MÁSTER EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA.

TRABAJO FIN DE MÁSTER

ANÁLISIS PREDICTIVO SOBRE LA EVOLUCIÓN TURÍSTICA EN LA CIUDAD DE MADRID

Autor: Andrés Canalejo Oliva

Director: Alejandro Llorente Pinto

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Análisis predictivo sobre la evolución turística en la ciudad de Madrid

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2022/23 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

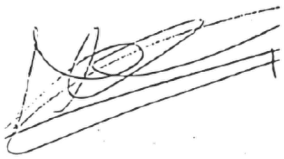
El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Andrés Canalejo Oliva Fecha: 06 / Junio / 2023

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO



Fdo.: Alejandro Llorente Pinto Fecha: 06 / Junio / 2023

MÁSTER EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA.

TRABAJO FIN DE MÁSTER

ANÁLISIS PREDICTIVO SOBRE LA EVOLUCIÓN TURÍSTICA EN LA CIUDAD DE MADRID

Autor: Andrés Canalejo Oliva

Director: Alejandro Llorente Pinto

Madrid

Agradecimientos

Expresar mi más sincero agradecimiento a mi tutor Alejandro Llorente Pinto por su constante atención y guía. Agradezco a mi familia y amigos por su apoyo incondicional durante el desarrollo de este proyecto. También agradezco a mis compañeros de clase y profesores por el conocimiento compartido. En resumen, a todos ustedes, gracias por su gran apoyo en mi Trabajo de Fin de Máster.

ANÁLISIS PREDICTIVO SOBRE LA EVOLUCIÓN TURÍSTICA EN LA CIUDAD DE MADRID

Autor: Canalejo Oliva, Andrés.
Director: Llorente Pinto, Alejandro.
Entidad Colaboradora: PiperLab – Business Data Science Differently.

RESUMEN DEL PROYECTO

Madrid es uno de los principales destinos turísticos de España y Europa. Por ello, en este proyecto se ha seleccionado a la capital de España como objeto de análisis para evaluar sus diferentes secciones censales de la ciudad en términos de problemas potenciales de sobreexplotación o subexplotación de la vivienda turística. Para lograr esto, se han utilizado técnicas de machine learning para poder comprender que factores tienen un mayor impacto tanto positivo como negativo en el mercado inmobiliario. Estos análisis permiten detectar aquellas zonas que aún no han evolucionado en términos de vivienda turística, pero que tienen el potencial de desarrollarse y alcanzar valores similares a los observados en otras áreas. Este estudio presenta finalmente sus resultados como una herramienta de Power BI que permite la interacción para ayudar a todas las posibles partes interesadas.

Palabras clave: Vivienda turística, Sobreexplotación, Evolución, Análisis explicativo, Secciones censales, Comunidad de Madrid, Impacto económico.

1. Introducción

Cada año, miles de visitantes deciden visitar la Comunidad de Madrid para disfrutar de su alta calidad de vida, su hospitalidad y su variedad de opciones culturales y gastronómicas. Como consecuencia del aumento constante de turistas y con la aparición de plataformas de alquiler turístico Peer to Peer, el mercado de la vivienda ha experimentado cambios significativos.

Aunque estos cambios han impulsado la economía local y han generado puestos de trabajo en el sector turístico, también han tenido efectos negativos, como un aumento en los costos de alquiler a largo plazo y una pérdida de identidad en las áreas altamente explotadas. Por ello es fundamental analizar y comprender estos cambios para identificar aquellas áreas que se encuentran bajo situación de sobreexplotación o subexplotación, y poder detectar los factores que contribuyen al crecimiento de la vivienda turística.

2. Definición del proyecto

Debido a la necesidad de analizar la evolución de la vivienda turística en la Comunidad de Madrid se decide realizar un análisis explicativo que permita determinar qué factores tienen un impacto positivo o negativo en el mercado inmobiliario, qué zonas se encuentran sobreexplotadas y qué secciones censales que aún no tienen una gran cantidad de viviendas turísticas pueden evolucionar hacia valores similares a los que se ven en otras zonas.

Por último, desarrollar una herramienta de visualización que permita analizar los resultados de una manera clara y comprensible para que sea utilizada por las partes interesadas permitiendo una toma de decisiones más inteligente.

3. Análisis y desarrollo

Mediante el uso de variables recogidas del Instituto Nacional de Estadística (INE) [1] y del catálogo de datos abiertos de la plataforma de AirBnB [2] se comienza a desarrollar el proyecto haciendo un análisis inicial detallado de la distribución de las variables más importantes para garantizar una buena comprensión de los conjuntos de datos que se van a utilizar. A continuación, se utilizan los datos recopilados de AirBnB para generar variables específicas que asignar a cada zona. De este modo se generan distintas variables de entrada de los modelos, así como la variable de salida, representada en la Figura 1 que explica el número de viviendas por cada sección censal. Para la creación de esta variable se transforman ambos conjuntos de datos al sistema de coordenadas geográficas y se utiliza la librería GeoPandas para ubicar cada vivienda dentro de cada sección censal.

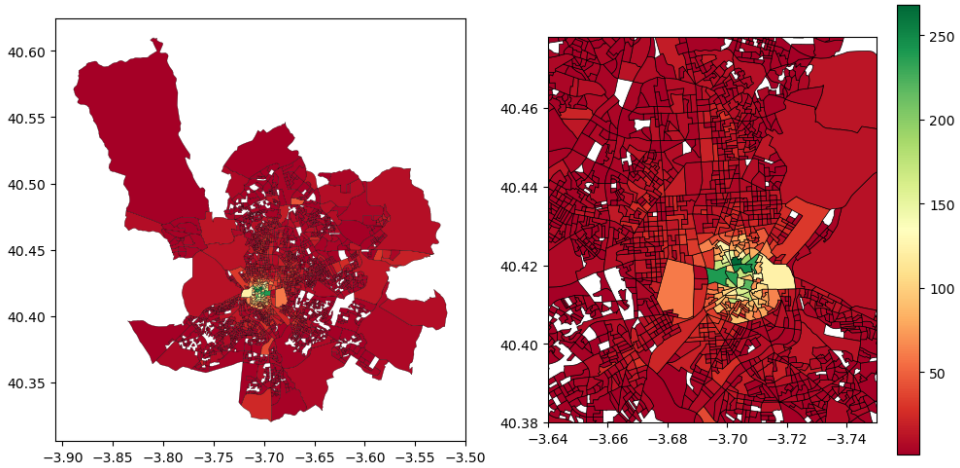


Figura 1. Mapa de la Comunidad de Madrid donde se ha representado el número total de viviendas en cada sección censal

De cara al proceso de modelado se han utilizado los algoritmos de CatBoost Regressor y Decision Tree Regressor como modelos explicativos para identificar aquellas secciones censales que requieren atención especial. Una vez se han implantado los modelos, se calcula la variable encargada de representar el grado de saturación de cada sección censal mediante la resta del valor real y el valor del modelo.

4. Resultados

Se ha utilizado la librería PyCaret para facilitar el proceso de modelado de datos, seleccionando dos modelos específicos, el CatBoost Regressor debido a su alta precisión y el Decision Tree Regressor por su capacidad explicativa. [3]

El CatBoost Regressor es el encargado de generar la variable que representa el grado de saturación de las secciones censales, reflejada en la Figura 2. La saturación de una

sección censal será mayor a medida que aumenta el valor de la variable y será menor cuanto menor sea dicho valor.

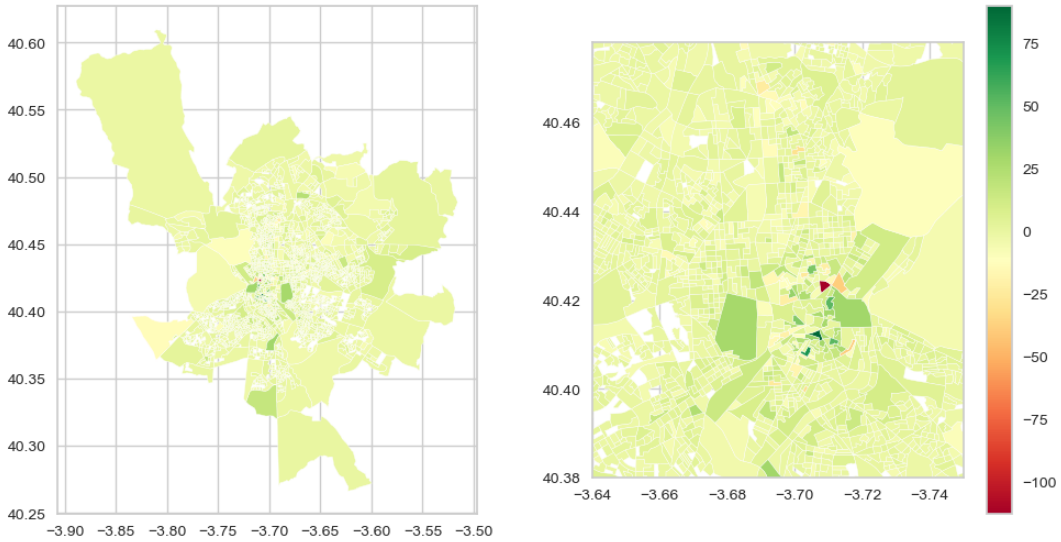


Figura 2. Distribución geográfica de la variable calculada que mide el grado de saturación

Además, se incorporan los valores shap para comprender la importancia de las variables en el modelo y determinar qué componentes tienen un impacto positivo o negativo en las viviendas turísticas. [4] Tanto el árbol de decisión como los valores shap consideran que las variables de "porcentaje de edad 0-24" y "precio medio por sección censal " son dos de las variables con mayor influencia en la variable de salida.

5. Conclusiones

Finalmente, se elabora un dashboard interactivo desarrollado en Power BI para poder ofrecer un análisis detallado y personalizado al usuario gracias a la interacción que permite esta herramienta de visualización. En primer lugar, se desarrolla una página que presenta un análisis de la relación entre la saturación y las características sociodemográficas de cada sección censal y en segundo lugar se elabora un dashboard mucho más interactivo y detallado, facilitando la identificación de patrones y oportunidades para una toma de decisiones más informada.

6. Referencias

[1] Instituto Nacional de Estadística (INE). (s.f.). https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm

[2] Inside AirBnB - Get the Data. (s.f.). http://insideairbnb.com/get-the-data/

[3] Equipo de desarrollo de PyCaret. Documentación de PyCaret 3.0. (2023). https://pycaret.gitbook.io/docs/

[4] Equipo de Desarrollo de SHAP. Documentación de SHAP. (2023). https://shap.readthedocs.io/en/latest/index.html

PREDICTIVE ANALYSIS OF TOURISM DEVELOPMENT IN THE CITY OF MADRID

Author: Canalejo Oliva, Andrés.
Supervisor: Llorente Pinto, Alejandro.
Collaborating Entity: PiperLab – Business Data Science Differently.

ABSTRACT

Madrid is one of the main tourist destinations in Spain and Europe. Therefore, in this project, the capital of Spain has been selected as the object of analysis to evaluate its different census tracts in terms of potential problems of over or under exploitation of tourist housing. To achieve this, machine learning techniques have been used to understand which factors have a greater impact, both positive and negative, on the real estate market. These analyses allow us to detect those areas that have not yet evolved in terms of tourist housing, but which have the potential to develop and reach values like those observed in other areas. This study finally presents its results as a Power BI tool that allows interaction to help all possible stakeholders.

Keywords: Tourist housing, Overexploitation, Evolution, Explanatory analysis, Census sections, Community of Madrid, Economic impact.

1. Introduction

Every year, thousands of visitors decide to visit the Community of Madrid to enjoy its high quality of life, its hospitality, and its variety of cultural and gastronomic options. As a result of the constant increase in tourists and with the emergence of Peer to Peer tourist rental platforms, the housing market has experienced significant changes.

While these changes have boosted the local economy and generated jobs in the tourism sector, they have also had negative effects, such as an increase in long-term rental costs and a loss of identity in highly exploited areas. It is therefore essential to analyze and understand these changes in order to identify those areas that are over or under exploited, and to detect the factors that contribute to the growth of tourist housing.

2. Project Definition

Due to the need to analyze the evolution of tourist housing in the Community of Madrid, it was decided to carry out an explanatory analysis to determine which factors have a positive or negative impact on the real estate market, which areas are overexploited and which census tracts that do not yet have a large amount of tourist housing can evolve towards values like those seen in other areas.

Finally, a visualization tool is developed to analyze the results in a clear and understandable way so it can be used by stakeholders for a more intelligent decision making.

3. Analysis and development

The project begins by using variables collected from the National Institute of Statistics (INE) [1] and from the AirBnB platform's open data catalog [2] to perform a detailed initial analysis of the distribution of the most important variables to ensure a good understanding of the datasets to be used. The data collected from AirBnB is then used to generate specific variables to assign to each zone. This way, different input variables of the models are generated, as well as the output variable, represented in Figura 3 that explains the number of dwellings per census tract. This variable is created by transforming both datasets to the geographic coordinate system and using the GeoPandas library to locate each residence within each census tract.

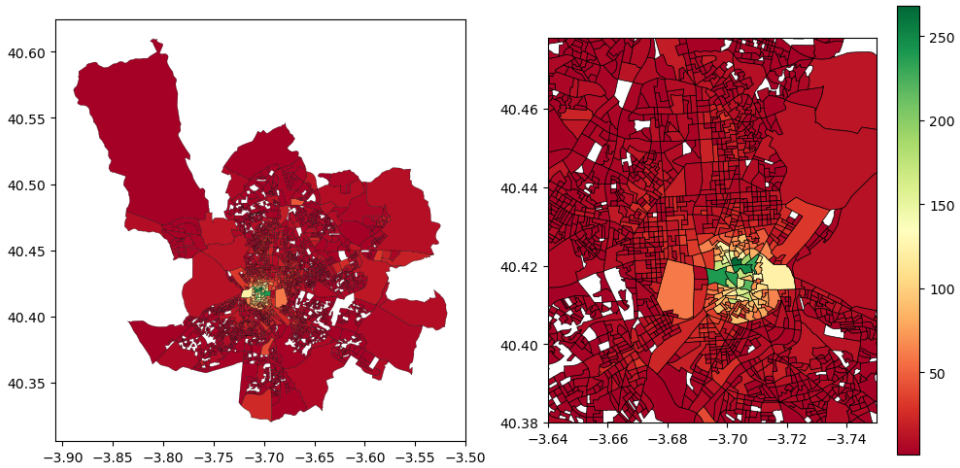


Figura 3. Map of the Community of Madrid showing the total number of properties in each census section.

For the modeling process, the CatBoost Regressor and Decision Tree Regressor algorithms have been used as explanatory models to identify those census tracts that require special attention. Once the models have been implemented, the variable representing the degree of saturation of each census tract is calculated by subtracting the real value and the model value.

4. Results

The PyCaret library has been used to simplify the data modeling process, selecting two specific models, the CatBoost Regressor due to its high accuracy and the Decision Tree Regressor for its explanatory capacity. [3]

The CatBoost Regressor is in charge of generating the variable that represents the degree of saturation of the census tracts, reflected in Figura 4. The saturation of a census tract is considered higher as the value of the variable increases, while it will be less the lower the value.

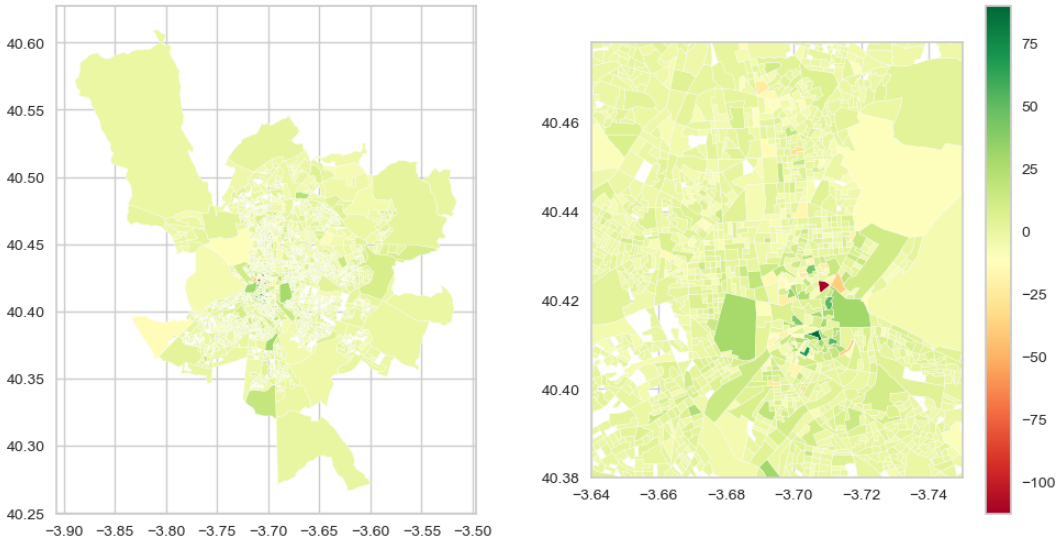


Figura 4. Geographical distribution of the calculated variable measuring the degree of saturation

In addition, shap values are incorporated to understand the importance of the variables in the model and to determine which components have a positive or negative impact on tourist housing. [4] Both the decision tree and the shap values consider that the variables "percentage of age 0-24" and "average price per census tract" are two of the variables with the greatest influence on the output variable.

5. Conclusions

Finally, an interactive dashboard developed in Power BI is elaborated to offer a detailed and personalized analysis to the user through the interaction that this visualization tool allows. Firstly, it is developed a page that presents an analysis of the relationship between saturation and the sociodemographic characteristics of each census section and secondly, a much more interactive and detailed dashboard is developed, allowing the identification of patterns and opportunities for a more informed decision making.

6. References

[1] National Institute of Statistics (INE). (n.d.). https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm

[2] Inside AirBnB - Get the Data. (n.d.). http://insideairbnb.com/get-the-data/

[3] PyCaret development team. PyCaret 3.0 Documentation. (2023). https://pycaret.gitbook.io/docs/

[4] SHAP Development Team. SHAP documentation. (2023). https://shap.readthedocs.io/en/latest/index.html

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

El turismo siempre ha sido una de las principales industrias de España, y Madrid, la capital del país, no es una excepción. Cada año, miles de personas visitan la capital española para disfrutar de la alta calidad de vida, la amabilidad de sus habitantes y la variedad de opciones culturales y gastronómicas. Como resultado del aumento constante de los turistas y la aparición de plataformas de alquiler turístico como Idealista o AirBnB (las conocidas como plataformas P2P o peer to peer), el mercado de la vivienda ha experimentado cambios significativos en los últimos años.

Actualmente, la mayor parte de las secciones censales de la comunidad de Madrid cuentan con viviendas de alquiler turístico. Esto permite obtener grandes beneficios, incluida la mejora de la economía local y la creación de nuevos empleos en el sector turístico, lo que permite a los propietarios aumentar sus ingresos y aprovechar el potencial turístico de la ciudad. Además, al alojarse en barrios tradicionales y áreas residenciales en lugar de estar restringidos a opciones hoteleras más convencionales, los visitantes pueden experimentar la ciudad de una manera más auténtica.

Sin embargo, es importante entender que la sobreexplotación de estas casas de alquiler turístico puede causar problemas. La subida de los costes de los alquileres de larga duración es uno de los principales inconvenientes. Debido a la alta demanda de alquileres turísticos, cada vez hay menos opciones de viviendas asequibles disponibles para su compra. Además, las áreas altamente explotadas suelen perder su identidad vecinal debido al ruido que el turismo masivo puede causar para los residentes.

La importancia de comprender el desarrollo de estas viviendas de alquiler turístico es evidente teniendo en cuenta los aspectos mencionados. Es fundamental comprender los efectos de las viviendas de alquiler en cada sección censal, ya que esto permite localizar e identificar áreas que están siendo poco explotadas y áreas que están siendo excesivamente

explotadas. También puede ser útil conocer los factores que contribuyen al aumento de la vivienda turística.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

Para facilitar la lectura y comprensión del proyecto se describen a continuación dos de las distintas tecnologías utilizadas en él. Estas son la biblioteca PyCaret y los valores SHAP.

2.1 LIBRERÍA PYCARET

PyCaret es una biblioteca de aprendizaje automático de Python de código abierto que ofrece una interfaz simple y de alto nivel para llevar a cabo operaciones clásicas de aprendizaje automático. [1]

La principal ventaja de PyCaret es su enfoque en la automatización y la productividad. PyCaret permite que los usuarios completen rápidamente tareas complejas de aprendizaje automático con solo unas pocas líneas de código reduciendo el tiempo y el esfuerzo necesarios para desarrollar modelos de aprendizaje automático. Además, PyCaret ofrece una amplia variedad de algoritmos de aprendizaje automático predefinidos, así como herramientas integradas para la selección de características y la evaluación del rendimiento del modelo. [1]

2.2 VALORES SHAP

Shap es una librería popular ampliamente usada en la comunidad para mejorar la interpretación de modelos. En muchos casos, al aplicar un modelo a nuestro problema, buscamos encontrar el equilibrio entre interpretabilidad y precisión, ya que los modelos altamente interpretables no suelen tener una alta precisión y los modelos altamente precisos con frecuencia son cajas negras que resultan difíciles de interpretar. Shap entonces aparece como una herramienta útil para comprender cómo funcionan los modelos de caja negra en este contexto.

La contribución de Shap destaca por su capacidad para explicar cómo afecta cada variable al resultado final del modelo, lo que nos permite comprender cómo se toman las decisiones. Para lograr esto, Shap emplea el método matemático conocido como "valores Shapley", que asigna de manera justa la contribución de cada variable. El uso de Shap en el proyecto ha permitido comprender mejor el Catboost Regressor, lo que ayuda a tomar decisiones más inteligentes. [2]

En los valores shap las variables de entrada del modelo se ordenan según su importancia en el impacto del resultado del modelo. Cada variable se representa con barras de color rojo para los valores altos o de color azul para los valores bajos. Estas se extienden a lo largo del eje hacia la derecha o izquierda para indicar un impacto positivo o negativo. [3]

Capítulo 3. ESTADO DE LA CUESTIÓN

En el presente capítulo se realiza un análisis de los trabajos y soluciones existentes en el desarrollo de la vivienda turística en la ciudad de Madrid, junto con el análisis predictivo correspondiente. El primer paso que debe darse siempre antes de empezar a desarrollar un proyecto es preguntarse si existen proyectos similares en el mercado y si hay investigaciones previas que hayan resultado relevantes para el objetivo. Este análisis previo ayudará a justificar la relevancia y la necesidad del proyecto.

3.1 TRABAJOS PREVIOS E INVESTIGACIONES

En los últimos años, debido al aumento constante del turismo y la creciente demanda de viviendas turísticas, una gran variedad de disciplinas académicas y profesionales han estudiado el sector del turismo.

Entre las investigaciones existentes que tratan el desarrollo de la vivienda turística y el análisis predictivo en ciudades y contextos similares destaca el artículo elaborado por Perles-Ribes, JF, Ramón-Rodríguez, AB, Moreno-Izquierdo, L y Such-Devesa, MJ donde se explica cómo predecir situaciones de sobre turismo en los destinos turísticos de España utilizando técnicas de machine learning. [4] Otro ejemplo, es un proyecto de la Comunidad de Madrid en el que se realiza un estudio de campo de carácter puntual (estado del arte) y específico para el Distrito Centro sobre la evolución de las viviendas destinadas a uso turístico. [5] Un último ejemplo es el elaborado por Casado Buesa, MP donde mediante el uso de técnicas de machine learning realiza una investigación con el propósito principal de definir la relación entre la expansión del turismo masivo y el empobrecimiento y deterioro de la calidad de vida de los habitantes de Barcelona, especialmente en lo que respecta a la vivienda. [6]

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

Aunque se han realizado investigaciones y trabajos previos sobre el desarrollo de la vivienda turística, es importante señalar que cada ciudad tiene características y dinámicas únicas. El análisis de la evolución de la vivienda en Madrid es importante debido a su posición como un destino turístico significativo y su posible repercusión en los precios de la vivienda y otros bienes básicos. Es importante tener en cuenta que, al realizar este análisis con técnicas de machine learning, se crea un proyecto único que puede generar mucho valor al sector del turismo.

Además, uno de los puntos a destacar es el desarrollo de una herramienta que permite una visualización clara y accesible a los resultados del proyecto. Esta herramienta permite la interacción con los datos, lo que permite que cada persona llegue a sus propias conclusiones y aborde de manera efectiva sus intereses particulares.

Tener una herramienta que permita analizar y sacar conclusiones de manera fácil y rápida es muy valioso, especialmente cuando se trata de un tema tan relevante como el desarrollo de viviendas turísticas. El potencial impacto y utilidad de esta herramienta se incrementará aún más si se difunde a otras ciudades y países.

4.2 OBJETIVOS

Realizar un análisis explicativo sobre la evolución de la vivienda turística en la ciudad de Madrid, determinando qué factores son los más relacionados con el aumento de este tipo de actividad económica y qué áreas que aún no tienen una gran cantidad de viviendas turísticas pueden evolucionar hacia valores similares a los que se ven en otras zonas, lo que acarrea una subida generalizada de los precios de la vivienda o de otros bienes básicos en esas zonas.

Finalmente, desarrollar una herramienta la cual permita visualizar los resultados de manera clara y comprensible. De forma que aquellas partes interesadas como por ejemplo los inversores inmobiliarios puedan tomar decisiones informadas mediante la ayuda de esta herramienta.

4.3 METODOLOGÍA

Para la consecución de estos objetivos, el presente proyecto seguirá los siguientes pasos:

Identificación de fuentes de información

Se han recopilado datos de dos fuentes principales para llevar a cabo este proyecto. Primero, se utilizaron los datos de viviendas de AirBnB de la página web Inside AirBnB, que proporciona estadísticas históricas trimestrales por región. Se han elegido dos conjuntos de datos correspondientes a diciembre de 2021 y 2022. [7]

Por otro lado, se ha utilizado el catálogo de datos abiertos proporcionado por el Instituto Nacional de Estadística (INE) para complementar esta información. Este catálogo ha seleccionado un conjunto de datos que contiene información sociodemográfica detallada para cada sección censal. [8]

Adaptación y transformación de los datos

Ambos conjuntos han pasado por una exhaustiva limpieza de datos. Los errores se han corregido y la información incompleta se ha corregido, lo que nos garantiza un conjunto de datos de alta calidad. Las coordenadas también se han cambiado al sistema de referencia geográfica EPSG 4326, ampliamente utilizado en aplicaciones geográficas para mostrar con precisión la ubicación de los objetos en la superficie terrestre.

Contabilización de viviendas por cada sección censal

Cada vivienda turística se asocia con su sección censal correspondiente al cruzar el archivo de secciones censales con los dos archivos de AirBnB. Para ello, se utiliza la ubicación de

las viviendas para ubicarlas de manera precisa en el mapa y contar el número de viviendas en cada sección. Finalmente se muestra una representación detallada de la distribución geográfica de la vivienda turística en la Comunidad de Madrid.

Análisis descriptivo del problema

Para comparar la evolución en la vivienda turística de Madrid entre los dos periodos, se ha realizado un análisis de las características de los anuncios de AirBnB, que incluyen precios, barrios y tipos de vivienda. El objetivo de este análisis inicial es comprender los datos y poder identificar tendencias y patrones en la evolución de la vivienda turística en la ciudad.

Generación de variables

Se han creado variables explicativas en este paso del análisis para mejorar la predicción y la comprensión del modelo y se van a usar como variables de entrada y de salida de los modelos. Estas variables permiten una mejor comprensión de los factores que afectan a la vivienda turística, lo que mejora la precisión de las predicciones.

Entrenamiento y generación de predicciones

Se entrenan varios modelos explicativos basados en la tendencia general de los datos y las variables sociodemográficas con el objetivo de predecir de manera precisa las variables de salida y lograr los objetivos establecidos en el proyecto.

Evaluación de la saturación de secciones censales

Para determinar la saturación en cada sección censal en términos del número de viviendas turísticas, obtendremos la diferencia entre el número real de viviendas por zona y su valor predicho.

Visualizador de resultados

Finalmente, se crea un dashboard en Power BI para mostrar de manera atractiva los resultados obtenidos y proporcionar una herramienta fácil de interactuar con ella.

4.4 PLANIFICACIÓN Y ESTIMACIÓN ECONÓMICA

Se muestra en esta sección del capítulo la planificación del presente proyecto. Es importante tener en cuenta que la duración de cada actividad podría verse modificada según las distintas necesidades del trabajo.

A continuación, se presenta en la Figura 5 un calendario tipo contemplando las actividades de las tres etapas, preprocesamiento, modelado y diseño de la herramienta, así como la priorización y planificación de dichas iniciativas.

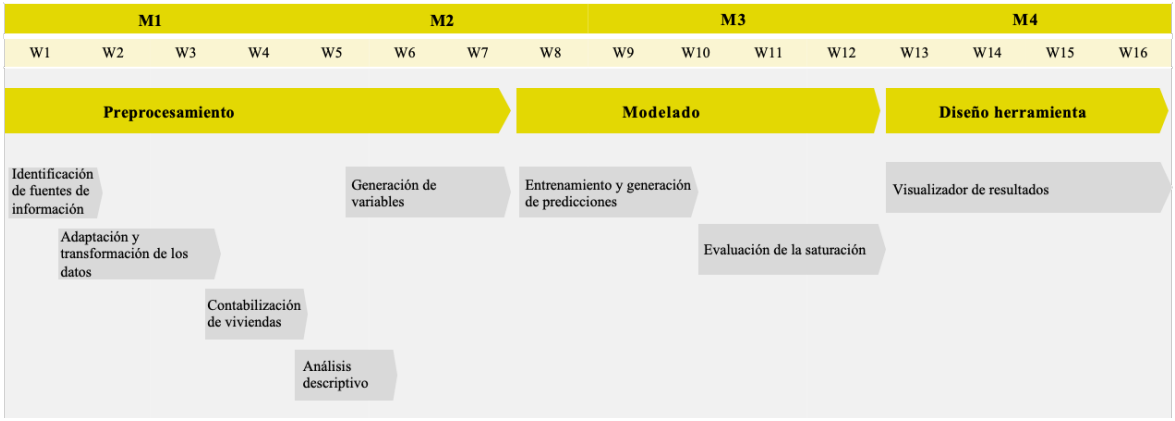


Figura 5. Diseño de la estrategia

Debido al tipo de proyecto y al tipo de contrato, es difícil estimar el costo del desarrollo. La estimación debe tener en cuenta el trabajo de un Data Scientist en prácticas, el soporte diario del tutor del proyecto, los costos de desarrollo de software y las herramientas específicas del proyecto. Además, debe tener en cuenta los costos indirectos, así como los riesgos y posibilidades que puedan surgir durante el proceso de desarrollo del proyecto.

Aun así, es importante tener en cuenta que un proyecto de este tipo en el mundo laboral debe contar con un Project Manager responsable de gestionar, coordinar y supervisar las actividades relacionadas con el proyecto, un Data Architect encargado de la construcción y mantenimiento de la infraestructura necesaria, un Data Engineer que desarrolle y mantenga los procesos de extracción, transformación y carga de datos (ETL), un Data Scientist encargado del análisis y modelado de los datos y por último un Business Analyst que se

enfoque en explotar los resultados del modelado con el desarrollo de una herramienta de visualización.

Capítulo 5. ANÁLISIS Y DESARROLLO

El capítulo actual comienza realizando un análisis inicial en el que se explican las fuentes de información y las variables utilizadas, prestando especial atención a las variables críticas del proyecto. Posteriormente, se examinan los procedimientos realizados desde este análisis inicial hasta la implementación de los modelos seleccionados.

5.1 FUENTES DE DATOS

Esta sección describe las fuentes de datos que se utilizaron durante el desarrollo del proyecto.

Variables que analizar	*Fuentes de datos*
Datos sociodemográficos de las secciones censales	Datos abiertos del Instituto Nacional de Estadística (INE). [8]
Datos obtenidos de la plataforma de AirBnB	Catálogo de datos abiertos de AirBnB. [7]

Tabla 1. Fuentes de datos

5.2 VARIABLES QUE ANALIZAR

Para el presente proyecto, se ha propuesto analizar las siguientes variables.

5.2.1 VARIABLES SOCIODEMOGRÁFICAS DE LAS SECCIONES CENSALES

Se han utilizado variables sociodemográficas para garantizar una comprensión profunda y una precisión óptima en los modelos. Es importante tener en cuenta que cada una de las variables proporciona datos sobre su sección censal:

Variables sociodemográficas	*Definición correspondiente.*
Código de sección censal	Código numérico identificador.
Área	Tamaño de la zona geográfica.
Ingresos per cápita	Cantidad promedio de ingresos que genera una persona en un año.
Media de edad	Edad promedio de las personas.
Porcentaje de edad 0-24	Porcentaje de personas que tienen entre 0 y 24 años.
Porcentaje de edad 25-39	Porcentaje de personas que tienen entre 25 y 39 años.
Porcentaje de edad 40-49	Porcentaje de personas que tienen entre 40 y 49 años.
Porcentaje de edad 50-59	Porcentaje de personas que tienen entre 50 y 59 años.
Porcentaje de edad 60-69	Porcentaje de personas que tienen entre 60 y 69 años.
Porcentaje de edad +70	Porcentaje de personas que tienen entre 70 años o más.
Porcentaje de españoles	Porcentaje de personas que tienen nacionalidad española.
Gasto en el hogar	Percentil en el que se sitúa cada sección censal con respecto a los gastos en bienes y servicios del hogar.
Población	Número total de personas que viven en la zona.
Densidad de población	Número de personas por unidad de área geográfica.
Número de farmacias	Número de farmacias que se encuentran dentro del área.

Número de colegios	Número de colegios que se encuentran dentro del área.
Número de puntos de la TTP	Número de puntos donde se puede acceder al transporte público.
Geometría	Describe la forma y la ubicación de la sección censal en el espacio geográfico.

Tabla 2. Variables sociodemográficas

Los siguientes gráficos muestran la distribución de las variables sociodemográficas más importantes utilizadas en los modelos.

5.2.1.1 Media de edad

La Figura 6 muestra una distribución con una forma aproximadamente normal y una mediana de 43.48 años al analizar la variable "Media de edad", que indica la edad promedio de los individuos de cada sección censal. Además, se puede apreciar que la mayoría de los datos se concentran en un rango estrecho alrededor de la media, lo que indica una distribución bastante homogénea.

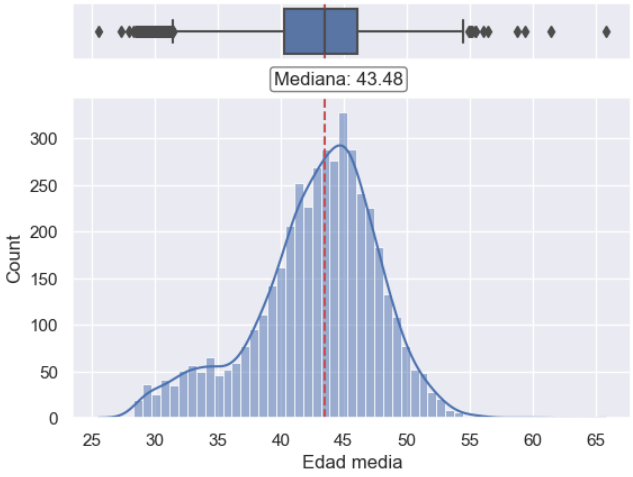


Figura 6. Distribución "Edad media"

En cuanto a la distribución geográfica, se puede observar en la Figura 7 que las secciones censales que se encuentran en las afueras de Madrid tienen una media de edad joven, lo que las convierte en las zonas preferidas para aquellas personas con una edad comprendida entre los 20 y 40 años. Por otra parte, al explorar la región central de la Comunidad de Madrid, se puede observar un incremento en la edad media de los individuos, situándose entre los 45 y la edad máxima.

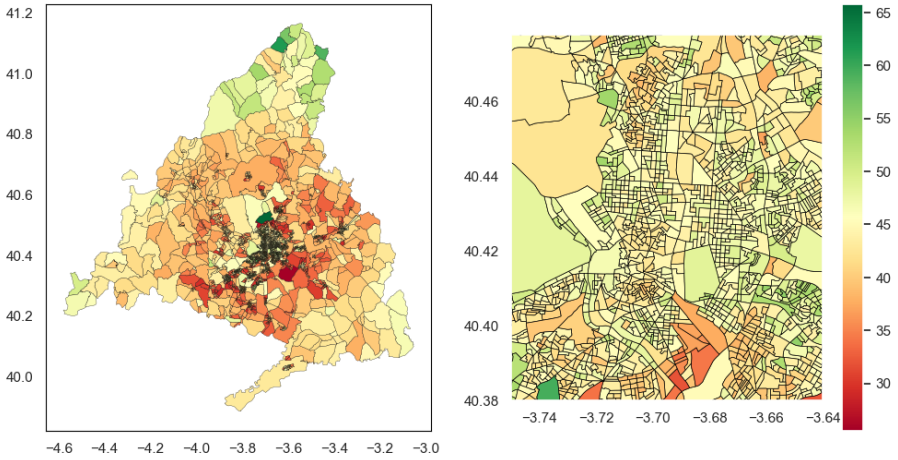


Figura 7. Distribución geográfica "Edad media"

5.2.1.2 Ingresos per cápita

La Figura 8 muestra la distribución de la variable, "Ingresos per cápita", la cual se encuentra sesgada hacia la derecha, con la mayoría de los datos concentrados en los rangos de ingresos más bajos. La gráfica muestra ciertos valores atípicos en el extremo superior, lo que sugiere que algunas personas tienen ingresos significativamente mayores que la gran parte del conjunto de datos.

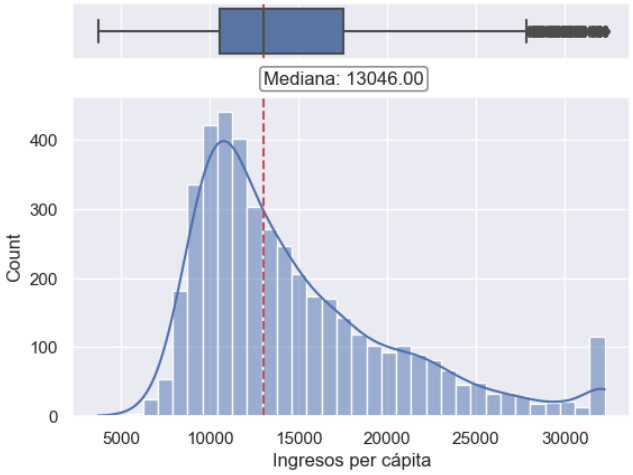


Figura 8. Distribución "Ingresos per cápita"

Al analizar la distribución geográfica de esta variable, visible en la Figura 9, se puede observar una fuerte correlación con la distribución geográfica de la edad media, en la Figura 7. En este mapa, se puede apreciar que las zonas más céntricas presentan ingresos per cápita superiores al resto, en contraste con las afueras, donde los ingresos son significativamente inferiores. En general, se puede extraer que las personas de edad avanzada tienen mayores ingresos per cápita y, por lo tanto, tienen más posibilidades de residir en el centro de la Comunidad de Madrid.

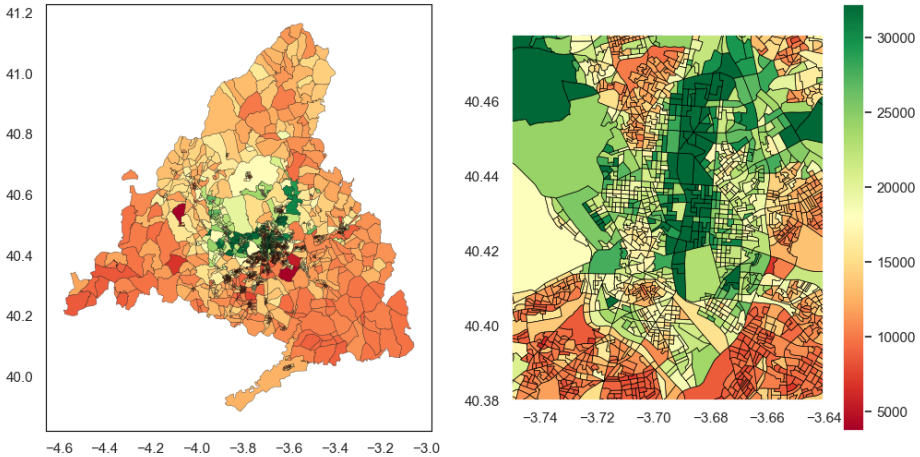


Figura 9. Distribución geográfica "Ingresos per cápita"

5.2.1.3 Porcentaje de españoles

Se muestra una distribución ligeramente sesgada a la izquierda en la Figura 10 correspondiente a la variable, "Porcentaje de españoles", donde se encuentran la mayoría de los datos concentrados en valores superiores al 80%. El valor mínimo se sitúa en torno al 50-60%, lo que puede indicar que hay algunas áreas censales con una proporción relativamente baja de españoles. Aunque el boxplot muestra algunos valores bajos atípicos, en general no hay valores extremos muy lejanos del resto de los datos.

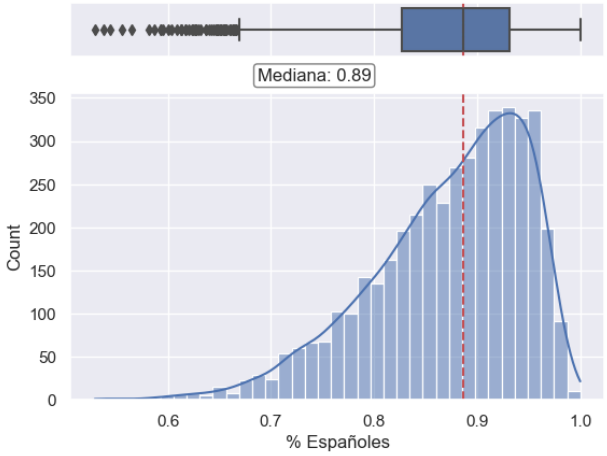


Figura 10. Distribución "Porcentaje de españoles"

Tal y como se acaba de explicar en el gráfico de distribución de la Figura 10, la distribución geográfica del porcentaje de españoles de la Figura 11 muestra que la mayoría de las secciones censales tienen un porcentaje de españoles superior al 80%. Sin embargo, en el mapa también se puede observar la presencia de varios valores atípicos bajos detectados en el boxplot, los cuales están distribuidos en pequeñas áreas del mapa.

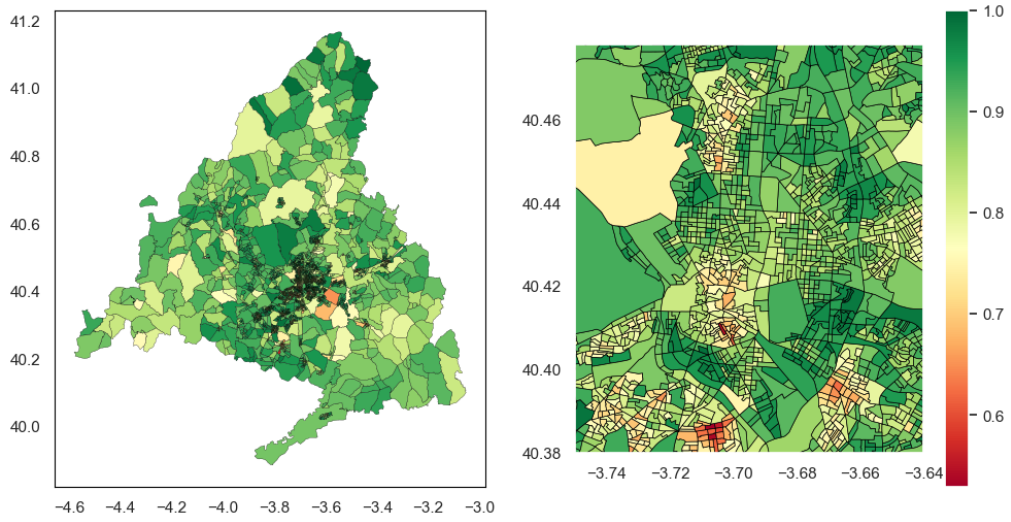


Figura 11. Distribución geográfica "Porcentaje de españoles"

5.2.1.4 Población

En la gráfica de la Figura 12, se puede observar que la variable "Población" presenta una distribución algo sesgada hacia la derecha, en la que la mayoría de los valores se encuentran en la parte inferior del rango.

En el boxplot se muestran varios valores atípicos en el lado derecho, lo que confirma de nuevo que la distribución está sesgada a la derecha. En general, hay una gran variabilidad en la población por sección censal, con algunos sectores contando con un número reducido de personas y otros con muchas más. Es importante señalar que la mayoría de las áreas censales cuentan con una población que oscila entre las 1000 y 2000 personas.

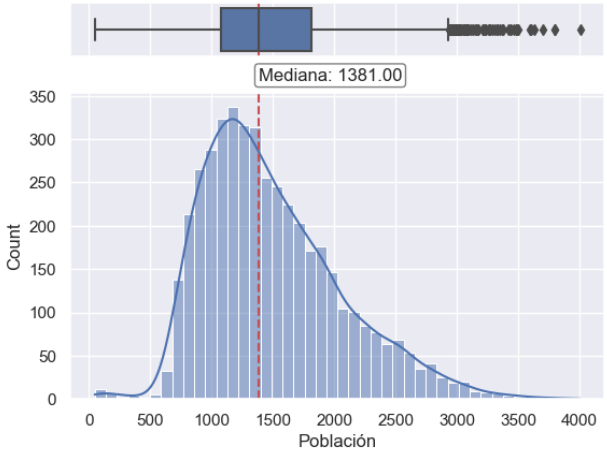


Figura 12. Distribución "Población"

En cuanto a la distribución geográfica de la población, se puede observar en la Figura 13 que en la región central hay una menor cantidad de personas por sección censal, lo cual se puede deber a que estas secciones censales tienen un tamaño mucho más pequeño que las áreas de las afueras de la Comunidad de Madrid. Asimismo, es relevante destacar que las áreas con un número muy reducido de individuos que se detectaron previamente se encuentran en la parte superior del mapa, específicamente en la zona de la sierra de Madrid.

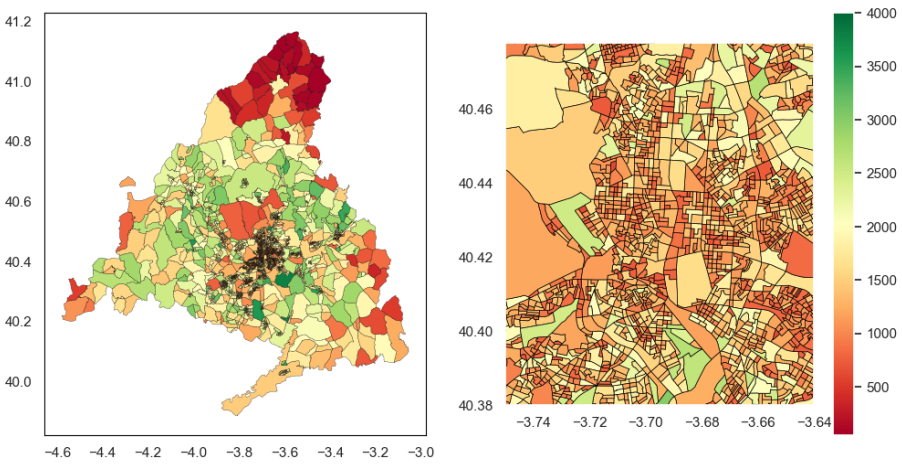


Figura 13. Distribución geográfica "Población"

5.2.1.5 Gasto en el hogar

La variable "Gasto en el hogar" representada en la Figura 14 hace referencia al percentil de la variable. Por lo tanto, lo que se observa en la gráfica es la distribución de los hogares según su posición relativa en términos de gasto en comparación con el conjunto de datos. Los percentiles más altos indican hogares que gastan más, mientras que los percentiles más bajos indican hogares que gastan menos.

La gran variabilidad alrededor del conjunto de datos puede ser resultado de una variedad de factores. Algunos hogares pueden tener ingresos altos y optar por gastar mucho dinero en cosas como la vivienda, el transporte o la comida. Estos hogares tendrán percentiles de gastos más altos. Por otro lado, aquellos hogares con ingresos más bajos pueden tener menos dinero para gastar en asuntos relacionados con el hogar. Estos hogares por lo tanto se situarán en los percentiles de gastos más bajos.

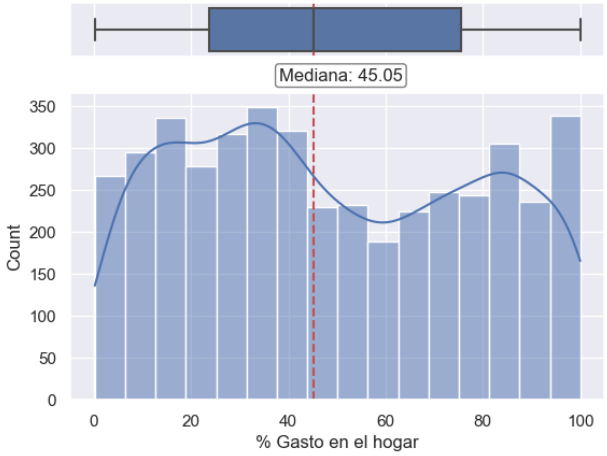


Figura 14. Distribución "Gasto en el hogar"

Se puede observar al analizar la distribución geográfica de la Figura 15 que las áreas más céntricas gastan más en este tipo de bienes. Por otro lado, el gasto en vivienda es significativamente menor en las zonas periféricas debido a los costos de vivienda más bajos.

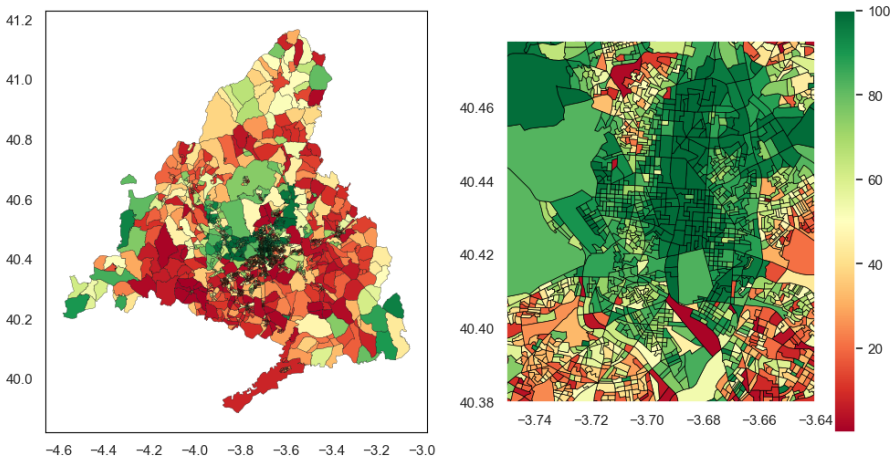


Figura 15. Distribución geográfica "Gasto en el hogar"

5.2.2 VARIABLES OBTENIDAS DE LA PLATAFORMA DE AIRBNB

En vista del objetivo del proyecto, que trata de realizar un análisis de la vivienda turística en la ciudad de Madrid, se han recogido datos correspondientes a los meses de diciembre de 2021 y diciembre de 2022 de la plataforma de alquiler AirBnB. A partir de estos datos recopilados, se han creado nuevas variables enfocadas en las secciones censales, tales como el precio medio por sección censal y la evolución del número de viviendas entre los dos períodos mencionados. Estas variables son cruciales para el proyecto y se describen con más detalle en los capítulos posteriores.

5.2.2.1 Número de viviendas en el último periodo

En la Comunidad de Madrid, se habían registrado un total de 20,764 anuncios publicados en diciembre de 2022. Debido a que esta variable es relevante para nuestros modelos, se requiere un análisis exhaustivo de su distribución. La Figura 16 muestra un mapa de colores que asigna diferentes colores según el valor de la variable a analizar para mostrar claramente dónde se encuentran los anuncios.

Este tipo de representación gráfica facilita la comprensión de la distribución de los anuncios, lo que puede ser beneficioso para nuestra investigación.

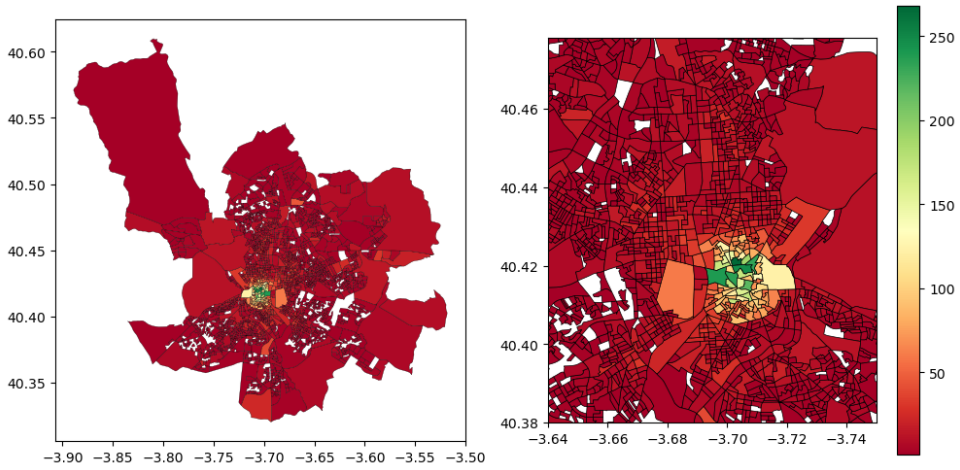


Figura 16. Mapa de la Comunidad de Madrid donde se ha representado el número total de viviendas en cada sección censal

En general, la mayoría de las áreas cuentan con menos de 50 viviendas turísticas. Sin embargo, se pueden detectar muchas secciones censales con más de 100 viviendas turísticas en la zona céntrica de la región.

La Figura 17 muestra su distribución más detallada para una mejor comprensión de esta variable. Esto nos da una visión más clara y detallada de los patrones espaciales en la distribución de las viviendas turísticas en la comunidad.

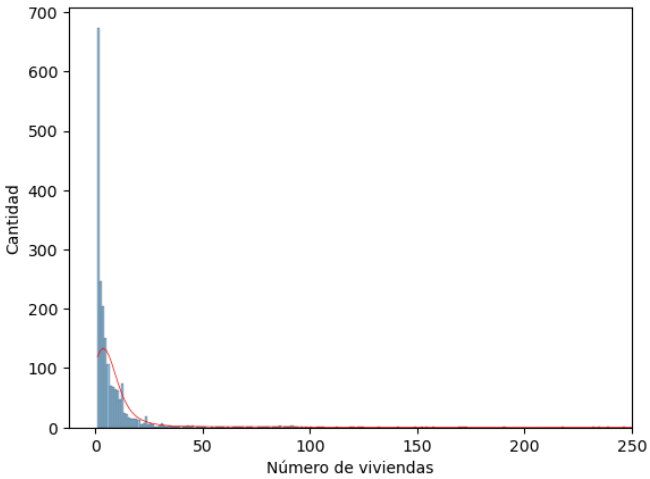


Figura 17. Distribución del número de viviendas en cada sección censal

El gráfico de distribución muestra que alrededor del 90% de las secciones censales tienen menos de 25 viviendas turísticas. A partir de ese momento, la curva comienza a aplanarse y, en algunas áreas, incluso alcanza un valor máximo de 250 viviendas.

5.2.2.2 Evolución de la vivienda turística

Se ha creado una variable que refleja la evolución de cada sección censal durante el último año utilizando datos de viviendas turísticas en dos periodos de tiempo diferentes. Al igual que con la variable anterior, se realizan representaciones visuales para comprender mejor la distribución de esta variable.

Se puede observar en la Figura 18 que las zonas muy céntricas, que son las que se encuentran más influenciadas por el turismo, han aumentado su número de viviendas de manera significativa, con un máximo de 53 viviendas. Por otro lado, la sección con la evolución más negativa experimentó una disminución de 21 viviendas. A pesar de esto, el promedio de evolución es favorable, con un aumento de 1,45 viviendas.

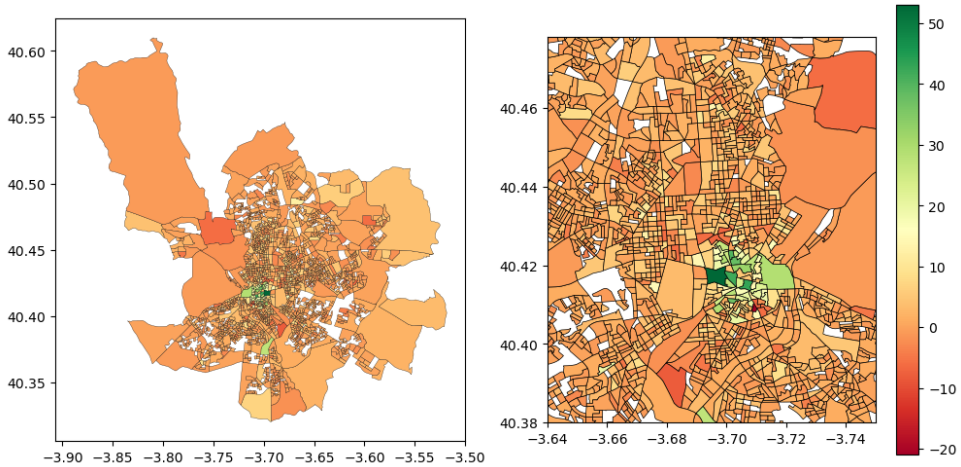


Figura 18. Mapa de la Comunidad de Madrid donde se ha representado la evolución de la vivienda turística

Con el fin de obtener una mejor comprensión de esta variable, se representa el gráfico de la Figura 19 que permite visualizar de manera más clara la distribución de la evolución de las secciones censales.

Lo más destacable es que durante el último año se ha registrado una evolución positiva en un mayor número de viviendas que aquellas que han experimentado una evolución negativa.

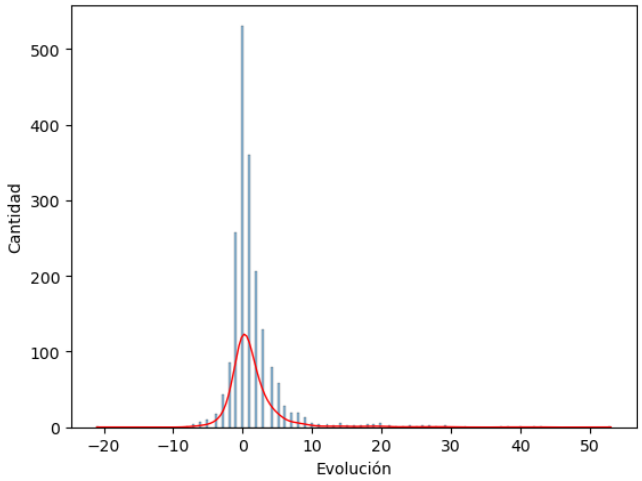


Figura 19. Distribución de la evolución de la vivienda turística

5.2.2.3 Precio medio por sección censal

Esta variable generada a raíz de los datos recopilados se utiliza como variable de entrada para los modelos para ayudar en la modelización. Esta variable muestra el precio medio por noche de las casas por cada sección censal. Es importante tener en cuenta que cuantas más viviendas tenga la sección, la media será más realista. El precio medio de una sección será el valor de esa vivienda si la sección solo tiene una.

La distribución geográfica de la variable se puede comprender mejor a través del mapa que se muestra a continuación en la Figura 20.

En el mapa se detecta que las áreas más económicas se ubican en las zonas exteriores de la Comunidad de Madrid, mientras que las áreas más costosas se ubican en la región central. El precio promedio por noche es de 126,47 euros, con un rango que oscila entre un máximo de 3550 euros y un mínimo de 9 euros.

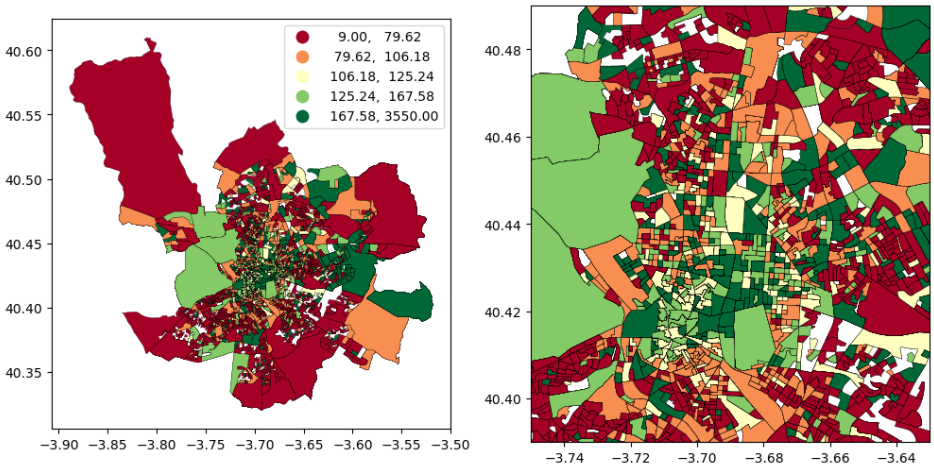


Figura 20. Mapa de la Comunidad de Madrid donde se ha representado el precio medio por sección censal

Se ha decidido utilizar de nueva el gráfico de distribución de la Figura 21 para mostrar estos datos con mayor claridad. Como se puede ver en el gráfico, alrededor del 90% de las secciones censales tienen un precio que se sitúa entre los 0 y los 250 euros. Desde ese momento, la curva se aplana y se observan precios muy altos hasta alcanzar su punto máximo.

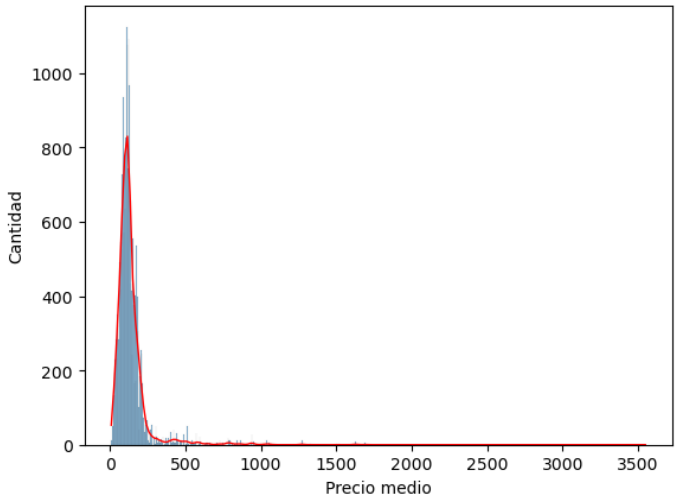


Figura 21. Distribución del precio medio por sección censal

5.3 DISEÑO

La presente sección describe el diseño utilizado para lograr los objetivos del proyecto. A continuación, se describen los pasos abordados.

5.3.1 UBICACIÓN DE LAS VIVIENDAS DENTRO DE SUS SECCIONES CENSALES

En este primer paso, se utilizan los datos de las secciones censales para asignar a cada una de las secciones censales, características extraídas del conjunto de datos de AirBnB. Esto nos permitirá obtener información detallada sobre las propiedades turísticas de cada área, lo que nos permitirá ajustar un modelo que sea capaz de estimar cuáles son las secciones donde se espera un aumento de la vivienda turística. Para lograrlo, es esencial conocer la cantidad de viviendas presentes en cada sección censal. Para obtener esta información, seguiremos los siguientes pasos.

Después de descargar e importar el archivo CSV que contiene las secciones censales de la Comunidad de Madrid, el conjunto de datos se transforma en GeoPandas para incluir la geometría de los polígonos de las secciones censales. GeoPandas es una biblioteca de Python que amplía las funcionalidades de Pandas para trabajar con datos geoespaciales, lo que la convierte en la mejor opción para este caso.

Se establece el sistema de coordenadas geográficas EPSG:4326 al realizar la transformación. Este paso es fundamental para estandarizar la proyección cartográfica de los datos en formato geoespacial y garantizar su compatibilidad con otros sistemas y aplicaciones. La elección de EPSG:4326 se debe a que es uno de los sistemas de referencia de coordenadas más utilizados a nivel mundial, lo que facilita la interoperabilidad de los datos geoespaciales. [9] Además, esta transformación permite representar los datos en un formato legible por humanos porque utiliza las unidades de longitud y latitud que la mayoría de las personas conocen.

Por otro lado, una vez importados los archivos CSV relacionados con las viviendas turísticas de AirBnB se han convertido a GeoPandas especificando que las columnas “latitud” y

"longitud" contienen información de su ubicación geoespacial. De esta manera, se ha creado en la conversión una columna con la geometría de este conjunto de datos. Esta columna contiene objetos de tipo punto que guardan información sobre la ubicación de cada vivienda turística en la Comunidad de Madrid. Es importante destacar que, al igual que en el conjunto de datos de las secciones censales, se ha configurado el sistema de coordenadas geográficas que se utilizará, que es el EPSG:4326 para garantizar la comparabilidad y precisión de los análisis posteriores. [9]

Una vez conocida la geometría de los polígonos correspondientes a las secciones censales y la ubicación geoespacial de las viviendas turísticas para cada uno de los dos periodos del análisis, se debe ubicar cada vivienda dentro de su respectiva sección. Para lograrlo, se emplea un "spatial join" que permite examinar todo el conjunto de datos de AirBnB y ubicar cada vivienda dentro de la sección censal correspondiente. [10] De esta manera, cada vivienda de los periodos de diciembre de 2021 y diciembre de 2022 se ubica dentro de su área específica.

5.3.2 MODELO EXPLICATIVO

Este apartado tiene como objetivo explicar la razón detrás de la utilización de modelos de Machine Learning en el proyecto. Los modelos de aprendizaje automático que se utilizan en este proyecto son modelos explicativos que permiten analizar y comprender lo que está sucediendo en el mercado de viviendas turísticas en la Comunidad de Madrid, a diferencia de los modelos predictivos, que se utilizan para anticipar futuros resultados.

Estos modelos explicativos ayudan a determinar las áreas que pueden estar siendo explotadas de manera excesiva o insuficiente. Esto se logra mediante el uso del error de predicción del modelo para identificar anomalías superiores o inferiores a la media, lo que permite identificar aquellas áreas que se encuentran en circunstancias extremas y requieren atención especial.

El proceso de elaboración del modelo comienza dividiendo el conjunto de datos en un 80 % para entrenamiento y un 20 % para prueba. Para mejorar el análisis, se lleva a cabo una

normalización de los datos y se eliminan aquellas variables que muestran una alta correlación.

Para el modelo, se utiliza un algoritmo de validación cruzada con 10 pliegues para evaluar su rendimiento. Este algoritmo permite obtener una estimación precisa del rendimiento del modelo y detectar posibles problemas de sobreajuste.

Una vez que se ha obtenido el modelo óptimo con los hiperparámetros ajustados, se realizan predicciones tanto para los datos de prueba como para el conjunto completo de datos. Esto permite calcular de manera sencilla la variable que representa el grado de saturación de las secciones censales. Como se muestra en la Ecuación 1, el valor del modelo y el valor real de los datos se restan para calcular la variable.

$$Saturación = Valor\ real - Valor\ del\ modelo$$

Ecuación 1. Cálculo de la variable saturación

Como se ha explicado, se utiliza la variable del grado de saturación para identificar aquellas secciones censales que presentan valores anómalos, ya sea por exceder el nivel de saturación o por estar subexplotadas. Los datos que se desvían significativamente de lo que predice el modelo indican que la sección censal está en una situación de sobreexplotación o subexplotación.

Se presentan a continuación en la Tabla 3 y en la Figura 22 dos secciones censales con características sociodemográficas similares, pero con diferentes números de viviendas para ilustrar lo que se busca en este análisis. El objetivo es demostrar que, al aplicar los modelos a estas dos secciones censales, que tienen variables de entrada similares, pero valores diferentes en la variable de salida, es sencillo detectar si una de ellas está sobreexplotada o subexplotada. De esta manera, se puede obtener una mejor comprensión de cómo funcionan las variables y cómo se relacionan entre sí, lo que ayudará a explicar las diferencias en los resultados.

Sección censal	*Precio medio*	*Edad media*	*Población*	*N.º de viviendas*
2807901042	88,22	41,87	1107	49
2807901019	87,12	45,21	1207	16

Tabla 3. Secciones censales seleccionadas para la demostración del modelo explicativo utilizado

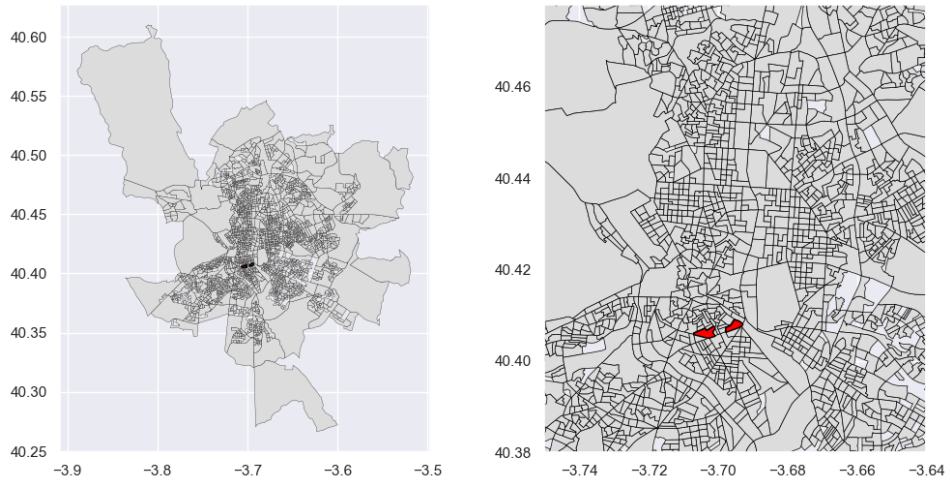


Figura 22. Ubicación de las secciones censales seleccionadas para la demostración del modelo explicativo utilizado

5.4 ALGORITMOS

El objetivo de esta sección es crear un modelo explicativo que ayude a comprender la tendencia general de los datos. A diferencia de los modelos puramente predictivos que buscan métricas precisas, en este caso se busca identificar los factores que influyen en el resultado y detectar posibles anomalías en los datos. El enfoque se centra en la comprensión profunda del problema y la interpretación de los resultados en lugar de enfocarse en la precisión de la predicción. Por lo tanto, el objetivo principal no es hacer predicciones precisas, sino comprender cómo se relacionan las variables.

Para el proceso de modelado se ha utilizado como variable de salida la generada anteriormente que representa el número total de viviendas por sección. Se ha empleado la

librería Pycaret que incluye 25 tipos diferentes de algoritmos para realizar comparaciones entre ellos, para finalmente, seleccionar dos para un análisis más exhaustivo. El primer modelo se considera el modelo ideal para resolver el problema, mientras que el segundo se encarga de aportar explicabilidad al problema.

5.4.1 CATBOOST REGRESSOR

Para resolver este problema de regresión, primero se ha implantado el modelo CatBoost Regressor, que mejora la capacidad del modelo mediante árboles de decisión y algoritmos de boosting.

En el contexto del proyecto actual, el uso de CatBoost Regressor es beneficioso debido a su capacidad para administrar varias variables de entrada y detectar interacciones entre ellas. Esto ayuda al modelo a comprender mejor los factores que influyen en el número de viviendas en cada sección censal y, por lo tanto, puede ofrecer estimaciones más precisas sobre las tendencias observadas en los datos. Además, el modelo puede ajustarse constantemente y mejorar su comprensión a medida que se agregan más datos al usar el boosting. [11] Por lo tanto, desde una perspectiva explicativa, CatBoost Regressor es una excelente opción para este problema de análisis del número de viviendas en cada sección censal.

La configuración del modelo incluye el ajuste de los siguientes hiperparámetros:

Parámetro	*Definición*	*Valor seleccionado*
Iterations	Número de iteraciones para ajustar el modelo.	1000
Learning_rate	Tasa de aprendizaje.	0,0489
Depth	Máxima profundidad de los árboles de decisión.	6

Max_leaves	Número máximo de hojas permitidas en cada árbol.	64
Boosting_type	Tipo de boosting a utilizar.	Plain
L2_leaf_reg	Regularización L2 aplicada a los pesos de los árboles.	3

Tabla 4. Configuración de los parámetros del modelo CatBoost Regressor

5.4.2 DECISION TREE REGRESSOR

El modelo de Decision Tree Regressor de la librería Scikit-Learn permite crear un árbol de decisión para representar la relación entre una variable de salida y varias variables de entrada. [12]

El motivo por el cual se ha seleccionado el modelo de árboles de decisión para este proyecto es para que el proyecto pueda tener más explicabilidad. Aunque puede ser menos preciso que otros modelos más complejos, los árboles de decisión son más fáciles de interpretar y visualizar, a diferencia de modelos como CatBoost Regressor. Estos árboles de decisión ayudan a comprender cómo se utilizan las diferentes variables de entrada, lo que permite identificar qué variables de entrada tienen un mayor impacto en la variable de salida.

En el presente modelo se han ajustado los siguientes hiperparámetros:

Parámetro	*Definición*	*Valor seleccionado*
Max_depth	Número máximo de niveles en el árbol.	4
Min_samples_leaf	Número mínimo de muestras que se requiere para ser considerado como una hoja en el árbol.	8

Min_samples_split	Número mínimo de muestras que se requieren para hacer la división de un nodo en dos subnodos.	2
Splitter	Estrategia utilizada para elegir la división en cada nodo.	Best

Tabla 5. Configuración de los parámetros del modelo Decision Tree Regressor

Capítulo 6. ANÁLISIS DE RESULTADOS

A continuación, se muestra el proceso de modelado que ha sido utilizado en el proyecto, con los resultados obtenidos y una explicación de las razones por las que se tomaron estas decisiones.

A lo largo de este proceso, se ha utilizado la librería Pycaret, que se ha ido presentando a través de las secciones anteriores. Esta librería permite utilizar una función muy útil que permite simplificar el proceso de preparación de los datos antes de entrenar modelos. Esta función se ha utilizado para dividir el conjunto de datos en 80 % para entrenamiento y 20 % para prueba. Como resultado, el conjunto de entrenamiento cuenta con 3089 registros y el conjunto de prueba con 1325 registros. Además, se ha utilizado un parámetro para eliminar automáticamente variables con una correlación muy alta, y se ha aplicado una normalización a los datos mediante el método "robust", que es menos sensible a valores atípicos en los datos.

A continuación, se utiliza una función de Pycaret que utiliza una variedad de métricas de evaluación comunes para comparar el rendimiento de diferentes modelos de aprendizaje automático. Esta función toma como entrada el conjunto de datos de entrenamiento y mediante la validación cruzada k-fold, evalúa cada modelo de forma automática. El resultado se presenta en la siguiente Tabla 6 mostrando las métricas de rendimiento correspondientes a los 10 mejores modelos y el Decision Tree Regressor.

Modelo	*MAE*	*MSE*	*RMSE*	*R2*
CatBoost Regressor	**2,38**	**56,46**	**7,28**	**0,78**
Extra Trees Regressor	2,50	63,47	7,73	0,76
Extreme Gradient Boosting	2,58	65,52	7,78	0,73

Gradient Boosting Regressor	2,76	69,29	8,06	0,72
Random Forest Regressor	2,54	74,53	8,44	0,69
Light Gradient Boosting Machine	2,81	79,44	8,56	0,69
K Neighbors Regressor	2,77	99,20	9,54	0,65
AdaBoost Regressor	7,94	123,98	11,00	0,44
Linear Regression	6,44	186,36	13,22	0,32
Ridge Regression	6,43	186,39	13,22	0,32
Decision Tree Regressor	3,54	165,75	12,62	0,30

Tabla 6. Comparación de modelos utilizando PyCaret

Tras analizar la Tabla 6, se puede observar que el modelo Catboost Regressor ofrece excelentes resultados y, por lo tanto, ha sido elegido como el modelo óptimo para resolver el problema en cuestión. Es importante destacar que, a pesar de su bajo rendimiento, el modelo Decision Tree Regressor puede ser útil para aportar la explicabilidad y transparencia que faltan al Catboost. Por estas razones, se ha decidido utilizar ambos modelos.

Una vez que se han seleccionado los modelos a desarrollar en profundidad, se han intentado mejorar los resultados utilizando una técnica de búsqueda de hiperparámetros conocida como Random Grid Search. Esta técnica permite examinar una variedad de combinaciones de hiperparámetros para encontrar los mejores.

Sin embargo, a pesar de ajustar los hiperparámetros del modelo Catboost Regressor, la Tabla 7 muestra que el rendimiento del modelo no ha logrado mejorar. Esto podría ser resultado del efecto de la validación cruzada, ya que esta introduce cierta variabilidad en el proceso al

seleccionar de forma aleatoria diferentes subconjuntos de datos para entrenar y validar el modelo en cada iteración. Es decir, aunque se ajusten los hiperparámetros para una iteración en particular, es posible que no sean los mejores para otras iteraciones, lo que puede afectar el rendimiento general del modelo.

Modelo	*Datos*	*MAE*	*MSE*	*RMSE*	*R2*
CatBoost Regressor	Train sin hiperparámetros	2,38	56,46	7,28	0,78
	Train con hiperparámetros	2,65	61,63	7,64	0,75
	Test sin hiperparámetros	2,48	47,52	6,89	0,79

Tabla 7. Resultados CatBoost Regressor

Respecto al ajuste de hiperparámetros en el Decision Tree Regressor, sí que se ha logrado una mejor significativa en el rendimiento del modelo al aplicar directamente el modelo con Scikit-Learn en lugar de hacerlo con PyCaret. Se muestran los resultados y las diferencias entre los modelos en la siguiente Tabla 8.

Modelo	*Datos*	*MAE*	*MSE*	*RMSE*	*R2*
Decision Tree PyCaret	Train	3,54	165,75	12,62	0,30
Decision Tree Scikit-Learn	Train	2,73	84,12	9,17	0,69
	Test	2,50	59,39	7,71	0,74

Tabla 8. Resultados Decision Tree Regressor

Para poder comprender de manera clara como ha funcionado la toma de decisiones del modelo, se utiliza el Decision Tree Regressor.

Estos árboles de decisión brindan una representación visual y explícita de cómo el modelo llega a sus decisiones, por ello son útiles para comprender la toma de decisiones del modelo. Es decir, se puede seguir el flujo de decisiones y ver qué características del conjunto de datos son las más importantes para el modelo a través del árbol de decisión.

En este caso concreto, se puede ver en la Figura 23, que el modelo ha selecciona la variable de "Porcentaje de edad 0-24" para dividir los datos en dos ramas de decisión y luego ha seleccionado "Porcentaje de edad +70" y "Precio medio" para crear nuevas ramas de decisión. Lo que además indica que las tres variables son cruciales para el modelo.

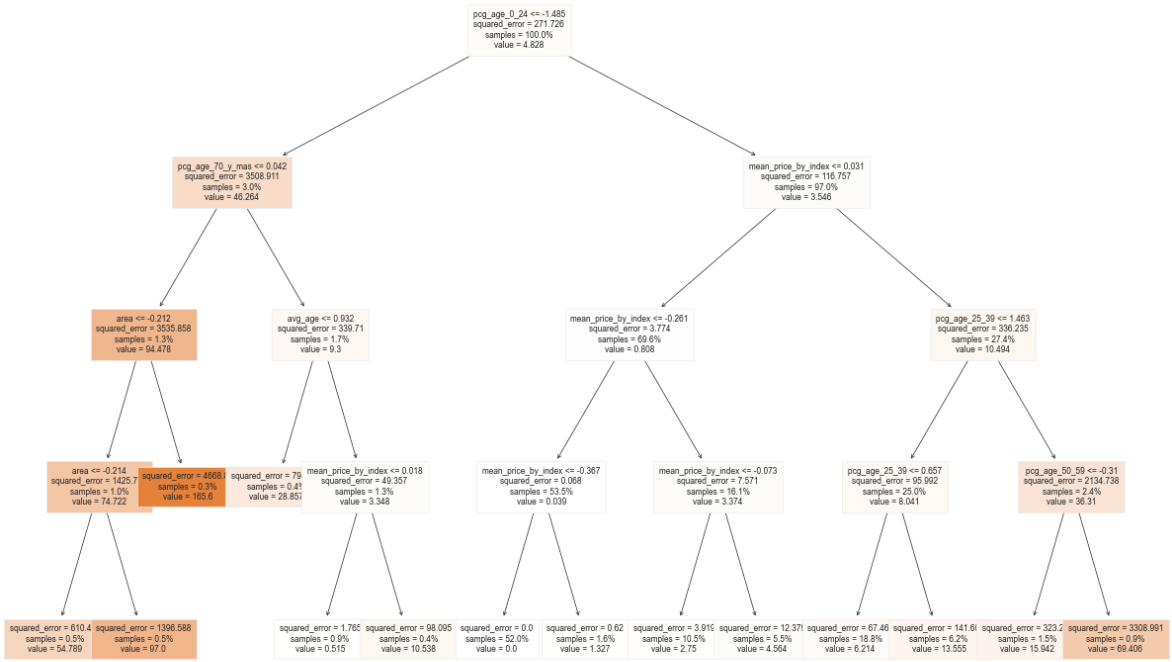


Figura 23. Representación del árbol del Decision Tree

Por otro lado, el encargado de generar la variable que mide el grado de saturación de las secciones censales es el CatBoost Regressor. Después de observar los resultados anteriores, se decidió utilizar el Catboost Regressor sin los parámetros ajustados para realizar la predicción sobre todo el conjunto de datos, como se muestra en la Tabla 9.

Modelo	*MAE*	*MSE*	*RMSE*	*R2*
CatBoost Regressor sin hiperparámetros	1.19	13.38	3.66	0.94

Tabla 9. Resultados del CatBoost Regressor sobre el conjunto total de los datos

La variable objetivo se obtiene calculando la diferencia entre el valor real y las predicciones del modelo. De esta manera, el grado de saturación aumenta con el valor de la variable obtenida, lo que indica que la sección está sobreexplotada. De manera similar, si el grado de saturación es menor, se puede inferir que la sección está siendo poco explotada.

Para comprender mejor la operación descrita, se muestra el gráfico de la Figura 24 que explica el error de predicción del modelo. El gráfico representado muestra la capacidad del modelo para predecir los valores reales. La línea "best fit", es la línea de regresión que se ajusta a los puntos en el gráfico y representa la tendencia general de los errores de predicción del modelo. Los puntos dispersos en el gráfico son cruciales para poder identificar aquellas secciones censales que podrían estar en una situación de sobresaturación o subexplotación.

Para lograrlo, se calcula la distancia vertical entre la línea "best fit" y el punto correspondiente en el gráfico, esta distancia representa el error de predicción, que es la diferencia entre el valor real de la variable y la variable objetivo que el modelo ha calculado. Entonces al calcular el error de predicción para cada punto, se obtiene una medida del grado de precisión del modelo. Esta medida se utilizará para medir la saturación y descubrir áreas potenciales de mejora.

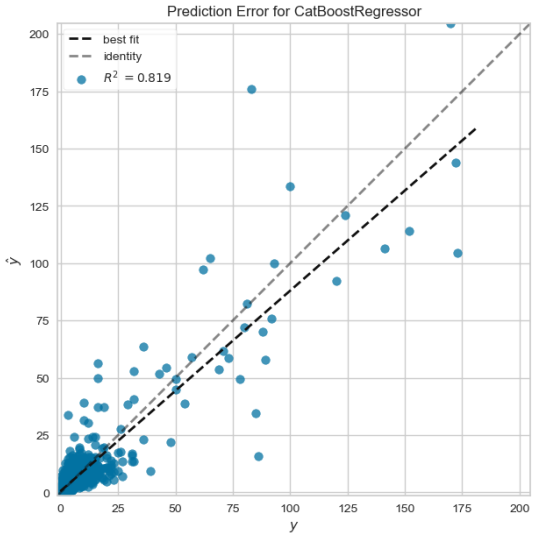


Figura 24. Gráfico del error de predicción del Catboost Regressor

Después de calcular la variable objetivo, se usa la Figura 25 para analizar su distribución. En base a los datos, las secciones censales presentan una estabilidad ligeramente negativa, con un valor medio de -0,0155.

Sin embargo, es importante destacar que la desviación estándar de la variable es bastante alta, con un valor de 3.66, lo que indica una gran variabilidad en los datos y que algunos valores están muy por alejados del valor medio. Además, se encontraron valores máximos de hasta 70.16 y valores mínimos de hasta -93.08.

En resumen, la variable muestra una gran variabilidad. Algunos de sus valores muestran una posible sobresaturación o subexplotación significativa en ciertas secciones censales, mientras que otros valores muestran una saturación moderada o incluso baja en otras secciones censales.

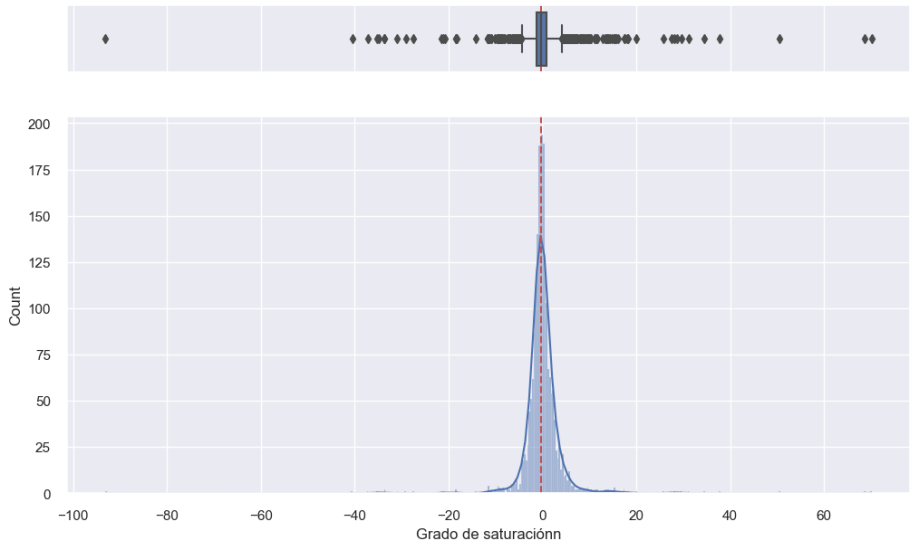


Figura 25. Distribución de la variable calculada que mide el grado de saturación

Al analizar la distribución geográfica de la Figura 26, se puede observar que la mayoría de las secciones censales presentan una estabilidad moderada o incluso ligeramente positiva. Sin embargo, es sencillo identificar áreas con valores extremos, tanto positivos como negativos, que indican una sobresaturación o subexplotación significativa en esas áreas.

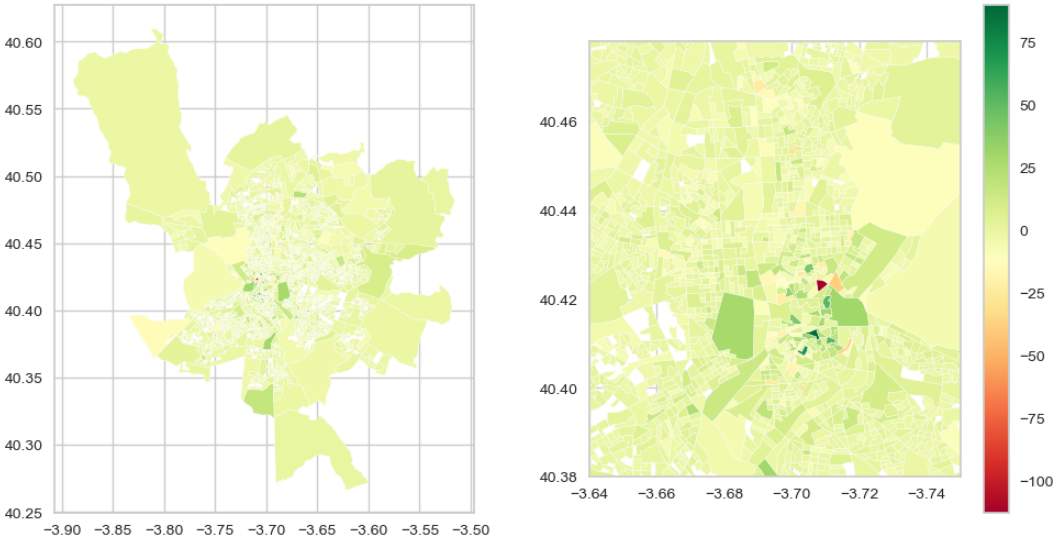


Figura 26. Distribución geográfica de la variable calculada que mide el grado de saturación

Por otro lado, como se ha ido mencionando en distintos capítulos, se han utilizado los valores shap para comprender la importancia de las variables en el modelo y determinar qué componentes tienen un impacto positivo o negativo en las viviendas turísticas.

Según la Figura 27, la variable con mayor influencia en el modelo es el "porcentaje de edad 0-24" (pcg_age_0_24). Esta tiene un impacto muy positivo en la variable de salida cuando los valores de la variable son bajos. Esto significa que cuando una sección censal tiene un porcentaje bajo de personas de 0 y 24 años, hay más viviendas turísticas. La siguiente variable, el "precio medio por sección censal" (mean_price_by_index), indica que la cantidad de viviendas turísticas en la zona aumenta con el precio medio de la sección. Luego está el "porcentaje de españoles" (spanish), donde la cantidad de viviendas turísticas es mayor cuanto menor es el porcentaje de españoles en la sección.

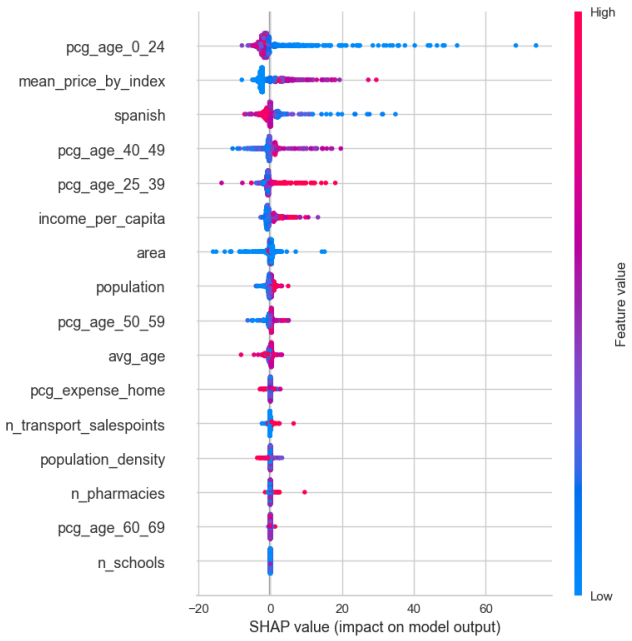


Figura 27. Valores shap de las variables de entrada del modelo

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

Para maximizar el valor total del proyecto, este capítulo de conclusiones utilizará un dashboard interactivo creado en Power BI. Este dashboard permite una visualización clara y concisa de los resultados obtenidos, lo que ayuda a comprender mejor los datos. Además, se pueden realizar análisis más detallados y personalizados gracias a la interacción que ofrece esta herramienta.

El dashboard está compuesto por dos páginas: la primera analiza la relación entre la saturación y sus características sociodemográficas, y la segunda se centra en permitir el ajuste de parámetros para un análisis más detallado e interactivo. Se facilita la identificación de patrones y oportunidades en esta segunda opción, lo que permite una toma de decisiones más informada.

7.1 SATURACIÓN DE LAS SECCIONES CENSALES

El objetivo de esta página del dashboard es realizar un estudio de la saturación de las secciones censales y su relación con las características sociodemográficas.

Como resultado, la Figura 28 muestra un mapa interactivo que tiene cuatro indicadores en la parte superior. La escala de colores utilizada en el mapa va del rojo al verde, donde el rojo indica una alta saturación (grado de saturación positivo) y el verde indica una baja saturación (grado de saturación negativo). A su vez, un gráfico circular que muestra el porcentaje del grupo de edad según la sección y tres filtros ajustables se encuentran en la zona central del dashboard. Finalmente, dos gráficos de barras describen las distribuciones de los ingresos per cápita y los gastos por hogar en la parte derecha del dashboard, donde la recta de color rojo representa la media de la variable representada.

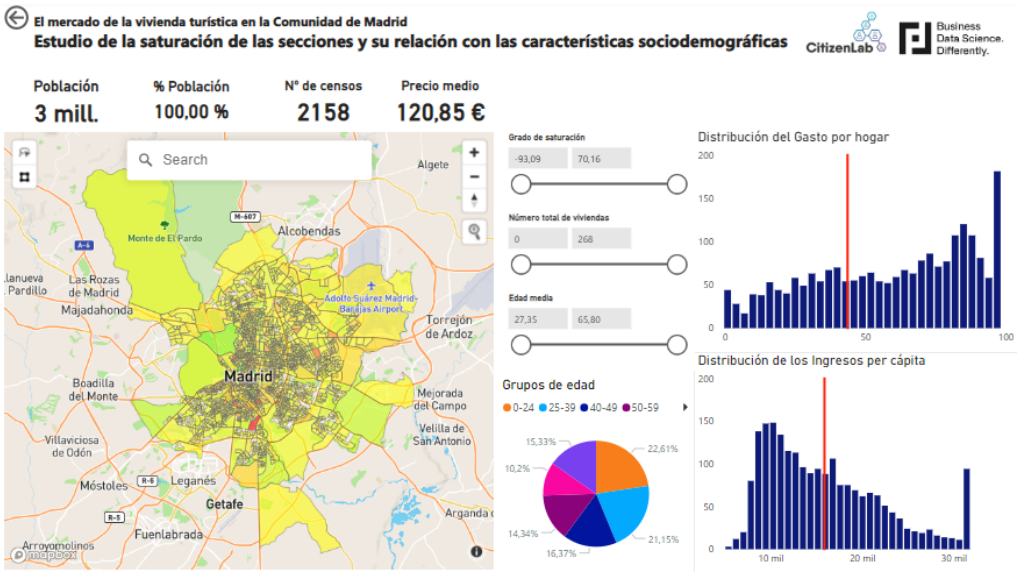


Figura 28. Hoja 1 del dashboard de Power BI

En la Figura 29, al ajustar el filtro del grado de saturación a valores negativos, se han encontrado 634 secciones censales poco explotadas, que representan el 30% de la población y tienen un precio medio por sección ligeramente más alto que el promedio. Estas secciones se encuentran por toda la comunidad y no tienen una ubicación clara.

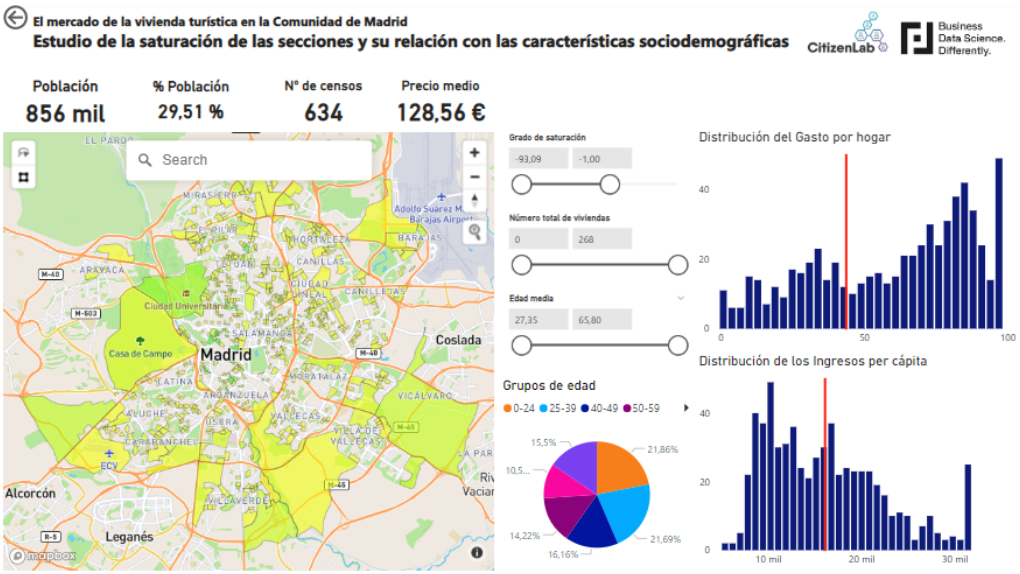


Figura 29. Hoja 1 del dashboard de Power BI representando las secciones censales poco explotadas

Al reducir aún más el grado de saturación de la variable, se obtienen en la Figura 30 las 10 secciones censales menos saturadas de la Comunidad de Madrid, ubicadas en la zona central y con el percentil de gasto por hogar más alto. Sin embargo, los ingresos per cápita varían significativamente en toda la zona. Es importante destacar que estas diez secciones tienen una media de edad más baja en el rango de 0 a 24 y una media de edad más alta en el rango de 25 a 39 en comparación con las edades medias del conjunto total de datos.

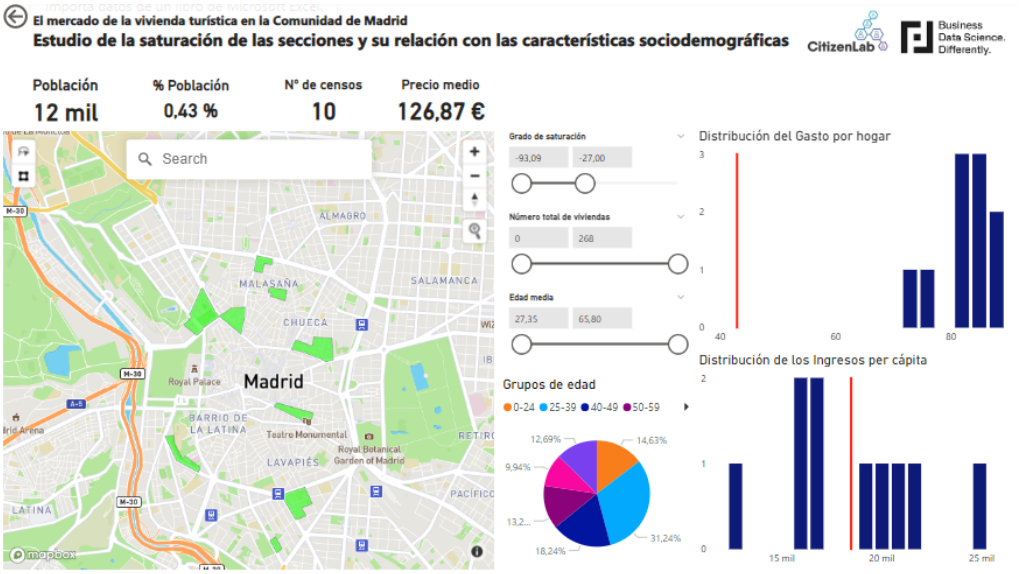


Figura 30. Hoja 1 del dashboard de Power BI representando las 10 secciones censales menos explotadas

Por otro lado, ajustando el grado de saturación a valores positivos, se pueden observar en la Figura 31, la presencia de 523 secciones censales sobreexplotadas, lo que equivale al 25% de la población.

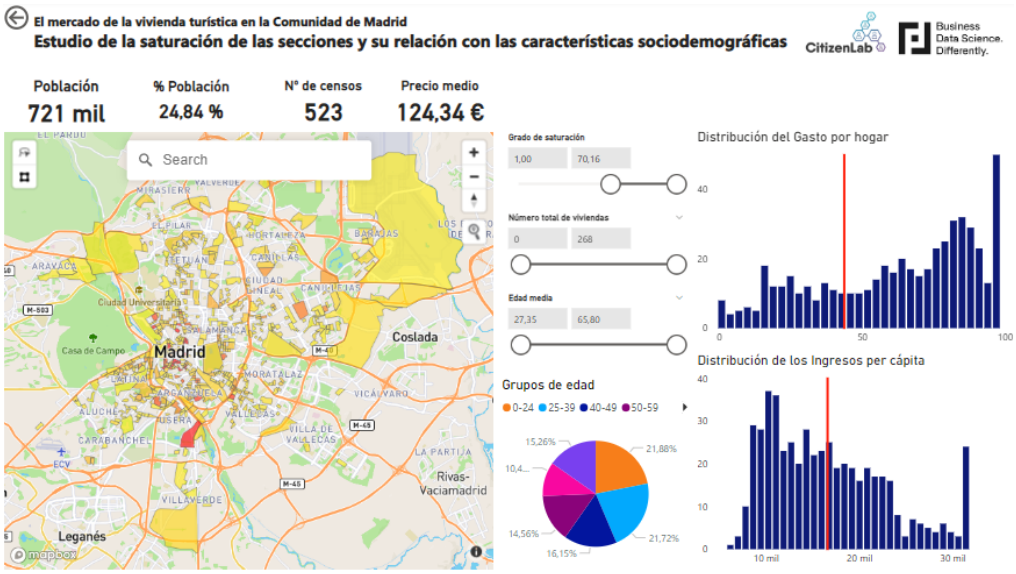


Figura 31. Hoja 1 del dashboard de Power BI representando las secciones censales explotadas

Al aumentar aún más la saturación se logra detectar en la Figura 32 las 10 secciones con una mayor sobreexplotación de la vivienda turística. Aunque la distribución es similar a la de la Figura 30, el precio promedio por área censal ha aumentado significativamente. Respecto a las edades, es destacable un aumento del 10% en el rango de 25 a 39 años, mientras que el rango de 0 a 24 años ha experimentado una disminución del 4%.

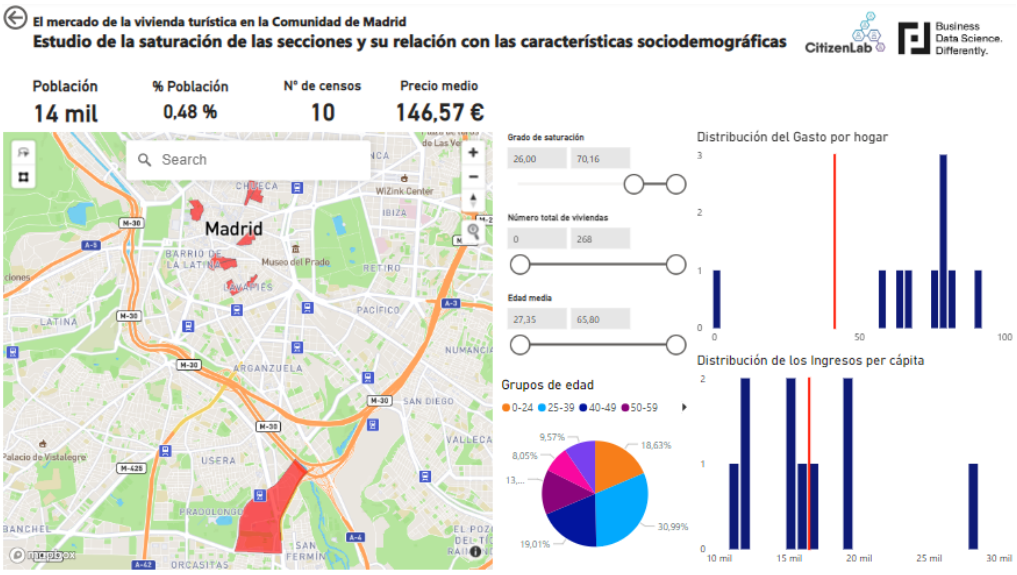


Figura 32. Hoja 1 del dashboard de Power BI representando las 10 secciones censales más explotadas

7.2 EVOLUCIÓN DE LAS SECCIONES CENSALES

En este caso, la página del dashboard tiene como objetivo permitir un análisis más profundo y detallado de los datos para identificar tendencias y oportunidades en el mercado inmobiliario con mayor seguridad.

La Figura 33 muestra el dashboard desarrollado para lograrlo. El dashboard cuenta con un mapa interactivo con cuatro indicadores en su parte superior. En este caso, la escala de colores del mapa representa la evolución de las secciones censales, con las secciones más verdes indicando una evolución positiva y las secciones más rojas indicando una evolución negativa. Además, el dashboard cuenta con cinco filtros ajustables en la zona central que permiten buscar las áreas deseadas con mayor precisión. También se presentan dos gráficos de dispersión en la parte derecha que indican la evolución frente al grado de saturación y al precio medio.

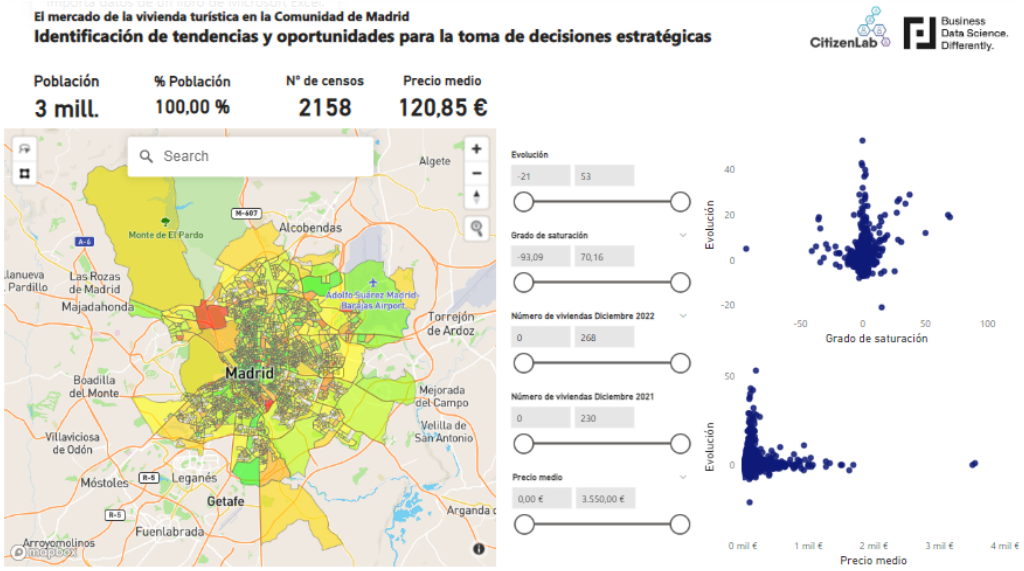


Figura 33. Hoja 2 del dashboard de Power BI

Al permitir una mayor interacción, se han ajustado los filtros para la Figura 34 para detectar aquellas secciones con una evolución positiva y un grado de saturación bajo.

Tras ajustar los filtros, se han detectado 282 secciones censales que representan el 13% de la población a analizar. Las zonas mencionadas tienen un precio promedio ligeramente inferior al promedio general y se pueden encontrar puntos extremos en sus distribuciones, los cuales serán examinados a continuación.

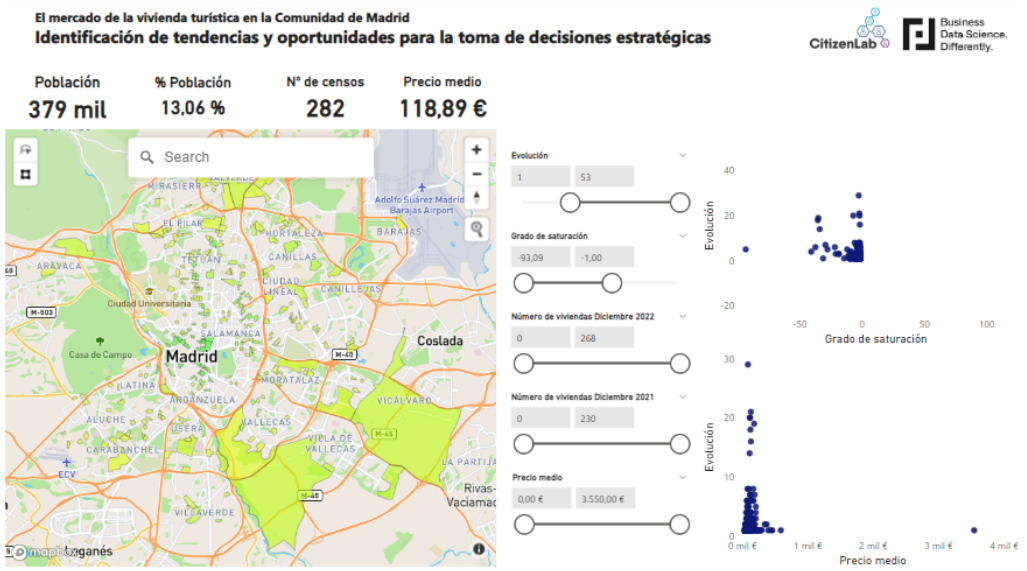


Figura 34. Hoja 2 del dashboard de Power BI representando todas las secciones censales que presentan evolución positiva y poca saturación

Lo que es realmente interesante para el proyecto es poder identificar aquellas secciones censales que tienen una saturación baja y una tendencia positiva en su evolución. Un buen ejemplo se puede observar en el gráfico de dispersión de la Figura 34, donde una de las muestras parece tener una evolución levemente ascendente y un valor mínimo de saturación.

Al seleccionar ese punto en el gráfico, se accede automáticamente a la zona deseada permitiendo explorar la sección censal en cuestión con mayor detalle, Figura 35.

La sección censal ubicada en plena Gran Vía, esquina con Plaza de España, tiene un precio promedio mucho más alto que el promedio, lo que demuestra su gran atractivo turístico. Dado que su nivel de saturación indica que se encuentra escasamente explotada (-93,08) y ha experimentado una evolución positiva (+5) en el último año, es interesante profundizar en el análisis de esta sección censal. La zona sigue en proceso de evolución y podría alcanzar

valores aún más altos en el futuro, ya que pasó de 78 viviendas en diciembre de 2021 a 83 en diciembre de 2022.

Figura 35. Hoja 2 del dashboard de Power BI representando la sección censal con menos saturación y con una evolución en tendencia positiva

Una vez examinadas las zonas que muestran una tendencia positiva, se han modificado de nuevo los filtros del dashboard para explorar ahora las secciones censales que presentan una sobreexplotación y una evolución negativa, como se muestra en la Figura 36.

En esta situación se han encontrado 82 secciones censales con precios medios ligeramente por encima del promedio total. Se puede observar un amplio rango de precios a lo largo del eje en los gráficos de dispersión, mientras que, al examinar la relación entre la evolución y el grado de saturación, la mayoría de los puntos están muy concentrados.

Figura 36. Hoja 2 del dashboard de Power BI representando todas las secciones censales que presentan evolución negativa y sobreexplotación

Puesto que previamente se ha analizado la sección censal que cuenta con una mayor tendencia positiva, a continuación, se va a examinar la zona con la peor tendencia. En ambos gráficos de dispersión, esta vivienda se encuentra en los puntos que se encuentran fuera de la media, con una disminución de 21 viviendas.

La zona mencionada se ha representado en la Figura 37 y esta se encuentra en el vecindario de Embajadores, ubicado entre las estaciones de metro de La Latina y Puerta de Toledo. A pesar de que su precio medio es significativamente inferior a la media, lo más notable es su evolución, que muestra una disminución de 21 viviendas en un año, pasando de 90 en diciembre de 2021 a 69 en diciembre de 2022. Además, el grado de saturación de +15,5 indica una sobreexplotación de la región.

Figura 37. Hoja 2 del dashboard de Power BI representando la sección censal con la peor tendencia de la Comunidad de Madrid

7.3 TRABAJOS FUTUROS

A pesar de los avances logrados en este proyecto, todavía existen varias áreas de investigación y desarrollo que pueden mejorar y ampliar sus resultados. A continuación, se describen las principales rutas con las que se podría continuar el proyecto.

1. Limitaciones en los datos: Inside AirBnB, la plataforma que se ha utilizado para recoger la información representa la mayor parte de la vivienda turística, pero no la totalidad. Con lo cual, sería un gran avance poder contar con el mercado turístico al completo recogiendo datos de todas las plataformas "Peer to Peer".
2. Datos en tiempo real: Actualmente, el proyecto utiliza datos recopilados hasta una fecha concreta. Sin embargo, sería muy beneficioso investigar la posibilidad de incorporar datos en tiempo real para obtener información actualizada sobre el mercado inmobiliario de la vivienda turística.
3. Análisis de precios: Se podría realizar un modelo de análisis del precio de la vivienda turística, considerando variables como la ubicación, las características específicas de cada propiedad y la demanda turística en cada zona. Además, se podría identificar la

correlación entre el precio de la vivienda y los costes de bienes básicos. Esto podría complementar el proyecto realizado aportando más información a las partes interesadas para facilitar las decisiones.

4. Herramienta web: Finalmente, se propone crear una herramienta en línea que elimine la dependencia de Power BI y sea completamente automatizada. La elección de esta opción garantiza que la herramienta esté disponible en línea sin necesidad de intervención humana para que funcione correctamente.

Capítulo 8. BIBLIOGRAFÍA

[1] Equipo de desarrollo de PyCaret. Documentación de PyCaret 3.0. (2023). https://pycaret.gitbook.io/docs/

[2] Equipo de Desarrollo de SHAP. Documentación de SHAP. (2023). https://shap.readthedocs.io/en/latest/index.html

[3] Luvsandorj, Z. "Explaining Scikit-Learn Models with SHAP". En: Towards Data Science. (2021). https://towardsdatascience.com/explaining-scikit-learn-models-with-shap-61daff21b12a

[4] Perles-Ribes, JF, Ramón-Rodríguez, AB, Moreno-Izquierdo, L, Such-Devesa, MJ. Machine learning techniques as a tool for predicting overtourism: The case of Spain. Int J Tourism Res. (2020). 22: 825– 838. https://doi.org/10.1002/jtr.2383

[5] Comunidad de Madrid. ANÁLISIS DEL IMPACTO DE LAS VIVIENDAS DE USO TURÍSTICO EN EL DISTRITO CENTRO. (2017). .https://www.madrid.es/UnidadesDescentralizadas/UDCMedios/noticias/2017/05Mayo/05viernes/Notasprensa/ficheros/Informe_final_5_mayo%20vivendas%20uso%20turístico%20(1).pdf

[6] Casado Buesa, M. P. El impacto del turismo en el acceso a la vivienda: el análisis de los barrios de Barcelona. (2017). [Tesis de maestría]. (Universitat Politècnica de Catalunya).

[7] Inside AirBnB - Get the Data. (s.f.). http://insideairbnb.com/get-the-data/

[8] Instituto Nacional de Estadística (INE). (s.f.). https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm

[9] Spatial Reference. (s.f.). https://spatialreference.org/ref/epsg/wgs-84/

[10] Equipo de Desarrollo de GeoPandas. Documentación de Spatial join. (2023). https://geopandas.org/en/stable/gallery/spatial_joins.html

[11] Equipo de Desarrollo de CatBoost. Documentación de CatBoost. (2023). https://catboost.ai/en/docs/concepts/python-reference_catboostregressor

[12] Equipo de Desarrollo de Scikit-Learn. Documentación de Decision Tree Regressor. (2023). https://scikit-learn.org/stable/modules/tree.html#regression

ANEXO I

Las Naciones Unidas han establecido una serie de Objetivos de Desarrollo Sostenible (ODS), y el presente proyecto ayudará a lograrlos. Este proyecto logra a través de la evaluación de los problemas de sobreexplotación o subexplotación de la vivienda turística, identificar a las áreas con potencial de desarrollo y evitar una expansión descontrolada, lo que contribuye al desarrollo sostenible de las ciudades y comunidades (ODS 11). Además, el proyecto impulsa la innovación y la aplicación de nuevas tecnologías en el sector al utilizar técnicas de aprendizaje automático para su desarrollo (ODS 9). Por último, la toma de decisiones más informadas e inteligentes pueden conducir a un crecimiento económico más sostenible y responsable (ODS 8).