Facultad de Ciencias Económicas y Empresariales
ICADE

TURISMO RURAL EN ESPAÑA: MODELOS MACHINE LEARNING Y OTRAS TÉCNICAS CUANTITATIVAS

Autor: Ángela Vallejo Mengod
Director: Luis Ángel Calvo

RESUMEN

En este TFG se analiza el comportamiento del turista rural en España utilizando técnicas de machine learning y minería de datos sobre una muestra de más de 140.000 turistas residentes según el marco temporal seleccionado (INE, s.f.). Los resultados han permitido identificar, por un lado, las características que mejor definen al turista rural, y por otro, los factores que predicen un mayor gasto medio diario entre estos. Se ha observado que variables como alto gasto total, el motivo del viaje (turismo de naturaleza), el tipo de transporte (vehículo particular), el uso de plataformas digitales para reservar alojamiento, la realización de actividades como el senderismo, la adquisición de paquetes turísticos y un nivel de estudios superior se asocian de manera significativa con el perfil de turista rural. Además, entre los turistas rurales, aquellos que están casados, viajan en grupos reducidos, y participan en actividades gastronómicas y deportivas presentan un mayor nivel de gasto. Estos hallazgos permiten mejorar la segmentación del mercado y ofrecen una base sólida para el diseño de estrategias personalizadas que fomenten un turismo rural más rentable y sostenible.

Palabras clave: turismo rural, sostenibilidad, gasto turístico, perfil turístico, machine learning, minería de datos.

ABSTRACT

This Final Degree Project analyzes the behavior of rural tourists in Spain using machine learning techniques and data mining on a sample of more than 140,000 domestic travelers from the sample selected (INE, s.f.). The results have made it possible to identify, on the one hand, the characteristics that best define the rural tourist, and on the other, the factors that predict higher average daily spending within this group. Variables such as higher total expenditure, travel motivation (nature tourism), mode of transport (private vehicle), use of digital platforms to book accommodation, participation in activities such as hiking, the acquisition of tour packages, and a higher level of education are significantly associated with the rural tourist profile. Furthermore, among rural tourists, those who are married, travel in small groups, and engage in gastronomic and sports-related activities tend to show higher spending levels. These findings enhance market segmentation and provide a solid foundation for the development of personalized strategies aimed at promoting a more profitable and sustainable rural tourism model.

Keywords: rural tourism, sustainability, tourism spending, tourist profile, machine learning, data mining.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1. Contexto

1.1.1. Importancia del sector turístico en España

El turismo es uno de los **sectores más relevantes para la economía** española, representando en 2023 un 12,3% del Producto Interior Bruto (PIB), con una cifra total de 184.002 millones de euros, según datos del Instituto Nacional de Estadística (INE, 2023). España, además, se posiciona como la segunda potencia mundial en la recepción de turistas internacionales, con aproximadamente 85 millones de visitantes en el mismo año (Statista, 2023). Este sector también tiene un **impacto directo en el empleo**, concentrando el 13,4% de la ocupación total del país (Ministerio de Industria, Comercio y Turismo, 2020).

En los últimos años, el turismo ha experimentado una **diversificación** hacia modalidades como el **turismo rural**, que en 2022 atrajo a 3,5 millones de turistas nacionales y más de 850.000 internacionales (Statista, 2023). Esta modalidad representó el 11,9% del gasto turístico total en 2023, consolidándose como una **herramienta clave para dinamizar las economías locales y afrontar retos globales** como la despoblación y la sostenibilidad (CaixaBank Research, 2023).

1.1.2. Problemática del despoblamiento rural

La despoblación es uno de los **grandes desafíos de las áreas rurales en España**. Según datos del Instituto Nacional de Estadística (INE), más de la mitad de los municipios españoles ha perdido población en las últimas décadas, afectando especialmente a comunidades autónomas como Castilla y León y Aragón. Este fenómeno no solo provoca el abandono de tierras y tradiciones, sino que también impacta negativamente en el tejido económico y social de las regiones afectadas.

El turismo rural se presenta como una herramienta clave para frenar este problema. Al promover actividades económicas sostenibles y diversificadas, se generan **nuevas oportunidades de empleo**, especialmente para jóvenes, mujeres y personas mayores. Además, fomenta la **revitalización de infraestructuras y servicios** como transporte, ocio y educación, fortaleciendo el bienestar comunitario y contribuyendo a fijar la población en estos territorios.

1.1.3. Turismo rural como herramienta del desarrollo sostenible

El turismo rural está estrechamente relacionado con el **desarrollo sostenible**, definido por (2015) como "*una forma de entender el mundo como la interacción compleja entre sistemas económicos, sociales, ambientales y políticos*". Esta nueva modalidad turística promueve un **uso equilibrado de los recursos naturales y culturales de las áreas rurales**, funcionando como una alternativa al turismo masivo. Según Loscertales (1999), además de dinamizar la economía local, impulsa la conservación del entorno al convertir el patrimonio en un activo económico. La actividad turística favorece la generación de empleos (como se demostró en la sección anterior de los datos de empleo), incorporando a jóvenes, mujeres y personas mayores en roles activos, lo que fortalece el tejido social y contribuye a fijar la población en territorios tradicionalmente afectados por el éxodo rural. Esto permite alcanzar un equilibrio entre el **crecimiento económico, la inclusión social y la protección ambiental.**

Sin embargo, como advierte Martín Gil (2014), pese a su potencial de generación de empleo, diversificación de la economía y revalorización de los recursos naturales y culturales, su expansión descontrolada puede generar conflictos con actividades tradicionales como la agricultura o la ganadería. Por ello, una adecuada planificación es fundamental para equilibrar el aprovechamiento económico y la conservación ambiental, evitando la degradación del entorno y la sobrecarga de los servicios locales.

1.2. Objetivos

1.2.1. Definición de las variables objetivo

Antes de presentar los objetivos específicos del estudio, introduciremos la definición de las variables objetivo utilizadas en cada modelo. Estas variables se consideran proxy, ya que no se encuentran de forma explícita e los datos originales, sino que han sido construidas a partir de variables existentes según el estudio de la literatura y criterios propios.

Por una parte, queremos clasificar al **turista en rural o no rural**, y para ello se ha definido la variable objetivo **"A_RURAL**", que toma valor 1 si el turista es rural y 0 si el turista es no rural. En este modelo se han procesado todos los registros que componen la base de datos seleccionada. Esta variable se ha determinado en función del valor de la variable **"ALOJAPRIN"** que indica el alojamiento principal utilizado en el viaje.

Cuando esta era igual a 5, que representa "Alojamiento turismo rural" (INE, 2017), "A_RURAL" se definía como 1 (rural) y cuando era igual a otro número, "A_RURAL" se definía como 0 (no rural).

La decisión de la manera en la que se ha definido la variable que identifica a un turista rural se justifica mediante la propia definición oficial del **Instituto Nacional de Estadística (INE)**, que describe el alojamiento rural como:

"*Establecimientos o viviendas destinados al alojamiento turístico mediante precio, con o sin otros servicios complementarios, y que estén inscritos en el correspondiente Registro de cada Comunidad Autónoma. Estos establecimientos suelen presentar unas características determinadas:*
a) Estar situados en medio rural.
b) Ser edificaciones con tipología arquitectónica propia de la zona o estar situados en fincas que mantengan activas explotaciones agropecuarias (agroturismo).
c) Ofrecer un límite de plazas y habitaciones para el alojamiento de huéspedes y reunir ciertos requisitos de infraestructura y dotaciones básicas." (INE, s.f.)

Esta definición refuerza la validez del criterio adoptado, ya que el tipo de alojamiento elegido implica una experiencia turística asociada de forma directa al entorno rural, su arquitectura, su entorno económico y sus servicios. Así, la construcción de la variable A_RURAL se encuentra alineada con los estándares oficiales del sector y permite una identificación adecuada del perfil de turista rural.

En segundo lugar, para predecir el **nivel de gasto diario del turista rural**, se ha definido la variable objetivo **"CLASIF2_GASTO",** que toma valor 0 si el nivel de gasto diario es bajo y 1 cuando el nivel de gasto diario es alto. Para este modelo, se ha aplicado un método de muestreo de los datos, seleccionando únicamente aquellas encuestas donde la variable A_RURAL = 1, es decir, el conjunto de datos se limita exclusivamente a los turistas rurales, contando en total con 3374 encuestas de turistas.

En primer lugar, se ha creado una nueva variable llamada **"GASTO_DIARIO"**, que representa el gasto medio diario por persona. Ha sido calculada como el gasto total del viaje dividido entre el número de pernoctaciones. No obstante, siguiendo la metodología propuesta por el Instituto Nacional de Estadística (INE) en su Guía para el tratamiento de los ficheros de microdatos de la Encuesta de Turismo de Residentes (ETR) (INE, 2017), el gasto diario por persona se ha calculado con la siguiente fórmula:

Figura 1: *Fómula para el Cálculo del Gasto Medio DIario por Persona*

$$GASTO\ MEDIO\ DIARIO\ POR\ PERSONA = \frac{suma\left(GASTOFI\_TOTAL * FACTORGAS\_TOT\right)}{suma\ (NPERNOC\_CORR * FACTORVI\_TOT)}$$

Fuente: INE (2017)

Donde:

- GASTOFI_TOTAL: Gasto total del viaje.
- FACTORGAS_TOT: Factor de elevación del gasto total.
- NPERNOC_CORR: Número de pernoctaciones corregido.
- FACTORVI_TOT: Factor de elevación de viajes.

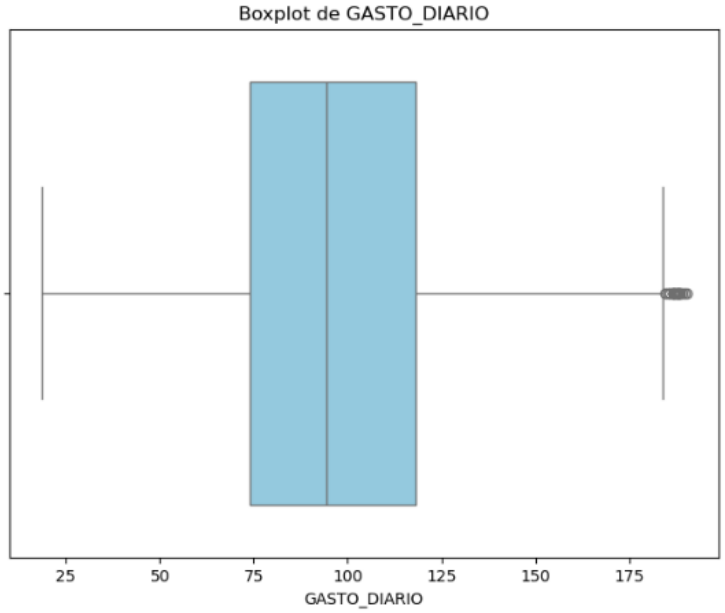
Posteriormente, se han eliminado los outliers de la nueva variable utilizando el método del rango intercuartílico (IQR). Seguidamente, hemos representado la distribución de los valores de la variable, que se puede encontrar en la Figura 2, lo que ha permitido determinar los límites de los niveles de gasto.

Tomando como referencia la mediana del gasto diario, que en este caso es de 96.17, se ha determinado que:

- Gasto diario < 96: Se clasifica como "bajo" (CLASIF2_GASTO = 0).
- Gasto diario ≥ 96: Se clasifica como "alto" (CLASIF2_GASTO = 1).

Este criterio garantiza que la clasificación del gasto diario del turista rural se base en un umbral objetivo, alineado con la metodología del INE y considerando la distribución real de los datos.

Figura 2: *Boxplot de la variable "GASTO_DIARIO" donde en el eje X muestra el gasto total del viaje. El primer y tercer cuartil se sitúan aproximadamente entre 74 y 119 euros, y la media es aproximadamente 97*



Fuente: Elaboración propia con python usando la librería pyplot

1.2.2. Objetivos específicos del estudio

Seguidamente, procedemos a definir las variables específicas de nuestro estudio. Para ello, previamente hemos realizado un análisis preliminar para comprender el panorama actual del turismo rural en España, su impacto en el territorio y las aplicaciones que se han realizado con la tecnología en este ámbito; es decir, un estudio de la literatura existente (Capítulo 1 – Introducción). Además de definir claramente qué estamos tratando de predecir.

Esta fase inicial es fundamental para identificar las oportunidades de investigación aún no exploradas y establecer una base sólida para el desarrollo del estudio.

A partir de este punto, se ha definido como objetivo principal de este TFG **desarrollar modelos de Machine Learning para mejorar la comprensión del turismo rural y optimizar la gestión de este segmento.** En concreto, los objetivos específicos que se plantean son los siguientes:

- Identificar el perfil del turista rural mediante un modelo de clasificación supervisado, que permita conocer qué variables lo caracterizan y entender su

influencia a través de técnicas de interpretación como SHAP (*Shapley Additive explanations*). (Capítulo 2 – Modelos de Machine Learning)

- Analizar los factores que afectan el nivel de gasto diario de los turistas rurales, mediante un modelo predictivo de clasificación, e interpretando el comportamiento de las variables mediante SHAP para obtener una comprensión más precisa del impacto de cada factor. (Capítulo 2 – Modelos de Machine Learning)
- Descubrir patrones y combinaciones de variables que caracterizan perfiles turísticos rurales, aplicando técnicas de minería datos con algoritmos reglas de asociación. (Capítulo 3 – A-Rules Data mining)

Todos estos objetivos permitirán una **comprensión más profunda del comportamiento del turista rural**, generando recomendaciones que contribuyan a mejorar la gestión de los alojamientos rurales y a fomentar su desarrollo sostenible.

1.3. Metodología

Para alcanzar los objetivos expuestos de este Trabajo de Fin de Grado, se ha utilizado una metodología específica, que se detallará a continuación. La metodología de definición de las variables ya ha sido comentada en el apartado 1.2 de Conclusiones.

1.3.1. Recopilación de datos

Los datos utilizados para este análisis provienen de la **Encuesta de Turismo de Residentes (ETR)** realizada por el Instituto Nacional de Estadística (INE) de España. Esta encuesta recoge información detallada sobre los viajes realizados por los residentes en España, tanto dentro del país como al extranjero. La ETR incluye un total de **117 variables** relacionadas con las características socioeconómicas de los viajeros, como edad, nivel de ingresos y situación laboral, así como datos específicos de los viajes, como el motivo del viaje, el tipo de alojamiento utilizado, el medio de transporte principal, gasto asociado al viaje, número de noches pernoctadas, el destino (a nivel de comunidad autónoma, provincia o país) o las actividades realizadas durante el mismo.

El periodo de datos para el análisis elegido comprende **desde enero de 2022 hasta junio de 2024,** coincidiendo con la fase de recuperación del turismo en España tras la pandemia de COVID-19. Según el informe de Perspectivas Turísticas de Exceltur, el año 2022 marcó la "total recuperación" de la actividad turística, alcanzando un PIB de 159.000

millones de euros, lo que representa un 1.4% más que en 2019, superando así las cifras previas al coronavirus. Además, Exceltur (2023) destaca que el 61% del crecimiento de la economía española en 2022 fue impulsado por el sector turístico. Este marco temporal refleja una etapa de normalización en los patrones turísticos, proporcionando una base sólida y representativa para el análisis de tendencias y comportamiento en el turismo. Además, eligiendo este marco temporal para nuestro estudio, recogemos un total de **142.843 encuestas** de turistas residentes en España.

1.3.2. Preprocesamiento de los datos

Como todo dataset, se requería un preprocesamiento de los datos para su posterior uso en la creación de los modelos predictivos. En esta fase hemos realizado las siguientes tareas:

- **Tratamiento de valores nulos:** se han identificado aquellas variables que contasen con un alto porcentaje de valores nulos (70%), y se ha procedido a descartarlas. En cuanto al resto de variales que contaban con valores nulos, se ha realizado una imputación de estos con la media de la variable para las numéricas, y la moda de la variable para las categóricas. Cabe mencionar que otra estrategia posible para el tratamiento de valores nulos es la imputación basada en correlaciones con otras variables, como el uso de modelos de regresión. Sin embargo, esta técnica no ha sido empleada en este estudio, ya que podría alterar la distribución original de los datos y afectar la validez de pruebas clave como Mutual Information y los métodos de selección de características.
- **Codificación de variables categóricas**: se han transformado aquellas variables categóricas no binarias (más de dos clases) en representaciones numéricas para permitir su procesamiento. Para ello, se ha utilizado la función *OneHotEncoder*, que permite que cada categoría única de una variable se codifique como una columna separada con valores 0 o 1, indicando la presencia o ausencia de esa categoría. Según Zhu et al. (2020), el uso de representaciones codificadas como una sola categoría asegura que las relaciones entre las clases se mantengan claras y ayuda a mejorar la precisión del modelo.

1.3.3. División del dataset en train test

Para la creación de los modelos, se ha utilizado un ratio de 70% train y 30% test para dividir el dataset, usando el primero para entrenar al modelo y el segundo para evaluar su predicción.

1.3.4. Creación modelos machine learning

Se han desarrollado dos tipos de modelos:

- **Modelo de clasificación de turista rural y no rural**: se han utilizado los algoritmos de árbol de decisión, random forest y XGBoost. Además, se han aplicado varias técnicas de balanceo de clases como class_*weight*, *undersampling* y *oversampling*; ya que había muchos más registros de turistas no rurales que de turistas rurales.
- **Modelo de clasificación del nivel de gasto diario del turista rural:** se han utilizado a su vez los algoritmos de árbol de decisión, random forest y XGBoost. Sin embargo, en este caso no ha sido necesario utilizar técnicas de balanceo de clases ya que el número de registros de las clases de la variable objetivo era similar.

Además, en el segundo modelo, una vez seleccionado el que mejor resultados ofrecía, se ha procedido a hacer **optimización bayesiana de hiperparámetros**. Este es un enfoque basado en modelos probabilísticos para optimizar funciones costosas de evaluar, como el ajuste de hiperparámetros en modelos de aprendizaje automático. Emplea un criterio para seleccionar de manera estratégica los próximos puntos a evaluar. Esto permite un equilibrio eficiente entre exploración y explotación, reduciendo el número de evaluaciones necesarias para encontrar configuraciones óptimas (Snoek et al., 2012).

1.3.5. Creación modelos reglas de asociación

Además de los modelos supervisados, se ha implementado una fase de **minería de de datos** con reglas de asociación utilizando el **algoritmo *Apriori***, con el objetivo de identificar combinaciones frecuentes de características entre los turistas rurales. Este enfoque no supervisado ha permitido extraer reglas significativas que revelan patrones de comportamiento asociados al perfil del turista rural y a aquellos que realizan viajes de trabajo.

Se han creados dos modelos, uno para la variable de turista rural y uno para la variable de nivel de gasto diario del turista rural. Reciclando para cada uno de ellos las variables objetivo y las variables seleccionadas de cada uno de los modelos de machine learning correspondientes (clasificación turista rural y clasificación gasto medio por persona).

1.3.6. Evaluación de los modelos

Para ambos **modelos de clasificación de machine learning** se se han utilizado las siguientes métricas de evaluación:

- **Precisión (Precision):** Mide la proporción de predicciones positivas correctas sobre el total de predicciones positivas realizadas. Es clave cuando los falsos positivos tienen un impacto importante, ya que evalúa cuán confiables son las predicciones positivas del modelo (Powers, 2020).
- **Recall (Sensibilidad):** Indica la capacidad del modelo para identificar correctamente todas las instancias positivas. Es útil cuando los falsos negativos son críticos, aunque no mide la exactitud de esas predicciones (Powers, 2020).
- **F1-score:** Combina precisión y recall en un solo valor mediante su promedio armónico. Es especialmente útil en problemas con clases desbalanceadas, ya que equilibra ambas métricas para evitar sesgos en la clasificación (Sitarz, 2022).
- **Accuracy:** Evalúa la proporción de predicciones correctas sobre el total de instancias. Aunque es fácil de interpretar, puede ser engañosa en conjuntos de datos desbalanceados, por lo que en este estudio tiene menor peso en la selección del modelo.

Además, para los **modelos de reglas de asociación** se han utilizado las siguientes métricas:

- **Soporte:** Esta métrica se calcula como la proporción de observaciones que contienen tanto el antecedente como el consecuente, respecto al total de observaciones del conjunto de datos. Indica la frecuencia con la que aparece una regla en el dataset. (Kotsiantis & Kanellopoulos, 2006)
- **Confianza:** Es la proporción de observaciones que contienen el consecuente entre aquellas que contienen el antecedente. Mide la fiabilidad de la regla, es decir, la probabilidad de que ocurra el consecuente dado que ha ocurrido el antecedente. (Kotsiantis & Kanellopoulos, 2006)

- **Lift:** esta métrica sirve para evaluar la calidad de las reglas de asociación. Se calcula como el cociente entre la confianza y el soporte del consecuente. Esta medida permite interpretar la relación entre el antecedente y el consecuente de una regla: un valor mayor que 1 indica una correlación positiva, es decir, la presencia del antecedente aumenta la probabilidad del consecuente; un valor menor que 1 sugiere una correlación negativa, donde el antecedente reduce dicha probabilidad; y un valor igual a 1 implica que no existe una relación significativa entre ambos elementos (Hussein et al., 2015).

1.3.7. Herramientas y recursos utilizados

Para la creación de los modelos, se ha utilizado el lenguaje de programación **"Python**", y sus correspondientes librerías como scikit-learn, pandas, numpy, pyplot, shap o apriori. Además, el código se ha ejecutado en el entorno de **Jupyter Notebook.**

1.4. Antecedentes

El turismo ha evolucionado significativamente en las últimas décadas, adoptando **herramientas tecnológicas avanzadas** para analizar grandes volúmenes de datos. En este contexto, el Machine Learning se presenta como una solución eficaz para mejorar la gestión turística y optimizar la toma de decisiones. Es por ello por lo que, antes de aportar nuevas ideas y crear los nuevos modelos de predicción, es importante conocer de qué forma se ha implementado la tecnología y el machine learning con anterioridad en este sector, lo que se analizará en esta sección.

Los **algoritmos de machine learning tiene una gran versatilidad**, ya que, gracias a su enfoque de creación de un modelo a partir de datos históricos, nos permite entrenarlo en función de un objetivo específico, lo que facilita su implementación en diversas áreas/sectores (Hermitaño Castro, 2022). En esta introducción, mostraré ejemplos en tres ámbitos diferentes. En el ámbito de la educación, donde se hacen estudios sobre la percepción docente y estudiantil, análisis del rendimiento académico, identificación de factores asociados a la deserción escolar y el fomento del pensamiento computacional (Forero-Corba & Negre Bennasar, 2024). En el ámbito y financiero, con la mejora de la gestion del riesgo de crédito, evaluando el perfil de los solicitantes de crédito y previendo incumplimientos de pago (Hermitaño Castro, 2022). En el ámbito sanitario, tanto para realizar pronósticos cínicos, como para la optimización de la gestion de servicios de salud

como en la mejora de la asignación de recursos o la planificación de atención de pacientes (Pedrero, Reynaldos-Grandón, Ureta-Achurra & Cortez-Pinto, 2021).

Una vez identificadas diferentes aplicaciones del Machine Learning, nos centraremos en la aplicación en el tema de estudio de este trabajo: el turismo rural. La **aplicación de la inteligencia artificial (IA)** en el turismo rural ha demostrado ser una **herramienta eficaz para incrementar la visibilidad y las visitas a municipios menos conocidos** (Smart Travel News, 2023). El machine learning permite analizar grandes volúmenes de datos de manera autónoma, identificando patrones y realizando predicciones sobre el comportamiento turístico (Zaara, 2024).

Entre las implementaciones realizadas, cabe destacar la plataforma **"MyStreetBook"**, una inteligencia artificial para recomendación de rutas personalizadas. Tal y como se indica en su web, su plataforma ofrece un sistema integral disponible en aplicación móvil, web y herramientas inteligentes, diseñado para crear recorridos personalizados según las preferencias del usuario. Su tecnología permite planificar viajes detalladamente o improvisar rutas sobre la marcha, integrando elementos de turismo, cultura y gastronomía. Su objetivo es proporcionar una experiencia auténtica y fluida, haciendo que el visitante se sienta parte del entorno local. Además, se destaca por su flexibilidad y enfoque en la diversidad, creando itinerarios accesibles para todo tipo de públicos, resaltando la riqueza y particularidad de cada destino ("INICIO - MyStreetBook," 2024). La implementación de esta IA ha permitido redistribuir el flujo turístico, logrando que localidades con menor notoriedad incrementen sus visitas en un 36% cada una. Este enfoque no solo promueve el desarrollo económico de estas áreas, sino que también contribuye a una **distribución más equilibrada del turismo**, aliviando la presión sobre destinos saturados y fomentando prácticas más sostenibles. Es por esto por lo que la empresa ha sido reconocida por la Organización Mundial del Turismo por su contribución al turismo sostenible mediante el uso de tecnologías innovadoras. Su modelo ha alcanzado más de 200.000 usuarios en 2022 y se ha implementado en 180 municipios de España, evidenciando la escalabilidad y eficacia de estas soluciones tecnológicas en la promoción del turismo rural (Smart Travel News, 2023).

Por otra parte, debemos mencionar la aplicación de la IA para la creación **de "destinos turísticos inteligentes" (DTI)**. Son espacios innovadores que utilizan tecnología avanzada para **gestionar de manera eficiente el entorno, integrando factores ambientales, culturales y socioeconómicos**. Gracias a sistemas inteligentes que

recopilan y analizan datos en tiempo real, estos destinos mejoran la interacción de los visitantes con el entorno y optimizan la toma de decisiones de los gestores, contribuyendo a una experiencia turística de mayor Calidad (López de Ávila & García, n.d.). Buscan integrar tecnologías de vanguardia, como Big Data, nanotecnología y sensores, en los comportamientos de los consumidores para acompañarlos en todas las fases de su viaje, y con el ojetivo de ofrecerles experiencias únicas y competitivas que destaquen frente a otros destinos (Fernández Alcantud et al., 2017). Si estos destinos consiguen transformarse en "inteligentes", estará permitiendo la comunicación e interacción con los usuarios, facilitando el acceso a información sobre el destino antes del viaje, promoviendo la planificación y la investigación, y creando un ecosistema de aplicaciones y recursos conectados que enriquezcan la experiencia del visitante (Zaara, 2024). Un ejemplo de la aplicación de inteligencia artificial en el sector turístico son los **sistemas implementados en hoteles**, como aplicaciones personalizadas que facilitan procesos como el *check-in* y *check-out*, la recepción de información útil, la contratación de traslados y la evaluación de servicios. Asimismo, el **Internet de las Cosas (IoT)** permite la interconexión de dispositivos para ofrecer al viajero servicios adaptados, como habitaciones configuradas según sus preferencias o pulseras electrónicas que agilizan el acceso al alojamiento. Además, los robots han comenzado a desempeñar un papel destacado en tareas como la atención al cliente, mejorando la eficiencia y la experiencia del usuario (Zaara, 2024).

En el context de DTIs, también encontramos las **"smart villages",** una iniciativa que forma parte del enfoque más amplio de Áreas Rurales Inteligentes y Competitivas, desarrollado por la Red Europea de Desarrollo Rural (ENRD) entre 2017 y 2020 (Unión Europea, s.f.). Estos comparten el uso de tecnologías avanzadas y la participación comunitaria para mejorar la Calidad de vida y hacer más sostenibles los servicios. Además, siguen la misma premisa de que la digitalización y la gestion de datos pueden aplicarse para optimizar la planificación de servicios turísticos, prever flujos de visitantes y adaptar las actividades a las necesidades locales.

Aunque el Machine Learning ha demostrado ser eficaz en diferentes áreas del turismo, su aplicación específica en **la perfilación de viajeros que realizan turismo rural** no ha sido explotada. Este Trabajo de Fin de Grado busca llenar este vacío, proponiendo modelos que descubran patrones que permitan la mejora de gestión del turismo rural.

1.5. Estructura del TFG

El Trabajo de Fin de Grado se organiza en cinco capítulos principales, diseñados para guiar al lector desde la contextualización del problema hasta los resultados obtenidos y las conclusiones. La estructura se explica a continuación.

En primer lugar, el Capítulo 1, que engloba la **Introducción.** Este capítulo establece las bases del trabajo, e incluye los apartados presentados anteriormente. Entre estos encontramos: el contexto, donde se describe el panorama actual del turismo rural, el problema del despoblamiento rural y su relevancia al desarrollo sostenible; los objetivos del estudio, la metodología que se ha seguido para realizar el trabajo, los antecedentes donde se comentan referencias previas relevantes sobre el turismo rural y el uso del machine learning entre otros, y la estructura del tfg, explicada en este apartado.

En segundo lugar, el capítulo 2 explica los **modelos de machine learning** creados. Este capítulo se divide en tres partes: Introducción a los modelos de Machine Learning, el modelo de clasificación de turista rural y el modelo de clasificación de predicción de gasto. Cada una de las secciones de los modelos incluye: selección de variables, creación y evaluación de modelos, visualización del mejor modelo elegido, visualización del árbol de decisión y conclusión.

En tercer lugar, el Capítulo 3 está dedicado a la minería de datos mediante **reglas de asociación**, utilizando el algoritmo *Apriori* como técnica de minería de datos no supervisada. Este apartado tiene como objetivo descubrir combinaciones frecuentes de variables que permiten identificar patrones de comportamiento turístico, especialmente en el contexto del turista rural y del perfil de mayor gasto medio por persona

En cuarto lugar, se encuentra el capítulo de **conclusiones y recomendaciones generales**, donde se sintetizan los hallazgos principales, las implicaciones prácticas y las recomendaciones para la gestión del turismo rural.

Seguidamente, el capítulo de **futuras líneas de investigación**, que se hace referencia a qué estudios o técnicas futuras se podrían aplicar al análisis del turismo rural en España.

Y por último se encuentra la **Bibliografía** y la **Declaración de uso de ChatGPT.**

2. Modelos Machine Learning

El *Machine Learning* (ML) es un campo de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de la experiencia, sin ser programados explícitamente para realizar tareas específicas (Bishop, 2006).

En el contexto de este Trabajo de Fin de Grado, el Machine Learning se utiliza para abordar los objetivos específicos detallados en el capítulo anterior relacionados con el turismo rural: la clasificación de turistas como rurales o no rurales y la clasificación del nivel de gasto diario del turista rural. Ambos problemas corresponden a técnicas de aprendizaje supervisado, una subcategoría de ML en la que los modelos son entrenados con datos etiquetados para realizar predicciones sobre nuevos datos (Raschka & Mirjalili, 2020).

Los dos modelos creados permiten:

- **Clasificación de turistas rurales**: Identificar a los turistas que eligen alojamientos rurales, un segmento crucial para desarrollar estrategias de promoción específicas.
- **Clasificación del gasto diario del turista rural**: Analizar factores que influyen en el gasto diario de los turistas rurales, proporcionando información valiosa para optimizar las políticas de precios y mejorar la oferta turística.

2.1. Modelo de clasificación turista rural y no rural

2.1.1. Selección de variables

En primer lugar, se ha llevado a cabo un **análisis estadístico** de las variables, con el objetivo de identificar aquellas que presentaban diferencias significativas entre el grupo "rural" y "no rural".

Para las **variables numéricas**, primero se ha realizado la prueba de **distribución Normal**, para identificar qué método estadístico era procedente utilizar. Sin embargo, se ha descubierto que ninguna de las variables seguía una distribución normal. Por consiguiente, se ha procedido a realizar la **prueba de Wilcoxon-Mann-Whitney**, como alternativa a la t de Student. Esta prueba es una herramienta estadística no paramétrica utilizada para comparar dos muestras independientes, particularmente útil cuando no se cumplen los supuestos de normalidad en los datos. Nos sirve para evaluar si una de las

muestras tiende a generar valores más altos o bajos que la otra (Alberto Sánchez Turcios, 2015)

Y para las **variables categóricas**, se ha utilizado la **prueba del chi-cuadrado**. Esta prueba es una herramienta estadística no paramétrica utilizada para evaluar si existe una asociación significativa entre dos variables categóricas. Compara las frecuencias observadas en una tabla de contingencia con las frecuencias esperadas bajo la hipótesis nula de independencia entre las variables. Si las diferencias entre las frecuencias observadas y esperadas son lo suficientemente grandes, se rechaza la hipótesis nula, indicando una posible relación entre las variables analizadas (Cerda L & Villarroel Del P, 2007)

En segundo lugar, se ha utilizado el método de ***mutual information*** para identificar variables relevantes. Esta es una medida estadística que cuantifica la dependencia entre dos variables. Evalúa la cantidad de información que una variable proporciona sobre otra, considerando tanto relaciones lineales como no lineales. La MI es igual a cero cuando las variables son estadísticamente independientes, lo que significa que una no aporta información sobre la otra (Vergara & Estévez, 2015).

Utilizando las variables identificadas como más relevantes en los pasos anteriores, se ha procedido a la búsqueda de la mejor combinación de variables para optimizar la clasificación. Tras este proceso, se han encontrado las 8 variables que permitían clasificar mejor a los turistas entre rurales y no rurales. Estas variables se explican en la siguiente tabla:

Tabla 1: *Variables seleccionadas para la creación del modelo de clasificación de turista ruralTabla 1Variables seleccionadas para la creación del modelo de clasificación de turista rural*

VARIABLE	TIPO DE VARIABLE	DEFINICIÓN
MES	Numérica	Mes del año en e que se ha realizado el viaje.
MIEMV_15MAS	Numérica	Número de miembros del hogar mayores de 15 años que participaron en el viaje
RESERV_ALOJA_5.0	Categórica binaria	La reserva de alojamiento se ha realizado “A través de página web

		especializada (AirBnb, Homeaway, Homelidays, Niumba, Rentalia, Housetrip, Wimdu, Interhome, Friendly Rentals)" (valor 1) o por otro medio (valor 0).
GASTOFI_TOTAL	Numérica	Gasto total del viaje
TRANSPRIN_5	Categórica binaria	El transporte principal utilizado en el viaje fue "Automóvil u otros vehículos particulares propios o cedidos" (valor 1) u otro diferente (valor 0).
MOTIV_3	Categórica binaria	El motivo de realización del viaje era "Turismo de naturaleza" (valor 1) u otro (valor 0).
ACTI_SENDER_1.0	Categórica binaria	Se han realizado actividades de senderismo en el viaje (valor 1) o no (valor 0).
VIAJA_HIJOS_1.0	Categórica binaria	Indica si el viaje se ha realizado acompñado de hijos (valor 1) o no (valor 0).

La selección de estas variables ha sido crucial para el desarrollo del modelo, ya que nos permite identificar patrones significativos en el comportamiento de los turistas y su relación con el turismo rural.

2.1.2. Análisis estadístico de las variables seleccionadas

Con el objetivo de ofrecer un mejor entendimiento de las variables seleccionadas para el modelo, incluimos esta sección de **analisis estadistico de variables**. En ella podremos ver la **distribución de las variables**, comparando ambos grupos: turistas rurales y no rurales.

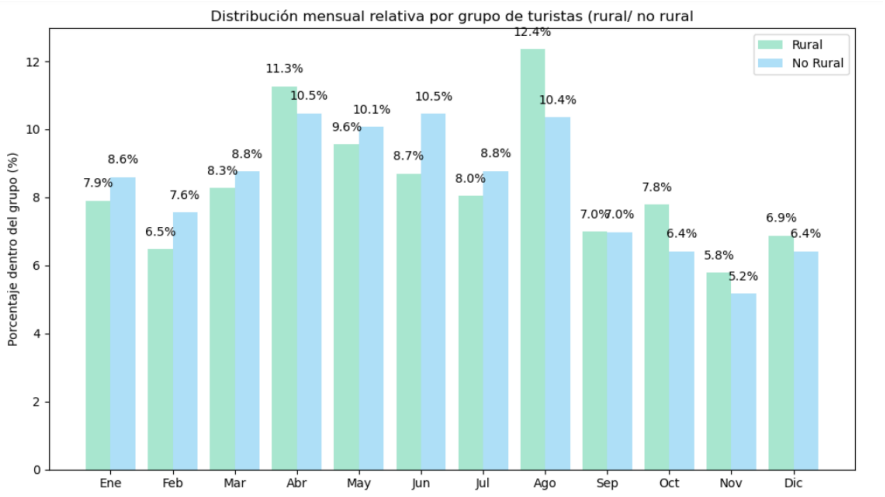
Como se ha explicado en el apartado anterior, para las variables numéricas se ha realizado la prueba de Wilcoxon y para las categóricas la de Chi-cuadrado, con el objetivo de

determinar si existía una diferencia significativa entre grupos respecto a la variable observada. Todas las seleccionadas mostraban un resultado positivo a este respecto.

Además, procedemos a realizar un análisis de la distribución entre grupos de estas variables, para visualizar cómo se comportan entre turistas rurales y no rurales. Empezamos con la variable que refleja el mes del año en el que se realiza el viaje, la cual representamos mediante un diagrama de barras. Seguidamente, para las variables numéricas representamos un gráfico de boxplot, y para las categóricas la tabla de contingencia.

La distribución de la variable **MES** la encontramos en la Figura 3. A lo largo del año, los turistas rurales muestran una mayor concentración en los meses de primavera y verano, destacando especialmente **agosto (12,4%) y abril (11,3%)** como los meses con mayor presencia relativa, manteniéndose este porcentaje el resto de meses por debajo de 10%. Los turistas no rurales presentan una distribución similar, con picos notables en primavera y verano, especialmente en **abril (10,5%), junio (10,5%) y agosto (10,4%)**. Esto sugiere que no hay una diferencia significativa entre ambos grupos.

Figura 3: *Distribución mensual del porcentaje de turistas según el grupo al que pertenecen (rural – barra verde, no rural – barra azul). El eje Y representa el porcentaje relativo de turistas dentro de cada grupo, y el eje X, el mes del año.*

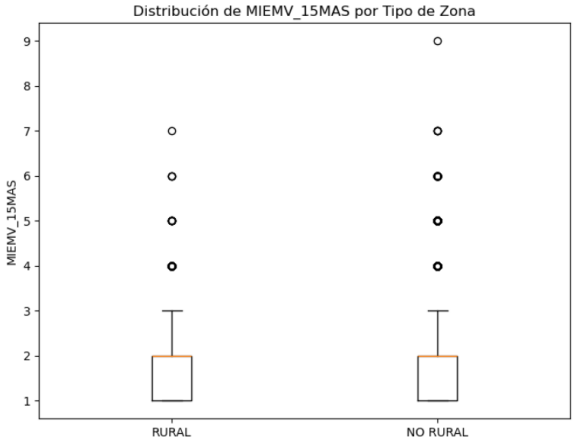


Fuente: Elaboración propia con la librería pyplot de python

La variable **MIEMV_15MAS**, también presenta una **distribución similar** en ambos grupos (Figura 4), además de **presencia de valores atípicos**. En ambos grupos, la mediana se encuentra en 2 miembros, con una distribución similar y presencia de valores atípicos, aunque en los turistas no rurales se observan algunos casos con hasta 9

miembros, indicando mayor variabilidad en este grupo. En este caso, la media en el grupo rural es 1.79 y en el grupo no rural 1.65.

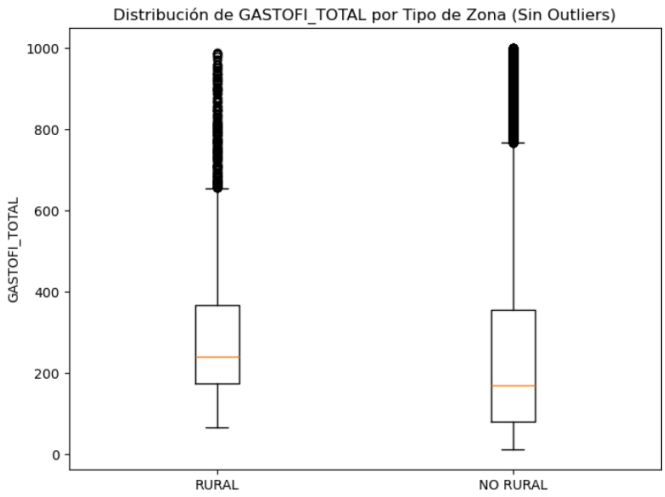
Figura 4: *Boxplot de la variable MIEMV_15MAS, donde el eje X muestra el grupo de turistas (rural y no rural) y el eje y el número de miembros del hogar mayores de 15 años que participaron en el viaje, ilustra la distribución de los datos de la variable en ambos*



Fuente: Elaboración propia con la librería pyplot de python

La variable **GASTOFI_TOTAL** muestra una distribución con dispersión considerable tanto en turistas rurales como no rurales. Para mejorar la visualización y reducir el impacto de valores atípicos extremos, se ha aplicado un límite superior de **1000 €** en ambos grupos. Aun con esta limitación, se observa que los turistas no rurales presentan una mayor variabilidad en el gasto, con valores mínimos más bajos y una mediana ligeramente inferior en comparación con los turistas rurales.

Figura 5: *Boxplot de la variable GASTOFI_TOTAL, donde el eje X muestra el grupo de turistas (rural y no rural) y el eje y el gasto total del viaje, ilustra la distribución de los datos de la variable en ambos grupos.*



Fuente: Elaboración propia con la librería pyplot de python

En cuanto a la variable **RESERV_ALOJA_5.0**, vemos, en la Tabla 2, cómo entre el grupo de turistas rurales, hay una mayor igualdad de proporciones de de las categorías de la variable que en el de no rurales. Entre los no rurales, la mayoría de los turistas no reservan el alojamiento a través de plataformas especializadas. Mientras que el 42% de los turistas rurales sí lo hacen.

Tabla 2: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable RESERV_ALOJA_5.0, e indica el porcentaje de turistas que presentan cada combinación de características*

	A_RURAL	
RESERV_ALOJA_5.0	**0**	**1**
0	95.32%	57.99%
1	4.68%	42.01%

Fuente: Elaboración propia con la librería pandas de python

Para la variable **TRANSPRIN_5**, observamos que, en el grupo de turistas rurales, la gran mayoría (90.70%) utiliza el automóvil o vehículos particulares como medio de transporte principal, mientras que, en los turistas no rurales, este porcentaje es menor (73.35%). En cambio, entre los no rurales, un 26.65% opta por otros medios de transporte, en comparación con solo un 9.30% en el grupo rural.

Tabla 3: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable TRANSPRIN_5, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
TRANSPRIN_5	**0**	**1**
0	26.65%	9.30%
1	73.35%	90.70%

Fuente: Elaboración propia con la librería pandas de python

En cuanto a la variable **MOTIV_3**, se aprecia una diferencia notable entre ambos grupos. Mientras que en los turistas no rurales el 94.79% no viaja por motivos de turismo de naturaleza, en el caso de los turistas rurales este porcentaje baja considerablemente al 62.49%. Por otro lado, el 37.51% de los turistas rurales sí declaran realizar su viaje por este motivo, en comparación con apenas el 5.21% de los turistas no rurales.

Tabla 4: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable MOTIV_3, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
MOTIV_3	0	1
0	94.79%	62.49%
1	5.21%	37.51%

Fuente: Elaboración propia con la librería pandas de python

Respecto a la variable **VIAJA_HIJOS_1.0**, se observa que un 35.28% de los turistas rurales viajan acompañados de sus hijos, lo que representa un porcentaje superior al 23.37% de los turistas no rurales. A su vez, la mayoría de los turistas en ambos grupos no viajan con hijos, aunque la proporción es ligeramente mayor en los no rurales (74.63%) en comparación con los rurales (64.72%).

Tabla 5: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable VIAJA_HIJOS_1.0, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
VIAJA_HIJOS_1.0	0	1
0	74.63	64.72%
1	23.37%	35.28%

Fuente: Elaboración propia con la librería pandas de python

2.1.3. Creación y selección de modelos

Tras la selección de las variables del modelo, explicaremos el proceso llevado a cabo para identificar el algoritmo más adecuado. Con el objetivo de encontrar el modelo que ofrezca la mejor capacidad de clasificación para los turistas, se han evaluado diversas alternativas, incluyendo **árbol de decisión, random forest y XGBoost.**

Antes de explorar las diferentes alternativas inspeccionadas, cabe destacar que la variable objetivo de nuestro dataset estaba muy desbalanceada: 3506 de 142843 turistas estaban clasificados como rurales; es decir, **2,49% del total.** Es por ello por lo que hemos procedido a ejecutar los modelos tanto aplicando **técnicas de balanceo de clases** como no, para investigar con qué técnica obteniamos mejores resultados. Se han probado las siguientes tres técnicas de balanceo:

- **Class_weight:** Esta técnica asigna diferentes pesos a las clases en función de la frecuencia de las muestras, aumentando el peso de las clases minoritarias para mitigar la desventaja causada por su baja representación. Esto mejora el rendimiento en la clasificación de la clase minoritaria sin necesidad de modificar los datos originales (Bakırarar & Elhan, 2023)
- **Undersampling:** Este enfoque reduce la cantidad de muestras de la clase mayoritaria seleccionando un subconjunto representativo, lo que puede simplificar el modelo y equilibrar la distribución, aunque existe el riesgo de perder información valiosa (Shelke et al., 2017)
- **Oversampling:** Consiste en incrementar las muestras de la clase minoritaria mediante duplicación o generación de datos sintéticos, como el uso de técnicas como SMOTE, para aumentar la representación de la clase y mejorar la precisión del modelo (Shelke et al., 2017)

Dado que no existe un único modelo que funcione mejor en todos los contextos, se han probado distintos algoritmos de clasificación, cada uno con características particulares:

- **Árbol de decisión:** es un algoritmo de clasificación que divide los datos en nodos de decisión, seleccionando en cada paso la variable que mejor separa las clases. Es fácil de interpretar, pero puede sobreajustarse si el dataset es ruidoso o complejo (Prajapati et al., 2019).
- **Random forest:** es un algoritmo de clasificación basado en la combinación de múltiples árboles de decisión, donde cada árbol se construye a partir de un subconjunto aleatorio de datos y variables. Esta estrategia reduce el sobreajuste y mejora la capacidad de generalización del modelo. Además, el método es robusto al ruido en los datos y permite evaluar la importancia de las variables de manera interna, lo que facilita la interpretación del modelo (Breiman, 2001).
- **XG Boost:** es un algoritmo de boosting optimizado para eficiencia y escalabilidad. Construye árboles de decisión secuenciales, corrigiendo errores previos y utilizando regularización para evitar sobreajuste. (Chen & Guestrin, 2016).

Teniendo en cuenta tanto las diferentes técnicas de desbalanceo y los distinto algoritmos de machine learning, hemos probado diferentes combinaciones que se muestran a continuación:

- **Árbol de decisión:**
    - Sin balancear (**"DT_unbalanced"**).
    - Balanceado utilizando la técnica de *class_weight* (**"DT_cw"**).
    - Balanceado mediante *undersampling* (**"DT_balanced_under"**).
    - Balanceado mediante *oversampling* (**"DT_balanced_over"**).
- **Random Forest:**
    - Sin balancear (**"RF_unbalanced"**).
    - Balanceado utilizando la técnica de *class_weight* (**"RF_cw"**).
    - Balanceado mediante *undersampling* (**"RF_balanced_under"**).
    - Balanceado mediante *oversampling* (**"RF_balanced_over"**).
- **XGBoost:**
    - Sin balancear (**"XGB_unbalanced"**).
    - Balanceado utilizando la técnica de *class_weight* (**"XGB_balanced"**).

En todos ellos, se ha utilizado la misma lista de variables seleccionadas mostradas en la sección anterior, y las mismas métricas de evaluación de modelos. A continuación, se puede observar una tabla resumen de los resultados de cada uno de ellos:

Figura 6: *Resultados comparativos de los modelos de clasificación de turista rural aplicados con distintas técnicas de balanceo de datos*

Classification Report for DT_unbalanced:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	41801
1	0.67	0.13	0.22	1052
accuracy			0.98	42853
macro avg	0.82	0.57	0.61	42853
weighted avg	0.97	0.98	0.97	42853

Classification Report for RF_unbalanced:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	41801
1	0.30	0.21	0.25	1052
accuracy			0.97	42853
macro avg	0.64	0.60	0.62	42853
weighted avg	0.96	0.97	0.97	42853

Classification Report for DT_cw:

	precision	recall	f1-score	support
0	0.99	0.86	0.92	41801
1	0.11	0.72	0.20	1052
accuracy			0.85	42853
macro avg	0.55	0.79	0.56	42853
weighted avg	0.97	0.85	0.90	42853

Classification Report for RF_cw:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	41801
1	0.31	0.21	0.25	1052
accuracy			0.97	42853
macro avg	0.64	0.60	0.62	42853
weighted avg	0.96	0.97	0.97	42853

Classification Report for DT_balanced_under:

	precision	recall	f1-score	support
0	0.99	0.85	0.92	41801
1	0.11	0.73	0.20	1052
accuracy			0.85	42853
macro avg	0.55	0.79	0.56	42853
weighted avg	0.97	0.85	0.90	42853

Classification Report for RF_balanced_under:

	precision	recall	f1-score	support
0	0.99	0.77	0.87	41801
1	0.08	0.77	0.14	1052
accuracy			0.77	42853
macro avg	0.54	0.77	0.50	42853
weighted avg	0.97	0.77	0.85	42853

Classification Report for DT_balanced_over:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	41801
1	0.23	0.24	0.23	1052
accuracy			0.96	42853
macro avg	0.60	0.61	0.61	42853
weighted avg	0.96	0.96	0.96	42853

Classification Report for RF_balanced_over:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	41801
1	0.30	0.23	0.26	1052
accuracy			0.97	42853
macro avg	0.64	0.61	0.62	42853
weighted avg	0.96	0.97	0.97	42853

Classification Report for XGB_balanced:

	precision	recall	f1-score	support
0	0.99	0.86	0.92	41801
1	0.12	0.73	0.20	1052
accuracy			0.86	42853
macro avg	0.55	0.80	0.56	42853
weighted avg	0.97	0.86	0.90	42853

Classification Report for XGB_unbalanced:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	41801
1	0.54	0.14	0.22	1052
accuracy			0.98	42853
macro avg	0.76	0.57	0.60	42853
weighted avg	0.97	0.98	0.97	42853

Fuente: Elaboración propia con python

Las métricas de evaluación son herramientas esenciales para analizar el desempeño de los modelos en tareas de clasificación, ofreciendo distintas perspectivas sobre la capacidad del modelo para realizar predicciones correctas. Hemos empleado las mencionadas en la metodología.

En concreto, para este estudio, nuestro indicador principal y determinante en la elección del modelo es el **F1-score de la clase 1 (turista rural)**, ya que nuestro objetivo principal es maximizar la capacidad del modelo para identificar correctamente a los turistas rurales mientras mantenemos un bajo número de falsos positivos en esta clase. Dado que la clase 1 es significativamente minoritaria en comparación con la clase 0, el F1-score proporciona una evaluación equilibrada entre la precisión y el recall, lo que es fundamental en este contexto.

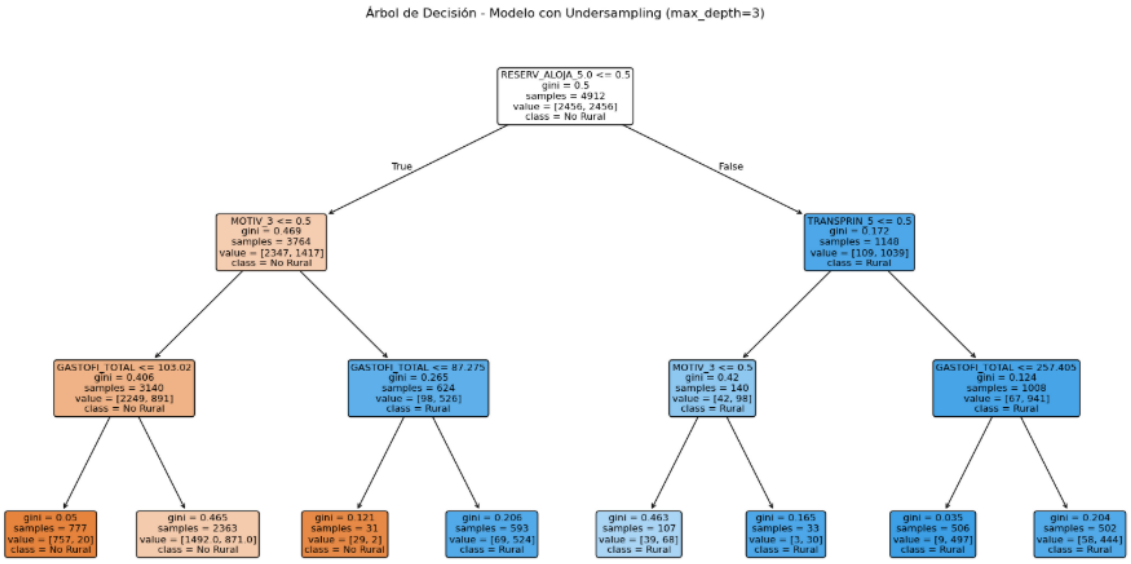
Es por ello por lo que, decidimos seleccionar el modelo **"XGB_balanced"**, es decir, el modelo XG Boost balanceado con class_weight para este estudio. Este modelo ofrece mejor equilibrio de resultados para nuestro trabajo: f1-score de la clase 1 igual a 0.20. Además, su *accuracy* total es de 0.86, lo que indica un buen desempeño general.

2.1.4. Visualización árbol de decisión

Nuestro modelo elegido como el mejor para nuestro estudio es de tipo XG Boost, pero este es poco interpretable en sus predicciones. Es por ello por lo que, a modo complementario, mostraremos inicialmente el **modelo de árbol de decisión** con las variables más importantes de nuestro modelo, para simplificar su visión y poder tener una idea más generalizada de cómo funciona la clasificación de un turista en rural y no rural.

Entre los diferentes algoritmos de árbol de decisión probados (desbalanceado, balanceado con class_weight, con undersampling y oversampling), observamos que el modelo de árbol de decisión balanceado con undersampling es el que mejor resultados nos ofrece. Este utilizaba las 8 variables originales seleccionadas y un valor para el hiperparámetro max_depth igual a 5. Sin embargo, dado que en esta sección pretendemos ofrecer una visión más clara y concisa de nuestro modelo explicativo, hemos decidido reducir la profundiad del árbol (mx_depth) a 3. Además de reducir el número de variables utilizadas por el modelo. Para esta selección, hemos tomado aquellas que presentaban un efecto más determinante en la Figura 7 de los valores shap. En concreto, nos quedamos con las siguientes variables: **GASTOFI_TOTAL, MOTIV_3, TRANSPRIN_5 y RESERV_ALOJA_5.0.** A continuación, podemos observar el gráfico en la Figura 7.

Figura 7: *Árbol interpretable con el número de variables reducido que refleja un modelo de clasificación realizado con Decision Tree y con técnica de balanceo de datos undersampling, longitud 4, accuracy 0.91.*



Fuente: Elaboración propia con la librería scikit-learn de python.

Como primera variable utilizada en el modelo de clasificación encontramos la variable "**RESERV_ALOJA_5.0",** como podemos observar en la Figura 11. El modelo la divide en si es menor o igual a 0.5 o mayor que 0.5; es decir, si la reserva del alojamiento se ha realizado "a través de página web especializada (AirBnb, Homeaway, Homelidays, Niumba, Rentalia, Housetrip, Wimdu, Interhome, Friendly Rentals)" es valor 1 y en caso contrario valor 0 ya que estamos ante una variable binaria. Este nodo empieza con 4912 turistas, donde hay un equilibrio perfecto entre rural y no rural, hay la misma proporción entre las clases. En el caso de que la variable para un turista sea igual a 0, lo procesará por la ruta de la izquierda, y en caso contrario, por la ruta derecha.

En el caso de la izquierda (RESERV_ALOJA_5.0 igual a 0), los turistas son clasificados mayoritariamente como No Rural, a excepción del caso de 593 turistas clasificados como Rural, que presentan las siguientes características: el motivo principal de su viaje era turismo deportivo (MOTIV_3 igual a 1) y el gasto total del viaje ha sido mayor que 87,275€. Además, observamos que, en esta ruta, no entra en acción la variable de transporte principal.

Y en el caso de la derecha (RESERV_ALOJA_5.0 igual a 1), todos los turistas son clasificados como Rural. Además, en esta ruta sí se tiene en cuenta la variable el transporte principal utilizado en el viaje.

Con este modelo obtenemos los siguientes resultados de las métricas de evaluación:

Figura 8: *Resultados de las métricas de evaluación del Árbol de decisión de clasificación de turista rural realizado con el número de variables reducidas y con técnica de balanceo de datos undersampling*

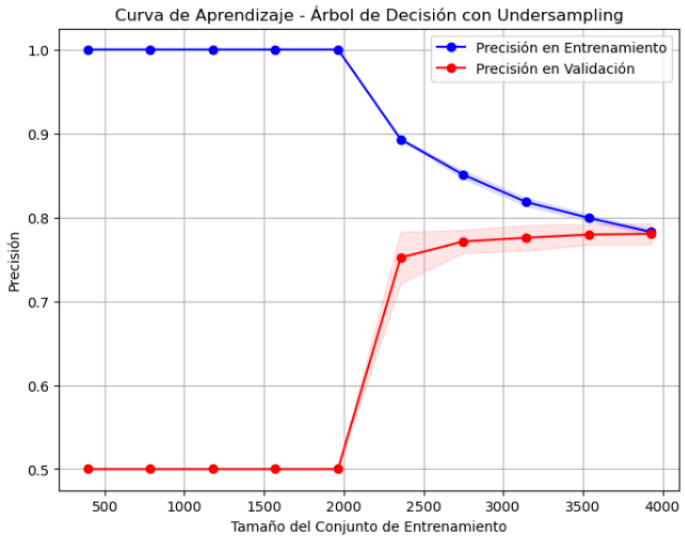
	precision	recall	f1-score	support
0	0.99	0.92	0.95	41803
1	0.16	0.61	0.25	1050
accuracy			0.91	42853
macro avg	0.57	0.77	0.60	42853
weighted avg	0.97	0.91	0.93	42853

Fuente: Elaboración propia con la librería pyplot de python

Además, hemos estudiado las curvas de aprendizaje de este modelo. Estas son una herramienta clave en el análisis del desempeño de los modelos de aprendizaje automático, ya que permiten visualizar cómo evoluciona la precisión o el error del modelo a medida que se incrementa el tamaño del conjunto de entrenamiento (Ospina-Gutiérrez & Aristizábal, 2021). Estas curvas muestran dos tendencias principales: una para los datos de entrenamiento y otra para los datos de validación.

En la Figura 9 se puede observar el gráfico de curvas de aprendizaje para nuestro modelo de árbol de decisión reducido. Dado que se observa una convergencia entre ambas curvas (precisión en entrenamiento y precisión en validación) conforme se añaden más datos, podemos afirmar que nuestro modelo está bien ajustado.

Figura 9: *Curvas de Aprendizaje que muestran la precisión en el set de entrenamiento y de validación del Árbol de Decisión con undersampling con el número de variables reducidas, elaborado con sklearn, librería de python.*



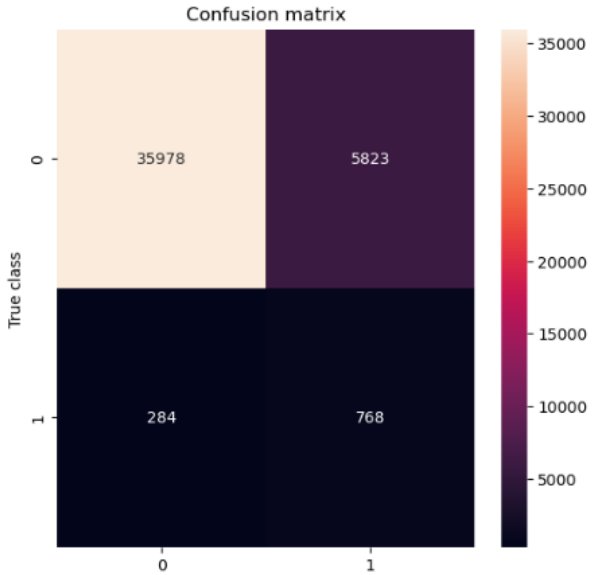
Fuente: Elaboración propia con las librerías numpy y pyplot de python

2.1.5. Visualización del mejor modelo elegido

Una vez hemos analizado un modelo sencillo e interpretable, y tenemos unas nociones básicas de los resultados nuestro análisis, procederemos a analizar el modelo que mejor resultados nos ofrecía: xgboost balanceado.

En primer lugar, debemos observar la matriz de confusión generada con este modelo. La matriz de confusión es una herramienta clave en la evaluación del desempeño de modelos de clasificación. Nos permite visualizar de manera directa las predicciones correctas e incorrectas realizadas por el modelo, separándolas en cuatro categorías: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Estas métricas son fundamentales para entender cómo el modelo maneja cada clase y para identificar posibles sesgos o deficiencias (IBM, 2024). En la figura 10, se muestra la matriz de confusión obtenida con el modelo "XGB_balanced", es decir, el modelo XG Boost balanceado con class_weight.

Figura 10: *Matriz de confusión y tabla de resultados de las métricas de evaluación del modelo "XGB_balanced" realizado con XGBClassifier y balanceo de datos con scale_pos_weight, longitud 6, accuracy 0.86*



	precision	recall	f1-score	support
0	0.99	0.86	0.92	41801
1	0.12	0.73	0.20	1052
accuracy			0.86	42853
macro avg	0.55	0.80	0.56	42853
weighted avg	0.97	0.86	0.90	42853

Fuente: Elaboración propia con la librería pyplot de python

En esta matriz se presenta el desempeño del modelo en términos de predicciones correctas e incorrectas para cada clase; y el correspondiente resultado de las diferentes métricas de evaluación explicadas anteriormente.

En cuanto a la clase 0 (no rural), 35.978 turistas fueron correctamente clasificados como clase 0, y 5.823 turistas fueron incorrectamente clasificados como clase 1 (rural). Y en cuanto la clase 1 (rural), 768 turistas fueron correctamente clasificados como clase 1(rural), y 284 turistas fueron incorrectamente clasificados como clase 0 (no rural).

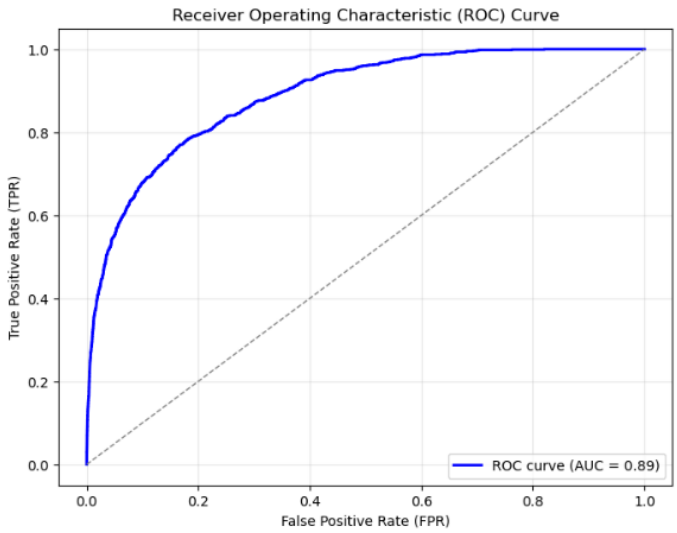
Estas clasificaciones nos permiten calcular las métricas de evaluación. Dado que nuestro objetivo principal es clasificar correctamente a los turistas rurales (clase 1), es importante destacar los resultados para esta clase. En este caso, el modelo obtuvo un *recall* de 0.12 y un *f1-score* de 0.20 para la clase 1.

El *recall* mide la capacidad del modelo para identificar correctamente a los turistas rurales (clase 1) dentro de todas las instancias que realmente pertenecen a esta clase. Un valor de 0.12 indica que el modelo pudo reconocer correctamente el 12% de los turistas rurales.

El *f1-score*, por otro lado, es una métrica que combina el *recall* y la *precisión*, proporcionando una medida equilibrada del desempeño del modelo. Un valor de 0.20 refleja que, aunque el modelo tiene un rendimiento limitado para detectar a los turistas rurales, este también está influido por su baja precisión en la clasificación de esta clase.

Otra forma efectiva de evaluar el desempeño de un modelo de clasificación binaria es mediante el uso de la curva ROC (Receiver Operating Characteristic). Esta herramienta gráfica permite representar la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) a diferentes umbrales de decisión. La curva ROC facilita la interpretación del modelo al destacar su capacidad para distinguir entre las dos clases. Una métrica clave derivada de esta curva es el Área Bajo la Curva (AUC), la cual proporciona un resumen cuantitativo del rendimiento general del modelo. Un AUC de 0.5 indica un desempeño equivalente al azar, mientras que un valor de 1.0 representa un modelo con discriminación perfecta (Marzban, 2004). En la figura 11 se puede observar la ilustración la curva ROC y el cálulo del AUC de nuestro modelo de clasificación.

Figura 11: *Curva ROC del modelo "XGB_balanced" que evalúa el rendimiento de la claficiación del modelo*



Fuente: Elaboración propia con las librerías sklearn.metrics y pyplot de python

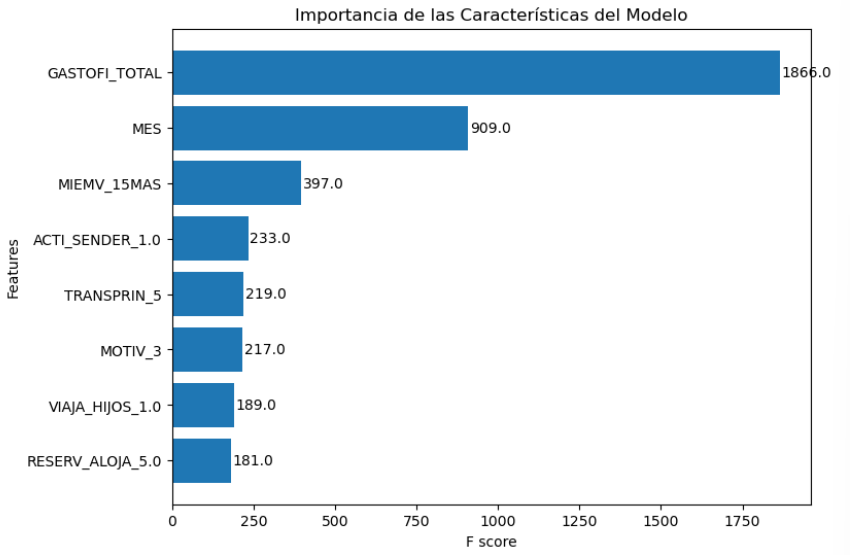
Podemos comentar que la curva ROC muestra un buen comportamiento, ya que está significativamente inclinada hacia la esquina superior izquierda, lo que indica que el modelo tiene un buen equilibrio entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Además, el AUC es de **0.89**, lo que representa un excelente desempeño general del modelo. Esto implica que, en promedio, el modelo tiene un 89% de probabilidad de clasificar correctamente una instancia positiva por encima de una negativa. Este es un indicador de que el modelo funciona bien, a pesar del desbalance de clases mencionado anteriormente.

Una vez evaluada la capacidad de clasificación del modelo, podemos inspeccionar cómo influyen las variables elegidas en la predicción.

En primer lugar, podemos observar la importancia de las variables del modelo a la hora de explicar la variable objetivo en la Figura 12 basada en el *F score*. Este gráfico muestra cuántas veces cada variable fue utilizada durante la construcción de los árboles de decisión en el modelo XGBoost, proporcionando una medida de su relevancia en las predicciones realizadas (Chen & Guestrin, 2016).

En este caso, la variable “GASTOFI_TOTAL" se destaca claramente como la más importante, con un *F score* de 1866.0, lo que indica que juega un papel crucial en la clasificación de los datos. Le sigue la variable "MES", con un *F score* de 909.0, sugiriendo que el momento temporal del viaje también tiene un impacto significativo la clasificación de turista rural. Otras variables, como "MIEMV_15MAS" y ACTI_SENDER_1.0", muestran un uso moderado, mientras que las restantes tienen un menor peso relativo.

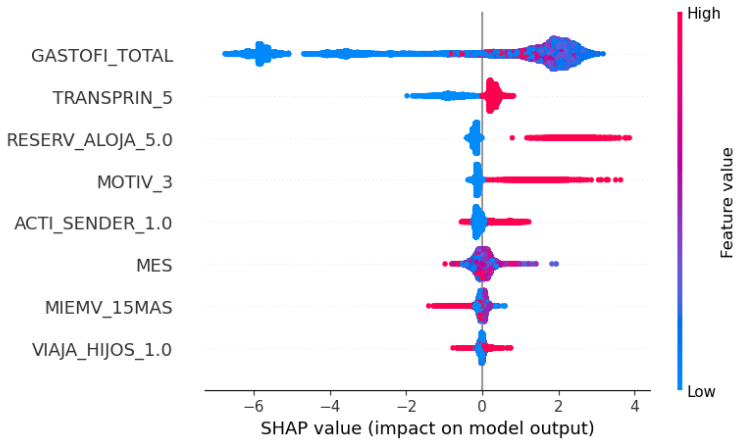
Figura 12: *Importancia de las variables del modelo "XGB_balanced" calculada según el F score, que mide la frecuencia con la que cada variable es utilizada en las divisiones del modelo para clasificar al turista rural*



Fuente: Elaboración propia con las librerías xgboost y pyplot de python

Además, dado que estamos trabajando con un modelo complejo (XG Boost) que tiene poca interpretabilidad, decidimos utilizar el método SHAP para entender cómo están influyendo las variables a la hora de hacer la clasificación. El método SHAP (*SHapley Additive exPlanations*) es una técnica de interpretación basada en la teoría de juegos que calcula la contribución de cada característica a las predicciones de un modelo, permitiendo tanto interpretaciones globales como locales de su funcionamiento (Ekanayake, Meddage, & Rathnayake, 2022). En la figura 13 se encuentra el gráfico de los valores SHAP de nuestro modelo, el cual muestra el impacto de cada variable en las predicciones del modelo, evaluando cómo los valores altos y bajos de cada variable influyen en el resultado del modelo.

Figura 13: *Gráfico de valores SHAP del modelo "XGB_balanced", que muestra la influencia de cada variable sobre las predicciones del modelo de clasificación de turista rural*



Fuente: Elaboración propia con la librería shap de python

Como se ha explicado, el gráfico de valores SHAP muestra el impacto individual de las características en las predicciones del modelo. A continuación, explicaremos la interpretabilidad de los resultados de cada variable:

- **GASTOFI_TOTAL**: Representa el gasto total del viaje y es una de las variables más relevantes. Se puede observar cómo el efecto de esta variable es muy extenso; sin embargo, no se puede distinguir bien si este es positivo o negativo, ya que los puntos rojos (altos) y azules (bajos) están distribuidos no uniformemente a lo largo del eje Y. Sí podemos decir que los valores altos de la variable están más presentes en el efecto positivo, lo que indica una mayor probabilidade de la clase 1 (rural).
- **TRANSPRIN_5**: Valores altos tienen un impacto positivo, lo que sugiere que el uso de vehículos particulares está relacionado con las predicciones del modelo, lo que aumenta la probabilidad de ser clasificado como rural.
- **RESERV_ALOJA_5.0**: Indica si la reserva de alojamiento se realizó a través de páginas web especializadas como Airbnb o similares (valor 1) o por otro medio (valor 0). Cuando los valores son altos (reserva realizada por estas plataformas) tienen un impacto positivo en la clasificación de turista rural, mientras que valores bajos reducen la probabilidad del resultado.
- **MOTIV_3:** Valores altos (el motivo del viaje es turismo de naturaleza) tienen un impacto positivo en el modelo, tiende a clasificar como rural, mientras que valores bajos (otros motivos) influyen negativamente.
- **ACTI_SENDER_1.0:** Aunque su impacto es menor, valores altos; es decir, cuando sí se han realizado actividades de senderismo, están asociados con un efecto positivo en las predicciones.

- **MES**: Representa el mes del año en el que se realizó el viaje. Su impacto es limitado, pero los meses más altos (valores en rojo) tienden a influir positivamente en las predicciones, mientras que los meses más bajos (valores en azul) tienen un impacto más neutral o ligeramente negativo.
- **MIEMV_15MAS**: Indica el número de miembros del hogar mayores de 15 años que participaron en el viaje. Su influencia es moderada, aunque se observan pocos valores con impacto significativo en las predicciones.
- **GASTOFI_TOTAL**: Representa el gasto total del viaje y es una de las variables más relevantes. Valores altos de gasto (en rojo) tienen un impacto negativo significativo en las predicciones, lo que indica que disminuye la probabilidad de ser clasificado como rural; mientras que valores bajos (en azul) contribuyen positivamente.
- **VIAJA_HIJOS_1.0:** Su impacto es reducido, y su rango de influencia en el modelo es más neutral; por lo que no podemos determinar su impacto específico en el modelo.

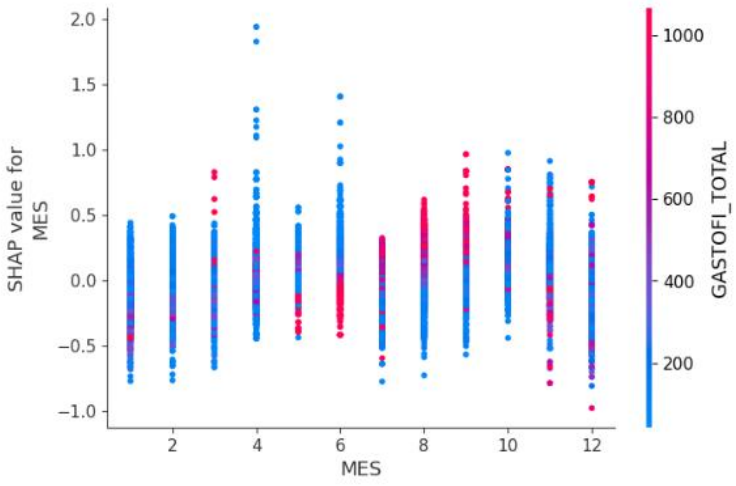
En resumen, las variables que más impacto tienen en el modelo son "GASTOFI_TOTAL" (gasto total del viaje), "RESERV_ALOJA_5.0" (reserva del alojamiento a través de plataformas especializadas) y "TRANSPRIN_5" (transporte principal vehículos de particulares). Mientras que las que menos impacto presentan son "MES" (mes del año que se realiza el viaje), "VIAJA_HIJOS_1.0" (si viaja con hijos o no) y "MIEMV_15MAS" (numéro de miembros del hogar que tienen mayores de 15 años" (INE, 2017)). Pese a su bajo impacto, consideramos que era relevante incluirlas en el modelo dado que estas variables podrían capturar patrones sutiles o interacciones con otras características, que, aunque no sean directamente determinantes, podrían aportar información complementaria para mejorar la precisión y generalización del modelo.

Para analizar con más detalle la influencia de estas variables, hemos realizado gráficos de dependencia para las variables mencionadas con menor impacto. Estos gráficos muestran cómo los valores individuales de cada variable afectan la predicción del modelo, identificando patrones claros entre los valores específicos de las características y sus contribuciones al resultado. Según Ekanayake, Meddage, y Rathnayake (2022), los gráficos de dependencia en SHAP no solo destacan el impacto promedio de una variable en las predicciones, sino también cómo interactúa con otras características del modelo. Esto permite comprender tanto los efectos individuales como las interacciones complejas

entre las variables, proporcionando una perspectiva detallada y práctica del comportamiento del modelo.

En la figura 14, se muestra el gráfico de dependencia de la variable “MES”. Este gráfico evidencia que su contribución es relevante en ciertos meses del año. En particular, hay meses como abril y junio en los que esta variable tiene un impacto positivo significativo en las predicciones del modelo; mayor posibilidad de clasificación como rural. Además, el impacto de MES se ve influenciado por el gasto total del viaje (GASTOFI_TOTAL), ya que los meses con gastos más altos presentan mayores contribuciones positivas, como julio, agosto y septiembre. Excepto junio, donde valores más altos de gasto total (puntos rojos) están relacionados con valores SHAP negativos. Esto nos indica que la interacción entre estas dos variables es relevante para el comportamiento del modelo; específicamente en meses con patrones de gasto elevados.

Figura 14: *Gráfico de dependencia SHAP que muestra el impacto de la variable “MES” en las predicciones del modelo de clasificación de turista rural, con la variable “GASTOFI_TOTAL” como valor interactivo*

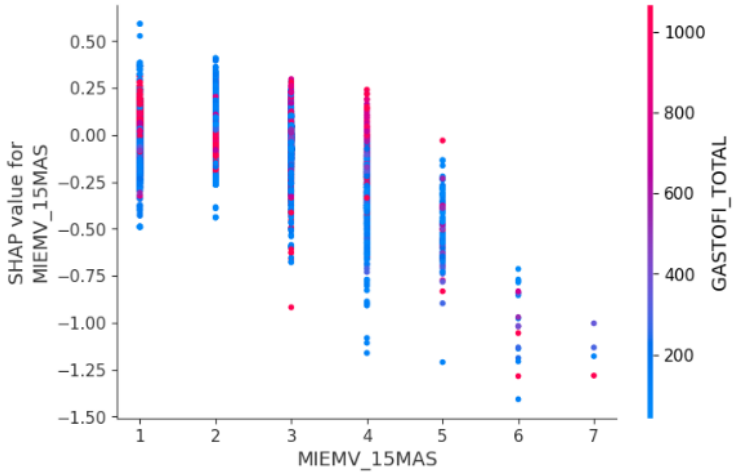


Fuente: Elaboración propia con la librería shap de python

En la figura 15, se genera el mismo gráfico para la variable “MIEMV_15MAS”, mostrándonos su dependencia con la variable “GASTOFI_TOTAL”. Por una parte, a medida que los valores de la variable; es decir, el número de miembros mayores de 15 años aumenta, los valores SHAP tienden a disminuir, lo que sugiere que afecta negativamente a la clasificación del turista como rural. En concreto, los hogares con 1 o 2 miembros mayores de 15 años, el impacto es generalmente positivo o neutro, mientras que, a partir de 3 miembros, se reduce la influencia positiva. Y por otra parte, encontramos

que en hogares con menos miembros mayores de 15 años, el gasto elevado tiene un mayor impacto positivo en las predicciones. Mientras que los gastos bajos (puntos azules) están más dispersos a lo largo de los diferentes tamaños de hogar, pero generalmente con valores más negativos; lo que sugiere que la combinación de un gasto bajo y más miembros reduce la influencia positiva en la clasificación.

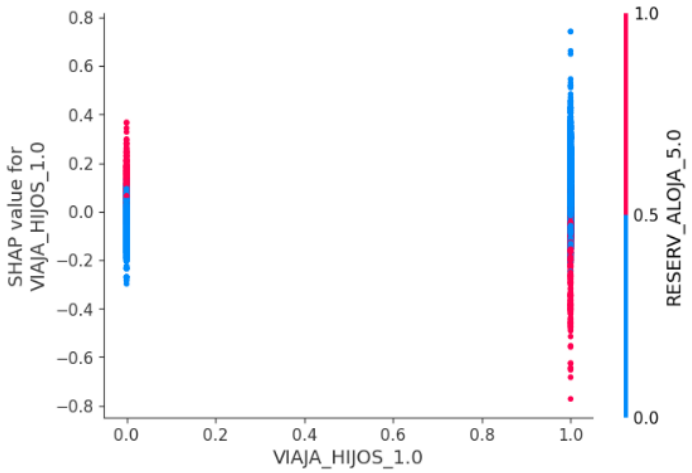
Figura 15: *Gráfico de dependencia SHAP que muestra el impacto de la variable "MIEMV_15MAS" en las predicciones del modelo de clasificación de turista rural, con la variable "GASTOFI_TOTAL" como valor interactivo*



Fuente: Elaboración propia con la librería shap de python

Por último, la figura 16 muestra el gráfico de dependencia de la variable "VIAJA_HIJOS_1.0"; donde la relaciona con la variable "RESERV_ALOJA_5.0". Por una parte, cuando la variable toma valor 0 (viaja sin hijos), los valores SHAP están alrededor de 0, lo que indica que tiene un impacto neutral o ligeramente negativo en las predicciones. Y cuando toma valor 1 (viaja con hijos), estos valores son más positivos. Esto sugiere que, si el viaje se ha realizado con hijos, tiende más a clasificar como rural. Por otra parte, en cuanto a la interacción entre variables, descubrimos que cuando "VIAJA_HIJOS_1.0" es 1 y "RESERV_ALOJA_5.0" es 0 (en azul), obtenemos valores SHAP más positivos. En otras palabras, el viajar con hijos y no realizar la reserva a través de plataformas especializadas tiene un impacto positivo más significativo en la clasificación como rural. Este análisis resalta la interacción clave entre estas dos variables en el comportamiento del modelo y su capacidad para predecir la clase objetivo.

Figura 16: *Gráfico de dependencia SHAP que muestra el impacto de la variable "VIAJA_HIJOS_1.0" en las predicciones del modelo de clasificación de turista rural, con la variable "GASTOFI_TOTAL" como valor interactivo*



Fuente: Elaboración propia con la librería shap de python

2.1.6. Conclusiones

El modelo de clasificación desarrollado ha permitido identificar **patrones relevantes que diferencian a los turistas rurales de los no rurales.** Para lograrlo, se han probado distintos algoritmos de Machine Learning, como Árbol de Decisión, Random Forest y XGBoost, con diferentes estrategias de balanceo de clases. Tras un análisis comparativo, se seleccionó XGBoost balanceado con class_weight ("XGB_balanced"), al ofrecer el mejor equilibrio entre precisión y recall, con un F1-score de la clase 1 (turista rural) de 0.20 y una precisión global (accuracy) del 86% (Figura 10). Además, el modelo alcanzó un AUC de 0.89, lo que indica un buen desempeño general del modelo.

En cuanto a la importancia de las variables, la figura 17 muestra que la variable "RESERV_ALOJA_5" (**reserva de alojamiento a través de plataformas digitales**) es la más influyente en la clasificación de turista rural, le sigue la variable "GASTOFI_TOTAL" (**gasto total del viaje**), después que el estado civil sea casado("ECIVIL_2"), que el **motivo del viaje sea deportivo** ("MOTIV_3") y el tipo de **transporte utilizado sea vehículo particular** ("TRANSPRIN_5"). El resto de las variables ("MIEMV_15MAS", "ACTI_SENDER_1.0", "VIAJA_HIJOS_1.0") tienen un efecto menor, pero siguen siendo influyentes en la clasificación mediante la relación con las otras, como explicamos con los gráficos de dependencia (Figura 14, Figura 15 y Figura 16).

Finalmente, como complemento para mejorar la interpretabilidad del modelo, se generó una representación gráfica de un árbol de decisión con las cuatro variables más influyentes mencionadas en el párrafo anterior (Figura 7). Este árbol permite visualizar

de manera más clara el proceso de clasificación, mostrando cómo las variables seleccionadas influyen en la predicción del turismo rural.

2.2. Modelo predictivo de gasto medio diario por persona

Tal y como se ha descubierto en el modelo de clasificación deturista rural (sección anterior), **el gasto total del viaje es un factor principal a la hora de realizar tal clasificación.** En concreto, aquellos turistas que presentan un gasto total del viaje mayor tienen más probabilidad de ser clasificados como rurales. Es por ello que, para satisfacer en mayor medida los objetivos de este trabajo, hemos decidido crear un segundo modelo de machine learning, que nos permita clasificar el nivel de gasto medio diario por turista entre aquellos que son clasificados como rurales. Con ello, pretendemos descubrir características clave que definan a los turistas que emplean una mayor o menor cantidad de dinero en sus viajes.

Como ya se ha explicado en la metodología, la variable a predecir de este modelo es "CLASIF2_GASTO", que refleja si el gasto medio diario del turista rural es alto (valor 1) o bajo (valor 0).

2.2.1. Selección de variables

A la hora de realizar la selección de variables, primero de todo se han descartado aquellas variables que reflejasen un gasto, por la alta correlación con la variable. Nuestro objetivo es explicar el nivel de gasto a través de otras variables diferentes que no estén relacionadas.

Seguidamente, se ha llevado a cabo un análisis estadístico de las variables, con el objetivo de identificar aquellas que presentaban diferencias significativas entre el grupo "gasto bajo" y "gasto alto".

Para ello, se ha seguido el mismo procedemiento explicado en el modelo anterior. Primero hemos realizado un análisis estadístico de las variables, utilizando la prueba de Wilcoxon-Mann-Whitney para las variables numéricas y la prueba del chi-cuadrado para las variables categóricas. Y segundo, se ha aplicado el método de *mutual information* para identificar variables relevantes.

Utilizando las variables identificadas como más relevantes en los pasos anteriores, se ha procedido a la búsqueda de la mejor combinación de variables para optimizar la clasificación. Tras este proceso, se han encontrado las 8 variables que permitían clasificar

mejor a los turistas entre rurales y no rurales. Estas variables se explican en la siguiente tabla:

Tabla 6: *Variables seleccionadas para la creación del modelo de clasificación de nivel de gasto medio diario del turista rural*

VARIABLE	**TIPO DE VARIABLE**	**DEFINICIÓN**
MIEMV	Numérica	Número de miembros del hogar mayores que participaron en el viaje
MOTIV_5	Categórica binaria	El motivo de realización del viaje era "Turismo deportivo" (valor 1) u otro (valor 0).
ACTI_GASTRO_1.0	Categórica binaria	Se han realizado actividades gastronómicas en el viaje (valor 1) o no (valor 0).
TRANSPRIN_5	Categórica binaria	El transporte principal utilizado en el viaje fue "Automóvil u otros vehículos particulares propios o cedidos" (valor 1) u otro diferente (valor 0).
ECIVIL_2	Categórica binaria	El estado civil del turista encuestado es "Casado/a" (valor 1) u otro (valor 0).

La selección de estas variables ha sido crucial para el desarrollo del modelo, ya que nos permite identificar patrones significativos en el comportamiento de los turistas y su relación con el turismo rural.

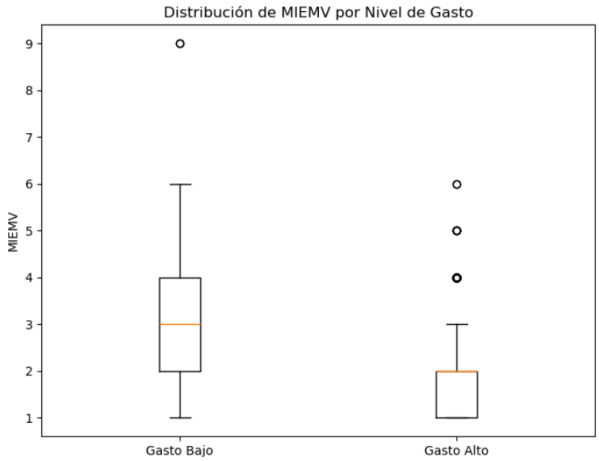
2.2.2. Análisis estadístico de las variables

De igual forma que hemos realizado en el primer modelo, analizaremos más en detalle las variables seleccionadas. Seguimos el mismo proceso y criterio de análisis.

La variable **MIEMV** muestra diferencias en la cantidad de miembros del hogar que participan en el viaje según el nivel de gasto. En el grupo de gasto bajo la mediana es más alta y la distribución más amplia, con valores atípicos que alcanzan hasta 9 miembros. En contraste, en el grupo de gasto alto, la mediana es menor y la mayoría de los valores se concentran entre 1 y 2 miembros, con menos valores atípicos. Además, la media de la

variable en el grupo gasto bajo es 2.75 y la del grupo gasto alto es 1.74. Esto sugiere que los viajes con menor gasto suelen realizarse en grupos más grandes, mientras que los de mayor gasto tienden a involucrar hogares más reducidos.

***Figura 17**: Boxplot de la variable MIEMV, donde el eje X muestra el grupo de turistas (rural y no rural) y el eje y el número de miembros del hogar que participaron en el viaje, ilustra la distribución de los datos de la variable en ambos grupos.*



Fuente: Elaboración propia con la librería pyplot de python

Para la variable **MOTIV_5**, se observa que la gran mayoría de los turistas, tanto rurales como no rurales, no realizan su viaje por motivos relacionados con esta categoría. Sin embargo, el porcentaje de quienes sí lo hacen es mayor en los turistas rurales (6.08%) en comparación con los no rurales (2.58%).

***Tabla 7:** Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable MOTIV_5, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
MOTIV_5	**0**	**1**
0	97.42%	93.92%
1	2.58%	6.08%

Fuente: Elaboración propia con la librería pandas de python

En cuanto a **ACTI_GASTRO_1.0**, se aprecia que la participación en actividades gastronómicas es más frecuente entre los turistas rurales (12.22%) en comparación con los no rurales (8.02%), aunque en ambos grupos la mayoría no realiza este tipo de actividades.

Tabla 8: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable ACTI_GASTRO_1.0, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
ACTI_GASTRO_1.0	**0**	**1**
0	91.98%	87.78%
1	8.02%	12.22%

Fuente: Elaboración propia con la librería pandas de python

Respecto a **TRANSPRIN_5**, se evidencia una diferencia en el uso del automóvil como transporte principal, siendo más frecuente en los turistas no rurales (95.59%) que en los rurales (87.91%). Esto sugiere que los turistas rurales pueden optar en mayor medida por otros medios de transporte.

Tabla 9: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable TRANSPRIN_5, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
TRANSPRIN_5	**0**	**1**
0	4.41%	12.09%
1	95.59%	87.91%

Fuente: Elaboración propia con la librería pandas de python

Finalmente, en **ECIVIL_2**, se nota que los turistas rurales tienen una mayor proporción de individuos pertenecientes a la categoría 0 (49.36%) en comparación con los no rurales (36.05%). Por otro lado, los turistas no rurales presentan un porcentaje más alto en la categoría 1 (63.95%) frente a los rurales (50.64%), lo que sugiere diferencias en la composición del estado civil entre ambos grupos.

Tabla 10: *Tabla de contingencia donde el eje X representa los valores de la variable A_RURAL y el eje Y los valores de la variable ECIVIL_2, e indica el porcentaje de turistas que presentan cada combinación de características.*

	A_RURAL	
ECIVIL_2	**0**	**1**
0	36.05%	49.36%
1	63.95%	50.64%

Fuente: Elaboración propia con la librería pandas de python

2.2.3. Creación y selección de modelos

En este caso, también hemos empleado diferentes algoritmos, para poder encontrar el que mejor se ajustase a nuestros objetivos. Hemos empleado tres tipos, entrenándolos con los mismos datos y la misma lista de variables:

- Árbol de decisión ("DT")
- Random Forest ("RF")
- XGBoost("XGB")

A continuación, se muestra una tabla resumen del resultado de cada uno de ellos.

Figura 18: *Resultados comparativos de los modelos de clasificación de nivel de gasto medio diario por persona del turista rural*

```
Classification Report for DT:
              precision    recall  f1-score   support

           0       0.75      0.62      0.68       524
           1       0.66      0.78      0.71       489

    accuracy                           0.70      1013
   macro avg       0.70      0.70      0.69      1013
weighted avg       0.70      0.70      0.69      1013

============================================================
Classification Report for RF:
              precision    recall  f1-score   support

           0       0.76      0.62      0.68       524
           1       0.66      0.79      0.72       489

    accuracy                           0.70      1013
   macro avg       0.71      0.70      0.70      1013
weighted avg       0.71      0.70      0.70      1013

============================================================
Classification Report for XGB:
              precision    recall  f1-score   support

           0       0.76      0.62      0.68       524
           1       0.66      0.79      0.72       489

    accuracy                           0.70      1013
   macro avg       0.71      0.70      0.70      1013
weighted avg       0.71      0.70      0.70      1013
```

Fuente: Elaboración propia

Se han calculado las mismas métricas de evaluación explciadas en el capítulo anterior: precisión, recall, accuracy y f1score. En este caso, nuestro objetivo es encontrar un modelo que prediga con la misma eficacia ambas clases. Observamos que los modelos "RF" y "XGB" nos dan los mejores resultados; por lo que decidimos seleccionar este último como modelo para nuestro estudio, ya que ofrece una complejidad mayor.

A continuación, realizamos un proceso de búsqueda de hiperparámetros a través del método de **optimización bayesiana**. Este método es una estrategia eficiente para

encontrar combinaciones óptimas de hiperparámetros en modelos de aprendizaje automático. Emplea un modelo probabilístico para estimar la relación entre los hiperparámtros y el rendimiento del modelo (Wu et al., 2019).

Los hiperparámetros optimizados y utilizados en el modelo son los siguientes, con el valor encontrado como el mejor predictor entre paréntesis:

- **colsample_bytree (0.5893):** Indica la fracción de características seleccionadas aleatoriamente para cada árbol (Wang & Ni, 2019).
- **gamma (0.1610):** Controla la reducción mínima de pérdida necesaria para hacer una nueva partición en un nodo, evitando divisiones innecesarias y ayudando a reducir la complejidad del modelo. Cuanto mayor, menos complejo es el modelo (Wang & Ni, 2019).
- **learning_rate (0.0903):** Define el tamaño del paso en cada actualización del modelo, regulando la velocidad de aprendizaje y permitiendo un ajuste más preciso de los parámetros (Wang & Ni, 2019).
- **max_depth (9):** Representa la profundidad máxima de los árboles, lo que afecta la capacidad del modelo para capturar patrones más detallados en los datos, pero también aumenta el riesgo de sobreajuste (Wang & Ni, 2019).
- **n_estimators (354):** Es el número total de árboles en el modelo que se construirán durante el entrenamiento del modelo., lo que influye en la capacidad del modelo para mejorar su precisión sin comprometer la eficiencia computacional, pero existe riesgo de sobreajuste (Brownlee, 2016).
- **subsample (0.5581):** Determina la fracción de muestras utilizadas para entrenar cada árbol, lo que introduce aleatoriedad y ayuda a reducir la varianza del modelo (Wang & Ni, 2019).

Estos valores permiten un equilibrio entre rendimiento y generalización, optimizando la capacidad predictiva del modelo sin aumentar el riesgo de sobreajuste.

2.2.4. Visualización árbol de decisión

Tal y como hemos presentado modelo de clasificación de turista rural, analizaremos primero el modelo de árbol de decisión, ya que nos ofrece un entendimiento más claro del modelo. Al contrario que en el primer modelo de clasificación, no reduciremos el

número de variables para esta visualización ya que ya contamos con número reducido (5 variables en total), donde todas ellas ofrecen información determinante para la clasificación. Pero sí reduciremos la profundidad del árbol de 5 a 4, para facilitar la visualización de este.

Al realizar este cambio, obtenemos resultados similares al árbol de decisión inicial presentado en el resumen de modelos de la Fugura 18, tan solo la accuracy disminuye 0.01. Estos resultados se pueden observar en la Figura 19.

Figura 19: *Resultados de las métricas de evaluación del Árbol de decisión de longitud 4 del nivel de gasto medio diario del turista rural*

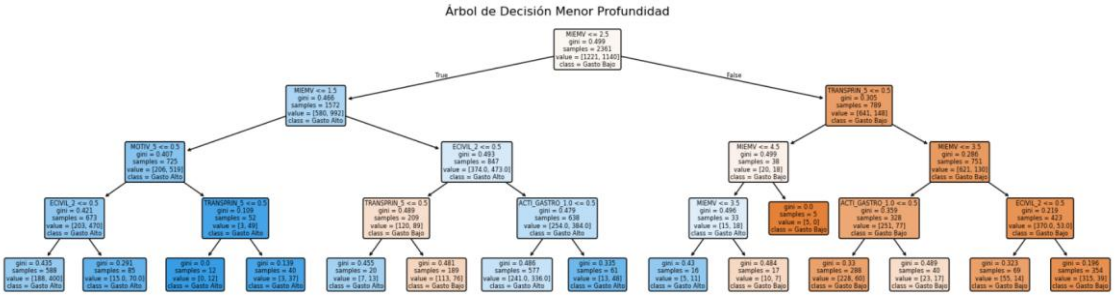
Predicted class

	precision	recall	f1-score	support
0	0.75	0.62	0.68	524
1	0.66	0.78	0.71	489
accuracy			0.69	1013
macro avg	0.70	0.70	0.69	1013
weighted avg	0.70	0.69	0.69	1013

Fuente: Elaboración propia

A continuación, analizaremos las ramas del árbol desde su representación gráfica en la Fgura 20.

Figura 20: *Árbol interpretable con menor profundidad (longitud 4) que refleja un modelo de clasificación realizado con Decision Tree del niverl de gasto medio diario del turista rural, accuracy 0.69, elaborado con scikit-learn*



Fuente: Elaboración propia

En el **nodo raíz** encontramos la variable MIEMV (Número de miembros del hogar mayores que participaron en el viaje), lo que indica que es la más relevante para el modelo. En este caso, marca el umbral como MIEMV ≤ 2.5, si esta condición se cumple, el modelo dirige los datos a la rama de la izquierda, y si no se cumple, los dirige a la rama de la derecha. Si nos fijamos en este nodo, observamos que el Índice gini es igual a 0.499,

lo que indica que las muestras en este nodo están bastante equilibradas entre gasto alto y gasto bajo. En concreto, 1221 turistas son clasificados como nivel de gasto bajo, y 1140 como nivel de gasto alto. Este nodo raíz establece la base para las siguientes divisiones en el árbol de decisión

En cuanto a la **ruta izquierda** del árbol de decisión corresponde a los turistas que cumplen la condición MIEMV ≤ 2.5, es decir, aquellos que viajaron con 2 o menos miembros del hogar. A partir de este punto, el árbol sigue dividiendo a los turistas en función de otras variables clave. En esta ruta se oberva cómo, la mayoría de los turistas que indiquen como número de miembros del hogar 2 o menos, son clasificados como "Gasto Alto"; sin embargo, encontramos algunas excepciones.

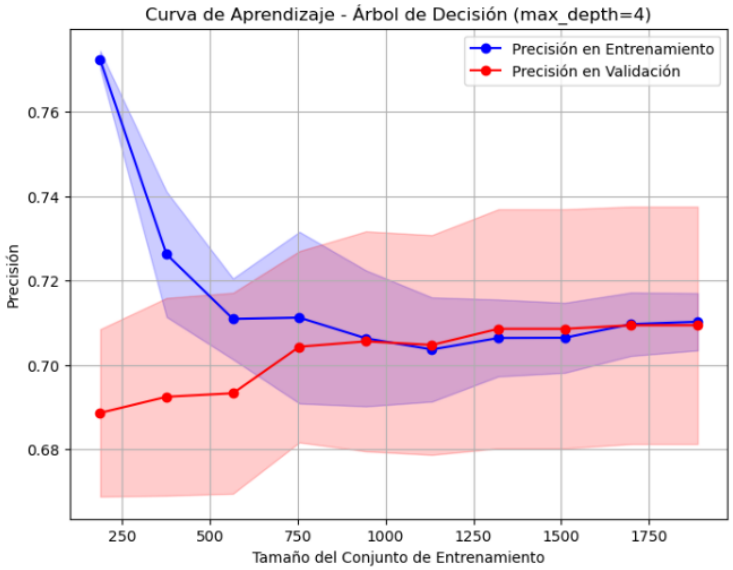
Por una parte, 209 turistas son clasificados como Gasto Bajo, y estos presentan las siguientes características: el número de miembros del hogar es 2 (mayor que 1.5) y el estado civil del turista es distinto de "Casado/a" (ECIVIL_2 igual a 0.

Y, por otra parte, hay 189 turistas también clasificados como Gasto Bajo que presentan esas mismas características, además con la característica de que el transporte principal utilizado en el viaje fue "Automóvil u otros vehículos particulares propios o cedidos" (TRANSPRIN_5.0 igual a 1).

En cuanto a la **ruta derecha** del árbol, cuando el número de miemberos del hogar es mayor o igual a 3, todos los turistas son clasificados como "Gasto Bajo", a excepción de dos nodos de 33 y 16 turistas. Pese a que es un número muy reducido, cabe destacar el valor de las variables que lo determinan. En el primer nodo destacado, el transporte principal utilizado en el viaje fue distinto de "Automóvil u otros vehículos particulares propios o cedidos" (TRANSPRIN_5 igual a 0) y el número de miembros del hogar que participaron es menor o igual que 4 (MIEMV <=4.5). Y en el segundo, la variable TRANSPRIN_5 toma el mismo valor, y la variable MIEMV es menor o igual que 3.5, es decir, el número de miembros es menor o igual que 3.

De forma complementarua, y tal y como hemos hecho en el anterior modelo de clasificación, mostraremos las curvas de aprendizaje. En la Figura 21 se puede observar la representación gráfica de las curvas de aprendizaje del modelo de árbol de decisión con profundidad 4. Podemos concluir que el modelo es adecuado ya que las líneas de precisión en Entrenamiento y en Validación convergen.

Figura 21: *Curvas de Aprendizaje que muestran la precisión en el set de entrenamiento y de validación del Árbol de decisión de longitud 4 del nivel de gasto medio diario del turista rural, elaboradas con sklearn, librería de python.*



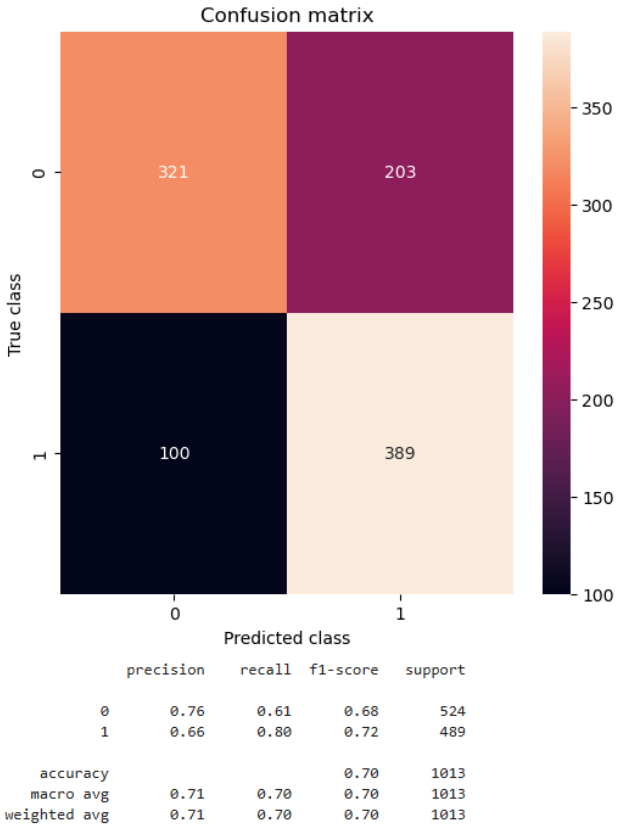
Fuente: Elaboración propia

2.2.5. Visualización del mejor modelo elegido

En esta sección, inspeccionaremos con más profundidad el resultado de nuestro modelo elegido por ofrecer los mejores resulrados: XGB con optimización bayesiana.

En primer lugar, visualizamos la matriz de confusión creada con las predicciones del modelo.

Figura 22: *Matriz de confusión y tabla de resultados de las métricas de evaluación del modelo realizado con XGBClassifier y con optimización bayesiana que clasifica el nivel de gasto medio diario del turista rural, longitud 9.*



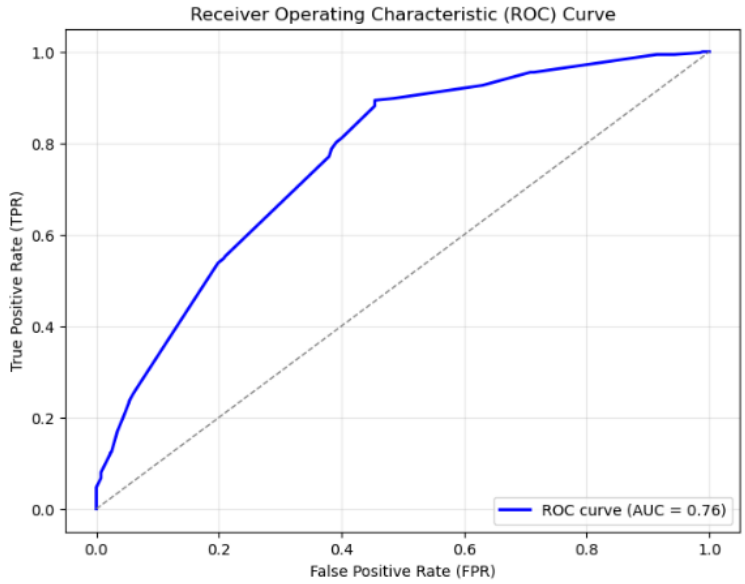
	precision	recall	f1-score	support
0	0.76	0.61	0.68	524
1	0.66	0.80	0.72	489
accuracy			0.70	1013
macro avg	0.71	0.70	0.70	1013
weighted avg	0.71	0.70	0.70	1013

Fuente: Elaboración propia

El modelo de clasificación presenta un desempeño con una precisión global del 71% en la predicción de las clases. A partir de la matriz de confusión, se observa que la clase 0 tiene 321 verdaderos negativos y 203 falsos positivos, mientras que la clase 1 tiene 389 verdaderos positivos y 100 falsos negativos. En términos de métricas, la clase 0 tiene una precisión de 76% y una sensibilidad (recall) de 61%, mientras que la clase 1 muestra una precisión de 66% y una sensibilidad de 80%. Esto indica que el modelo es ligeramente mejor identificando la clase 1 que la clase 0 en términos de sensibilidad. El F1-score para la clase 0 es 0.68 y para la clase 1 0.72, lo que sugiere un equilibrio aceptable entre precisión y recall.

Otra métrica de evaluación del modelo es la curva ROC y el valor AUC. La curva de nuestro modelo se encuentra por encima de la línea diagonal de referencia (que representa una clasificación aleatoria), lo que indica que el modelo tiene capacidad predictiva. El área bajo la curva (AUC) es de 0.76, lo que sugiere un desempeño moderado del modelo. En términos generales, un AUC de 0.76 indica que el modelo tiene una capacidad decente para distinguir entre las clases positivas y negativas.

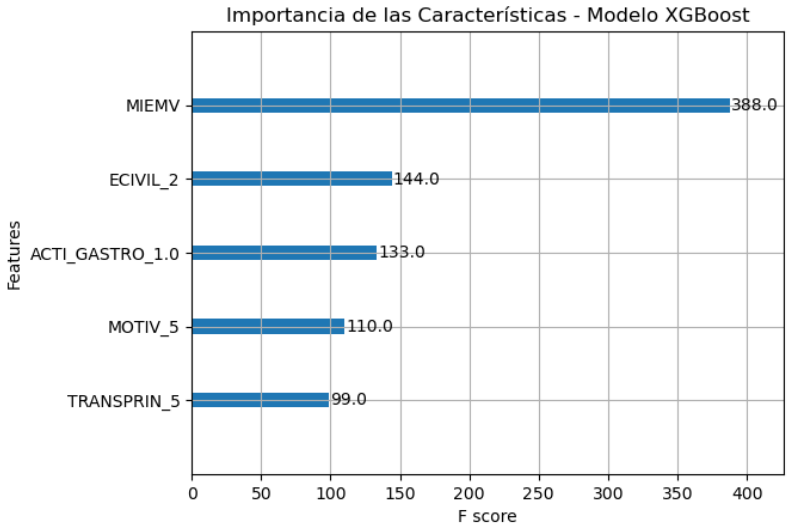
Figura 23: *Curva ROC del modelo XGB que evalúa el rendimiento de la clasificación de nivel de gasto medio diario del turista rural, elobrada con sklearn, librería de python.*



Fuente: Elaboración propia

Por otra parte, es interesante entender qué papel juegan las variables utilizadas a la hora de predecir el nivel de gasto. Para ello, se investiga la importancia de las variables en el modelo, que podemos ver en la figura 24.

Figura 24: *Importancia de las variables del modelo "XGB" calculada según el F score, que mide la frecuencia con la que cada variable es utilizada en las divisiones del modelo, elaborada con la librería de python xgboost*

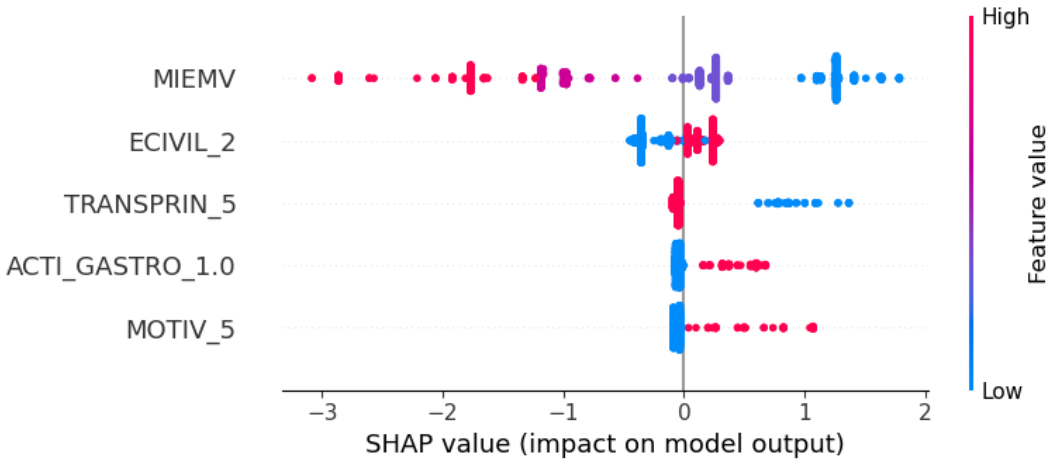


Fuente: Elaboración propia

El gráfico muestra la importancia de las características en un modelo XGBoost, evaluada mediante el F-score. Se observa que la variable MIEMV tiene la mayor importancia, con un valor de 388, lo que sugiere que es el predictor más influyente en el modelo. A una distancia considerable le siguen ECIVIL_2 (144) y ACTI_GASTRO_1.0 (133), lo que indica que estas variables también tienen un impacto relevante, aunque menor. Finalmente, MOTIV_5 (110) y TRANSPRIN_5 (99) presentan la menor importancia relativa dentro de las características destacadas. Estos resultados sugieren que el modelo XGBoost otorga un peso significativamente mayor a MIEMV en comparación con el resto de las variables

Para un mayor entendimiento de cómo predice el modelo utilizando las variables seleccionadas, se ha utilizado el método SHAP nuevamente. En la figura 25 se pueden ver el resultado de los valores SHAP de cada variable a la hora de predecir la clase 1 (nivel de gasto alto).

Figura 25: *Gráfico de valores SHAP del modelo XGB, que muestra la influencia de cada variable sobre las predicciones del modelo de clasificación del niverl de gasto medio diario del turista rural, elobrado con la librería shap de Python.*



Fuente: Elaboración propia

Gracias a este gráfico, podemos entender el efecto de los valores de cada variable del modelo:

- **MIEMV (Número de miembros del hogar mayores que participaron en el viaje)**
    - Valores altos (rojo): Disminuyen la probabilidad de un gasto alto.
    - Valores bajos (azul): Se asocian con mayor gasto.

    - A mayor número de miembros del hogar que participaron en el viaje, menor es la probabilidad de un gasto alto. Esto puede indicar que los viajes en grupo familiar tienden a ser más moderados en términos de gasto individual.
- **MOTIV_5 (El motivo del viaje fue "Turismo deportivo")**
    - Valores altos (rojo, turismo deportivo): Aumentan la probabilidad de un gasto alto.
    - Valores bajos (azul, otro motivo de viaje): Se asocian con menor gasto.
    - Los viajes cuyo motivo principal es el turismo deportivo parecen estar asociados con un mayor nivel de gasto. Esto podría indicar que los viajes deportivos están organizados con costos mayores.
- **ACTI_GASTRO_1.0 (Se realizaron actividades gastronómicas en el viaje)**
    - Valores altos (rojo, realizaron actividades gastronómicas): Aumentan la probabilidad de un gasto alto.
    - Valores bajos (azul, no realizaron actividades gastronómicas): Se asocian con un gasto bajo.
    - Participar en actividades gastronómicas es un fuerte predictor de un nivel de gasto alto. Esto sugiere que los turistas que incluyen experiencias gastronómicas en sus viajes tienden a gastar más.
- **TRANSPRIN_5 (El transporte principal fue "Automóvil u otros vehículos particulares")**
    - Valores altos (rojo, uso de transporte privado): Disminuyen la probabilidad de un gasto alto.
    - Valores bajos (azul, uso de otro medio de transporte): Se asocian con mayor gasto.
    - Los turistas que utilizan transporte privado propio o cedido tienden a gastar menos. Esto puede deberse a que evitan costos adicionales como billetes de avión, tren o alquiler de vehículos.
- **ECIVIL_2 (El turista encuestado es casado/a)**
    - Valores altos (rojo, casado/a): Aumentan la probabilidad de un gasto alto.
    - Valores bajos (azul, otro estado civil): Disminuyen la probabilidad de un gasto alto.
    - Los turistas que están casados tienden a gastar más en comparación con aquellos que tienen otro estado civil, lo que podría estar relacionado con

una mayor estabilidad económica, la planificación de viajes en pareja o en familia, y una predisposición a realizar actividades que impliquen un mayor desembolso, como estancias en alojamientos más cómodos, cenas en restaurantes de mayor categoría o la participación en experiencias turísticas más exclusivas.

2.2.6. Conclusiones

El modelo de clasificación desarrollado para predecir el nivel de gasto diario del turista rural ha permitido identificar los principales factores que influyen en el comportamiento de consumo de estos viajeros. Utilizando diferentes algoritmos de Machine Learning, se ha determinado que el modelo XGBoost optimizado con búsqueda bayesiana ofrece el mejor rendimiento predictivo, con una precisión global del 71% y un F1-score equilibrado entre ambas clases (Figura 15). Además, presenta un valor AUC de 0.76 (Figura 16), lo que refleja un desempeño moderado en la predicción de las dos clases.

El análisis de importancia de variables (Figura 28) reveló que los principales determinantes del nivel de gasto son:

1. MIEMV (Número de miembros del hogar en el viaje): Es la variable más influyente, mostrando que, a mayor número de miembros en el hogar, menor es el gasto diario por persona. Esto sugiere que los viajes en familia tienden a ser más moderados en términos de gasto individual.

2. ECIVIL_2 (Estado civil del turista): Los turistas casados tienden a gastar más que otros grupos, posiblemente debido a mayores preferencias por alojamientos más cómodos o actividades premium.

3. ACTI_GASTRO_1.0 (Participación en actividades gastronómicas): Se observó que aquellos turistas que incluyen experiencias gastronómicas en su viaje presentan un mayor gasto diario, lo que resalta la importancia de este tipo de oferta en la industria rural.

4. MOTIV_5 (Turismo deportivo): Los turistas cuyo motivo principal es el turismo deportivo muestran un mayor gasto.

5. TRANSPRIN_5 (Tipo de transporte utilizado): El uso de transporte privado o cedido se asocia con un menor nivel de gasto, en contraste con otros medios de transporte que implican gastos adicionales.

Para entender mejor cómo estas variables afectan las predicciones del modelo, se ha empleado el método SHAP (Figura 29), que permite visualizar el impacto individual de cada característica en la clasificación del nivel de gasto. Además, como apoyo a la interpretabilidad del modelo, se ha generado un árbol de decisión de profundidad 4 (Figura 24), en el que se observa que la variable MIEMV es el nodo raíz, confirmando su importancia en la segmentación del gasto. Y el resto de las variables confirmando las conclusiones mencionadas anteriormente.

3. A-rules Data Mining

Además de los modelos de Machine Learning, hemos estudiado una técnica avanzada de data mining: algoritmo a-rules (*association rules*). En primer lugar, data mining consiste en encontrar información relevante a través del análisis de datos. (Kumbhare & Chobe, 2014). Dentro de este campo, encontramos el algoritmo de reglas de asociación, cuyo objetivo principal es descubrir relaciones significativas entre conjuntos de variables o atributos dentro de grandes bases de datos. Este algoritmo funciona en dos fases: primero, identifica conjuntos frecuentes de ítems, que superen un umbral mínimo de soporte; y segundo, a partir de ellos se generan reglas de asociación, las cuales se evalúan según su nivel de confianza (Kotsiantis, S., & Kanellopoulos, 2006). En este trabajo en concreto, hemos utilizado el algoritmo a priori.

Una regla de asociación se compone del antecedente y el consecuente, por lo que la regla se lee de la siguiente manera: que se de el antecedente implica que se de el consecuente (Kotsiantis, S., & Kanellopoulos, 2006).

Hemos generado dos modelos de reglas de asociación: uno para la variable de clasificación de turista rural, y otro para la clasificación de nivel de gasto diario del turista rural.

3.1. Reglas de asociación para turista rural

Empezamos con el modelo asociado a la clasificación de turista rural. Nuestro objetivo es encontrar reglas concretas que nos permitan identifcar cuando un individuo va a realizar turismo rural, y qué características presentará en consecuencia.

3.1.1. Selección y preparación de los datos

En primer lugar, comenzamos con la selección de variables a usar en el algoritmo. Haciendo uso del previo análisis realizado de mutual information para los modelos de

machine learning, hemos seleccionado aquellas variables que superan el threshold de 0.025. Sin embargo, hemos eliminado de esta lista aquellas variables que hemos considerado como no relevantes por lo que representan, gracias al estudio inicial que realizamos de las variables. Además, hemos añadido, en caso de que no estuviesen entre las de alta mutual information, las variables utilizadas en el modelo de clasificación de nivel de gasto, y la variable objetivo que queremos explicar (A_RURAL).

Tras este proceso, la lista final de variables que hemos utilizado para nuestro algoritmo es la siguiente:

Tabla 11: *Variables utilizadas en el modelo de reglas de asociación de clasificación de turista rural*

Variable
PAQUETE_6.0
ACTI_FAMILIA_1.0
'TRANSPRIN_5'
'VIVSEC_6'
GASTO_ACT_6
GASTO_BIENDUR_6
RELAECON_1.0
GASTOFI_ALOJA
NIVELEST_4.0
MES
MIEMV_15MAS
RESERV_ALOJA_5.0
GASTOFI_TOTAL
MOTIV_3
ACTI_SENDER_1.0
VIAJA_HIJOS_1.0
A_RURAL

Una vez seleccionadas las variables, procedemos a realizar su transformación para poder utilizarlas en el algoritmo a priori. Para ello, transformamos las variables numéricas a categóricas, entre ellas se encuentran: GASTOFI_TOTAL, MIEMV_15MAS, MES.

Además, cabe destacar que hemos filtrado la variable GASTOFI_TOTAL por valores menores a 10000€ dado la presencia de outliers en la variable, como observamos en la sección de metodología de este trabajo. Para la transformación, dividimos los valores de cada variable en cuatro grupos. A continuación, en las Tablas 12, 13 y 14 se puede observar en qué rangos se divide cada variable.

Tabla 12: *División variable GASTOFI_TOTAL en su transformación a variable categórica*

Grupo	**Rango**
Grupo 1	1.19 - 2489.45
Grupo 2	2489.45 - 4967.81
Grupo 3	4967.81 - 7446.16
Grupo 4	7446.16 - 9924.51

Tabla 13: *División variable MIEMV_15MAS en su transformación a variable categórica*

Grupo	**Rango**
Grupo 1	0.99 – 3.00
Grupo 2	3.00 - 5.00
Grupo 3	5.00 - 7.00
Grupo 4	7.00 - 9.00

Tabla 14: *División variable MES en su transformación a variable categórica*

Grupo	**Rango**
Grupo 1	0.99 – 3.75
Grupo 2	3.75 – 6.50
Grupo 3	6.50 – 9.25
Grupo 4	9.25 - 12.00

Después de esta transformación, nuestro dataset contiene 14.843 registros y 26 variables. Dado que este tamaño es demasiado grande para poder procesarlo en nuestro algoritmo, pocedemos a reducirlo. El dataset está muy desbalanceado en cuestión de la variable

objetivo A_RURAL, ya que hay 139337 turistas clasificados como no rurales y 3506 como rural. Es por ello que, en primer lugar, reducimos el número de registros no rurales al mismo número de registros rurales. Además, volvemos a reducir cada una de las secciones de rural y no rural al 50% cada uno. Finalmente, nos quedamos con un total de 3506 registros. Por otra parte, en cuanto a las variables, eliminamos aquellas que sean positivas (igual a 1) en al menos el 1% de los registros. Eliminado así las siguientes cinco variables: GASTOFI_TOTAL_cat_Grupo_2, GASTOFI_TOTAL_cat_Grupo_3, GASTOFI_TOTAL_cat_Grupo_4, MIEMV_15MAS_cat_Grupo_3, MIEMV_15MAS_cat_Grupo_4. Esto nos ha permitido la eficiente ejecución del algoritmo.

3.1.2. Algoritmo a priori y generación de reglas

Para ejecutar nuestro algoritmo, llevamos a cabo una búsqueda exhaustiva de los hiperparámetros utilizados en el algoritmo: soporte y confianza mínimos. Estos dos parámetros, ya mencionados anteriormente, son clave, ya que determinan qué combinaciones de elementos se consideran frecuentes y qué tan fiables son las reglas generadas. Al evaluar diferentes valores posibles, el algoritmo se adapta mejor a las características del conjunto de datos y logra identificar el mayor número posible de asociaciones relevantes. Con este proceso, encontramos que la mejor combinación de hiperparámetros es la siguiente: soporte mínimo igual a 0.05 y confianza mínima igual a 0.3. Con estos valores, el algoritmo encuentra un total de 1799511 reglas.

De entre ellas, filtramos aquellas reglas que contengan la variable A_RURAL en el consecuente, ya que es la variable que estamos tratando de explicar. Quedándonos con un total de 390831 reglas. Entre ellas, analizaremos las 10 reglas que mayor valor de lift, o también conocido como fuerza de la regla, tengan. Estas se muestran en la tabla 15.

Tabla 15: *Top 10 reglas de asociación que explican A_RURAL y variables relacionadas*

Antecedents	Consequents	Support	Confidence	Lift
MOTIV_3, GASTO_BIENDUR_6, GASTOFI_TOTAL_cat_Grupo_1, PAQUETE_6.0, RELAECON_1.0	MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.320	3.266
MOTIV_3, GASTO_BIENDUR_6,	MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL,	0.053	0.319	3.260

RELAECON_1.0, PAQUETE_6.0	ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0			
MOTIV_3, GASTO_BIENDUR_6, RELAECON_1.0, PAQUETE_6.0	GASTOFI_TOTAL_cat_Grupo_1, MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.319	3.260
MOTIV_3, RELAECON_1.0, PAQUETE_6.0, GASTOFI_TOTAL_cat_Grupo_1	GASTO_BIENDUR_6, MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.318	3.255
MOTIV_3, RELAECON_1.0, PAQUETE_6.0, GASTOFI_TOTAL_cat_Grupo_1	MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.318	3.255
MOTIV_3, GASTO_BIENDUR_6, GASTOFI_TOTAL_cat_Grupo_1, MIEMV_15MAS_cat_Grupo_1, PAQUETE_6.0, RELAECON_1.0	NIVELEST_4.0, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5	0.053	0.331	3.250
MOTIV_3, GASTO_BIENDUR_6, MIEMV_15MAS_cat_Grupo_1, PAQUETE_6.0, RELAECON_1.0	GASTOFI_TOTAL_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.330	3.244
MOTIV_3, GASTO_BIENDUR_6, MIEMV_15MAS_cat_Grupo_1, PAQUETE_6.0, RELAECON_1.0	NIVELEST_4.0, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5	0.053	0.330	3.244
MOTIV_3, RELAECON_1.0, PAQUETE_6.0	GASTOFI_TOTAL_cat_Grupo_1, MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.317	3.244

MOTIV_3, RELAECON_1.0, PAQUETE_6.0	GASTO_BIENDUR_6, GASTOFI_TOTAL_cat_Grupo_1, MIEMV_15MAS_cat_Grupo_1, GASTO_ACT_6, A_RURAL, ACTI_SENDER_1.0, TRANSPRIN_5, NIVELEST_4.0	0.053	0.317	3.244

Fuente: Elaboración propia con la librería pyplot de python

Todas estas reglas son una combinación de variables que, cuando se presentan conjuntamente en un turista, permiten identificar un perfil específico con alta probabilidad de estar asociado al turismo rural. Estas combinaciones revelan patrones de comportamiento y características sociodemográficas que, de manera conjunta, incrementan significativamente la probabilidad de que el individuo pertenezca a este segmento. Así, estas reglas no solo reflejan relaciones individuales entre variables, sino que muestran cómo ciertos factores actúan en conjunto para definir perfiles turísticos relevantes, lo cual resulta de gran utilidad para el diseño de estrategias de promoción, segmentación de mercado y toma de decisiones en el ámbito del turismo rural.

En cuanto a las variables que aparecen en estas 10 mejores reglas, encontramos 10 de las 21 utilizadas inicialmente:

- ACTI_SENDER_1.0
- A_RURAL
- GASTOFI_TOTAL_cat_Grupo_1
- GASTO_ACT_6
- GASTO_BIENDUR_6
- MIEMV_15MAS_cat_Grupo_1
- MOTIV_3
- NIVELEST_4.0
- PAQUETE_6.0
- RELAECON_1.0
- TRANSPRIN_5

La variable A_RURAL aparece en todas las reglas en la parte del consecuente, ya que así hemos decidido filtrar las reglas. Es por ello por lo que omitiremos esta variable a la hora

de describirlas. A continuación, se encuentra una descripción de cada una de las reglas mostradas en la Tabla 15.

Regla 1

Causa: El motivo principal del viaje es el turismo de naturaleza, se ha realizado gasto en bienes duraderos, el gasto total es inferior a 2.489,45 €, se ha adquirido un paquete turístico y el turista se encuentra económicamente ocupado.

Efecto: Es probable que en el viaje participaran menos de tres miembros del hogar mayores de 15 años, haya realizado gastos en actividades, haya realizado actividades de senderismo, se haya desplazado en vehículo particular y el nivel de estudios del turista sea superior.

Regla 2

Causa: El motivo del viaje es realizar turismo de naturaleza, se ha hecho gasto en bienes duraderos, el turista está económicamente ocupado y ha adquirido un paquete turístico.

Efecto: Es probable que en el viaje participaran menos de tres miembros del hogar mayores de 15 años, haya realizado gastos en actividades, haya realizado actividades de senderismo, se haya desplazado en vehículo particular y posea estudios superiores.

Regla 3

Causa: Cuando el motivo del viaje es turismo de naturaleza, hay gasto en bienes duraderos, se ha adquirido un paquete turístico y su situación económica es "ocupado/a".

Efecto: Es probable que gasto total del viaje se sitúe por debajo de 2.489,45 €, que hayan participado en el viaje menos de tres miembros del hogar mayores de 15 años, haya realizado gasto en actividades y haya realizado actividades de senderismo, que el transporte principal utilizado en el viaje se un vehículo particular, y su nivel de estudios sea superior.

Regla 4

Causa: El motivo del viaje es turismo de naturaleza, el turista está ocupado económicamente, ha adquirido un paquete turístico, y haya realizado un gasto total menor de 2.489,45€.

Efecto: Es muy probable que se realicen gastos en bienes duraderos y en actividades, se

practiquen actividades de senderismo, y se cumplan condiciones como transporte en vehículo particularm y nivel de estudios superior.

Regla 5

Causa: El motivo del viaje es turismo de naturaleza, el turista está ocupado económicamente, ha adquirido un paquete turístico, y haya realizado un gasto total menor de 2.489,45€.

Efecto: Es muy probable que se realicen gastos en actividades, se practiquen actividades de senderismo, y se cumplan condiciones como transporte en vehículo particularm y nivel de estudios superior.

Regla 6

Causa: El motivo principal del viaje es turismo de naturaleza, se ha realizado gasto en bienes duraderos, se ha realizado un gasto total menor de 2.489,45€, han participado en el viaje menos de tres miembros del hogar mayores de 15 años, se ha adquirido un paquete turístico y el turista está económicamente ocupado.

Efecto: Es probable que el nivel de estudios del turista sea superior, haya realizado gasto en actividades, haya realizado actividades de senderismo y el transporte principal utilizado en el viaje sea un vehículo particular.

Regla 7

Causa: El motivo principal del viaje es turismo de naturaleza, se ha realizado gasto en bienes duraderos, han participado en el viaje menos de tres miembros del hogar mayores de 15 años, se ha adquirido un paquete turístico y el turista está económicamente ocupado.

Efecto: Es probable que se ha realizado un gasto total menor de 2.489,45€, y se ha realizado gasto en actividades, se han realizado actividades de senderismo, el transporte principal utilizado en el viaje es un vehículo particular, y su nivel de estudios es superior.

Regla 8

Causa: El motivo principal del viaje es turismo de naturaleza, se ha realizado gasto en bienes duraderos, han participado en el viaje menos de tres miembros del hogar mayores de 15 años y se ha adquirido un paquete turístico.

Efecto: El nivel de estudios del turista es superior, ha realizado gasto en actividades, ha realizado actividades de senderismo y el transporte principal utilizado en el viaje es un vehículo particular.

Regla 9

Causa: El motivo principal del viaje es turismo de naturaleza, el turista se encuentra económicamente ocupado y se ha adquirido un paquete turístico.

Efecto: Es probable que se ha realizado un gasto total menor de 2.489,45€, han participado en el viaje menos de tres miembros del hogar mayores de 15 años, se ha realizado gasto en actividades, se han realizado actividades de senderismo, el transporte principal utilizado en el viaje es un vehículo particular y el nivel de estudios del turista es superior.

Regla 10

Causa: El motivo principal del viaje es turismo de naturaleza, el turista se encuentra económicamente ocupado y se ha adquirido un paquete turístico.

Efecto: Es probable que se ha realizado gasto en bienes duraderos, el gasto total del viaje es menor de 2.489,45€, han participado en el viaje menos de tres miembros del hogar mayores de 15 años, se ha realizado gasto en actividades, se han realizado actividades de senderismo, el transporte principal utilizado en el viaje es un vehículo particular y el nivel de estudios del turista es superior.

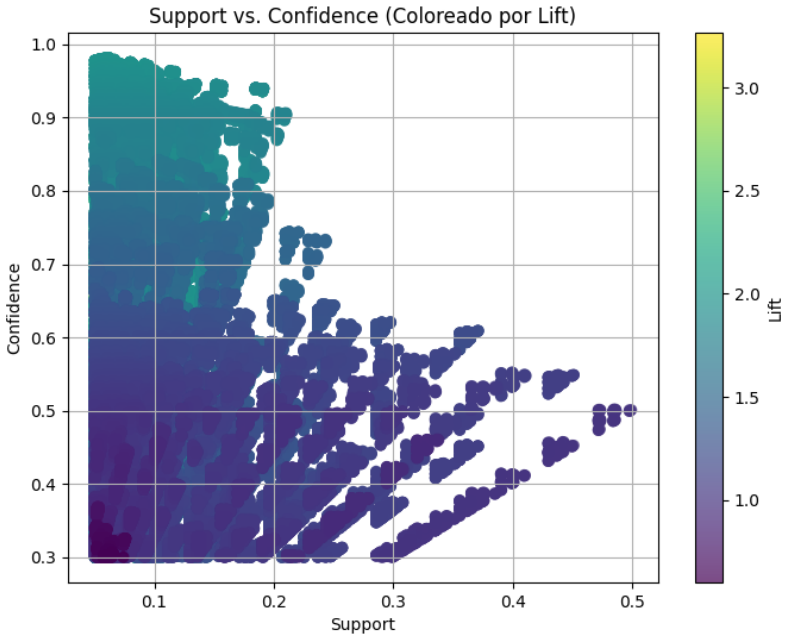
Todas las reglas obtenidas presentan un soporte constante de 0.053, lo que indica que aproximadamente el 5.3 % del total de las observaciones del conjunto de datos cumplen simultáneamente con las condiciones del antecedente y del consecuente. En cuanto a la confianza, los valores oscilan entre 0.317 y 0.331, lo que refleja una probabilidad moderadamente alta de que el consecuente se cumpla cuando el antecedente está presente. Por último, el lift de todas las reglas se sitúa entre 3.24 y 3.26, lo que indica que estas asociaciones tienen una fuerza significativa: la ocurrencia conjunta de los elementos del antecedente y el consecuente es más de tres veces más probable que si fueran eventos independientes. Estas métricas en conjunto confirman la relevancia, consistencia y utilidad de las reglas identificadas para explicar patrones en el perfil del turista rural.

Finalmente, de manera complementaria, en la Figura 30 se muestra una representación visual de las reglas generadas con el algoritmo. El gráfico representa un diagrama de

dispersión que muestra las reglas de asociación obtenidas en el análisis, donde el eje X indica el soporte (frecuencia relativa con la que ocurre una regla en el conjunto de datos), el eje Y refleja la confianza (probabilidad de que ocurra el consecuente dado el antecedente), y el color de los puntos representa el lift (indicador de la fuerza de la regla) utilizando una escala de colores desde morado (bajo lift) hasta amarillo (alto lift).

Se observa una alta concentración de reglas con bajo soporte, pero alta confianza; es decir, muchas reglas se cumplen con alta probabilidad, pero en un subconjunto reducido de casos. Además, los valores de lift no superan el valor de 3.5, lo que indica que, si bien hay reglas interesantes, no todas tienen un poder predictivo fuerte.

***Figura 26**: Gráfico de dispersión de las reglas de asociación generadas sobre la variable A_RURAL mediante el algoritmo Apriori. El eje X representa el soporte, el eje Y la confianza, y el color de los puntos el valor de lift. Esta visualización permite identificar las reglas más interesantes según las métricas de evaluación mencionadas.*



Fuente: Elaboración propia con la librería pyplot

3.2. Reglas de asociación nivel de gasto turístico rural

En segundo lugar, estudiamos las reglas de asociación relacionadas con la variable de gasto turístico rural medio.

3.2.1. Selección y preparación de los datos

Para la selección de variables, seguimos el mismo proceso explicado en el modelo anterior. Sin embargo, en este caso, el umbral que hemos marcado para seleccionar las

variables con alta mutual information es de 0.015. Y la variable objetivo de este modelo ha sido CLASIF2_GASTO.

Habiendo realizado estos pasos, nuestra lista final de variables para la creación de reglas de asociación sobre el nivel de gasto turístico rural medio se encuentran en la siguiente tabla:

Tabla 16: *Variables utilizadas en el modelo de reglas de asociación del nivel de gasto medio diario del turista rural*

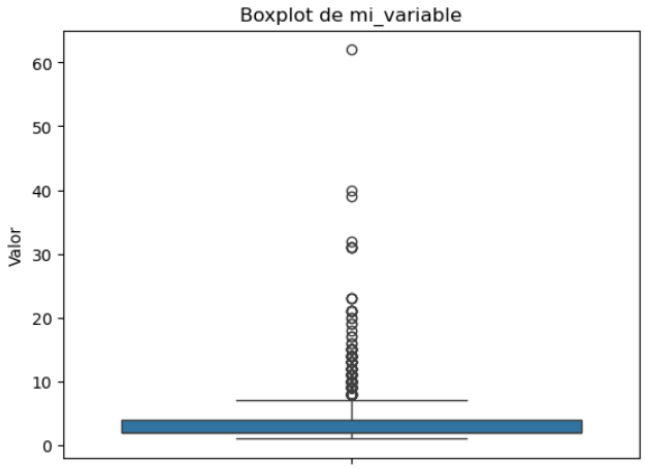
Variable
MIEMV
VIAJA_HIJOS_1.0
TIPHOGAR_4
NPERNOC
CONV_3
TIPHOGAR_1
DESTESP_6
ACTI_OTROSNAUTICOS_1.0
ACTI_JUEGOS_1.0
ECIVIL_2
VIAJA_SOLO_1.0
PROVDEST_37
CCAA_RESIDENCIA_15
PAISNACIO_2
TIPOVIAJ_3
TRANSPRIN_5
CLASIF2_GASTO

Para este modelo, también hemos procedido a la transformación de las variables para su uso en el algoritmo.

En primer lugar, comenzamos con la transformación de variables numéricas (MIEMV y NPERNOC). Para ello, creamos cuatro categorías para cada una de ellas. En concreto, para la variable NPERNOC hemos realizado una previa transformación, ya que esta contaba con outliers por encima de las 30 noches (como se observa en la Figura 27). Para

ello, hemos transformado todos los valores mayores de 30 al mayor valor por debajo de 30 (23 noches). Con este proceso, se crean nuevas variables llamadas MIEMV_cat y NPERNOC_trunc_cat, y procedemos a eliminar las originales. La división de las variables en categóricas la mostramos en las Tablas 17 y 18.

Figura 27: *Boxplot de la variable NPERNOC donde en el eje X se muestra el número de pernoctaciones del viaje. El primer y el tercer cuartil se sitúan entre 2 y 4 noches, y la media es 3.29 noches por viaje.*



Fuente: Elaboración propia con la librería pyplot de python

Tabla 17: *División variable MIEMV en su transformación a variable categórica*

Grupo	Rango
Grupo 1	0.99 - 3.00
Grupo 2	3.00 - 5.00
Grupo 3	5.00 - 7.00
Grupo 4	7.00 - 9.00

Tabla 18: *División variable NPERNOC en su transformación a variable categórica*

Grupo	Rango
Grupo 1	0.98 - 6.50
Grupo 2	6.50 - 12.00
Grupo 3	12.00 - 17.50
Grupo 4	17.50 - 23.00

A continuación, convertimos estas variables categóricas en dicotómicas mediante one-hot encoding. Y, por último, trasformamos todas las variables a tipo booleano (True/False).

3.2.2. Algoritmo a priori y generación de reglas

Para este algoritmo realizamos también la búsqueda de hiperparámetros, y encontramos como la mejor combinación soporte mínimo de 0.001 y confianza mínima de 0.3. Con estos valores, el algoritmo encuentra un total de 18509.

De entre ellas, filtramos aquellas reglas que contengan la variable CLASIF2_GASTO en el consecuente, reduciendo el número de reglas a 4276. A continuación, en la tabla 19, mostramos las 10 reglas que mayor valor de lift presentan:

Tabla 19: *Top 10 reglas que mejor explican CLASIF2_GASTO y variables relacionadas*

Antecedents	Consequents	Support	Confidence	Lift
TIPHOGAR_1, NPERNOC_trunc_cat_Grupo_1, TIPOVIAJ_3	CLASIF2_GASTO, VIAJA_SOLO_1.0, TRANSPRIN_5, CONV_3	0.002	0.583	30.753
TIPHOGAR_1, NPERNOC_trunc_cat_Grupo_1, MIEMV_cat_Grupo_1, TIPOVIAJ_3	CLASIF2_GASTO, VIAJA_SOLO_1.0, TRANSPRIN_5, CONV_3	0.002	0.583	30.753
TIPHOGAR_1, NPERNOC_trunc_cat_Grupo_1, TIPOVIAJ_3	CLASIF2_GASTO, MIEMV_cat_Grupo_1, TRANSPRIN_5, VIAJA_SOLO_1.0, CONV_3	0.002	0.583	30.753
NPERNOC_trunc_cat_Grupo_1, CONV_3, TRANSPRIN_5, TIPOVIAJ_3	TIPHOGAR_1, VIAJA_SOLO_1.0, CLASIF2_GASTO	0.002	0.500	28.593
NPERNOC_trunc_cat_Grupo_1, MIEMV_cat_Grupo_1, TRANSPRIN_5, TIPOVIAJ_3, CONV_3	TIPHOGAR_1, VIAJA_SOLO_1.0, CLASIF2_GASTO	0.002	0.500	28.593
NPERNOC_trunc_cat_Grupo_1, CONV_3, TRANSPRIN_5, TIPOVIAJ_3	TIPHOGAR_1, VIAJA_SOLO_1.0, MIEMV_cat_Grupo_1, CLASIF2_GASTO	0.002	0.500	28.593
TIPHOGAR_1, TIPOVIAJ_3	CLASIF2_GASTO, MIEMV_cat_Grupo_1, TRANSPRIN_5, VIAJA_SOLO_1.0, CONV_3	0.002	0.538	28.387
TIPHOGAR_1, TIPOVIAJ_3	CLASIF2_GASTO, NPERNOC_trunc_cat_Grupo_1, TRANSPRIN_5, VIAJA_SOLO_1.0, CONV_3	0.002	0.538	28.387

TIPHOGAR_1, MIEMV_cat_Grupo_1, TIPOVIAJ_3	CLASIF2_GASTO, VIAJA_SOLO_1.0, TRANSPRIN_5, CONV_3	0.002	0.538	28.387
TIPHOGAR_1, TIPOVIAJ_3	CLASIF2_GASTO, NPERNOC_trunc_cat_Grupo_1, MIEMV_cat_Grupo_1, TRANSPRIN_5, VIAJA_SOLO_1.0, CONV_3	0.002	0.538	28.387

Fuente: Elaboración propia con la librería pyplot de python

En cuanto a las variables que aparecen en estas 10 mejores reglas, encontramos 8 de las 23 utilizadas inicialmente:

- TIPOHOGAR_1
- NPERNOCT_trunc_cat_Grupo_1
- TIPOVIAJ_3
- CLASIF2_GASTO
- VIAJA_SOLO_1.0
- TRANSPRIN_5
- CONV_3
- MIEMV_cat_Grupo_1

A continuación, describiremos las reglas generadas. En este caso, todas hechas contienen la variable CLASIF2_GASTO en el consecuente; es decir, todas tienen como consecuencia que el nivel de gasto del turista rural es alto. Es por ello que omitiremos esta variable en la descripción de las reglas.

Regla 1

Causa: El hogar es unipersonal, el número de pernoctaciones del viaje está entre 1 y 6 noches, y el tipo de viaje es de trabajo;
Efecto: El turista viaja solo, el transporte principal es un vehículo particular y no convive en pareja.

Regla 2

Causa: El hogar es unipersonal, el número de pernoctaciones del viaje está entre 1 y 6

noches, el número de miembros del hogar que participaron en el viaje está entre 1 y 3 y el tipo de viaje es de trabajo.
Efecto: El turista viaja solo, el transporte principal es un vehículo particular y no convive en pareja.

Regla 3

Causa: El hogar es unipersonal, el número de pernoctaciones del viaje está entre 1 y 6 noches, y el tipo de viaje es de trabajo;
Efecto: El número de miembros del hogar que participaron en el viaje está entre 1 y 3, el turista viaja solo, el transporte principal es un vehículo particular y no convive en pareja.

Regla 4

Causa: El número de pernoctaciones del viaje está entre 1 y 6 noches, el turista no convive en pareja, el transporte principal es un vehículo particular y no convive en pareja;
Efecto: El hogar es unipersonal y el turista viaja solo.

Regla 5

Causa: El número de pernoctaciones del viaje está entre 1 y 6 noches, el número de miembros del hogar que participaron en el viaje está entre 1 y 3, el transporte principal utilizado en el viaje es un vehículo particular, el viaje es de tipo trabajo, y el turista no convive en pareja;
Efecto: El hogar es unipersonal y el turista viaja solo.

Regla 6

Causa: El número de pernoctaciones del viaje está entre 1 y 6 noches, no convive en pareja, el transporte principal utilizado en el viaje es un vehículo particular y el viaje es de tipo trabajo;
Efecto: El hogar es unipersonal, el turista viaja solo y el número de miembros del hogar que participaron en el viaje está entre 1 y 3.

Regla 7

Causa: El hogar del turista es unipersonal y el viaje es de tipo trabajo; **Efecto:** El número de miembros del hogar que participaron en el viaje está entre 1 y 3, el transporte principal utilizado en el viaje es un vehículo particular, el turista viaja solo y no convive en pareja.

Regla 8

Causa: El hogar del turista es unipersonal y el viaje es de tipo trabajo; **Efecto:** El número de pernoctaciones del viaje está entre 1 y 6 noches, el transporte principal utilizado en el viaje es un vehículo particular, el turista viaja solo y no convive en pareja.

Regla 9

Causa: El hogar del turista es unipersonal, el número de miembros del hogar que participaron en el viaje está entre 1 y 3, y el viaje es de tipo trabajo. **Efecto:** El turista viaja solo, el transporte principal utilizado en el viaje es un vehículo particular y no convive en pareja.

Regla 10

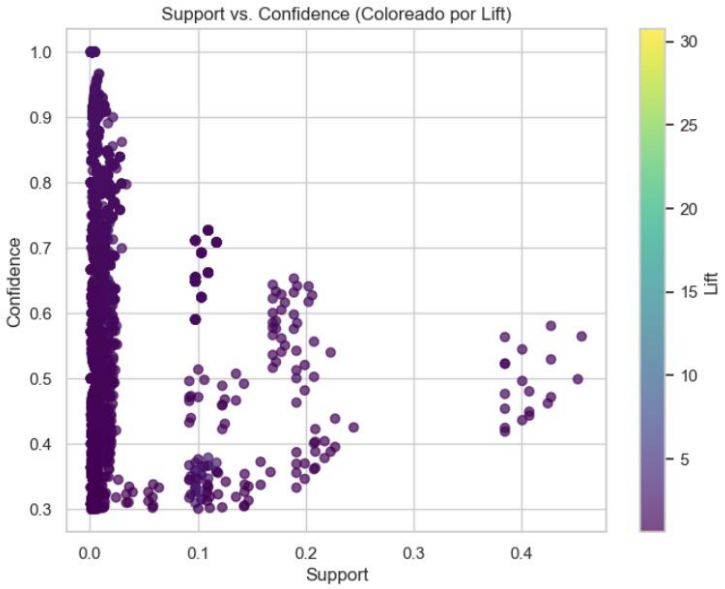
Causa: Cuando el hogar del turista es unipersonal y el viaje es de tipo trabajo. **Efecto:** El número de pernoctaciones del viaje está entre 1 y 6 noches, el número de miembros del hogar que participaron en el viaje está entre 1 y 3, el transporte principal utilizado en el viaje es un vehículo particular, el turista viaja solo y no convive en pareja.

Las **reglas 1, 2 y 3** están presentes en el 0.2 % de los datos; cuando se dan los antecedentes, en un 50 % de los casos también se da el consecuente, y tienen una fuerte asociación, ya que el *lift* de todas ellas es igual a **28.593** (muy por encima de 1). Las **reglas 4, 5 y 6** están presentes en el 0.2 % de los datos; cuando se dan los antecedentes, en un 58.3 % de los casos también se da el consecuente, y tienen una fuerte asociación, ya que el *lift* de todas ellas es igual a **30.752** (muy por encima de 1). Y las **reglas 7, 8, 9 y 10**, están presentes en el 0.2% de los datos. Cuando se dan sus antecedentes, en un

53.8% de los casos también se da el consecuente. Y tienen una fuerte asociación ya que el lift de todas ellas es igual a 28.387 (muy por encima de 1).

Además, mostramos una visualizaciçon de las 4276 reglas generadas que contienen A_RURAL en el consecuente. La Figura 28 muestra un diagrama de dispersión donde el eje X representa el soporte, y el eje Y la confianza, utilizando el color de los puntos para indicar el lift (impulso) de cada regla. Este gráfico nos permite identificar visualmente las reglas más relevantes: aquellas con mayor confianza (más arriba en el eje Y) indican una mayor probabilidad de que el consecuente ocurra dado el antecedente, mientras que un mayor soporte (más a la derecha) señala reglas más frecuentes en el conjunto de datos. El color más claro (amarillo) representa un lift alto, lo que significa que la regla tiene mayor capacidad de predicción.

Figura 28: *Gráfico de dispersión de las reglas de asociación generadas sobre la variable CLASIF2_GASTO mediante el algoritmo Apriori. El eje X representa el soporte, el eje Y la confianza, y el color de los puntos el valor de lift. Esta visualización permite identificar las reglas más interesantes según las métricas de evaluación mencionadas.*



Fuente: Elaboración propia con la librería pyplot

4. Conclusiones y recomendaciones generales

Este Trabajo de Fin de Grado ha tenido como objetivo principal la aplicación de modelos de Machine Learning para mejorar la comprensión del turismo rural en España y optimizar su gestión. Como se observa en el Capítulo 2, los modelos creados han

mostrado una capacidad razonable para clasificar a los turistas como rurales o no rurales (*ver Figura 10*) y para predecir su nivel de gasto diario (*ver Figura 22*)

El turismo rural en España ha emergido como una **estrategia efectiva para dinamizar las zonas rurales y frenar la despoblación**. Su impacto económico es significativo, contribuyendo al desarrollo local y generando oportunidades laborales en sectores complementarios como la hostelería y el comercio. A través del análisis realizado, se ha evidenciado que **la digitalización y la integración de tecnologías avanzadas** pueden jugar un papel crucial en la expansión del turismo rural.

En cuanto a nuestro estudio, en primer lugar, el primer modelo de clasificación implementado ha permitido definir un conjunto de variables clave que diferencian al turista rural del no rural (*ver Figuras 12 y 13, Capítulo 2*). Entre los principales resultados podemos destacar:

- **Gasto total del viaje**: Los turistas rurales tienden a realizar un mayor gasto total en comparación con los turistas no rurales.
- **Medio de transporte**: El uso de vehículo particular es una característica predominante en los turistas rurales, lo que sugiere que la accesibilidad a los destinos influye en la decisión de viaje.
- **Reserva del alojamiento**: La mayoría de los turistas rurales utilizan plataformas especializadas como Airbnb o Booking para reservar su alojamiento, lo que resalta la importancia de la digitalización en el sector.
- **Actividades realizadas:** Los turistas rurales suelen participar en actividades relacionadas con la naturaleza, como senderismo, lo que confirma el atractivo de estos entornos para este tipo de viajero.

En segundo lugar, el modelo de clasificación del nivel de gasto dentro de los turistas rurales ha identificado variables determinantes que influyen en el comportamiento de consumo del turista rural (*ver Figuras 24 y 25, Capítulo 3*). Entre ellas encontramos:

- **Estado civil:** Se ha encontrado que los turistas casados tienden a gastar más en comparación con otros grupos, posiblemente debido a viajes en pareja con mayor nivel de confort y gasto.

- **Actividades gastronómicas y deportivas**: La participación en experiencias gastronómicas y turismo deportivo está estrechamente relacionada con un mayor gasto.
- **Medio de transporte:** Los turistas que utilizan transporte privado propio o cedido tienden a tener un gasto menor, mientras que aquellos que viajan en otros medios de transporte incurren en mayores costos.
- **Número de miembros del hogar:** Se ha detectado que los viajes en pareja o en pequeños grupos tienden a generar un mayor nivel de gasto diario por persona en comparación con viajes en grupos grandes.

Además de los modelos supervisados, la incorporación de **modelos de reglas de asociación** en el análisis turístico ha supuesto un valor añadido complementario (*ver Capítulo 3*). Mientras que los algoritmos de clasificación y predicción permiten estimar resultados concretos a partir de ciertas variables, las reglas de asociación aportan una perspectiva más exploratoria y descriptiva, al identificar combinaciones de características que aparecen frecuentemente de forma conjunta en determinados perfiles de turistas. Este enfoque permite descubrir relaciones no evidentes entre variables, y genera conocimiento útil para la segmentación y la toma de decisiones estratégica

En primer lugar, el modelo de reglas de asociación sobre la variable clasificatoria de turista rural ha permitido identificar combinaciones frecuentes de variables sociodemográficas y comportamentales que caracterizan a este segmento. Se han detectado reglas en las que intervienen factores como el **motivo del viaje (turismo de naturaleza), la adquisición de paquetes turísticos, la realización de actividades como el senderismo, el nivel educativo y el transporte utilizado**, entre otros (*ver Tabla 15, Capítulo 3*). Estas reglas presentan valores de *lift* de hasta 3.26, lo que indica una asociación relevante entre los conjuntos de antecedentes y consecuentes. Además, estas asociaciones permiten perfilar mejor al turista rural y ofrecen información práctica para diseñar estrategias de promoción y experiencias personalizadas, especialmente en el ámbito de los alojamientos rurales y la oferta de actividades.

En segundo lugar, en relación con el modelo de predicción del gasto medio por persona de los turistas rurales, se concluye que existen determinadas variables que permiten estimar con cierta precisión el nivel de gasto alto. Entre las variables que predominan en esta clasificación encontramos **viajes unipersonales, estacias cortas, no convivencia en**

pareja y transporte privado. (*ver Tabla 19, Capítulo 3*) Las reglas extraídas presentan valores de *lift* superiores a 28, lo cual refleja una gran fuerza de las reglas. Este tipo de análisis resulta especialmente valioso para agentes del sector rural que buscan maximizar el impacto económico del turismo en sus territorios.

Los resultados de este TFG refuerzan algunas ideas ya presentes en la literatura previa, pero también aportan una perspectiva metodológica innovadora y conclusiones no abordadas hasta ahora de forma empírica. En primer lugar, los antecedentes recogidos en la introducción subrayan el papel del turismo rural como motor de desarrollo económico, herramienta frente al despoblamiento y vehículo para un modelo de turismo más sostenible (CaixaBank Research, 2023; Loscertales, 1999; Martín Gil, 2014). En este sentido, las conclusiones de nuestro trabajo son coherentes con dichas ideas, ya que identifican un perfil de turista rural vinculado a actividades en la naturaleza, el uso del transporte privado y una mayor disposición al gasto, características que pueden potenciar el desarrollo local.

No obstante, **este estudio va más allá de los enfoques recogidos en los antecedentes**, incorporando por primera vez técnicas específicas de machine learning supervisado y reglas de asociación una base de datos oficial como la Encuesta de Turismo de Residentes (ETR). A diferencia de otros trabajos que destacaban el uso de tecnologías en la mejora de la experiencia turística (como las plataformas de recomendación tipo MyStreetBook o los Destinos Turísticos Inteligentes (DTI)), este análisis adopta un enfoque orientado al **perfilado predictivo del turista rural**, lo cual representa una aportación metodológica nueva y complementaria.

Con base a los resultados obtenidos, se proponen las siguientes recomendaciones para mejorar la **gestión de los alojamientos rurales**:

- **Impulsar la digitalización y fortalecer la presencia online**. Dado que la mayoría de los turistas rurales reservan su alojamiento a través de plataformas digitales, es fundamental mejorar la visibilidad en estos canales (por ejemplo, en plataformas como Airbnb o Booking) y optimizar la experiencia del usuario en los sitios web de los alojamientos. Además, invertir en publicidad en redes sociales o en portales de viajes puede aumentar la visibilidad y atraer a más turistas.

- **Diseñar paquetes de actividades complementarias**. La integración de paquetes turísticos que incluyan actividades gastronómicas, deportivas y en la naturaleza no solo permite personalizar la oferta, sino que también incrementa los ingresos de los alojamientos rurales, dado que estos factores están asociados con un mayor gasto por parte de los visitantes.
- **Segmentar el público objetivo para estrategias de marketing.** Los viajeros en pareja o en grupos reducidos han demostrado un mayor gasto diario en comparación con grupos numerosos. Diseñar campañas de marketing dirigidas a estos segmentos representa una oportunidad de crecimiento para el sector. Un ejemplo de estrategia podría ser la promoción de experiencias exclusivas bajo el lema: *"Escapada romántica para desconectar"* o *"Fin de semana gourmet en la naturaleza"*.
- **Crear productos turísticos específicos para el viajero individual.** Como, por ejemplo, paquetes personalizados, actividades de bienestar o escapadas cortas pensadas para la desconexión. Esto servirá para ajustarse al perfil identificado de turista con gasto elevado.

A **nivel estratégico**, es necesario que el sector público apoye e impulse el turismo rural a través de iniciativas concretas, como:

- **Programas de digitalización.** Implementar políticas que faciliten la transformación digital de los alojamientos rurales mediante programas de subvenciones y formación para mejorar su presencia en plataformas digitales y redes sociales.
- **Fomento del turismo en períodos de baja demanda.** Para mantener un flujo constante de visitantes durante todo el año, se recomienda diseñar estrategias que incentiven el turismo en temporadas bajas. Algunas iniciativas incluyen:
    - Descuentos y promociones especiales dirigidos a distintos segmentos de turistas (familias, parejas o grupos).
    - Creación de eventos temáticos y actividades culturales, como festivales gastronómicos, ferias de artesanía, conciertos y competiciones deportivas.
    - Desarrollo de experiencias exclusivas para atraer turistas interesados en el bienestar y la desconexión, como retiros de yoga, enoturismo o rutas de senderismo guiadas.

Además, es fundamental que todas estas estrategias se diseñen bajo **un enfoque de sostenibilidad**, promoviendo un turismo responsable que respete el medioambiente y contribuya al desarrollo equilibrado de las comunidades locales. Para ello, se debe fomentar el uso de recursos naturales de manera eficiente, minimizar el impacto ecológico de las actividades turísticas y potenciar iniciativas que integren a la población local en la gestión y beneficio del turismo rural. La implementación de prácticas como la promoción de alojamientos respetuoso con el medioambiente, la organización de eventos con criterios de sostenibilidad y la sensibilización de los visitantes sobre la importancia de la conservación del entorno garantizará un crecimiento turístico armonioso, alineado con los principios del turismo sostenible.

5. Futuras líneas de investigación

Este trabajo ha permitido identificar características clave del turista rural y del de su nivel de gasto medio diario utilizando modelos de machine learning y minería de datos. No obstante, existe una oportunidad muy fuerte en la investigación del turismo rural en España, donde encontramos múltiples líneas futuras que podrían enriquecer y ampliar los hallazgos obtenidos.

Este trabajo ha permitido identificar características clave del turista rural y de su nivel de gasto medio diario utilizando modelos de *machine learning* y minería de datos. No obstante, existe una oportunidad muy fuerte en la investigación del turismo rural en España, donde encontramos múltiples líneas futuras que podrían enriquecer y ampliar los hallazgos obtenidos. Cabe señalar que, debido a limitaciones de tiempo y recursos, no ha sido posible implementar algunas de las técnicas más avanzadas o explorar enfoques alternativos, por lo que se plantean como propuestas para futuros estudios.

En primer lugar, una **ampliación de la fuente de datos sería clave**. Contar con información procedente de **empresas del sector turístico rural**, como alojamientos, agencias de viajes o plataformas de reserva (por ejemplo, Airbnb), permitiría realizar un análisis interno y específico del sector. Estos datos aportarían variables adicionales y más detalladas, como el género del turista, su comportamiento de compra, valoraciones de servicios (comentarios cualitativos y cuantitativos), frecuencia de visitas, o datos relativos al canal de reserva. Esta información enriquecería enormemente la calidad del análisis y permitiría explorar dimensiones actualmente no cubiertas en los microdatos oficiales..

Además, disponer de estos datos facilitaría la realización de **análisis por localización específica**, centrados en zonas concretas del país. Esto nos permitiría llevar a cabo análisis ad-hoc para territorios concretos, lo que permitiría pasar de una visión general a una visión operativa, real y personalizada.

En segundo lugar, la incorporación de **técnicas más avanzadas de machine learning** ampliaría la capacidad predictiva y descriptiva del modelo. El uso de **técnicas de aprendizaje no supervisado**, como el **clustering jerárquico o K-means**, permitiría descubrir grupos de turistas con características similares, sin necesidad de definir previamente los segmentos. Tal y como demuestra el estudio de Rodríguez et al. (2018), este método es especialmente útil para identificar patrones de visita, duración de la estancia y tipo de actividad realizada, lo que permite a los destinos turísticos diseñar estrategias diferenciadas y tomar decisiones más informadas.

Otra línea de investigación especialmente relevante sería el uso de **algoritmos de series temporales.** Aplicar modelos como **ARIMA, SARIMA o Prophet** permitiría analizar la evolución del turismo rural a lo largo del tiempo y predecir su comportamiento futuro en función de datos mensuales agregados, como el número de turistas, el gasto medio por persona o la ocupación media de alojamientos. Estas técnicas permiten identificar patrones estacionales, tendencias de crecimiento o caída, e incluso anomalías vinculadas a eventos externos como crisis económicas o sanitarias. Además, al combinar estas series con variables explicativas adicionales (como el precio medio, festividades locales o condiciones meteorológicas), podrían construirse modelos multivariantes que ofrecieran un análisis mucho más robusto. Esto facilitaría tanto la **toma de decisiones operativas** como la **planificación estratégica a medio y largo plazo**.

De hecho, estudios recientes, como el de Dong et al. (2023), han propuesto modelos que mejoran la predicción del turismo usando series temporales con mecanismos de atención, es decir, modelos que "aprenden" a centrarse en las partes más importantes de los datos. Estos enfoques superan a los métodos clásicos y funcionan especialmente bien cuando la demanda turística es irregular o cambiante, por lo que podrían ser muy útiles para analizar el turismo rural en España.

En conjunto, estas líneas de investigación contribuirían a una comprensión más profunda, precisa y actualizada del turismo rural en España, generando conocimiento valioso para empresas, administraciones y comunidades locales.

6. Bibliografía

Bakırarar, B., & Elhan, A. H. (2023). Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research. *Turkiye Klinikleri Journal of Biostatistics, 15*(1), 19-29. https://doi.org/10.5336/biostatic.2022-93961

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. Retrieved from https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016, September 6). How to Tune the Number and Size of Decision Trees with XGBoost in Python - MachineLearningMastery.com. Retrieved February 25, 2025, from MachineLearningMastery.com website: https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/

CaixaBank Research. (2023). *El auge del turismo rural en España: una oportunidad para el desarrollo rural*. Recuperado de https://www.caixabankresearch.com/es/analisis-sectorial/agroalimentario/auge-del-turismo-rural-espana-oportunidad-desarrollo-rural

Cerda L, J., & Villarroel Del P, L. (2007). Interpretación del test de Chi-cuadrado (X^2) en investigación pediátrica. *Revista Chilena de Pediatría*, *78*(4), 414–417. https://doi.org/10.4067/S0370-41062007000400010

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Dong, Y., Xiao, L., Wang, J., & Wang, J. (2023). A time series attention mechanism based model for tourism demand forecasting. *Information Sciences, 628*, 269–290. https://doi.org/10.1016/j.ins.2023.01.095

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials, 16*, e01059. https://doi.org/10.1016/j.cscm.2022.e01059

Exceltur. (2023). *Informe Perspectivas Turísticas N.º 83: Balance del año 2022 y expectativas para 2023*. Exceltur. Recuperado de https://www.exceltur.org/wp-content/uploads/2023/01/Informe-Perspectivas-N83-Balance-del-ano-2022-y-expectativas-para-2023.pdf

Forero-Corba, W., & Negre Bennasar, F. (2024). Técnicas y aplicaciones del Machine Learning e inteligencia artificial en educación: una revisión sistemática. *RIED-Revista*

Iberoamericana de Educación a Distancia, *27*(1), 209–253. https://doi.org/10.5944/ried.27.1.37491

Fernández Alcantud, J. I., López Belmonte, J., & Cantón Mayo, I. (2017). *Destinos turísticos inteligentes: Innovación, tecnología y sostenibilidad*. En ICE. Innovación y turismo inteligente (pp. XX-XX). Universidad de Alicante. Recuperado de https://rua.ua.es/dspace/bitstream/10045/68402/1/2017_Fernandez-Alcantud_etal_ICE.pdf

Gao, J. (2023). R-Squared (R^2): How much variation is explained? *Research Methods in Medicine & Health Sciences,* *5*(4), 104–109. https://doi.org/10.1177/26320843231186398

Hermitaño Castro, J. A. (2022). Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática. Interfases, (15), 160-178. https://doi.org/10.26439/interfases2022.n015.5898

Hussein, N., Alashqur, A., & Sowan, B. (2015). *Using the interestingness measure lift to generate association rules*. Journal of Advanced Computer Science & Technology, 4(1), 156–162. https://www.researchgate.net/publication/276102262_Using_the_interestingness_measure_lift_to_generate_association_rules

IBM. (2024, January 24). *What is a confusion matrix?* Recuperado de https://www.ibm.com/mx-es/topics/confusion-matrix

INICIO - MyStreetBook. (2024, September 12). https://mystreetbook.es

Instituto Nacional de Estadística (INE). (2017). *Encuesta de Turismo de Residentes (ETR): Cuestionario*. https://ine.es/daco/daco42/etr/etr_cuestionario.pdf

Instituto Nacional de Estadística (INE). (2017). *Encuesta de Turismo de Residentes (ETR/FAMILITUR) Guía para el tratamiento de los ficheros de microdatos*. Recuperado de https://ine.es/daco/daco42/etr/etr_aprov_ficheros.pdf

Instituto Nacional de Estadística (INE). (2023). *Cuenta Satélite del Turismo de España. Revisión estadística 2023*. Recuperado de https://ine.es/dyngs/Prensa/es/CSTE2023.htm

Instituto Nacional de Estadística (INE). (s.f.). *Concepto seleccionado: Alojamiento de turismo rural*. https://www.ine.es/DEFIne/es/concepto.htm?c=4775

Instituto Nacional de Estadística (INE). (s.f.). *Encuesta de turismo de residentes (ETR/FAMILITUR). Microdatos*. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176990&menu=resultados&idp=1254735576863

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Kumbhare, T. A., & Chobe, S. V. (2014). An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies*, *5*(1), 927-930

López de Ávila, A., & García, M. F. (n.d.). *Destinos turísticos inteligentes: Nuevas perspectivas de gestión y desarrollo*. Ministerio de Industria, Energía y Turismo. Recuperado de https://www.mintur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/395/LOPEZ%20DE%20AVILA%20y%20GARCIA.pdf

Loscertales, B. (1999). *El turismo rural como forma de desarrollo sostenible. El caso de Aragón*. Revista Geographicalia, 37, 123-138. Universidad de Zaragoza.

Martín Gil, F. (2014). *Problemas de sostenibilidad del turismo rural en España*. Recuperado de https://www.researchgate.net/profile/Fernando-Martin-5/publication/270840220_Problemas_de_sostenibilidad_del_turismo_rural_en_Espana/links/54d2126a0cf25ba0f0424b3d/Problemas-de-sostenibilidad-del-turismo-rural-en-Espana.pdf

Marzban, C. (2004). The ROC Curve and the Area under It as Performance Measures. *Weather and Forecasting*, *19*(6), 1106–1114. https://doi.org/10.1175/825.1

Ministerio de Industria, Comercio y Turismo. (2020, 6 de febrero). *2019 cierra con la creación de 93.850 empleos en el sector turístico y un incremento del 3,6%*. Turespaña. Recuperado de https://www.mincotur.gob.es/es-es/GabinetePrensa/NotasPrensa/2020/Paginas/200206Np-empleo-turismo.aspx.

Ospina-Gutiérrez, J. P., & Aristizábal, E. (2021). Aplicación de inteligencia artificial y técnicas de aprendizaje automático para la evaluación de la susceptibilidad por movimientos en masa. *Revista Mexicana de Ciencias Geológicas*, *38*(1), 43–54. https://doi.org/10.22201/cgeo.20072902e.2021.1.1605

Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., & Cortez-Pinto, E. (2020). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista Médica De Chile*, *149*(2). Recuperado a partir de https://www.revistamedicadechile.cl/index.php/rmedica/article/view/8467

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. Recuperado de https://arxiv.org/abs/2010.16061

Prajapati, P., Shah, M., & Patel, V. (2019). *Study and Analysis of Decision Tree-Based Classification Algorithms*. Retrieved from https://www.researchgate.net/profile/Purvi-

Prajapati/publication/330138092_Study_and_Analysis_of_Decision_Tree_Based_Classification_Algorithms/links/5d2c4a91458515c11c3166b3/Study-and-Analysis-of-Decision-Tree-Based-Classification-Algorithms.pdf

Raschka, S., & Mirjalili, V. (2020). *Python Machine Learning*. Packt Publishing.

Rodríguez, J., Semanjski, I., Gautama, S., Van de Weghe, N., & Ochoa, D. (2018). Unsupervised Hierarchical Clustering Approach for Tourism Market Segmentation Based on Crowdsourced Mobile Phone Data. *Sensors*, *18*(9), 2972. https://doi.org/10.3390/s18092972

Sachs, J. D., & Vernis, R. V. (2015). *La era del desarrollo sostenible* (Vol. 606). Barcelona: Deusto.

Sánchez Turcios, R. A. (2015). Prueba de Wilcoxon-Mann-Whitney: mitos y realidades. *Revista Mexicana de Endocrinología, Metabolismo y Nutrición, 2*(1), 18-21. Recuperado de https://www.endocrinologia.org.mx

Sánchez Turcios, Reinaldo Alberto. (2015). t-Student: Usos y abusos. *Revista mexicana de cardiología*, *26*(1), 59-61. Recuperado en 08 de marzo de 2025, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-21982015000100009&lng=es&tlng=es

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering & Research, 3*(4), 444-449. https://doi.org/10.23883/IJRTER.2017.3168.0UWXM

Sitarz, M. (2022). Extending F1 metric, probabilistic approach. Recuperado de https://arxiv.org/abs/2210.11997

Smart Travel News. (2023, diciembre 2). *El potencial de aplicar la IA al turismo rural: los pueblos con menos visibilidad aumentan sus visitas un 36 por ciento*. Recuperado de https://www.smarttravel.news/el-potencial-de-aplicar-la-ia-al-turismo-rural-los-pueblos-con-menos-visibilidad-aumentan-sus-visitas-un-36-por-ciento/

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms*. Advances in Neural Information Processing Systems, 25. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Statista. (2023). *Ranking de países con más llegadas de turistas extranjeros en el mundo en 2023*. Recuperado de https://es.statista.com/estadisticas/596659/ranking-de-paises-con-mas-llegadas-de-turistas-extranjeros-en-el-mundo

Statista. (2023). *Número anual de viajeros en alojamientos de turismo rural en España*. Recuperado de https://es.statista.com/estadisticas/511447/alojamientos-de-turismo-rural-numero-anual-de-viajeros-en-espana/

Unión Europea. (s.f.). *Smart and competitive rural areas*. Comisión Europea. https://ec.europa.eu/enrd/enrd-thematic-work/smart-and-competitive-rural-areas_es.html

Vergara, J. R., & Estévez, P. A. (2015). A review of feature selection methods based on mutual information. *arXiv preprint arXiv:1509.07577*.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Wang, Y., & Ni, X. S. (2019). *A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization*. arXiv:1901.08433. Recuperado de https://arxiv.org/pdf/1901.08433

Zaara, M. H. (2024). *La Inteligencia Artificial aplicada al territorio y al turismo 4.0: ¿Puede contribuir al incremento de la resiliencia?* Geo UERJ, 46, e87241. https://doi.org/10.12957/geouerj.2024.87241

Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., … Wang, Y. (2020). So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, *8*(3), 76–89. https://doi.org/10.1109/mgrs.2020.2964708

7. Declaración de uso de ChatGPT

Por la presente, yo, Ángela Vallejo Mengod, estudiante de Doble grado Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Modelo Machine Learning Aplicado al Turismo Rural", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.

2. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.

3. Interpretador de código: Para realizar análisis de datos preliminares.

4. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.

5. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

6. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

7. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

8. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: [Fecha]

Firma: _____Ángela Vallejo Mengod_____