



Facultad de Ciencias Económicas y Empresariales
ICADE

TURISMO SOSTENIBLE EN CANARIAS: MODELOS DE MACHINE LEARNING Y ANÁLISIS CUANTITATIVO

Autor: Casilda De La Nuez Sánchez
Director: Luis Ángel Calvo Pascual

MADRID | Abril 2025

RESUMEN

Este trabajo analiza el perfil del turista sostenible en Canarias y las tensiones sociales derivadas del actual modelo turístico. Se utilizaron microdatos de la Encuesta de Gasto Turístico (T2 2024) con una muestra de 8.311 turistas. A través de análisis estadísticos y modelos de Machine Learning, se identifican variables clave como la autenticidad, la tranquilidad o la seguridad. Los resultados muestran que es posible predecir comportamientos sostenibles, que no dependen solo del gasto o la edad, sino de motivaciones específicas. Además, se identifican reglas frecuentes que describen al turista sostenible como alguien que valora experiencias auténticas y un consumo más consciente. Por otro lado, el análisis de más de 4.000 comentarios de residentes en vídeos de YouTube sobre turismo en Canarias revela un creciente malestar con el turismo masivo, especialmente por su impacto en la vivienda y la calidad de vida. Se han analizado todos los comentarios disponibles para evitar sesgos. La coincidencia entre estos perfiles y las preocupaciones locales refuerza la necesidad de avanzar hacia un modelo más justo y equilibrado.

ABSTRACT

This study analyses the profile of the sustainable tourist in the Canary Islands and the social tensions linked to the current tourism model. It uses microdata from the Tourist Expenditure Survey (Q2 2024), with a sample of 8,311 tourists. Through statistical analysis and Machine Learning models, key variables are identified, such as authenticity, tranquillity, and safety. Results show that sustainable behaviours can be predicted, and they are driven more by specific motivations than by spending or age. Frequent patterns also depict sustainable tourists as those who seek authentic experiences and adopt more conscious consumption habits. Additionally, the analysis of over 4,000 resident comments on YouTube videos about tourism in the Canary Islands reveals growing discontent with mass tourism, especially due to its impact on housing and quality of life. All available comments were analysed to avoid selection bias. The overlap between tourist profiles and local concerns highlights the need for a fairer and more balanced tourism model.

Tabla de contenido

1. Introducción y Metodología.....	5
1.1 El turismo en Canarias.....	5
1.2 Objetivo	6
1.3 Metodología.....	7
2. Definición y Creación del Perfil del Turista Sostenible	9
2.1 Revisión de la literatura de un turista sostenible	9
2.2 Revisión de la metodología de Machine Learning en el turismo	10
2.3 Creación de la variable Turista Sostenible	11
2.4 Distribución de Turistas Sostenibles en el Dataset.....	11
3. Selección de Variables mediante Mutual Information.....	12
3.1 Introducción del Mutual Information	12
3.2 Fundamentos Teóricos de Mutual Information	12
3.3 Aplicación en la Selección de Variables	13
3.4 Resultados de Mutual Information	14
3.5 Conclusiones:	16
4. Análisis Estadístico Descriptivo de las Variables Seleccionadas en Mutual Information	16
4.1 Distribución de las Variables Seleccionadas	17
4.2 Interpretación de los Resultados.....	17
4.3 Distribución de las Variables Seleccionadas	19
4.3.1 Variables de Importancia en la Experiencia del Turista.....	19
4.3.2 Variables económicas	20
4.3.3 Otras Variables de Relevancia.....	20
4.4 Comparación entre Turistas Sostenibles y No Sostenibles	21
4.5 Prueba T-Student para Comparación de Medias	23
4.6 Matriz de Correlación.....	25
4.7 Conclusiones.....	26
5 Modelos de clasificación y predicción.....	27
5.1 Random forest	28
5.2 XGBoost: Fundamentos y Aplicaciones	29
5.3 Regresión logística	29
5.4 Support Vector Machines (SVM).....	30
5.5 Resultados y comparación de modelos.....	31
5.6 Discusión de Resultados	35
6. Reglas de Asociación para la Sostenibilidad Turística	35
6.1 Reglas de Asociación en el Estudio del Turismo	35

6.2 Resultados de las Reglas de Asociación.....	38
7. Análisis de la Percepción Local del Turismo (NLP).....	40
7.1 Introducción al análisis de comentarios locales	40
7.2 Extracción de bigramas frecuentes	40
7.3 Clustering con WordCloud.....	42
7.4 Resultados e interpretación de temas emergentes	44
8. Conclusiones y Recomendaciones	45
9. Bibliografía.....	49

Tabla de ilustraciones

Ilustración 1: Importancia de Variables según Mutual Information	15
Ilustración 2: Resumen estadístico de las variables de MI	18
Ilustración 3: Distribución de variables de percepción	18
Ilustración 4: Distribución de variables de percepción	19
Ilustración 5: Distribución de variables de MI	21
Ilustración 6: Comparación de variables entre Turistas Sostenibles y No Sostenibles	23
Ilustración 7: Comparación de Medias: Prueba T-Student	24
Ilustración 8: Matriz de Correlación de Variables Seleccionadas	26
Ilustración 9: Precisión y AUC-ROC antes y después de la optimización	31
Ilustración 10: Interpretabilidad de Random Forest	33
Ilustración 11: Interpretabilidad de XGBoost	33
Ilustración 12: Interpretabilidad de Regresión logística	34
Ilustración 13: Interpretabilidad de SVM	34
Ilustración 14: Representación de las reglas de asociación que predicen turistas sostenibles.	38
Ilustración 15: Bigramas más frecuentes en comentarios locales	41
Ilustración 16: Visualización de las nubes de palabras de los Clusters	43
Ilustración 17: Visualización de las nubes de palabras de los Clusters	43

1. Introducción y Metodología

1.1 El turismo en Canarias

El turismo desempeña un papel crucial en la economía de Canarias, representando el 35,5% del PIB regional en 2022, con un crecimiento del 9,3% respecto a los niveles pre-pandemia (Exceltur & Gobierno de Canarias, 2023). En términos de empleo, el sector turístico generó 344.358 puestos de trabajo, equivalentes al 39,7% del empleo total en Canarias, lo que demuestra su gran importancia en la economía del archipiélago.

Históricamente, Canarias ha dependido de un modelo de turismo de masas basado en el sol y playa. Sin embargo, desde la década de 2000, se ha promovido una diversificación de la oferta turística, con un mayor énfasis en productos de mayor valor añadido como el turismo rural y el ecoturismo (Parra López & Calero García, 2006). No obstante, el archipiélago sigue enfrentando desafíos como la presión sobre los recursos naturales, la congestión en zonas turísticas y la necesidad de una mayor regulación en el acceso a espacios protegidos.

Diversos estudios han alertado de los impactos negativos derivados del modelo de turismo masivo. El crecimiento explosivo del número de turistas —de menos de 100.000 en 1960 a más de 13 millones en 2015— ha provocado un aumento del 120% en la población residente en islas como Lanzarote y Fuerteventura, muchas de ellas declaradas Reserva de la Biosfera (Hernández Luis et al., 2017). Esto ha supuesto una intensificación del consumo de recursos como el agua y la energía, la generación masiva de residuos, la ocupación del territorio y la alteración de paisajes protegidos.

A esta situación se suma una creciente exposición a fenómenos climáticos extremos vinculados al cambio climático, como lluvias torrenciales e inundaciones. En destinos turísticos del sur de Gran Canaria, como San Agustín, la urbanización sobre laderas escarpadas ha generado riesgos de deslizamientos, cortes de carreteras y aislamiento de zonas turísticas en episodios de lluvia intensa, poniendo en evidencia la fragilidad del modelo urbanístico ligado al turismo de masas (Ley Bosch et al., 2024).

Además, el turismo de masas ha generado un contexto de vulnerabilidad socio-territorial en las islas, con problemas como el aumento de la pobreza en algunas zonas turísticas, la

especulación del suelo y el encarecimiento del acceso a la vivienda para los residentes (Hernández Luis et al., 2017; Bianchi, 2004).

En este contexto, es imprescindible repensar el modelo turístico de Canarias hacia uno más equilibrado y resiliente, que no comprometa los recursos naturales ni el bienestar de la población residente. Esto implica integrar criterios de sostenibilidad, adaptar la planificación urbana a los riesgos derivados del cambio climático y promover formas de turismo con menor huella ecológica y mayor beneficio social.

1.2 Objetivo

El objetivo principal de este trabajo es analizar el perfil del turista sostenible durante el segundo trimestre de 2024 en Canarias y predecir su comportamiento mediante modelos de Machine Learning. Para ello, en primer lugar, se define el concepto de ‘Turista Sostenible’ a partir de la literatura académica y de las variables disponibles en el conjunto de datos, lo que permite construir una variable binaria representativa para su análisis.

A partir de esta base, se plantean los siguientes objetivos específicos:

1. Seleccionar las variables más relevantes para predecir la sostenibilidad turística utilizando el método de Mutual Information.
2. Desarrollar modelos de clasificación que permitan predecir si un turista será sostenible o no, evaluando su rendimiento mediante métricas como precisión y AUC.
3. Interpretar los modelos entrenados mediante técnicas como SHAP, con el fin de analizar la contribución de cada variable en la predicción del perfil sostenible.
4. Desarrollar un sistema de reglas comprensibles que permitan identificar de forma explícita qué combinaciones de comportamientos caracterizan a un turista sostenible.
5. Analizar la percepción local del turismo mediante técnicas de procesamiento de lenguaje natural (NLP), a partir de más de 4.000 comentarios de residentes en Canarias.

Finalmente, a partir de los resultados obtenidos, se plantean conclusiones y recomendaciones que pueden servir de apoyo en la toma de decisiones en materia de turismo sostenible.

1.3 Metodología

La metodología seguida combina un enfoque cuantitativo basado en el análisis de microdatos turísticos y un enfoque cualitativo orientado al análisis de la percepción social mediante comentarios públicos.

En primer lugar, se trabajó con una base de datos oficial del Gobierno de Canarias correspondiente al segundo trimestre de 2024, que incluye 8.311 registros y 99 variables. Estas variables abarcan desde características demográficas, motivaciones de viaje y hábitos de gasto, hasta aspectos relacionados con la sostenibilidad (uso de transporte público, participación en actividades culturales, valoración del entorno) y la experiencia turística (satisfacción, intención de volver, recomendación del destino).

Se llevó a cabo una limpieza y tratamiento de los datos, eliminando aquellas variables que estaban mayoritariamente vacías. Para los valores nulos restantes, se imputó la media del resto de observaciones, con el objetivo de trabajar con una base de datos completa y evitar problemas de procesamiento. Además, las variables categóricas de tipo ordinal, que son aquellas que recogen respuestas en rangos como *nada*, *poco*, *bastante* o *mucho*, se transformaron a formato numérico, asignándoles valores del 0 al 3 respectivamente, para facilitar su análisis.

Tras esta fase, se definió la variable objetivo ‘Turista Sostenible’, asignando un valor de 1 a aquellos visitantes que cumplían al menos uno de los siguientes criterios: uso de transporte público, participación en actividades culturales o alta valoración del entorno. A partir de ahí, se aplicaron técnicas de selección de variables mediante Mutual Information, y se entrenaron cuatro modelos de clasificación: Random Forest, XGBoost, Regresión Logística y SVM. Los modelos se evaluaron mediante métricas de precisión y AUC-ROC, y se interpretaron con la técnica SHAP (Shapley Additive Explanations).

Estos algoritmos de Machine Learning, especialmente Random Forest y XGBoost, se consideran modelos de tipo “caja negra”, ya que su lógica interna es compleja y poco transparente. Para solventar esta limitación y comprender cómo influyen las variables seleccionadas en las predicciones, se aplicó SHAP, una técnica basada en la teoría de juegos que descompone cada predicción en la contribución individual de cada variable. Como destaca Li (2022), SHAP permite abrir la caja negra de los modelos

proporcionando explicaciones locales y coherentes del comportamiento del modelo, incluso en presencia de relaciones no lineales o interacciones complejas. Esta capacidad resulta especialmente útil en contextos como el turismo sostenible, donde se busca no solo predecir, sino también comprender qué factores están impulsando el comportamiento sostenible de los turistas.

Además, como complemento a los modelos predictivos, se aplicaron técnicas de reglas de asociación para identificar patrones de comportamiento sostenibles recurrentes entre los turistas. Estas reglas permiten generar conclusiones fácilmente interpretables que pueden ser de utilidad para diseñar estrategias de promoción turística alineadas con la sostenibilidad.

Paralelamente, se llevó a cabo un análisis de procesamiento de lenguaje natural (NLP) a partir de más de 4.000 comentarios de residentes extraídos de vídeos de YouTube relacionados con el turismo en Canarias. Dado que muchos de estos comentarios contenían ruido, spam o no eran relevantes para el análisis, se aplicó un filtrado riguroso para conservar únicamente aquellos con contenido potencialmente útil.

Se seleccionaron comentarios que cumplían dos condiciones: (1) contener al menos 8 palabras, y (2) incluir alguna palabra clave relacionada con el turismo o la sostenibilidad, como “turismo”, “canarias”, “alquiler”, “vivienda”, “vacacional”, “residente”, “inmigración”, “gobierno”, entre otras. Este proceso permitió construir un subconjunto más representativo del discurso social vinculado al turismo, con un total de 4.000 comentarios.

Posteriormente, se aplicó una limpieza textual que eliminó signos de puntuación, emojis y palabras vacías (stopwords), manteniendo únicamente el contenido léxico significativo. Los comentarios fueron tokenizados para permitir su análisis posterior.

En conjunto, la metodología adoptada proporciona una visión amplia y robusta del turismo en Canarias, integrando datos cuantitativos sobre los turistas con percepciones cualitativas de los residentes.

2. Definición y Creación del Perfil del Turista Sostenible

2.1 Revisión de la literatura de un turista sostenible

El concepto de turismo sostenible ha ganado relevancia en la literatura académica como una respuesta a los impactos negativos del turismo en el medio ambiente y las comunidades locales. A medida que aumenta la preocupación por el cambio climático, la sobreexplotación de los destinos turísticos y la degradación cultural surge la necesidad de identificar y fomentar un perfil de turista que contribuya a la sostenibilidad. Sin embargo, la definición de turista sostenible no es uniforme y varía según los enfoques teóricos y metodológicos empleados en la investigación académica.

Este capítulo tiene como objetivo establecer una definición operativa del turista sostenible basada en la revisión de la literatura científica. Para ello, se analizan diferentes perspectivas y segmentaciones existentes en la investigación previa, con el fin de adaptar una definición que se ajuste a las variables disponibles en el análisis de datos.

La literatura académica ha abordado el concepto de turista sostenible desde distintas perspectivas, integrando elementos de comportamiento, intención y consumo responsable. Juvan & Dolnicar (2016) definen al turista sostenible como aquel que minimiza su impacto negativo en el medio ambiente y, en algunos casos, toma medidas activas para protegerlo, destacando que la intención y el comportamiento observado pueden diferir debido a sesgos en la autoevaluación de los turistas.

Por otro lado, Budeanu (2007) enfatiza que la sostenibilidad en el turismo no solo se limita a la elección de productos responsables, sino que involucra un conjunto de decisiones a lo largo del viaje, incluyendo transporte, alojamiento, consumo local y comportamiento en el destino. La autora argumenta que, a pesar de una alta conciencia ambiental declarada por los turistas, sólo una fracción de ellos actúa de manera coherente con estos valores.

Desde una perspectiva segmentada, Gupta & Chopra (n.d.) identifican tres tipos de turistas sostenibles en función de su comportamiento proambiental:

1. Eco-visionarios: aquellos que activamente participan en la conservación del medio ambiente y apoyan iniciativas ecológicas.

2. Conservadores del medio ambiente: quienes buscan reducir su huella ecológica, pero sin modificar sustancialmente su experiencia turística.
3. Minimizadores del desperdicio: turistas que adoptan prácticas de sostenibilidad principalmente relacionadas con la reducción de residuos y el consumo eficiente de recursos.

A nivel teórico, el modelo de comportamiento ambientalmente significativo establece que el turista sostenible es aquel que posee normas personales claras hacia la sostenibilidad, toma decisiones basadas en la reducción del impacto ambiental y se siente responsable del entorno que visita (Gupta & Chopra, n.d.) Además, He et al. (2018) refuerzan la importancia de los incentivos externos, como regulaciones gubernamentales y certificaciones ecológicas, en la promoción del turismo sostenible.

Por tanto, en este estudio se adoptará la siguiente definición: Un turista sostenible es aquel que, de manera intencionada, busca minimizar su impacto ambiental y social en el destino, eligiendo alojamientos sostenibles, utilizando medios de transporte con menor huella ecológica, consumiendo productos locales y adoptando comportamientos responsables que contribuyan a la conservación del entorno y la cultura local.

2.2 Revisión de la metodología de Machine Learning en el turismo

En los últimos años, el uso de técnicas de Machine Learning (ML) ha experimentado un notable auge en el ámbito del turismo. Según la revisión sistemática realizada por Núñez et al. (2024), los algoritmos de aprendizaje supervisado —especialmente redes neuronales (ANN, CNN, LSTM), árboles de decisión y SVM— son los más utilizados, representando más del 75 % de los casos analizados. Estas metodologías se han aplicado con éxito en la predicción de demanda turística, clasificación de perfiles, análisis de satisfacción o segmentación de mercado. Otros estudios complementarios también han empleado modelos como Random Forest, Naive Bayes, XGBoost o regresión lineal para predecir gasto turístico, intención de visita o impacto del overtourism, con especial interés en la sostenibilidad (Aparicio Castillo et al., 2022; Madzík et al., 2023).

Asimismo, se observa una tendencia creciente hacia modelos interpretables, sistemas híbridos y técnicas aplicadas a datos abiertos (redes sociales, encuestas, plataformas de reservas). No obstante, la mayoría de los estudios se centran en flujos agregados de turistas o en variables de negocio, y no en el análisis del comportamiento individual desde

un enfoque de sostenibilidad. Por tanto, este trabajo se sitúa en un enfoque novedoso al definir explícitamente la figura del “turista sostenible” y combinar técnicas de clasificación, interpretación (SHAP) y reglas de asociación para identificar patrones de conducta responsables.

2.3 Creación de la variable Turista Sostenible

Para aplicar esta definición de turista sostenible, se han seleccionado variables que reflejan comportamientos alineados con la sostenibilidad, asegurando que cada turista pueda clasificarse en función de su nivel de compromiso con estas prácticas. Se han considerado las siguientes dimensiones:

- Uso del transporte público como indicador de reducción de emisiones contaminantes.
- Valoración del entorno ambiental, reflejando la conciencia ecológica del turista.
- Participación en actividades culturales, indicando un interés en la conservación y respeto por la cultura local.

A partir de estos criterios, se ha definido un turista sostenible como aquel que cumple al menos una de las siguientes condiciones:

- Uso de transporte público ($\text{Desglose_Transp_Publico} > 0$).
- Alta valoración del entorno ambiental ($\text{Importancia_Entorno_Ambiental} \geq 3$).
- Participación en actividades culturales ($\text{Desglose_Activ_Cultural} > 0$).

Si un turista cumple al menos una de estas condiciones, se le asigna el valor 1 (Sostenible); de lo contrario, se le clasifica como 0 (No sostenible).

2.4 Distribución de Turistas Sostenibles en el Dataset

Tras la aplicación de los criterios de clasificación y la limpieza de datos, se obtuvo la siguiente distribución de turistas sostenibles y no sostenibles en el dataset:

- Número total de turistas analizados: 8.311.
- Turistas sostenibles: 3.606 (43,4% del total).
- Turistas no sostenibles: 4.705 (56,6% del total).

Estos valores reflejan que una parte significativa de los turistas en el dataset exhibe al menos un comportamiento alineado con la sostenibilidad. Sin embargo, la mayoría (más del 56%) aún no cumple con ninguno de los criterios establecidos, lo que sugiere oportunidades para fomentar prácticas turísticas más responsables en Canarias.

El proceso de construcción de la variable ‘Turista Sostenible’ permite categorizar de manera objetiva a los turistas en función de comportamientos medibles y alineados con la sostenibilidad. La selección de variables responde a criterios teóricos y empíricos que permiten evaluar el compromiso de los visitantes con prácticas responsables en su estancia en Canarias. Este criterio de clasificación servirá como base para el posterior análisis de factores determinantes de la sostenibilidad turística y la predicción de perfiles de turistas sostenibles.

3. Selección de Variables mediante Mutual Information

3.1 Introducción del Mutual Information

En el ámbito del aprendizaje automático y la selección de características, el Mutual Information (MI) se ha consolidado como una técnica fundamental para evaluar la dependencia entre variables. Su principal ventaja radica en su capacidad para capturar tanto relaciones lineales como no lineales entre características y la variable objetivo (Bennasar et al., 2015). En este capítulo, se describe la aplicación de MI en el contexto del turismo sostenible en Canarias, seleccionando las variables más relevantes para predecir si un turista puede considerarse sostenible.

3.2 Fundamentos Teóricos de Mutual Information

El Mutual Information (MI) mide la cantidad de información compartida entre dos variables aleatorias XX y YY . Se basa en la entropía, que representa la incertidumbre de una variable. Matemáticamente, MI se expresa como:

$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$ donde:

- $H(X)$ es la entropía de XX ,
- $H(Y)$ es la entropía de YY ,
- $H(X|Y)$ es la entropía condicional de XX dado YY .

Si XX y YY son completamente independientes, entonces $I(X;Y)=0$. En cambio, si existe una fuerte dependencia, MI toma valores más altos (Kraskov et al., 2004)

El uso de MI en selección de características es especialmente útil en problemas donde las variables pueden estar correlacionadas de manera no lineal con la variable objetivo, lo que la hace superior a métricas como la correlación de Pearson en ciertos casos (Bennasar et al., 2015).

3.3 Aplicación en la Selección de Variables

Para la selección de características en nuestro análisis, se utilizó Mutual Information (MI) para medir la relevancia de cada variable con respecto a la variable objetivo "Turista Sostenible". Sin embargo, con el fin de evitar sobreajustar el modelo y garantizar que las variables seleccionadas reflejan relaciones genuinas con la sostenibilidad del turista, se decidió excluir del análisis aquellas variables que ya formaban parte de la definición de sostenibilidad. Específicamente, se eliminaron del cálculo las variables relacionadas con el uso de transporte público, la valoración del entorno ambiental y la participación en actividades culturales, así como cualquier otra variable que directamente influye en la clasificación de los turistas sostenibles. Esta decisión se tomó para evitar que el modelo identificara como relevantes características que, por definición, ya determinan la variable objetivo.

Para calcular la Mutual Information, se utiliza la función `mutual_info_classif` de la librería Scikit-learn, la cual permite evaluar la dependencia entre variables predictoras y una variable objetivo en problemas de clasificación. Sin embargo, el cálculo de MI en una única iteración puede verse afectado por ruido en los datos, especialmente cuando las variables presentan distribuciones desbalanceadas o relaciones no lineales. Para mitigar este problema, se decidió realizar 500 repeticiones del cálculo de MI , siguiendo un enfoque de promedio estadístico, tal como se recomienda en la literatura para mejorar la robustez de la selección de variables. Este enfoque permite reducir la variabilidad en la estimación de MI , obteniendo un ranking más estable y confiable de las variables más relevantes.

Las principales librerías utilizadas en la implementación fueron:

- Scikit-learn (mutual_info_classif): Para calcular la información mutua entre cada variable predictora y la variable objetivo.
- NumPy: Para realizar operaciones matemáticas y almacenar los resultados en múltiples iteraciones.
- Pandas: Para manipulación y almacenamiento de los datos en estructuras tipo DataFrame.
- Matplotlib y Seaborn: Para la visualización de los resultados mediante gráficos de barras.

3.4 Resultados de Mutual Information

Tras aplicar esta metodología con 500 repeticiones, las variables con mayor Mutual Information (MI) en relación con la sostenibilidad turística fueron:

1. Importancia_Autenticidad (0,050)
2. Importancia_Trquilidad (0,040)
3. Importancia_Gastronomía (0,035)
4. Importancia_Seguridad (0,032)
5. Importancia_Exotismo (0,030)
6. Coste_Vuelos_Euros (0,028)
7. Coste_Paquete_Euros (0,026)
8. Coste_Aloj_Euros (0,025)
9. Importancia_Viaje_Sencillo (0,023)
10. Importancia_Precio (0,021)

Ilustración 1: Importancia de Variables según Mutual Information

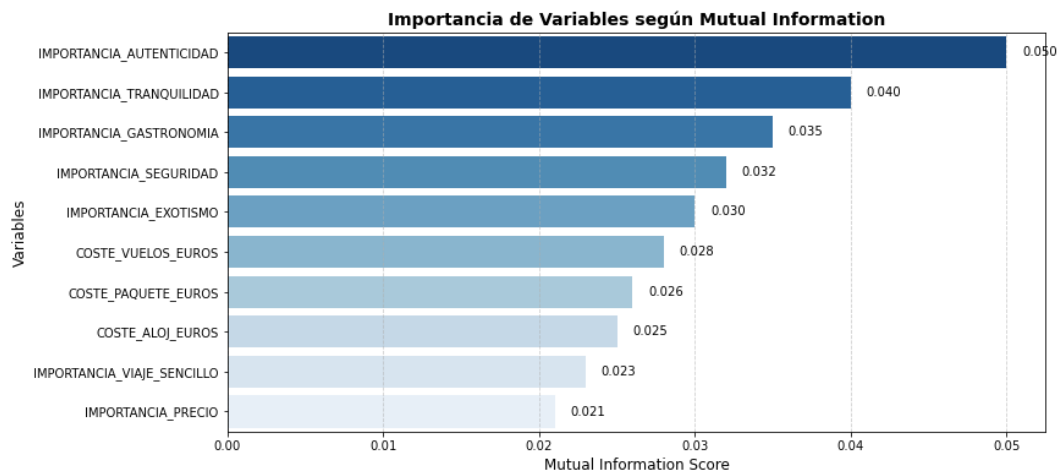


Gráfico de barras que muestra la relevancia de cada variable predictora en la clasificación de turistas sostenibles. El valor se obtiene a partir del promedio de 500 repeticiones del cálculo de mutual information con la función `mutual_info_classif` de `scikit-learn`, lo que permite obtener un ranking más estable. Elaboración propia en Python con `scikit-learn`, `numpy`, `pandas`, `matplotlib` y `seaborn`.

Los resultados muestran que las variables `Importancia_Autenticidad` e `Importancia-Tranquilidad` son las más informativas para determinar si un turista es sostenible. La alta puntuación de estas variables sugiere que los turistas que buscan autenticidad en su experiencia de viaje y valoran la tranquilidad del destino tienen una mayor propensión a comportamientos sostenibles. Esto puede estar relacionado con una preferencia por destinos menos masificados y una mayor consideración por el impacto ambiental de su viaje.

Además, factores económicos como `Coste_Vuelos_Euros`, `Coste_Aloj_Euros` y `Coste_Paquete_Euros` aparecen como elementos clave en la clasificación de sostenibilidad. La relación entre el gasto turístico y la sostenibilidad podría estar vinculada a la elección de opciones más responsables, como alojamientos certificados por prácticas ecológicas o paquetes turísticos que priorizan experiencias sostenibles y con un menor impacto ambiental.

Por otro lado, variables como `Importancia_Gastronomía` e `Importancia_Exotismo` también han mostrado una relación con la sostenibilidad. Esto sugiere que los turistas que valoran la gastronomía local y buscan experiencias exóticas pueden estar más dispuestos

a participar en prácticas de consumo responsable, como la preferencia por productos locales o actividades culturales que fomenten el desarrollo comunitario.

Asimismo, la variable *Importancia_Viaje_Sencillo* indica que aquellos turistas que buscan una experiencia de viaje sin complicaciones pueden tener un perfil diferenciado en cuanto a su sostenibilidad, ya sea evitando el turismo de lujo o prefiriendo opciones más accesibles y respetuosas con el entorno.

Por último, la variable *Importancia_Precio* han mostrado menor influencia en la sostenibilidad del turista, lo que sugiere que la sensibilidad al precio no son factores tan determinantes en este contexto.

3.5 Conclusiones:

La percepción de autenticidad y tranquilidad en el destino son aspectos clave en la sostenibilidad del turista, indicando que aquellos que buscan experiencias transparentes y relajadas podrían estar más predispuestos a adoptar prácticas sostenibles. Además, el gasto turístico, especialmente en alojamiento y paquetes turísticos, está relacionado con la sostenibilidad, lo que sugiere que ciertos patrones de consumo pueden estar alineados con prácticas más responsables.

El interés por la gastronomía y la búsqueda de experiencias exóticas pueden estar asociados a una mayor predisposición hacia el consumo local y la valoración de la cultura del destino.

Mutual Information ha permitido identificar qué variables influyen más en la sostenibilidad turística, facilitando el desarrollo de estrategias basadas en datos para fomentar el turismo responsable.

4. Análisis Estadístico Descriptivo de las Variables Seleccionadas en Mutual Information

El análisis estadístico descriptivo permite examinar la distribución de las variables más influyentes en la clasificación del Turista Sostenible. La selección de estas variables se realizó mediante Mutual Information, lo que permitió identificar aquellas que más información aportan a la distinción entre turistas sostenibles y no sostenibles.

4.1 Distribución de las Variables Seleccionadas

El análisis de medidas estadísticas como la media y la desviación estándar permite comprender la tendencia central y la dispersión de las variables seleccionadas en el estudio. Adicionalmente, el análisis de dispersión a través de diagramas de caja permite identificar la presencia de valores atípicos y la variabilidad de los datos.

A continuación, se presentan las medidas estadísticas clave de las variables analizadas:

4.2 Interpretación de los Resultados

- Variables de importancia subjetiva (como autenticidad, tranquilidad, gastronomía y seguridad) tienen medias cercanas a 2 en una escala de 0 a 3. Esto indica que la mayoría de los turistas consideran estos aspectos como moderadamente importantes.
- La desviación estándar de estas variables es relativamente baja ($\sim 0,75-0,95$), lo que sugiere que la mayoría de los turistas tienen percepciones similares sobre estos aspectos.
- Variables económicas (coste de vuelos, alojamiento y paquete turístico) presentan medias bajas en comparación con sus desviaciones estándar. Esto indica que, aunque la mayoría de los turistas gastan cantidades moderadas, existen algunos valores extremadamente altos que elevan la dispersión de los datos.
- Los diagramas de caja revelan una fuerte asimetría en las variables económicas, con múltiples valores atípicos en Coste_Vuelos_Euros, Coste_Aloj_Euros y Coste_Paquete_Euros. Estos valores reflejan la existencia de turistas con gastos muy elevados en comparación con la media del grupo.
- El coste de los paquetes turísticos y alojamientos muestra la mayor dispersión (2009,11 y 1367,08 , respectivamente), lo que evidencia una gran diferencia entre los turistas que optan por opciones económicas y aquellos que realizan un mayor gasto en estos servicios. El análisis de dispersión y medidas estadísticas permite identificar patrones clave en la muestra y ayuda a contextualizar la información sobre el comportamiento de los turistas en relación con su sostenibilidad.

Ilustración 2: Resumen estadístico de las variables de MI

	Media	Desviación Estándar	Mediana
Importancia_Autenticidad	1,82	0,95	2
Importancia_Tranquilidad	2,38	0,75	3
Importancia_Gastronomía	1,85	0,97	2
Importancia_Exotismo	1,28	1	1
Importancia_Seguridad	2,38	0,8	3
Coste_Vuelos_Euros	766,04	860,05	570,53
Coste_Aloj_Euros	1184,58	1367,08	899,99
Importancia_Viaje_Sencillo	2,08	0,91	2
Importancia_Precio	2,16	0,83	2
Coste_Paquetes_Euros	1343,16	2009,11	577,22

Resumen de estadísticos descriptivos (media, desviación estándar y mediana) para las variables perceptivas y de coste analizadas. Elaboración propia en Python con pandas, y exportación de resultados a Excel para su presentación.

Ilustración 3: Distribución de variables de percepción

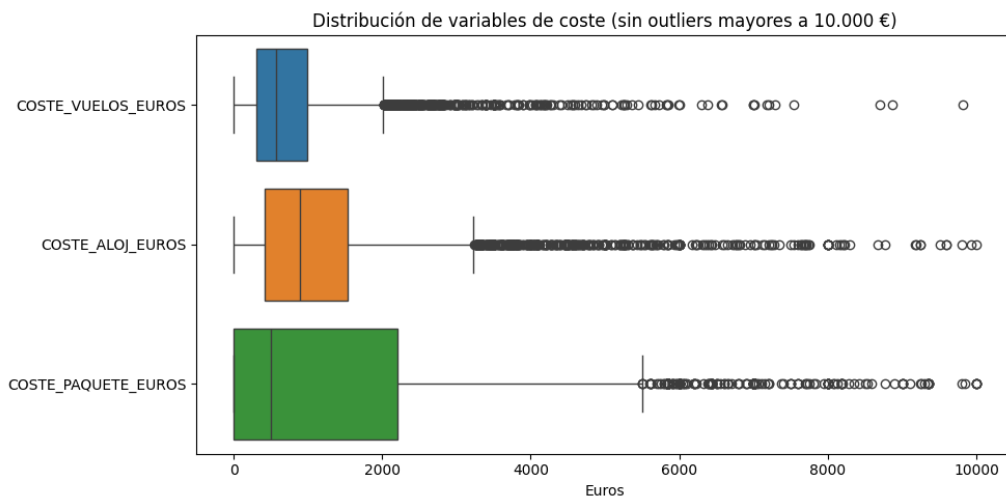
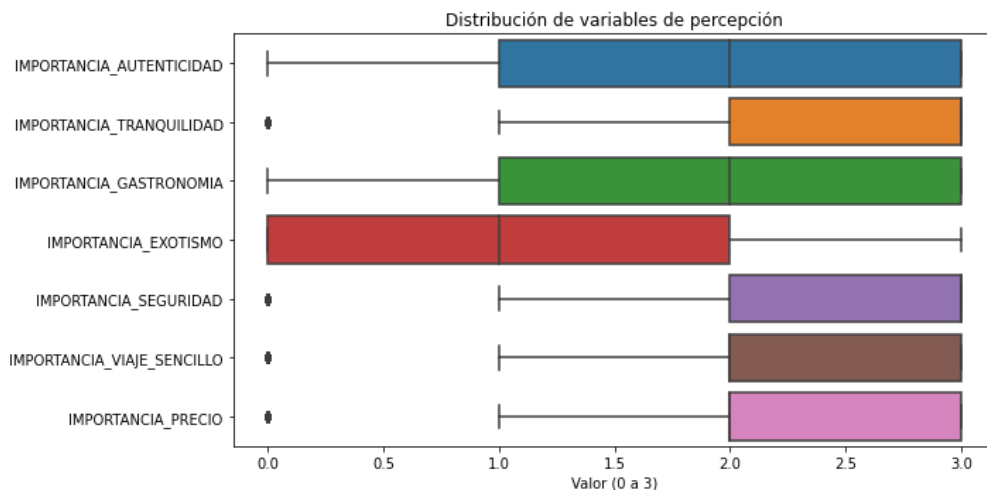


Ilustración 4: Distribución de variables de percepción



Gráficos exploratorios que muestran la distribución de las variables perceptivas (Tabla 3) y las variables de coste (Tabla 4) incluidas en el análisis de sostenibilidad turística. En la Ilustración tres se han eliminado los outliers poniendo como tope 10.000 euros. Se observan diferencias en la escala y dispersión entre ambos grupos, lo que justifica su tratamiento diferenciado en etapas posteriores. Elaboración propia en Python utilizando pandas y matplotlib

4.3 Distribución de las Variables Seleccionadas

El análisis de la distribución de las variables seleccionadas permite visualizar la frecuencia de los valores observados en la muestra, proporcionando información clave sobre patrones de comportamiento y tendencias dentro del grupo de turistas analizados.

4.3.1 Variables de Importancia en la Experiencia del Turista

Las variables Importancia_Autenticidad, Importancia-Tranquilidad, Importancia_Gastronomía, Importancia_Exotismo e Importancia_Seguridad presentan distribuciones multimodales con concentraciones de valores en distintos puntos de la escala de 0 a 3. Esto indica la existencia de diferentes segmentos de turistas con valoraciones diferenciadas sobre estos aspectos.

- Importancia de la autenticidad y seguridad: Se observa que la mayoría de los turistas consideran estos aspectos como "bastante" o "mucho importancia", reflejando una tendencia hacia experiencias auténticas y entornos seguros.
- Importancia de la gastronomía y exotismo: Muestran una distribución más equilibrada, lo que indica que algunos turistas priorizan estos elementos, mientras que otros les dan menos relevancia.

- Importancia de la tranquilidad: Se distribuye de manera similar a la autenticidad y seguridad, con una fuerte concentración en los valores más altos, lo que sugiere que muchos turistas buscan destinos relajantes.

4.3.2 Variables económicas

Las variables económicas Coste_Vuelos_Euros, Coste_Aloj_Euros y Coste_Paquete_Euros presentan distribuciones sesgadas a la derecha, con una alta concentración de valores bajos y una cola larga de valores más altos.

- Coste de los vuelos: se observa una asimetría positiva que indica que la mayoría de los turistas gastan cantidades relativamente bajas en vuelos, pero existen valores extremos que indican turistas con gastos significativamente más altos.
- Coste del alojamiento: Similar al caso de los vuelos, hay una mayoría de turistas con un gasto moderado, pero también se presentan valores elevados que corresponden a turistas que optan por opciones de alojamiento más costosas.
- Coste de los paquetes turísticos: Es la variable con la mayor dispersión, con valores que van desde opciones económicas hasta paquetes de lujo con un costo elevado.

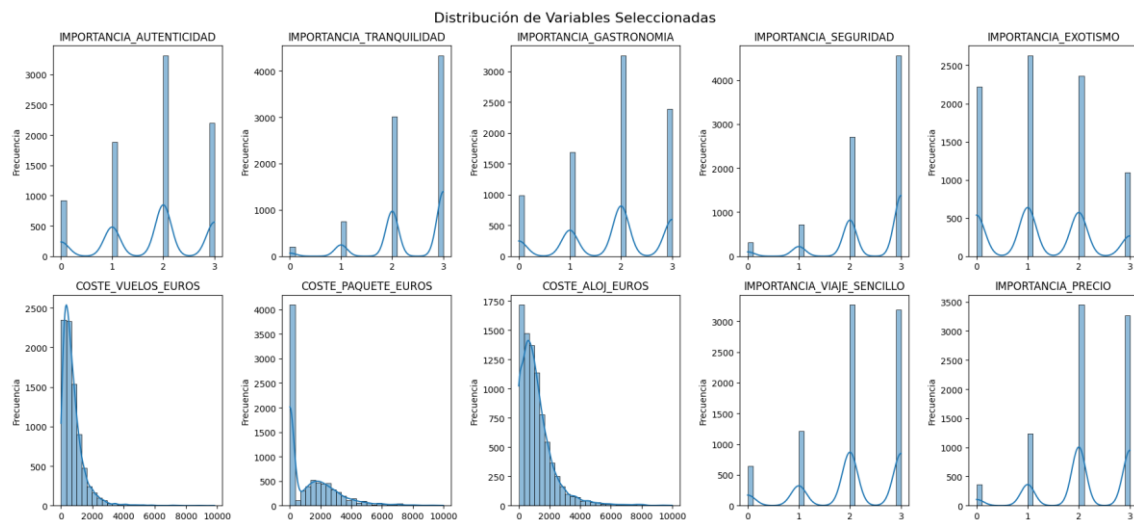
4.3.3 Otras Variables de Relevancia

- Importancia del viaje sencillo: Su distribución muestra una tendencia hacia valores altos, lo que indica que muchos turistas priorizan la facilidad en la planificación y ejecución de su viaje.
- Importancia del precio: Presenta una estructura similar a otras variables de importancia subjetiva, sugiriendo que el costo es un factor clave en la toma de decisiones de muchos turistas.

Interpretación General

El análisis de distribución refleja que los turistas valoran principalmente la autenticidad, seguridad y tranquilidad en sus experiencias, y que los factores económicos presentan una alta variabilidad, con algunos turistas dispuestos a gastar considerablemente más que la media. Estas diferencias en la percepción y gasto pueden influir en la sostenibilidad del turismo y la adopción de prácticas responsables.

Ilustración 5: Distribución de variables de MI



Histogramas que muestran la frecuencia de valores para cada una de las variables seleccionadas en el análisis de sostenibilidad turística. Permiten observar su comportamiento general y detectar posibles sesgos o asimetrías. Elaboración propia en Python con seaborn y matplotlib.

4.4 Comparación entre Turistas Sostenibles y No Sostenibles

El análisis de los boxplots permite identificar diferencias relevantes en las preferencias y comportamientos de los turistas sostenibles frente a los no sostenibles. En términos generales, se pueden distinguir varios patrones clave en relación con la importancia otorgada a ciertos aspectos del viaje, así como en los niveles de gasto en transporte y alojamiento.

Preferencias y Valores del Viaje

Los turistas sostenibles tienden a valorar más la autenticidad y la tranquilidad en su experiencia turística. Esto sugiere que buscan una conexión más genuina con el destino y prefieren ambientes menos masificados. La importancia de la gastronomía también es ligeramente mayor en este grupo, lo que refleja un posible interés en la cultura culinaria local.

En cuanto a la seguridad, ambos grupos la consideran relevante, sin mostrar diferencias significativas. Sin embargo, se observa una diferencia notable en la percepción del exotismo: los turistas no sostenibles le otorgan mayor importancia, mientras que los sostenibles tienden a valorarlo menos. Esto sugiere que la sostenibilidad no se asocia tanto

con la búsqueda de experiencias llamativas o fuera de lo común, sino más bien con vivencias auténticas, culturales y respetuosas con el entorno.

Patrones de Gasto

Los turistas sostenibles muestran una ligera tendencia a gastar más en vuelos y alojamiento en comparación con los no sostenibles. No obstante, estas variables presentan una gran dispersión, con algunos turistas que realizan un gasto considerablemente elevado. Esto sugiere que la sostenibilidad no está exclusivamente ligada a un perfil económico concreto, sino que puede encontrarse en distintos niveles de gasto. Por otro lado, el coste de los paquetes turísticos muestra una variabilidad similar en ambos grupos, sin una tendencia clara que permita diferenciarlos.

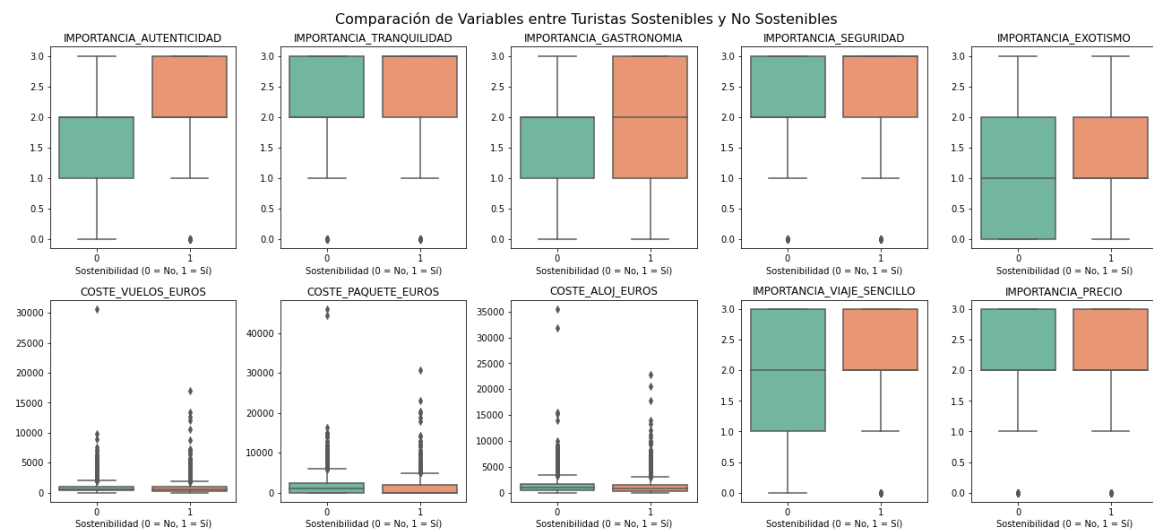
Simplicidad y Organización del Viaje

Un aspecto relevante es que los turistas sostenibles parecen dar mayor importancia a la sencillez en la planificación del viaje. Esto podría estar relacionado con una preferencia por experiencias bien organizadas, con menor impacto ambiental y mayor facilidad de movilidad dentro del destino. En contraste, la preocupación por el precio no muestra diferencias significativas entre ambos grupos, lo que indica que el coste del viaje no es un factor determinante en la adopción de comportamientos turísticos sostenibles.

Conclusiones

Los resultados indican que los turistas sostenibles priorizan experiencias auténticas y tranquilas, lo que puede estar relacionado con una mayor conciencia sobre el impacto de su viaje en la cultura local y el entorno. Además, aunque tienden a gastar más en transporte y alojamiento, este comportamiento no es uniforme, ya que existe una gran variabilidad en los valores. Finalmente, la preferencia por un viaje sencillo sugiere que la sostenibilidad podría estar relacionada con una planificación más eficiente y menos dependiente de actividades turísticas masificadas.

Ilustración 6: Comparación de variables entre Turistas Sostenibles y No Sostenibles



Boxplots comparativos para evaluar diferencias en la distribución de variables según si los turistas fueron clasificados como sostenibles ($SOSTENIBLE = 1$) o no ($SOSTENIBLE = 0$). Elaboración propia en Python utilizando seaborn y matplotlib.

4.5 Prueba T-Student para Comparación de Medias

Se aplicó la prueba T-Student para evaluar si las diferencias observadas entre los turistas sostenibles y no sostenibles son estadísticamente significativas. Se encontraron valores de $p < 0.05$ en la mayoría de las variables analizadas, lo que indica que las diferencias en sus medias no son producto del azar.

Resultados de la Prueba T-Student

Los resultados de la prueba T-Student muestran diferencias significativas en las medias de varias variables entre turistas sostenibles y no sostenibles. La dirección y magnitud del estadístico t permiten identificar qué aspectos son más valorados por cada grupo.

Variables con valores negativos en el estadístico t:

Las variables importancia de la autenticidad, tranquilidad, gastronomía, seguridad, viaje sencillo y precio presentan valores negativos, lo que indica que los turistas sostenibles otorgan mayor importancia a estos aspectos en comparación con los no sostenibles. Esto sugiere que los turistas con mentalidad sostenible tienden a buscar experiencias más auténticas, relajadas, culturalmente enriquecedoras y valoran también la simplicidad y

accesibilidad del viaje. Aunque la diferencia en la importancia del precio es más moderada, se aprecia una mayor sensibilidad relativa en los turistas sostenibles.

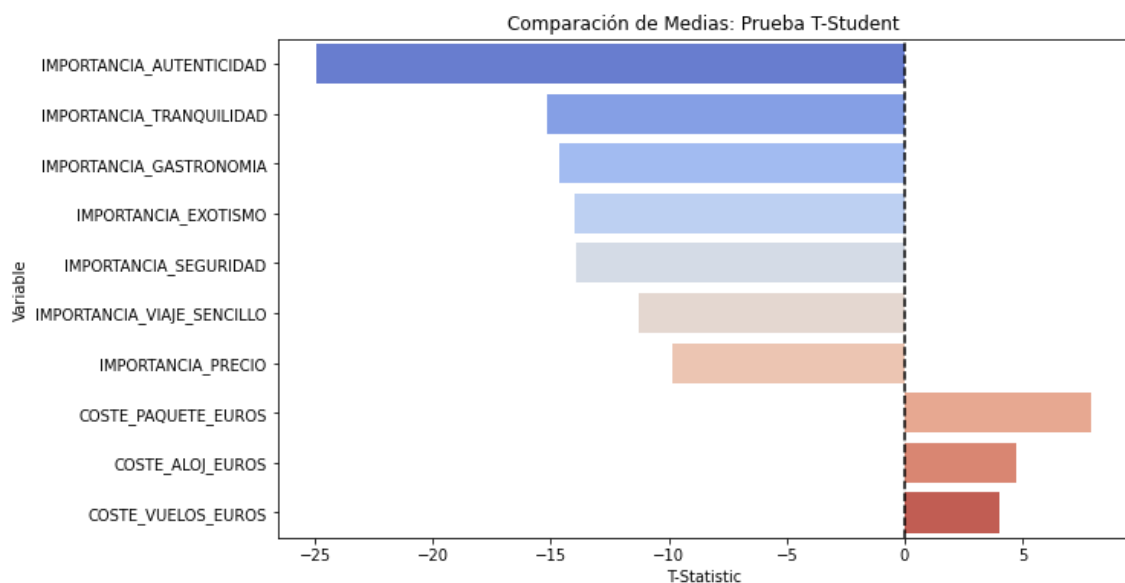
En el caso de la variable exotismo, aunque también presenta un valor negativo del estadístico t, su interpretación es distinta: indica que los turistas sostenibles le otorgan menos importancia que los no sostenibles. Esto sugiere que la sostenibilidad no está necesariamente asociada a la búsqueda de experiencias llamativas o fuera de lo común, sino más bien a una conexión respetuosa y coherente con el entorno local.

Variables con valores positivos en el estadístico t:

Las variables coste de vuelos, alojamiento y paquetes turísticos obtienen valores positivos, lo que implica que los turistas no sostenibles presentan un mayor gasto medio en estos elementos. Este patrón podría reflejar un perfil de turista menos preocupado por el impacto ambiental y más enfocado en la comodidad o el lujo, independientemente del coste económico.

En conjunto, los resultados reflejan que los turistas sostenibles priorizan aspectos culturales, auténticos y organizados de forma eficiente, mientras que los no sostenibles tienden a realizar un mayor gasto económico en sus viajes. Esta diferenciación aporta evidencia empírica sobre los valores y comportamientos que caracterizan a cada perfil.

Ilustración 7: Comparación de Medias: Prueba T-Student



Resultados del test t de Student aplicados a cada variable para comprobar si existen diferencias estadísticamente significativas entre turistas sostenibles y no sostenibles. Se representa el valor de la t-estadística ordenado por significancia. Elaboración propia en Python con scipy, pandas y seaborn.

4.6 Matriz de Correlación

La matriz de correlación de Spearman permite evaluar las relaciones entre las variables seleccionadas en el análisis de Mutual Information. Este tipo de correlación, no paramétrica, es adecuada para detectar relaciones monótonas, incluso cuando no son lineales.

Uno de los principales objetivos de esta matriz es comprobar si existe colinealidad entre variables, es decir, si dos o más variables están altamente correlacionadas y, por tanto, aportan información redundante al modelo. En este caso, se observa que la mayoría de las correlaciones son bajas o moderadas, lo que indica una relativa independencia entre las variables. Esto es deseable desde una perspectiva analítica, ya que sugiere que pueden ser incluidas conjuntamente en modelos predictivos sin riesgo de multicolinealidad.

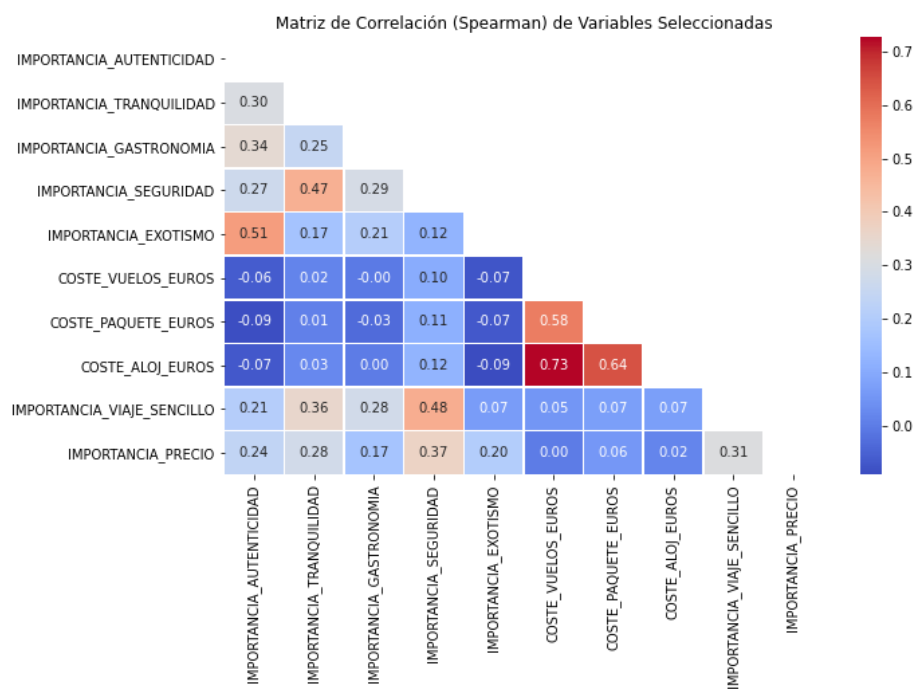
Entre los hallazgos más destacados:

- Se aprecia una correlación alta entre el coste del paquete turístico y el coste de alojamiento ($r = 0,73$), y una correlación moderada entre el coste de vuelos y el coste de alojamiento ($r = 0,64$). Esto sugiere que los turistas que gastan más en un componente del viaje tienden también a gastar más en otros, reflejando un perfil con mayor disposición al gasto y probablemente orientado a calidad y comodidad.
- La importancia de la autenticidad se correlaciona de forma moderada con el exotismo ($r = 0,51$), lo que sugiere que los turistas que valoran experiencias auténticas también tienden a buscar elementos exóticos o diferenciadores del destino.
- También destaca la relación entre la importancia del viaje sencillo y la importancia del precio ($r = 0,31$), apuntando a un perfil pragmático que valora la eficiencia tanto en organización como en costes.
- Por otro lado, las correlaciones entre las variables de percepción (autenticidad, seguridad, tranquilidad, etc.) y las variables económicas (coste del viaje) son

generalmente bajas o nulas. Esto indica que las motivaciones del turista no están directamente condicionadas por su nivel de gasto.

En resumen, esta matriz no solo permite detectar algunas relaciones interpretables entre variables, sino que también confirma que no existe una colinealidad alta, lo cual refuerza la validez de incluir estas variables en el análisis de Mutual Information y en modelos predictivos posteriores.

Ilustración 8: Matriz de Correlación de Variables Seleccionadas



Mapa de calor que representa la correlación de Spearman entre las variables empleadas en el análisis, mostrando solo la mitad superior de la matriz para mayor claridad. Elaboración propia en Python con pandas, numpy, seaborn y matplotlib.

4.7 Conclusiones

El análisis descriptivo de las variables seleccionadas mediante Mutual Information ha permitido identificar patrones significativos en el comportamiento de los turistas y su relación con la sostenibilidad. Se ha observado que variables como la importancia de la autenticidad, la tranquilidad y la seguridad presentan diferencias notables entre turistas sostenibles y no sostenibles, lo que indica que aquellos con mayor conciencia ambiental tienden a valorar más estos aspectos del destino.

Asimismo, el coste del viaje, en particular el de los vuelos, el alojamiento y los paquetes turísticos, muestra cierta capacidad para discriminar entre los grupos, aunque con un peso menor en comparación con las percepciones subjetivas. La distribución de muchas de estas variables, especialmente las relacionadas con la valoración del destino, es multimodal, lo que sugiere la existencia de diferentes perfiles de visitantes con motivaciones diversas.

La prueba T-Student confirmó que las diferencias observadas en las variables clave son estadísticamente significativas, respaldando la idea de que ciertos factores resultan más determinantes en la sostenibilidad del turista. Por su parte, la matriz de correlación de Spearman mostró una relación lógica entre las variables económicas, como el coste del vuelo y del alojamiento, mientras que las percepciones del destino mantienen una correlación baja con los costes, lo que indica que el nivel de gasto no determina necesariamente el perfil sostenible del turista.

En conjunto, este análisis descriptivo proporciona una base sólida para la siguiente fase del estudio, en la que se aplicarán modelos de aprendizaje automático con el fin de predecir y clasificar turistas sostenibles a partir de las variables identificadas.

5 Modelos de clasificación y predicción

El uso de técnicas de Machine Learning en el turismo ha experimentado un crecimiento exponencial en los últimos años, convirtiéndose en una herramienta clave para afrontar algunos de los desafíos más relevantes del sector, como la predicción de la demanda, la personalización de experiencias, la planificación territorial, y la sostenibilidad (Núñez et al., 2024). Estas técnicas, que permiten construir modelos a partir de datos históricos y extraer patrones complejos sin intervención humana directa, son especialmente útiles en entornos donde el comportamiento del usuario es heterogéneo, dinámico y difícil de anticipar con métodos tradicionales.

La literatura actual revela que el aprendizaje supervisado es el enfoque más ampliamente utilizado en aplicaciones turísticas, representando más del 45% de los estudios revisados (Núñez et al., 2024). Este tipo de algoritmos se emplea en tareas de clasificación y predicción donde los datos están etiquetados, lo cual encaja perfectamente con la finalidad de este trabajo: predecir si un turista será o no sostenible a partir de sus características y comportamientos. Dentro de este marco, los modelos seleccionados —Random Forest,

XGBoost, Regresión Logística y SVM— han demostrado un rendimiento sólido en estudios similares, tanto por su capacidad predictiva como por su utilidad en contextos reales de toma de decisiones (Wijayawardhana, 2024).

Además, estos modelos permiten no solo obtener predicciones precisas, sino también interpretar qué factores contribuyen a esas predicciones, lo cual resulta fundamental en el contexto del turismo sostenible, donde las decisiones deben ser justificadas ante múltiples actores institucionales, económicos y sociales. El uso combinado de métricas de rendimiento (precisión, AUC) con técnicas de interpretabilidad como SHAP contribuye a esta necesidad, proporcionando transparencia y explicaciones comprensibles sobre la influencia de cada variable en el modelo.

5.1 Random forest

El algoritmo Breiman (2001) es uno de los modelos de clasificación más utilizados en el ámbito del aprendizaje automático, especialmente por su capacidad para manejar datos heterogéneos y su resistencia al sobreajuste. Se basa en un enfoque de ensamblado que construye múltiples árboles de decisión a partir de diferentes subconjuntos de datos y variables aleatorias, lo que permite mejorar la precisión y la generalización del modelo. Además, su estructura interna permite estimar la importancia relativa de las variables predictoras, lo cual resulta de gran utilidad para tareas interpretativas en contextos aplicados (Schauberger et al., 2024).

En el ámbito del turismo, Random Forest ha sido ampliamente empleado para clasificar perfiles de turistas, predecir niveles de gasto, o analizar la satisfacción del visitante. Según la revisión sistemática de Núñez et al. (2024) es uno de los modelos más frecuentes en estudios que aplican Machine Learning al turismo, destacando por su rendimiento en tareas de clasificación, así como por su adaptabilidad a distintos tipos de variables y problemas. Su uso se ha extendido particularmente en investigaciones sobre sostenibilidad, donde ha permitido identificar patrones de comportamiento turístico responsable de forma eficiente.

Su capacidad para evaluar la importancia de las variables predictoras lo convierte en una herramienta muy útil cuando se requiere interpretar los factores determinantes del comportamiento turístico. En estudios recientes, como el de Núñez et al. (2024) el Random Forest ha sido identificado como uno de los modelos más frecuentes y efectivos

en la clasificación de conductas turísticas responsables, debido a su adaptabilidad a distintos tipos de variables y su facilidad de implementación.

5.2 XGBoost: Fundamentos y Aplicaciones

Li et al. (2020) explican que XGBoost representa una evolución de las técnicas de boosting, combinando árboles de decisión de manera secuencial para corregir errores y maximizar la precisión. Su diseño incorpora técnicas de regularización, control del sobreajuste, y una alta eficiencia computacional. Estas características lo han posicionado como uno de los algoritmos más utilizados en competencias de ciencia de datos y en investigaciones aplicadas.

En el contexto del turismo, XGBoost se ha utilizado para tareas como la predicción de cancelaciones, el análisis de satisfacción, o la identificación de patrones en el comportamiento de los viajeros (Louati et al., 2024). En particular, su capacidad para detectar relaciones no lineales y modelar interacciones entre variables resulta muy útil cuando se estudian perfiles de turistas sostenibles, que pueden depender de múltiples dimensiones (económicas, culturales, actitudinales). Su compatibilidad con técnicas de optimización bayesiana también facilita su ajuste en proyectos como el presente, centrado en el análisis predictivo de sostenibilidad turística.

5.3 Regresión logística

La regresión logística es un modelo estadístico clásico que sigue siendo ampliamente utilizado en contextos donde la variable dependiente es binaria. Su popularidad se debe a su simplicidad, interpretabilidad y a la robustez de su estimación a través de máxima verosimilitud (Stoltzfus, 2011). Permite calcular probabilidades de pertenencia a una clase y analizar los efectos de cada predictor mediante los conocidos odds ratios, lo cual resulta muy útil en estudios donde se desea interpretar la influencia relativa de cada variable.

En el ámbito del turismo, este modelo ha sido ampliamente empleado para analizar factores como la intención de volver al destino, la satisfacción general, o la probabilidad de adoptar comportamientos sostenibles (LaValley, 2008). Aunque su capacidad predictiva puede ser inferior a modelos más complejos, su valor interpretativo la convierte

en una herramienta indispensable para estudios de sostenibilidad, donde se necesita comunicar los resultados de manera clara a instituciones y actores no técnicos.

5.4 Support Vector Machines (SVM)

Support Vector Machine (SVM) es un algoritmo supervisado que destaca por su capacidad para encontrar fronteras óptimas de separación entre clases, maximizando el margen entre los datos más cercanos (vectores de soporte). Esta propiedad lo hace especialmente robusto en contextos donde las clases no son fácilmente separables de forma lineal, como suele ocurrir en perfiles de comportamiento turístico (Otchere et al., 2021).

Mediante el uso de funciones Kernel, SVM puede proyectar los datos en espacios de mayor dimensión donde los patrones se vuelven más evidentes. En turismo, se ha utilizado para predecir satisfacción, clasificar reseñas de usuarios, y segmentar turistas según motivaciones (Zhang & Mahdi, 2023). Su principal ventaja es la alta precisión con datasets de tamaño moderado y su buena capacidad de generalización, aunque puede requerir ajustes finos en sus hiperparámetros.

Los modelos seleccionados en este estudio no solo se encuentran entre los más utilizados en la literatura sobre Machine Learning aplicado al turismo, sino que ofrecen ventajas complementarias: Random Forest y XGBoost proporcionan alto rendimiento y análisis de importancia de variables; Regresión Logística aporta claridad interpretativa; y SVM ofrece precisión en contextos complejos y no lineales. Esta combinación permite abordar el problema de la predicción del turista sostenible desde distintos enfoques metodológicos, equilibrando rendimiento y comprensión de los resultados. Además, refuerza la relevancia del uso de modelos de Machine Learning como herramienta analítica y estratégica para avanzar hacia un turismo más inteligente y sostenible.

5.5 Resultados y comparación de modelos

Con el objetivo de predecir qué turistas presentan comportamientos sostenibles, se entrenaron cuatro modelos de clasificación: Random Forest, XGBoost, Regresión Logística y SVM. En una primera etapa, se evaluaron estos modelos con sus parámetros por defecto. Posteriormente, se aplicó un proceso de optimización de hiperparámetros utilizando búsqueda bayesiana (Bayesian Optimization), lo que permitió mejorar de

forma significativa el rendimiento de los modelos y ajustarlos mejor a las características del conjunto de datos.

La tabla siguiente muestra los resultados de precisión y AUC-ROC antes y después de la optimización:

Ilustración 9: Precisión y AUC-ROC antes y después de la optimización

MODELO	Precisión (sin optimizar)	AUC (sin optimizar)	Precisión (optimizados)	AUC (optimizados)
Random Forest	0,6278	0,6566	0,659	0,6996
XGBoost	0,6236	0,6447	0,6638	0,7033
Regresión Logística	0,6597	0,6895	0,6602	0,6895
SVM	0,6633	0,6784	0,6704	0,6974

Comparación del rendimiento de cuatro modelos de clasificación aplicados al análisis de turistas sostenibles: Random Forest, XGBoost, Regresión Logística y SVM. Se muestran los valores de precisión y AUC antes y después de aplicar una optimización de hiperparámetros. Elaboración propia en Python utilizando scikit-learn, xgboost y shap, con organización y presentación de resultados en Excel.

El mejor modelo en términos de precisión es SVM, con un valor de 0,663, seguido por la regresión logística con 0,659. En cuanto al AUC, la regresión logística obtuvo el valor más alto (0,689), lo que indica que es el modelo con mejor capacidad para diferenciar entre turistas sostenibles y no sostenibles.

En términos generales, todos los modelos mejoraron su rendimiento tras la optimización, excepto la regresión logística, cuyo rendimiento se mantuvo prácticamente constante. El modelo SVM fue el que obtuvo la mayor precisión tras la optimización (0,6704), mientras que el modelo XGBoost logró el mayor valor de AUC (0,7033), lo que indica una mayor capacidad para discriminar entre turistas sostenibles y no sostenibles.

En el caso de Random Forest, los hiperparámetros más efectivos incluyen una profundidad máxima de 6 y un mínimo de 5 muestras por nodo para dividir, con 300 árboles. Para XGBoost, la combinación óptima consistió en un learning rate bajo (0,01), profundidad reducida (2), y una tasa de muestreo parcial (subsample) de 0.5, lo que refleja la necesidad de controlar el sobreajuste. En el caso de SVM, los mejores resultados se obtuvieron con un kernel radial (rbf), una penalización C elevada (2,5) y una gamma

reducida. Todos los procesos de optimización se realizaron utilizando búsqueda bayesiana con validación cruzada para asegurar la generalización de los resultados.

Además del rendimiento cuantitativo, se consideró esencial dotar a los modelos de interpretabilidad. Para ello, se aplicó la técnica SHAP, que permite entender la contribución de cada variable a las predicciones individuales realizadas por los modelos.

Interpretación con SHAP

El análisis con SHAP permitió entender mejor qué variables tienen más peso a la hora de predecir si un turista se comporta de forma sostenible. En los cuatro modelos analizados, la variable "importancia de la autenticidad" fue la que tuvo un impacto más positivo, lo que sugiere que los turistas que buscan experiencias auténticas tienden a ser más sostenibles.

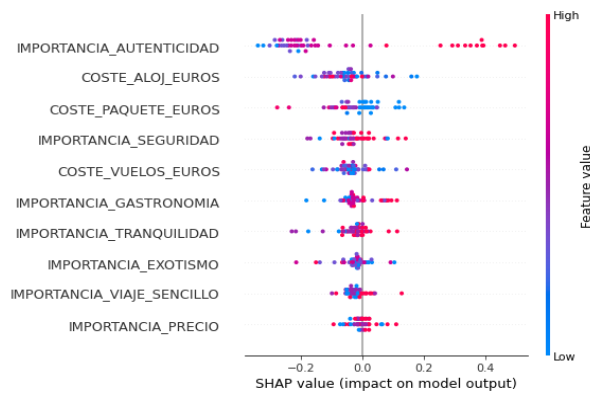
En el modelo de Random Forest, además de la autenticidad, destacaron variables económicas como el coste del paquete turístico, el alojamiento y la seguridad. El modelo XGBoost dio resultados similares, aunque pareció más sensible al coste de los vuelos y a la gastronomía como factor diferenciador.

La regresión logística fue el modelo que resultó más fácil de interpretar. En su gráfico SHAP se aprecia con más claridad cómo cada variable influye en las predicciones, con una separación más evidente entre los valores altos (en rosa) y bajos (en azul). En este caso, además de los costes, también tienen peso variable la tranquilidad o la seguridad, y se observan valores negativos cuando estas se relacionan con perfiles menos sostenibles.

El modelo SVM, por su parte, también dio mucho peso a la autenticidad, y además destacó variables como el precio y la tranquilidad, mientras que el impacto de los costes económicos (vuelos y alojamiento) fue menor en comparación con los otros modelos.

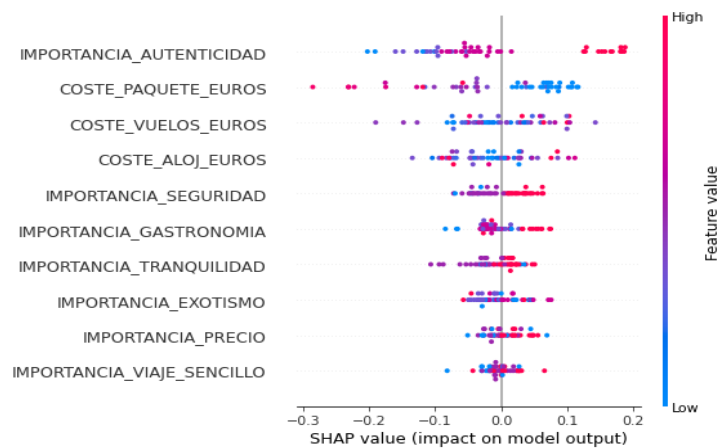
Los gráficos SHAP que se incluyen a continuación permiten ver cómo cada variable afecta a la predicción en cada modelo. El color indica si el valor de la variable es alto (rosa) o bajo (azul), y la posición muestra si empuja la predicción hacia un comportamiento más o menos sostenible. Gracias a estos gráficos, se puede ver que las percepciones del destino (como autenticidad o tranquilidad) suelen ser importantes en todos los modelos, mientras que los costes varían más según el caso.

Ilustración 10: Interpretabilidad de Random Forest



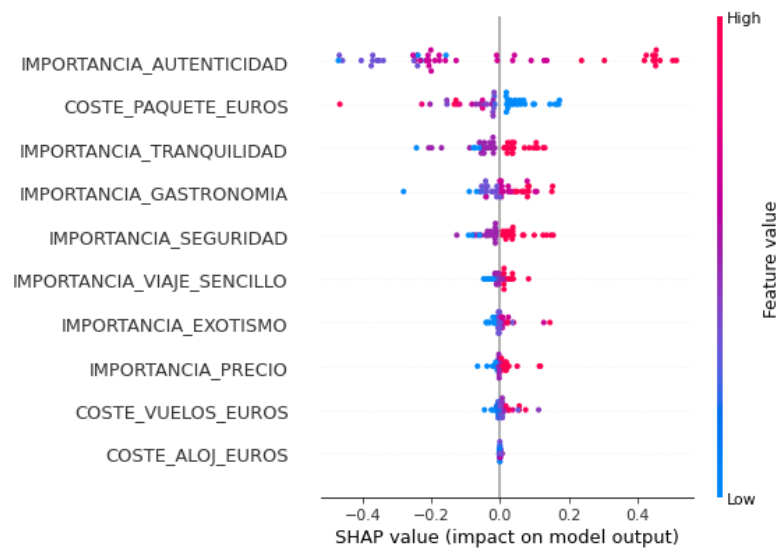
Visualización del impacto de cada variable sobre la predicción de sostenibilidad turística, utilizando valores SHAP. Las variables empleadas fueron previamente seleccionadas por su relevancia según Mutual Information. Elaboración propia en Python con shap, scikit-learn y matplotlib.

Ilustración 11: Interpretabilidad de XGBoost



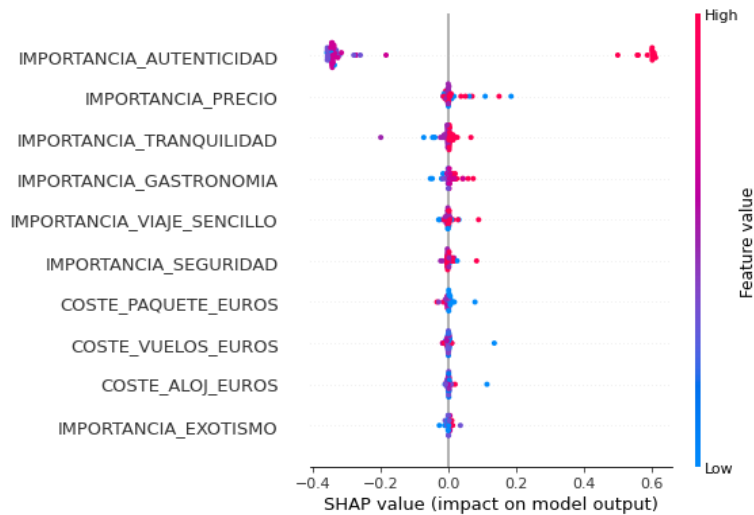
Visualización del impacto de cada variable sobre la predicción de sostenibilidad turística, utilizando valores SHAP. Las variables empleadas fueron previamente seleccionadas por su relevancia según Mutual Information. Elaboración propia en Python con shap, xgboost y matplotlib

Ilustración 12: Interpretabilidad de Regresión logística



Visualización del impacto de cada variable sobre la predicción de sostenibilidad turística, utilizando valores SHAP. Las variables empleadas fueron previamente seleccionadas por su relevancia según Mutual Information. Elaboración propia en Python con shap, scikit-learn y matplotlib.

Ilustración 13: Interpretabilidad de SVM



Visualización del impacto de cada variable sobre la predicción de sostenibilidad turística, utilizando valores SHAP. Las variables empleadas fueron previamente seleccionadas por su relevancia según Mutual Information. Elaboración propia en Python con shap, scikit-learn y matplotlib.

5.6 Discusión de Resultados

Este análisis permite sacar varias conclusiones interesantes sobre el perfil del turista sostenible en Canarias. Por un lado, aunque modelos como XGBoost y SVM obtienen mejores resultados en términos de precisión y AUC, también son más complejos y sensibles a los parámetros, lo que puede hacerlos más propensos al sobreajuste.

Por otro lado, la regresión logística, aunque tiene una precisión algo menor, destaca por ser el modelo más fácil de interpretar, especialmente gracias a los gráficos SHAP. Esto lo convierte en una opción muy útil cuando lo que se busca es entender bien cómo influyen las distintas variables en las predicciones, no solo acertar más.

Además, los resultados confirman que las variables más importantes no son solo económicas. Factores como la autenticidad, la tranquilidad o la seguridad tienen un peso clave, lo que sugiere que el comportamiento sostenible va muy ligado a una forma concreta de ver y vivir el destino.

Por último, el uso de SHAP ha sido muy útil para entender mejor los modelos y sacar conclusiones claras. Esta información puede servir de base para que administraciones y empresas diseñen estrategias más enfocadas en atraer a turistas que valoren atributos sostenibles, y no solo precios bajos o paquetes turísticos cerrados.

6. Reglas de Asociación para la Sostenibilidad Turística

6.1 Reglas de Asociación en el Estudio del Turismo

Además del uso de modelos predictivos complejos, este trabajo explora una vía alternativa basada en la extracción de reglas de asociación mediante el algoritmo Apriori, con el objetivo de identificar patrones claros y fácilmente interpretables que caractericen a los turistas sostenibles. A través de esta técnica, es posible detectar combinaciones frecuentes de atributos (antecedentes) que conducen a una determinada conclusión (consecuente), como en este caso, la probabilidad de que un turista adopte comportamientos sostenibles.

Una de las principales ventajas de este enfoque es su transparencia. A diferencia de modelos más sofisticados como Random Forest o XGBoost, que si bien ofrecen alta precisión pueden resultar opacos para la toma de decisiones públicas, las reglas de asociación permiten generar conclusiones del tipo: “si un turista valora mucho la autenticidad y gasta poco en el paquete turístico, entonces es probable que sea sostenible”.

Esta lógica simple facilita la interpretación de resultados por parte de gestores turísticos, entidades públicas y comunidades locales.

Para implementar esta técnica, se ha utilizado la librería `mlxtend` de Python, en concreto el módulo `frequent_patterns`, que proporciona una implementación eficiente del algoritmo Apriori. Las variables continuas más influyentes (identificadas previamente mediante Mutual Information) se transformaron en variables categóricas ("bajo", "medio", "alto") y posteriormente en variables binarias mediante codificación one-hot.

Durante la ejecución del algoritmo se establecieron varios hiperparámetros que condicionan la generación de reglas:

- `Min_support = 0,05`: umbral mínimo de frecuencia para que un conjunto de atributos sea considerado relevante.
- `Metric = "confidence"` y `min_threshold = 0,6`: se utilizaron como criterios de selección al generar las reglas, asegurando que solo se incluyeran relaciones con un mínimo nivel de fiabilidad.

Las reglas obtenidas fueron evaluadas mediante tres métricas clave:

1. Soporte (support): proporción de registros en los que aparece simultáneamente el antecedente y el consecuente. Mide la frecuencia.
2. Confianza (confidence): probabilidad de que el consecuente ocurra dado que el antecedente ha ocurrido. Evalúa la fuerza de la relación.
3. Elevación (lift): cuantifica la mejora en la predicción del consecuente al conocer el antecedente, comparado con una predicción aleatoria. Un valor superior a 1 indica una asociación positiva.

Con el fin de facilitar su interpretación, se ha generado un gráfico de elaboración propia que representa cada una de las reglas obtenidas. En este gráfico, el eje X muestra el soporte, el eje Y el lift, y el tamaño y color del punto representan el valor de confianza. Esta representación permite identificar de forma visual qué reglas, aunque no sean las más frecuentes, tienen un mayor poder predictivo sobre el comportamiento sostenible.

La elaboración del gráfico se realizó en Python utilizando la librería `matplotlib`. Cada regla fue representada como un punto en un diagrama de dispersión, donde el tamaño del punto está proporcionalmente escalado según su nivel de confianza y el color varía

también en función de este valor, utilizando una escala de color (colormap) tipo “viridis”. Este gráfico fue generado a partir de los datos exportados en Excel desde las reglas obtenidas, filtrando únicamente aquellas cuyo consecuente incluye ‘Turista Sostenible’=1.

Este enfoque ya ha demostrado su utilidad en contextos reales. Por ejemplo, Xiao (2022) aplicó Apriori para identificar factores impulsores del turismo rural sostenible en China, analizando cómo se combinan aspectos como la demanda, el entorno natural y el apoyo institucional para guiar el desarrollo turístico en zonas rurales. Asimismo, Arreeras et al. (2019) utilizaron reglas de asociación sobre datos masivos recogidos mediante sensores Wi-Fi para identificar patrones de movilidad turística en Hokkaido. Estos resultados sirvieron para planificar de forma más sostenible la distribución de los flujos de visitantes y mejorar la gestión del destino. Por su parte, Clark & Doyle (2022) argumentan que el uso de algoritmos interpretables como Apriori es especialmente adecuado en dominios donde la transparencia y la trazabilidad de las decisiones son clave, como ocurre en el ámbito de la sostenibilidad turística, la planificación participativa o la gobernanza pública.

En resumen, el uso de reglas de asociación no solo complementa a los modelos predictivos más complejos, sino que aporta un valor diferencial: permite traducir los datos en reglas fácilmente comunicables y aplicables, reforzando así el vínculo entre análisis técnico y toma de decisiones basada en evidencia.

Ilustración 14: Representación de las reglas de asociación que predicen turistas sostenibles.

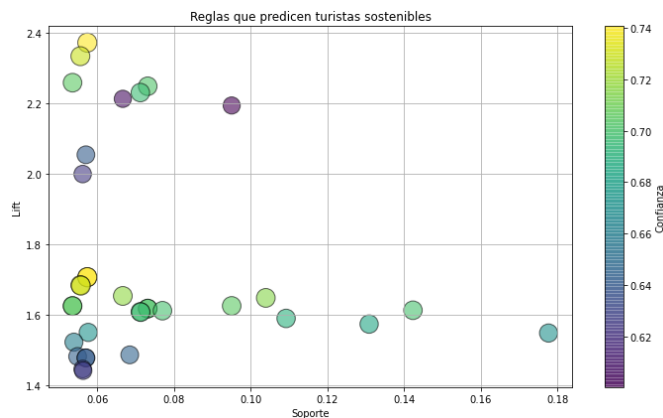


Gráfico de dispersión que muestra la relación entre el soporte (frecuencia relativa) y el lift (capacidad predictiva) de cada regla generada, siendo el tamaño y el color de los puntos proporcionales a su confianza (probabilidad condicional). Elaboración propia realizada en Python utilizando las librerías mlxtend y matplotlib.

6.2 Resultados de las Reglas de Asociación

A continuación, se analizan las cinco reglas de asociación con mayor lift, es decir, aquellas que mejor predicen la sostenibilidad turística en comparación con una ocurrencia aleatoria:

Regla 1

- Antecedentes: Importancia alta de la autenticidad, importancia media de la tranquilidad, gasto bajo en alojamiento.
- Interpretación: Turistas que valoran una experiencia auténtica y relajada, y que además gastan poco en alojamiento, tienen un 74,1% de probabilidad de ser sostenibles. Esta combinación multiplica por 2,37 la probabilidad de sostenibilidad respecto al promedio.

Regla 2

- Antecedentes: Igual que la anterior, pero sustituyendo el alojamiento por gasto bajo en vuelos.

- Interpretación: Confirma que los turistas que priorizan la autenticidad y la tranquilidad, y que además tienen un gasto contenido en transporte, también presentan un perfil sostenible, con una probabilidad del 72,9% y un lift de 2,33.

Regla 3

- Antecedentes: Importancia alta de la autenticidad y gastos bajos tanto en vuelos como en alojamiento.
- Interpretación: El turista que combina una motivación cultural fuerte con una actitud de ahorro en transporte y alojamiento se asocia a una probabilidad del 70,5% de ser sostenible.

Regla 4

- Antecedentes: Importancia alta de la autenticidad + gasto bajo en alojamiento.
- Interpretación: La autenticidad, por sí sola, combinada con decisiones económicas moderadas, ya predice la sostenibilidad con un 70,2% de confianza.

Regla 5

- Antecedentes: Importancia alta de la autenticidad + gasto bajo en vuelos.
- Interpretación: Similar a la anterior, pero con ahorro en transporte. Esto sugiere que el perfil sostenible no es necesariamente el de mayor gasto, sino el de decisiones conscientes, tanto en motivaciones como en presupuesto.

Además de estas reglas frecuentes y sólidas, el gráfico de dispersión de reglas (soporte vs. confianza, coloreado por *lift*) permite identificar relaciones menos comunes, pero altamente fiables. En particular, destaca un punto amarillo ubicado en la esquina superior izquierda del gráfico, correspondiente a una regla con soporte bajo (5 %), pero confianza perfecta (1,0) y *lift* elevado (1,50). Esta regla indica que los turistas que otorgan alta importancia al precio, baja a la gastronomía y que tienen un coste bajo tanto en alojamiento como en el paquete turístico, cumplen siempre con el perfil de sostenibilidad definido. Aunque se trata de un perfil poco frecuente, su fiabilidad absoluta la convierte en una regla de especial interés para diseñar estrategias dirigidas a nichos muy específicos.

Los resultados refuerzan la idea de que la sostenibilidad turística está asociada a turistas con valores personales fuertes (como la búsqueda de autenticidad y tranquilidad), y no necesariamente con niveles de gasto elevados. Este tipo de hallazgos puede ser especialmente útil para orientar políticas de promoción y diseño de experiencias turísticas más responsables.

7. Análisis de la Percepción Local del Turismo (NLP)

7.1 Introducción al análisis de comentarios locales

Una vez analizadas las características del turismo en Canarias, los perfiles de los visitantes y el grado de sostenibilidad de sus comportamientos, resulta fundamental incorporar la perspectiva de quienes conviven diariamente con este modelo turístico: la población local.

El análisis de percepción local permite complementar la evaluación de sostenibilidad, ya que los residentes son los primeros en experimentar los efectos positivos y negativos del turismo en su territorio. Por tanto, conocer sus discursos, preocupaciones y valoraciones es clave para comprender si el modelo actual está generando tensiones sociales, así como para identificar posibles medidas que alineen el turismo con el bienestar colectivo.

En este apartado se ha aplicado procesamiento de lenguaje natural (NLP) a una base de más de 4.000 comentarios de personas residentes en Canarias. Estas opiniones, recogidas de comentarios de Youtube, abordan libremente temas vinculados al turismo, la vivienda, la economía o la calidad de vida.

7.2 Extracción de bigramas frecuentes

Con el objetivo de identificar temas relevantes y patrones recurrentes en los comentarios de los residentes locales, se realizó una extracción de bigramas, es decir, combinaciones de dos palabras que aparecen consecutivamente con alta frecuencia en el texto. Este tipo de análisis permite captar conceptos más específicos y contextuales que las palabras individuales no siempre revelan.

Para este análisis se utilizó Python, con las librerías nltk, collections, pandas, matplotlib y seaborn. Los bigramas fueron extraídos utilizando la función Counter sobre una lista de tokens previamente preprocesada (que incluía eliminación de signos, stopwords y lematización). Se seleccionaron los 15 bigramas más frecuentes, sin aplicar técnicas de

ponderación ni filtrado avanzado, dado que el objetivo era una representación exploratoria y accesible.

El gráfico de barras muestra los bigramas más frecuentes en el corpus. Se observa una alta aparición de combinaciones como "inmigración ilegal", "gran canaria", "contra turismo", "inmigrantes ilegales" o "turismo masivo", lo que refleja una preocupación por la presión migratoria y los efectos del turismo en las islas. También destacan términos vinculados a la calidad de vida y el acceso a la vivienda, como "calidad vida" o "alquiler vacacional". Este análisis sugiere que los discursos locales en torno al turismo están interconectados con otros temas sociales y económicos.

Este análisis no incluye optimización de hiperparámetros, ya que la elección del número de bigramas a representar se realizó de forma empírica, basada en criterios de claridad visual y cobertura temática. Aun así, los resultados son superiores a un análisis trivial de palabras sueltas, al permitir detectar relaciones contextuales como "inmigración ilegal", "contra turismo" o "turismo calidad", que resultan mucho más informativas que tokens individuales aislados.

Ilustración 15: Bigramas más frecuentes en comentarios locales

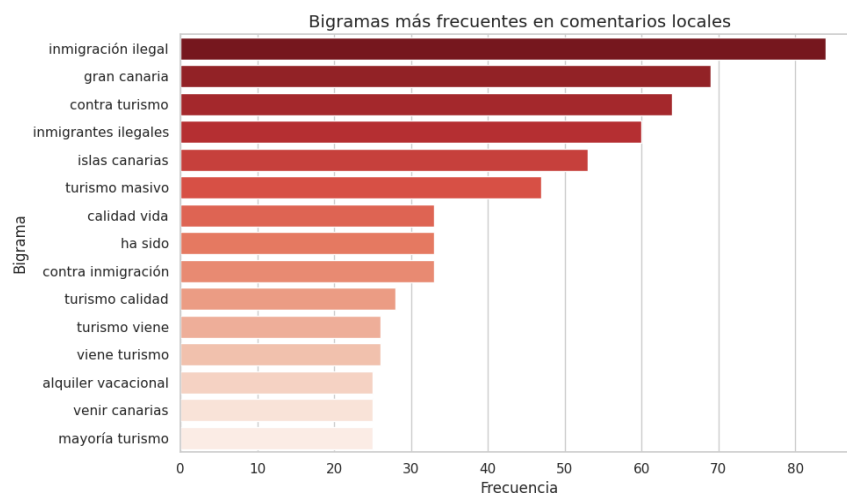


Gráfico de barras que muestra las combinaciones de dos palabras más repetidas en el corpus de comentarios de residentes canarios. Permite identificar los temas más presentes en la conversación social en torno al turismo. Elaboración propia con Python, utilizando las librerías collections, pandas, matplotlib y seaborn.

7.3 Clustering con WordCloud

Para explorar más a fondo los temas latentes en los comentarios, se aplicó una técnica de agrupamiento (Clustering) basada en el contenido semántico de los textos. El proceso constó de las siguientes fases:

- Preprocesamiento del texto: Se limpiaron los comentarios eliminando signos de puntuación, stopwords, mayúsculas, y se aplicó lematización para reducir las palabras a su forma base.
- Vectorización: Los textos preprocesados se transformaron en vectores numéricos utilizando la técnica de TF-IDF (Term Frequency – Inverse Document Frequency), que permite identificar la relevancia de cada palabra en un documento con respecto al conjunto total de comentarios.
- Aplicación del algoritmo de Clustering: Se utilizó el algoritmo KMeans con `n_clusters=2` y `random_state=42`. La elección del valor de `k` se basó en una exploración inicial, donde se observó una división preliminar entre dos grandes grupos temáticos: uno centrado en la inmigración y otro en la masificación turística. El modelo obtuvo un Silhouette Score de 0,021, lo que indica una separación modesta entre clusters, suficiente para capturar diferencias discursivas relevantes en un contexto exploratorio.
- Visualización con nubes de palabras: Para facilitar la interpretación de los clusters generados, se creó una nube de palabras para cada grupo, donde el tamaño de cada término es proporcional a su frecuencia relativa dentro del cluster correspondiente.

El vectorizador TF-IDF se configuró con los siguientes parámetros: `max_features=1000` (para limitar el número de términos a los 1.000 más relevantes), `max_df=0,5` (para excluir palabras demasiado frecuentes que aparecen en más del 50 % de los documentos) y `min_df=2` (para eliminar términos que aparecen en un solo documento). Aunque no se optimizaron automáticamente, estos hiperparámetros se eligieron siguiendo buenas prácticas en análisis exploratorio de texto.

Además, se extrajeron las palabras con mayor peso medio en cada cluster. En el Cluster 0, las más representativas fueron: `contra` (0,182), `inmigración` (0,135), `ilegal` (0,122), `ilegales` (0,076), `inmigrantes` (0,065). En el Cluster 1 destacaron: `problema` (0,016),

en Python, utilizando las librerías *scikit-learn* (para vectorización con *TfidfVectorizer* y agrupamiento con *KMeans*) y *Wordcloud* para la representación visual.

7.4 Resultados e interpretación de temas emergentes

El análisis de Clustering permite dividir los comentarios en dos grupos temáticamente diferenciados, lo que refleja preocupaciones distintas por parte de los residentes.

En el primer grupo (**Cluster 0**), los comentarios giran principalmente en torno a la inmigración. Son frecuentes palabras como *inmigración*, *ilegal*, *patera*, *manifestación*, *invasión* o *ilegales*. Aunque no todos los mensajes mencionan directamente el turismo, sí aparecen expresiones como *contra turismo* o *contra inmigración*, lo que sugiere que algunas personas vinculan ambos fenómenos. En varios casos, se transmite una sensación de saturación, de pérdida de control sobre el espacio y los recursos, y de rechazo hacia lo externo.

Este tipo de comentarios refleja malestar con la situación actual en las islas y, en algunos casos, se mezcla con enfado hacia las instituciones o con la percepción de que la población local está siendo desplazada o ignorada. No se trata solo de una crítica al turismo o a la inmigración, sino a cómo se están gestionando ambos fenómenos. Muchas de las expresiones que aparecen en este cluster también se encuentran entre los bigramas más frecuentes (como *inmigración ilegal*, *contra turismo* o *inmigrantes ilegales*), lo que refuerza la presencia de este discurso en el conjunto de comentarios.

En el segundo grupo (**Cluster 1**), el foco está mucho más puesto en el turismo, pero desde una visión crítica hacia sus efectos sobre la vida cotidiana. Aparecen con frecuencia palabras como *vivienda*, *alquiler vacacional*, *dinero*, *gobierno*, *vivir* o *playa*. Muchos comentarios hacen referencia a los problemas para acceder a una vivienda, a la sensación de que el turismo beneficia solo a unos pocos, y a la falta de acción por parte de las instituciones.

Este tipo de mensajes no rechaza el turismo como tal, sino que critica cómo se está desarrollando. Lo que se cuestiona es el modelo actual, basado en la masificación, el aumento del alquiler turístico y la falta de regulación. Las quejas también coinciden con varios de los bigramas más frecuentes, como *calidad vida*, *turismo masivo* o *turismo calidad*, lo que indica que estos temas están muy presentes en el debate público.

En conjunto, el análisis sugiere que el descontento con el turismo no es uniforme, pero sí generalizado. Mientras que un grupo pone el foco en cuestiones identitarias o de seguridad, el otro se centra en el acceso a derechos básicos como la vivienda. En ambos casos, aparece una demanda clara de mayor control y planificación. Esto refuerza la idea de que avanzar hacia un modelo de turismo más sostenible no solo es necesario desde una perspectiva ambiental, sino también social. Los resultados apuntan a que cualquier estrategia turística futura debería tener en cuenta estas preocupaciones locales si se quiere lograr un equilibrio real entre desarrollo económico y bienestar de la población.

8. Conclusiones y Recomendaciones

Uno de los objetivos principales de este trabajo era **definir el perfil del turista sostenible en Canarias y determinar qué variables permiten predecir su comportamiento**. A partir de la revisión teórica (Budeanu, 2007; Juvan & Dolnicar, 2016), se estableció una definición operativa del turista sostenible basada en comportamientos observables, como el uso del transporte público, la participación en actividades culturales o la valoración del entorno. Esta definición fue trasladada al análisis empírico mediante una variable binaria, lo que permitió aplicar técnicas de selección de características, clasificación e interpretación de resultados.

El uso de **Mutual Information** permitió identificar las variables que más influyen en la sostenibilidad turística. Los resultados destacan especialmente la importancia otorgada a la autenticidad, la tranquilidad y la seguridad, situándolas por encima de aspectos económicos como el coste del alojamiento o de los vuelos. Estos hallazgos coinciden con la literatura previa (Gupta & Chopra, n.d.; He et al., 2018), y también se reflejan en los resultados del **Capítulo 3, Figura 1**, donde se observa que las variables relacionadas con motivaciones personales presentan una mayor puntuación de importancia que las económicas. Esto refuerza la idea de que la sostenibilidad turística depende más de las motivaciones del viajero que de su nivel de gasto.

Otro de los objetivos del estudio era **evaluar la capacidad predictiva de distintos modelos de Machine Learning** para anticipar el comportamiento sostenible del turista. En este sentido, los modelos de clasificación aplicados han demostrado una **precisión de 0,6704 y 0,664 en SVM y XGBoost**, respectivamente, los que ofrecieron mejores métricas, aunque la **regresión logística** resultó especialmente útil por su claridad interpretativa. La técnica **SHAP** permitió, además, entender cómo influye cada variable

en las predicciones, reforzando la idea de que las percepciones del destino, especialmente la autenticidad, son determinantes, mientras que los factores económicos presentan un peso más variable.

Con el objetivo de aportar interpretabilidad complementaria, se desarrolló un sistema de **reglas de asociación** que permitió identificar **combinaciones frecuentes de atributos vinculadas al comportamiento sostenible**. Las reglas con mayor lift muestran que los turistas que valoran la autenticidad y que presentan un gasto contenido en alojamiento o transporte tienen una mayor propensión a adoptar prácticas sostenibles. Este enfoque, además de reforzar los resultados anteriores, ofrece una herramienta sencilla y útil para la toma de decisiones en políticas de promoción turística.

Por último, se abordó el análisis de la **percepción local** del turismo mediante técnicas de procesamiento del lenguaje natural. Este análisis reveló dos grandes líneas de malestar entre los residentes: una centrada en cuestiones identitarias, como la inmigración, y otra más estructural, relacionada con el acceso a la vivienda, la masificación y la desigualdad. Lo más relevante es que estos discursos no expresan un rechazo al turismo como tal, sino una crítica al modelo actual y a su impacto sobre la vida cotidiana. Este hallazgo refuerza la necesidad de incluir la voz de la población local como **indicador esencial de sostenibilidad**.

En conjunto, los resultados permiten concluir que **la sostenibilidad turística no responde a un perfil económico concreto**, sino que está asociada a un conjunto de **motivaciones personales, percepciones del destino y decisiones conscientes**. La coincidencia entre lo que valoran los turistas sostenibles y las preocupaciones expresadas por los residentes apunta a una posible convergencia hacia un modelo turístico más equilibrado, justo y resiliente.

Recomendaciones

A partir de los hallazgos obtenidos a lo largo del análisis, se proponen las siguientes líneas de acción orientadas a reforzar la sostenibilidad turística en Canarias:

- **Fomentar un perfil de turista sostenible**, mediante productos turísticos que prioricen la autenticidad, la cultura local, la tranquilidad y el respeto por el entorno. Esta recomendación se basa en el hecho de que los turistas sostenibles valoran especialmente

estos atributos, como se ha demostrado en la selección de variables mediante Mutual Information y en los modelos de clasificación.

- **Regular y equilibrar el alquiler vacacional**, con el objetivo de garantizar el acceso a la vivienda de los residentes. Esta medida responde a una preocupación recurrente detectada en el análisis de comentarios locales, donde el aumento de los alquileres aparece como uno de los efectos más negativos del modelo turístico actual.
- **Reducir la masificación turística**, no solo por sus impactos ambientales, sino también para mitigar el malestar social. Tanto los turistas sostenibles como los residentes muestran rechazo hacia la masificación: los primeros prefieren experiencias tranquilas, y los segundos denuncian su efecto sobre la calidad de vida y los servicios.
- **Reforzar la conexión entre turismo y comunidad local**, asegurando que los beneficios económicos se distribuyan de forma más justa. Este punto emerge del análisis de la percepción local, donde se critica que los beneficios del turismo no se traducen en mejoras para la población residente.
- **Incorporar de forma sistemática la percepción de los residentes** en la planificación turística, estableciendo mecanismos formales que permitan recoger su voz y evaluar la sostenibilidad desde una perspectiva social. La explotación de los comentarios cualitativos ha demostrado el valor de esta dimensión para complementar las métricas tradicionales.

Estos resultados se alinean con estudios previos como los de Hernández Luis et al. (2017), Bianchi (2004) y Exceltur (2023), pero aportan una visión más integrada basada en **datos actuales** y en **técnicas analíticas avanzadas**. Además, demuestran que los perfiles sostenibles no responden a un único patrón, sino que surgen de **distintas combinaciones de motivaciones y decisiones de consumo**.

Finalmente, por motivos de tiempo, este trabajo no ha podido abordar ciertas líneas de investigación que se proponen como futuras extensiones. Entre ellas destacan el uso de **redes KAN** y **procesos gaussianos profundos**, el análisis longitudinal con datos de varios años, o la integración de distintas fuentes (económicas, ambientales y censales) para enriquecer el análisis. Estas propuestas podrían aportar herramientas valiosas para avanzar hacia una planificación turística verdaderamente sostenible.

En definitiva, para que el turismo en Canarias sea sostenible, no basta con transformar el comportamiento de los turistas: **también es necesario transformar el modelo**. Esto

implica escuchar a quienes viven en el territorio, entender lo que valoran los visitantes más conscientes y garantizar que el desarrollo turístico contribuya de forma clara al bienestar de toda la comunidad

9. Bibliografía

- Arreeras, T., Arimura, M., Asada, T., & Arreeras, S. (2019). Association rule mining tourist-attractive destinations for the sustainable development of a large tourism area in Hokkaido using Wi-Fi tracking data. *Sustainability (Switzerland)*, 11(14). <https://doi.org/10.3390/su11143967>
- Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42(22), 8520–8532. <https://doi.org/10.1016/j.eswa.2015.07.007>
- Bianchi, R. V. (2004). *Tourism restructuring and the politics of sustainability: A critical view from the European periphery (the Canary Islands)*. *Journal of Sustainable Tourism*, 12(6), 495–529. <https://doi.org/10.1080/09669580408667251>
- Breiman, L. (2001). *Random Forests* (Vol. 45).
- Budeanu, A. (2007). Sustainable tourist behaviour – a discussion of opportunities for change. *International Journal of Consumer Studies*, 31(5), 499–508. <https://doi.org/10.1111/j.1470-6431.2007.00606.x>
- Clark, R., & Doyle, T. E. (2022). A Priori Quantification of Transfer Learning Performance on Time Series Classification for Cyber-Physical Health Systems. *Canadian Conference on Electrical and Computer Engineering*, 2022-September, 458–464. <https://doi.org/10.1109/CCECE49351.2022.991827>
- Exceltur & Gobierno de Canarias. (2023). *Impactur Canarias 2022. Estudio del Impacto Económico del Turismo sobre la Economía y el Empleo de Canarias*. <https://www.exceltur.org/wp-content/uploads/2023/12/Impactur-Canarias-2022.pdf>
- Gupta, G., & Chopra, P. (n.d.). *Eco-tourists and Environment Protection: A Pro-Environment Behavioural Segmentation Approach*.
- He, P., He, Y., & Xu, F. (2018). Evolutionary analysis of sustainable tourism. *Annals of Tourism Research*, 69, 76–89. <https://doi.org/10.1016/j.annals.2018.02.002> Impactur-Canarias-2022-. (n.d.).
- Hernández Luis, J.Á., González Morales, A., Hernández Torres, S., & Ramón Ojeda, A. (2017). *The impact of mass tourism in the Canary Islands in the context of world biosphere reserves*. *Cuadernos de Turismo*, (40), 685–688. <https://revistas.um.es/turismo/article/view/312921>
- Juvan, E., & Dolnicar, S. (2016). Measuring environmentally sustainable tourist behaviour. *Annals of Tourism Research*, 59, 30–44. <https://doi.org/10.1016/j.annals.2016.03.006>
- Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(6), 16. <https://doi.org/10.1103/PhysRevE.69.066138>
- LaValley, M. P. (2008). Logistic regression. In *Circulation* (Vol. 117, Issue 18, pp. 2395–2399). <https://doi.org/10.1161/CIRCULATIONAHA.106.682658>
- Ley Bosch, P., de Castro González, Ó., & García Sánchez, F. (2024). *Mass tourism urban destinations and climate change in small islands: Resilience to extreme rainfall in the Canary Islands*. *Environment, Development and Sustainability*, 26, 10765–10785. <https://doi.org/10.1007/s10668-023-03406-7>
- Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems*, 35(3), 52–61. <https://doi.org/10.1109/MIS.2020.2972533>
- Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>

- Louati, A., Louati, H., Alharbi, M., Kariri, E., Khawaji, T., Almubaddil, Y., & Aldwsary, S. (2024). Machine Learning and Artificial Intelligence for a Sustainable Tourism: A Case Study on Saudi Arabia. *Information*, 15(9), 516. <https://doi.org/10.3390/info15090516>
- Núñez, J. C. S., Gómez-Pulido, J. A., & Ramírez, R. R. (2024). Machine learning applied to tourism: A systematic review. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. John Wiley and Sons Inc. <https://doi.org/10.1002/widm.1549>
- Otchere, D. A., Arbi Ganat, T. O., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. In *Journal of Petroleum Science and Engineering* (Vol. 200). Elsevier B.V. <https://doi.org/10.1016/j.petrol.2020.108182>
- Parra López, E., & Calero García, F. J. (2006). Agrotourism, sustainable tourism and Ultraperipheral areas: The Case of Canary Islands. *PASOS Revista de Turismo y Patrimonio Cultural*, 4(1), 85–97. <https://doi.org/10.25145/j.pasos.2006.04.006>
- Schauberger, G., Klug, S. J., & Berger, M. (2024). Random forests for the analysis of matched case–control studies. *BMC Bioinformatics*, 25(1). <https://doi.org/10.1186/s12859-024-05877-5>
- Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, 18(10), 1099–1104. <https://doi.org/10.1111/j.1553-2712.2011.01185.x>
- Wijayawardhana, A. J. (2024). Forecasting Future Tourism Demand to Popular Tourist Attractions to Asses Overtourism: A Machine Learning Modeling Approach. 2024 *International Conference on Engineering and Emerging Technologies (ICEET)*, 1–6. <https://doi.org/10.1109/ICEET65156.2024.10913889>
- Xiao, H. (2022). Algorithm of Apriori-Based Rural Tourism Driving Factors and Its System Optimization. *Mobile Information Systems*, 2022. <https://doi.org/10.1155/2022/3380609>
- Zhang, M., & Mahdi, W. A. (2023). Development of SVM-based machine learning model for estimating lornoxicam solubility in supercritical solvent. *Case Studies in Thermal Engineering*, 49. <https://doi.org/10.1016/j.csite.2023.103268>

ANEXO

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, [Nombre completo del estudiante], estudiante de [nombre del título] de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "[Título del trabajo]", 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 [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]:

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. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
4. Interpretador de código: Para realizar análisis de datos preliminares.
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: 8 de abril de 2025

Firma: Casilda De La Nuez Sánchez