COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales (ICADE)

PREDICCIÓN DEL VALOR DE VENTA DE PISOS MADRILEÑOS CON DEEP LEARNING Y ENSEMBLES, TRABAJO FIN DE GRADO

Autor: Gonzalo V. Egea Hernández
Director: Eduardo César Garrido Merchán

MADRID | Junio 2023

RESUMEN

En el presente trabajo se lleva a cabo un estudio preliminar del mercado inmobiliario, analizando los diferentes factores que influyen en el precio de los inmuebles, así como los movimientos resultantes en la oferta y la demanda. Posteriormente, se realiza una revisión teórica de los modelos de predicción a implementar; bosques aleatorios, xgboost y un nuevo modelo, TabPFN, caracterizado a priori por tener obtener resultados robustos en muestras de datos reducidas, así como el uso de machine learning y datos en formato de tabla. Finalmente, se describe en detalle el proceso de implementación de estos modelos, incluyendo un análisis preliminar de las variables, la limpieza y preparación previa de los datos, y posterior comparación de los resultados obtenidos por cada modelo. Además, se desarrolla una herramienta de predicción que facilita la clasificación de futuros pisos, y se propone su entrenamiento y prueba con muestras más grandes de datos.

Palabras clave: Sector inmobiliario, inmueble, valoración inmobiliaria, Machine learning, aprendizaje supervisado, estandarización, datos tabulares, modelos de predicción, precisión del modelo.

ABSTRACT

In this paper, a preliminary study of the real estate market is carried out, analyzing the different factors that influence the price of real estate, as well as the resulting movements in supply and demand. Subsequently, a theoretical review of the prediction models to be implemented is carried out; Random forests, Xgboost and a new model, TabPFN, characterized a priori by having to obtain robust results in reduced data samples, as well as the use of machine learning and data in table format. Finally, the implementation process of these models is described in detail, including a preliminary analysis of the variables, data cleaning and preparation, and subsequent comparison of the results obtained by each model. In addition, a prediction tool is developed to facilitate the classification of future floors, and its training and testing with larger data samples is proposed.

Key words: Real estate, housing, Machine learning, supervised learning, standardization, tabular data, predictive models, Accuracy.

Índice de la memoria

Sección 1: Introducción 5
Sección 2: Estado del arte 7
Sección 3: Real State en España y análisis de datos tabulares con Machine Learning 9
3.1.- El mercado de la vivienda y su evolución temporal 9
3.2.- Factores determinantes del precio de la vivienda 15
3.3.- Datos tabulares con ML 18
Sección 4: Marco de la tesis 21
4.1.- Asunciones, objetivos, hipótesis y restricciones 21
4.2.- Planificación del trabajo 22
Sección 5: Metodología 25
5.1.-Análisis cualitativo 25
5.1.1.- Extracción de datos 25
5.1.2.- Limpieza y preparación de los datos 29
5.1.3.- Análisis exploratorio de los datos 33
5.1.4.- Machine Learning 39
5.1.5.- Descripción analítica de los modelos. 42
Sección 6: Experimentos 48
6.1.- Análisis cuantitativo 48
6.1.1.- Random Forest 50
6.1.2.- XGBOOST 53
6.1.3.- TabPFN 55
6.1.4.- Extra: KNN 56
6.1.5.- Herramienta para la predicción 57
Sección 7: Conclusiones y trabajo futuro 59
Sección 8: Bibliografía 61
Sección 9: Anexos 68

Índice de figuras

Figura 1: Variación Trimestral del IPV. Total Nacional 9
Figura 2: Índice de Precios de la Vivienda (IPV), Viviendas segunda mano
Figura 3: Índice de Precios de la Vivienda (IPV), Vivienda nueva 10
Figura 4: Precio de la vivienda libre en 2019. Porcentaje del máximo anual histórico. 10
Figura 5: Tasas de variación del IPV último trimestre 2022 11
Figura 6: Valor tasado de la vivienda libre en España desde el primer trimestre de 2015 hasta el primer trimestre de 2021 12
Figura 7: Evolución del precio de la vivienda y el PIB (Tasa de crecimiento interanual, en porcentaje) 13
Figura 8: Gráfico PIB e inversión sector construcción 13
Figura 9: Crecimiento medio anual crisis financiera vs pandemia 14
Figura 10: Estadística de Transición de Derechos de la propiedad, Total Nacional, Vivienda Libre 15
Figura 12: Esquema de un mercado inmobiliario y sus componentes de oferta. corto plazo 16
Figura 13: Características de los modelos Multidimensionales y Tabulares 19
Figura 14: Gráfico resumen del proyecto 24
Figura 15: Captura de pantalla, cuestionario solicitud acceso API idealista 25
Figura 16: Tabla descriptiva de los diferentes filtros aplicables en la API 26
Figura 17: Datos necesarios de acceso API 27
Figura 18: Ejemplificación archivo resultados extraído 29
Figura 19: Ejemplificación error datos extraídos 30
Figura 20: Ejemplificación columna parkingSpace 31
Figura 21: Histograma variable distancia 35
Figura 23: Gráfico de dispersión entre variables priceByArea y distance 36
Figura 24: Gráfico de correlación entre las distintas variables 38
Figura 25: Gráfico de Correlación de variables con p-valores 39
Figura 26: Proceso de construcción de un modelo de Machine Learning 41
Figura 27: Representación de un ensemble 43
Figura 28: Representación de un Bosque Aleatorio 45
Figura 29: Información del dataset 49
Figura 30: Matriz de confusión del modelo Random Forest 51
Figura 31: Tabla de variables y pesos asignados por Random Forest 52
Figura 32: Matriz de confusión del modelo Xgboost 54
Figura 33: Matriz de confusión del modelo TabPFN 55
Figura 34: Niveles de precisión para distintos valores de k 57

Sección 1: Introducción

El mercado inmobiliario es muy complejo dado que se encuentra diariamente afectado por factores como son la oferta y demanda de casas, precios de los materiales, recalificaciones de los terrenos, nueves leyes aplicables a este mercado, tendencias de crecimiento de las ciudades y núcleos de población…etc.

Desde el punto de vista de la inmobiliaria, es crítico estar informado y tener en cuenta todos estos factores para poder encontrar mejores oportunidades de inversión y asesorar de una manera más eficiente a sus clientes con el objetivo de incrementar sus márgenes de beneficio y realizar ventas más rápidas.

Desde el punto de vista del comprador, es importante estar informados de las posibles tendencias y movimientos del mercado para poder realizar la compra eficiente de un inmueble, tanto si se quiere destinar a uso personal como método de inversión.

En el mercado podemos encontrar mucha información sobre el precio de los pisos de acuerdo con sus características, pero son las inmobiliarias las que, en mayor medida, manejan dichos datos debido a que son los principales encargados de recopilar la misma para poder, posteriormente, realizar ofertas o estudiar los mismos con el objetivo de encontrar posibilidades de inversión.

Para los clientes finales, recopilar toda esta información suele ser un trabajo muy tedioso, en el que se suele emplear mucho tiempo y que suele caracterizarse por la búsqueda de información en diferentes portales inmobiliarios y empresas destinadas a dichos servicios. Es muy difícil saber si el precio que están recibiendo en la oferta dista en gran proporción del precio de inmuebles con características similares o están ante una gran oportunidad para llevar a cabo una inversión.

La problemática a estudiar es la gran dificultad por parte del cliente final de poder tomar una decisión eficiente de acuerdo a la información que dispone y es suministrada por parte de las empresas inmobiliarias e intentar solventar la misma mediante la creación de una herramienta en la que, introduciendo una gran cantidad de pisos con sus características principales (número de habitaciones, baños, barrio, distancia al centro…etc), pueda ayudar a este cliente a tomar una decisión más ajustada a sus necesidades de compra y que le permita poder encontrar posibilidades de inversión en dichos activos.

Por ello se empezará realizando un análisis de la actual situación en la que se encuentra dicha problemática, así como los estudios y análisis más recientes con similar temática. Posteriormente, se llevará a cabo un estudio del mercado de la vivienda, los factores que afectan al mismo y la evolución que ha ido experimentando a nivel nacional. Del mismo modo, se realizarán tanto un análisis cualitativo de los modelos a aplicar y la preparación y limpieza de los datos objeto de estudio como un análisis cuantitativo de los resultados obtenidos mediante las distintas implementaciones, así como la construcción de una herramienta para el futuro uso y aplicación, acabando todo ello con las distintas conclusiones obtenidas a lo largo del estudio.

Finalmente, la motivación de este trabajo viene dada principalmente por varios motivos. El primero el estudio e implementación, debido la ausencia, de una herramienta o servicio accesible al que poder acudir e introducir los datos de un piso en concreto que una persona haya encontrado y que dicha herramienta pueda clasificar si dicho inmueble se encuentra por encima o por debajo del precio de mercado de sus homólogos.

En segundo lugar, mi inquietud sobre el mercado inmobiliario y el gran protagonismo que ha ido adquiriendo en los últimos años debido a la búsqueda de métodos de inversión alternativos a la bolsa de valores por parte de algunos inversores influenciado por la gran volatilidad que ha experimentado la misma favorecida por la guerra de Ucrania y la oferta, demanda del sector. Del mismo modo, me encuentro actualmente en período de búsqueda de piso junto a dos compañeros y hemos podido observar como el precio de este tipo de activos he ido escalando de manera significante y nos gustaría poder tomar una decisión de alquiler lo más eficiente posible.

Por último, intentar descubrir las posibilidades y restricciones de este tipo de análisis aplicado a los activos inmuebles dado que he observado que es un tema que teóricamente se ha abordado de múltiples formas e incluse se han aplicado anteriormente algoritmos y modelos de predicción a este tipo de problemáticas, pero no he encontrado una herramienta o software de código libre destinado a las personas que pudiera ofrecerles este servicio.

Sección 2: Estado del arte

En esta sección, se realiza un pequeño análisis de distintos estudios que se han realizado durante los últimos años con un base similar al proyecto a desarrollar, más en concreto, aquellos que combinan el machine learning con la valoración de inmuebles.

En 2018, un grupo de ingenieros científicos de la universidad de Chicago presentaron un artículo focalizado en la predicción del precio de la vivienda mediante la implementación de *Deep Learning*. Tras la aplicación de tres distintos modelos; regresión logística, una red neuronal de convolución (CNN) y un modelo de red neuronal de memoria a corto plazo (LSTM), descubrieron que el error de predicción era demasiado grande debido principalmente a la herramienta de rastreo empleada, *web scrapping*, y la poca representación de los factores generales para tener en cuenta; tanto temporales como macroeconómicos (Yu, et al., 2018).

Dos años después, en 2020, destacan dos proyectos caracterizados por la aplicación de modelos de valoración automatizada (AVMs); programas informáticos que emplean cálculos matemáticos para predecir el valor de un inmueble. En el primero, los AVMs proporcionaron un mayor ahorro económico y de tiempo para valorar carteras con muchos inmuebles, sin embargo, cada AVM estaba basado en un modelo diferente y por ello, la gran diversidad de predicciones vino dada por el algoritmo implementado (Agrasar, 2020). En el segundo, se plantea una visión alejada de los métodos de valoración tradicionales y busca hacer uso del *Big Data* y los métodos de entrenamiento de máquinas; más en concreto los AVMs, debido a las posibilidades que aportan a nivel de captura, almacenamiento, gestión y tratamiento de datos masivos en dicho contexto en desarrollo (San José, 2020).

Ese mismo año, destaca otro proyecto que tiene como objetivo reducir el error en la predicción del precio de un inmueble mediante la aplicación de siete modelos de regresión entre los que destacan; regresión linear, bosques aleatorios, KNN y árboles de decisión, entre otros, frente al modelo empleado por una empresa inmobiliaria; *Generalized Boosted Regression*. Los resultados obtenidos mostraron que el *Random Forest* es el modelo con mayor precisión frente al de la empresa, sin embargo, este último es un mejor modelo predictivo para casas y apartamentos con características más similares. Como

conclusión, insta a la empresa a aplicar métodos de optimización paramétrica con el objetivo de minimizar el error asociado (Bozanic, 2020).

En 2021, destacan principalmente la implementación de un modelo de machine learning para la estimación del valor del metro cuadrado de un inmueble, así como un modelo de predicción del precio de la vivienda con Machine Learning. Este primero se caracteriza por implementar los modelos de regresión lineal, árbol de decisión potencializado y bosque de decisión, sin embargo, su principal restricción es la cantidad de variables utilizadas para la implementación, lo que afectó negativamente a la precisión (García, 2021). Por otro lado, el segundo paper se caracteriza por implementarse tanto de técnicas de aprendizaje supervisado; regresión lineal múltiple y una regresión *Ridge*, como ensembles; *Random forest y Xgboost*, a un dataset de 2.481 registros donde los principales resultados ejemplificaron unos mejores niveles de precisión obtenidos por los modelos de ensembles (Soto & David, 2021).

Finalmente, en 2022, destaca un proyecto de investigación enfocado tanto para el cliente común como los profesionales del sector, basado en el desarrollo de una aplicación móvil entrenada con recomendaciones de expertos pertenecientes al mercado inmobiliario cuya implementación se base en el Machine Learning. Los resultados obtenidos permiten reducir el tiempo de recopilación de los datos por las personas interesadas como también disminuir el tiempo empleado en investigación y búsqueda de mano de obra en dicho sector (Hernández, 2022).

Sección 3: Real State en España y análisis de datos tabulares con Machine Learning

3.1.- El mercado de la vivienda y su evolución temporal

El mercado de la vivienda viene definido por el conjunto de acciones de oferta y demanda de bienes inmuebles pudiendo ser la naturaleza de estos muy distinta; bienes residenciales, comerciales, industriales…etc (Alves & Urtasun, 2019).

Es importante diferenciar entre vivienda libre y vivienda protegida, debido a su distinta naturaleza, para llevar a cabo un estudio realista y consistente de los datos.

La vivienda protegida es aquella que ha sido construida con cualquier tipo de subvención y que presenta ciertas limitaciones de superficie y precio, mientras que la Vivienda libre es aquella en la que no es vivienda protegida en el momento de la tasación (Raya, 2020)

El Índice de Precio de Venta (IPV) es uno de los índices de referencia para comprar y estudiar los movimientos del precio de la vivienda, el cual tiene como objetivo "medir la evolución de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano a lo largo del tiempo" (de Tudela y Torres, 2019).

A nivel nacional, el IPV trimestral ha experimentado grandes variaciones, destacando una subida de 3,2 puntos porcentuales durante el último trimestre de 2020 y el segundo trimestre de 2021 favorecidas por la salida y estabilización de la pandemia, así como una contracción de 3,4 puntos porcentuales durante el año 2022, como se puede observar en la Figura 1.

Figura 1: Variación Trimestral del IPV. Total Nacional

Fuente: Instituto Nacional de Estadística, INE (2023)

Del mismo modo, no hay diferencias significativas en el índice de precios de la vivienda disgregado entre vivienda nueva o vivienda de segunda mano a nivel nacional. Aunque si es remarcable un mayor incremento del precio de la vivienda nueva durante los periodos de 2020 y 2022, como destacan las Figura 2 y 3.

Figura 2: Índice de Precios de la Vivienda (IPV), Viviendas segunda mano

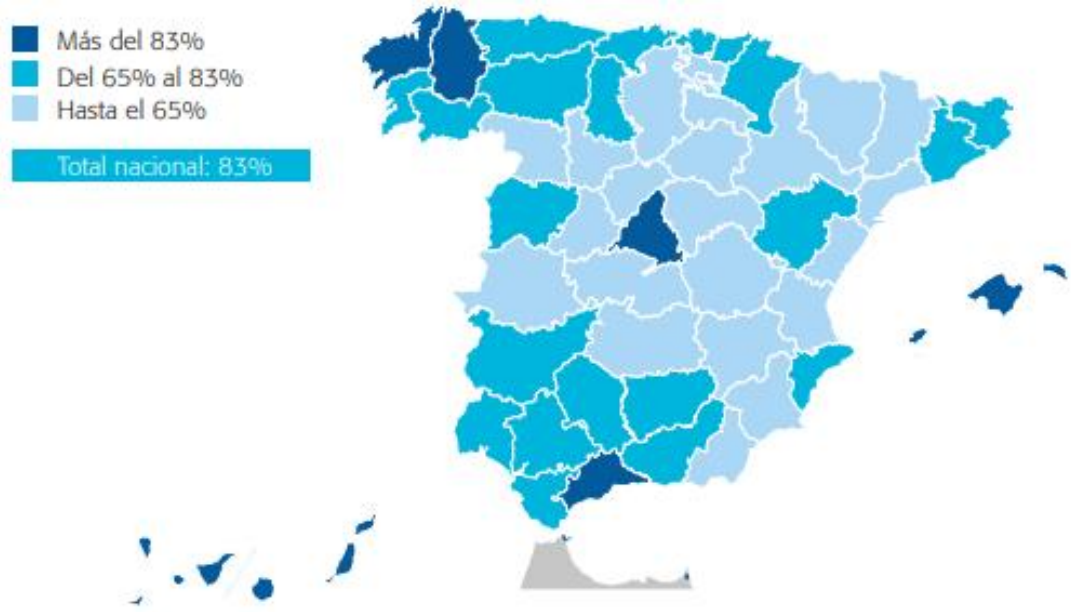
Fuente: Instituto Nacional de Estadística, INE (2023)

Figura 3: Índice de Precios de la Vivienda (IPV), Vivienda nueva

Fuente: Instituto Nacional de Estadística, INE (2023)

A nivel regional, el precio de la vivienda libre en 2019 experimentó un gran crecimiento, siendo este del 83%. Del mismo modo, las comunidades con un mayor incremento fueron Madrid, Galicia, Ceuta, Melilla, Canarias y Baleares (Figura 4).

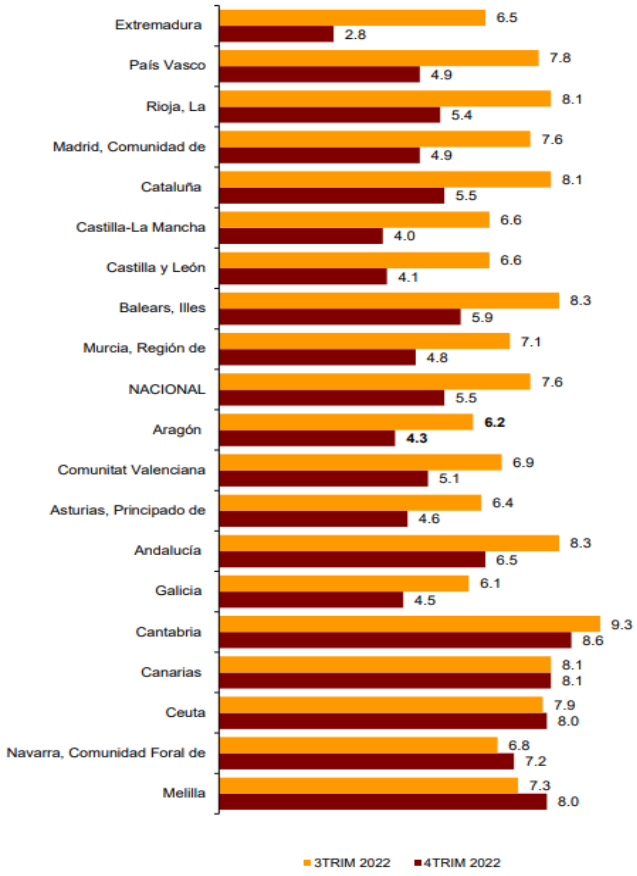
Figura 4: Precio de la vivienda libre en 2019. Porcentaje del máximo anual histórico.



Fuente: Ministerio de Fomento (2022)

Mientras que en 2022 el IPV trimestral fue inferior en el último trimestre en todas las comunidades menos en las de Ceuta, Melilla y Navarra donde el incremento es positivo y en Canarias donde el crecimiento del índice se mantuvo estable; reflejando una mayor compraventa de pisos libres en dichas comunidades y cambiando la tendencia que venía experimentando en años anteriores (Figura 5).

Figura 5: Tasas de variación del IPV último trimestre 2022

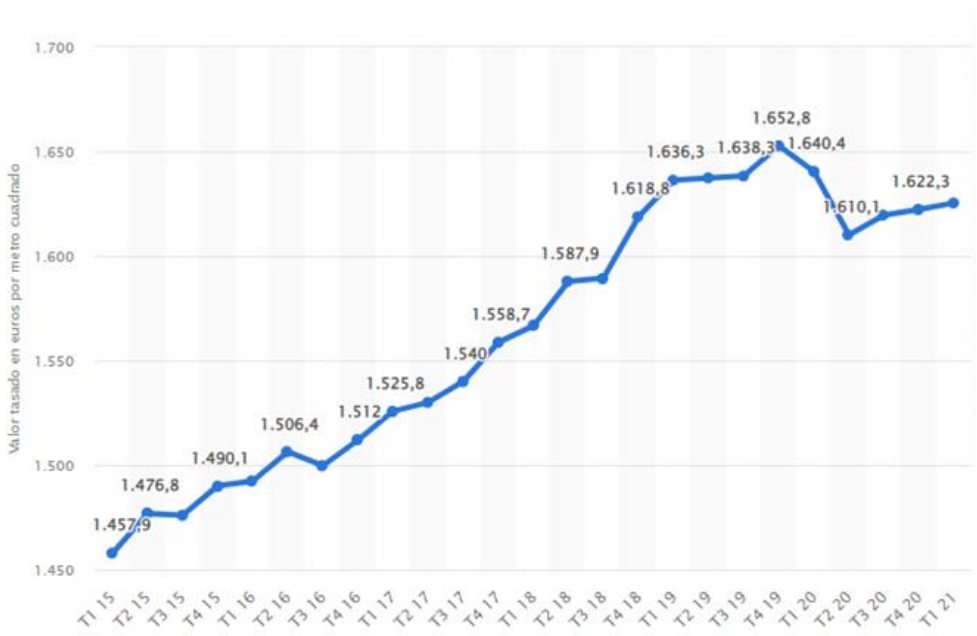


Fuente: Instituto Nacional de Estadística, INE (2023)

Otro índice de referencia a la hora de estimar movimientos en los precios de acuerdo con los m2 de las viviendas es el Valor Tasado de Vivienda. Dicho índice está formado por todas las viviendas que han sido valoradas por empresas de tasación en un determinado trimestre, pudiendo llegar a presentar estas una gran varianza de acuerdo con las tasaciones que se les apliquen. (Ministerio de Fomento, 2022). En la figura 6 se observa como dicho Valor ha experimentado un crecimiento sostenible durante cinco años, desde

2015 hasta finales de 2019, coincidiendo la contracción de este con el comienzo del período experimentado de pandemia.

Figura 6: Valor tasado de la vivienda libre en España desde el primer trimestre de 2015 hasta el primer trimestre de 2021

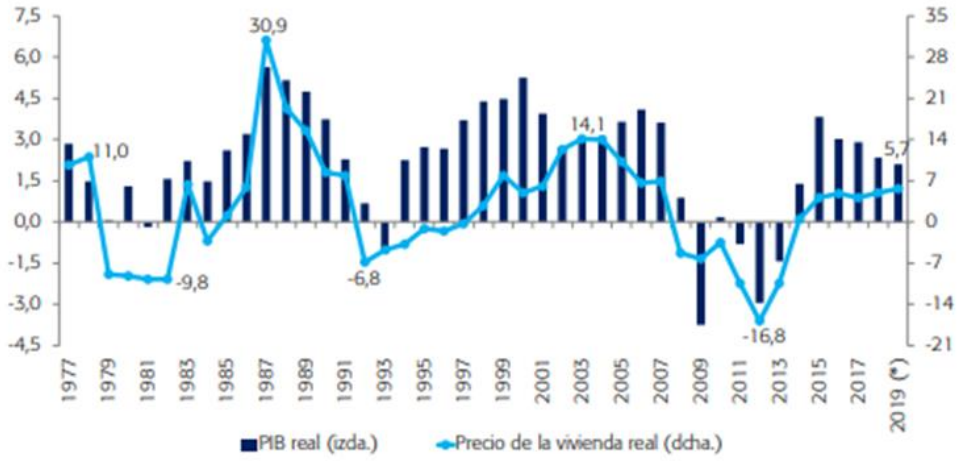


Fuente: Ministerio de Fomento (2022)

Posteriormente, para entender de una manera mejor qué cambios ha experimentado el precio de la vivienda y las situaciones que han producido movimientos en el mismo es lógico comparar el precio de la vivienda con el Producto Interior Bruto Real (PIB) del país en términos reales (de Tudela y Torres, 2019).

Destacan dos periodos en el que el precio de la vivienda, así como el PIB, experimentan una gran contracción. El primero viene determinado por la segunda crisis del petróleo durante los años 70, mientras que el segundo se caracteriza por el resultado de la crisis financiera del año 2008 que tuvo una mayor afección al mercado inmobiliario y cuya recuperación no se empezó a hacer efectiva hasta el año 2014 (Figura 7).

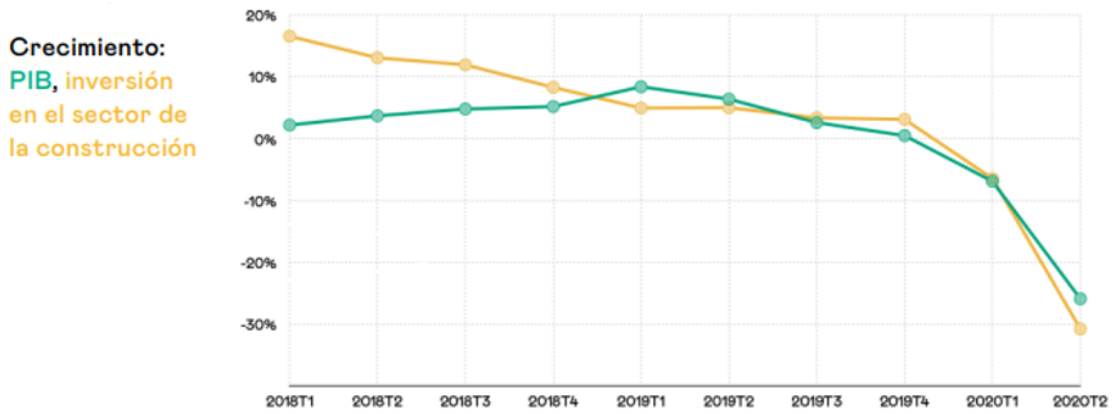
Figura 7: Evolución del precio de la vivienda y el PIB (Tasa de crecimiento interanual, en porcentaje)



Fuente: BIS, INE y Funcas (enlace de series)

A estos dos periodos anteriores, es necesario remarcar es la pandemia de Covid-19. Dicha situación tuvo una gran repercusión tanto a nivel económico como poblacional. Los sectores de la construcción y la intermediación de vivienda se vieron fuertemente azotados, así como el comienzo de la desaceleración del Producto Interior Bruto (PIB) del país dando como resultado una reducción histórica del 30.8% (Raya, 2020) en el capital en viviendas (Figura 8).

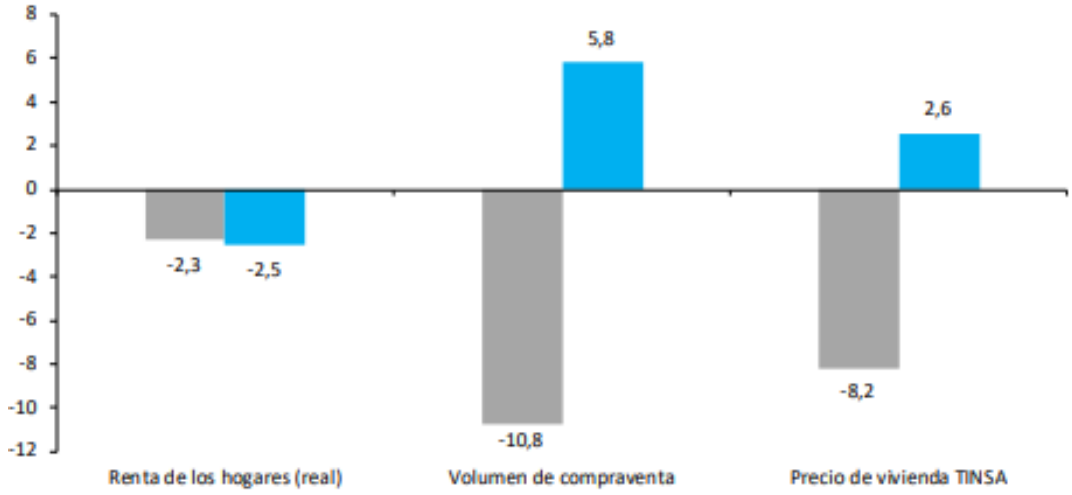
Figura 8: Gráfico PIB e inversión sector construcción



Fuente: Raya, J.M. (2020)

La pandemia ha tenido un efecto muy distinto a la crisis del 2008 comentada anteriormente, debido a que, en 2021, el volumen de transacciones en el sector inmobiliario experimentó un gran rebote (Funcas, 2022). Esta se debió principalmente a dos factores; en primer lugar, las familias se encontraban con un exceso de ahorros tras la pandemia debido al cese de sus actividades habituales, así como una reducción de los costes asociados a los mismas. El segundo factor se caracteriza por la bajada de los tipos de interés por parte del Banco Central Europeo (BCE) lo que permitió acceder a una financiación más barata y favoreció la inversión en este tipo de activos (Figura 9).

Figura 9: Crecimiento medio anual crisis financiera vs pandemia

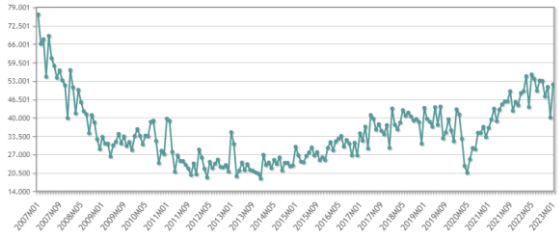


Fuente: Funcas (2022)

En último lugar, a niveles de propiedad, acudimos a la diferencia entre la vivienda protegida y la vivienda libre con el objetivo de estudiar la implicación de los factores económicos en dichos mercados y analizar la tendencia de esta y su correlación con el mercado de la vivienda (Funcas, 2022). Podemos observar en las figuras 10 y 11 como en el período posterior a la crisis financiera de 2008, el número de viviendas protegidas aumentó con considerablemente a diferencia de las viviendas libres. Experimentando posteriormente tendencias similares.

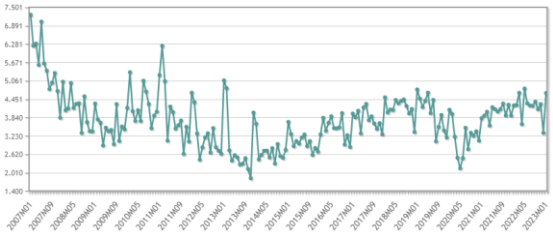
A continuación, estudiaremos los factores que ejercen mayor influencia a la hora de establecer el precio de la vivienda, así como aquellos cuya aplicación puede llegar a afectar en una mayor o menor medida los datos comentados en esta sección.

Figura 10: Estadística de Transición de Derechos de la propiedad, Total Nacional, Vivienda Libre



Fuente: Instituto Nacional de Estadística (2023)

Figura 11: Estadística de Transición de Derechos de la propiedad, Total Nacional, Vivienda protegida



Fuente: Instituto Nacional de Estadística (2023)

3.2.- Factores determinantes del precio de la vivienda

Para entender los factores que determinan el precio en un mercado, recurrimos a los principios de la ley de la oferta y la demanda en los que se establece que, manteniéndose todos lo demás constante (Ceteris Paribus), la demanda de un bien disminuye cuando sube su precio. Por otro lado, manteniendo también lo demás constante, la cantidad ofrecida de un aumenta cuando sube su precio (Mankiw, 2012).

En cuanto al desarrollo y análisis de una función que resuma los principales costes asociados a la oferta, las aplicaciones han sido variadas de acuerdo con la selección de diferentes factores a la hora de realizar posteriores análisis.

Es imprescindible mencionar el término del *stock* de un bien cuando acudimos al estudio de la oferta y demanda de dicho bien dentro de un mercado específico. El mismo, puede ser definido como la cantidad de viviendas que hay en un territorio en un periodo, más las existentes en el periodo anterior menos las destruidas por demoliciones y otras causas (Taltavull, 2001). Del mismo modo, el concepto de depreciación es necesario para entender la reducción del valor de una vivienda y su posterior amortización en el largo tiempo con objetivo de mantener constante el nivel de calidad de las unidades. Con todo esto, el modelo de Taltavull (2001) se presenta como:

$$ht{+}1 - ht = xt - dht$$

Donde:

ht+1 = Stock de viviendas en el momento t+1

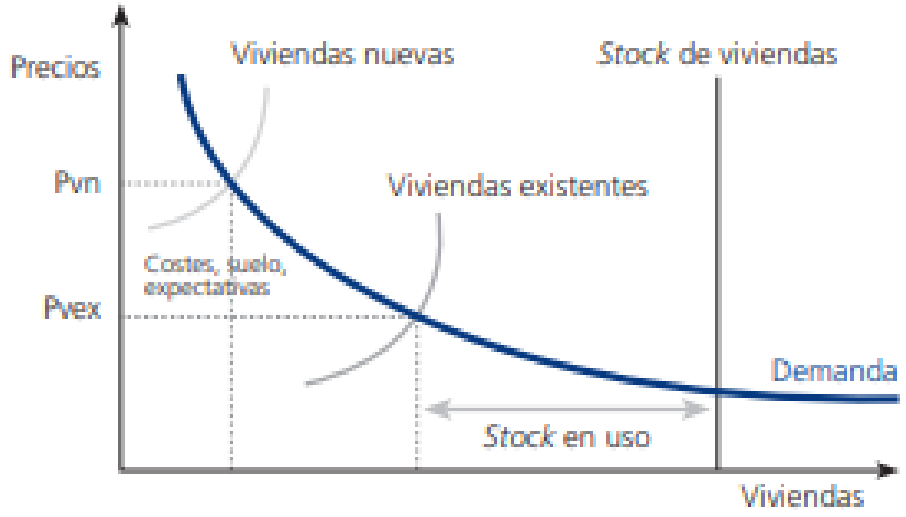
ht = Stock de viviendas en el momento t

xt = Total de viviendas nuevas terminadas en el mercado

dht = Depreciación del stock existente en el momento t

Además, la oferta de casas de nueva construcción y casas de segunda mano juegan un papel importante debido a que los excesos de oferta que puedan ir apareciendo el mercado, reducirán la construcción de nuevos inmuebles, así como el aumento del precio de las casas de segunda mano puede favorecer a un incremento del precio de las viviendas de nueva construcción (Taltavull, 2006). Un mercado con ausencia de presión de demanda, en el que la oferta se focaliza principalmente en la construcción de ofertas nuevas (Figura 12) verá incrementado sus niveles de oferta, mientras que un mercado con una alta presión de demanda verá reducida en gran proporción su nivel de stock hasta el punto en el que el precio se vea forzado a subir para cubrir dicha escasez de viviendas.

Figura 12: Esquema de un mercado inmobiliario y sus componentes de oferta. corto plazo



Fuente: Taltavull de la Paz, (2006)

Sin embargo, fórmula mencionada anteriormente de oferta debería recoger también diferentes componentes que puede hacer incrementar o disminuir en una mayor medida dicho stock de viviendas en un mercado, como son los costes relacionados con el mercado laboral, los tipos de interés o los precios del mercado de dichos bienes. Por ello, la función agregada quedaría de la siguiente forma (Taltavull, 2001):

$$HS = s(R, L, i, Ph)$$

donde:

L = Coste por unidad de trabajo y tiempo, o tipo de salario real en construcción.

p¡, = Precio de mercado al cual son vendidas las viviendas.

R = Precio de alquiler de estas.

i = Tipo de interés del mercado, como medida de coste de oportunidad del capital empleado en la producción de nuevas viviendas.

Por todo ello, los factores principales a la hora de establecer la cantidad ofertada de una vivienda en el mercado estarán directamente relacionados con las variables de precio de nueva vivienda, coste de producir la misma y el stock de viviendas en dicho momento (López, 2009).

En cuanto a la definición de un modelo de demanda que pueda reflejarse de forma efectiva es muy complejo debido al gran número de factores influyentes a la hora de demandar una vivienda. Siendo algunos de los más importantes (Mankiw, 2012):

- Factores financieros: entre los que podemos destacar el precio de la vivienda, el tipo de interés del mercado a la hora de realizar la compra, así como el tipo aplicable a los créditos hipotecarios, impuestos, renta…etc
- Bienes Sustitutivos: el principal bien sustitutivo en este mercado son las viviendas de alquiler.
- Variables sociodemográficas: el incremento de los costes financieros relacionados con la adquisición de una nueva vivienda se traslada en un periodo más tardío de emancipación por parte de las poblaciones más jóvenes. Así como la introducción de capital extranjero para la adquisición de nuevas viviendas puede conllevar a grandes implicaciones en la demanda relacionadas con incrementos o disminuciones en el stock de estas.
- Variables geográficas/psicológicas: relacionadas con la ubicación, accesibilidad y conexión con los medios de transporte urbanos e interurbanos, estatus, ruido, aglomeraciones.

El modelo explicativo presentado por la profesora Taltavull (2001) refleja principalmente la relación entre la inflación y los precios reales de las viviendas, siendo este:

$$C = [(1 - \theta)i - \pi + \mathrm{d}]\mathrm{p}\ h/p22$$

Siendo:

$ph/p22$ = Precio real de una vivienda

Θ = Tipo impositivo

i = Tipo de interés nominal al que se paga un bono

π = Tipo de inflación esperada, es decir, tipo al cual se deprecia el bono

d = Tipo al que se deprecia la vivienda

C = Coste de uso de la vivienda

La demanda de los bienes en un mercado como la vivienda se ve principalmente afectado por las variables financieras, las cuales tienen un gran impacto que se traduce en posibles movimientos en la función de esta. Del mismo modo, la economía del país objeto de estudio tiene una gran influencia en el mercado inmobiliario y en el establecimiento de medidas debido a que el propio estado puede modificar los precios, así como las cantidades de oferta y demanda mediante la construcción de viviendas protegidas o el control de las calificaciones de tierra y diferentes licencias (López, 2009).

A continuación, se estudian las características de los datos representados en formato de tabla; tabular, así como su implementación al Machine learning.

3.3.- Datos tabulares con ML

Los datos en formato tabular son aquellos que se caracterizan por utilizar y estar representados en tablas y relaciones estructuradas cuyo enfoque está, en mayor medida, destinado a los usuarios finales. Dicho formato ha buscado la efectividad y eficiencia en la exploración y visualización de los datos, es por ello por lo que dicho modelo es más efectivo que otros como, por ejemplo, el modelo multidimensional (basado en los conceptos de cubos y dimensiones), dado que reduce el tiempo de computación de las bases de datos, sin embargo, este modelo es poco conveniente si se dispone de poca memoria RAM dado que grandes volúmenes de datos exigen una mayor capacidad de procesamiento (Sanchez etal., 2015).

Un ejemplo de las principales diferencias entre estos dos modelos se puede observar en la Figura 13.

Figura 13: Características de los modelos Multidimensionales y Tabulares

Característica	Multidimensional	Tabular
RAM	Menos (16/32 Gb)	Bastante (64/128Gb)
Velocidad de RAM	Es importante	Es crucial
Número de CPU	4 / 8 / 16	4 / 8 / 16
Velocidad de CPU	Menos importante	Es crucial
Utilización de disco de estado sólido (SSD)	Fuertemente recomendado	No se utiliza
Velocidad de la red	Importante	Importante

Fuente: Sanchez etal., (2015)

El concepto de Machine Learning, aprendizaje de máquinas o automático, hace referencia a la parte de la ciencia de datos que tiene como objetivo realizar predicciones o construir modelos con una intervención humana reducida. Para alcanzar dicho objetivo, es necesario el uso e implementación de grandes volúmenes de datos (Zhou, 2021).

Los dos tipos de datos principalmente utilizados son los datos estructurados (aquellos que se encuentran en formato tabular; como, por ejemplo, las bases de datos) y los no estructurados (imágenes, videos, audios…etc). Entre estos dos grupos, podemos destacar también los datos semiestructurados, caracterizados por no tener un esquema definido y estar organizados mediante etiquetas que permiten agruparlos y crear jerarquías (Sotoquirá, 2021). Un ejemplo de estos últimos son las bases de datos NoSQL o el estándar JSON.

Los usuarios como las empresas se encuentran diariamente produciendo datos de todo tipo; estructurados (p. ej. creando bases de datos), no estructurados (p. ej. sacando videos y fotos) y semiestructurados (p. ej. desarrollando de código de programación), por lo que es muy importante utilizar herramientas que faciliten su almacenamiento y simplificar los procesos.

La estructuración de los datos en tablas y columnas facilita tanto su análisis con la implementaciones y desarrollo de modelos debido a su estructura organizada en tablas y columnas. Por ello, a veces es necesario llevar a cabo procesos de preprocesamiento y limpieza de las bases de datos, con el objetivo de simplificar su posterior análisis e intentar eliminar posibles sesgos o incorrecciones a las que los modelos se pueden

adecuar. Del mismo modo, el Machine Learning termina, en cierta parte, con el posible sesgo que puede presentarse debido a la introducción de la parte humana que lleve a cabo el análisis de ciertos volúmenes de datos dado que ese tipo de modelos "aprende" tras el procesamiento de una cantidad de datos suficiente (Aigneren, 2004).

Finalmente, una vez analizado el sector inmobiliario, así como las diferentes aplicaciones e influencias del Machine Learning en los datos, pasamos a establecer el marco de la tesis, núcleo de esta memoria, así como el análisis en profundidad de los posteriores modelos.

Sección 4: Marco de la tesis

4.1.- Asunciones, objetivos, hipótesis y restricciones

Aunque algunos de los modelos no han tenido gran implementación en periodos anteriores al objeto de estudio de este proyecto, como es el caso de la red neuronal *TabPFN*, dicha implementación puede obtener mejores resultados que los obtenidos con otros modelos de predicción como los *Random forest* o el modelo de aprendizaje supervisado *Xgboost*.

Del mismo modo, dichos resultados obtenidos en una población de 1.000 inmuebles pueden ser aplicables a poblaciones más extensas lo que aumentaría el nivel de exactitud de los modelos mediante un mayor entrenamiento de los modelos.

Los principales objetivos de este trabajo son:

- La implementación y posterior estudio de la aplicación de modelos de predicción a una base de datos de pisos actualizada de la que poder extraer conclusiones y verificaciones acerca del grado de exactitud con el que dichos modelos predicen el precio de los pisos, de acuerdo con sus distintas características
- La creación de una base de datos fiable para su posterior análisis, así como el estudio de la *accurary* del modelo de red neural *TabPFN* y la semejanza o diferencia con la opinión de la comunidad científica que ha estado analizando el mismo desde su implementación.

Por ello, la principal hipótesis de este proyecto es la aplicación de modelos de clasificación de manera conjunta; *ensembles*, obtienen un resultado fiable sobre la predicción del precio de una vivienda de acuerdo con sus características.

En cuanto a las restricciones que se presentan en la elaboración del trabajo podemos destacar las siguientes:

- El tiempo que emplear para la extracción de los datos. La descarga de datos a través de la API de idealista está limitadas a 100 pisos por petición mensual, lo que incrementa el tiempo de espera entre peticiones e imposibilita el llegar a tener una amplia cantidad y variedad de pisos para utilizar como población de estudio. Es cierto que dicha restricción puede ser solventada mediante el pago por mayor

número de datos, sin embargo, dicha opción podría incrementar mucho los costes del estudio si las poblaciones necesarias fueran muy grandes.

- La reciente implementación del modelo de red neuronal *TabPFN*, el cual hace que las asunciones y conclusiones que se puedan extraer del estudio no sean totalmente fiables y pueden llevar al investigador a caer en sesgos o asunciones erróneas sobre la exactitud de los resultados obtenidos por los modelos.
- El escaso acceso a bases de datos amplias y estructuradas dado que las mejores bases de datos; bases de datos amplias y actualizadas a tiempo real, son propiedad de empresas privadas (portales inmobiliarios y empresas inmobiliarias) y La Ley Orgánica de Protección de Datos y Garantías de los Derechos Digitales (LOPDGDD o LOPD, España) les imposibilita el suministro de información sin la anterior aprobación por parte del propietario del inmueble a la que se encuentran suscritas las inmobiliarias. Esto acaba resultando en el uso de bases de datos poco actualizadas o incluso modificadas, en los posteriores análisis e implementación de modelos predictivos.

Por todo ello, es necesario realizar una previa planificación y estructuración con el objetivo de alcanzar los plazos establecidos para el desarrollo de la memoria, así como solventar posibles imprevistos que vayan apareciendo al desarrollo de la misma.

4.2.- Planificación del trabajo

El trabajo se estructura en cinco partes principales. Para empezar a abordar el mismo se desarrolló un esquema orientativo con las fechas para ir alcanzando los objetivos y pasos para la realización del proyecto, como se puede observar en la figura 14.

Posteriormente, y una vez seleccionado el tema, se llevó a cabo una reunión para poder establecer los objetivos y desarrollar en mayor profundidad el esquema con el índice preliminar y los temas a tratar, correspondientes al Capítulo 1 del proyecto.

Una vez establecidos estos objetivos preliminares, los tiempos de desarrollo se distribuyeron de la siguiente forma:

En los primeros meses se intentó contactar vía email a un gran número de inmobiliarias y portales web inmobiliarios con el objetivo de conseguir una base de datos actualizada y fiable que estuviera compuesta por diferentes pisos (la muestra poblacional objetivo era

de 1.000 pisos madrileños) con sus características específicas para poder estudiar posteriormente. Dichas empresas fueron:

- **Inmobiliarias**: vivalta, tecnocasa, remax, mipisoenmadrid, globaliso, gilmar, unicainmobiliaria, optimacasa, martinagenciainmobiliaria, ambassador.
- **Portales web**: idealista, redpisos, fotocasa, pisos.com, habitaclia.com

La lógica empleada a la hora de contactar con las diferentes empresas fue que aquellas empresas de menor tamaño (las inmobiliarias) iban a ser más propensas a responder a nuestro correo debido a que el tamaño reducido permite que haya menos trabas y burocracia a la hora de suministrar ciertos datos y el correo que suelen tener en sus áreas de contacto de la página web suele ser el de un agente inmobiliario, lo que facilita la comunicación directa con un profesional por ello que el número de inmobiliarios contactadas supera el doble que el número de portales web.

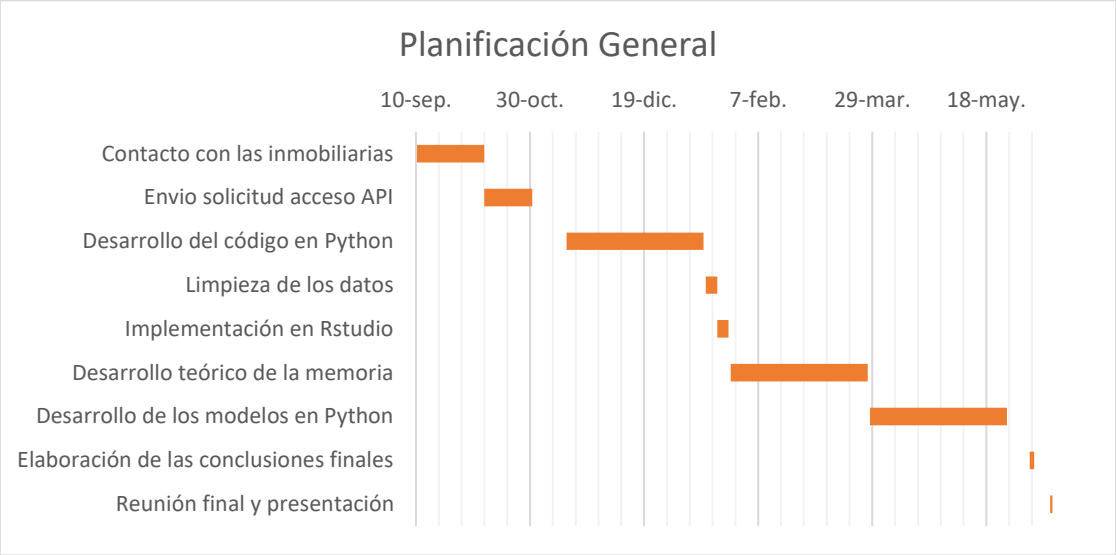
Tras la espera aproxima da unos dos meses, solamente se recibió contestación por parte de dos trabajadores de inmobiliarias. Uno de ellos señaló que era imposible el envío de los datos solicitados debido a que la ley de protección de datos se lo prohibía y que para poder compartir con otras personas los datos de un piso en concreto debía ser el dueño el que diera aprobación explícita y personal, lo que complicaría mucho el trabajo y los tiempos estableces. El otro trabajador señaló puntos parecidos, sin embargo, mencionó que el portal de pisos Idealista contaba con una herramienta que permitía descargar de forma gratuita 100 pisos al mes, complementados con todo tipo de información, para poder destinarlos al estudio y análisis. Del mismo modo, nos facilitó el acceso directo a dicha herramienta.

Para seguir adelante con nuestro estudio se cumplimentaron los datos necesarios para poder acceder a la misma y tras tres semanas de espera, Idealista facilitó facilitaron las claves de acceso necesarias.

Desde el quinto al séptimo mes y una vez obtenido el acceso a la herramienta, se llevó a cabo el desarrollo del código en Python, necesario para poder extraer datos de la herramienta, se realizó la limpieza y estudio de los datos mediante Excel y RStudio y se desarrollaron e implementaron los códigos de Python propios de los distintos modelos de Machine Learning que se aplicaron a nuestro estudio; los cuales se comentarán en un mayor detalle en la sección 5.

Finalmente, una vez realizado y obtenido la gran parte de análisis del proyecto, del séptimo mes hasta la finalización y entrega del proyecto, se desarrolló el estudio cualitativo de la situación del Real State en España, propio de la sección 3, así como el desarrollo y traslado de las principales conclusiones del análisis de los resultados, correspondientes a la sección 7, mientras se iban complementando la bibliografía y el anexo a lo largo de todo el periodo de desarrollo, correspondientes a los capítulos 8 y 9.

Figura 14: Gráfico resumen del proyecto



Fuente: Elaboración propia

En las secciones posteriores, se realiza una revisión tanto teórica como práctica de los contenidos, así como la implementación de los códigos y modelos desarrollados.

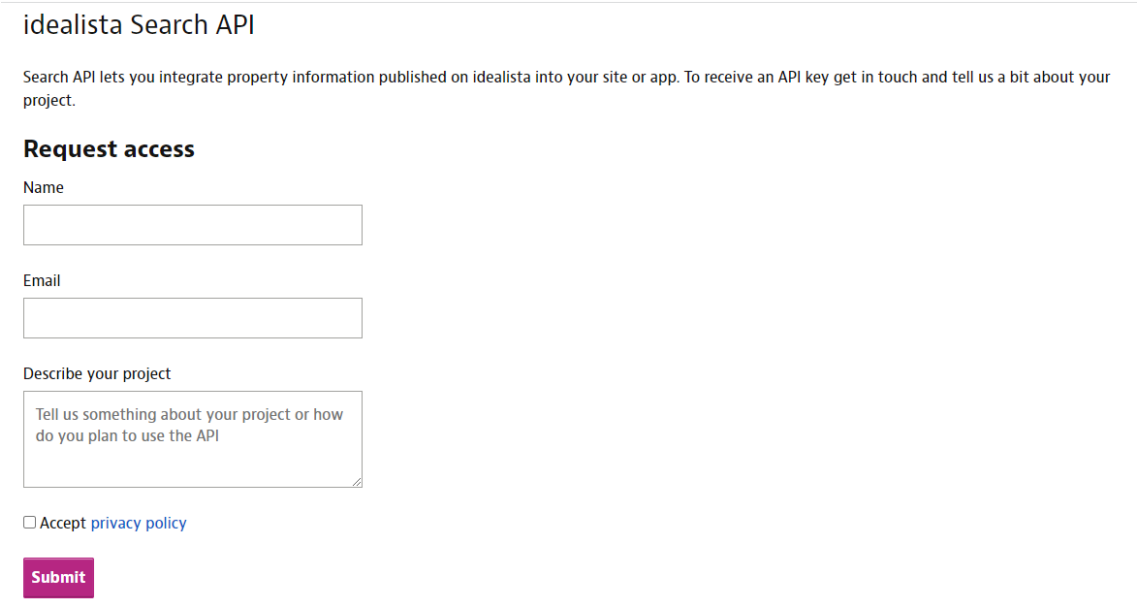
Sección 5: Metodología

5.1.-Análisis cualitativo

5.1.1.- Extracción de datos

La recopilación de datos de distintos pisos se llevó a cabo a través de la API de Idealista mediante la implementación de código con Python, pero para poder llevar a cabo dicho proceso era necesario cumplimentar con anterioridad el formulario que se expone en la Figura 15 en el que se pedía información del solicitante, así como la descripción del proyecto que se iba a desarrollar, con el objetivo de obtener la posterior aprobación por parte de esta empresa.

Figura 15: Captura de pantalla, cuestionario solicitud acceso API idealista



Fuente. Idealista Search API. https://developers.idealista.com/access-request.

Una vez rellenados los datos, si se obtenía el visto bueno por parte de Idealista, se facilitaba al usuario la siguiente información:

- **Documento de autentificación**: se caracteriza por describir las parámetros, respuestas y errores que podían aparecer en el proceso de autentificación por parte

del usuario para acceder a la API. Del mismo modo, destacan tres conocimientos necesarios para poder realizar de forma correcta dicha autentificación:

  - La URL codifica la clave de API y el secreto según RFC 1738
  - Concatenar la clave API codificada, dos puntos ":" y el secreto en una sola cadena
  - Codificar la cadena del pase anterior en base Base64

- **Documento de búsqueda de propiedades**: se caracteriza por contener la URL específica en la que poder realizar las llamadas, la descripción de todos los filtros específicos que se pueden aplicar a las distintas llamadas (tipo de propiedad, distancia al centro, precio máximo…etc) así como un ejemplo de una llamada y una respuesta a la API. Del mismo modo, no solamente se especifica solo el nombre de las variables si no el tipo de dato que son y una descripción e información adicional como se observa en la Figura 16.

Figura 16: Tabla descriptiva de los diferentes filtros aplicables en la API

Filters

Allowed parameters:

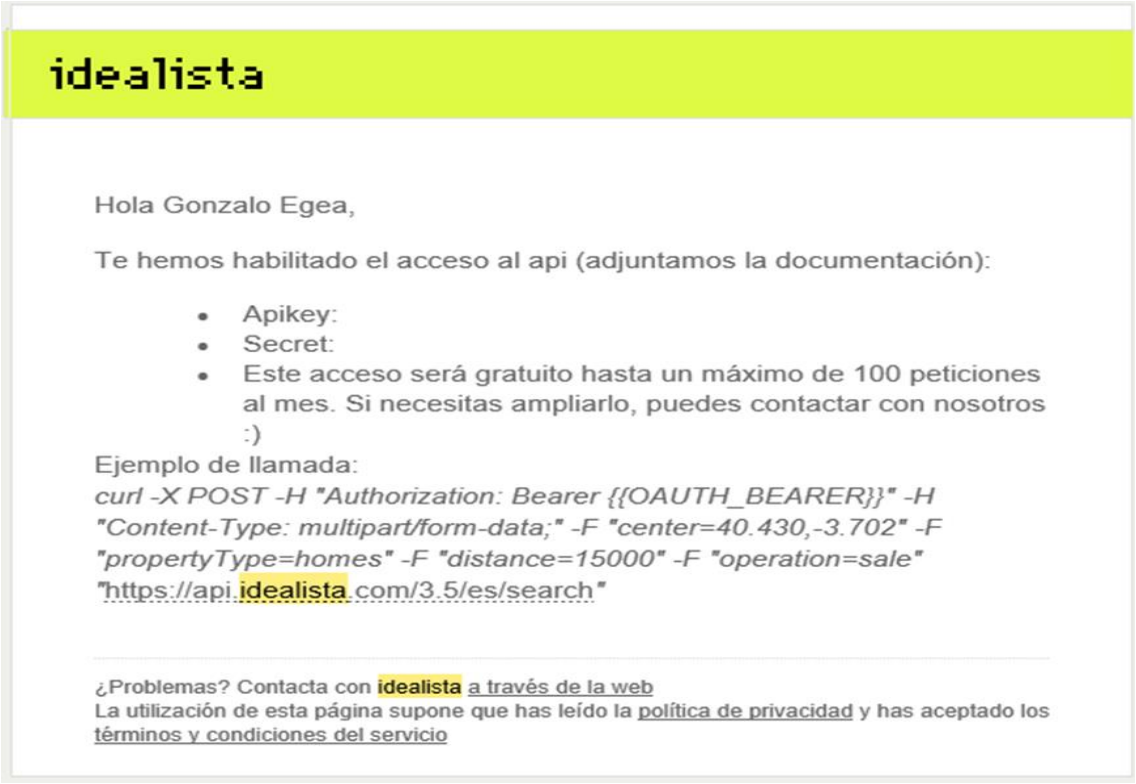
name	data type	description	additional info
country (required)	string	• es - idealista.com • it - idealista.it • pt - idealista.pt	`values: es, it, pt`
operation (required)	string		`values: sale, rent`
propertyType (required)	string		`values: homes, offices, premises, garages, bedrooms`
center*	string	geographic coordinates (WGS84) (latitude, longitude)	example: "40.123,-3.242"
locale	string	search language for summary	`values: es, it, pt, en, ca`
distance*	double	distance to center, in metres (ratio)	
locationId*	string	idealista location code	
maxItems	integer	items per page	50 as maximun allowed

Fuente: Documentación extraída de la API de Idealista

Finalmente, el correo termina con la información propia del usuario (Apikey y Secret) necesarios para poder acceder de manera individual a la herramienta, como se observa en la Figura 17, así como un ejemplo de llamada y un apartado en el que se especifica que

el número máximo de peticiones gratuitas al mes es de 100 pero existe la opción de aumentar dicho número de peticiones; con su coste correspondiente.

Figura 17: Datos necesarios de acceso API


idealista

Hola Gonzalo Egea,

Te hemos habilitado el acceso al api (adjuntamos la documentación):

- Apikey:
- Secret:
- Este acceso será gratuito hasta un máximo de 100 peticiones al mes. Si necesitas ampliarlo, puedes contactar con nosotros :)

Ejemplo de llamada:

curl -X POST -H "Authorization: Bearer {{OAUTH_BEARER}}" -H "Content-Type: multipart/form-data;" -F "center=40.430,-3.702" -F "propertyType=homes" -F "distance=15000" -F "operation=sale" "https://api.idealista.com/3.5/es/search"

¿Problemas? Contacta con idealista a través de la web
La utilización de esta página supone que has leído la política de privacidad y has aceptado los términos y condiciones del servicio

Fuente: Documentación extraída de la API de Idealista

Tras realizar las consultas y obtener toda la información necesaria, comentada anteriormente, el siguiente paso consistía en la elaboración del código de Python necesario para poder acceder a la API y realizar las distintas llamadas de forma correcta con el objetivo de trasladar todos esos datos posteriormente a un Excel, unificarlos y realizar un posterior estudio y manipulación de estos.

Una vez realizado e implementado el código de Python se llevaron a cabo 30 peticiones (25 primeras peticiones y 5 peticiones finales con el objetivo de complementar los datos adicionales) a la API caracterizadas por estar compuestas de 50 pisos cada una.

En dichas llamadas se utilizaron los siguientes filtros con el objetivo de obtener una muestra lo más homogénea y representativa del posterior estudio a realizar:

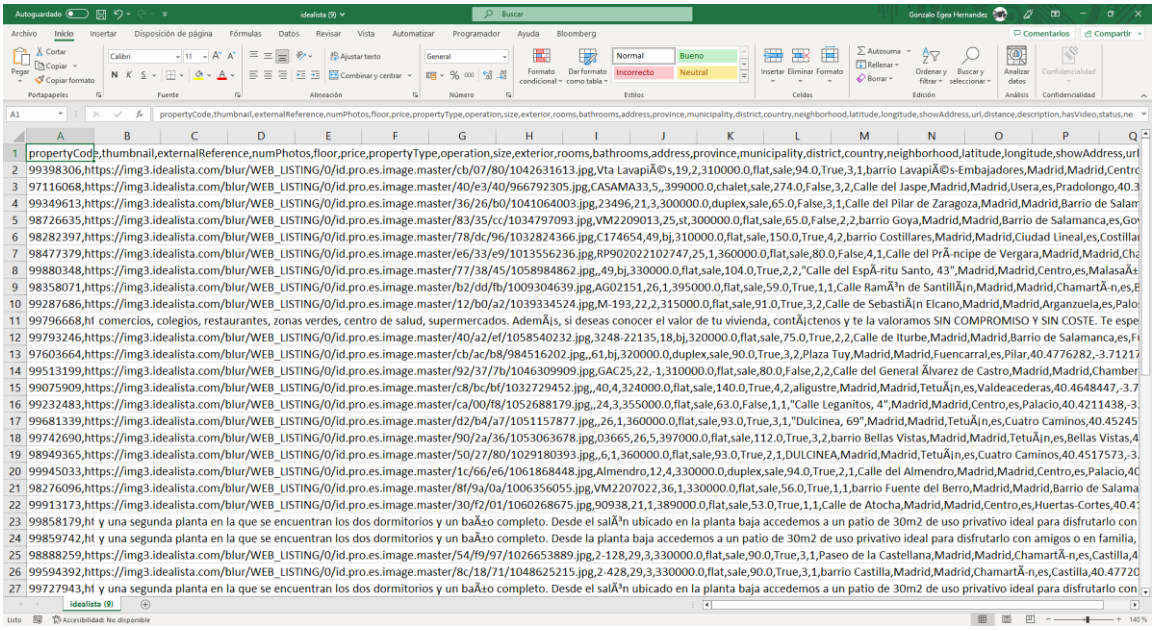
- **Country:** filtro en el que se especificaba el país del que se quiere realizar la extracción de pisos (dicho filtro solo puede tomar los valores es (españa), it (itailia) y pt (portugal))
- **Language**: filtro característico del idioma en el que aparece la información extraída, puede tomar diferentes valores dependiendo si se busca información en español, italiano, portugués, inglés o catalán (es, it, pt, en, ca).
- **Max_items**: filtro en el que especifica el máximo número de ítems que se extraen por llamada (50 en nuestro caso)
- **Operation**: filtro en el que especificar si los pisos objeto de estudio están en venta o en alquiler (rent, sale)
- **Property_type**: filtro específico del tipo de propiedad que se quiere extraer, pudiendo tomar los valores: homes, offices, premises, garages, bedrooms. (homes en este caso)
- **Order**: filtro en el que se especificaba el orden en el que se quieren obtener los resultados, tomando el valor "priceDown" en este caso.
- **Center**: filtro en el que se indican las coordenadas del lugar a tomar como referencia de centro en el estudio a realizar (en nuestro estudio dichos valores fueron; 40.4167, -3.70325, característicos de la Puerta del Sol, Madrid).
- **Distance**: filtro en el que se especifica la distancia al centro (este iba siendo modificado de mil en mil con el objetivo de abarcar cada vez una mayor zona)
- **bankOffer:** filtro que tomaba los valores True o False, dependiendo si el piso era propiedad de un banco o no.
- **max/min Price:** filtro en el que se estable el rango de precios de los inmuebles objeto de estudio y cuya variación aplicada fue de 50.000€ por cada llamada.

Finalmente, tras los descargar todos los distintos archivos resultantes de las diferentes consultas realizadas en formato csv, se unificaron las mismas en un Excel conjunto con el objetivo de limpiar y preparar todos los datos para su posterior análisis exploratorio e implementar los modelos objeto de estudio.

5.1.2.- Limpieza y preparación de los datos

Cada vez que se hacía una llamada de resultados a la API de idealista se obtenía un archivo con una estructura como el que aparece reflejado en la Figura 18, al que eran necesarios aplicarle una serie de procedimientos y modificaciones con el objetivo de ordenar y estructurar los datos.

Figura 18: Ejemplificación archivo resultados extraído

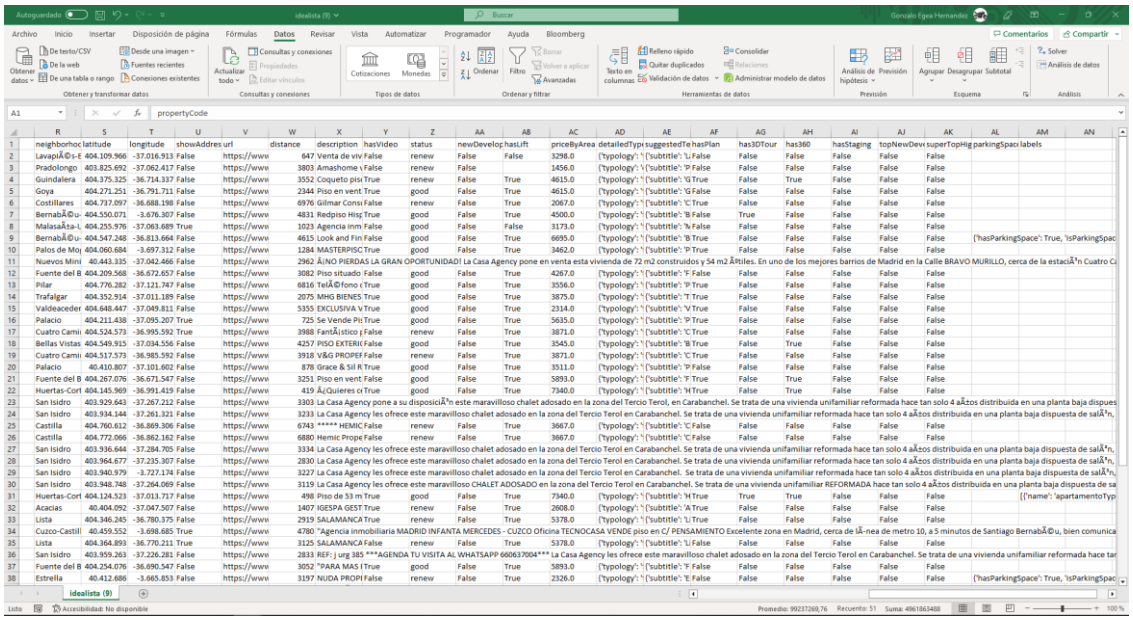


Fuente: Información extraída de la API de Idealista

En primer lugar, es necesario estructurar los datos mediante la herramienta de datos, texto en columnas. Para llevar a cabo este proceso, se seleccionan todas las columnas del archivo, se especifica el tipo de archivo que es; delimitado, y la separación que aplicar al mismo; separados por comas, en este caso.

Tras realizar dicho paso algunos pisos experimentaban lo reflejado en la Figura 19, es decir, la fila de descripción se acoplaba al resto de filas que había a su derecha por lo que todos esos datos se eliminaban. Por ello, fue necesario eliminar dichos pisos incompletos y complementarlos con nuevas peticiones de datos para poder cubrir los huecos (de aquí que el número de llamadas pasara de 25 a 30 como comentamos anteriormente en la sección de extracción de datos).

Figura 19: Ejemplificación error datos extraídos



Fuente: Información extraída de la API de Idealista

Una vez eliminados dichos pisos, se fueron acoplando los datos de los pisos de llamada en llamada a un Excel general utilizado como agregador para obtener una base de datos final. A la hora de realizar este paso, es necesario eliminar los datos duplicados dado que algunos pisos con características similares podían aparecer duplicados en dos llamadas distintas. Un ejemplo de este problema sería realizar una petición de pisos que estén a dos mil metros del centro y otra igual, pero con pisos que se encuentren a una distancia de tres mil. Ante esto, se daba la situación en la que había pisos de ese primer grupo (2.000m del centro) que se encontraban dentro del segundo grupo (3.000m del centro) debido a que la distancia se agregaba, sin embargo, el número total de pisos que se encontraban duplicados era muy pequeño debido al alto número de pisos que tiene Idealista registrado, pero era necesario eliminarlos con el objetivo de eliminar duplicidades.

En tercer lugar, era necesario aplicar una serie de modificaciones a ciertas columnas de los datos con el objetivo de estructurarlos de una mejor forma.

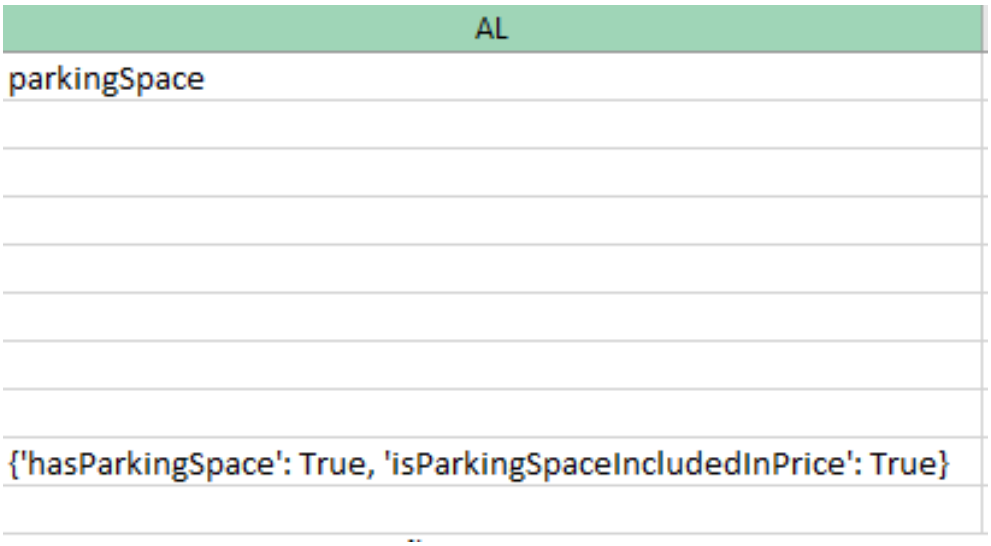
Los datos de las columnas *prize*, *prizebyarea* y *size* se volcaron con un formato distinto al numérico (ejemplo extraído del precio de un piso con dicho formato es el siguiente: “300000.0”) y era muy difícil su transformación al mismo debido a que se encontraban en formato texto y el punto que tenían no los identificaba correctamente. Por ello, se

utilizó la función de Excel "=extrae", a la que se le indica el texto al que aplicarla, el número de caracteres a extraer y la posición inicial en la que extraer. Debido a que los números de los distintos pisos tenían una longitud similar, se podían transformar los datos en una nueva columna de forma rápida y efectiva.

Posteriormente, se asignó el valor 0 a aquellos pisos que tenían asignados un valor negativo en la columna *floor* por tratarse de un bajo, con el objetivo de tener todos los datos en valores positivo, del mismo modo se asignó 0 a aquellos pisos que no tenían valor en dicha columna.

Las columnas *parkingSpace* y *Labels* se caracterizaban por estar en blanco aquellas en las que no había datos acerca del piso y en aquellas en las que sí había, aparecía un texto señalando si el piso tenía dicha característica o no, como se observar en la Figura 20.

Figura 20: Ejemplificación columna parkingSpace



Fuente: Información extraída de la API de Idealista

Para solventar dicha situación, se optó por aplicar la fórmula de Excel "=Si" en la que se le especificaba que, si en la celda había texto, se clasificara bajo el texto "True" mientras que, si la misma carecía de este, se clasificara bajo el texto "False".

En cuarto lugar, se realizó un análisis de la variable *Price* con el objetivo de estudiar la dispersión de dicha variable y poder identificar si había grandes diferencias entre los valores de los primeros y los últimos pisos. Para ello se multiplicó por 1.5 el rango

intercuartílico, es decir, aquellos pisos que pertenecían al cuartil 3 menos los que pertenecían al cuartil 1. Del mismo modo se realizó el mismo procedimiento, pero multiplicando el rango intercuartílico por -1.5. Con ello descubríamos qué pisos podían clasificarse como atípicos para, posteriormente, ser eliminados.

Tras dicho análisis, se descubrió que había un número de pisos que pertenecían a la zona de El Viso y cuyos precios disparaban el precio medio total, situándola en torno los 900.000€. Mientras que por debajo del cuartil 1 no había gran número de atípicos. Una vez eliminados estos, el precio medio total se estableció cerca de los 430.000€, haciendo este un valor más representativo.

En quinto lugar, analizando cómo podía mejorarse el estudio posterior de las variables de localización *adress, neighborhood* y *district*, se decidió construir una variable categórica, debido a que aportaría una mayor información en el análisis exploratorio de los datos. Dicha variable consistía en la división de los pisos en cuatro grandes grupos (top 75, top 50, top 25 y resto) dependiendo de la renta media del barrio al que pertenecían de acuerdo con el mapa interactivo (Andrino, B., Llaneras, K., & Grasso, D. (2021)). Del mismo modo, se eliminó la variable categórica *Country y Language* dado que no aportaban información al estudio posterior.

En sexto lugar, y con el objetivo de clasificar los distintos pisos de acuerdo con su precio, se realizaron los cálculos del percentil 33 y del percentil 66. Dichos cálculos buscaban la clasificación de los inmuebles de la siguiente manera:

- Baratos: aquellos pisos cuyo precio se encontrará por debajo del predio correspondiente al percentil 33.
- Medio: aquellos pisos cuyo precio se encontrará entre los valores del percentil 33 y el percentil 66.
- Caro: aquellos pisos cuyo precio se encontrará por encima del predio correspondiente al percentil 66.

Dicha clasificación se guardó en la columna de *Resultado*, con el objetivo de ser utilizada e implementada en los posteriores modelos de clasificación.

Finalmente, una vez implementados todos estos cambios, se decidió acudir a todas las distintas variables que tenían datos categóricos en sus filas para convertirlas en numéricas

y poder simplificar así el análisis a la vez que se permitía el uso de estas para entrenar los distintos modelos. Todas estas modificaciones se pueden observar en la parte del Anexo bajo el nombre; modificación de las variables, en las que se exponen aquellas variables modificadas y los valores finales asignados.

Con todo ello, terminamos limpiando el dataset y nos quedamos con el mayor número de variables numéricas representativas y la variable categórica de Clasificación (top 75, top 50…etc) para realizar el análisis exploratorio de los datos.

5.1.3.- Análisis exploratorio de los datos

Una vez obtenida la base de datos limpia y estructura, se inicia el análisis exploratorio de los datos con el objetivo de estudiar y aplicar diferentes métodos de visualización y poder así obtener ciertas conclusiones a cerca de la distribución de las distintas variables. Para realizar dicho procedimiento, se empleó el entorno de desarrollo RStudio, mediante la programación en R, con el objetivo de llevar a cabo un estudio analítico y representativo en gráficos.

Cabe destacar que la base de datos utilizada para el estudio está compuesta principalmente por dos tipos de variables:

- **Variables categóricas**: son aquellas variables que pueden tomar valores característicos de las propiedades cualitativas de los sujetos objetos de estudio. Se caracterizan también debido que, para su estudio, se utilizan instrumentos de observación y no de medición (Flores, 2007). En nuestra base de datos, la variable categórica es:
    - propertyType
    - exterior
    - neighborhooh
    - Clasificacion
    - showAddress
    - Url
    - hasVideo
    - status
    - newDevelopment

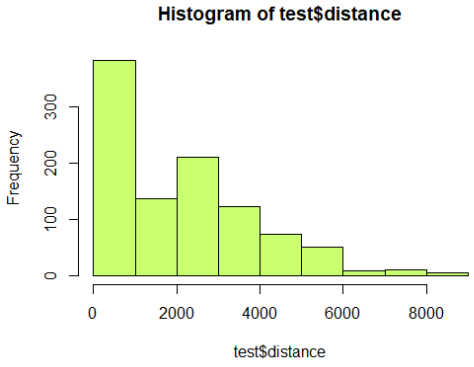
- haslift
- hasPlan
- has3DTour
- has360
- hasStaging
- parkingSpace
- Labels
- BankOffer

La gran mayoría de estas toman los valores “True” o “False” de acuerdo con si el piso tiene o no dicha característica, menos las variables de *clasificacion* que está dividida en 4 grupos como comentamos con anterioridad, la variable *propertyType* que tomas los valores: flat, penthouse, dúplex, studio o chalet, de acuerdo al tipo de vivienda que es la propiedad. La variable *neighborhood* que especifica el barrio en el que se encuentra la vivienda. *Url* que representa el identificador de la dirección web del inmueble. La variable *Status* que toma los valores: Good, newdevelopment o renew en base al estado en el que se encuentra la vivienda y la variable *laberls* que presenta ciertos valores de acuerdo a las etiquetas que aparecen en el anuncio del piso (Villa, Lujo, Casa Baja… o Nan para aquellos en los que no hay datos).

- **Variables Continuas**: son aquellas variables que tomar valores numéricos y pueden variar en cualquier cantidad (Flores, 2007). En nuestra base de datos, las variables numéricas son:
    - numPhotos
    - Price
    - Size
    - Rooms
    - Bathrooms
    - Latitude
    - Longitude
    - Distance
    - priceByArea
    - floor

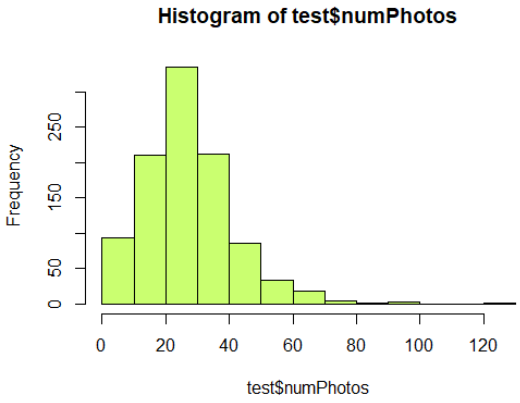
En primer lugar, y una vez cargada la base de datos a RStudio, comenzamos representado las variables continuas *Distance* y *numphotos* en un histograma. Un histograma es una representación gráfica compuesta por intervalos de igual tamaño donde se representan la frecuencia relativa de una variable en el eje de las ordenadas y los distintos valores que toma la misma en el eje de las coordenadas (Behar, 2018). Al realizar dicha representación, observamos como la gran mayoría de pisos de nuestra base de datos se distribuyen en torno a los 1.000 y los 2.000 metros de distancia al centro (Figura 21), mientras que la gran mayoría de estos se distribuyen en torno a 30 fotos en sus publicaciones (Figura 22).

Figura 21: Histograma variable distancia



Fuente: Elaboración propia

Figura 22: Histograma variable número de fotos

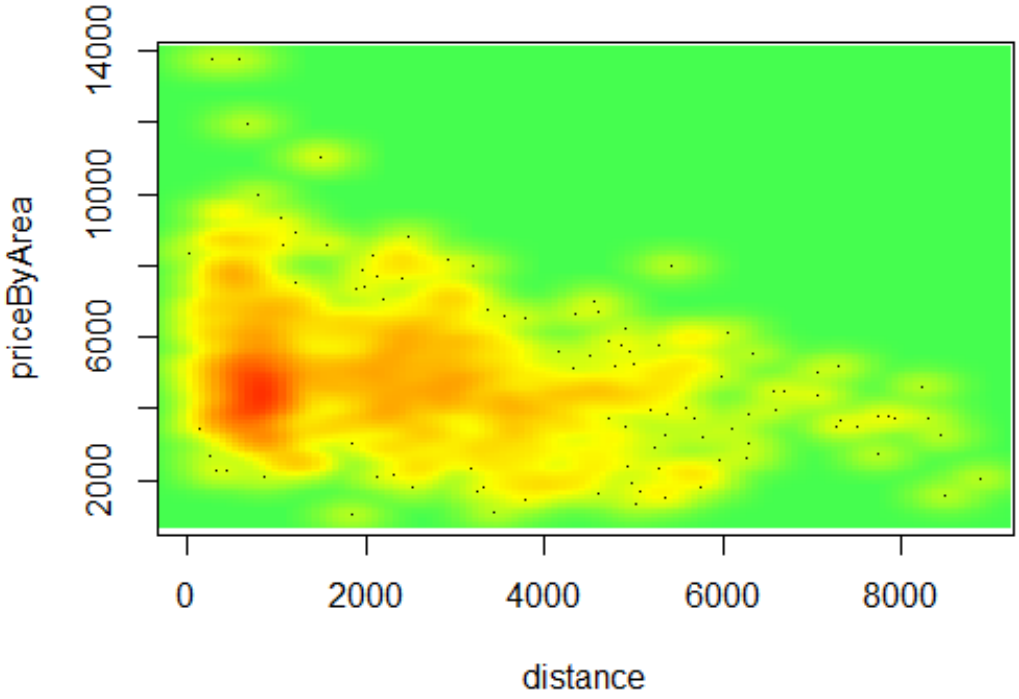


Fuente: Elaboración propia

Posteriormente, pasamos a estudiar la dispersión que presentaban algunas variables. Entendiendo por dispersión aquella medida de la desviación de los datos respecto a una medida de tendencia central (Estepa y Pino, 2013). Con este análisis intentamos entender el grado de desviación que presentaban dos variables concretas y si hay patrones de aglomeración significativos entre las mismas. Para ello, se emplearon las variables *priceByArea* y *distance,* con el objetivo de estudiar si el precio del suelo los pisos que se encontraban más cercanos al centro de Madrid presentaban un mayor precio a diferencia de los que se encontraban a distancias más largas (Figura 23). Sin embargo, tras estudiar los resultados, observamos que la gran mayoría de pisos que componen nuestra base de

datos y que se encuentran cercanos al centro, se concentran en un valor “bajo” de coste del metro cuadrado, comparado con el resto.

Figura 23: Gráfico de dispersión entre variables priceByArea y distance



Fuente: Elaboración propia

En tercer lugar, realizamos un estudio a cerca de la covarianza y la correlación de variables con el objetivo de estudiar el grado de dependencia de esta y los posibles aumentos o disminuciones que se producen en una variable como resultados del incremento o reducción de la otra. Por ello, la covarianza permite analizar el signo de independencia entre dos variables estadísticas mientras que la correlación permite mediar la intensidad de dicha dependencia (Gea y Roa ,2014).

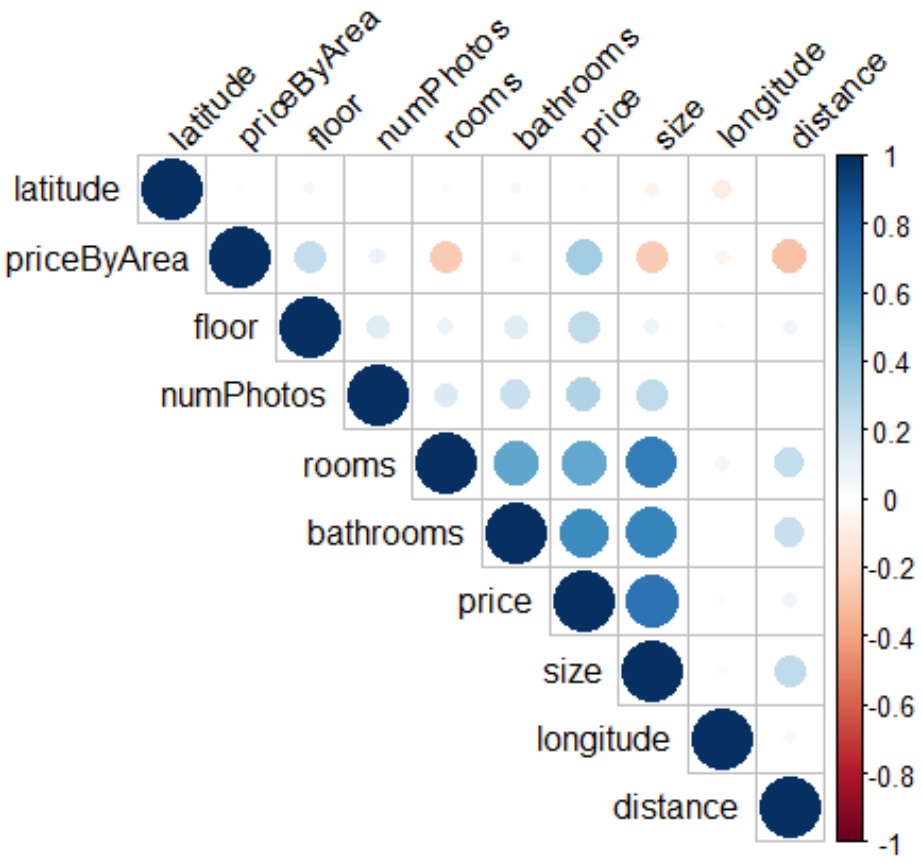
Para llevar a cabo dicho análisis, se implementaron diferentes métodos de cálculo y visualizaciones.; el primero consistió en la selección de la variable *Price* y el estudio de su covarianza y correlación con respecta a las variables: *numPhotos, size, rooms, bathrooms, distance, priceByArea* y *floor*. Mediante la implementación de un bucle, dichos cálculos se automatizaron y los resultados obtenidos en todas las variables fueron

de una covarianza y correlación positiva, por lo que un aumento de las dichas variables mencionadas se puede traducir en un aumento del precio del inmueble. Destacan sobre todo las correlaciones de las variables size (0.73), rooms (0.52) y bathrooms (0.63). así como la correlación entre el precio y la variable distancia por ser muy débil (0.02), lo que vuelva a confirmar el análisis realiza anteriormente.

Posteriormente, se estudiaron los valores de la correlación mediante el estudio de los p-valores obtenidos en el cálculo de cada una de las variables con el resto y la implementación de gráfico representativo sobre dichos cálculos (Figura 24 y Figura 25). En el mismo, podemos observar cómo hay una correlación positiva y alta entre las variables *size* y *rooms*, *bathrooms* y *Price*; siendo las dos primeras algo intuitivo debido a que un mayor número de habitaciones y baños está directamente correlacionado con un aumento del tamaño del piso, sin embargo, podemos señalar que un aumento del tamaño está positivamente relacionado con un aumento del precio del piso; verificando así los resultados obtenidos anteriormente mediante el bucle. Una situación similar ocurre la variable *Price* y las variables *rooms* y *bathrooms*, las cuales presentan una correlación positiva media/fuerte, ejemplificando que un aumento del número de baños tiene una relación positiva con respecto al precio del inmueble. Es necesario remarcar que las estrellas que aparecen encima de los números de los coeficientes de correlación en la Figura 25 corresponden a los p-valores, siendo las tres estrellas el mayor nivel de significación.

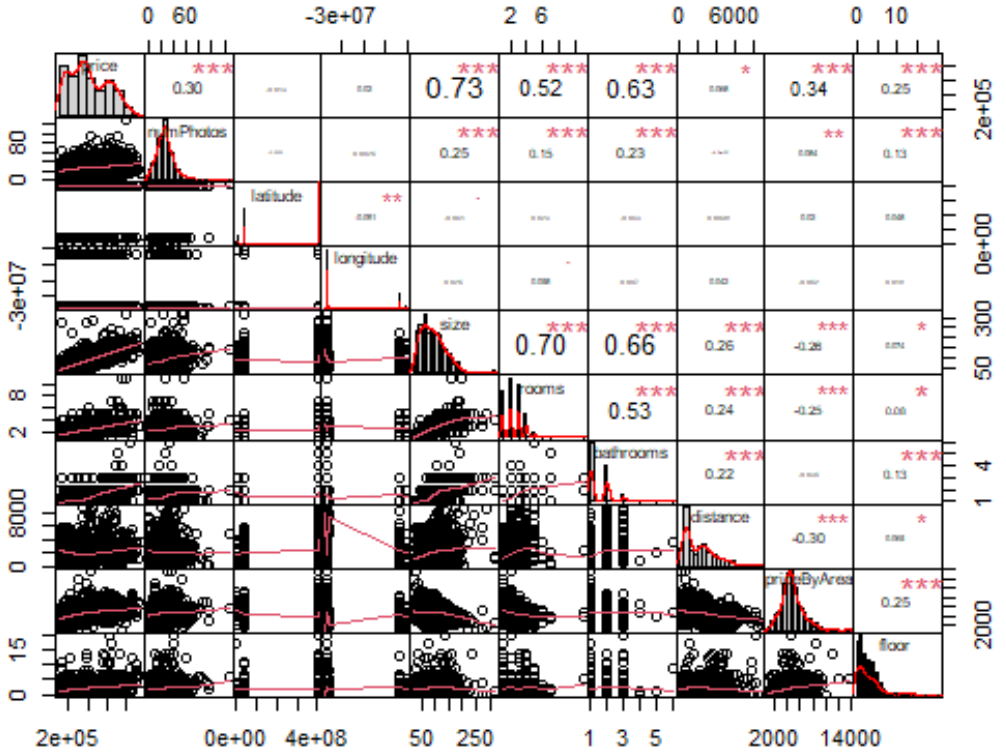
Finalmente, podemos observar como la variable *Price* tiene una relación de dependencia positiva con las variables *priceByArea* y *floor* pese a ser dicha relación débil. Con lo que podemos extraer que un aumento de la planta en la que se encuentra el piso tiene una relación positiva con el precio de dicho inmueble.

Figura 24: Gráfico de correlación entre las distintas variables



Fuente: Elaboración propia

Figura 25: Gráfico de Correlación de variables con p-valores



Fuente: Elaboración propia

5.1.4.- Machine Learning

El origen de este término se remonta a 1943 cuando el matemático Walter Pitts y el neurofisiólogo Warren MacCulloch realizaron un estudio del celebro humano con el objetivo de analizarlo como un organismo computacional y posterior creación de computadoras de funcionamiento similar, dando a conocer el término de inteligencia artificial. Más en concreto, el Machine Learning es la materia o herramienta informática aplicada de esta (Hinestroza, 2018).

Sin embargo, no es hasta el periodo de 2006 en el que grandes empresas como Microsoft o IBM empiezan a destinar grandes cantidades de dinero a su uso, implementación y distribución a nivel mundial.

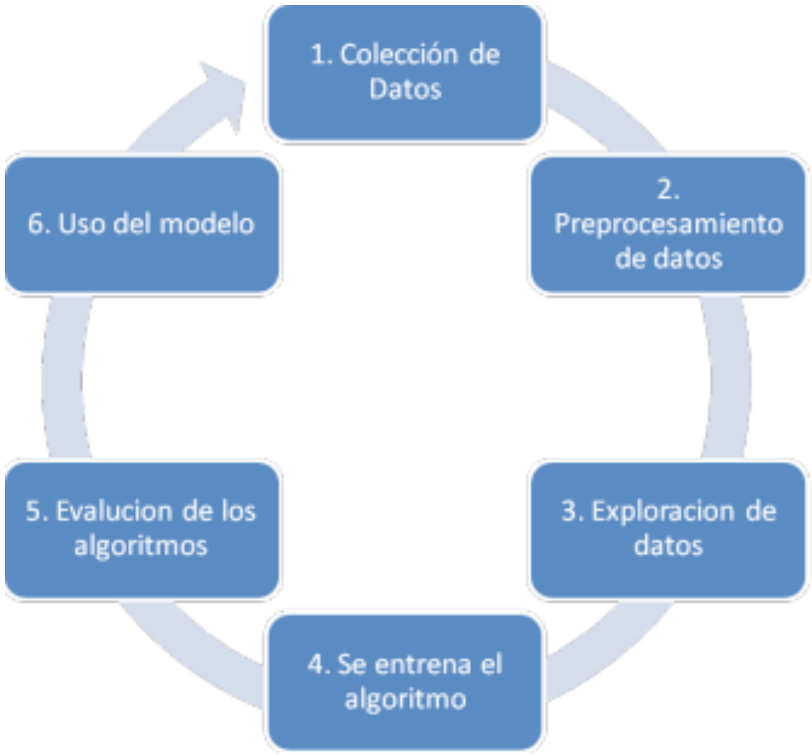
Dentro del campo del Machine Learning, los algoritmos de aprendizaje se clasifican en tres grupos:

- **Aprendizaje supervisado**: caracterizado por ser el tipo de aprendizaje automático en el que el algoritmo aprende a predecir mediante datos etiquetados. A través de la implementación de datos de entrada (conocidos como etiquetas y objetivos) y salida (conocidos como características o predictores), el algoritmo aprende a mapear dichos datos con el objetivo de predecir correctamente futuros datos de entrada. Su principal uso recae en las tareas de clasificación o regresión, distinguiendo estas si el resultado es una etiqueta o categoría o el valor continuo. Algunos ejemplos de los principales algoritmos de aprendizaje supervisado son la regresión lineal, los árboles de decisión, la regresión logarítmica o las redes neuronales (Mueller & Massaron, 2021).
- **Aprendizaje no supervisado**: caracterizado por ser el propio algoritmo el que estudia posibles patrones o semejanzas entre los datos etiquetados; sin saber con anterioridad el verdadero valor o relación entre los mismos. Algunos ejemplos de algoritmos son la agrupación k-means, los autocodificadores o el análisis de componentes principales (ACP). Y su implementación adquiere gran relevancia en grandes volúmenes de conjuntos de datos caracterizados por un alto grado de complejidad (Rojas, 2020).

- **Aprendizaje reforzado**: este tipo de aprendizaje de Machine Learning se caracteriza por aprender de acuerdo con el entorno a través de prueba-error, recibiendo recompensas o penalizaciones en el proceso con el objetivo de minimizar el error a largo plazo. Algunos ejemplos de estos algoritmos son Q-learning, SARSA y los métodos de gradiente de políticas (Lopez, Lopez & Díaz, 2005).

Tras la selección del algoritmo, la implementación de los mismo se caracteriza por la elaboración de un proceso muy similar en los distintos modelos a aplicar (Figura 26).

Figura 26: Proceso de construcción de un modelo de Machine Learning



Fuente: Rojas, E.M. (2020)

En primer lugar, es necesario desarrollar un proceso de recolección de datos necesarios para el estudio a realizar. Dichos datos pueden caracterizarse por estar estandarizados o no, provenir de diversas fuentes como papers, información extraída de APIS u otros tipos y tener o no relación causal, dado que dependerá del objeto de estudio y los posibles *insights* que se obtengan tras su tratamiento. Del mismo modo, dependiendo de las características y métodos de extracción empleados, dichos datos necesitarán en mayor o menor medida la aplicación de un preprocesamiento con el objetivo de estructurarlos y asemejarlos a las necesidades y características de los algoritmos a aplicar.

Posteriormente, la complejidad y dificultad del problema a estudiar hace necesaria una previa exploración de los datos con el objetivo de obtener distintas conclusiones iniciales y conocer posibles distribuciones de los datos para decidir si la implementación del algoritmo posterior es correcta, así como el análisis de posibles valores atípicos que pueden estar presentes o la necesidad de recopilar nuevos datos (López, R. 2015).

Los datos estructurados son utilizados para entrenar el algoritmo aplicando diferentes tipos de análisis de su nivel de predicción, como pueden ser el estudio de la *Accuracy* o cálculos de errores probabilísticos (MSE, RMSE…etc.) con el objetivo de poder extraer conclusiones objetivas, hacer predicciones correctas y evitar sesgos o correlaciones en los resultados.

Finalmente, tras la implementación y evaluación de los distintos pasos anteriores, el modelo puede ser usado de forma correcta e implementado a nuevos volúmenes de datos tratando así de solucionar problemas similares o resolver nuevos.

Sin embargo, uno de los principales problemas que pueden aparecer a la hora de implementar este tipo de algoritmos es el caracterizado por el ajuste de los modelos a características específicas de los datos distintas al objeto de estudio, también conocido como sobre entrenamiento o sobreajuste. Es por ello, que la aplicación de estrategias como la validación cruzada o la retención de datos adquieren importancia y se hacen necesarias para poder desarrollar estudios estadísticamente representativos (López, R. 2015).

En los últimos años, la implementación de dichos algoritmos ha sido muy representativa, así como en la industria 4.0, en la que han representado un impacto directo mediante la mejora de la eficiencia de los sistemas productivos, la calidad de los productos y la seguridad de las personas (Seebo, 2019). Del mismo modo, los modelos de Machine Learning pueden ser desarrollados en diferentes lenguajes de programación; como R o Python, facilitando así su aplicación a una gran variedad de problemáticas, lo que ha aumentado su importancia e implementación en los problemas cotidianos. Más en concreto, en los desarrollados y aquellos que son objeto de estudio de esta memoria como se comentará a continuación.

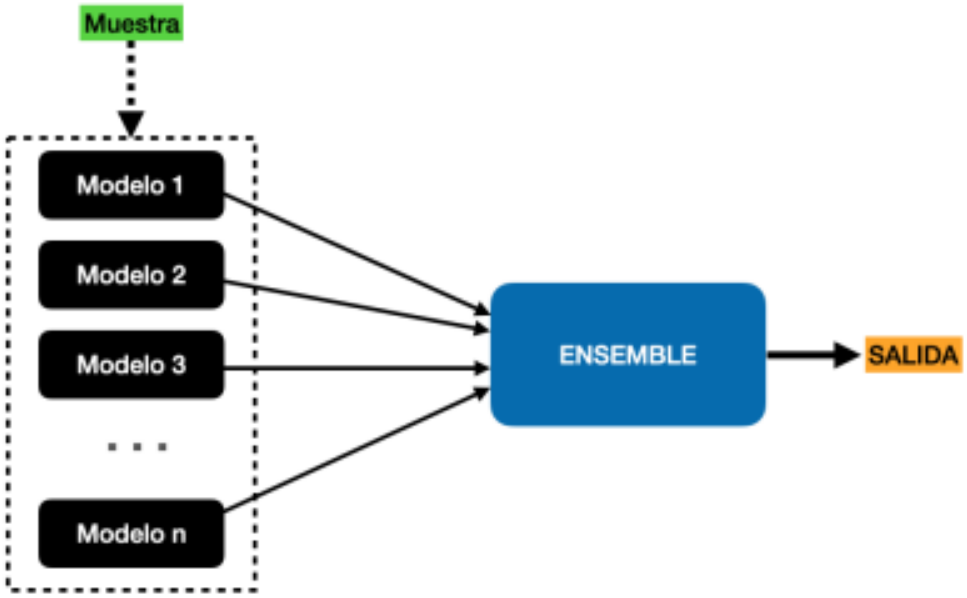
5.1.5.- Descripción analítica de los modelos.

Los ensembles hacen referencia a la unión de varios modelos de predicción de Machine Learning con el objetivo de obtener mejores predicciones que las resultantes de un modelo individual (Figura 27). Este tipo de agrupación se puede llevar a cabo tanto en problemas de clasificación como regresión, así como seleccionar diferentes combinaciones de

modelos como árboles de decisión, redes neuronales, *Supporting Vector Machine* (SVM), etcétera.

El uso de estos métodos puede resultar muy beneficioso en situaciones en las que exista una gran diversidad entre los datos y modelos que se van a combinar. Esto permite entrenar los modelos con diferentes distribuciones o generar muestras de entrenamiento con variaciones entre ellas (Pérez P.J. et al, 2015).

Figura 27: Representación de un ensemble



Fuente: Molina, M. (2022)

De acuerdo con cómo interactúen entre sí los diferentes modelos propios del ensemble, así como la jerarquía que puede llegar a establecerse, podemos destacar (Molina, M. 2022):

- **Voting:** se caracteriza por otorgar la misma capacidad de decisión a los diferentes modelos que forman el ensamble. En el caso de un problema de clasificación, se devuelve la clase más votada, mientras que, en un problema de regresión, se devuelve la media de las predicciones. Dentro de esta categoría, también se pueden distinguir dos técnicas:

    - Hard Voting: Esta técnica utiliza la moda de las predicciones como resultado y prioriza el modelo que tenga una mayor precisión en caso de empate.
    - Soft Voting: En este caso, se utilizan las probabilidades de cada clase en lugar de las predicciones. Se calcula la media de estas probabilidades y se selecciona la clase con la mayor media como resultado final.
- **Stacking**: dicho modelo implica el uso de una estructura jerárquica en la que las salidas de los primeros modelos se utilizan como entradas de los siguientes, creando así un metamodelo. Posteriormente, se selecciona la media o la moda de las predicciones, según el problema.
- **Bagging**: La principal característica de este modelo es que se entrenan múltiples modelos con diferentes conjuntos de datos de entrenamiento. Para lograr esto, se crean particiones aleatorias de los datos, generando subdivisiones distintas. La predicción final se obtiene mediante el promedio o la moda de las predicciones realizadas por los diferentes modelos.
- **Boosting**: Este modelo sigue una estructura jerárquica en la que se busca mejorar las predicciones de los modelos previos. En primer lugar, se entrena un modelo y el modelo siguiente intenta reducir el error del modelo anterior para lograr finalmente alcanzar un umbral de tolerancia aceptable.

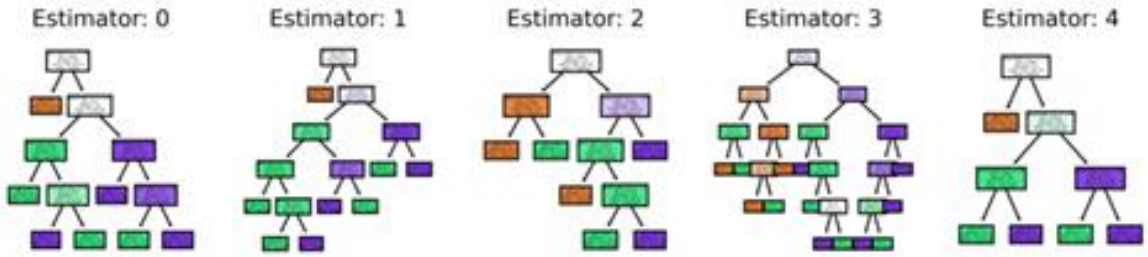
Para nuestro estudio en específico, emplearemos tres modelos distintos de Machine Learning; Random Fofest, XGBoost y TabPFN. En un primer lugar, realizaremos los cálculos oportunos para estudiar la precisión de cada modelo por separado y posteriormente, crearemos un ensemble con los mismos para poder comparar qué modelo realiza una mejor predicción y si el ensemble construido mejora dicha predicción individual.

El modelo de Random Forest se basa en el entrenamiento de múltiples árboles de decisión utilizando subconjuntos diferentes de los datos de entrenamiento. Cada árbol genera sus propias predicciones y estas son combinadas mediante el promedio o votación para obtener una predicción final (Figura 28). Al utilizar un conjunto completo de árboles de decisión, se ayuda a evitar problemas de sobreajuste. Además, este modelo se caracteriza por implementar el proceso de bagging mencionado anteriormente. (Liaw, A. & Wiener, M. 2002).

Dicho modelo depende principalmente de dos parámetros, el número de árboles que forman el bosque y el número de variables (p) seleccionadas en cada nodo. Con respecto a la tasa de error de estos, se establece que el reducir el número de variables (p), se reduce la correlación entre los árboles, pero también la precisión de cada árbol. Por ello, los valores recomendados para un problema de clasificación son (García Ruíz de León, 2018):

$$\sqrt{p}$$

Figura 28: Representación de un Bosque Aleatorio



Fuente: Liaw, A. & Wiener, M. (2002)

El método XGBoost, también conocido como *Extreme Gradient Boosting*, utiliza árboles de decisión como el bosque aleatorio y el reforzamiento de gradientes. Este modelo emplea el proceso de *boosting* mediante la creación de árboles de decisión utilizando una puntuación de similitud de los residuales. A continuación, se calcula la ganancia y la similitud de los datos, y cada árbol aprende de los árboles anteriores sin tener un peso asignado. Dicho modelo es característico del proceso de *boosting* mencionado anteriormente y se utiliza en grandes conjuntos de datos. (Chen, T., & Guestrin, C. 2016).

Dicho método se caracteriza por (Espinosa-Zúñiga, 2020):

- En primer lugar, se obtiene un árbol inicial F0 para predecir una variable "y" y el resultado se asocia con un residual (y-F0)
- Se obtiene un nuevo árbol h1 que se ajusta al error del previo
- Los resultados de los dos árboles se combinan para obtener un árbol total F1, donde el error cuadrático medio será inferior que el de F0.

$$F1x < - F0x + h1(X)$$

- Por último, este proceso se repite hasta que el error se minimice lo máximo posible.

$$Fmx < - Fm\text{-}1x + hm(x)$$

Con respecto al modelo *TabPFN*; a finales de noviembre de 2022 un grupo de investigadores de la Universidad de Freiburg, Alemania formado por Noah Hollman, Samuel Müller, Katharina Eggensperger y Frank Hutter, presentaron un modelo entrenado para realizar clasificación supervisada sobre conjuntos de datos tabulares pequeños caracterizado por emplear un tiempo de procesamiento muy reducido, compitiendo contra los métodos de clasificación más avanzados. Para ello, realizaron un artículo científico en el que presentaron el modelo mediante la implementación en un ejemplo de datos real, así como la descripción de las principales características del mismo (Hollmann et all. 2022).

Los autores de este trabajo optan por no abordar las complicaciones que surgen al entrenar un modelo de Deep Learning en un conjunto de datos tabulares. En su lugar, utilizan un transformador de gran tamaño que ha sido preentrenado en un conjunto de datos tabulares previo para resolver tareas de clasificación probabilística. De esta manera, integran los principios de simplicidad y razonamiento causal en su enfoque.

El método se basa en redes neuronales que se ajustan a datos previos (PFN) y aprenden el algoritmo de entrenamiento y predicción por sí mismas. Las PFN aproximan la inferencia bayesiana utilizando cualquier dato a priori del que se pueda obtener una muestra y proporcionan directamente la distribución predictiva posterior (PPD).

$$p(y|x, D) \propto \int_{\Phi} p(y|x, D, \phi)p(D|\phi)p(\phi)d\phi.$$

El estudio se centra en conjuntos de datos pequeños debido a que, estos conjuntos de datos se encuentran a menudo en el mundo real, así como los métodos de Deep Learning existentes son más limitados en estos ámbitos. Por ello se aplican a datos reales de hasta 1000 muestras de entrenamiento, 100 características y 10 clases, incluyendo tipos de características mixtas, datos perdidos y objetivos desequilibrados.

Una característica principal es que en dicho modelo se utilizan distribuciones en lugar de estimaciones puntuales para todos los hiperparámetros. Los autores enuncian: “En nuestra

priorización definimos los hiperparámetros que describen nuestras hipótesis a priori mediante distribuciones de probabilidad, por ejemplo, una distribución Uniforme de escala logarítmica para el número medio de nodos en los MEC que generan datos. El PPD resultante modela implícitamente la incertidumbre sobre estos hiperparámetros, ponderando los hiperparámetros en las explicaciones de los datos por su verosimilitud dados los datos y su probabilidad a priori y no requiere validación cruzada ni selección de modelos" (Hollman et all, 2022, p.2)

Aunque los ensambles de diferentes arquitecturas e hiperparámetros pueden ofrecer una idea aproximada de los hiperparámetros adecuados, las redes ajustadas a datos anteriores (PFN) permiten abordar estos de manera completamente bayesiana en una sola pasada. Por lo tanto, mediante la definición de una probabilidad a priori sobre el espacio de los hiperparámetros, la distribución predictiva posterior (PPD) aproximado por *TabPFN* integra conjuntamente el espacio de estos y los pesos del modelo en una sola pasada, sin necesidad de realizar inferencia o descubrimiento causal. De esta forma, se resuelve la tarea de predicción descendente directamente (Hollmann et all. 2022).

$$\mathcal{L}_{PFN} = - \mathop{\mathbb{E}}_{\{D_{test} \cup D_{train}\} \sim p(D)} [q_\theta(y_{test} | x_{test}, D_{train})]$$

Sin embargo, algunas de las limitaciones que presentaba dicho modelo son el tiempo de ejecución y el uso de la memoria por parte de la arquitectura PFN.

En este modelo se emplea un transformador que se escala de forma cuadrática en función del número de muestras de entrenamiento, lo que dificulta la inferencia de secuencias que superen los cien mil datos en las GPU actuales. Por este motivo, se ha limitado el número de características a 100 y el número de clases a 10. A pesar de ello, existen métodos que intentan abordar esta limitación, escalando el número de entrada de forma lineal. Por ello, los autores sugieren que estos métodos podrían aplicarse en la arquitectura PFN.

Una vez analizados los distintos modelos, así como las características asociadas a cada uno, se introduce el desarrollo y posterior aplicación de los códigos necesarios en Python con el objetivo de obtener conclusiones significativas así como revisar la hipótesis y objetivos establecidos con anterioridad.

Sección 6: Experimentos

En dicho capítulo se exponen los distintos algoritmos utilizados, así como los resultados obtenidos por cada uno de ellos. Además, se incluye una nueva implementación con el objetivo de estudiar si los resultados obtenidos son suficientemente significativos como para compararlos con el resto de los modelos. Finalmente, se expone la herramienta desarrollada para poder predecir en un futuro nuevos pisos mediante la implementación de los distintos modelos estudiados.

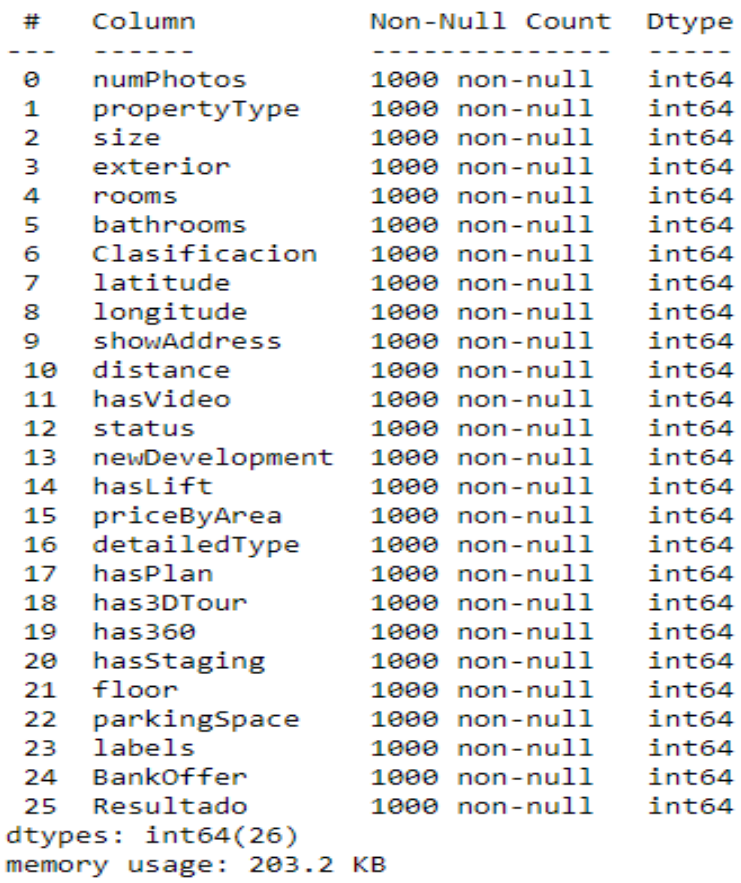
6.1.- Análisis cuantitativo

En cuanto al análisis desarrollado en cada uno de los distintos modelos, el punto de partida ha sido el mismo.

En primer lugar, se descargaban las librerías y paquetes necesarios para llevar a cabo los distintos cálculos y predicciones. Destaca la necesidad de instalar *mlxtend, xgboost, joblib y tabpfn* para poder seguir adelante con el trabajo, además de los distintos paquetes propios de los algoritmos de clasificación como las distintas métricas empleadas de la librería *sklearn.*

Posteriormente, se realiza el leído y carga del archivo Excel para su almacenamiento como *dataframe* y posteriormente su uso para los datos de entrenamiento y test. Como se puede observar en la Figura 29, se llevó a cabo una descripción de los datos objeto de estudio preliminar para poder conocer si había datos nulos o alguna variable guardada en formato distinto al numérico para poder así arreglarlo antes de proseguir y evitar posibles errores futuros.

Figura 29: Información del dataset



```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   numPhotos       1000 non-null   int64
 1   propertyType    1000 non-null   int64
 2   size            1000 non-null   int64
 3   exterior        1000 non-null   int64
 4   rooms           1000 non-null   int64
 5   bathrooms       1000 non-null   int64
 6   Clasificacion   1000 non-null   int64
 7   latitude        1000 non-null   int64
 8   longitude       1000 non-null   int64
 9   showAddress     1000 non-null   int64
 10  distance        1000 non-null   int64
 11  hasVideo        1000 non-null   int64
 12  status          1000 non-null   int64
 13  newDevelopment  1000 non-null   int64
 14  hasLift         1000 non-null   int64
 15  priceByArea     1000 non-null   int64
 16  detailedType    1000 non-null   int64
 17  hasPlan         1000 non-null   int64
 18  has3DTour       1000 non-null   int64
 19  has360          1000 non-null   int64
 20  hasStaging      1000 non-null   int64
 21  floor           1000 non-null   int64
 22  parkingSpace    1000 non-null   int64
 23  labels          1000 non-null   int64
 24  BankOffer       1000 non-null   int64
 25  Resultado       1000 non-null   int64
dtypes: int64(26)
memory usage: 203.2 KB
```

Fuente: Elaboración propia

En tercer lugar, y dado que nuestro problema se caracteriza por llevar a cabo un análisis de la eficacia de la clasificación de los distintos modelos, eliminamos la variable *Resultado* del dataset así como la variable precio debido a la dependencia entre las mismas y poder conseguir así la independencia entre las distintas las variables. Sin embargo, la variable Resultado, dependiente, se almacenó en una lista para poder utilizarla a la hora de realizar la división entre los datos de entrenamiento y test.

Dicha división está caracterizada por el establecimiento de un 75 por ciento de los datos destinados al entrenamiento y el 25 por ciento restante, destinado a la prueba.

Finalmente, una vez empleado todos estos cambios, se empezaron a implementar los cálculos e implementaciones de cada modelo en concreto con el objetivo de realizar predicciones y estudiar el nivel medio de precisión de cada uno de ellos.

6.1.1.- Random Forest

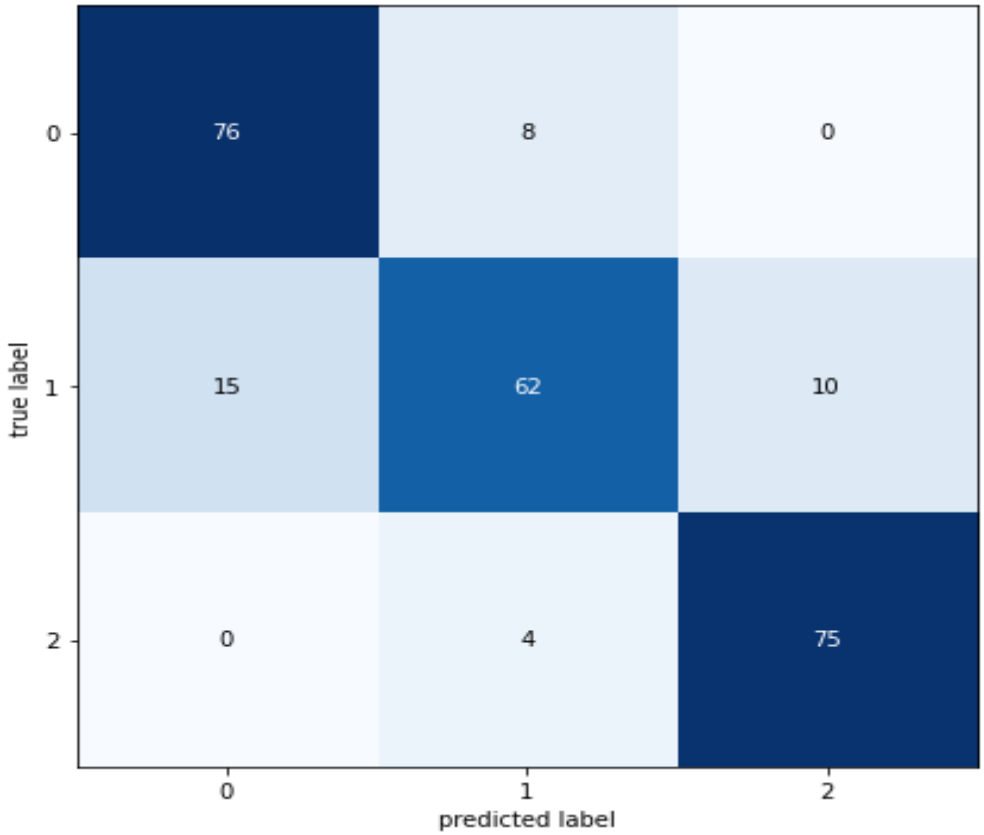
En primer lugar, implementamos el modelo de Bosques aleatorios caracterizado por el uso del método de *Bagging*, o el voto del resultado por mayoría, como comentamos con anterioridad. Para ello, decidimos acudir a la teoría y creamos un modelo de Bosques aleatorios con las siguientes características:

- N_estimators = 19
- Random State = 2016
- Min_Samples_Leaf = 8

La primera variable hace referencia al número de árboles de decisión que se quieren incluir en total en el modelo, en nuestro caso inicial fueron 19. La segunda variable hace referencia a la semilla aleatoria para que los resultados sean reproducibles, que en nuestro caso inicial se estableció en 2016 y la tercera variable hace referencia al número mínimo de observaciones que debe tener cada uno de los nodos hijos para que se produzca la división (hojas).

Con dichas características, el resultado de precisión medio que obtuvimos fue de 0.848 (84.8%). Del mismo modo, como podemos observar en la figura 30, la matriz de confusión resultante se caracterizó por clasificar de forma correcta gran parte de los datos, sin embargo, el modelo clasificó de forma errónea, en mayor medida, los datos como 0 cuando realmente estaban clasificados como 1.

Figura 30: Matriz de confusión del modelo Random Forest



Fuente: Elaboración propia

Una vez obtenidos dichos resultados, se decidió llevar a cabo dos implementaciones con el objetivo de mejorar los resultados anteriores.

En primer lugar, se estudiaron los distintos pesos que el modelo había asignado a las diferentes variables, con el objetivo de eliminar las menos relevantes, así como estudiar la asignación de pesos del resto de modelos e intentar dar solución a una peor clasificación. En la figura 31 podemos observar que el modelo basado en *Random Forest* asignó una mayor importancia a la variable tamaño; *size*, con un 0.395 mientras que no asignó importancia a las variables *hasStaging*, *newDevelopment* y *BankOffer*. Por ello, se decidió eliminar dichas variables del modelo y volver a entrenarlo para analizar si los resultados obtenidos con esta modificación mejoraban la precisión media de nuestro modelo.

Figura 31: Tabla de variables y pesos asignados por Random Forest

	feature	importance
2	size	0.395271
15	priceByArea	0.158345
7	latitude	0.067793
4	rooms	0.060772
6	Clasificacion	0.056868
10	distance	0.052156
5	bathrooms	0.044703
14	hasLift	0.037001
21	floor	0.027060
23	labels	0.021503
3	exterior	0.020730
0	numPhotos	0.020131
8	longitude	0.018769
17	hasPlan	0.004400
18	has3DTour	0.004093
22	parkingSpace	0.002818
16	detailedType	0.002318
11	hasVideo	0.001529
12	status	0.001120
19	has360	0.001015
1	propertyType	0.000989
9	showAddress	0.000616
20	hasStaging	0.000000
13	newDevelopment	0.000000
24	BankOffer	0.000000

Fuente: Elaboración propia

Tras implementar dichos cambios y reentrenar el modelo, obtuvimos una precisión media más alta, de 0.912 (91.2%). Lo que evidenció la necesidad de eliminación de dichas variables, sin embargo, se decidió llevar a cabo un segundo análisis, basado en la optimización y búsqueda de los mejores parámetros para entrenar nuestro modelo; *cross*

validation (hiperparameter tuning). Para ello, se decidió aplicar el método de búsqueda exhaustiva o *Gridsearch* caracterizado por la prueba de todas las posibles combinaciones de valores que se proporcionan como parámetros. Una vez implementado el mismo, se estuvieron que el mejor modelo a implementar era aquel caracterizado por un número mínimo de hojas de 16, una semilla aleatoria establecida de 42 y un total de estimadores de 20.

Los resultados de la precisión media de este modelo fueron de 0.856 (85.6%) con todas las variables iniciales y 0.928 (92.8%) mediante la eliminación de estas tres variables anteriores que menos importancia tenían, representando así una mejor optimización y resultados mediante dicha implementación.

6.1.2.- XGBOOST

En segundo lugar, se implementó el modelo de *Xgboost*, caracterizado por la implementación de *Boosting* o el intento de arreglo de los errores cometidos por los modelos anteriores por parte de los nuevos modelos.

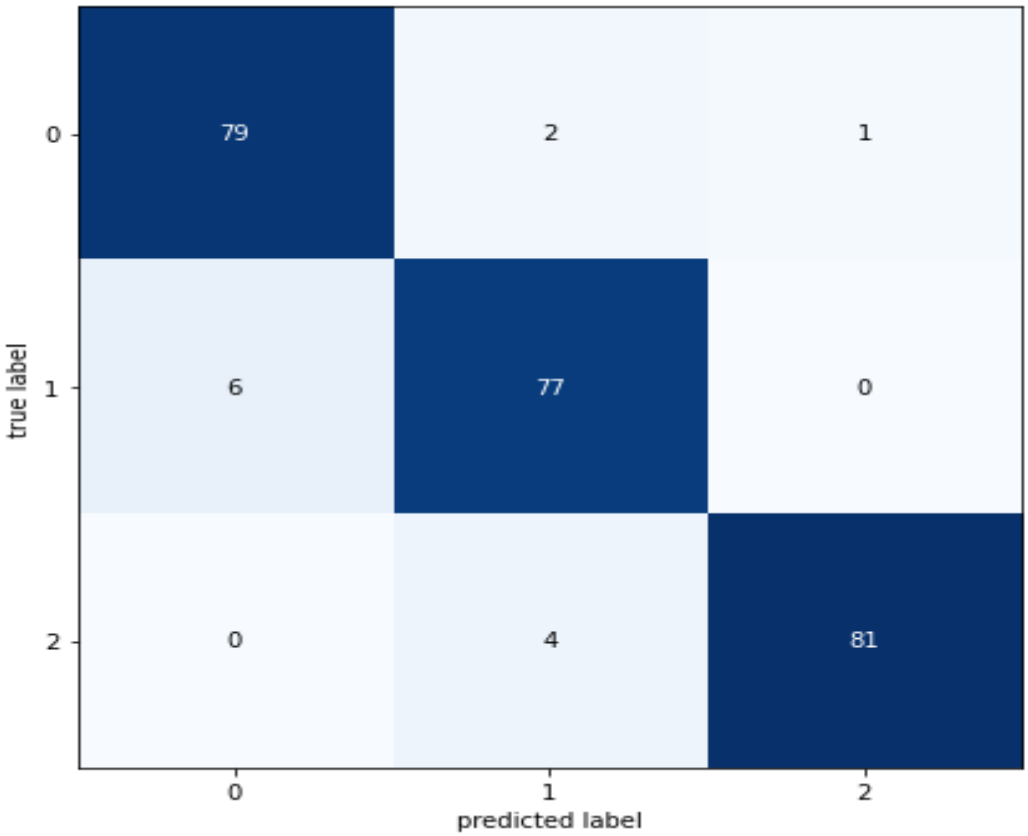
La primera implementación que se hizo se caracterizó por tener las siguientes características:

- N_estimators = 20
- Learning_rate = 0.1
- Max_depth = 7

El número de árboles de decisión totales a implementar fue de 20, la ratio de aprendizaje de 0.1 (un valor alto llega antes al mínimo de la función objetivo, sin embargo, puede llegar a pasarse) y un número total de hojas de 7.

Con dichas características, el modelo entrenado obtuvo una precisión media de 0.948 (94.8%), representando una matriz de confusión como la que se ejemplifica en la figura 32. En este caso, y como pasaba en el anterior, la mayoría de los errores que ha cometido el modelo ha sido a la hora de clasificar el resultado como 0 cuando realmente el valor era 1.

Figura 32: Matriz de confusión del modelo Xgboost



Fuente: Elaboración propia

Con respecto a la importancia que el modelo ha asignado a las distintas variables, se puede destacar que la variable con mayor importancia ha vuelto a ser el tamaño; *size*, sin embargo, la variable que mejor importancia ha establecido el modelo ha sido *showAddress*. Por ello, eliminando dicha variable del modelo, se obtiene una precisión media de 0.96 (96%), mejorando así el resultado obtenido con anterioridad.

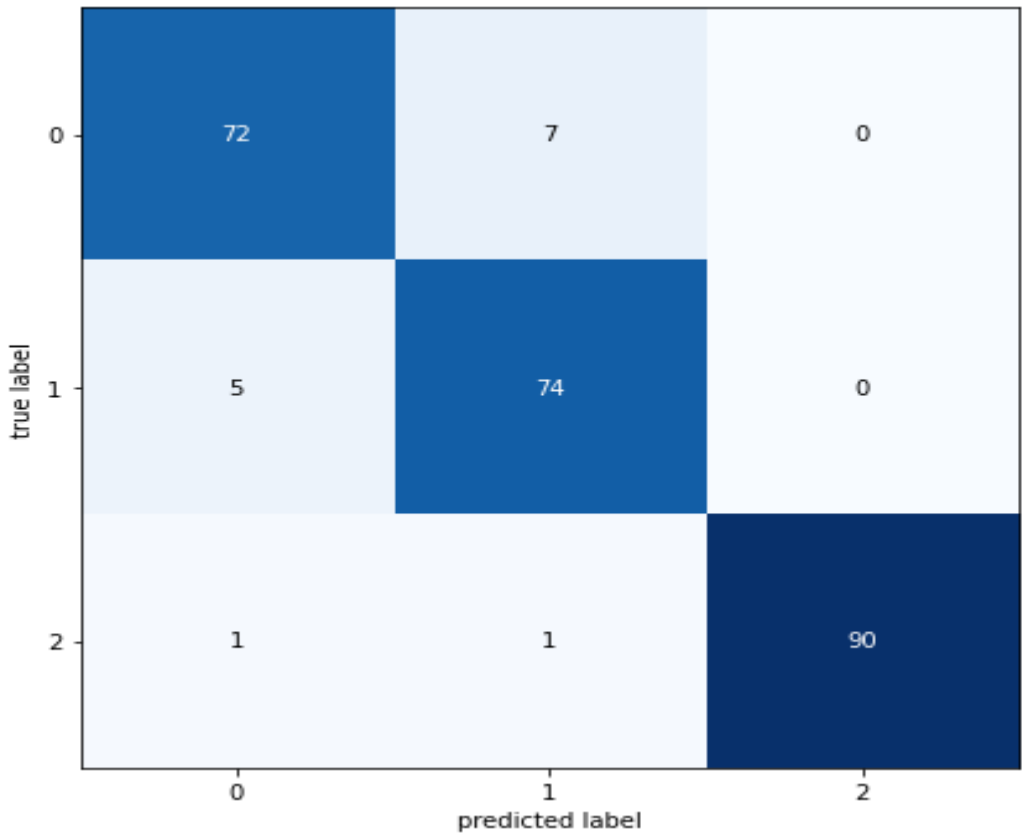
Finalmente, se decidió aplicar el mismo proceso que con anterioridad de búsqueda de mejores parámetros para poder implementar e intentar obtener el modelo óptimo, sin embargo, los resultados obtenidos no llegan a mejorar la precisión media obtenida con anterioridad, posicionando ambas en 0.960 (96%).

6.1.3.- TabPFN

En tercer lugar, para el modelo *TabPFN* no se han podido establecer ciertas variables con anterioridad debido a su opacidad por lo que la única variable a establecer ha sido la de semilla aleatoria, eligiendo el número 42, como en los modelos anteriores.

Una vez entrenado dicho modelo, la precisión media obtenida ha sido de 0.944 (94.4%) y la matriz de confusión resultante se ha caracterizado por clasificar erróneamente en mayor medida aquellos valores que realmente eran 0 pero el modelo los clasificaba como 1, a diferencia del resto de modelos (Figura 33).

Figura 33: Matriz de confusión del modelo TabPFN



Fuente: Elaboración propia

Esta opacidad mencionada ha imposibilitado también la implementación del código para estudiar la importancia que el modelo asigna a las diferentes variables, por lo que se ha decidido eliminar las mismas variables que se eliminaron en un primer lugar;

BankOffer, *hasStaging* y *newDevelopment*. Tras realizar dicha implementación, la precisión media se ha visto incrementada, llegando hasta 0.972 (97.2%), obteniendo el mejor resultado.

6.1.4.- Extra: KNN

En el estudio realizado de clasificación mediante los tres modelos con anterioridad, así como las características de las variables utilizadas para el mismo se pueden realizar ciertas asunciones.

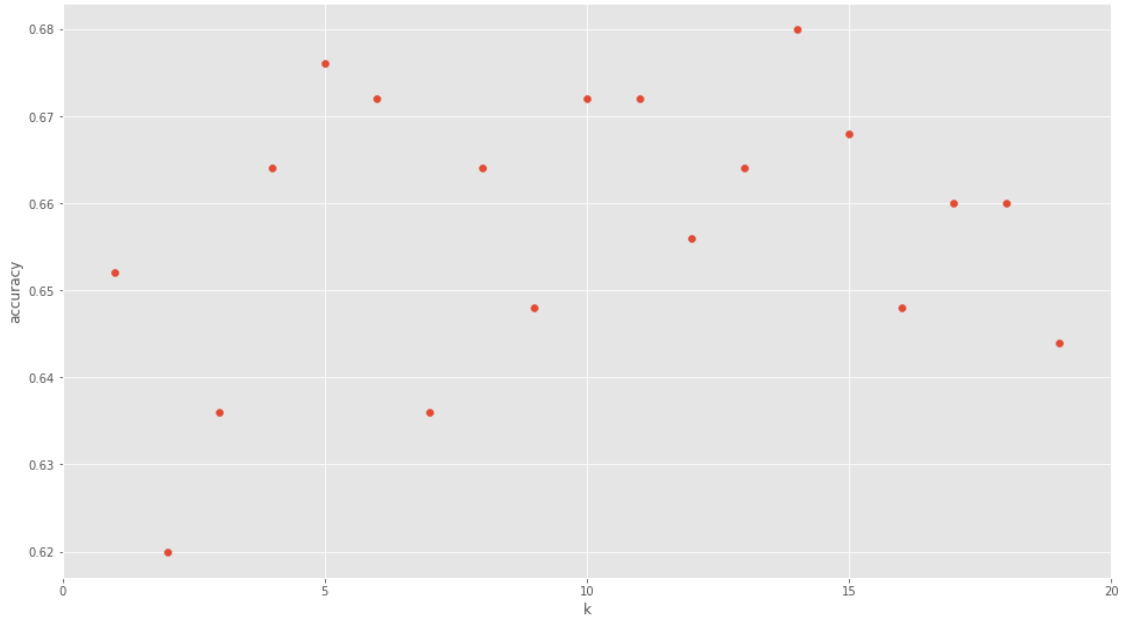
En primer lugar, la variable resultados se calcula de acuerdo con la variable precio, del mismo modo, tenemos las variables de clasificación, en la que se clasifica el inmueble de acuerdo con la renta medio de la zona en la que se encuentra sí como la variable *priceByArea* que representa el precio medio del área en la que se encuentra el inmueble. Entre dichas variables podemos pensar que hacer cierta relación dado que los pisos que se encuentren en zonas cercanas deberían de tener un precio por área parecido, una clasificación de acuerdo con su barrio parecida, un precio similar y, por ende, un resultado similar.

Por ello se ha decidido implementar un modelo de *K-nearest neighbors* (KNN) en el que se intenta estudiar si dichas asunciones anteriores se cumplen y, como resultado, este modelo presenta unas mejores predicciones de clasificación. En definitiva, se intenta estudiar si los pisos se encuentran agrupados principalmente en grupos homogéneos, en nuestro caso 3 debido a la variable Resultado.

En primer lugar, se implementaron las mismas características que en los modelos anteriores de acuerdo con la división entre datos de entrenamiento y datos de prueba. De acuerdo con las asunciones realizas con anterioridad y aplicando el método del estudio de la cercanía mediante la distancia euclídea, se decidió elegir 3 vecinos (k).

Los resultados obtenidos no fueron los esperados. El nivel de precisión medio del modelo fue de 0.72 (72%) así como un mayor error a la hora de predecir los valores. Por ello, se decidió estudiar los distintos niveles de precisión que obtendría el modelo de acuerdo con los distintos k asignados y, como se puede observar en la figura 34, los mayores valores de precisión se obtienen con unos valores de k de entre 10 y 15, por lo que podemos afirmar que los pisos no se encuentran agrupados homogéneamente en tres grupos.

Figura 34: Niveles de precisión para distintos valores de k



Fuente: Elaboración propia

6.1.5.- Herramienta para la predicción

Finalmente, tras los resultados obtenidos, se decidió desarrollar una herramienta que permitiera a los usuarios poder introducir las características específicas de un piso y que dicho modelo conjunto diera como resultado una clasificación lo más precisa posible. Por ello, se juntaron los tres modelos principales caracterizados por estar implementados con las características que mayor precisión promedio dieron en su estudio (ver Anexo).

Para el estudio, se utilizaron los datos de un piso que ya se conocía para observar si el modelo lo clasificaba correctamente. Posteriormente, se introdujeron diversos pisos y el modulo pudo clasificarlos correctamente. Mediante la combinación de los tres modelos iniciales, se consiguió un nivel de precisión medio de 0.945 (94.5%), el cual supera al obtenido por el *Random Forest* y se sitúa muy cercano al obtenido por los otros dos restantes. Sin embargo, hay que tener en cuenta que dicho estudio se ha llevado a cabo mediante la implementación de 1000 observaciones porque sería recomendable que dicho se incrementaran el número de observaciones a utilizar, así como entrenar a los distintos modelos.

El objetivo final sería que una persona pudiera introducir las variables de un piso que ha encontrado y tiene interés por el mismo y pueda saber si se encuentra ante una oportunidad de inversión, la cual se ejemplificaría si dicho piso obtiene un resultado de 1 dado que se encuentra por debajo de la media del resto de pisos de su misma zona. Del mismo modo, si el valor de resultado fuera de 2, el cliente estaría ante un piso con un precio superior al resto de pisos con características similares y si el valor resultante fuera de 2 se encontraría ante unas condiciones de mercado normales.

Con todo ello, podemos destacar los resultados de la figura 35, obtenidos por cada modelo.

Figura 35: Resultados totales de precisión por modelo

Modelos	**Precisión Inicial**	**Precisión Final**
Random Forest	0.848	0.928
Xgboost	0.948	0.960
TabPFN	0.944	0.972
Modelo Combinado	0.945	

Fuente: elaboración propia

Sección 7: Conclusiones y trabajo futuro

Tras los resultados obtenidos de los análisis podemos destacar ciertas conclusiones;

En primer lugar, la hipótesis que se estableció a principio del trabajo de que la aplicación de modelos de clasificación de manera conjunta, ensembles, obtienen resultados fiables acerca de las características de una vivienda se ha visto reforzada. Mas en concreto, con respecto a los resultados de los distintos modelos, así como el modelo combinado, podemos destacar el modelo *TabPFN* como aquel que ha obtenido finalmente un mayor nivel de precisión, así como el *Xgboost*, cuyas presiones finales les han permitido posicionarse por delante del modelo de *Random Forest* y del modelo combinado.

De acuerdo con los desarrollados del modelo *TabPFN*, ha obtenido mejores resultados que el resto para la solución de problemas tabulares con *dataset* muy reducidos; inferiores a mil observaciones, sin embargo, si reciente implementación hace necesario el continuo estudio y uso de este en diferentes proyectos con el objetivo de llegar a unas conclusiones más específicas a cerca de la fiabilidad de este.

Por otro lado, dicha implementación sigue presentando ciertas restricciones. En primer lugar, la difícil accesibilidad a las bases de datos actuales con las características de los pisos que se encuentran en el mercado, así como la poca homogeneidad entre ellas y la necesidad de llevar a cabo estudio de tipo universitario o de investigación para poder recibir el acceso a las mismas. En segundo lugar, el amplio tiempo necesario para poder obtener bases de datos amplias a través de la API de Idealista debido a la única posibilidad de descarga de cien observaciones por mes y la poca flexibilidad que la misma compañía, como el resto de las empresas inmobiliarias ofrecen.

En tercer lugar, la escasa homogeneidad que presentan los datos de los pisos implementados, de acuerdo con los resultados obtenidos por la implementación del modelo KNN, ejemplifican que el mercado inmobiliario es muy complejo y ciertas características de un piso pueden hacer que se diferencie mucho del resto de pisos que se encuentren cerca suyo, lo que acaba principalmente con posibles sesgos en los datos o cierta dependencia.

El número de observaciones empleadas en dicho estudio hace necesaria la futura extracción periódica, limpieza e implementación en los modelos empleados con el

objetivo de seguir entrenando los mismos a la vez que se intenta obtener unos resultados más fieles a la realidad así como una mejor predicción de cara al uso de dicha herramienta desarrollada por parte de clientes finales que decidan llevar a cabo estrategias de inversión en el mercado inmobiliario y quieran hacer uso de dicha herramienta a modo de ayuda en su toma de decisiones.

Por ello, dicho trabajo puede seguir implementándose en el futuro basándose principalmente en dos grandes puntos.

En primer lugar, la necesidad actualizar e ir introduciendo más pisos en la base de datos con el objetivo de entrenar los algoritmos y poder así realizar mejores predicciones para poder hacer así que estas sean cada vez más representativas de la actualidad dado que a un mayor número de pisos de entrada, mejor predicción podremos obtener de salida.

En segundo lugar, dicho análisis podría maximizarse mediante la creación de una herramienta o página web con el objetivo de facilitar la interacción entre los algoritmos y el usuario final. Esta podría caracterizarse por un uso sencillo en la parte del *front-end* en la que los usuarios puedan introducir la información que han recopilado de uno o varios pisos que quieran analizar para saber si es una buena oportunidad de inversión o no y que en la parte *del back-end* se realicen y apliquen todos los filtros necesarios para llevar a cabo dicha predicción. Es decir, desarrollar en mayor medida el punto de "Herramienta para la predicción" que se expone con anterioridad haciéndolo más interactivo y fácil para el usuario final.

Finalmente, podrían estudiarse nuevos algoritmos con el objetivo de obtener mejores predicciones, así como aplicar diferentes metodologías de extracción de datos de los pisos como *web scraping* en diversas páginas y portales web para poder así obtener una *database* más grande.

Con todo ello, y una vez los algoritmos llegaran a obtener niveles de predicción muy satisfactorios, dicha herramienta podría ampliarse para el estudio de inversiones inmobiliarias en distintas ciudades, así como en distintos países.

Sección 8: Bibliografía

Agrasar González, B. (2020). Valoración inmobiliaria a través de Automated Valuation Models (AVMs).

Aigneren, M. (2004). Análisis de datos tabulados. *La Sociología en sus Escenarios*, (10).

Alves, P., & Urtasun, A. (2019). *Evolución reciente del mercado de la vivienda en España* (pp. 1-11). Banco de España.

Andrino, B., Llaneras, K., & Grasso, D. (2021, 30 abril). *El mapa de la renta de los españoles, calle a calle.* ElPaís., de https://elpais.com/economia/2021-04-29/el-mapa-de-la-renta-de-los-espanoles-calle-a-calle.html Recuperado 15 febrero de 2023.

Arora, A. (2017). Specific Cross Validation with Random Forest. *StackOverflow* de https://stackoverflow.com/questions/38151615/specific-cross-validation-with-random-forest Recuperado 20 de mayo de 2023

Bagnato, J. I. (2018). Clasificar con K-Nearest Neighbor ejemplo en Python. de https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/ Recuperado 15 de mayo de 2023

Barreda Agusti, L. (2022, 25 septiembre). Idealista data extraction. Api utils., de https://github.com/laurabarredaagusti/idealista_data_extraction/tree/main/SRC Recuperado 15 de diciembre de 2022.

Behar, R. (2018). Histograma: mucho más que una representación gráfica.

Bozanic Leal, M. S. (2020). Sistema de predicción de precios-venta de inmuebles en el mercado del sector inmobiliario de la Región Metropolitana de la República de Chile con el uso de algoritmos de machine learning.

Cardellino, F. (2021). Tutorial para un clasificador basado en bosques aleatorios: cómo utilizar algoritmos basados en árboles para el aprendizaje automático. *FreeCodeCamp.* Recuperado 20 de mayo de 2023 de

https://www.freecodecamp.org/espanol/news/random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning/

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

de Tudela, C. O. P., & Torres, R. (2019). El mercado de la vivienda: situación y perspectivas a corto plazo. *Cuadernos de Información económica*, (273), 1-9.

Diaz, R. (2023). Random Forest – Bagging y árboles de decisión. *The Machine Learners*. de https://www.themachinelearners.com/random-forest-python/ Recuperado 28 de mayo de 2023

El mercado de la vivienda: evolución reciente y perspectivas. Dirección de Coyuntura, Funcas (2022). de https://www.funcas.es/documentos_trabajo/el-mercado-de-la-vivienda-evolucion-reciente-y-perspectivas/ Recuperado 22 marzo 2023

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, *21*(3).

Estepa, A., & Pino, J. D. (2013). Elementos de interés en la investigación didáctica y enseñanza de la dispersión estadística. Números. Revista de Didáctica de las Matemáticas, 83, 43-63.

Flores, M. I. N. (2007). Las variables: estructura y función en la hipótesis. Investigación educativa, 11(20), 163-182.

García Camargo, P. A. (2021). Implementación de un modelo machine learning para la estimación del valor del metro cuadrado de un inmueble ubicado en Cundinamarca.

García Ruiz de León, M. (2018). Análisis de Sensibilidad Mediante Random Forest.

Gea, M., Batanero, C., & Roa, R. (2014). El sentido de la correlación y regresión. Números, 87, 25-35.

Gómez, F. P. (2020). El mercado de la vivienda: diferencias territoriales en su recuperación. Cuadernos de Información económica, (274), 17-24.

Hernández Chávez, N. D. (2022). Diseño de una aplicación móvil que utiliza sistemas de filtrado de información mediante Machine Learning para identificar, recomendar y describir los datos de contacto de personas que laboran en las distintas fases de proyectos de construcción de viviendas familiares.

Hinestroza Ramírez, D. (2018). El Machine Learning a través de los tiempos, y los aportes a la humanidad.

Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second.

Instituto Nacional de Estadística, INE. (2023) Estadística de Transmisión de Derechos de la Propiedad, Total Nacional, Vivienda Protegida., de https://www.ine.es/jaxiT3/Datos.htm?t=6150#!tabs-grafico Recuperado 15 de marzo de 2023

Instituto Nacional de Estadística, INE (2023). Índice de Precios de la Vivienda (IPV). Nacional, Vivienda Nueva. INE., de https://www.ine.es/jaxiT3/Datos.htm?t=25171#!tabs-grafico Recuperado 15 de marzo de 2023

Instituto Nacional de Estadística, INE (2023). Índice de Precios de la Vivienda (IPV). Nacional, Vivienda Segunda Mano. INE., de https://www.ine.es/jaxiT3/Datos.htm?t=25171#!tabs-grafico Recuperado 15 de marzo de 2023

Instituto Nacional de Estadística, INE (2023). Índice de Precios de Vivienda (IPV). Base 2015 Cuarto trimestre de 2022 [Comunicado de prensa]. https://www.ine.es/daco/daco42/ipv/ipv0422.pdf

Instituto Nacional de Estadística, INE. (2023). Variación Trimestral del IPV. Total Nacional. de https://www.ine.es/consul/serie.do?d=true&s=IPV949&c=2& Recuperado 15 de marzo de 2023

Kabacoff, R. (2020). Data visualization with R. Quantitative Analysis Center: Wesleyan University.

Kabacoff, R. (2022). R in Action: Data Analysis and Graphics with R and Tidyverse. Simon and Schuster.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Lopez Briega, R. E. (2015). Machine Learning con Python.

López Boada, M., López Boada, B., & Díaz López, V. (2005). Algoritmo de aprendizaje por refuerzo continuo para el control de un sistema de suspensión semi-activa.

López Rodríguez, D., & Matea Rosa, M. D. L. L. (2019). Evolución reciente del mercado del alquiler de vivienda en España. Boletín económico/Banco de España [Artículos], n. 3, 18 p.

López Serrano, A. (2019). Evolución de precios en el mercado inmobiliario (vivienda).

Mankiw, N. G. (2012). Principios de economía/N. Gregory Mankiw (No. 330 M5Y 2009.).

Martín, A., Lopez, J.M. (2009). Efecto Darwin en el mercado inmobiliario. Aproximación a un modelo de supervivencia empresarial. Revista de Economía y Derecho, 6(4).

Martinez, J. (2019). Ensembles: voting, bagging, boosting, stacking. *IArtificil*. de https://www.iartificial.net/ensembles-voting-bagging-boosting-stacking/ Recuperado 20 de mayo de 2023

Ministerio de Fomento (2022). Valor tasado de la vivienda. de https://www.fomento.gob.es/BE2/?nivel=2&orden=35000000 Recuperado 16 marzo de 2023

Molina Pérez, M. (2022). Ensembles dinámicos de modelos machine learning para regresión.

Mueller, J. P., & Massaron, L. (2021). Machine learning for dummies. John Wiley & Sons.

Pérez Gallego, P. J., Quevedo Pérez, J. R., & Coz Velasco, J. J. D. (2015). Uso de ensembles en problemas con cambios de distribución caracterizables. In Actas de la XVI Conferencia de la Asociación Española para la Inteligencia Artificial, CAEPIA 2015. Asociación Española para la Inteligencia Artificial.

Raya, J. M. (2020). El impacto de la pandemia en el mercado de la vivienda en España: diagnóstico y políticas.

Reiz, A. N., de la Hoz, M. A., & García, M. S. (2019). Big data analysis y machine learning en medicina intensiva. Medicina Intensiva, 43(7), 416-426.

Revelo, D. (2021). Machine Learning. Bosques aleatorios. *Github*. de https://github.com/DavidReveloLuna/Machine-Learning/blob/master/3_4_BosquesAleatorios.ipynb Recuperado20 Recuperado 18 marzo de 2023

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E28), 586-599.

Sánchez, M. T., Cervantes, Y. E., Cuevas, A. S., Hernández, L. G., & Cuevas, A. J. S. (2015). Modelación Tabular: una alternativa sugerente para el análisis de los datos. Ciencias de la Información, 46(1), 3-10.

Santana, A., Hernandez, C. N. (2016). Gráficos en R: Introducción. *Estadística-dma*. de https://estadistica-dma.ulpgc.es/cursoR4ULPGC/9a-graf-Intro.html Recuperado el 20 de enero de 2023

San José Cabrero, A. (2020). El futuro de la valoración inmobiliaria: Big Data y modelos AVM.

Santos, J. (2022). Invalid classes inferred from unique values of `y`. Expected: [0 1 2 3 4 5], got [1 2 3 4 5 6] *Stackoverflow*. de https://stackoverflow.com/questions/71996617/invalid-classes-inferred-from-unique-values-of-y-expected-0-1-2-3-4-5-got Recuperado 15 de enero de 2023

Sanz, F. (2023). Búsqueda de hiperparámetros – Hyperparameter Tuning. *The Machine Learners*. de https://www.themachinelearners.com/busqueda-hiperparametros/#Hiperparametro Recuperado 28 de mayo de 2023

Sotaquirá, M. (2021). ¿Se requiere SQL para trabajar en Machine Learning?. *Codificandobits*. de https://www.codificandobits.com/blog/sql-machine-learning/#qu%C3%A9-relaci%C3%B3n-hay-entre-el-machine-learning-y-sql Recuperado 15 de febrero de 2023

Soto Hincapié, R. A., & David Rodríguez, E. (2021). Modelo de predicción del precio de la vivienda en el Valle de San Nicolás.

Subdirección General de Estudios y Estadísticas. (2017). Estadística de Valor Tasado de Vivienda. de https://www.mitma.gob.es/recursos_mfom/pdf/B0E2BE62-28EF-41A8-B9D4-CCBD92A28643/144522/MetodValorVivienda.pdf Recuperado 15 de marzo de 2023

Taltavull de La Paz, P. (2001). Economía de la construcción. Civitas.

Taltavull de la Paz, P. (2006). La oferta de viviendas y el mercado inmobiliario en España. Papeles de Economía Española, 109, 156-181.

Tarrés Benet, L. (2019). *Clasificación de lesiones en la piel con un ensemble de redes neuronales residuales* (Bachelor's thesis, Universitat Politècnica de Catalunya).

Vega, J. B. M. Universidad Nacional Autónoma de México (UNAM). "R para principiantes". Libro en bookdown. org. 2016.

Vinuesa, P. (2016). Correlación: teoría y práctica. de https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.html Recuperado 20 de marzo de 2023

Wei, T., & Simko, V. (2017). An introduction to corrplot Package. R package version.

Yu, L., Jiao, C., Xin, H., Wang, Y., & Wang, K. (2018). Prediction on housing price based on deep learning. *International Journal of Computer and Information Engineering*, *12*(2), 90-99.

Zhou, Z. H. (2021). *Machine learning*. Springer Nature.

Sección 9: Anexos

Modificación de las variables:

Variable	Valores Asignados
propertyType	4; flat 3; penthouse 2; duplex 1; chalet 0; studio
exterior	1; True 0; Flase
Clasificacion	3; top 75 2; top 50 1; top 25 0; resto
showAdress	1; True 0; Flase
hasVideo	1; True 0; Flase
status	2; good 1; renew 0; newdevelopment
newDevelopment	1; True 0; Flase
hasLift	1; True 0; Flase
detailedType	4; flat

	3; penthouse 2; duplex 1; chalet 0; studio
hasPlan	1; True 0; Flase
Has3DTour	1; True 0; Flase
Has360	1; True 0; Flase
hasStaging	1; True 0; Flase
parkingSpace	1; True 0; Flase
Labels	6; Lujo 5; Nan 4; fina 3; Apartamento 2; Casa Baja 1; True 0; Buhardilla
BankOffer	1; True 0; Flase

Todos los códigos de Python implementados en esta memoria se encuentran en el siguiente enlace:

GitHub - ZombieKiller9/TFG_Analytics: Este repositorio cuenta con los códigos utilizados y desarrollados en el desarrollo de la memoria