UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

GRADO EN MÁSTER DATA SCIENCE

PROYECTO FIN DE GRADO

PREDICCIÓN INTELIGENTE DE PRECIOS EN AIRBNB MADRID

KATHERINE LÓPEZ RAMÍREZ

Dirigido por la

DIRECTORA: CARMEN ALONSO

CURSO 2024-2025

TÍTULO: PREDICCIÓN INTELIGENTE DE PRECIOS AIRBNB MADRID

AUTOR: KATHERINE LÓPEZ RAMÍREZ

TITULACIÓN: MÁSTER DATA SCIENCE

DIRECTORA DEL PROYECTO: INGENIERA CARMEN ALONSO

FECHA: SEPTIEMBRE 2025

RESUMEN

El crecimiento sostenido del turismo en Madrid y la consolidación de plataformas como Airbnb han generado un mercado altamente competitivo, donde la fijación de precios se convierte en un desafío complejo. Los anfitriones deben equilibrar factores como la ubicación, las características del alojamiento y la estacionalidad, mientras que turistas e inversores requieren referencias fiables para tomar decisiones informadas.

El presente proyecto aborda este problema mediante el desarrollo de un modelo predictivo basado en técnicas de aprendizaje automático. Utilizando datos públicos de Inside Airbnb, que incluyen características de los anuncios y precios diarios, se realizó un proceso completo de limpieza, análisis exploratorio y eliminación de valores atípicos, garantizando la calidad de la información. Posteriormente, se entrenaron y compararon distintos algoritmos (CatBoost, XGBoost y Random Forest), siendo este último el que ofreció los mejores resultados.

El modelo final de Random Forest optimizado alcanzó un rendimiento sobresaliente (**MAE ≈ 6,34 €, RMSE ≈ 8,75 €, R^2 ≈ 0,975**), lo que demuestra su capacidad para realizar predicciones altamente precisas y robustas. Además, se desarrolló una interfaz de usuario sencilla e intuitiva que permite realizar consultas prácticas, facilitando la interpretación de los resultados.

En conclusión, el proyecto aporta una solución útil y escalable que contribuye a optimizar la gestión de precios en el sector turístico, beneficiando a anfitriones, turistas e inversores.

Palabras clave: Airbnb, predicción de precios, aprendizaje automático, Random Forest, estacionalidad, Data Science

ABSTRACT

The sustained growth of tourism in Madrid and the consolidation of platforms such as Airbnb have generated a highly competitive market, where price setting becomes a complex challenge. Hosts must balance factors such as location, accommodation characteristics, and seasonality, while tourists and investors require reliable references to make informed decisions.

This project addresses this problem through the development of a predictive model based on machine learning techniques. Using public data from Inside Airbnb, which includes listing characteristics and daily prices, a complete process of data cleaning, exploratory analysis, and outlier removal was carried out to ensure data quality. Subsequently, different algorithms (CatBoost, XGBoost, and Random Forest) were trained and compared, with the latter achieving the best results.

The final optimized Random Forest model achieved outstanding performance (**MAE ≈ €6.34, RMSE ≈ €8.75, R^2 ≈ 0.975**), demonstrating its ability to generate highly accurate and robust predictions. In addition, a simple and intuitive user interface was developed to allow practical queries, facilitating the interpretation of results.

In conclusion, the project provides a useful and scalable solution that contributes to optimizing price management in the tourism sector, benefiting hosts, tourists, and investors alike.

Keywords: Airbnb, price prediction, machine learning, Random Forest, seasonality, Data Science

AGRADECIMIENTOS

Deseo expresar mi más sincero agradecimiento a la ingeniera Carmen Alonso, quien dedicó parte de su tiempo extra para acompañarme en este proceso. Su constante apoyo, cercanía y diligencia han sido fundamentales para culminar este Trabajo de Fin de Máster con éxito. Más allá de la orientación académica, valoro profundamente la energía positiva, el compromiso y la dedicación con los que me ha acompañado en cada etapa.

Agradezco también a mi director de programa Álvaro Sánchez, por su comprensión y paciencia, así como por las facilidades otorgadas en los plazos para la entrega de este trabajo. Su apoyo ha sido clave para poder adaptarme a las circunstancias y finalizar este proyecto.

Finalmente, extiendo mi gratitud a todos los profesores del máster, quienes con sus enseñanzas y conocimientos hicieron posible que adquiriera las competencias necesarias para desarrollar este proyecto.

DEDICATORIA

Este trabajo está dedicado, en primer lugar, a mi hijo Cristian. Gracias, hijo, porque incluso en los momentos más difíciles siempre encontraste la manera de animarme, “mamá, recuerda terminar tu trabajo”. Durante todo el máster fuiste mi inspiración constante, impulsándome a estudiar, a presentar cada entrega y a no rendirme. Hoy, aunque me acompañas desde el cielo, sigues dándome la fuerza para levantarme y culminar este proyecto que también es tuyo.

Dedico este logro a mi esposo Juan y a mis padres, quienes con amor, paciencia y apoyo incondicional me sostuvieron cuando sentí que no podía continuar. En cada instante me brindaron la motivación necesaria para seguir adelante, aun cuando la vida pesaba demasiado.

A mi hermosa familia, a los que están y a Cris, que me mira desde el cielo, este trabajo es para ustedes.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Katherine López Ramírez
Título del proyecto:	Predicción inteligente de precios en Airbnb Madrid
Directores del proyecto:	Álvaro Sánchez Pérez Carmen Alonso
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto:	SI
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Desarrollar un modelo predictivo en Machine Learning para estimar con precisión los precios de Airbnb en Madrid, considerando ubicación, características del alojamiento y estacionalidad.
Fuente de código:	Repositorio GitHub – Predicción de precios Airbnb Madrid

Índice

Índice de Figuras

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El turismo en Madrid ha experimentado un crecimiento sostenido en los últimos años, lo que ha impulsado el uso de plataformas como Airbnb. Sin embargo, la fijación de precios en un mercado dinámico y competitivo representa un desafío.

Este proyecto surge como respuesta a dicha necesidad, ofreciendo una herramienta que contribuye a:

- Anfitriones: ajustar precios de forma óptima según zona y estacionalidad.
- Turistas: identificar mejores alternativas de alojamiento en función de precio y localización.
- Inversores: tomar decisiones estratégicas sobre dónde invertir y qué rendimiento esperar.

1.2 Planteamiento del problema

La ausencia de un modelo robusto que considere múltiples factores (barrio, características del alojamiento, estacionalidad, entre otros) limita la capacidad de los actores del mercado para fijar precios justos. Esto genera incertidumbre, pérdida de competitividad y oportunidades de mejora en la experiencia turística.

1.3 Objetivos del proyecto

Objetivo general

Desarrollar un modelo de predicción que estime el precio óptimo de alojamientos en Airbnb Madrid, incorporando variables contextuales y temporales.

Objetivos específicos

- Analizar y limpiar las bases de datos de Airbnb.
- Explorar y visualizar la información relevante sobre barrios, precios y estacionalidad.
- Entrenar y comparar modelos de Machine Learning (CatBoost, XGBoost y Random Forest).
- Seleccionar el modelo más preciso y generar una herramienta de predicción práctica.

1.4 Resultados obtenidos

El modelo seleccionado, Random Forest optimizado con codificación ordinal, alcanzó métricas sobresalientes en la predicción de precios de alojamientos en Madrid:

- MAE: 6.34 €

- RMSE: 8.75 €
- R^2: 0.975

Estos resultados evidencian un alto grado de precisión y robustez, lo que permite ofrecer estimaciones confiables y realistas. Gracias a este desempeño, el modelo es capaz de proporcionar recomendaciones personalizadas de precios, comparativas entre barrios de Madrid y un apoyo sólido para la toma de decisiones de anfitriones, turistas e inversores.

1.5 Estructura de la memoria

La memoria se organiza de la siguiente forma:

- Capítulo 1: Introducción. Contexto, justificación, planteamiento del problema y estructura del documento.
- Capítulo 2: Estado del arte. Revisión de literatura y enfoques de pricing en plataformas P2P.
- Capítulo 3: Objetivos. Objetivo general, objetivos específicos y beneficios esperados.
- Capítulo 4: Desarrollo del proyecto. Metodología (CRISP-DM), datos, limpieza, EDA, modelado, evaluación e interfaz.
- Capítulo 5: Discusión. Interpretación crítica de resultados, decisiones tomadas y limitaciones.
- Capítulo 6: Conclusiones. Logros, aportaciones y validación del cumplimiento de objetivos.
- Capítulo 7: Líneas futuras. Mejoras, escalabilidad y extensiones del sistema.
- Capítulo 8: Referencias. Fuentes bibliográficas y documentales utilizadas.
- Capítulo 9: Anexos. Material complementario (disponible en el repositorio de GitHub).

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

En los últimos años, el auge de plataformas de economía colaborativa como Airbnb ha transformado el sector turístico. Este fenómeno ha generado un vasto campo de investigación en torno a la fijación dinámica de precios, la ocupación y la competitividad del mercado de alojamientos.

La literatura revisada destaca tres enfoques principales:

- Modelos descriptivos, que analizan tendencias de precios y demanda a partir de datos históricos.
- Modelos de predicción, que emplean técnicas de Machine Learning para estimar precios en función de características del alojamiento y su contexto.
- Modelos de optimización, que incorporan estrategias de pricing dinámico para maximizar ingresos.

Diversos estudios han demostrado que factores como la localización, el tipo de habitación, la estacionalidad y la reputación del anfitrión influyen de manera significativa en el precio. Sin embargo, aún existe una brecha en el desarrollo de modelos que integren estas variables de manera conjunta y que resulten aplicables en un caso práctico para la ciudad de Madrid.

2.2 Contexto y justificación

El crecimiento del turismo en Madrid y la fuerte presencia de Airbnb plantean la necesidad de herramientas que faciliten la toma de decisiones en un mercado altamente competitivo.

El presente proyecto busca aportar un modelo predictivo de precios que tenga utilidad en distintos ámbitos:

- Anfitriones: optimizar sus tarifas en función de la zona, la temporada y las características de su propiedad.
- Turistas: identificar barrios y periodos con precios más convenientes.
- Inversores: evaluar zonas con mayor potencial de rentabilidad.

2.3 Planteamiento del problema

La fijación de precios en Airbnb Madrid presenta limitaciones importantes:

- La información disponible es abundante, pero dispersa y con presencia de valores atípicos.

- No existen modelos unificados que integren factores espaciales (barrios), temporales (estacionalidad) y características del alojamiento en una predicción coherente.
- La ausencia de estas herramientas genera dificultades tanto para anfitriones como para turistas e inversores, afectando la competitividad del mercado y la transparencia de la oferta.

Este proyecto surge como respuesta a dicha problemática, con el objetivo de construir un modelo predictivo sólido que aproveche los datos disponibles y genere valor práctico para los distintos actores del ecosistema turístico.

Capítulo 3. OBJETIVOS

3.1 Objetivo general

Analizar y predecir los precios de alojamientos en Airbnb Madrid en función de la zona geográfica, la estacionalidad turística y las características del inmueble, aplicando técnicas de machine learning.

3.2 Objetivos específicos

Para alcanzar el objetivo general, se han definido los siguientes objetivos concretos:

- Analizar la información de Inside Airbnb para comprender los factores que influyen en el precio de los alojamientos.
- Depurar y transformar los datos, eliminando valores atípicos y nulos, e integrando la dimensión temporal de la estacionalidad.
- Implementar y comparar diferentes algoritmos de Machine Learning (CatBoost, XGBoost y Random Forest) para identificar el modelo con mejor desempeño predictivo.
- Validar el modelo seleccionado mediante métricas de error (MAE, RMSE) y ajuste (R^2), asegurando precisión y robustez.
- Desarrollar una funcionalidad práctica que permita consultar precios estimados y compararlos entre barrios, facilitando la toma de decisiones.

3.3 Beneficios del proyecto

El desarrollo de este proyecto aporta beneficios en tres dimensiones:

- Para los anfitriones: optimizar la fijación de precios en función de la zona y la estacionalidad, aumentando competitividad y rentabilidad.
- Para los turistas: identificar barrios alternativos con precios más accesibles y ajustados, favoreciendo una mejor planificación del viaje.
- Para los inversores: evaluar con mayor precisión qué zonas de Madrid presentan mayor potencial de rentabilidad en el mercado de alquiler turístico.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

El proyecto se desarrolló siguiendo una secuencia lógica de fases que garantizó una construcción sólida del modelo predictivo y un análisis exhaustivo de los datos:

1. ***Estudio del estado del arte y definición del problema***
    - Revisión de la literatura sobre pricing en plataformas de economía colaborativa.
    - Identificación del reto de fijación de precios en Airbnb como un problema multivariable con fuerte dependencia de factores geográficos, estructurales y de estacionalidad.
2. ***Obtención y exploración de datos***
    - Descarga de datasets públicos de Inside Airbnb: listings.csv (características de los anuncios y anfitriones) y calendar.csv (precios diarios).
    - Exploración inicial para identificar inconsistencias, nulos y outliers.
3. ***Limpieza y preparación de datos***
    - Selección de variables relevantes.
    - Transformación de precios a formato numérico.
    - Tratamiento de valores nulos y codificación de variables categóricas.
    - Eliminación de outliers a nivel global y por barrio para asegurar distribuciones realistas.
4. ***Análisis exploratorio y visualizaciones***
    - Distribución de precios a nivel global y por barrio.
    - Comparativa por tipo de alojamiento.
    - Mapas de calor para visualizar diferencias espaciales.
    - Boxplots de precios promedio mensuales en barrios con mayor número de registros.
5. ***Modelado y evaluación***
    - Entrenamiento y comparación de tres modelos representativos: CatBoost, XGBoost y Random Forest.
    - Selección del modelo óptimo en función de métricas de error y capacidad de generalización.
6. ***Resultados y conclusiones***
    - Análisis de métricas de rendimiento.
    - Identificación de variables más influyentes.
    - Pruebas de predicción práctica y utilidad del sistema.

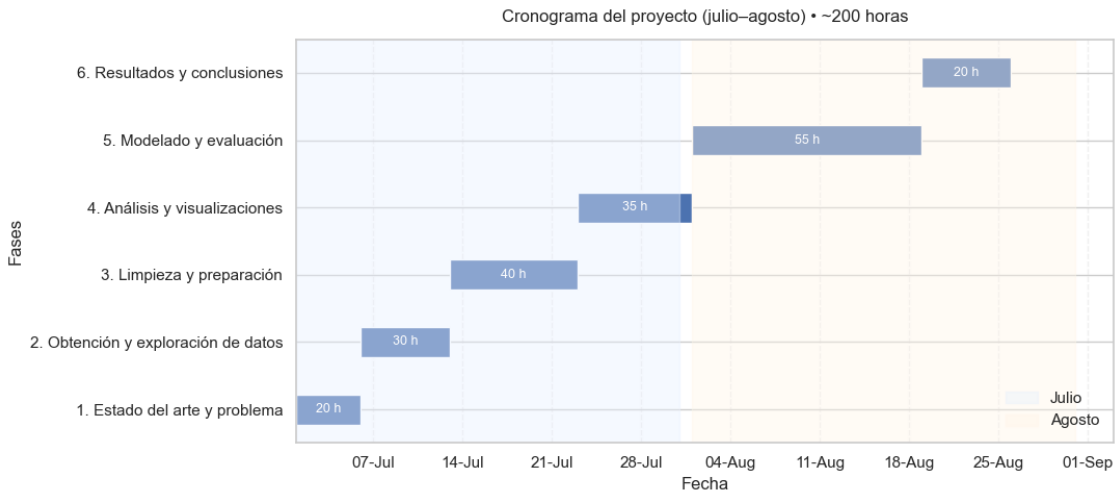

Ilustración 1. Cronograma del proyecto de predicción de precios en Airbnb Madrid, con fases y horas estimadas (julio–agosto). Fuente: elaboración propia.

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Metodología de trabajo

El desarrollo del proyecto se ha basado en la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), un estándar ampliamente utilizado en proyectos de ciencia de datos. Esta metodología propone un enfoque estructurado y cíclico que permite avanzar de manera ordenada desde la definición del problema hasta la implementación de soluciones, garantizando rigor analítico y flexibilidad en la adaptación a los datos disponibles.

Las fases principales de CRISP-DM y su aplicación en este trabajo han sido:

- Comprensión del problema: análisis del contexto de la fijación de precios en Airbnb Madrid, identificando la necesidad de predecir precios óptimos considerando factores geográficos, estructurales y estacionales.
- Preparación de datos: integración de los datasets públicos de Inside Airbnb, limpieza de outliers, imputación de valores nulos y codificación de variables categóricas.
- Análisis exploratorio: estudio de la distribución de precios, visualización de diferencias por barrio y tipo de alojamiento, y detección de patrones estacionales.
- Modelado y comparación de algoritmos: entrenamiento de tres modelos representativos —CatBoost, XGBoost y Random Forest— aplicando técnicas de encoding específicas para cada uno.
- Evaluación y selección: comparación de métricas (MAE, RMSE y R^2) para seleccionar el modelo óptimo, priorizando precisión y capacidad de generalización.
- Interpretación de resultados: análisis de las variables más influyentes en el precio, visualización de errores, y prueba de predicciones en escenarios prácticos para demostrar la utilidad del sistema.

.

4.2.2 Herramientas utilizadas

En cuanto a las herramientas empleadas, se ha trabajado en Python 3.13.2 dentro de un entorno Jupyter Notebook, apoyado en las siguientes librerías principales:

- **pandas, NumPy →** limpieza, transformación y análisis de datos.
- **matplotlib, seaborn, plotly →** generación de visualizaciones estáticas e interactivas.
- **scikit-learn →** preprocesamiento, métricas y modelado con Random Forest.
- **xgboost y catboost →** implementación de modelos de gradient boosting.
- **category_encoders →** codificación de variables categóricas con Target Encoding.
- **Streamlit →** despliegue web y visualización interactiva del modelo (UI con selectores, métricas y tablas en tiempo real).
- **PyDeck (deck.gl) →** visualización geográfica: capas GeoJSON para resaltar barrios (elegido y alternativas) en el mapa.

Este enfoque metodológico ha permitido garantizar la trazabilidad del trabajo, replicabilidad de los resultados y claridad en cada fase del proceso de desarrollo.

Ilustración 2. Tecnologías y librerías utilizadas en el desarrollo del proyecto. Fuente: elaboración propia.

4.2.3 Limpieza y preparación de datos

La preparación de los datos fue una de las fases más críticas del proyecto, ya que de ella dependía la calidad del modelo predictivo. Se trabajó inicialmente con dos bases proporcionadas por Inside Airbnb:

- **listings.csv →** información de los anuncios y características de anfitriones.
- **calendar.csv →** precios diarios de los alojamientos, que posteriormente permitieron calcular la estacionalidad.

Proceso de limpieza realizado

1. ***Selección de variables relevantes***
   Se redujo el dataset listings.csv que tenia 79 columnas a 26 columnas clave, priorizando aquellas que aportaban valor al modelo: características del inmueble (número de habitaciones, baños, capacidad), métricas del anfitrión (superhost, tasas de respuesta/aceptación), variables de reseñas y localización (barrio).
2. ***Transformación de precios***
   - Conversión de los precios desde formato texto (ej. "$120") a valores numéricos.
   - Creación de la variable precio_promedio_mensual, que pasó a ser la variable objetivo del modelo.
3. ***Manejo de nulos***
   - Eliminación de registros sin reseñas.
   - Eliminación de entradas con datos críticos vacíos (ej. bedrooms, bathrooms), que resultaban fundamentales para la predicción.
4. ***Normalización de campos del host***
   - Tasas de respuesta y aceptación → transformadas a valores numéricos entre 0 y 100.
   - Variables booleanas → transformadas a binario (1 para “sí”, 0 para “no”).
5. ***Tratamiento de outliers***
   El mayor reto de esta fase fueron los precios atípicos. Se aplicó una eliminación en dos niveles:

   - Global (dataset listings) → detección de precios extremos superiores a 20.000 €, claramente irreales.
   - Por barrio (dataset unificado con calendar) → ajuste más fino, eliminando valores desproporcionados dentro de cada zona.

4.2.3.1 Variables utilizadas, unión y descartes

El proyecto trabajó con dos fuentes principales de datos:

- listings.csv, que contiene información detallada de cada anuncio y anfitrión.
- calendar.csv, que registra la disponibilidad y el precio diario de cada alojamiento.

Variables iniciales (listings): En la primera fase, se seleccionaron de listings.csv aquellas variables con mayor potencial explicativo para la predicción de precios:

- **Identificador:** id.
- **Ubicación:** latitude, longitude, neighbourhood_cleansed.

- **Alojamiento:** room_type, accommodates, bedrooms, beds, bathrooms.
- **Reseñas:** review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, number_of_reviews.
- **Anfitrión:** host_is_superhost, host_response_time, host_response_rate, host_acceptance_rate, host_identity_verified.
- **Reservas y disponibilidad:** availability_365, minimum_nights, instant_bookable.
- **Precio:** price (en formato texto, posteriormente normalizado).

Estas variables permiten capturar tanto las características estructurales del alojamiento como aspectos relacionados con la reputación del anfitrión y la experiencia del huésped.

Unión con calendario: Posteriormente, se unió la información de listings con calendar.csv. Esta unión permitió enriquecer los datos con dos variables adicionales:

- mes: obtenido a partir de la fecha, útil para capturar la estacionalidad.
- precio_promedio_mensual: calculado como la media de precios diarios por anuncio y mes, y definido como la variable objetivo del modelo.

El dataset resultante tras la unión incluyó las siguientes columnas:

id, latitude, longitude, neighbourhood_cleansed, room_type, accommodates, bedrooms, beds, bathrooms, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, number_of_reviews, host_is_superhost, host_response_time, host_response_rate, host_acceptance_rate, host_identity_verified, availability_365, minimum_nights, instant_bookable, price, mes, precio_promedio_mensual.

Variables eliminadas para el modelado: Antes del entrenamiento de modelos predictivos, se eliminaron aquellas columnas que no aportaban información útil o podían generar ruido:

- **id →** identificador sin valor predictivo.
- **latitude y longitude →** aunque reflejan la ubicación, se decidió mantener únicamente neighbourhood_cleansed, que captura la dimensión espacial de forma más manejable e interpretable.
- **price →** se eliminó porque ya estaba representado en la variable agregada precio_promedio_mensual, evitando duplicación y sobreajuste.

Con este proceso de selección y depuración, se construyó un dataset final robusto y coherente, con las variables más relevantes y una variable objetivo estable, adecuada para la predicción de precios a nivel mensual.

4.2.3.2 Análisis de outliers en precios

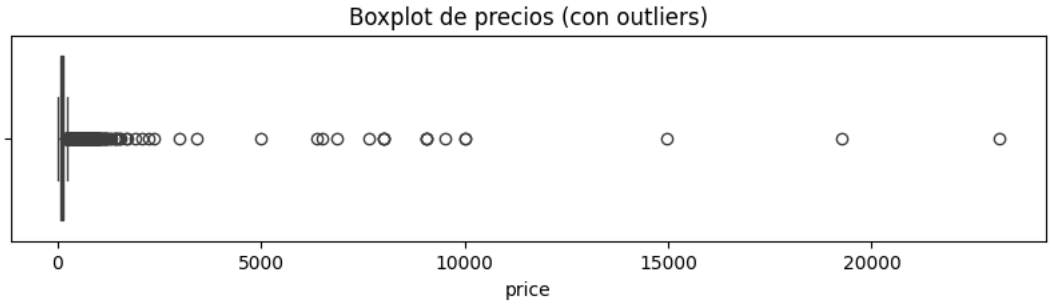

Ilustración 3. Boxplot de precios de los alojamientos (con outliers). Fuente: elaboración propia.

La (Ilustración 3), muestra precios llegando a más de 20.000 € por noche

El análisis inicial mostró que:

- Precio mínimo: 8 €
- Precio máximo: 23.124 €
- Media: 130 €
- Mediana: 96 €
- Límite superior estadístico: 248 €
- Se identificaron 1.035 precios extremos que sobrepasaban los límites normales.

Conclusión: la existencia de precios tan desorbitados distorsiona la media y genera distribuciones irreales, por lo que fue necesario eliminarlos.

4.2.3.3 *Distribución de precios por barrio*

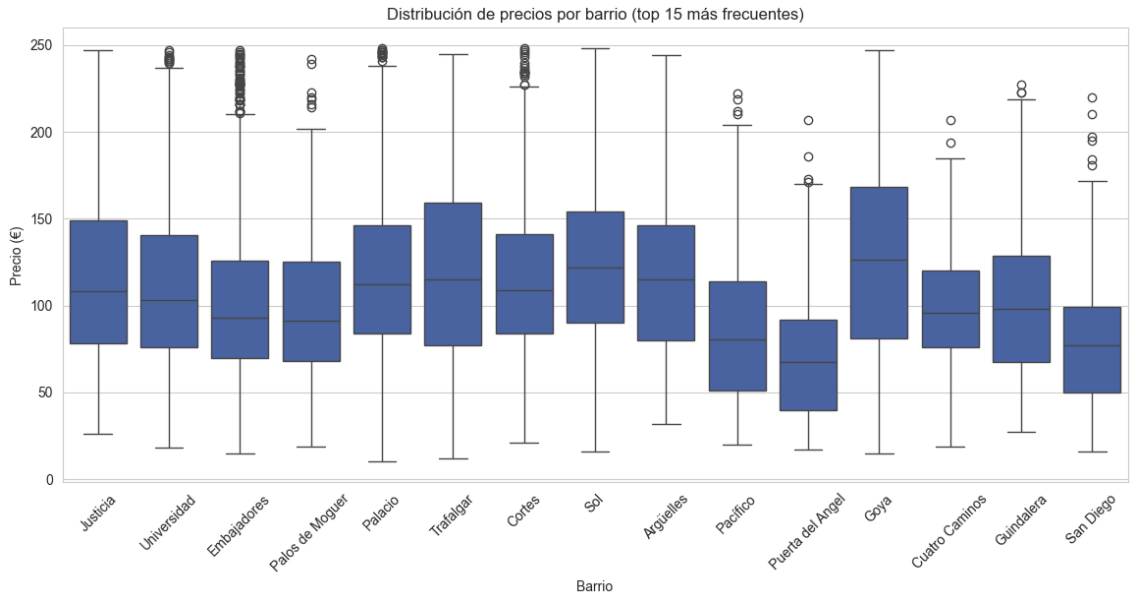

Ilustración 4. Boxplot de precios en los 15 barrios con mayor número de alojamientos. Fuente: elaboración propia.

El gráfico confirma que los precios dependen fuertemente del barrio:

- Zonas como Goya, Sol y Trafalgar presentan precios más altos y mayor dispersión.
- En contraste, Puerta del Ángel, San Diego o Pacífico muestran precios más bajos y estables.

Esta observación justificó mantener la variable neighbourhood_cleansed como una de las más influyentes del modelo.

4.2.3.4 *Distribución de precios por tipo de alojamiento*

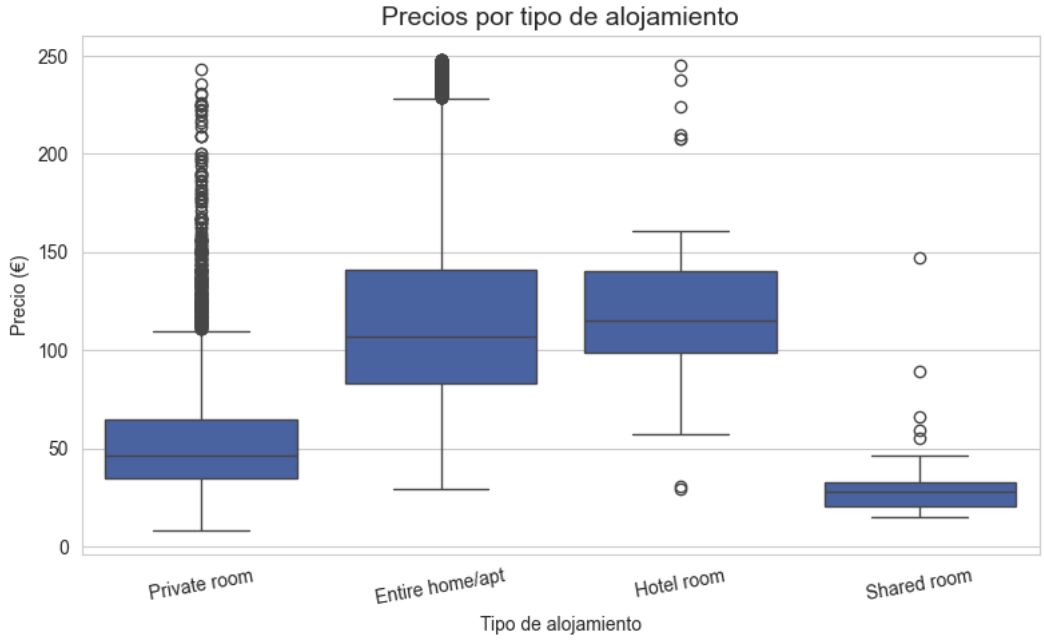

Ilustración 5. Distribución de precios por tipo de alojamiento en Madrid. Fuente: elaboración propia.

El análisis mostró un patrón coherente:

- Entire home/apt (apartamento entero) → mayor mediana (~100 €).
- Private room → precios intermedios, mayoría entre 30–60 €.
- Shared room → opción más barata (mediana ~25 €).
- Hotel room → precios intermedios-altos con mayor dispersión.

No obstante, también se detectaron outliers (ej. habitaciones privadas con precios superiores a apartamentos completos), atribuibles a errores de carga o anuncios poco representativos.

4.2.3.5 *Impacto de la limpieza de outliers*

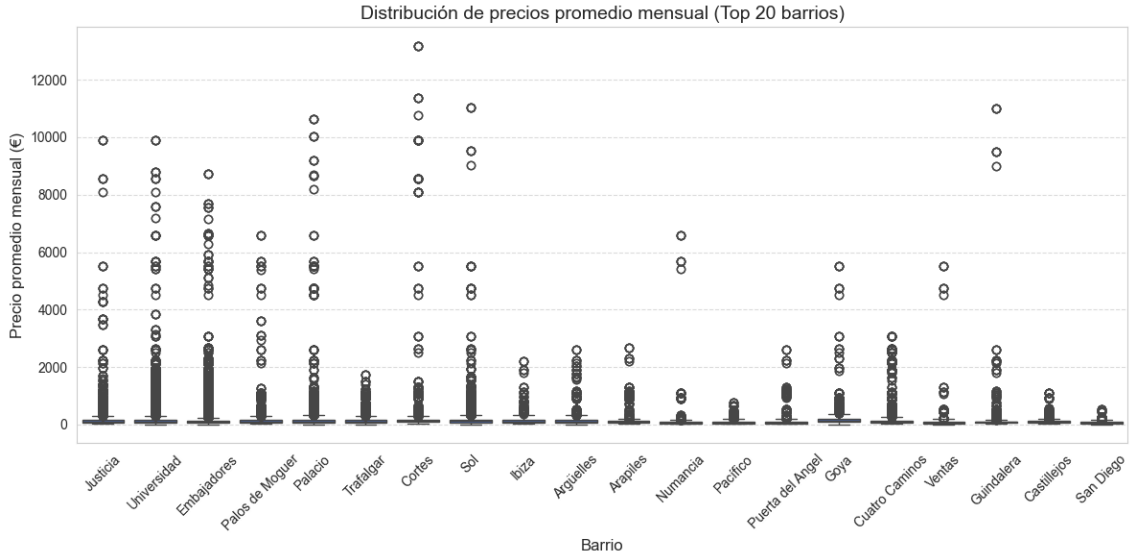

Ilustración 6. Impacto de los outliers en los precios promedio mensuales de los barrios más frecuentes. Fuente: elaboración propia.

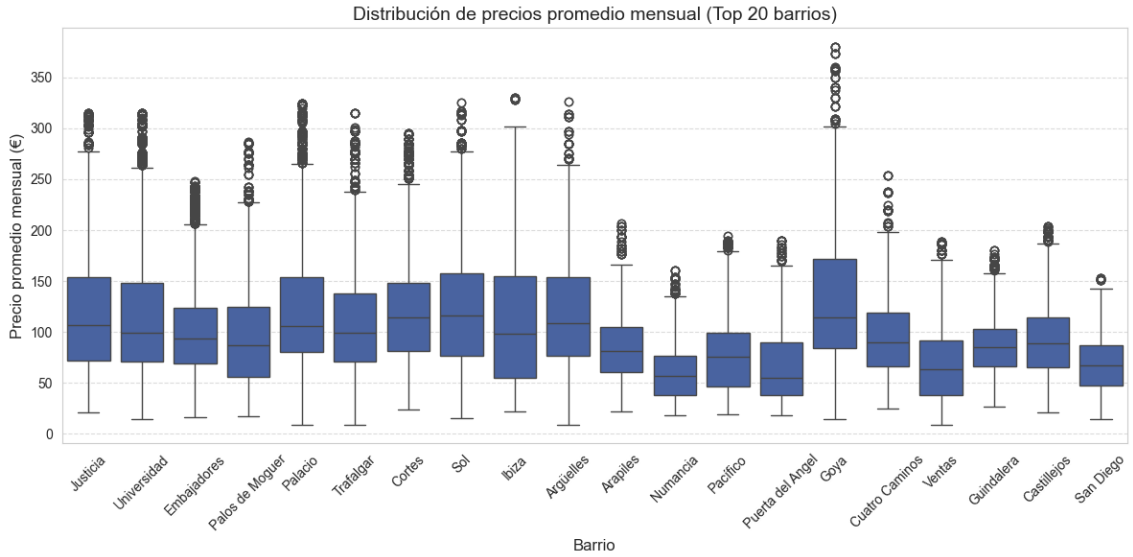

Ilustración 7. Distribución de precios promedio mensual por barrio (top 20) tras la limpieza de outliers. Fuente: elaboración propia.

- Antes de limpiar por barrio: se observaron en la (Ilustración 6), precios irreales después de la unión de listings con calendar (ej. más de 12.000 €/mes en barrios concretos).
- Después de limpiar por barrio: como se ve en la (Ilustración 7), la distribución se volvió coherente, manteniendo diferencias lógicas entre zonas pero eliminando valores irreales.

El gráfico confirma que los precios dependen fuertemente del barrio:

- Zonas como Goya, Sol y Trafalgar presentan precios más altos y mayor dispersión.
- En contraste, Puerta del Ángel, San Diego o Pacífico muestran precios más bajos y estables.

Esta observación justificó mantener la variable neighbourhood_cleansed como una de las más influyentes del modelo.

4.2.3.6 Conclusión de la preparación de datos

La unión de los datasets (listings + calendar), junto con la eliminación de outliers globales y por barrio, permitió obtener una base de datos realista y robusta. Esto fue clave para:

- Evitar que los modelos se vieran sesgados por valores irreales.
- Capturar patrones geográficos y estacionales reales.
- Mejorar la calidad y fiabilidad de las predicciones.

Eliminación de outliers por barrio: Tras la integración de ambos datasets, se aplicó un segundo filtrado, esta vez a nivel de barrio, con el fin de eliminar valores atípicos locales y garantizar que los precios reflejaran tendencias realistas en cada zona de Madrid.

Finalmente, se generó el archivo df_modelo.csv, con el dataset limpio, preparado y listo para el modelado. De esta forma, se obtuvo un dataset final robusto, coherente y listo para el análisis exploratorio y la fase de modelado predictivo.

4.2.4 Análisis exploratorio

Como parte del análisis exploratorio de datos (EDA, Exploratory Data Analysis), se realizó una visualización espacial mediante un mapa de calor que representa el precio promedio por barrio en la ciudad de Madrid.

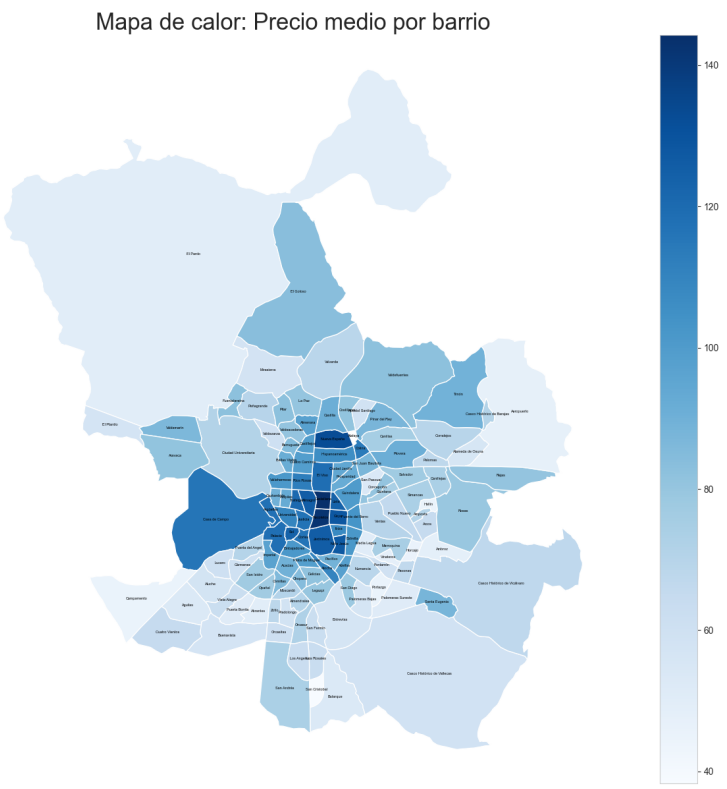

Ilustración 8. Mapa de calor por precios promedio de los alojamientos en Madrid. Fuente: elaboración propia.

En el mapa (Ilustración 8) se observa claramente cómo la ubicación geográfica influye en el valor de los alojamientos:

- Los barrios céntricos, como Justicia, Sol, Embajadores y Palacio, aparecen en tonos más oscuros, lo que refleja precios medios significativamente más altos. Esto coincide con la alta demanda turística y la proximidad a los principales atractivos de la ciudad.
- A medida que se avanza hacia la periferia, los precios descienden, representados en tonos más claros, lo que indica una mayor accesibilidad económica en esas zonas.
- La escala lateral de colores permite identificar un rango de valores que va aproximadamente de 40 € a más de 140 € por noche en promedio, lo que demuestra la gran dispersión de precios dentro del mercado de Airbnb en Madrid.

Conclusión: Este análisis exploratorio confirma que la variable neighbourhood_cleansed es clave en el modelo predictivo, ya que la localización del inmueble representa uno de los factores más determinantes en la fijación de precios. La dimensión espacial obtenida a partir del mapa de calor permite identificar claramente zonas premium frente a zonas más asequibles, aportando un insumo fundamental tanto para anfitriones como para inversores y turistas.

4.2.5 Modelados predictivos

Para evaluar el rendimiento del sistema de predicción, se entrenaron y compararon tres algoritmos representativos de machine learning: CatBoost, XGBoost y Random

Forest, todos aplicados sobre el dataset limpio y enriquecido con información de estacionalidad mensual.

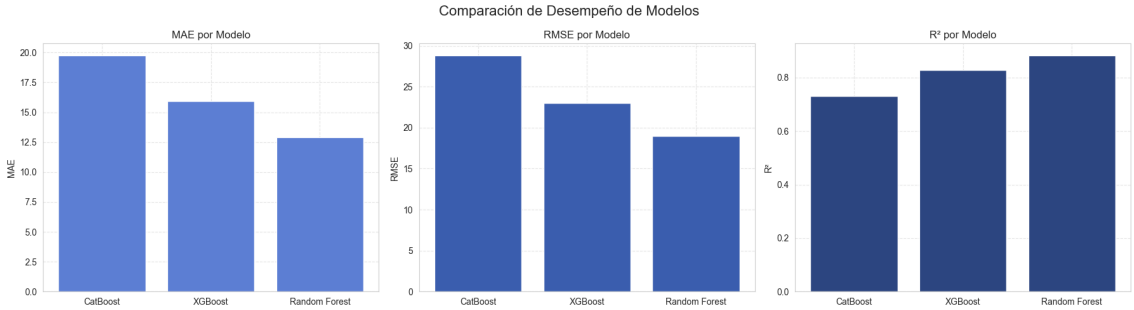

Ilustración 9. Comparación de desempeño de modelos de predicción (CatBoost, XGBoost y Random Forest). Fuente: elaboración propia.

Los resultados obtenidos en el conjunto de prueba fueron los siguientes:

- CatBoost → MAE: 19,75 €, RMSE: 28,81 €, R^2: 0,731
- XGBoost → MAE: 15,91 €, RMSE: 23,04 €, R^2: 0,828
- **Random Forest → MAE: 12,90 €, RMSE: 18,99 €, R^2: 0,883**

La comparación gráfica (Ilustración 9) y numérica evidencia que el Random Forest es el modelo con mejor desempeño global, obteniendo los valores más bajos de error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE), junto con el mayor coeficiente de determinación (R^2).

Estrategia de codificación

Cada modelo requirió un tratamiento diferenciado de las variables categóricas, acorde con sus características internas:

- **CatBoost:** permite trabajar directamente con variables categóricas, incorporando un mecanismo interno de codificación eficiente.
- **XGBoost:** no admite variables categóricas de forma nativa, por lo que se aplicó Target Encoding, transformando las categorías en valores numéricos representativos en función de la media del target.
- **Random Forest:** se utilizó Ordinal Encoding, una codificación simple que mantiene un buen equilibrio entre eficiencia y rendimiento, evitando problemas de alta cardinalidad que podrían surgir con OneHot o Target Encoding.

Decisión final

La estrategia de comparación mostró que Random Forest no solo ofrece el mejor rendimiento cuantitativo, sino también la mayor estabilidad y robustez. Esto lo convierte en la opción más adecuada para el objetivo del proyecto: predecir de forma

precisa y consistente los precios de alojamientos en Airbnb Madrid, considerando factores espaciales y estacionales.

4.2.6 Interpretación del modelo Random Forest

Ajuste de Hiperparámetros: Con el fin de optimizar el rendimiento del modelo, se aplicó un proceso de búsqueda de hiperparámetros utilizando GridSearch con validación cruzada (CV = 3).

En este procedimiento se evaluaron diferentes combinaciones de parámetros fundamentales:

- **n_estimators:** número de árboles en el bosque.
- **max_depth:** profundidad máxima de cada árbol.
- **min_samples_split:** número mínimo de observaciones necesarias para dividir un nodo.
- **min_samples_leaf:** número mínimo de observaciones en una hoja terminal.
- **max_features:** número de variables consideradas al dividir un nodo.

Tras evaluar múltiples configuraciones, los hiperparámetros óptimos fueron los siguientes:

- **max_depth:** 25
- **max_features:** None
- **min_samples_leaf:** 2
- **min_samples_split:** 4
- **n_estimators:** 314

Esta configuración permitió un equilibrio entre complejidad y capacidad predictiva, evitando tanto el sobreajuste como el subajuste del modelo.

Rendimiento en Conjunto de Entrenamiento y Prueba: Con los hiperparámetros ajustados, el modelo alcanzó un rendimiento sobresaliente:

- **Conjunto de entrenamiento (Train):**
  - MAE: 3,88 €
  - RMSE: 5,35 €
  - R^2: 0,991

- **Conjunto de prueba (Test):**
  - MAE: 6,34 €
  - RMSE: 8,75 €
  - R^2: 0,975

Estos resultados evidencian que el modelo no solo logra un ajuste muy preciso en los datos de entrenamiento, sino que también mantiene un alto nivel de generalización en datos no vistos, reduciendo el riesgo de sobreajuste.

Validación Cruzada: Para reforzar la robustez del modelo, se aplicó una validación cruzada con 5 particiones. Los valores de R^2 obtenidos en cada fold fueron:

- Fold 1: 0,9700
- Fold 2: 0,9685
- Fold 3: 0,9693
- Fold 4: 0,9683
- Fold 5: 0,9705
- Media: 0,9693
- Desviación estándar: 0,0008

Estos valores muestran que el modelo es estable y consistente a lo largo de diferentes subconjuntos de datos, lo que confirma su fiabilidad en la predicción de precios de alojamientos en Madrid.

Importancia de las variables: La (Ilustración 10) muestra las 15 variables más influyentes en el modelo Random Forest optimizado. Se observa que:

- room_type es, con diferencia, la variable más determinante, explicando más del 25% de la variabilidad del precio. Este resultado refleja la fuerte diferencia de tarifas entre apartamentos completos, habitaciones privadas y compartidas.
- review_scores_location y neighbourhood_cleansed aparecen como factores clave relacionados con la ubicación, confirmando que tanto la percepción de los usuarios sobre la localización como el barrio en el que se encuentra el alojamiento inciden directamente en el precio.
- availability_365 y number_of_reviews también tienen un peso significativo. La disponibilidad anual refleja la estrategia del anfitrión en el mercado, mientras que el número de reseñas puede estar asociado a la confianza del cliente y la demanda del anuncio.
- Variables estructurales como bedrooms, bathrooms y accommodates aportan explicaciones adicionales, dado que la capacidad y el nivel de comodidad son elementos que naturalmente impactan en el precio.
- Finalmente, las reseñas detalladas (limpieza, comunicación, check-in, valor, puntuación general) junto con características del anfitrión (ej. host_acceptance_rate) tienen un efecto complementario, mostrando que la calidad percibida y la fiabilidad también permiten justificar precios más altos.

Este análisis demuestra que el precio de un alojamiento en Airbnb Madrid depende de una combinación equilibrada de factores estructurales (tipo de alojamiento, capacidad), espaciales (barrio, localización), temporales (disponibilidad) y sociales (reseñas y confianza en el anfitrión).

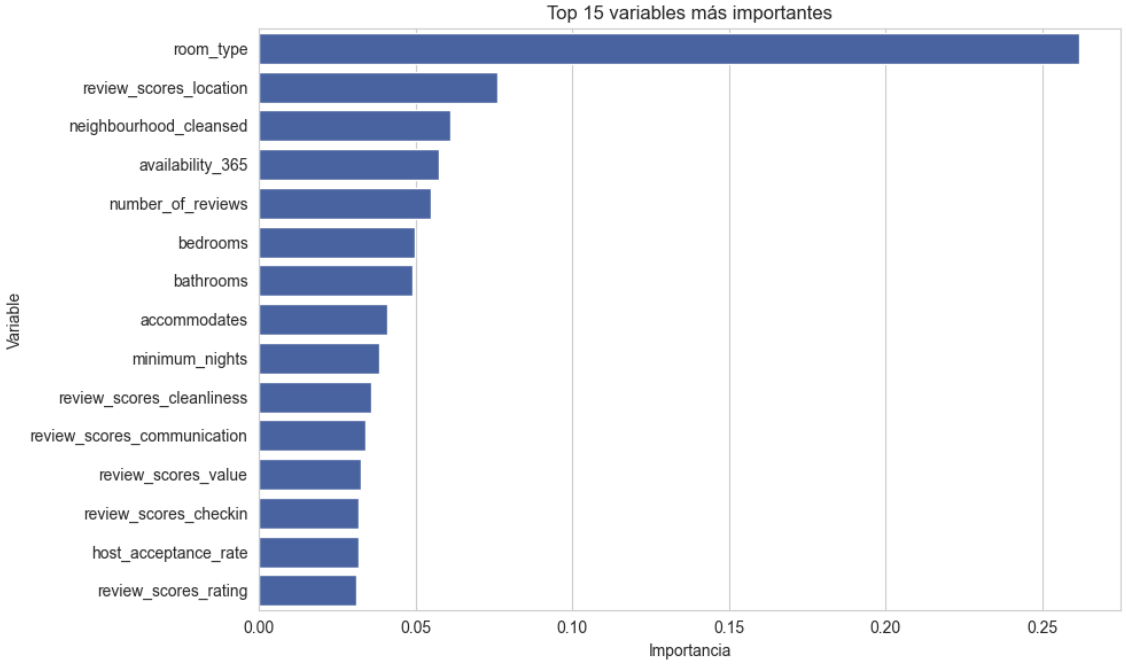

Ilustración 10. Importancia de las 15 variables principales en el modelo de predicción. Fuente: elaboración propia.

Matriz de correlación: La (Ilustración 11) presenta la matriz de correlación entre las variables numéricas utilizadas en el modelo Random Forest optimizado. Este análisis permite identificar relaciones lineales entre los distintos predictores y detectar posibles redundancias o multicolinealidad.

- **Relaciones esperadas entre variables estructurales:** se observan correlaciones positivas altas entre accommodates, bedrooms y beds. Esto es consistente, ya que a mayor número de habitaciones y camas, mayor capacidad de alojamiento.
- **Calidad de reseñas:** existe una fuerte asociación entre las variables de puntuaciones (review_scores_accuracy, cleanliness, checkin, communication, value), lo que indica que los usuarios tienden a valorar de manera conjunta la experiencia global del alojamiento.
- **Disponibilidad y noches mínimas:** variables como availability_365 y minimum_nights muestran correlaciones bajas con el resto, lo cual confirma que aportan información complementaria sin redundancia fuerte con otras características.
- **Factores del anfitrión:** variables como host_response_time o host_response_rate presentan correlaciones bajas con la mayoría de predictores, lo que refleja que influyen de forma independiente en la predicción del precio.

El análisis confirma que no existe una multicolinealidad crítica entre las variables numéricas, lo que valida su uso en el modelo Random Forest. Además, resalta la importancia de considerar un conjunto diverso de factores (estructurales, de reseñas, de anfitrión y de disponibilidad) para capturar adecuadamente la complejidad de los precios en Airbnb Madrid.

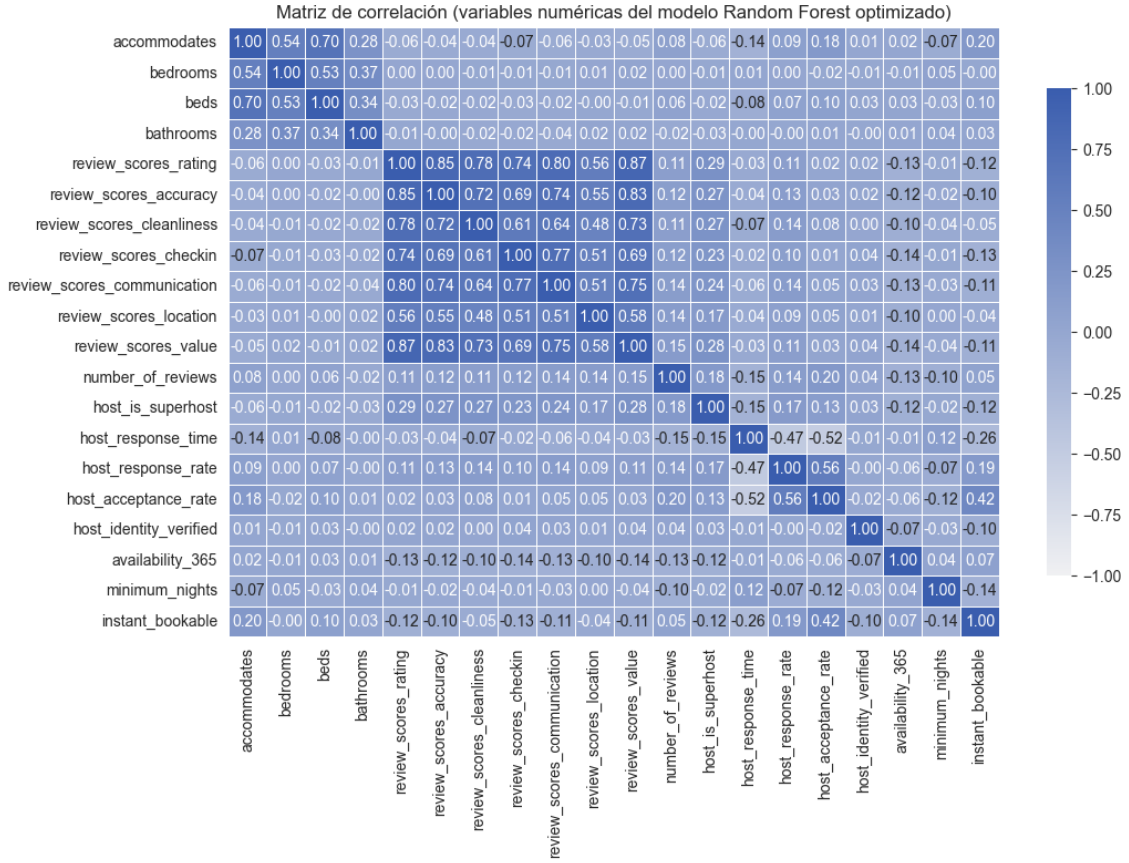

Ilustración 11. Matriz de correlación de variables numéricas utilizadas en el modelo Random Forest. Fuente: elaboración propia

Relación entre precios reales y predichos: La (Ilustración 12) muestra un gráfico de dispersión que compara los precios reales de los alojamientos en Madrid con los precios predichos por el modelo Random Forest optimizado. La línea roja discontinua representa la línea de referencia ideal (y = x), donde el valor predicho coincide exactamente con el valor real.

- Alta alineación con la línea de referencia: la mayoría de los puntos se concentran muy próximos a la línea roja, lo que indica que el modelo predice con gran precisión los precios de la mayoría de los alojamientos.
- Desempeño robusto en el rango principal: para precios entre 0 € y 300 €, que representan la mayor parte de la oferta, las predicciones son consistentes y ajustadas al valor real.

- Menor precisión en valores extremos: en los precios más altos (por encima de 400 €), se observan algunas desviaciones, con predicciones que tienden a estar por debajo de los valores reales. Esto sugiere una ligera subestimación en los alojamientos de lujo o de gama muy alta.

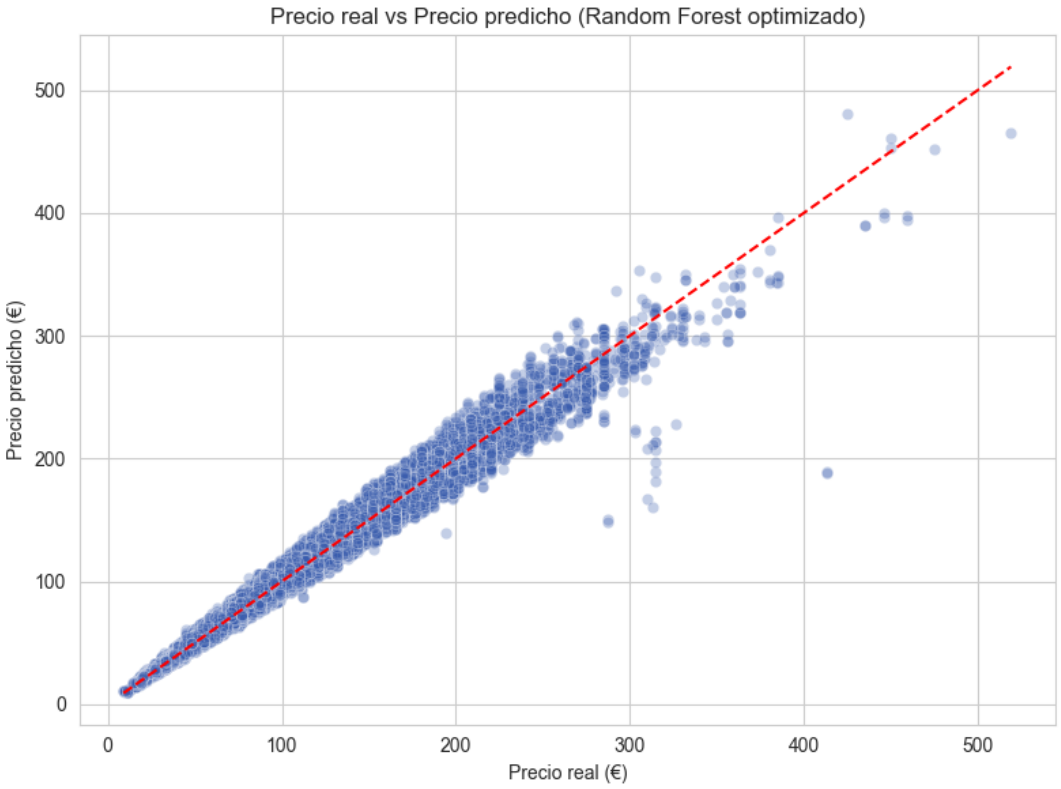

Ilustración 12. Comparación entre precio real y precio predicho por el modelo Random Forest. Fuente: elaboración propia.

Por su parte, la (Ilustración 13) representa la distribución del error absoluto en las predicciones del modelo Random Forest optimizado. Se observa que la mayoría de los errores se concentran en valores inferiores a 20 €, lo que indica que el modelo es altamente preciso en la gran mayoría de los casos.

Asimismo, la curva muestra que a medida que aumenta el error, la frecuencia disminuye de manera pronunciada, confirmando que las predicciones con errores elevados son poco comunes. En conjunto, esta distribución respalda la robustez del modelo y su capacidad de generar estimaciones confiables para la mayoría de los anuncios, destacando únicamente algunas excepciones en valores atípicos.

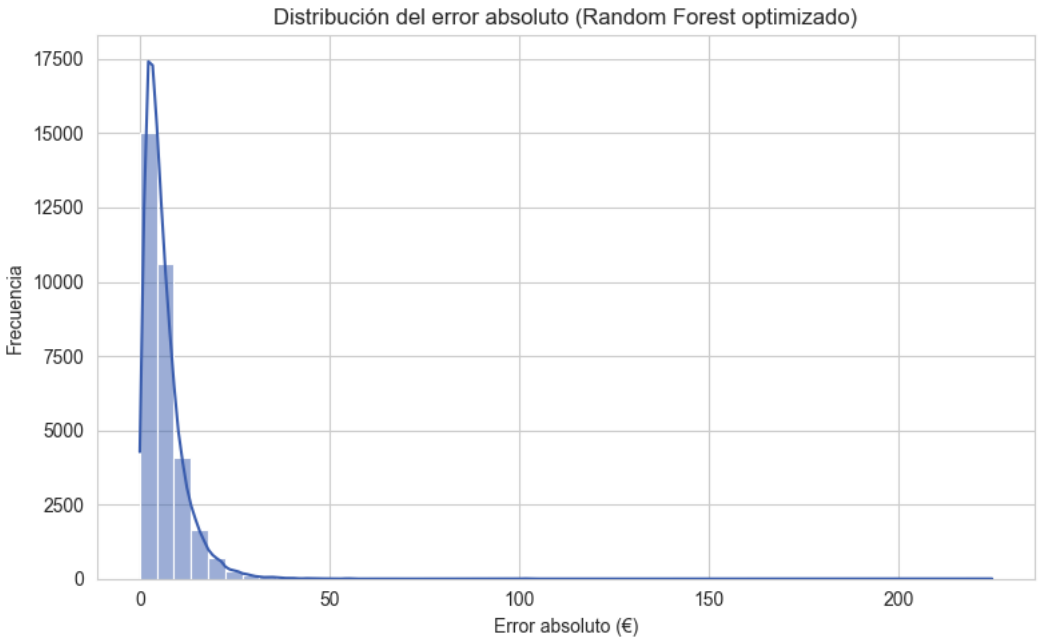

Ilustración 13. Distribución del error absoluto en la predicción de precios. Fuente: elaboración propia.

Desempeño por barrios: Por otro lado, la Ilustración 14 muestra el Top 15 de barrios con mayor error absoluto promedio en las predicciones del modelo Random Forest optimizado. Se observa que los barrios con precios más altos y mayor heterogeneidad en la oferta, como Jerónimos (11,65 €), Lista (10,82 €) y Goya (10,23 €), presentan los errores más elevados.

Este comportamiento puede explicarse porque en estas zonas existen alojamientos muy diversos en cuanto a tamaño, servicios y exclusividad, lo que dificulta al modelo capturar todos los matices que influyen en la fijación de precios.

En contraste, barrios como Fuente del Berro (7,43 €), Cortes (7,54 €) e Ibiza (7,64 €) presentan errores más reducidos, reflejando un mercado más homogéneo y, por lo tanto, más predecible.

En general, este análisis confirma que el modelo mantiene un desempeño sólido en todos los barrios, aunque la complejidad y variabilidad del mercado en áreas premium de Madrid genera ligeros aumentos en el error absoluto promedio.

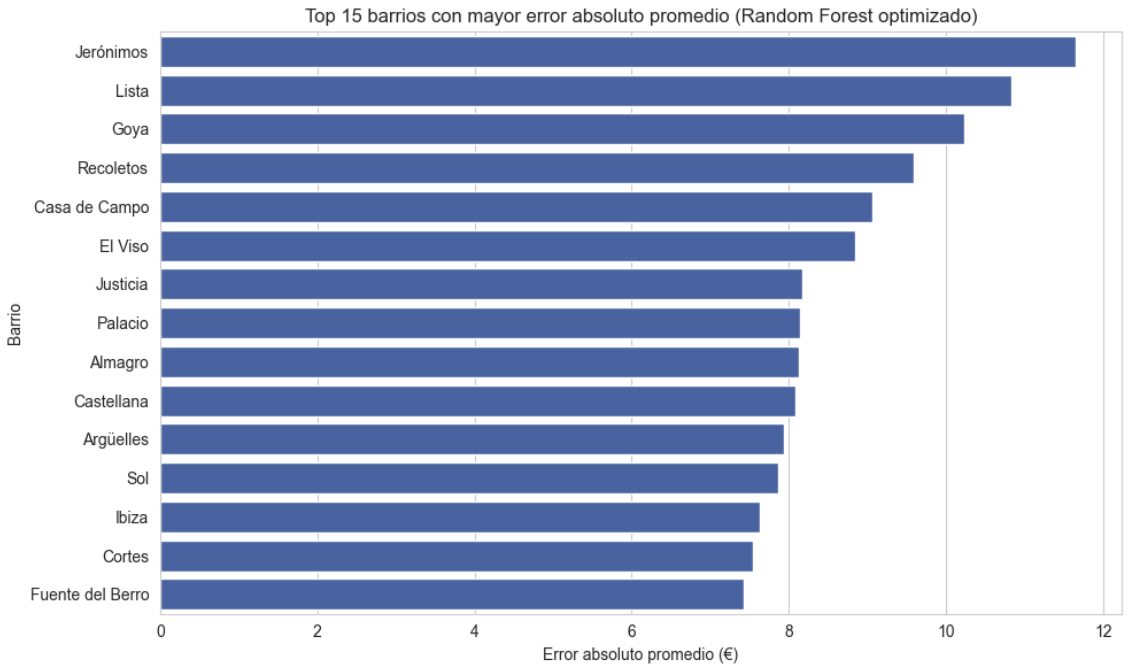

Ilustración 14. Top 15 barrios con mayor error absoluto promedio en la predicción. Fuente: elaboración propia.

Evaluación global del modelo: La optimización de hiperparámetros permitió obtener un modelo de Random Forest con un rendimiento altamente competitivo en la predicción de precios de Airbnb en Madrid. Los resultados alcanzados, con un MAE de 6,34 €, un RMSE de 8,75 € y un R^2 de 0,975, confirman la capacidad del modelo para realizar estimaciones muy cercanas a los valores reales y explicar de forma robusta la variabilidad de los precios.

El análisis de importancia de variables reveló que factores como el tipo de alojamiento (room_type), la ubicación (neighbourhood_cleansed) y ciertos indicadores de calidad percibida por los usuarios (por ejemplo, review_scores_location y review_scores_cleanliness) son determinantes en la fijación de precios. Esto demuestra que los precios no dependen únicamente de la ubicación, sino también de las características estructurales del alojamiento y de la percepción del servicio.

Asimismo, los análisis de correlación y de desempeño por barrios permitieron identificar patrones clave:

- En barrios con mercados más homogéneos, el error de predicción es menor.
- En zonas premium como Jerónimos, Lista o Goya, el error absoluto promedio es mayor debido a la gran heterogeneidad y exclusividad de la oferta, lo cual representa un desafío natural para los modelos predictivos.

Síntesis final del modelo: El modelo Random Forest optimizado no solo alcanzó un desempeño sobresaliente en métricas cuantitativas, sino que también aportó insights cualitativos valiosos

sobre los factores que más inciden en el precio de los alojamientos y las dinámicas de distintos barrios de Madrid.

Este resultado sienta una base sólida para su aplicación práctica en la toma de decisiones de anfitriones, inversores y usuarios, y abre la puerta a futuras extensiones que incorporen nuevas fuentes de datos externas o enfoques de modelado más avanzados.

4.2.7 Interfaz de usuario

Como parte final del desarrollo, se implementó una interfaz interactiva utilizando Streamlit, con el objetivo de hacer el modelo accesible para cualquier usuario sin necesidad de conocimientos técnicos.

La interfaz permite:

- Seleccionar las características del alojamiento (barrio, tipo de alojamiento, número de habitaciones, baños y capacidad de personas).
- Consultar el precio estimado mensual para un mes específico.
- Comparar el precio con otros meses, capturando la estacionalidad.
- Obtener alternativas por barrio, destacando tanto las opciones con precios similares como las más baratas.
- Visualizar en un mapa dinámico las zonas de Madrid según los precios estimados, lo que facilita el análisis espacial.

El diseño sigue una lógica sencilla y user-friendly:

- En el panel lateral (sidebar) se ingresan las características del inmueble.
- En el panel central se muestran los resultados: precio estimado, comparativas mensuales y alternativas por barrio.
- El mapa interactivo añade una dimensión visual y geográfica al análisis, permitiendo ubicar fácilmente las diferencias de precio entre zonas de la ciudad.

Esta estructura garantiza una navegación intuitiva y rápida, incluso para usuarios no expertos en Data Science. Además, el lenguaje empleado es claro, evitando términos técnicos complejos.

- (Ilustración 15). Interfaz principal de la aplicación con predicción de precios, comparativa mensual y alternativas por barrio.
- (Ilustración 16). Mapa dinámico de barrios de Madrid resaltando la zona seleccionada y alternativas.

La interfaz convierte el modelo en una herramienta práctica y aplicable en escenarios reales:

- Para anfitriones, permite ajustar precios de forma competitiva.
- Para turistas, facilita comparar opciones y elegir la mejor alternativa.

- Para inversores, ofrece una visión clara de las zonas con mayor potencial de rentabilidad.

En resumen, la aplicación no solo demuestra la viabilidad técnica del modelo, sino que también lo transforma en un producto funcional, capaz de integrarse en el ecosistema de toma de decisiones del mercado turístico.

Ilustración 15. Interfaz de la aplicación para la predicción de precios en Airbnb Madrid. Fuente: elaboración propia.

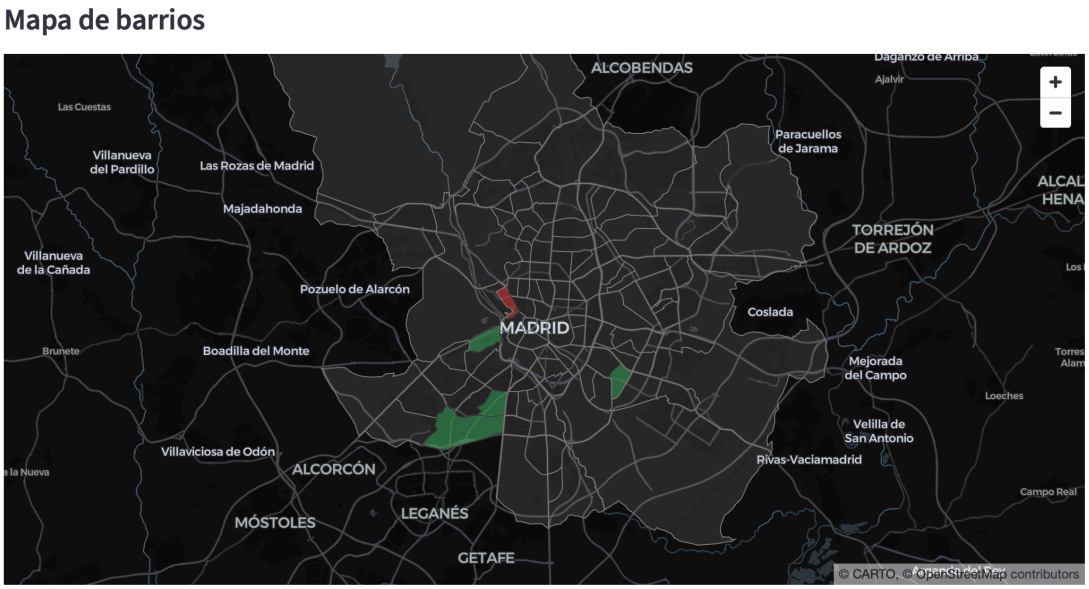

Ilustración 16. Visualización geográfica de barrios de Madrid en la herramienta interactiva. Fuente: elaboración propia.

4.3 Recursos requeridos

Para la ejecución del proyecto se utilizaron los siguientes recursos:

- Equipo técnico: ordenador portátil personal con capacidad suficiente para el procesamiento de datos y entrenamiento de modelos de Machine Learning.
- Software: ecosistema open-source, principalmente Python y librerías de Data Science (pandas, NumPy, matplotlib, seaborn, plotly, scikit-learn, XGBoost, CatBoost, category_encoders), además del entorno de desarrollo Jupyter Notebook.
- Datos: conjuntos de datos públicos procedentes de la plataforma Inside Airbnb, incluyendo archivos listings.csv y calendar.csv.
- Visualización geográfica: uso de archivos geojson públicos de los barrios de Madrid para la representación espacial de precios.
- Tiempo de trabajo: aproximadamente 200 horas distribuidas en las fases de recolección, limpieza, análisis, modelado, interpretación de resultados y redacción de la memoria.

Este conjunto de recursos permitió garantizar un desarrollo eficiente, sin necesidad de inversión económica adicional, al apoyarse en software libre y en fuentes de datos abiertas.

4.4 Presupuesto

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto (≈ 200)	4.000 €	Estimación a 20 €/h como valor de mercado de trabajo de data scientist junior.
Equipo técnico utilizado	1.200 €	Portátil personal, MacBook Pro.
Software utilizado	0 €	Python, librerías y Streamlit (gratuitos).
Materiales y documentación	0 €	Uso de recursos digitales y datasets públicos.
Total estimado	**5.200€**	

4.5 Viabilidad

El proyecto resulta económicamente y técnicamente viable, dado que:

- **Costes reducidos:** los gastos fueron mínimos gracias al uso de software open-source (Python, librerías de Data Science, Jupyter Notebook, Streamlit), y a la utilización de datasets públicos de Inside Airbnb. Únicamente se requirió equipo personal (ordenador portátil), lo que evita inversiones adicionales significativas.
- **Beneficios potenciales:** el modelo ofrece un alto valor añadido para distintos actores:

- ⇒ **Anfitriones:** pueden optimizar sus ingresos ajustando precios en función de zona, estacionalidad y características de la propiedad.
- ⇒ **Turistas:** acceden a una referencia objetiva de precios, lo que les facilita identificar oportunidades y evitar sobrecostes.
- ⇒ **Inversores y gestores inmobiliarios:** disponen de una herramienta de análisis que contribuye a tomar decisiones estratégicas más fundamentadas sobre dónde y cuándo invertir.

- **Escalabilidad y sostenibilidad:** la metodología es replicable en otras ciudades o plataformas de alquiler vacacional, lo que multiplica su aplicabilidad. Asimismo, el uso de tecnologías abiertas garantiza que el sistema pueda mantenerse y evolucionar a futuro sin dependencia de licencias de pago.

En conjunto, la relación coste/beneficio es altamente favorable, con una inversión reducida (5.200 € estimados), se logra un modelo robusto, de gran utilidad práctica y con un horizonte de aplicación sostenible en el tiempo.

4.6 Resultados del proyecto

El proyecto no solo alcanzó métricas sólidas en términos de predicción, sino que también permitió extraer conocimiento valioso sobre el comportamiento del mercado de alojamientos turísticos en Madrid.

Uno de los hallazgos más relevantes fue la confirmación del peso de tres dimensiones clave en la fijación de precios:

1. Estructura del alojamiento (tipo de vivienda, número de habitaciones y capacidad).
2. Ubicación geográfica (diferencias significativas entre barrios céntricos y periféricos).
3. Estacionalidad (variaciones en función del mes, asociadas a la dinámica turística de la ciudad).

Más allá de los indicadores numéricos, el modelo demostró que es posible construir un sistema de apoyo a la decisión que:

- Sugiere precios realistas a anfitriones para maximizar ocupación y rentabilidad.
- Permite a turistas identificar zonas más asequibles o con mejor relación calidad-precio.
- Ofrece a inversores una herramienta objetiva para valorar barrios con mayor potencial de rentabilidad.

Además, el proceso de análisis exploratorio y limpieza evidenció que el mercado presenta anomalías y errores de registro (ejemplo: anuncios con precios irreales superiores a 20.000 € por noche). Detectar y corregir estos casos no solo mejora la calidad del modelo, sino que también aporta una visión más transparente del ecosistema Airbnb en Madrid.

En términos metodológicos, el proyecto mostró la importancia de combinar fuentes y enfoques: la unión de listings con calendar enriqueció el dataset y permitió capturar la estacionalidad, mientras que la comparación de algoritmos evidenció fortalezas y limitaciones de cada enfoque, consolidando a Random Forest como la mejor opción por su balance entre precisión, estabilidad y facilidad de interpretación.

En definitiva, los resultados finales no deben entenderse únicamente como un modelo predictivo exitoso, sino como la base de una solución práctica y escalable, con impacto directo en la gestión del mercado turístico y con capacidad de extenderse a otros destinos urbanos.

Capítulo 5. DISCUSIÓN

El desarrollo de este proyecto ha permitido obtener un modelo predictivo robusto para estimar precios de alojamientos en Airbnb Madrid. No obstante, más allá de los resultados cuantitativos alcanzados, es necesario reflexionar sobre los alcances, limitaciones y aprendizajes obtenidos durante el proceso.

5.1 Utilidad y pertinencia de la metodología

La metodología CRISP-DM aplicada demostró ser adecuada para estructurar el proyecto en fases claras: comprensión, preparación, análisis, modelado y evaluación. Esta organización permitió mantener una línea de trabajo coherente y replicable.

- La fase de preparación de datos fue especialmente crítica, dado el volumen de inconsistencias, outliers y datos nulos presentes en los datasets originales de Inside Airbnb.
- El enfoque modular de CRISP-DM permitió introducir cambios sin perder coherencia, como la decisión de realizar eliminaciones de outliers tanto globales como por barrio.

En términos generales, la metodología inicial ha sido útil y aplicable, aunque requirió ajustes para adaptarse a la naturaleza particular del dataset.

5.2 Limitaciones del estudio

Aunque el modelo Random Forest optimizado alcanzó métricas sobresalientes en la predicción de precios de Airbnb en Madrid, es importante reconocer ciertas limitaciones que enmarcan el alcance del trabajo:

- **Cobertura temporal reducida:** los datos analizados corresponden a un periodo específico y no incluyen históricos suficientemente amplios para capturar de manera robusta la evolución de los precios en el largo plazo ni los ciclos completos del turismo en Madrid.
- **Ausencia de variables exógenas:** factores externos como eventos culturales y deportivos, variaciones macroeconómicas, o situaciones excepcionales (ej. la pandemia) no fueron incorporados en el modelo. Estos elementos tienen un impacto directo en la fijación de precios y podrían enriquecer las predicciones futuras.
- **Heterogeneidad en barrios premium:** en zonas exclusivas como Jerónimos, Lista, Goya o El Viso, el modelo presentó mayor error absoluto promedio. Esto se debe tanto a la menor disponibilidad de registros representativos como a la alta dispersión de precios, propia de mercados de lujo donde los valores varían significativamente según características diferenciales del alojamiento.

Estas limitaciones no desmerecen los resultados obtenidos, pero sí señalan líneas de mejora para futuros desarrollos, permitiendo avanzar hacia un modelo aún más generalizable, robusto y adaptable a dinámicas cambiantes del mercado turístico.

5.3 Limitaciones tecnológicas

El desarrollo del proyecto se llevó a cabo íntegramente con herramientas open-source, como Python, scikit-learn, XGBoost, CatBoost, Random Forest y la plataforma Streamlit para la interfaz. Este enfoque garantizó la viabilidad, accesibilidad y replicabilidad del trabajo, aunque también presenta ciertas limitaciones desde el punto de vista tecnológico:

- **Escalabilidad:** el modelo ha demostrado un excelente rendimiento al trabajar con datos de Madrid, pero para extenderlo a múltiples ciudades o entornos de predicción en tiempo real sería necesario optimizar tanto la gestión de datos como la eficiencia computacional.
- **Infraestructura:** el entrenamiento se realizó en entornos locales y gratuitos, sin acceso a recursos avanzados de computación en la nube. La incorporación de plataformas cloud (como AWS, GCP o Azure) permitiría acelerar los procesos de entrenamiento, manejar volúmenes de datos mayores y facilitar un despliegue más sofisticado.
- **Automatización de actualizaciones:** aunque la plataforma desarrollada permite consultar precios de manera práctica, los datos provienen de extracciones puntuales. Integrar procesos automáticos de recolección y actualización de datos (mediante APIs o pipelines ETL) supondría un avance significativo para mantener el modelo actualizado en todo momento.

En conjunto, estas limitaciones marcan oportunidades de mejora tecnológica que pueden reforzar la usabilidad y escalabilidad del modelo en futuros desarrollos.

5.4 Adaptaciones y cambios en el desarrollo

Durante el proyecto se realizaron ajustes relevantes respecto a los objetivos iniciales:

- **Reducción del alcance:** se acotó el trabajo a un modelo enfocado exclusivamente en la predicción de precios, descartando otros objetivos planteados inicialmente como la segmentación por capacidad o el análisis diferenciado de tipologías de usuarios.
- **Implementación de una interfaz interactiva en Streamlit:** esta decisión transformó el proyecto de un análisis puramente académico en una herramienta práctica y accesible para distintos perfiles de usuario.

Estos cambios fueron necesarios y enriquecieron el producto final, alineándolo mejor con las necesidades reales de anfitriones, turistas e inversores.

5.5 Impacto y valor del proyecto

El resultado del proyecto tiene un impacto directo en distintos actores del mercado turístico:

- Para anfitriones, el modelo facilita fijar precios competitivos que maximizan la ocupación.
- Para turistas, mejora la transparencia y permite comparar alternativas con facilidad.
- Para inversores, ofrece una herramienta que ayuda a identificar zonas con mayor potencial de rentabilidad.

Más allá de lo técnico, el proyecto refleja la capacidad de aplicar Data Science a un problema real, aportando valor tanto académico como práctico.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

El objetivo general del proyecto fue desarrollar un sistema de predicción inteligente de precios de alojamientos en Airbnb Madrid, integrando variables relacionadas con la ubicación, características del inmueble, anfitriones y factores estacionales.

Los resultados obtenidos permiten concluir que:

- El modelo de Random Forest optimizado alcanzó un desempeño sobresaliente, con un MAE de 6,34 €, un RMSE de 8,75 € y un R^2 de 0,975, lo que lo posiciona como la mejor opción frente a CatBoost y XGBoost.
- El modelo es capaz de predecir precios con un alto nivel de fiabilidad, ofreciendo estimaciones precisas y consistentes para la toma de decisiones en un mercado altamente competitivo como el de Madrid.
- La metodología aplicada (limpieza de datos, eliminación de outliers, incorporación de estacionalidad, ajuste de hiperparámetros y comparación de algoritmos) ha demostrado ser efectiva para garantizar la calidad y robustez de los resultados.
- El sistema desarrollado beneficia a distintos actores: anfitriones, que pueden fijar precios más competitivos; turistas, que disponen de referencias más justas; e inversores, que pueden identificar oportunidades de rentabilidad en diferentes barrios.

En síntesis, se ha cumplido plenamente con el objetivo general y los objetivos específicos del proyecto, validando la utilidad de modelos de Machine Learning para optimizar la gestión de precios en plataformas de economía colaborativa.

6.2 Conclusiones personales

El desarrollo de este trabajo ha representado un reto académico y profesional significativo. La complejidad de los datos, la necesidad de adaptar metodologías y la comparación de distintos modelos han enriquecido mi experiencia como futura profesional en Data Science.

De forma personal, este proyecto me ha permitido:

- Consolidar conocimientos técnicos en análisis de datos, modelado predictivo y validación de resultados.
- Desarrollar habilidades de gestión de proyectos, organización del trabajo y toma de decisiones ante imprevistos.
- Confirmar la relevancia de aplicar la ciencia de datos en contextos reales, donde las soluciones no solo generan valor económico, sino que también tienen un impacto social en la experiencia de usuarios y en la sostenibilidad del turismo.

En lo personal, este trabajo ha sido un proceso de aprendizaje profundo y de crecimiento, reforzando mi interés por seguir desarrollándome en el ámbito del análisis predictivo y la inteligencia artificial aplicada a problemas del mundo real.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Durante el desarrollo del presente proyecto se identificaron diversas oportunidades de mejora y ampliación que, por cuestiones de alcance y tiempo, no pudieron abordarse, pero que representan caminos prometedores para potenciar el valor del modelo desarrollado:

- **Integración de datos en tiempo real:** conectar la solución con la API de Airbnb u otras fuentes de datos externas para actualizar continuamente precios, disponibilidad y tendencias del mercado.
- **Ampliación del alcance geográfico:** extender el modelo a otras ciudades españolas o europeas con alto flujo turístico, permitiendo comparativas más amplias y robustas.
- **Incorporación de variables adicionales:** incluir factores externos como festividades locales, grandes eventos, datos de transporte o nivel de contaminación, que pueden influir en la demanda y los precios.
- **Desarrollo de un sistema de recomendación completo:** evolucionar el prototipo hacia una herramienta interactiva para anfitriones, inversores y turistas, con funcionalidades de simulación de escenarios y recomendaciones personalizadas.

En conclusión, este trabajo abre la puerta a un amplio espectro de líneas de investigación y desarrollo, que podrían consolidar una solución más integral, escalable y con mayor aplicabilidad práctica en la toma de decisiones dentro del mercado de alquiler turístico.

Capítulo 8. REFERENCIAS

[1] Inside Airbnb. Get the Data. Disponible en: http://insideairbnb.com/get-the-data.html [Accedido: 15-jul-2025].

[2] Scikit-learn. Machine Learning in Python. Disponible en: https://scikit-learn.org/stable/ [Accedido: 20-jul-2025].

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.

[4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin. "CatBoost: unbiased boosting with categorical features." Advances in Neural Information Processing Systems (NeurIPS), 2018.

[5] J. Friedman, T. Hastie, R. Tibshirani. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[6] W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference (SciPy), pp. 56–61, 2010.

[7] Seaborn. Statistical Data Visualization. Disponible en: https://seaborn.pydata.org/ [Accedido: 20-jul-2025].

[8] Plotly. Plotly Python Graphing Library. Disponible en: https://plotly.com/python/ [Accedido: 25-jul-2025].

[9] Streamlit. The fastest way to build and share data apps. Disponible en: https://streamlit.io/ [Accedido: 10-ago-2025].

[10] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1145, 1995.

[11] M. Wirth, "CRISP-DM: Towards a Standard Process Model for Data Mining," Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD), 2000.

[12] AENOR, Norma UNE 157001: Criterios generales para la elaboración de proyectos. Madrid: AENOR, 2002.

[13] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[14] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[15] K. Jordahl et al., “GeoPandas: Python tools for geographic data,” Journal of Open Source Software, 5(51), 2157, 2020.

Capítulo 9. ANEXOS

Todo el material complementario del proyecto (código fuente, notebooks, datasets procesados y visualizaciones adicionales) se encuentra disponible en el repositorio de GitHub asociado a este trabajo:

🔗 Repositorio GitHub – Predicción de precios Airbnb Madrid