UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER DE FORMACIÓN PERMANENTE EN BUSINESS ANALYTICS

PROYECTO FIN DE MÁSTER

MODELIZACIÓN Y PREDICCIÓN DEL GASTO TURÍSTICO DESDE EL ANÁLISIS DE DATOS: EL CASO DE CANARIAS

PAULA HERNÁNDEZ RODRÍGUEZ

ERNESTO RODRÍGUEZ GONZÁLEZ

Dirigido por

CARLOS NIETO LÓPEZ VILLALÓN

CURSO 2023-2024

TÍTULO: Modelización y predicción del gasto turístico desde el análisis de datos: el caso de Canarias

AUTOR: PAULA HERNÁNDEZ RODRÍGUEZ y ERNESTO RODRÍGUEZ GONZÁLEZ

TITULACIÓN: Máster de Formación Permanente en Business Analytics

DIRECTOR/ES DEL PROYECTO: CARLOS NIETO LÓPEZ VILLALÓN

FECHA: [JULIO] de 2024

RESUMEN

El objetivo principal de este trabajo es desarrollar un modelo predictivo del gasto turístico en la Comunidad Autónoma de Canarias por parte de turistas provenientes de países de la Eurozona, utilizando técnicas de análisis de datos y Machine Learning. En primer lugar, se realizó una descripción demográfica, climática y económica de Canarias, luego se llevaron a cabo los análisis univariante y bivariante y por último se procedió al modelo predictivo. En este sentido, usando datos de la encuesta de gasto turístico (EGATUR), del Banco de España y de EUROSTAT desde octubre de 2015 hasta marzo de 2024 se evaluó la capacidad predictiva -mediante validación cruzada- de tres algoritmos: Linear Regression, K-Nearest Neighbors y Random Forest, y se analizó la importancia de las variables en el modelo preferido. Con los resultados se concluye que en este estudio el modelo que tiene un mejor desempeño es el Random Forest, aunque las métricas evaluativas (RMSE, MAE y R^2) en los tres modelos resultan ser similares. Por su parte, los descriptores con mayor relevancia para explicar el gasto turístico son el número de pernoctaciones y el tipo de alojamiento mientras que las otras variables, como el EURIBOR, el transporte y el motivo del viaje, presentan una importancia marginal. Es importante destacar que este período incluye eventos significativos como la pandemia mundial y fenómenos naturales como la erupción del volcán de La Palma, los cuales han tenido un impacto notable en el sector turístico y en los patrones de gasto de los turistas.

Palabras clave: Machine Learning, turismo, Canarias, gasto total

ABSTRACT

The main objective of this work is to develop a predictive model of tourism expenditure in the Autonomous Community of the Canary Islands by tourists from Eurozone countries, using data analysis and Machine Learning techniques. First, a demographic, climatic and economic description of the Canary Islands was carried out, then univariate and bivariate analyses were performed and finally the predictive model was developed. In this sense, using data from the tourism expenditure survey (EGATUR), the Bank of Spain and EUROSTAT from October 2015 to March 2024, the predictive capacity was evaluated -by cross validation- of three algorithms: Linear Regression, K Nearest Neighbors and Random Forest, and the importance of the variables in the preferred model was analyzed. With the results it is concluded that in this study the best performing model is the Random Forest, although the evaluative metrics (RMSE, MAE and R^2) in the three models turn out to be similar. On the other hand, the descriptors with the greatest relevance in explaining tourism expenditure are the number of overnight stays and type of accommodation, while the other variables, such as EURIBOR, transportation and reason for the trip, are of marginal importance. It is important to note that this period includes significant events such as the world pandemic and natural phenomena such as the eruption of the La Palma volcano, which have had a notable impact on the tourism sector and tourist spending patterns.

Keywords: Machine Learning, tourism, Canary Islands, total spending

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Paula Hernández Rodríguez y Ernesto Rodríguez González
Título del proyecto:	Modelización Y Predicción Del Gasto Turístico Desde El Análisis De Datos: El Caso De Canarias
Directores del proyecto:	Carlos Nieto López Villalón
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto:	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SÍ
Objetivo general del proyecto:	El objetivo general del proyecto es desarrollar un modelo predictivo del gasto turístico en Canarias por parte de turistas provenientes de países de la Eurozona, utilizando técnicas de análisis de datos y Machine Learning.

Índice

Índice de Imágenes

Índice de Tablas

Índice de Gráficos

Índice de Abreviaturas

Siglas	**Aclaración terminológica**
C.A	Comunidad Autónoma
DL	Deep Learning (Aprendizaje profundo)
KNN	K-Nearest Neighbors (K Vecinos más Cercanos)
MAE	Mean Absolute Error (Error Absoluto Medio)
RF	Random Forest
LR	Linear Regression (Regresión Lineal)
RMSE	Root Mean Squared Error (Raíz del Error Cuadrático Medio)
SVM	Support Vector Machine (Máquina de Vectores de Soporte)

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El proyecto de investigación se enmarca en la significativa contribución económica del turismo en la Comunidad Autónoma de Canarias, representando el 35,5% del PIB regional y empleando al 39,7% de la fuerza laboral. Esta dependencia económica destaca la necesidad de desarrollar modelos avanzados de predicción del gasto turístico, especialmente utilizando técnicas de Machine Learning y siendo fundamentales para asegurar la sostenibilidad y el crecimiento continuo del sector. Por ejemplo, los sistemas de recomendación personalizados pueden analizar los datos de preferencias de los turistas para sugerir actividades y atracciones acordes a sus intereses. Además, los algoritmos de Machine Learning pueden optimizar la gestión de recursos en hoteles, prediciendo la ocupación y ajustando la asignación de personal.

1.2 Planteamiento del problema

El turismo en Canarias es vital para su economía, pero la mayoría de la literatura académica ha priorizado el estudio de las llegadas de turistas sobre el análisis del gasto turístico. Aunque algunas investigaciones han aplicado técnicas econométricas clásicas para predecir el gasto, hay una notable carencia de estudios que emplean técnicas avanzadas de Machine Learning, que integren variables macroeconómicas y que se centren exclusivamente en Canarias. Esta brecha de conocimiento subraya la necesidad de desarrollar un modelo predictivo robusto que comprenda mejor los determinantes del gasto turístico en la región, ofreciendo así herramientas innovadoras para la gestión estratégica y la toma de decisiones en el sector turístico canario. Dichas herramientas innovadoras, apoyadas en la gestión de los datos gracias al big data, incluyen técnicas avanzadas como Random Forest, Linear Regression y K-Nearest Neighbors. Estas técnicas permiten analizar grandes volúmenes de datos y extraer patrones significativos, lo que nos proporciona una comprensión más profunda de los comportamientos y preferencias de los turistas. Con estos conocimientos, podemos optimizar la oferta turística, mejorar la satisfacción del cliente y, en última instancia, impulsar el crecimiento sostenible del sector turístico en Canarias.

1.3 Objetivos del proyecto

El proyecto tiene como objetivo principal desarrollar un modelo predictivo del gasto turístico en Canarias por parte de turistas de países de la Eurozona, utilizando técnicas avanzadas de análisis de datos y Machine Learning.

1.4 Resultados obtenidos

Los modelos comparados (Random Forest, Linear Regression y KNN) tuvieron resultados similares, sin embargo, destacó el modelo RF como el más efectivo. Además, se identificó que el número de pernoctaciones y el tipo de alojamiento son los principales impulsores del gasto turístico.

1.5 Estructura de la memoria

El trabajo está estructurado de la siguiente manera: en el capítulo 2 se presenta una revisión de la literatura y se justifica la temática del estudio. En el capítulo 3 se explican los objetivos, tanto el general como los específicos. En el capítulo 4 se describe la base de datos, las variables y los modelos utilizados y se muestran los resultados de los modelos estimados. A continuación, en el capítulo 5 se discuten los mismos. En el capítulo 6 se encuentran las conclusiones del estudio. Finalmente, en el capítulo 7 se indican las futuras líneas de trabajo.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

El turismo es la actividad económica más importante de las Islas Canarias. Para comprender mejor este fenómeno y el comportamiento de los turistas, este estudio se enfoca en analizar las variables que afectan el gasto turístico en las Islas Canarias.

La mayoría de la literatura académica aborda el estudio de la demanda turística en términos de llegadas (Albaladejo et al., 2016; Álvarez-Díaz et al., 2014; Álvarez-Díaz et al., 2016; Sánchez & Sánchez, 2021) Sin embargo, los análisis de la demanda turística centrados específicamente en el gasto, que es el objeto de este trabajo, son considerablemente menos comunes. Además, la gran mayoría de los estudios emplean técnicas estadísticas/econométricas avanzadas, pero el uso de técnicas de aprendizaje automático (Machine Learning) es notablemente escaso. No obstante, entre las investigaciones que intentan modelizar el gasto turístico con ML, destacan dos en particular:

Piña (2018) utilizando datos de la base de datos de la Encuesta de Gasto Turístico (EGATUR) desde octubre de 2015 hasta agosto de 2017, utiliza algoritmos de Linear Regression (un método estadístico que modela relaciones entre variables), Random Forest (que construye múltiples árboles de decisión para mejorar la precisión), Support Vector Machines (que encuentra el mejor hiperplano para clasificar los datos), Deep Learning (redes neuronales con múltiples capas que capturan patrones complejos) y KNN (clasifica los datos en función de sus vecinos más cercanos) para predecir el gasto turístico. En ese estudio, los resultados del desempeño de los modelos predictivos indicaron que siguen el siguiente orden según RMSE:

$$DL<RF<SVM<KNN<LR$$

Donde:

DL es Deep Learning, RF es Random Forest, SVM es Support Vector Machine, KNN es K Nearest Neighbors y LR es Linear Regression.

Además, la autora señala que, según el análisis de la importancia de diferentes descriptores, el principal factor que influye en el gasto del turista es el número de pernoctaciones. Le sigue en importancia el país de procedencia y con menor importancia el tipo de alojamiento y la Comunidad Autónoma visitada. Otros descriptores muestran una relevancia marginal para predecir el gasto.

Por su parte y en línea con estos hallazgos, Moreno (2023), utilizando datos de EGATUR de los años 2019 y 2022 y empleando técnicas de Machine Learning tanto de regresión como de clasificación, encontró que el número de pernoctaciones es el factor de mayor relevancia en el estudio del gasto turístico. Para ello, se utilizaron técnicas como feature selection (selección de características, un proceso que identifica las variables más importantes) y mutual information (información mutua, que mide la dependencia entre variables). Además, mediante la técnica de

selección de características, se observó que, para ambos períodos (2019 y 2022), tanto las pernoctaciones como las variables "país" y "medio de transporte" son las más relevantes para explicar el gasto turístico.

Sin embargo, hasta donde llega nuestro conocimiento, ningún estudio anterior ha examinado el gasto turístico incluyendo a la encuesta de gasto turístico variables macroeconómicas y centrándose solo en la Comunidad Autónoma de Canarias.

2.2 Contexto y justificación

2.2.1. Contexto espacial: Las Islas Canarias

Las Islas Canarias, un archipiélago compuesto por ocho islas situadas en el Océano Atlántico (Imagen 1), son una región ultraperiférica de la Unión Europea y una Comunidad Autónoma de España. Localizadas a 1.300 km de la Península Ibérica y a unos 100 km de la costa africana, forman el mayor y más oriental de los cinco archipiélagos de la Macaronesia.

Imagen 1. Islas Canarias

Fuente: https://es.wikipedia.org/wiki/Geograf%C3%ADa_de_Canarias

Con una superficie de 7.447 km² y una población de 2.172.944 habitantes (INE, 2021), Tenerife y Gran Canaria son las islas más pobladas, representando en torno al 80% de la población total.

Desde un punto de vista medioambiental, las Islas Canarias son un territorio único en el mundo, caracterizado por (Hernández et al., 2023):

- Condiciones climáticas suaves: La temperatura media durante el día es de 22,1°C en invierno y 26°C en verano. El clima en las islas está influenciado principalmente por su condición insular y su ubicación geográfica cercana al trópico de Cáncer. La presencia del mar contribuye significativamente a que el clima canario sea distinto del que se

esperaría por su latitud, que de otro modo sería extremadamente seco. Otro factor determinante en el clima del archipiélago es la influencia de los vientos alisios. Estos vientos, originados cerca de las islas Azores, son moderados y regulares, cálidos y secos en su origen, pero adquieren humedad y se refrescan al cruzar el océano. Durante el verano, las Islas Canarias están particularmente bajo la influencia de estos vientos.

- Origen volcánico: Las islas Canarias son la única región de España con volcanismo activo, donde ha habido erupciones volcánicas recientes y donde existe riesgo de que ocurran más en el futuro. Tenerife, La Palma, Lanzarote y El Hierro han tenido erupciones en los últimos siglos y son volcánicamente activas, con la erupción más reciente en la isla de La Palma en 2021.
- Riqueza ambiental y diversidad de hábitats: Casi la mitad del territorio canario está protegido por la red de áreas de conservación de la Unión Europea, la Red Natura 2000.
- Paisaje variado y contrastante: Las islas más orientales, Lanzarote y Fuerteventura, son más áridas, mientras que las occidentales presentan paisajes montañosos y diversos.
- Diversidad biológica: Canarias es el centro de diversidad biológica más importante de la Unión Europea y uno de los 25 puntos calientes de biodiversidad en el mundo, con aproximadamente el 38% de las especies de fauna endémicas del archipiélago y más de la mitad de los endemismos vegetales de España. Además, alberga cerca del 80% de las especies de cetáceos del Atlántico Norte, como ballenas, delfines y zifios.

En concreto, las especies vegetales en los ecosistemas terrestres de Canarias están organizadas en comunidades que se distribuyen en diferentes niveles de vegetación según el clima y la altitud. Los principales ecosistemas que se pueden encontrar en las islas incluyen (Del Arco, 2006):

1. Matorral costero: Desde el nivel del mar hasta los 350 m de altitud, este ecosistema está representado principalmente por especies como las tabaibas y los cardones, que son arbustos y matorrales suculentos.
2. Bosques termófilos: Situados por encima del matorral costero, estos bosques incluyen especies como las sabinas, la palmera canaria y los dragos.
3. Laurisilva: Este es el ecosistema más complejo de Canarias. La laurisilva se encuentra en Tenerife (macizos de Anaga y Teno), La Palma (Los Tilos), El Hierro (escarpes del Golfo), y La Gomera, donde el fayal-brezal predomina en la zona central de la isla (bosque del Cedro). Las zonas superiores de la laurisilva están dominadas por fayas y brezos.
4. Pinar: Estos bosques están dominados por el pino canario, una especie que puede rebrotar de cepa y regenerarse rápidamente tras un incendio. Esta formación es altamente adaptable a condiciones adversas (suelos ácidos, erosionados o pedregosos) y crece en zonas con climas diversos, soportando bien tanto las altas como las bajas temperaturas (incluidas las heladas), así como diferentes niveles de precipitaciones.
5. Matorral de cumbre: Este ecosistema se encuentra solo en las dos islas más altas, Tenerife y La Palma, y se localiza por encima de los 2.000 metros de altitud. Las especies vegetales dominantes en este ecosistema son las retamas y los codesos.

En cuanto a la infraestructura y a la conectividad de las islas (Hernández et al., 2023):

- Canarias cuenta con 8 aeropuertos: el de Gran Canaria, el de Tenerife Sur Reina Sofía, el de Lanzarote - César Manrique, el de Tenerife Norte, el de Fuerteventura, el de La Palma, el de El Hierro y el de La Gomera. Y con 9 puertos de pasajeros: el de S/C de Tenerife, el de Los Cristianos, el de S/C de La Palma, el de San Sebastián de La Gomera, el de la Estaca en El Hierro, el de Las Palmas, el de Agaete, el de Arrecife y el de El Rosario.
- Principales aerolíneas: Ryanair es la principal aerolínea que opera en las islas por volumen de pasajeros internacionales, seguida de Jet2 y el Grupo TUI.
- Destinos turísticos inteligentes: Canarias cuenta con siete destinos turísticos inteligentes que forman parte de la Red de Destinos Turísticos Inteligentes. Tenerife, Puerto de la Cruz, Las Palmas de Gran Canaria y Arona se unieron en 2019, mientras que Tías, Santa Cruz de Tenerife y Puerto del Rosario se incorporaron en 2020.

Con respecto a los datos socioeconómicos relevantes:

- Canarias tiene una elevada tasa de paro, siendo una de las más altas de España y Europa.
- Canarias se encuentra entre las regiones españolas con menor PIB per cápita, esto está relacionado con el alto nivel de desempleo, la falta de cualificación educativa, el escaso emprendimiento y las dificultades para integrar el mercado interior y promover actividades económicas más allá del turismo. Por lo tanto, a pesar del liderazgo del sector turístico canario en el mercado europeo, los indicadores de renta, pobreza y empleo han situado tradicionalmente a las islas entre las regiones más pobres de España y Europa.
- La comunidad autónoma realiza un bajo gasto por habitante en I+D y muestra una alta dependencia de universidades y la Administración Pública en este ámbito. Además, es una de las regiones con menor capacidad de atracción y retención de talento.
- Canarias mantiene una alta dependencia de energía procedente del exterior y basada en combustibles fósiles.

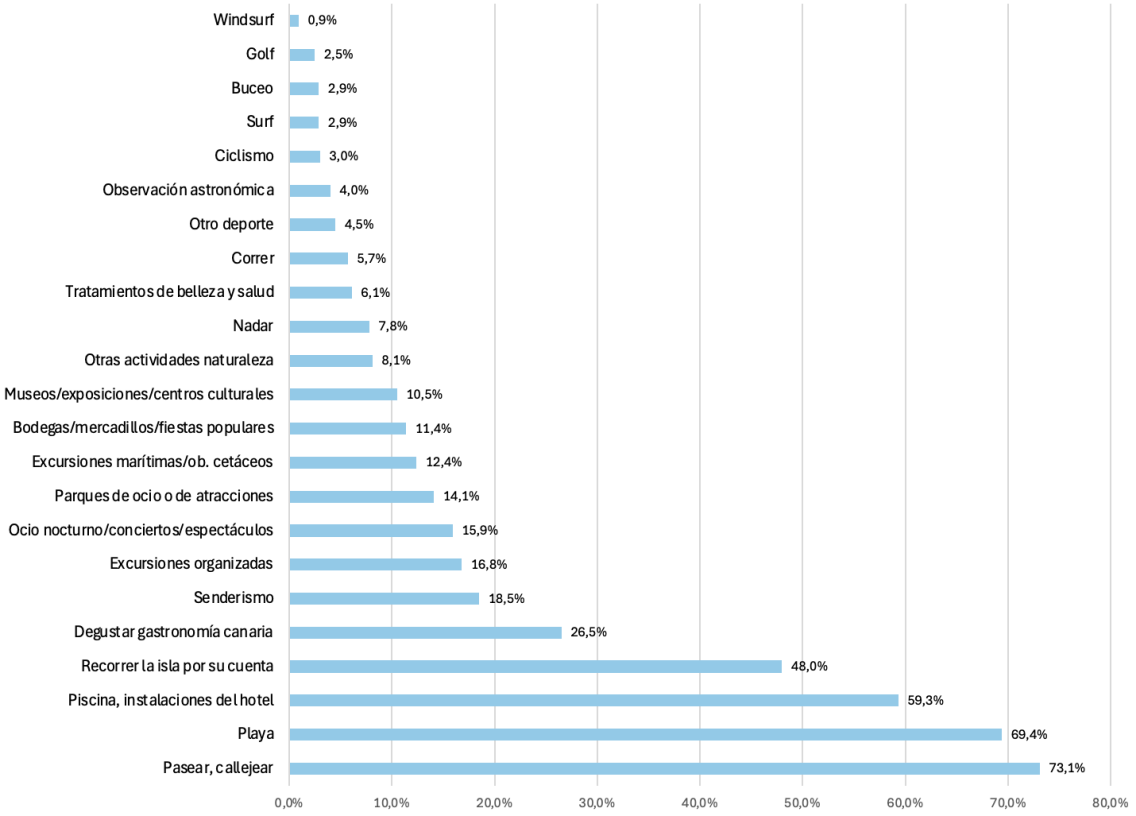
2.2.2. Contexto turístico y justificación

La contribución histórica del turismo en Canarias ha sido significativa, impulsada por su capacidad de impacto en otros sectores productivos de la isla. Según el informe más reciente de IMPACTUR, el PIB turístico representó el 35,5% de la economía canaria en 2022, equivalentes a 16.961 millones de euros, y el empleo turístico alcanzó el 39,7% del total del empleo en la región, con 344.358 puestos de trabajo. Además, IMPACTUR señala que, por cada 100 euros de valor añadido generado directamente por la demanda turística, se aportan 44,9 euros en otros sectores (efectos indirectos), y por cada 100 empleos creados en actividades directamente relacionadas con el turismo, se generan 38 empleos adicionales en otros sectores productivos (EXCELTUR, 2023). Estos números destacan la dependencia de Canarias del turismo y la necesidad de comprender este sector para asegurar su sostenibilidad y crecimiento.

El modelo turístico de Canarias ha dependido en gran medida de sus recursos naturales y ha integrado poco la inteligencia y el conocimiento en su desarrollo. Esta región ha visto su crecimiento turístico impulsado por condiciones climáticas excepcionales a nivel mundial, así

como por su riqueza natural y paisajística. Tradicionalmente, el crecimiento del turismo en Canarias se ha basado en la producción a gran escala y en atraer un alto volumen de visitantes (Simancas et al., 2016). Destaca, entonces, el turismo de sol y playa, un segmento del sector turístico que se centra en ofrecer experiencias relacionadas con el disfrute del sol, el mar y las playas (Gráfico 1).

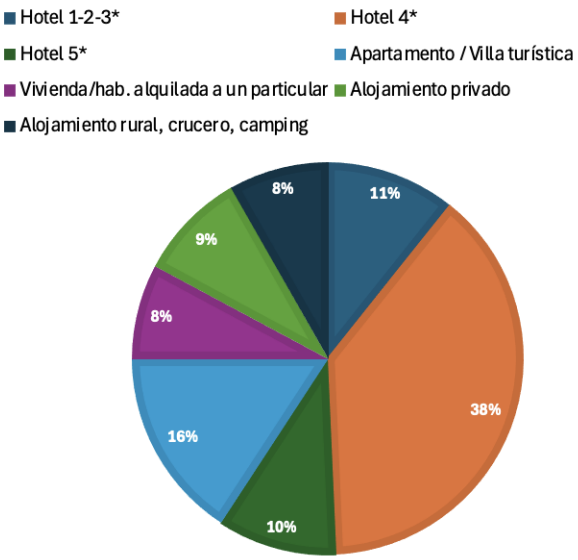
Gráfico 1. Actividades realizadas por los turistas durante su estancia en 2023



Fuente: PROMOTUR (2023)

Este tipo de turismo es particularmente popular en destinos que poseen climas cálidos y soleados durante gran parte del año, así como playas atractivas y bien mantenidas. Canarias, con su clima favorable, playas exóticas y paisajes naturales únicos, es uno de los destinos más destacados para el turismo de sol y playa. Esto también se ve reflejado en el tipo de alojamiento que contratan los turistas (Gráfico 2), en el que casi el 50% se aloja en hoteles de 4 o 5 estrellas.

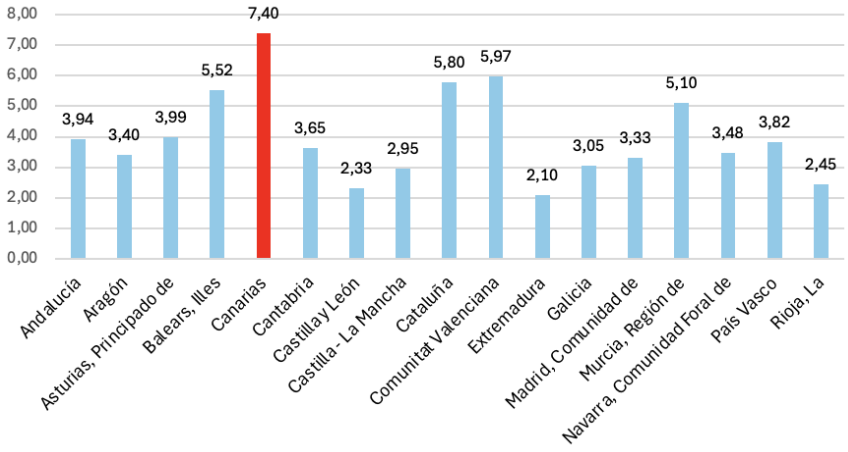
Gráfico 2. Tipo de alojamiento del turista



Fuente: PROMOTUR (2023)

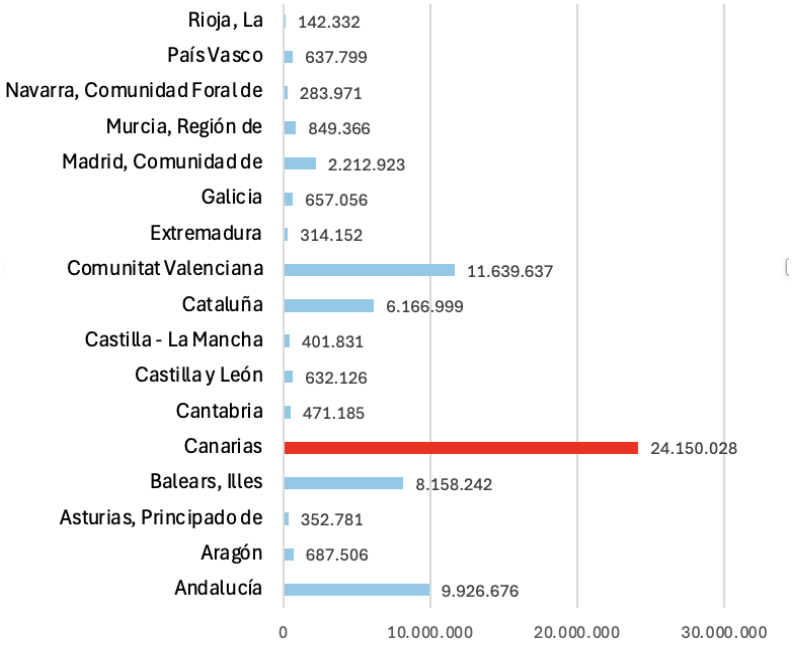
En este contexto, en el ámbito nacional, Canarias lidera la estancia media (Gráfico 3) y el número de pernoctaciones de los viajeros (Gráfico 4) con datos de 2022.

Gráfico 3. Estancia media de los turistas por CC.AA 2022



Fuente: INE (2022).

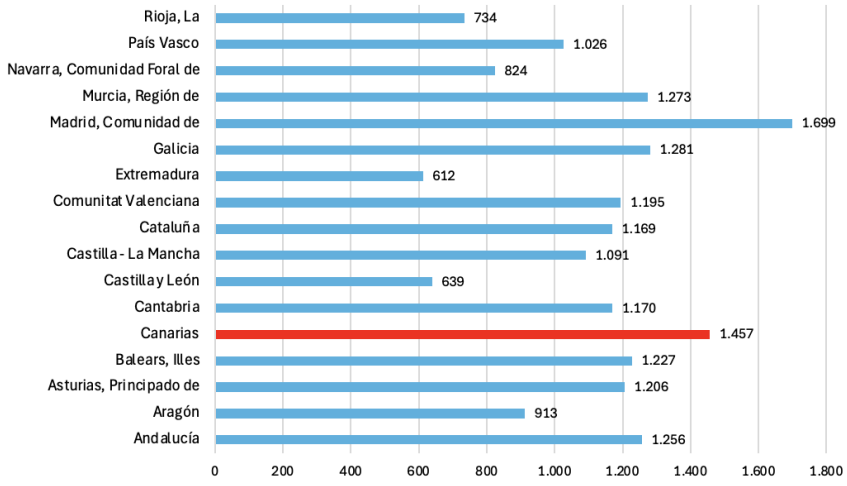
Gráfico 4. Pernoctaciones de los viajeros por CC.AA en 2022



Fuente: INE (2022).

Por su parte, en cuanto al gasto medio por turista en 2023, Canarias se encuentra en la segunda posición nacional, después de Madrid (Gráfico 5).

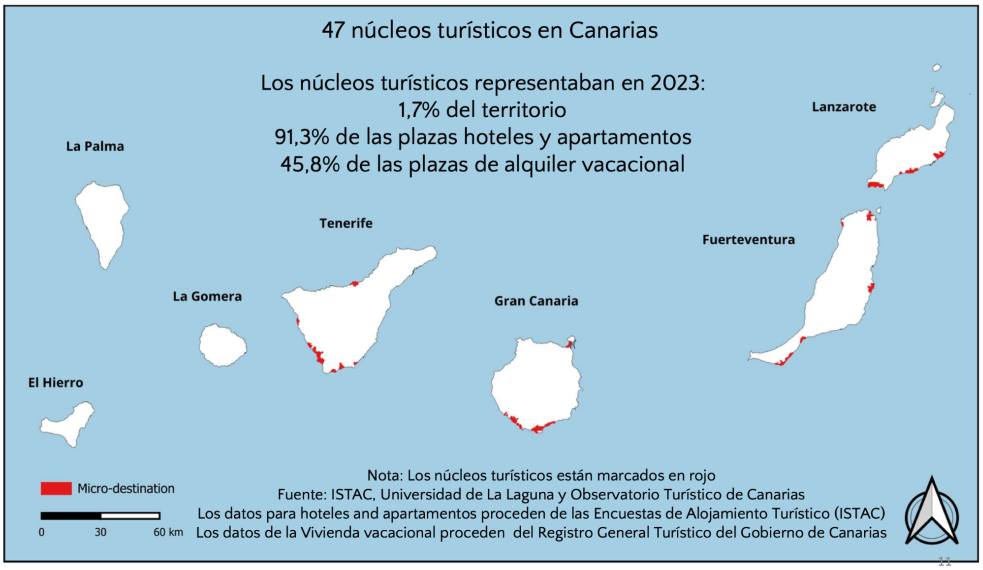
Gráfico 5. Gasto medio por turista por CC.AA en 2023



Fuente: INE (2023).

En cuanto a la distribución y concentración del alojamiento turístico en Canarias, Hernández et al. (2024) indican que:

Gráfico 6. Distribución alojamiento turístico en Canarias



Fuente: Hernández et al. (2024)

Los núcleos se concentran mayoritariamente en las islas orientales (Fuerteventura, Lanzarote y Gran Canaria) y en el sur de Tenerife.

Por lo tanto, dado los anteriores datos y visto que el turismo es un motor económico vital para la región, es esencial disponer de modelos predictivos precisos que permitan a los responsables de políticas y a los actores del sector turístico tomar decisiones informadas. Además, existe una notable carencia de investigaciones centradas específicamente en el gasto turístico, especialmente utilizando técnicas de Machine Learning. Este proyecto abordará esa brecha, proporcionando una perspectiva novedosa y metodológicamente avanzada sobre el tema.

2.3 Planteamiento del problema

El turismo representa un pilar fundamental para la economía de Canarias, siendo crucial para su desarrollo socioeconómico. A pesar de su significativa contribución, la mayoría de los estudios han enfocado tradicionalmente su atención en la cantidad de visitantes, dejando de lado un aspecto igualmente crucial: el gasto turístico. Este último aspecto es fundamental no solo para comprender mejor el impacto económico del turismo en la región, sino también para optimizar las estrategias de gestión y promoción turística.

El análisis del estado del arte revela que existen limitaciones significativas en cuanto a la comprensión del comportamiento de gasto de los turistas en Canarias. Aunque se han realizado estudios utilizando técnicas estadísticas y econométricas avanzadas, el uso de métodos de aprendizaje automático, que podrían ofrecer una visión más profunda y predictiva, es escaso. Además, la mayoría de las investigaciones previas se han centrado en variables específicas como las pernoctaciones y el país de origen, sin explorar la influencia de variables macroeconómicas más amplias que podrían estar afectando el comportamiento de gasto de los turistas.

Por lo tanto, el planteamiento del problema radica en la necesidad de desarrollar un modelo predictivo robusto del gasto turístico en Canarias, utilizando técnicas avanzadas de análisis de datos y Machine Learning. Este modelo no solo permitirá llenar el vacío actual de conocimiento sobre los determinantes del gasto turístico en la región, sino que también proporcionará herramientas efectivas para mejorar la planificación estratégica y la toma de decisiones en el sector turístico canario.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo general del proyecto es desarrollar un modelo predictivo del gasto turístico en Canarias por parte de turistas provenientes de países de la Eurozona, utilizando técnicas de análisis de datos y Machine Learning. Al implementar estas técnicas innovadoras, este estudio busca proporcionar una herramienta útil para la planificación y gestión del turismo en Canarias, permitiendo una mejor toma de decisiones basada en datos sólidos y mejorando la capacidad de anticiparse a las necesidades y tendencias del mercado turístico.

3.2 Objetivos específicos

Los objetivos específicos de este proyecto son los siguientes:

- Analizar la correlación entre las variables principales y el gasto turístico en el período comprendido entre 2015 y 2024, con el fin de identificar patrones y relaciones significativas que puedan influir en el comportamiento de gasto de los turistas.
- Encontrar los modelos predictivos que mejor representen el gasto turístico, considerando las variables más explicativas, y evaluar su rendimiento y precisión para asegurar que se elige la metodología más adecuada para este contexto.
- Identificar la importancia relativa de cada variable en el modelo predictivo, determinando en qué medida cada factor contribuye a la predicción del gasto turístico y proporcionando así una comprensión más profunda de los determinantes clave del gasto en el sector turístico de Canarias.

3.3 Beneficios del proyecto

El desarrollo de un modelo predictivo del gasto turístico en Canarias por parte de turistas provenientes de países de la Eurozona ofrecerá múltiples beneficios en relación con los objetivos planteados.

En primer lugar, al desarrollar un modelo predictivo del gasto turístico utilizando técnicas avanzadas de análisis de datos y Machine Learning, se facilita una planificación más precisa y efectiva. Esto permite a los gestores y responsables de políticas turísticas adaptar estrategias que maximicen el impacto económico positivo del turismo en la región.

Además, el análisis detallado de la correlación entre las variables principales y el gasto turístico proporcionará *insights* significativos sobre los factores que influyen en las decisiones de gasto de los turistas. Este conocimiento no solo permitirá una mejor comprensión de los comportamientos del consumidor en el contexto turístico de Canarias, sino que también facilitará la identificación de oportunidades para optimizar la experiencia del turista y aumentar su satisfacción.

Por último, al abordar una brecha notable en la literatura existente sobre el gasto turístico mediante el uso de técnicas de Machine Learning, este proyecto contribuirá al enriquecimiento

del conocimiento académico, proporcionando una metodología avanzada que podrá ser utilizada y mejorada en estudios futuros, tanto en Canarias como en otras regiones turísticas.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

La planificación del proyecto ha sido la siguiente:

1. Análisis de variables influyentes para el gasto turístico (Duración estimada: 4 semanas)

1.1. Revisión bibliográfica sobre variables que influyen en el gasto turístico.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Revisión de literatura académica y estudios previos para identificar variables relevantes.
- Consumo de recursos previsto: Acceso a bases de datos académicas y bibliotecas.
- Dependencias: Ninguna.

1.2. Análisis exploratorio de datos.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Exploración inicial de los datos para identificar tendencias, distribuciones y posibles relaciones entre variables.
- Consumo de recursos previsto: Software de análisis de datos (Python).
- Dependencias: 1.1.

2.1. Análisis de correlación entre variables principales y gasto turístico.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Estudio de la relación entre las variables identificadas y el gasto turístico.
- Consumo de recursos previsto: Software de análisis de datos (Python)
- Dependencias: 1.2.

3. Desarrollo y evaluación de modelos predictivos (Duración estimada: 4 semanas)

3.1. Desarrollo de modelos predictivos.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Implementación de modelos de Machine Learning para predecir el gasto turístico.
- Consumo de recursos previsto: Software de análisis de datos (Python)
- Dependencias: 2.1.

3.2. Evaluación de modelos.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González

- Descripción: Evaluación de la precisión y el rendimiento de los modelos desarrollados.
- Consumo de recursos previsto: Software de análisis de datos (Python).
- Dependencias: 3.1.

4. Documentación y presentación de resultados (Duración estimada: 2 semanas)

4.1. Redacción de informe final.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Documentación detallada de los resultados obtenidos y conclusiones del proyecto.
- Consumo de recursos previsto: Software de procesamiento de texto.
- Dependencias: 3.2.

4.2. Preparación de presentación.

- Responsable: Paula Hernández Rodríguez y Ernesto Rodríguez González
- Descripción: Preparación de diapositivas para presentar los resultados del proyecto.
- Consumo de recursos previsto: Software de presentación (PowerPoint).
- Dependencias: 4.1.

5. Revisión final y entrega (Duración estimada: 1 semana)

5.1. Revisión final del informe y presentación.

- Responsable: Paula Hernández Rodríguez, Ernesto Rodríguez González, Carlos Nieto López Villalón
- Descripción: Revisión exhaustiva del informe final y de la presentación para garantizar su calidad.
- Consumo de recursos previsto: Software de procesamiento de texto y presentación.
- Dependencias: 4.1., 4.2.

Tabla 1. Diagrama de Gantt

Tarea/Subtarea	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10
1. Análisis de variables influyentes	**X**	**X**	**X**	**X**						
- 1.1. Revisión bibliográfica	**X**	**X**								
- 1.2. Análisis exploratorio de datos			**X**	**X**						
2. Estudio de correlación				**X**						
- 2.1. Análisis de correlación				**X**						
3. Desarrollo y evaluación de modelos predictivos					**X**	**X**	**X**	**X**		
- 3.1. Desarrollo de modelos predictivos					**X**	**X**	**X**			
-3.2. Evaluación de modelos							**X**	**X**		
4. Documentación y presentación de resultados								**X**	**X**	**X**

- 4.1. Redacción de informe final								X	X	
- 4.2. Preparación de presentación										X
5. Revisión final y entrega										X
- 5.1. Revisión final del informe y presentación.										X

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1. Base de datos y variables

Los datos empleados en este estudio provienen, principalmente, de la Encuesta de gasto turístico (EGATUR) publicada por el Instituto Nacional de Estadística, desde octubre de 2015 hasta marzo de 2024, que tiene como objetivo principal conocer el gasto turístico de los visitantes extranjeros a su salida de España, recogiendo, además, la principal información y características del viaje.

En nuestro estudio las variables seleccionadas de esta encuesta han sido: gasto total del viaje, vía de acceso, motivo del viaje, país de origen, número de pernoctaciones, tipo de alojamiento y paquete turístico.

- El gasto turístico es la variable de interés de este estudio y se obtiene de la suma del gasto en paquete turístico, gasto en alojamiento, gasto en transporte y otros gastos. Por lo tanto, el término de gasto turístico en este estudio se define de acuerdo con estándares internacionales como el total de pagos realizados por la adquisición de bienes y servicios de consumo, así como de objetos de valor, tanto antes como durante un viaje turístico. Este gasto abarca los desembolsos efectuados por los propios visitantes, así como aquellos cubiertos por terceros.
- La vía de acceso es el modo en el que se entra al país y se incluyen las siguientes opciones: carretera, aeropuerto, puerto o tren.
- El alojamiento principal se define como el tipo de alojamiento en el que se pasa el mayor número de noches durante el viaje. Están divididos en tres categorías: hoteles y similares, alojamientos de no mercado y resto de mercado.
    - Los alojamientos de no mercado incluyen vivienda en propiedad (secundarias), vivienda de familiares o amigos y otro alojamiento no de mercado (vivienda de

uso compartido, viviendas intercambiadas, coche, refugios de montaña, en la playa,...).
    - Por su parte, la categoría de resto de mercado hace referencia a alojamiento de alquiler, camping, alojamiento de turismo rural, crucero, entre otros.
- El motivo principal de un viaje se define como el motivo sin el cual el viaje no habría tenido lugar. Está clasificado en tres categorías: ocio/vacaciones, negocios y otros.
    - En la categoría de ocio/vacaciones se incluyen los viajes realizados para visitar lugares de interés turístico, ya sean naturales, patrimonio cultural, ciudades, etc.; la asistencia a eventos deportivos o culturales; los viajes orientados a la práctica no profesional de un deporte; ir a la playa, piscinas o a cualquier instalación de entretenimiento o recreo; los cruceros; los viajes a casinos; asistencia a campamentos de verano; descanso; lunas de miel, viajes gastronómicos, a balnearios, spas u otros establecimientos especializados en tratamientos de relax o belleza; estancias e viviendas vacacionales propias, cedidas o alquiladas, etc.
    - Por su parte, se consideran viajes por motivos profesionales los viajes realizados para atender actividades de trabajo o negocios. Se incluyen en esta categoría, por ejemplo, la asistencia a: reuniones, conferencias, congresos, convenciones o ferias; para impartir charlas, dar conciertos o actuar en obras de teatro u otros espectáculos; los viajes de promoción, compra o venta de bienes o servicios en representación de productores no residentes en el lugar visitado; la misiones desempañadas por personal diplomático, militar o perteneciente a organizaciones internacionales fuera de su lugar de destino; los viajes para participar en misiones de organizaciones no gubernamentales; las estancias de investigación académica o científica; los viajes por guías u otros profesionales del sector turístico para programar y preparar viajes o actividades turísticas, como la contratación de servicios de alojamiento o transporte en el lugar visitado; la participación profesional en actividades deportivas; la asistencia a cursos de formación relacionados con el desempeño profesional; los viajes de los trabajadores especializados en medios de transporte privados (yates, aviones privados, ...).
    - Por último, la categoría de otros motivos incluye los motivos que no han podido ser clasificados en las categorías anteriores.
- El país hace referencia al país de origen del turista y en este estudio este puede ser: Alemania, Bélgica, Francia, Irlanda, Italia, Países Bajos y Portugal.
- Se entiende por paquete turístico la reserva previa del viaje que incluye al menos alojamiento y transporte.
- Para la duración del viaje se mide el número de noches que se han pasado durante el viaje en España, es decir, el número de pernoctaciones.

De manera complementaria, se han obtenido datos del Banco de España sobre el EURIBOR y de EUROSTAT sobre tasas de paro e IPC de los países de origen de los turistas seleccionados en el estudio.

- El EURIBOR es el tipo de interés del mercado monetario del euro que surgió desde 1999. Es el tipo al que un banco principal ofrece depósitos a plazo interbancarios en euros a otro banco principal.
- Las personas desempleadas son las personas de 15 a 74 años que:
    - Han estado sin trabajo durante la semana de referencia.
    - Están actualmente disponibles para el trabajo, es decir, estaban disponibles para el empleo remunerado o el trabajo por cuenta propia antes del final de las dos semanas siguientes a la semana de referencia.
    - Buscan activamente trabajo, es decir, han tomado medidas específicas en el período de cuatro semanas que termina con la semana de referencia para buscar trabajo remunerado o por cuenta propia o que encontró un trabajo para comenzar más tarde, es decir, dentro de un período de, como máximo, tres meses.
- Por último, el IPC es un indicador para medir la inflación, es un índice con base en el año 2015.

Tabla 2. Resúmenes variables

VARIABLE	INTERPRETACIÓN
A1	Vía de salida: 1: carretera, 2: aeropuerto, 3: puerto, 4: tren
País	01:Alemania. 02:Bélgica. 03:Francia. 04: Irlanda. 05: Italia. 06: Países Bajos. 07: Portugal
A13	Total pernoctaciones
Aloja	Alojamiento principal: 1: Hoteles y similares, 2: Resto de mercado, 3: Alojamiento no de mercado
Motivo	1: Ocio/vacaciones, 2: Negocios, 3: Resto
A16	Paquete turístico: 1: Sí, 6: No
gastototal	Gasto total del viaje/excursión
IPC	Índice de precios al consumo
Desempleo	Tasa de paro
EURIBOR	Tipo de interés

4.2.2. Metodología

Primeramente, se procede a la unión de la base de datos a utilizar, dividida en dos pasos:

1. Unificación de los csv individuales de cada mes de EGATUR desde octubre de 2015 hasta marzo de 2024.
2. Unión de los datos relativos a las variables macroeconómicas (EURIBOR, IPC y tasa de paro) a la base de datos unificada en el paso anterior con el objetivo de facilitar el posterior análisis y estudio de las características de las características del turismo.

En segundo lugar, se filtra solo por países de la Eurozona y para la C.A. de Canarias (ámbito espacial de este estudio).

En tercer lugar, se lleva a cabo la limpieza y el preprocesamiento de los datos que incluye el manejo de datos faltantes y la codificación de variables categóricas. El total de observaciones de la base de datos es, finalmente, de 42383.

Luego, se divide el conjunto de datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo de manera adecuada, en este caso, mediante validación cruzada K-Fold (5 iteraciones).

Para iniciar el estudio de los datos, se realiza un análisis univariante, que se centra en la exploración de cada variable individualmente, permitiendo comprender sus características básicas y distribuciones. El objetivo de este análisis es identificar patrones, detectar valores atípicos y obtener una visión general del comportamiento de cada variable en el conjunto de datos. Para ello se emplean gráficos de barras en todas las variables, tanto categóricas como numéricas, con los valores porcentuales de cada una de ellas. Además, en las variables numéricas se estudian diversas medidas descriptivas como la media y la desviación estándar.

Posteriormente se procede a realizar un análisis bivariante, que se enfoca en el estudio de la relación entre la variable de interés (gasto total) y cada una de las variables del modelo. Este análisis es esencial para entender las interacciones y posibles correlaciones entre las variables. En este caso, para las variables categóricas se emplean boxplot y para las numéricas, diagramas de dispersión.

Para el análisis multivariante se ha utilizado: Linear Regression, KNN y Random Forest. En este sentido, para el estudio de los algoritmos:

- Se seleccionan los hiperparámetros fundamentales para cada algoritmo. En concreto, en el caso del algoritmo KNN, se obtuvo el valor de K óptimo mediante un proceso de ajuste utilizando validación cruzada con 5 iteraciones. Este procedimiento abarcó varios valores de K dentro del rango de 2 a 60, con el objetivo de minimizar los valores promedio de MAE y RMSE en el conjunto de prueba, además de maximizar el coeficiente de determinación R^2.
- Por su parte, en el Random Forest, en el caso del número de descriptores a utilizar en cada árbol, se optó por emplear la raíz cuadrada del número total de descriptores, una metodología comúnmente empleada en la literatura debido a sus resultados positivos, mientras que para encontrar el número de árboles óptimos a utilizar se estimaron el MAE, RMSE y R^2 para varios valores del número de árboles en el rango de 10 a 200. También con el objetivo de minimizar los valores promedio de MAE y RMSE en el conjunto de prueba y de maximizar el coeficiente de determinación R^2. Los resultados para cada valor del número de árboles se estimaron utilizando validación cruzada de 5 iteraciones.
- Una vez seleccionado los valores óptimos de los mismos, se estiman el RMSE, MAE y R^2 para las predicciones utilizando validación cruzada de 5 iteraciones.

La validación cruzada proporciona una evaluación robusta del rendimiento del modelo al dividir los datos en múltiples particiones de entrenamiento y prueba. En este trabajo se ha utilizado una validación cruzada de 5 iteraciones. Una vez entrenado el modelo, se evalúan sus parámetros predictivos, en este caso, las métricas seleccionadas han sido: el Error Cuadrático Medio (RMSE), el Desvío Medio Absoluto (MAE) y el coeficiente de determinación (R^2), que son

las métricas relacionadas con análisis de regresión que se emplean con más frecuencia (Alaminos-Fernández, 2022). El RMSE y el MAE son dos medidas de evaluación de errores. Por su parte el R^2 mide la bondad del ajuste, indica qué tan bien se ajusta un modelo a los datos observados. Específicamente, el R^2 indica la proporción de la varianza total de la variable dependiente que es explicada por las variables independientes en el modelo (Stock et al., 2012).

4.2.3. Algoritmos

Los algoritmos empleados en este estudio son tres: Linear Regression, K-Nearest Neighbors y Random Forest, que se explican a continuación:

El análisis Linear Regression destaca como una de las herramientas estadísticas más empleadas en la investigación actual para estudiar la relación entre una variable de interés y una o más variables que la explican. Este método, conocido como estimación por mínimos cuadrados, se atribuye a Legendre en 1805 y sigue siendo uno de los enfoques más frecuentemente utilizados para el ajuste de modelos (Dolores et al., 2019).

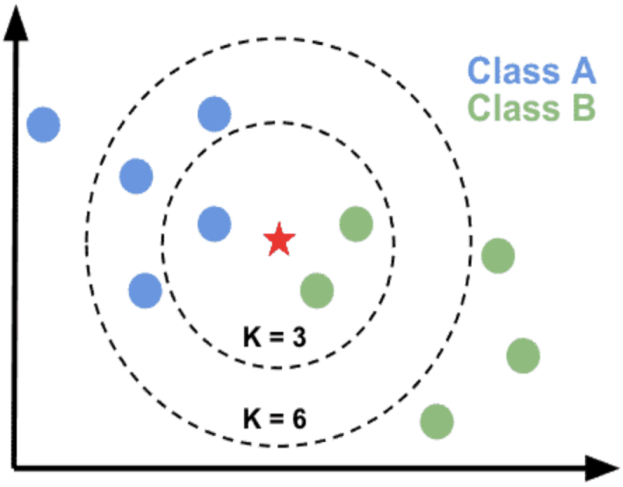
El modelo Linear Regression, múltiple en nuestro caso, se especifica de la siguiente forma:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i + \cdots + \beta_p X_p + \varepsilon_i \quad (1)$$

Donde i son las diferentes observaciones, Yi es la variable dependiente, Xi las variables independientes, α es la constante, ε el término de error aleatorio no observable y βi los parámetros de regresión.

Por su parte, el algoritmo KNN (K-Nearest Neighbors) es un tipo de algoritmo no paramétrico, considerado como uno de los de mejor desempeño en el ámbito del aprendizaje automático, que se puede emplear para problemas predictivos de regresión y clasificación (Narváez et al., 2022). Sirve para predecir tanto variables categóricas como numéricas con naturaleza no paramétrica siendo el único parámetro que hay que definir, el número de vecinos más cercanos. Su funcionamiento se basa en identificar los k puntos de datos más cercanos en el conjunto de entrenamiento cuando se presenta un nuevo punto. Si todos estos k vecinos más cercanos comparten la misma categoría, se infiere que el nuevo punto también posee características similares y comparte los mismos atributos que estos vecinos.

Imagen 2. KNN



La imagen 2 representa un ejemplo de funcionamiento del algoritmo KNN. Los círculos azules y verdes son las muestras de las dos clases distintas del conjunto de entrenamiento y la estrella roja es el nuevo dato que se quiere clasificar. Para k=3 la clase asignada a la estrella roja es la clase verde ya que hay dos círculos verdes frente a uno azul. En cambio, para k=6 la clase asignada a la estrella roja es la azul, puesto que hay cuatro círculos azules frente a dos verdes.

Por último, Random Forest es un poderoso algoritmo de aprendizaje automático basado en ensambles que combina múltiples árboles de decisión (Breiman, 2001). Esta técnica se ha destacado por su eficacia en una amplia gama de problemas de clasificación y regresión. En lugar de utilizar un solo árbol de decisión, Random Forest construye una "foresta" de árboles de decisión durante el entrenamiento. Cada árbol en el bosque se entrena de forma independiente utilizando una muestra aleatoria del conjunto de datos (con reemplazo, conocido como bagging) y seleccionando aleatoriamente un subconjunto de características en cada división del árbol. Esto introduce aleatoriedad y diversidad en los árboles individuales, lo que ayuda a reducir la varianza del modelo y mejorar la generalización.

4.2.4. Herramientas

En el desarrollo de este estudio se llevó a cabo el procesamiento y análisis de datos utilizando el lenguaje de programación Python. Las bibliotecas fundamentales empleadas fueron pandas y NumPy para la manipulación eficiente de datos y cálculos numéricos. Además, se utilizaron las bibliotecas scikit-learn, matplotlib y seaborn para realizar predicciones mediante técnicas de aprendizaje automático y para la visualización detallada y comprensión de los datos.

Particularmente relevante para la realización de predicciones precisas fue la utilización de scikit-learn, una biblioteca de aprendizaje automático de código abierto diseñada específicamente para análisis predictivo de datos. Construida sobre NumPy, SciPy y matplotlib, scikit-learn se integra fácilmente con otras bibliotecas populares de Python utilizadas en ciencia de datos, proporcionando acceso a una amplia gama de algoritmos tanto supervisados como no supervisados, desde Linear Regression y clasificación hasta clustering y reducción de dimensionalidad, en este sentido, la biblioteca soporta algoritmos de última generación como k-Nearest Neighbors (KNN), XGBoost, Random Forest y Support Vector Machines (SVM), entre otros (Scikit-learn, 2024).

Las importaciones específicas de scikit-learn necesarias para realizar el ajuste y predicción de datos incluyeron:

Imagen 3. Código scikit-learn

```
In [ ]:
from sklearn.model_selection import KFold, cross_validate, cross_val_score
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import make_scorer, mean_squared_error, mean_absolute_error, r2_score
```

Estas importaciones proporcionaron las herramientas esenciales para la construcción de modelos predictivos robustos, la evaluación de su desempeño y la preparación de los datos mediante transformaciones y preprocesamientos adecuados.

4.3 Recursos requeridos

En la ejecución del proyecto se utilizaron diversos recursos técnicos y dispositivos, que se detallan a continuación:

Recursos técnicos:

- Lenguaje de programación: Python.
- Bibliotecas y herramientas de software:
  - Pandas (para la manipulación de datos).
  - NumPy (para cálculos numéricos).
  - Scikit-learn (para técnicas de aprendizaje automático).
  - Matplotlib (para visualización de datos).
  - Seaborn (para visualización detallada y comprensión de datos)
- Dispositivos y equipos: Computadora de alto rendimiento para procesamiento de datos y entrenamiento de modelos.

Material de datos:

- Datos de la Encuesta de Gasto Turístico (EGATUR) del Instituto Nacional de Estadística (octubre 2015 - marzo 2024).
- Datos macroeconómicos del Banco de España y EUROSTAT (EURIBOR, tasas de paro, IPC).

4.4 Presupuesto

El presente presupuesto proporciona una evaluación económica detallada del proyecto de investigación llevado a cabo. Se consideran todos los recursos utilizados, incluyendo el tiempo dedicado por el personal involucrado, el equipo técnico empleado y el consumo de servicios como Internet y electricidad.

Tabla 3. Presupuesto

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	300h	Paula Hernández Rodríguez 150 h y Ernesto Rodríguez González 150h
Equipo técnico utilizado	4000€	Se incluyen dos ordenadores valorados aproximadamente en 2000 euros cada uno, sumando un total de 4000 euros. Esto representa el costo que tendría adquirir estos ordenadores nuevos en el mercado.
Software utilizado	0€	Software de código abierto utilizado sin costo de licencia.

Consumo de electricidad	50€	Estimación del consumo de electricidad para alimentar el ordenador durante el período del proyecto.
Consumo de Internet	30€	Costo asociado al uso de Internet para acceder a recursos y comunicación necesaria durante la investigación.

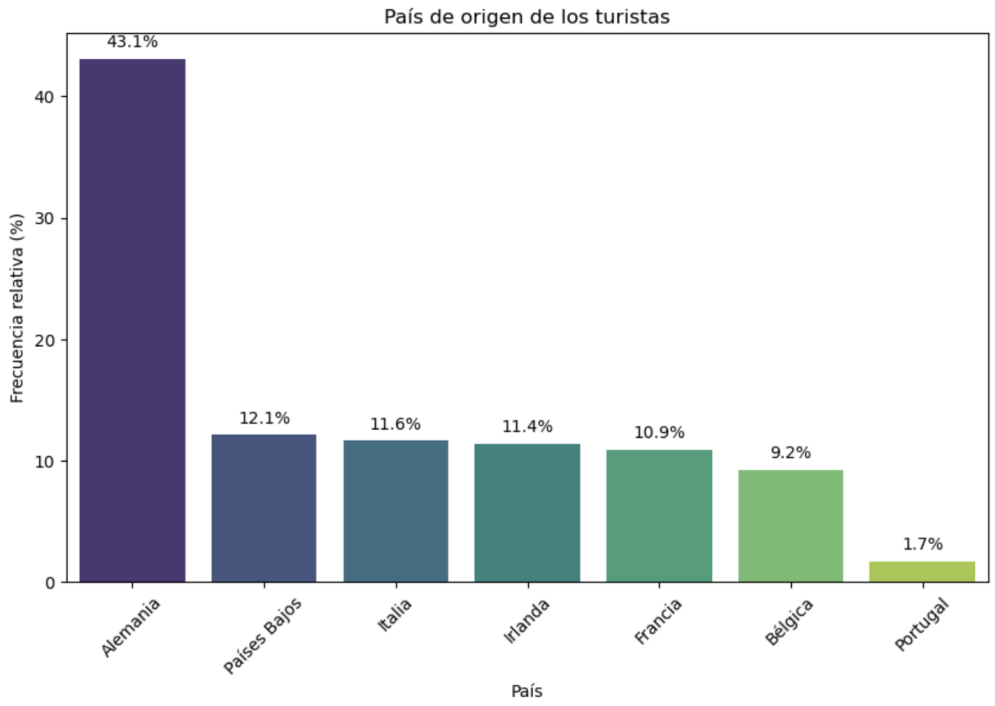
4.5 Resultados del proyecto

En esta sección se presentan los resultados obtenidos del análisis de datos y la modelización predictiva del gasto turístico en Canarias.

4.5.1. Análisis univariante

El objetivo principal del análisis univariante es comprender las características básicas de cada variable individual en nuestro conjunto de datos. Se pretende describir las distribuciones y características de las variables categóricas y numéricas para obtener una visión general del comportamiento de los datos.

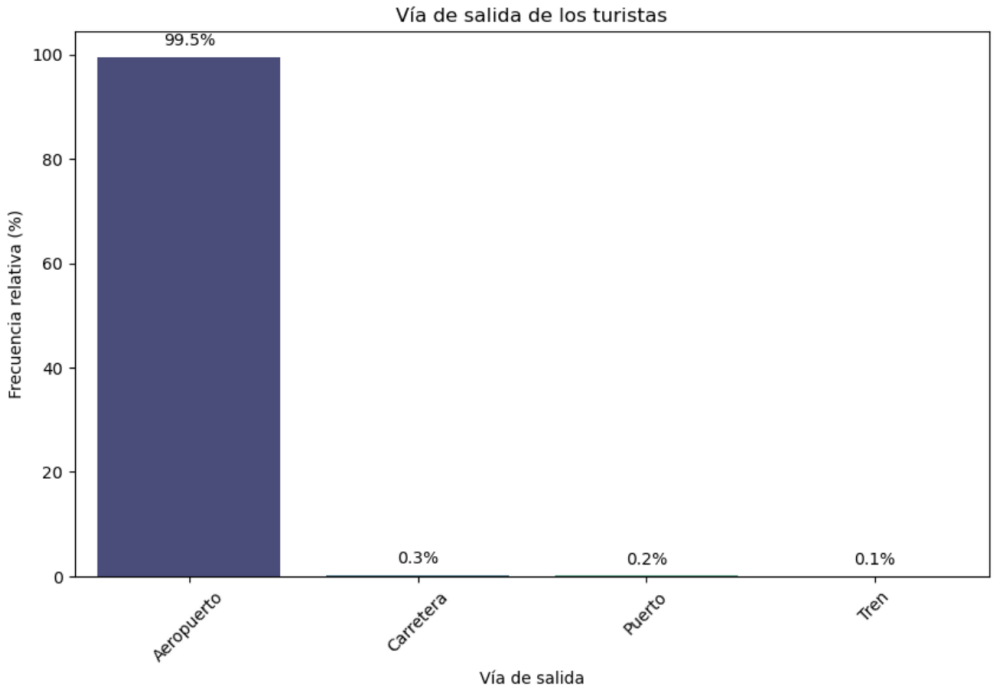
Gráfico 7. Análisis univariante del país de origen



El análisis univariante del conteo de individuos por país de procedencia en Canarias (Gráfico 7) revela que Alemania es el principal país de origen en términos de número de turistas, con más del 40% de individuos para el periodo estudiado, aproximadamente unos 17.500, triplicando prácticamente la cantidad del resto de países de la Eurozona que se han introducido en el modelo, por orden: Países Bajos, Italia, Irlanda, Francia, Bélgica y, con una cantidad muy inferior de turistas, Portugal.

Este predominio alemán destaca la importancia de enfocar estrategias de marketing y servicios turísticos hacia este mercado específico, optimizando las conexiones aéreas y adaptando las ofertas a las preferencias de los turistas alemanes. La considerable diferencia en los números también sugiere la necesidad de diversificar los esfuerzos promocionales para atraer a más turistas de otros países de la Eurozona.

Gráfico 8. Análisis univariante de la vía de salida

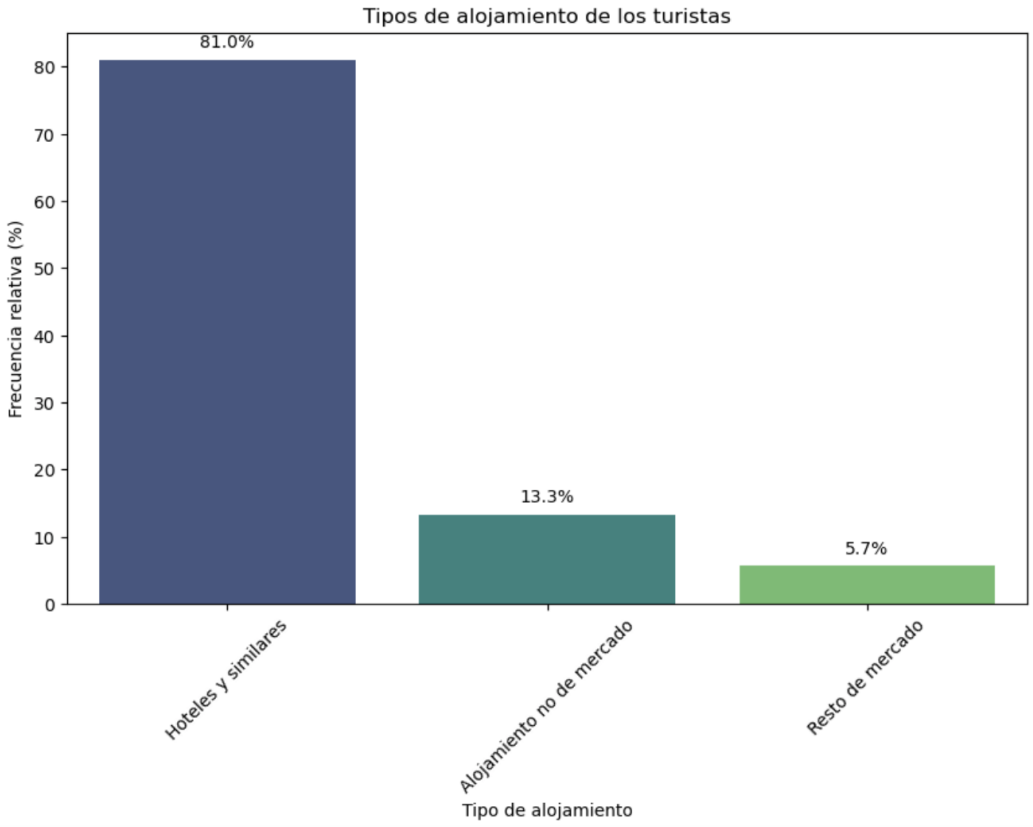


Por su parte, el gráfico 8 confirma la preponderancia del transporte aéreo en la movilidad de los visitantes siendo prácticamente el 100% de los turistas europeos los que utilizan esta vía de llegada, subrayando algunos aspectos importantes:

Por un lado, la infraestructura aeroportuaria debe ser capaz de manejar un alto volumen de pasajeros, garantizando eficiencia y comodidad para los viajeros. Esto incluye no solo la capacidad física de los aeropuertos, sino también los servicios adicionales como el transporte terrestre, la seguridad, el control de fronteras y las instalaciones comerciales.

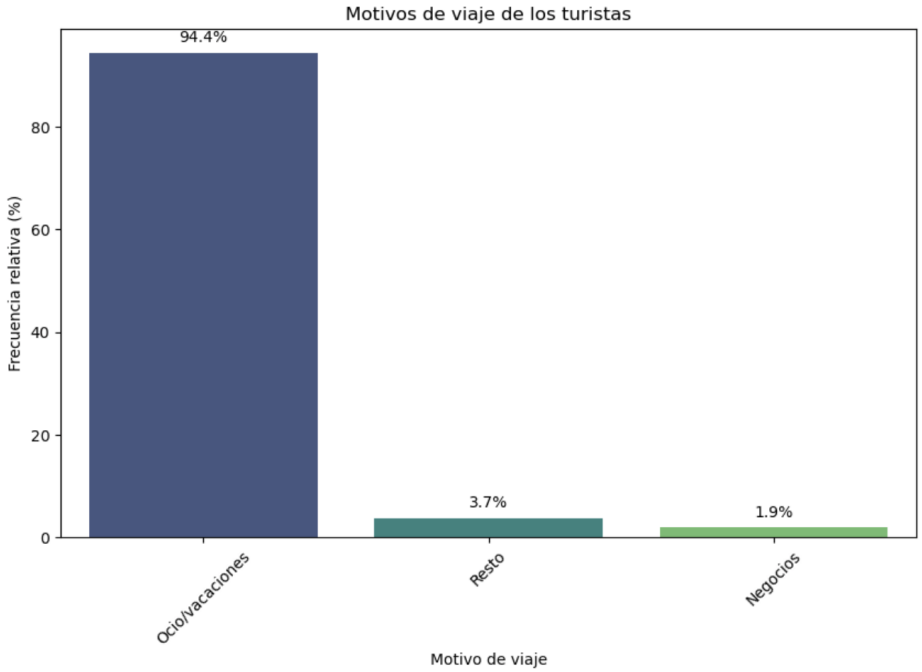
Por otro lado, la conectividad aérea se convierte en un factor determinante y es por ello por lo que las aerolíneas y los acuerdos de rutas juegan un papel vital en el mantenimiento y la expansión del flujo turístico.

Gráfico 9. Análisis univariante del tipo de alojamiento



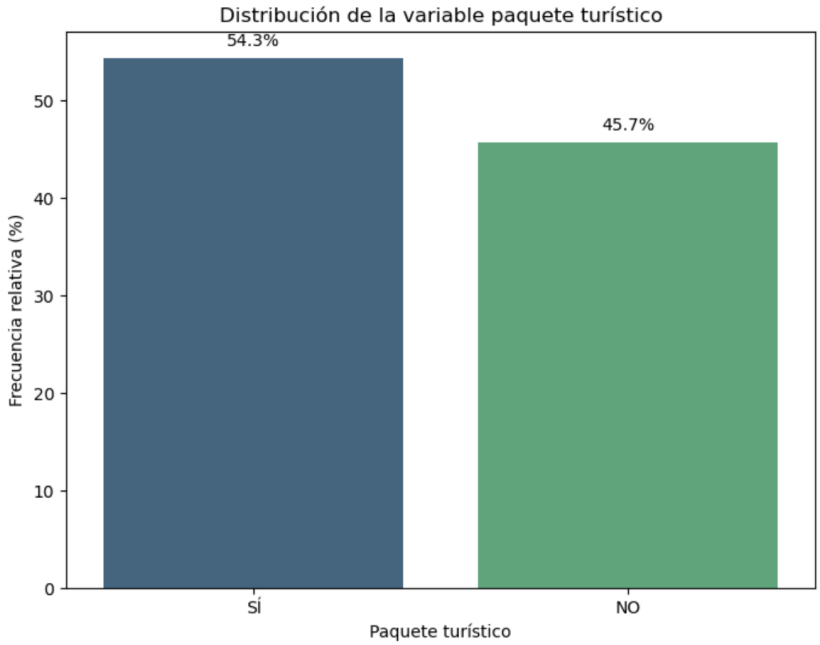
En cuanto a los datos del tipo de alojamiento (Gráfico 9), se evidencia una clara dominancia de la categoría de "Hoteles y Similares," con el 81%. Esta preferencia destaca la importancia de la infraestructura hotelera en Canarias. En contraste, los "Alojamientos de No Mercado" y los del "Resto de Mercado" muestran cifras considerablemente menores, indicando que son opciones menos utilizadas por los turistas. Esta disparidad sugiere que mientras los hoteles siguen siendo la opción preferida, existe un potencial sin explotar en los otros tipos de alojamiento, que podrían beneficiarse de mayores esfuerzos promocionales y mejoras en sus servicios para atraer a una mayor proporción de visitantes.

Gráfico 10. Análisis univariante del motivo de viaje



Con respecto al motivo del viaje (Gráfico 10), la gran mayoría de los visitantes llegan a Canarias para disfrutar de su tiempo libre, siendo casi el 95%. Este dato es coherente con la reputación del archipiélago, conocido por su amplia oferta de actividades recreativas y su atractivo natural. En contraste, los viajes por negocios y otros motivos representan una fracción pequeña del total, lo que refuerza la imagen de Canarias como un destino predominantemente orientado al turismo de ocio. Esta tendencia subraya la importancia de seguir invirtiendo en infraestructuras y servicios turísticos que potencien la experiencia de los visitantes que buscan relajación y recreación.

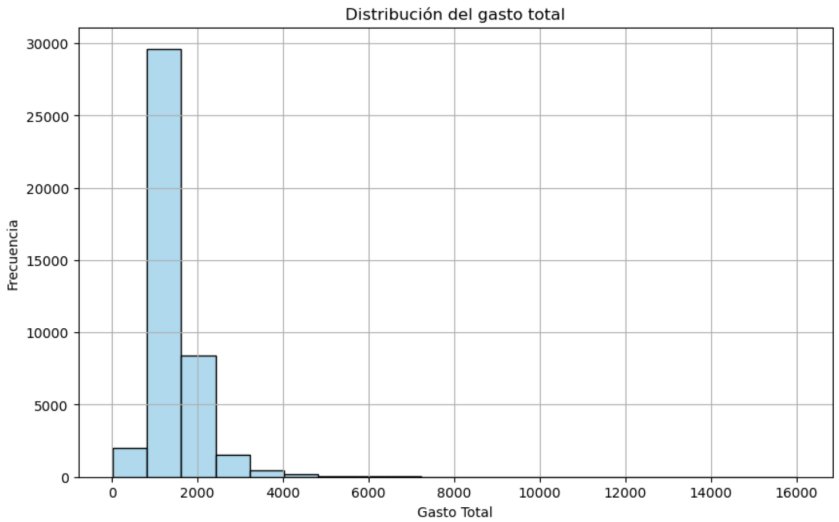
Gráfico 11. Análisis univariante del paquete turístico



El Gráfico 11 subraya la importancia de los paquetes turísticos en la industria del turismo en Canarias indicando una fuerte demanda por estos productos organizados. Este dato refleja la preferencia de muchos turistas por la comodidad y la conveniencia de los paquetes que incluyen: alojamiento, transporte y actividades.

Al mismo tiempo, el notable número de turistas que prefieren no usar paquetes turísticos sugiere que hay una significativa oportunidad para las empresas turísticas de atender a este segmento del mercado. Ofrecer opciones personalizadas y flexibles puede satisfacer las necesidades y preferencias de estos viajeros, quienes son capaces de buscar experiencias más independientes o hechas a medida. Este enfoque dual, combinando la oferta de paquetes organizados con alternativas personalizadas, puede potenciar aún más el atractivo de Canarias como destino turístico y maximizar la satisfacción de una variedad más amplia de visitantes.

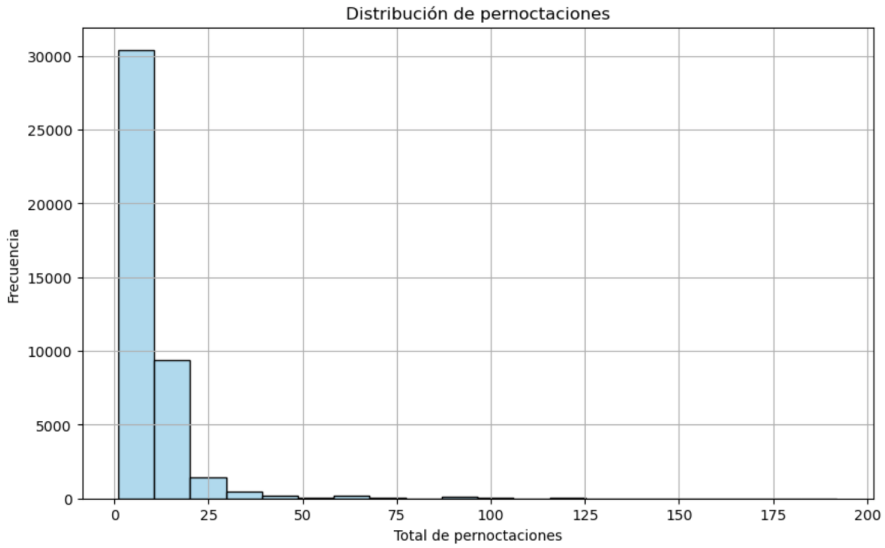
Gráfico 12. Análisis univariante del gasto total




```
count    42383.000000
mean      1442.973065
std        652.368656
min         17.849627
25%       1084.893327
50%       1287.777220
75%       1629.734751
max      16038.275845
Name: gastototal, dtype: float64
```

El histograma del gráfico 12 revela que la mayoría de los turistas tienen un gasto total concentrado en el rango de 1.100 a 1.600 euros, con un máximo de 16.038 euros y un mínimo de 17 euros. La distribución es asimétrica, con una larga cola hacia la derecha, lo que indica la presencia de algunos turistas que incurren en gastos significativamente mayores. La mayoría de los datos se agrupan alrededor del promedio de 1.442 euros, con una desviación estándar de 652, lo que sugiere una variabilidad considerable en el gasto turístico. Por lo tanto, este gráfico indica que la mayoría de los turistas tienen un gasto total moderado, mientras que unos pocos incurren en gastos mucho mayores.

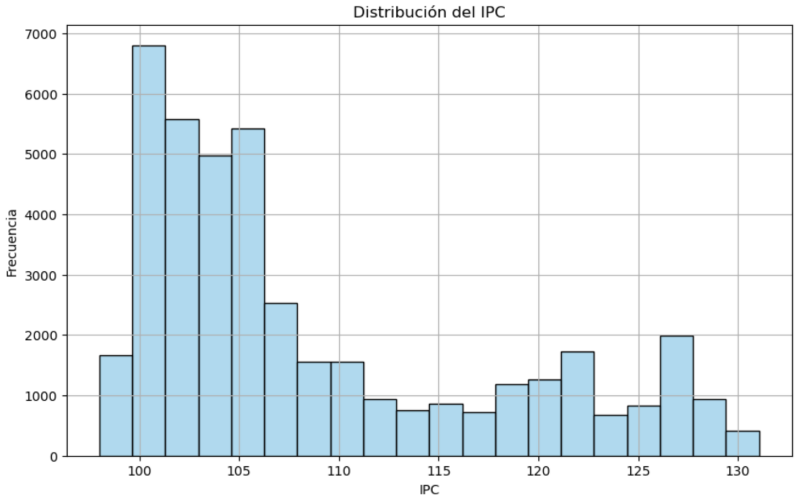
Gráfico 13. Análisis univariante de las pernoctaciones




```
count    42383.000000
mean        10.339853
std          9.275416
min          1.000000
25%          7.000000
50%          7.000000
75%         12.000000
max        192.000000
Name: A13, dtype: float64
```

El gráfico 13 muestra la distribución del total de pernoctaciones de los turistas en Canarias. En general, los turistas se hospedan entre 7 y 12 noches, con un máximo de 192 noches y un mínimo de 1 noche. La media de pernoctaciones es de 10 noches, lo que indica que el turista medio suele hospedarse un poco más de una semana. Este patrón de estancia sugiere que la mayoría de los visitantes optan por una duración de vacaciones que permite una experiencia completa del destino sin extenderse demasiado en el tiempo, lo cual es coherente con los paquetes turísticos tradicionales que suelen ofrecer estancias de una o dos semanas. Esta tendencia también refleja las preferencias de los turistas europeos, quienes, a menudo, planifican sus vacaciones en función de períodos de una o dos semanas, posiblemente influenciados por las políticas de vacaciones laborales y escolares en sus países de origen.

Gráfico 14. Análisis univariante del IPC

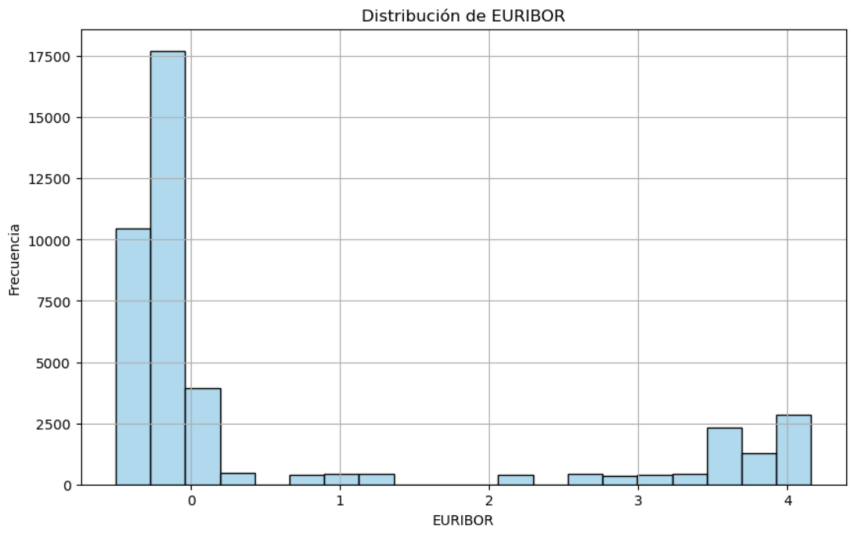



```
count    42383.000000
mean       108.788302
std          8.873758
min         98.000000
25%        101.900000
50%        105.400000
75%        114.600000
max        131.050000
Name: IPC, dtype: float64
```

El gráfico 14 muestra el análisis de la distribución del IPC (Índice de Precios al Consumo), -índice con base en 2015-. La mayoría de los valores se encuentran concentrados alrededor de 100, dentro del rango de 98 a 105. Esta agrupación sugiere períodos de baja inflación o estabilidad de precios, indicando posiblemente políticas económicas efectivas y condiciones macroeconómicas favorables durante gran parte del período analizado.

Sin embargo, se observa una larga cola hacia la derecha del histograma, con valores del IPC superiores a 120. Esto puede reflejar períodos de alta inflación, probablemente asociados con la recuperación económica post-pandemia y la crisis energética, que afectaron los precios de manera significativa en ciertos momentos. Esta variabilidad en el IPC subraya la importancia de monitorear de cerca las fluctuaciones económicas para entender su impacto en el sector turístico y otras industrias en Canarias.

Gráfico 15. Análisis univariante del EURIBOR

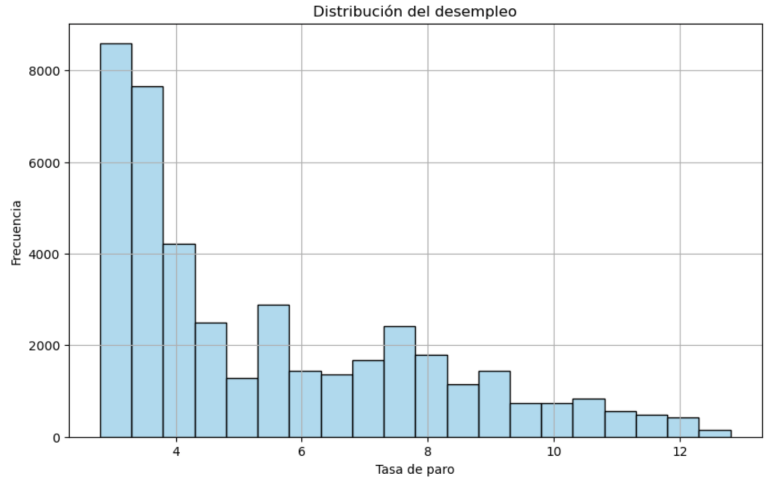



```
count    42383.000000
mean         0.593276
std          1.552077
min         -0.505000
25%         -0.266000
50%         -0.134000
75%          0.079000
max          4.160000
Name: EURIBOR, dtype: float64
```

En el gráfico 15 se observa la distribución del EURIBOR a lo largo del periodo de estudio de 2015 a 2024. La mayoría de los valores se concentran alrededor del 0%, indicando una política monetaria expansiva del Banco Central Europeo (BCE) desde aproximadamente el año 2012. Este enfoque tiene como objetivo estimular la economía reduciendo las tasas de interés para facilitar el acceso al crédito y promover la inversión.

Se pueden observar pequeños picos en los valores cercanos al 4%, particularmente hacia finales del año 2022 hasta la actualidad. Estos picos pueden estar relacionados con ajustes en las políticas monetarias del BCE para controlar la inflación, que se intensificaron después de la pandemia de COVID-19 y la recuperación económica posterior. Además, la crisis financiera de 2008 y sus repercusiones económicas también jugaron un papel crucial en mantener las tasas de interés bajas durante períodos prolongados para apoyar la recuperación económica.

Gráfico 16. Análisis univariante del desempleo



count	42383.000000
mean	5.416691
std	2.473074
min	2.800000
25%	3.400000
50%	4.300000
75%	7.300000
max	12.800000

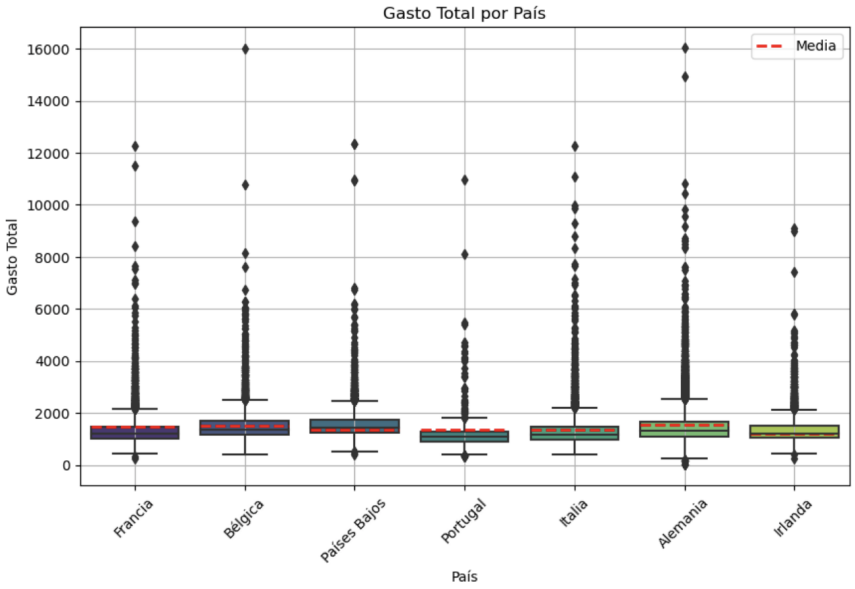
El gráfico 16 muestra la distribución de la tasa de desempleo a lo largo del período analizado. La mayoría de las observaciones se concentran entre el 3% y el 7%, indicando tasas relativamente bajas de desempleo. A medida que aumenta la tasa de desempleo, la frecuencia de observaciones disminuye gradualmente. Se observa una cola hacia la derecha, lo que indica que existen algunas observaciones con tasas de desempleo más altas, alcanzando un máximo de 12.8%, pero siendo menos frecuentes.

La media de la tasa de desempleo del 5.41% y la mediana del 4.8% refuerzan la observación de que la mayoría de los años analizados muestran tasas de desempleo relativamente bajas. Esta distribución puede explicarse por varios factores, incluyendo la recuperación económica post-crisis, las políticas de estímulo económico implementadas, así como el impacto y la posterior recuperación de la pandemia de COVID-19. Además, el desarrollo de innovaciones tecnológicas y la adaptación hacia mercados laborales más flexibles también podrían haber contribuido a mantener las tasas de desempleo en niveles bajos-moderados.

4.5.2. Análisis bivariante

El análisis bivariante es una técnica estadística que nos permite explorar la relación entre cada una de las variables consideradas en un estudio y el gasto total en el sector turístico de Canarias. Este enfoque nos proporciona una comprensión detallada de cómo cada variable individualmente impacta en el gasto total de los turistas.

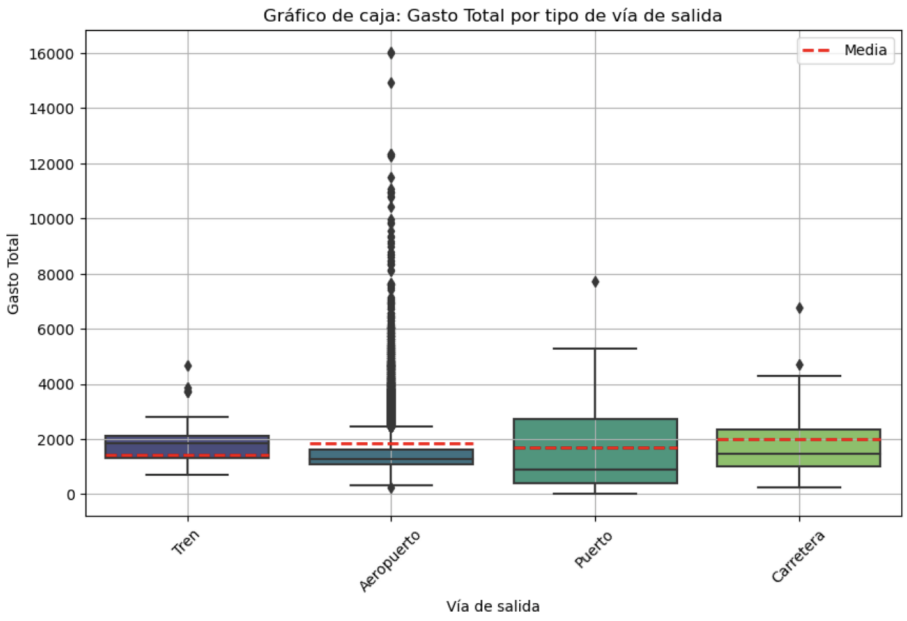
Gráfico 17. Análisis bivariante del gasto total - país de origen



El análisis bivariante del gasto total por país de origen (gráfico 17) revela que la media del gasto se mantiene relativamente constante alrededor de los 1500-2000 euros para todos los países estudiados en Canarias. Esto sugiere que el gasto turístico en la región es en general homogéneo, aunque hay una notable variabilidad con numerosos valores atípicos. Los outliers indican que un segmento de turistas gasta significativamente más que la media, lo cual puede ser crucial para estrategias dirigidas a personalizar ofertas turísticas para estos visitantes de alto gasto. En orden de mayor a menor gasto medio, los países son: Alemania, Bélgica, Francia, Países Bajos, Italia, Portugal e Irlanda.

En resumen, mientras el gasto turístico en Canarias muestra una relativa homogeneidad en términos de medias, la presencia de valores atípicos subraya la importancia de la personalización y la segmentación en las estrategias de marketing para optimizar la experiencia del turista y maximizar el impacto económico del sector turístico en la región.

Gráfico 18. Análisis bivariante del gasto total - vía de salida

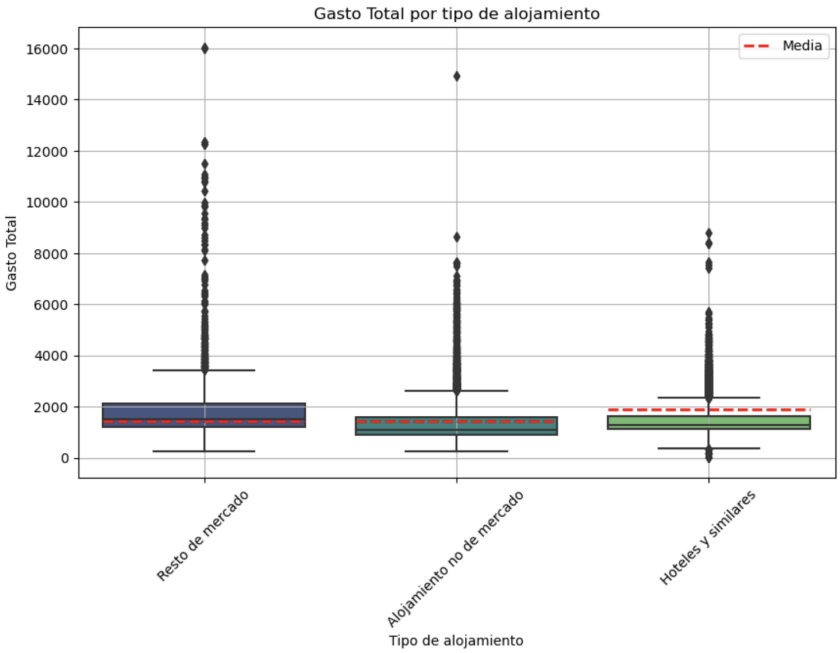


El gráfico 18 muestra que la mayoría de los visitantes que llegan a España lo hacen a través del aeropuerto, teniendo una distribución de gasto total que varía ampliamente. La media del gasto para aquellos que utilizan el aeropuerto como vía de salida es de aproximadamente 2000 euros, pero existe una cantidad significativa de valores atípicos que indican gastos mucho más altos, llegando hasta los 16000 euros. Esto sugiere la presencia de un segmento de turistas de alto poder adquisitivo que influye considerablemente en la media.

Por otro lado, los otros modos de llegada, como el tren, el puerto y la carretera, representan casos aislados y muestran distribuciones de gasto más controladas y menos significativas en términos de cantidad de turistas. Esto puede indicar que estos modos de transporte son utilizados por una minoría de visitantes, que se desplazan internamente dentro de las islas, y no tanto por turistas internacionales con estancias más largas y mayores gastos.

En términos generales, el aeropuerto se destaca como la principal vía de salida de los turistas desde sus países de origen hacia España, con una variabilidad notable en el gasto total entre los visitantes.

Gráfico 19. Análisis bivariante por tipo de alojamiento

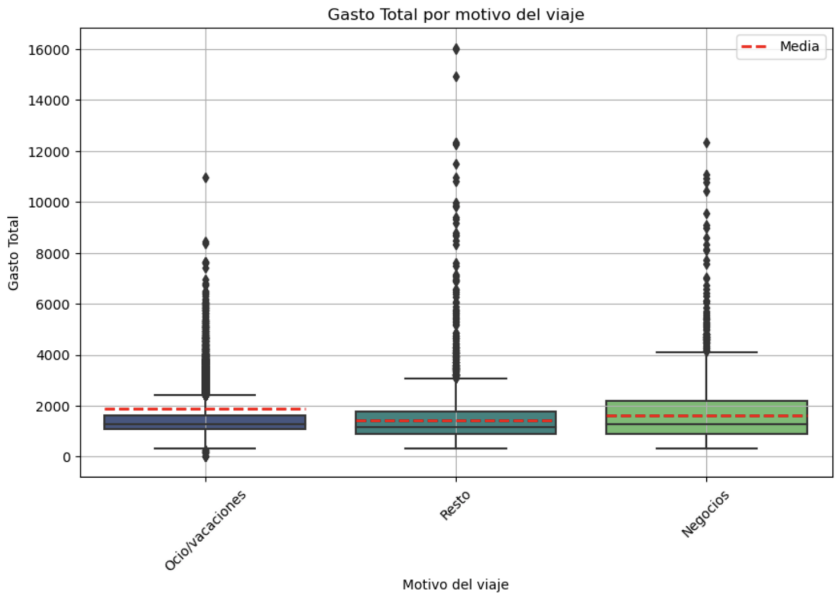


El análisis bivariante del gasto total en relación con el tipo de alojamiento en Canarias (Gráfico 19) revela varias tendencias importantes. En promedio, los turistas que se hospedan en "Resto de mercado" y en "Alojamiento no de mercado" tienen un gasto total que se sitúa entre los 1500 y 2000 euros. Por otro lado, el grupo que elige "Hoteles y similares" muestra una media de gasto muy cercana a los 2000 euros, con una distribución más concentrada y menos valores atípicos. Esto sugiere que los turistas que optan por hoteles y establecimientos similares tienden a tener un gasto más consistente y predecible, posiblemente debido a tarifas estándar y servicios incluidos.

En contraste, los turistas que eligen otras opciones de mercado muestran una mayor variabilidad en sus gastos, lo que puede reflejar una variedad más amplia de opciones de precios y servicios dentro de estos alojamientos menos convencionales.

Para las empresas y destinos turísticos en Canarias, esta información puede ser crucial para adaptar estrategias de gestión empresarial y mejorar la oferta de servicios, asegurando que se satisfagan las necesidades y expectativas de diferentes segmentos de mercado con respecto al tipo de alojamiento preferido.

Gráfico 20. Análisis bivariante del gasto total - motivo del viaje

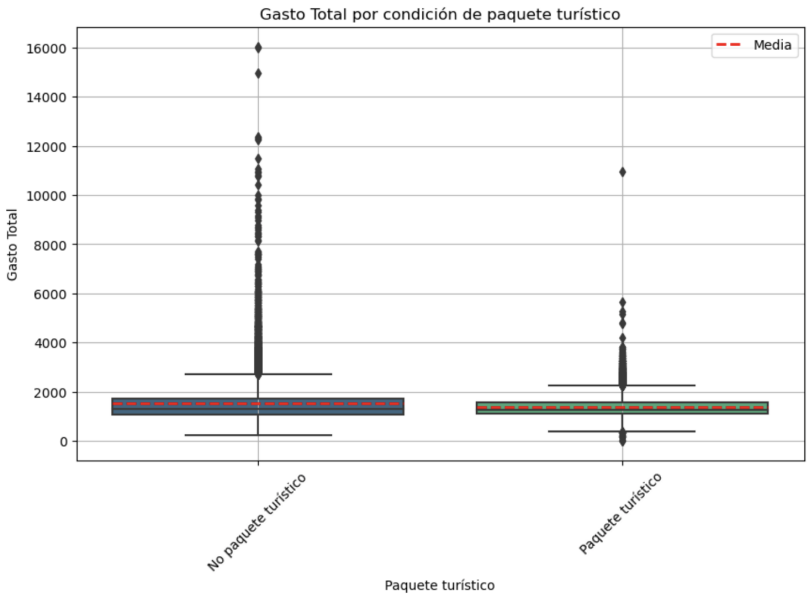


En el gráfico 20 se observa una diferencia en los patrones de gasto entre los motivos "ocio vacaciones" y "negocios" respecto al motivo "resto". Para el motivo "ocio vacaciones", la media de gasto se sitúa cercana a los 2000 euros. Esto sugiere que los turistas que visitan Canarias por vacaciones tienden a gastar más, posiblemente debido a la duración prolongada de sus estancias y a la participación en actividades recreativas y de disfrute.

En contraste, para el motivo "negocios", la media de gasto es menor y queda por debajo de los 2000 euros. Esto podría reflejar que los viajeros de negocios tienen estancias más cortas y gastos limitados en comparación con los turistas en vacaciones, centrando sus recursos principalmente en necesidades empresariales específicas.

Por otro lado, el motivo "resto" muestra una distribución más dispersa en cuanto a los valores de gasto total, aunque la media también se acerca a los 2000 euros. Esto indica que factores como la duración de la estancia, las preferencias personales y las actividades planificadas durante la visita hacen variar el gasto.

Gráfico 21. Análisis bivariante del gasto total - paquete turístico

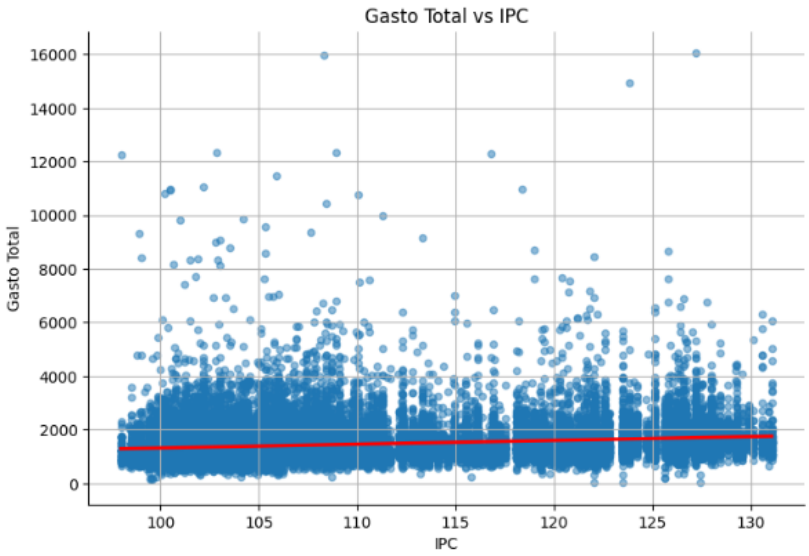


El gráfico 21 muestra que la media de gasto total es similar entre los turistas que tienen paquete turístico y los que no lo tienen. Sin embargo, se observa una mayor variabilidad en los montos de gasto entre los turistas sin paquete turístico, con la presencia de outliers que indican gastos significativamente más altos. En contraste, el grupo con paquete turístico muestra una distribución más homogénea alrededor de la media.

Esto puede entenderse porque los viajeros que organizan su viaje sin un paquete turístico tienen la libertad de tomar decisiones más flexibles y variadas en cuanto a dónde y cómo gastar su dinero. Algunos pueden optar por invertir más en experiencias exclusivas o actividades específicas, lo que explica la mayor variabilidad en sus gastos y la presencia de valores atípicos más altos.

Por otro lado, los viajeros que eligen un paquete turístico generalmente optan por una opción que incluye un precio fijo por un conjunto de servicios, como: alojamiento, transporte y algunas actividades. Esto tiende a estandarizar los gastos dentro del grupo, explicando por qué la distribución de gastos entre estos viajeros tiende a ser más homogénea alrededor de la media, con menos variabilidad y menos presencia de valores atípicos.

Gráfico 22. Análisis bivariante del gasto total - IPC

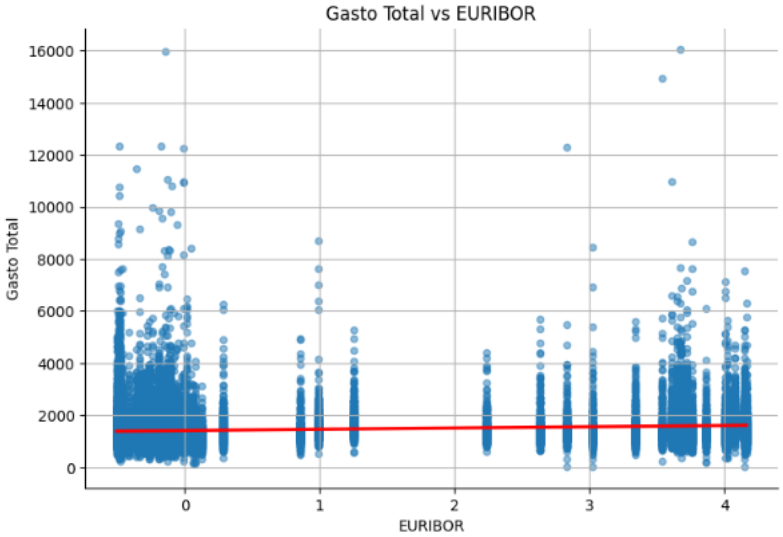


En el gráfico 22 se observa que la densidad de puntos a lo largo del eje X (IPC) es bastante uniforme, lo que sugiere que el gasto total de los turistas en Canarias se mantiene relativamente constante independientemente del valor del Índice de Precios al Consumo (IPC). Esto indica que no hay un patrón claro o una tendencia evidente de aumento o disminución del gasto total en relación con las fluctuaciones del IPC.

La constancia en el gasto total a pesar de las variaciones en el IPC puede ser atribuible a varios factores. Por ejemplo, los turistas a menudo planifican y presupuestan sus vacaciones con anticipación, este presupuesto suele estar determinado por factores personales y económicos específicos del hogar, más que por las fluctuaciones a corto plazo del IPC en su país de origen.

En conclusión, la falta de una relación clara entre el gasto total y el IPC en el gráfico 16 sugiere que otros factores pueden estar más influyentes en las decisiones de gasto de los turistas en Canarias, más allá de las variaciones en los precios al consumidor.

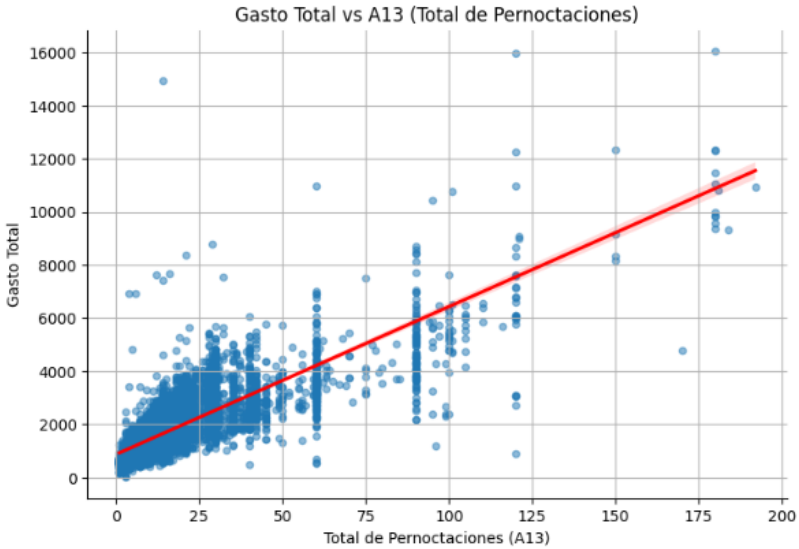
Gráfico 23. Análisis bivariante del gasto total - EURIBOR.



En el gráfico 23 se observa una relación débil o inexistente entre el gasto total y el EURIBOR. La dispersión de datos es notable, especialmente en valores altos de este último, lo que sugiere que no hay una tendencia clara o fuerte correlación entre el gasto total de los turistas en Canarias y las fluctuaciones en las tasas del tipo de interés del mercado monetario.

La línea de media indica que, en promedio, el gasto total no varía significativamente con cambios en el EURIBOR. Esto podría implicar que los turistas no ajustan su gasto de manera significativa en respuesta a las variaciones del indicador macroeconómico. Sin embargo, se observan valores atípicos con gastos totales más altos, lo que sugiere la presencia de ciertos segmentos de turistas que podrían verse menos influenciados por las condiciones representadas por este último.

Gráfico 24. Análisis bivariante del gasto total - pernoctaciones



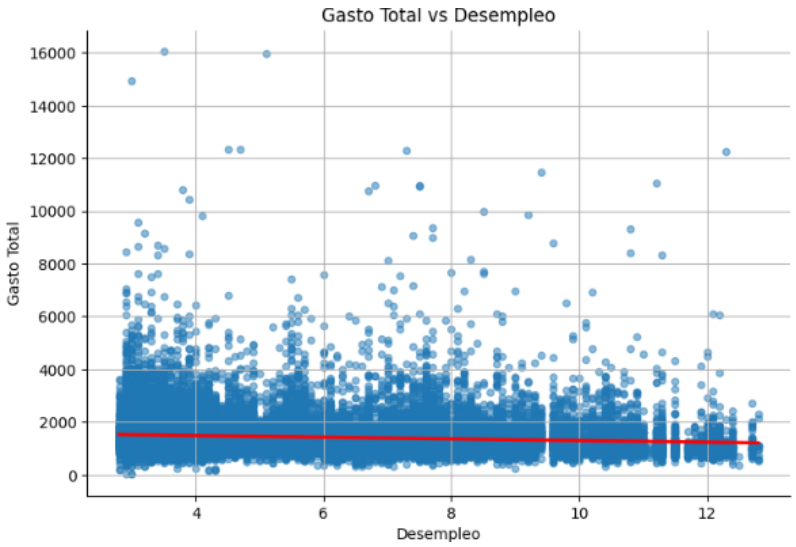
En el gráfico 24 se observa una fuerte correlación positiva entre el gasto total y el total de pernoctaciones en Canarias. A medida que aumenta el número de pernoctaciones, el gasto total

también tiende a aumentar, como se muestra claramente por la inclinación ascendente de la línea de tendencia roja.

Esta relación sugiere que los turistas que optan por alojarse más tiempo en Canarias tienden a gastar más durante su estancia. Esto puede deberse a varios factores, como una mayor participación en actividades turísticas, una exploración más extensa de la región, o simplemente una mayor inversión en comodidades y servicios durante una estancia prolongada.

La fuerte correlación positiva entre el gasto total y el número de pernoctaciones destaca una tendencia importante para la industria turística de Canarias. Este patrón indica que los turistas que permanecen más tiempo tienden a gastar más, lo cual puede ser aprovechado por los operadores turísticos y los servicios locales para crear paquetes y promociones que alienten a los visitantes a extender su estancia.

Gráfico 25. Análisis bivariante del gasto total - tasa de paro



El gráfico 25 muestra que no hay una relación significativa entre la tasa de desempleo y el gasto total de los turistas en Canarias. La mayoría de los puntos se concentran en niveles bajos de desempleo, especialmente entre el 4% y el 8%. A medida que la tasa de desempleo aumenta, hay menos puntos representados en el gráfico, particularmente a partir del 10% de desempleo. Aunque existen algunos puntos con gasto total elevado, estos son excepcionales y no siguen un patrón claro de aumento o disminución con respecto a la tasa de paro.

La línea de tendencia horizontal refuerza esta observación, indicando que el gasto total de los turistas en Canarias parece estar impulsado por otros factores que no están directamente relacionados con las variaciones del indicador macroeconómico. Esto puede sugerir que los turistas no ajustan significativamente su gasto en respuesta a las condiciones locales de empleo, como podrían hacerlo en relación con otros indicadores económicos o sociales.

Gráfico 26. Mapa de calor de correlación de variables respecto al Gasto total

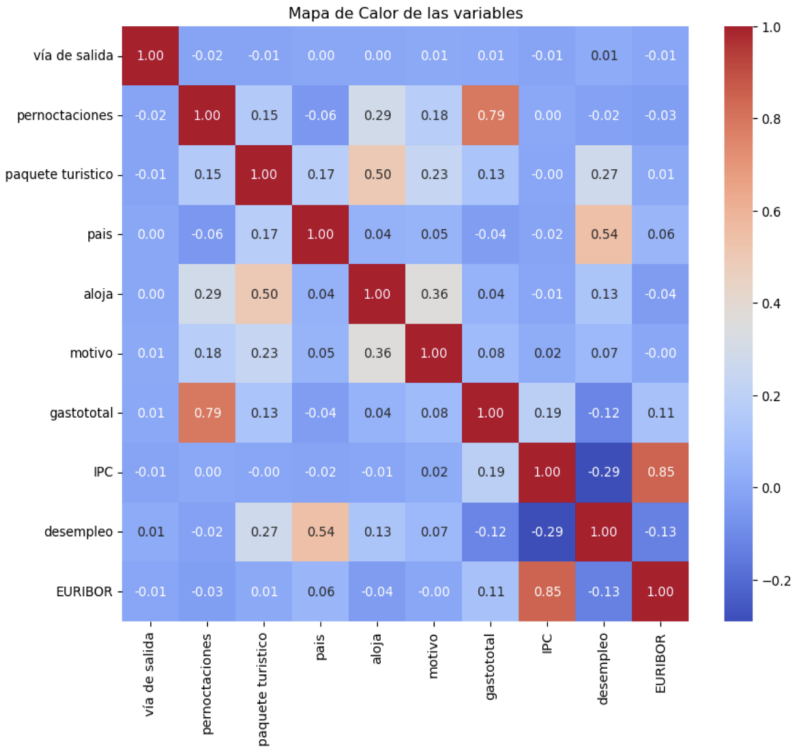


Tabla 4. Ranking correlación variables con gasto total

	VARIABLE	VALOR
1	Pernoctaciones	0.79
2	Alojamiento	0.36
3	Paquete turístico	0.23
4	IPC	0.19
5	Tasa de paro	-0.12
6	EURIBOR	0.11
7	Motivo	0.08
8	País	0.04
9	Vía de salida	0.01

La matriz de correlaciones muestra las relaciones que existen entre las variables que se van a incluir en el modelo. Ordenando las variables de mayor a menor correlación con el gasto total (variable de interés), encontramos lo siguiente: pernoctaciones (0.79), aloja (0.36), paquete turístico (0.23), IPC (0.19), desempleo (-0.12), EURIBOR (0.11), motivo (0.08), país (0.04), vía de salida (0.01).

4.5.3. Comparación de modelos

Para la evaluación de los tres modelos aplicados en este trabajo (Linear Regression, KNN y Random Forest), se analizan las siguientes métricas clave: MAE (Error Absoluto Promedio), RMSE (Raíz Cuadrada del Error Cuadrático Medio) y R^2 (proporción de la varianza explicada por el modelo). Para ello, antes de realizar las predicciones y la evaluación con validación cruzada en los tres modelos, se determinó el valor óptimo de K (el número de vecinos más cercanos a

considerar) para el modelo KNN y el mejor hiperparámetro de número de árboles para el Random Forest. Ajustar estos parámetros es crucial para mejorar la precisión y estabilidad de los modelos, asegurando que las predicciones sean lo más acertadas posible. En cuanto al número de descriptores a utilizar en cada árbol del modelo Random Forest, se decidió dejarlo como la raíz cuadrada del número total de descriptores. Esta práctica es ampliamente utilizada en la literatura debido a su efectividad y los buenos resultados que generalmente produce.

- En el caso del algoritmo KNN, se obtuvo el valor de K óptimo mediante un proceso de ajuste utilizando validación cruzada con 5 iteraciones. Este procedimiento abarcó varios valores de K dentro del rango de 2 a 60, con el objetivo de minimizar los valores promedio de MAE y RMSE en el conjunto de prueba, además de maximizar el coeficiente de determinación R^2. Se puede observar en los Gráficos 27, 28 y 29 que el valor óptimo de K es 9, de acuerdo a las métricas MAE, RMSE y R^2. Por lo tanto, se seleccionó K=9 como el parámetro óptimo para las predicciones con el modelo KNN.

Gráfico 27. MAE vs valor de K

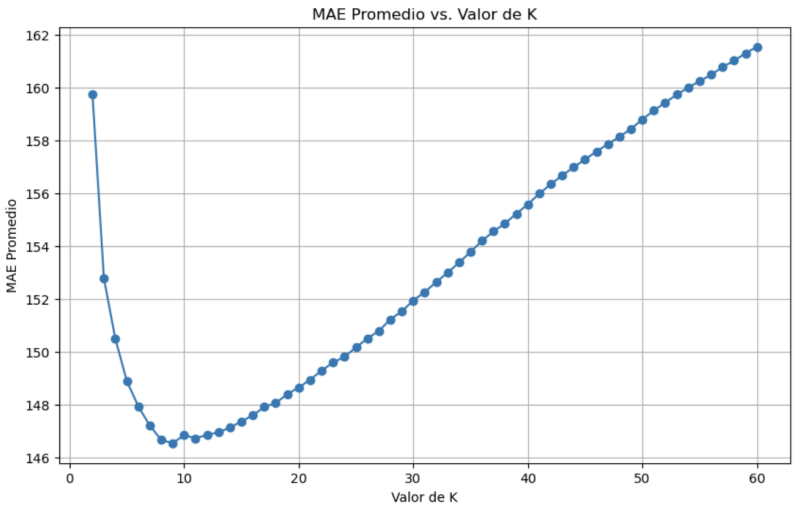


Gráfico 28. RMSE vs valor de K

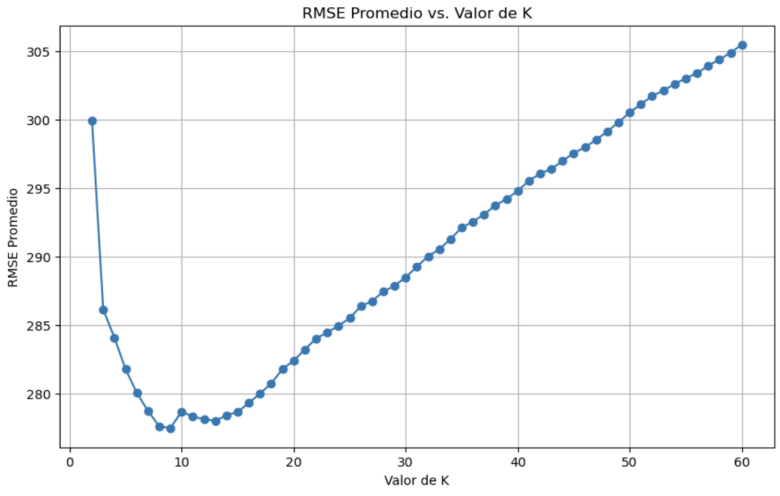
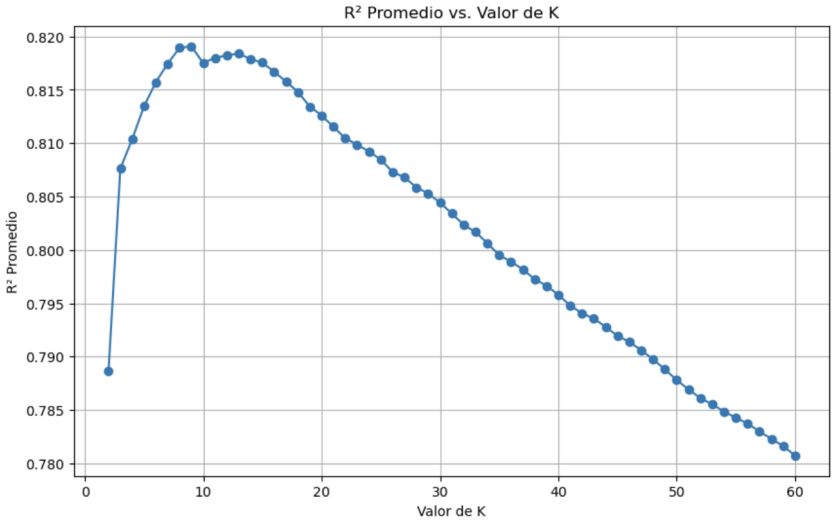


Gráfico 29. R^2 vs valor de K



- Por su parte, para determinar el número óptimo de árboles a utilizar en el modelo Random Forest, se calcularon las métricas MAE, RMSE y R^2 para varios valores de árboles en el rango de 10 a 200. Estos cálculos se realizaron utilizando validación cruzada con 5 iteraciones.

  Puede observarse un rápido descenso del MAE y del RMSE a medida que aumenta el número de árboles, seguido de un comportamiento casi asintótico (al revés ocurre con el R^2). Considerando que un mayor número de árboles incrementa el costo computacional, se seleccionó un valor del número de árboles en la zona donde la curva de MAE y RMSE deja de disminuir rápidamente y donde la curva del R^2 deja de aumentar rápidamente. En el gráfico de MAE (Gráfico 30), esta zona se encuentra entre 40 y 100 árboles. En los gráficos de RMSE y R^2 (Gráfico 31 y 32), la zona de cambio de pendiente se encuentra en el rango de 40 a 70 árboles. Con base en esta observación, se escogió 40 como el número óptimo de árboles.

Gráfico 30. MAE vs nº de árboles

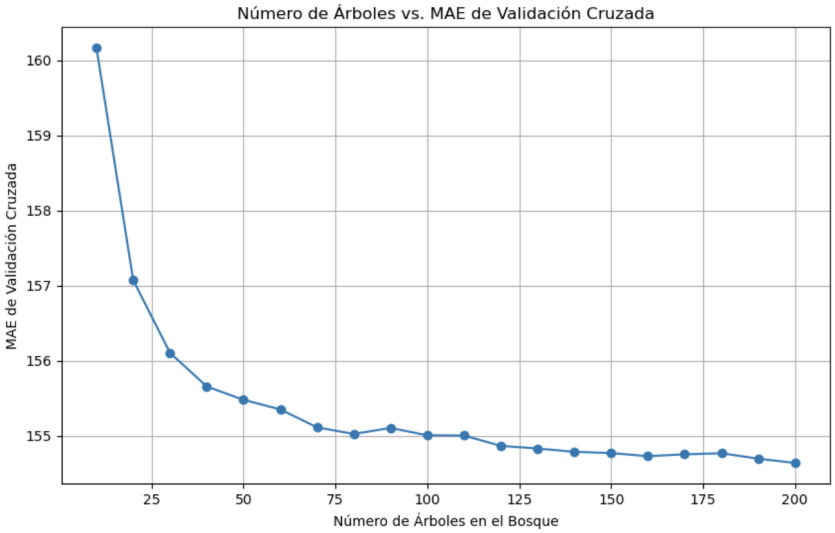


Gráfico 31. RMSE vs nº de árboles

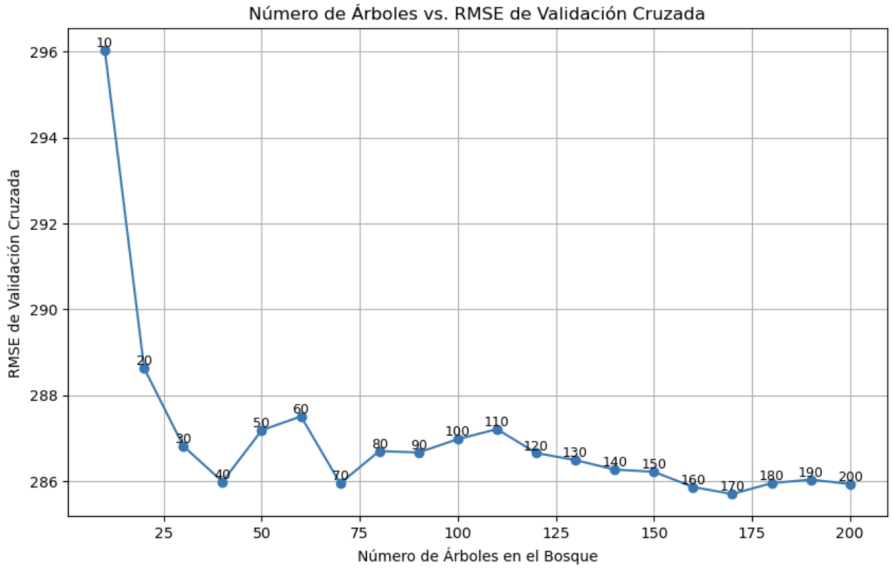
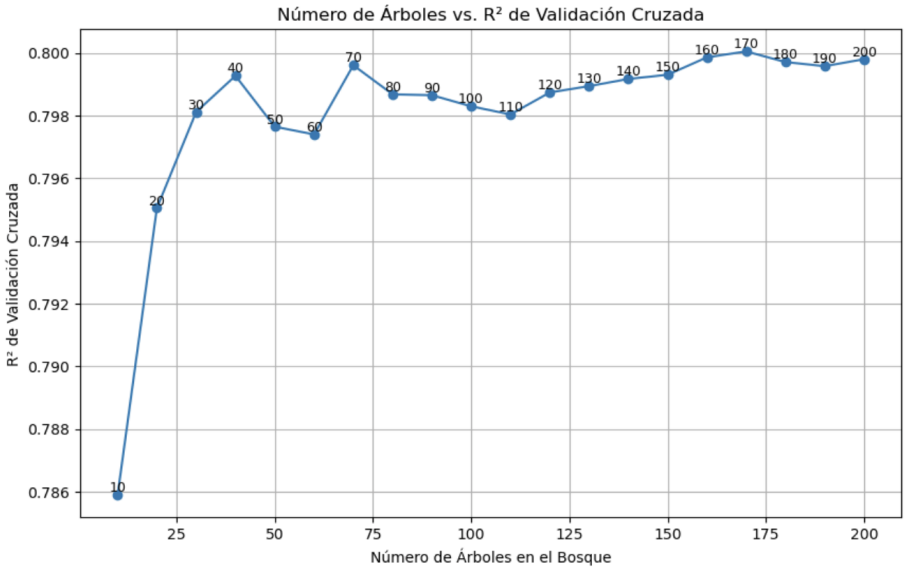


Gráfico 32. R^2 vs nº de árboles



Una vez determinados los valores óptimos para KNN y Random Forest, se procede a realizar las predicciones y las evaluaciones de los tres modelos comparando las métricas obtenidas. A continuación, se presentan las métricas de evaluación para cada modelo:

Tabla 5. Resultados de las métricas MAE, RMSE y R^2

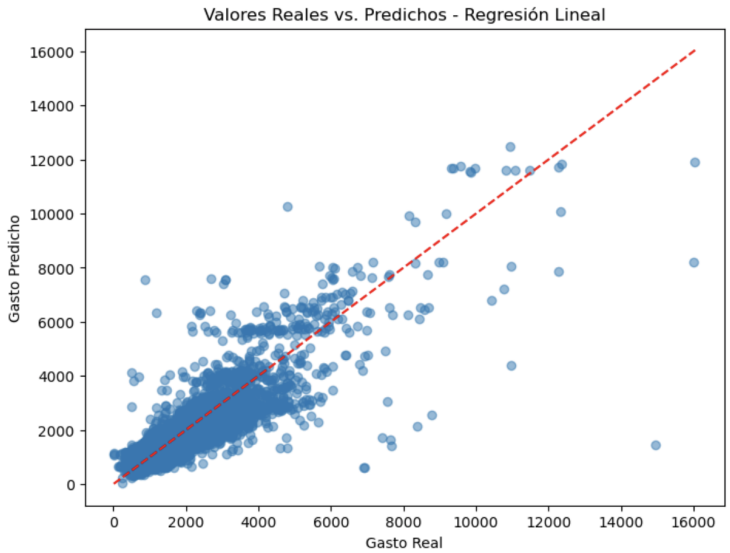
Modelo	Linear Regression	KNN	Random Forest
MAE	201.23	146.41	140.07
RSME	333.68	276.97	264.96
R^2	0.74	0.82	0.83

De acuerdo con las métricas de evaluación, el modelo Random Forest supera a los otros dos modelos (Linear Regression y KNN) en términos de MAE, RMSE y R^2. El Random Forest tiene el menor MAE (140.07), el menor RMSE (264.96) y el mayor R^2 (0.83). Estas métricas indican que el modelo Random Forest ofrece las predicciones más precisas y con menor error.

El menor valor de MAE del modelo Random Forest sugiere que, en promedio, las predicciones del modelo están más cerca de los valores reales en comparación con los otros modelos. El menor valor de RMSE indica que el modelo Random Forest tiene un error de predicción global más bajo, considerando tanto la varianza como el sesgo. Además, el mayor valor de R^2 muestra que el modelo Random Forest explica mejor la variabilidad en los datos en comparación con Linear Regression y el KNN.

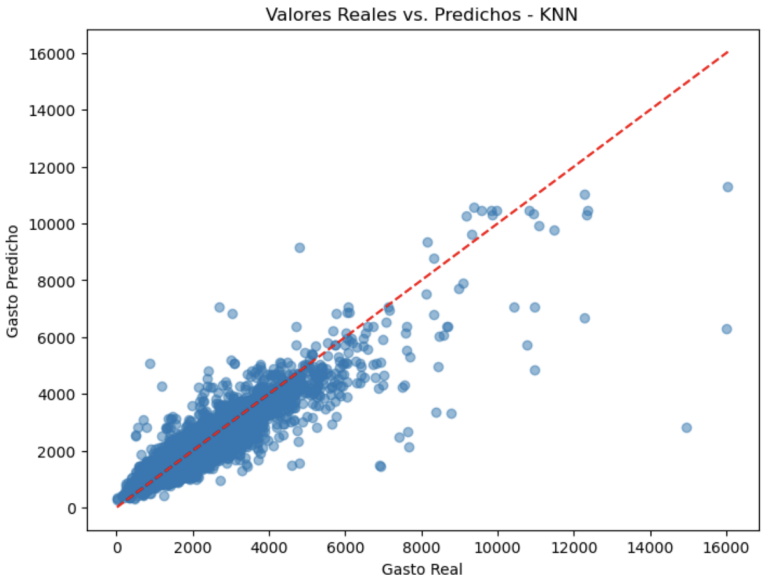
Por lo tanto, se recomienda utilizar el modelo Random Forest para este conjunto de datos, ya que proporciona un mejor equilibrio entre precisión y capacidad de generalización, superando tanto al algoritmo Linear Regression como al KNN en todos los aspectos evaluados. Los gráficos 33, 34 y 35, que apoyan las métricas analizadas, presentan una comparación visual de los valores reales versus los valores predichos utilizando tres modelos de predicción diferentes: Linear Regression, K-Nearest Neighbors (KNN) y Random Forest. Cada uno de los tres gráficos incluye una línea roja que representa la línea ideal donde los valores predichos serían exactamente iguales a los valores reales (es decir, la línea y = x).

Gráfico 33. Gasto total real vs gasto total predicho para LR



En el gráfico 27, que muestra los resultados del modelo Linear Regression, se observa que los puntos están bastante dispersos alrededor de la línea roja ideal. A medida que los valores reales aumentan, la dispersión de los puntos también se incrementa. Aunque se puede ver una tendencia general que sigue la línea ideal, las desviaciones son considerables, especialmente para valores altos de gasto. Esto sugiere que el modelo Linear Regression puede no ser suficientemente robusto para capturar todas las complejidades inherentes en los datos.

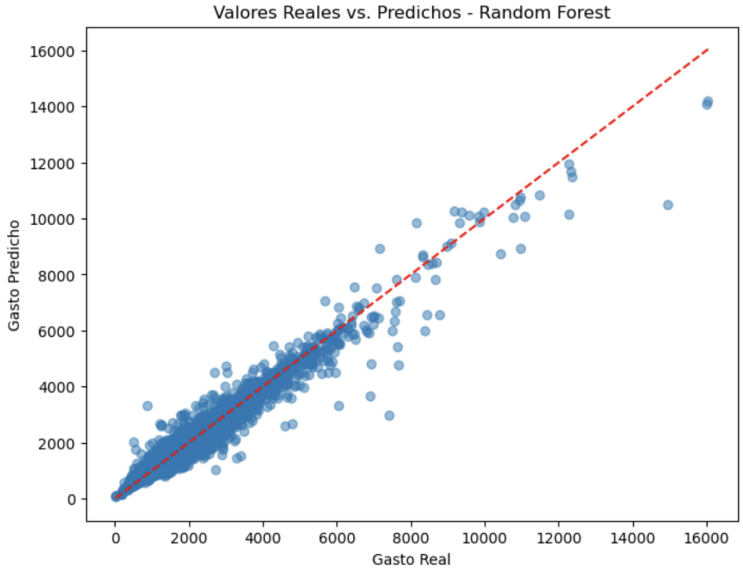
Gráfico 34. Gasto total real vs gasto total predicho para KNN para K=9



El gráfico 28 presenta los resultados del modelo KNN. En este caso, los puntos están más cerca de la línea roja en comparación con el modelo Linear Regression. Aunque aún se observa cierta dispersión, es significativamente menor. El ajuste del modelo KNN parece ser mejor, particularmente para los valores medianos de gasto. Esto indica que KNN tiene una mayor

capacidad para capturar la relación entre los valores reales y predichos, aunque todavía puede haber áreas para mejorar.

Gráfico 35. Gasto total real vs gasto total predicho para RF con 80 árboles



Finalmente, el gráfico 29 muestra los resultados del modelo Random Forest. Aquí, los puntos están más concentrados alrededor de la línea roja, y la dispersión es la menor entre los tres modelos evaluados. Los valores predichos por el modelo Random Forest siguen la tendencia de los valores reales con mayor precisión, lo que sugiere que este modelo tiene la mejor capacidad para capturar la relación subyacente en los datos. La menor dispersión y la mayor cercanía de los puntos a la línea ideal reflejan un ajuste más preciso y confiable.

Estos resultados respaldan las métricas analizadas previamente, indicando que el modelo Random Forest es el más adecuado para predecir los valores en este conjunto de datos, seguido por el KNN y finalmente Linear Regression.

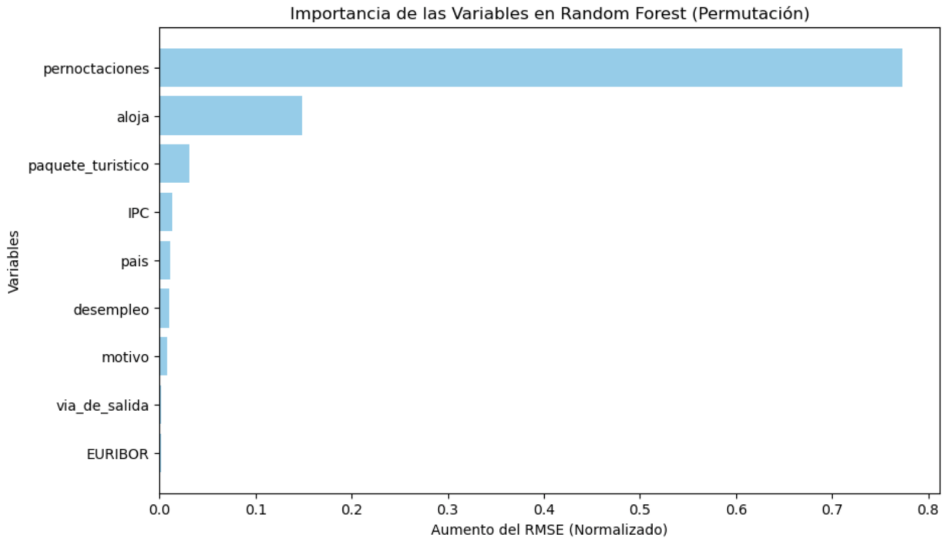
4.5.4. Análisis de la importancia de las variables

La siguiente tabla (Tabla 6) y el siguiente gráfico (Gráfico 36) muestra un análisis detallado de la importancia de diferentes variables en un modelo de Random Forest diseñado para predecir el gasto turístico en Canarias. La representatividad de cada variable se determina mediante un método de permutación, que evalúa el aumento del error cuadrático medio (RMSE) al permutar los valores de cada variable.

Tabla 6. Importancia de las variables Random Forest

FEATURE	IMPORTANCE	NORMALIZED IMPORTANCE
pernoctaciones	375,1862	0,7736
aloja	71,9498	0,1484
paquete_turistico	15,2657	0,0315
IPC	6,2522	0,0129
pais	5,5168	0,0114
desempleo	5,0330	0,0104
motivo	3,8038	0,0078
via_de_salida	1,0261	0,0021
EURIBOR	0,9481	0,0020

Gráfico 36. Importancia de las variables Random Forest



La variable más significativa, con una gran diferencia respecto a las demás, es el número de pernoctaciones. Esto indica que el número de noches que los turistas pasan en su alojamiento tiene el mayor impacto en el gasto turístico en Canarias. Este resultado es intuitivo, ya que una mayor cantidad de noches generalmente implica mayores gastos en alojamiento, comidas y otras actividades turísticas.

La segunda variable en importancia es el tipo o categoría del alojamiento. Este factor también afecta significativamente el gasto turístico, lo cual es plausible porque diferentes categorías de alojamiento (por ejemplo, hoteles de lujo versus hostales) pueden tener precios muy variados, influyendo directamente en el gasto total del turista.

La tenencia de un paquete turístico es otra variable relevante, aunque su importancia es mucho menor en comparación con las pernoctaciones y el tipo de alojamiento.

El Índice de Precios al Consumidor (IPC) ocupa el cuarto lugar en términos de importancia. Aunque su influencia es pequeña en comparación con las variables mencionadas anteriormente, es notable. El IPC refleja la variación en los precios de bienes y servicios, lo que puede afectar el costo de la vida y, por ende, el gasto turístico.

Por último, las demás variables, que incluyen el país de origen de los turistas, la tasa de desempleo, el motivo del viaje, la vía de salida y el EURIBOR, tienen un impacto mucho menor en el gasto turístico.

Capítulo 5. DISCUSIÓN

El principal objetivo del trabajo era desarrollar un modelo predictivo del gasto turístico en Canarias por parte de turistas provenientes de países de la Eurozona, utilizando técnicas de análisis de datos y Machine Learning. Con datos obtenidos de EGATUR, Banco de España y EUROSTAT se obtuvo que el mejor modelo para predecir este caso de estudio es el Random Forest, seguido del KNN y, por último, Linear Regression. Esta superioridad del Random Forest puede explicarse por su capacidad para manejar relaciones complejas y no lineales en los datos, lo que es una limitación de la regresión lineal tradicional. A pesar de las diferencias en los modelos, todos ellos mostraron desempeños similares en términos de las métricas utilizadas: Error Medio Absoluto (MAE), Raíz del Error Cuadrático Medio (RMSE) y el coeficiente de determinación (R^2). Estos resultados coinciden con el estudio de Piña (2018), donde se encontró que los modelos más efectivos fueron, en orden, Deep Learning, Random Forest, Support Vector Machines, KNN y Linear Regression.

Además del rendimiento de los modelos, se llevó a cabo un análisis exhaustivo de la importancia de las variables predictoras. Nuestro estudio reveló que el factor más influyente en el gasto turístico es el número de pernoctaciones, lo cual es consistente con los hallazgos de Piña (2018). Sin embargo, nuestro análisis difiere en la identificación de la segunda variable más relevante. Mientras que Piña (2018) destacó el país de procedencia como el segundo factor más importante, nuestros resultados indican que el tipo de alojamiento ocupa esta posición. Otras variables significativas en nuestro modelo fueron el paquete turístico y el Índice de Precios al Consumo (IPC), relegando al país de procedencia a una quinta posición en términos de relevancia. Por su parte, Moreno (2023) encontró que las variables más relevantes para explicar el gasto turístico son el número de pernoctaciones, el país de origen y el medio de transporte. Esta discrepancia puede deberse a diferencias en los períodos de estudio y las metodologías utilizadas, subrayando la importancia de contextos y enfoques específicos en los estudios de gasto turístico.

En este contexto, resulta relevante comentar que al entender los factores que influyen en el gasto turístico, como el tipo de alojamiento, las autoridades locales pueden planificar mejor el uso de recursos y la infraestructura turística, incluyendo la gestión sostenible del agua, energía y otros recursos naturales. Así, identificar qué tipos de turismo generan mayor gasto puede fomentar prácticas comerciales más sostenibles, como promover alojamientos eco-friendly y actividades que respeten el entorno natural y cultural de las Islas Canarias.

Por último, es fundamental mencionar que la metodología inicialmente planeada resultó ser muy útil para proporcionar un marco estructurado y coherente para la recolección y análisis de datos. No obstante, a lo largo del desarrollo del proyecto, nos enfrentamos a varios desafíos que requirieron adaptaciones y ajustes. Por ejemplo, en nuestra planificación inicial, consideramos el uso de modelos más complejos como Deep Learning y Support Vector Machines (SVM). Sin embargo, debido a limitaciones de tiempo y a la capacidad computacional disponible, tomamos la decisión estratégica de centrarnos en modelos más accesibles y menos exigentes computacionalmente, como Random Forest y K-Nearest Neighbors (KNN). Estos modelos,

además de ser más manejables dentro de nuestras restricciones, demostraron ser altamente efectivos para el propósito de nuestro estudio.

Además, es importante señalar que el estudio no está exento de limitaciones. Por un lado, debemos reconocer las limitaciones comunes a todos los estudios basados en encuestas, que incluyen el sesgo en la selección de la muestra pudiendo no ser completamente representativa de la población objetivo, y las posibles inconsistencias o inexactitudes en las respuestas de los encuestados, como destaca Alvira (2011). Por otro lado, la falta de inclusión de otras variables en el modelo puede influir en la variabilidad del gasto turístico; aspectos como las diferencias socioeconómicas podrían estar afectando a la variable objetivo.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

El presente estudio tuvo como objetivo desarrollar un modelo predictivo del gasto turístico en Canarias por parte de turistas provenientes de países de la Eurozona. Utilizando técnicas de análisis de datos y Machine Learning se analizaron diversas variables y se compararon tres modelos diferentes: Linear Regression, K-Nearest Neighbors (KNN) y Random Forest. A continuación, se resumen las principales conclusiones del estudio:

Los modelos construidos, una vez entrenados, se evaluaron utilizando RMSE, MAE y R^2. Entre ellos, el modelo Random Forest presentó las mejores métricas, estableciendo un orden de desempeño como sigue:

RF > KNN > LR

La robustez del Random Forest en el manejo de datos complejos y su capacidad para capturar relaciones no lineales entre las variables lo convierten en la opción más adecuada para predecir el gasto turístico en Canarias.

Por su parte, en el análisis preliminar bivariante de las variables, se observó una fuerte dependencia entre el número de pernoctaciones y el gasto total. El número de pernoctaciones realizadas (o la duración de la estancia) demostró tener una gran relevancia al estudiar el gasto turístico, mostrando un impacto considerable en comparación con otras variables. Esto se confirmó con el análisis de la importancia de las variables en el modelo de Random Forest que reveló que las pernoctaciones y el tipo de alojamiento son los factores más determinantes en el gasto turístico en Canarias. La presencia de un paquete turístico y el Índice de Precios al Consumidor también tienen un impacto significativo, aunque menor. Las demás variables estudiadas tienen un efecto relativamente reducido.

6.2 Conclusiones personales

Desarrollar este proyecto sobre el gasto turístico en Canarias por parte de turistas de la Eurozona ha sido una experiencia educativa y significativamente práctica. Desde el principio, nos intrigó la complejidad del tema debido al importante impacto económico del turismo en la región. A lo largo del estudio, pudimos apreciar la diversidad de factores que influyen en las decisiones de gasto de los turistas, desde variables como el número de pernoctaciones hasta el tipo de alojamiento y factores socioeconómicos adicionales.

Uno de los aspectos más destacados fue enfrentar los desafíos metodológicos y técnicos durante el proyecto. La necesidad de ajustar nuestras herramientas y enfoques debido a limitaciones de tiempo y recursos fue una lección valiosa en términos de flexibilidad y adaptabilidad en la investigación científica.

Desde una perspectiva profesional, este proyecto ha fortalecido nuestras habilidades en análisis de datos y modelado predictivo, proporcionándonos herramientas prácticas para abordar problemas complejos en el futuro. La aplicación de técnicas avanzadas como Random Forest y KNN y la comparación entre ellos amplió nuestro entendimiento sobre las capacidades y limitaciones de cada método. Esto no solo nos prepara mejor para enfrentar desafíos similares en nuestra carrera profesional, sino que también nos sensibiliza sobre la importancia de la

innovación y la adaptabilidad en el campo del análisis predictivo y la toma de decisiones estratégicas.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Para futuras líneas de trabajo se sugiere considerar la incorporación de datos a nivel individual, como: estado civil, ingresos mensuales, nivel educativo, edad y género, entre otros. Esto permitiría una segmentación más detallada de los turistas según criterios como edad, género y tipo de viaje (familiar, individual, en pareja), lo cual podría revelar patrones de gasto más específicos y personalizados. Sería interesante también la inclusión de variables adicionales que podrían influir en el gasto turístico, como la oferta de actividades turísticas locales, eventos específicos y la calidad de los servicios turísticos.

Además, sería pertinente analizar el impacto de factores ambientales y climáticos, dado el contexto cálido de Canarias como destino turístico. En adición, explorar el uso de modelos de Machine Learning más avanzados, como redes neuronales profundas, que podría mejorar significativamente la precisión de las predicciones del gasto turístico.

Estas líneas de trabajo tienen el potencial de ampliar y enriquecer los resultados obtenidos hasta ahora, proporcionando una visión más completa y precisa del comportamiento del gasto turístico en Canarias, que no solo podría beneficiar la planificación estratégica del sector turístico, sino también contribuir al desarrollo económico sostenible y a una mejor experiencia para los visitantes en la región.

Capítulo 8. REFERENCIAS

Alaminos-Fernández, A. F. (2022). *Árboles de decisión en r con random forest.* Universidad de Alicante. Obets Ciencia Abierta. Alicante: Limencop.

Albaladejo, I. P., González-Martínez, M. I., & Martínez-García, M. P. (2016). Nonconstant reputation effect in a dynamic tourism demand model for Spain. *Tourism Management*, *53*, 132-139.

Álvarez-Díaz, M., González-Gómez, M., & Otero-Giráldez, M. S., & Iglesias, A. B. T. (2014). Modelización econométrica de la demanda de turistas británicos a España. *Documentos de traballo do Departamento de Economía Aplicada*, (4), 1.

Álvarez-Díaz, M., González-Gómez, M., & Otero-Giráldez, M. S. (2016). La modelización de la demanda de turismo de economías emergentes: el caso de la llegada de turistas rusos a España. *Cuadernos de Economía*, *39*(110), 112-125.

Alvira, F. M. (2011). *La encuesta: una perspectiva general metodológica* (Vol. 35). CIS.

Del Arco, M. J., Wildpret, W., Pérez-de-Paz, P. L., Rodríguez, O., Acebes, J. R., García, A., ... & García, S. (2006). Mapa de vegetación de Canarias. *GRAFCAN, Santa Cruz de Tenerife*.

EXCELTUR (2023). IMPACTUR CANARIAS 2023. Estudio del impacto económico del turismo sobre la economía y el empleo de las Islas Canarias. https://www.exceltur.org/wp-content/uploads/2023/12/Impactur-Canarias-2022-.pdf

Hernández Martín, R., León González, C., Baute Díaz, N., Simancas Cruz, M. R., Padrón Fumero, N., Herrera Priano, F. Á., ... & Perdomo Santana, M. F. (2023). *Sostenibilidad del Turismo en Canarias. Informe 2023.* Observatorio Turístico de Canarias. Resumen ejecutivo.

Hernández Martín, R., Viera González, J. M., & Toledo Bordón, P. S. (2024). *Distribución y concentración del alojamiento turístico en Canarias. Plazas hoteleras, en apartamentos, vivienda vacacional y población.* Observatorio Turístico de Canarias.

INE (2021). Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero. https://www.ine.es/jaxiT3/Tabla.htm?t=2915&L=0

INE (2022). Encuesta de ocupación en apartamentos turísticos 2022. https://www.ine.es/dynt3/inebase/es/index.htm?padre=9978&capsel=9986

INE (2023). Encuesta de gasto turístico 2023. https://www.ine.es/dynt3/inebase/index.htm?padre=3620&capsel=3620

Dolores, M. P. H., Basulto, J. S., & Camúñez, J. A. R. (2019). Un antecedente histórico de regresión lineal: la estimación mediana propuesta por Boscovich. *Gaceta de la Real Sociedad Matemática Española*, *22*(2), 351-364.

Moreno Bastante, M. (2023). *Comparación de modelos Machine Learning para el gasto turístico tras el COVID-19,* Trabajo Fin de Máster, Universidad Pontificia Comillas. https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/68678/TFG%20MORENO%20BASTANTE%20MARIA.pdf?sequence=1

Narváez, M. E., Arias, R. J. M. A. M., Campana, D. B. P., & Guamán, G. A. I. (2022). Predicción de clientes potenciales utilizando K vecino más cercano en el área de negocios de la Cooperativa Riobamba. *Revista Perspectivas*, *4*(1), 21-26.

Piña Miranda, L. (2018). *Modelación y predicción del gasto de turistas en España enfocado desde el análisis de datos*, Trabajo de Fin de Máster, Universidad de Islas Baleares. https://dspace.uib.es/xmlui/handle/11201/149300

PROMOTUR (2023). Perfil del turista Canarias e Islas 2023. https://investigacion.turismodeislascanarias.com/sites/default/files/2024-03/Promotur_canarias%20e%20islas_2023.pdf

Sánchez, F. J. S., & Sánchez, A. M. S. (2021). Modelos de predicción para el sector turístico andaluz mediante métodos estadísticos avanzados. *Cuadernos de turismo*, (48), 183-208.

Scikit-learn (2024). *Machine Learning in Python.* https://scikit-learn.org/stable/index.html

Stock, J. H., Watson, M. W., & Larrión, R. S. (2012). *Introducción a la Econometría*.

Capítulo 9. ANEXOS

UNIÓN DE LA BASE DE DATOS DE EGATUR

In [ ]:
```python
import pandas as pd
import os

# Ruta principal donde se encuentran las carpetas "2015", "2016", ..., "2024"
ruta_archivos = r'/Users/paulahernandezrodriguez/Desktop/BBDD'

# Lista de nombres de carpetas (años)
carpetas = [str(year) for year in range(2015, 2025)]

# Crear una lista para almacenar los DataFrames de cada archivo
dfs = []

# Iterar sobre cada carpeta
for carpeta in carpetas:
    folder_path = os.path.join(ruta_archivos, carpeta)
    # Asegurarse de que la carpeta exista
    if os.path.exists(folder_path):
        # Depuración: imprimir la carpeta actual
        print(f"Explorando la carpeta: {folder_path}")
        # Iterar sobre cada archivo en la carpeta
        for file_name in os.listdir(folder_path):
            if file_name.endswith('.txt'):
                file_path = os.path.join(folder_path, file_name)
                # Depuración: imprimir el archivo que se está leyendo
                print(f"Leyendo archivo: {file_path}")
                # Leer el archivo TXT con tabulaciones como delimitador
                df = pd.read_csv(file_path, delimiter='\t')
                dfs.append(df)  # Agregar el DataFrame a la lista
    else:
        print(f"Carpeta no encontrada: {folder_path}")

# Verificar si se encontraron y leyeron archivos TXT
if not dfs:
    print("No se encontraron archivos TXT en las carpetas especificadas.")
else:
    # Concatenar los DataFrames en uno solo
    datos_combinados = pd.concat(dfs, ignore_index=True)
    # Guardar los datos combinados en un nuevo archivo CSV
    ruta_salida = os.path.join(ruta_archivos, 'datos_combinados.csv')
    datos_combinados.to_csv(ruta_salida, index=False)
    print(f"Archivo CSV combinado guardado en: {ruta_salida}")


```

ARREGLO DE LAS BASES DE DATOS DE IPC, TASA DE DESEMPLEO Y EURIBOR

EURIBOR

In [ ]:
```python
# Cargar el archivo CSV
datos_euribor = pd.read_csv('datos_euribor.csv')

# Mostrar los primeros registros para verificar los datos
datos_euribor.head()
```

In [ ]:
```python
# Cambiar el formato de la columna 'mm_aaaa' a MMYYYY
datos_euribor['mm_aaaa'] = pd.to_datetime(datos_euribor['mm_aaaa'], format='%Y-%m').dt.strftime('%m%Y')

# Eliminar ceros a la izquierda en el mes
datos_euribor['mm_aaaa'] = datos_euribor['mm_aaaa'].str.lstrip('0')
```

In [ ]:
```python
# Guardar los datos en un nuevo archivo CSV
datos_euribor.to_csv('datos_euribor_modificado.csv', index=False)
```

IPC

In [ ]:
```python
# Cargar el archivo CSV
datos_ipc = pd.read_csv('prc_hicp_midx_page_linear.csv')

# Mostrar los primeros registros
print(datos_ipc.head())
```

In [ ]:
```python
# Reemplazar los valores en la columna 'geo'
datos_ipc['geo'] = datos_ipc['geo'].replace({'BE': 2, 'FR': 3, 'DE': 1, 'IE': 4, 'IT': 5, 'NL': 6, 'PT': 7})

# Mostrar los primeros registros para verificar los cambios
print(datos_ipc.head())
```

In [ ]:
```python
datos_ipc = datos_ipc[['geo', 'TIME_PERIOD', 'OBS_VALUE']]
datos_ipc['TIME_PERIOD'] = pd.to_datetime(datos_ipc['TIME_PERIOD']).dt.strftime('%m%Y')
datos_ipc['TIME_PERIOD'] = datos_ipc['TIME_PERIOD'].str.lstrip('0')
```

In [ ]:
```python
datos_ipc.rename(columns={'geo': 'pais', 'TIME_PERIOD': 'mm_aaaa', 'OBS_VALUE': 'IPC'}, inplace=True)
```

In [ ]:
```python
# Guardar los datos en un nuevo archivo CSV
datos_ipc.to_csv('datos_ipc_modificado.csv', index=False)
```

DESEMPLEO

In [ ]:
```
# Cargar el archivo CSV
datos_empleo = pd.read_csv('ei_lmhr_m_page_linear.csv')
```

In [ ]:
```
#Auedarnos solo con las columnas que nos interesan
datos_empleo = datos_empleo[['geo', 'TIME_PERIOD', 'OBS_VALUE']]

# Reemplazar los valores en la columna 'geo'
datos_empleo['geo'] = datos_empleo['geo'].replace({'BE': 2, 'FR': 3, 'DE': 1, 'IE': 4, 'IT': 5, 'NL': 6, 'PT':
datos_empleo.rename(columns={'geo': 'pais', 'TIME_PERIOD': 'mm_aaaa', 'OBS_VALUE': 'desempleo'}, inplace=True)
```

In [ ]:
```
datos_empleo['mm_aaaa'] = pd.to_datetime(datos_empleo['mm_aaaa']).dt.strftime('%m%Y')
datos_empleo['mm_aaaa'] = datos_empleo['mm_aaaa'].str.lstrip('0')
```

In [ ]:
```
# Guardar los datos en un nuevo archivo CSV
datos_empleo.to_csv('datos_desempleo_modificado.csv', index=False)
```

UNIÓN DE LAS VARIABLES MACROECONÓMICAS A LA BASE DE DATOS DE EGATUR

EURIBOR

In [ ]:
```
datos_euribor = pd.read_csv('datos_euribor_modificado.csv')
datos_combinados = pd.read_csv('datos_combinados.csv')
```

In [ ]:
```
datos_combinados2 = pd.merge(datos_combinados, datos_euribor, on=['mm_aaaa'], how='left')
datos_combinados2.to_csv('datos_combinados2.csv', index=False)
```

IPC

In [ ]:
```
datos_combinados2 = pd.read_csv('datos_combinados2.csv')
datos_ipc_modificado = pd.read_csv('datos_ipc_modificado.csv')
```

In [ ]:
```
datos_combinados3 = pd.merge(datos_combinados2, datos_ipc_modificado, on=['mm_aaaa'], how='left')
datos_combinados3.to_csv('datos_combinados3.csv', index=False)
```

DESEMPLEO

In [ ]:
```
datos_combinados3 = pd.read_csv('datos_combinados3.csv')
datos_desempleo_modificado = pd.read_csv('datos_desempleo_modificado.csv')
```

In [ ]:
```
datos_combinados4= pd.merge(datos_combinados3, datos_desempleo_modificado, on=['mm_aaaa', 'pais'], how='left')
datos_combinados4.to_csv('datos_combinados4.csv', index=False)
```

FILTRACIÓN POR CCAA CANARIAS Y POR PAÍSES DE LA EUROZONA

In [ ]:
```
dfd = pd.read_csv('datos_combinados4.csv')
df_filtrado = dfd[(dfd['ccaa'] == 5) & (dfd['pais'].isin([1, 2, 3, 4, 5, 6, 7]))]
df_filtrado.to_csv('datos_definitivos.csv', index=False)
```

In [ ]:
```
# Cargar el archivo CSV con ';' como delimitador
df = pd.read_csv('datos_definitivos.csv', delimiter=',')

# Analizar los Nan's
print(df.isna().any())
null_counts= df.isna().sum()
print(null_counts)
```

ANÁLISIS UNIVARIANTE

PAÍS

In [2]:

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Crear un diccionario para mapear los códigos de país a los nombres de país
codigo_pais = {
    1: 'Alemania',
    2: 'Bélgica',
    3: 'Francia',
    4: 'Irlanda',
    5: 'Italia',
    6: 'Países Bajos',
    7: 'Portugal'
}

# Reemplazar los códigos de país en la columna 'pais' con los nombres de país correspondientes
df['pais'] = df['pais'].map(codigo_pais)

# Calcular la frecuencia relativa (en %) de cada país
df_relative = df['pais'].value_counts(normalize=True).reset_index()
df_relative.columns = ['pais', 'relative_freq']
df_relative['relative_freq'] *= 100  # Convertir a porcentaje

# Ordenar el DataFrame por la frecuencia relativa
df_relative = df_relative.sort_values(by='relative_freq', ascending=False)

# Graficar la frecuencia relativa de cada país con la misma gama de colores
plt.figure(figsize=(10, 6))
barplot = sns.barplot(data=df_relative, x='pais', y='relative_freq', palette='viridis')
plt.title('País de origen de los turistas')
plt.xlabel('País')
plt.ylabel('Frecuencia relativa (%)')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Añadir porcentajes encima de cada barra
for p in barplot.patches:
    barplot.annotate(f'{p.get_height():.1f}%',
                     (p.get_x() + p.get_width() / 2., p.get_height()),
                     ha='center', va='center',
                     xytext=(0, 10),
                     textcoords='offset points')

plt.show()
```

ALOJAMIENTO

In [6]:

```python
# Crear un diccionario para mapear los códigos de alojamiento a sus nombres
codigo_aloja = {
    1: 'Hoteles y similares',
    2: 'Resto de mercado',
    3: 'Alojamiento no de mercado'
}

# Reemplazar los códigos de alojamiento en la columna 'aloja' con los nombres correspondientes
df['aloja'] = df['aloja'].map(codigo_aloja)

# Calcular la frecuencia relativa (en %) de cada tipo de alojamiento
df_relative_aloja = df['aloja'].value_counts(normalize=True).reset_index()
df_relative_aloja.columns = ['aloja', 'relative_freq']
df_relative_aloja['relative_freq'] *= 100  # Convertir a porcentaje

# Ordenar el DataFrame por la frecuencia relativa
df_relative_aloja = df_relative_aloja.sort_values(by='relative_freq', ascending=False)

# Graficar la frecuencia relativa de cada tipo de alojamiento con la misma gama de colores
plt.figure(figsize=(10, 6))
ax = sns.barplot(data=df_relative_aloja, x='aloja', y='relative_freq', palette='viridis')
plt.title('Tipos de alojamiento de los turistas')
plt.xlabel('Tipo de alojamiento')
plt.ylabel('Frecuencia relativa (%)')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Agregar etiquetas con los valores encima de cada barra
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.1f') + '%',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha = 'center', va = 'center',
                xytext = (0, 10),
                textcoords = 'offset points')

plt.show()

```

VÍA DE SALIDA

In [43]:

```
# Crear un diccionario para mapear los códigos de la variable 'A1' a los nombres correspondientes
codigo_A1 = {
    1: 'Carretera',
    2: 'Aeropuerto',
    3: 'Puerto',
    4: 'Tren'
}

# Reemplazar los códigos de la variable 'A1' con los nombres correspondientes
df['A1'] = df['A1'].map(codigo_A1)

# Calcular la frecuencia relativa (en %) de cada categoría de la variable 'A1'
df_relative_A1 = df['A1'].value_counts(normalize=True).reset_index()
df_relative_A1.columns = ['A1', 'relative_freq']
df_relative_A1['relative_freq'] *= 100  # Convertir a porcentaje

# Ordenar el DataFrame por la frecuencia relativa
df_relative_A1 = df_relative_A1.sort_values(by='relative_freq', ascending=False)

# Graficar la frecuencia relativa de cada categoría de la variable 'A1' con la misma gama de colores
plt.figure(figsize=(10, 6))
ax = sns.barplot(data=df_relative_A1, x='A1', y='relative_freq', palette='viridis')
plt.title('Vía de salida de los turistas')
plt.xlabel('Vía de salida')
plt.ylabel('Frecuencia relativa (%)')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Agregar etiquetas con los valores encima de cada barra
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.1f') + '%',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha = 'center', va = 'center',
                xytext = (0, 10),
                textcoords = 'offset points')

plt.show()

```

MOTIVO DE VIAJE

In [9]:

```
# Crear un diccionario para mapear los códigos de motivo a sus nombres
codigo_motivo = {
    1: 'Ocio/vacaciones',
    2: 'Negocios',
    3: 'Resto'
}

# Reemplazar los códigos de motivo en la columna 'motivo' con los nombres correspondientes
df['motivo'] = df['motivo'].map(codigo_motivo)

# Calcular la frecuencia relativa (en %) de cada tipo de motivo
df_relative_motivo = df['motivo'].value_counts(normalize=True).reset_index()
df_relative_motivo.columns = ['motivo', 'relative_freq']
df_relative_motivo['relative_freq'] *= 100  # Convertir a porcentaje

# Ordenar el DataFrame por la frecuencia relativa
df_relative_motivo = df_relative_motivo.sort_values(by='relative_freq', ascending=False)

# Graficar la frecuencia relativa de cada tipo de motivo con la misma gama de colores
plt.figure(figsize=(10, 6))
ax = sns.barplot(data=df_relative_motivo, x='motivo', y='relative_freq', palette='viridis')
plt.title('Motivos de viaje de los turistas')
plt.xlabel('Motivo de viaje')
plt.ylabel('Frecuencia relativa (%)')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Agregar etiquetas con los valores encima de cada barra
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.1f') + '%',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha = 'center', va = 'center',
                xytext = (0, 10),
                textcoords = 'offset points')

plt.show()


```

PAQUETE TURÍSTICO

In [10]:
```
# Crear un diccionario para mapear los códigos de la variable 'A16' a los nombres correspondientes
codigo_A16 = {
    1: 'SI',
    6: 'NO'
}

# Reemplazar los códigos de la variable 'A16' con los nombres correspondientes
df['A16'] = df['A16'].map(codigo_A16)

# Calcular la frecuencia relativa (en %) de cada categoría de la variable 'A16'
df_relative_A16 = df['A16'].value_counts(normalize=True).reset_index()
df_relative_A16.columns = ['A16', 'relative_freq']
df_relative_A16['relative_freq'] *= 100  # Convertir a porcentaje

# Ordenar el DataFrame por la frecuencia relativa
df_relative_A16 = df_relative_A16.sort_values(by='relative_freq', ascending=False)

# Graficar la frecuencia relativa de cada categoría de la variable 'A16' con la misma gama de colores
plt.figure(figsize=(8, 6))
ax = sns.barplot(data=df_relative_A16, x='A16', y='relative_freq', palette='viridis')
plt.title('Distribución de la variable paquete turístico')
plt.xlabel('Paquete turístico')
plt.ylabel('Frecuencia relativa (%)')
plt.xticks(rotation=0)  # No rotar etiquetas del eje x

# Agregar etiquetas con los valores encima de cada barra
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.1f') + '%',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center',
                xytext=(0, 10),
                textcoords='offset points')

plt.show()
```

PERNOCTACIONES

In [63]:
```
# Calcular estadísticas descriptivas
descripcion = df['A13'].describe()

# Visualizar la distribución de las pernoctaciones
plt.figure(figsize=(10, 6))
sns.histplot(df['A13'], bins=20, color='skyblue')
plt.title('Distribución de pernoctaciones')
plt.xlabel('Total de pernoctaciones')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()
# Imprimir estadísticas descriptivas
print(descripcion)
```

GASTO TOTAL

In [62]:
```
# Calcular estadísticas descriptivas
descripcion_gasto = df['gastototal'].describe()

# Visualizar la distribución del gasto total
plt.figure(figsize=(10, 6))
sns.histplot(df['gastototal'], bins=20, color='skyblue')
plt.title('Distribución del gasto total')
plt.xlabel('Gasto Total')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()

# Imprimir estadísticas descriptivas
print(descripcion_gasto)
```

IPC

In [11]:
```
# Calcular estadísticas descriptivas
descripcion_gasto = df['IPC'].describe()

# Visualizar la distribución del gasto total
plt.figure(figsize=(10, 6))
sns.histplot(df['IPC'], bins=20, color='skyblue')
plt.title('Distribución del IPC')
plt.xlabel('IPC')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()

# Imprimir estadísticas descriptivas
print(descripcion_gasto)
```

DESEMPLEO

In [67]:
```
# Calcular estadísticas descriptivas
descripcion_gasto = df['desempleo'].describe()

# Visualizar la distribución del gasto total
plt.figure(figsize=(10, 6))
sns.histplot(df['desempleo'], bins=20, color='skyblue')
plt.title('Distribución del desempleo')
plt.xlabel('Tasa de paro')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()

# Imprimir estadísticas descriptivas
print(descripcion_gasto)
```

EURIBOR

In [12]:
```
# Calcular estadísticas descriptivas
descripcion_euribor = df['EURIBOR'].describe()

# Visualizar la distribución de EURIBOR
plt.figure(figsize=(10, 6))
sns.histplot(df['EURIBOR'], bins=20, kde=False, color='skyblue')
plt.title('Distribución de EURIBOR')
plt.xlabel('EURIBOR')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()

# Imprimir estadísticas descriptivas
print(descripcion_euribor)
```

ANÁLISIS BIVARIANTE

VÍA DE SALIDA

In [5]:
```
# Crear un diccionario para mapear los códigos numéricos a los nombres correspondientes
codigo_A1 = {
    1: 'Carretera',
    2: 'Aeropuerto',
    3: 'Puerto',
    4: 'Tren'
}

# Reemplazar los códigos numéricos en la columna 'A1' con los nombres correspondientes
df['A1'] = df['A1'].map(codigo_A1)

plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='A1', y='gastototal', palette='viridis')
plt.title('Gráfico de caja: Gasto Total por tipo de vía de salida')
plt.xlabel('Vía de salida')
plt.ylabel('Gasto Total')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Calcular y agregar la media de gasto total para cada categoría de A1
mediagastototal_por_A1 = df.groupby('A1')['gastototal'].mean()

# Dibujar línea punteada para representar la media dentro de cada caja
for i, value in enumerate(mediagastototal_por_A1):
    plt.plot([i-0.4, i+0.4], [value, value], color='red', linestyle='--', linewidth=2, label='Media' if i == 0

plt.legend()  # Mostrar leyenda
plt.grid(True)
plt.show()

```

PAÍS

In [7]:

```
# Crear un diccionario para mapear los códigos numéricos a los nombres de los países
codigo_pais = {
    1: 'Alemania',
    2: 'Bélgica',
    3: 'Francia',
    4: 'Irlanda',
    5: 'Italia',
    6: 'Países Bajos',
    7: 'Portugal'
}

# Reemplazar los códigos numéricos en la columna 'pais' con los nombres correspondientes
df['A0_7'] = df['A0_7'].map(codigo_pais)

sns.boxplot(data=df, x='pais', y='gastototal', palette='viridis')
plt.title('Gasto Total por País')
plt.xlabel('País')
plt.ylabel('Gasto Total')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Calcular y agregar la media del gasto total para cada país
media_gastototal_por_pais = df.groupby('pais')['gastototal'].mean()

# Dibujar línea punteada para representar la media dentro de cada caja
for i, value in enumerate(media_gastototal_por_pais):
    plt.plot([i-0.4, i+0.4], [value, value], color='red', linestyle='--', linewidth=2, label='Media' if i == 0

plt.legend()  # Mostrar leyenda
plt.grid(True)
plt.show()

```

TIPO DE ALOJAMIENTO

In [33]:

```
# Crear un diccionario para mapear los códigos numéricos a los nombres correspondientes
codigo_aloja = {
    1: 'Hoteles y similares',
    2: 'Resto de mercado',
    3: 'Alojamiento no de mercado'
}

# Reemplazar los códigos numéricos en la columna 'aloja' con los nombres correspondientes
df['aloja'] = df['aloja'].map(codigo_aloja)

plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='aloja', y='gastototal', palette='viridis')
plt.title('Gasto Total por tipo de alojamiento')
plt.xlabel('Tipo de alojamiento')
plt.ylabel('Gasto Total')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Calcular y agregar la media de gasto total para cada categoría de aloja
mediagastototal_por_aloja = df.groupby('aloja')['gastototal'].mean()

# Dibujar línea punteada para representar la media dentro de cada caja
for i, value in enumerate(mediagastototal_por_aloja):
    plt.plot([i-0.4, i+0.4], [value, value], color='red', linestyle='--', linewidth=2, label='Media' if i == 0

plt.legend()  # Mostrar leyenda
plt.grid(True)
plt.show()

```

MOTIVO DE VIAJE

In [ ]:

```
# Crear un diccionario para mapear los códigos numéricos a los nombres correspondientes
codigo_motivo = {
    1: 'Ocio/vacaciones',
    2: 'Negocios',
    3: 'Resto'
}

# Reemplazar los códigos numéricos en la columna 'motivo' con los nombres correspondientes
df['motivo'] = df['motivo'].map(codigo_motivo)

plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='motivo', y='gastototal', palette='viridis')
plt.title('Gasto Total por motivo del viaje')
plt.xlabel('Motivo del viaje')
plt.ylabel('Gasto Total')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Calcular y agregar la media de gasto total para cada categoría de motivo
mediagastototal_por_motivo = df.groupby('motivo')['gastototal'].mean()

# Dibujar línea punteada para representar la media dentro de cada caja
for i, value in enumerate(mediagastototal_por_motivo):
    plt.plot([i-0.4, i+0.4], [value, value], color='red', linestyle='--', linewidth=2, label='Media' if i == 0

plt.legend()  # Mostrar leyenda
plt.grid(True)
plt.show()

```

PAQUETE TURÍSTICO

In [39]:

```
# Crear un diccionario para mapear los códigos numéricos a los nombres correspondientes
codigo_A16 = {
    1: 'Paquete turístico',
    6: 'No paquete turístico'
}

# Reemplazar los códigos numéricos en la columna 'A16' con los nombres correspondientes
df['A16'] = df['A16'].map(codigo_A16)

plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='A16', y='gastototal', palette='viridis')
plt.title('Gasto Total por condición de paquete turístico')
plt.xlabel('Paquete turístico')
plt.ylabel('Gasto Total')
plt.xticks(rotation=45)  # Rotar etiquetas del eje x para mejor visualización

# Calcular y agregar la media de gasto total para cada categoría de A16
mediagastototal_por_A16 = df.groupby('A16')['gastototal'].mean()

# Dibujar línea punteada para representar la media dentro de cada caja
for i, value in enumerate(mediagastototal_por_A16):
    plt.plot([i-0.4, i+0.4], [value, value], color='red', linestyle='--', linewidth=2, label='Media' if i == 0

plt.legend()  # Mostrar leyenda
plt.grid(True)
plt.show()

```

IPC

In [ ]:
```
# Gráfico de dispersión con línea de tendencia para gastototal e IPC
sns.lmplot(x='IPC', y='gastototal', data=df, aspect=1.5, scatter_kws={'s':20, 'alpha':0.5}, line_kws={'color':'
plt.title('Gasto Total vs IPC')
plt.xlabel('IPC')
plt.ylabel('Gasto Total')
plt.grid(True)
plt.show()
```

EURIBOR

In [ ]:
```
# Gráfico de dispersión con línea de tendencia para gastototal y EURIBOR
sns.lmplot(x='EURIBOR', y='gastototal', data=df, aspect=1.5, scatter_kws={'s':20, 'alpha':0.5}, line_kws={'colo
plt.title('Gasto Total vs EURIBOR')
plt.xlabel('EURIBOR')
plt.ylabel('Gasto Total')
plt.grid(True)
plt.show()
```

PERNOCTACIONES

In [ ]:
```
# Gráfico de dispersión con línea de tendencia para gastototal y A13
sns.lmplot(x='A13', y='gastototal', data=df, aspect=1.5, scatter_kws={'s':20, 'alpha':0.5}, line_kws={'color':'
plt.title('Gasto Total vs A13 (Total de Pernoctaciones)')
plt.xlabel('Total de Pernoctaciones (A13)')
plt.ylabel('Gasto Total')
plt.grid(True)
plt.show()
```

DESEMPLEO

In [ ]:
```
# Gráfico de dispersión con línea de tendencia para gastototal e inflacion
sns.lmplot(x='desempleo', y='gastototal', data=df, aspect=1.5, scatter_kws={'s':20, 'alpha':0.5}, line_kws={'co
plt.title('Gasto Total vs Desempleo')
plt.xlabel('Desempleo')
plt.ylabel('Gasto Total')
plt.grid(True)
plt.show()
```

ALGORITMOS

In [ ]:
```
from sklearn.model_selection import KFold, cross_validate, cross_val_score
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import make_scorer, mean_squared_error, mean_absolute_error, r2_score
```

KNN

In [20]:

```
# Separar variables independientes y dependientes
X = df.drop('gastototal', axis=1)
y = df['gastototal']

# Las variables categóricas
categorical_features = ['A1', 'pais', 'aloja', 'motivo', 'A16']

# Las variables numéricas
numeric_features = ['A13', 'IPC', 'EURIBOR', 'desempleo']

# Preprocesamiento para variables numéricas y categóricas
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(drop='first'), categorical_features)
    ])

# Definir el rango de valores de K a probar
k_range = range(2, 61)

# Almacenar los resultados
mae_scores = []
rmse_scores = []
r2_scores = []

# Realizar la validación cruzada
kf = KFold(n_splits=5, shuffle=True, random_state=42)

for k in k_range:
    # Pipeline para KNN
    pipeline_knn = Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', KNeighborsRegressor(n_neighbors=k))
    ])

    # Predicciones usando validación cruzada
    y_pred = cross_val_predict(pipeline_knn, X, y, cv=kf)

    # Calcular métricas
    mae = mean_absolute_error(y, y_pred)
    rmse = np.sqrt(mean_squared_error(y, y_pred))
    r2 = r2_score(y, y_pred)

    mae_scores.append(mae)
    rmse_scores.append(rmse)
    r2_scores.append(r2)

# Encontrar el valor de K óptimo basado en MAE
optimal_k_mae = k_range[np.argmin(mae_scores)]
optimal_k_rmse = k_range[np.argmin(rmse_scores)]
optimal_k_r2 = k_range[np.argmax(r2_scores)]

print(f'El valor óptimo de K basado en MAE es: {optimal_k_mae}')
print(f'El valor óptimo de K basado en RMSE es: {optimal_k_rmse}')
print(f'El valor óptimo de K basado en R^2 es: {optimal_k_r2}')

# Graficar MAE vs K
plt.figure(figsize=(10, 6))
plt.plot(k_range, mae_scores, marker='o', linestyle='-')
plt.xlabel('Valor de K')
plt.ylabel('MAE Promedio')
plt.title('MAE Promedio vs. Valor de K')
plt.grid(True)
plt.show()

# Graficar RMSE vs K
plt.figure(figsize=(10, 6))
plt.plot(k_range, rmse_scores, marker='o', linestyle='-')
plt.xlabel('Valor de K')
plt.ylabel('RMSE Promedio')
plt.title('RMSE Promedio vs. Valor de K')
plt.grid(True)
plt.show()

# Graficar R^2 vs K
plt.figure(figsize=(10, 6))
plt.plot(k_range, r2_scores, marker='o', linestyle='-')
plt.xlabel('Valor de K')
plt.ylabel('R² Promedio')
plt.title('R² Promedio vs. Valor de K')
plt.grid(True)
plt.show()
```

RANDOM FOREST

In [ ]:

```
# Separar variables independientes y dependientes
X = df.drop('gastototal', axis=1)
y = df['gastototal']

# Las variables categóricas
categorical_features = ['A1', 'pais', 'aloja', 'motivo', 'A16']

# Las variables numéricas
numeric_features = ['A13', 'IPC', 'EURIBOR', 'desempleo']

# Preprocesamiento para variables numéricas y categóricas
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(drop='first'), categorical_features)
    ])

# Definir el rango de números de árboles a probar
num_trees_range = range(10, 210, 10)

# Almacenar los resultados
mae_scores = []
rmse_scores = []
r2_scores = []

# Definir los scorers para MAE, RMSE y R²
mae_scorer = make_scorer(mean_absolute_error, greater_is_better=False)
rmse_scorer = make_scorer(mean_squared_error, greater_is_better=False, squared=False)
r2_scorer = make_scorer(r2_score)

# Iterar sobre el rango de números de árboles
for n_trees in num_trees_range:
    # Pipeline para Random Forest
    pipeline_rf = Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', RandomForestRegressor(n_estimators=n_trees, random_state=42))
    ])

    # Calcular MAE de validación cruzada (convertimos a positivo porque make_scorer lo hace negativo)
    mae = -cross_val_score(pipeline_rf, X, y, cv=5, scoring=mae_scorer).mean()
    mae_scores.append(mae)

    # Calcular RMSE de validación cruzada (convertimos a positivo porque make_scorer lo hace negativo)
    rmse = -cross_val_score(pipeline_rf, X, y, cv=5, scoring=rmse_scorer).mean()
    rmse_scores.append(rmse)

    # Calcular R² de validación cruzada
    r2 = cross_val_score(pipeline_rf, X, y, cv=5, scoring=r2_scorer).mean()
    r2_scores.append(r2)

# Crear el gráfico para MAE
plt.figure(figsize=(10, 6))
plt.plot(num_trees_range, mae_scores, marker='o', linestyle='-')
plt.xlabel('Número de Árboles en el Bosque')
plt.ylabel('MAE de Validación Cruzada')
plt.title('Número de Árboles vs. MAE de Validación Cruzada')
plt.grid(True)
plt.show()

# Crear el gráfico para RMSE
plt.figure(figsize=(10, 6))
plt.plot(num_trees_range, rmse_scores, marker='o', linestyle='-')
plt.xlabel('Número de Árboles en el Bosque')
plt.ylabel('RMSE de Validación Cruzada')
plt.title('Número de Árboles vs. RMSE de Validación Cruzada')
plt.grid(True)

# Etiquetar cada punto con su valor en el gráfico de RMSE
for i, rmse in enumerate(rmse_scores):
    plt.text(num_trees_range[i], rmse, f'{num_trees_range[i]}', fontsize=9, ha='center', va='bottom')
plt.show()

# Crear el gráfico para R²
plt.figure(figsize=(10, 6))
plt.plot(num_trees_range, r2_scores, marker='o', linestyle='-')
plt.xlabel('Número de Árboles en el Bosque')
plt.ylabel('R² de Validación Cruzada')
plt.title('Número de Árboles vs. R² de Validación Cruzada')
plt.grid(True)

# Etiquetar cada punto con su valor en el gráfico de R²
for i, r2 in enumerate(r2_scores):
    plt.text(num_trees_range[i], r2, f'{num_trees_range[i]}', fontsize=9, ha='center', va='bottom')
plt.show()
```

In [2]:

```python
# Definir modelos con hiperparámetros óptimos
models = {
    'Linear Regression': Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', LinearRegression())
    ]),
    'KNN': Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', KNeighborsRegressor(n_neighbors=9))
    ]),
    'Random Forest': Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', RandomForestRegressor(n_estimators=40, random_state=42))
    ])
}

# Definir métricas de evaluación
scoring = {
    'MAE': 'neg_mean_absolute_error',
    'RMSE': make_scorer(lambda y_true, y_pred: np.sqrt(mean_squared_error(y_true, y_pred))),
    'R2': 'r2'
}

# Definir validación cruzada
kf = KFold(n_splits=5, shuffle=True, random_state=42)

# Evaluar cada modelo usando cross_validate
results = {}
predictions = {}
for model_name, model in models.items():
    cv_results = cross_validate(model, X, y, cv=kf, scoring=scoring, return_train_score=True)
    results[model_name] = {
        'MAE': -cv_results['test_MAE'].mean(),  # convertir de nuevo a positivo
        'RMSE': cv_results['test_RMSE'].mean(),  # ya es positivo por el make_scorer
        'R2': cv_results['test_R2'].mean()
    }
    # Obtener predicciones para graficar
    model.fit(X, y)  # Entrenar en todos los datos para obtener predicciones
    y_pred = model.predict(X)
    predictions[model_name] = y_pred

# Mostrar resultados
for model_name, scores in results.items():
    print(f"Modelo: {model_name}")
    print(f"  MAE : {scores['MAE']:.2f}")
    print(f"  RMSE: {scores['RMSE']:.2f}")
    print(f"  R²  : {scores['R2']:.2f}")
    print()

# Graficar valores predichos vs. reales para cada modelo
plt.figure(figsize=(16, 12))

# Regresión Lineal
plt.subplot(3, 1, 1)
plt.scatter(y, predictions['Linear Regression'], alpha=0.5)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.title('Valores Reales vs. Predichos - Regresión Lineal')
plt.xlabel('Gasto Real')
plt.ylabel('Gasto Predicho')

# KNN
plt.subplot(3, 1, 2)
plt.scatter(y, predictions['KNN'], alpha=0.5)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.title('Valores Reales vs. Predichos - KNN')
plt.xlabel('Gasto Real')
plt.ylabel('Gasto Predicho')

# Random Forest
plt.subplot(3, 1, 3)
plt.scatter(y, predictions['Random Forest'], alpha=0.5)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.title('Valores Reales vs. Predichos - Random Forest')
plt.xlabel('Gasto Real')
plt.ylabel('Gasto Predicho')

plt.tight_layout()
plt.show()
```

IMPORTANCIA VARIABLES

In [22]:

```
X = df.drop('gastototal', axis=1)
y = df['gastototal']

# Renombrar columnas
X.rename(columns={'A13': 'pernoctaciones', 'A16': 'paquete_turistico', 'A1': 'via_de_salida'}, inplace=True)

# Definir preprocesamiento
categorical_features = ['via_de_salida', 'pais', 'aloja', 'motivo', 'paquete_turistico']
numeric_features = ['pernoctaciones', 'IPC', 'EURIBOR', 'desempleo']

preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(drop='first'), categorical_features)
    ])

# Definir el modelo de Random Forest con el preprocesamiento
rf_pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=40, random_state=42))
])

# Validación cruzada para obtener el RMSE original
kf = KFold(n_splits=5, shuffle=True, random_state=42)
y_pred_original = cross_val_predict(rf_pipeline, X, y, cv=kf)
rmse_original = np.sqrt(mean_squared_error(y, y_pred_original))

# Calcular RMSE permutado para cada variable
permuted_rmse = {}

for feature in X.columns:
    X_permuted = X.copy()
    X_permuted[feature] = np.random.permutation(X_permuted[feature])

    y_pred_permuted = cross_val_predict(rf_pipeline, X_permuted, y, cv=kf)
    rmse_permuted = np.sqrt(mean_squared_error(y, y_pred_permuted))

    permuted_rmse[feature] = rmse_permuted

# Calcular la importancia relativa
importance = {feature: permuted_rmse[feature] - rmse_original for feature in permuted_rmse}

# Crear un DataFrame para mostrar la importancia de cada variable
importance_df = pd.DataFrame(list(importance.items()), columns=['Feature', 'Importance'])

```

```
# Filtrar las variables con importancia mayor a cero y quitar las variables no deseadas
features_to_remove = ['A0_7', 'mm_aaaa', 'A0', 'A0_1', 'ccaa', 'factoregatur']
importance_df = importance_df[(importance_df['Importance'] > 0) & (~importance_df['Feature'].isin(features_to_r
importance_df = importance_df.sort_values(by='Importance', ascending=False).reset_index(drop=True)

# Normalizar las importancias para que la suma sea 1
importance_df['Normalized Importance'] = importance_df['Importance'] / importance_df['Importance'].sum()

# Mostrar la tabla de importancia normalizada
print("Importancia de las Variables (Normalizada):")
print(importance_df)

# Crear un gráfico de barras de la importancia de las variables normalizadas
plt.figure(figsize=(10, 6))
plt.barh(importance_df['Feature'], importance_df['Normalized Importance'], color='skyblue')
plt.xlabel('Aumento del RMSE (Normalizado)')
plt.ylabel('Variables')
plt.title('Importancia de las Variables en Random Forest (Permutación)')
plt.gca().invert_yaxis()  # Invertir el eje y para mostrar la variable más importante arriba
plt.show()

```