COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

# LA INCIDENCIA DEL CORONAVIRUS SOBRE EL SECTOR TURÍSTICO ESPAÑOL

Análisis cuantitativo de la evolución del sector turístico en temporada alta antes y después de la crisis sanitaria

Autor: Laura Mateos Rojas

Director: Luis Ángel Calvo Pascual

MADRID | Junio 2023

## RESUMEN

La crisis sanitaria causada por el coronavirus ha afectado a diversos ámbitos de la sociedad. Tomando como base los datos obtenidos en las encuestas EGATUR de los meses de temporada alta anteriores y posteriores a la pandemia, este trabajo estudia la incidencia de la pandemia sobre el sector turístico español y, en concreto, sobre el gasto de los turistas internacionales que visitan España. Se analiza con diversas técnicas cuantitativas esta variable con el fin de comprobar si el gasto diario de los turistas es significativamente diferente antes y después de la crisis del coronavirus. Se observa diferencias significativas en las variables más relevantes, así como una reducción considerable en el gasto de los turistas debido a la pandemia, si bien en regresión discontinua se aprecia una recuperación progresiva en los meses posteriores.



**Palabras clave:** turismo, análisis cuantitativo, feature selection, aprendizaje automático, aprendizaje supervisado, regression discontinua, coronavirus, pandemia.



## ABSTRACT

The health crisis caused by the coronavirus has affected various areas of society. Based on data obtained from EGATUR surveys in the high season months before and after the pandemic, this paper studies the impact of the pandemic on the Spanish tourism sector and, specifically, on the expenditure of international tourists visiting Spain. This variable is analyzed with various quantitative techniques in order to check whether the daily expenditure of tourists is significantly different before and after the coronavirus crisis. Significant differences are observed in the most relevant variables, as well as a considerable reduction in tourist spending due to the pandemic, although in regression discontinuity a progressive recovery is observed in the following months.



**Key words:** tourism, quantitative analysis, feature selection, machine learning, supervised learning, regression discontinuity, coronavirus, pandemic.

# ÍNDICE

**LISTADO DE ABREVIATURAS**

EGATUR – Encuesta de Gasto Turístico

FRONTUR – Estadística de Movimientos Turísticos en Frontera

FS – *Feature Selection*

INE – Instituto Nacional de Estadística

MI – *Mutual Information*

ML – *Machine Learning*

OMS – Organización Mundial de la Salud

P33 – Percentil 33

P66 – Percentil 66

RD – Regresión Discontinua o *Regression Discontinuity*

RFR – *Random Forest Regression*

RNA – Redes Neuronales

RSV – *Regresión de Soporte Vectorial*

SVM – *Support Vector Machine* o Máquina de Soporte Vectorial

UCI – Unidad de Cuidados Intensivos

# 1. INTRODUCCIÓN Y CONTEXTO

## 1.1. Periodo histórico y económico

En este trabajo se analizará la historia y evolución del turismo en España a lo largo del tiempo con la finalidad de dilucidar el impacto que la crisis del Covid-19 ha tenido sobre este sector. Este estudio se centra en los periodos de junio, julio y agosto, desde el año 2018 a el año 2022, lo que nos permitirá conocer la situación del turismo en España en dos tramos distintos:

i. La etapa prepandémica, con anterioridad a la emergencia sanitaria generada por el coronavirus,
ii. La etapa postpandemia, que comprende el periodo durante la crisis y posterior a la crisis, es decir, la actualidad.

Uno de los sectores que más se ha visto afectado por la pandemia provocada por el coronavirus ha sido el sector turístico. La Organización Mundial del Turismo de las Naciones Unidas muestra en su compilación realizada "UNWTO tourism dashboard" sobre el impacto del coronavirus en el turismo, que España está en el primer puesto en la lista de destinos europeos de mayor vulnerabilidad en términos de PIB, correspondiendo el turismo a un 12% del PIB total del país.

### *1.1.1. El turismo en España*

El turismo se trata de una actividad económica en la que España se trata se sitúa como uno de los países lideres a nivel internacional y cuya popularidad y crecimiento ha sido progresivo a lo largo de las últimas décadas. A este respecto, de acuerdo con el Instituto Nacional de Estadística (s.f.) y con Briones (2023), actualmente se trata de una actividad económica de gran relevancia para el país: supuso un 12,2% del PIB en 2018, 12,6% en 2019, 5,8% en 2020, 8% en 2021 y 12,2% en 2022.

El turismo en España ha ido evolucionando estratégicamente a lo largo de las décadas. Si bien en España ha predominado el turismo de "sol y playa", en la actualidad se busca impulsar estrategias centradas en la calidad (Ministerio de Industria, Comercio y Turismo, s.f.).

A pesar de la popularidad de España como destino turístico en la actualidad, su conformación como destino turístico internacional de surgimiento relativamente reciente.

En esta línea, si bien el inicio del turismo en España puede establecerse a finales del siglo XIX, cuando el turismo era un privilegio, este potencial únicamente comenzó a verse a partir del siglo XX, especialmente, en su segunda mitad, cuando se consolida lo que conocemos como "turismo de masas". El turismo se transforma de una actividad "con motivación con motivación diversificada alrededor de los paisajes nuevos y exóticos, la riqueza artística y patrimonial, las playas, los balnearios o estaciones termales y, en general, los espacios para la representación del prestigio social" (Vallejo Pousada, 2002, p.205) a un turismo masivo en torno al sol y la playa. En este sentido, desde la República, bajo el eslogan "*Spain is different"*, pasando por el periodo del franquismo y la fijación de España como destino turístico, hasta la llegada de la democracia y la campaña "España, todo bajo el sol" la cual ha sido una de las más reputadas y afamadas, el turismo ha sido una actividad de impacto significativa para la economía española. (Moreno Garrido y Villaverde, 2019).

Para el desarrollo y popularización del turismo fue esencial el desarrollo de los medios de transporte como el ferrocarril, el automóvil y el avión. Estos eventos "fueron decisivos en la irrupción del turismo como un fenómeno sociológico y económico de importancia creciente" (Vallejo Pousada, 2002).

Siguiendo a Vallejo Pousada (2002), el turismo ha pasado de ser un bien lujoso y al alcance de unos pocos, a poder considerarse incluso de primera necesidad. Por ello, a pesar de que la demanda turística es elástica, a partir de cierto nivel de renta, las personas no están dispuestas a renunciar completamente a ello. Además, las motivaciones turísticas se han diversificado, por lo que ahora hay muchos y muy distintos tipos de turismo.

De acuerdo con Deloitte (s.f.), tras la pandemia, el sector turístico se enfrenta a varios retos estructurales, entre ellos: "la creciente demanda social de un modelo turístico sostenible" (siendo un 83% los turistas que considera un asunto prioritario la sostenibilidad en el turismo), "la alta estacionalidad y concentración geográfica"; "la evolución hacia un turismo de mayor valor", y la "ausencia de estrategia transversal y coordinada a nivel nacional", entre varios otros.

### *1.1.2. La crisis del coronavirus*

La crisis sanitaria provocada por la enfermedad del coronavirus en 2019, comúnmente conocido como Covid-19, ha supuesto un impacto trascendente en la sociedad humana en todos los ámbitos de esta. De acuerdo con la Organización Mundial de la Salud (en adelante, OMS) (s.f., "Coronavirus"), se trata de una "enfermedad infecciosa causada por el virus SARS-CoV-2". Se trata de lo que en economía se denomina un cisne negro (en anglosajón, *black swan*), habiendo sido comparado con el escenario económico de la Segunda Guerra Mundial (Nicola et al., 2020). El primer caso conocido de esta enfermedad se dio en Wuhan, China, el 31 de diciembre de 2019, fecha desde la cual la enfermedad fue expandiéndose globalmente (OMS, s.f., "Brote de enfermedad por coronavirus (COVID-19)"). A raíz de ello, el 11 de marzo de 2020, el Dr. Tedros Adhanom, Director General de la OMS declaraba la situación como pandemia durante una rueda de prensa (Sevillano, 2020).

Desde un punto de vista **sanitario**, durante los primeros meses del brote, el coronavirus tuvo una rápida propagación, causando distintos grados de enfermedad. España en concreto ha sido uno de los países más afectados por esta crisis sanitaria ya que, de acuerdo con Casas-Rojo et al. (2020), "es uno de los países del mundo con mayor número de pacientes con infección por SARS-CoV-2". En España, el primer caso se registró el 31 de enero de 2020 y, desde entonces, a 2 de junio de 2023 se han registrado un total de casi catorce millones de casos confirmados notificados en España y un total de 121.416 casos fallecidos notificados (Ministerio de Sanidad, s.f., "Situación actual"). Siguiendo el estudio clínico llevado a cabo por Casas-Rojo et al. (2020), el mayor número de casos se encontraban en el rango de edad entre los 30 y los 64 y los 65 y los 79 años. Muchos de los casos más graves requirieron ingreso en las unidades de cuidados intensivos (en adelante, UCI), requiriendo incluso sistemas de soporte ventilatorio.

El coronavirus, además de las personas afectadas, también ha afectado a nivel institucional del sistema de salud. En este sentido, la crisis sanitaria ha generado una presión extrema sobre los sistemas de salud a nivel mundial. Ha supuesto la saturación de los hospitales debido a la alta demanda de atención primaria médica y de urgencias. En este sentido, el colapso de las UCIs trajo consigo también el planteamiento de numerosos debates éticos en relación con la necesidad de establecer criterios de selección para acceder a estas unidades. Estos dilemas éticos trataron como posibles criterios de

corte para el ingreso o no en las UCIs la edad, la gravedad de la enfermedad o la probabilidad de supervivencia (*cfr*. Benito, 2020 y Esteban y Méndez, 2020).

Asimismo, esta situación tuvo como consecuencia un impulso significativo en el ámbito de la investigación científica, así como la inversión de grandes cantidades de recursos económicos y humanos, con el fin de avanzar en la búsqueda de tratamientos efectivos y en el desarrollo de vacunas contra la enfermedad. En este aspecto, la urgencia de poner una solución científica a la pandemia llevó a una movilización sin precedentes de la comunidad científica, suponiendo la implicación de numerosos expertos de diversas disciplinas en un trabajo conjunto con el objetivo de comprender mejor la enfermedad y sus implicaciones (transmisión, síntomas, afecciones, tratamientos, efectos a largo plazo/secuelas, entre muchas otras) (*cfr*. Parlamento Europeo, s.f.)

En lo que a las vacunas respecta, como se venía mencionando, se realizaron esfuerzos masivos con el fin de acelerar el desarrollo de las vacunas. De este proceso resultaron numerosas vacunas desarrolladas por distintas compañías farmacéuticas. Siguiendo al Gobierno de España (s.f., "Preguntas y respuestas: ¿Qué vacunas tendremos disponibles es España?") dichas vacunas tuvieron que someterse a los procedimientos de la Agencia Europea de Medicamentas para determinar su seguridad y eficacia para el uso. Son un total de siete vacunas las que la Comisión Europea ha negociado con dichas farmacéuticas entre ellas, la vacuna "Comirnaty" de BioNTech y Pfizer, la vacuna "Spikevax" de Moderna, la vacuna "Vaxzevria" de AstraZeneca y la vacuna "Jcovden" de Janssen (AEMPS, s.f., "Información de vacunas autorizadas"), son algunas de las más conocidas. Gracias a la promoción de medidas preventivas, las campañas de vacunación y los grandes esfuerzos en todos los ámbito, la OMS declaró el fin de la emergencia internacional sanitaria generada por el Covid-19 el 5 de mayo de 2023 (Linde, 2023). Actualmente se han administrado más de 105 millones de dosis y hay más de 40 millones de personas con la pauta vacunal completa en España (Ministerio de Sanidad, s.f., "Situación actual").

Por otra parte, desde un punto de vista **social** y psicológico, la enfermedad causada por el coronavirus ha tenido un impacto transcendente en la vida de las personas a nivel de la social en numerosos aspectos. Por un lado, las medidas de confinamiento, distanciamiento social y restricciones, no solo ha sido transcendente a nivel económico, sino que también lo ha sido para la vida cotidiana de los ciudadanos y sus interacciones sociales (*cfr*. Paredes et al., 2021). La huella de la pandemia es visible en la salud mental de muchos

ciudadanos en todo el mundo a causa de ese aislamiento social obligado con motivo del confinamiento y, posteriormente, por las restricciones e incluso el temor a la enfermedad sobrevenido. La situación de incertidumbre vivida también ha tenido un impacto negativo en la salud mental de las personas. Paredes et al. (2021) afirma que el nivel de amenaza percibido, este temor, se ve influido por numerosos factores como la probabilidad de vulnerabilidad o de contagio, así como la dureza de las consecuencias de la enfermedad en caso de infección. A mayor temor de la persona a la enfermedad, mayor impacto negativo sobre la salud mental de este.

En este sentido, se han realizado multitud de estudios en relación con lo expuesto, así Cullen et al. (2020) defienden que hay una creciente evidencia de que la pandemia causada por el Covid-19 y las medidas adoptadas para paliar y frenar sus efectos puede acrecentar la severidad de enfermedades mentales pre-existentes e incluso generar nueves síntomas en aquellos individuos que no tenían enfermades mentales con anterioridad. Este mismo estudio menciona una declaración formal de la OMS por la que advierte concretamente del riesgo al que se enfrenta personal sanitario de sufrir ansiedad, depresión o estrés post-traumático.

Siguiendo a Paredes et al. (2021), las investigaciones anteriores han demostrado que los acontecimientos inesperados como pueden los desastres naturales o, como es el caso, las pandemias, generan efectos emocionales significativos sobre las personas, en detrimento de su bienestar general. En este aspecto, estos autores alegan que la pandemia ha elevado el número de casos de depresión, ansiedad, frustración, aislamiento, ira, entre otros trastornos.

Por otra parte, desde el punto de vista **económico** la pandemia tenido repercusiones en diversos ámbitos. Muchos trabajadores quedaron expuestos a la pérdida de ingresos y a la inestabilidad e inseguridad laboral, al mismo tiempo que muchas empresas y sectores vieron su actividad desvanecida o extremadamente reducida. Como consecuencia, esto tuvo un efecto directo sobre el empleo, la producción y las cadenas de suministro (Parlamento Europeo, 2023).

Así, en el ámbito laboral, siguiendo a Benavides y Silva-Peñaherrera (2022), la crisis sanitaria ha incrementado exponencialmente el teletrabajo (definido por estos autores como “el trabajo realizado en domicilio utilizando equipos electrónicos”). En este sentido, la Organización Mundial del Trabajo estimó en base a una encuesta realizada que el 17,4% de los individuos a nivel mundial teletrabajaron durante la pandemia; para el

caso europeo, fue el 37% y en España, el 30%. Como añadido a todo ello, resulta importante resaltar el hecho de que la crisis sanitaria supuso una aceleración de la digitalización y una transformación de los espacios de trabajo y, en consecuencia, trajo consigo numerosas oportunidades y retos para empresas, trabajadores y territorios (Parlamento Europeo, 2023) y, con ello, el impulso del teletrabajo.

Sin embargo, de acuerdo con Nicola et al. (2020), para muchas funciones dentro de una empresa como puede ser una manufacturera, el teletrabajo no es una opción viable y lo mismo ocurre con muchas otras empresas de distintos sectores en los que la presencialidad es requerida.

De acuerdo con Benavides y Silva-Peñaherrera (2022), la crisis del coronavirus también ha elevado los niveles de desempleo, especialmente al inicio de la pandemia, suponiendo a nivel mundial más de 300 millones de despidos de empleos a jornada completa; en España concretamente, de acuerdo con el titular del periódico El País, se destruyeron 622.600 empleos y la tasa de desempleo ascendió hasta el 16,3% en 2020 (Gómez, 2021).

Si bien todos los sectores han sufrido el impacto del coronavirus y de las medidas impuestas como consecuencia de este, cada uno de ellos ha sido afectado por dicho impacto de forma diferente. Esto se refleja en un descenso del PIB del 10,8% en 2020 para el caso de España que, de acuerdo con Benavides y Silva-Peñaherrera (2022), es un "hecho solo superado por el registrado durante la Guerra Civil española".

En este sentido, de acuerdo en todo ello con Nicola et al. (2020), en el sector primario, la agricultura ha sufrido la caída de demanda de los servicios de restauración y hostelería, provocando una caída del 20% en los precios de los productos agrícolas.

Por otra parte, en el sector secundario, más del 80% de las empresas encuestadas de Reino Unido preveía un descenso del volumen de negocio en los dos trimestres siguientes y el 98% declaró su preocupación por el impacto negativo que la pandemia podría llegar a tener en las operaciones empresariales.

En tercer lugar, en el sector terciario también se ha sufrido un impacto significativo a raíz de la pandemia. Las medidas de confinamiento y de distanciamiento social han afectado al sector educativo en muchos ámbitos: las aquellas familias con acceso a recursos tecnológicos y a internet han podido garantizar la continuación de la educación de sus hijos durante estos periodos de aislamiento, mientras que las que no disponían de estas herramientas bien no han podido o han tenido que enfrentarse a mayores obstáculos; han

aumentado las tasas de abandono escolar. Además, otro ejemplo de influencia del Covid-19 en el sector educativo está en el impacto de este en la enseñanza universitaria y la suspensión por parte de las comunidades de investigación académicas de aquellas investigaciones sobre temas no relacionado con el Covid-19.

El sector financiero ha sufrido la pandemia al ver una caída dramática en los mercados bursátiles mundiales, creando un entorno muy volátil y con unos niveles de liquidez críticos. Para solventar este problema, muchos Bancos centrales han tenido que intervenir con el fin de garantizar el mantenimiento de la liquidez y tratar de mitigar el impacto económico. En Europa concretamente, se han desarrollado varios planes y programas de recuperación como son el "Pandemic Emergency Purchase Programme (PEPP)" y los fondos "NextGenerationEU" (Banco Central Europeo, s.f. y Comisión Europea, s.f.).

Si bien son muchos otros sectores los que se han visto afectados por la pandemia, resulta relevante poner especial atención en el **sector turístico** y de hostelería a efectos de esta investigación. En este sentido, la Universidad Carlos III de Madrid (s.f.) recalca numerosos aspectos en los que la pandemia ha afectado al sector turístico: por un lado, ha tenido un impacto económico debido a las restricciones de viaje y los cierres de fronteras, suponiendo una reducción drástica en la demanda turística y generando grandes pérdidas económicas para empresas y trabajadores del sector (en 2020, la llegada de turistas a España disminuyo en un 77% en comparación con 2019, si bien algunas comunidades autónomas sufrieron descensos aún mayores).

Asimismo, Sigala (2020) señala el cambio de mentalidad de los turistas a causa del coronavirus. La pandemia ha tenido un impacto significativo en las actitudes, intenciones y comportamientos futuros a la hora de viajar.

De acuerdo con Nicola et al. (2020), el sector turístico y hostelero son unos de los sectores más afectados por la pandemia, con repercusiones tanto en la oferta como en la demanda de viajes. En este sentido, grandes cadenas hoteleras han tenido que pedir grandes préstamos a modo preventivo y, el sector en general ha sufrido distorsiones internacionales e importantes caídas.

Con todo ello, de acuerdo con Naciones Unidas (2020), la recuperación económica en todo el mundo tras la crisis del coronavirus será un proceso largo y lento. Se trata de una recuperación que se desarrolla en un contexto incierto debido a que "las distorsiones en las cadenas de suministros, el repunte de la inflación y, ya en 2022, la guerra en Ucrania,

han obstaculizado la recuperación de la economía mundial tras la fase más aguda de la pandemia" (Banco de España, 2022, p.55).

### 1.2. Objetivos

En este estudio se trabajará con diversas metodologías cuantitativas las variables que afectan al turismo y, en especial, el gasto diario de los turistas durante sus visitas. Asimismo, resulta pertinente analizar los posibles cambios que pueda haber sufrido el modelo de turismo que recibe el país con motivo de la pandemia.

Estudiaremos estos aspectos en dos contextos temporales distintos: del 2018 al 2019, del 2020 al 2022. Para ello, los métodos que se emplearán son: análisis estadístico, *Feature Selection*, *Machine Learning* y Regresión Discontinua.

Estas técnicas nos permitirán conocer cuáles son los factores más importantes que influyen en el turismo y los modelos que mejor se ajustan a estos factores. Así, con todo ello, se podrá responder a la pregunta de si el modelo de turismo y la cantidad de gasto del turista ha cambiado tras la crisis sanitario o si, por el contrario, no se ha visto alterado.

De esta forma, los objetivos concretos de este estudio se concretan en los siguientes cuatro objetivos:

- Objetivo 1: realizar un análisis descriptivos de los datos, así como un contraste de hipótesis con el fin de comprobar si ha habido diferencias significativas entre el gasto turístico diario de los turistas antes y después de la pandemia causada por el Covid-19.
- Objetivo 2: establecer cuáles son las variables más influyentes en la explicación del gasto diario de los turistas, así como comprobar si existen diferencias entre los dos periodos antes mencionados.
- Objetivo 3: se busca estudiar los modelos de *Machine Learning* para la variable gasto turístico diario empleando las variables identificadas como más relevantes, de cara a elegir el modelo que mejor se ajusta y explica esta variable antes y después de la pandemia y estudiar su significado.
- Objetivo 4: emplear la técnica de *Regression Discontinuity* con el fin de medir el efecto causal de la pandemia provocada por el coronavirus antes y después en el gasto de los turistas.

## 1.3. Metodología y estructura

### *1.3.1. Metodología*

Con el fin de lograr los objetivos expuestos anteriormente, se ha estudiado la situación del turismo en España y los estudios cuantitativos antecedentes sobre esta materia mediante una investigación bibliográfica por medio de artículos de investigación, revistas científicas y libros y manuales, entre otros. Por otro lado, para alcanzar los objetivos también se han llevado a cabo prácticas informáticas y reuniones con el director.

Para realizar el análisis cuantitativo, se ha utilizado una base de datos cuyo contenido ha sido extraído del INE y se han empleado herramientas de software matemático como Python, Matlab, SPSS y Excel.

Los datos empleados para este estudio han sido extraídos del Instituto Nacional de Estadística (en adelante, INE), concretamente, se han obtenido de la Estadística de Movimientos Turísticos en Frontera (en adelante, FRONTUR) y Encuesta de Gasto Turístico (en adelante, EGATUR). Estas encuestas estadísticas tienen por objeto: "medir el número de visitantes no residentes en España que llegan a nuestro país cada mes", "conocer las características principales de los viajes que realizan dichos visitantes", y determinar el gasto que los viajeros y turistas no residentes en España realizan durante sus viajes a este país (ya sea con o sin pernoctaciones) o durante su estancia en el territorio nacional español, dentro de sus fronteras (INE, 2022).

De acuerdo con el informe del INE sobre la metodología seguida en estos estudios, con motivo de estas estadísticas, tienen consideración de visita turística aquellos desplazamientos realizados fuera del entorno de la residencia habitual del individuo cuya duración sea inferior a un año, cuando el motivo de este viaje (negocios, recreo u otras motivaciones particulares) sea otro distinto de un empleo en empresa con domicilio en el lugar de visita. Por otro lado, dentro de las visitas turísticas se distinguen dos categorías: (1) viaje turístico, el cual se dará en los casos que haya, como mínimo, una pernoctación; y (2) excursión, cuando la visita fuera del municipio no incluya pernoctación alguna, si bien ha de durar un mínimo de tres horas, y no es parte de una rutina diaria (la frecuencia ha de ser inferior a una vez por semana).

El conjunto de datos contiene 26.325 observaciones con 10 variables distintas que recogen información relativa a distintos aspectos relacionados con el viaje: mes y año del viaje, el tipo de encuestado, la vía de salida, el país de residencia habitual, la Comunidad

Autónoma de destino principal del viaje, el total de pernoctaciones, el alojamiento principal, el motivo principal del viaje, uso de paquete turístico y el gasto diario del viaje.

A propósito de este estudio, se ha optado analizar los datos de los meses de junio, julio y agosto de los años comprendidos entre 2018 y 2022 ambos inclusive. Se ha seleccionado estos meses ya que estos forman la temporada alta de viaje en verano. También, esta selección temporal se encuentra relacionada con el objetivo del trabajo: estudiar el impacto del virus del coronavirus. Por ello, se distinguirán dos fases, una anterior a la crisis y otra posterior.

### *1.3.2. Estructura del TFG*

El presente trabajo está compuesto por seis capítulos diferenciados a lo largo de los cuales se dará respuesta a la cuestión planteada de acuerdo con la metodología expuesta en el apartado anterior.

En el primer capítulo de estudio, llamado "Análisis estadístico primario", se responde al Objetivo 1. Para ello, realizaremos un estudio de las principales variables que conforman la base de datos empleada para ese estudio. Se realiza un análisis de los principales estadísticos y distribuciones para conocer la información con la que se va a trabajar, así como para extraer información inicial que puedan ser provechosa para el estudio. Para ello se llevarán a cabo las pruebas de Levene y Kruskal-Wallis para estudiar distintos contrastes de hipótesis respecto de la varianza, la media y mediana de la variable objetivo.

En el segundo capítulo de "Feature selection", se busca alcanzar al Objetivo 2. Con este propósito, se emplearán tanto técnicas tradicionales como técnicas modernas para estudiar cuáles son las variables más relevantes para la predicción de la variable objetivo (el gasto diario del turista) y cuál es el peso de estas. Esta selección es fundamental de cara a la optimización del modelo predictivo que se elaborará en el capítulo siguiente a este.

Posteriormente, con el fin de dar respuesta al Objetivo 3, en el capítulo tercero, denominado "Machine Learning", se estudiará, cuál es el modelo predictivo de clasificación óptimo. Para la búsqueda de ambos modelos se emplearán las variables que concluyeron ser que mejor explican el modelo.

Finalmente, en el capítulo cuarto se llevará a cabo una regresión discontinua de cara a responder al Objetivo 4. Se buscará comprobar si existen o no diferencias significativas

en el gasto diario de los turistas entre ambas fases, esto es, si esta variable ha sufrido un cambio.

### 1.4. Antecedentes

Resulta necesario enmarcar este estudio dentro de su contexto, el turismo, y para ello resulta necesario tener en consideración los antecedentes de la presente investigación.

Entre los artículos académicos precedentes a destacar, encontramos el publicado en 2016 por la Revista de Economía Aplicada, "Modelling tourism demando to Spain with Machine Learning techniques. The impact of forecast horizon on model selection" de los autores Claveria, Torra y Monte. En este otro estudio, los modelos empleados para la predicción son, por un lado, las llamadas Regresión de Soporte Vectorial (en adelante, RSV) y, por otro, las Redes Neuronales (en adelante, RNA). La predicción mediante Machine Learning se focaliza en el turismo exterior que recibe España regionalmente, obteniendo como modelo óptimo la RSV curtida con un núcleo de base radial gaussiana. Asimismo, concluye que estos sistemas mejoran la precisión predictiva frente a modelos lineales a mayor horizonte de predicción.

Los autores de este mismo estudio publicaron en el mismo año el artículo "Modelling cross-dependencies between Spain's regional tourism markets with an extension of the Gaussian process regression model" en la revista SERIEs. En este estudio se amplía el modelo de regresión del proceso gaussiano del artículo expuesto anteriormente. Este estudio refleja excelencia de este modelo de predicción y la notable mejora que proporciona en la precisión de los pronósticos.

Por otro lado, encontramos el artículo de Mishra et al. publicado en la revista International Journal of Advanced Computer Science and Applications en 2021, "Machine Learning based Forecasting Systems for Worldwide International Tourists Arrival". En este *artículo* se emplea sobre datos de turistas extraídos desde 2010 hasta 2020 la técnica de *Feature Selection* con el fin de identificar las variables más significativas. Con estas variables principales, el estudio antecedente busca predecir la llegada de turistas internacionales a nivel global con herramientas como el *Machine Learning* y técnicas de regresión. Las técnicas de regresión empleadas fueron RSV y *Random Forest Regression* (en adelante, RFR). Sin embargo, para este estudio emplearemos el Su*pport Vector Machine* (en adelante, SVM) por ser este modelo el que mejor se ajusta a la variable objetivo de este trabajo, el gasto diario.

Resulta relevante la nota de investigación de Perles Ribes et al. publicada en 2017 por la revista Tourism Economics, "Economic crises and market performance—A machine learning approach", en la cual se señala la relevancia de los eventos críticos y periodos de crisis en el desenvolvimiento normal del turismo en España, generando graves alteraciones que afectan a la posición en el mercado turístico de los destinos. Este estudio se centra en el periodo comprendido entre 1970 y 2013, por lo que comprende distintos momentos clave y momentos de crisis en la historia española (resaltamos como ejemplo relevantes, la Transición española y la crisis financiera del 2008), mientras que el presente estudio se centra en la crisis a nivel global que supuso la pandemia por el Covid-19 y que tuvo importante repercusión en España y, por consiguiente, en la demanda de turismo.

En contraposición con lo expuesto, si bien se encuentran diversa bibliografía relacionada con la temática de este estudio (el turismo) con aplicación de técnicas de *machine learning* y otras técnicas de análisis cuantitativo de la información, se encuentran escasos estudios que apliquen métodos como es el *Regression Discontinuity*. En este sentido, se destaca el artículo en relación con el turismo y la sostenibilidad de Dong et al. (2019), "Estimating the Impact of Air Pollution on Inbound Tourism in China: An Analysis Based on Regression Discontinuity Design". En este artículo se emplea la herramienta de regresión discontinua con el fin de estudiar el impacto de la contaminación atmosférica en el turismo de China. Es por ello por lo que resulta tan relevante el empleo de este método para este trabajo, ya que la escasez de referencias de ello en la investigación supone una oportunidad de innovación.

# 2. ESTADÍSTICA DESCRIPTIVA

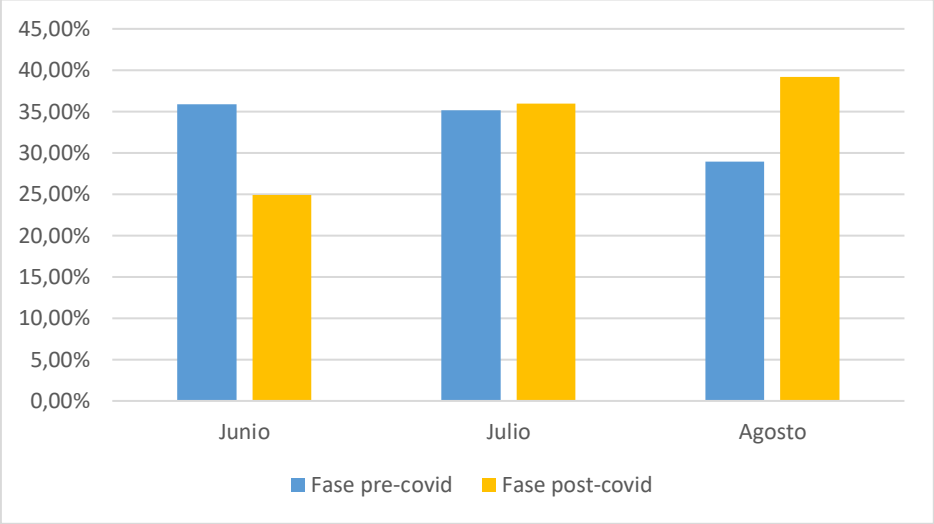
## 2.1. Descripción del conjunto de datos

A continuación, se introducen las variables empleadas en este estudio y las características básicas de las mismas. Como se introducía anteriormente, para este estudio, los datos se han dividido en dos fases: una fase anterior a la crisis sanitaria y una fase posterior.

### *2.1.1. Variable "mm_aaaa"*

La variable "mm_aaaa" que recoge el mes y año de referencia en el que se obtuvieron los datos de ese elemento. Se trata de una variable numérica discreta.

Gráfico 1. Diagrama de barras de los turistas encuestados según mes y fase.



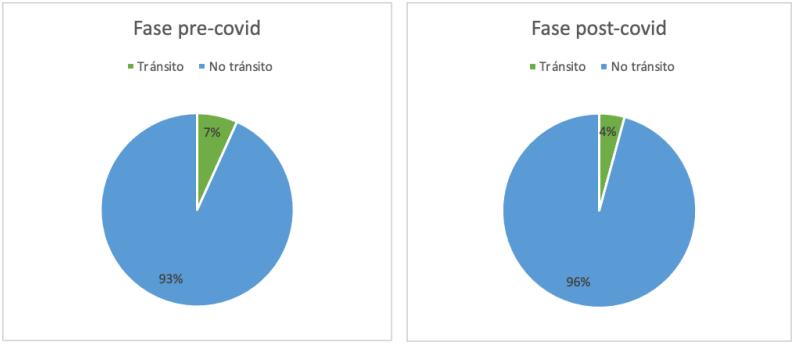

Elaboración propia. Fuente: INE

### *2.1.2. Variable "tipo"*

La variable "tipo" se trata de una variable categórica dicotómica. Esta variable informa sobre si el turista encuestado ha pasado por España con otro país de destino (tránsito) o si, por el contrario, viene específicamente a España (no tránsito).

Un ejemplo de turista en tránsito puede ser un residente en Portugal que pasa por España para llegar a Francia por carretera.

Gráfico 2. Diagrama de tarta de los turistas encuestados según si estaba en tránsito por España o no.




Elaboración propia. Fuente: INE.

### *2.1.3. Variable "vía"*

La variable "vía" recoge la vía de salida de los turistas. Se trata de una variable categórica de escala nominal la cual se concreta en cuatro modalidades: aeropuerto, carretera, puerto y tren.

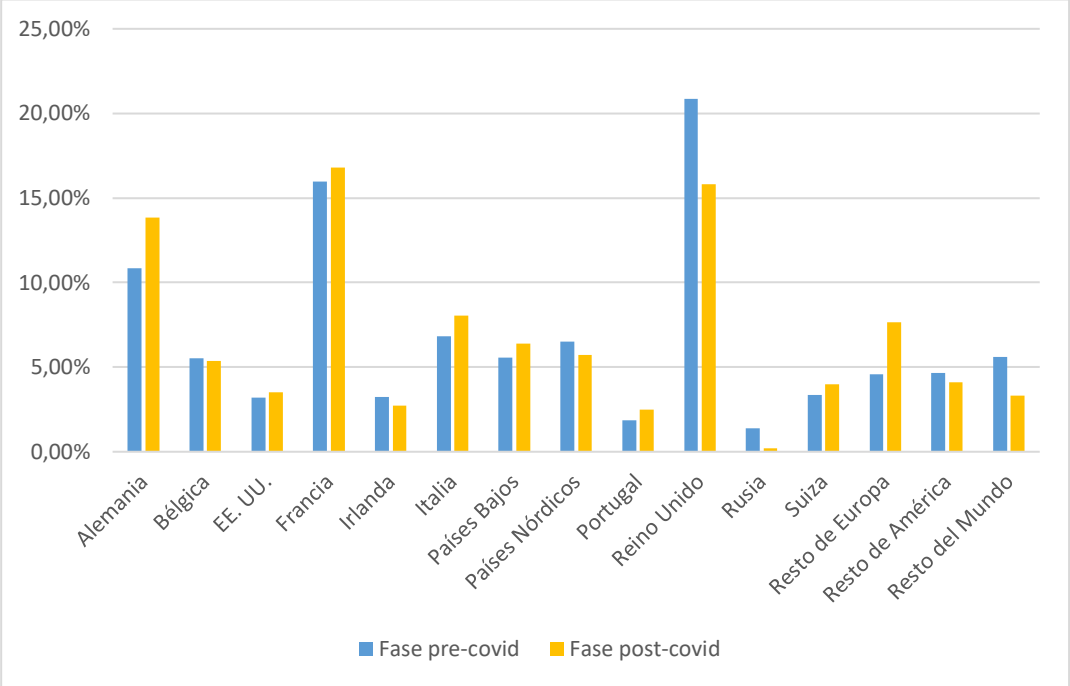
Gráfico 3. Diagrama de tarta de los turistas encuestados según vía de salida y fase.


Elaboración propia. Fuente: INE.

### *2.1.4. Variable "país"*

La variable "país" es una variable categórica nominal que recoge información sobre el país de residencia habitual del turista encuestado. Esta variable tiene quince modalidades distintas: Alemania, Bélgica, Francia, Irlanda, Italia, Países Bajos, Portugal, Reino Unido, Suiza, Rusia, Países Nórdicos, Resto de Europa, Estados Unidos (EE. UU.), Resto de América y Resto del Mundo.

Gráfico 4. Diagrama de barras de los turistas encuestados según país de residencia y fase.

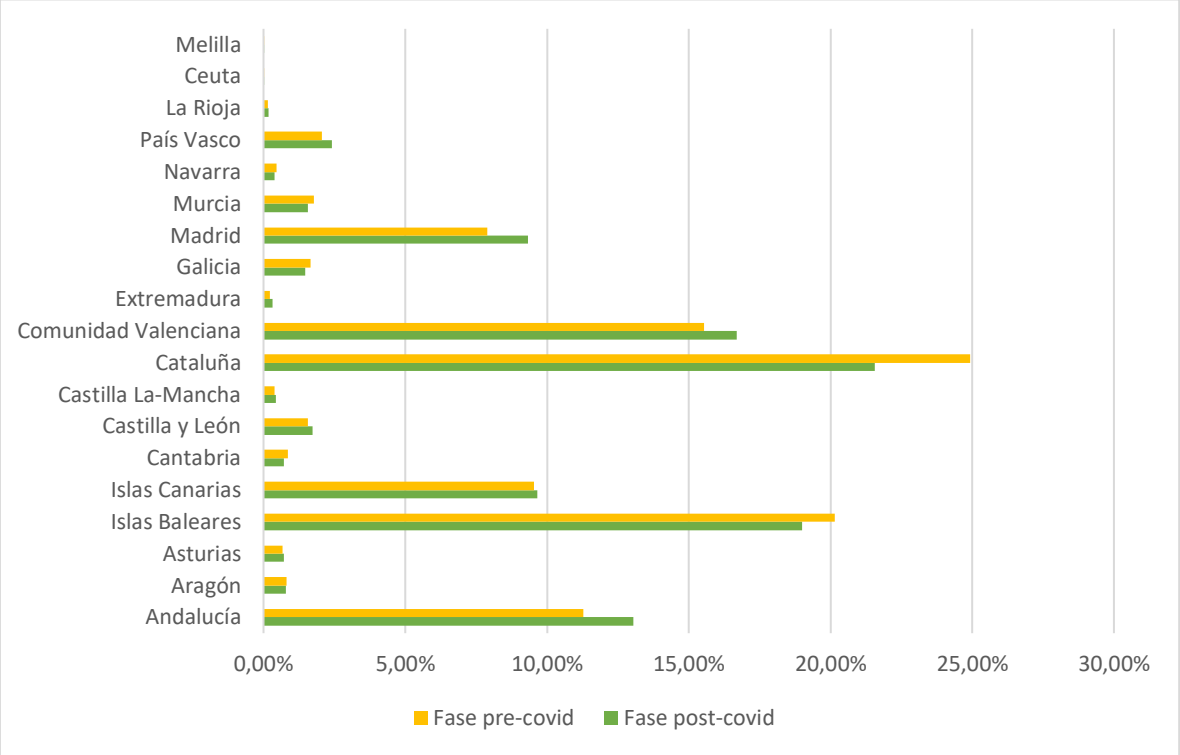



Elaboración propia. Fuente: INE.

### *2.1.5. Variable "CCAA"*

La variable "CCAA" recopila los datos relativos a la Comunidad Autónoma de destino principal del viaje del turista encuestado. Se trata de una variable categórica nominal con diecinueve modalidades distintas (una para cada Comunidad Autónoma, además de Ceuta y Melilla).

Gráfico 5. Diagrama de barras de los turistas encuestados según Comunidad Autónoma de destino y fase.




Elaboración propia. Fuente: INE.

### *2.1.6. Variable "totalpernocta"*

Esta variable recoge el número total de pernoctaciones realizadas por el turista encuestado en su viaje. Se trata de una variable cuantitativa la cual sigue una escala de razón.

Tabla 1. Estadísticos de las pernoctaciones totales de los turistas encuestados.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Media** | 9,35 | 10,69 |
| **Moda** | 7 | 7 |
| **Desviación típica** | 10,78 | 12,88 |
| **Varianza** | 116,17 | 165,89 |
| **Valor mínimo** | 1 | 1 |
| **Valor máximo** | 181 | 181 |

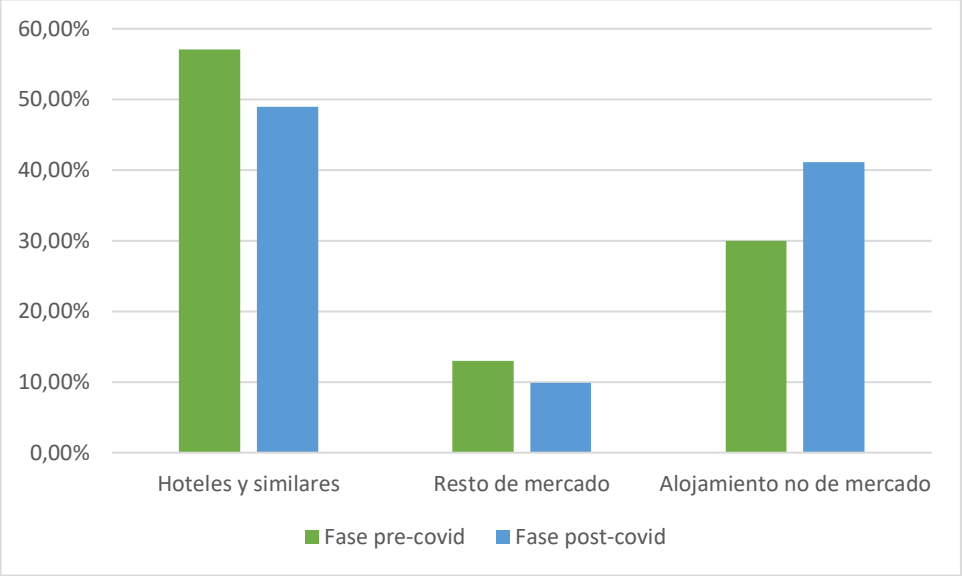
Elaboración propia. Fuente: INE.

### *2.1.7. Variable "alojamiento"*

La variable "alojamiento" recoge aquellos datos indicativos del tipo de alojamiento principal que ha escogido el turista encuestado para su estancia. Se trata de una variable categórica nominal que distingue tres modalidades: hoteles y similares, resto de mercado y alojamiento no de mercado.

La diferencia entre estas modalidades es la existencia de transacción monetaria o no. Así, para las dos primeras existirán transacciones monetarias al tratarse de alojamientos de pago como hoteles, pensiones, apartamentos turísticos, alquiler de habitaciones, entre otros. Sin embargo, se considera alojamiento no de mercado aquellos que no son de pago, como pueden ser, viviendas en propiedad, vivienda de familiares y amigos (si la estancia es gratuita), caravana estacionada fuera de un *camping*, entre otros.

Gráfico 6. Diagrama de barras de los turistas encuestados según tipo de alojamiento y fase.

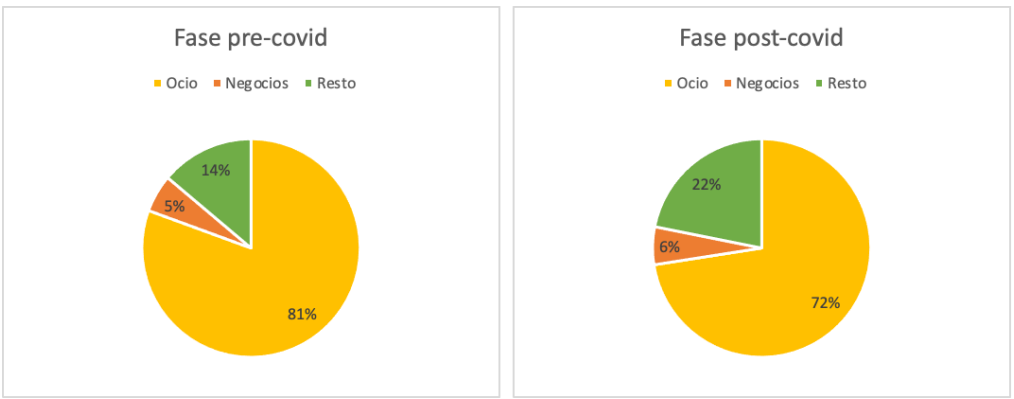



Elaboración propia. Fuente: INE.

### *2.1.8. Variable "motivo"*

La variable "motivo" se trata de una variable categórica nominal que recoge la información relativa al motivo principal del viaje del turista encuestado. Entre las tres categorías existentes se incluye: ocio, negocios y resto de motivos (aquellos que no han podido ser clasificados en las modalidades anteriores).

Gráfico 7. Diagrama de tarta según motivo del viaje y fase.

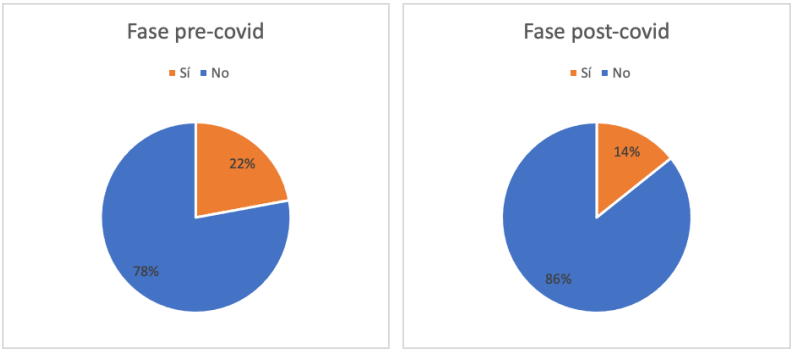



Elaboración propia. Fuente: INE.

### *2.1.9. Variable "paquete"*

La variable "paquete" se trata de una variable categórica dicotómica la cual recoge información relativa a la contratación o no de un paquete turístico. A efectos de esta encuesta, se entiende que se ha contratado uno cuando se ha realizado una reserva por medio de una agencia de una oferta que incluya alojamiento y transporte de manera conjunta.

Gráfico 8. Diagrama de tarta en función de la contratación o no de un paquete turístico por fase.




Elaboración propia. Fuente: INE.

### *2.1.10. Variable "gastodiario"*

Esta variable recoge el gasto diario en euros individual del turista encuestado, resulta de dividir el gasto total entre el número de pernoctaciones para cada turista encuestado. Se trata de una variable cuantitativa la cual sigue una escala de razón

Tabla 2. Estadísticos de los gastos diarios de los turistas encuestados.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Media** | 191,48 | 181,93 |
| **Moda** | 158,07 | 207,86 |
| **Desviación típica** | 170,12 | 159,21 |
| **Varianza** | 28.939,58 | 25.347,78 |
| **Valor mínimo** | 12,08 | 4,67 |
| **Primer cuartil (25%)** | 104,79 | 96,13 |
| **Segundo cuartil (50%)** | 155,70 | 154,81 |
| **Tercer Cuartil (75%)** | 215,17 | 213,20 |
| **Valor máximo** | 4.856,95 | 5.984,70 |

Elaboración propia. Fuente: INE.

## 2.2. Contraste de hipótesis

Con el fin de evitar la heterocedasticidad en las varianzas, resulta oportuno realizar la **prueba de Levene**. De acuerdo con, DATAtab (2023, "Prueba de Levene"), la prueba de Levene parte de la hipótesis nula de que las varianzas en las distintas muestras son las mismas, esto es, son homogéneas. En este caso, la hipótesis alternativa implica que los grupos tendrán varianzas diferentes.

> $H_0$ = la fase previa a la pandemia y la fase posterior a la pandemia tienen varianzas significativamente iguales.
>
> $H_1$ = la fase previa a la pandemia y la fase posterior a la pandemia tienen varianzas significativamente diferentes.

Para esta prueba, también tiene importancia el p-valor puesto que, si este es superior a 0,05, implica que las varianzas no son significativamente diferentes entre sí. En cambio, si es menor que 0,05, entonces sí existe una diferencia significativa que nos lleva a rechazar la hipótesis nula (*Ibid*.).

Para realizar la prueba de Levene se ha empleado el software de Python y SPSS. Los resultados obtenidos en Python tras realizar la prueba arrojan un p-valor muy pequeño (p-valor = 0,00000010575483712938135). Esto implica que existe una diferencia significativa entre las varianzas de los gastos de la fase anterior a la pandemia y la fase posterior a la pandemia y que, por lo tanto, se ha de rechazar la hipótesis nula.

En SPSS los resultados obtenidos se reflejan a continuación:

Tabla 3. Resumen de la prueba de Levene.

| | | Levene's Test for Equality of Variances | |
|---|---|---|---|
| | | F | Sig. |
| gastodiario | Equal variances assumed | 86.808 | <.001 |
| | Equal variances not assumed | | |

Elaboración propia.

Como muestran los resultados, en concordancia con los resultados obtenidos con Python, se debe rechazar la hipótesis nula.

Tabla 4. Resumen de la prueba t de igualdad de medias.

| | | t-test for Equality of Means | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Significance | | | | 95% Confidence Interval of the Difference | |
| | | t | df | One-Sided p | Two-Sided p | Mean Difference | Std. Error Difference | Lower | Upper |
| gastodiario | Equal variances assumed | 9.638 | 110576 | <.001 | <.001 | 9.54570053 | .990461699 | 7.60441063 | 11.4869904 |
| | Equal variances not assumed | 9.608 | 107818.159 | <.001 | <.001 | 9.54570053 | .993544677 | 7.59836750 | 11.4930336 |

Elaboración propia.

Los resultados mostrados para la prueba t muestran que sí existe una diferencia el gasto diario medio de los turistas en la fase anterior y posterior a la pandemia.

Dado que existe heterocedasticidad, no procedería realizar un contraste de hipótesis en relación con la variable objetivo (el gasto diario).

Por esto mismo, se realizará paralelamente la **prueba de Kruskal-Wallis** o prueba H. Esta prueba "es una prueba de hipótesis para muestras múltiples independientes, que se utiliza cuando no se cumplen los supuestos de un análisis de varianza de un factor" (DATAtab, 2023, "Prueba de Kruskal-Wallis"). De acuerdo con IBM (2021), esta prueba "es el análogo no paramétrico de análisis de varianza de un factor y detecta diferencias en la ubicación de distribución". En este sentido, esta prueba estudia las medianas de los grupos por lo que se plantea un nuevo contraste de hipótesis sobre las medianas:

> $H_0$ = la fase previa a la pandemia y la fase posterior a la pandemia tienen la misma tendencia central y, en consecuencia, proceden de la misma población.
>
> $H_1$ = la fase previa a la pandemia y la fase posterior a la pandemia no tienen la misma tendencia central, por lo que proceden de poblaciones diferentes.

La prueba H se ha realizado con el software SPSS, con el cual se ha obtenido los siguientes resultados:

Tabla 5. Resumen de la prueba Kruskal-Wallis.

| | |
|---|---|
| Total N | 110578 |
| Test Statistic | 63.428[a,b] |
| Degree Of Freedom | 1 |
| Asymptotic Sig.(2-sided test) | <.001 |

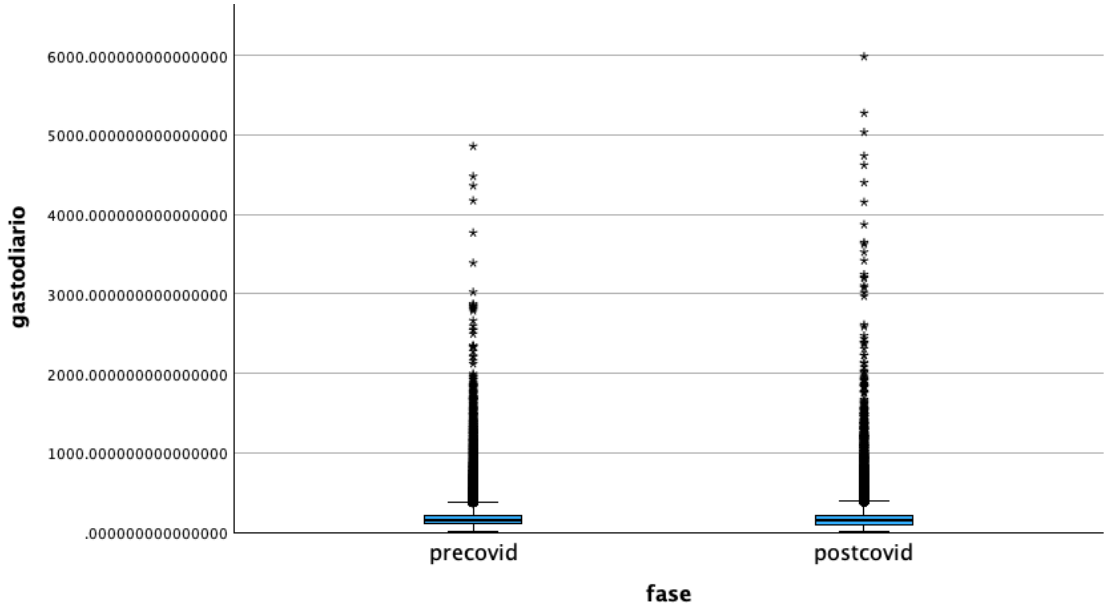
a. The test statistic is adjusted for ties.

b. Multiple comparisons are not performed because there are less than three test fields.

Elaboración propia.

Como se puede comprobar, el p-valor de la prueba H es inferior a 0,001. Esto implica que se debe rechazar la hipótesis nula, lo cual quiere decir que las distintas fases tienen distinta mediana. La mediana de la fase posterior a la pandemia es significativamente superior a la de la fase anterior, como puede apreciarse en los gráficos mostrados *infra*.

Gráfico 9. Diagrama de cajas y bigotes de la prueba Kruskal-Wallis.




Elaboración propia.

Los resultados obtenidos en estas pruebas serán ampliados y estudiados en mayor detalle a lo largo del presente trabajo mediante las técnicas de *machine learning* y *regression discontinuity*,

# 3. FEATURE SELECTION

De acuerdo en todo lo expuesto a continuación con Venkatesh y Anuradha (2019), Li et al. (2017) y Kumar y Minz (2014), el *feature selection* (en adelante, FS) es una estrategia de pre-procesamiento de datos la cual es de gran utilidad de cara a conocer cuáles son las variables más importantes de la base de datos del estudio, esto es, cuáles son las variables que mejor ayudan a predecir la variable objetivo. El objetivo de esta herramienta es poder llevar a cabo modelos más sencillos y así mejorar la predicción de estos.

El porqué de la utilidad de reducir el número de variables está en que, con un exceso de variables, los modelos tienden a hacer sobreajuste (en inglés, *overfitting*), esto es, que el modelo predictivo se ajusta perfecta o casi perfectamente a el conjunto de datos de entrenamiento, sin embargo, de cara a realizar predicciones este fenómeno debe evitarse ya que presentará un bajo porcentaje de acierto. En consecuencia, el FS permite mejorar el rendimiento del modelo de aprendizaje, así como la eficiencia de la predicción al permitirnos seleccionar únicamente los factores más relevantes entre los existentes.

Para llevar a cabo este análisis, pueden acudirse a técnicas de FS tradicionales o a aquellas más modernas. A modo de ejemplo, entre las técnicas tradicionales encontramos el chí cuadrado y el F-Test. Por otro lado, entre las técnicas más modernas de FS encontramos la técnica de *mutual information*. Para este estudio, emplearemos y contrastaremos la técnica de F-test y la de *mutual information* para conocer las variables que posteriormente emplearemos para crear un modelo de Machine Learning.

## 3.1. F-test

El método del *F-test* se emplea para estudiar el modelo que mejor se ajusta para predecir la variable objetivo.

La fórmula clásica para realizar este estudio se lleva a cabo mediante el contraste del estadístico "F" de significación conjunta. De acuerdo con De Ibarreta Zorita et al. (2019), mediante este contraste simultáneo, se puede conocer si un modelo es en su conjunto significativo o no. El estadístico de contraste que se emplea para ello es el siguiente:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Donde el numerador indica cuánto sobre la variable dependiente es explicado por el modelo y el denominador cuánto no se explica con el modelo.

A partir de esta base se desarrolla el modelo de *F-test* empleado para este estudio. Concretamente, se ha trabajado con un modelo de selección univariante en Python (método de "SelectKBest") por el cual se seleccionan las mejores variables de acuerdo con pruebas estadísticas univariantes. Con el cálculo de la covarianza entre cada variable y su conversión a un valor-F para, posteriormente, la obtención de un p-valor, se eliminan todas las variables excepto el número seleccionado de ellas (en este caso, cuatro) con mayor puntuación. Con ello se identifican aquellas variables relevantes para el modelo (Scikit-learn, s.f.).

Se realizará este análisis para ambas fases de estudiadas (tanto la fase pre-covid como la fase post-covid) para comprobar si resulta oportuno unificar las fases en un único modelo o si, por el contrario, resulta necesario desarrollar dos modelos distintos. Esto último sucedería en el caso de que del *feature selection* de cada una de las fases resulte que las mejores variables no son las mismas para ambas fases.

#### *3.1.1. Fase pre-covid*

Considerando el gasto diario la variable objetivo en este análisis, los resultados obtenidos como mejores variables en la fase anterior al covid, de acuerdo con el *F-test* llevado a cabo, son: la variable concerniente a la vía de salida de los turistas, la variable referente al país de residencia habitual, la variable relativa al número total de pernoctaciones y la variable indicativa del tipo de alojamiento.

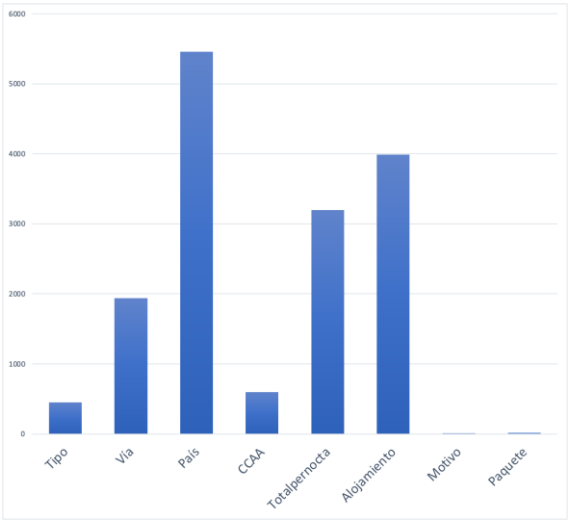
A continuación, se ha cuantificado la puntuación de las variables para el modelo, eso es, su nivel de relevancia. El resultado de esta cuantificación es el siguiente:

Tabla 6. Puntuación de las variables de la fase pre-covid para el *F-test*.

| Variable | Puntuación |
|---|---|
| Tipo | 450,83 |
| Vía | 1939,51 |
| País | 5459,87 |
| CCAA | 598,75 |
| Totalpernocta | 3196,54 |
| Alojamiento | 3986,72 |
| Motivo | 8,20 |
| Paquete | 17,57 |

Elaboración propia. Fuente: INE.

Gráfico 10. Histograma de puntuaciones de las variables de la fase pre-covid para el *F-test*.




Elaboración propia. Fuente: INE.

Por lo tanto, las cuatro variables más relevantes para el modelo en el caso de la fase pre-covid, por orden de relevancia serían: en primer lugar, la variable "país"; en segundo lugar, la variable "alojamiento"; después, la variable "totalpernocta"; y, por último, la variable "vía".

Si bien el siguiente paso de este estudio es contrastar estos resultados con aquellos obtenidos mediante el método de *Mutual Information*, en primera instancia estas cuatro variables son las que se emplearían para elaborar el modelo de *Machine Learning* para la fase pre-covid, descartando las demás variables que no han resultado ser relevantes.

### *3.1.2. Fase post-covid*

Se mantiene la consideración a lo largo de todo el estudio del gasto diario como la variable objetivo. De esta manera, los resultados obtenidos como mejores variables en la fase posterior al covid, de acuerdo con el *F-test* llevado a cabo, son: la variable concerniente al tipo de visita de los turistas, la variable relativa al país de residencia, la variable concerniente al número total de pernoctaciones y la variable indicativa del tipo de alojamiento.

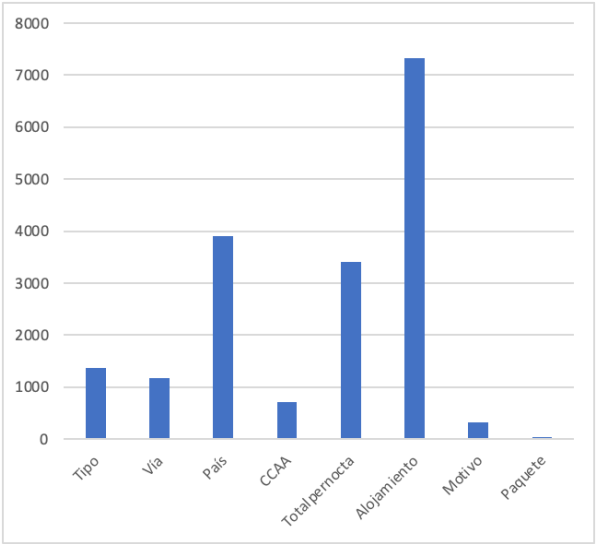
A continuación, se ha cuantificado la puntuación de las variables para el modelo, eso es, su nivel de relevancia. El resultado de esta cuantificación es el siguiente:

Tabla 7. Puntuación de las variables de la fase post-covid para el *F-test*.

| Variable | Puntuación |
|---|---|
| Tipo | 1369,06 |
| Vía | 1166,93 |
| País | 3911,30 |
| CCAA | 722,16 |
| Totalpernocta | 3404,31 |
| Alojamiento | 7332,24 |
| Motivo | 326,92 |
| Paquete | 38,035 |

Elaboración propia. Fuente: INE.

Gráfico 11. Histograma de puntuaciones de las variables de la fase post-covid para el *F-test*.




Elaboración propia. Fuente: INE.

Por ello, las cuatro variables más relevantes para el modelo en el caso de la fase post-covid, por orden de relevancia serían: en primer lugar, la variable "alojamiento"; en segundo lugar, la variable "país"; después, la variable "totalpernocta"; y, por último, la variable "tipo".

Observamos, en primera instancia, que las variables relevantes en el *F-test* no solo han cambiado respecto de aquellas resultantes en la fase pre-covid, sino que también se ha alterado su orden de importancia, su peso para el modelo.

Asimismo, como se indicaba también para el estudio de la fase pre-covid, se ha de contrastar los resultados obtenidos en esta cuantificación con aquellos resultantes del método de *Mutual Information*, de cara a elaborar el modelo de *Machine Learning* para la fase post-covid, descartando las demás variables que no han resultado ser relevantes.

### 3.2. Mutual information

El concepto de *Mutual Information* (en adelante, MI) fue introducido por Claude Shannon en 1948 en relación con la teoría de la información clásica (Cardona y Velásquez, 2006). La teoría de la información es una rama de la informática y de las matemáticas que trata

el estudio de la transmisión, el procesamiento y la medición de los datos y de la información (Peiró, 2021). De acuerdo con Zhang (2020), la MI es uno de los bloques esenciales de la teoría de la información.

Uno de los elementos principales de esta teoría es la entropía. Para esta rama del conocimiento, la entropía se entiende como una medida para cuantificar la cantidad de incertidumbre presente en un conjunto de datos concreto (Cover y Thomas, 2005).

De acuerdo con Williams y Li (2009), encontramos tres tipos de entropía: la entropía conjunta, la condicionada y la marginal. El promotor del concepto general de la entropía de la información, Shannon, en su publicación "A Mathematical Theory of Communication" desarrolla para explicar estos tres tipos de entropía la denominada "función H":

$$H = -K \sum_{i=1}^{k} p_i \log p_i$$

A partir de esta fórmula, y considerando los tres tipos de entropía mencionados, se desarrolla la Regla de la cadena para la entropía, la cual se formula de la siguiente forma:

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

Aquí es donde la MI tiene un importante papel, ya que facilita la reducción de la incertidumbre al examinar la entropía que se pierde de una variable cuando se conoce la otra. Se trata de una medida de la información que se comparte entre dos variables, así como de la dependencia entre ellas (cfr. MacKay, 2003). Es una medida no paramétrica de la relevancia de las variables que mide tanto la correlación lineal como la no lineal (Williams y Li, 2009). De esta forma, si hay una alta correlación entre dos variables, la información que proporciona una de las variables puede ser empleada para la mejor predicción de los valores de la otra variable.

Así mismo, esta medida guarda relación con la teoría de la probabilidad puesto que la MI se basa en el concepto de la probabilidad de relevancia de las variables (Cardona y Velásquez, 2006), cuantificando la relación entre las dos variables. La MI puede emplearse en la medición de la discrepancia cuando se trata de probabilidades condicionadas o de probabilidades conjuntas (Cover y Thomas, 2005).

Esta teoría fundamental respalda el presente estudio y los resultados obtenidos en este. Así, conociendo el marco teórico se ha procedido a medir cuál es la MI de las variables

de este estudio, de acuerdo con la variable objetivo, para así realizar la selección de aquellas más relevantes.

### *3.2.1. Fase pre-covid*

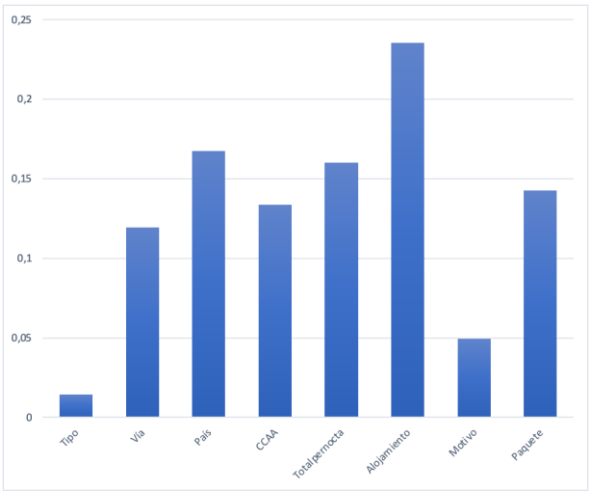
A continuación, se muestran los valores obtenidos tras estudiar la MI de las variables para la fase pre-covid del modelo:

Tabla 8. Valores obtenidos con *mutual information* para la fase pre-covid.

| Variable | Puntuación |
|---|---|
| Tipo | 0,014628 |
| Vía | 0,119317 |
| País | 0,16747 |
| CCAA | 0,13364 |
| Totalpernocta | 0,160131 |
| Alojamiento | 0,235369 |
| Motivo | 0,049395 |
| Paquete | 0,142594 |

Elaboración propia. Fuente: INE.

Gráfico 12. Histograma de los valores obtenidos con *mutual information* para la fase pre-covid.




Elaboración propia. Fuente: INE.

De acuerdo con el estudio de las variables más relevantes para el modelo con la técnica de MI, observamos que, por orden de relevancia, estas serían: en primer lugar, la variable "alojamiento"; en segundo lugar, la variable "país"; después, la variable "totalpernocta"; y, por último, la variable "paquete".

Comprobamos que, si bien las tres primeras variables más relevantes en MI son también las tres primeras variables más relevantes para el F-test (aunque con orden de relevancia distintos), la cuarta variable más relevante no coincide para ambas técnicas. Es por ello por lo que se ha de priorizar una técnica frente a otra; en este caso, se prioriza el resultado obtenido mediante MI con motivo de lo expuesto a la hora de teorizar esta técnica, esto es, que la MI atiende todo tipo de relaciones entre las variables del modelo, inclusive aquellas relaciones no lineales e inversas.

Por lo tanto, las cuatro variables a emplear para elaborar el modelo de *Machine Learning* son: “alojamiento”, “país”, “totalpernocta” y “paquete”. En cuanto al resto de variables, como se adelantaba en el apartado relativo al F-test, estas serán descartadas.

### *3.2.2. Fase post-covid*

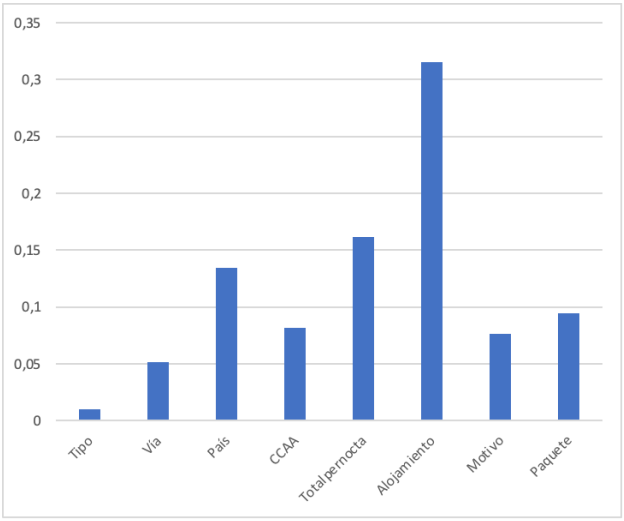
A continuación, se muestran los valores obtenidos tras estudiar la MI de las variables para la fase post-covid del modelo:

Tabla 9. Valores obtenidos con *mutual information* para la fase post-covid.

| Variable | Puntuación |
|---|---|
| Tipo | 0,009999 |
| Vía | 0,051217 |
| País | 0,134195 |
| CCAA | 0,081946 |
| Totalpernocta | 0,161272 |
| Alojamiento | 0,315178 |
| Motivo | 0,076465 |
| Paquete | 0,094774 |

Elaboración propia. Fuente: INE.

Gráfico 13. Histograma de los valores obtenidos con *mutual information* para la fase post-covid.




Elaboración propia. Fuente: INE.

De acuerdo con el estudio de las variables más relevantes para el modelo con la técnica de MI, observamos que estas coinciden con las resultantes del análisis la fase pre-covid si bien con el orden de peso de las variables para el modelo diferente. Así, por orden de relevancia, las variables más importantes serían: la variable “alojamiento”, la variable “totalpernocta”, la variable “país”, y la variable “paquete”.

Se observan también para la fase post-covid una variación en las variables resultantes en ambos métodos de *feature selection*. Pese a ello, como se exponía con anterioridad al seleccionar las variables de la fase pre-covid, la técnica de MI prima sobre el *F-test*. Por ende, las variables seleccionadas serán las expuestas en el párrafo anterior, las resultantes de la técnica de MI. El resto de las variables no se tendrán en consideración para el modelo.

### 3.3. Conclusiones del *Feature Selection*

En virtud de lo expuesto en los apartados anteriores, se ha priorizado el método de MI frente al F-test para la selección de las variables relevantes para el modelo.

Teniendo esto en cuenta, variables resultantes para ambas fases, si bien en distinto orden de importancia, resultan coincidir. En este sentido, las variables en cuestión son: "alojamiento", "totalpernocta", "país", y "paquete". Por lo tanto, estas serán las variables empleadas para la creación del modelo único.

Se elaborará un modelo único para ambas fases, unificando las mismas, para una mayor eficiencia del estudio debido a que para ambas fases las variables relevantes resultan ser las mismas.

Por otro lado, también resulta interesante considerar el motivo de la relevancia de dichas variables sobre la variable objetivo. En primer lugar, es lógico considerar que el tipo alojamiento elegido para un viaje adquiere una gran relevancia en el gasto diario de dicho viaje, dado que los costes de dicho alojamiento varían en función de la categoría de alojamiento seleccionado (el coste de alojarse en un hotel no va a ser el mismo que el coste de alojarse en un alojamiento no de mercado como puede ser una vivienda en propiedad o en la vivienda de familiares o amigos del encuestado).

Por otra parte, el país de residencia habitual del encuestado es relevante en el gasto diario debido al poder adquisitivo del mismo y, en consecuencia, los distintos presupuestos de gasto que pueda tener en función del país. Probablemente, aquellos países con salarios medios más altos gastarán más.

## 4. MODELO *MACHINE LEARNING*

Una vez se conocen las variables más relevantes para explicar la variable objetivo a raíz del *feature selection* elaborado en el apartado anterior, se elaborará un modelo de *machine learning* (en adelante, "ML"). Como se adelantaba en el apartado anterior, dado que las variables más relevantes han resultado ser coincidentes para ambas fases, se ha unificado los datos de ambas fases de cara al desarrollo de esta fase. Por ello, la base de datos resultante para esta fase está formado por los datos de todas las fases para las variables "alojamiento", "totalpernocta", "país", y "paquete", así como la variable objetivo "gasto diario".

De acuerdo con Zhou (2021), *machine learning* es la técnica por la cual se mejora el rendimiento de un sistema por medio del aprendizaje experiencial mediante métodos informáticos. Esta experiencia, para el caso de los sistemas informáticos, se obtiene en forma de datos.

En este estudio, el modelo de ML que se desarrollará es un modelo de clasificación. Los modelos ML de clasificación encuadran dentro de los modelos de aprendizaje supervisado. De acuerdo con Kotsiantis (2007), el aprendizaje supervisado se da cuando los datos empleados en ML se presentan con una serie de características o variables conocidas, esto es, con una denominación concreta. Con estos datos, el aprendizaje se lleva a cabo mediante el reconocimiento de unas reglas o ejemplos en los datos de entrenamiento, creando una clasificación o modelo predictivo que pueda ser empleado de manera general para nuevos casos.

Por lo tanto, mediante ML se busca calcular u obtener la función que mejor se ajusta a las variables seleccionadas como más relevantes, sin que ello conlleve *overfitting*.

La variable objetivo en este estudio es el gasto diario de los turistas, variable la cual, en primera instancia, se trata de una variable continua. Sin embargo, con vistas a la simplificación del modelo y a un mejor ajuste, dicha variable continua se transformará a una discreta a efectos de este apartado. Para ello, se ha tomado los datos de la variable objetivo y se ha dividido en tres categorías distintas para distintos niveles de gastos: bajo (0), medio (1) y alto (2).

Para realizar la división de la manera más homogénea posible, se ha calculado el percentil 33 y el percentil 66 del conjunto de datos, con el fin de dividir la muestra en estos tres grupos. En este sentido, el percentil 33 (en adelante, P33) corresponde con un nivel de

gasto diario de 119,49 euros y el percentil 66 (en adelante, P66) con un nivel de gasto diario de 188,54 euros aproximadamente. Por lo tanto, los tres grupos se formarían de la siguiente forma: el gasto diario "bajo" compuesto por aquellos turistas con un gasto diario inferior a la cifra del P33; el gasto diario "medio", formado por aquellos turistas con un gasto diario que se encuentra entre el P33 y el P66; y, el grupo del gasto diario "alto", representado por aquellos turistas cuyo gasto diario supera la cifra del P66.

Como resultado de esta división, el grupo del gasto diario "bajo" cuenta con 36.492 observaciones, el grupo del gasto diario "medio" con 36.490 y el grupo del gasto diario "alto" con 37.596 observaciones.

Dado que la variable objetivo es ahora discreta, la función en cuestión se deberá buscar mediante el método de ML de clasificación que se introducía en los párrafos anteriores.

El mejor modelo que se ajuste a las variables para predecir el nivel de gasto va a ser proporcionado por el software de Matlab, con la aplicación "Classification Learner". Esta herramienta trata de encontrar el menor error en el modelo al mismo tiempo que busca el mejor ajuste si bien evitando el *overfitting*. Para medir la precisión de un modelo de clasificación se emplean dos herramientas: la matriz de confusión y la curva ROC. En este sentido, el software empleado en este estudio emplea la matriz de confusión de cara a evaluar la precisión del modelo e indicar cuál es el mejor modelo.

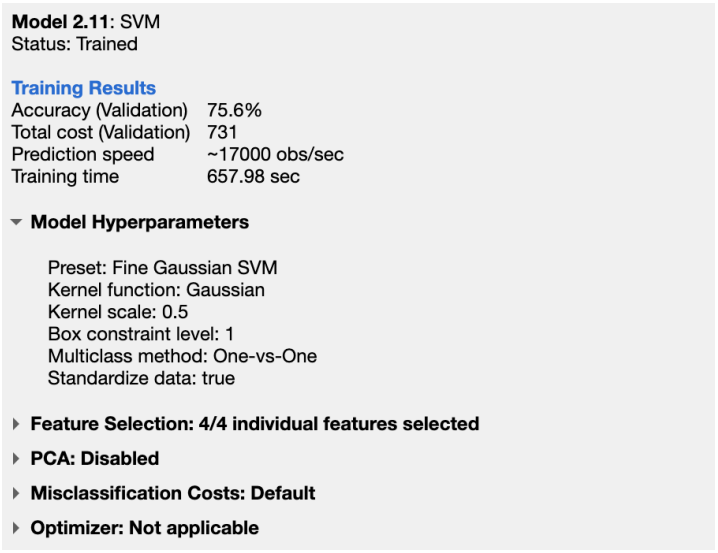
Para evitar el *overfitting*, se realizará un *cross-validation* de 10 pliegues. De acuerdo con la herramienta Matlab, este sistema de validación protege frente al sobreajuste con el estudio de la precisión de cada una de las divisiones o pliegues. De acuerdo en todo ello con Scikit-learn (s.f., "Cross-validation: evaluating estimator performance"), se emplea este método en ML debido a que, si se entrena el modelo con los datos empleados para desarrollar la función y se pone el modelo a prueba con esos mismos datos, se estaría cometiendo un error metodológico. Por ello, la práctica común es separar una parte de los datos del conjunto para empelarla como conjunto de prueba y otra parte para emplearla como conjunto de validación, resultando así en tres divisiones del conjunto de datos. Sin embargo, esto puede resultar en una escasa cantidad de observaciones disponibles para emplear en el modelo de aprendizaje. Es aquí donde interesa este sistema de *cross-validation*, con el cual ya no se requerirá el conjunto de validación.

A continuación, se ha realizado el entrenamiento del modelo con todos los tipos de métodos disponibles en el software de cara a encontrar aquel método con el menor error.

Entre los métodos entrenados encontramos: árboles de decisión, SVM, RNA, ensemble, entre otros. De cara a encontrar el método más preciso de la forma más eficiente posible, se ha seleccionado de forma aleatoria una muestra significativa de 4.000 datos empleando la función "random" de Matlab. De todos estos datos, se ha reservado 1.000 para realizar un estudio posterior con el mejor modelo resultante, de cara a probar su calidad de ajuste. El mejor modelo ML de clasificación será elegido mediante optimización bayesiana por ser este el mejor enfoque para la optimización en ML de entre un total de 31 modelos ML.

Como resultado de este análisis, se deduce que el modelo con menor error de entre todos los estudiados es el SVM (*Support Vector Machine*) y, más concretamente, el *fine gaussian* SVM. En cuanto a los resultados de entrenamiento de este modelo, este cuenta con una precisión de validación (*validation accuracy*) del 75,6% y una velocidad de predicción de aproximadamente 17.000 observaciones por segundo. No se ha empleado PCA. En mayor detalle, los hiperparámetros de este modelo son los siguientes:

Imagen 1. Resumen del modelo *fine gaussian* SVM.



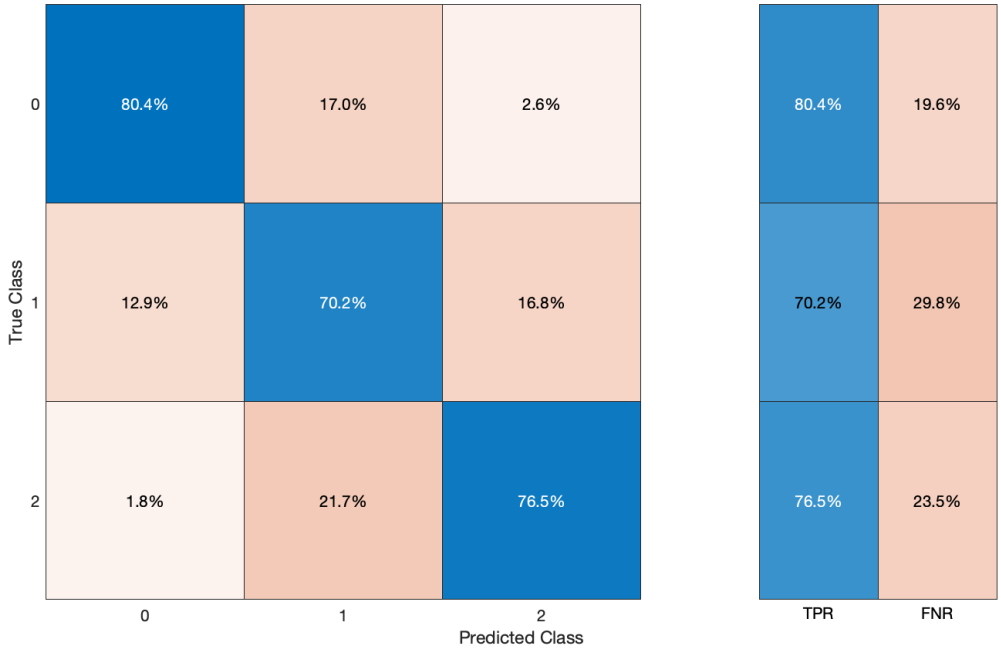

Elaboración propia.

De acuerdo con Nyuytiymbiy (2020), los hiperparámetros en ML parámetros cuyos valores controlan el proceso de aprendizaje y establecen los valores de los parámetros del modelo que un algoritmo de aprendizaje acaba aprendiendo.

### 4.1. Matriz de confusión

La matriz de confusión se emplea para evaluar la precisión de la clasificación realizada por el modelo. Así, de acuerdo con Blanco (2020), la matriz de confusión mide tanto la cantidad de registros que ha identificado correcta e incorrectamente como el número de aciertos y fallos y en qué sentido lo han sido. Esta matriz arroja cuatro posibles combinaciones: verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. En este caso, Matlab emplea esta herramienta de cara a evaluar la precisión del modelo e indicar cuál es el mejor modelo.

Gráfico 14. Matriz de confusión TPR-FNR del modelo *fine gaussian* SVM.




Elaboración propia.

La matriz de confusión para el modelo *fine gaussian* SVM mostrada en el gráfico 14 refleja la relación entre la tasa de verdaderos positivos (*true positive rate* – TPR) y la tasa de falsos negativos (*false negative rate* – FNR). En este sentido, la TPR muestra la proporción de observaciones clasificadas correctamente como "positivas" aquellas que efectivamente son "positivas" mientras que, la FNR, muestra la proporción de observaciones clasificadas incorrectamente como "negativas" cuando eran "positivas" (Mathworks, s.f., "Visualize and Assess Classifier Performance in Classification Learner").

Para el presente caso, observamos que el 80,4% de los datos para el gasto diario bajo (0) fueron correctamente clasificados en este tipo, el 70,2% de los datos para el gasto diario medio (1) también fueron correctamente clasificados por el modelo dentro de su correspondiente tipo y, por último, el 76,5% de los datos para el gasto diario alto (2) fueron correctamente clasificados en su tipo.

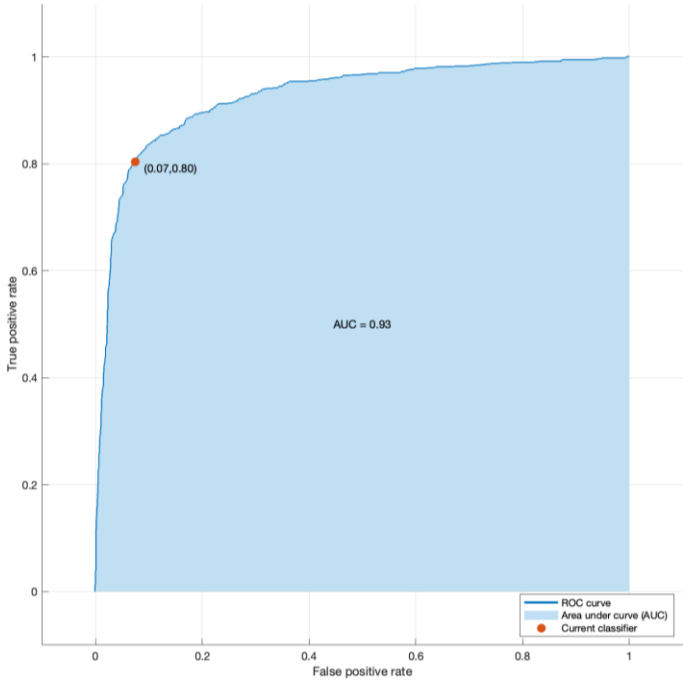
Sin embargo, para el tipo de gasto diario bajo (0), fue clasificado incorrectamente un 17% como gasto diario medio y un 2,6% como gasto diario alto, lo cual supone una clasificación incorrecta total de 19,6%. Asimismo, para el gasto diario medio (1), fue clasificado de manera incorrecta un 12,9% como gasto diario bajo y un 16,8% como gasto diario alto, lo que hace un total de clasificación incorrecta del 29,8%. En última instancia, para el gasto diario alto (2), tuvo lugar una clasificación incorrecta del 1,8% de los datos como gasto bajo y un 21,7% como gasto medio, lo que supone un 23,5% de clasificación incorrecta total.

### 4.2. La Curva ROC

Resulta relevante considerar la curva ROC para este modelo. Siguiendo a Sigrist (2022), la curva ROC representa la tasa de verdaderos positivos frente a la tasa de falsos positivos para todos los umbrales posibles. Así, cuanto menor sea el umbral, mayor será la tasa de verdaderos positivos, así como la de falsos positivos. A efectos de determinar la precisión del modelo, resulta relevante el área bajo la curva ROC (conocida como AUC o AUROC) dado que, a mayor área, mejor resultará ser el modelo. Esta área toma valores entre 0 y 1. La AUC tiene la interpretación de qué probabilidad hay de que dos puntos de datos de prueba elegidos al azar sean clasificados correctamente por el sistema clasificador.

A continuación, se muestra la curva ROC para cada una de las variables del gasto diario.

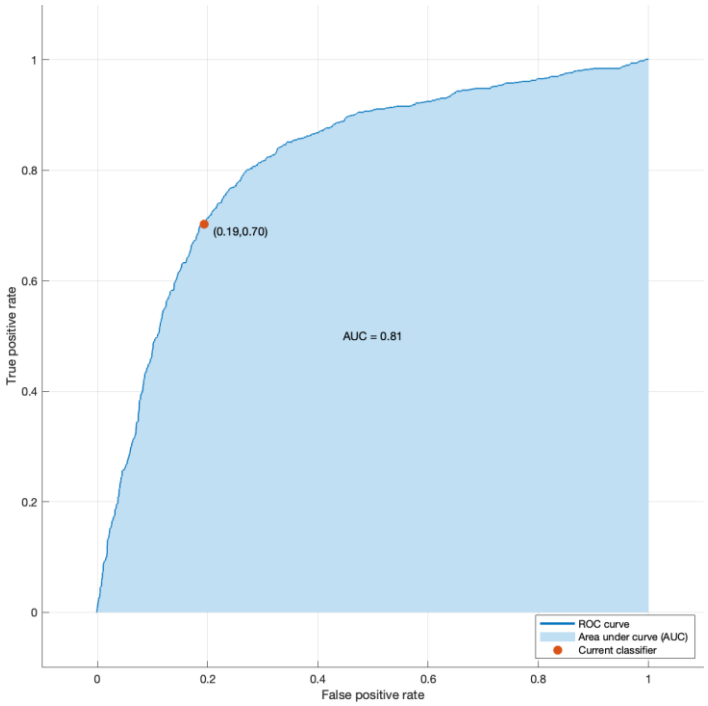
Gráfico 15. Curva ROC para la variable gasto diario en su categoría “bajo”.




Elaboración propia.

Centrando el análisis en el área bajo la curva ROC, para un gasto diario bajo esta es de 0,93. Por lo tanto, podemos concluir que la calidad del clasificador es alta puesto que, a mayor AUC, mejor es el clasificador.

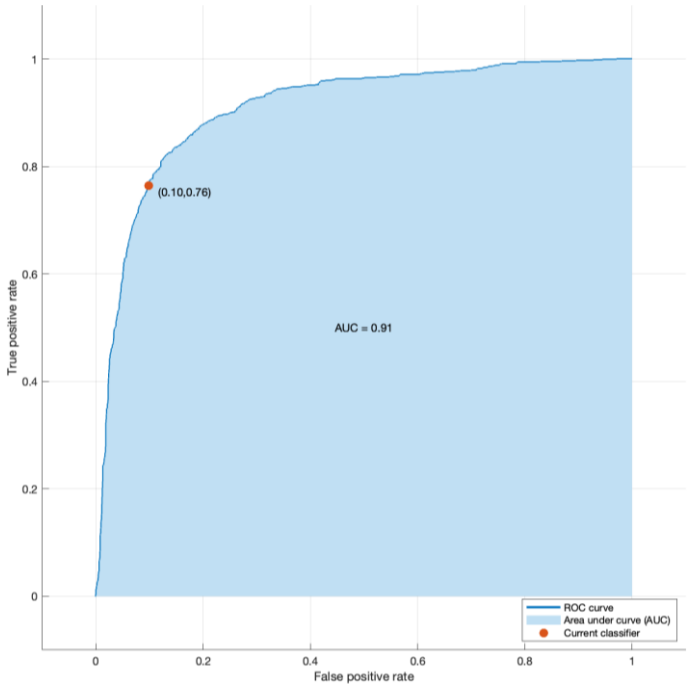
Gráfico 16. Curva ROC para la variable gasto diario en su categoría “medio”.




Elaboración propia.

Para el gasto diario medio, la AUC es de 0,81, lo cual sigue indicando una buena calidad del clasificador, si bien para el gasto diario bajo era mejor.

Gráfico 17. Curva ROC para la variable gasto diario en su categoría "alto".




Elaboración propia.

Para un gasto diario alto, la AUC es de 0,91. Este área es alta, lo que indica que el clasificador es de buena calidad para esta clase.

### 4.3. Ajuste del modelo

Una vez se ha seleccionado el mejor modelo ML y se ha estudiado las características de este, resulta oportuno poner a prueba dicho modelo para ver su calidad de ajuste.

En el subapartado anterior, se seleccionó el mejor modelo ML, el cual resultó ser un modelo SVM. Este modelo genera la función de clasificación que ha de emplearse para predecir, es un modelo entrenado fruto de la aplicación de Matlab de "Classification learner".

De cara a predecir los valores del gasto diario sobre un conjunto de datos, también se ha empleado el software de Matlab. Para usar la función del modelo, se introduce en el espacio de trabajo el archivo con los datos (en este archivo se ha eliminado la información de la variable "gasto diario", esta es, la variable objetivo a predecir, si bien se ha guardado otra copia del archivo en el que no se ha eliminado esta información, es decir, los datos

reales, de cara a compararlos posteriormente). A continuación, se ejecuta la función para que realice la clasificación. En último lugar, se compararán los resultados obtenidos de la función con los datos reales de la variable.

### *4.3.1. Ajuste en un subconjunto general de datos*

Como se introducía anteriormente, se ha empleado una muestra de 1.000 datos seleccionada de manera aleatoria con la función "random" de Matlab. Con ello, se ha empleado la función del modelo para predecir el gasto diario.

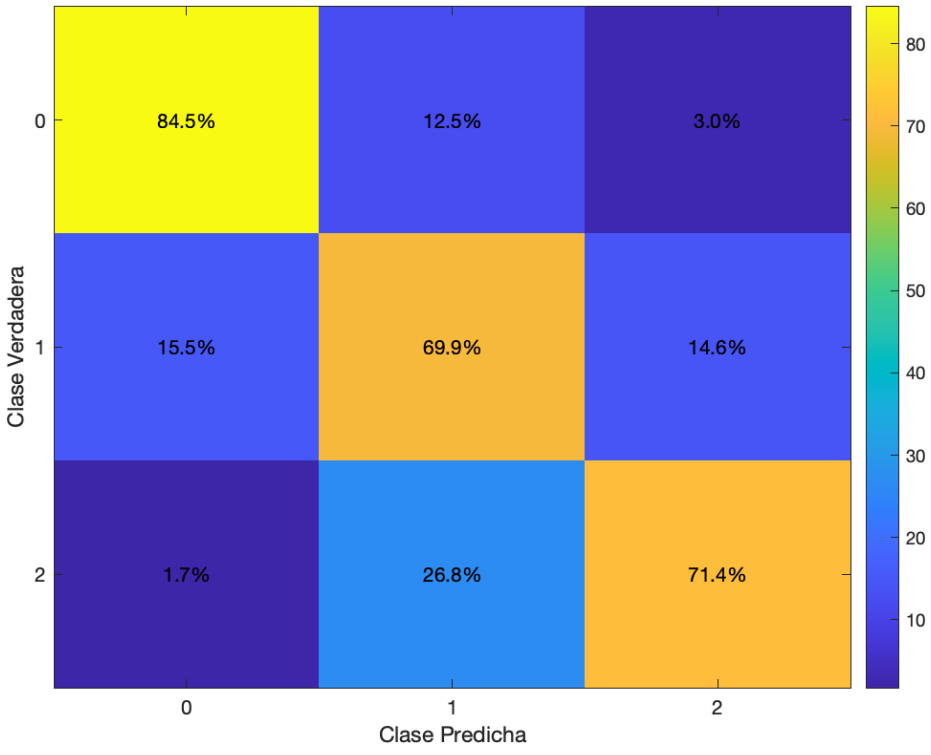
Los resultados son favorables, ya que la función del modelo ha clasificado correctamente el 75,30% de los datos y ha clasificado incorrectamente el 24,70% de los datos.

Tabla 10. Resultados de la clasificación en el subconjunto general de datos.

| Datos clasificados correctamente | Datos clasificados incorrectamente |
|---|---|
| 75,30% | 24,70% |

Elaboración propia.

Gráfico 18. Matriz de confusión para el subconjunto general.




Elaboración propia.

Se muestra la matriz de confusión para el subconjunto general de clasificación. En ella se muestran los resultados para las diferentes clases de gasto. En este sentido, la matriz refleja que se ha clasificado correctamente como gasto "bajo" (0) el 84,50%, si bien incorrectamente un 12,50% como gasto "medio" y un 3% como gasto "alto".

Respecto al gasto "medio" (1), se ha clasificado correctamente como tal gasto un 69,90% e incorrectamente un 15,50% como gasto "bajo" y un 14,60% como gasto "alto".

Por último, en cuanto al gasto "alto" (2), se ha clasificado correctamente un 71,40% de los datos. Sin embargo, se ha clasificado erróneamente un 1,70% como gasto "bajo" y un 26,80% como gasto "medio".

### *4.3.2. Ajuste en un subconjunto concreto de datos*

Por otro lado, de la muestra seleccionada de 4.000 datos se han filtrado y copiado en un archivo separado aquellos datos relacionados con un subconjunto específico, de cara a ver el comportamiento de nuestro modelo a la hora de predecir sobre este.

Concretamente, el subconjunto seleccionado ha sido filtrado en función de la variable "país", seleccionándose aquellos turistas cuyo país de residencia habitual es el Reino Unido.

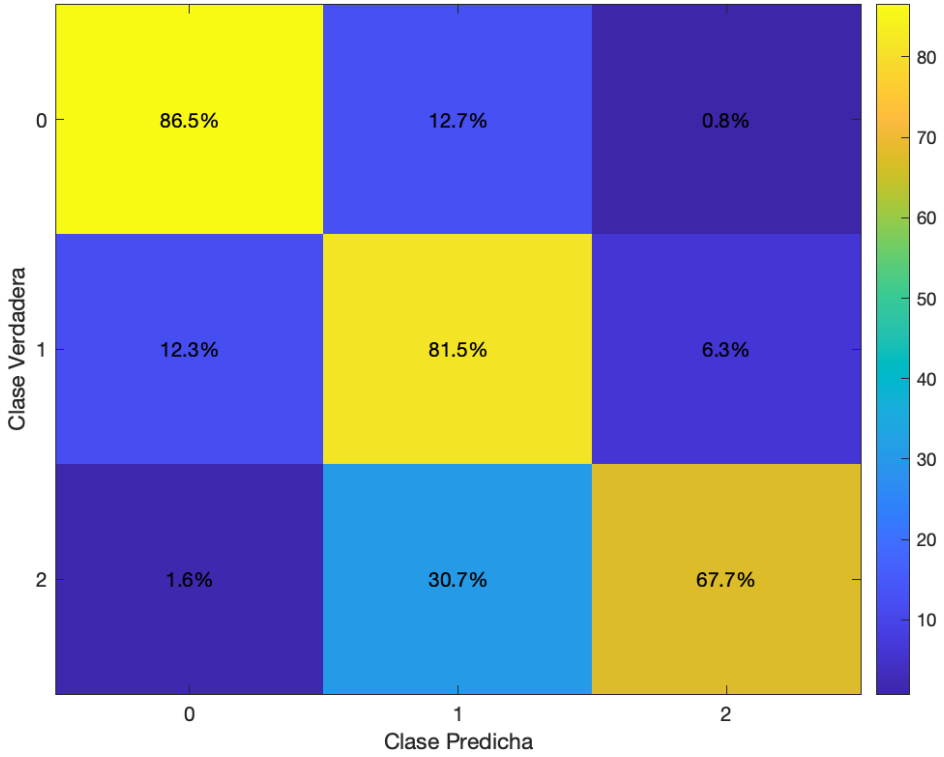
Los resultados de la función de predicción son positivos ya que ha clasificado correctamente el 79,73% de los datos y ha clasificado incorrectamente el 20,27% de ellos.

Tabla 11. Resultados de la clasificación en el subconjunto del Reino Unido.

| Datos clasificados correctamente | Datos clasificados incorrectamente |
|---|---|
| 79,73% | 20,27% |

Elaboración propia.

Gráfico 19. Matriz de confusión para el subconjunto del Reino Unido.




Elaboración propia.

Por lo que respecta a la matriz de confusión para el subconjunto del Reino Unido, se observan los siguientes resultados mostrados en el gráfico 19. De esta forma, se ha clasificado correctamente como gasto "bajo" (0) el 86,50% de los casos, como gasto "medio" (1) hay un 81,50% de los casos clasificados correctamente, y para el gasto "alto" (2), se ha clasificado de forma acertada el 67,70% de los casos.

En contraste, se ha clasificado incorrectamente como gasto "medio" un 12,70% de los datos y como gasto "alto" un 0,80% cuando realmente se trataba de gastos de tipo "bajo". Por otro lado, el modelo ha clasificado de manera incorrecta un 12,30% de los casos como gasto "bajo" y un 6,30% como gasto "alto" cuando, en realidad, se trataban de gastos del grupo "medio". Por último, se clasificaron incorrectamente como gasto "bajo" un 1,60% y un 30,70% de los casos los cuales, verdaderamente, eran gastos de tipo "alto".

## *5. REGRESSION DISCONTINUITY*

La regresión discontinua (en inglés, *regresión discontinuity*) (en adelante, RD) es un análisis no experimental o cuasiexperimental empleado para medir el impacto de un suceso. Este método fue introducido por primera vez por Donald L. Thistlethwaite y Donald T. Campbell con el fin de estimar los efectos de un tratamiento en un contexto no experimental considerando un punto de corte conocido. (Lee y Lemieux, 2010, p.281). Pese a la publicación por parte de Thistlethwaite y Campbell, "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment", en 1960, ha comenzado a llamar la atención de los economistas hace relativamente poco. Ha sido en las últimas décadas cuando el diseño de RD ha crecido significativamente, pasando a ser una herramienta de uso frecuente en económica, ciencias políticas, educación, criminología y otras muchas disciplinas (Cattaneo et al., 2019, p.4).

Siguiendo a Lee y Lemieux (2010, p.286), los creadores de este método sugieren la atribución de un salto discontinuo en el momento de corte como causa del suceso objeto de estudio (en el caso de estos autores, la obtención de premios al mérito de los estudiantes).

En este sentido, en RD se estudian dos variables: X e Y, donde la primera recibirá un tratamiento concreto, cuyos efectos repercutirán sobre la Y. Ese punto de corte representa el punto de inflexión, la diferencia, entre el conjunto que ha recibido el tratamiento y aquel que no lo ha recibido y con lo que se podrá observar la perturbación causada por el tratamiento en la variable que lo recibe.

A efectos de este trabajo, dicho punto de inflexión es el 2020, año de inicio de la pandemia, por lo que consideraríamos como tratamiento el estar ante una situación anterior a la pandemia o posterior.

El sistema de RD tiene distintas variantes, si bien para el caso que nos concierne, interesa la variante "Sharp" del RD. De acuerdo con Cattaneo et al. (2019, p.8), esta variante supone que la condición de tratamiento asignada es idéntica a la condición de tratamiento realmente recibida por todas las unidades. Si esta asignación fuese imperfecta, estaríamos ante la variante "Fuzzy".

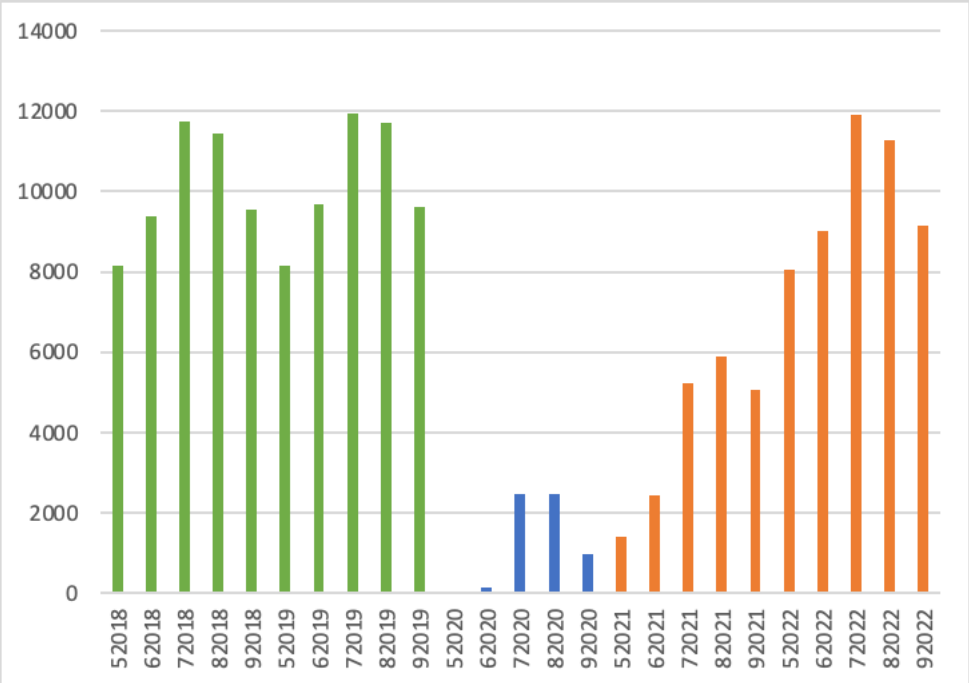
En esta línea, de acuerdo con estos autores, el S*harp* RD supone realizar la gráfica correspondiente a la probabilidad condicionada de recibir el tratamiento para distintos valores de la variable X. Matemáticamente la función del *Sharp* RD sería la siguiente:

$$\mathbb{P}(T_i = 1 \,|X_i = x)$$

Para elaborar la RD de este estudio concreto, se parte de una nueva base de datos. En esta nueva base de datos se considera el gasto total de los turistas en España en millones de euros. Se trata de datos recopilados del INE, los cuales tienen una periodicidad mensual. Consideraremos, además de los meses de temporada alta (junio, julio y agosto), también los meses de mayo y septiembre. Asimismo, se tendrá en cuenta los años relativos a la etapa anterior a la crisis sanitaria, así como los años de la etapa posterior a esta.

En una primera instancia observamos la siguiente distribución de los datos del gasto total. Observamos en verde los gastos totales de la etapa anterior a la crisis sanitaria, en azul los relativos al año 2020 y en naranja aquellos de la etapa posterior a la pandemia.

Gráfico 20. Diagrama de barras del gasto total en millones de euros de los turistas en España en los meses de mayo a septiembre de 2018 a 2022.



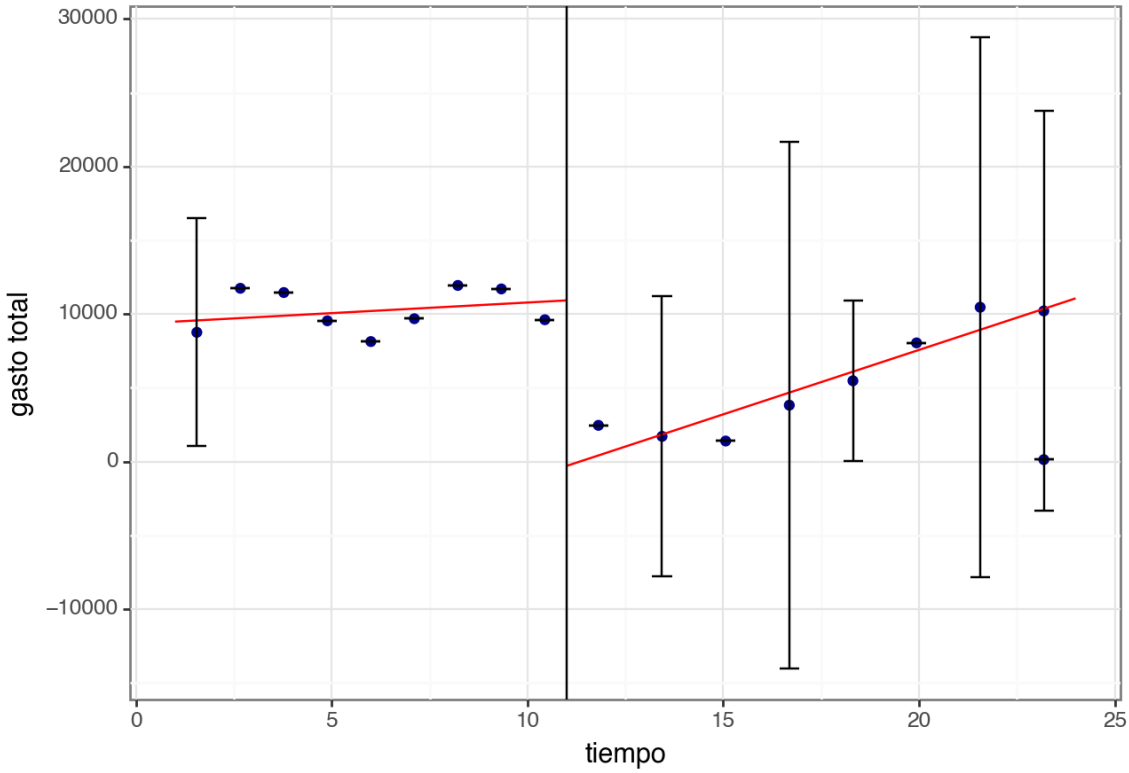

Elaboración propia. Fuente: INE.

Una vez se conocen los datos con los que se trabaja y su distribución, se ha procedido a elaborar la RD correspondiente, tal y como se venía introduciendo. Para la elaboración de esta, se ha empleado como herramienta principal el software de Python. En dicho software, se ha empleado el paquete "rdrobust", un paquete focalizado en este tipo de regresión.

Una vez instalado e importado este paquete, se ha procedido a elaborar el gráfico de RD con un intervalo de confianza del 95% y con el punto de corte en el inicio de la pandemia, tal y como se mencionó con anterioridad.

A continuación, se muestra la gráfica de RD resultante.

Gráfico 21. Gasto total de los turistas en España en millones de euros.




Elaboración propia.

RD para las variables tiempo y gasto, siendo el umbral de corte el inicio de la crisis sanitaria.

Se observa, como era esperado, una diferencia entre la etapa anterior a la pandemia y la etapa posterior. El salto en el punto de corte (el año de la crisis del coronavirus) se aprecia con claridad, este salto puede deberse a la globalidad de circunstancias existentes en dicho momento: por un lado, la falta de predisposición a viajar causada por el temor a contraer el virus; por otro lado, el cierre de fronteras de muchos países y en consecuencia, la imposibilidad de viajar desde dichos países; también, las restricciones internas a nivel nacional de España que obstaculizaban el viaje y por ello, podía provocar desmotivación a la hora de elegir España como destino.

En este sentido, siguiendo al Consejo de Ministros (2020), respecto a esto último, si bien el confinamiento en España finalizó el 4 de mayo con el inicio de la "fase 0" de la "desescalada", la transición hacia la llamada "nueva normalidad" implicaba el mantenimiento de un elevado número de restricciones. Si bien estas restricciones en cada fase eran más permisivas, no dejaban de limitar el libre albedrío de los turistas (porcentajes máximos de ocupación en restauración, la necesidad de un "certificado covid" a nivel europeo para probar que se la persona se había vacunado y sin el cual no se permitía a la persona viajar o acceder a ciertos establecimientos, entre otras restricciones que, a fin y al cabo, obstaculizaban la llegada del turismo a España.

En el gráfico se observa cómo antes de la pandemia, la tendencia del gasto era bastante regular, sin apenas variaciones más allá de las consecuentes de la estacionalidad del turismo. Sin embargo, con posterioridad a la crisis sanitaria, la tendencia es en alza, de recuperación hacia los niveles de antes de la pandemia, pero aún con notables diferencias.

Imagen 2. Resultados del análisis con *regression discontinuity*.

```
Call: rdplot
Number of Observations:                  24
Kernel:                             Uniform
Polynomial Order Est. (p):                1

                                 Left      Right
-------------------------------------------------
Number of Observations             10         14
Number of Effective Obs            10         14
Bandwith poly. fit (h)             10         13
Number of bins scale                1          1
Bins Selected                       9          8
Average Bin Length              1.111      1.625
Median Bin Length               1.111      1.625
IMSE-optimal bins                 9.0        8.0
Mimicking Variance bins           4.0       22.0

Relative to IMSE-optimal:
Implied scale                     1.0        1.0
WIMSE variance weight             0.5        0.5
WIMSE bias weight                 0.5        0.5
```

Elaboración propia.

# 6. CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

Al inicio del presente trabajo se introducía como pregunta de investigación y, por lo tanto, el objeto de este trabajo, si el modelo de turismo y la cantidad de gasto del turista ha cambiado tras la crisis sanitario o si, por el contrario, no se ha visto alterado.

Para responder a esta cuestión y a los objetivos del trabajo se ha realizado estudios a distintos niveles. En primer lugar, se ha alcanzado el primer objetivo llevando a cabo distintas pruebas estadísticas y contrastes de hipótesis con el fin de estudiar si existían diferencias entre las varianzas, las medias y las medianas entre los dos grupos (antes y después de la pandemia) respecto a la variable objetivo, el gasto diario. Se concluyó con estas pruebas que había diferencias significativas entre los dos grupos en los tres parámetros. En base a ello, resultó pertinente la continuación del estudio con el fin de estudiar dichas diferencias.

Puesto que la variable objetivo a estudiar en este trabajo es el gasto diario, se analizaron en un análisis estadístico primario todas las variables que afectaban al mismo para ver su comportamiento inicial. Con ello, respondiendo al segundo objetivo, se analizó qué variables de entre estas diez iniciales eran las más determinantes a la hora de explicar el gasto diario de los turistas. Al respecto, como resultado de este *feature selection* las variables más relevantes para antes y después de la crisis sanitaria para explicar el gasto diario no cambiaron, coincidiendo en las variables del país de residencia habitual, tipo de alojamiento, número total de pernoctaciones y contratación o no de paquete turístico.

Para dar respuesta al tercer objetivo, se procedió a elaborar un modelo *machine learning* que explicase el gasto diario con el fin también de estudiar el ajuste de este y su significado. Para una mayor eficiencia del modelo, se transformó la variable del gasto diario en una variable discreta con tres clases de gasto: bajo, medio y alto. El modelo resultante, un *fine Gaussian* SVM, aportaba buenos resultados tanto en su matriz de confusión como en su curva ROC. Se procedió a estudiar el ajuste de este modelo sobre dos muestras, un subconjunto general extraído aleatoriamente del conjunto de datos y un subconjunto concreto de los turistas del Reino Unido cuyos resultados en ambos casos fueron satisfactorios.

En último lugar, se respondió al cuarto objetivo mediante una regresión discontinua para estudiar el efecto causal que ha tenido la crisis sanitaria, teniendo en cuenta el antes y el después de esta, sobre el gasto de los turistas. Mediante la regresión discontinua se

comprobó la existencia de diferencias en el gasto total de los turistas en las dos fases al observar un salto considerable en el punto de corte (el año de la pandemia).

Es por ello por lo que se puede concluir y responder a la pregunta de investigación afirmando la existencia de una diferencia en el gasto de los turistas entre la fase anterior y la fase posterior a la pandemia. Si bien la tendencia es de recuperación del nivel del gasto, aún se observa diferencias en el gasto.

Por lo que respecta a las posibles futuras líneas de investigación, observar la evolución y tendencia de recuperación del gasto de los turistas en España resulta relevante e interesante. Por otro lado, como consecuencia del estallido de la guerra en Ucrania y, en consecuencia, la subida de precios de muchos bienes y servicios, resulta interesante estudiar la influencia de este fenómeno social sobre el turismo. En este sentido, puesto que Rusia es uno de los principales exportadores de petróleo, su protagonismo directo en la guerra influye sobre el precio de este *commodity*. En consecuencia, un aumento del precio de esta materia prima genera aumentos en todos los productos y servicios que requieran de ellas, entre ellos, el turismo.

De la misma forma, otra línea de investigación relevante sería emplear los modelos ML de manera predictiva para una cadena hotelera con el fin de cuantificar los gastos o el tipo de gasto de los turistas.

Asimismo, el estudio objeto del presente trabajo podría ampliarse en mayor detalle y mejorarse introduciendo más variables de carácter económico en los modelos de ML como pueden ser la inflación, el índice de seguridad política y el índice VIX, entre otros. Puesto que estas variables afectan a la idiosincrasia de los países, son relevantes para el estudio de la evolución de un sector como es el turístico.

7. BIBLIOGRAFÍA

**DATOS**

Instituto Nacional de Estadística. (s.f.). *Encuesta de Gasto Turistico. Resultados. Microdatos.*https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177002&menu=resultados&idp=1254735576863#!tabs-1254736195390

**REFERENCIAS**

Agencia Española de Medicamentos y Productos Sanitarios. (s.f.). *Información de vacunas autorizadas*. Recuperado el 2 de junio, 2023 de https://www.aemps.gob.es/la-aemps/ultima-informacion-de-la-aemps-acerca-del-covid%E2%80%9119/vacunas-contra-la-covid%E2%80%9119/informacion-de-vacunas-autorizadas/#

Banco Central Europeo. (s.f.). *Pandemic emergency purchase programme (PEPP)*. Recuperado el 3 de junio de https://www.ecb.europa.eu/mopo/implement/pepp/html/index.en.html

Banco de España. (2022). *Informe anual 2021*. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesAnuales/InformesAnuales/21/Fich/InfAnual_2021_Cap1.pdf

Benavides, F.G. y Silva-Peñaherrera, M. (2022). Datos y evidencias del teletrabajo, antes y durante la pandemia por COVID-19. *Archivos de prevención de riesgos laborales*, *25*(2), 133-146. https://orcid.org/0000-0003-0747-2660

Benito, M. (1 de abril, 2020). El dilema ético de decidir quién entra en la UCI. *La Razón.* https://www.larazon.es/salud/20200401/oswzy2iytrevhlfoyfbxmphmku.html

Blanco, M. (27 de abril, 2020). Confusion matrix. *LinkedIn*. https://www.linkedin.com/pulse/confusion-matrix-miguel-blanco/?originalSubdomain=es

Briones, E. (17 de enero, 2023). El PIB turístico en España superó en 2022 en un 1,4% el nivel prepandemia. *Europa Press*. https://www.europapress.es/turismo/nacional/noticia-turismo-20230117135150.html

Cardona, C.A. y Velásquez, J.D. (2006). Selección de características relevantes usando información mutua. *Dyna*, 73 (149), 149-163. http://www.scielo.org.co/pdf/dyna/v73n149/a13v73n149.pdf

Casas-Rojo, J. M., Antón-Santos, J., Millán-Núñez-Cortés, J., Lumbreras-Bermejo, C., Ramos, J., Roy-Vallejo, E., Artero-Mora, A., Arnalich-Fernández, F., García-Bruñén, J. M., Vargas-Núñez, J. A., Freire-Castro, S. J., Manzano-Espinosa, L., Perales-Fraile, I., Crestelo-Viéitez, A., Puchades-Gimeno, F., Rodilla-Sala, E., Solís-Marquínez, M. N., Bonet-Tur, D., Fidalgo-Moreno, M. P., . . . Gómez-Huelgas, R. (2020). Características clínicas de los pacientes hospitalizados con COVID-19 en España: resultados del Registro SEMI-COVID-19. *Revista Clínica Española*, 220(8), 480-494. https://doi.org/10.1016/j.rce.2020.07.003

Cattaneo, M.D., Idrobo, N. y Titiunik, R. (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108684606

Claveria, O., Torra, S. y Monte, E. (2016). Modelling cross-dependencies between Spain's regional tourism markets with an extension of the Gaussian process regression model. *SERIEs*, 7, 341-357. https://doi.org/10.1007/s13209-016-0144-7

Claveria, O., Torra, S. y Monte, E. (2016). Modelling tourism demand to Spain with Machine Learning techniques. The impact of forecast horizon on model selection. *Revista de Economía Aplicada*, 24(72), 109-132. http://www.redalyc.org/articulo.oa?id=96949059006

Comisión Europea. (s.f.). *Plan de recuperación para Europa*. Recuperado el 3 de junio de https://commission.europa.eu/strategy-and-policy/recovery-plan-europe_es

Consejo de Ministros. (2020). El Gobierno aprueba un plan de desescalada que se prolongará hasta finales de junio. *La Moncloa*. Recuperado el 6 de junio de https://www.lamoncloa.gob.es/consejodeministros/resumenes/Paginas/2020/280420-consejo_ministros.aspx

Cover, T.M. y Thomas, J.A. (2005). *Elements of Information Theory*. Wiley-Interscience. https://ia801400.us.archive.org/30/items/ElementsOfInformationTheory2ndEd/Wiley_-_2006_-_Elements_of_Information_Theory_2nd_Ed.pdf

Cullen, W., Gulati, G. y Kelly, B.D. (2020). Mental health in the COVID-19 pandemic. *QJM: An International Journal of Medicine*, 113 (5), 311-312. https://doi.org/10.1093/qjmed/hcaa110

DATAtab. (2023). *Prueba de Levene*. Recuperado el 7 de junio de https://datatab.es/tutorial/levene-test

DATAtab. (2023). *Prueba de Kruskal-Wallis*. Recuperado el 7 de junio de https://datatab.es/tutorial/kruskal-wallis-test

De Ibarreta Zorita, C.M., Álvarez Fernández, C., Borrás Palá, F., Budría Rodríguez, S. Curto González, T. y Escobar Torres, L.S. (2021). Tema 6. Validación del

modelo. En C.M. De Ibarreta Zorita (Coord.), *Modelos Cuantitativos para la Economía y la Empresa en 101 ejemplos* (67-78). EV Services.

Deloitte. (s.f.). *Plan Horizonte Sector Turismo*. Recuperado el 3 de junio de https://www2.deloitte.com/es/es/pages/consumer-business/articles/plan-horizonte-sector-turismo.html

Dong, D., Xu, X. y Wong, Y.F. (2019). Estimating the Impact of Air Pollution on Inbound Tourism in China: An Analysis Based on Regression Discontinuity Design. *Sustainability*, 11 (6), 1-18. https://doi.org/10.3390/su11061682

Esteban, P., y Méndez, R. (25 de marzo, 2020). El colapso llega: "La UCI no es para el más grave, sino para el que más años puede vivir". *El Confidencial*. https://www.elconfidencial.com/espana/madrid/2020-03-25/coronavirus-colapso-uci-pacientes-grave-anos-puede-vivir_2516151/

Gobierno de España. (s.f.). *Preguntas y respuestas: ¿Qué vacunas tendremos disponibles en España?*. Recuperado el 2 de junio, 2023 de https://www.vacunacovid.gob.es/preguntas-y-respuestas/que-vacunas-tendremos-disponibles-en-espana

Gómez, M.V. (28 de enero, 2021). España destruyó 622.600 empleos y la tasa de paro aumentó hasta el 16,13% en el año de la pandemia de coronavirus. *El País*. https://elpais.com/economia/2021-01-28/espana-destruyo-622600-empleos-y-la-tasa-de-paro-aumento-hasta-el-1613-en-el-ano-de-la-pandemia-de-coronavirus.html#

IBM. (2021). *Prueba de Kruskal-Wallis*. Recuperado el 8 de junio de https://www.ibm.com/docs/es/spss-statistics/beta?topic=tests-kruskal-wallis-test

Instituto Nacional de Estadística. (2022). *Estadística de Movimientos Turísticos en Frontera y Encuesta de Gasto Turístico (FRONTUR-EGATUR). Metodología.* https://www.ine.es/daco/daco42/frontur/frontur_egatur_metodologia.pdf

Instituto Nacional de Estadística. (s.f.). Aportación del turismo al PIB de la economía española: por PIB y sus componentes, valor absoluto/porcentaje/índice y periodo. https://www.ine.es/jaxi/Datos.htm?path=/t35/p011/rev19/serie/l0/&file=03001.px#!tabs-grafico

Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. En I. Maglogiannis et al. (Eds), *Emerging Artificial Interlligence Applications in Computer Engeneering* (3-24). IOS Press. Recuperado de: https://books.google.es/books?hl=es&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=machine+learning+classification&ots=CZswxvXBoo&sig=yTAbuiZZeBaG_OduvuzMAAhmkSc#v=onepage&q=machine%20learning%20classification&f=false

Kumar, V., & Minz, S. (2014). Feature Selection: A literature Review. *The Smart Computing Review*, 4(3). https://faculty.cc.gatech.edu/~hic/CS7616/Papers/Kumar-Minz-2014.pdf

Lee, D. S., y Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281–355. https://www.princeton.edu/~davidlee/wp/RDDEconomics.pdf

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Computing Surveys*, 50(6), 94. https://doi.org/10.1145/3136625

Linde, P. (5 de mayo, 2023). La OMS decreta el fin de la emergencia internacional por la covid. *El País*. https://elpais.com/sociedad/2023-05-05/la-oms-decreta-el-fin-de-la-emergencia-internacional-por-la-covid.html

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. https://www.inference.org.uk/itprnn/book.pdf

Mathworks (s.f.). *Visualize and Assess Classifier Performance in Classification Learner*. https://es.mathworks.com/help/stats/assess-classifier-performance.html

Ministerio de Industria, Comercio y Turismo. (s.f.) Estrategia de Turismo Sostenible de España 2030. https://turismo.gob.es/es-es/estrategia-turismo-sostenible/paginas/index.aspx

Ministerio de Sanidad. (s.f.). *Situación actual*. Recuperado el 3 de junio, 2023 de https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm

Mishra, R.K., Urolagin, S., Arul Jothi, J.A., Nawaz, N. y Ramkisoon, H. (2021). Machine Learning based Forecasting Systems for Worldwide International Tourist Arrival. *International Journal of Advanced Computer Science and Applications*, 12 (11), 55-64. https://dx.doi.org/10.14569/IJACSA.2021.0121107

Moreno Garrido, A. y Villaverde, J. (2019). De un sol a otro: turismo e imagen exterior española (1914-1984). *Ayer. Revista De Historia Contemporánea*, 114(2), 95–121. https://doi.org/10.55509/ayer/114-2019-04

Murthy, S., Archambault, P., Atique, A., Carrier, F. M., Cheng, M. P., Codan, C., Daneman, N., Dechert, W., Douglas, S. A., Fiest, K. M., Fowler, R. A., Goco, G., Gu, Y.,

Guerguerian, A., Hall, R. L., Hsu, J., Joffe, A. R., Jouvet, P., Kelly, L., . . . Wood, G. (2021). Characteristics and outcomes of patients with COVID-19 admitted to hospital and intensive care in the first phase of the pandemic in Canada: a national cohort study. *CMAJ Open,* 9(1), 181-188. https://doi.org/10.9778/cmajo.20200250

Naciones Unidas. (6 de octubre, 2020). La recuperación de la crisis del COVID-19 será más larga y lenta de lo previsto, advierte la CEPAL. *Noticias ONU*. https://news.un.org/es/story/2020/10/1481922

Nicola, M., Alsafi, Z., Sohrabi, C. Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. y Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery, 78*, 185-193. https://doi.org/10.1016/j.ijsu.2020.04.018

Nyuytiymbiy, K. (30 de diciembre, 2020). Parameters and hyperparameters in Machine Learning and Deep Learning. *Medium*. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac#:~:text=Hyperparameters%20are%20parameters%20whose%20values,parameters%20that%20result%20from%20it.

Organización Mundial de la Salud. (s.f.). *Brote de enfermedad por coronavirus (COVID-19)*. Recuperado el 2 de junio, 2023 de https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019

Organización Mundial de la Salud. (s.f.). *Coronavirus*. Recuperado el 2 de junio, 2023 de https://www.who.int/es/health-topics/coronavirus#tab=tab_1

Organización Mundial del Turismo. (s.f.). *Turismo internacional y COVID-19*. https://www.unwto.org/es/datos-turismo/turismo-internacional-covid-19

Paredes, M.R., Apaolaza, V., Fernández-Robin, C., Hartmann, P. y Yañez-Martinez, D. (2021). The impact of the COVID-19 pandemic on subjective mental well-being: The interplay of perceived threat, future anxiety and resilience. Personality and Individual Differences, 170. https://doi.org/10.1016/j.paid.2020.110455

Parlamento Europeo. (s.f.). *COVID-19: impulso de la investigación, inversión de futuro.* Recuperado el 2 de junio de https://www.europarl.europa.eu/news/es/press-room/20200319IPR75304/covid-19-impulso-de-la-investigacion-inversion-de-futuro

Parlamento Europeo. (2023). *Social and Economic Consequences of COVID-19*. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740071/IPOL_STU(2023)740071_EN.pdf

Peiró, R. (1 de mayo, 2021). Teoría de la información. *Economipedia*. https://economipedia.com/definiciones/teoria-de-la-informacion.html?nab=0

Perles Ribes, J.F., Ramón Rodríguez, A.B., Moreno Izquierdo, L. y Sevilla Jiménez, M. (2017). Economic crises and market performance—A machine learning approach. *Tourism economics*, 23 (3), 692-696. https://doi.org/10.5367/te.2015.0536

Scikit-learn. (s.f.) *Cross-validation: evaluating estimator performance*. https://scikit-learn.org/stable/modules/cross_validation.html

Scikit-learn (s.f.). *Feature Selection*. https://scikit-learn.org/stable/modules/feature_selection.html

Sevillano, E.G. (11 de marzo, 2020). La OMS declara el brote de coronavirus pandemia global. *El País*. https://elpais.com/sociedad/2020-03-11/la-oms-declara-el-brote-de-coronavirus-pandemia-global.html

Sigala, M. (2020). Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research. *Journal of Business Research*, 177, 312-321 https://doi.org/10.1016/j.jbusres.2020.06.015

Sigrist, P. (25 de enero, 2022). Demystifying ROC and precision-recall curves. *Medium*. https://towardsdatascience.com/demystifying-roc-and-precision-recall-curves-d30f3fad2cbf

Universidad Carlos III de Madrid. (s.f.) *Covid-19 Impact on Tourism in Spain*. Recuperado el 3 de junio de https://madiuc3m.com/covid-19-impact/

Vallejo Pousada, Rafael. (2002). Economía e historia del turismo español del siglo XX. *Historia contemporánea*, (25), 203-232. https://doi.org/10.1387/hc.5934

Venkatesh, B., y Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, 19(1), 3-26. https://doi.org/10.2478/cait-2019-0001

Williams, J. y Li, Y. (2009). Estimation of Mutual Information: A Survey. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds), *Rough Sets and Knowledge Technology* (389-396). Springer. https://doi.org/10.1007/978-3-642-02962-2_49

Zhang, Z. (2020). Generalized Mutual Information. *Stats*, 3(2), 158-165. https://doi.org/10.3390/stats3020013

Zhou, Z. (2021). *Machine Learning*. Springer. Recuperado de: https://books.google.es/books?hl=es&lr=&id=ctM-

EAAAQBAJ&oi=fnd&pg=PR6&dq=machine+learning&ots=oZQmX2XwYp&sig=iW2nr9Ao828CkTa_3dOhSLA6upw#v=onepage&q=machine%20learning&f=false

# 8. ÍNDICE DE TABLAS, GRÁFICOS E IMÁGENES

## 8.1. Tablas



## 8.2. Gráficos

### 8.3. Imágenes

# 9. ANEXOS

**Anexo I**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según mes y año.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Junio** | 35,91% | 24,92% |
| **Julio** | 35,16% | 35,93% |
| **Agosto** | 28,93% | 39,15% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo II**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según si el turista estaba en tránsito por España o no.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Tránsito** | 6,77% | 4,23% |
| **No tránsito** | 93,23% | 95,77% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo III**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según vía de salida y año

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Aeropuerto** | 77,09% | 83,06% |
| **Carretera** | 16,72% | 11,86% |
| **Puerto** | 4,35% | 2,89% |
| **Tren** | 1,84% | 2,18% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo IV**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según país de residencia y año.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Alemania** | 10,86% | 13,86% |
| **Bélgica** | 5,54% | 5,36% |
| **EE. UU.** | 3,20% | 3,51% |
| **Francia** | 15,98% | 16,80% |
| **Irlanda** | 3,22% | 2,73% |
| **Italia** | 6,84% | 8,05% |
| **Países Bajos** | 5,56% | 6,39% |
| **Países Nórdicos** | 6,53% | 5,71% |
| **Portugal** | 1,87% | 2,48% |
| **Reino Unido** | 20,84% | 15,81% |
| **Rusia** | 1,38% | 0,21% |
| **Suiza** | 3,37% | 4,00% |
| **Resto de Europa** | 4,57% | 7,64% |
| **Resto de América** | 4,64% | 4,11% |
| **Resto del Mundo** | 5,60% | 3,33% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo V**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según Comunidad Autónoma de destino y año.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Andalucía** | 11,29% | 13,04% |
| **Aragón** | 0,81% | 0,78% |
| **Asturias** | 0,67% | 0,72% |
| **Islas Baleares** | 20,15% | 19,01% |
| **Islas Canarias** | 9,54% | 9,67% |
| **Cantabria** | 0,87% | 0,71% |
| **Castilla y León** | 1,56% | 1,72% |
| **Castilla La-Mancha** | 0,39% | 0,44% |
| **Cataluña** | 24,93% | 21,56% |
| **Comunidad Valenciana** | 15,55% | 16,70% |
| **Extremadura** | 0,22% | 0,31% |
| **Galicia** | 1,65% | 1,48% |
| **Madrid** | 7,90% | 9,32% |
| **Murcia** | 1,78% | 1,56% |
| **Navarra** | 0,45% | 0,40% |
| **País Vasco** | 2,06% | 2,41% |
| **La Rioja** | 0,16% | 0,17% |
| **Ceuta** | 0,002% | 0,009% |
| **Melilla** | 0,008% | 0,010% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo VI**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según tipo de alojamiento y año.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Hoteles y similares** | 57,06% | 48,97% |
| **Resto de mercado** | 12,95% | 9,91% |
| **Alojamiento no de mercado** | 29,99% | 41,11% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo VII**. Tabla de frecuencia relativa en porcentaje de turistas encuestados según motivo del viaje.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Ocio** | 80,58% | 72,54% |
| **Negocios** | 5,57% | 5,60% |
| **Resto** | 13,85% | 21,86% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE

**Anexo VIII**. Tabla de frecuencia relativa en porcentaje de turistas encuestados en función de la contratación o no de un paquete turístico por año.

| | **Fase pre-covid** | **Fase post-covid** |
|---|---|---|
| **Sí** | 22,11% | 14,31% |
| **No** | 77,88% | 85,69% |
| *Total* | *100%* | *100%* |

Elaboración propia. Fuente: INE