posteriormente, las predicciones de cada árbol se combinan mediante votación o promedio para obtener la predicción final (Cheng, et al, 2004).

Figura 24. Árbol de decisión

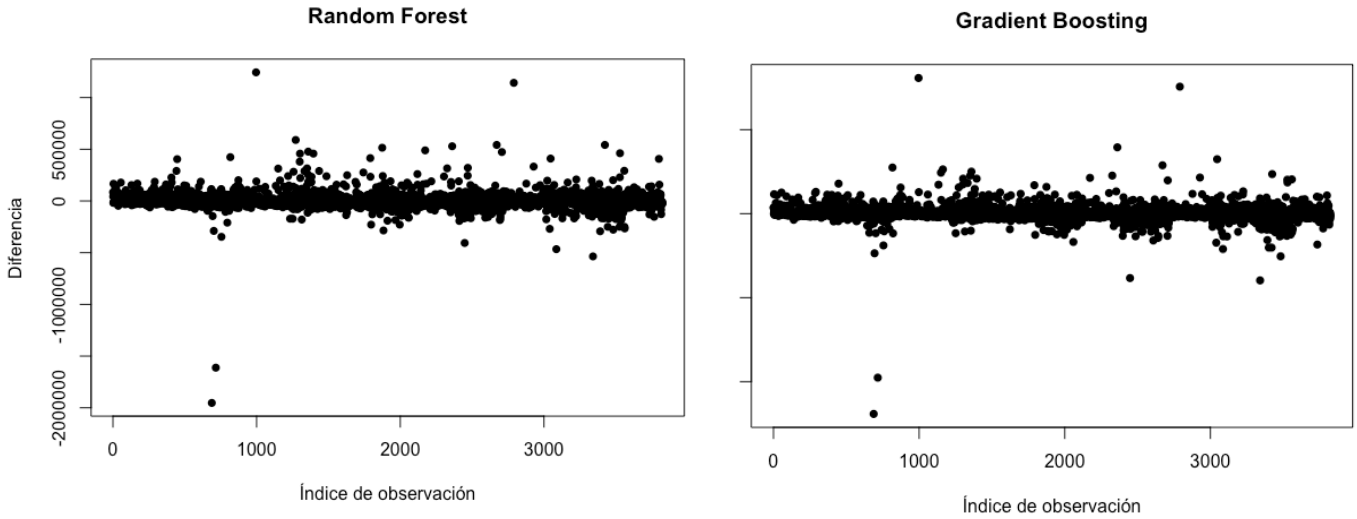


(Elaboración propia a partir de datos de Martínez, et al, 2009)

Conocidos los modelos que se van a implementar, estos **se entrenan**. El primer paso es determinar los datos que se van a usar para entrenar el modelo y los datos de prueba. El porcentaje de datos de entrenamiento debe ser mayor que el de prueba. Entrenar el modelo se refiere al proceso de ajustar los parámetros o coeficientes del modelo utilizando un conjunto de datos de entrenamiento. En otras palabras, implica alimentar al modelo con datos de entrada conocidos y las correspondientes salidas o etiquetas para que el modelo aprenda a hacer predicciones o clasificaciones correctas. Es importante en esta fase elegir un número óptimo de árboles para evitar el *overfitting.* Este problema afecta a la precisión de los algoritmos, que deja de mejorar después de cierto punto, o incluso empeora debido al aprendizaje con ruido.

Al finalizar el entrenamiento del modelo y elaborar las predicciones, se someten los modelos a un testeo para que puedan ser evaluados y comparados. Los resultados brutos obtenidos en la fase de testeo de los modelos se pueden apreciar en la **siguiente figura.** La evaluación de un modelo de aprendizaje automático es un proceso esencial para comprender su rendimiento y su capacidad para generalizar a nuevos datos. Permite obtener una medida objetiva de cómo se comporta el modelo en la práctica y si es capaz de hacer predicciones precisas y confiables (Rama, et al, 2021). Como se aprecia en la figura, ambos modelos presentan figuras parecidas y difieren prácticamente en las predicciones de los mismos elementos, lo cual pueden ser outliers de nuestro modelo que no hemos eliminado.

Figura 25. Gráficos de diferencias en la predicción

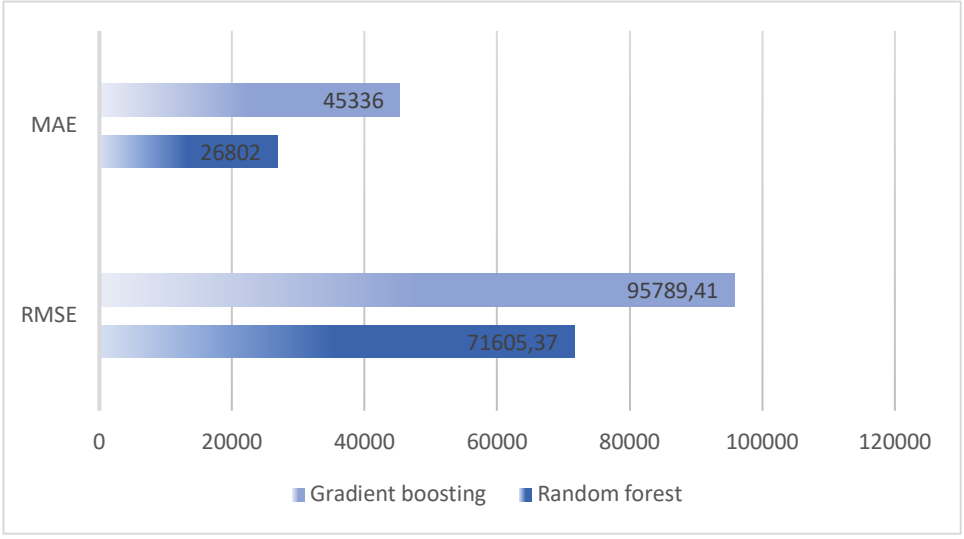


(Elaboración propia a partir de datos de Kaggle)

La comparación a través de la diferencia bruta entre las predicciones y los datos de prueba a partir de gráficas puede dar lugar a la subjetividad. Una medida que permite comparar distintos modelos objetivamente es el **RMSE** (*Root Mean Square Error*), que representa la raíz cuadrada del promedio de los errores al cuadrado entre los valores predichos por el modelo y los valores reales; y **MAE** (*Mean Absolute Error),* una medida de error que calcula la media de las diferencias absolutas entre los valores predichos por un modelo y los valores reales. Una limitación de este último es que no considera la dirección de la equivocación, distorsionando la diferencia (Chai, Draxler, 2014). Ambas medidas muestran el error medio, por lo que un menor valor en la medida implicará un modelo más preciso.

Si se observa la **siguiente figura**, se ve claramente que el modelo de *random forest* ha obtenido mejores resultados en la predicción de inmuebles, teniendo una media de error menor en las dos variables. De igual manera, los modelos obtenidos no se pueden interpretar para su uso comercial debido a su gran desviación sobre los valores, pues 71.600 euros de error en un piso barato constituyen un gran error. El modelo sí serviría para dar una idea general de sobre qué precio se podría partir para empezar una negociación. Al analizar la **figura 26**, se aprecia que la mayoría de las predicciones se sitúan alrededor de su valor real, sin embargo, hay algunos que difieren en gran medida de estos (diferencia mayor a 1.000.000 de euros), por lo que dichos valores no comunes en la predicción distorsionan estas medidas.

Figura 26. RMSE y MAE



(Elaboración propia a partir de datos de Kaggle)

Al finalizar el modelo y una vez se obtienen los resultados, se inicia la fase de perfeccionamiento del modelo. Existen diversas técnicas que consiguen que un modelo ya ejecutado mejore su rendimiento y precisión. Por ejemplo, se podría llevar a cabo un Ensamblado de modelos, que consiste en combinar las predicciones de varios modelos individuales para obtener una predicción más precisa y robusta (Dietterich, 2000); o se podría someter al modelo a un proceso de ajuste de hiperparámetros, mediante la búsqueda en cuadrícula o optimización bayesiana (Bergstra, Bengio, 2012).

5.3.5 Conclusiones.

La oferta inmobiliaria de Madrid es inmensa y variada, aunque se puede agrupar en tres grandes grupos heterogéneos: casas exclusivas de barrios caros, inmuebles baratos de barrios baratos y viviendas de barrios baratos pero que presentan mejores características.

El análisis exploratorio realizado sobre el precio de la vivienda en Madrid ha revelado que existen variables clave que influyen significativamente en el precio. Entre las variables más relevantes se encuentran los metros construidos de la vivienda, el número de baños y el barrio donde se encuentra. Asimismo, en los barrios más baratos, las casas tienden a ser más pequeñas, al igual que la vivienda de obra nueva, donde los constructores tienden actualmente a elaborar viviendas de menor superficie. La actividad económica inmobiliaria tiene mayor participación en el centro de la ciudad, donde existe una mayor oferta y donde se encuentran también los barrios más costosos por m2. Por su parte, los áticos también cuentan con un mayor precio por área, siendo los dúplex los más baratos, ya que, aunque sean más grandes, se encuentran fuera de la ciudad, ocasionando la duda en las familias sobre si vivir en una casa más modesta en el centro de la ciudad buscando la comodidad, o una casa de mayores dimensiones a las afueras, ambas por un precio parecido.

El modelo desarrollado para predecir los precios de los inmuebles en la ciudad de Madrid ha sido sometido a un proceso de preparación de datos, selección de variables relevantes y entrenamiento utilizando diferentes algoritmos de aprendizaje automático. Durante este proceso, se han aplicado técnicas como la normalización de variables y y eliminación de muchas de ellas para evitar *overfitting.*

Se han utilizado varios algoritmos de aprendizaje automático, incluyendo árboles de decisión: gradient boosting, y random forest, con el objetivo de encontrar el modelo que mejor se ajuste a los datos y que ofrezca las mejores predicciones.

Además, se ha realizado una evaluación exhaustiva del modelo utilizando métricas como el RMSE (Root Mean Squared Error) y el MAE (Mean Absolute Error) para medir la precisión de las predicciones, concluyendo que el *random forest* se adaptaba mejor a nuestra necesidad ya que tenía un menor error en ambas medidas. Aunque no alcanzaron una precisión óptima, los modelos proporcionaron una estimación inicial del valor de las propiedades. Estos modelos nos brindan una referencia para tener una idea aproximada del precio al que podrían venderse o comprarse las viviendas en la ciudad.

6. Conclusión.

En conclusión, este trabajo ha puesto de manifiesto la creciente importancia del Big Data en el sector inmobiliario, especialmente en el contexto de la revolución Proptech y el cambio hacia la digitalización. A medida que la industria se adentra en una nueva era impulsada por la tecnología, el análisis de datos se ha convertido en un factor clave para la toma de decisiones informadas y la maximización de oportunidades.

En primer lugar, la revolución Proptech ha revolucionado la forma en que se realizan las transacciones inmobiliarias y se gestionan los activos. Los startups y las empresas tecnológicas han introducido innovaciones disruptivas que han transformado los procesos tradicionales, desde la búsqueda de propiedades hasta la firma de contratos y la gestión de la propiedad. Este cambio ha generado una gran cantidad de datos en todas las etapas del ciclo de vida inmobiliario.

El Big Data, por su parte, proporciona las herramientas y técnicas necesarias para aprovechar al máximo esta avalancha de información. El análisis de grandes volúmenes de datos permite identificar patrones, tendencias y insights que de otra manera serían difíciles de percibir. Esto permite a los profesionales del sector inmobiliario tomar decisiones más fundamentadas, basadas en datos concretos, en lugar de depender únicamente de la intuición o la experiencia.

En el sector inmobiliario, el uso estratégico del Big Data puede tener un impacto significativo. Permite comprender mejor las preferencias y necesidades de los clientes, optimizar la fijación de precios de las propiedades, predecir tendencias del mercado y realizar evaluaciones de riesgo más precisas. Además, el análisis exploratorio de datos y la construcción de modelos predictivos, como se realizó en el caso práctico presentado, brindan una visión más profunda de los factores que influyen en el precio de las viviendas y permiten realizar pronósticos más precisos.

La digitalización del sector inmobiliario y la adopción del Big Data representan una oportunidad para mejorar la eficiencia operativa, la personalización de servicios y la satisfacción del cliente. Al aprovechar la información disponible y utilizarla de manera estratégica, las empresas inmobiliarias pueden ofrecer experiencias más personalizadas, optimizar sus procesos internos y mejorar su posición competitiva en el mercado.

En resumen, el Big Data juega un papel fundamental en la revolución Proptech y en la transformación del sector inmobiliario hacia la digitalización. Proporciona una base sólida para la toma de decisiones basada en datos, mejora la comprensión del mercado y

las preferencias de los clientes, y ofrece oportunidades para optimizar la eficiencia y brindar servicios más personalizados. En un entorno en constante evolución, aquellos actores del sector inmobiliario que se adapten y aprovechen el potencial del Big Data estarán mejor posicionados para el éxito a largo plazo.

7. Bibliografía

Aguilar, L. J. (2016). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.

Aguilar, L. E. B. (2019). Diferencias en la estimación del coeficiente de curtosis en diferentes softwares estadísticos. e-Agronegocios, 5(2). https://revistas.tec.ac.cr/index.php/eagronegocios/article/view/4456/4953

Alamer, M and Almaiah, M.A. "Cybersecurity in Smart City: A Systematic Mapping Study," *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan, 2021, pp. 719-724, doi: 10.1109/ICIT52682.2021.9491123. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9491123

Barkham, R., Mellott, D., & Raaum, C. (2, feb 2022). "2021 Global Investment Volume Hits Record Level", en *Cbre.es*, https://www.cbre.es/insights/briefs/2021-global-investment-volume-hits-record-level.

Base de datos: Madrid Real Estate Market https://www.kaggle.com/datasets/mirbektoktogaraev/madrid-real-estate-market

Baum, A. (2017). PropTech 3.0: the future of real estate. University of Oxford https://www.sbs.ox.ac.uk/sites/default/files/2018-07/PropTech3.0.pdf

Batrinca, B., Treleaven, P.C. (2015). "Social media analytics: a survey of techniques, tools and platforms", en *AI & Soc* 30 pp. 89–116, en https://link.springer.com/article/10.1007/s00146-014-0549-4.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Bouskela, M., Casseb, M., Bassi, S., De Luca, C., & Facchina, M. (2016). *La ruta hacia las smart cities: Migrando de una gestión tradicional a la ciudad inteligente*. en https://books.google.es/books?hl=es&lr=&id=TdB3DwAAQBAJ&oi=fnd&pg=PA20&dq=la+ruta+hacia+smart+cities&ots=fSDPA5DcpR&sig=zRElbTrlvdlaazlUvKKU3jKgHQw#v=onepage&q=la%20ruta%20hacia%20smart%20cities&f=false.

Braesemann, F., & Baum, A. (2020). PropTech: Turning real estate into a data-driven market?. Available at SSRN 3607238. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3607238

Braun, T., Fung, B. C., Iqbal, F., & Shah, B. (2018). "Security and privacy challenges in smart cities" en Sustainable *cities and society* Vol. 39, pp. 499-507, https://www.sciencedirect.com/science/article/abs/pii/S2210670717310272?casa_token=TxcUn3d4etUAAAAA:cYf7dE_UuyBtuF86XqPqrr8KnBO64KnLWoV1tF-_N05Motls9HtVG-c-vCym1TX_D9ohJUpj9A.

Cappiello, A. (2020). The technological disruption of insurance industry: A review. *International Journal of Business and Social Science*, *11*(1), 1-11. https://www.researchgate.net/profile/Antonella-Cappiello

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534. https://gmd.copernicus.org/preprints/7/1525/2014/gmdd-7-1525-2014.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Cohen, P. R., & Jensen, D. (1997, January). Overfitting explained. In Sixth International Workshop on Artificial Intelligence and Statistics (pp. 115-122). PMLR. http://proceedings.mlr.press/r1/cohen97a/cohen97a.pdf

Daniel Glez-Peña and others. "Web scraping technologies in an API world", en *Briefings in Bioinformatics*, Vol. 15, Issue 5, 2014, pp. 788–79, d https://academic.oup.com/bib/article/15/5/788/2422275.

Danyang Du, Aihua Li, Lingling Zhang (2014). "Survey on the Applications of Big Data in Chinese Real Estate Enterprise", en *Procedia Computer Science*, Vol. 30, pp. 24-33, https://www.sciencedirect.com/science/article/pii/S1877050914005547.

De Mauro, A., Greco, M., y Grimaldi, M. (2015, February). "What is big data? A consensual definition and a review of key research topics", en *AIP conference proceedings* (Vol. 1644, No. 1, pp. 97-104). American Institute of Physics.

Donner, H., Eriksson, K., & Steep, M. (2018). "Digital cities: Real estate development driven by big data", en https://www.researchgate.net/profile/Herman-Donner/publication/325253311_Digital_Cities_Real_Estate_Development_Driven_by_Big_Data/links/5cdbac94458515712eac2286/Digital-Cities-Real-Estate-Development-Driven-by-Big-Data.pdf

Eckerson, W. W. (2007). Predictive analytics. Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, 1, 1-36. http://download.101com.com/pub/tdwi/files/pa_report_q107_f.pdf

"El ecosistema PropTech en España. (2021)", en Cbre.es, https://www.cbre.es/insights/books/informe-proptech/el-ecosistema-proptech-en-espana.

Espinoza Freire, E. E. (2018). Las variables y su operacionalización en la investigación educativa. Parte I. Conrado, 14, 39-49. http://scielo.sld.cu/pdf/rc/v14s1/1990-8644-rc-14-s1-39.pdf

Fung, G. (2001). A comprehensive overview of basic clustering algorithms. https://sites.cs.ucsb.edu/~veronika/MAE/clustering_overview_2001.pdf

Hernández, C., Puigdevall, A., López, G. (2021). *La revolución PropTech: una reflexión sobre la transformación e innovación en el mercado inmobiliario*. Gestion 2000.

Itay Goldstein and others (2021). "Big Data in Finance", en *The Review of Financial Studies*, Vol. 34, Issue 7, pp. 3213–3225, https://academic.oup.com/rfs/article/34/7/3213/6210658.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.

Jankowska, B., Götz, M., & Główka, C. (2017). Intra-cluster cooperation enhancing SMEs' competitiveness-the role of cluster organisations in Poland. Investigaciones Regionales-Journal of Regional Research, (39), 195-214.

Kaile Zhou, Chao Fu, Shanlin Yang (2016). "Big data driven smart energy management: From big data to big insights", en *Renewable and Sustainable Energy Reviews*, Vol. 56, pp. 215-225, https://www.sciencedirect.com/science/article/pii/S1364032115013179.

Khatoun R. y Zeadally S. "Cybersecurity and Privacy Solutions in Smart Cities," en *IEEE Communications Magazine*, vol. 55, Nº3, pp. 51-59, 2017, https://ieeexplore.ieee.org/document/.

Kok, Nils & Koponen, Eija-Leena & Martínez-Barbosa, Carmen Adriana. (2017). "Big Data in Real Estate? From Manual Appraisal to Automated Valuation" en *The Journal of Portfolio Management*, Vol. 43, pp. 202-211, https://www.pm-research.com/content/iijpormgmt/43/6/202.

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. https://www.researchgate.net/profile/Vlad-Krotov

Lahura, E. (2003). El coeficiente de correlación y correlaciones espúreas (Vol. 218). Pontificia Universidad Católica del Perú, Departmento de Economía. https://gc.scalahed.com/recursos/files/r161r/w24082w/S7_01.pdf

Lies, J. (2019). "Marketing intelligence and big data: Digital marketing techniques on their way to becoming social engineering techniques in marketing", https://www.researchgate.net/profile/Abderahman-Rejeb/publication/339630258_Potential_of_Big_Data_for_Marketing_A_Literature_Re

view/links/5e5d67a24585152ce8010820/Potential-of-Big-Data-for-Marketing-A-Literature-Review.pdf.

Lingqiang Kong, Zhifeng Liu, Jianguo Wu. "A systematic review of big data-based urban sustainability research: State-of-the-science and future directions", en *Journal of Cleaner Production*, Vol. 273, 2020, 123142, https://doi.org/10.1016/j.jclepro.2020.123142.

Ma, S. (2021). *Technological obsolescence* (No. w29504). National Bureau of Economic Research. https://www.nber.org/papers/w29504

Maltby, D. (2011, October). Big data analytics. In 74th Annual Meeting of the Associationfor Information Science and Technology (ASIST) (pp. 1-6) https://d1wqtxts1xzle7.cloudfront.net

Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. D. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 9(2), 19-24. http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Martin-Sanchez, F., & Verspoor, K. (2014). "Big data in medicine is driving big changes ", en *Yearbook of Medical Informatics*, 9(01), pp. 14–20, https://www.thieme-connect.de/products/ejournals/abstract/10.15265/IY-2014-0020.

Masdeu, J. (2020). "Rastreados por el móvil para frenar al coronavirus" en *LaVanguardia.com,* https://www.lavanguardia.com/vida/20200327/48109834703/coronavirus-tecnologia-rastreo-movil-telecomunicaciones-datos.html.

Maté Jiménez, C. (2014). "Big data. Un nuevo paradigma de análisis de datos", en *Revista Anales de Mecánica y electricidad,* Nº7, Vol. 90, pp. 10-16.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Houghton Mifflin Harcourt.

Mayer-Schönberger, V, Ingelsson, E. (2018). "Big Data and medicine: a big deal?", en *Review Symposium* Vol. 283, Issue 5 , pp. 418– 429, https://onlinelibrary.wiley.com/doi/epdf/10.1111/joim.12721.

McAfee, A., & Brynjolfsson, E. (2012). "Big data: the management revolution", en *Harvard Business Review*, *90*(10), 60–66, 68, 128, https://hbr.org/2012/10/big-data-the-management-revolution.

Mei, Y., Gao, L., Zhang, J. et al. "Valuing urban air quality: a hedonic price analysis in Beijing, China", en *Environ Sci Pollut Res* Vol. 27, 2020, pp. 1373–1385, https://link.springer.com/article/10.1007/s11356-019-06874-5.

Moro Visconti, R. (2020). "Fintech valuation", en *SSRN Electronic Journal,* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3533869.

Munawar, H.S.; Ullah, F.; Qayyum, S.; Shahzad. D. (2022). "Big Data in Construction: Current Applications and Future Opportunities" en Big *Data Cogn. Comput*, Vol. 6, Issue 1, https://doi.org/10.3390/bdcc6010018.

Olayonwa, G.O. "Adaptation of Sensory Evaluation Technique to Collection of Real Estate Data" en *International Review of Social Sciences and Humanities* Vol. 4, Nº. 2, 2013, pp. 13-21, https://web.archive.org/web/20180409221614id_/http:/www.irssh.com/yahoo_site_admin/assets/docs/2_IRSSH-425-V4N2.44203203.pdf.

Oluwunmi, A. O., Role, B. A., Akinwale, O. M., Oladayo, O. P., & Afolabi, T. O. (2019). "Big data and real estate: A review of literature", en Journal *of physics. Conference series*, Vol. 1378, Issue 3, https://iopscience.iop.org/article/10.1088/1742-6596/1378/3/032015.

Pérez, R. (2010). Nociones básicas de estadística. Rigoberto Perez.

Pons, L. (2023, febrero 18). "Software CRM: Qué son, cómo funcionan y cuáles usar en 2023", en *Holded*, https://www.holded.com/es/blog/software-crm.

Puyol, J., Excedencia, E., & Abogado. (2014). "UNA APROXIMACIÓN A BIG DATA" en *Revista de Derecho Uned* nº14, http://e-spacio.uned.es/fez/eserv/bibliuned:rduned-2014-14-7150/Documento.pdf.

Quinlan, J. R. (1993). Program for machine learning. C4. 5. https://link.springer.com/article/10.1007/BF00993309

Rama-Maneiro, E., Vidal, J., & Lama, M. (2021). Deep learning for predictive business process monitoring: Review and benchmark. IEEE Transactions on Services Computing. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9667311

Rawat, D. B., & Ghafoor, K. Z. (Eds.). (2018). *Smart cities cybersecurity and privacy.* Elsevier. https://books.google.es/books?hl=es&lr=&id=xnN9DwAAQBAJ&oi=fnd&pg=PP1&dq=cybersecurity+in+smart+cities&ots=miaG4hDxn2&sig=WKwb_Ypbc1dpSDHrAUchDnjC-x8#v=onepage&q=cybersecurity%20in%20smart%20cities&f=false.

Rousseau, C. (2021). "ConTech: la tecnología que busca modernizar la construcción", en *Linkedin.com*, https://www.linkedin.com/pulse/contech-la-tecnolog%C3%ADa-que-busca-modernizar-carlos-rousseau/?originalSubdomain=es

Sanders, N. R. (2016). "How to use big data to drive your supply chain", en California Management Review, Vol. 58, Issue 3, pp. 26-48, https://journals.sagepub.com/doi/pdf/10.1525/cmr.2016.58.3.26?casa_token=ljobyk4Dr

JMAAAAA:SlNFfylpWINivBjmWhXGqD7Mppi1P3Ru4lQpoc1aVyxDW27n1uKvlUqBnQ6ubBc9cFlo5tk9IWSs.

Saputra, D. M., & Oswari, L. D. (2020, May). Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In Sriwijaya International *Conference on Information Technology and Its Applications* (SICONIAN 2019) (pp. 341-346). Atlantis Press.

Siniak, N, et al, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 869 062041https://iopscience.iop.org/article/10.1088/1757-899X/869/6/062041/pdf

Tabares, L. F., & Hernández, J. F. (2014) "Big Data Analytics: Oportunidades, Retos y Tendencias", en *Scalahed.com,* https://gc.scalahed.com/recursos/files/r161r/w25569w/Big%20Data%20Analytics.pdf.

Urrea Ramos (2022). "Big Data, el nuevo adn de las organizaciones", https://repository.unimilitar.edu.co/bitstream/handle/10654/43627/UrreaRamosNicolas2022.pdf?sequence=1&isAllowed=y.

Villamar, V. P., Morillo, P. A. L., & Arevalo, F. G. (2022). Análisis estadístico general de mediciones de parámetros de la red celular en Quito. Revista de Investigación en Tecnologías de la Información: RITI, 10(21), 120-131. https://dialnet.unirioja.es/servlet/articulo?codigo=8652911

Wadhwani, Khushboo & Wang, Dr. (2017). Big Data Challenges and Solutions. https://www.researchgate.net/publication/313819009_Big_Data_Challenges_and_Solutions

Wei C, Fu M, Wang L, Yang H, Tang F, Xiong Y. "The Research Development of Hedonic Price Model-Based Real Estate Appraisal in the Era of Big Data" en *Land* Vol. 11, Issue 3, 2022, https://www.mdpi.com/2073-445X/11/3/334.

Xiaojie Xu, Yun Zhang, House price forecasting with neural networks, Intelligent Systems with Applications, Volume 12, 2021, 200052, ISSN 2667-3053, https://doi.org/10.1016/j.iswa.2021.200052.

Yebra del Puerto, G. I. (2018). ¿ Qué sabe internet de nosotros?.

https://openaccess.uoc.edu/bitstream/10609/81252/6/yebraTFM0618memoria.pdf

Yin, C., Xiong, Z., Chen, H., Wang, J., Cooper, D., & David, B. (2015). "A literature survey on smart cities", en *Sci. China Inf. Sci.*, Vol. 58, Issue10, pp. 1-18, https://www.researchgate.net/profile.

Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf

Yovanof, G.S., Hazapis, G.N (2019). "An Architectural Framework and Enabling Wireless Technologies for Digital Cities & Intelligent Urban Environments" en Wireless *Pers Commun* Vol. 49, 2009, pp. 445–463, https://link.springer.com/article/10.1007/s11277-009-9693-4.

Yu M, Yang C, Li Y. (2018). "Big Data in Natural Disaster Management: A Review" en *Geosciences* Vol. 8, Issue 5, 2018, 165, https://www.mdpi.com/2076-3263/8/5/165.

Zhao, Bo (2017). "Web scraping", en *Springer International Publishing AG, Encyclopedia of big data* (2017), https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf.

Zhao M, Cheng W, Zhou C, Li M, Wang N, Liu Q. "GDP Spatialization and Economic Differences in South China Based on NPP-VIIRS Nighttime Light Imagery" en Remote Sensing, Vol. 9, Issue 7, 2017, https://www.mdpi.com/2072-4292/9/7/673.

(2022), PwC.com: *Global PropTech Confidence Index* https://www.pwc.com/us/en/industries/financial-services/library/pdf/pwc-year-end-2022-global-proptech-confidence-index.pdf

8. Anexo

Variables base de datos de Kaggle:
https://www.kaggle.com/datasets/mirbektoktogaraev/madrid-real-estate-market

Variable	Explicación	Variable2	Explicación2
id	identificación del inmueble	is_new_development	obra nueva
title	titulo	built_year	año de construcción
subtitle	barrio	has_central_heating	calefacción central
sq_mt_built	metros construidos	has_individual_heating	calefacción invidual
sq_mt_useful	metros aprovechables	are_pets_allowed	permite mascotas
n_rooms	número de habitaciones	has_ac	tiene ac
n_bathrooms	número de baños	has_fitted_wardrobes	armarios empotrados
n_floors	número de plantas	has_lift	ascensor
sq_mt_allotment	metros adjudicados	is_exterior	exterior
latitude	latitud	has_garden	jardín
longitude	longitud	has_pool	piscina
raw_address	Dirección	has_terrace	terraza
is_exact_address_hidden	Se da la dirección exacta	has_balcony	balcón
street_name	nombre de la calle	has_storage_room	trastero
street_number	número	is_furnished	está amueblado
portal	portal	is_kitchen_equipped	cocina equipada
floor	planta	is_accessible	accesible
is_floor_under	es un bajo	has_green_zones	tiene zonas verdes
door	puerta	energy_certificate	certificado energético
neighborhood_id	identificación del barrio	has_parking	tiene parking
operation	operación	has_private_parking	parking privado
rent_price	precio de alquier	has_public_parking	parking publico
rent_price_by_area	precio de alquiler/m2	is_parking_included_in_price	parking incluido en precio
is_rent_price_known	se conoce el precio de alquiler	parking_price	precio del parking
buy_price	precio	is_orientation_north	orientación norte
buy_price_by_area	precio/m2	is_orientation_west	orientación oeste
is_buy_price_known	se conoce el precio de alquiler	is_orientation_south	orientaión sur
house_type_id	tipo de casa	is_orientation_east	orientación este
is_renewal_needed	necesita reparación		