COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

EL NUEVO SECTOR INMOBILIARIO (PropTech): caminando hacia el Big Data

Aplicación práctica en una inmobiliaria familiar de la ciudad de Valladolid

Clave: 201506051

MADRID | Junio, 2020

RESUMEN

El PropTech comienza a convertirse en la nueva realidad que impera en el sector inmobiliario. El espectacular auge de empresas inmobiliarias cuyos cimientos empiezan a construirse desde la innovación tecnológica augura un cambio sin parangón. Con estas raíces, sobresalen por encima del resto los nuevos modelos de negocio en torno al Big Data inmobiliario. Modelos de negocio para los que el tradicional *modus operandi* de la industria es ya parte del pasado. Se realiza un análisis teórico, exhaustivo y profundo de este nuevo entorno y de su reciente evolución, que incluye una revisión de la muy reciente literatura existente. A continuación, mediante un ejercicio práctico de análisis de la oferta residencial de la ciudad de Valladolid se pondrá de manifiesto cómo estas innovaciones pueden dictaminar un antes y un después en el sector inmobiliario; y se concluirá con la idea de que la adopción de la innovación tecnológica no solo es recomendable por las ventajas que entraña a nivel de resultados, sino sobre todo que su implantación, a día de hoy, está lejos de ser un imposible para cualquier empresa que opere en el sector como puede ser el caso de una inmobiliaria tradicional o familiar.

PALABRAS CLAVE

PropTech, Big Data inmobiliario, Sector Inmobiliario, Machine Learning, tecnología, análisis de datos, modelos predictivos, empresa familiar

ABSTRACT

PropTech is starting to become the new reality in the real estate sector. The spectacular rise of real estate companies whose foundations are beginning to be built on technological innovation augurs well for an unparalleled change. With these roots, the new business models around the real estate Big Data stand out from the rest. Business models for which the traditional modus operandi of the industry is now a thing of the past. A theoretical, exhaustive and in-depth analysis of this new environment and its recent evolution is carried out, including a review of the very recent existing literature. Then, through a practical exercise of analysis of the residential offer of the city of Valladolid, it will be shown how these innovations can determine a before and after in the real estate sector; and it will be concluded with the idea that the adoption of the technological innovation is not only recommendable by the advantages that it entails at level of results, but mainly that its implantation, nowadays, is far from being an impossible one for any company that operates in the sector as it can be the case of a traditional or familiar real estate agency.

KEYWORDS

PropTech, Big Data, Real Estate, Machine Learning, technology, data analysis, predictive models, family business

ABREVIATURAS

API	Apliccation Programing Interface

EDA	Exploratory Data Analysis
M^2	Metro cuadrado
ML	Machine Learning
MSE	Mean Square Error
NA	Valor Ausente
RICS	Royal Institution of Chartered Surveyors
RMSE	Root Mean Square Error

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Objetivos

El presente trabajo tiene por objetivo estudiar la transformación tecnológica que está experimentando el sector inmobiliario (*PropTech*), y en especial, el impacto que está teniendo el *Big Data* en el análisis de los mercados inmobiliarios, tanto desde un punto de vista teórico, como desde una perspectiva más visual y práctica.

En concreto, tres serán los objetivos técnicos a lograr en la parte práctica:

1. Crear un **sistema automatizado de obtención de datos**, referidos a inmuebles residenciales que estén en venta en Valladolid y que, al tiempo de su obtención, estén publicados en el portal inmobiliario Idealista.
2. Obtener de forma automatizada una **imagen global de la situación del mercado inmobiliario de la ciudad de Valladolid**.
3. Lograr desarrollar un **modelo algorítmico que consiga predecir los precios** de los inmuebles en venta con la máxima precisión posible y en el que aplicaremos técnicas propias del aprendizaje automático -*Machine Learning*-.

En último término se trata de transmitir la idea de que la adopción de la innovación tecnológica no solo es recomendable por las ventajas que entraña a nivel de resultados, sino sobre todo que su implantación, a día de hoy, está lejos de ser un imposible para cualquier empresa que opere en el sector inmobiliario como puede ser el caso de una inmobiliaria tradicional o familiar.

1.2. Justificación del tema objeto de estudio y motivación

En una sociedad con indudable orientación tecnológica, en la que todas y cada una de las industrias están transformando sus modelos de negocio, el sector inmobiliario deja de ser una excepción. La industria del inmueble, a pesar de sus reticencias y aversión a los cambios, ha emprendido una nueva etapa de la mano del PropTech. Si bien, por su propia naturaleza, la influencia tecnológica difiere notablemente en función de la región en la que nos encontremos.

Dentro del movimiento PropTech una de sus ramificaciones más comunes y destacadas es la del Big Data inmobiliario. Tradicionalmente, las empresas del sector han fundamentado sus decisiones en una combinación de intuición e información de corte tradicional. Sin embargo, hoy, como bien explica Mckinsey (2018), "*es posible combinar grandes bases de datos para predecir el precio por metro cuadrado de un alquiler de tres años en la ciudad de Seattle*, *por poner un ejemplo*". Para mayor abundancia, la cantidad de datos que genera esta industria y sus usuarios hace presuponer que las empresas que se especialicen en el Big Data no solo no tengan ningún problema a la hora de adentrarse en el mercado, sino que disfrutarán de una muy relevante ventaja competitiva sobre empresas ya establecidas (Deloitte, 2018, p. 4). Comprender y actuar en función del

estado de tal cuestión se antoja una decisión crucial para cualquier empresa o usuario inmerso en el mercado inmobiliario.

Por su parte, los motivos que me llevan a abordar esta temática, y sobre todo de esta manera, tienen un notorio trasfondo personal. Por un lado, mi cercana relación al sector del inmueble, pues mi padre dirige una inmobiliaria pequeña y familiar en la ciudad de Valladolid, con la que intento colaborar siempre que puedo. Y, por otro lado, mi especial interés por la industria del Big Data y sus impresionantes aplicaciones de negocio, a raíz de cursar el Diploma en Fundamentos en Business Analytics que imparte la Universidad Pontificia Comillas

1.3. Metodología

Respecto de la parte teórica, la metodología que se ha seguido es la del análisis *top-down*, tratando de ir desde una cuestión más general (PropTech) a una más particular (Big Data inmobiliario). Con tal fin, se ha realizado una exhaustiva revisión de la literatura existente sobre ambas áreas de estudio.

Por su parte, la parte práctica ha requerido, debido a su carácter más técnico, de un mayor esfuerzo a nivel personal. Varios cursos de programación después y el no menos importante estudio personal han resultado en un proyecto analítico de datos que, mediante la utilización de herramientas y técnicas de Business Analytics y Big Data, además, ha tenido su base en mi ciudad natal, Valladolid.

1.4. Estructura

Con ánimo de construir una efectiva comprensión de lo que se quiere plantear, el trabajo se dividirá en cinco capítulos (además de este primero de introducción) y una conclusión final.

El **primer capítulo** abordará el concepto del PropTech, haciendo especial énfasis en los pilares que lo conforman, las diferentes etapas que ha tenido el movimiento a largo del tiempo, y su actual dispersión geográfica. Seguidamente, el **capítulo segundo,** a través de un examen riguroso de su significado, contextualizará el término de Big Data, y la evolución que ha sufrido hasta llegar a convertirse en Smart Data.

Los **capítulos tercero** y **cuarto** permitirán al lector, conocer cómo está impactando el uso del Big Data en la operativa diaria de la industria inmobiliaria, esclarecer cuáles son las ventajas y desafíos inherente a ello, y comprender, en profundidad, la metodología del análisis predictivo que se aplica actualmente en el sector inmobiliario.

Finalmente, el **capítulo quinto** acogerá una demostración práctica de la utilización que cualquiera, incluso la inmobiliaria con menos recursos, puede dar a las herramientas enfocadas al Big Data y Business Analytics en el análisis de datos del sector inmobiliario. Con ese propósito, y utilizando el lenguaje de programación R, seremos capaces de obtener, gracias a la API del portal inmobiliario Idealista, una base de datos de viviendas pertenecientes a la ciudad de Valladolid con la que realizaremos un análisis puramente de negocio sobre los patrones que informan el mercado inmobiliario vallisoletano, para finalmente diseñar un modelo de *Machine Learning* que prediga los precios de las viviendas.

2. EL FENÓMENO DEL *PROPTECH*

2.1. Definición y características

La vertiginosa transformación tecnológica que nuestra sociedad está presenciando afecta a todos y a cada uno de los sectores. En este contexto, el sector inmobiliario, a pesar de su más que conocido carácter tradicional, no es una excepción. Ya sea como activo o como industria, el sector del inmueble no es ajeno a ello. La diversidad de operadores que conviven en la industria inmobiliaria (grandes, pequeños, familiares, etc.) comienzan a ser testigos de la repercusión que entraña la transformación tecnológica y el

impacto directo que ésta tiene sobre sus modelos de negocio. La transformación digital en la que está inmerso el sector inmobiliario acoge el término o concepto de "*PropTech*[1]".

Sin ánimo de entrar en definiciones complejas que abarquen numerosos ámbitos, de la lectura de varios autores e informes corporativos -Baum & Dearseley (2017), PropTech house (2019), entre otros- se puede determinar que el PropTech es cualquier innovación digital que afecta al sector inmobiliario. Por tanto, vemos que es un concepto amplio, por no decir amplísimo. Quizá, la aproximación menos ambigua, y al mismo tiempo más acertada y concreta podría ser la de "*aquel movimiento que dirige un cambio de mentalidad, en el sector y sus consumidores, en relación con el proceso tecnológico transformador del uso de datos, transacciones y diseño de edificio y ciudades*" (Baum, Saul, & Braesemann, 2020, p. 5).

Como vemos, y relevante a los efectos del presente documento, el uso y tratamiento de datos es una de las claves que dirigen el fenómeno PropTech, pero en clave teórica, datos y el fenómeno a estudiar, son conceptos distintos, aunque conexos. De hecho, se retroalimentan. El uso de las nuevas tecnologías ayuda a una mayor recopilación de datos e información, y es el análisis del dato el que faculta que el progreso tecnológico de transformación sea más rápido, principalmente por la posibilidad de tomar decisiones más informadas. Probablemente esta última afirmación tenga un mayor calado en esa relación de retroalimentación pues son las transformaciones digitales centradas en el almacenamiento, análisis y visualización de datos, las que en todo caso informan el devenir de todas y cada una de las capas que actualmente desprende el PropTech, y las que puedan aparecer. Es, por ello, que esta matización es importante dejarla clara para evitar confusiones a lo largo del texto. De igual importancia en su consideración es distinguir la figura del PropTech de otras afines o coligadas, especialmente por estar también acompañadas de la nomenclatura *tech*. Hablo de términos como ContructTech, referido a las nuevas formas de construcción inmobiliaria, LegalTech, basado en el uso de *smart contracts* y facilitador de las transacciones, o incluso algunos prácticamente idénticos como es el caso del término RealTech. Relacionado con esto, probablemente la más importante distinción que se debe hacer es la que existe entre PropTech y el término *FinTech*.

[1] De aquí en adelante se utilizará el término PropTech sin cursiva en aras de una mayor uniformidad del documento

Con la palabra FinTech se concibe la *implementación del uso de las nuevas tecnologías y modelos de negocio disruptivos en el campo de los servicios financieros* (World Economic Forum, 2015). Como todo proceso de cambio se caracteriza por tener una serie de pilares angulares que sostienen y fundamentan el mismo. Pues bien, muchos de ellos, como *Lending tech*, *Blockchain* o *capital markets tech*, por decir algunos, son variables que también afectan al sector inmobiliario. De hecho, muchas compañías dedicadas a éste combinan buena parte de esos pilares con los propios del PropTech. En este sentido, dentro de ambas industrias existe una esfera en la que comparten cualidades como si de un círculo conexo se tratase, y la cual recibe el nombre de *Real Estate FinTech*.

Ahora bien, esta circunstancia no siempre se da. La realidad PropTech no tiene por qué estar en connivencia con el FinTech. Por un lado, y como diferencia obvia, el FinTech desarrolla sus efectos sobre un tipo concreto de servicios que son los financieros, y, por tanto, distintos que los activos inmobiliarios. De esta forma, la construcción de un edificio inteligente en el que la obtención y análisis de datos sea un hecho no es de ninguna manera FinTech. Pero por otro lado, como diferencia no tan notoria a simple vista, el PropTech atrae, precisamente por la tipología de sus activos, a la economía colaborativa o *sharing economy* (Baum, 2017, p. 8). Ejemplo por antonomasia de ello es el de Airbnb, plataforma de alquiler turístico entre particulares, que no incorpora de manera clara en su modelo de negocio ninguna de las características del FinTech.

2.2. Los pilares del PropTech

Ya estrictamente dentro de lo que es PropTech, y partiendo de la definición dada en el apartado anterior, es menester definir cuáles son los principales pilares que efectivamente guían el PropTech. Desde un punto de vista taxonómico (**Pilares verticales y horizontales del PropTech**) el PropTech se divide en 3 grandes áreas o subsectores, a saber: *Real Estate FinTech*, *Shared Economy* (economía colaborativa), y *Smart Real Estate* (*ConstructTech*).

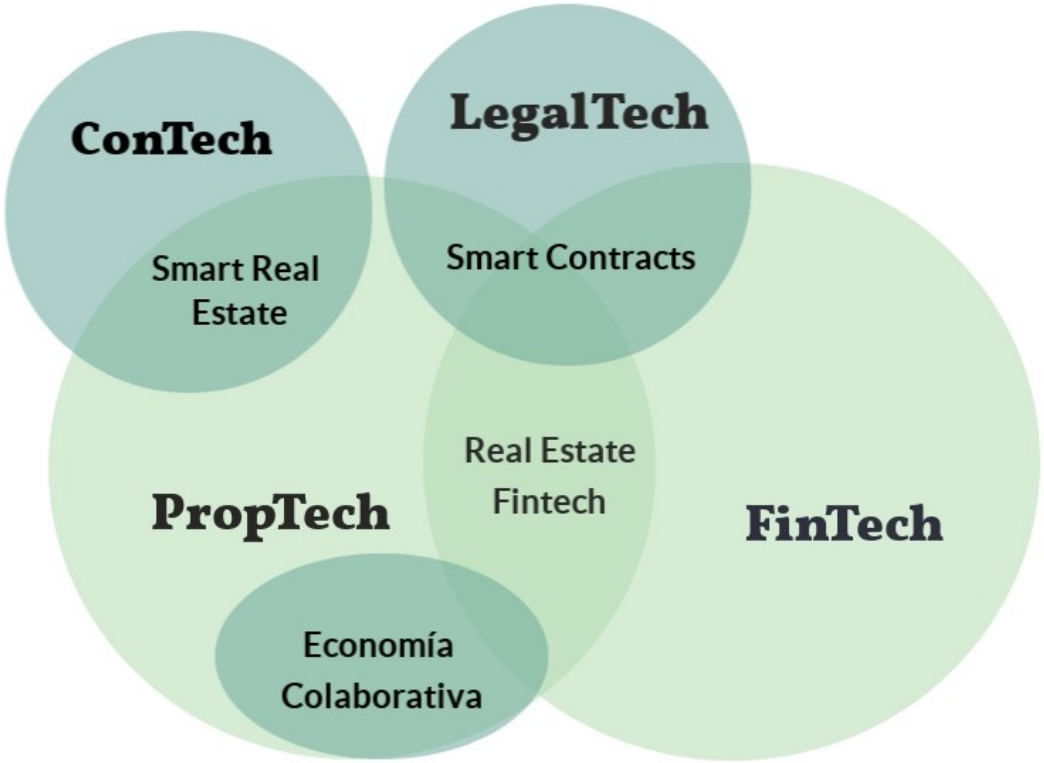
El primero de ellos, ***Real Estate Fintech*** hace referencia a cualquier plataforma tecnológica que facilite la negociación de activos inmobiliarios. El paso del tiempo ha hecho que el concepto haya desembocado en diferentes formas operativas. De entre ellas, las más relevantes son las plataformas dedicadas a la búsqueda de inmuebles (Idealista,

Fotocasa, etc.), que ponen en contacto a las distintas partes interesadas y que encarnan el paso de la tradicional operativa *offline* a la *online*, y aquellas cuyo ámbito de actuación es el enfoque del activo inmobiliario como inversión, en vez de como propiedad. Por su parte, el subsector de la **economía colaborativa (***Shared Economy)* despliega sus efectos estrictamente en el uso que se hace de los inmuebles en sí mismos, desde un punto de vista social. Antes hablábamos del ejemplo de Airbnb como referente capital en este campo, pero tampoco podemos olvidar otras formas también dedicadas a la gestión de inmuebles, aunque no residenciales como WeWork que pone a disposición el uso compartido de espacios entre diferentes empresas. Por último, ***Smart Real Estate*** es el resultado de diferentes plataformas que, basadas en el análisis de la información procedente del mismo activo, tienen por objeto conseguir un uso más eficiente del inmueble, especialmente en el desempeño energético[2]. Expectativas de futuro auguran, no solo un aumento del interés de los inversores en hacerse con este tipo de activos, sino también en invertir en activos inmobiliarios construidos para almacenar la generación masiva de datos – *data centers*.

Una mejor ejemplificación de este encuadre se puede apreciar en la **Figura 1**. En ella se contiene la distinción entre el sector PropTech y FinTech, y su conexión con el Real Estate FinTech. Adicionalmente, se recogen las diferentes formas o subsectores en los que PropTech toma cuerpo e infunde efectos. Un análisis rápido de la misma permite identificar la ambigua y difícil estratificación de los distintos ámbitos del PropTech y el FinTech.

[2] Ejemplo reciente de Smart Real Estate es el de la startup Poween: https://poween.com/

Figura 1: Los subsectores del PropTech



Fuente: Elaboración propia a partir de Baum, Saul & Braesemann (2020)

Por su parte, de forma transversal (pilares horizontales) y a consecuencia de la nueva era digital en la que vivimos, el PropTech queda afectado en todos sus ámbitos por la mayor obtención de **información** o datos, la facilidad para realizar **transacciones** y la mayor eficiencia en el **manejo y control** de los activos, todo ello beneficiado por una mayor capacidad de procesamiento de la información que en el pasado (**Tabla 1**).

Tabla 1: Pilares verticales y horizontales del PropTech

	Real Estate FinTech	Economía colaborativa	Smart Real Estate
Información	Sí	Sí	Sí
Transacciones	Sí	Sí	
Manejo y control			Sí

Fuente: Elaboración propia a partir de Baum (2017)

2.3.Evolución en el tiempo: las discutidas tres eras del PropTech

El sector inmobiliario siempre ha sido un sector que, por las características de sus activos, ha tendido a rechazar o al menos a aplazar la influencia del proceso tecnológico. Sin embargo, podemos inferir que el PropTech no es un fenómeno nuevo. Concretar el punto de partida de este fenómeno no es cuestión sencilla, y de hecho, es bastante discutida. La dificultad estriba principalmente en que la adopción tecnológica siempre ha sido algo característico del ser humano a lo largo toda su historia, por lo que precisar el momento exacto encuentra sus dificultades. Aun así, tomando en consideración la definición dada *supra*, y la importancia de los datos, se puede determinar que el PropTech tiene su primera aparición a mediados de la década de 1980, dando lugar al llamado ***PropTech 1.0***. Por entonces el poder computacional, que se había venido desarrollando, permitió la introducción de diferentes herramientas de análisis, como las hojas de cálculo, y consecuentemente estandarizar el uso de plataformas orientadas al análisis y organización de la información. Todo ello, unido con la introducción y avance del ordenador personal (PC), dio lugar a un impacto en la operativa del sector inmobiliario. Sin embargo, este, como es obvio, careció de un efecto relevante sobre el mercado inmobiliario que pudiese anunciar un cambio de mentalidad (Baum, Saul, & Braesemann, 2020, pp. 7-8).

El inicio de siglo despidió al *PropTech 1.0* y dio la bienvenida al ***PropTech 2.0.*** Según Baum (2017), el factor decisivo entre uno y otro es el paso de un mercado inmobiliario residencial tradicional a otro online. Si nos paramos a pensar, esto es algo consecuente con el momento que se estaba viviendo. Desde mediados de los años 90 la efervescencia tecnológica estaba abriendo paso a nuevos modelos de negocio, y aunque desembocó en la burbuja de las "punto com", permitió el nacimiento de los primeros portales inmobiliarios. No obstante, el verdadero crecimiento tuvo lugar a raíz de la crisis financiera global que surge en el año 2007. He ahí el nacimiento del cambio decisivo que vive el sector inmobiliario. La pérdida de credibilidad en los métodos de inversión tradicionales y la tremenda disrupción tecnológica liderada principalmente por la generalización en el uso de Internet móvil facilitaron el acceso a información inmobiliaria de forma instantánea, y sobre todo, a un coste menor. La consecuencia principal de todo esto fue la multiplicación exponencial de la cantidad de datos y el crecimiento constante en la creación de empresas dedicadas a la manipulación y extracción de valor de esa información. Entre ellas, como no, se encontraban las dedicadas al PropTech.

El número de empresas PropTech que se fundaron en el período que va desde la crisis hasta los años 2014/2015 no paró de crecer. Alcanzado ese punto de inflexión el mercado PropTech ha venido viviendo un descenso generalizado en el número de empresas fundadas llegando incluso, en el año 2018, a niveles propios de inicios de la crisis (Baum, Saul, & Braesemann, 2020, pp. 10-11). Esto nos hace pensar que quizá la era PropTech 2.0 está dando ya claros síntomas de que su final está cerca, y de hecho es así. Sin embargo, no podemos pasar por alto que desde el 2015 hasta ahora lo que ha ocurrido no es tanto un descenso de la operativa de este mercado, sino una orientación más estratégica, financiera y de obtención de capital. Así, si se analiza comparativamente las **Tabla 2** y **Tabla 3** se puede comprobar cómo, desde los años 2014-2015, han existido dos realidades en el mercado PropTech. O, por un lado, empresas dedicadas expresamente al PropTech han decido seguir un proceso de expansión a través de la adquisición de otras que también tengan base tecnológica, logrando así, embarcarse en un *cuasi-proceso* de consolidación dentro del mercado del PropTech, construir plataformas tecnológicas más atractivas por el surgimiento de sinergias, y sobre todo, atraer a un mayor número de inversores[3]. O, en dirección distinta, que no opuesta, aquellas empresas que, dedicadas al sector inmobiliario tradicional pero con una clara orientación a la transformación tecnológica, han decidido implementar tecnología a través de la adquisición de distintas PropTech de menor tamaño[4] (Goodwin, 2019).

El resultado de lo anterior es un mercado de crecimiento pero en clara fase de consolidación. La explosión de empresas PropTech es ya una cuestión del pasado, y quizá también lo sea el PropTech 2.0. Cuestión distinta es si hemos entrado ya o no en la era ***PropTech 3.0*.**

En humilde opinión del que escribe decir que el concepto de PropTech tiene muchas variantes en su análisis, especialmente si se enfoca desde el punto de vista económico y de dinámica de mercado, o si por el contrario se enfoca en el sentido de la innovación tecnológica. Si nos decantamos por el primero no será desencaminado afirmar una continuación de la tendencia antes expuesta de consolidación de mercado. Por su parte, si el enfoque escogido es el tecnológico (el más adecuado a mi juicio), podríamos ya estar entrando en esa era 3.0. La consolidación y madurez de tecnologías nuevas como

[3] *Tech/Tech M&A*
[4] Real Estate/Technology Company Combinations

el *Blockchain* o la Inteligencia Artificial, así como el mayor interés de aplicación de estas en distintos sectores, pueden ser la primera piedra de un nuevo tiempo para el PropTech. Por último, y sin ánimo de aventurar circunstancias que puede que no ocurran, el futuro del PropTech no puede dejar de ser analizado con una nueva variable que en pocos años será sinónimo de ventaja competitiva para algunos, y de caída para otros: el cambio climático.

Tabla 2: Operaciones M&A Tech/Tech

Comprador	Vendedor	Año	Precio	Segmento
States Title	Captive Title North America	2019	Undisclosed	**Analytics platform**
Amazon	Eero	2019	Undisclosed	**Tech e-commerce**
RealPage	ClickPay	2018	$219M	**Property management and payment platform**
Autodesk	PlanGrid	2018	$875M	**Construction productivity software**
Oracle	Aconex	2017	$1.2B	**Cloud software and construction platform**
Oracle	Textura	2016	$663M	**Cloud software construction platform**
Expedia	HomeAway	2015	$3.9B	**Hospitality**
Zillow	Trulia	2015	$3.5B	**Online real estate databases**

Tabla 3: Operaciones M&A Real Estate/Tech

Comprador	Vendedor	Año	Precio	Segmento
Prologis	DTC Industrial	2018	$8.5B	**Industrial / e-commerce**
Sumitomo Forestry	Crescent Communities	2018	$370M	**Smart building technology**
CoStar Group	ForRent	2017	$385M	**Tech-based multifamily information platform**
Industrious	Pivotdesk	2017	Undisclosed	**Digital platform for flex office**
CBRE Group, Inc.	Floored Inc.	2017	Undisclosed	**Interactive tech-based leasing**
AccorHotels	Onefinestay	2016	$167M	**Hospitality**

Fuentes: Elaboración propia a partir de Goodwin (2019)

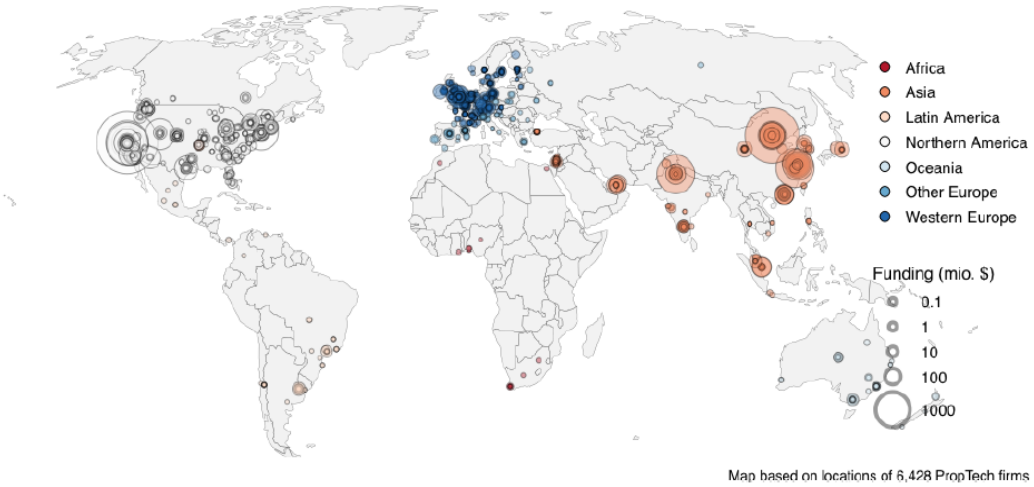
2.4. Dispersión geográfica

2.4.1. Internacional: Estados Unidos y China

Hablar de PropTech implica hablar también de un fenómeno de escala globlal. La combinacion del activo inmobiliario, independientemente de la forma, con la innovacion tecnológica es un hecho presente en todos los lugares. Sin embargo, esto requiere de matización porque el PropTech no concurre con la misma fuerza en todas las regiones. Así, existen ciertos lugares en los que el desarrollo PropTech es expectacular – Estados Unidos, Europa Occidental, y selectas metropolis asiáticas – pero fuera de estas zonas el desarrollo del sector se situa muy por debajo (**Figura 2**). Dicho de otra manera, se produce lo que se conoce como "*clusterización*" del PropTech (Baum, Saul, & Braesemann, 2020, p. 21).

Concretando más habría que detener el análisi en Estados Unidos. Es con diferencia la region geográfica más avanzada en este sentido. En términos numéricos, si actualmente hay aproximadamente 7000 compañias PropTech[5], más de 2000 se localizan en el país americano. Uno de los factores más importantes a su favor es precisamente el enorme desarrollo tecnologico, y es que, 7 de las 10 empresas más tecnológicas del mundo, tienen a Estados Unidos como lugar de origen[6]. Por otro lado, y esto es un punto a su favor, la extension territorial y su poblacion permiten que su mercado inmobiliario sea increiblemente rico, tanto en número como en variedad.

Figura 2: Distribución mundial de firmas PropTech



Fuente: Crunchbase, Unissu y FoRE

[5] Base de datos de Unissu
[6] Microsoft, Amazon, Apple, Google, Facebook, Intel y Oracle

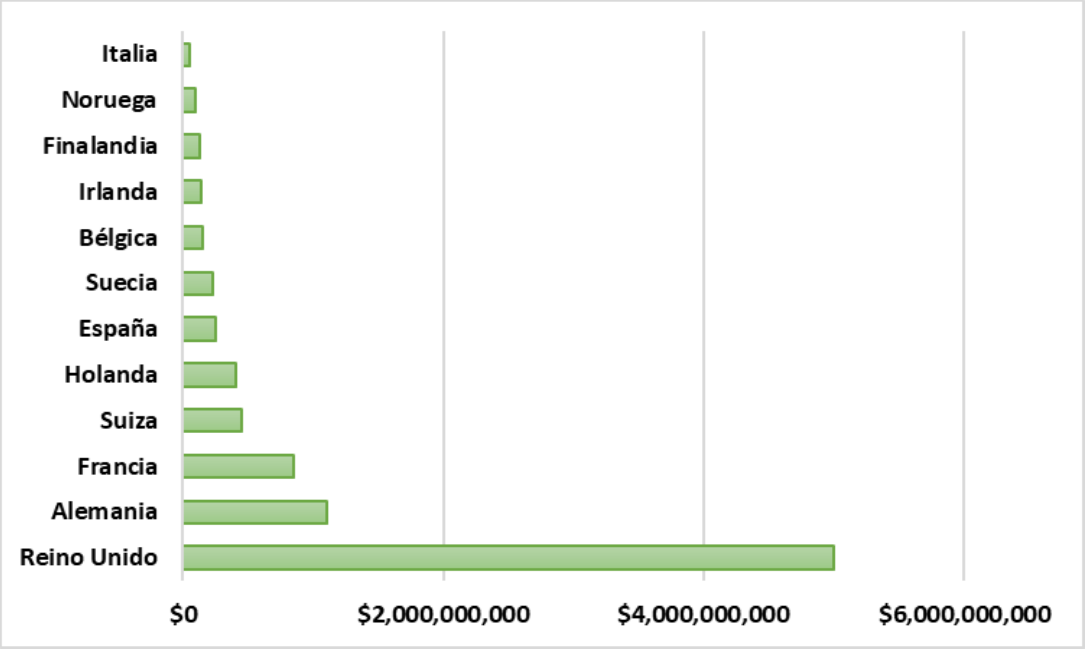
Al otro lado del globo, y siguiendo la misma lógica argumentativa que para Estados Unidos, nos encontramos con China que es la segunda gran potencia PropTech. A diferencia del anterior, su distribución está mucho más focalizada en las grandes ciudades probablemente por las difíciles características orográficas del país, y la mayor desigualdad entre ciudades desde el punto de vista tecnológico.

2.4.2. Europa

Completando el pódium estaría el continente europeo cuya dominancia es copada por Reino Unido. Según la base datos especializada en PropTech (Unissu), son aproximadamente 3300 las firmas PropTech presentes en Europa, y es de remarcar que 8 países europeos tienen más de 100 compañías PropTech. En este contexto, España se sitúa en tercer lugar por detrás de Francia con aproximadamente 330 firmas.

Para evitar que estos números lleven a un malentendido es necesario verlo desde la perspectiva de capital atraído en inversiones. En estos términos, Reino Unido sobresale con mucha diferencia como se puede comprobar en la **Figura 3**, y España deja de estar en las tres primeras posiciones.

Figura 3: Financiación total de compañías PropTech por país (Europa)



Fuente: Elaboración propia a partir de Unissu (2019)

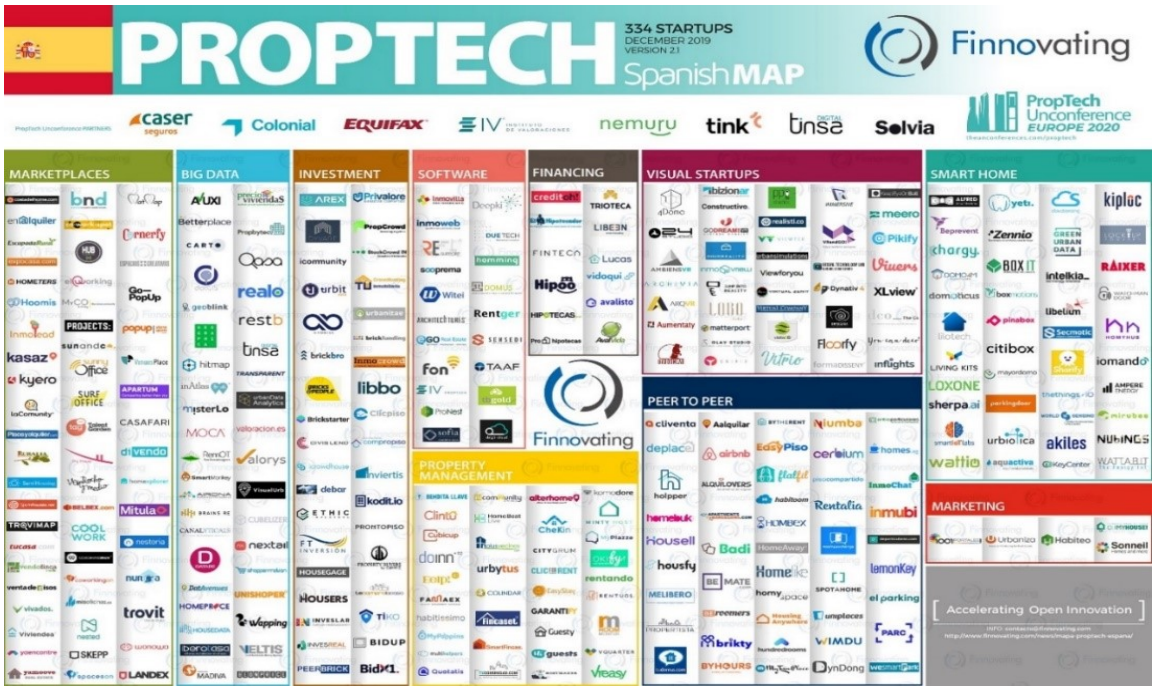
La conclusión que podemos sacar de ello es que actualmente en Europa, si bien existe un mercado bastante consolidado por el número de compañías, en términos de

movimiento de capital aún hay muchos países con altos márgenes de mejora. De hecho, es especialmente interesante nuestro caso, teniendo en cuenta la importancia que tiene el sector del ladrillo en España.

2.4.3. España

De acuerdo con Finnovating, plataforma digital dedicada al ecosistema *start-up*, en diciembre de 2019 España contaba con un total de 334 compañías PropTech. En el año 2017, concretamente en octubre, el número de empresas PropTech era de 160. Es decir, en poco más de dos años el mercado español dedicado al PropTech se ha duplicado. Probablemente, uno de los factores que podría haber ayudado bastante en esta expansión es la diversificación interna que tiene el PropTech. A la hora de conceptualizar el fenómeno PropTech hacíamos referencia a las capas que desprendía o los distintos subsectores de los que se componía. Pues bien, actualmente, el mercado español cuenta con varios permitiendo esa expansión diversificada tan remarcable: *market places*, *Big Data, investment*, *software*, *financing*, *visual startups*, *smart home*, *marketing*, *property management*, y *peer to peer*. El número y tipo de divisiones es variable por lo que no es extraño encontrarse con clasificaciones distintas e incluso discrepancias en el número de empresas PropTech creadas a lo largo de los años.

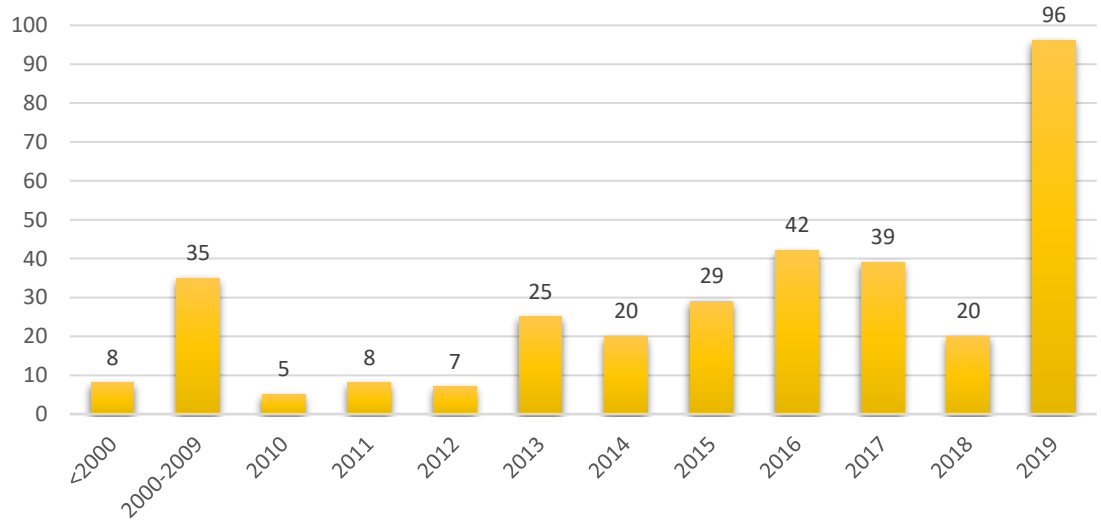
Figura 4: Mapa del PropTech en España (diciembre 2019)



Fuente: Finnovating

Desde un punto de vista estratégico, y de perspectiva, se podría afirmar con total rotundidad que el mercado español está en una fase de expansión, y que por ello, España todavía no se encuentra en una fase de consolidación como en otros focos internacionales. Esto puede entrar en contradicción con las previsiones señaladas para el año 2019. Según la consultora inmobiliaria Savills Aguirre Newman, en su Informe PropTech del 2019, el mercado español se adentraba en una fase de consolidación por el descenso de nuevas empresas *start-up* dedicadas al ámbito PropTech en el año 2018. Desde el año 2014 no se alcanza un número tan bajo (**Figura 5**). Sin embargo, la realidad ha sido totalmente distinta. Si tenemos en cuenta los datos de Finnovating (**Figura 5** y **Figura 6**), que sí contiene datos del año 2019, en un año se han sumado aproximadamente 100 empresas, por lo que no existe clarividencia de que realmente el mercado español esté efectivamente adentrándose en una espiral de consolidación, ya sea por desaparición de algunas o por la fusión o adquisición de otras. Aun así, en este contexto, sí que se debe tener en cuenta que el movimiento de capital generado por el mercado español llegó en el año 2018 a un total de 150 millones de dólares (Savills Aguirre Newman, 2019, p. 3). De ahí, que no sea tan descabellada la idea de un mercado que comienza ya a dar los primeros síntomas de una futura consolidación en el corto plazo.

Figura 5: Evolución del número de empresas PropTech en España



Fuente: Elaboración propia a partir de datos de Finnovating y Proptech.es[7]

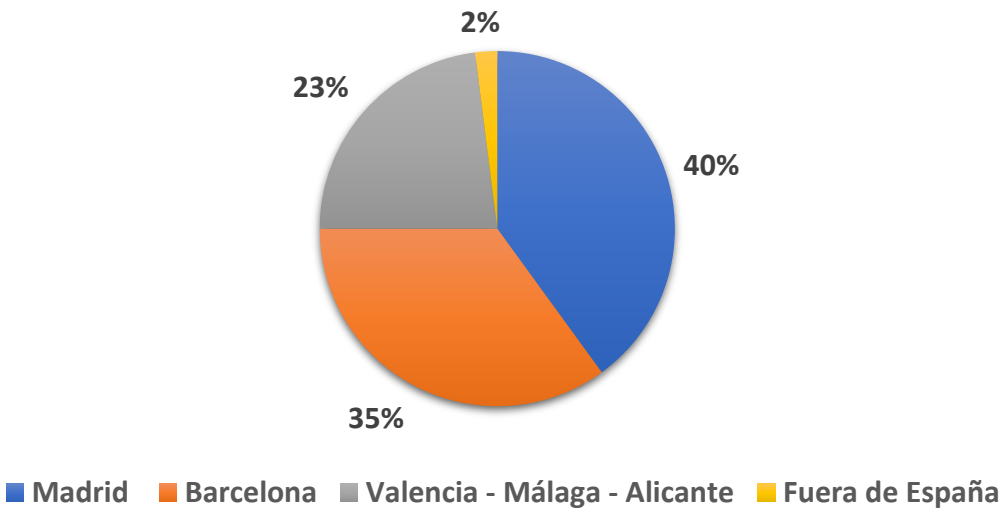
Geográficamente hablando no hay prácticamente discusión. Madrid y Barcelona contienen más del 75% de empresas. Si a estas dos les añadimos el grupo perseguidor –

[7] https://proptech.es/

Valencia, Málaga y Alicante – prácticamente completaríamos el mercado español. Por su parte, encontramos un número reducido de empresas que, con sede fuera de España, sí que operan en nuestro país.

Ello permite reflexionar acerca de las características que presenta el actual tejido inmobiliario en la mayor parte de España, pero también nos demuestra la espectacular cantidad de oportunidades que todavía existen para nuevos proyectos inmobiliarios cuya punta de lanza sea la aplicación de medios tecnológicos. A este último respecto da buena fe el apartado número 6 (**APLICACIÓN PRÁCTICA**) en el cual se explica de manera minuciosa cómo es posible desarrollar, en la ciudad de Valladolid, un proyecto inmobiliario de naturaleza PropTech enfocado en el análisis de datos.

Figura 6: Localización de las empresas PropTech en España



Fuente: Elaboración propia a partir de datos de Proptech.es

3. BIG DATA Y DATA ANALYTICS

Como adelantábamos en el anterior apartado, el PropTech tiene una gran variedad de ramificaciones, cada una de ellas merecedora de un estudio concreto y separado pues, por lo general, todos responden a la aplicación de distintas tecnologías en el sector inmobiliario. De hecho, a la hora de clasificar las firmas PropTech, suele ser práctica común tomar como referencia la tipología tecnológica en la que se especializa cada empresa (**Figura 4**).

Dentro de esas clasificaciones siempre nos encontramos con la de Big Data. Independientemente de su consideración como fenómeno que informa al sector PropTech, como una variable clasificatoria de las empresas o simplemente como una tendencia más del PropTech, este fenómeno se ha ganado con creces un sitio dentro del sector inmobiliario, pero su relevancia aún está por determinar.

3.1. El concepto de Big Data

A pesar de la falta de consenso en su definición, Oracle[8], a la hora de definir Big Data, parte de la definición dada por Doug Laney, analista de Gartner, en el año 2001: *"Big Data son datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior"*. Es decir, Big Data no solo atiende al volumen generado de datos, sino también a la variedad, complejidad y diversidad de origen de estos. Diez años más tarde, Sam Madden, experto en ciencia computacional, reformulaba la definición combinándola con el contexto de las herramientas existentes por entonces para el procesamiento de datos. Así, opinaba que, para él, la expresión Big Data se debía utilizar para aquel grupo de datos que es "*demasiado grande, complejo, y veloz para ser procesado con las herramientas* existentes"[9] (Madden, 2012, p. 4). Aunque la gran mayoría, especialmente por su carácter más académico, utiliza la primera de ellas para analizar el concepto de Big Data, creo que la segunda es mucho más representativa en cuanto a la magnitud de dicho fenómeno.

La información que deriva del Big Data es tan grandilocuente en todos los sentidos, que no solo supone tener más datos, sino que provoca cambiar la manera en la

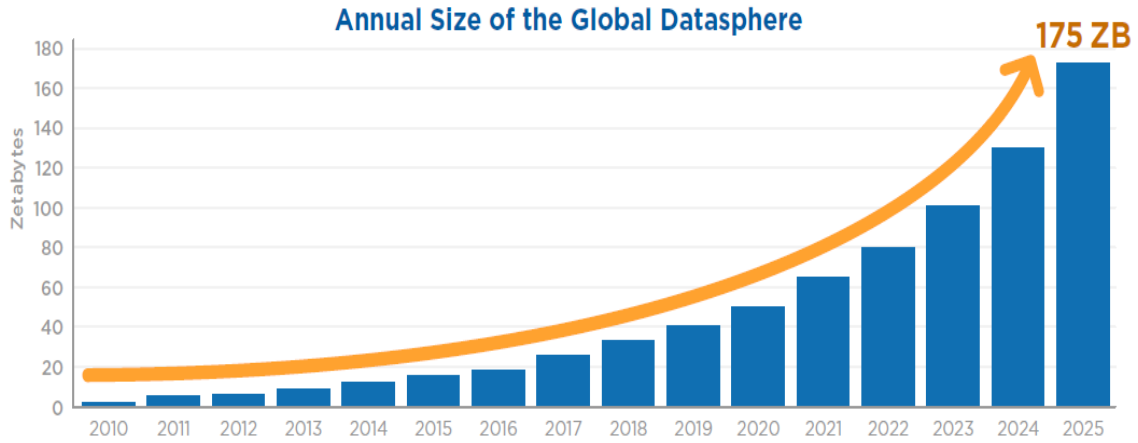
[8] https://www.oracle.com/es/big-data/guide/what-is-big-data.html
[9] *"Data that's too big, too fast, or too hard for existing tools to process"*

que pensamos (Boyd & Crawford, 2012). Además de la mentalidad, Big Data también abarca el cambio que han sufrido las herramientas computacionales en este sentido. Las tradicionales herramientas de procesamiento de datos pasan a ser historia dejando paso a otras nuevas y confeccionando un panorama totalmente distinto.

Volviendo a las características más académicas, la definición de Laney antes enunciada contiene tres de las características que actualmente conforman el concepto de Big Data. Con el paso del tiempo dichas características se fueron consolidando, posibilitando incluso la llegada de otras cuatro para acabar por componer el ya ampliamente conocido como esquema de las "7 Vs[10]":

- **Volumen**. Actualmente, no podemos hablar de una gran cantidad de datos si se compara con las proyecciones de futuro. De forma diaria nuestra dependencia tecnológica lleva consecuentemente una generación de datos, y resultado de lo anterior es que año tras año el volumen generado rompe records **(Figura 7**). De acuerdo con la Corporación Internacional del Dato (2018), se calcula que para 2025 la creación de datos llegará a los 175 Zettabytes[11]. En términos más visuales, si quisiéramos almacenar la cantidad la cantidad total de datos en DVDs, necesitaríamos tal torre que podríamos llegar hasta la luna una 23 veces o dar la vuelta a la Tierra 222 veces (Reinsel, Gantz, & Rydning, 2018, p. 7).

Figura 7: Evolución histórica y previsión del tamaño global de los datos generados anualmente a nivel mundial



Fuente: International Data Corporation

[10] En cualquier caso, la definición técnica estándar de IBM es la de las cuatro primeras: volumen, velocidad, variedad y velocidad https://www.ibm.com/analytics/hadoop/big-data-analytics

[11] 1 Zettabyte = 10^8 Gigabytes

- **Velocidad**. Directamente relacionado con el volumen está la velocidad a la que se generan los datos. Con este término nos referimos al ritmo al que se reciben, y , en su caso, se analizan. Simplemente con echar un pequeño vistazo a la **Figura 7** uno se puede dar cuenta del poco tiempo que nuestra sociedad ha necesitado para lograr tal descomunal abundancia de datos. Las causas de ello giran principalmente en torno a la enorme cantidad de fuentes que dan origen a los mismos, y a las innumerables posibilidades de almacenamiento, que van desde los descomunales centros de datos hasta cualquier dispositivo móvil u ordenador portátil de cada uno de nosotros.

- **Variedad**. La creación de datos no es uniforme. Los datos pueden aparecer de formas muy diversas, que es lo que de verdad brinda de complejidad a la mezcla (Khan, Uddin, & Grupta, 2014). La principal clasificación es de carácter bicéfalo. Por un lado, los datos estructurados que se crean conforme a una serie de reglas y de patrones comunes, mientras que, contrariamente, los datos no estructurados no responden a ninguna regla y carecen de cualquier estructura interna reconocible. La utilización de estos últimos estará sujeta a un proceso previo de tratamiento y limpieza. Entre medias de ambos tipos se encuentran los datos semiestructurados que contienen características de los otros dos.

- **Veracidad**. Con este término uno se refiere a como de creíble o veraz es ese grupo de datos (Khan, Uddin, & Grupta, 2014). También a los distintos errores cometidos a la hora de producir los datos como puede ser una palabra erróneamente escrita en un tweet, lenguaje informal o simplemente un acrónimo (Winson-Geideman & Krause, 2016, p. 3). El hecho de que gran parte de los datos nazcan de las personas individuales priva en muchos casos de que los datos tengan un mayor beneficio. Por esta razón, la limpieza de datos (*datacleaning*) a la hora de analizar los mismos se ha tornado un proceso esencial, y la necesidad de herramientas más sofisticadas en su análisis una obviedad (Winson-Geideman & Krause, 2016, p. 3).

- **Validez**. Aunque estrechamente relacionado con el anterior no deben confundirse. Por validez entendemos la disposición que tiene un determinado grupo de datos con el objetivo que se persigue. De esta forma, es perfectamente posible tener una serie de datos, que sin ningún problema de veracidad no sean óptimos ni adecuados para el

problema que se plantea. Esto, eso sí, no obsta para que tengan su virtualidad de cara planteamientos distintos (Khan, Uddin, & Grupta, 2014).

- **Volatilidad**. De las 7 Vs quizá es la más abstracta. Una serie de datos determinados pueden ser beneficiosos en un espacio de tiempo determinado, pero una vez este ha llegado a su fin, la utilidad de estos también. Es decir, la volatilidad de los datos es el espacio de tiempo en el que el dato es de relevante utilidad. Cuanto mayor es el volumen, velocidad y variedad (3 primeras Vs) de los datos, mayor es la volatilidad de los mismos.

- **Valor**. Es la variable más omnipresente. A diferencia de otras, no se trata de una característica propia del ente conocido como dato, que ya tiene un valor *per se*, en cuanto a que su generación implica ya un coste en forma de factores productivos, sino que se refiere al valor obtenido o extraído de su procesamiento o análisis (Khan, Uddin, & Grupta, 2014). Así, este último tiene que ser contrastado con el coste efectivo contraído que posibilita ese procesamiento, y que viene a ser primeramente el coste de almacenamiento, no solo entendido como espacio, sino como calidad del mismo (seguridad del almacenamiento).

3.2. Del Big Data al Smart Data

La aglomeración de datos es algo intrínseco a toda industria. El uso y aprovechamiento de la información en forma de datos es el ingrediente esencial que hace posible su movimiento y, en consecuencia, seguir creciendo. Es más, alguno incluso habla de los datos como el petróleo del siglo XXI, aunque con la gran diferencia de que los primeros se encuentran en todas partes (Catella Research, 2015, p. 3). Esta última década, como se ha explicado *supra*, ha constatado una generación de datos sin precedentes, dando lugar al nacimiento del Big Data. Por tanto, y evitando entrar en detalles específicos de cada sector, ahora es posible encontrar ese petróleo informático en cualquier lado. Cuestión distinta será el estado en el que se encuentre. Pero, entonces, si tomamos en consideración lo anterior, la pregunta que surge es la siguiente: ¿por qué es tan interesante su estudio, si al final es una simple cuestión de volumen?

No, no es simplemente una cuestión de volumen, aunque su término dé pie a pensarlo. El término Big Data implica necesariamente la existencia de un ente generador

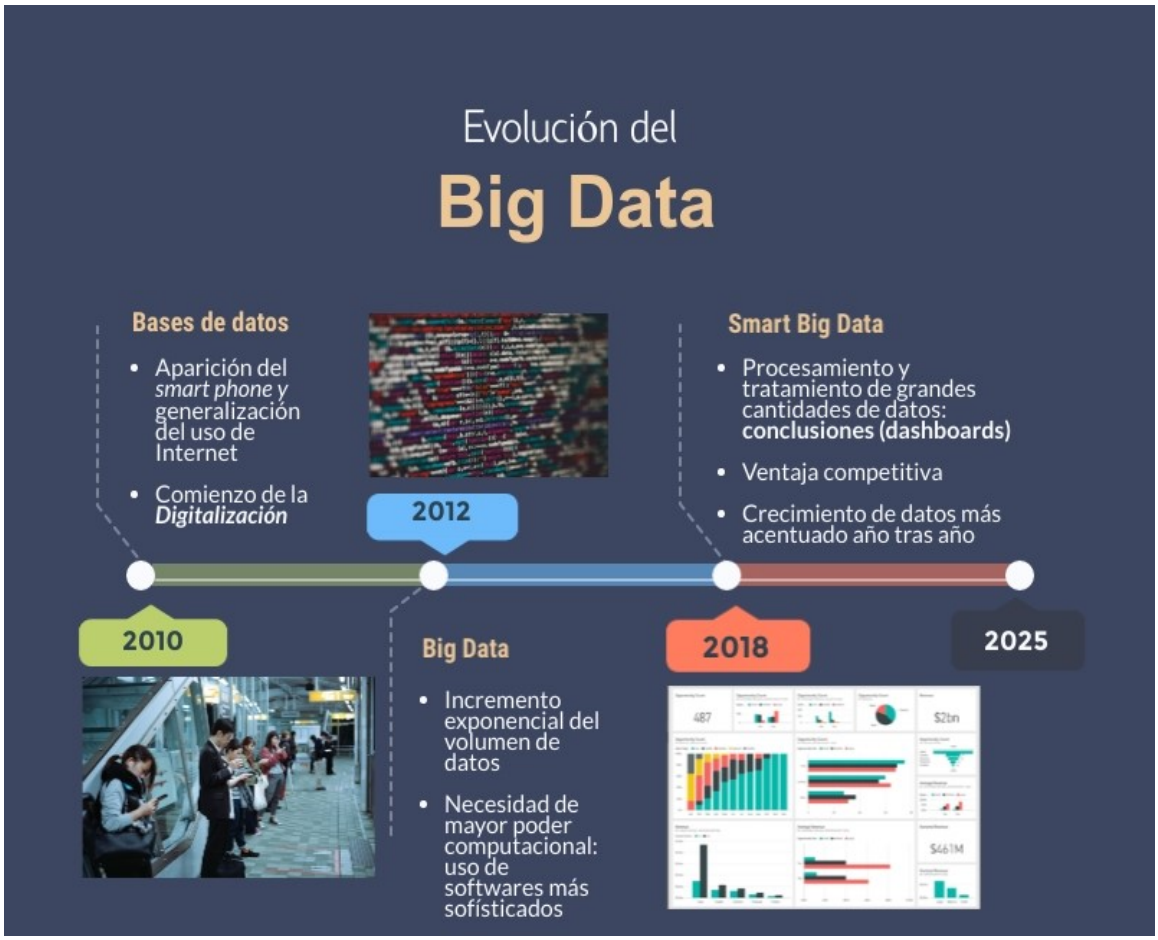
que se desvincule de todo lo manual. De esta forma, el Big Data se aleja de la digitalización[12] basada en la conversión de la copia física a la copia digital[13], para poner el foco en la digitalización puramente digital. Dicho de otra manera, el Big Data entiende la digitalización como el paso de cualquier información, nacida ya en formato digital, a un formato que sea digitalmente interpretable (Baum, Saul, & Braesemann, 2020). Por tanto, una primera aproximación, a raíz de la existencia de ese volumen, es que el Big Data necesita, como factor para su nacimiento, el soporte digital. De ahí, que el Big Data sea también una **cuestión de tecnología digital**. Un ejemplo práctico de lo explicado, y asociado al sector inmobiliario, sería la publicación de inmuebles en venta o en alquiler en los distintos portales inmobiliario (Idealista, Fotocasa, Pisos.com, etc.). A partir de esa información, se generan ficheros interpretables que posibilitan, en último término, un posterior análisis del que se pueda sacar provecho.

También es una **cuestión de procesamiento**. El Big Data pierde su razón de ser cuando de él no se puede sacar fruto. Distintas razones, como la falta de capacidad computacional, o simplemente la dificultad de su procesamiento para convertirlo en información susceptible de utilización, han enarbolado una barrera entre lo que se puede entender como Big Data de volumen o escala, y lo que se entiende por Big Data inteligente (***Smart Big Data* o *Smart Data)*** que es el que aquí interesa. Smart Big Data es el resultado de "*la transformación de una gran cantidad de datos, inicialmente no estructurado, y a través de un proceso inteligente de tratamiento, en conocimiento*" (Lenk, Bonorden, Hellmans, Roedder, & Jaehnichen, 2015). Así, el Big Data para poder convertirse en información (*Smart Data*) necesita de un proceso de transformación previo, al igual que las materias primas, como el petróleo, necesitan del suyo. Este proceso recibe el nombre de ciclo de vida del dato[14]. Eso sí, será clave que, este ciclo de vida de transformación, sea lo suficientemente óptimo como para dejar atrás las limitaciones propias del enorme volumen de información digital y "*permitir, así, extraer de la gigante masa de datos recibidos, su verdadero valor e interés*" (Souissi & El Arass, 2018). Si a lo explicado, se le aplica un análisis con perspectiva temporal (**Figura 8**), la conclusión es clara. Si a inicios de la década del 2010, éramos testigos de la evolución desde las bases de datos hacia el Big Data[15], desde hace algunos años, se puede comprobar

[12] El término anglosajón más adecuado para este tipo de digitalización sería el de *digitisation*
[13] Nótese que este tipo de digitalización no es suficiente por sí sola para crear el fenómeno del Big Data
[14] Traducción del término anglosajón *Data Lifecycle*
[15] Madden en el año 2012 publicaba un artículo con el siguiente título: *From Databases to Big Data*

el paso del Big Data al Smart Big Data[16]. De esta forma, al igual que la acumulación descomunal de datos, impulsada por esa digitalización, ha labrado el Big Data, el seguimiento de un ciclo de vida del dato óptimo consigue que el Big Data no solo sea conocido por su gran escala, sino por los grandes beneficios que provee a quien los analiza.

Figura 8: Evolución temporal del Big Data



Fuente: Elaboración propia

Por tanto, y volviendo a la pregunta que formulaba al inicio de este apartado, se puede ratificar la respuesta de que el Big Data no es solo una cuestión de cantidad o de volumen. Es la disrupción tecnológica, primero, la que ha permitido generar una cantidad asombrosa de datos, y son los procesos de tratamiento y optimización, en segundo lugar, los que permiten sacar conclusiones de ellos, y en consecuencia entrar en la nueva era del Smart Data.

[16] Souissi y El Arass (2018) confirman este hecho titulando su artículo Data Lifecycle: *From Big Data to Smart Data*

4. EL BIG DATA EN EL SECTOR INMOBILIARIO

4.1. Estado de la cuestión: ¿por qué es importante?

Ya se ha adelantado a lo largo del presente texto que la adopción de la disrupción tecnológica es algo que, estando presente en el sector inmobiliario, se ha caracterizado más por su lentitud que por su verdadero efecto, provocando, consecuentemente, que muchos datos pertenecientes a la industria aun requieran de un proceso de digitalización (Baum, Saul, & Braesemann, 2020, p. 27). Sin embargo, este hecho no niega la realidad. El sector inmobiliario, por las características de sus activos negociados, siempre ha sido aproximado desde el análisis de datos – rentabilidad del alquiler, características de un determinado edificio o inmueble, tipos de interés, índice de precios, etc.

Consecuencia de lo anterior es la situación en la que nos encontramos. La mezcla de una adopción tardía de la tecnología y la existencia de una gran cantidad de datos, inherente a la operativa propia del sector, provocan distintas realidades respecto de gran parte del conjunto de datos que recogen información inmobiliaria. Una primera posibilidad es que estén almacenados en bases de datos no estructuradas, que necesiten de un tratamiento previo para su maximización (Royal Institution of Chartered Surveyors, 2019). La otra posibilidad es que directamente esos datos no tengan acceso digital por contenerse aún en formato físico (Baum, Saul, & Braesemann, PropTech 2020: the future of real estate, 2020, p. 27). Es probable que gran parte de la culpa radique en el tradicional *modus operandi* del sector inmobiliario por el cual, la información de los mercados se encontraba preferentemente en manos de economistas, agentes y *brokers*, reticentes a la adopción tecnológica. Pues bien, a pesar de ello, la *Royal Institution of Chartered Surveyors* (RICS) -cuerpo profesional creador de los estándares internacionales que cubren la práctica del sector- asegura que son tres las circunstancias que, matizadas por el que escribe, están propiciando un cambio en todo este entorno:

1. **Mayor disponibilidad de datos estructurados sobre el mercado inmobiliario**. Como consecuencia de la aparición de datos ya digitalizados, son muchas las compañías, eminentemente de carácter PropTech, las que los depuran y ponen a disposición de sus clientes para un uso inmediato. Por tanto, no es que haya más datos disponibles, sino que la mayor cantidad de datos disponibles tienen un formato del

que se puede sacar partido[17]. Y aquí existirían dos circunstancias. O, son datos que naciendo en formato no estructurado o semiestructurado se han depurado y se han convertido para su uso, o, por su parte, se trata de datos que, a raíz de más avanzados procesos, se almacenan directamente en formatos estructurados.

2. **Aparición de nuevos competidores sin experiencia previa en el sector inmobiliario**. Si bien se ha dicho que el nuevo enfoque del sector es el análisis de datos, la carencia de habilidades para ese cometido por el personal inmobiliario ha propiciado que expertos de esta materia (analistas de grandes cantidades de datos) hayan visto dentro de esta industria una gran oportunidad de generación de riqueza. Esto no es más que otra muestra de la falta del desarrollo de aplicaciones tecnológicas propias por parte del sector inmobiliario (Baum, Saul, & Braesemann, 2020, p. 27). El efecto principal de ello es la generalizacion de un contexto con competidores muy dispares en cuanto a la adopcion tecnológica del análisis de datos. Desde agencias inmobiliarias totalmente tradicionales y alejadas de esta realidad[18], hasta nuevas empresas cuya operativa ni siquiera necesita de un contacto físico con el activo inmobiliario -se valen del análisis desde la distancia-. Logicamente, el paso del tiempo permitirá un ajuste de este desequilibrio entre los operadores.

3. **Automatización de procesos manuales**. Seguramente sea la transformación más genérica del sector. La adopción del fenómeno tecnológico puede ser más o menos grandilocuente, pero en términos generales, las tareas y procesos más simples están automatizados, entre ellos, el almacenamiento de datos a nivel básico. De ello deriva que, en el sector inmobiliario, hasta los operadores más rezagados, cuentan con herramientas susceptibles de guardar y proveer de información más precisa que en el pasado. Un ejemplo simple y práctico de esto es que actualmente, cualquiera puede hacerse con un disco duro de 1 *terabyte*, que ocupa menos que un libro de bolsillo y que su precio de adquisición no llega a los 50€.

[17] RICS no matiza expresamente que sean datos estructurados
[18] En España, numerosas agencias fuera de las zonas PropTech funcionan así

Un análisis omnicomprensivo de lo anterior lleva a corroborar posibles tendencias en el corto-medio plazo. En primer lugar, una mayor disponibilidad de datos estructurados permitirá que haya más y más operadores con acceso a este factor productivo y que haya más competencia e igualdad en el mercado. Probablemente las diferencias generadas entre competidores tenderán a reducirse. Ahora bien, en aquellas zonas todavía aversas a la adopción tecnológica, será motivo de importante ventaja competitiva. En segundo lugar, la entrada de usuarios expertos en el análisis de datos y que resultan beneficiados de la escasez de habilidades tecnológicas de quienes operan normalmente en el sector provocará que otros sigan el mismo camino, pero que quienes carecen de ellas busquen adaptarse. Así, la tendencia es claramente hacia un mercado más multifuncional, en el que el experto en inmuebles no se encuentre incómodo en el proceso de análisis de grandes cantidades de datos, y que el experto en análisis de datos cada vez tenga mayor conocimiento del entorno inmobiliario. En tercer y último lugar, la generalización del proceso tecnológico en las funcionalidades más básicas permite que el sector inmobiliario haya asentado las bases para continuar en esa dirección e implementar esos avances en otras facetas y actividades más determinantes y complejas desde el punto de vista empresarial como es el uso del Big Data. Así, lo más seguro, es que la información en formato físico, con el tiempo, tienda a desaparecer siendo la nueva regla general la información de tipo digital, ya sea en formato estructurado o no estructurado.

4.2. Tipos de datos en el sector inmobiliario

Una de las Vs que explicábamos a la hora de contextualizar el concepto de Big Data era la de variedad. Y es que el Big Data no solo es gran volumen, sino que ese volumen muchas veces también tiene su reflejo en una amplia variedad y diversidad de datos. Así, enfrentarse a este apartado, especialmente en un sector tan plural como el inmobiliario, no es tarea nada sencilla. Para evitar caer en una explicación sin final, es preciso abordar la cuestión planteando los tipos más generales y tradicionales de datos que existen en el sector, para luego encontrar subespecies de estos, que siendo cada más utilizados, ayudan a completar a los primeros.

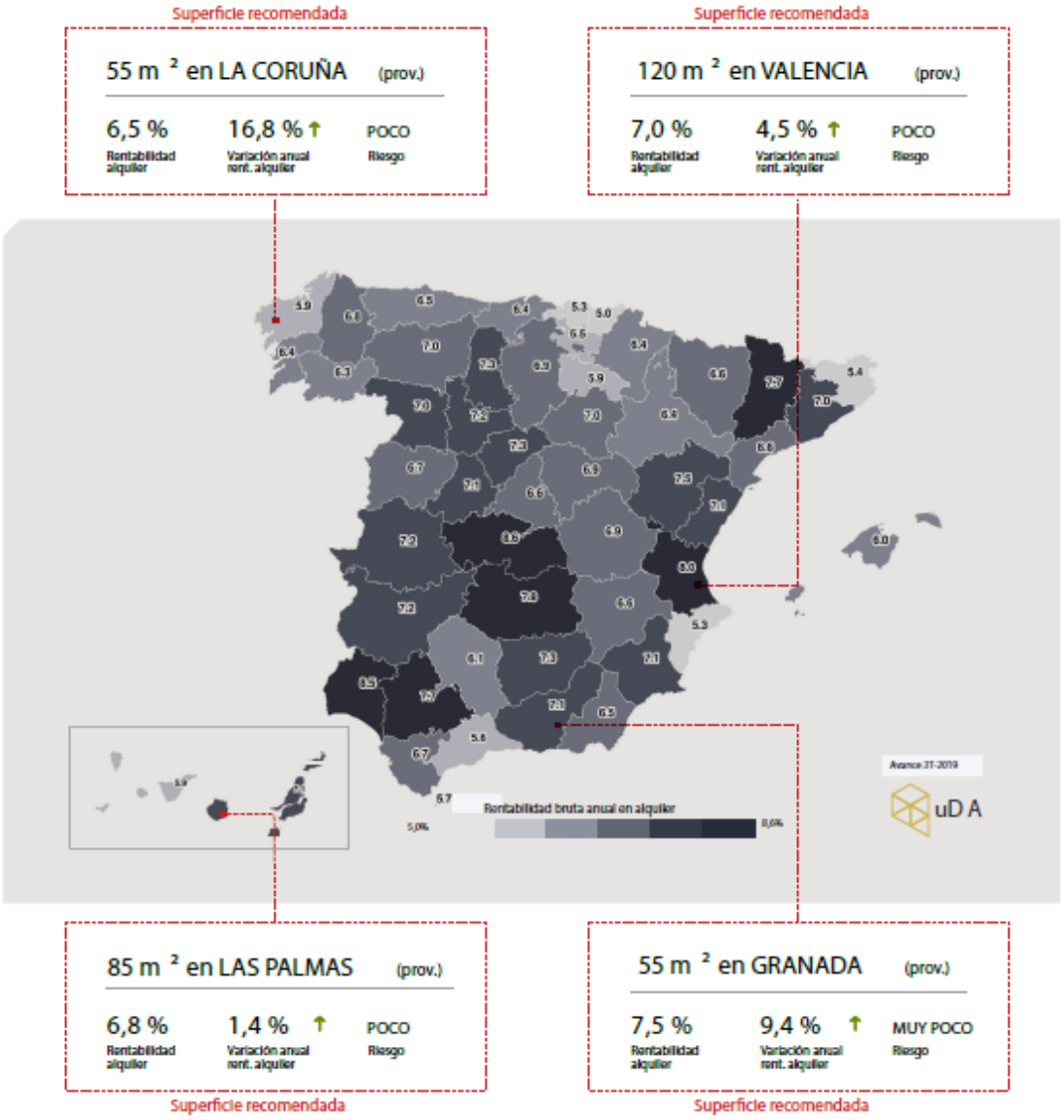
En esta materia Winson-Geideman & Krause (2016, pp. 6-7) hacen una gran labor a la hora de diferenciar los tipos de datos que se pueden encontrar en el sector inmobiliario. Según ellos, el núcleo de los datos (*core data*) que existe en el sector está conformado por tres tipos de datos, a saber:

- **Financiero**. Incluye información relacionada con Socimis[19] y todo lo relativo al mercado de capitales de tipología inmobiliaria. A pesar de estar incluidos, no se profundiza en ellos.
- **Transacciones**. Aunque relacionado con el anterior, este ha de ser correctamente entendido. Incluye una amplia variedad de ámbitos, entre ellos el financiero, pero en genérico se refiere a la información relativa a las transacciones de activos inmobiliarios –precio de compra, día de la misma, alquileres, rentabilidades, ingresos, costes impuestos, hipotecas, préstamos etc.-. Generalmente, a partir de este tipo de información se pretende estudiar el estado del mercado en general, tanto presente como futuro, en términos más macroeconómicos. Ejemplos de ello pueden ser los índices de precios, la tendencia de la actividad inversora o una radiografía de las zonas urbanas con mejores rentabilidades por decir algunos. En nuestro país, es posible que la manera más fácil y rápida de obtener conclusiones rápidas desde un punto de vista práctico sea acudiendo a la página web de Idealista en su apartado *data*[20]. En ella uno puede encontrar indicadores rápidos y sencillos de entender sobre los mercados, incluso a nivel provincial.
- **Comercial**. Engloba información sobre el activo inmobiliario en concreto. Está mucho más enfocado en las características estructurales que tiene un determinado inmueble -metros cuadrados, número de baños, habitaciones, distribución, localización, etc.-. Es con diferencia el tipo de datos menos desarrollado porque al final no todos los inmuebles tienen su registro en una base de datos, pero, sí que es verdad, que es el que más disrupción está generando y el que mejores expectativas de crecimiento presenta. El hecho de que muchos de esos datos no susceptibles de obtención sino a partir de entablar contacto con su propietario, o que su actualización suele ser bastante lenta, provocan dicha situación. Los denominaremos comerciales porque creo que es el conjunto de datos que se suele facilitar a la hora de intentar vender o alquilar un inmueble.

En general, los datos que recogen información transaccional y comercial tienden a integrarse. De hecho, es una tendencia cada vez más usual encontrar informes que incluyen métricas de ambos conjuntos de datos.

[19] Sociedades anónimas cotizadas cuya actividad principal es la adquisición, promoción y rehabilitación de activos de naturaleza urbana para su arrendamiento (elEconomista)
[20] https://www.idealista.com/data/

Figura 9: Rentabilidad bruta del alquiler en España



Fuente: Urban Data Analytics[21]

La **Figura 9** representa perfectamente esta unión. En ella, no solo se explica qué rentabilidades brutas presentan las provincias españolas en el mercado del alquiler, -datos de tipo transaccional- sino que también aportan conclusiones más precisas en cuanto a la tipología del inmueble como cuál es tamaño (metros cuadrados) óptimo para lograr las mejores rentabilidades en esas zonas -datos comerciales-. De este modo, las virtualidades de ambos conjuntos de datos permiten sacar conclusiones más acertadas para el propósito que se plantee.

[21] Obtenida de (Sociedad de Tasación, 2019, p. 13)

Además de esta tendencia cada vez más arraigada, hay que considerar una serie de datos que en los últimos años están entrando con fuerza en el análisis inmobiliario, y que tienen una notable relación con los datos de tipo comercial. Me refiero a los datos obtenidos por los Sistemas de Información Geográfica y que recogen información espacial relativa a las zonas circundantes del inmueble en cuestión. Reciben el nombre de **datos espaciales** (*static spatial data)* y son "*datos que cuantifican cómo es la relación existente entre el activo inmobiliario y las características del exterior más próximo*" (Winson-Geideman & Krause, 2016, p. 6). Por tanto, su distinción con los dos anteriores reside en que no contiene datos sobre el inmueble en sí mismo, sino sobre los alrededores. Dentro de ellos, uno se puede encontrar desde métricas físicas, como la distancia que existe entre ese inmueble y el hospital más cercano, hasta variables demográficas o socioculturales como el PIB *per capita* de ese distrito o su tasa de crimen. Incluso, y esto ya entraría en una categoría distinta, actualmente, se puede conseguir información, sobre los patrones de movimientos poblacionales en una determinada ciudad, y así, determinar, junto con otras métricas, la demanda para un barrio o distrito determinado, su valor real, etc. A estos últimos los conocemos como **datos periféricos** (*peripheral data)*, en general se refiere a datos que incluyen información sobre una determinada zona en cuestión pero que no son estáticos en el tiempo, sino cambiantes, y suelen ser medidos en tiempo real como el tráfico que hay, los ya mencionados movimientos poblacionales de ese distrito o incluso el sentimiento de sus vecinos a través del análisis de sus tweets.

Toda esta arquitectura permite, entre otras cosas, que las empresas del sector inmobiliario tengan en su poder información poderosísima para mejorar su toma de decisiones (Winson-Geideman & Krause, 2016, pp. 6-7). Para ello, resultará decisivo entender las diferentes relaciones existentes entre cada tipología de datos, y aplicarlas correctamente. Por último, y a raíz de lo explicado, es perfectamente posible afirmar que, el poder que podía tener un agente inmobiliario por su conocimiento de una determinada zona en el pasado, disminuya drásticamente, por el cada vez más fácil acceso a los datos y la mayor variedad de los mismo.

En cualquier caso, sigo creyendo que el conocimiento y experiencia del tejido inmobiliario local es un valor que en ningún caso podrá ser sustituido. Por eso, para el que ya es un profesional conocedor de una determinada zona, el análisis de los datos inmobiliarios en masa va a lograr cambios tales como prever con antelación cambios de comportamiento en esa zona, conocer en mayor profundidad lo que ya se conocía o

entender relaciones, que ya existentes, no se habían podido identificar por la falta de análisis.

Tabla 4: Tipos de datos en el sector inmobiliario

Core	Espaciales	Periféricos
Transacciones de venta	Información demográfica	Tráfico
Transacciones de alquiler	Indicadores económicos espaciales	Conexiones a Internet
Características físicas del inmueble	Información censal	Movimientos geoespaciales
Mercado de valores	Servicios disponibles (colegios, hospitales,etc)	

Fuente: Elaboración propia a partir de ilustración contenida en Winson-Geideman & Krause (2016)

4.3. Ventajas de utilizar el Big Data en el sector inmobiliario

Acertadamente Catella Research (2015), cuando aborda las implicaciones que tiene el Big Data en el sector inmobiliario, no duda en afirmar desde un inicio las notables virtualidades que puede tener, pero hace algo que me parece muy relevante para este apartado. Parte de la premisa de que el sector inmobiliario es muy amplio y diverso, no solo en activo, sino también en usuarios. Por ello, no es de extrañar que la proyección que ha tenido y que va a tener el Big Data sobre el sector sea difícil de contextualizar en pocas líneas, y que ello no deje por el camino alguna laguna.

Precisamente, debido a lo anterior, el planteamiento que, a mi juicio, hay que mostrar tiene que venir desde una aproximación más genérica que específica, logrando así abarcar la mayor parte de la operativa inmobiliaria. En este sentido, se ha realizado una revisión de la literatura a partir de Donner, et al. (2018), Royal Institution of Chartered Surveyors (2019) y Catella Research (2015), y de la que derivan las siguientes conclusiones, cada una de ellas acompañada por ejemplos reales para su mejor visualización:

1. **Identificación y minimización de riesgos**. En el momento que un operador tiene a su disposición una mayor cantidad de información, el riesgo que implica tomar decisiones disminuye, y la anticipación es más sencilla. Esto tiene su reflejo en muchos y diversos ámbitos del sector inmobiliario, por lo que las distintas posibilidades de mitigación de riesgo adoptarán distintas formas en función del ámbito que abordemos. De esta forma, un inversor que compra para alquilar tenderá a hacerlo en aquella zona que tenga unas perspectivas optimas desde el punto de vista

de la rentabilidad, pero quizá decida no hacerlo en esa zona porque el tiempo medio necesitado para conseguir venderlo en un futuro es demasiado largo para el caso de así hacerlo. Esto, en otras palabras, significa que el Big Data da la posibilidad, gracias a la variedad de sus métricas (datos core, espaciales y periféricos) adoptar decisiones que minimizan el riesgo en función de muchos más parámetros. No solo voy a obtener una rentabilidad mayor en ese lugar, sino que voy a poder vender el inmueble más rápido para así dar otro uso al capital invertido. Mismo razonamiento para un fondo de inversión que busca repetir dicha operación, pero a una escala mayor.

2. **Mejor proyección de preferencias**. Igual que si de una tienda se tratase. Controlar las tendencias del mercado se antoja clave en cualquier sector. Pues bien, si esto lo llevamos a un sector como el inmobiliario, donde las necesidades de capital son notorias, contar con proyecciones más precisas que aportan valiosa información será un pilar clave (Donner, Eriksson, & Steep, 2018, p. 14). Para este cometido, el Big Data se presta como una herramienta perfecta, a partir de la cual, encontrar patrones de comportamiento y de preferencias se convierte en algo mucho más asequible. De igual manera que con la anterior ventaja, los beneficiados por esta ventaja son muchos y distintos. Desde promotores eligiendo el lugar adecuado para edificar, hasta el caso del propietario individual que deciden cambiar la modalidad de alquiler en función de la movilidad que presentan potenciales inquilinos, pasando por entidades financieras a la hora de proceder con las proyecciones de entrada de capital.
3. **Valoraciones más rápidas y acertadas**. Cuando hablamos de valoración no solo nos referimos a predecir el precio de un inmueble, que también, por su notable importancia en las transacciones. Nos referimos, más bien, al valor intrínseco que desprende esa propiedad en particular. En este campo, los métodos tradicionales de valoración suelen pecar de subjetividad, lo cual es algo normal. Sin embargo, aquí el Big Data viene a jugar un papel muy relevante permitiendo a quien realiza la valoración tener una imagen más clara del valor actual y ayudar, así, a poder predecir, por ejemplo, el precio que tiene esa vivienda, ahora y en el futuro. Otra ventaja importante es que el proceso de búsqueda de información para llevar a cabo dichas valoraciones necesitará de mucho menos tiempo. De esto, se pueden inferir varias conclusiones. Por un lado, el Big Data posibilita aumentar el conocimiento, sin sustituir el que ya existe. Pero, por otro lado, al conseguir que los procesos ganen en automatización, es posible que "*el número de profesionales dedicados a este campo*

tienda a descender con el paso del tiempo" (Royal Institution of Chartered Surveyors, 2017, p. 20).

4. **Búsquedas más rápidas y precisas**. Tratándose de una de las actividades más cotidianas del entorno inmobiliario, es factible que sea de las más beneficiadas. Contar con una mayor cantidad de datos y sobre todo, con una mayor diversidad de los mismos, permitirá que las búsquedas se caractericen por una alta personalización y adaptación al cliente. En mi opinión, esto resulta en una ventaja muy asimilable por compañías más pequeñas como las inmobiliarias tradicionales. Muchas de estas llevan a cabo un trabajo de campo *in situ* muy sacrificado, y contar con información previa que permita desde un principio filtrar las propiedades que pueden ser de interés resulta muy ventajoso. Es más, ellos mismos, los agentes, a la hora conformar su cartera de inmuebles (aquellos que pueden poner a la venta o alquiler como intermediarios), lo tendrán mucho más fácil. En otras palabras, los agentes inmobiliarios tendrán la posibilidad de resaltar de forma rápida y asequible qué inmuebles pueden encajar con los requisitos de su cliente, o de ellos mismos.

Aun siendo posible hacer más profundo este análisis, por lo explicado al inicio de este apartado, estructurar de forma clara y concisa las principales ventajas que de forma transversal trae el Big Data al sector inmobiliario parece la manera más adecuada de comprender sus efectos.

4.4. Desafíos futuros

Siendo las ventajas "un claro un pilar" de lo que significa aplicar el Big Data en el sector inmobiliario, no se puede ni mucho menos dejar de lado las perspectivas futuras y los desafíos que estas puedan entrañar.

Es necesario partir de la premisa que ya hemos alertado antes. Con el paso del tiempo la disponibilidad y obtención de los datos comienza a ser mucho más asequible provocando, entre otras cosas, que el hecho de ostentarlos o de simplemente tener acceso a los mismos haya perdido valor, y que el verdadero interés ahora resida en ser capaz de utilizarlos (Royal Institution of Chartered Surveyors, 2019, p. 25)

Es, precisamente ahí, donde actualmente reside el gran desafío para el sector. Ostentar las habilidades necesarias para convertir una gran masa de datos inmobiliarios en información con la que se pueda trabajar y sacar partido es la clave que informa a la

industria en este sentido. Aun así, para muchos de los operadores de la industria, ni siquiera la disponibilidad de estos es sencilla. Por estas razones, el *status quo* del sector inmobiliario es guiado, principalmente, por importantes dificultades en el uso de los datos, pero también, aunque ya venido a menos, por los obstáculos que sigue habiendo para obtener los datos con los que poder discernir el verdadero valor que hay dentro de ellos.

Teniendo en cuenta lo anterior, y guiado por la última actualización realizada por la prestigiosa *Royal Institution of Chartered Surveyors*, los desafíos a los que de forma transversal se enfrenta la industria inmobiliaria en relación con el uso del Big Data se podrían resumir en los siguientes:

1. **Alto coste de los datos**. A pesar de la facilidad de obtener datos sobre propiedades, es muy cierto que su calidad no siempre es la más adecuada. Por esta razón, muchas empresas, especialmente en Estados Unidos[22] han sacado partido de esta situación planteando nuevos modelos de negocio basados en el procesamiento de datos en bruto, y dando lugar a información inmobiliaria muy valiosa (Royal Institution of Chartered Surveyors, 2019, p. 25). Un ejemplo de ello es Zillow, el portal inmobiliario líder del mercado americano que, a partir de información disponible para el público, provee de un listado de valoraciones inmobiliarias. En otras palabras, que los datos sean costosos no es tanto una cuestión de accesibilidad, sino de procesamiento de los mismos. Esta circunstancia plantea claramente un desafío a corto plazo para el sector, que no es otro que reducir el coste de los datos, ya sea por un acercamiento de los demás operadores a estos nuevos modelos de negocio que haga reducir la dependencia que hay en ellos, o por el surgimiento o fundación de más empresas que hagan aumentar la oferta existente de este tipo de servicios. En cualquier caso, la realidad es muy dispar, y dependiendo de la región en la que nos movamos, esta situación tendrá una mayor o menor profundidad.
2. **Aumento del conocimiento analítico de datos**. Tradicionalmente, dentro del sector inmobiliario, el análisis de variables de carácter macroeconómico, y en menor medida el análisis de datos, generalmente contenidos en bases con cantidades de inmuebles muy reducidas, han conformado la hoja de ruta de los profesionales inmobiliarios (Royal Institution of Chartered Surveyors, 2019, p. 25). De ello se deduce un notable

[22] Un ejemplo claro de ello es Zillow, el portal inmobiliario líder en Estados Unidos

estancamiento y falta de flexibilidad para adaptarse a las bases de datos tipo que existen hoy en día, y las cuales son mucho más variadas en contenido, y mucho más cuantiosas. Paliar esta situación tiene dos respuestas bastante simples y obvias en la teoría, pero algo más difíciles en la práctica. Una primera es que los profesionales que actualmente se encuentran presentes en la industria comiencen un proceso de adaptación y entrenamiento que los lleve a reducir esa brecha. Y una segunda es que el sector gire sus miras, en la atracción de talento, hacia colectivos con formación más científica, tecnológica, ingenieril o matemática, que de negocio. Respecto de esta segunda posibilidad, remarcar, la tendencia que ha venido surgiendo estos últimos años en el ámbito académico, y que consiste en mezclar la formación enfocada en el ámbito empresarial con conocimientos propios del análisis de datos y de las nuevas realidades digitales[23]. Nuevas perspectivas como esta permitirán que este desafío desaparezca, y que lo haga más pronto que tarde.

3. **Mejorar en la calidad de los datos generados**. Uno de los grandes problemas a los que ese enfrenta la industria inmobiliaria es disponer de datos ya aptos para su uso, que no necesiten de un esfuerzo excesivo en su procesamiento o limpieza. Consecuencia de ello es que actualmente muchos de los profesionales dedicados al sector inmobiliario sean aversos a adentrarse en esta nueva aventura que es el Big Data inmobiliario, y por el contrario, prefieran continuar trabajando como han hecho hasta ahora (Royal Institution of Chartered Surveyors, 2019, p. 26). Lidiar con este desafío no es tarea sencilla. Al final grandes cantidades de datos son generadas a través de portales inmobiliarios, y almacenados en bases de datos una vez los inmuebles son listados por el propietario o arrendador. He ahí el origen del problema. Por un lado, quien procede a publicar el inmueble puede introducir métricas erróneas o ni siquiera completar todos los apartados que se piden aminorando la veracidad y calidad de la base de datos (Paluri, 2016). Y por otro lado, los propios portales inmobiliarios no siempre coinciden en las variables requeridas a quien quiere listar una propiedad, es decir, existe una falta de estandarización entre ellos que implica que el acceso a información relevante de una manera eficiente -sin un procesado excesivo- no sea tarea sencilla (Menin Machado, 2019, p. 79). En otras palabras, aquel que se

[23] Un ejemplo de ello son los Dobles Grados en Análisis de Negocio/ Business Analytics, y Derecho o ADE impartidos por la Universidad Pontificia Comillas y que permiten complementar el perfil profesional con aptitudes propias del análisis de datos y del mundo digital.

disponga a poner en común datos cuya procedencia sea distinta necesitará de un trabajo de campo previo realmente exhaustivo.

4. **Distintas formas de almacenamiento**. Aun siendo una cuestión bastante genérica es obvio que el sector inmobiliario también se verá afectado. La generación exponencial de datos precisará en el futuro de un poder de almacenamiento mucho mayor al que tenemos hoy en día. Sin embargo, esto hace que se cuestione el modelo de almacenamiento que tenemos hoy en día, especialmente en términos de sostenibilidad. La creación de una base de datos común y centralizada, en la que todos los operadores y actores contribuyan y saquen partido, parece la opción más adecuada al desafío que se nos plantea (Menin Machado, 2019, p. 81). No solo se gana en sostenibilidad, requiriendo de un menor coste de mantenimiento, sino que también parece la mejor opción en cuanto a diversificación, precisión y cantidad de los datos (Menin Machado, 2019). Incluso me atrevo a vaticinar que esta posibilidad puede ganar enteros si se realiza por regiones geográficas, paliando así las posibles carencias que pueden tener los pequeños operadores a la hora de obtener datos. Aun así, salvo que alguna de las situaciones anteriores se produzca, saber discernir entre aquellos datos que pueden seguir teniendo una utilidad y aquellos que han dejado de tenerla se convertirá en un tarea esencial del proceso.

5. ANÁLISIS PREDICTIVO EN EL SECTOR INMOBILIARIO

5.1. *Predictive Analytics*

Una vez los datos son recogidos, las labores de organización y análisis de los mismos juegan una suerte de etapa esencial en el cribado que nos permite en última instancia obtener la información que deseamos. Mediante ese proceso, disgregado en otras etapas como son el filtrado, la visualización o la simulación de escenarios futuros, la industria en su conjunto tiene la capacidad de conocer cuáles son las tendencias del mercado, predecir con mayor o menor certeza los potenciales resultados y en consecuencia, lograr el objetivo último de ser capaz de tomar decisiones mejor contrastadas (Chaillou, Fink, & Gonçalves, 2017, p. 41).

Probablemente la capacidad de estimar con gran certeza escenarios futuros se haya conformado como la gran disrupción que ha sufrido el sector inmuboliario desde que el Big Data comenzase a ser la norma general en la industria. A las fases de agregación y análisis de datos se suma la predictiva (comunmente conocida como ***predictive analytics***) que en términos más coloquiales viene a ser el verdadero "salto de calidad" que, eso sí, proyecta sus efectos tanto en el sector inmobiliario como en muchos otros. Con ello se trata de hacer ver, que el *predictive analytics* se conforma como una parte muy importante del abrumador contexto del Big Data, que supone,"*no tanto una revolución de cantidad o de escala (volumen de datos, poder computacional), como una disrupcion de inteligencia"* (Chaillou, Fink, & Gonçalves, 2017, p. 42). Es en esos términos como mejor se puede definir y entender la llegada del análisis predictivo. La tranformaciones de escala y cantidades masivas de datos han dejado paso al análisis inteligente, que de verdad proporciona un verdadero valor añadido.

Centrando más el foco, y a pesar de haber sido usualmente estigmatizada como poco ambiciosa y conservadora, la industria del inmueble se ha volcado en el entorno del análisis predictivo. Actualmente el mercado inmobiliario, a través del análisis predictivo, es capaz de pronosticar una infinidad de ámbitos, a saber: predecir el precio de un inmueble, pronosticar la rotación de arrendatarios, o simplemente plantear las probabilidad con la que un inquilino determinado podrá hacer frente al pago del alquiler mensual. Una gran variedad de campos, todos ellos entroncados dentro del sector, cuyas predicciones ganan importancia cuanto mayor es la amplitud de los rangos de tiempo

pronosticados y cuanto mayor es el nivel de detalle que presentan los datos -granularidad del dato- (Chaillou, Fink, & Gonçalves, 2017, p. 42). Eso sí, de las infinitas posibilidades que entraña el análisis predictivo es importante matizar que hay una que ha ganado enteros por encima del resto: la valoración del inmueble para predecir su precio de venta. Un desarrollo que incluso ha traído consigo todo una movimiento en este sentido, conocido con el nombre de "Modelos de Valoración Automática" (*Automated Valuation Models)* el cual se puede definir como *"la aplicación de una o más técnicas matemáticas, para obtener en una fecha específica la valoración estimada de un inmueble, eso sí, acompañada de una métrica que determine la confianza de la precisión de dicho resultado y sin requerir, una vez iniciada, la intervención humana"* (Royal Institution of Chartered Surveyors, 2017, p. 24).

5.2. Tecnologías subyacentes y su funcionamiento

Aunque no es ni mucho menos objeto de este trabajo explicar las notas científicas que caracterizan a los modelos predictivos, sí que es relevante, para un mejor entendimiento, conocer cuál es su armazón y los cimientos tecnológicos sobre los que se asienta. De forma genérica, el mundo del análisis predictivo en aplicación al sector inmobiliario puede dividirse en tres grandes técnicas: *Machine Learning* (ML), *Neural Networks* (NNs), y *Deep Learning* (DL). Son tres métodos avanzados de modelado estadístico que siguen, como norma general, el mismo patrón, -predecir el futuro mediante el análisis de sucesos presentes y pasados- principalmente porque cada uno de ellos es un subconjunto más pequeño y específico del anterior (Nguyen, et al., 2019, p. 79). Por esta razón, a continuación se describen brevemente las fases que constituyen el proceso del análisis predictivo desde el punto de vista de la técnica del ML.

1. Recolección de datos. Parece una fase lógica y obvia, pero es importante escoger los datos correctos en función de lo que se quiere observar, identificar, y predecir. De este modo, se puede acudir a distintas bases de datos que contengan esa información. Los datos pueden encontrarse en distinto formato -estructurado, semiestructurado o no estructurado-. Por ejemplo, para la aplicación práctica del apartado 6 se han recopilado inmuebles en venta de la base de datos que tiene el portal inmobiliario Idealista.
2. Análisis exploratorio y amasado. Los datos recopilados deben ser analizados mediante un análisis exploratorio de datos (*Exploratory Data Analysis)*, que nos dé

una idea de cómo son. Gráficas, distribuciones y métricas estadísticas son herramientas muy comunes en esta fase. Gracias a esta parte del proceso, se podrá proseguir con el "amasado" de datos. Consiste en convertir los datos obtenidos en origen en datos aptos para ser utilizado por las herramientas analíticas (*future engineering)*. Es un proceso arduo, pero a la vez esencial y determinante en la efectividad o no de las predicciones (Kumar & Garg, 2018, p. 32). Aplicado al sector inmobiliario consistiría, entre otros, en determinar qué variables son las más significativas en la predicción del precio, proceder a "enjugar" las variables con valores ausentes (NA), o incluso realizar transformaciones del tipo de dato que conforma cada variable (por ejemplo, *one-hot encoding)*.

3. Modelado (entrenamiento). Una vez se tiene la estructura de datos óptima comienza la fase de entrenamiento. En esta parte del proceso se aplica a los datos de entrenamiento -obtenidos de una división previa de todos los datos- distintos algoritmos de forma repetitiva, para poder encontrar en cada uno de ellos los parámetros más acertados en la predicción del precio del inmueble (*fine-tuning*). Es decir, se trata de una fase en la que "entrenamos" al algoritmo para que "aprenda" la complejidad existente en el set de datos, y consiga ponderar el impacto que tiene cada variable -localización, número de habitaciones, tamaño, etc.- en nuestra variable respuesta -precio del inmueble- (Chaillou, Fink, & Gonçalves, 2017, p. 42). De ahí, que se llame *Machine Learning* (maquina de aprendizaje). Al final, el objetivo primario de este apartado es conseguir encontrar, mediante la intermediacionn de distintos modelos algorítmicos, los patrones de comportamiento de las características de los inmuebles que permitan predecir con mayor precisión su precio. Por esta razón, cuanto más grande sea la base de datos, mayor terreno de aprendizaje, y por ende, mejor resultado se obtendrá en la predicción.
4. Evaluación o testeo. De la partición de los datos siempre se reserva una para verificar que el algoritmo, previamente entrenado, funciona también en otros datos de los que desconoce su variable respuesta, pues no ha tenido un proceso de aprendizaje previo con ellos, como sí sucede con el set de entrenamiento. Es en esta fase en la que podemos determinar cómo de preciso es nuestro modelo en la predicción del precio del inmueble. Estas dos fases, la de entrenamiento y testeo se intercalan una con otra, y es muy normal realizarlas de forma repetitiva hasta encontrar la calibración ideal del algoritmo.

5. Implementación. Viene a ser como su salida a la “vida profesional”. Igual que un trabajador, antes de saltar al mercado laboral, primero estudia y trabaja durante sus años como estudiante, sin tener un contacto total con la realidad profesional, los modelos predictivos primero se entrenan, testean y calibran, antes de utilizarse para explotar sus funcionalidades en casos reales, como sería el de una inmobiliaria asesorando a un cliente sobre el tipo de inmueble a comprar, el precio adecuado a pagar o las probabilidades de revaloración o devaluación, por poner algunos ejemplos.

5.3. El fin del dilema tiempo-precisión

Con anterioridad al advenimiento del ML, las predicciones de precios solían llevarse a cabo mediante simples modelos de regresión lineal que normalmente se enfrentaban a la cuestión de lo que me gusta denominar el dilema tiempo-precisión del análisis predictivo. O bien, la predicción era relativamente precisa, pero limitada a rangos temporales reducidos de unos pocos meses, o bien, se prefería reducir la precisión en la predicción, a cambio de incrementar el espacio temporal predicho.

Pues bien, la perspectiva de dicho dilema se transforma radicalmente con la venida del ML, aunque, si bien es verdad, el núcleo de la cuestión sigue siendo la precisión. Mediante el entrenamiento automatizado de los algoritmos se consigue no solo una mayor precisión del precio, sino además lograrla en espacios temporales más amplios (Chaillou, Fink, & Gonçalves, 2017, p. 43)

Sin emabrgo, incluso con el ML sigue existiendo cierto dilema entre el espacio de tiempo, y la precisión del precio, que se cierne en torno a una propiedad inherente a la industria inmobiliaria como es la tipología (vivienda, garaje, local, piso, chalet, etc). Si se quiere ampliar el rango temporal de prediccion sin sacrificar precisión, es más sensato solo utilizar datos sobre un tipo de inmueble, pero ello conllevará logicamente no poder entender la posible interrelación entre los distintos tipos de propiedades (Chaillou, Fink, & Gonçalves, 2017, p. 43).

6. APLICACIÓN PRÁCTICA

Una vez se ha detallado el impacto que ha tenido la llegada del Big Data inmobiliario, en el presente apartado se dará un paso más. Muchas de las cuestiones que se han planteado *supra* tendrán ahora su plasmación práctica de tal forma que el lector pueda obtener una visión total y panorámica de lo que es ya para muchas empresas el nuevo sector inmobiliario y lo que acabará siendo para otras muchas, que aún no han emprendido el camino de la transformación digital en su totalidad. De hecho, el desarrollo de este proyecto tiene su razón de ser en ese último grupo de empresas, que en su gran mayoría vienen a ser inmobiliarias de pequeño tamaño localizadas en ciudades donde el movimiento PropTech está lejos de ser la norma general.

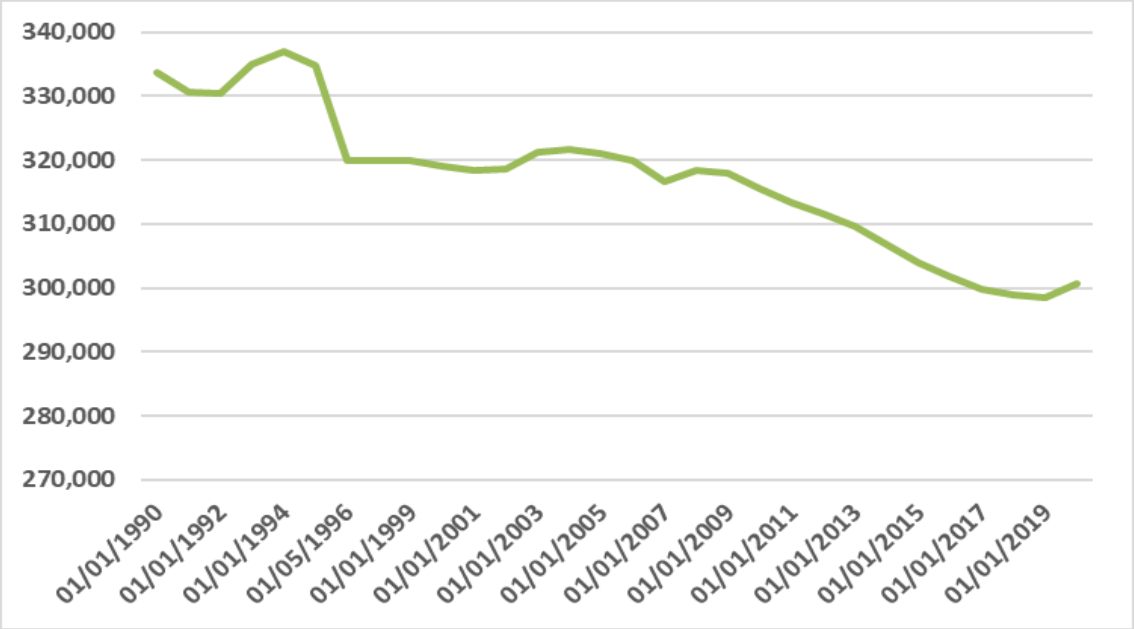
Eso sí, lejos de un ánimo científico en su puridad, de lo que se tratará de plantear es, desde un punto de vista pragmático, cómo el acogimiento de las nuevas tecnologías para estos operadores no es un imposible, y transmitir la convicción de que las oportunidades de negocio inherentes en su adopción son innumerables, estando algunas de ellas aún por descubrir. En definitiva, un despliegue cercano, con el que los no expertos en estos campos se sientan identificados y desafiados, y que sirva a más de uno como punto de partida.

6.1. Explicación y objetivos

El presente proyecto tiene por objeto analizar la oferta de inmuebles residenciales que existe actualmente (marzo-abril 2020) en la ciudad de Valladolid como si de una inmobiliaria pequeña con pocos recursos se tratara. La ciudad de Valladolid, capital de la Comunidad Autónoma de Castilla y León, ha sufrido, aproximadamente, un descenso de la población del 10% en los últimos treinta años, dejando el número de habitantes en poco más de 300.000. Es decir, estamos hablando de una ciudad, de tamaño medio, que actualmente ocupa la posición decimotercera en la clasificación de ciudades españolas por número población de acuerdo con el INE[24](**Figura 10**), y que como bien se explicó en la sección del PropTech en España (ver apartado **España**), no concentra de ningún modo un punto geográfico desdeñable desde la perspectiva de inmobiliaria tecnológica.

[24] https://libretilla.com/ciudades-mas-grandes-espana/

Figura 10: Evolución poblacional de la ciudad de Valladolid (Serie 1990-2020)



Fuente: Elaboración propia a partir de datos del Ayuntamiento de Valladolid

A mayores, es importante señalar que un miembro de mi familia es un profesional ligado a este sector (agente inmobiliario) regentando ciudad de Valladolid una inmobiliaria. De dicha inmobiliaria resaltar que no tiene punto físico de atención al público en general, y que toda la operativa se realiza de forma digital, y a distancia (redes sociales, portales inmobiliarios, etc.). Así, los recursos con los que cuenta no son más que un ordenador, conexión a Internet, y la experiencia de estar presente en el mercado vallisoletano desde el año 2015.

Mediante este planteamiento previo se pretende poner al lector en la posición de quien tiene una inmobiliaria con características análogas -pequeña, con pocos recursos, y sin aversión tecnológica- y cuya localización sea la de una ciudad asimilable en términos poblacionales a la de Valladolid. Esto último es muy importante, pues, en función de la población, así será la oferta que nos encontremos. No obstante, con ello no quiere decirse que lo que se plantee aquí no tenga una practicidad en otros escenarios, pero sí que los resultados que se obtengan puedan diferir bastante en función del lugar en el que se encuentre sita la inmobiliaria.

El análisis, como ya se ha avanzado, se llevará a cabo mediante la utilización de técnicas propias del análisis de datos que cualquiera puede adquirir con esfuerzo e interés,

y a partir de una base de datos obtenida de Idealista (el portal inmobiliario online, líder en compraventa y alquiler de inmuebles en España) de forma totalmente gratuita.

Por su parte, tres serán los objetivos técnicos a lograr en este proyecto:

1. Crear un **sistema automatizado de obtención de datos**, siempre con las limitaciones impuestas por Idealista (se explicará más adelante). Los datos en este caso estarán referidos a inmuebles residenciales que estén en venta en Valladolid y que, al tiempo de su obtención, estén publicados en el portal inmobiliario.
2. Obtener de forma automatizada una **imagen global de la situación del mercado inmobiliario de la ciudad de Valladolid**. Para ello, se crearán distintas gráficas a partir de los registros obtenidos de Idealista, que permitirán dar una idea general, tanto en términos económicos, como puramente técnicos de la oferta residencial en Valladolid. Esto es lo que denominaremos, con un regusto más propio de la jerga del análisis de datos, como análisis descriptivo -*Descriptive Analytics-,* el cual se desarrollará en la parte del EDA (*Exploratory Data Analysis)*.
3. Desarrollar un **modelo algorítmico que consiga predecir los precios** de los inmuebles en venta con la máxima precisión posible. Es en esta parte, donde se desplegará en gran medida lo explicado en el apartado de análisis descriptivo aplicado al sector inmobiliario, y en el que aplicaremos técnicas propias del aprendizaje automático -*Machine Learning*-.

A partir de esos tres objetivos, lo que se pretende, en última instancia, es transmitir la idea de que un negocio pequeño y de poca base tecnológica, con un poco de esfuerzo y dedicación, puede lograr dar un primer paso muy importante para futuros desafíos en los que, sí o sí, la tecnología va a ser la "punta de lanza". Además, al tratar un sector que aún padece, en muchas partes de la geografía española, de una palmaria falta de inmersión tecnológica, lograr un paso adelante de estas características supondría una gran oportunidad -quizá incluso única- de poder consolidar una enorme ventaja competitiva con respecto de otros competidores de ese mismo entorno.

6.2. Metodología

Como se ha preavisado, la realización de este proyecto, en su totalidad, tendrá en consideración la partitura ficticia de una inmobiliaria pequeña y con pocos recursos. Tanto es así, que el proyecto se desarrollará sin coste alguno, con un solo ordenador, y

sin el requerimiento de un conocimiento amplísimo en la materia. Requisitos, todos ellos, perfectamente asequibles para una inmobiliaria de esas características.

La nota más técnica y científica la pondrá el lenguaje de programación estadístico R o R software, el cual será ejecutado en el entorno de desarrollo integrado **RStudio** (*integrated development environment*). Si extrapolamos esta relación a un ejemplo sencillo de un vehículo quedaría de la siguiente manera: mientras que R vendría a ser el motor del coche, RStudio sería el control de mandos, el navegador, los pedales o la palanca de cambios para que ese "motor" que es R pueda funcionar y circular. Para evitar complicar la cuestión nos referiremos a él indistintamente como R o RStudio.

Figura 11: Logos de R y Rstudio

Icon for R

Icon for RStudio

Las razón por la que utilizamos R en vez de otros lenguajes de programación como pueden ser Python, C++ o JavaScript, por poner algunos, estriba principalmente en el hecho de tener un conocimiento más avanzado del primero. Además, R cubre todas las necesidades de un proyecto de estas características. Provee de una gran variedad de técnicas estadísticas y herramientas de representación gráfica que son óptimas para el fin que persigue este proyecto. Además, téngase en cuenta que R, y siguiendo la tesis de "proyecto a coste cero", es un software computacional de código fuente abierto -*open-source*-, que implica que sea públicamente accesible (gratuito), y que pueda ser desarrollado y distribuido de una manera descentralizada por la comunidad en su conjunto, en vez de por un solo autor (Red Hat, s.f.). Dicho de otra forma, sin ser expertos

en programación podemos utilizar todo el potencial de R gracias al desarrollo que han realizado otras personas, y además, sin coste alguno.

De modo esquemático, el proceso que seguiremos será el siguiente:

1. **Recolección de datos**. Los datos serán obtenidos a partir de la API de Idealista. Una API, que son las siglas anglosajonas de "*apliccation programing interface",* será algo así como el intermediario que permitirá la conexión entre nuestro entorno de programación (Rstudio) y la aplicación determinada con la que queramos iniciar una conexión (base de datos de Idealista).
2. **Preprocesado previo de la base de datos**. Es quizá la parte más monótona y rudimentaria del proyecto. Algunas variables provistas por Idealista son descartadas por diferentes motivos que no hacen óptimo su uso para el proyecto y otras son transformadas.
3. **Análisis exploratorio de datos** (EDA). Este apartado nos ayudará en dos circunstancias determinadas. Por un lado, desde la perspectiva más de negocio, nos permitirá mediante distintas visualizaciones entender las tendencias actuales que hay en el mercado inmobiliario residencial de Valladolid (algo muy valioso para cualquier inmobiliaria pequeña). Por otro lado, nos asistirá para ir conociendo el comportamiento de las variables (características de los inmuebles) y saber las claves que deberán guiar la conversión de los datos obtenidos en origen en datos aptos para ser utilizados por las herramientas analíticas (*future engineering*).
4. ***Future engineering***. Algunas de las variables se someterán a distintos procesos de transformación para que el dataset (conjunto de datos) sea compatible con los requisitos que tienen los algoritmos de ML, y también, para que mejore su poder predictivo del precio de los inmuebles.
5. **Selección de predictores**. Con ello se trata de conocer cuál es el número de variables óptimo que debe tener nuestro modelo antes de ejecutar los algoritmos, y, especialmente, de identificar aquellas más importantes respecto de la predicción del precio de las viviendas.
6. **Entrenamiento de modelos predictivos y modelado.** Dividiremos el dataset en dos partes -entrenamiento y test-. Sobre la primera de las partes ejecutaremos varios algoritmos de ML. Es la parte más crítica, pues se debe encontrar, no solo el modelo que mejor se adapte a los datos que tenemos, sino también su configuración idónea. Se entrenarán 4 algoritmos diferentes (*Support Vector Machine*, *K-Nearest Neighbor*,

Random Forest y *Gradient Boosting*) cuyo funcionamiento se explicará a medida que se vayan aplicando.

7. **Evaluación**. Las predicciones entrenadas del conjunto de entrenamiento son ejecutadas sobre el conjunto de test. Esto nos permitirá contrastar la capacidad de nuestro modelo predictivo con unos inmuebles que "no ha visualizado" previamente.
8. **Comparación de modelos**. Aunque en este punto ya conoceremos qué modelo algorítmico es el que mejor resultado ofrece en la predicción de precios, sí que es bueno comparar visualmente su capacidad predictiva. Esto se logra a partir del error medio que cometa cada algoritmo en la predicción. Cuanto menor sea el error, mejor será el modelo predictivo.

6.3.Aplicación práctica en una inmobiliaria de la ciudad de Valladolid

En este apartado, explicaremos en detalle cada una de las fases del proceso analítico que hemos seguido para el tratamiento de los datos inmobiliarios de Idealista. Las conclusiones que se obtengan tratarán, en la medida de lo posible, de ser utilizadas para dar respuesta a problemas de negocio que se encontraría en el día a día nuestra inmobiliaria ficticia. No obstante, al adentrarnos en el ámbito de uso de las nuevas tecnologías, también acudiremos a explicaciones algo más técnicas pero que, en ningún caso, eviten entender las bases del proyecto.

6.3.1. Obtención de datos

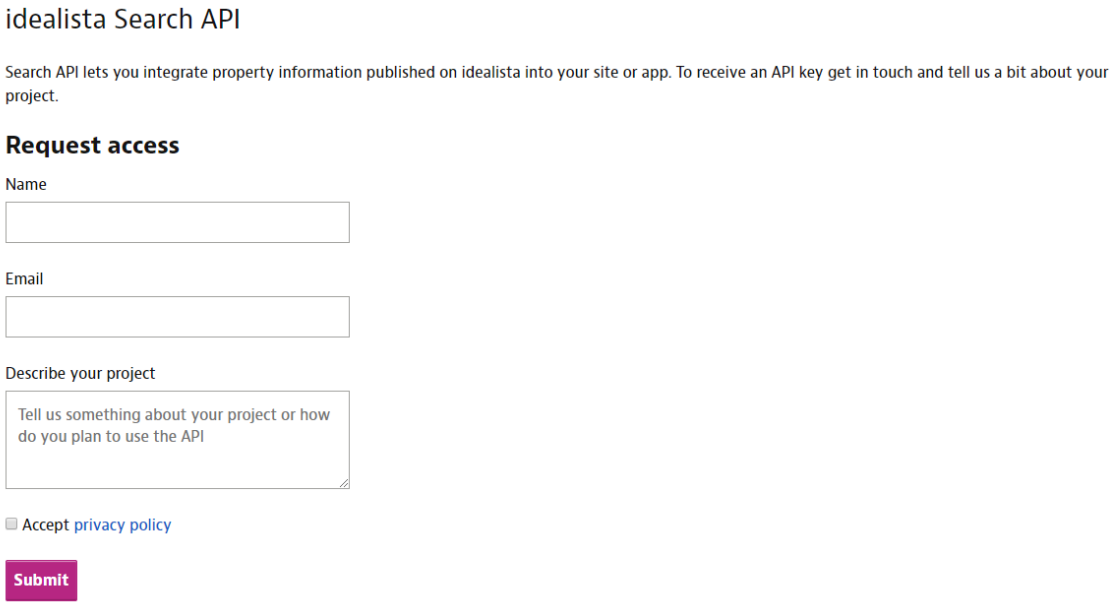
El dataset sobre el que se trabajará se ha obtenido a través de la API de Idealista durante los meses de marzo y abril de 2020. Los datos que Idealista proporciona son de aquellos inmuebles que, en el momento de la solicitud, se encuentran publicados. Es decir, si en mayo de 2020 quisiéramos volver a solicitar esos mismos datos podrían haber sufrido cambios, como consecuencia de una alteración del precio, por ejemplo, o desaparecido, porque se ha logrado su venta, o porque se ha desistido de ella.

Para que Idealista facilite este tipo de datos es necesario solicitar un acceso de entrada en su propia web[25] **(Figura 12**). En nuestro caso, el mismo día de la solicitud nos dieron acceso, pero como su API es de arquitectura tipo RESTful (utiliza HTTP para procesar las solicitudes), con protocolo de seguridad *OAuth 2.0 Authentication*, también

[25] http://developers.idealista.com/access-request

se nos facilitó una clave -*API key*- y una contraseña -*secret*-. El hecho de que accedamos al conjunto de datos mediante ese tipo de protocolo es porque nuestro acceso está limitado. En concreto, podemos tramitar hasta 100 solicitudes al mes, pero en cada una de ellas se devuelve un conjunto de datos con un máximo de 50 propiedades. En el supuesto de necesitar más solicitudes es tan sencillo como ponerse en contacto con Idealista.

Figura 12: Solicitud de acceso a la API de Idealista



Fuente: Idealista.

Además de clave y contraseña, también se nos facilitó la documentación necesaria para proceder a diseñar el código de llamada a la API, y así obtener los datos que se tratarán en el proyecto. Para realizar dicha llamada fuimos ayudados por una consulta realizada en 2017 en la comunidad de Stack Overflow[26]. En ella, explícitamente se contenía un código en R para realizar solicitudes a la API de Idealista. Aun siendo éste antiguo, simplemente tuvimos que cambiar clave y contraseña y realizar las búsqueda que queríamos en función de unos filtros determinados[27]. La API de Idealista tiene dos tipos de filtros. Unos generales para todo tipo de inmuebles (país, tipo de operación, tipo de propiedad, localización de la propiedad, rangos de precios, etc.) y otros más específicos en función de la tipología del inmueble (garajes, locales, viviendas, oficinas, y habitaciones).

[26] Comunidad abierta para cualquier persona que programe
[27] https://stackoverflow.com/questions/46668113/cannot-make-request-from-get-in-r-to-idealista-api

Nosotros realizamos la siguiente búsqueda: viviendas en venta en un radio de 6 kilómetros desde la plaza mayor de Valladolid. Utilizar como punto de partida la plaza mayor de Valladolid (en coordenadas) tiene su razón de ser por dos motivos. De un lado, se entiende que es la zona más céntrica de la ciudad, y por tanto, nos permitirá obtener la gran mayoría de inmuebles en venta del municipio de Valladolid al mismo tiempo que se limita, en todo lo posible, la extracción de aquellos no pertenecientes al municipio. De otro lado, al utilizar el filtro de la plaza mayor, concebido como punto neurálgico de referencia de la ciudad, podremos visualizar las relaciones existentes entre la distancia (en metros) que hay desde la localización de la vivienda a la plaza mayor y el precio de la vivienda. El sentido común nos dice que la relación debería de ser negativa, y que a mayor distancia menor será el precio de la vivienda. Pero en efecto, el sentido común, a diferencia del análisis de datos, dista en muchas ocasiones de ser preciso.

Una vez realizamos todas las llamadas hasta agotar el número de viviendas obtenemos un dataset compuesto por un total de 39 variables. Las variables habidas son de distinto tipo. Unas relacionadas con la localización -coordenadas de latitud y longitud, el distrito o barrio, o la distancia a la plaza mayor-; otras con el precio -precio de venta, y precio por metro cuadrado-; y finalmente, la gran mayoría relativas a las características propias de la vivienda -área en metros cuadrados, número de baños, habitaciones, tipo de propiedad, existencia de ascensor-. En el **Anexo A** se realiza un resumen de esas 39 variables obtenidas en origen.

6.3.2. Preprocesado de datos

Previamente a comenzar con el tratamiento de datos, es preciso hacer un "pequeño lavado de cara" a nuestro dataset. Este apartado vendría a ser una suerte de cribado previo para obtener de forma definitiva el conjunto de datos con el que trabajaremos. Los cambios son los siguientes:

a) Eliminación de variables sobrantes. De las 39 observaciones que nos ofrece Idealista, no todas son aprovechables. Ya sea por una cuestión formal (número excesivo de valores ausentes o falta de confianza en los datos) o por una cuestión de contenido (muchas de las variables eran referidas a las publicaciones en sí mismas), algunas de las 39 variables han sido desechadas de nuestro dataset final. El **Anexo A** contiene una relación de todas y cada una de las 39 variables, incluyendo una explicación de cada una, además de su aceptación o rechazo para el proyecto.

b) Filtrado de inmuebles pertenecientes al municipio de Valladolid. Como consecuencia de la distribución territorial de la provincia de Valladolid y de utilizar como filtro de búsqueda 6 kilómetros de radio desde la plaza mayor algunos de los inmuebles contenidos en el dataset no pertenecen al municipio de Valladolid, los cuales deben ser eliminados. Para llevar a cabo dicha eliminación se filtró la variable municipio de tal manera que solo quedaron incluidos aquellos inmuebles que sí se localizaban en el municipio.
c) Eliminar los inmuebles repetidos. Realizar varias llamadas en momentos temporales distintos conlleva el riesgo de contener inmuebles repetidos. Por esta razón, es necesario encontrar alguna forma mediante la cual se pueda diferenciar cada uno de ellos de forma individualizada y conocer, así, cuál de ellos estaba repetido y cuál no. En nuestro caso se ha utilizado la referencia que aporta Idealista a cada anuncio, la cual es única para ese anuncio (*propertyCode*). Mediante tal técnica se filtraron aquellos inmuebles que resultaban repetidos y se eliminaron del dataset.

Tras llevar a cabo las tres operaciones anteriores, el dataset resultante, susceptible de manipulación y tratamiento analítico, estará compuesto por 20 variables y 2657 observaciones. De ellas, no todas serán utilizadas en su totalidad, pues algunas solamente tienen carácter informativo, o simplemente han servido para realizar algunas operaciones del preprocesado de datos como es el caso del código de referencia.

6.3.3. Análisis Exploratorio de Datos (EDA)

Para evitar caer en un análisis falto de orden y rigor, realizaremos un EDA dividido en dos partes. Primero, nos centraremos en el área más relacionada con el negocio inmobiliario. Se confeccionarán distintas gráficas y visualizaciones con las que obtendremos información básica sobre las tendencias del mercado en la ciudad de Valladolid. Segundo, como actividad previa a las distintas transformaciones que tendremos que realizar en las variables (*future engineering*) se examinan los datos con un componente más estadístico y técnico, incluyendo un análisis de la importancia de cada variable en la predicción del precio de los inmuebles.

a) Mercado inmobiliario de la ciudad de Valladolid

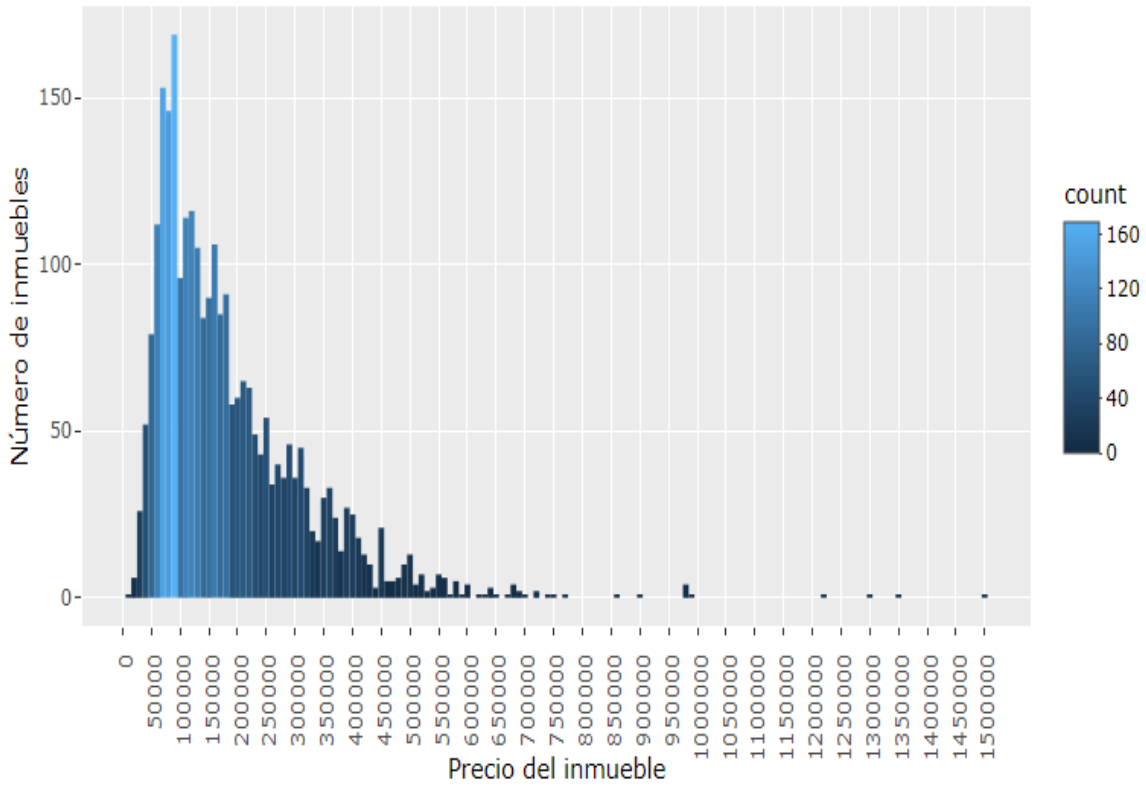
Por ser uno de los pilares sobre los que se asienta este trabajo y conformar la variable, por excelencia, más descriptiva de la situación en la que se encuentra el sector

inmobiliario, el estudio del mercado vallisoletano comenzará con el análisis de la variable precio. La **Figura 13** nos muestra datos básicos de cómo es su distribución. A partir de los datos de la mediana y tercer cuartil ya es posible inferir que la gran parte de los inmuebles tienen un precio por debajo de los 250.000€, y más en concreto, entre los 80.00€ y los 170.000€, como muestra la **Figura 14**. De hecho, tal es así la concentración, que el número de viviendas con un precio superior a 400.000€ es notoriamente pequeño. En el apartado del modelado predictivo será esencial tener muy en cuenta valores extremadamente grandes que puedan distorsionar el poder predictivo.

Figura 13: Estadísticas de la variable *price*

Min	Q1	Mediana	Media	Q3	Max
13,900.00 €	93,000.00 €	154,000.00 €	189,688.00 €	250,000.00 €	1,502,200.00 €

Figura 14: Distribución de la variable *price*

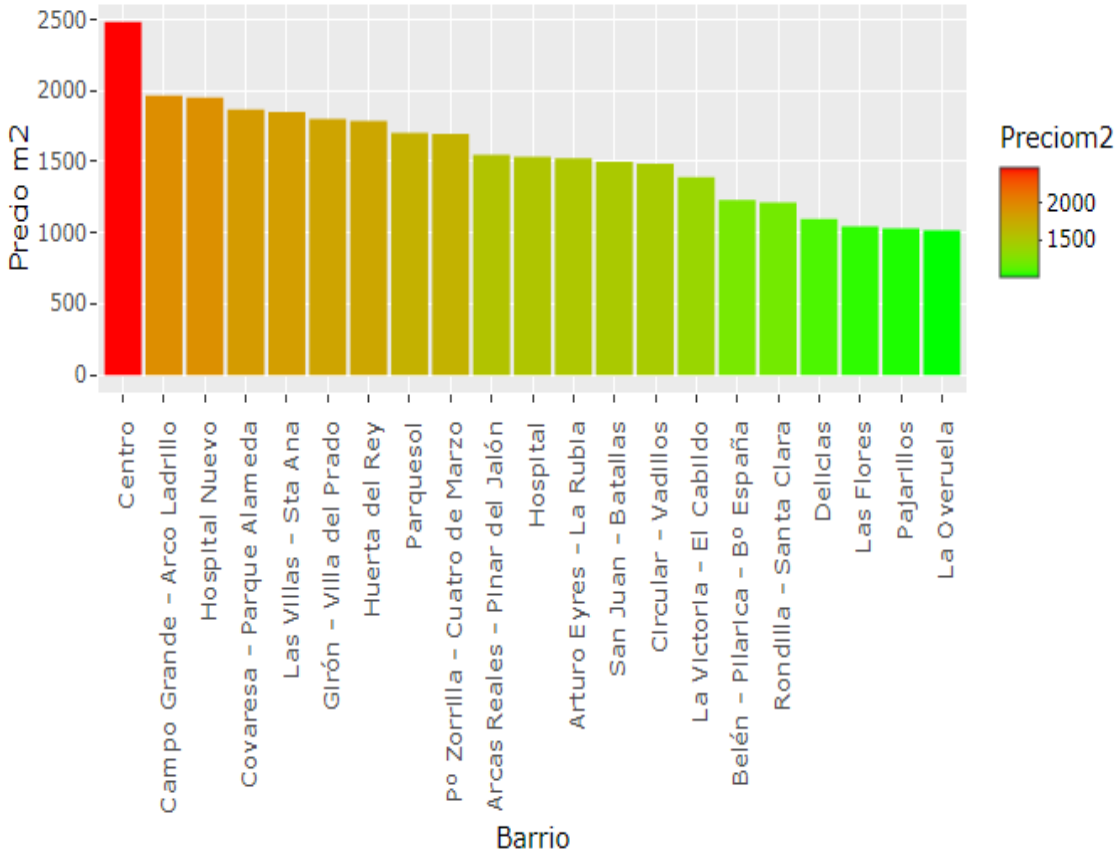


Fuentes: Elaboración propia a partir de datos de Idealista

Tener una perspectiva general es adecuado pero lógicamente insuficiente, ya que a efectos prácticos el precio de un inmueble debe estar contextualizado. No es lo mismo una vivienda en el centro de la ciudad, donde suelen existir precios más elevados, que

aquella situada en suburbios más periféricos. Perfectamente lo explica el siguiente diagrama de barras (**Figura 15**). El precio por metro cuadrado más alto se localiza en el centro de la ciudad (rozando los 2.500€/m^2), mientras que el más bajo -aproximadamente 1.000€- se localiza en el barrio residencial de La Overuela. El precio medio para toda la ciudad se sitúa en torno a los 1.000€ por metro cuadrado. Una visión más completa se consigue cuando se interrelaciona la información económica del precio por metro cuadrado en cada barrio y la localización geográfica de cada uno de ellos (**Anexo B**).

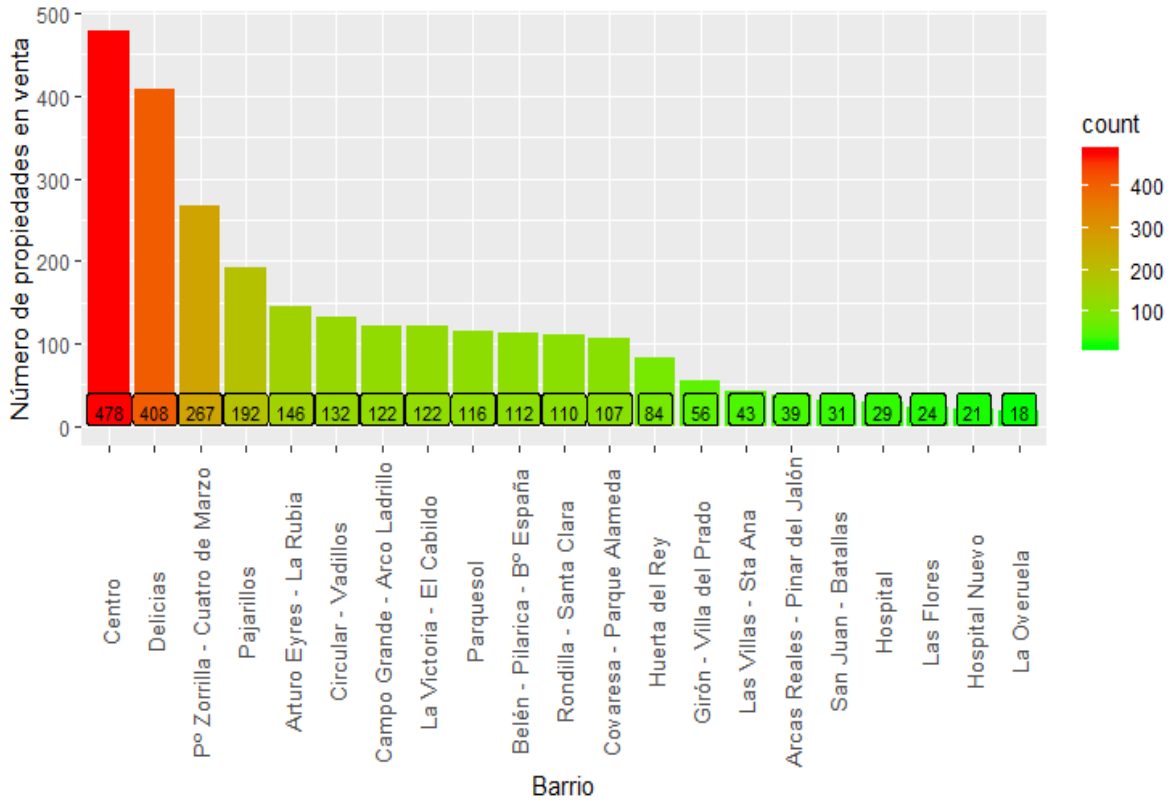
Figura 15: Precio del metro cuadrado por barrio



Fuente: Elaboración propia a partir de datos de Idealista

Si profundizamos más en las particularidades propias de cada zona, observamos el distinto grado de oferta en función de cada barrio (**Figura 16**) Así, con notable diferencia, la zona Centro y el barrio de las Delicias son, actualmente, las zonas urbanas con mayor número de inmuebles a la venta. Por su contra, hasta 7 zonas residenciales no llegan a las 50 viviendas en venta.

Figura 16: Propiedades en venta en función del barrio

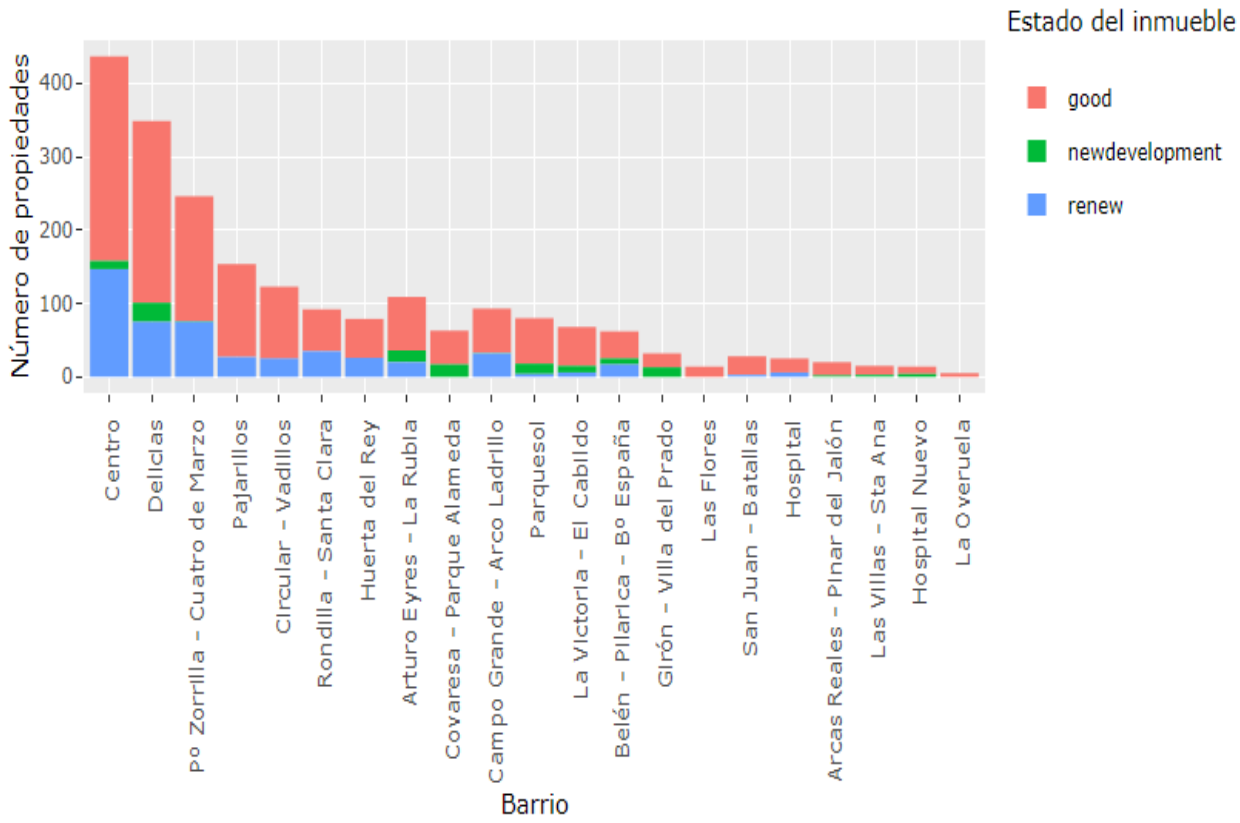


Fuente: Elaboración propia a partir de datos de Idealista

Con un carácter más pragmático se puede proceder también al análisis a partir del estado que presenta el inmueble en cuestión. En ese sentido, suele ser bastante común que determinados clientes pregunten por viviendas en buenas condiciones pero que requieran de una rehabilitación permitiendo con ello lograr un precio más bajo, o que la vivienda goce de ese "toque personal" que toda persona quiere otorgar a su casa A tal objetivo podría dar respuesta la **Figura 17**. De ella se denota, en general, que la situación del parque inmobiliario en Valladolid necesita de una renovación (*good*). Por este motivo, creo honestamente que una posible y atractiva idea de negocio sería que una inmobiliaria tornase su operativa ordinaria en posibilitar ventas que incluyan paquetes de reformas, más o menos integrales, de las viviendas. Como añadido a estas conclusiones, denotar, a su vez, la escasez de inmuebles de "obra nueva" (*newdevelopment*), aunque esto parece que se debe, en gran parte, a que las promociones de nueva construcción no se venden a través de estos portales, sino a través de los propios intermediarios/inmobiliarias con los

que cuenta cada promotora. Es en esta última idea donde brilla por sí mismo el valor añadido que puede otorgar una persona con experiencia en el mercado inmobiliario local a la información obtenida del análisis de datos.

Figura 17: Estado de los inmuebles por barrios

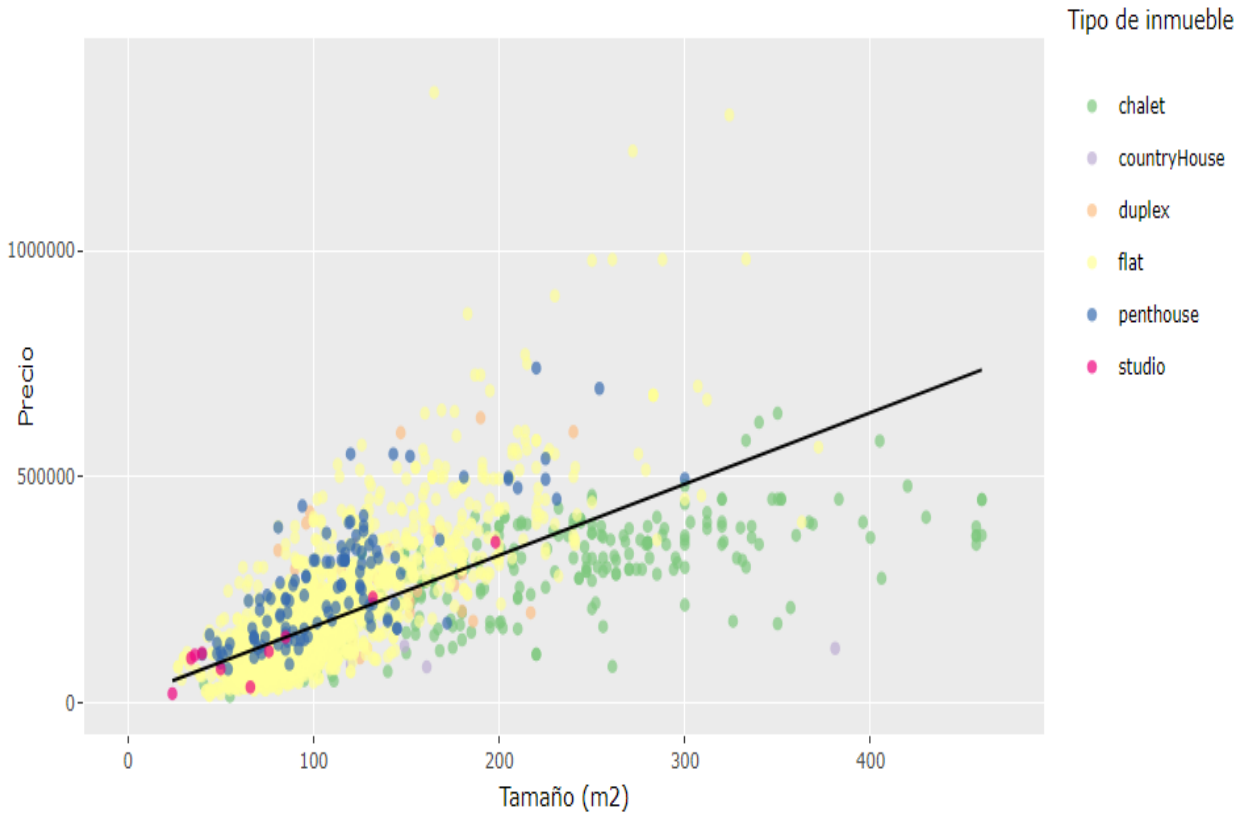


Fuente: Elaboración propia a partir de datos de Idealista

Para finalizar se presentan a continuación una serie de relaciones que dan buena muestra de las características que presenta la oferta inmobiliaria en la ciudad de Valladolid. Una primera sería la que se obtiene de comparar el tamaño de las viviendas y su precio, señalando de forma complementaria el tipo de vivienda (**Figura 18**). Como era de esperar, es la categoría de *chalet* la tipología de vivienda con mayor tamaño de entre la ofertadas (puntos verdes), mientras que el tamaño estándar de las viviendas oscila entre los 50 y los 180 metros cuadrados, principalmente debido a que la gran parte de la oferta está formada por pisos. En cuanto a la relación tamaño-precio, es, como no, positiva, aunque aminora desde el momento que comienza la mencionada categoría de chalet. Transponiéndolo a un ejemplo práctico, si un cliente pregunta por una vivienda superior a los 200m^2, la recomendación iría dirigida a buscar un chalet, pero sin descartar otro tipo

de viviendas que también presentan esos tamaños. Esto posibilita, por tanto, una atención mucho más personal y detallada.

Figura 18: Relación entre precio, tamaño y tipo de inmueble

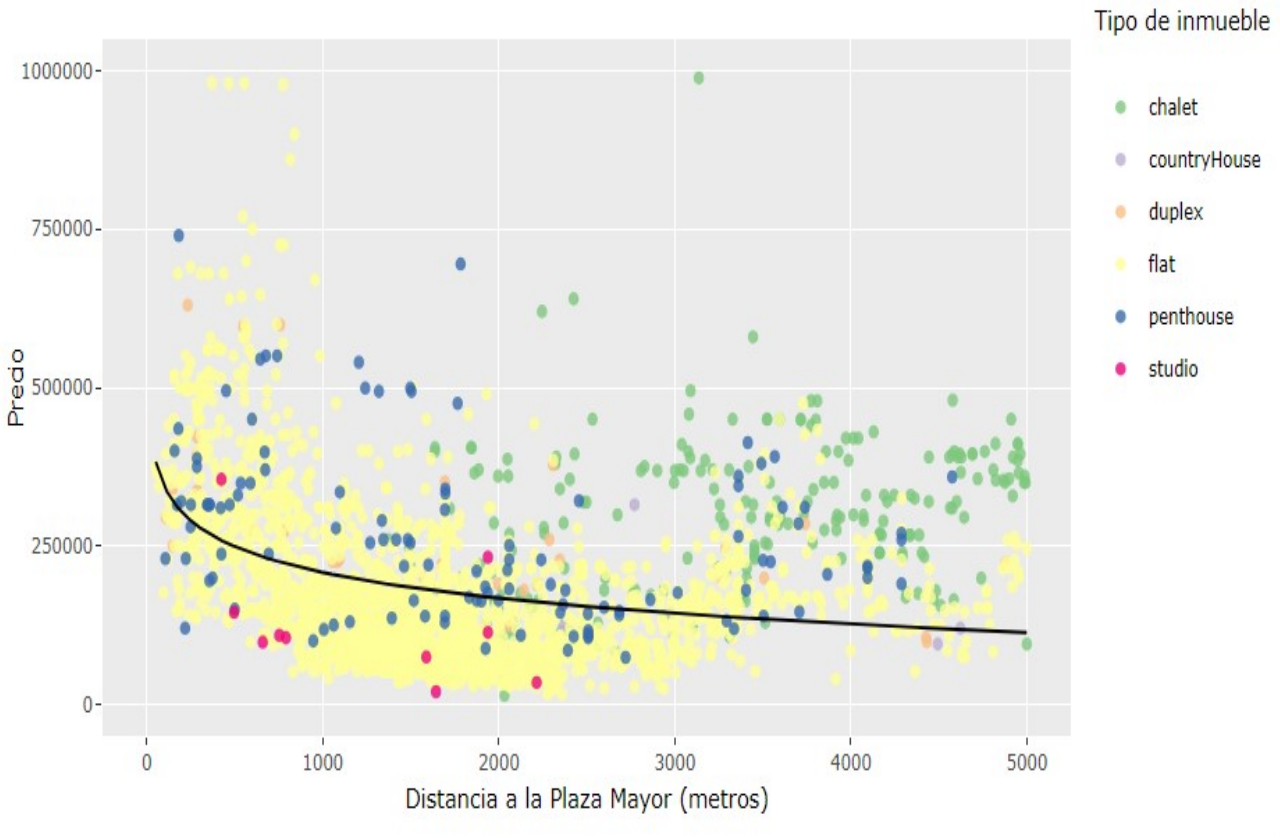


Fuente: Elaboración propia a partir de datos de Idealista

Una segunda relación susceptible de ser mostrada es la que constituyen el precio y la distancia a la Plaza Mayor de Valladolid, complementada también con el tipo de inmueble (**Figura 19**). Las viviendas cuya distancia a la Plaza Mayor es inferior a un kilómetro tienen en general un precio más elevado, mientras que a partir de una distancia de 500 metros la diferencia de precios no es tan significativa. Una información harto importante que tiene su reflejo en las decisiones de compra de los clientes. De comprar una vivienda dentro del radio de 500 metros de distancia a hacerlo más allá puede entrañar un descenso muy relevante del precio, lo cual suele tener mucha incidencia cuando lo que se está buscando es adquirir un número de viviendas mayor del común. Ejemplo propio de esta parte sería la compra de un edificio o grupo de viviendas para ofrecerlas en régimen de alquiler. En este tipo de situaciones, el análisis de datos es esencial pues puede permitir al cliente ahorrar una cantidad importante sin desviar demasiado las intenciones originales que este pudiese tener en cuanto a localización de las viviendas. Finalmente,

reseñar cómo, en función del tipo de inmueble, las distancias son totalmente distintas. Mientras que los pisos (*flat*) están presentes en todas las distancias, los chalets, por necesidades obvias de espacio, tienden a localizarse a partir de los dos kilómetros de distancia a la Plaza Mayor.

Figura 19: Relación entre precio, distancia a la Plaza Mayor y tipo de inmueble



Fuente: Elaboración propia a partir de datos de Idealista

De todo lo anterior se puede comprobar el ingente potencial que tiene el Big Data para este sector, y la "facilidad" con la que uno puede acceder a ello. De igual forma que se ha hecho aquí, es perfectamente posible que una inmobiliaria tenga disponibilidad de acceso a este tipo de recursos, más si cabe, cuando el coste de este proyecto asciende a los 0 euros. Es más, sigo haciendo hincapié en la idea de que en aquellas ciudades españolas donde no se ha visto aún una disrupción en el uso de este tipo de herramientas, dar un paso al frente antes que los demás, puede suponer un "golpe importante encima de la mesa" en lo que a términos de competencia se refiere.

Como broche final a este subapartado del análisis exploratorio, me he permitido realizar una pequeña representación geoespacial de la oferta inmobiliaria analizada mediante un gráfico de *clusters* que clarifica visualmente cómo se distribuye la oferta en torno a la ciudad de Valladolid (**Anexo C**). Dicha representación ha sido diseñada con el paquete *leaflet*, el cual está especializado en este tipo de representaciones. Otra muestra clara y evidente del espectacular potencial inherente al mundo del dato y la programación.

a) Análisis estadístico de las variables

A continuación, se realiza el análisis estadístico de las variables que estamos utilizando en nuestro proyecto. A tal efecto, solo se examinarán aquellas que se utilicen o al menos sean potencialmente útiles a la hora de diseñar nuestro modelo predictivo. Para ello dividiremos nuestro conjunto de datos separando las variables numéricas de las categóricas.

Comenzando por las **variables numéricas**, la **Tabla 5** recoge su análisis estadístico. De ella, merecen comentarios los mínimos y máximos de las variables *price* y *size*. Como se señaló en el subapartado anterior, la variable precio contiene observaciones muy altas llegando hasta el millón y medio de euros, y que nosotros consideraremos como *outliers* (observación que difiere significativamente del resto). Parece que se puede decir lo mismo del tamaño. Ratifican dicha tesis las métricas del coeficiente de asimetría[28] y de curtosis[29], pues sus valores son propios de distribuciones con largas colas y por ende, con numerosos *outliers*. A mayores, ambas variables muestran la existencia de una gran concentración de valores en torno a la media (curtosis > 3) y la falta de simetría en sus distribuciones.

Respecto de las demás variables no parece haber mayores problemas salvo por la constatación de valores extraños en las variables de *rooms* y *bathrooms*. El sentido común nos dice que no es posible una vivienda sin habitaciones y menos aún, una en la que no haya baño. Esto nos hace "encender las alarmas" para verificar más adelante a qué se deben estas anomalías.

[28] Medida que pone de manifiesto la simetría de la distribución de una variable respecto a la media aritmética
[29] Medida que indica la cantidad de datos que hay cercanos a la media

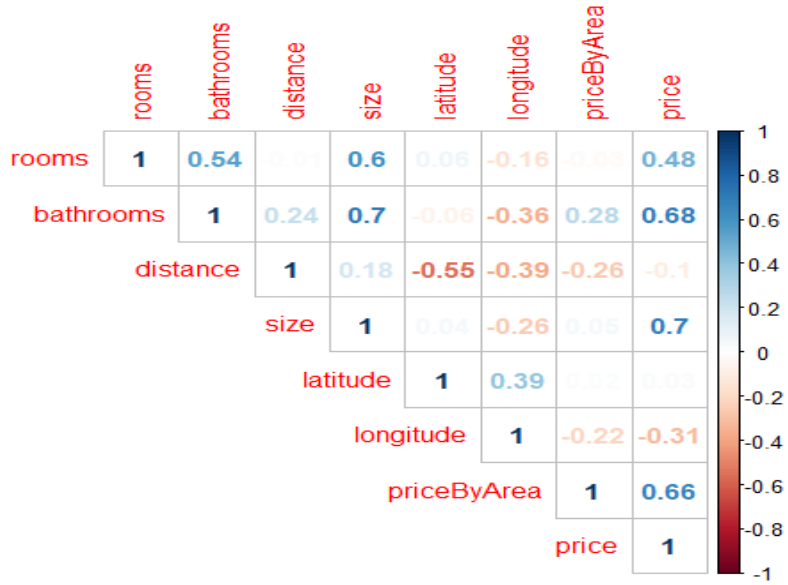
Tabla 5: Resumen estadístico de las variables numéricas

Variables	Rol	n	Media	Sd	Min	Mediana	Máx	Rango	Asimetría	Curtosis
rooms	Input	2657	3.09	1.01	0.00	3.00	8.00	8.00	0.56	2.31
bathrooms	Input	2657	1.59	0.75	0.00	1.00	5.00	5.00	1.22	1.58
PriceByArea	Input	2657	1635.59	743.93	253.00	1523.00	8182.00	7929.00	1.25	3.49
distance	Input	2657	1858.71	1164.74	52.00	1636.00	5882.00	5830.00	0.96	0.53
size	Input	2657	114.51	68.27	24.00	95.00	1266.00	1242.00	4.28	40.44
latitude	Input	2657	41.64404	0.01389	41.60290	41.64511	41.69303	0.09013	0.04630	0.94131
longitude	Input	2657	-4.72847	0.01504	-4.77265	-4.72634	-4.68915	0.08350	-0.41184	0.04919
price	Objetivo	2657	189688	136892	13900	154000	1502200	1488300	2.28	10.62

Fuente: Elaboración propia a partir de datos de Idealista

Para finalizar con las variables numéricas mostramos su matriz de correlaciones (**Figura 20**). Como se puede apreciar, las variables de *bathrooms* y *size* tienen una correlación media-alta con la variable objetivo (*price*). Por su parte, es de destacar que la variable *distance* no presenta una correlación muy significativa, aunque como se comprobó en la **Figura 19**, la relación entre *distance* y *price* no seguía una regresión totalmente lineal, sino más bien de tipo logarítmico[30]. Entre las propias variables explicativas, son reseñables las relaciones positivas que existen entre *size* y *bathroom* (0.7) y entre *size* y *rooms* (0.6).

Figura 20: Matriz de correlaciones (variables numéricas)



Fuente: Elaboración propia a partir de datos de Idealista

[30] El precio se encarecía enormemente cuando la distancia era inferior a 0,5 km

El restante grupo de variables está compuesto por variables categóricas, que indican la existencia de varios niveles o clases, y por variables binarias -también conocidas como *dummies*- las cuales sirven para expresar la presencia o inexistencia de un determinado elemento. Para nuestro proyecto serán variables binarias: *status, exterior, hasLift* y *hasPlan*.

A juzgar por el contenido de la **Tabla 6**, las variables *floor* y *district* presentan un alto número de niveles que podrían entrañar problemas en la confección de nuestro modelo predictivo, por lo que deberán ser tenidas en cuenta en apartados siguientes. Además, y esto es especialmente importante, tres variables cuentan con valores ausentes (NA) que tendrán que tratarse más adelante de una manera específica, buscando eso sí, un balance entre la representatividad del conjunto de datos y su veracidad.

Tabla 6: Resumen estadístico de las variables categóricas y binarias

Variables	Rol	n	Niveles	Ausentes	Moda	Frecuencia	Proporción
floor	Input	2270	18	387	1	526	0.232
district	Input	2657	21	0	Centro	478	0.18
propertyType	Input	2657	6	0	flat	2231	0.84
status	Input	2433	3	224	good	1705	0.701
exterior	Input	2657	2	0	TRUE	1931	0.727
hasLift	Input	2348	2	309	TRUE	1934	0.824
hasPlan	Input	2657	2	0	FALSE	2003	0.754

Fuente: Elaboración propia a partir de datos de Idealista

Para finalizar la parte más técnica de nuestro EDA, se procede a elaborar un estudio de la importancia que tiene cada uno de los predictores[31] a la hora de explicar la variable respuesta del precio. Es una parte, por tanto, con un enfoque mucho más ligado al *predictive analytics* y que necesita de un conocimiento más profundo de estas técnicas para quien lo quiera poner en práctica.

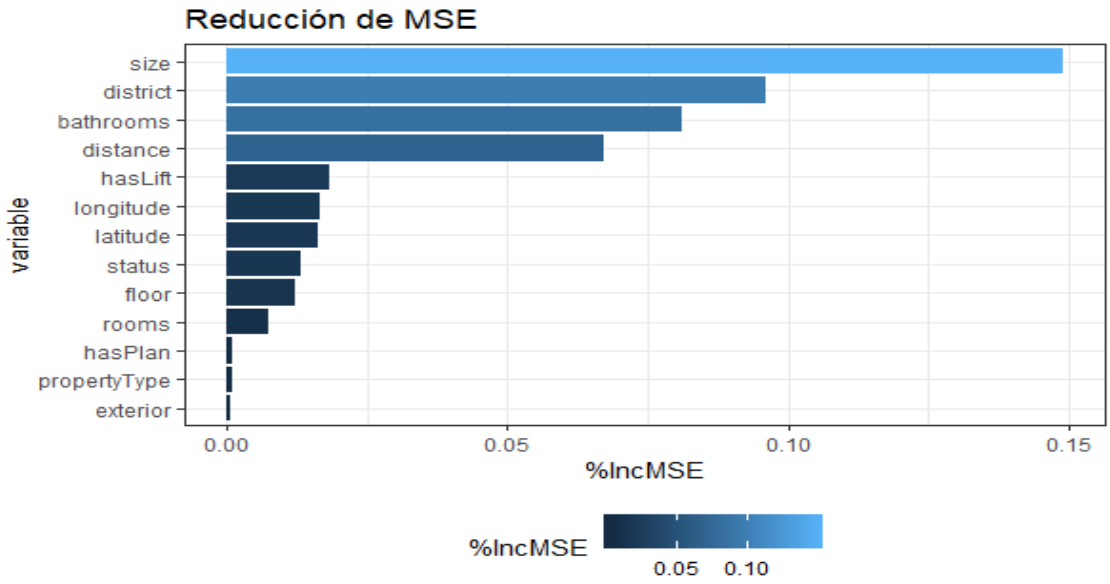
Nosotros aplicaremos sobre el conjunto de datos una de las técnicas más extendidas y más sencilla de entender: *Random Forest*[32]. Se trata de un método estadístico de aprendizaje perteneciente a la familia de los árboles de decisión cuyo funcionamiento se asienta sobre la construcción de un gran número de árboles individuales que se

[31] Desde este momento se llamarán a las variables indistintamente como variables o predictores
[32] Para un mejor explicación de los árboles de decisión visite: https://rpubs.com/Joaquin_AR/255596

asemejan a un bosque (Amat Rodrigo, 2018). La idea que persiste detrás de todo este proceso es conocer, ya en la fase del EDA, cuáles van a ser las variables predictoras que tengan más influencia sobre el precio. Téngase en cuenta que, cuando se entrena un modelo, obtener resultados positivos depende, en gran parte, de incluir aquellos predictores que sí estén realmente relacionados con la variable respuesta (Amat Rodrigo, 2018).

El resultado obtenido se puede apreciar en la **Figura 21**. En ella se expone una de las medidas más utilizadas para cuantificar la importancia de los predictores cuando la variable respuesta es continua: "*Reducción de MSE (Mean Square Error)*". Esta medida muestra cuál sería el impacto esperado que tendría en el modelo (incluyendo todas las variables) una permutación de ese predictor en concreto por otro distinto. Dicho impacto es medido a través del incremento que sufre el MSE, o error medio de test en la predicción del precio y que, sensu *contrario,* se interpreta tal que, la presencia de un cierto predictor o variable en el modelo permite reducir el error de test en un porcentaje determinado. Así pues, a mayor porcentaje de reducción, mayor influencia tendrá esa variable sobre el modelo de predicción de la variable objetivo (*price*).

Figura 21: Importancia de cada predictor



Fuente: Elaboración propia a partir de datos de Idealista

Observando los resultados, las variables *size*, *district*, *bathrooms* y *distance* muestran una influencia muy alta sobre la predicción de los precios de las viviendas. Contrariamente, *hasPlan*, *propertyType* y *exterior* son predictores cuya aportación al

modelo parece ser prácticamente nula. Predictores con resultados intermedios -hasLift, longitude, *latitude*, *status*, y *floor*- deberán ser examinados para así determinar si su aportación a la hora de predecir la variable *price* es lo suficientemente alta. Para estos casos habrá que tener en cuenta que los modelos de ML deben regirse por el **principio de parsimonia**: será mejor aquel modelo que, obteniendo resultados iguales, necesite de un menor número de predictores para lograr ese resultado (Amat Rodrigo, 2018).

6.3.4. Future engineering

Una de las fases que tiene prácticamente todo proyecto de Big Data, sea cual sea su enfoque, es aquella en la que los datos pasan por un proceso de transformación. Una circunstancia que se agrava cuando los datos han sido obtenidos de una fuente externa, como es el caso. Por ello, la fase de *future engineering* viene a resolver este problema. Para este proyecto las transformaciones y demás operaciones que se realicen sobre el dataset tienen como objetivo principal facilitar el trabajo a los algoritmos que van a ser utilizados en la fase de entrenamiento, y que los algoritmos, como contrapartida, consigan mejores resultados que los que se hubiesen obtenido en caso de no pasar por este paso.

a) Tratamiento de la variable *floor*

Esta variable presentaba un problema en la manera en la que se almacenaban los datos. Si, por norma general cuando hablamos de viviendas en edificios hacemos referencia a la planta en la que se encuentra mediante cifras (0, 1º, 2º, etc.), en el conjunto de datos se encontraron ciertas observaciones que no se correspondían con cifras, a saber: "bj", "en" y "st". Siendo entradas con una proporción y frecuencia muy bajas, y entendiendo que se refieren a los conceptos de "bajo", "entrada" y "sótano", se han transformado y agrupado con las viviendas que se encuentran en la planta baja, o piso 0 de los edificios. Muchas de esas viviendas, como es lógico, respondían a la tipología de *chalets*, pero también existían viviendas que en los propios edificios se localizan en dicha planta baja pudiendo, así, causar confusión.

b) Eliminación de viviendas sin baño, sin habitaciones o sin ambos espacios

Ya en el EDA se hizo referencia a esta circunstancia (**Tabla 5**). Según los registros de Idealista hay viviendas que, o bien, no tienen habitaciones, o bien, no tienen baño, y otras que incluso no cuentan con ninguno de los dos espacios. Esta circunstancia es cuanto

menos extraña, y lejos de querer entrar en el detalle de cuál es la razón que provoca ello (error en la introducción de datos, viviendas con baño compartido, viviendas diáfanas como los estudios, etc.), se proceden a eliminar los registros con esas características. Otra razón que permite seguir la tesis de la eliminación es que el número de viviendas que son excluidas es de 15 y la consecuencia de la eliminación es una reducción del número de observaciones desde 2657 a 2642.

c) Tratamiento de valores ausentes (NA)

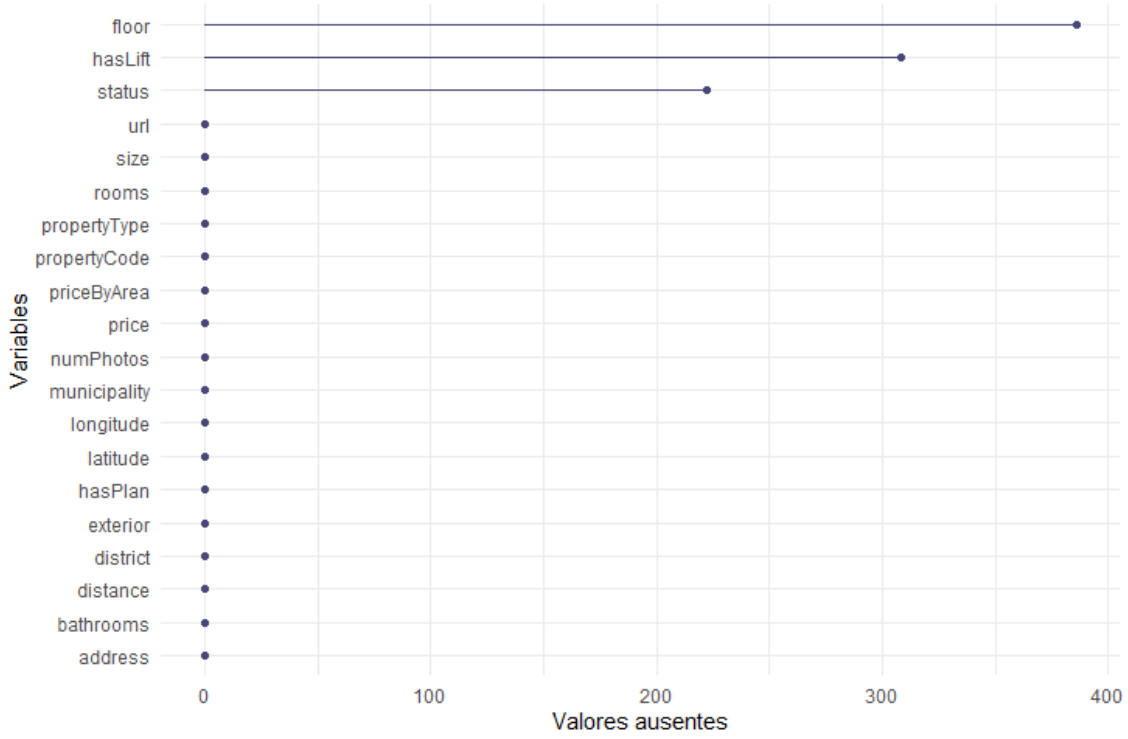
Seguramente se trate de una de las cuestiones más importantes que comprende esta fase. Varios de los algoritmos que se emplean en ML no aceptan conjuntos de datos con valores ausentes por lo que se requiere de una solución. De acuerdo con Amat Rodrigo (2018), la resolución de este conflicto pasa por las siguientes tres posibilidades:

- Eliminar las **observaciones** con valores ausentes
- Eliminar las **variables** que contengan los valores ausentes
- Intentar estimar los valores ausentes a partir del resto de información que nos brinda el conjunto de datos (imputación)

Elegir entre una u otra depende mucho de la casuística que exista, pero el énfasis tiene que ponerse, sobre todo, en la influencia o grado de aportación que tenga en el modelo esa variable que contiene valores ausentes. Dicho lo cual, las dos primeras opciones, aun siendo las más sencillas de aplicar, son las que pueden entrañar una pérdida de información mayor. Por su parte, la imputación cuenta con el *handicap* de que la transformación se realiza a partir de una estimación, y por tanto, no existe certeza de que esos datos representen la realidad.

La **Figura 22** representa una imagen general de los valores ausentes que tiene cada variable. Son tres las variables que tienen observaciones ausentes -*floor*, *hasLift*, y *status*-, y siendo realistas no se trata de un número bajo. De esta forma, para decidir qué técnica seguir nos guiaremos por el sentido común inmobiliario y la importancia de cada variable en el modelo (**Figura 21**).

Figura 22: Variables con valores ausentes (NA)



Fuente: Elaboración propia a partir de datos de Idealista

La variable *floor* tiene una importancia relativamente baja, y además es difícil conocer exactamente a qué se debe que no haya un registro para la planta en la que se encuentra una vivienda. Es más, de las viviendas sin especificación de planta, un total de 220 eran chalets (siempre son bajos) mientras que 140 eran pisos (pueden encontrarse en cualquier planta). Esto, a mi parecer, hace inviable ejecutar una imputación por lo que directamente no utilizaremos la variable *floor* en nuestros modelos predictivos.

En el caso de la variable de *hasLift*, al tener una importancia más grande trataremos de imputarla. Por lo general, no especificar si se cuenta con ascensor quiere decir implícitamente que no existe tal elemento. No tiene lógica alguna no mostrar si se cuenta o no con ascensor cuando en realidad sí que se tiene. Así pues, la imputación de esta variable consistirá en transformar todos los NA en la categoría de FALSE, que significa que, el edificio en el que se encuentra la vivienda, no tiene ascensor.

Por último, trataremos la variable de *status* que muestra el estado en el que se encuentran las viviendas (bien, nueva obra o renovada). Su importancia es algo mayor que la de *floor*, y no parece descabellado pensar que quien no especifica el estado del inmueble es, en efecto, porque su estado no es el mejor. Entonces, si seguimos la lógica

que utilizamos para la imputación de *hasLift*, aquellas observaciones que en la variable de *status* tuvieran un valor ausente ahora contendrán el nivel que indica el peor estado, es decir, *good*.

El resultado de estas transformaciones se puede apreciar en el resumen de la **Tabla 7**. Las cifras sombreadas en amarillo han sufrido variaciones como consecuencia de la imputación de valores ausentes.

Tabla 7: Resultado de las transformaciones de los valores ausentes (NA)

Variables	Rol	n	Niveles	Ausentes	Moda	Frecuencia	Proporción
status	Input	2642	3	0	good	1920	0.727
hasLift	Input	2642	2	0	TRUE	1926	0.729

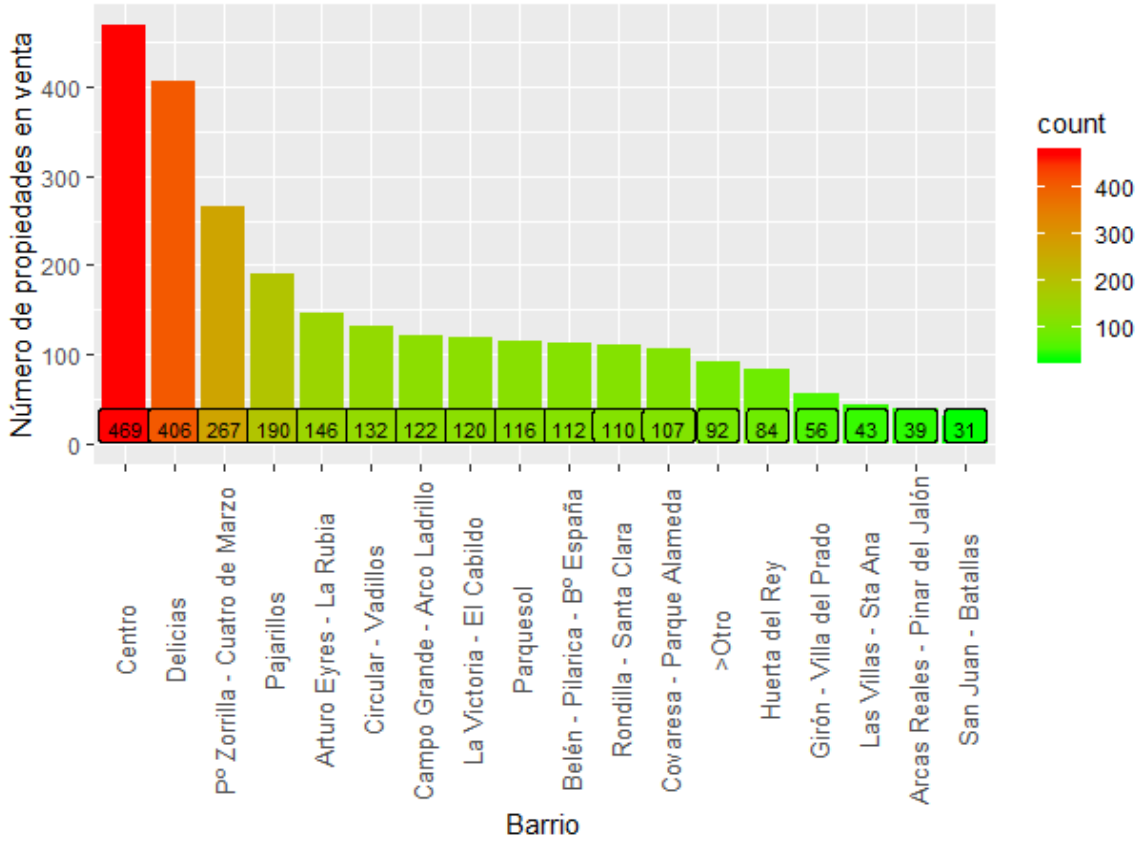
Fuente: Elaboración propia a partir de los datos de Idealista

d) Agrupación o eliminación de niveles de poca proporción

El dataset contiene ciertas variables categóricas cuyo número de niveles es muy alto. A esto hay que añadir que ciertos niveles tienen una frecuencia muy baja, es decir, una varianza próxima a 0. Para solucionarlo es recomendable, además de aplicar la técnica de *one-hot-encoding* que se explicará más adelante, agrupar esos niveles que presentan proporciones muy reducidas. Este es el caso de las variables *floor*, *propertyType* y *district*.

En cuanto a la primera, como hemos decidido no emplearla en los modelos de ML, no practicaremos ninguna variación sobre ella. La variable *propertyType* tiene 5 niveles (*flat*, *chalet*, *penthhouse*, *duplex*, *countryHouse*) y el 90% de las entradas corresponde con el nivel *flat*. Esta es una de las razones por la que la **Figura 21** señala que la importancia del predictor *propertyType* dentro de un modelo con todas las variables es muy baja. Por este motivo, no contaremos con ella en el modelo pero, en caso contrario, se recomienda la eliminación del nivel *countryHouse* pues solo existen 6 observaciones que responden a esa tipología de vivienda. Finalmente, respecto de la variable *district*¸ se agruparán los barrios que tengan menos incidencia, concretamente aquellos cuyo número de viviendas sea inferior a 30 -*Hospital*, *Las Flores*, *Hospital Nuevo*, y *La Overuela*-. El nuevo nivel se reformulará bajo el nombre de "otros", y estará conformado por 92 viviendas (**Figura 23**).

Figura 23: Número de viviendas por barrios tras agrupación de niveles



Fuente: Elaboración propia a partir de datos de Idealista

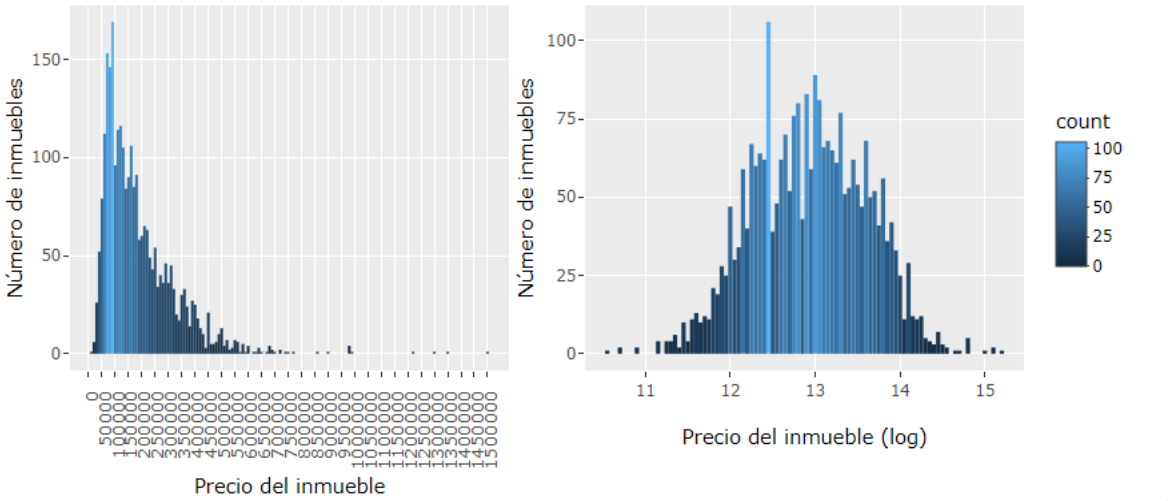
e) Cambios en la variable precio

Con el propósito de encontrar los mejores resultados posibles sobre la variable objetivo o variable respuesta (*price*) se harán las siguientes operaciones:

1. Como comprobamos en la **Figura 14**, la variable precio no seguía una distribución normal, sino que su distribución era de tipo logarítmico, la cual constataba problemas de asimetría. En general, los algoritmos empleados en ML mejoran sustancialmente cuando la distribución de la variable respuesta tiende a caracterizarse como normal. Para lograr tal situación, es necesario realizar sobre la variable dependiente una transformación logarítmica cuyo resultado se puede apreciar en la **Figura 24**.
2. A pesar de haber tratado la variable *price* con la transformación logarítmica, siguen existiendo observaciones que se alejan mucho de la mediana (*outliers*). Esto incide directamente en el poder predictivo que pueden presentar los modelos, y por ende, se procede a su eliminación. No obstante, el problema no queda todavía resuelto. Cabe recordar que el 75% de los inmuebles registrados en los datos tenían un precio inferior

a 250.000€. Por tanto, aunque se eliminen aquellas entradas que se reconozcan como *outliers*, sigue habiendo un espectro muy importante de ellos que se aleja bastante de la mediana y que, seguramente, supondrá un empeoramiento del resultado predictivo de los modelos. La clave en este punto reside en ser capaces de encontrar un balance óptimo entre el resultado y la representatividad de los datos. Para ello, se probará indistintamente eliminaciones parciales de viviendas con precios por encima de los 250.000€, siempre y cuando, el número de registros con el que operemos no sea inferior a 2.000 observaciones.

Figura 24: Distribución de la variable precio sin transformar y con transformación



Fuente: Elaboración propia a partir de datos de Idealista

6.3.5. División de datos en test y entrenamiento

Para poder comprobar la capacidad predictiva "real" que presenta un determinado modelo es necesario contrastar los resultados obtenidos con un conjunto de observaciones que, aun conociéndose los registros de la variable que queremos predecir, no haya tenido ningún contacto anterior con el modelo (Amat Rodrigo, 2018). A este propósito aplica la "división de todos los datos en entrenamiento y test". El subconjunto de entrenamiento será utilizado, como su propio nombre indica, para entrenar el modelo de forma reiterada tratando de encontrar el ajuste que maximice el poder predictivo. Una vez entrenados todos los modelos estos serán puestos a prueba con el subconjunto de test.

Dicho esto, proceder de forma adecuada con la partición del conjunto de datos es altamente importante. Primero, es necesario encontrar una proporción adecuada. Es una

cuestión cuyo núcleo reside en la cantidad de datos disponibles, pero la norma general suele dotar al subconjunto de entrenamiento con un 70-80% de los datos, dejando el restante 30-20% al subconjunto de test. Para el presente proyecto utilizaremos la división 70%-30% pues los datos con los que contamos, en términos comparativos con otros proyectos de Big Data, no son excesivamente grandes. Segundo, no se puede pasar por alto que hay que garantizar una similar representatividad en ambos subconjuntos, lo cual se logra mediante el comando de R *createDataPartition*.

Procedemos con la partición, y obtenemos dos subconjuntos. Uno, el de entrenamiento, formado por 1537 observaciones y 20 variables, y dos, el de test, que cuenta con 656 observaciones y las mismas 20 variables. Téngase en cuenta que se realizó sobre un dataset ya disminuido -20193 viviendas- puesto que se ha procedido a desechar los inmuebles con un precio superior a 300.001€[33].

6.3.6. Preparación de datos para los modelos de Machine Learning

El propósito de esta etapa del proceso es llevar a cabo una serie de transformaciones para poder ejecutar sobre ellos diferentes algoritmos de ML, y que, además, traten de mejorar sus resultados. Es cierto que varias de las operaciones que fueron llevadas a cabo en el apartado de **Future engineering** podrían haberse realizado aquí, pero, sin embargo, esta parte se diferencia en que las transformaciones en las variables desdeñan un carácter más grupal, y no tanto singular, como se pudo observar en aquel apartado.

La premisa que debe primar en todo este apartado es que las transformaciones deben tener efecto sobre ambos subconjuntos, pero éstas tendrán que estar guiadas previamente por el aprendizaje obtenido del subconjunto de entrenamiento. Un paquete realmente efectivo para este tipo de tratamiento es el de *recipes*, cuyo funcionamiento sigue el patrón de una receta de cocina (Barter, 2019). A continuación, se explican las etapas fundamentales que ha seguido esta preparación de los datos inmobiliarios:

1. Obtenemos los ingredientes. Especificamos la variable respuesta -*price*- y las variables predictoras -*rooms*, *size*, *bathrooms*, *district*, *hasLift*, *distance*, *status*, *exterior*, *longitude* y *latitude*-.

[33] Para lograr un mejor poder predictivo

2. Escribimos la receta. Aquí determinamos todas las transformaciones con las que queremos proceder:

    a) Centrado. Es un tipo de tratamiento típico en predictores numéricos. Logra que en cada predictor seleccionado la media sea de 0. Esto se consigue restando a cada registro la media de todas las observaciones de la variable (Amat Rodrigo, 2018).
    b) Estandarización (normalización) y escalado. También enfocado en los predictores numéricos, la estandarización permite colocar a todos los predictores numéricos en la misma escala. Por su parte, el escalado normaliza las variables numéricas dentro del rango [0,1].
    c) Binarización de variables cualitativas (*one-hot-enconding*). Muchos algoritmos de ML necesitan para su funcionamiento que tanto los predictores como la variable objetivo sean numéricas. Esto significa que las variables cualitativas tienen que convertirse en variables numéricas. Así, la operación que se realiza es crear a partir de cada uno de los niveles que presenta esa variable cualitativa en concreto una variable *dummy* o binaria (Brownlee, 2017). De este modo, por ejemplo, en la variable *district* cada uno de los barrios se convierte en una nueva variable con el valor de 0 si la vivienda no se encuentra en ese barrio o de 1 si efectivamente pertenece a él.
    d) Eliminación de variables con varianza próxima a cero. Puede suceder que determinadas variables solo contengan un único valor (varianza cero), o que tomen varios valores, y algunos de ellos aparezcan con una frecuencia muy baja. El problema que nace a raíz de ello es una degeneración de sus distribuciones, y en consecuencia, un empeoramiento del poder predictivo de algunos algoritmos de ML (Lin & Li, 2020, pp. 95-96). Por esta razón, dentro del paquete *recipes* tendremos la opción, una vez binarizadas las variables cualitativas, de eliminar aquellos predictores que tengan este problema.

3. Preparación de la receta. Se indica el subconjunto de datos con el que se va a preparar la transformación (entrenamiento). Básicamente, consiste en lograr que R aprenda cuáles son las variables a las que puede aplicar la receta.
4. Hornear. Es el último paso, y consiste en aplicar las transformaciones que se han preparado a los subconjuntos de entrenamiento y test.

Un resumen de todas las operaciones puede encontrarse en la **Figura 25**. De ellas, es destacable que 11 variables que correspondían a ciertos barrios han sido eliminadas por tener una varianza próxima a 0.

Figura 25: Resumen del preparado de datos

```
Data Recipe

Inputs:

      role #variables
   outcome          1
 predictor         10

Training data contained 1537 data points and no missing data.

Operations:

Centering for rooms, size, bathrooms, distance, longitude, latitude [trained]
Scaling for rooms, size, bathrooms, distance, longitude, latitude [trained]
Range scaling to [0,1] for rooms, size, bathrooms, distance, longitude, latitude [trained]
Dummy variables from district, status [trained]
Sparse, unbalanced variable filter removed 11 items [trained]
```

Fuente: Obtenido de la consola de R

6.3.7. Selección de predictores

Incluir únicamente aquellos predictores que de verdad tengan influencia sobre la variable respuesta puede significar la diferencia entre crear un modelo bueno y otro muy bueno (Amat Rodrigo, 2018). De hecho, uno de los riesgos que siempre asoman en estas etapas es que el modelo predictivo sea muy bueno prediciendo el subconjunto de entrenamiento pero no tanto respecto del subconjunto de test que, al final, es el que más interesa (*overfitting*). De este modo, saber cuáles son los predictores que deben participar de los modelos de aprendizaje se antoja una cuestión de notoria importancia.

Tras las transformaciones del anterior apartado, el número de variables predictoras será de 16 (**Tabla 8**).

Tabla 8: Información de las variables una vez transformadas

Variable	Tipo	Rol	Fuente
rooms	Numérico	Predictor	Original
size	Numérico	Predictor	Original
bathrooms	Numérico	Predictor	Original
hasLift	Binario	Predictor	Original
distance	Numérico	Predictor	Original
exterior	Binario	Predictor	Original
longitude	Numérico	Predictor	Original
latitude	Numérico	Predictor	Original
district_Arturo.Eyres...La.Rubia	Numérico	Predictor	Derivado
district_Centro	Numérico	Predictor	Derivado
district_Circular...Vadillos	Numérico	Predictor	Derivado
district_Delicias	Numérico	Predictor	Derivado
district_Pajarillos	Numérico	Predictor	Derivado
district_Pº.Zorrilla...Cuatro.de.Marzo	Numérico	Predictor	Derivado
status_newdevelopment	Numérico	Predictor	Derivado
status_renew	Numérico	Predictor	Derivado
price	Numérico	Respuesta	Original

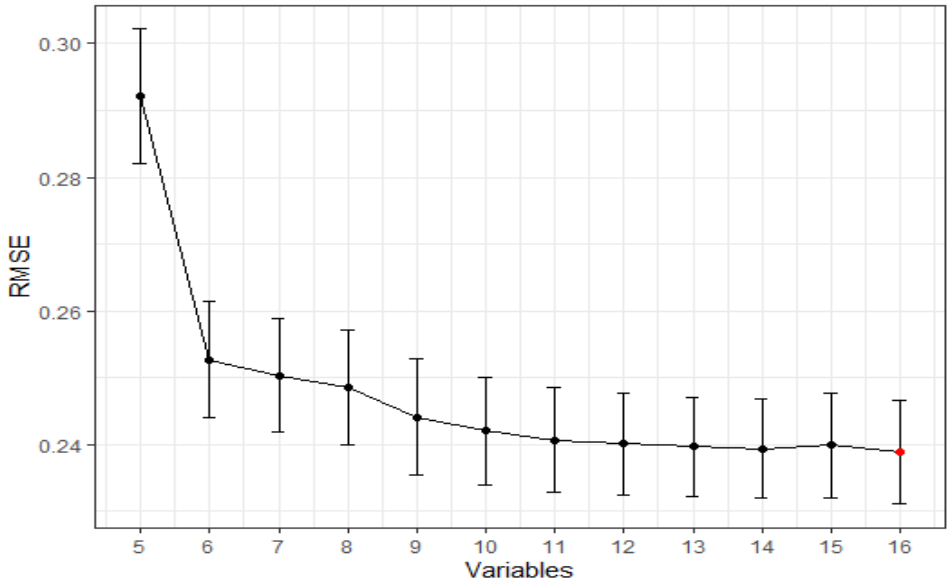
Fuente: Elaboración propia a partir de datos de Idealista

Se procede a continuación a operar uno de los métodos más comunes en la selección de predictores: eliminación recursiva de variables[34]. Es decir, se prueba un modelo con todas las variables predictoras en él, y se introduce la variable *price* en términos logarítmicos. Para determinar cuál es el mejor conjunto de variables es necesario utilizar alguna de las métricas de validación de modelos apta para problemas de regresión o cuando la variable respuesta es continua. La más utilizada es la media de los errores elevados al cuadrado (*Mean Square Error*), pero al estar en unidades cuadradas, y en aras de una interpretación más sencilla, se utiliza su raíz cuadrada: *Root Mean Square Error* (RMSE).

Si nos detenemos en la **Figura 26** podemos observar que el modelo que contiene todas las variables (16) es el que más reduce el RMSE. Sin embargo, es de reseñar que la mejora a partir de 12-13 variables es muy reducida, por lo que siguiendo el principio de parsimonia se podría prescindir de algunas variables. Entonces, la pregunta que surge es: ¿qué variables eliminamos o serían susceptibles de eliminación?

[34] Una explicación detallada de su funcionamiento: https://rpubs.com/Joaquin_AR/383283

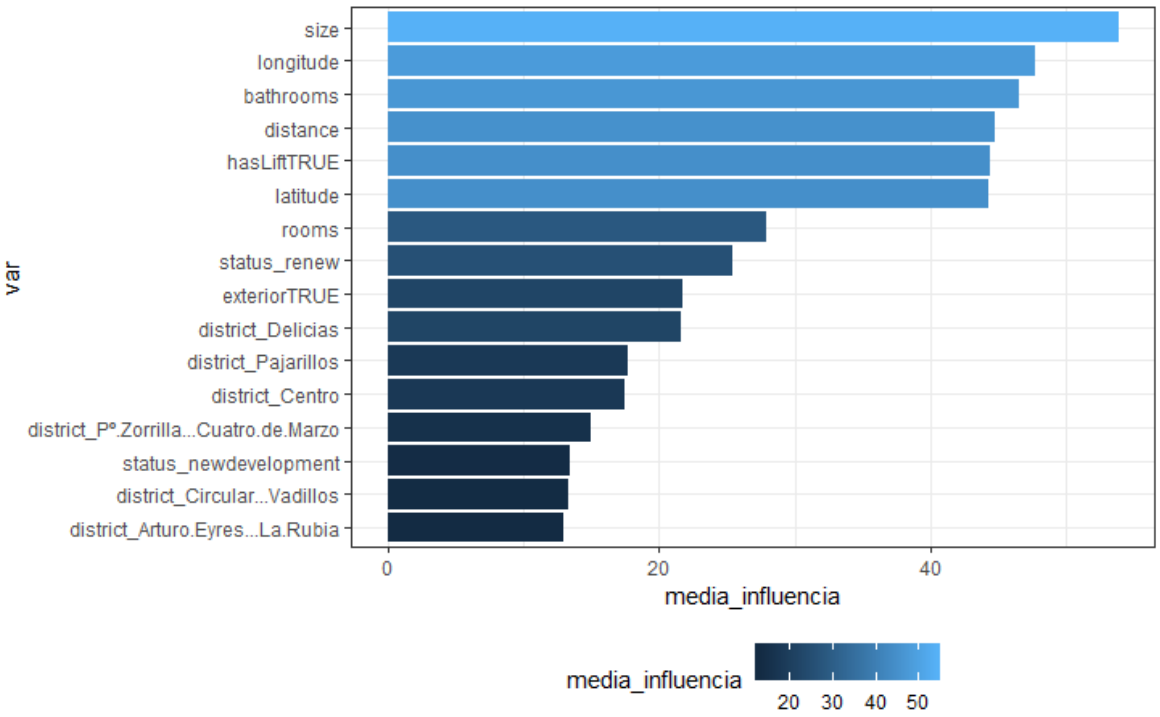
Figura 26: Resultados de la eliminación recursiva de variables



Fuente: Elaboración propia a partir de datos de Idealista

Para dar respuesta a la pregunta anterior se representa la influencia media que tiene cada variable del mejor modelo en la reducción del RMSE (**Figura 27**). Nosotros utilizaremos las 16.

Figura 27: Influencia media de cada variable del mejor modelo



Fuente: Elaboración propia a partir de datos de Idealista

6.3.8. Entrenamiento de modelos predictivos y modelado (fine-tuning)

Una vez los datos han sido transformados y analizados, nos adentramos en el núcleo del proceso de ML: "*emplear un algoritmo capaz de representar los patrones presentes en los datos de entrenamiento y generalizarlos a nuevas observaciones*" (Amat Rodrigo, 2018). Es decir, en esta fase se aplicarán diferentes algoritmos de ML sobre el conjunto de los datos de entrenamiento para que el modelo aprenda validando ese aprendizaje mediante su contrastación con el subconjunto de test. Entre tanto, al tener cada uno de los algoritmos de ML sus propios parámetros estos se ajustarán progresivamente hasta encontrar el mejor modelo. De esta forma, el propósito en este apartado es el de encontrar el modelo que tenga mejor capacidad predictiva de entre los múltiples modelos que cada algoritmo va a entrenar (*fine-tuning*). El mejor modelo será aquel que presente menor RMSE al predecir la variable *price*. Además, al introducirse la variable *price* con transformación logarítmica, los resultados que se obtienen respecto del error también se expresarán en términos logarítmicos. Será en el siguiente apartado, destinado a la predicción de nuevas observaciones, cuando mostremos el error en los términos originales del precio de las viviendas

Por último, como medida de validación de los modelos se hace uso de la técnica de *Repeated Cross-Validation* que implica ajustar y evaluar los modelos un número determinado de veces, cada vez con una partición distinta, incluyendo un último ajuste con todo el subconjunto de entrenamiento (Amat Rodrigo, 2018).

Los algoritmos que se emplean son:

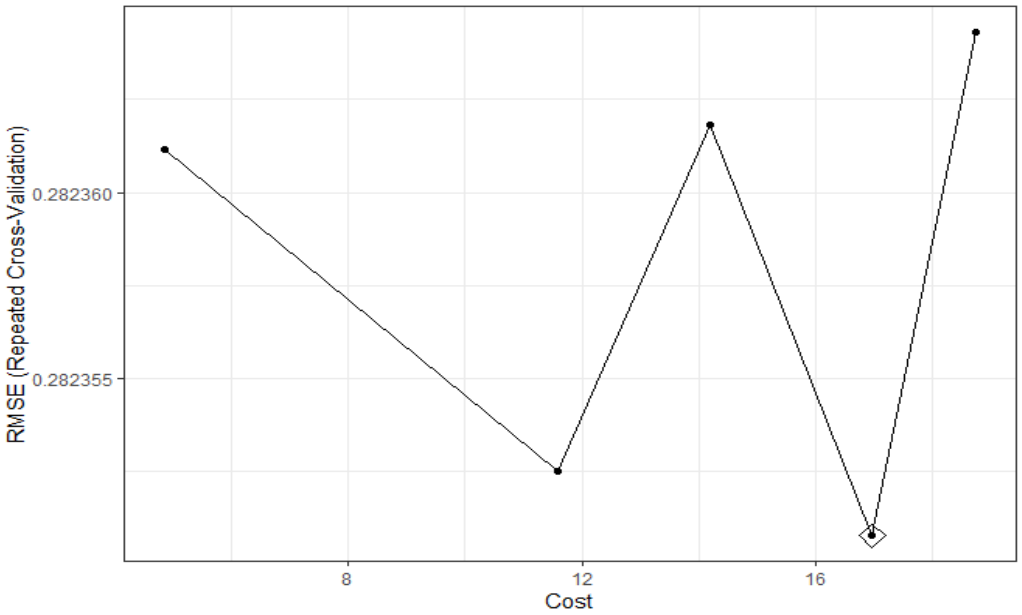
- ***Support Vector Machine***[35] (SVM). Válido tanto para problemas de clasificación como de regresión. Se fundamenta en la idea de encontrar el hiperplano que mejor divida un conjunto de datos en dos clases (Bambrick, 2016). El parámetro a ajustar de este algoritmo será C -el coste asociado a una clasificación errónea del hiperplano- (Toscano Pinel, 2019).

En este caso se entrenan varios modelos con el algoritmo *Support Vector Machine* de tipo lineal, de ahí que se utilice el método *svmLinear*. En concreto, se programa el modelo para que el factor de coste C se ajuste entre los valores 0,001 y 20. El modelo que más

[35] Máquinas de Vector de Soporte

logra reducir el RMSE en la predicción del precio es aquel cuyo factor de coste es de 16,96 (**Figura 28**). Aun así, las diferencias entre modelos son muy reducidas.

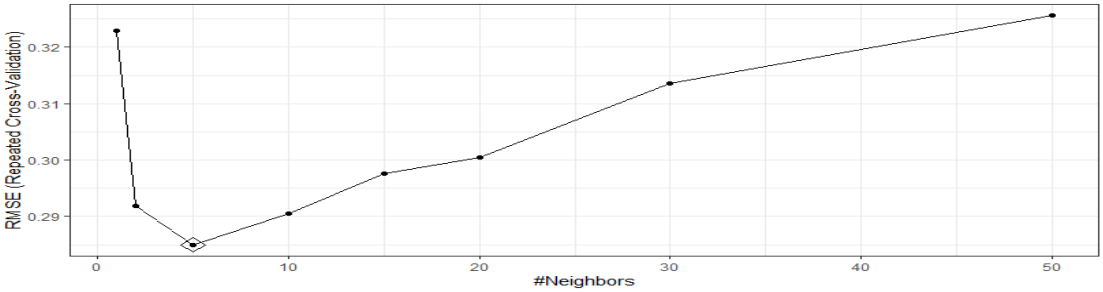
Figura 28: Evolución de RMSE en función del factor C (SVM)



Fuente: Elaboración propia a partir de datos de Idealista

- ***K-Nearest Neighbor*** (KNN). Es con diferencia el más sencillo. Trata de identificar las observaciones del subconjunto de test que se asemejen a las del subconjunto entrenamiento para otorgarles como valor la clase predominante. (Amat Rodrigo, 2018). Además, es un algoritmo que solo contiene un único hiperparámetro: K -número de observaciones vecinas-. Concretamente, para el entrenamiento de este algoritmo el número de observaciones vecinas que emplearemos estará en el rango [1,50].

Figura 29: Evolución de RMSE en función del nº de observaciones vecinas (KNN)

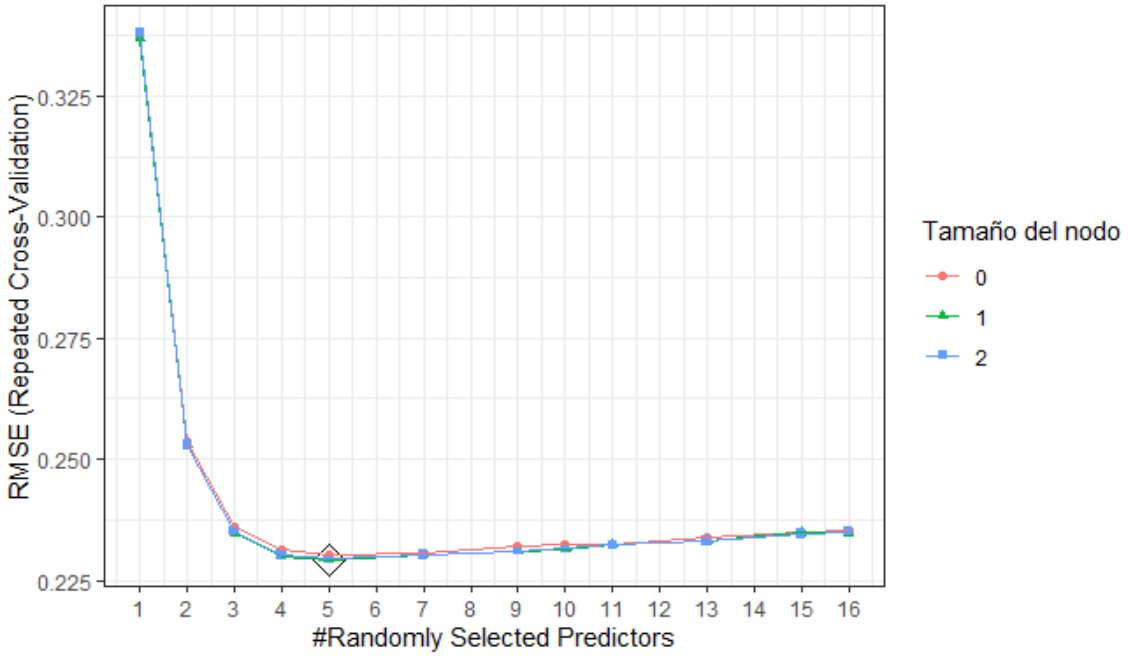


Fuente: Elaboración propia a partir de datos de Idealista

Como muestra la **Figura 29**, el número de observaciones vecinas que consigue un mejor poder predictivo en el entrenamiento de modelos es el de 5.

- ***Random Forest*** (RF). Perteneciente a la familia de algoritmos basados en los árboles de decisión, este algoritmo combina los resultados obtenidos de varios árboles de decisión pero con la gran diferencia de que en cada iteración se introducen una serie de variables de forma aleatoria (Toscano Pinel, 2019). Los parámetros ajustados son: *mtry* -número de predictores utilizados aleatoriamente-, *min.node.size* -tamaño mínimo de cada nodo para efectuar la división de los árboles- y *splitrule* -criterio de la división-. A mayores, habrá que determinar el número de árboles óptimo para evitar un excesivo coste computacional.

Figura 30: Evolución de RMSE en función de los parámetros (*Random Forest*)



Fuente: Elaboración propia a partir de datos de Idealista

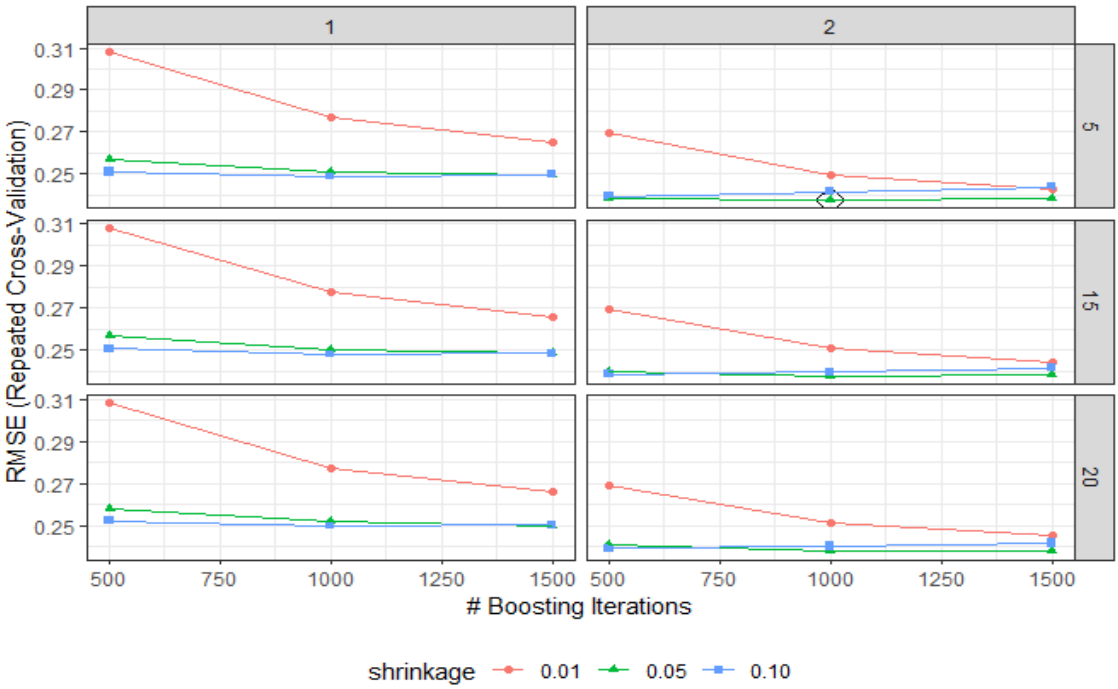
A la luz de los resultados, el modelo de *Random Forest* que devuelve un mejor comportamiento en la predicción de *price* tiene un nodo de tamaño 1 y un total de 5 variables aleatoriamente escogidas (**Figura 30**).

- ***Gradient Boosting*** (GB). Hablamos de una versión posterior a *Random Forest* dentro de la familia de árboles de decisión. La gran novedad reside en el factor de regularización, pues permite minimizar el error de predicción gracias a la corrección recíproca que hacen unos árboles de los otros respecto de dicho error. Será el algoritmo con mayor número de parámetros (4): el número de árboles -*n.trees*-, el número de divisiones realizado a cada árbol -*interaction.depth*-, el parámetro de regularización -*shrinkage*- y número mínimo de

observaciones que un nodo tiene que tener para producirse la división -*n.minobsinnode*- (Amat Rodrigo, 2018).

Para ajustar el algoritmo de *Gradient Boosting* partiremos de tres cifras distintas de árboles (500, 1000 y 1500), así como dos cifras diferentes para las divisiones de cada árbol (1 y 2). Por su lado, el parámetro de regularización *shrinkage* tendrá tres cifras distintas de divisiones (5, 15 y 20), y el número de observaciones para cada nodo podrá ser de 5, 15 y 20.

Figura 31: Evolución de RMSE en función de los parámetros (*Gradient Boosting*)



Fuente: Elaboración propia a partid de datos de Idealista

La **Figura 31** muestra que para el algoritmo *Gradient Boosting* la combinación de parámetros que minimiza el error para predecir el precio de la viviendas es la que engloba 2 divisiones en cada árbol, 1000 árboles, y un factor de regularización de 0,05.

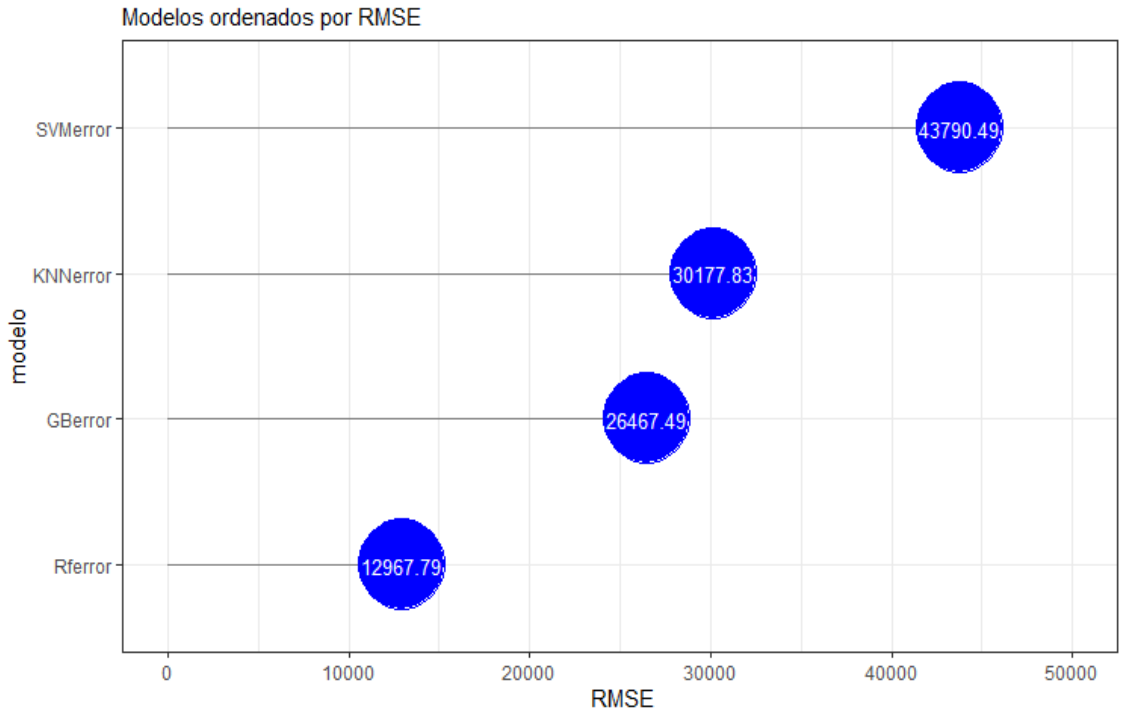
6.3.9. Evaluación comparativa de todos los modelos en función del poder predictivo

La idea fundamental que descansa en todo proyecto de ML es que los modelos tengan capacidad para predecir lo más acertadamente el precio de viviendas con las que no se haya entrenado. Para ello, los modelos entrenados con el subconjunto de entrenamiento ahora son testados con el subconjunto de test. Aquel modelo que presente

un error de test más bajo en la predicción del precio será el más adecuado de entre los entrenados. Dicho esto, para conseguir una mejor visualización de los resultados se realiza una evaluación comparativa del mejor ajuste que presenta cada algoritmo de los anteriormente explicados.

En primer lugar, comprobamos los resultados predictivos que presenta cada algoritmo en la predicción del precio de las viviendas del subconjunto entrenamiento. A priori, el modelo debería de comportarse mejor sobre el conjunto de datos que sobre el subconjunto de test pues son las observaciones sobre las que ha podido aprender los patrones que explican el precio de las viviendas.

Figura 32: Error medio de predicción del subconjunto de entrenamiento



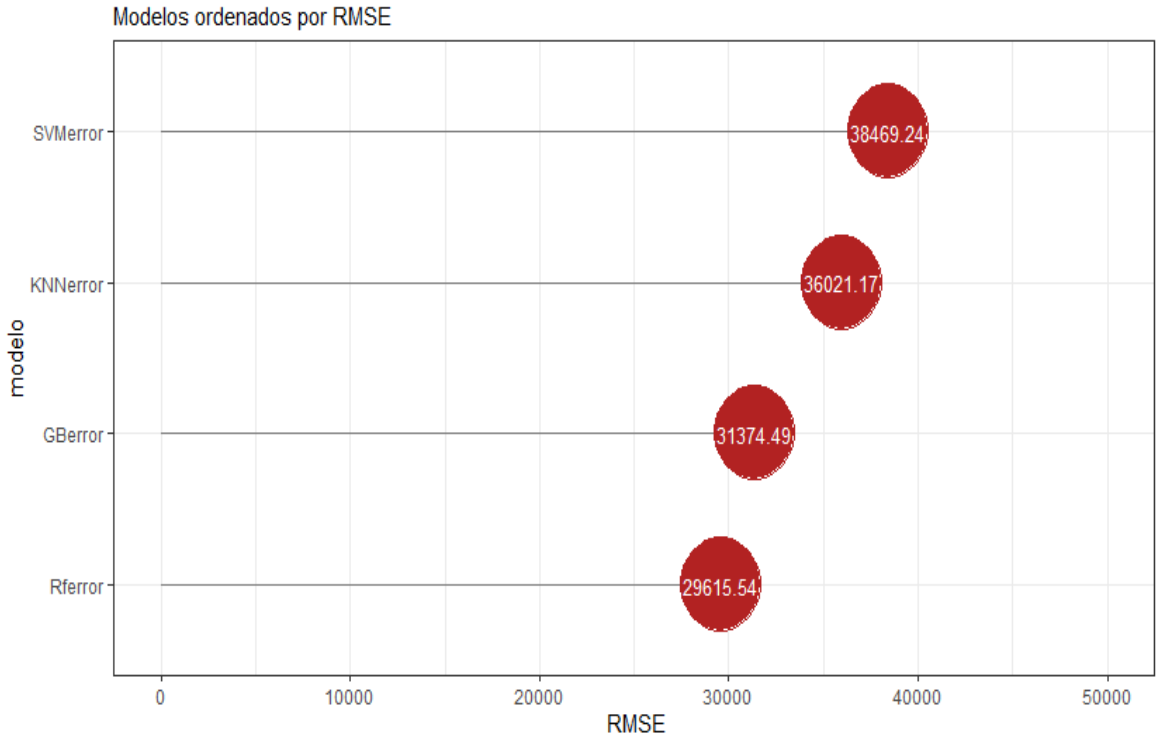
Fuente: Elaboración propia a partir de datos de Idealista

Como se puede observar (**Figura 32**), *Random Forest* es, con notoria diferencia, el algoritmo que mejor predice el precio de los inmuebles en el subconjunto de entrenamiento -alrededor de 13.000€-. En cambio, el algoritmo *Support Vector Machine* tiene un error mucho mayor, sobre los 43.000€.

Estos resultados, aun teniendo cierta fiabilidad, deben ser contrastados con observaciones nuevas, es decir, con el subconjunto de test. Mediante esta técnica es fácil

comprobar si en el aprendizaje ha habido *overfitting* o por el contrario, *underfitting*. La primera situación implica un sobre aprendizaje del modelo sobre el conjunto de entrenamiento, lo que impide llevar a cabo una generalización de las predicciones a otras viviendas no observadas previamente. Por su parte, el concepto de *underfitting* es, en términos sencillos, que el modelo no trabaja del todo bien con el conjunto de datos porque sus características no mezclan bien con el algoritmo utilizado.

Figura 33: Error medio de predicción del subconjunto de test



Fuente: Elaboración propia a partir de datos de Idealista

A este respecto, parece que claramente *Random Forest* padece de un problema de *overfitting*. El error varía desde los 13.000€ del subconjunto de entrenamiento a los casi 30.000€ de la predicción del subconjunto test. Razonamiento, además, extrapolable a los modelos de *Gradient Boosting* y *K-Nearest Neigbors* pero en ambos casos la diferencia de RMSE que hay para uno y otro subconjunto de datos es asimilable (alrededor de los 5.000€). Finalmente, el modelo de *Support Vector Machine*, además de tener predicciones bastante insatisfactorias, presenta síntomas de *underfitting*. Consigue reducir en mayor medida el error medio de predicción cuando trabaja con el subconjunto test que cuando lo hace con el de entrenamiento.

En general, el poder predictivo de todos los modelos no es excesivamente bueno. Un error medio de 30.000€ a la hora de predecir el precio de una vivienda no puede servir, desde luego, como fundamento para asesorar en alguna compra el precio al que se debería de adquirir un inmueble en concreto o, en una venta, a qué precio se debería vender para obtener una rentabilidad, por poner algunos ejemplos. Pero, en cambio, sí que debe servir como punto de partida para encontrar el modelo que más optimice los patrones de comportamiento de los precios en una ciudad como Valladolid. Por los resultados, tanto el algoritmo de *Support Vector Machine* como el de *K-Nearest Neighbor* serían descartados en futuros intentos de optimización, mientras que *Gradient Boosting* y *Random Forest*, al comportarse mejor que los otros dos, ganan enteros para próximos proyectos.

Por último, notar que este proyecto ha tenido, respecto de la precisión predictiva, una limitación importante en el número de datos. No hay que dejar de lado el hecho de que los algoritmos de ML, cuanto mayor campo de aprendizaje tienen, mejor es el resultado predictivo que generan sobre nuevas observaciones. Contar con 2657 observaciones es una gran aproximación a lo que podría consistir un proyecto real y profesional, pero lograr resultados más esperanzadores requiere de un mayor número de datos.

6.4. Futuros proyectos

Aunque ha quedado demostrado cuán es la importancia y las diferentes posibilidades que otorga el Big Data inmobiliario a cualquier operador, probablemente a más de uno a lo largo de la lectura del documento se le habrán ocurrido distintas ideas a las que aquí se han plasmado. Es aquí donde realmente reside el verdadero valor añadido que cada uno puede dotar al análisis de datos. Para finalizar la parte práctica, se exponen dos ideas de análisis que surgen a raíz de este proyecto:

1. Combinar los registros de inmuebles de distintos portales inmobiliarios. Además de utilizar las observaciones que Idealista provee, existen otros portales que también ponen a disposición de los usuarios parte de sus datos. De hecho, gracias al trabajo de investigación perpetrado para este proyecto se han descubierto nuevos lugares de los que obtener datos como es el caso de Realo[36] o la comunidad

[36] https://www.realo.es/en/about

de ROpenSpain[37] que pone a disposición distintos códigos de R que permiten, por ejemplo, descargar datos de la Sede Electrónica del Catastro de España. Algunos proyectos más avanzados incluirían la utilización de técnicas de *web scrapping* (extracción de contenidos y datos de las páginas web sin necesidad de API).

2. Diseñar un modelo conjunto (*ensemble model*) que integre las predicciones acertadas de cada algoritmo. Es decir, aunque un algoritmo no sea capaz de predecir toda la verdad, sí que lo puede hacer para una parte de ella. Por ello, si cada uno de esos modelos consigue descubrir partes de esa verdad, al emplearlos de manera conjunta, esas carencias predictivas que tenía cada uno de los modelos por separados se consiguen minimizar (Amat Rodrigo, 2018).

Estas son solo algunas de las ideas que han surgido con motivo de la realización de este proyecto, sin perjuicio, como es obvio, de todas las posibles mejoras con las que puede contar este proyecto.

[37] https://ropenspain.es/paquetes/

7. CONCLUSIÓN

Es innegable que el uso de las nuevas tecnologías ha llegado al sector inmobiliario para quedarse. El sector del ladrillo comienza a vislumbrar una nueva era (**PropTech**), donde el desconocimiento y la aversión a la innovación tecnológica dejan paso a un entramado enteramente digital. Aun así, la inmersión de la innovación tecnológica en la industria inmobiliaria tiene todavía mucho camino por recorrer en gran parte de los mercados internacionales, incluido el español: un mercado eminentemente *clusterizado* por las metrópolis urbanas de Madrid y Barcelona que deja entrever importantes carencias desde el punto de vista de la adopción tecnológica, pero que también pone de manifiesto el importante número de oportunidades inmobiliarias que existen fuera de esas dos ciudades.

En este contexto, una de las ramificaciones del PropTech que más fuerza ha ganado es la del Big Data. Consecuencia directa de sus singularidades, el sector inmobiliario se convierte en una industria muy proclive para el despliegue de la ciencia de los datos, con las transformaciones que ello implica. Poder proyectar las preferencias de los clientes o minimizar el riesgo intrínseco de cualquier inversión inmobiliaria son solo algunas de las virtualidades que lleva consigo el análisis masivo de datos dentro del sector. Es más, determinadas áreas del Big Data como es el *predictive analytics* o el *descriptive analytics* han creado verdaderos movimientos en torno a ellos, logrando desmarcarse del tradicional análisis inmobiliario. No obstante, la ventaja más disruptiva que esconde el Big Data es, sin lugar a dudas, su poder democratizador. No entiende de grandes, medianos o pequeños, ni siquiera de entendidos o desconocedores, pues todos y cada uno de nosotros tenemos acceso a los datos, como aquí se ha demostrado. Una democratización que, además, impacta de lleno en el sector inmobiliario gracias a la puesta a disposición que hacen distintas plataformas, portales e instituciones de sus datos.

Adentrarse en este nuevo contexto que dibujan la innovación tecnológica y el Big Data inmobiliario debería conformar la nueva hoja de ruta de gran parte de los operadores de la industria. Formar o no formar parte de ello puede significar la diferencia entre seguir o quedarse por el camino. Una realidad que, sin lugar a dudas, duele más si cabe cuando "subirse al carro" de la transformación digital nunca ha sido tan sencillo ni ha estado al alcance de tantos

De forma práctica, a través del análisis de la oferta residencial de la ciudad de Valladolid y sus conclusiones predictivas (**APLICACIÓN PRÁCTICA**), hemos demostrado como estos cambio están también al alcance de cualquier empresa (independientemente de sus características), incluso para aquellas que piensan que el cambio hacia el Big Data inmobiliario es un imposible o algo redundante para sus modelos de negocio. Ser una inmobiliaria familiar y tradicional que cuenta con pocos recursos y que nunca antes ha probado este tipo de herramientas, no puede ser, de ningún modo, razón tal para evitar dicha transformación. Entender y comprender esta circunstancia tiene que conformar el núcleo sobre el que se asiente la nueva era del sector. En definitiva, una nueva que trae consigo un nuevo sector.

8. BIBLIOGRAFÍA

Amat Rodrigo, J. (abril de 2017). *Máquinas de Vector de Soporte (Support Vector Machine)* https://rpubs.com/Joaquin_AR/267926

Amat Rodrigo, J. (abril de 2018). *Machine Learning con R y Caret* https://rpubs.com/Joaquin_AR/383283

Bambrick, N. (21 de junio de 2016). *Support Vector Machine for dummies; A simple Explanation* https://blog.aylien.com/support-vector-machines-for-dummies-a-simple-explanation/

Barter, R. (6 de junio de 2019). *Using the recipes package for easy pre-processing* http://www.rebeccabarter.com/blog/2019-06-06_pre_processing/

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

Baum, A., & Dearsley, J. (2017). *What is PropTech?* https://www.unissu.com/proptech-resources/what-is-proptech

Baum, A., Saul, A., & Braesemann, F. (2020). *PropTech 2020: the future of real estate.* University of Oxford , Saïd Business School. https://www.sbs.ox.ac.uk/sites/default/files/2020-02/proptech2020.pdf

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: provocations for a cultural, technological and sholarly phenomenon. En *Information, Communication, and Society* (Vol. 15, pp. 662-679). https://doi.org/10.1080/1369118X.2012.678878

Brownlee, J. (28 de julio de 2017). *Why One-Hot Encode Data in Machine Learning?* https://machinelearningmastery.com/why-one-hot-encode-data-in-machine learning/

Catella Research. (2015). *Big data in the real estate sector - a big opportunity or a big threat?* https://mb.cision.com/Public/6412/9849658/9f9655d86b42e977.pdf

Chaillou, S., Fink, D., & Gonçalves, P. (2017, diciembre 31). *Urban Tech on the Rise: Machine Learning Disrupts the Real Estate Industry.* http://journals.openedition.org/

Deloitte. (febrero de 2018). Data is the new gold: the future of real estate service providers. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Public-Sector/gx-real-estate-data-new-gold.pdf

Donner, H., Eriksson, K., & Steep, M. (2018, enero). *Digital Cities: Real Estate Development Driven by Big Data.* https://www.researchgate.net/publication/325253311

Goodwin. (4 de Marzo de 2019). *The PropTech Industry is set for a wave of consolidation*. https://www.goodwinlaw.com/publications/2019/03/03_04-proptech-pulse

Khan, A., Uddin, M., & Grupta, N. (2014). *Seven V's of Big Data: Understanding Big Data to extract Value.* Institute of Electrical and Electronics. Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1) http://www.asee.org/documents/zones/zone1/2014/Professional/PDFs/113.pdf

Kumar , V., & Garg, M. (2018, julio). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Science, 182*(1), pp. 31-37. https://www.researchgate.net/publication/326435728

Lenk, A., Bonorden, L., Hellmans, A., Roedder, N., & Jaehnichen, S. (2015). Towards a Taxonomy of Standards in Smart data. Proceedings of 2015 IEEE Conference on Big Data. https://www.researchgate.net/publication/283548444_Towards_a_Taxonomy_of_Stanards_in_Smart_Data

Lin, H., & Li, M. (12 de febrero de 2020). Sparse Variables. En *Introduction to Data Science* (pp. 95-96). Próxima publicación en CRC Press. https://scientistcafe.com/ids/IDS.pdf

Madden, S. (2012, mayo/junio). From databases to big data. *IEEE Computer Society*, pp. 4-6. https://pdfs.semanticscholar.org/e275/f643c97ca1f4c7715635bb72cf02df928d06.pdf

Menin Machado, G. (2019). *Big Data analytics for Real Estate Asset Management.* Trabajo Fin de Máster en Gestión del Entorno Construido Politecnico di Milano, Shool of Architecture, Urban Planning, and Construction Engineering, Milan. https://www.politesi.polimi.it/bitstream/10589/148475/1/Thesis%20-%20Gabriela%20Menin%20Machado.pdf

Morgan Asaftei, G., Doshi, S., Means, J., & Sanghovi, A. (8 de octubre de 2018). *Getting ahead of the market: How big data is transforming real estate*. https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., . . . Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-sacale data mining: a survey. *Artificial Intelligence Review*, 52, pp. 77-124. https://doi.org/10.1007/s10462-018-09679-z

Oluwunmi, A., Role, A., Akinwale, O., Oladayo, O., & Afolabi, T. (2019). Big Data and Real Estate: A Review of Literature. *Journal of Physics: Conference Series* 1378: 032015. https://www.researchgate.net/publication/338013966_Big_Data_And_Real_Estate_A_Review_Of_Literature

Oracle. (marzo de 2020). *¿Qué es big data?* https://www.oracle.com/es/big-data/guide/what-is-big-data.html

Paluri, S. (26 de agosto de 2016). *Solving the Challenges of Public Records Data*. https://www.zillow.com/tech/public-data-challenges/

PropTech house. (Julio de 2019). *Demystifying PropTech - a complete overview of the European PropTech ecosystem.* (I. Goossens, Ed.) https://www.proptech.nl/wp-content/uploads/2019/06/eBook_Demystifying-PropTech_a-complete-overview-of-the-European-PropTech-ecosystem_2019.pdf

Red Hat. (s.f.). *What is open source?* https://www.redhat.com/en/topics/open-source/what-is-open-source

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World - From Edge to Core.* International Data Corporation.White Paper #US444133118. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Royal Institution of Chartered Surveyors. (2017, noviembre). *The Future of Valuations.* https://www.rics.org/globalassets/ricswebsite/media/knowledge/research/insights/future-of-valuations-insights-paper-rics.pdf

Royal Institution of Chartered Surveyors. (2019, abril). *The use and value of commercial property data.* https://www.rics.org/globalassets/ricswebsite/media/knowledge/research/insights/value-and-commercial-property-data-rics.pdf

Savills Aguirre Newman. (2019, marzo). *Informe PropTech 2019.* https://proptech.es/wp-content/uploads/informe-proptech-spain-2019.pdf

Sociedad de Tasación. (septiembre de 2019). *Tendencias del Sector Inmobiliario.* https://www.sttasacion.es/ext/pdf/estudios/dic19/1Informe_de_Tendencias_del_Sector_Inmobiliario-Enero_2020.pdf

Souissi, N., & El Arass, M. (2018). Data Lifecycle: From Big Data to Smart Data. Conference Paper 5THEDITION INTERNATIONAL IEEE CONGRESS on INFORMATION SCIENCE and TECHNOLOGY (CiSt'18), Marruecos. https://www.researchgate.net/publication/328769944

Toscano Pinel, P. (septiembre de 2019). Rentalbility - Predicting the profitability of rental of properties in Madrid, a kick off for a tool to help small investors. Trabajo Fin de Máster en minería de datos e inteligencia de negocios. Universidad Complutense de Madrid. https://eprints.ucm.es/57354/1/TFM_Rentalbility_PriscillaToscano_VF_160919.pdf

Unissu. (agosto de 2019). *Global PropTech Analysis: Europe*. https://www.unissu.com/proptech-resources/proptech-europe

Winson-Geideman, K., & Krause, A. (2016). *Transformation in Real Estate Research: The Big Data Revolution.* University of Melbourne. Proceedings of the 22nd Annual Pacific-Rim Real Estate Society Conference, Queensland, Australia, pp. 17-20. http://www.prres.net/papers/Geideman_Transformations_in_RE_Research.pdf

World Economic Forum. (2015). *The Future of FinTech: A Paradigm Shift in Samall Business Finance.* http://www3.weforum.org/docs/IP/2015/FS/GAC15_The_Future_of_FinTech_Paradigm_Shift_Small_Business_Finance_report_2015.pdf

9. ANEXO

Anexo A: Descripción de variables originales obtenidas de Idealista

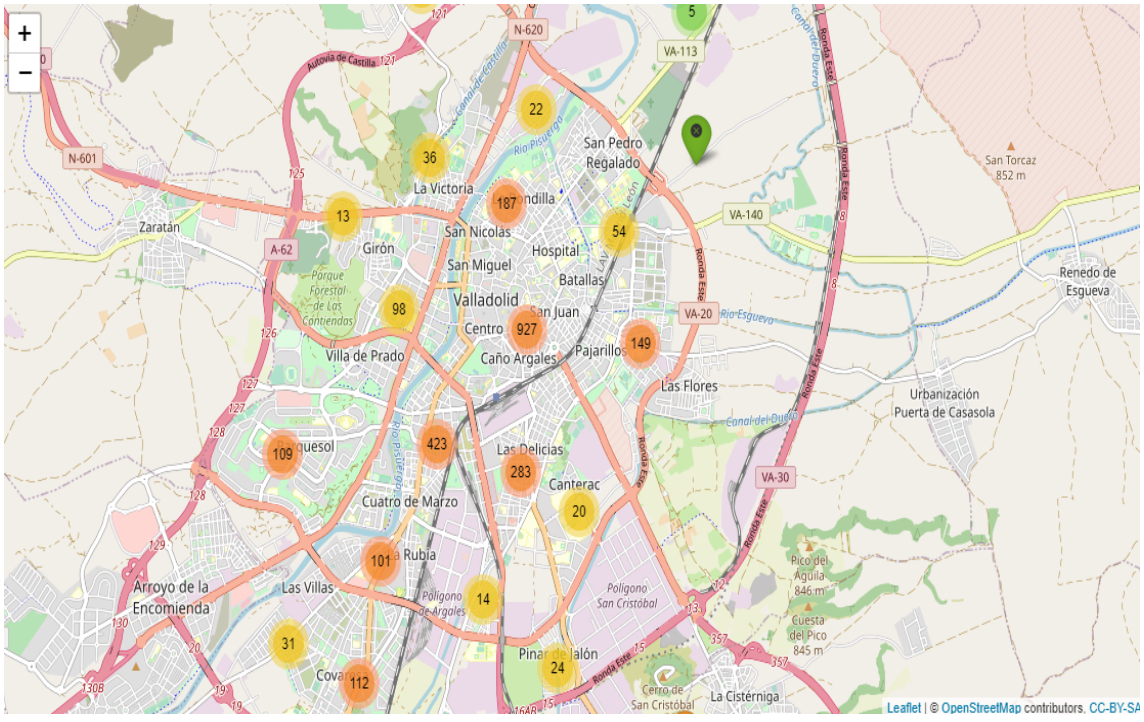
#	Variable	Descripción	Tipo de Variable	Rol
1	***propertyCode***	Código de Identificación de Idealista	Character	Aceptado
2	***thumbnail***	URL fotos	Character	Rechazado
3	***External Reference***	Referencia externa	Character	Rechazado
4	***numPhotos***	Número de fotos disponibles	Integer	Aceptado
5	***floor***	Piso	Character	Aceptado
6	***price***	Precio	Double	Aceptado
7	***propertyType***	Tipo de propiedad	Character	Aceptado
8	***operation***	Venta o alquiler	Character	Rechazado
9	***size***	Tamaño	Double	Aceptado
10	***exterior***	Exterior	Logical	Aceptado
11	***rooms***	Nº habitaciones	Integer	Aceptado
12	***bathrooms***	Nº baños	Integer	Aceptado
13	***address***	Dirección	Character	Aceptado
14	***province***	Provincia	Character	Rechazado
15	***municipality***	Municipio (Valladolid)	Character	Aceptado
16	***district***	Distrito/Barrio	Character	Aceptado
17	***country***	País	Character	Rechazado
18	***latitude***	Coordenadas (latitud)	Double	Aceptado
19	***longitude***	Coordenadas (longitud)	Double	Aceptado
20	***showAddress***	Tiene dirección	Logical	Rechazado
21	***url***	URL	Character	Aceptado
22	***distance***	Distancia a la Plaza Mayor	Character	Aceptado
23	***hasVideo***	Tiene vídeo	Logical	Rechazado
24	***status***	Estado	Character	Aceptado
25	***newDevelopment***	Nueva construcción	Logical	Rechazado
26	***hasLift***	Tiene ascensor	Logical	Aceptado
27	***priceByArea***	Precio por metro cuadrado (m2)	Double	Aceptado
28	***hasPlan***	Tiene plano	Logical	Aceptado
29	***has3DTour***	Tiene vídeo 3D	Logical	Rechazado
30	***has360***	Tiene visión 360	Logical	Rechazado
31	***topNewDevelopment***	NA	Logical	Rechazado
32	***newDevelopmentFinished***	Nueva construcción finalizada	Logical	Rechazado
33	***parkingSpace.hasParkingSpace***	Tiene plaza de garaje	Logical	Rechazado
34	***parkingSpace.isParkingSpaceIncludedInPrice***	Garaje incluido en precio	Logical	Rechazado
35	***parkingSpace.parkingSpacePrice***	Precio plaza de garaje	Double	Rechazado
36	***detailedType.typology***	Tipología detallada	Character	Rechazado
37	***detailedType.subTypology***	Subtipología detallada	Character	Rechazado
38	***suggestedTexts.subtitle***	Subtítulo sugerido	Character	Rechazado
39	***suggestedTexts.title***	Título sugerido	Character	Rechazado

Fuente: Elaboración propia a partir de datos de Idealista

Anexo B: Distribución geográfica de los barrios de la ciudad de Valladolid

Fuente: Idealista

Anexo C: Representación geoespacial de la oferta residencial en Valladolid



Fuente: Elaboración propia a partir de datos de Idealista