Facultad de Ciencias económicas y Empresariales

COMPARACIÓN DE MODELOS MACHINE LEARNING PARA EL GASTO TURÍSTICO TRAS EL COVID-19

Nombre: María Moreno Bastante
Tutor: Luis Ángel Calvo Pascual
Grado: E2 Analytics

Madrid | Marzo 2023

ÍNDICE

ÍNDICE DE FIGURAS

CAPÍTULO 1

CAPÍTULO 2

CAPITULO 3

CAPITULO 4

CAPITULO 5

RESUMEN

En este trabajo se ha investigado la evolución del gasto turístico en los años 2019 y 2022 para conocer el impacto de la pandemia sobre esta variable y sobre el sector. Para ello, se ha analizado la variable gasto turístico extraída de encuestas de turismo como EGATUR, que toman como población de estudio turistas no residentes en el país. Se han empleado diferentes técnicas cuantitativas como *feature selection* donde se ha obtenido como resultado que no ha habido variación en los factores o variables más determinantes del gasto turístico de un periodo a otro. Dichas variables han sido el número de pernoctaciones, el país de procedencia y el medio de transporte. Por otro lado, se han construido a partir de estos resultados modelos de regresión y clasificación para predecir el gasto turístico. Como mejor modelo de regresión se ha obtenido un proceso gaussiano GPR, mientras que de los modelos de clasificación ninguno ha demostrado suficiente capacidad predictiva debido a un aparente desequilibrio en el volumen de observaciones durante el proceso de categorización de la variable. Como consecuencia, no se obtuvieron resultados concluyentes donde se pudiera distinguir una clara distribución para las categorías de gasto alto, medio y bajo, y por ello, se ha realizado un análisis complementario usando *imbalance data* para detectar el turismo de lujo. En este último análisis se ha obtenido como algoritmo que mejor modeliza los datos un *XGBboost* que ha mostrado una precisión del 65% al clasificar a los turistas de lujo.

ABSTRACT

This paper has investigated the evolution of tourism expenditure in the years 2019 and 2022 in order to know the impact of the pandemic on this variable and on the sector. For this purpose, we have analyzed the tourism expenditure variable extracted from tourism surveys such as EGATUR, which take non-resident tourists in the country as the study population. Different quantitative techniques have been used as feature selection where it has been obtained as a result that there has been no variation in the most determining factors or variables of tourist expenditure from one period to another. These variables were the number of overnight stays, the country of origin and the means of transport. On the other hand, regression and classification models have been constructed from these results to predict tourist spending. A Gaussian GPR process was obtained as the best regression model, while none of the classification models showed sufficient predictive capacity due to an apparent imbalance in the volume of observations during the process of categorizing the variable. As a consequence, no conclusive results were obtained where a clear distribution could be distinguished for the high, medium and low expenditure categories, and therefore, a complementary analysis has been carried out using imbalance data to detect luxury tourism. In this last analysis, an XGBboost algorithm has been obtained as the algorithm that best models the data, showing an accuracy of 65% and therefore a good capacity to classify luxury tourists.

Capítulo 1

CONTEXTO

1.1. El sector del turismo en España

España representa uno de los destinos turísticos más importantes del mundo. El turismo es uno de los principales motores económicos en España ya que juega un papel muy importante en la generación de empleo e ingresos del país. Tal es así, que la actividad turística tiene una contribución de 154´487 millones de euros, lo que representa un 12´4% sobre el PIB español y un 12´9% del total de empleos de la economía del país ([1])[1], por encima de otros sectores relevantes como la construcción y el comercio. Según la Organización Mundial del Comercio (OMC) ([2]), España se encuentra actualmente en la cuarta posición en cuanto a número de visitas turísticas internacionales de entre los países europeos, tan solo detrás de países como Francia y Estados Unidos, siendo el tercer país en cuanto a nivel de ingresos generados con la actividad turística. El hecho de que España ocupe una de las primeras posiciones como destino escogido por los turistas, explica que el sector sea una de las principales fuentes de ingresos y su capacidad de producir empleo y riqueza para el país, convirtiéndole en un referente. El desarrollo turístico en España se ha visto beneficiado tanto por factores internos como externos. Por un lado, los factores internos, principalmente se refieren al hecho de que el país presenta características favorables para el desarrollo de la actividad turística como puede ser la accesibilidad, el clima, la diversidad y su reciente modernización y desarrollo. Por otro lado, en cuanto a factores externos, encontramos principalmente cuestiones históricas que han resultado en el desarrollo socioeconómico del país, el progreso de los medios de transporte y la popularización de las vacaciones en Europa tras la segunda Guerra Mundial ([3]).

En España, y a pesar de la crisis actual causada por el Covid-19, el turismo sigue siendo una actividad en crecimiento y, según la encuesta de turismo FRONTUR ([4]), entre enero y noviembre de 2022, ha habido un incremento en la recepción de turistas internacionales en casi un 139% con respecto a 2021, una cifra que refleja recuperación para el sector.

[1] Se han usado como referencia las estadísticas de 2019 ya que son los datos más recientes despreciando los años 2020 y 2021 cuyas cifras no son representativas ya que se ven afectadas por el impacto de la pandemia

Además, cada vez hay más destinos turísticos, lo que ha fomentado la construcción de nuevas infraestructuras y alojamientos vacacionales dando lugar a un tipo de turismo *low cost*. Como consecuencia de esto y de la aparición de nuevas herramientas para reservar estancias vacacionales ajustadas por calidad-precio como *Booking* o *Airbandb*, el turismo ha incrementado exponencialmente en los últimos años, especialmente tras la crisis de 2008 ([5]). En España[2], se reciben de media un total de 57´4 millones de turistas internacionales anualmente, con una evolución significativa y una tendencia que ha ido creciendo desde el año 2000 hasta el año 2019 (obviando el año 2008-2009 debido al impacto de la crisis económica) y con una caída significativa del 77´3% en los dos últimos años de 2020 y 2021 con respecto al año anterior debido al impacto de la pandemia y sus consecuencias. La cantidad máxima de turistas internacionales se experimentó en el año 2019 con un crecimiento del 79´9% con respecto al inicio de periodo, y una tasa interanual del 0´84% con respecto al año anterior. España tiene gran afluencia de turistas internacionales que proceden de partes de todo el mundo, y los turistas europeos prevalecen frente al resto debido no solo a la proximidad geográfica entre las regiones sino también a la utilización de una moneda común, lo que facilita a los turistas tener una mejor percepción de los precios y da lugar a una mayor facilidad para viajar. De entre los diferentes países de residencia de turistas europeos que viajan a España, destaca Francia, con un 18´67% de turistas, Alemania con el 16´71% y Reino Unido con el 13´8% ([5]).

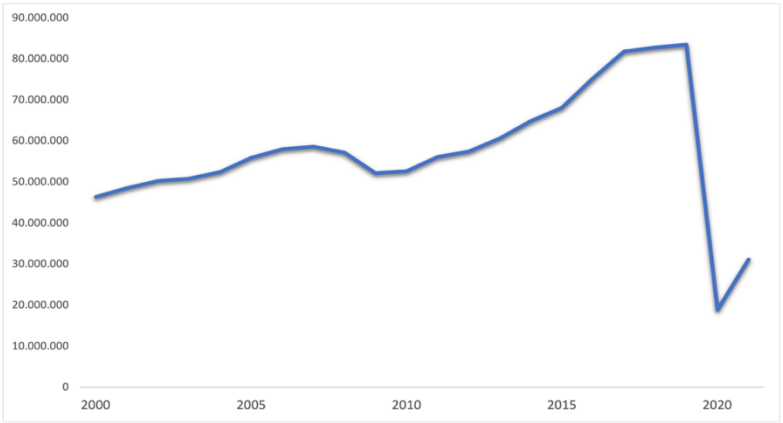

Figura 1.1. Evolución del turismo internacional anual desde el 2000 hasta el 2021.
Fuente de datos: Datosmacro ([5]). Elaboración propia.

[2] Los siguientes datos han sido extraídos de ([5]).

País	Turistas anuales (en millones)
Francia	5.822.671
Alemania	5.208.894
Reino Unido	4.302.634
Países Bajos	2.048.853
Italia	1.703.423
Bélgica	1.464.091
Portugal	1.193.649
Suiza	945.710
Estados Unidos	797.844
Irlanda	631.314
México	249.732
Brasil	139.937
Canadá	139.449
Turquía	139.427
Rusia	134.242
Emiratos Árabes Unidos	112.888
Israel	79.908
Argentina	64.287
Arabia Saudita	62.267

Figura 1.2. Principales países de procedencia de los turistas que llegan a España (2021). Fuente de datos: Datosmacro([5]). Elaboración propia.

El turismo cuenta con diversas actividades que implican la prestación de servicios a las personas que visitan un lugar. España es un país muy atractivo debido a su diversa cultura, geografía e historia, lo que lo convierte en un destino vacacional muy favorable para visitantes de todo el mundo e incluye numerosas actividades culturales, artísticas, y comercios especializados. Debido a esta diversidad, podemos distinguir dos tipos principales de turismo en la región:

- **Turismo cultural.** España es conocido por ser una gran fuente de historia y cultura ya que cuenta con fuertes tradiciones arraigadas y un gran patrimonio que incluye importantes monumentos, obras de arte y reliquias. Los extranjeros que viajan al país con motivo cultural buscan aprender y sumergirse en la historia del lugar y vivir experiencias únicas trasladándose a diferentes épocas. Este tipo de turismo tiene como misión promover la cultura y mantener las costumbres e instituciones ([6]).

  España cuenta con alrededor de 30400 bienes de interés cultural, de los cuales 48 han sido reconocidos como Patrimonio de la Humanidad convirtiéndose en el

tercer país del mundo con más lugares declarados Patrimonio de la Humanidad. Dentro de esta categoría turística, se encuentran ciudades como Córdoba y Toledo que reflejan la historia de España mediante la influencia de distintas épocas. Como podemos apreciar en el siguiente gráfico, el turismo cultural es cada vez más relevante y presenta una significante evolución en los últimos años ([7]).

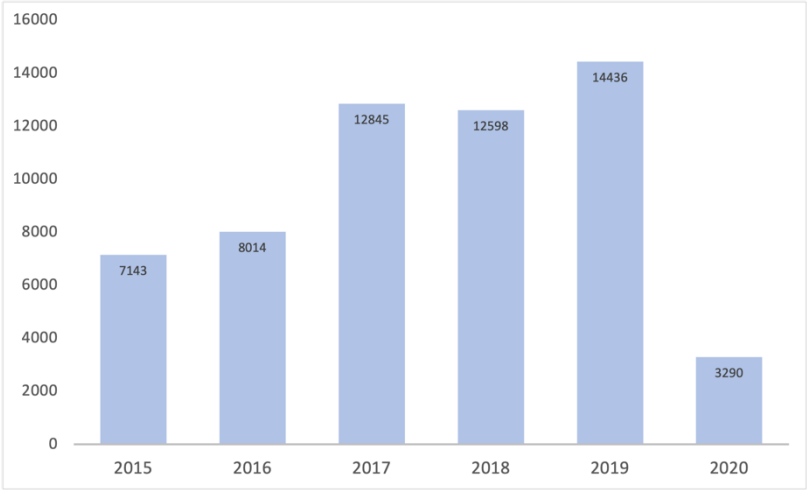

Figura 1.3. Número de turistas internacionales que visitaron España por motivos culturales entre 2015 y 2020 (en miles). Fuente de datos: Statista([8]). Elaboración propia.[3]

- **Turismo de sol y playa**. Este tipo de turismo es el preferido por los turistas extranjeros, pues supone un 75% de la oferta y demanda turística del país. Se caracteriza por la realización de actividades recreativas en la zona del litoral mediterráneo, que debido a sus favorables condiciones climáticas supone una opción de gran atractivo para los turistas. Este tipo de turismo se caracteriza por tener mayor afluencia en los meses de verano, y es una opción muy popular. En las costas de España, el turismo de sol y playa supone un 70´5% del total del PIB y un 62% del total de empleo directo, siendo las ciudades del mediterráneo como Valencia y las Islas los destinos más concurridos ([9]). De hecho, los destinos españoles de las islas y el mediterráneo, generan por sí mismos unos ingresos superiores a los de sus respectivos competidores del sector en turismo de sol y playa.

[3] Los datos de 2020 no son representativos pues se ven afectados por el cierre de fronteras y restricciones de movilidad como consecuencia a la llegada de la pandemia del COVID-19, lo que explica la disminución del turismo general.

Sin embargo, no solo es importante el número de turistas que llegan a España y su origen, también lo es el impacto económico que generan en el país. Los turistas que llegan a España suponen una fuente de ingresos elevada debido al gasto turístico generado durante su estancia en actividades de ocio, consumo, gastronomía y alojamiento principalmente. El gasto turístico es una variable muy relevante a la hora de analizar el turismo y su impacto un país, pues es lo que genera los ingresos no solo para el sector, sino para toda la economía del país. La Encuesta de Gasto Turístico (EGATUR) se creó en el año 2002 fundada en colaboración entre el Instituto Nacional de Estadística (INE) y el Banco de España con el objetivo de analizar y medir dicho gasto turístico incurrido por los visitantes no residentes durante su visita, pero también para medir el gasto turístico generado por los miembros residentes en los viajes que realizan a otros países en el extranjero ([10]).

Según esta encuesta, (datos extraídos de ([11])) encontramos que el gasto total medio de los turistas no residentes en España durante noviembre de 2022 fue de 5387´32 millones de euros siendo el gasto total medio por turista de 1241€ y el gasto medio diario por turista de 165€. Por otro lado, si desglosamos el gasto medio según el origen de procedencia de los turistas, encontramos que, en el año 2021, aquellos visitantes procedentes de Alemania serían los que aportaron un nivel de ingresos superior, siendo su gasto total de 6034´58 millones de euros, seguidos por Reino Unido, con un gasto total de 4773´64 millones de euros.

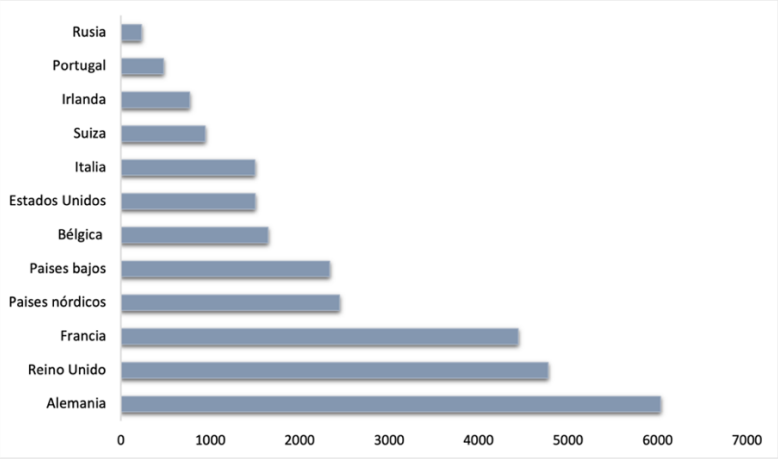

Figura 1.4. Gasto total de los turistas visitantes en España por país de origen en 2021 (en millones de euros). Fuente de datos: Statista(ver[12]). Elaboración propia.

1.2. Impacto de la crisis del COVID-19 sobre el turismo español

El origen del Covid-19 se remonta a finales del año 2019, donde por primera vez en China se detectaron numerosos casos en un mercado situado al lado de un relevante laboratorio en Wuhan. Se trata de una enfermedad infecciosa con síntomas normalmente moderados que en algunos casos pueden resultar en grandes complicaciones derivando incluso la muerte en personas con patologías previas o sistemas inmunológicos más débiles como las personas de tercera edad. La alta capacidad de contagio del virus hizo que este se expandiera a gran velocidad llegando a principios de 2020 a Europa y otros países del mundo. En marzo de este mismo año, la situación ya descontrolada por la expansión del virus hizo que la OMS (Organización Mundial de la Salud) lo declarase situación mundial de pandemia. En España, el gobierno ejecutó el estado de alarma con el objetivo de incrementar la seguridad y protección en el país, de tal forma que se produjo el cierre de fronteras y de varios negocios, así como el aislamiento total de la población en sus hogares. Estas mismas medidas se ejecutaron en mayor o menor medida en numerosos países para evitar la propagación del virus.

En este periodo de confinamiento de la población, que duró hasta varios meses, todo tipo de actividad económica salvo los bienes y servicios de primera necesidad, quedaron cancelados. El sector turístico, fue sin lugar a duda uno de los más afectados por la pandemia. Esto se debe a la fragilidad de la industria y su dependencia de la economía, que se ve gravemente afectado ante situaciones de crisis donde, además, debido esta situación, las restricciones y los efectos de la pandemia se alargaron durante el año completo y parte del siguiente como consecuencia del cierre de fronteras y las numerosas restricciones de desplazamiento entre comunidades autónomas del país. Otros factores igual de relevantes que afectaron al sector como consecuencia de los acontecimientos, fueron el miedo al contagio, las restricciones de aforo en lugares públicos y la disminución de la renta de la población como consecuencia del cierre de negocios y el parón económico durante los meses de confinamiento.

Las restricciones se iban ajustando por el gobierno y según la comunidad autónoma en función de los nuevos contagios con el objetivo de controlar la expansión del virus y aliviar la saturación de los hospitales. En España, hubo restricciones muy duras respecto

a la movilidad entre provincias, horarios con toque de queda y cierre de comercios y locales cuya actividad no fuera esencial como por ejemplo el ocio nocturno. Es importante destacar que, aunque en toda Europa había restricciones, España fue uno de los países europeos donde las medidas tomadas fueron más estrictas. Pese a que estas restricciones buscaban el bienestar y la buena salud de la población, en España se registraron a principios de 2021, dos millones de contagios y 51000 fallecimientos ([13]), que, sumado a los efectos sobre la situación económica, dieron lugar a una gran crisis para el país. Como se ha mencionado anteriormente, en España, el turismo es el motor económico principal, con una gran contribución sobre el PIB, por lo que el cierre de fronteras y las restricciones de movilidad, hicieron que este sector se viera gravemente afectado y por consiguiente, la situación general económica del país. Como consecuencia, la pandemia afectó tanto a la demanda, debido al menor poder adquisitivo de los turistas, además de las restricciones impuestas y del riesgo al contagio como a la oferta, debido a un incremento de los costes y a la reducción de la capacidad de los servicios (debido a las medidas de distanciamiento de seguridad). Estos dos factores combinados, dieron lugar a un significante cambio sobre la estructura de precios del sector, que también se vio afectada por la situación ([14]).

En este contexto, la contribución del sector turístico sobre el PIB español sufrió una caída de un 11´3% en 2020, ya que durante los primeros meses del año 2020 (de marzo a junio), la llegada de turistas extranjeros disminuyó en un 77% ([15]), convirtiéndose en el peor año de la historia del sector lo que supuso una pérdida de 43´4 millones de euros en actividad económica, que es un 88% por debajo del mismo periodo el año anterior ([16]).

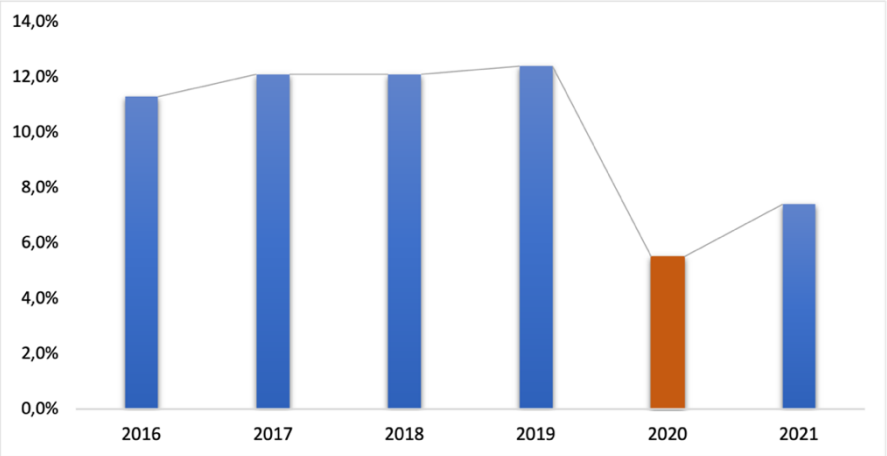

Figura 1.5. Evolución del porcentaje del PIB aportado por el sector turístico en España 2016-2021. Fuente de datos: Statista([17]). Elaboración propia.

Como es evidente, este fenómeno tuvo consecuencias adicionales, entre ellos, el aumento de la tasa de desempleo del país y el cierre de muchos negocios del sector. En este contexto, según el EPA (Encuesta de Población Activa), el empleo aportado por sector turístico pasó de un 13´4% en 2019 a un 11´6% en 2020, una caída del 1´8%, que tuvo graves consecuencias implicando el despido de muchas personas y la creación de ERTES (expediente de regulación temporal de empleo) llegando a alcanzar 1´1 millones. Sin embargo, según Eurostat, los trabajadores en ERTE se clasificarían como ocupados, pues existe la garantía de su reincorporación al trabajo una vez se reanude la actividad y se finalice el periodo de suspensión ([18]).

Otra variable gravemente afectada por esta situación fue el gasto turístico. Como consecuencia del descenso de turismo exterior durante la pandemia y de la pérdida de poder adquisitivo, el nivel de gasto turístico sufrió una caída importante descendiendo un 78´5% en 2020 con respecto al año anterior ([19]).

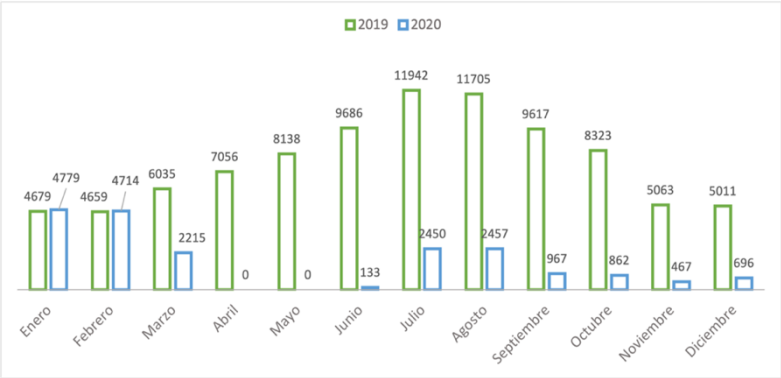

Figura 1.6. Comparativa gasto turístico 2019-2020 (en millones de euros). Fuente: INE ([19]). Elaboración propia

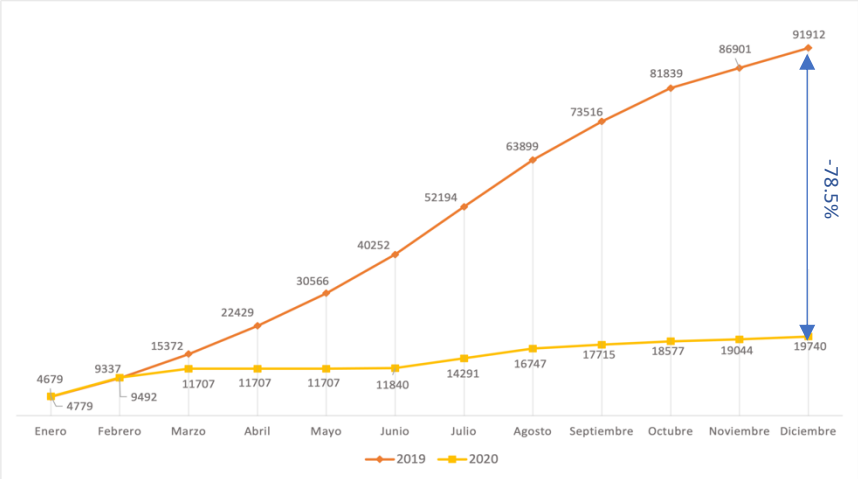

Figura 1.7. Comparativa acumulada del gasto turístico internacional 2019-2020 (en millones de euros). Fuente de datos: INE ([19]). Elaboración propia.

Sin embargo, a pesar de la situación de incertidumbre y descenso económico vivido los dos últimos años con la pandemia, el año 2022, presentaba perspectivas optimistas dando lugar a una recuperación económica para el sector, debido a las cifras alcanzadas en este periodo y las perspectivas de crecimiento de los consiguientes años que analizaremos en el siguiente apartado.

1.3. Situación actual y perspectivas para el sector turístico en España

La pandemia ha dejado una situación complicada para el sector del turismo en España. Tras los malos resultados del año 2020, se pudo apreciar una relevante pero pequeña recuperación en el siguiente año, 2021, que, según los datos publicados por el Instituto Nacional de Estadística, el gasto total de turistas no residentes en España durante este año fue de 34´816 millones de euros, lo que supone un incremento del 76% respecto al año 2020 ([19]). Esta recuperación se debe en parte a la relajación de las medidas de seguridad impuestas por el gobierno con respecto a la pandemia. Aun así, se puede apreciar que las cifras experimentadas en este periodo son lejanas a las que se presentaban en periodos anteriores previos a la pandemia. Esto, se debe a factores como la inseguridad ante la posibilidad de contagio y la crisis económica. Como consecuencia de estos hechos, aparece un nuevo factor determinante para el sector: la seguridad sanitaria, donde los turistas buscan su seguridad para evitar el contagio y disminuir la exposición ante el virus.

Sin embargo, puesto que el turismo es uno de los factores principales de la economía española, su recuperación es crucial y por ello ha sido necesario la intervención del Gobierno para establecer una serie de medidas que ayuden a favorecer la recuperación del sector, aunque garantizando la seguridad al mismo tiempo. Es por ello, que, para viajar, se impuso de forma obligatoria la realización de pruebas de detección de Covid-19 y la adopción de una serie de protocolos y mecanismos sanitarios como el uso obligatorio de mascarilla en los transportes y otros lugares públicos. Por otro lado, el Gobierno ha tratado de favorecer al sector mediante ayudas económicas, laborales y fiscales durante todo el periodo de la pandemia hasta ahora como por ejemplo mediante la inyección de liquidez en las empresas mediante diferentes mecanismos, el aplazamiento del pago de impuestos o la ampliación de la línea ICO del sector turístico[4]. Al mismo tiempo, se ha querido impulsar el turismo en las regiones más afectadas por la pandemia mediante medidas urgentes y financiación de emergencia para la promoción del turismo local mediante bonos de patrocinio[5] ([20]).

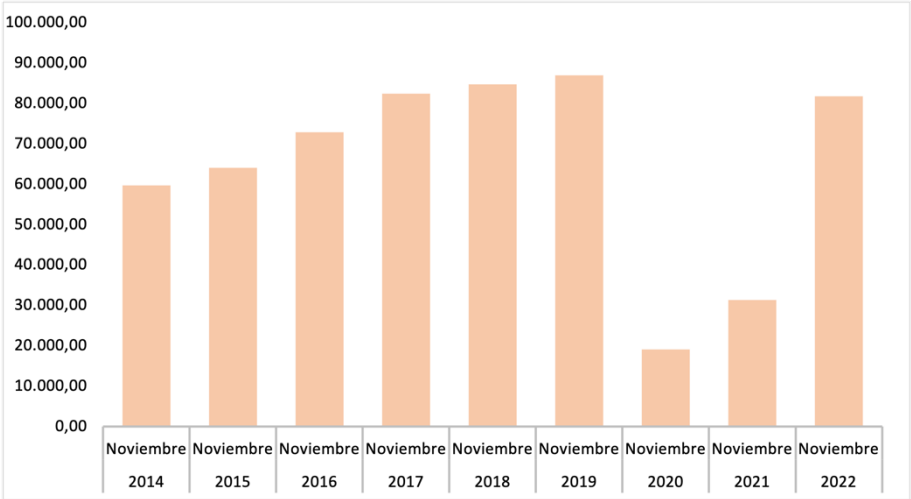

Figura 1.8. Evolución del gasto total acumulada de turistas extranjeros en España (en millones de euros). Fuente de datos: INE. Elaboración propia. ([21])

[4] La línea ICO del sector turístico es una medida impuesta por el gobierno para aliviar el impacto de la pandemia con el objetivo de facilitar la financiación de autónomos, pymes y empresas del sector garantizando así la liquidez.

[5] Los bonos de patrocinio son bonificaciones facilitadas a los turistas, sujeto al cumplimiento de ciertos requisitos, con el objetivo de cubrir determinados gastos relacionados con el viaje como las pernoctaciones hasta un importe máximo. El objetivo de estos bonos es incentivar y reactivar el turismo en el país.

Estas medidas nos llevan a un escenario más positivo en el año 2022, donde sí que se puede apreciar un crecimiento significativo del sector, presentando cifras muy superiores a los años del periodo de pandemia y más cercanas a las de 2019 previas al inicio de la esta. Según el Ministerio de Industria, Comercio y Turismo, el año 2022 será recordado como "el año de recuperación del turismo tras la pandemia" pues en este año, el turismo consiguió alcanzar cifras superiores a los años anteriores y más cercanas al contexto pre-pandemia. En concreto, España recibió en noviembre de 2022 aproximadamente un total de 4´34 millones de turistas internacionales, lo que supone un incremento del 29´2% con respecto al mismo mes del año anterior ([22]). Por otro lado, el gasto turístico internacional para el año 2022, alcanzó durante los 11 primeros meses del año los 81´821 millones de euros, lo que supone un incremento del 160´9% con respecto al mismo periodo del año anterior. Sin embargo, aunque estas cifras son positivas y muestran indicios de recuperación para el sector con respecto a los dos últimos años donde se produjo una pérdida en la actividad de más de 160.000 millones de euros, la industria está aún lejos de alcanzar las cifras obtenidas durante los años anteriores a la pandemia. Es por ello, que es de vital importancia que no se ignore el peso que tiene este sector sobre la economía, y que se sigan implementando medidas de protección mediante programas efectivos de recuperación del turismo en las diferentes regiones del país ([23]).

Finalmente, es importante mencionar también, que, pese al inicio de la recuperación del sector turístico tras las duras consecuencias producidas por los efectos de la pandemia, durante el año 2022, hubo otros factores que impactaron sobre la economía del país como es el caso del comienzo de una guerra entre Rusia y Ucrania. Este fenómeno, ha tenido graves impactos a nivel europeo agravando la crisis ya iniciada en la pandemia como consecuencia de la falta de suministro energético, lo que ha supuesto un crecimiento de la inflación impactando sobre el nivel general de precios. En concreto, esto es relevante para el sector turístico pues el incremento del precio sobre los combustibles supondrá un incremento sobre el gasto de viaje, debido a un mayor coste de los transportes, especialmente el aéreo. Por otro lado, otra consecuencia bilateral de la guerra sobre el turismo español tiene que ver con la disminución de turistas procedentes de Rusia y Ucrania, lo cual es de gran relevancia considerando que los turistas procedentes de Rusia generan un elevado gasto turístico por persona en comparación con otros visitantes extranjeros. Según los datos de la encuesta EGATUR 2021, el gasto turístico por viajero procedente de Rusia fue de 1704€, en comparación con 1119€ de gasto medio de turistas

internacionales procedentes de otros. Sin embargo, pese a las circunstancias expuestas, podemos afirmar que se espera una situación de esperanza y recuperación para el sector, que, debido a la relajación de medidas de seguridad, a la disminución de los contagios como consecuencia de las vacunas y a las ayudas aportadas por el Gobierno, se verá beneficiado de un crecimiento significativo en los próximos años llegando a alcanzar las cifras pre-pandemia o incluso superarlas. De hecho, según los datos aportados por EXCELTUR, a pesar de la aparente situación de recesión global en la economía, en 2023 se podría alcanzar un valor de 168 mil millones de euros en la actividad turística, lo que supone un incremento del 7´1% del PIB español con respecto a los datos de 2019, cumpliendo así con una completa recuperación contra los efectos de la pandemia ([24]).

1.4. Antecedentes al estudio

Como se ha mencionado en los apartados anteriores, el turismo es una industria con potencial de crecimiento y con gran impacto en la economía española. Por este motivo, la industria turística posee gran cantidad de datos e información generados a través de diferentes fuentes lo que junto con el desarrollo de las tecnologías y gracias a técnicas de *machine learning*, hoy en día se pueden realizar exhaustivos análisis a partir de estos datos para poder extraer conclusiones que permitan mejorar diferentes aspectos de la industria como la segmentación de clientes o predecir la demanda para un periodo. Los estudios en este campo se han centrado en una variedad de aplicaciones como el análisis de sentimientos, el reconocimiento de imágenes y el modelado predictivo entre otros, para obtener información y mejorar la toma de decisiones para las empresas turísticas. La integración de ML en el turismo se considera como una de las innovaciones clave que darán forma al futuro de la industria.

En este contexto, encontramos estudios como ([25]) donde se utilizan modelos predictivos con el objetivo de analizar el gasto turístico en España a partir de datos de EGATUR. La autora modela diferentes algoritmos de predicción como Regresión lineal o *Random Forest* ajustando los hiperparámetros correspondientes para cada uno de ellos con el objetivo de identificar cuál se ajusta mejor a los datos y genera un resultado más preciso. En este caso, se obtuvo un mejor resultado en los algoritmos de SVM (Máquina de vectores de soporte) y DL (Aprendizaje profundo). Por otro lado, se analizó también la importancia de las variables donde se obtuvo en primer lugar el número de

pernoctaciones y en segundo lugar el país de origen. En un segundo estudio, ([26]), se utiliza *machine learning* para analizar cuáles son los factores que más afectan a la demanda turística. En este caso, se entrenan varios algoritmos y se compara el resultado obtenido comparando dos modelos: Regresión Lineal y *Random Forest* y a partir de las métricas de *Root Mean Square Error* (RMSE) y R^2 para determinar en qué medida las diferentes variables incluidas en el modelo pueden contribuir mejor a la predicción de la demanda turística. La conclusión obtenida fue que el clima, la estabilidad política, el precio y especialmente los gastos de promoción fueron las variables más significativas para la demanda turística siendo el modelo de *Random Forest* el que obtuvo mejores predicciones y mejor ajuste.

Por otro lado, en ([27]), se implementaron técnicas de *machine learning* para estudiar el impacto del Covid sobre el sector del turismo en Italia. Los datos en este artículo se obtuvieron de repositorios como Google Trends, Drago y Hoxhalli. En este caso, se usó un algoritmo de validación de agrupamiento o *clustering validation* conocido como *K-means* con el objetivo de evaluar la calidad del agrupamiento realizado por el algoritmo de *clustering* donde se han clasificado varias regiones de Italia en tres grupos. Para dicha validación se usaron índices como la suma del error al cuadrado y el coeficiente de *Silhouette*. El resultado obtenido fue que las regiones del centro y norte de Italia son las más afectadas por los efectos de la pandemia mientras que las regiones del sur son menos afectadas. Sin embargo, el estudio concluye con que, a pesar de la situación, no se ha perdido el interés por ir de vacaciones entre los ciudadanos siendo las regiones del norte las que más temprano se recuperarán de la crisis de la pandemia.

Como hemos mencionado previamente, gracias a las técnicas de *machine learning*, las empresas turísticas pueden ganar *insights* relevantes para la toma de decisiones y llevar a cabo estrategias. En ([28]), se estudia el impacto económico que tienen las empresas turísticas de cruceros sobre la industria en España. Con datos obtenidos a partir de EGATUR, se hace una correlación con los datos de las propias empresas de cruceros, llevando a cabo un análisis de regresión para medir la relación entre rentabilidad de las empresas turísticas locales de los diferentes destinos como consecuencia de la afluencia de visitantes generada por los cruceros.

Finalmente, en otro estudio ([29]), se aplican técnicas de *machine learning* con el objetivo de predecir el volumen turístico por comunidad autónoma en España, teniendo en cuenta el impacto generado por la pandemia del Covid-19. Para ello, se emplean por un lado datos del periodo de pandemia para predecir el número de fallecimientos en los próximos meses, y por otro, se usan los datos históricos del turismo en España por comunidad autónoma extraídos del INE (Instituto Nacional de Estadística), para que, teniendo en cuenta el resultado del análisis anterior, predecir el volumen turístico del año por comunidad autónoma. Para cada uno de los análisis se emplean técnicas de *machine learning* diferentes, se comparan y se escoge aquellos modelos que mejor se ajustan para cada *dataset.* Para el análisis relacionado con la predicción de las defunciones, el modelo de *Random Forest* obtuvo mejor precisión mientras que para el segundo análisis, se emplea el algoritmo de KNN.

A pesar de que estos estudios se aproximan en cierto modo al análisis que se va a llevar a cabo en este trabajo, ninguno de estos trabajos contienen un análisis comparativo entre el periodo pre-pandemia y post-pandemia utilizando técnicas de *machine learning* capaces de predecir una variable objetivo ya que, los estudios que contienen análisis de *machine learning* no tratan temas relacionados con el impacto del Covid-19 y el turismo, y los que tratan este tema, no lo abordan mediante técnicas de *machine learning*.

1.5. Objetivos

Descritos los apartados anteriores, se procede a definir los objetivos comprendidos en este estudio.

- Analizar las variables más influyentes para el gasto turístico antes y después de la pandemia
- Estudiar correlación entre las variables principales y el gasto turístico para extraer aquellas que mejor explican la variable objetivo "gastototal"
- Encontrar los modelos de regresión y clasificación que mejor modelizan el gasto turístico teniendo en cuenta las variables más explicativas seleccionadas
- Evaluar el desempeño de los modelos construidos seleccionando un subconjunto de datos determinado
- Encontrar un modelo usando *imbalance* para predecir el turismo de lujo en turistas cuyo gasto supera los 3000€

1.6. Metodología

Em este trabajo se realizará un estudio sobre el impacto de la crisis del Covid-19 en el sector turístico español, haciendo una comparativa de datos respectivos a periodos pre y post-pandemia respectivamente para estudiar el impacto sobre la variable gasto turístico. Para ello, se van a extraer datos de la principal fuente respecto al gasto turístico en España, la encuesta turística EGATUR, correspondientes al periodo inmediatamente anterior e inmediatamente posterior a la pandemia. Se emplearán además dos herramientas principales de trabajo para tratar y manipular los datos obtenidos: Python y Matlab y diferentes algoritmos y técnicas de *machine learning* tanto de regresión como de clasificación. Se utilizarán también diferentes soportes como repositorios de programación, prácticas y códigos proporcionados por el tutor, y bibliografía relevante para soportar los argumentos e información del estudio.

1.7. Estructura del TFG

El siguiente trabajo está compuesto por cuatro capítulos principales. El primer capítulo consiste en una contextualización inicial del estudio que se va a llevar a cabo donde se presentan los datos relevantes de la industria, así como estudios precedentes relevantes para la realización de este trabajo. En el segundo capítulo se lleva a cabo un análisis preliminar de los datos y se estudian procesos de selección de variables que permitirán simplificar procesos posteriores. A continuación, en el capítulo tres, se describe la elaboración de modelos de *machine learning* construidos bajo diferentes técnicas y procesos tanto de regresión como de clasificación y se evalúa su desempeño mediante diferentes enfoques y medidas. En el capítulo cuatro se ha realizado un estudio de *imbalanced data* con el objetivo de construir modelos de clasificación capaces de predecir el turismo de lujo en función de aquellos turistas con gastos muy elevados. Finalmente, este trabajo se compone de un último capítulo donde se recopila la información más relevante descrita a lo largo del trabajo y se exponen las conclusiones principales obtenidas a partir del análisis realizado en los capítulos anteriores.

Capítulo 2

DESCRIPCIÓN DE LOS DATOS Y SELECCIÓN DE VARIABLES

2.1. Justificación de los datos

Para llevar a cabo nuestro análisis sobre el impacto de la pandemia de 2020 causada por el Covid-19 sobre el gasto turístico español, vamos a utilizar un conjunto de datos extraídos de la fuente de información EGATUR. Esta base de datos pertenece al gobierno español, en concreto del Ministerio de Comercio y Turismo, y son encuestas llevadas a cabo por el Instituto Nacional de Estadística (INE), que permite acceder a datos y cifras relacionadas con el gasto turístico que realizan visitantes extranjeros en España, recogiendo la principal información y características del viaje.

En nuestro caso concreto, vamos a seleccionar para nuestro estudio los datos referentes al gasto turístico internacional en el periodo de temporada alta de vacaciones que coincide con los meses de verano en España: junio, julio y agosto. Por este motivo, usaremos dos bases de datos separadas y trabajaremos por un lado con datos correspondientes al periodo comprendido entre junio y agosto de 2019 (datos pre-pandemia), y por otro, usaremos los datos para este mismo periodo del año 2022 (datos post-pandemia), con el objetivo de analizar el impacto generado sobre la variable del gasto turístico por el efecto de la pandemia durante los años de 2020 y 2021.

De forma preliminar, se puede observar en la Figura 2.1 que en 2019 existe una distribución superior de observaciones, o lo que es lo mismo, un mayor número de turistas extranjeros que visitaron España ese año en comparación con 2022, donde la cifra de visitantes difiere de los valores pre-pandemia, siendo el volumen de turistas un 10% inferior aproximadamente. Esto demuestra la prolongación en el tiempo por los efectos de la pandemia y la actual crisis económica no solo del país, sino de toda Europa como consecuencia de otros factores como la guerra de Ucrania.

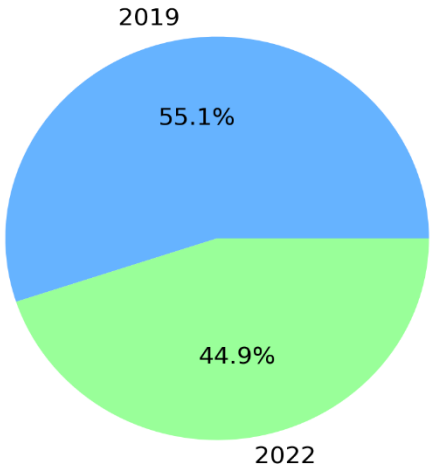

Figura 2.1. Distribución de turistas extranjeros en España en 2019 vs 2022.
Elaboración propia.

A pesar de que vamos a utilizar dos conjuntos de datos diferentes para poder hacer una comparativa, al descargar todos los datos de la misma fuente de datos (EGATUR), estos poseen la misma estructura y tipología, por lo ambas bases de datos poseen una estructura de variables equivalente**.** En primer lugar, es importante analizar las variables en cuanto a significado y tipología lo que nos permitirá entender mejor el significado de los datos para su posterior tratamiento.

Variable	Interpretación
mm_aaaa	Fecha en formato mes y año
A0	Encuesta de procedencia de los datos
A0_1	Identificador de cuestionario
A0_7	2:Turista no residente en no tránsito y 8: turista no residente en tránsito
A1	Método de transporte empleado siendo 1: carretera, 2: aeropuerto, 3: puerto, 4: tren
País	País de procedencia del turista con valores 01:Alemania. 02:Bélgica. 03:Francia. 04: Irlanda. 05: Italia. 06: Países Bajos. 07: Portugal. 08: Reino Unido. 09: Suiza. 10:Rusia. 11: Países Nórdicos (Dinamarca, Finlandia, Noruega, Suecia). 12: Resto de Europa. 13:
ccaa	Comunidad autónoma de destino de viaje con valores 01: Andalucía. 02: Aragón. 03: Principado de Asturias. 04: Illes Balears. 05: Canarias. 06: Cantabria. 07: Castilla y León. 08: Castilla-La Mancha. 09: Cataluña. 10: Comunitat Valenciana. 11: Extremadura. 12: Galicia. 13: otros
A13	Total de pernoctaciones
aloja	Tipo de alojamiento con valores 1: Hoteles y similares, 2: Resto de mercado, 3: Alojamiento no de mercado
motivo	Motivo principal del viaje con valores 1: Ocio/vacaciones, 2: Negocios, 3: Resto
A16	Si se ha contratado algún paquete turístico siendo 1: Sí, 6: No
gastototal	Gasto total incurrido por el turista durante el viaje
factoregatur	Factor de elevación de la encuesta EGATUR

Tabla 2.1. Relación de variables, tipología y significado. Fuente de datos: EGATUR.
Elaboración propia

Por otro lado, es importante realizar un análisis preliminar para obtener medidas estadísticas básicas como la media o la desviación típica e identificar cuántas observaciones y variables tenemos con el fin de obtener una aproximación sobre la estructura de los datos.

En primer lugar, vamos a realizar un análisis previo de la estructura de los datos para obtener una idea preliminar de los mismos. El *dataset* correspondiente a 2019 está compuesto por 26370 observaciones y 13 variables que se corresponden con las descritas en la Tabla 2.1. Sin embargo, para trabajar con los datos, se han eliminado las variables "mm_aaaa", "A0", "A0_1" y "factoregatur" ya que no nos aportan una información relevante para nuestro análisis y de esta forma obtendremos menos complejidad computacional.

A continuación se han obtenido las medidas estadísticas principales para cada una de las variables para cada año con el objetivo de comprobar su distribución:

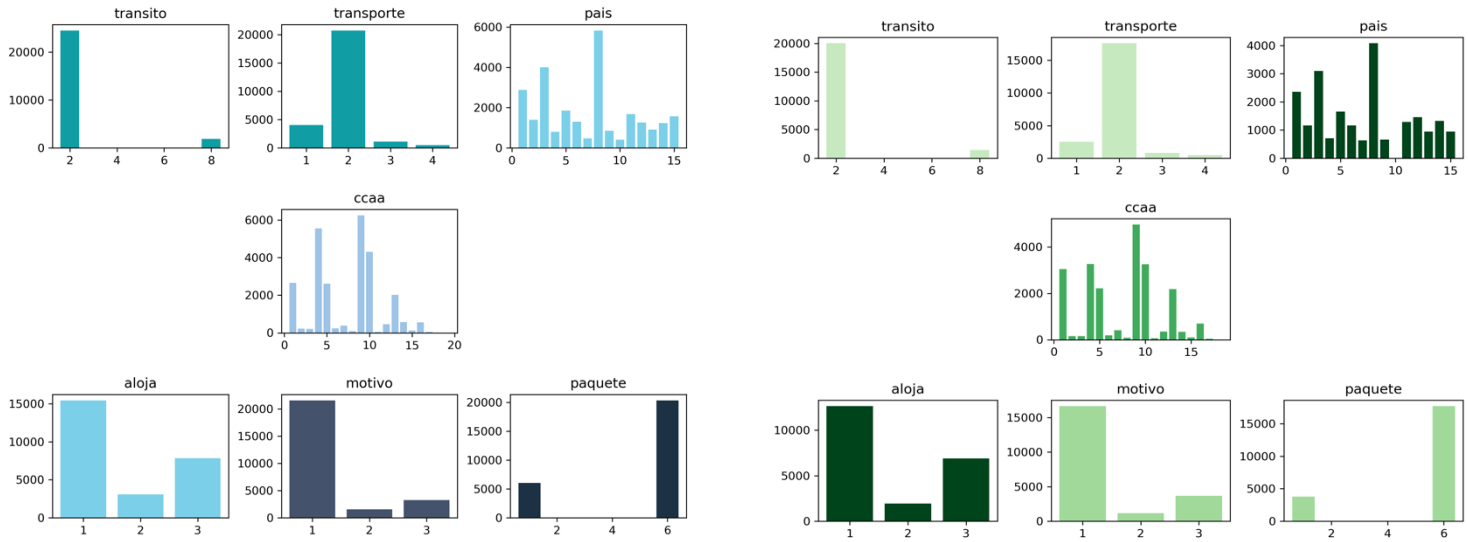

Figura 2.2. Análisis descriptivo de las variables categóricas de 2019 vs 2022. Elaboración propia

	pernoct	gastototal
mean	9.33	1297.97
std	10.19	972.85
min	1.00	40.45
max	181.00	24481.45

	pernoct	gastototal
mean	9.62	1452.24
std	11.94	1247.65
min	1.00	17.88
max	181.00	22307.97

Tabla 2.2. Medidas estadísticas principales de las variables cuantitativas 2019 vs 2022. Elaboración propia

Como podemos observar, puesto que la mayoría de las variables son de tipología categórica, no tiene tanto sentido obtener medidas estadísticas cuantitativas como la media, el mínimo o el máximo pues en general no aportan demasiada información sobre dichas variables. Por este motivo, en las variables cualitativas nos centramos en analizar la distribución en el número de categorías existentes para cada variable. Se puede apreciar en la Figura 2.2. que en general el patrón es similar para ambos periodos donde coincide la tendencia de las variables analizadas siendo el país con mayor volumen de visitantes Gran Bretaña (8), seguido de Francia (3). También se puede observar que, entre los tipos de alojamiento, predomina la categoría de hoteles frente a otro tipo de alojamientos vacacionales, siendo el ocio/vacaciones el motivo principal de viaje de los turistas extranjeros en nuestro país. Además, los destinos con mayor concurrencia por turistas se encuentran en las comunidades autónomas de Cataluña (10) e Islas Baleares (5).

Por otro lado, en las variables cuantitativas, como vemos en la Tabla 2.2, observamos en cuanto al número de pernoctaciones (A13), que los turistas realizan una media de 9 noches de estancia en sus viajes y se gastan alrededor de 1298€ en 2019 y 1452€ en 2022. Esto nos muestra un primer impacto del efecto de la pandemia y su consecuente crisis económica en el país y en el sector, que se ve reflejada en la subida de precios en la hostelería ya que, en el año 2022 por el mismo número de noches, los turistas gastan unos 155€ más de media que en 2019.

A continuación, se ha estudiado la correlación existente entre las variables. Como se puede observar en la Figura 2.3 obtenemos que las variables más correlacionadas entre sí son, en primer lugar, "gastototal" con la variable "A13" (total de pernoctaciones) , donde observamos una fuerte correlación y, en segundo lugar, "A16" (contratación o no de paquete turístico) con la variable "aloja" o alojamiento.

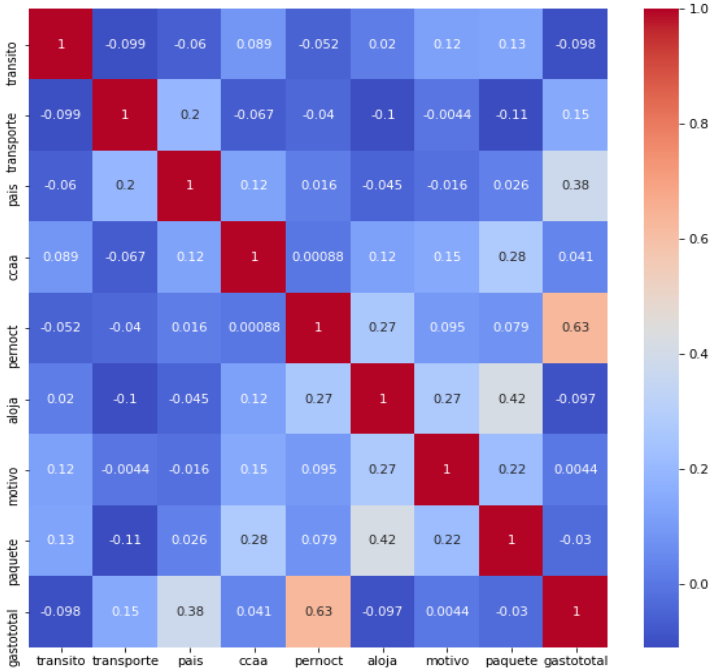

Figura 2.3. Correlación entre variables, Datos 2019. Elaboración propia

En el caso de los datos de 2022, se han obtenido un total de 21497 observaciones y 13 variables (la tipología y número de variables son equivalentes a los datos de 2019). De nuevo, para trabajar con los datos, se han eliminado las variables "mm_aaaa", "A0", "A0_1" y "factoregatur" al igual que en el *dataset* de 2019, ya que estas variables no nos aportan una información relevante para nuestro análisis y de esta forma obtendremos menor complejidad computacional.

En cuanto a la correlación de las variables en 2022, encontramos que al igual que en 2019, las variables con mayor correlación son en primer lugar, "gastototal" con "A13" o número de pernoctaciones y en segundo lugar, las variables con mayor correlación son "aloja" con "A16" (contratación o no de paquete turístico).

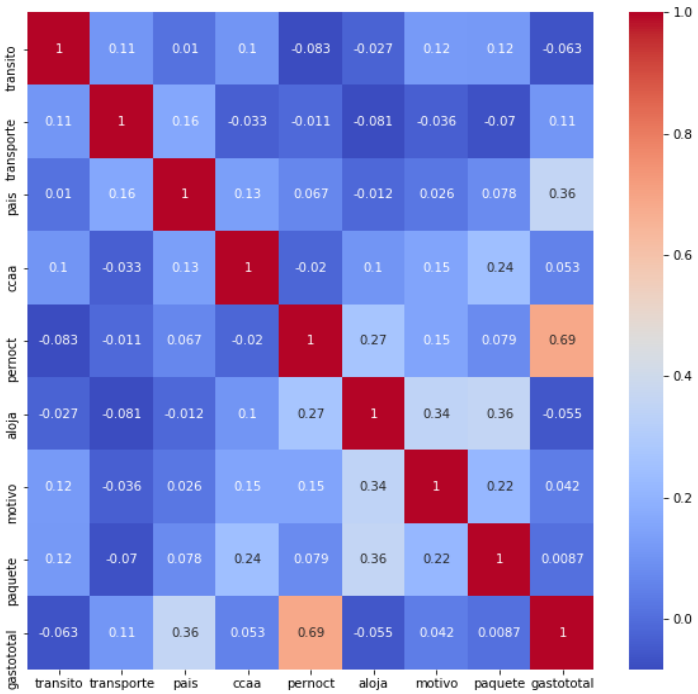

Figura 2.4. Correlación entre variables, Datos 2022. Elaboración propia

2.2. Selección de variables explicativas

2.2.1. Marco teórico

En esta primera parte de nuestro análisis, se busca encontrar el número de variables más para la variable objetivo seleccionada, con el objetivo de construir un modelo predictivo de *machine learning* a posteriori que sea capaz de predecir la variable "gastototal de la manera más precisa posible. El objetivo de este proceso consiste en reducir el número de variables que comprenden nuestros datos para facilitar los procesos posteriores de construcción del modelo de *machine learning* y sobre todo para conocer el impacto de cada una de las variables sobre la variable definida como variable objetivo o "*target*", que en este caso es el gasto turístico (variable "gastototal" en nuestro *dataset*). Para este proceso, vamos a utilizar dos técnicas estadísticas diferentes donde se comparará el resultado obtenido con cada método mediante una medida de cálculo del error cometido apropiada y se estudiará la información proporcionada por los mismos, escogiendo aquel que nos proporcione información más precisa. El objetivo de este análisis, por tanto, es conocer qué variables explican mejor nuestra variable objetivo "gastototal" disminuyendo así el número de variables de nuestro conjunto de datos, para reducir así el coste computacional y encontrar el mejor modelo predictivo posible.

2.2.2. *Feature Selection*

El primer método que se va a emplear para la selección de variables se conoce como *feature selection*. Este proceso consiste en elegir aquellas variables más relevantes dentro de un gran conjunto de variables para poder explicar de la forma más precisa posible la variable objetivo previamente definida. Para esto, debemos excluir aquellas variables que no aportan ningún tipo de información. Este proceso es crucial para poder construir a posteriori un modelo predictivo de *machine learning* ya que permite:

- **Reducir el sobreajuste u *overfitting***: este fenómeno se produce cuando un modelo es demasiado complejo, y por tanto se ajusta demasiado al *training set* o datos de entrenamiento pero es incapaz a posteriori de generalizar bien, o lo que es lo mismo, no captura el patrón de los datos correctamente y por tanto su desempeño frente a datos diferentes a los trabajados en el entrenamiento no es bueno a la hora de predecir datos nuevos. Mediante la eliminación de aquellas variables poco relevantes podemos reducir el sobreajuste de los datos.

- **Incrementar el rendimiento del modelo**: reducir el número de variables haciendo una selección previa, puede mejorar el rendimiento del modelo predictivo, incrementar la eficiencia y conseguir mayor precisión.

- **Simplificar la interpretación del modelo**: un modelo con un número menor de variables es más fácil de interpretar y comprender, lo que puede ser útil para tomar decisiones y explicar las predicciones del modelo.

- **Ahorrar recursos computacionales**: el uso de un menor número de variables, ayuda a disminuir la carga computacional que se necesita para entrenar y evaluar un modelo, lo que puede ser de gran relevancia para conjuntos de datos de gran volumen y complejidad.

Sin embargo, este método tiene la limitación de que solo captura la relación entre variables cuando existe dependencia lineal. Por este motivo, y con el propósito de evitar discriminaciones entre posibles relaciones entre las variables de tipo no lineal, en el

apartado siguiente se empleará otra técnica de selección de variables que a diferencia de *feature selection*, sí que es capaz de capturar dependencias no lineales entre las variables.

a) *Feature selection* - Datos de 2019

Una vez realizado el análisis preliminar anterior, procedemos a realizar *feature selection* usando estos datos para comprobar la dependencia entre las variables y qué variables son más explicativas en cuanto a nuestra variable objetivo "gasto turístico".

En primer lugar, para realizar *feature selection* hemos escogido el método "f-test", que es uno de los muchos que se pueden emplear. F-test permite realizar *feature selection* de forma más precisa cuando la variable objetivo es de carácter numérico ya que utiliza un sistema de puntuación en función del análisis de varianzas (ANOVA), con el objetivo de medir la dependencia lineal entre las variables mediante la comparación de medias entre grupos. Para ello, calcula un p-valor para cada variable indicando la probabilidad de que la hipótesis nula sea cierta, o lo que es lo mismo, que la variable estudiada no explique de forma relevante la variable objetivo. En nuestro caso, al realizar f-test hemos obtenido que las variables más relevantes y que mejor explican la variable "gastototal" según los datos de 2019, son principalmente la variable "A13" (número de pernoctaciones) con una puntuación f-test de 13868´98 (resultado que coincide con nuestro análisis preliminar de correlación entre las variables), seguida de la variable "pais" con una puntuación de 3590´19 y finalmente, aunque con menos peso, la variable A1 (medio de transporte). El resto de las variables han obtenido unas puntuaciones muy bajas por lo que las consideramos como poco relevantes para nuestra variable objetivo.

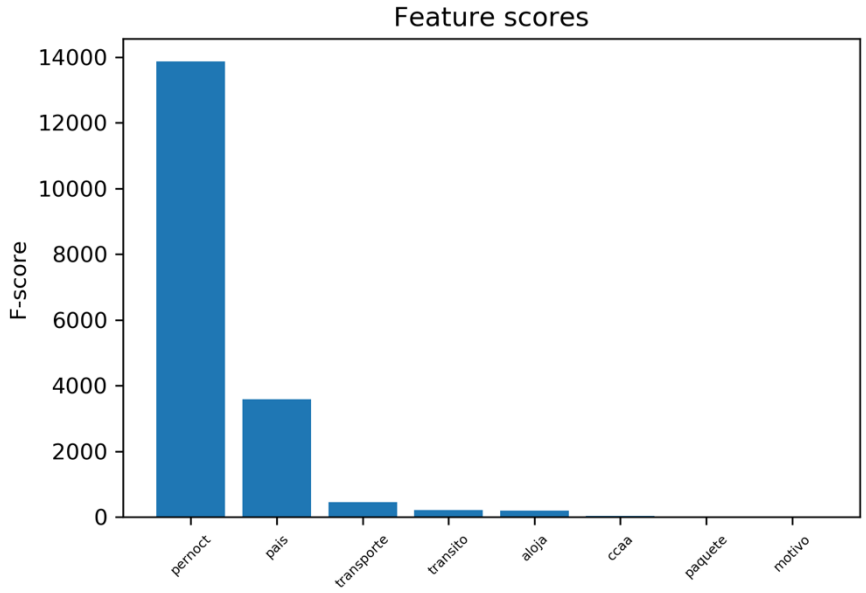

Figura 2.5. *Feature selection* datos 2019. Elaboración propia

b) *Feature selection* - Datos 2022

En este apartado, se vuelve a realizar el proceso descrito en el apartado anterior (a) pero esta vez empleando los datos correspondientes al verano de 2022 para evaluar con estos datos qué variables influyen más sobre nuestra variable objetivo "gastototal" y observar posibles diferencias a consecuencia de la pandemia. Realizando el mismo procedimiento que con los datos de 2019, hemos obtenido en este caso que las variables que mejor explican nuestra variable objetivo son "A13" o número de pernoctaciones, seguida de "país" y "transporte", que coincide con los resultados que obtenemos en el análisis de *feature selection* para 2019. Sin embargo, en este caso, encontramos que la puntuación f-test que determina el grado de correlación entre las variables, es algo superior en la variable "A13" de 2022 que en la de 2019 mientras que en el caso de la variable "país" ocurre lo contrario También se pueden observar diferencias en el grado de dependencia del resto de variables de un periodo a otro.

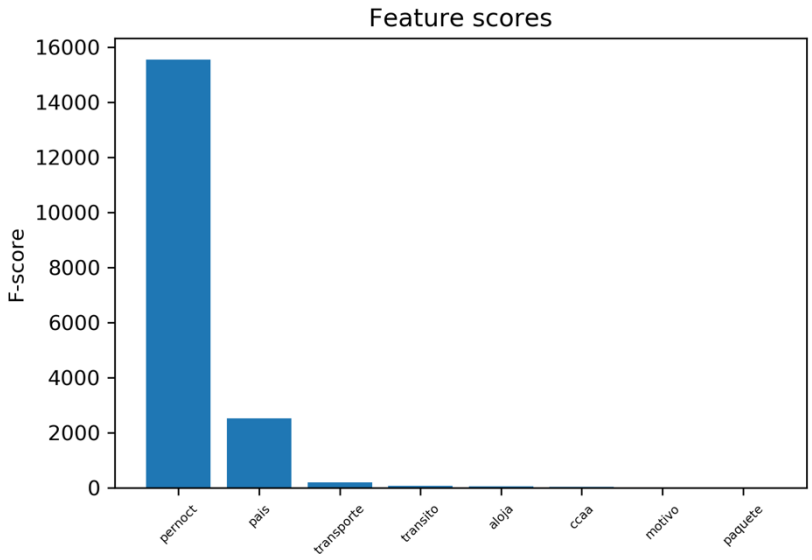

Figura 2.6. *Feature selection* datos 2022. Elaboración propia

	A13	pais	A1	A0_7	aloja	ccaa	A16	motivo
Datos 2019	13868.9	3590.19	455.85	209.12	193.34	38.83	16.73	0.67
Datos 2022	15539.4	2521.59	205.29	64.69	52.20	44.86	1.19	24.31

Tabla 2.3. Comparativa puntuaciones F-test en *feature selection*. Elaboración propia.

2.2.3. *Mutual information*

Esta técnica es una medida empleada para hallar el grado de dependencia entre variables siendo capaz a diferencia de *feature selection*, de hallar dependencias no lineales. Este método permite explicar la relación entre las variables y la importancia e impacto de las mismas sobre la variable objetivo definida en nuestros datos. En el caso de *mutual information,* la relación entre dos variables X e Y se define como la reducción de la incertidumbre sobre Y una vez conocido X, o lo que es lo mismo, esta técnica permite conocer la información que una variable aporta sobre otra. Cuanto mayor sea la información encontrada entre dos variables, más relevante se considera esta variable para la variable objetivo.

Este método es de gran utilidad especialmente cuando la relación entre la variable objetivo y otras variables es no lineal. Sin embargo, aunque estemos describiendo *mutual information* por un lado y *feature selection* por otro, ambos se consideran técnicas complementarias o mutualmente exclusivas, ya que están estrechamente relacionadas y

lo ideal es combinarlas y compararlas entre sí, para obtener un resultado más preciso, pues la información obtenida con *mutual information*, puede ser de gran utilidad para la selección de variables de *feature selection* como medida de cuantificación de la relación entre una variable X y la variable objetivo Y. Por tanto, *mutual information* es una herramienta de gran importancia que puede ser de muy útil para seleccionar variables relevantes cuya relación con la variable objetivo no es lineal, si no que capta cualquier influencia entre las variables.

a) *Mutual information* - Datos 2019

Procedemos con el análisis de *mutual information*. Esta metodología de estudio de relación entre variables difiere de la empleada anteriormente en *feature selection,* ya que no se limita a capturar la dependencia lineal entre las variables, si no que puede asumir cualquier tipo de relación entre las mismas como se ha mencionado anteriormente. Este método evalúa las variables en función de la información sobre la variable objetivo que contengan mediante la asignación de puntuaciones en las variables. El resultado que se ha obtenido al aplicar *mutual information* en nuestros datos del periodo correspondiente a 2019 ha sido el siguiente:

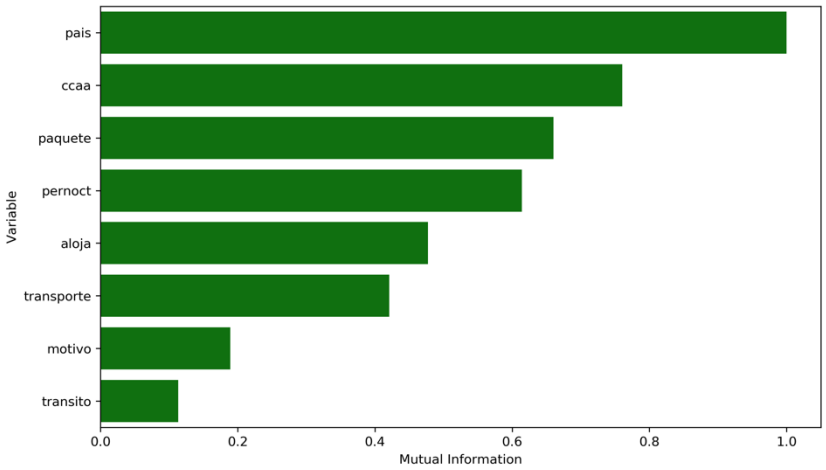

Figura 2.7. *Mutual information*, datos 2019. Elaboración propia

Por tanto, a partir de esta información, comprobamos que según el análisis de *mutual information* la variable "pais" tiene completa correlación con la variable objetivo "gasto total" pues su puntuación es de 1, seguida de la variable "ccaa" o comunidad autónoma donde se viaja. Después encontramos que las variables "A16" que determina la contrata -ción de paquete turístico o no y la variable"A13" que es el número de pernoctaciones también aportan relevancia a la hora de predecir la variable objetivo. Por tanto, para el periodo anterior a la pandemia, las variables que mayor peso presentan frente gasto total de un turista son el país de procedencia del turista, la comunidad autónoma donde se viaje, si se contrata o no un paquete turístico y el número de noches de alojamiento.

b) *Mutual information* - Datos 2022

Usando los mismos procesos que hemos empleado en los datos de 2019, calculamos los valores de *mutual information* para las variables, pero esta vez con los datos de 2022 para comparar si existen diferencias entre ambos y observar si la importancia de las variables sobre el gasto total varía de alguna forma como consecuencia del efecto de la pandemia sobre el sector y la economía del país.

En este segundo caso, obtenemos al contrario que en 2019, que las variables que han aportado más información sobre la variable objetivo han sido en primer lugar la variable "pais" seguida de "A13" (número de pernoctaciones):

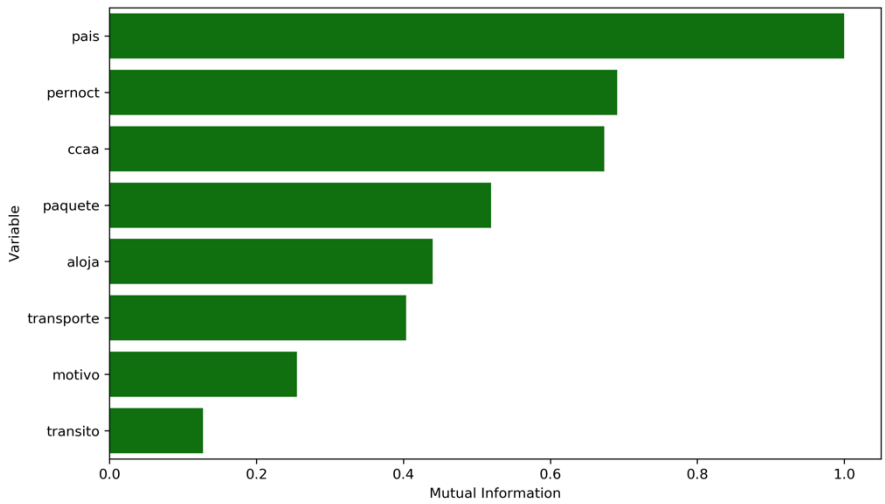

Figura 2.8. *Mutual information*, datos 2022. Elaboración propia

En este caso, sí que podemos apreciar diferencias, no solo con respecto a los datos de *feature selection*, sino también con respecto a los datos de *mutual information* del periodo anterior a la pandemia estudiado en el apartado anterior. Si comparamos los resultados con respecto a lo obtenido en *mutual information* de 2019, podemos observar que en ambos la variable que más impacta sobre el gasto total es el país de procedencia del turista. Sin embargo, la segunda variable más importante en 2022 resulta ser el número de pernoctaciones mientras que en 2019 fue la comunidad autónoma donde se viaja. Esto se puede deber al incremento de precios en la hostelería como consecuencia de la crisis económica y los efectos de la pandemia sufrida en el sector a lo largo de los últimos años. Sin embargo, si comparamos los resultados obtenidos con respecto a *feature selection* de este mismo año, observamos que al contrario que en *mutual information*, la variable más relevante para el gasto turístico es el número de pernoctaciones (A13), seguida de país y el medio de transporte empleado (A1).

2.3. Comparación de resultados

Como hemos mencionado anteriormente, *feature selection* y *mutual information* son dos técnicas de selección de variables. Sin embargo, puesto que emplean metodologías e hipótesis diferentes para determinar la relación entre las variables, los resultados que se obtienen con cada uno no tienen por qué coincidir y, de hecho, la mayoría de las veces difieren entre sí. Es importante, por tanto, para determinar qué resultado escoger, averiguar qué metodología ha proporcionado un nivel de precisión mayor, con el objetivo de poder obtener a posteriori un mejor modelo de *machine learning*. Para ello, analizamos el nivel de precisión obtenido en cada método empleado. Para poder evaluar de una forma comparable la precisión de los resultados obtenidos en ambos métodos, se debe usar un sistema común de evaluación de desempeño. En este caso, se ha utilizado el método de *root mean square error* (RMSE), ya que esta medida es la más común para medir el desempeño en problemas regresión para predecir variables continuas. Se calcula obteniendo la raíz cuadrada del promedio de las diferencias al cuadrado entre los valores reales y predichos.

Los resultados obtenidos según RMSE han sido los siguientes:

	2019	2022
Feature selection	599.76	740.45
Mutual information	873.18	791.08

Tabla 2.4. Comparativa RMSE 2019 y 2022 en *feature selection* y *mutual information* Elaboración propia.

Para interpretar estos resultados, se debe tener en cuenta que, cuanto menor valor de RMSE, menores son las diferencias, menor error estamos cometiendo y por tanto mejor es el resultado obtenido (existe mayor nivel de *accuracy*). Por lo tanto, observamos que el resultado obtenido tanto para los datos de 2019 como para los de 2022, *feature selection* ha sido mejor en comparación con el obtenido en *mutual information*. Esto puede deberse a que los datos en ambos casos presenten dependencias lineales y, por tanto, *feature selection* es capaz de identificar mejor las relaciones y los patrones entre los mismo que *mutual information*. Teniendo en cuenta este resultado, se optará por tomar como válidas las variables que han resultado ser de mayor relevancia mediante *feature selection* de cara a su implementación para la obtención del modelo de *machine learning* en ambos periodos.

Desde el punto de vista del sector turístico respecto al impacto del Covid-19, se han podido observar ligeras diferencias entre el gasto turístico de 2019 y el de 2022, ya que se ha podido apreciar que, como consecuencia de la crisis y el incremento de los precios, los turistas en 2022 han gastado alrededor de 155€ adicionales que en 2019 para el mismo número de noches de estancia. A la hora de estudiar qué factores afectan más a la variable del gasto, no existen discrepancias entre ambos periodos pues mediante *feature selection* hemos obtenido que tanto e 2019 como 2022 las variables más significativas con respecto al gasto han sido el número de pernoctaciones, el país de procedencia del turista y el medio de transporte empleado respectivamente. Sin embargo, pese a no existir diferencias en cuanto a los factores que más influyen sobre esta variable en ambos periodos, sí que existen diferencias en cuanto a la proporción de importancia, pues en 2022, el f-test para las pernoctaciones ha sido de 15539´4 versus 13868´9 en 2019. Para las variables país y medio de transporte por el contrario, las puntuaciones f-test son superiores en 2019 versus 2022, lo que implica que la pandemia ha afectado más sobre el gasto turístico debido a factores internos (subida de precios del sector) que externos como el país de procedencia.

Capítulo 3

MACHINE LEARNING: Optimización de modelos

En esta sección, se procede a realizar un análisis de los datos mediante la construcción de modelos de *machine learning* (todos los conceptos descritos en el capítulo pueden encontrarse en ([30]).

Para este proceso, emplearemos las herramientas *Regression learner* y C*lassification learner* integradas en la plataforma MATLAB, que permiten construir y entrenar diferentes modelos de regresión y clasificación respectivamente en nuestros datos hasta alcanzar el que sea óptimo, o lo que es lo mismo, el que mejor prediga nuestra variable objetivo "gastototal". *Regression learner* como su propio nombre indica, permite llevar a cabo procesos de regresión para predecir variables con valor numérico mediante aprendizaje supervisado[6] en función de una serie de características. Se pueden emplear diferentes tipos de algoritmos de regresión en *machine learning*, en función del tipo de datos y de análisis que se quiera llevar a cabo. Entre los diferentes tipos de regresión encontramos algunos como:

- **Regresión lineal**: este tipo de regresión mide la relación entre una variable dependiente y una o varias variables independientes mediante el uso de ecuaciones lineales. Se emplea con datos que tienen dependencias lineales entre las variables.
- **Regresión polinómica**: es una extensión de la regresión lineal, pues incluye algo más de complejidad añadiendo términos de grado superior a la ecuación. Se emplea para datos cuyas variables poseen relaciones no lineales entre sus variables.
- **Regresión logística**: es un tipo de regresión que se emplea para predecir variables de tipo binario midiendo la relación entre esta variable y una o varias variables

[6] El aprendizaje supervisado es una técnica de *machine learning* que consiste en el aprendizaje automático del algoritmo a través de unos datos de entrada, que aprende a predecir el *output* en función de estos datos iniciales a través de patrones y relaciones que encuentra en los datos.

predictivas. En la regresión logística se estudia la probabilidad de una variable con valores en una escala entre 0 y 1.

Estos tipos de regresión descritos son algunos ejemplos de los muchos que podemos encontrar. El aplicativo de *Regression learner* es una interfaz que permite construir, entrenar y comparar diferentes modelos de regresión como regresión lineal, redes neuronales o árboles de decisión entre otros de forma sencilla. Los datos se pueden cargar de diversas fuentes, y la herramienta permite seleccionar aquellas variables que queremos incluir en el modelo a construir, así como definir la variable objetivo que queremos predecir.

Por otro lado, usaremos *Classification learner* para construir modelos de clasificación con el objetivo de poder predecir si el gasto total incurrido por los turistas se considera alto, medio o bajo. Esta herramienta funciona de la misma manera que *Regression learner*, con la diferencia de que *Classification learner* se emplea para construir modelos de clasificación en lugar de regresión.

Ambas herramientas permiten seleccionar entre diferentes mecanismos de validación para medir la calidad de los modelos construidos y evitar *overfitting*.[7] Entre los mecanismos de validación encontramos:

- ***Cross-fold validation***: esta técnica permite dividir los datos en varios grupos o *folds*, utilizando una parte como datos de entrenamiento o *train* y la otra parte como datos de test. Existen diferentes tipos de *cross-validation*, siendo el más común k-fold cross validation, donde el parámetro k, hace referencia al número de subconjuntos en los que se va a dividir el *dataset*. Elegir un valor u otro de k, dependerá del tamaño de los datos. Normalmente se emplea un valor de K de entre 5 y 10 y una forma de encontrar el valor óptimo de K, es ir probando con diferentes valores hasta reducir el error. Este método se utiliza principalmente cuando el *dataset* no es demasiado grande y se quiere encontrar un modelo que tenga un buen desempeño a la hora de adaptarse a datos nuevos.

[7] El *overfitting* es un fenómeno que se da en los problemas de *machine learnin*g cuando un modelo se ajusta muy bien a los datos de entrenamiento pero, sin embargo, no es capaz de generalizar bien con datos distintos a los empleados en el entrenamiento del modelo.

- ***Hold-out validation***: esta técnica divide los datos en dos segmentos: *train set* y *test set*. El conjunto de entrenamiento es donde se entrena el modelo y el conjunto de test, es donde se estudia cómo se ajusta el modelo al resto de los datos que no han sido entrenados previamente. Normalmente se hace una partición de los datos de 80% (*train*) - 20% (*test*) pero esta proporción puede variar según el tipo de datos que tengamos y el análisis que se quiera realizar. Este método es apropiado cuando se tienen grandes *datasets* y suficientes recursos computacionales.

- ***Resubstitution validation***: en este tipo de validación, a diferencia de los anteriores, no se hace ningún *split* en los datos, si no que se utiliza el conjunto entero tanto para el entrenamiento como para la parte de test. Esta técnica, sin embargo, puede ser algo imparcial, ya que el modelo aprende como encajar los datos en la parte de entrenamiento, lo que supone un mejor desempeño en la parte de test y por tanto mostrará un resultado con un error muy bajo y muy optimista pero poco preciso, pues no proporciona una buena estimación de la capacidad de generalización del modelo a la hora de ajustarse a nuevos datos.

Teniendo en cuenta lo descrito, las mejores opciones a considerar como método de validación serán *cross-fold validation* y *hold-out validation*. Elegir uno u otro, dependerá del tamaño y naturaleza del dataset, así como de la complejidad. Teniendo esto en cuenta, puesto que en nuestro caso contamos con un *dataset* de gran tamaño en ambos periodos se ha seleccionado como método de validación *hold-out validation,* con una proporción del 80%-20% (se empleará el 80% de los datos como entrenamiento y el 20% como test) que son los valores estándar para construir modelos suficientemente robustos y precisos capaces de generalizar correctamente con nuevos datos.

Una vez entrenados y validados los datos, para encontrar el modelo óptimo, se calcula para cada uno una medida de error que permite comparar y evaluar los modelos construidos. En el caso de los modelos de regresión, se emplea RMSE (*root mean square error)*, tal que, una vez entrenados los modelos, se escoge el que tiene menor valor de RMSE, pues este será el que mejor se ajuste a los datos y consiga mejor desempeño. Aunque RMSE es la medida que la herramienta emplea por defecto ya que es más

interpretable, también permite escoger otras medidas estadísticas para determinar el error de los modelos como MSE *(mean squared error*)[8] y MSA (*mean absolute error)*[9] entre otras. En el caso de los modelos de clasificación, se emplea una medida de desempeño conocida como *accuracy*. Esta medida hace referencia al % de observaciones que el modelo ha sido capaz de clasificar correctamente, por lo que a mayor nivel de *accuracy*, mejor será el modelo prediciendo la variable objetivo. El cálculo de esta medida es bastante simple, consiste en dividir el número de observaciones correctamente clasificadas entre el número total de observaciones del *dataset*.

Finalmente, en cuanto a los modelos de clasificación, es importante aclarar que las observaciones se clasifican en cuatro posibles categorías, que tendrán una proporción superior o inferior en función de la capacidad de clasificación del modelo. Estas categorías son: Verdadero positivo (TP)**,** Falso positivo (FP)**,** Verdadero negativo (TN) **y** Falso negativo (FN)**.**

3.1. Modelo de regresión

3.1.1. Resultados obtenidos en los datos de 2019

Como se ha mencionado antes, se ha decidido construir todos los modelos existentes en el aplicativo con el objetivo de encontrar aquel que proporcione unos resultados más precisos. Se han escogido para la construcción del modelo las variables obtenidas en el proceso de *feature selection*, ya que respecto a *mutual information*, obtuvo un menor error. Por este motivo, se han escogido para construir el modelo las variables: "país", "A1" o medio de transporte empleado y "A13" o número de pernoctaciones siendo "gastototal" la variable objetivo.

[8] *Mean Squared Error* (MSE) es un método de validación empleado en *machine learning*, empleado para medir el desempeño en modelos de regresión. Se calcula con el promedio de las diferencias al cuadrado entre valores reales y predichos. Aunque ambas medidas son similares, RMSE es preferible frente MSE ya que al proporcionar el error en las mismas unidades que la variable objetivo, es más interpretable.

[9] *Mean Absolute Error* (MSA), es una medida que calcula la diferencia en valores absolutos del promedio entre los valores reales y predichos de la variable objetivo dando así el mismo peso a todos los errores.

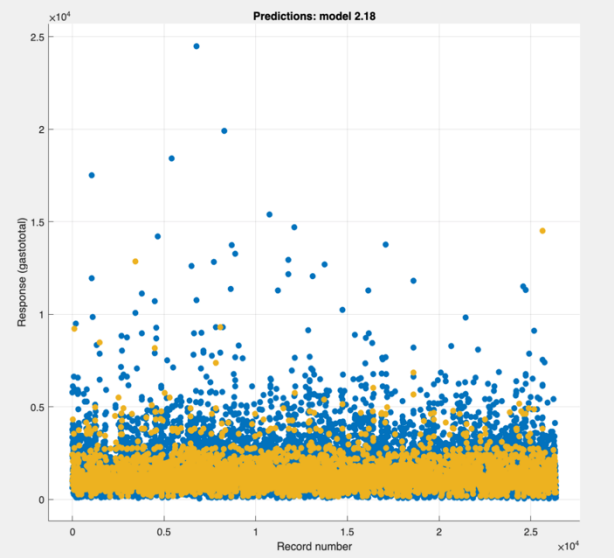

Figura 3.1. Proceso de regresión Gaussiana (GPR). Elaboración propia

Como se muestra en la Figura 3.1, el modelo de *machine learning* que se ha obtenido consiste en un proceso de regresión gaussiana, siendo los puntos representados en color azul los valores reales de la variable y los puntos amarillos los valores predichos. Este tipo de regresión está fuertemente relacionado con la optimización bayesiana, una técnica de optimización de parámetros donde se utiliza un modelo matemático de la función que se va a optimizar para hacer predicciones sobre los valores de la variable objetivo. La regresión gaussiana, por tanto, es uno de los modelos que se puede emplear como modelo matemático en la optimización bayesiana, siendo un enfoque estadístico para el análisis de regresión que utiliza procesos gaussianos, o lo que es lo mismo, distribuciones normales.

En este algoritmo, se representa la relación entre las variables utilizando una distribución de probabilidad que viene definida por una función media y una función de covarianza. La función media se encarga de captar la tendencia general de los datos, mientras que la función de covarianza modela como se relacionan los valores de las variables entre sí. Para ello, la variable objetivo se modela en este caso como una función gaussiana, permitiendo un modelado más flexible de la relación entre las variables. Este tipo de regresión es especialmente útil a la hora de hallar relaciones no lineales y complejas entre las variables sin realizar fuertes suposiciones sobre la forma funcional de la relación entre las variables.

El programa ha construido alrededor de 30 modelos diferentes, pero el proceso de regresión gaussiana ha resultado ser el más apropiado para nuestros datos, ya que el RMSE obtenido en el proceso de validación ha sido de 505´71, siendo el error más bajo encontrado con respecto al resto de modelos.

Model 2.18: Gaussian Process Regression

Training Results

RMSE (Validation)	505.71
R-Squared (Validation)	0.70
MSE (Validation)	2.5574e+05
MAE (Validation)	306.07
Prediction speed	~1600 obs/sec
Training time	1466 sec

Figura 3.2. Resultados en la validación del modelo GPR. Elaboración propia.

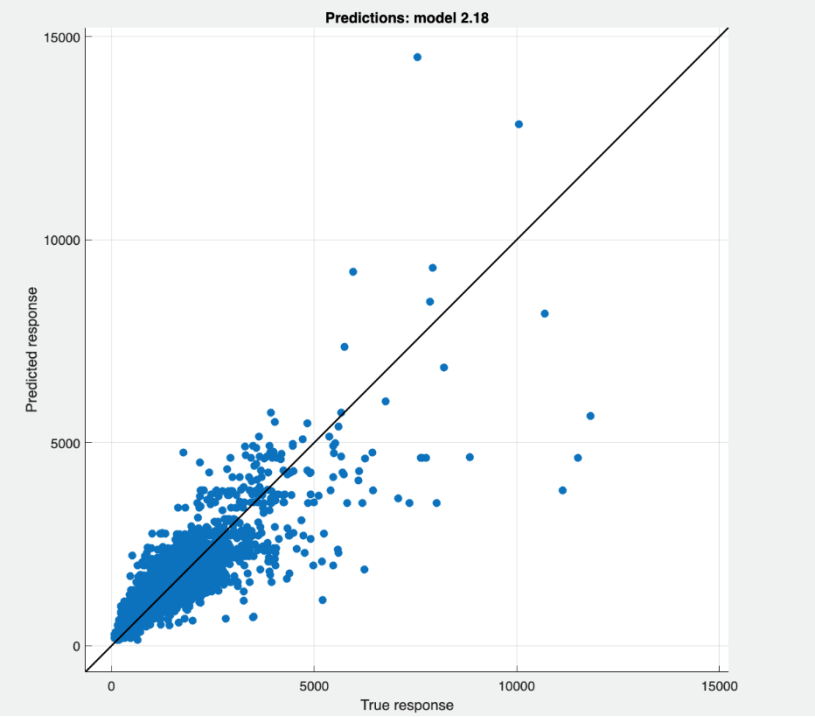

Figura 3.3. Comparativa entre los valores reales y predichos. Elaboración propia

Sin embargo, como se puede observar en la Figura 3.3, el modelo no se ajusta a la perfección a los datos, pues los valores predichos, no se adaptan exactamente con los valores reales, aunque si se acercan bastante ya que en general todos los puntos se encuentran muy cercanos a la línea de regresión y por tanto el modelo captura bastante bien la tendencia. Podemos comprobar también la calidad del modelo a partir de los residuos generados, que, en este caso, podemos ver que el modelo no comete errores

demasiado elevados salvando algunos *outliers*, que son valores atípicos en los datos y por tanto más difíciles de predecir.

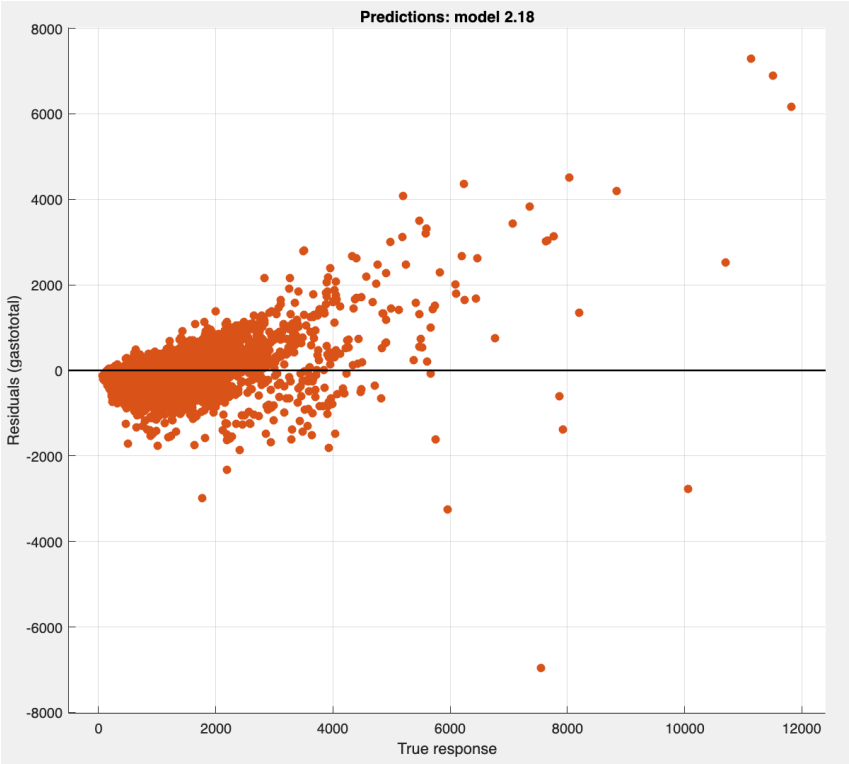

Figura 3.4. Representación de los residuos. Elaboración propia

Esto indica, que el modelo de regresión gaussiana predice con bastante acierto el gasto total incurrido por un turista en función del resto de variables representadas y por tanto será capaz de predecir el gasto total por turista si se empleasen otros datos diferentes. Otro indicativo de que el modelo presenta un buen desempeño es el valor obtenido en la medida de R^2, que es de 0´70. Esta medida se usa para determinar qué proporción de varianza de la variable dependiente (variable objetivo), queda explicada por las variables independientes. Cuanto mayor sea su valor, mejor es el modelo, pues mayor proporción de la variable objetivo queda explicada por el resto de las variables siendo el modelo capaz de explicar el 70% de la varianza de la variable dependiente.

3.1.2. Resultados obtenidos en los datos de 2022

Siguiendo la misma línea que con los datos de 2019, en los datos de 2022, hemos seleccionado de cara a la construcción de modelos de *machine learning*, las variables obtenidas como más significativas en los resultados de *feature selection*, que al igual que en el caso anterior, esta técnica obtuvo un menor RMSE que *mutual information* a la hora de definir qué variables explican mejor la variable objetivo "gastototal". Como se pudo observar en la Figura 2.6, la variable "A13" o número de pernoctaciones, es la más representativa al igual que en los datos de 2019, sin embargo, en este caso, la variable obtiene un *f-score* más elevado, eclipsando a las siguientes variables que han obtenido en comparación puntuaciones muy inferiores. De igual forma que en el apartado anterior, tomaremos como variables más representativas las tres primeras: "A13" (número de pernoctaciones), "pais", y "A1" (medio de transporte). Del mismo modo que en 2019, hemos utilizado el método *hold-out validation* con una proporción de 80%-20%.

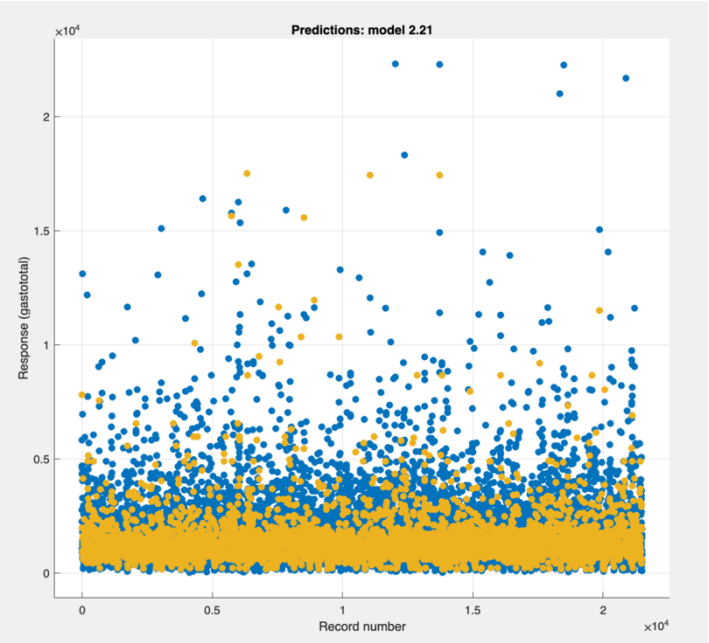

Figura 3.5. Red Neuronal media (*neural network*). Elaboración propia

Como se puede apreciar en la Figura 3.3, el mejor modelo de *machine learning* obtenido de entre un total de 32 modelos entrenados, ha sido una red neuronal de tipo medio. Las redes neuronales son un tipo de modelos de *machine learning*, inspirados en el funcionamiento del cerebro humano, que se caracterizan por ser capaces de representar relaciones complejas y no lineales entre las variables de entrada y de salida, lo que hace

que sea un modelo muy potente capaz de resolver diferentes tipos de problemas. Las redes neuronales se pueden emplear para resolver problemas tanto de regresión como de clasificación y funcionan mediante la interconexión de nodos llamados neuronas, que procesan la información y realizan predicciones en función de los datos de entrada.

Cuando se utiliza para la regresión, la red neuronal se entrena para predecir un valor numérico continuo basado en un conjunto de variables de entrada y analizando un conjunto de datos de entrenamiento previamente. Durante el entrenamiento, la red ajusta los pesos (hiperparámetros) que conectan las neuronas entre sí para reducir el error entre los valores reales y predichos. Una de las ventajas de las redes neuronales para la regresión es su capacidad para capturar patrones poco evidentes en los datos que otros modelos de regresión podrían no tener en cuenta. El hecho de que la red neuronal se caracterice como "medio" simplemente hace referencia al número de capas que se han utilizado para el procesamiento de los datos, o lo que es lo mismo, el número de neuronas empleadas.

Model 2.21: Neural Network

Training Results

RMSE (Validation)	638.07
R-Squared (Validation)	0.75
MSE (Validation)	4.0713e+05
MAE (Validation)	380.65
Prediction speed	~670000 obs/sec
Training time	1076.6 sec

Figura 3.6. Resultados de la validación en el modelo de red neuronal. Elaboración propia.

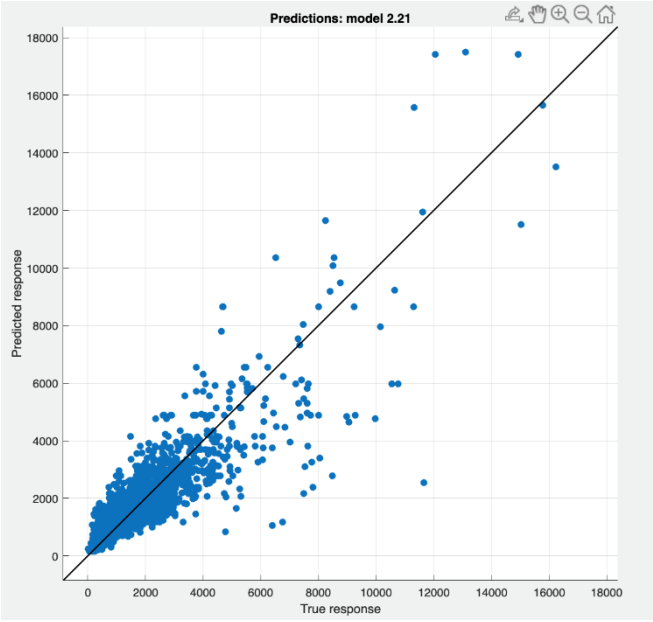

Figura 3.7. Comparativa entre valores reales y predichos. Elaboración propia.

Como se puede apreciar en la Figura 3.7, el modelo se ajusta con cierta precisión a los datos reales y su predicción es relativamente buena ya que los puntos quedan bastante cercanos a la línea de regresión. De hecho, al realizar la validación, el RMSE o error obtenido entre los valores reales y predichos en el modelo ha sido de 638´07, y el valor de R^2 obtenido de 0´75. Estas métricas son superiores que las obtenidas en el modelo GPR de 2019 si comparamos tanto RMSE como R^2, por lo que este modelo es mejor en comparación con el anterior y por tanto se ajusta con mayor precisión a los datos por lo que será capaz de predecir bastante mejor exactitud el gasto total a la hora de utilizar datos nuevos así como de cometer menos errores en la predicción.

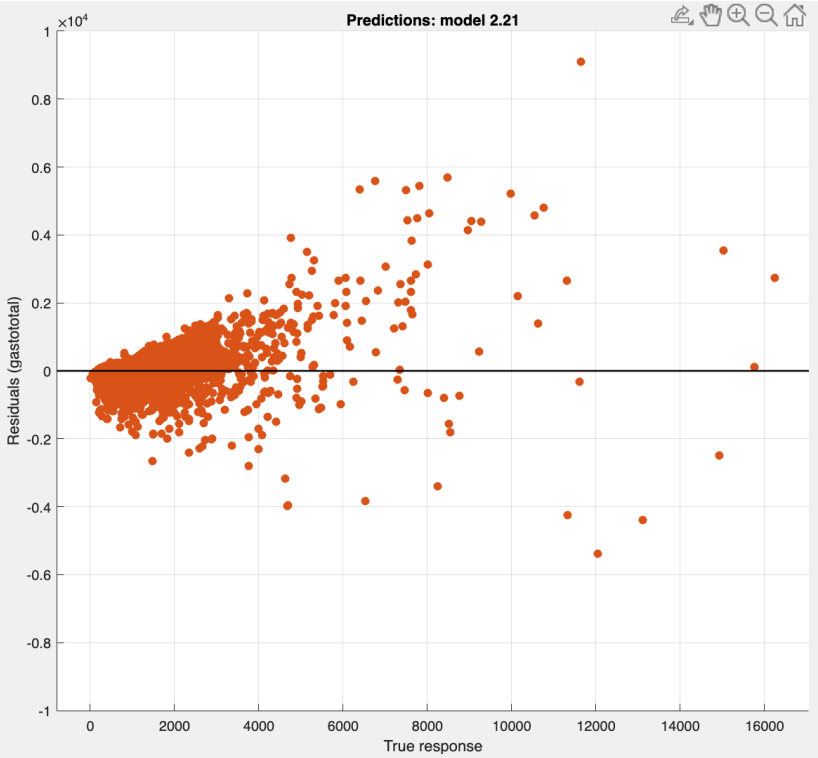

Figura 3.8. Representación de los residuos. Elaboración propia

3.2. Modelo de Clasificación

De la misma manera que en la realización de los modelos de regresión anteriores, para los modelos de clasificación se emplearán las tres variables más significativas resultantes de *feature selection* ("pais", "A13" y "A1"), empleando como método de validación *hold-out* con una proporción de 80%-20%.

3.2.1. Resultados obtenidos en los datos de 2019

Para la realización de los modelos de regresión anteriores, nuestra variable objetivo era la variable original del conjunto de datos "gastototal", pues se trata de una variable numérica continua. Sin embargo, para poder llevar a cabo un proceso de clasificación, ha sido necesario modificar esta variable, de tal forma que sea de tipo categórico y podamos predecirla empleando modelos de clasificación. Para ello, se ha transformado la variable original "gastototal" en una variable categórica con valores 1, 2 y 3 que representan si el gasto es bajo, medio o alto respectivamente.

Para definir unos umbrales lógicos a la hora de clasificar el gasto como alto, medio o bajo, se han estudiado los percentiles de la variable "gastototal". Para ello, se han calculado concretamente los percentiles 33 y 67, donde se ha obtenido, como se puede observar en la Figura 3.9, el valor medio de la distribución es de 1100€ aproximadamente. Por otro lado, los valores por debajo de 800€ quedarían en el 33% de la distribución mientras que los valores por debajo de 1500€ representan aproximadamente el 67%.

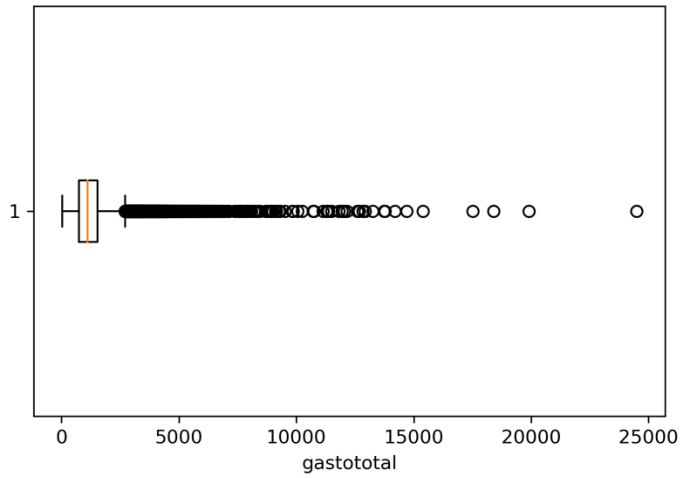

Figura 3.9. Percentiles de la variable "gastototal". Elaboración propia.

Gracias a este análisis, podemos identificar cómo dividir y clasificar el gasto en las categorías mencionadas. La división que se ha realizado ha sido la siguiente:

- Si el gasto es <800€, el gasto se clasifica como bajo (valor 1)
- Si el gasto se encuentra entre 800€ y 1500€, el gasto se clasifica como medio (valor 2)
- Si el gasto se encuentra por encima de 1500€, el gasto se clasifica como alto (valor 3)

Una vez transformada la variable "gastototal" a variable categórica, se han calculado varios tipos de modelos de clasificación usando *Classification learner* donde se ha obtenido como mejor modelo el siguiente:

Figura 3.10. Matriz de confusión - Modelo de Red Neuronal (*neural network*) de clasificación 2019. Elaboración propia.

Se trata de nuevo, al igual que en la regresión de 2022, de un modelo de red neuronal, esta vez de clasificación. En este caso, se ha representado mediante una matriz de confusión que proporciona información relevante sobre el nivel de precisión con el que el modelo ha clasificado los datos. Como se puede observar, un total de 906 observaciones que en los datos reales se habían considerado como gasto bajo, han sido clasificadas correctamente por el modelo. Para el gasto medio, el modelo ha clasificado correctamente

un total de 1908 observaciones y en el caso de gasto alto, ha clasificado correctamente 1036 observaciones. El nivel de precisión de clasificación del modelo se puede observar con mejor detalle en la tabla situada junto al modelo, donde podemos encontrar las ratios de clasificación de valores correctos e incorrectos para cada categoría. En el caso de valores positivos verdaderos (TPR), el modelo ha clasificado correctamente el 60% de gasto como bajo, el 80% de gasto medio y el 75´1% de gasto alto donde el 40%, el 20% y el 24´9% representan la tasa de falsos negativos (FNR) o lo que es lo mismo, los valores positivos que el modelo ha clasificado de forma errónea como negativos.

Finalmente, cabe destacar que el nivel de *accuracy* del modelo ha sido del 73%. Esto significa que el modelo es capaz de clasificar correctamente a al 70% de las observaciones lo que supone un resultado aceptable en nuestro caso. Para contrastar este dato, se ha obtenido también la curva ROC, que es una representación de los TPR (tasa de valores positivos) versus FPR (tasa de falsos positivos). La curva ROC es de bastante utilidad a la hora de visualizar y comparar el rendimiento de diferentes modelos de clasificación.

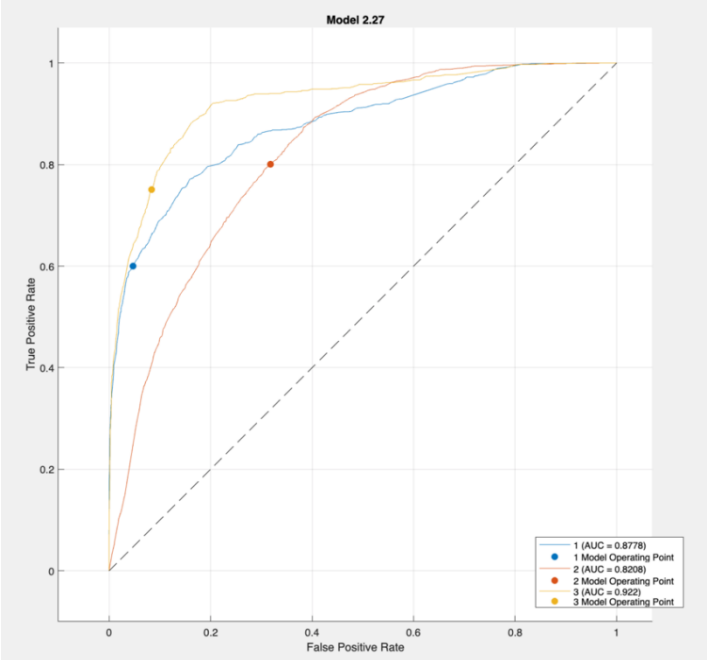

Figura 3.11. Curva ROC del modelo de clasificación red neuronal 2019. Elaboración propia

Aunque según la tasa obtenida de *accuracy*, nuestro modelo tan solo parecía aceptable a la hora de clasificar los valores de las variables, según el umbral marcado por la curva ROC, podemos observar que tenemos una tasa bastante elevada de verdaderos positivos

en las tres categorías representadas, lo cual es un buen indicativo. El AUC significa área bajo la curva, que es la métrica de la curva ROC. Generalmente, a mayor puntuación AUC, mejor es el rendimiento del modelo a la hora de clasificar. Por tanto, puesto que nuestros valores de AUC son todos superiores a 0´8, donde la categoría que mejor clasifica es el gasto alto, podemos concluir con que el modelo obtenido es bastante preciso a la hora de clasificar el gasto total de los turistas en nuestros datos de entrenamiento.

3.2.2. Resultados obtenidos en los datos de 2022

Para trabajar con los datos de este periodo, se ha llevado a cabo el mismo proceso que el que se ha explicado en la sección anterior con los datos de 2019. En primer lugar, se ha transformado la variable "gastototal" en categórica para poder trabajar con el modelo de clasificación. En esta ocasión la distribución de los percentiles salía aproximadamente igual que en 2019, por lo que, para poder contrastar ambos modelos, se han utilizado los mismos umbrales de clasificación del gasto.

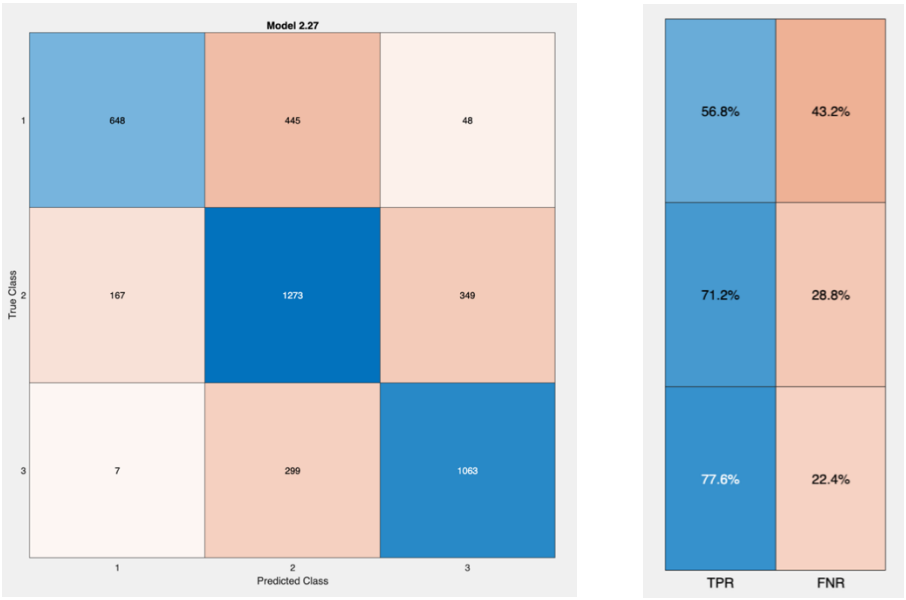

Figura 3.12. Matriz de confusión - Modelo de Red Neuronal (*neural network*) de clasificación 2022. Elaboración propia.

Como se puede observar en la Figura 3.12, se ha obtenido de nuevo un modelo de red neuronal como óptimo de entre todos los entrenados. Sin embargo, en este caso el modelo presenta un *accuracy* algo inferior, del 69´4% lo que supone un 3´6% menos que en el modelo obtenido para el periodo anterior. En cuanto a la clasificación de las diferentes categorías, podemos observar que 648 observaciones han sido correctamente clasificadas

como gasto bajo, 1273 como gasto medio y 1053 como gasto alto. En consecuencia, los valores positivos verdaderos correctamente clasificados (TPR), son del 56´8% para gasto bajo, el 71´2% para gasto medio y del 77´6% para el gasto alto con un ratio de falsos negativos (FNR) del 43´2%, el 28´8% y el 22´4% respectivamente.

Si comprobamos los resultados obtenidos con la curva ROC, podemos apreciar que este modelo en comparación con el anterior tiene una precisión de clasificación inferior, pues los valores de AUC son inferiores con respecto a los de 2019, aunque siguen siendo valores aceptables pues el AUC para las tres categorías está alrededor del 0´8 e incluso lo supera en dos de ellas siendo el grupo 3 (el gasto alto) la categoría clasificada con mayor precisión.

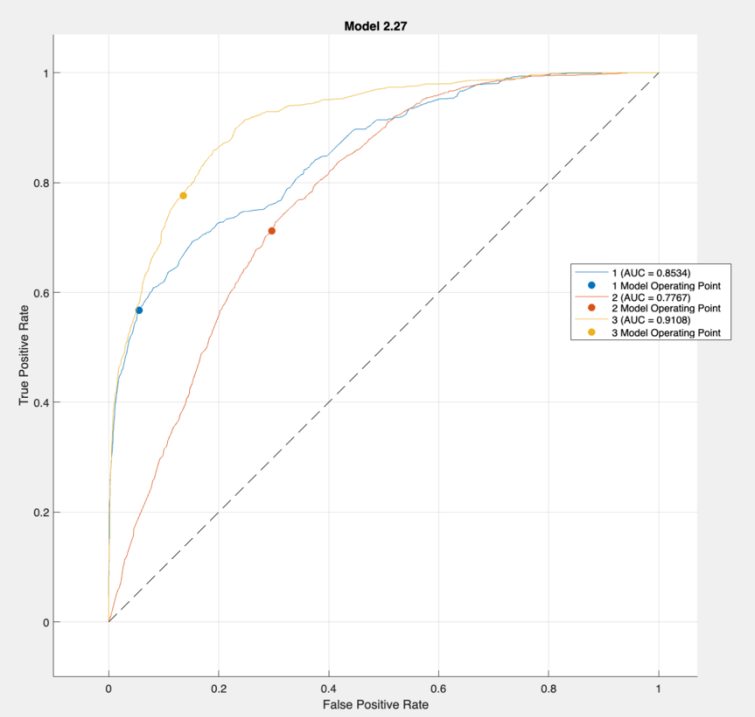

Figura 3.13. Curva ROC del modelo de clasificación red neuronal 2022. Elaboración propia

	2019	2022
Modelo de regresión *(RMSE)*	505´71	638
Modelo de clasificación *(accuracy)*	73%	69´4%

Tabla 3.1. Resumen de la medición de desempeño de cada modelo según el proceso de entrenamiento. Elaboración propia

Finalmente, como podemos apreciar en la Figura 3.14, existe una capacidad ligeramente superior de predicción en los datos de 2019 versus los de 2022 respecto a la fase de entrenamiento, ya que se ha obtenido tanto un mejor modelo de clasificación como de regresión. Este fenómeno se puede atribuir a que los datos de 2019 presentan un número algo superior de observaciones, lo que resulta beneficioso a la hora de construir un modelo de *machine learning* ya que se cuenta con más datos de entrenamiento, lo que facilita encontrar patrones y características en los datos reduciendo el riesgo de *overfitting*. Para descubrir realmente cómo de bien son capaces de predecir la variable objetivo "gastototal" estos modelos, se debe llevar a cabo un nuevo procesamiento de estos, pero usando un conjunto de datos diferente a los empleados en el entrenamiento, ya que esto es lo que permitirá observar si verdaderamente el modelo es capaz de generalizar bien y obtener buenas predicciones.

3.2. Desempeño de los modelos: Análisis del turismo en la región de Cataluña

En esta sección se quiere profundizar más aún en la evaluación de desempeño y capacidad de predicción de los modelos entrenados anteriormente. Para ello, en lugar de utilizar como conjunto de test un nuevo conjunto de datos, se ha utilizado, un subconjunto de datos de los datos originales empleados para construir los modelos. En este caso, se ha querido observar la capacidad de predicción de la variable objetivo "gastototal" de los modelos en un subconjunto poblacional, donde para trabajar y visualizar el desempeño del modelo, se han seleccionado exclusivamente los valores correspondientes a los viajeros que han hecho turismo en la comunidad autónoma de Cataluña ya que en nuestro análisis preliminar se observó que era la más frecuente entre los destinos vacacionales.

El objetivo de este procedimiento es observar la capacidad de predicción de los modelos tanto de regresión como de clasificación, entrenándolos nuevamente, pero esta vez sobre este *dataset* reducido y en lugar de evaluar el modelo sobre las tres variables, en este caso evaluaremos el gasto sobre la variable más explicativa que es pernoctaciones. Para evaluar los modelos de regresión se ha empleado la métrica habitual RMSE empleada en el entrenamiento y se ha representado mediante una gráfica de regresión, mientras que los modelos de clasificación se han evaluado usando el *accuracy* y de igual manera que en el conjunto de entrenamiento, se han representado mediante una matriz de confusión.

En la Figura 3.15, podemos apreciar cómo se ha ajustado el modelo gaussiano (GPR) entrenado con los datos de 2019, a los datos del subconjunto filtrado por la comunidad de Cataluña y en función del número de pernoctaciones. En la gráfica, los datos reales de la variable gasto están representados por los puntos azules, mientras que los puntos en color verde y rojo representan las predicciones realizadas por el modelo GPR en función de la comprensión del sobre la relación entre la variable de entrada (representada en el eje x) y la variable de destino (representada en el eje y). En otras palabras, la gráfica nos muestra las predicciones realizadas por el modelo no como un único valor, si no como un rango o intervalo de confianza, dando lugar a un margen de incertidumbre en la predicción realizada. En este caso, podemos ver que los valores reales de la variable se encuentran comprendidos en el intervalo proporcionado siguiendo la misma tendencia, lo que indica que el modelo está capturando bien la estructura de los datos y por tanto se ajusta bastante a los valores reales haciendo buenas predicciones.

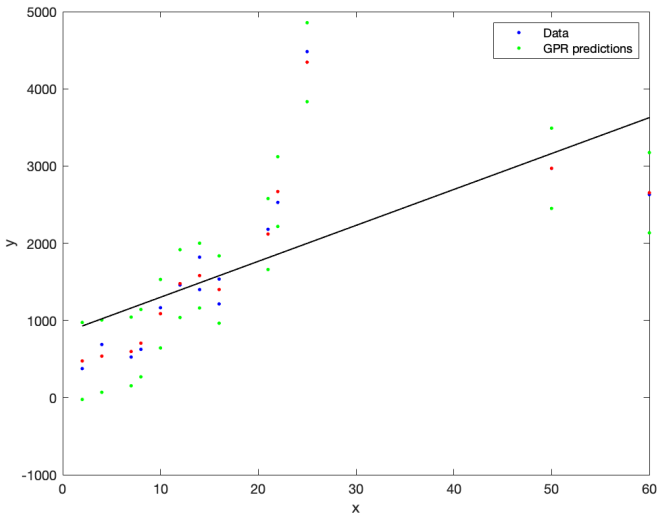

Figura 3.14. Gráfica de regresión del modelo GPR 2019. Elaboración propia

Para evaluar el modelo de regresión correspondiente a 2022 se ha llevado a cabo el mismo proceso, aunque a diferencia de 2019, en este caso el modelo que se obtuvo fue una red neuronal por lo que esta vez, a diferencia de GPR, no se representa una aproximación del valor esperado de la variable en intervalo de confianza, sino que directamente se han obtenido los predichos de la variable objetivo “gastototal” por el modelo NN.

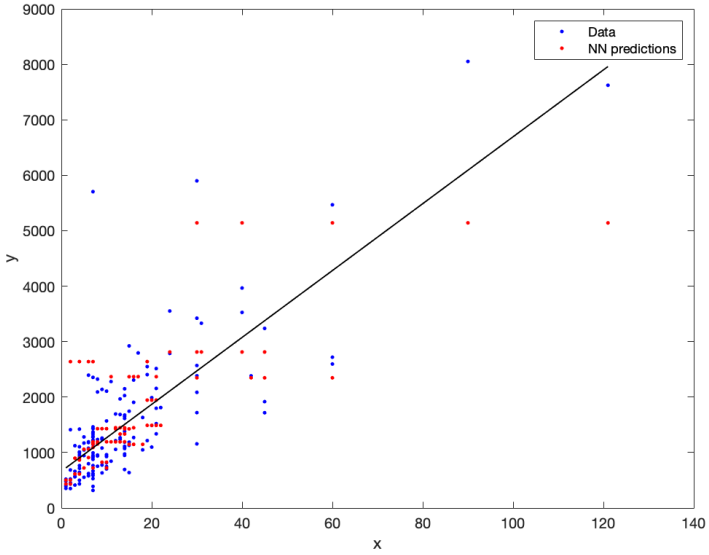

Figura 3.15. Gráfica de regresión del modelo red neuronal 2022.
Elaboración propia

Como se puede apreciar en la gráfica, mediante la línea de regresión, observamos que para el subconjunto de datos entrenado, el modelo es capaz de entender con cierta precisión el patrón en los datos. Adicionalmente, si observamos la representación de los puntos azules (valores reales) con los puntos rojos (valores predichos), se puede apreciar bastante proximidad entre ambos lo que indica que el modelo es capaz de predecir valores muy cercanos a los reales, lo que demuestra una buena capacidad predictiva.

	GPR	Red Neuronal
RMSE	7´36%	9´94%

Tabla 3.2. RMSE por modelo. Elaboración propia

Como se puede observar, ambos modelos han sido capaces de predecir muy bien la variable objetivo, y de ajustarse a los datos correspondientes a la comunidad autónoma de Cataluña. Sin embargo, si evaluamos ambos modelos objetivamente mediante métricas de evaluación el desempeño, observamos que el modelo que ha obtenido menor error y por tanto una mejor capacidad predictiva ha sido el GPR.

A continuación, para evaluar los modelos de clasificación, de igual manera se ha empleado el subconjunto filtrado por Cataluña como comunidad autónoma de destino vacacional, con la diferencia, como se ha mencionado antes, de que en este caso se evalúa la capacidad predictiva de modelos usando el *accuracy*. Para el modelo de 2019, se

obtuvo una red neuronal, y a la hora de evaluar el modelo usando este subconjunto de datos se ha obtenido un *accuracy* del 62´79%, que como es de esperar, es inferior que el 70% obtenido con los datos de entrenamiento, aunque todavía aceptable ya que el modelo está siendo capaz de clasificar correctamente a más del 60% de los datos.

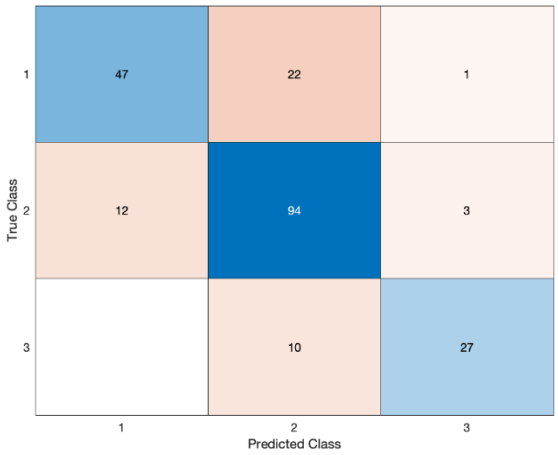

Figura 3.16. Matriz de confusión modelo de clasificación red neuronal 2019. Elaboración propia

El modelo ha sido capaz de clasificar correctamente 47 observaciones de 70 como gasto bajo, 94 observaciones de un total de 104 como gasto medio y 27 de 37 como gasto alto. Esto, se traduce en una tasa del 67% de verdaderos positivos clasificados y un 31% de falsos negativos clasificados por el modelo, resultados mejorables pero aceptables ya que el modelo está siendo capaz de clasificar gran parte de los datos dentro de su categoría correctamente. A pesar de que estas métricas indican que el modelo es aceptable, el AUC (área bajo la curva) obtenida, ha sido muy baja, con una puntuación de 0´18. Esto indica, al contrario que el *accuracy*, que el modelo modeliza bien los datos ni tiene buena capacidad de generalización. El hecho de tener un *accuracy* relativamente bueno y un AUC tan bajo se debe a que los datos pueden estar desequilibrados o sesgados hacia una de las categorías. Esto quiere decir que puede ocurrir que en nuestro *dataset*, existan muchas observaciones pertenecientes a una de las categorías como por ejemplo "gasto medio" y muy pocas asignadas en el resto, lo que impide que el modelo sea capaz de llevar a cabo una clasificación.

Finalmente, si comprobamos la validez del modelo de red neuronal de clasificación entrenado para el periodo de 2022, observamos en la Figura 3.18, en este caso, el modelo ha sido capaz de clasificar correctamente 24/40, 60/71 y 40/52 observaciones para cada una de las categorías (gasto bajo, medio y alto) respectivamente. El nivel de *accuracy*

obtenido ahora ha sido de 0´47, también inferior que en el caso de los datos de entrenamiento y en este caso, puesto que se encuentra por debajo del 50%, no podemos considerar que el modelo sea bueno como clasificador de nuestra variable objetivo.

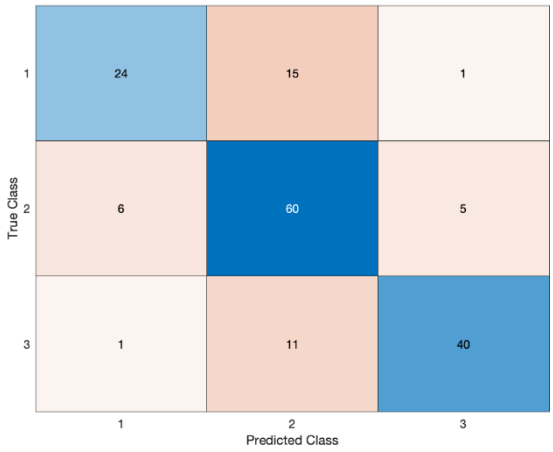

Figura 3.17. Matriz de confusión modelo de clasificación red neuronal 2022. Elaboración propia

Adicionalmente, los valores de TPR y FNR han sido de 0´60 y 0´38 respectivamente, por lo que la tasa de falsos negativos se considera demasiado elevada. Sin embargo, a pesar de existir una gran diferencia de *accuracy* entre este modelo y el anterior, el AUC ha sido de 0´17 por lo que difiere solo en una décima del anterior. Esto implica de nuevo, que nuestros datos podrían estar desequilibrados o sesgados hacia una de las categorías y por ello el modelo no estaría siendo capaz de predecir correctamente los valores de nuestra variable objetivo.

Capitulo 4

IMBALANCED DATA: modelos de clasificación

En este último capítulo del trabajo, se va a realizar un estudio de los *outliers* o valores menos frecuentes en nuestros datos. Para ello, se van a entrenar diferentes modelos de clasificación usando una técnica conocida como *imbalanced data*, que se emplea para evaluar la probabilidad de que se dé un evento determinado teniendo unos datos desequilibrados, lo cual ocurre en nuestros datos como se pudo observar en los modelos de clasificación obtenidos en el capítulo anterior. Los datos desequilibrados en el aprendizaje automático se refieren a una situación en la que la distribución de categorías entre los datos está desproporcionada. Esto significa que una categoría puede tener significativamente menos ejemplos que las otras por lo que mediante algoritmos de *imbalance data* se pretende construir modelos de clasificación y estudiar la probabilidad de que se den dichos valores. Esta técnica se emplea en situaciones especialmente relacionadas con el campo de la medicina, para predecir la probabilidad de padecer ciertas enfermedades como puede ser el cáncer.

Para llevar a cabo este proceso, se necesita convertir la variable objetivo "gastototal" en binaria (debe tomar valores 0 y 1). Para ello, se han tomado los datos usados para entrenar los modelos de clasificación anteriores en ambos periodos, pero esta vez, en lugar de clasificar la variable "gastototal" como gasto alto, medio o bajo, queremos observar aquellos valores que más atípicos en los datos por lo que se ha estudiado un percentil al 10%, como referencia a la hora de hacer una clasificación basada en 0 y 1. A partir de esto, se ha determinado hacer una clasificación binaria de los datos de la siguiente manera:

- Gasto total > 3000 se ha considerado valor 1
- Gasto total < 3000 se ha considerado valor 0

De esta forma, se pretende observar el comportamiento de los datos en la casuística de lo que sería para nuestro estudio concreto el turismo de lujo, que son aquellos turistas que gastan más de 3000€ y que suponen alrededor de un 10% de las observaciones de nuestro *dataset*. Es decir, se quiere entrenar modelos capaces de predecir la probabilidad de que

ocurra un evento concreto, que en nuestro caso es el turismo de lujo (asignado con valor 1). Se han entrenado para ambos periodos, un total de cuatro modelos con características diferentes:

- **Regresión Logística**. Es un tipo de algoritmo de *machine learning* para llevar a cabo regresiones en problemas de clasificación de tipo binario. Esta función, modela la probabilidad de que se den algunas de las dos clases estudiadas (0 o 1)
- **Árbol de decisión**. Es otro algoritmo que permite estudiar problemas tanto de regresión como de clasificación usando una partición jerárquica de los datos en subconjuntos más pequeños donde cada nodo del árbol representa una decisión para una característica determinada.

Adicionalmente, se han entrenado unas versiones avanzadas y más complejas del árbol de decisión conocidas como *XGBboost* y *Random Forest*. Estos métodos emplean también métodos jerárquicos de decisión con la diferencia de que *XGBboost* por ejemplo, combina las predicciones de diferentes árboles de decisión para crear un modelo más preciso utilizando un gradiente que va corrigiendo de forma iterativa los errores cometidos en los árboles anteriores consiguiendo así mejorar el modelo. *Random Forest*, aunque es muy similar a *XGBboost,* utiliza una técnica de optimización diferente conocida como *bagging*, que consiste en hacer predicciones independientes y después coger la media de las predicciones encontradas en todos los modelos.

Una vez entrenados los cuatro diferentes algoritmos, se ha obtenido para ambos periodos 2019 y 2022, que el algoritmo de *machine learning* capaz de clasificar el turismo de lujo mejor, ha sido el *XGBboost*. Para determinar el mejor modelo de entre todos los que se han entrenado, lo que se busca es encontrar aquel capaz de conseguir por un lado, un buen *accuracy,* y por otro, un *f1-score* [10] lo más alto posible para ambas categorías 0 y 1.

[10] F1-Score es una medida de precisión en la clasificación binaria donde su valor se encuentra en un rango entre 0 y 1, siendo el 1 una precisión perfecta y el 0, una mala capacidad predictiva. Esta métrica se calcula obteniendo una media entre precisión y sensibilidad, donde la precisión mide el número de verdaderos positivos de entre todas las predicciones realizadas por el modelo, y la sensibilidad la proporción de verdaderos positivos identificados correctamente de ente los casos reales positivos.

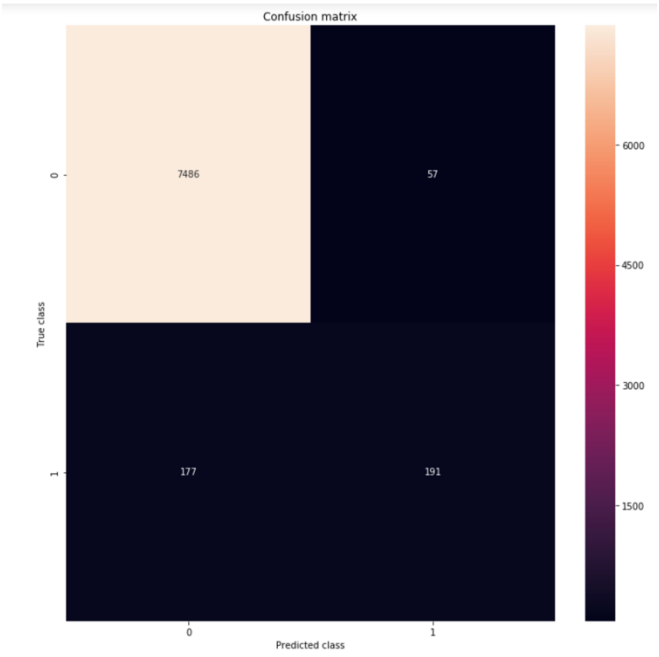

Figura 4.1. Modelo *XGBboost* de clasificación binaria 2019. Elaboración propia.

En este caso, el modelo ha clasificado correctamente 191 turistas de lujo (1), en comparación con 177 que no ha logrado identificar, y 7486 turistas de no lujo (0) con tan solo 57 observaciones asignadas de forma errónea. Adicionalmente, se ha conseguido un *accuracy* del 0´97%, que es un valor muy elevado e indicador de la buena capacidad predictiva del modelo, junto con unas *f1-scores* de 0´98 para los 0, y 0´62 para los 1, que igualmente son valores muy buenos que demuestran el buen desempeño del modelo como clasificador de turistas de lujo y una buena capacidad de modelar los datos.

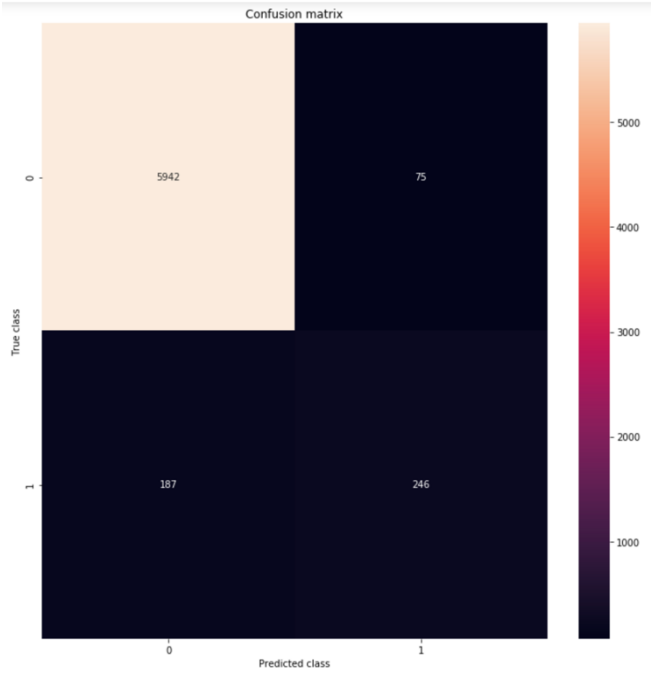

Figura 4.2. Modelo *XGBboost* de clasificación binaria 2022. Elaboración propia.

Finalmente, el modelo *XGBboost* de 2022, que también ha resultado ser el mejor de entre todos los modelos entrenados, ha sido capaz de predecir muy bien cada una de las categorías, donde ha clasificado correctamente a 246 turistas de lujo (1) versus 187 incorrectamente, y a un total de 5942 turistas de no lujo (0) clasificando tan solo 75 incorrectamente. Esto, se traduce en una *f1-score* de 0´98 para la categoría 0, y 0´65 para la categoría 1. Estos resultados obtenidos son bastante buenos y demuestran una buena capacidad de clasificación de los turistas de lujo donde se comete un error bajo en los 0 o turistas de no lujo. Del modelo, donde además se ha conseguido un *accuracy* del 96%, de tal forma que nuestro modelo es capaz de clasificar correctamente al 65% de los turistas de lujo correctamente cometiendo un error muy bajo en los turistas de no lujo.

Si comparamos ambos modelos, concluimos con que ambos modelos han obtenido resultados muy similares, demostrando un desempeño muy bueno y una gran capacidad predictiva a la hora de clasificar a lo que hemos designado como turistas de lujo dentro del sector en nuestro estudio, que son aquellos que en su estancia han tenido un gasto total igual o superior a 3000€. Sin embargo, podemos concluir con que el modelo obtenido para el periodo de 2022 es algo mejor, pues el *f1-score* obtenido para clasificar a los turistas de lujo es tres décimas superiores al de 2019.

Capítulo 5

CONCLUSIONES

En este trabajo, se ha analizado la relación entre variables relevantes para el turismo que mayor impacto suponen en el gasto turístico de visitantes extranjeros, con el fin de comparar si ha habido cambios significativos en dicha variable con respecto al periodo anterior y posterior a la pandemia. Para ello, se han obtenido los datos a partir de la encuesta de turismo EGATUR, correspondiente a los meses de verano para los años 2019 y 2022. Se han implementado técnicas de *feature selection* y *machine learning* para ser capaces de encontrar modelos tanto de regresión como de clasificación capaces de predecir el gasto turístico.

En el análisis preliminar de la variable, se ha podido observar mediante medidas estadísticas tradicionales que el gasto turístico medio antes y después de la pandemia realizando una estancia de duración similar equivalente a 9 días, ha sufrido un incremento de casi el 11%, lo cual es un resultado esperable debido la crisis del sector y la subida de precios como consecuencia de la inflación y la disminución de la demanda del sector. Por otro lado, se ha apreciado una fuerte dependencia entre esta variable y el número de pernoctaciones. El número de pernoctaciones realizadas (o lo que es lo mismo, la duración de la estancia), ha demostrado tener una gran relevancia a la hora de estudiar el gasto turístico y supone un gran impacto sobre la variable en comparación con el resto de las variables estudiadas tanto en 2019 como 2022. Con relación a lo mencionado, se ha observado también mediante la técnica de *feature selection* que, para ambos periodos además de las pernoctaciones, las variables "pais" y "medio de transporte" han demostrado ser las más relevantes para explicar el gasto turístico. Sin embargo, a pesar de que en ambos periodos (pre y post-pandemia), la variable del gasto está principalmente afectada por estas tres variables, sí que se aprecia una diferencia en cuanto al nivel de importancia que suponen para cada periodo (ver Figuras 2.5 y 2.6). En 2022, el número de pernoctaciones supone un peso mayor para el gasto que para 2019, y, por el contrario, el país de procedencia y el medio de transporte son más relevantes en 2019 que en 2022. Esto demuestra que el sector se ha visto afectado por la pandemia principalmente debido a factores intrínsecos del país como la inflación o las regulaciones gubernamentales

impuestas, y no tanto por factores extrínsecos como el país de procedencia del turista o el medio de transporte empleado. Por otro lado, para evaluar los modelos construidos, previamente entrenados usando estas tres variables para predecir el gasto, tanto de regresión como de clasificación, se utilizó un subconjunto de los turistas que viajaron a la comunidad autónoma de Cataluña. En este proceso, se obtuvieron buenos resultados en ambos modelos, pero el modelo GPR obtenido en 2019, consiguió un mejor desempeño y cometió menor error de predicción que el modelo de red neuronal de 2022. En el caso de los modelos de clasificación, se hizo una transformación de la variable del gasto en categorías 1,2 y 3 que representaban un gasto bajo, medio y alto respectivamente. Sin embargo, al contrario que en los modelos de regresión, los resultados obtenidos no fueron los esperados, ya que en el proceso de entrenamiento se obtuvieron buenos niveles de *accuracy*, mientras que al evaluarlos con el subconjunto de Cataluña, se consiguió un mal desempeño, con valores elevados de FNR y discrepancias significativas entre *accuracy* y AUC. Esto fue un indicativo para detectar que, al clasificar el gasto en dichas categorías, podía haber un problema de desequilibrio debido a que la mayoría de los valores se podían encontrar entre dos de las categorías siendo difícil de clasificar el gasto.

En consecuencia, se ha estudiado posteriormente el comportamiento de la variable del gasto turístico empleando la técnica de *imbalanced data,* técnica que permite analizar datos no balanceados, con el objetivo de estudiar el turismo de lujo, que en nuestros datos serían los turistas que tienen un gasto superior a 3000€. En este caso, se consiguieron modelos muy buenos de clasificación, capaces de asignar correctamente un elevado porcentaje de valores en cada categoría y con puntuaciones f1 sustanciales a la hora de clasificar. En concreto, se obtuvieron dos modelos XGBboost donde el correspondiente al periodo de 2022 resultó tener algo mejor de *performance* consiguiendo clasificar correctamente al 65% como turistas de lujo. Esta técnica, ha resultado interesante para obtener modelos capaces de clasificar los valores atípicos encontrados en nuestros datos, ya que se pudo observar que existían algunos *outliers* en la variable gasto, donde muchos turistas mostraban un gasto muy superior. Esto resulta especialmente útil para diferentes ramas del sector turístico como la hostelería o las agencias de viaje, para poder detectar estos perfiles de turistas de lujo, dispuestos a gastar más dinero, con el fin de poder ofrecerles servicios más exclusivos y establecer estrategias de acorde con este segmento.

Finalmente, quiero mencionar que, como futura línea de investigación, sería conveniente evaluar los modelos de clasificación empleando una categorización diferente donde se obtenga una distribución de los valores por categoría más equitativa. Por otro lado, mencionar también que hoy en día existen técnicas más complejas que se encuentran en proceso de investigación como *Deep Gaussian Processes* o procesos gaussianos profundos que podrían ser de gran utilidad para optimizar especialmente los modelos GPR así como mejorar tareas de predicción como las llevadas a cabo en este trabajo, dando lugar a mejores resultados, predicciones más precisas del gasto turístico, debido a su capacidad de encontrar patrones más complejos y no lineales en los datos y a su flexibilidad, pues requieren menos datos de entrenamiento para obtener buenos resultados. Construir modelos capaces de predecir el gasto turístico resulta de gran utilidad para el sector ya que de esta manera se pueden realizar estudios de mercado y segmentación de los clientes de forma más precisa, que permitan detectar mejor sus necesidades para ofreciendo servicios y paquetes turísticos acorde con sus perfiles.

	2019	2022
Gasto medio total	1,298 €	1,452 €
Variables más relevantes (feature selection)	Pernoctaciones, pais y medio de transporte	Pernoctaciones, pais y medio de transporte
Modelos de regresión	★ GPR	Red Neuronal
Modelos de clasificación	Red Neuronal	Red Neuronal
Modelos imbalanced	XGBboost	XGBboost ★

Figura 5.1. Resumen de resultados obtenidos. Elaboración propia

*Las estrellas muestran los modelos que mejor han optimizado la variable del gasto turístico en cada categoría, salvo en los modelos de clasificación donde el resultado no ha sido óptimo en ninguno.

BIBLIOGRAFÍA

[1] Instituto Nacional de Estadística, “Cuenta Satélite del Turismo de España (CSTE). Revisión estadística 2019: Serie 2016-2019”, 2020. Recuperado de https://www.ine.es/prensa/cst_2019.pdf

[2] DataEstur, “Llegadas de turistas internacionales a nivel mundial”, 2018. Recuperado de https://www.dataestur.es/general/llegadas-de-turistas-internacionales-a-nivel-mundial/

[3] M.S, Bravo Cabria, "La competitividad del sector turístico." Boletín económico/Banco de España, vol 19, n9, p. 91-106 , 2004.

[4] C.Y. Koorndijk, “El efecto de la pandemia del coronavirus en el sector del turismo. Estudio contrastivo a escala internacional”, tesis doctoral, Universitat Politénica de Valencia, 2021.

[5] Datosmacro, “Turismo internacional en España”, Recuperado de https://datosmacro.expansion.com/comercio/turismo-internacional/espana

[6] G. C. Valiente, J. M. P., Forga & A. B. Romero, “Turismo en España, más allá del sol y la playa. Evolución reciente y cambios en los destinos de litoral hacia un turismo cultural”. *Boletín de la Asociación de Geógrafos Españoles*, no.71, pp 151-172,2016.

[7] Epdata, “Bienes culturales y patrimonio en España”, 2021. Recuperado de https://www.epdata.es/datos/bienes-culturales-patrimonio-espana-datos-graficos/358

[8] Statista, “Turistas internacionales por motivos culturales en España desde 2010 hasta 2019 (en millones”, 2021. Recuperado de https://es.statista.com/estadisticas/508224/turistas-internacionales-por-motivos-culturales-en-espana/

[9] Exceltur, “Informe Solytur 2019: análisis y perspectivas del sector turístico español”, 2021. Recuperado de https://www.exceltur.org/wp-content/uploads/2021/03/Solytur-2019-Mar2021.pdf

[10] C. Romero Dexeus & J. Prado Mascuñano, “La medición del gasto turístico en la Europa del euro: el caso español”. *ICE, los retos de las estadísticas del sector exterior*. Recuperado de http://www.revistasice.com/index.php/ICE/article/view/328/328

[11] Instituto Nacional de Estadística (INE), “ EGATUR: Encuesta de gasto turístico Noviembre 2022”, 2023. Recupero de https://www.ine.es/daco/daco42/egatur/egatur1122.pdf

[12] A. Orús, Statista, “Gasto total de los turistas internacionales en España por país de residencia”, 2021. Recuperado de https://es.statista.com/estadisticas/476305/gasto-total-de-los-turistas-internacionales-en-espana-por-pais-de-residencia/

[13] A. Orús, Statista, “Número total de casos confirmados de coronavirus en España entre el 25 de febrero de 2020 y el 17 de marzo de 2023”, 2023. Recuperado de https://es.statista.com/estadisticas/1104275/casos-confirmados-de-covid-19-por-dia-espana/

[14] Calveras, A. Santana, M. (2022). Desafíos para el crecimiento de la economía española en el contexto post-pandemia. Universidad de Islas Baleares. Recuperado de https://www.funcas.es/wp-content/uploads/2022/11/PEE-173_Calveras.pdf

[11] Statista, "Porcentaje del PIB aportado por el sector turístico en España de 2010 a 2022", 2022. Recuperado de https://es.statista.com/estadisticas/1082929/sector-turistico-porcentaje-del-pib-aportado-espana/

[12] Exceltur, "Impacto del Coronavirus sobre el sector turístico español", Perspectivas turísticas Nº73, 2020. Recuperado de https://www.exceltur.org/wp-content/uploads/2020/10/Informe-Perspectivas-N73-Balance-empresarial-y-escenario-impacto-Covid-19.pdf

[13] Instituto Nacional de Estadística (INE), "Encuesta de gasto turístico Diciembre 2022 y año 2020", 2021. Recuperado de https://www.ine.es/daco/daco42/egatur/egatur1220.pdf

[14] A. Calveras, & M. Santana, "El turismo en España ante el Covid-19: El efecto frontera en el turismo internacional e interregional". *Papeles de Economía Española*, (173), 161-228, 2022.

[15] O. Arce, "La economía española: impacto de la pandemia y perspectivas" , *Dirección General de Economía y Estadística*, Banco de España, 2021. Recuperado de https://www.bde.es/f/webbde/GAP/Secciones/SalaPrensa/IntervencionesPublicas/DirectoresGenerales/economia/Arc/Fic/arce260521.pdf

[16] Exceltur, "Impacto del Coronavirus sobre el sector turístico español", Perspectivas turísticas Nº73, 2020. Recuperado de https://www.exceltur.org/wp-content/uploads/2020/10/Informe-Perspectivas-N73-Balance-empresarial-y-escenario-impacto-Covid-19.pdf

[17] Statista, "Porcentaje del PIB aportado por el sector turístico en España de 2010 a 2021", 2023. Recuperado de https://es.statista.com/estadisticas/1082929/sector-turistico-porcentaje-del-pib-aportado-espana/

[18] Tourspain, "Encuesta de población activa", Cuarto trimestre 2020 https://www.tourspain.es/es-es/ConocimientoTuristico/PoblacionActiva/epa4T20.pdf

[19] Instituto Nacional de Estadística (INE), "Encuesta de gasto turístico (EGATUR) diciembre 2021 y año 2021", 2022. Recuperado de https://www.ine.es/daco/daco42/egatur/egatur1221.pdf

[20] La Moncloa, "Plan de Impulso del Sector Turístico Español", 2020. Recuperado dehttps://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/industria/Documents/2020/20062020_PlanTurismo.pdf

[21] Epdata, "Gasto de los turistas extranjeros", actualización marzo 2023. Recuperado de https://www.epdata.es/datos/turistas-turismo-espana/64/espana/106

[22] Instituto Nacional de Estadística (INE) . "Estadística de Movimientos turísticos en frontera (FRONTUR) noviembre 2022", 2023. Recuperado de https://www.ine.es/daco/daco42/frontur/frontur1122.pdf

[23] Instituto Nacional de Estadística (INE), "Encuesta de gasto turístico noviembre 2022", 2023. Recuperado de https://www.ine.es/daco/daco42/egatur/egatur1122.pdf

[24] Exceltur, "Informe de Perspectivas Nº83: Balance del año 2022 y expectativas para 2023". 2023. Recuperado de https://www.exceltur.org/wp-content/uploads/2023/01/Informe-Perspectivas-N83-Balance-del-ano-2022-y-expectativas-para-2023.pdf

[25] Piña Miranda, L. (2018). "Modelación y predicción del gasto de turistas en España enfocado desde el análisis de datos". Recuperado de https://dspace.uib.es/xmlui/handle/11201/149300

[26] A. Musonera, E. Musabaganji & H. Musahara. “Determinants of tourism demand using machine learning techniques”, Vol 11, Nº 1, pp 770-780, 2022.

[27] S. Savaiano, & C. Drago, “Cluster Validation in Unsupervised Machine Learning with Application to the Analysis of the Tourism Demand in Italy after COVID-19 Lockdown”, 2021.

[28] A. Peláez Verdet, & P. Loscertales Sánchez, “Evaluación de la sostenibilidad económica del turismo de cruceros. Una aproximación metodológica en el Litoral Mediterráneo español”. *Revista de Ciencias de la Administración y Economía*, *8*(15), 101-115, 2018.

[29] R. Guerra Reyes, “Predicción de Turismo en España mediante Aprendizaje Automático Concatenado a Datos Globales de la Pandemia”, 2021.

[30] K.P. Murphy, “Machine Learning: a probabilistic perspective”, Morgan Kauffmann, 2012.