UNIVERSIDAD DE BURGOS

Programa de Doctorado
Ingeniería y Tecnologías Industrial, Informática y Civil

TÉCNICAS INTELIGENTES EN LA GESTIÓN DE LA INDUSTRIA TURÍSTICA

Tesis Doctoral
mención Doctorado Internacional

Autor: **Anita Krupskaia Herrera Vaca**

Directores: **Dr. Ángel Arroyo Puente**
Dr. Alfredo Jiménez Palmero

Burgos 2024

Declaración del Autor

La tesis titulada "Técnicas Inteligentes en la Gestión de la Industria Turística" que presenta Anita Krupskaia Herrera Vaca para optar al título de Doctora, ha sido realizada dentro del programa de doctorado en "Ingeniería y Tecnologías Industrial, Informática y Civil", bajo la dirección del Dr. Ángel Arroyo Puente de la Universidad de Burgos y el Dr. Alfredo Jiménez Palmero de Kedge Business School, Burdeos Francia.

Burgos, marzo 2024

Los Directores	**La Doctoranda**
Dr. Ángel Arroyo Puente Universidad de Burgos	Anita Krupskaia Herrera Vaca
Dr. Alfredo Jiménez Palmero Kedge Business School	

A mis padres, Laura y Edgar

A mis hermanas y sobrinos

Agradecimientos

A la Universidad de Burgos y el programa de doctorado en “Ingeniería y Tecnologías Industrial, Informática y Civil”, por proporcionar el entorno propicio para el desarrollo e impulso a la investigación.

Al doctor Dimitrios Tsagdis y a Kedge Business School por brindarme la oportunidad de realizar mi estancia de investigación, un período de crecimiento personal y académico.

De forma muy especial a mis directores de tesis, los doctores Ángel Arroyo Puente y Alfredo Jiménez Palmero cuyo amplio conocimiento y apoyo constante, han sido fundamentales en la elaboración de esta tesis y en mi formación doctoral.

Formato de la Tesis

La presente tesis doctoral ha sido elaborada con el formato de compendio de artículos/publicaciones.

Las publicaciones que presentan los resultados obtenidos con esta tesis son los siguientes:

- Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review, J.UCS Journal of Universal Computer Science, https://lib.jucs.org/article/101550/
- Exploratory techniques to analyze Ecuador's tourism industry, Logic Journal of the IGPL, (In press).
- Forecasting hotel cancellations through Machine Learning, Expert Systems.

Los artículos señalados han sido aceptados para su publicación en revistas científicas del ámbito de la especialidad del trabajo desarrollado en la presente tesis e indexadas en el Journal Citation Reports (JCR).

Además, con la presente tesis se persigue optar por la mención de “Doctor internacional”, por lo que esta se ajusta a las indicaciones establecidas en el Procedimiento para la Obtención de la Mención de “Doctor Internacional” en el Título de Doctor, por la Universidad de Burgos, así como a los cambios del Real Decreto para los estudios de doctorado RD 576/2023 de fecha 4 de julio (BOE Nº 170 de 18 de julio de 2023).

De acuerdo con lo anterior y considerando la motivación y objetivos de esta tesis, el resto del presente documento queda estructurado como se expone a continuación:

- **Parte I**. Introducción y Objetivos: En esta primera parte se presenta la motivación y objetivos generales de la tesis, así como la estructura del propio documento.
- **Parte II**. Trabajos seleccionados: Esta segunda parte presenta los artículos desarrollados por el autor y directores de la tesis y aprobados para publicación en revistas científicas del ámbito de estudio, enunciados anteriormente.
- **Parte III**. Discusión y Conclusiones: Esta parte analiza los resultados de los artículos seleccionados, presentando además las conclusiones finales, así como propuestas para el trabajo futuro de esta línea de investigación.

Resumen

La industria turística representa una oportunidad en el desarrollo de diferentes localidades, gracias a las inversiones realizadas en infraestructura y servicios, así como a la generación de empleo, factores que impulsan el crecimiento económico y social. Esta industria ha experimentado transformaciones rápidas y profundas, alcanzando mayor eficiencia en la gestión de los recursos, optimizando la planificación y mejorando la operación de los servicios turísticos, todo esto principalmente impulsado por la adopción de nuevas tecnologías. En este contexto, las técnicas de Machine Learning (ML) se muestran como un recurso prometedor frente a una industria que debe innovar de acuerdo a los requerimientos de los turistas. La integración de ML, permite analizar grandes conjuntos de datos para adaptarse a las demandas cambiantes del mercado y ofrecer servicios más eficientes, impulsando así la innovación y la competitividad de la industria turística. La presente tesis doctoral aborda el estudio de las técnicas de ML en el ámbito de la gestión turística, tratado en tres artículos de investigación que han sido aprobados para la publicación en revistas científicas indexadas en JCR.

1. El primer artículo se centra en la revisión y síntesis de investigaciones previamente publicadas, sobre la Inteligencia Artificial en el sector turístico. El estudio presenta una categorización de las aplicaciones de Inteligencia Artificial en diferentes áreas del turismo, reconociendo estudios y herramientas válidas para el crecimiento e innovación del sector y destacando la apropiación de la Inteligencia Artificial por parte de la industria turística
2. El siguiente estudio utiliza técnicas de Soft Computing para analizar variables relacionadas con la operación de las empresas turísticas de Ecuador, verificando la tendencia de la operación en diferentes años y generando una fuente de información válida para la toma de decisiones. En el estudio se aplican técnicas de reducción de dimensionalidad con el objeto de mejorar la interpretación minimizando la pérdida de información. Además, se aplican técnicas de agrupación para crear grupos acorde a la similaridad de las características y proporcionar una representación visual y numérica de la relación de los datos entre sí.
3. El tercer artículo se enfoca en el uso de las técnicas de ML para prever cancelaciones de reservaciones de hotel. El trabajo analiza e implementa pasos clave como el preprocesamiento de datos, la configuración de hiperparámetros y la evaluación de los modelos utilizando métricas y gráficas de rendimiento. El artículo incluye clasificadores base, clasificadores de conjunto y redes neuronales.

En los estudios analizados en esta tesis, se destaca la eficacia de las técnicas de ML en la generación de información valiosa para respaldar la toma de decisiones en la gestión turística. Al analizar las variables relacionadas con la operación de las empresas turísticas, es posible identificar la tendencia de la operación en diferentes períodos de tiempo, reconociendo además el efecto de factores externos. Así también, a partir de las técnicas de ML es factible

obtener modelos de pronóstico con alta precisión, muy útiles en la gestión para anticipar tendencias y optimizar la planificación en el sector turístico. Asimismo, una exhaustiva revisión de la literatura relacionada con la Inteligencia Artificial en la industria turística a través de aplicaciones, evidencia cómo estas tecnologías transforman la manera de ofrecer servicios, a la vez que enriquecen la experiencia del usuario, impulsando la innovación y desarrollo en el sector. En síntesis, estas técnicas son un recurso de gran ayuda, que permiten alcanzar mejores niveles de competitividad en un mercado en constante evolución.

Abstract

The tourism industry represents an opportunity for the development of different localities, thanks to the investments made in infrastructure and services, as well as the generation of employment, factors that drive economic and social growth. This industry has undergone rapid and profound transformations, achieving greater efficiency in the management of resources, optimizing planning and improving the operation of tourism services, all of this driven mainly by the adoption of new technologies. In this context, Machine Learning (ML) techniques are a promising resource for an industry that must innovate according to tourists' requirements. The integration of ML allows analyzing large datasets to adapt to changing market demands and offer more efficient services, thus boosting innovation and competitiveness of the tourism industry. This doctoral thesis addresses the study of ML techniques in the field of tourism management, addressed in three research articles that have been approved for publication in scientific journals indexed in Journal Citation Reports.

1. In the first article, Soft Computing techniques are used to analyze variables related to the operation of tourism companies in Ecuador, verifying the trend of the operation in different years and generating a valid source of information for decision making. In the study, dimensionality reduction techniques are applied to improve the interpretation, minimizing the loss of information. In addition, clustering techniques are applied to create groups according to the similarity of the characteristics and to provide a visual and numerical representation of the relationship of the data with each other.
2. The following study focuses on the review and synthesis of previously published research on Artificial Intelligence in the tourism sector. The study presents a categorization of the applications of Artificial Intelligence in different areas of tourism, recognizing valid studies and tools for the growth and innovation of the sector and highlighting the appropriation of Artificial Intelligence by the tourism industry.
3. The third paper focuses on the use of ML techniques to foresee hotel reservation cancellations. It discusses and implements key steps such as data preprocessing, hyperparameter settings, and model evaluation using performance metrics and graphs. The paper includes base classifiers, ensemble classifiers and neural networks.

The studies analyzed in this thesis demonstrate the effectiveness of ML techniques to generate valuable information to support decision-making in tourism management. By analyzing the variables related to the operation of tourism companies, it is possible to identify the trend of the operation in different periods of time, also recognizing the effect of external factors. Also, from ML techniques it is possible to obtain highly accurate forecasting models, which are very useful in management to anticipate trends and optimize planning in the tourism sector. Likewise, an exhaustive review of the literature related to Artificial Intelligence in the tourism industry through applications, shows how these technologies transform the way of offering services, while enriching the user's experience, driving innovation and

development in the sector. In short, these techniques are a very helpful resource to improve competitiveness levels in a constantly evolving market.

Índice

Parte I. Introducción y Objetivos 17

Capítulo I. Introducción 19

Antecedentes 19

Sobre la solución planteada 22

Metodología 24

Objetivos 25

Publicaciones seleccionadas 26

Referencias 27

Parte II. Artículos Seleccionados 31

Capítulo II. Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review 33

Resumen 33

Capítulo III. Exploratory techniques to analyze Ecuador's tourism industry 53

Resumen 53

Capítulo IV. Forecasting hotel cancellations through Machine Learning 71

Resumen 71

Parte III. Discusión y Conclusiones 95

Conclusiones Finales 97

Parte I. Introducción y Objetivos

Capítulo I. Introducción

Antecedentes

El turismo es una de las industrias más importantes a nivel mundial, por el número de empleos que genera y el aporte en el desarrollo de diferentes localidades. La Organización Mundial de Turismo (UNWTO) define al turismo como un fenómeno social, cultural y económico que supone el desplazamiento de personas a países o lugares fuera de su entorno habitual por motivos personales, profesionales o de negocios [1].

El origen del turismo se asocia con actividades de ocio, comercio y religión en las civilizaciones antiguas. Sin embargo, la identificación del turismo como actividad económica se dio en el siglo XIX con el desarrollo de nuevos medios de transporte, como ferrocarriles y barcos a vapor, la construcción de infraestructura, incluidos hoteles, carreteras y atracciones turísticas, así como la creación de agencias para organizar viajes y excursiones, lo cual generó la base para una industria turística moderna [2].

Desde su establecimiento, la industria turística experimentó un notable desarrollo y expansión. Es así que, durante las últimas décadas, había logrado un crecimiento sostenido, llegando a aportar más del 10% del PIB mundial y 334 millones de puestos de trabajo en el año 2019 [3], antes de la declaratoria de pandemia provocada por la Covid-19.

La pandemia generó un impacto negativo en el turismo, la operación de esta industria se redujo significativamente por las restricciones de movilidad, cierre de fronteras y el distanciamiento social, medidas adoptadas a nivel mundial, para disminuir los contagios y muertes por Covid-19. Según el World Travel & Tourism Council [4, 5], el aporte del sector de viajes y turismo disminuyó al 5.3% en el 2020, porcentaje que pasó al 6.1% en el 2021 y 7.6% en el año 2022. Así también, en el año 2020 se perdieron 62 millones de empleos en el sector turístico, recuperando 18,2 millones de empleos en el año 2021 y 22 millones de empleos en el año 2022.

En este escenario, el sector turístico ha debido explorar mecanismos para desarrollar una gestión eficiente, que permita la recuperación de esta industria. La tecnología se ha consolidado como un facilitador clave en la planificación y ejecución. Cada vez más, los usuarios recurren a plataformas con información actualizada sobre rutas, horarios, disponibilidad de servicios, recomendaciones, convirtiéndose así en turistas mejor informados, en una industria que se ha reinventado para adaptarse a una nueva realidad, que requiere de decisiones más precisas e inmediatas.

A su vez, dicha tendencia ha generado un incremento notable de interacciones en plataformas ONLINE como Booking, Airbnb, Uber, TripAdvisor, entre otras, diseñadas para ofrecer productos y servicios turísticos, lo que aporta una valiosa fuente de datos para analizar las opiniones y preferencias de los clientes, para comprender las nuevas expectativas del mercado a fin de ajustar las estrategias y servicios de manera óptima. Por ende, una eficiente gestión turística requiere obtener información válida a partir de los datos registrados en motores de búsqueda [6], plataformas de registros online [7], así como en los registros administrativos de los establecimientos turísticos [8].

En el ámbito de la gestión, la innovación y la implementación de técnicas inteligentes son aspectos fundamentales para el desarrollo exitoso de la industria turística. La innovación permite la adaptación de las empresas a los nuevos requerimientos de los clientes, generando ventajas competitivas y fomentando la eficiencia de las empresas. Mientras que la implementación de técnicas inteligentes contribuye a mejorar la operación de las empresas turísticas, al analizar datos que generan información relevante y útil para la toma de decisiones, creación de estrategias y formulación de políticas apropiadas.

Al aplicar técnicas inteligentes a conjuntos de datos relacionados con la gestión de la industria turística, se busca evidenciar la eficacia de estas técnicas para mejorar la visualización de datos de alta dimensión [9], descubrir patrones ocultos [10] y generar pronósticos con alta precisión [11]. Las técnicas de reducción de dimensionalidad y agrupamiento en la gestión turística, han sido objeto de diferentes estudios. En el documento de Gabor et *al*. [12], se analiza la competitividad de la industria turística de los países europeos frente a la competitividad de países que no pertenecen a esta zona, identificando los atributos más relevantes en el posicionamiento de los destinos turísticos. Claveria et *al*. [13] propone un modelo para agrupar los destinos turísticos de acuerdo a la evolución de los indicadores turísticos, proporcionando información importante para los gerentes y responsables de la formulación de políticas. En el estudio de Sztorc [14], se analizan variables relacionadas con la automatización de varios procesos en la cadena de valor de los hoteles, como estrategia durante la crisis generada por la pandemia del COVID-19. Para el tratamiento de los datos se utilizaron métodos descriptivos, análisis de componentes principales y análisis de conglomerados, y como resultado se observó que la pandemia obligó a los hoteleros a optimizar procesos para desarrollar estrategias de reacción ante crisis derivadas por agentes externos.

De igual manera, durante la reactivación de la industria turística, se ha reconocido la necesidad de cambiar el modelo turístico hacia un modelo más resiliente y sostenible, capaz de enfrentar una crisis. El estudio presentado por Neshat et *al.* [15] utiliza redes neuronales profundas para analizar experiencias del sector turístico para superar crisis en la industria turística. El caso de estudio toma los datos de Irán y los resultados se orientan al turismo interno con una planificación sostenible. El estudio de Imam and Ananda [16] utiliza técnicas de ML para identificar factores influyentes en el crecimiento del empleo turístico directo e indirecto en Sri Lanka. El estudio analiza información turística para identificar los factores influyentes que caracterizan el crecimiento del empleo en el sector turístico de Sri Lanka, concluyendo que los tomadores de decisiones deben prestar especial atención al crecimiento de las llegadas en el cuarto trimestre del año. Estudios de este tipo son de utilidad para los responsables de la definición de estrategias para el desarrollo del turismo en localidades con semejantes características de desarrollo económico.

Por otra parte, diversos estudios analizan las técnicas de Machine Learning (ML) [17] utilizadas en la construcción de modelos de pronósticos para el sector turístico. Estos modelos permiten mantener una adecuada planificación de recursos, alcanzando una gestión eficiente que evite problemas de sobreoferta o falta de capacidad en la atención a clientes [18]. Los pronósticos ayudan a las empresas a adaptarse a los cambios de la demanda, establecer estrategias de precios atractivos para los clientes, y maximizar los ingresos y la rentabilidad [19]. La precisión de los modelos es un elemento crítico en la toma de decisiones, un aspecto considerado en las investigaciones con el objetivo de generar información de mayor calidad [20].

En lo referente al pronóstico de la cancelación de reservaciones de servicios turísticos, estos proporcionan información válida para la creación de políticas de cancelación, cuyo principal objetivo es gestionar adecuadamente las cancelaciones [21]. La investigación realizada por Antonio et *al.* [22] aplica técnicas de ML en un volumen grande de datos relacionados con las reservaciones de diferentes hoteles. La investigación presenta mejores resultados al incluir variables que capturan las características y el entorno operativo de cada hotel, demostrando además la efectividad de las técnicas de ML para obtener pronósticos con alta precisión. A su vez, el estudio de Sánchez-Medina [23] obtiene el pronóstico de cancelaciones de reservaciones de hotel utilizando métodos de ML. En este caso, el conjunto de datos está formado por un número reducido de variables, específicamente aquellas que se corresponden con la información más utilizada en una reservación. Los modelos alcanzan un Accuracy del 80% para aquellos modelos basados en árboles o SVM y el 98% en una red neuronal. Por su parte, la investigación de Sánchez et *al.* [24] aplica métodos de ML en registros de nombres personales para pronosticar las cancelaciones en períodos cortos de tiempo previos a la llegada de los huéspedes al hotel. El pronóstico con la precisión más baja corresponde al período de 4 días antes del check-in, 60% con redes neuronales y 73% con un método de ensamble.

A diferencia de los trabajos previamente analizados, la presente tesis doctoral aborda el estudio de técnicas inteligentes en la gestión turística, enfocándose en el estudio de la operación de las empresas turísticas y el uso de modelos para prever la cancelación de las reservaciones en este sector, como elementos claves para la toma de decisiones estratégicas.

Sobre la solución planteada

En relación a la situación ya descrita, el presente trabajo doctoral plantea el uso de técnicas de Inteligencia Artificial [25] en la gestión turística. Estas técnicas analizan grandes volúmenes de datos, identifican patrones y relaciones complejas entre variables, para encontrar soluciones óptimas y proporcionar información valiosa en la toma de decisiones.

Las técnicas de reducción de dimensionalidad permiten reducir un conjunto de datos de cierta dimensión a uno de menor dimensión, con lo cual disminuye la complejidad de la alta dimensionalidad, mientras se mantiene la relación entre las variables. Es así que las técnicas de reducción de dimensionalidad facilitan la identificación de patrones de datos y la comprensión de la relación entre las variables [26] [27]. Las técnicas de reducción de dimensionalidad utilizadas en los trabajos presentados en esta tesis doctoral son: Principal Component Analysis (PCA) [28], Locally Linear Embedding (LLE) [29] y por último Isometric Mapping (ISOMAP) [30]. Al aplicar las técnicas de reducción de dimensionalidad en el conjunto de datos, se logra disminuir el número de variables, formando un nuevo conjunto de datos que mantiene la información relevante de los datos originales. La representación gráfica del nuevo conjunto de datos permite identificar la estructura de los datos y detectar valores anómalos o outliers.

En los trabajos considerados en el presente trabajo doctoral, también se utilizan las técnicas de agrupamiento [31-33], cuya principal característica es agrupar instancias de manera no supervisada. El agrupamiento se realiza según la similitud de los objetos, para descubrir patrones y la relación entre los objetos. Las técnicas de agrupamiento consideradas en el presente estudio son: *K*-means [34], *K*-medoids [35] y Hierarchical Clustering Analysis [36]. La efectividad del agrupamiento y el número de grupos se determina con métricas cuantitativas basadas en la cohesión interna de los grupos y la separación entre ellos [37]. Las métricas utilizadas en el presente trabajo son: Elbow Method [38], Silhouette Coefficient [39], Davies-Bouldin Index [40] y Dunn Index [41].

La aplicación de las técnicas de agrupamiento y reducción de dimensionalidad en variables relacionadas con la operación de empresas turísticas permite identificar patrones de comportamiento en los datos y relaciones entre ellos para extraer conocimiento, para la gestión de esta industria. Esto posibilita el reconocer la tendencia de la operación de las empresas en los años analizados, identificar el incremento o disminución de la operación. Esta información abre la posibilidad de desarrollar estrategias más efectivas para fortalecer y promover el crecimiento sostenible de esta industria.

Otra categoría de técnicas abordadas en este trabajo, corresponde a aquellas usadas en la creación de modelos de pronósticos mediante el uso de ML. Esta categoría incluye el aprendizaje supervisado, el cual se caracteriza por utilizar conjunto de datos etiquetados para entrenar algoritmos que clasifican datos o predicen con precisión, un resultado específico de interés [42]. Así también, esta categoría incluye modelos con estructuras más poderosas como las **R**edes **N**euronales **A**rtificiales RNA supervisadas [43], definidas como herramientas de IA cuya funcionalidad se basa en neuronas biológicas. Las RNA están formadas por n nodos o neuronas interconectados en una estructura de capas semejante al cerebro humano. Los modelos de pronósticos analizados en esta tesis doctoral incluyen los siguientes métodos de clasificación y RNA: Decision Tree Classifier [44], Random Forest Classifier [45], AdaBoost Classifier [46], XgBoost Classifier [47], Multilayer Perceptron Neural Network [48], Radial Basis Function Neural Network [49] y Deep Neural Network [50].

Con el objetivo de obtener alta precisión en los pronósticos, es fundamental realizar un adecuado preprocesamiento de los datos y la identificación de los mejores hiperparámetros. Las técnicas de preprocesamiento, aplicadas a los conjuntos de datos de este trabajo, incluyen la corrección de los valores nulos, eliminación de los outliers [51], identificación de los registros con datos incorrectos, codificación de las características categóricas, normalización y estandarización de las características y la eliminación de las características menos significativas. A su vez, la evaluación del desempeño de los modelos se realiza a partir de la matriz de confusión y las métricas de rendimiento como: accuracy, precision, recall, specificity y F1 [52]. De esta forma es posible identificar los mejores métodos para pronosticar las cancelaciones de las reservaciones de hotel.

Metodología

La construcción de conocimiento científico a través de una tesis por compendio requiere la articulación de metodologías que guíen la investigación y consoliden la validez de los resultados obtenidos. En este contexto, a continuación, se detallan las consideraciones comunes de las metodologías implementadas en los trabajos abordados en esta tesis doctoral:

- Caso de estudio: La definición del caso de estudio proporciona una base sólida para el desarrollo del trabajo, incluye el propósito de la investigación y cómo este será abordado. En los trabajos considerados en esta tesis doctoral, los casos de estudio están relacionados con la gestión turística, específicamente la operación de las empresas turísticas y el pronóstico de las cancelaciones de reservaciones de hotel. En el caso de estudio se define, además, el ámbito, la ubicación y el período de estudio.
- Conjunto de datos: Los datos corresponden a una fuente de datos pública, lo que permite la reproducción de los experimentos por otros investigadores. El tamaño del conjunto de datos es el adecuado tanto en número de variables como en número de registros. Así también, las variables consideradas en los estudios, incluyen información necesaria y relevante.
- Tratamiento de datos: Implementación de técnicas para el preprocesamiento de los datos, siendo importante realizar un Análisis de Datos Exploratorios (EDA) para identificar inconsistencias y aplicar las estrategias de corrección adecuadas.
- Técnicas inteligentes: La elección de técnicas depende de la naturaleza de los datos y de las características del problema abordado. La correcta elección de la técnica influye en la capacidad del modelo para identificar patrones relevantes en los datos, a la vez que impacta en la interpretación de los resultados.
- Análisis de los resultados: Los artículos incluidos en esta tesis analizan los resultados desde el ámbito de la IA y la gestión turística, identificando el comportamiento de la operación de las empresas turísticas en diferentes períodos de tiempo, así como la precisión de los pronósticos de la cancelación de servicios turísticos.

En los capítulos II al IV se detalla la metodología de cada artículo, proporcionando así un análisis minucioso de las decisiones metodológicas y los procedimientos implementados en cada instancia de la investigación. Finalmente, se presentan las conclusiones obtenidas desde el punto de vista de la gestión de la industria turística.

Objetivos

Esta tesis doctoral intenta contribuir al estudio de las técnicas de ML que ayudan a la toma de decisiones. Su objetivo general es validar la eficacia de las técnicas de ML para generar información relevante en la toma de decisiones en la gestión turística.

En el cumplimiento de este objetivo se han considerado las siguientes actividades:

- Revisar y analizar las investigaciones y conocimientos existentes sobre las técnicas y modelos de ML aplicadas a la gestión turística. Una revisión profunda permite identificar las tendencias y avances actuales, así como, determinar aspectos aún no explorados en la literatura académica y científica.
- Definir casos de estudio interesantes que promuevan una investigación relevante y novedosa, que añada valor a la aplicación de ML a la gestión turística.
- Identificar bases de datos públicas relacionadas con la gestión turística, las cuales deben incluir un número adecuado de registros y variables. A su vez, es importante incluir las variables más relevantes que permitan obtener resultados confiables.
- Aplicar diferentes técnicas de ML especializadas en la reducción de dimensionalidad, agrupamiento y pronósticos de cancelación de servicios turísticos, e identificar la técnica con mejor desempeño en escenarios específicos.
- Beneficiar la gestión de las empresas turísticas evidenciando la importancia de tomar decisiones basadas en el conocimiento generado mediante técnicas de ML.

El objetivo de esta tesis doctoral se alcanza mediante la aplicación de técnicas de IA en casos de estudio relacionados con la gestión turística, cada uno de los cuales engloba un contexto geográfico, ámbito de estudio e intervalo de tiempo específico. Referente al ámbito geográfico, se consideran datos de Ecuador en el Capítulo III y de Portugal en el Capítulo IV. En cuanto al campo de estudio, los trabajos se relacionan con la operación de las empresas turísticas en el Capítulo III, las reservaciones de hotel en el Capítulo IV y aplicaciones de IA en el turismo en el Capítulo II. Además, se considera el período de tiempo entre 2016 y 2023 en el Capítulo II, los años 2015, 2019 y 2020 en el Capítulo III, y el período entre 2015 y 2017 en el Capítulo IV.

Las publicaciones que forman parte de esta tesis doctoral se describen en el apartado *Publicaciones seleccionadas*.

Publicaciones seleccionadas

Las publicaciones que se listan a continuación, están relacionadas con el tema de estudio abordado en esta tesis doctoral.

1. Herrera, A., Arroyo, Á., Jiménez, A., & Herrero, Á. (2023). **Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review**. *J.UCS Journal of Universal Computer Science*, 29(12), 1439-1460. https://doi.org/10.3897/jucs.101550.
2. A. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero. (In Press). **Exploratory techniques to analyze Ecuador's tourism industry**. *Logic Journal of the IGPL.*
3. A. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero. **Forecasting hotel cancellations through Machine Learning**. *Expert Systems.* (2023). DOI: 10.22541/au.168175807.71371411/v1

Las tres publicaciones han sido consideradas para esta tesis doctoral por su relevancia y relación con el tema de estudio. El primer estudio "***Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review*",** recopila y analiza estudios sobre la IA en el sector turístico, la información se presenta categorizada según aplicaciones de IA. El segundo estudio titulado "***Exploratory techniques to analyze Ecuador's tourism industry***" analiza variables relacionadas con la operación de las empresas turísticas de Ecuador, identificando el comportamiento de las variables en diferentes períodos de tiempo y ante factores externos como protestas sociales y la pandemia generada por el Covid-19. La tercera publicación se denomina "***Forecasting hotel cancellations through Machine Learning*",** y presenta un estudio de modelos para pronosticar las cancelaciones de reservaciones de hotel, utilizando un dataset de reservaciones de hoteles ubicados en Portugal. El estudio incluye técnicas de preprocesamiento y la identificación de los mejores hiperparámetros para obtener pronósticos con mayor precisión.

4. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero. **Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Techniques**. *16 International Conference on SoftComputing Models in Industrial and Environmental Applications. Advances in Intelligent Systems and Computing*, vol 1401. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-87869-6_77

El estudio titulado "***Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Techniques***" es un inicio a la investigación y sirve de base a trabajos de investigación más ambiciosos.

Referencias

[1] UNWTO. (2023). *GLOSSARY OF TOURISM TERMS* [Online]. Available: https://www.unwto.org/glossary-tourism-terms.

[2] J. Towner and G. Wall, "History and tourism," *Annals of Tourism Research,* vol. 18, no. 1, pp. 71-84, 1991/01/01/ 1991.

[3] WTTC. (2023). *Economic Impact Research* [Online]. Available: https://wttc.org/research/economic-impact#:~:text=Prior%20to%20the%20pandemic%2C%20Travel,%24%2010%20trillion)%20in%202019.

[4] WTTC. (2022). *Travel & Tourism Economic Impact 2022* [Online]. Available: https://wttc.org/Portals/0/Documents/Reports/2022/EIR2022-Global%20Trends.pdf.

[5] WTTC. (2023). *Economic Impact Tourism* [Online]. Available: https://wttc.org/Research/Economic-Impact.

[6] I. Sabuncu, "Understanding Tourist Perceptions and Expectations During Pandemic Through Social Media Big Data," in *Handbook of Research on the Impacts and Implications of COVID-19 on the Tourism Industry*: IGI Global, 2021, pp. 330-350.

[7] G. Solazzo, Y. Maruccia, V. Ndou, and P. Del Vecchio, "How to exploit Big Social Data in the Covid-19 pandemic: the case of the Italian tourism industry," *Service Business,* vol. 16, no. 3, pp. 417-443, 2022/09/01 2022.

[8] R. Ramsawak and P. S. Mohan, "COVID-19 and Big DataBig Data Research: Techniques and Applications in the Caribbean," in *Interdisciplinary Perspectives on COVID-19 and the Caribbean, Volume 2: Society, Education and Human Behaviour*, S. Roberts, H. A. F. DeShong, W. C. Grenade, and D. Devonish, Eds. Cham: Springer Nature Switzerland, 2023, pp. 513-543.

[9] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion,* vol. 59, pp. 44-58, 2020/07/01/ 2020.

[10] P. Choudhury, R. T. Allen, and M. G. Endres, "Machine learning for pattern discovery in management research," *Strategic Management Journal,* vol. 42, no. 1, pp. 30-57, 2021.

[11] X. Li, R. Law, G. Xie, and S. Wang, "Review of tourism forecasting research with internet data," *Tourism Management,* vol. 83, p. 104245, 2021/04/01/ 2021.

[12] M. R. Gabor, L. C. Conţiu, and F. D. Oltean, "A Comparative Analysis Regarding European Tourism Competitiveness: Emerging Versus Developed Markets," *Procedia Economics and Finance,* vol. 3, pp. 361-366, 2012/01/01/ 2012.

[13] O. Claveria and A. Poluzzi, "Positioning and clustering of the world's top tourist destinations by means of dimensionality reduction techniques for categorical data," *Journal of Destination Marketing & Management,* vol. 6, no. 1, pp. 22-32, 2017/03/01/ 2017.

[14] M. Sztorc, "Autonomous Enterprise as a Model of Hotel Operation in the Aftermath of the COVID-19 Pandemic," *Sustainability,* vol. 14, no. 1, p. 97, 2022.

[15] N. Neshat, S. Moayedfar, K. Rezaee, and N. Amrollahi Biuki, "Sustainable planning of developing tourism destinations after COVID-19 outbreak: a deep learning approach," *Journal of Policy Research in Tourism, Leisure and Events,* pp. 1-21.

[16] T. Imam and J. Ananda, "Machine learning for characterizing growth in tourism employment in developing economies: an assessment of tourism employment in Sri Lanka," *Current Issues in Tourism,* vol. 25, no. 16, pp. 2695-2716, 2022/08/18 2022.

[17] Z.-H. Zhou, *Machine learning*. Gateway East, Singapore: Springer Nature, 2021.

[18] I. Ghalehkhondabi, E. Ardjmand, W. A. Young, and G. R. Weckman, "A review of demand forecasting models and methodological developments within tourism and passenger transportation industry," *Journal of Tourism Futures,* vol. 5, no. 1, pp. 75-93, 2019.

[19] Z. Sulong, M. Abdullah, and M. A. F. Chowdhury, "Halal tourism demand and firm performance forecasting: new evidence from machine learning," *Current Issues in Tourism,* pp. 1-17, 2022.

[20] H. Song, R. T. R. Qiu, and J. Park, "A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting," *Annals of Tourism Research,* vol. 75, pp. 338-362, 2019/03/01/ 2019.

[21] Z. Liu, P. Jiang, J. Wang, Z. Du, X. Niu, and L. Zhang, "Hospitality order cancellation prediction from a profit-driven perspective," *International Journal of Contemporary Hospitality Management,* vol. 35, no. 6, pp. 2084-2112, 2023.

[22] N. Antonio, A. de Almeida, and L. Nunes, "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior," *Cornell Hospitality Quarterly,* vol. 60, no. 4, pp. 298-319, 2019.

[23] A. J. Sanchez-Medina and E. C-Sanchez, "Using machine learning and big data for efficient forecasting of hotel booking cancellations," (in English), *International Journal of Hospitality Management,* Article vol. 89, p. 9, Aug 2020, Art no. 102546.

[24] E. C. Sánchez, A. J. Sánchez-Medina, and M. Pellejero, "Identifying critical hotel cancellations using artificial intelligence," *Tourism Management Perspectives,* vol. 35, p. 100718, 2020/07/01/ 2020.

[25] S. Russell and P. Norvig, P. Hall, Ed. *Artificial intelligence a modern approach*, Third Edition ed. Pearson Education, Inc., 2010.

[26] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res,* vol. 10, no. 66-71, p. 13, 2009.

[27] M. Garzon, C.-C. Yang, D. Venugopal, N. Kumar, K. Jana, and L.-Y. Deng, *Dimensionality Reduction in Data Science*. Springer Cham, 2022, pp. XI, 265.

[28] P. Sanguansat, *Principal Component Analysis: Engineering Applications*. BoD–Books on Demand, 2012.

[29] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, 2000.

[30] G. Pai, A. Bronstein, R. Talmon, and R. Kimmel, "Deep Isometric Maps," *Image and Vision Computing,* vol. 123, p. 104461, 2022/07/01/ 2022.

[31] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in) valid?," *Information Sciences,* vol. 581, pp. 620-636, 2021/12/01/ 2021.

[32] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing,* vol. 267, pp. 664-681, 2017/12/06/ 2017.

[33] Q. Xu, Q. Zhang, J. Liu, and B. Luo, "Efficient synthetical clustering validity indexes for hierarchical clustering," *Expert Systems with Applications,* vol. 151, p. 113367, 2020/08/01/ 2020.

[34] K. P. Sinaga and M. S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access,* vol. 8, pp. 80716-80727, 2020.

[35] E. Schubert and P. J. Rousseeuw, "Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms," Cham, 2019: Springer International Publishing, pp. 171-187.

[36] V. Cohen-addad, V. Kanade, F. Mallmann-trenn, and C. Mathieu, "Hierarchical Clustering: Objective Functions and Algorithms," *J. ACM,* vol. 66, no. 4, p. Article 26, 2019.

[37] F. E. Öztürk and N. Demirel, "Comparison of the methods to determine optimal number of cluster," vol. 6, no. 1, 2023.

[38] F. Liu and Y. Deng, "Determine the Number of Unknown Targets in Open World Based on Elbow Method," *IEEE Transactions on Fuzzy Systems,* vol. 29, no. 5, pp. 986-995, 2021.

[39] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients," *Pattern Recognition,* vol. 135, p. 109144, 2023/03/01/ 2023.

[40] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing,* vol. 528, pp. 178-199, 2023/04/01/ 2023.

[41] C.-E. Ben Ncir, A. Hamza, and W. Bouaguel, "Parallel and scalable Dunn Index for the validation of big data clusters," *Parallel Computing,* vol. 102, p. 102751, 2021/05/01/ 2021.

[42] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161-168.

[43] O. I. Abiodun *et al.*, "Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition," *IEEE Access,* vol. 7, pp. 158820-158846, 2019.

[44] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 2, pp. 612-619, 2020.

[45] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, Cham, 2019: Springer International Publishing, pp. 758-763.

[46] Y. Zhang *et al.*, "Research and Application of AdaBoost Algorithm Based on SVM," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 2019, pp. 662-666.

[47] T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2,* vol. 1, no. 4, pp. 1-4, 2015.

[48] J. Singh and R. Banerjee, "A Study on Single and Multi-layer Perceptron Neural Network," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 35-40.

[49] X. Li and C. A. Micchelli, "Approximation by radial bases and neural networks," *Numerical Algorithms,* vol. 25, no. 1, pp. 241-262, 2000/09/01 2000.

[50] S. Mittal, "A survey on modeling and improving reliability of DNN algorithms and accelerators," *Journal of Systems Architecture,* vol. 104, p. 101689, 2020/03/01/ 2020.

[51] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams," *Big Data and Cognitive Computing,* vol. 5, no. 1, p. 1, 2021.

[52] G. S. Handelman *et al.*, "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods," *American Journal of Roentgenology,* vol. 212, no. 1, pp. 38-43, 2019.

Parte II. Artículos Seleccionados

Capítulo II. Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review

Autores: Anita Herrera[1], Ángel Arroyo[1], Alfredo Jiménez[2] and Álvaro Herrero[1]

Afiliaciones:

[1] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela Politécnica Superior, Universidad de Burgos, Av.Cantabria s/n, 09006, Burgos, Spain, ahv1002@alu.ubu.es, {aarroyop, ahcosio}@ubu.es
[2] KEDGE Business School, 680 cours de la Liberation, Talence (Bordeaux) France. alfredo.jimenez@kedgebs.com

Resumen

La adopción de tecnologías en el turismo ha permitido mejorar sus servicios y orientarlo hacia una industria más personalizada e innovadora. Este estudio presenta una revisión bibliográfica de artículos científicos relacionados con el uso de la Inteligencia Artificial en el turismo, como una descripción de diferentes aplicaciones que contribuyen en el desarrollo del sector.

El estudio realiza una Revisión Sistémica de la Literatura (SLR) para identificar y analizar los artículos relevantes sobre la IA en el turismo, cuya fecha de publicación estén comprendidas entre los años 2016 y 2023. En la búsqueda se utilizaron combinaciones de palabras claves como Artificial Neuronal Network (ANN), Machine Learning (ML), Gradient Booster Regression Trees (GBRT), Random Forest, Naïve Bayes, Support Vector Machine (SVM), N-gram models, regression, clustering, tourism industry, tourism management, tourism innovation, entre otros.

Los trabajos seleccionados para el análisis, están desarrollados en inglés, publicados en revistas científicas de impacto, con un número de citas igual o superior a 8 (promedio de citas por artículo en los años de estudio, según Citation Report en Web of Science) y publicados en revistas indexadas relacionadas con las áreas de IA y turismo. El análisis bibliométrico de la producción científica identificada se realizó con la herramienta Bibliometric de R, y los resultados se visualizaron mediante diversas gráficas, revelando el volumen de producción científica en cada año del período analizado.

Como introducción al análisis de los documentos seleccionados, se revisa la evolución de las tecnologías de la información y comunicación en el turismo y su aporte desde la década de 1960, con la creación del primer sistema informático para reservas aéreas. A continuación, los estudios relacionados con la IA aplicada al turismo, se organizan en las siguientes categorías, definidas para abordar diferentes aspectos de este sector:

- Pronósticos en el sector turístico: Los pronósticos permiten anticipar las tendencias y comportamientos, proporcionando una base para la toma de decisiones estratégicas. Las investigaciones que analizan los modelos de IA utilizados en los pronósticos del sector turísticos, destacan la importancia de estos para disminuir la incertidumbre en la gestión y controlar los

inconvenientes generados por un número de arribos de turistas superior a la capacidad del destino. A su vez, los pronósticos de la demanda turística representan una fuente de información válida para tomar decisiones relacionadas con la administración de personal, estrategias de precios, marketing, desarrollo de productos y servicios, entre otros aspectos estratégicos.

El porcentaje de la producción científica sobre el uso de IA para los pronósticos en turismo (en el período de estudio), decreció en el año 2018 y 2020. Durante la pandemia, la operación de la industria turística cambió considerablemente y por ende fue necesario analizar un escenario diferente para determinar las variables a ser consideradas en el desarrollo de nuevos modelos para los pronósticos en esta industria.

- Planificación de viajes: Las técnicas de IA aplicadas a esta área, se enfocan en una planificación personalizada, garantizando experiencias más satisfactorias para los viajeros. Recientes investigaciones presentan el uso de IA para las sugerencias de viajes en grupos de turistas más grandes y la incorporación de variables como el tráfico, clima, transporte, así como el horario de atención, popularidad y tiempo de espera en las atracciones turísticas.

  La producción científica sobre la planificación de viajes experimenta un crecimiento más lento en el período 2019-2022, posiblemente vinculado a una disminución en el uso de estas plataformas debido a la reducción de operaciones en la industria turística.
- Marketing de los destinos turísticos: Busca promover de manera efectiva la captación de la atención de los clientes objetivo. En este contexto, destacan las investigaciones relacionadas con el metaverso, al ofrecer la oportunidad de experimentar un destino turístico sin viajar físicamente. Potenciales turistas pueden explorar un destino, sus instalaciones y servicios antes de decidir viajar. El metaverso es una herramienta importante para el marketing turístico y la planificación de viajes.

  La producción científica relacionada con este tema, decreció en el año 2020 lo que podría reflejar el replanteamiento del marketing turístico para un nuevo escenario de la industria turística.

- Análisis de los comentarios registrados en línea: El crecimiento acelerado en el uso de las plataformas en línea para el registro de comentarios relacionados con los viajes, genera una fuente de datos significativa, cuyo análisis brinda una ventaja competitiva en la gestión turística. La evaluación de la percepción de los turistas representa una valiosa oportunidad para los tomadores de decisiones, al adquirir conocimiento del mercado para mejorar los servicios e impulsar la innovación.

  Durante el período de estudio, la producción científica relacionada con la aplicación de la Inteligencia Artificial para analizar el registro de comentarios en línea ha experimentado un constante incremento, lo que refleja la importancia de aplicar estas técnicas para identificar las nuevas necesidades turísticas, principalmente durante la reactivación de la industria turística.
- Chatbots turísticos: facilitan la interacción en tiempo real, proporcionando asistencia instantánea y personalizada. Los estudios enfocados en la aplicación de la IA en los Chatbots turísticos, destacan los beneficios del uso de esta herramienta, al facilitar una interacción continua que mejora la comunicación entre el proveedor del servicio y los turistas.

  Las investigaciones analizadas en la presente revisión, resaltan los beneficios de los Chatbots que identifican emociones para ofrecer respuestas personalizadas de los servicios turísticos. Estos Chatbots utilizan algoritmos de procesamiento de lenguaje natural para analizar el texto de la conversación e identificar emociones. Así también, los Chatbots que usan algoritmos de reconocimiento de voz, tienen la capacidad de identificar las emociones según el tono de voz, ofreciendo respuestas aún más precisas.

La producción científica relacionada con el uso de la IA en los Chatbots turísticos ha experimentado un crecimiento constante durante el período de estudio, impulsado principalmente por el volumen de datos generados en la continua interacción con los clientes para atender las consultas.

La revisión de literatura realizada en este trabajo, identifica aplicaciones de IA importantes para la reactivación de la industria turística, especialmente para aquellos países cuya economía depende de este sector. La gestión pública y privada de turismo podrían beneficiarse de la IA para formular estrategias y tomar decisiones basadas en información.

Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review

Anita Herrera
(Applied Computational Intelligence Group (GICAP), Department of Digitalization,
Polytechnic School, University of Burgos, Av. Cantabria s/n, 09006, Burgos, Spain,
https://orcid.org/0000-0002-2655-412X, ahv1002@alu.ubu.es)

Ángel Arroyo
(Applied Computational Intelligence Group (GICAP), Department of Digitalization,
Polytechnic School, University of Burgos, Av. Cantabria s/n, 09006, Burgos, Spain,
https://orcid.org/0000-0002-3561-6257, aarroyop@ubu.es)

Alfredo Jiménez
(KEDGE Business School, 680 cours de la Liberation, Talence, Bordeaux, France,
https://orcid.org/0000-0001-7811-5113, alfredo.jimenez@kedgebs.com)

Álvaro Herrero
(Applied Computational Intelligence Group (GICAP), Department of Digitalization,
Polytechnic School, University of Burgos, Av. Cantabria s/n, 09006, Burgos, Spain,
https://orcid.org/0000-0002-2444-5384, ahcosio@ubu.es)

Abstract: The analysis of Artificial Intelligence techniques and models used in the tourism sector provides insightful information for the management and innovation of this industry. In this paper, we conduct a comprehensive review of the different techniques and models, in regards to Artificial Intelligence when applied to the tourism industry. Specifically, we present a categorization of Artificial Intelligence applications used in different areas of tourism. The results allow to recognize valid studies and useful tools for the activation and growth of the tourism sector, an industry that represents a significant increase in the Gross Domestic Product of various economies and supports the development of life conditions for their inhabitants. Artificial Intelligence applications generate more personalized travel experiences, improve the efficiency of tourism services and strengthen the tourism competitiveness of the destination.

Keywords: Artificial Intelligence, Machine Learning, tourism management, tourism innovation, tourism forecast, hospitality industry
Categories: G 3, H 2, I 2, I 7, J 4
DOI: 10.3897/jucs.101550

1 Introduction

Growth in the tourism industry represents an opportunity and a challenge in the economy and development of various locations, thanks to investments and creation quality job opportunities [Calero and Turner, 2020]. This industry is characterized by fast and deep changes that in recent years have included new technologies, and a preference for tourism based on experiences and nature, which have played a critical role in the reactivation of this industry after the pandemic generated by Covid-19 [UNWTO, 2023].

Innovation is, therefore, a fundamental key in the tourism industry, by satisfying the needs and expectations of tourists, in a process of permanent interaction. Success in tourism management lies in detecting changes in the customer's preferences and responding to them by reviewing and adapting the services, according to new requirements [Buhalis and Cooper, 2022]. According to the World Travel & Tourism Council, the travel and tourism sector contributed 10.3% to the global GDP in 2019; percentage that decreased to 5.3% in 2020, increased to 6.1% in 2021, and increased to 7.6% in 2022, this due to mobility restrictions given by the pandemic. Also, in 2019, the tourism industry generated 330 million jobs, while in 2020, 62 million jobs were lost, in 2021, 18.2 million jobs were recovered, and in 2022, 22 million jobs were recovered, changes generated during the pandemic and subsequent reactivation process [WTTC, 2022, WTTC, 2023].

In this context, the development of smart systems and Internet communications represent a significant opportunity for the tourism sector, due to travelers collecting and exchanging information and the contribution that this activity represents to the location's economy. This is where Artificial Intelligence (AI) plays a fundamental role, as a new stage in the tourism industry, with better opportunities for tourism service providers.

The objective of this paper is to contribute with a description of the different AI techniques that have paved the way for innovation and development of the tourism sector, considering that the consulted literature identifies reviews that focus on a specific AI technique and on a specific tourist activity (accommodation, transportation, food and drinks, etc.).

The remainder of this paper is structured as follows: Section 2, *Methodology* describes the criteria for selection and review of papers. In Section 3, *Review findings and discussions* pertains to a review on *Information and Communication Technologies in tourism*, based on the evolution of these tools and the support they have provided to the sector since the 1960s. In addition, different literature reviews on this topic, are presented. Furthermore, *Artificial Intelligence in tourism* is reviewed and analyzed, highlighting the impact of smart systems on innovation in the tourism industry, as well as their importance in the decision-making process. This analysis has been categorized according to the areas of tourism management. Finally, Section 4 includes the *Conclusions and Future Research* of the present work.

2 Methodology

The development of the present work considered a Systemic Literature Review (SLR) to identify and examine scientific articles related to AI in the tourism sector, summarizing the most relevant and current research. The SLR includes a valid, reliable, and repeatable protocol that helps to determine existing research and recognize information gaps to be explored. For the development of this work, three stages in the SLR are considered: planning the review, conducting it, and reviewing the results, as suggested by [Tranfield *et al*., 2003].

2.1 Planning the Review

The work focused on the search for applications of AI techniques important for market intelligence to improve the competitiveness of the tourism sector. The keywords used in this search were: Artificial Neuronal Network (ANN), Machine Learning (ML), Gradient Booster Regression Trees (GBRT), Random Forest, Naïve Bayes, Support Vector Machine (SVM), and N-gram models, regression, clustering, tourism industry, tourism management, tourism innovation, etc., using combinations of keywords related to AI and tourism, for each search. The papers identified were analyzed to confirm that they corresponded to the subject of the review.

2.2 The review

The initial search was carried out in Google Scholar, a database that offers results from important journals such as Elsevier, IEEE, Springer, etc. The queries used combinations of the previously mentioned keywords and the following search criteria:

- Articles developed in English and published in scientific journals of impact.
- Date of publication in the period between 2016 and 2023.
- Scientific papers with the highest number of citations or reviews, representing the highest impact in their scientific area. We limited the analysis to articles with at least 8 citations, except for articles published in 2022 and 2023, in which a minimum number of citations is not considered. The minimum number of citations was established according to the average citation per item (topic "artificial intelligence" AND “tourism”; in the study period), according to Citation Report in Web of Science. https://www.webofscience.com/wos/woscc/citation-report/115f8ad4-9ec6-4f3b-8b72-8356436977a4-8759b02c
- Publications in indexed journals corresponding to the areas of AI and tourism.

In addition, the Bibliometrix package of R was used to analyze the results of the Web of Science search to obtain data on the scientific literature.

2.3 Review results

The information search reflects the increase in the number of researches related to AI applications in tourism during the study period. This increase indicates the researchers' interest in analyzing the influence of AI on the growth of the tourism industry whose development is based on innovation and appropriation of technology.

3 Review findings and discussions

The literature presents different definitions of Information and Communication Technologies - ICT - according to the area of study which this concept is associated. However, all these concepts consider that ICT are technologies that provide access to information through telecommunications, which include the Internet, wireless networks, and cellular phones, among others.

The tourism industry has maintained for long a close relationship with ICT, since the creation of the first computerized airline reservation system in 1960. The Global Distribution System (GDS) in the 1980s, and the creation of the Internet in the '90s, are milestones that substantially modified the operation and strategy practices of the tourism sector [Leung, 2020].

ICTs have changed the way in which travelers' access information, plan, book tourism services, and share their experiences. The platforms that provide these services also collect data related to their preferences, age, nationality, etc. Those responsible for tourism management can make use of such data, even those in small towns that did not have this information before, in order to develop effective strategies to promote the tourist destinations in an effective and smart way. Therefore, how this data is used offers an extraordinary opportunity due to its ability to provide answers to practically any question that may be asked about the behaviors, opinions, and feelings of tourists [Mariani *et al.*, 2016].

The contribution that ICT has had in the tourism sector has grown and evolved thanks to new technologies such as the Internet of Things [Nitti *et al.*, 2017], smartphones [Kang *et al.*, 2020], portable devices [Castañeda *et al.*, 2019], new connectivity [Liang *et al.*, 2017], and Big Data [Demunter, 2017] among others. These concepts have increased the interest of several researchers, who now are trying to understand how technology helps in the search for travel information and in the decision-making process in order to develop better tools in a scenario in which information will continue to be an important and exciting topic [Xiang, 2018].

The Table 1 summarizes the most interesting review articles on ICT in tourism, as a topic of interest in different investigations. The Table 1 includes authors, review period, number of articles that were analyzed, as well as the area of knowledge to which it belongs.

Author / Year	**Period**	**Articles Reviewed**	**Area**
[Singh, 2015]	1981 – 2012	182	ICT Application in Tourism
[Leung *et al.*, 2015]	1996 – 2013	331	Marketing in Tourism
[Pesonen, 2013]	2000 – 2011	188	ICT and Market Segment in Tourism
[Leung *et al.*, 2013]	2007 – 2011	44	Social Media in Tourism
[Zeng and Gerritsen, 2014]	2007 – 2013	279	Social Media in Tourism
[Law *et al.*, 2014]	2009 – 2013	107	ICT Application in Tourism
[Khatri, 2019]	2009 – 2018	63	ICT Application in Tourism
[Molina-Collado *et al.*, 2022]	1988-2021	2424	ICT Application in Tourism
[Law *et al.*, 2019]	2014 – 2017	288	ICT Application in Tourism
[Marasco *et al.*, 2018]	- 2017	79	Innovation in Tourism
[Han and Bai, 2022]	2010-2019	575	Marketing in Tourism
[Verma *et al.*, 2022]	2000-2021	1652	ICT Application in Tourism

Table 1: Previously Reviewed Articles - ICT in Tourism

These publications, among others, reflect an increase in scientific articles focused on ICT in Tourism, as a result of the interest created by the incorporation of new technologies such as AI, which brings us closer to programs that simulate human intelligence.

Through the years, the concept of intelligence has been researched in various investigations, identifying it as the ability or faculty to understand, reason, know, learn and solve problems. In a manner of speaking, the concept of AI has also been the subject of different definitions, highlighting American scientist John McCarthy's research, who at the Dartmouth Conference in 1956 coined this term as *"the science and ingenuity of creating intelligent machines, specifically, intelligent computer programs"* [McCarthy, 2007].

As AI obtains greater applications with advanced algorithms and improved computing and storage capacity; it becomes an integrating element of digital systems and, more specifically, has a profound impact on human decision-making processes [Duan *et al*., 2019]. Smart systems work on areas such as knowledge representation [Ahmed *et al*., 2019], reasoning [Georgeff and Ingrand, 1989], ML [Hutter *et al*., 2019], perception in Natural Language Processing (NLP) [Young *et al*., 2018], as well as facial recognition [Chen and Jenkins, 2017]. Specifically, in the travel and tourism industry, smart systems provide innovations in recommendation systems [Abbasi-Moud *et al*., 2021, Thiengburanathum *et al*., 2016], emotional computing, group decisions, social networks and analysis, allowing them to design and experiment on business models, model-making of user decisions and use analysis, with the predominance of ICT-based platforms such as Airbnb, Uber and Online Travel Agencies (OTA) [Neidhardt and Werthner, 2018].

The tourism industry applies AI in OTA, air traffic controllers, hotel chains, tour operators, a sector that generates and uses large amounts of information with results that set a competitive advantage by being able to anticipate, predict and proactively satisfy customer' needs [Romero Dexeus, 2019]. The search in Web of Science with the keywords AI and tourism, between 2016 and 2022, was analyzed with Bibliometrix package of the R programming language, displaying the information in Fig. 1.

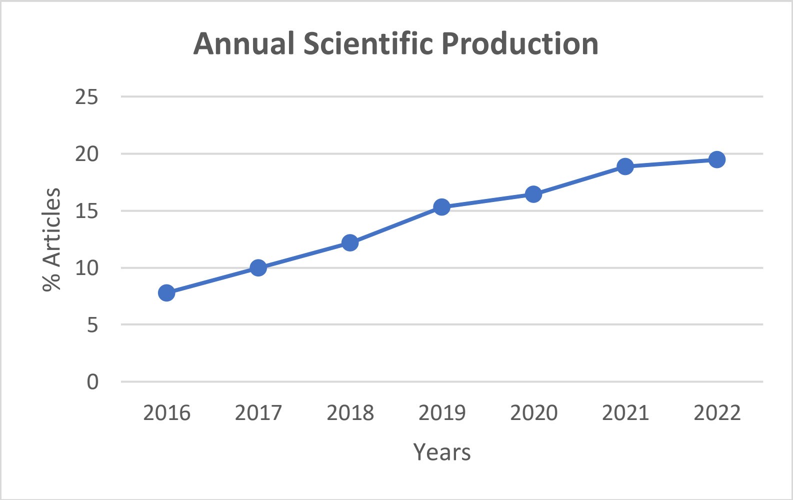

Figure 1: Annual Scientific Production
Keywords: Artificial Intelligence and tourism
Range of years: 2016 to 2022

Fig. 1 shows the scientific production between 2016 to 2022, related to AI in tourism, which shows the percentage of scientific production for each year, in relation to the total number of productions during the period 2016 - 2022. The scientific production in 2022 is higher than in previous years, which demonstrates the importance of the research topic.

In general, it can be seen that the number of works on AI in different applications for tourism, published between 2016 to 2022, maintains a growing trend, which coincides with the period of reactivation of the tourism industry (2020–2022), a space of analysis of the new preferences of tourists and reinvention of tourism services. The works developed have been of great importance for the generation of strategies in the tourism sector.

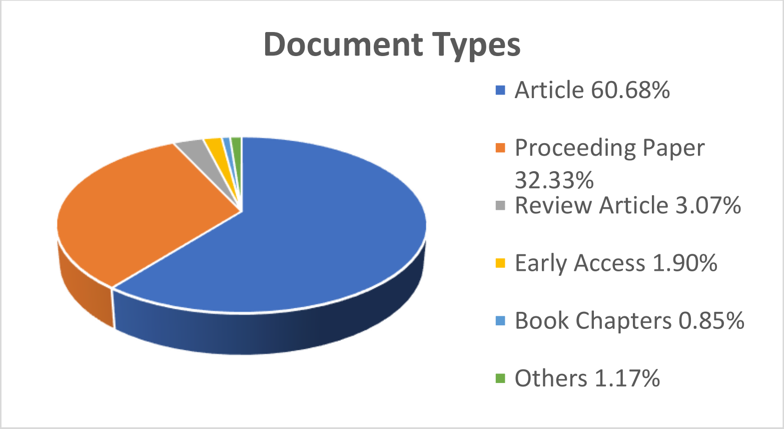

Figure 2: Document Types
Keywords: Artificial Intelligence and tourism
Range of years: 2016 to 2022

Fig. 2 shows the document types of scientific production between 2016 and 2022, related to AI in tourism. The highest percentage of scientific production corresponds to 60.68% of articles and 32.33% of proceedings papers.

Next, we will categorize the paper according to the following areas of study: 1. *Forecast in the Tourism Sector*, 2. *Travel Planning*, 3. *Tourist Destinations Marketing*, 4. *Analysis of Online Comment Log* and 5. *Touristic Chatbots*.

3.1 Forecast in the Tourism Sector

Being able to make accurate predictions about the behavior of tourism drastically reduces uncertainty, which will help us in the management of the different areas related to the tourism industry. Thus, AI-based forecasting has fomented great interest and focuses on models such as the fuzzy theory, gray theory, ANN, genetic algorithms, and expert systems [Wang, 2004], with algorithms that find user behavior from large databases generated with the online registration of users during searches, reservations, purchases and comments of tourist services on the Internet [Moro and Rita, 2016].

Tourism is perhaps one of the sectors that has benefited the most from ML, this forecasts tourist expenses, determines profiles, and predicts the number of arrivals. In the research by [Cankurt and Subasi, 2015] the series of tourist demand in Turkey in a period between 1996 and 2013 is analyzed. It is imperative that ML-based predictive models consider the different periods of the year when demand varies significantly, in order to make predictions that are more in line with the reality of the sector at the time. This paper uses the Principal Components Analysis (PCA) to decorrelate the input data, Back-Propagation Neural Network Architecture (BPNN), and the adaptive differential evolution algorithm (ADE).

By obtaining forecasts through ML and Internet search indexes, it is possible to predict tourist arrivals to a certain destination, performance that can be compared with the results obtained by other tools such as Google and Baidu [Sun *et al*., 2019]. In the research by [Li *et al*., 2018a] algorithms are created to forecast tourist arrivals. The objective of the research is to forecast the number of tourists arriving in Beijing in order to control the inconvenience generated by the arrival of a number of tourists exceeding the capacity of the destination. For this purpose, algorithms are designed using data on arrivals in Beijing in the period from January 2011 to December 2016. Using PCA, dimensionality is reduced. A hybrid model is then created using the ADE algorithm and the BPNN model. Comparing the results obtained with other techniques, it is identified that this model increases the prediction accuracy.

On the other hand, Gradient Boosted Regression Trees, Ranking SVM and ML techniques produce predictive models that when applied in tourism management allow different types of prediction, such as the following location that a tourist will have according to their travel history. Specifically, the technique called Gradient Boosted Regression is used for classification problems and model generation for predictions based on decision branches, while the Ranking SVM function is a method for learning to classify in the extraction of key phrases process [Muntean *et al*., 2015].

Also, research shows that the use of ANN is perfectly applicable to forecast tourism demand. ANN are usually applied for pattern recognition due to their association, memory, storage, and learning function [Li *et al*., 2009]. For example, research conducted by [Folgieri *et al*., 2017] evidences the importance of ANN in predicting tourist information. For the study, monthly data on tourist arrivals in Croatia were collected for the period from January 1, 2007, to December 31, 2012. The results indicate that ANN is a robust method in predicting tourist arrivals, a method that outperforms linear regression.

The forecast of tourist demand represents a valid source of information for decision-making related to personnel, capacity, resource management and pricing strategies. Therefore, a more accurate forecast will be a lot more useful to create strategies in tourism management [Jiao and Chen, 2019]. Tourism demand forecasting is an important element in the efficient planning and allocation of resources in the tourism industry. [Li and Jiao, 2020] conducts a literature review on tourism demand forecasting from 1960 to 2018, showing the relevance of this research topic.

The following figure shows the annual scientific production of Forecast in the Tourism Sector:

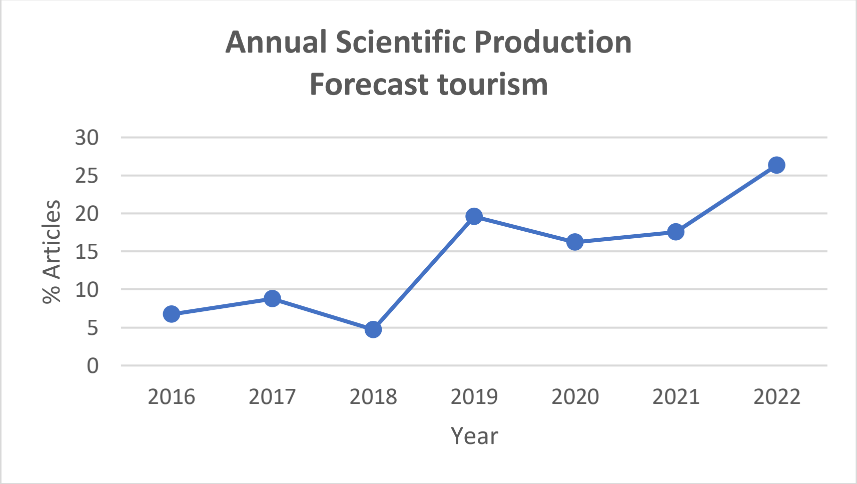

Figure 3: Annual Scientific Production
Keywords: Forecast tourism and Artificial Intelligence
Range of years: 2016 to 2022

Fig. 3 shows the percentage of scientific production using Artificial Intelligence in forecasts in tourism, for each year in relation to the total number of productions during the period 2016 – 2022. Fig. 3 shows that scientific production related to tourism forecasting decreased in the year 2018 and also in 2020. One possible explanation for the decrease in 2020 is that, during the pandemic, the tourism industry operation changed considerably due to mobility restrictions. In this sense, it was necessary to analyze the new scenario and determine the variables to be considered for the development of forecast models for the tourism industry.

3.2 Travel Planning

AI-based travel planning systems provide orderly and precisely categorized information to organize the trip, providing a personalized plan taking into account tourist's preferences. These systems take into account the user's demands and parameters such as the cost of the trip, traffic volume, weather and the time of the trip in the tourist destination [Dezfouli *et al*., 2018].

AI-based travel planning allows tourists a personalized and advanced offer of services, always looking for the best options, taking into account tastes and preferences. The travel recommendation systems use the haversine algorithm and Traveling Salesman Problem (TSP), they consider the user's preferences, as well as the traffic, weather, recommendations, transportation, etc. of a particular city. In addition, a recommendation system can consider the popularity of tourist attractions and possible waiting times at these attractions as part of personalized itinerary recommendations. Information about the suggested tourist sites is collected from websites and social networks through an AI retrieval process [Asaithambi *et al*., 2023].

Among the tools used for travel planning, clustering or grouping techniques [Saxena *et al*., 2017] stand out as unsupervised ML algorithms that represents an important issue to analyze. In the travel industry, the '*k*-means' clustering technique has several applications such as the algorithms used for travel itineraries by finding a complete route that connects all the nodes of a network, visiting each point once and minimizing the total travel time (TSP) [Rani *et al*., 2018].

[Ravi *et al*., 2019] proposed a hybrid travel recommendation system based on location, with swarm intelligence algorithms, which provides points of interest depending on the preferences and the users' needs, as registered on TripAdvisor. This system efficiently extracts, filter and present heterogeneous and geographically distributed tourist information from the Internet, such as the Intelligent Travel Planning (ITP), a travel planning system, whose objective is to find different useful tourism solutions for the users of the system.

Other researches such as that conducted by [Jiang *et al*., 2013] propose a tourist recommendation system based on the information provided by geographically tagged photographs on the Internet. It identifies tourist attractions according to their geographical location and determines popularity based on the number of user photographs. The textual tags in an attraction are grouped as a document and analyzed using the vector space model. According to the textual and visual information of said photographs, one can determine the user's personal interests.

A recommendation system allows to know the best options for travel, given the large amount of information and offer of tourist services on the Internet. It is also important that these systems consider the restrictions established during the pandemic generated by Covid-19. In this sense, the work of [Nilashi *et al*., 2021] presents travel recommendation based on social networks, the system stores updated information about Covid-19 and the tourist's recent trips, which is analyzed with ML techniques.

Some recommender systems analyze the Vacation Planning Problem (VPP), which is related to the design of tourist trips in a large geographical area, providing daily and personalized tourist routes, intermediate destinations along the trip and places of accommodation. The study presented by [Vathis *et al*., 2023], presents a heuristic model for the VPP. The model is based on the clustering of points of interest according to the distance of the points as well as the available transportation between them. This type of algorithm is also used in other areas such as search and rescue operations, clustering points of interest to locate or save people.

Several recommendation systems for tourism use Multi-Agent Systems (MAS) to improve the recommendation process. In the work done by [Alves *et al*., 2022], they present the modeling of a mobile Group Recommendation System (GRS) using the personality of tourists to predict the preference for tourist attractions, as well as intelligent agents with microservices, achieving to provide better group recommendations.

The following figure shows the scientific production related to Travel Planning:

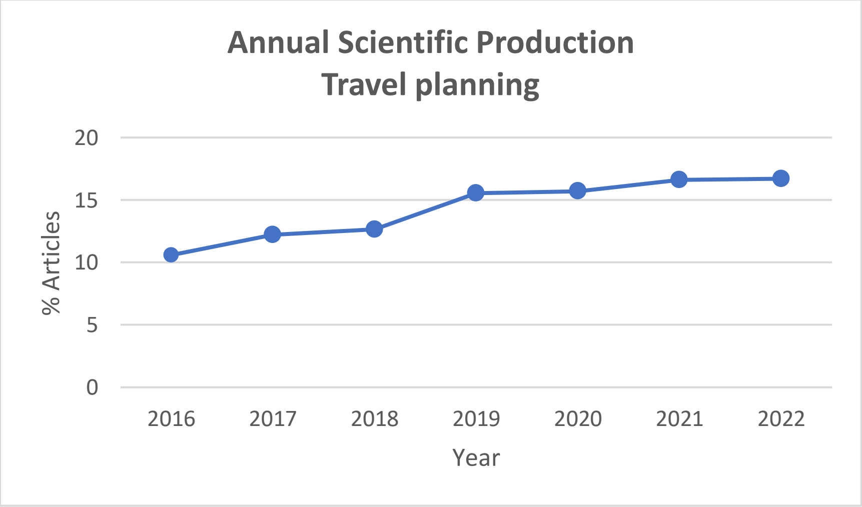

Figure 4: Annual Scientific Production
Keywords: Travel planning and Artificial Intelligence in tourism
Range of years: 2016 to 2022

Fig. 4 reflects a slow growth of the scientific production related to travel planning platforms, during the period from 2019 to 2022, compared to previous years. This could be related to the decrease in the use of travel recommendation platforms, and the identification of new requirements and preferences of tourists, according to the mobility and security restrictions generated by the pandemic.

3.3 Tourist Destinations Marketing

Marketing in the tourism industry was known for face-to-face interactions between service providers and tourists, which changed with the creation of platforms offering a wide range of tourism services. Currently, OTAs present various travel options and suggestions for tourists, where tourism providers contact potential travelers via the Internet, supporting to the evolution of this sector [Romero Dexeus, 2019]. Smart systems are responsible for defining prices in airlines and in the hotel industry according to the Revenue Management, they are systems that predict consumer behavior and aim at maximizing revenue growth.

Tourist Destinations Marketing is an area that has benefited from advanced data analysis, ANN and knowledge representation technologies. [Stalidis *et al*., 2015] presented in their research a smart support system for this area, which uses as input the data collected with surveys which were done to tourists who visited Thessaloniki in Greece between May and October 2013. The smart system includes a method of data analysis that identifies multivariate nonlinear relationships, the decomposition of complex phenomena into factors, and the definition of the characteristics of population groups. A second method focused on ANN to make smart decisions in new cases based on what has been learned; and a third knowledge model method that expresses the results in an understandable way.

ANNs can evaluate the capacities to select products and services in an e-tourism environment. These models are precise in classifying products and services in perspective for consumers, taking into consideration that from the launch of the first tourism website, this area has been one of the fastest growing segments in e-commerce. It is valuable to incorporate behavioral factors in e-tourism systems, proposing a multi-agent e-tourism system architecture for product intermediation, supplier negotiation and evaluation, which is based on the supplier reputation agent and the use of ANNs [Cao and Schniederjans, 2006].

The article by [Pyo *et al.*, 2002] takes as a case study the stagnation in the growth of the number of tourists and their spending on the island of Cheju in South Korea. By analyzing the variables of a large database from the island's airport, it looks to improve tourism promotion by building a smart model that determines the visitors, that once have returned home, would recommend visiting the destination.

The AI technique ‘fuzzy theory’ [Sugeno and Kang, 1998], combined with ANN, can be used for population projections that incorporates concepts such as probability, optimization, simulation and precision, along with policy decision-making advertising and tourism marketing. The article by [Li, 2000] proposes the development of a hybrid system for the promotion of a tourist destination based on marketing strategies. The hybrid system includes the features of an expert system, fuzzy logic and a forecasting model with ANN.

It is also interesting to analyze the benefit of the metaverse in tourism marketing. The metaverse offers the opportunity to experience a tourist destination without physically traveling. A potential tourist can explore the destination, tourist facilities and services, prior to travel. The metaverse benefits tourism marketing, travel planning [Monaco and Sacchi, 2023], and the metaverse can support sustainable tourism development [Go and Kang, 2023].

Fig. 5 shows the annual scientific production in Tourism destinations marketing:

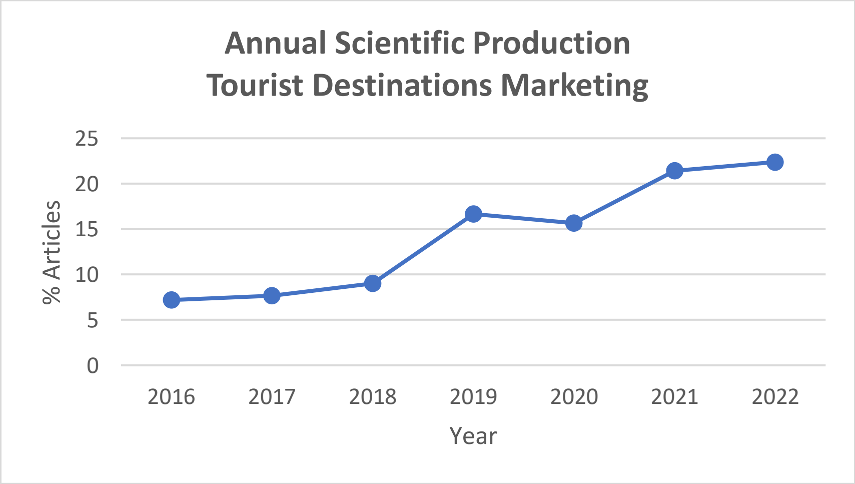

Figure 5: Annual Scientific Production
Keywords: Tourism Marketing and Artificial Intelligence
Range of years: 2016 to 2022

The scientific production in Tourist destinations marketing decreased in 2020, which could reflect the rethinking of tourism marketing, in agreement with new tourism services, as part of the revival of the tourism industry. The tourism sector was probably one of the most affected industries during the pandemic, with high economic losses, which limited the tourist destinations marketing.

3.4 Analysis of Online Comment Log

The analysis of comments registered in social networks and other online platforms (TripAdvisor, Booking, Kayak, Triage, Expedia, and Airbnb, among others) by means of AI techniques quantify visitor satisfaction towards the services provided in a tourist destination, making it possible to define tourists' requirements or preferences. This type of information is, in many cases, the basis for decision-making processes by potential tourists who organize their trip based on comment reviews made by other tourists.

Models such as Word2vec [Li *et al*., 2018b] and the Random Forest algorithm [Kurnia *et al*., 2020] can be used to classify user post-trip comments made on social networks or platforms designed for this particular purpose. Comments can be analyzed with deep learning models and NLP techniques [Ku *et al.*, 2019]. This information is a very valuable opportunity for decision-makers which want to improve services, gain knowledge about the market and innovate.

In reality, platform accelerated growth used to record trip-related personal comments has increased the amount of information now available for use by other tourists, or by those responsible for tourism management as an indicator of the topics that need improvement. The research conducted by [Ye *et al*., 2009] analyzes sentiment classification techniques of reviews recorded in travel blogs for seven popular places in the United States and Europe, comparing Naïve Bayes, SVM and N-gram models. The experiment indicates that if the training datasets have a large number of reviews, the models achieve an accuracy of at least 80%.

The development of technology has created new channels for the generation and consultation of information in the tourism sector. Currently, tourism service platforms generate a large volume of data, which are processed with different methods to provide quality information to decision-makers. In this context, opinion mining based on sentiment orientation allows interpreting tourist perceptions [Alaei *et al*., 2019].

AI techniques make it possible to derive insights from the analysis of multiple opinions, reviews and ratings of tourism services, recorded on social networks. This analysis helps to identify opportunities and represents a competitive advantage for decision makers, a topic that has generated the interest of different researchers [Chen *et al*., 2017, Gan *et al*., 2017, Geetha *et al*., 2017, Gitto and Mancuso, 2017, Kim *et al*., 2017, Lee *et al*., 2017, Luo *et al*., 2020, Ma *et al*., 2018, Ren and Hong, 2017].

In [Gyódi, 2022, Nilashi *et al.*, 2017, Silva *et al*., 2022, Sun *et al*., 2022, Yang and Han., 2021] effects of the COVID-19 pandemic on the hotel sector and new travelers' demands are identified through sentiment analysis. The studies consider the comments registered by tourists in an online platform, during and before the pandemic. These studies provide important information for the reactivation of the tourism industry [Gyódi, 2022].

Fig. 6 shows the annual scientific production on the analysis of online comment:

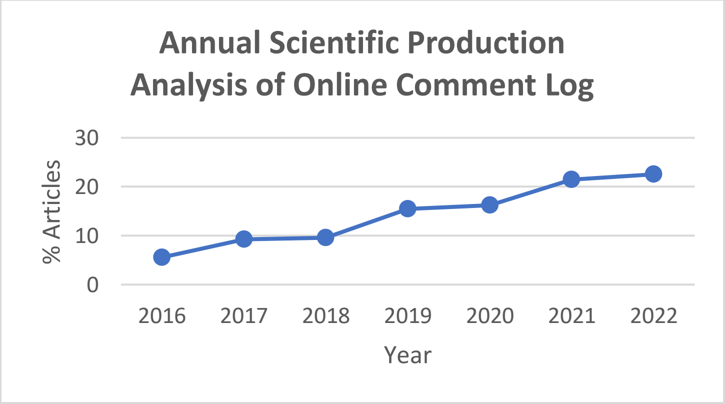

Figure 6: Annual Scientific Production – Analysis of Online Comment Log
Keywords: Text mining and sentiment analysis in tourism
Range of years: 2016 to 2022

Fig. 6 presents the increase in scientific production related to Artificial Intelligence in the analysis of online comment log, which would be related to the relevance of the analysis of comments to identify new tourism needs, mainly during the reactivation of the tourism industry.

3.5 Touristic Chatbots

Another form of application when it comes to AI that improves communication between service provider industries and their clients are AI-operated chatbots, which are intelligent solutions used by companies to provide a personalized service. In the tourism sector, hotels, airlines, tour operators, and others use chatbots to provide a better service to their customers.

A study done by [Zsarnoczky, 2017] focused on using chatbots to answer more than 300 questions related to accommodation, travel options, transportation, as well as available programs at the Matra Resort in Hungary. The results of said research indicated that tourists use chatbots for travel planning, finding them convenient to get information.

The research conducted by [Sano *et al.*, 2018], analyzes the generation of a knowledge base of tourist sites in the cities of Malang and Batu, by means of the hierarchical clustering algorithm AGNES (Agglomerative Nesting). The chatbot fed with this database provides information about the tourist sites that can be optimally visited by tourists with short stay times.

[Alotaibi *et al.*, 2020] presents a case study of the Smart Guidance chatbot that provides information about the city of Jeddah in Saudi Arabia for travel planning. The text received by the chatbot is analyzed in the NLP engine, determines keywords and answers the user's query.

Also, the paper presented by [Suanpang and Jamjuntret, 2021] presents the development of a chatbot that provides information to tourists interested in visiting the Active Beach area in Thailand. The design uses a Deep Learning model and demonstrates the usefulness of chatbots to improve the stay of tourists in different destinations.

The systemic review of literature elaborated by [Calvaresi *et al*., 2021], presents studies of the services provided by chatbots in hotels, airlines and travel agencies, highlighting their contribution to the competitiveness of the tourism industry, due to the speed of response to customers and the wide availability. The role of chatbots in the tourism industry generates the interest of different researchers [Li *et al.*, 2021, Melián-González *et al*., 2021, Mohamad Suhaili *et al*., 2021].

Chatbots or virtual assistants are implemented in the tourism sector to improve communication, user experience and empathy. The tourism sector uses travel chatbots, voice-based chatbots and emotion-based chatbots [Doborjeh *et al*., 2022]. An interesting implementation is the emotion chatbot, which can give a more personalized response. The emotion chatbot uses natural language processing algorithms to analyze the conversation and identify emotions, according to which it performs the best response [Lv *et al*., 2021].

The following figure shows the annual scientific production on Tourist Chatbots:

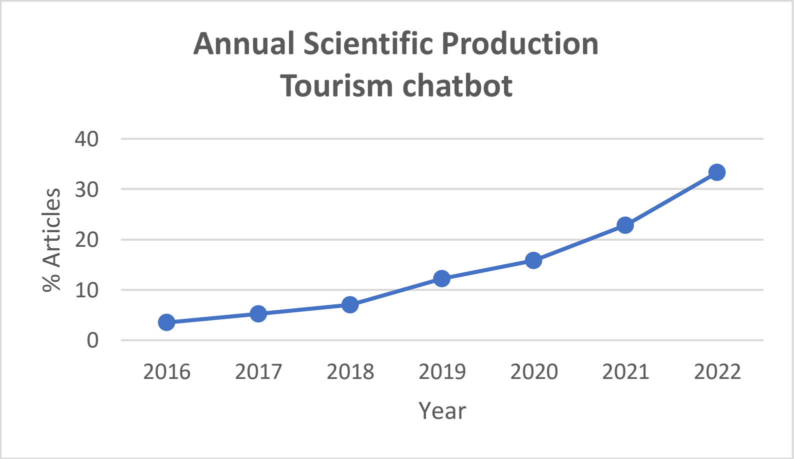

Figure 7: Annual Scientific Production
Keywords: Artificial Intelligence and tourism Chatbots
Range of years: 2016 to 2022

Fig. 7 shows a significant increase in the annual scientific production on tourism chatbots, which could be related to the importance of chatbots, mainly during the reactivation of the tourism industry. Chatbots allow to establish permanent communication with customers, resolving concerns and generating important data for management.

3.6 Other areas of study

As a field that is constantly advancing, its versatility to perform complex tasks makes it a valuable tool for the growth of various industries. The healthcare sector uses AI in: early disease detection, drug development, treatment personalization, epidemic outbreak prediction and so on [Schwalbe and Wahl, 2020]. Another important field of application is air quality and its influence on health, where AI makes it possible to develop forecasts to detect elevated concentrations of certain air pollutants so that measures can be taken to avoid negative effects [Méndez *et al*., 2023].

Industry has also benefited from AI through improved productivity, reduced operating costs, quality control and industrial robotics, among many other applications [Javaid *et al*., 2022]. In this context, the number of devices connecting to the Internet is increasing and the incorporation of AI into communication networks is beneficial. Also in communications in large cities, the AI methods can predict network traffic and improve network performance [Chen *et al*., 2021].

In the food field, AI enables crop management, pest control, crop forecasting, etc. Companies in the food sector also benefit from AI with sales forecasting to minimize expired products and avoid economic losses [Tsoumakas, 2019].

These are just some of the areas where AI has had a relevant impact.

4 Conclusions and Future Research

This article reviews the literature related to the contribution of AI in the tourism industry, an area characterized by its contribution to the economy and development of many countries' economies. In this context, the increased coverage and use of the Internet, social networks and the creation of smart systems demonstrate high levels of innovation in this sector.

In countries whose economy is committed to innovation and the development of the tourism industry, it is necessary for the public administration and the private sector to define joint strategies supported by AI, a tool that allows detecting trends and knowing the demands of tourists based on registration on platforms and social networks.

In the scientific production related to Artificial Intelligence applications in tourism, for the period from 2016 to 2022, it was possible to identify that 60.68% of the production corresponds to articles published in a journal and 32.33% were published in Conference Proceedings, being the types with the highest percentage, as shown in Fig 2. Review Article only represents 3.07%, and other types of documents less than 2%

The literature reviewed shows that ANN models are used to predict tourist arrivals, models that outperform linear regression techniques. The relevance of fuzzy theory models, gray theory, genetic algorithms, ML and expert systems was also identified, with results that support the development of the tourist offer of the localities.

Chatbots interact with tourists and allow the automation of different services. Deep learning models are used in their design and NLP engines identify keywords to answer user queries.

Large databases created from keywords that users register for search, reservations, purchases and from comments of tourist services on the Internet represent the main input when it comes to smart systems used in different areas of management for the tourism sector. This article reviews publications related to Forecasting in the Tourism Sector, a relevant topic due to the information it provides to those responsible for decision-making in the tourism sector, in aspects such as new investments or the forecast of logistical aspects. Smart systems for travel planning, tools used by travelers, who choose a destination and tourist services to be included in their next trip are also reviewed. Marketing of tourist destinations carried out by OTAs, platforms that hire different tourist services. Articles related to Tourism Marketing are also reviewed, smart systems that promote a tourist destination. Online records analysis provides information on the perception of tourists about the services received. Finally, chatbots, AI applications that are part of tourist services are also reviewed and considered.

Tourism is probably one of the areas most affected by the public health crisis created by the COVID-19 pandemic, having to reinvent itself while also attaining an economic reactivation. Therefore, it is the most opportune moment for science, through different research, to propose smart applications that accelerate the path towards the recovery of said industry.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[Abbasi-Moud *et al*., 2021] Z. Abbasi-Moud, H. Vahdat-Nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Systems with Applications,* vol. 167, p. 114324, 2021/04/01/ 2021.

[Ahmed *et al*., 2019] A. Ahmed *et al*., "Knowledge-Based Systems Survey," *International Journal of Academic Engineering Research (IJAER),* vol. 3, no. 7, pp. 1-22, 2019.

[Alaei *et al*., 2019] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment Analysis in Tourism: Capitalizing on Big Data," *Journal of Travel Research,* vol. 58, no. 2, pp. 175-191, 2019.

[Alotaibi *et al*., 2020] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamdi, "AI Chatbot for Tourist Recommendations: A Case Study in the City of Jeddah, Saudi Arabia," *International Association of Online Engineering,* 2020.

[Alves *et al*., 2022] P. Alves, D. Gomes, C. Rodrigues, J. Carneiro, P. Novais, and G. Marreiros, "Grouplanner: A Group Recommender System for Tourism with Multi-agent MicroServices," Cham, 2022: Springer International Publishing, pp. 454-460.

[Asaithambi *et al*., 2023] S. P. R. Asaithambi, R. Venkatraman, and S. Venkatraman, "A Thematic Travel Recommendation System Using an Augmented Big Data Analytical Model," *Technologies,* vol. 11, no. 1, p. 28, 2023.

[Buhalis and Cooper, 2022] D. Buhalis and C. Cooper, "Tourism Management," in *Encyclopedia of Tourism Management and Marketing*: Edward Elgar Publishing, 2022, pp. 441-444.

[Calero and Turner, 2020] C. Calero and L. W. Turner, "Regional economic development and tourism: A literature review to highlight future directions for regional tourism research," *Tourism Economics,* vol. 26, no. 1, pp. 3-26, 2020.

[Calvaresi *et al*., 2021] D. Calvaresi, A. Ibrahim, J.-P. Calbimonte, R. Schegg, E. Fragniere, and M. Schumacher, "The Evolution of Chatbots in Tourism: A Systematic Literature Review," Cham, 2021: Springer International Publishing, pp. 3-16.

[Cankurt and Subasi, 2015] S. Cankurt and A. Subasi, "Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components," *Balkan Journal of Electrical & Computer Engineering,* vol. 3, no. 1, pp. 42-49, 2015.

[Cao and Schniederjans, 2006] Q. Cao and M. J. Schniederjans, "Agent-mediated architecture for reputation-based electronic tourism systems: A neural network approach," *Information & Management,* vol. 43, no. 5, pp. 598-606, 2006/07/01/ 2006.

[Castañeda *et al*., 2019] J.-A. Castañeda, M.-J. Martínez-Heredia, and M.-Á. Rodríguez-Molina, "Explaining tourist behavioral loyalty toward mobile apps," *Journal of Hospitality and Tourism Technology,* vol. 10, no. 3, pp. 415-430, 2019.

[Chen and Jenkins, 2017] J. Chen and W. K. Jenkins, "Facial recognition with PCA and machine learning methods," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 973-976.

[Chen *et al*., 2017] F.-W. Chen, A. Guevara Plaza, and P. Alarcon Urbistondo, "Automatically extracting tourism-related opinion from Chinese social media," *Current Issues in Tourism,* vol. 20, no. 10, pp. 1070-1087, 2017.

[Chen *et al*., 2021] A. Chen, J. Law, and M. Aibin, "A Survey on Traffic Prediction Techniques Using Artificial Intelligence for Communication Networks," *Telecom,* vol. 2, no. 4, pp. 518-535, 2021.

[Demunter, 2017] C. Demunter, "Tourism statistics: early adopters of big data," *Publications Office of the European Union: Luxemburg,* 2017.

[Dezfouli *et al*., 2018] M. Dezfouli, M. Shahraki, and H. Zamani, "A Novel Tour Planning Model using Big Data," presented at the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, 2018. [Online]. Available: https://www.researchgate.net/publication/331422495_A_Novel_Tour_Planning_Model_using_Big_Data.

[Doborjeh *et al*., 2022] Z. Doborjeh, N. Hemmington, M. Doborjeh, and N. Kasabov, "Artificial intelligence: a systematic review of methods and applications in hospitality and tourism," *International Journal of Contemporary Hospitality Management,* vol. 34, no. 3, pp. 1154-1176, 2022.

[Duan *et al*., 2019] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *International Journal of Information Management,* vol. 48, pp. 63-71, 2019/10/01/ 2019.

[Folgieri *et al*., 2017] R. Folgieri, T. Baldigara, and M. Mamula, "Artificial neural networks-based econometric models for tourism demand forecasting," *Tourism in South East Europe,* vol. 4, pp. 169-182, 2017.

[Gan *et al*., 2017] Q. Gan, B. H. Ferns, Y. Yu, and L. Jin, "A text mining and multidimensional sentiment analysis of online restaurant reviews," *Journal of Quality Assurance in Hospitality & Tourism,* vol. 18, no. 4, pp. 465-492, 2017.

[Geetha *et al*., 2017] M. Geetha, P. Singha, and S. Sinha, "Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis," *Tourism Management,* vol. 61, pp. 43-54, 2017.

[Georgeff and Ingrand, 1989] M. Georgeff, P and F. Ingrand, "Decision-Making in an Embedded Reasoning System," in *International Joint Conference on Artificial Intelligence*, Detroit, United States, 1989, https://hal.laas.fr/hal-01980071/file/10.1.1.644.1321.pdf.

[Gitto and Mancuso, 2017] S. Gitto and P. Mancuso, "Improving airport services using sentiment analysis of the websites," *Tourism management perspectives,* vol. 22, pp. 132-136, 2017.

[Go and Kang, 2023] H. Go and M. Kang, "Metaverse tourism for sustainable tourism development: Tourism Agenda 2030," *Tourism Review,* vol. 78, no. 2, pp. 381-394, 2023.

[Gyódi, 2022] K. Gyódi, "Airbnb and hotels during COVID-19: different strategies to survive," *International Journal of Culture, Tourism and Hospitality Research,* vol. 16, no. 1, pp. 168-192, 2022.

[Han and Bai, 2022] W. Han and B. Bai, "Pricing research in hospitality and tourism and marketing literature: a systematic review and research agenda," *International Journal of Contemporary Hospitality Management,* vol. 34, no. 5, pp. 1717-1738, 2022.

[Hutter *et al*., 2019] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[Javaid *et al*., 2022] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Artificial intelligence applications for industry 4.0: A literature-based study," *Journal of Industrial Integration and Management,* vol. 7, no. 01, pp. 83-111, 2022.

[Jiang *et al*., 2013] K. Jiang, H. Yin, P. Wang, and N. Yu, "Learning from contextual information of geo-tagged web photos to rank personalized tourism attractions," *Neurocomputing,* vol. 119, pp. 17-25, 2013/11/07/ 2013.

[Jiao and Chen, 2019] E. X. Jiao and J. L. Chen, "Tourism forecasting: A review of methodological developments over the last decade," *Tourism Economics,* vol. 25, no. 3, pp. 469-492, 2019.

[Kang *et al*., 2020] S. Kang, L. W. Jodice, and W. C. Norman, "How do tourists search for tourism information via smartphone before and during their trip?," *Tourism Recreation Research,* vol. 45, no. 1, pp. 57-68, 2020/01/02 2020.

[Khatri, 2019] I. Khatri, "Information Technology in Tourism & Hospitality Industry: A Review of Ten Years' Publications," *Journal of Tourism and Hospitality Education,* vol. 9, pp. 74-87, 2019.

[Kim *et al*., 2017] K. Kim, O.-j. Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management," *Technological Forecasting and Social Change,* vol. 123, pp. 362-369, 2017.

[Ku *et al*., 2019] C. H. Ku, Y.-C. Chang, Y. Wang, C.-H. Chen, and S.-H. Hsiao, "Artificial intelligence and visual analytics: A deep-learning approach to analyze hotel reviews & responses," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[Kurnia *et al*., 2020] R. Kurnia, Y. Tangkuman, and A. Girsang, "Classification of user comment using word2vec and SVM classifier," *Int. J. Adv. Trends Comput. Sci. Eng,* vol. 9, no. 1, pp. 643-648, 2020.

[Law *et al*., 2014] R. Law, D. Buhalis, and C. Cobanoglu, "Progress on information and communication technologies in hospitality and tourism," *International Journal of Contemporary Hospitality Management,* 2014.

[Law *et al*., 2019] R. Law, D. Leung, and I. C. C. Chan, "Progression and development of information and communication technology research in hospitality and tourism," *International Journal of Contemporary Hospitality Management,* 2019.

[Lee *et al*., 2017] M. Lee, M. Jeong, and J. Lee, "Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website," *International Journal of Contemporary Hospitality Management,* 2017.

[Leung *et al*., 2013] D. Leung, R. Law, H. Van Hoof, and D. Buhalis, "Social media in tourism and hospitality: A literature review," *Journal of travel & tourism marketing,* vol. 30, no. 1-2, pp. 3-22, 2013.

[Leung *et al*., 2015] X. Y. Leung, L. Xue, and B. Bai, "Internet marketing research in hospitality and tourism: a review and journal preferences," *International Journal of Contemporary Hospitality Management,* 2015.

[Leung, 2020] X. Y. Leung, "Technology-enabled service evolution in tourism: a perspective article," *Tourism Review,* vol. 75, no. 1, pp. 279-282, 2020.

[Li and Jiao, 2020] G. Li and X. Jiao, "Tourism forecasting research: a perspective article," (in German), *Tourism Review of AIEST - International Association of Scientific Experts in Tourism,* vol. 75, no. 1, pp. 263-266, 2020 2020.

[Li *et al*., 2009] Y. Li, Y. Fu, H. Li, and S.-W. Zhang, "The Improved Training Algorithm of Back Propagation Neural Network with Self-adaptive Learning Rate," in *2009 International Conference on Computational Intelligence and Natural Computing*, 2009, vol. 1, pp. 73-76.

[Li *et al*., 2018a] S. Li, T. Chen, L. Wang, and C. Ming, "Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index," *Tourism Management,* vol. 68, pp. 116-126, 2018/10/01/ 2018.

[Li *et al*., 2018b] W. Li, L. Zhu, K. Guo, Y. Shi, and Y. Zheng, "Build a Tourism-Specific Sentiment Lexicon Via Word2vec," *Annals of Data Science,* vol. 5, no. 1, pp. 1-7, 2018/03/01 2018.

[Li *et al*., 2021] M. Li, D. Yin, H. Qiu, and B. Bai, "A systematic review of AI technology-based service encounters: Implications for hospitality and tourism operations," *International Journal of Hospitality Management,* vol. 95, p. 102930, 2021/05/01/ 2021.

[Li, 2000] S. Li, "The development of a hybrid intelligent system for developing marketing strategy," *Decision Support Systems,* vol. 27, no. 4, pp. 395-409, 2000/01/01/ 2000.

[Liang *et al*., 2017] S. Liang, M. Schuckert, R. Law, and L. Masiero, "The relevance of mobile tourism and information technology: an analysis of recent trends and future research directions," *Journal of Travel & Tourism Marketing,* vol. 34, no. 6, pp. 732-748, 2017/07/24 2017.

[Luo *et al*., 2020] J. Luo, S. Huang, and R. Wang, "A fine-grained sentiment analysis of online guest reviews of economy hotels in China," *Journal of Hospitality Marketing & Management,* pp. 1-25, 2020.

[Lv *et al*., 2021] X. Lv, Y. Liu, J. Luo, Y. Liu, and C. Li, "Does a cute artificial intelligence assistant soften the blow? The impact of cuteness on customer tolerance of assistant service failure," *Annals of Tourism Research,* vol. 87, p. 103114, 2021/03/01/ 2021.

[Ma *et al*., 2018] E. Ma, M. Cheng, and A. Hsiao, "Sentiment analysis – a review and agenda for future research in hospitality contexts," *International Journal of Contemporary Hospitality Management,* vol. 30, no. 11, pp. 3287-3308, 2018.

[Marasco *et al*., 2018] A. Marasco, M. De Martino, F. Magnotti, and A. Morvillo, "Collaborative innovation in tourism and hospitality: a systematic review of the literature," *International Journal of Contemporary Hospitality Management,* 2018.

[Mariani *et al*., 2016] M. M. Mariani, M. Di Felice, and M. Mura, "Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations," *Tourism Management,* vol. 54, pp. 321-343, 2016/06/01/ 2016.

[McCarthy, 2007] J. McCarthy, "WHAT IS ARTIFICIAL INTELLIGENCE?," 2007.

[Melián-González *et al*., 2021] S. Melián-González, D. Gutiérrez-Taño, and J. Bulchand-Gidumal, "Predicting the intentions to use chatbots for travel and tourism," *Current Issues in Tourism,* vol. 24, no. 2, pp. 192-210, 2021/01/17 2021.

[Méndez *et al*., 2023] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artificial Intelligence Review,* 2023/02/16 2023.

[Mohamad Suhaili *et al*., 2021] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," *Expert Systems with Applications,* vol. 184, p. 115461, 2021/12/01/ 2021.

[Molina-Collado *et al*., 2022] A. Molina-Collado, M. Gómez-Rico, M. Sigala, M. V. Molina, E. Aranda, and Y. Salinero, "Mapping tourism and hospitality research on information and communication technology: a bibliometric and scientific approach," *Information Technology & Tourism,* vol. 24, no. 2, pp. 299-340, 2022/06/01 2022.

[Monaco and Sacchi, 2023] S. Monaco and G. Sacchi, "Travelling the Metaverse: Potential Benefits and Main Challenges for Tourism Sectors and Research Applications," *Sustainability,* vol. 15, no. 4, p. 3348, 2023.

[Moro and Rita, 2016] S. Moro and P. Rita, "Forecasting tomorrow's tourist," *Worldwide Hospitality and Tourism Themes,* vol. 8, no. 6, pp. 643-653, 2016.

[Muntean *et al*., 2015] C.-I. Muntean, F.-M. Nardini, F. Silvestri, and R. Baraglia, "On Learning Prediction Models for Tourists Paths," *ACM Trans. Intell. Syst. Technol.,* vol. 7, no. 1, pp. 1-34, 2015.

[Neidhardt and Werthner, 2018] J. Neidhardt and H. Werthner, "IT and tourism: still a hot topic, but do not forget IT," *Information Technology & Tourism,* vol. 20, no. 1, pp. 1-7, 2018/12/01 2018.

[Nilashi *et al*., 2017] M. Nilashi, K. Bagherifard, M. Rahmani, and V. Rafe, "A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques," *Computers & Industrial Engineering,* vol. 109, pp. 357-368, 2017/07/01/ 2017.

[Nilashi *et al*., 2021] M. Nilashi *et al.*, "Recommendation agents and information sharing through social media for coronavirus outbreak," *Telematics and Informatics,* vol. 61, p. 101597, 2021/08/01/ 2021.

[Nitti *et al*., 2017] M. Nitti, V. Pilloni, D. Giusto, and V. Popescu, "IoT Architecture for a Sustainable Tourism Application in a Smart City Environment," *Mobile Information Systems,* vol. 2017, p. 9201640, 2017/01/30 2017.

[Pesonen, 2013] J. A. Pesonen, "Information and communications technology and market segmentation in tourism: a review," *Tourism Review,* 2013.

[Pyo *et al*., 2002] S. Pyo, M. Uysal, and H. Chang, "Knowledge Discovery in Database for Tourist Destinations," *Journal of Travel Research,* vol. 40, no. 4, pp. 374-384, 2002.

[Rani *et al*., 2018] S. Rani, K. N. Kholidah, and S. N. Huda, "A Development of Travel Itinerary Planning Application using Traveling Salesman Problem and *K*-Means Clustering Approach," presented at the Proceedings of the 2018 7th International Conference on Software and Computer Applications, Kuantan, Malaysia, 2018. [Online]. Available: https://doi.org/10.1145/3185089.3185142.

[Ravi *et al*., 2019] L. Ravi, V. Subramaniyaswamy, V. Vijayakumar, S. Chen, A. Karmel, and M. Devarajan, "Hybrid Location-based Recommender System for Mobility and Travel Planning," *Mobile Networks and Applications,* vol. 24, no. 4, pp. 1226-1239, 2019/08/01 2019.

[Ren and Hong, 2017] G. Ren and T. Hong, "Investigating online destination images using a topic-based sentiment analysis approach," *Sustainability,* vol. 9, no. 10, p. 1765, 2017.

[Romero Dexeus, 2019] C. Romero Dexeus, "The Deepening Effects of the Digital Revolution," in *The Future of Tourism: Innovation and Sustainability*, E. Fayos-Solà and C. Cooper, Eds. Cham: Springer International Publishing, 2019, pp. 43-69.

[Sano, 2018] A. V. D. Sano, T. D. Imanuel, M. I. Calista, H. Nindito, and A. R. Condrobimo, "The Application of AGNES Algorithm to Optimize Knowledge Base for Tourism Chatbot," in *2018 International Conference on Information Management and Technology (ICIMTech)*, 2018, pp. 65-68.

[Saxena *et al*., 2017] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing,* vol. 267, pp. 664-681, 2017/12/06/ 2017.

[Schwalbe and Wahl, 2020] N. Schwalbe and B. Wahl, "Artificial intelligence and the future of global health," *The Lancet,* vol. 395, no. 10236, pp. 1579-1586, 2020/05/16/ 2020.

[Silva *et al*., 2022] B. Silva, S. Moro, and C. Marques, "Sensing the Impact of COVID-19 Restrictions from Online Reviews: The Cases of London and Paris Unveiled Through Text Mining," Singapore, 2022: Springer Nature Singapore, pp. 223-232.

[Singh, 2015] P. Singh, "Role of geographical information systems in tourism decision making process: a review," *Information Technology & Tourism,* vol. 15, no. 2, pp. 131-179, 2015.

[Stalidis *et al*., 2015] G. Stalidis, D. Karapistolis, and A. Vafeiadis, "Marketing Decision Support Using Artificial Intelligence and Knowledge Modeling: Application to Tourist Destination Management," *Procedia - Social and Behavioral Sciences,* vol. 175, pp. 106-113, 2015/02/12/ 2015.

[Suanpang and Jamjuntret, 2021] P. Suanpang and P. Jamjuntr, "A Chatbot Prototype by Deep Learning Supporting Tourism," *Psychology and Education Journal,* vol. 58, no. 4, pp. 1902-1911, 2021.

[Sugeno and Kang, 1998] M. Sugeno and G. T. Kang, "Structure identification of fuzzy model," *Fuzzy Sets and Systems,* vol. 28, no. 1, pp. 15-33, 1988/10/01/ 1988.

[Sun *et al*., 2019] S. Sun, Y. Wei, K.-L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," *Tourism Management,* vol. 70, pp. 1-10, 2019/02/01/ 2019.

[Sun *et al*., 2022] S. Sun, F. Jiang, G. Feng, S. Wang, and C. Zhang, "The impact of COVID-19 on hotel customer satisfaction: evidence from Beijing and Shanghai in China," *International Journal of Contemporary Hospitality Management,* vol. 34, no. 1, pp. 382-406, 2022.

[Thiengburanathum *et al*., 2016] P. Thiengburanathum, S. Cang, and H. Yu, "Overview of personalized Travel Recommendation Systems," in *2016 22nd International Conference on Automation and Computing (ICAC)*, 2016, pp. 415-422.

[Tranfield *et al*., 2003] D. Tranfield, D. Denyer, and P. Smart, "Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review," *British Journal of Management,* vol. 14, no. 3, pp. 207-222, 2003.

[Tsoumakas, 2019] G. Tsoumakas, "A survey of machine learning techniques for food sales prediction," *Artificial Intelligence Review,* vol. 52, no. 1, pp. 441-447, 2019/06/01 2019.

[UNWTO, 2023] UNWTO. (2023). *Impact assessment of the covid-19 outbreak on international tourism* [Online]. Available: https://www.unwto.org/impact-assessment-of-the-covid-19-outbreak-on-international-tourism.

[Vathis *et al*., 2023] N. Vathis, C. Konstantopoulos, G. Pantziou, and D. Gavalas, "The Vacation Planning Problem: A multi-level clustering-based metaheuristic approach," *Computers & Operations Research,* vol. 150, p. 106083, 2023.

[Verma *et al*., 2022] S. Verma, L. Warrier, B. Bolia, and S. Mehta, "Past, present, and future of virtual tourism-a literature review," *International Journal of Information Management Data Insights,* vol. 2, no. 2, p. 100085, 2022/11/01/ 2022.

[Wang, 2004] C.-H. Wang, "Predicting tourism demand using fuzzy time series and hybrid grey theory," *Tourism Management,* vol. 25, no. 3, pp. 367-374, 2004/06/01/ 2004.

[WTTC, 2022] WTTC. (2022). *Travel & Tourism Economic Impact 2022* [Online]. Available: https://wttc.org/Portals/0/Documents/Reports/2022/EIR2022-Global%20Trends.pdf.

[WTTC, 2023] WTTC. (2023). *Economic Impact Tourism* [Online]. Available: https://wttc.org/Research/Economic-Impact.

[Xiang, 2018] Z. Xiang, "From digitization to the age of acceleration: On information technology and tourism," *Tourism Management Perspectives,* vol. 25, pp. 147-150, 2018/01/01/ 2018.

[Yang and Han., 2021] M. Yang and C. Han, "Revealing industry challenge and business response to Covid-19: a text mining approach," *International Journal of Contemporary Hospitality Management,* vol. 33, no. 4, pp. 1230-1248, 2021.

[Ye *et al*., 2009] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications,* vol. 36, no. 3, Part 2, pp. 6527-6535, 2009/04/01/ 2009.

[Young *et al*., 2018] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," *IEEE Computational Intelligence Magazine,* vol. 13, no. 3, pp. 55-75, 2018.

[Zeng and Gerritsen, 2014] B. Zeng and R. Gerritsen, "What do we know about social media in tourism? A review," *Tourism management perspectives,* vol. 10, pp. 27-36, 2014.

[Zsarnoczky, 2017] M. Zsarnoczky, "HOW DOES ARTIFICIAL INTELLIGENCE AFFECT THE TOURISM INDUSTRY?," *Journal of Management,* vol. 2, no. 31, 2017.

Capítulo III. Exploratory techniques to analyze Ecuador's tourism industry

Autores: Anita Herrera[1], Ángel Arroyo[1], Alfredo Jiménez[2] and Álvaro Herrero[1]

Afiliaciones:

[1] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela Politécnica Superior, Universidad de Burgos, Av.Cantabria s/n, 09006, Burgos, Spain, ahv1002@alu.ubu.es, {aarroyop, ahcosio}@ubu.es
[2] KEDGE Business School, 680 cours de la Liberation, Talence (Bordeaux) France. alfredo.jimenez@kedgebs.com

Resumen

El crecimiento de la industria turística y su impacto en la economía mundial ha impulsado la investigación de este sector, generando interés en su gestión. En el presente estudio, se aplican y analizan las técnicas de reducción de dimensionalidad y agrupamiento para evaluar el comportamiento de conjuntos de datos relacionados con la operación de las empresas turísticas de Ecuador.

Las técnicas de reducción de dimensionalidad utilizadas son Principal Component Analysis (PCA), Isometric Feature Mapping (ISOMAP) y Locally Linear Embedding (LLE). Así también se usan las técnicas de agrupamiento *k*-means, *k*-medoids y Hierarchical Clustering en conjunto con las técnicas Elbow Method, Silhouette Coefficient, Davies-Bouldin Index y Dunn Index, para evaluar el número óptimo de grupos.

Los datos analizados corresponden a estadísticas sectoriales provenientes de encuestas aplicadas por el Instituto Ecuatoriano de Estadísticas y Censos de Ecuador (INEC), a las empresas turísticas del país. Con este estudio se pretende identificar la tendencia de la operación de las empresas turísticas de Ecuador, así como reconocer el efecto de factores externos en la industria turística.

La metodología utilizada en este estudio se inicia con la selección de los datos correspondientes a los años 2015, 2019 y 2020. La elección de estos años se basa en consideraciones específicas: el año 2015 corresponde a un período con mayor asignación de presupuesto para el desarrollo turístico en Ecuador, el 2019 refleja el impacto al sector turístico de Ecuador durante fuertes protestas protagonizadas por el sector indígena del país y el año 2020 como período marcado por la pandemia generada por el Covid-19.

El análisis considera las variables evaluadas como relevantes, estadísticamente significativas y disponibles en los conjuntos de datos examinados. A continuación, se buscó la correlación entre las variables, obteniendo las dependencias entre estas para crear la matriz de similitud. Además, como parte del preprocesamiento, se normalizan los datos y se identifican los outliers, para gestionar valores

atípicos que podrían afectar la validez del estudio. El preprocesamiento permite asegurar la consistencia y la validez de los datos antes de aplicar las diferentes técnicas de reducción de dimensionalidad y agrupamiento.

Posteriormente, se aplicaron las técnicas de reducción de dimensionalidad ya mencionadas, facilitando la comprensión y visualización de los datos, mientras se minimiza la pérdida de información. El uso de estas técnicas no solo posibilitó entender la estructura de los datos, sino también identificar patrones relevantes. A continuación, mediante las técnicas de evaluación del número de grupos se identificó el valor óptimo de grupos, parámetro importante para continuar con el análisis de los datos mediante las técnicas de agrupamiento. Finalmente se identificaron los porcentajes de asignación de muestras para cada uno de los períodos estudiados, a fin de facilitar el análisis de los resultados.

Al evaluar las técnicas inteligentes aplicadas en este estudio, es importante reconocer la relevancia de aplicar las medidas de evaluación del número óptimo de grupos, previo al uso de las técnicas de agrupamiento. Los métodos coinciden en el número de grupos sugerido para cada período analizado. En cuanto a las técnicas de agrupamiento *k*-means y *k*-medoids presentan resultados similares con el uso de distancias euclidianas.

El estudio refleja que en el año 2015 existía un mayor número de empresas turísticas en el segmento de medianas y grandes empresas, esto considerando mejores condiciones en la operación de este sector, mientras que esta distribución cambia en el año 2020 cuando las pérdidas económicas por las restricciones del Covid-19 obligaron al cierre de las empresas turísticas, principalmente a las pequeñas y medianas empresas. Además, al analizar la operación de las empresas turísticas en los períodos seleccionados, es posible identificar que, factores externos como las protestas sociales y la pandemia afectaron significativamente la operación del sector turístico.

Exploratory techniques to analyze Ecuador's tourism industry

Anita Herrera[1] [0000-0002-2655-412X], Ángel Arroyo[1][0000-0002-3561-6257], Alfredo Jiménez[2][0000-0001-7811-5113] and Álvaro Herrero[1][0000-0002-2444-5384]

[1] Applied Computational Intelligence Group (GICAP), Department of Digitalization, Polytechnic School, University of Burgos, Av. Cantabria s/n, 09006, Burgos, Spain, ahv1002@alu.ubu.es, {aarroyop, ahcosio}@ubu.es

[2]KEDGE Business School, 680 cours de la Liberation, Talence (Bordeaux) France. alfredo.jimenez@kedgebs.com

Abstract. The analysis of the operation of tourism companies will provide valid information for the design of policies to reactivate the tourism industry, which has been strongly affected during the pandemic generated by Covid-19. The objective of this paper is to use Soft Computing techniques to analyze tourism companies in Ecuador. First of all, dimensionality reduction methods are applied: Principal Component Analysis, Isometric Feature Mapping, and Locally Linear Embedding, on data of tourism enterprises in Ecuador for the year 2015. In addition, to verify the trend of operational variables, the data of tourism companies in Ecuador in 2019 and 2020 are analyzed with dimensionality reduction methods that improve the interpretation by minimizing the loss of information. The datasets are analyzed with *k*-means, *k*-medoids, and Hierarchical Clustering, generating groups according to similar characteristics. The optimal number of clusters is determined with: Elbow Method, Silhouette Coefficient, Davies-Bouldin Index, and Dunn Index. In addition, an analysis of the operation of tourism companies in the year 2020 concerning previous years is included. The study allows exploring Soft Computing techniques to identify important information for the definition of strategies that contribute to an effective reactivation of the tourist industry of Ecuador.

Keywords: Machine learning, Soft Computing, dimensionality reduction, clustering techniques, tourism industry.

1 Introduction

The tourism sector is a key area of job creation and foreign exchange generation in many advanced and emerging economies, with consecutive growth in recent years [1]. However, this scenario changed with the declaration of a pandemic in March 2020. The value chain of the tourism industry was probably one of the first to come to a halt, facing the closure of tourism establishments and the loss of jobs worldwide.

In the case of Ecuador, tourism establishments were affected during the pandemic, in an industry that generated US$ 2.28 billion, with a contribution to the Gross Domestic Product (GDP) of 2.24% in 2019 [2]. Thus, Ecuador's tourism sector is betting on a reactivation focused on rethinking tourism services towards more sustainable and safer models [3].

Decision makers in the tourism sector in Ecuador have an opportunity in Soft Computing techniques to generate information that allows the design, implementation, and evaluation of new strategies, reducing the impact of decision-making in scenarios with a high degree of imprecision and uncertainty. Research should generate useful information on the necessary transformations in the tourism sector to face the crisis caused by the pandemic. In this context, it is valid to use Soft Computing techniques to analyze the economic structure and production of tourism enterprises in Ecuador, extending previous work [4], for which datasets of the operation of tourism companies in Ecuador corresponding to the years 2015, 2019 and 2020 are analyzed, applying dimensionality reduction techniques: Principal Component Analysis (PCA), Isometric Mapping (ISOMAP) and Locally Linear Embedding (LLE), metrics to evaluate clustering algorithms: Elbow Method, Silhouette Coefficient, Davies-Bouldin Index and Dunn Index, and clustering methods: *k*-means, and *k*-medoids, and Hierarchical clustering.

The study conducted by Penagos et *al.* [5], identifies tourism segments according to the perception of sustainability and reliability of the destination. For this purpose, it selects the most useful variables, applying a genetic algorithm, to a sample of 438 tourists from Chile and Ecuador over 17 years of age. The results of the research contribute to the analysis of the strategies of public and private institutions in tourist destinations.

Also, based on monthly data on the performance of airlines, hotels, and the online searches of potential tourists, Anguera-Torrell et *al*. [6] propose the creation of an urban tourism performance index, with data from 16 tourism cities collected since the initial outbreak of the COVID-19 pandemic in 2020. As part of the study, the data are treated with the Principal Component Analysis (PCA) technique, generating an index that can be used as a tool to analyze the recovery that these cities will experience.

The economy of the tourism sector presented serious problems during the pandemic, many of them with credit risk. Wang [7] proposes a model to accurately identify an institution's credit risk. The model uses financial indicators as input data and considers Isometric Mapping (ISOMAP) in the processing stage because financial data present a non-linear distribution.

The objective of this work is to use Soft Computing techniques in variables related to the operation of tourism companies in Ecuador, improving the performance of visualization and clustering, to identify valid information that contributes to the reactivation of this industry.

The continuation of this document includes Section 2 called *Applied Techniques*, which describes the dimensionality reduction techniques: Principal Component Analysis (PCA), Isometric Mapping (ISOMAP) and Locally Linear Embedding (LLE), the methods of clustering: *k*-means, *k*-medoids, and Hierarchical Cluster, and the metrics to evaluate clustering algorithms: Elbow Method, Silhouette Coefficient, Davies-Bouldin Index and Dunn Index. Section 3 describes a *Case Study* with data on the operation of tourism enterprises in Ecuador. Section 4 corresponds to the *Experiment and Results*. Finally, Section 5 includes the *Conclusions* of the present work.

2 Techniques Applied

Dimensionality reduction and clustering techniques have been applied to visualize and group the information into an efficient representation [8].

Dimensionality reduction is a set of techniques that allow reducing a dataset of a certain dimension to a smaller one. Namely, these techniques transform the data set X with dimensionality D into a new data set Y with dimensionality d (where d<D), preserving as much information as possible [9]. Also, allow an improvement in the visualization of high dimensional data, reduces the noise of the original dataset, and the possibility to discover hidden patterns in a dataset [10].

In turn, the objective of clustering is to discover a new set of categories, clustering groups the data into subsets such that similar data are grouped, while different data belong to other groups [11] [12]. In this context, the definition of the optimal number of clusters represents an important step in the application of clustering methods [13].

The dimensionality reduction, clustering methods, and methods for defining the optimal number of clusters, developed in the case study of this work are described below.

2.1 Principal Component Analysis

Principal Component Analysis (PCA), is a dimensionality reduction technique used in many disciplines to extract important information from a data set and represent it as a set of new orthogonal variables called principal components [14].

PCA reduces an n-dimensional data set to an m-dimensional subspace, where m is less than n, for this, PCA determines the linear combinations that best represent the original variables (X_1, ..., X_p) The *m* components *(*Z_1*, ...,* Z_m*)* will be identified from the linear combinations of the original p variables:

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm} X_j$$

Where $\emptyset_{1m}$, $\emptyset_{2m}$, ... $\emptyset_{pm}$ are the weights of the principal components: $\emptyset_{11}$ corresponds to the first load of the first primary component.

2.2 Isometric Feature Mapping

Isometric Mapping (ISOMAP) is a nonlinear dimensionality reduction method that considers the shortest distance between data in a dataset. For this, a neighborhood graph connecting each datum to its *k* nearest neighbors is constructed and the geodesic distance between each pair of points in the graph is calculated, using Dijkstra's method or Floyd's shortest path algorithm. The shortest distances form the pairwise geodesic distance matrix [9]. The solution is given by:

$$Y = (\sum \mathrm{Iso,k})^{1/2} \mathrm{U}_{\mathrm{Iso,k}}^{\mathrm{T}}$$

Where $\sum_{\mathrm{Iso,k}}$ is the diagonal matrix of the top *k* singular values of $\mathrm{K}_{\mathrm{Iso}}$ and $\mathrm{U}_{\mathrm{Iso,k}}$ are the associated singular vectors.

2.3 Locally Linear Embedding

Locally Linear Embedding (LLE) [15] is a nonlinear dimensionality reduction method, that identifies the *k* nearest neighbors of each data point and constructs a weights matrix, where each point is a linear combination of its neighbors.

$$E(W) = \sum_i (Xi - \sum_j WijXj\ )^2$$

X_i= position of the point of interest.
X_j= position of all the nearest neighbors

The cost function is solved to find the weights, where the sum of weights for each X_i is set to equal to 1. This method can be considered as a series of local principal component analyses that are compared globally to find the best nonlinear embedding.

Clustering techniques that allow grouping the samples into groups of similar characteristics are described below.

2.4 *K*-means

The *k*-means method is a clustering algorithm that seeks the formation of *k* well-defined groups. The method requires defining the number of clusters and will aim to identify the centroid and the members of each group by minimizing the sum of distances between the objects and the centroid of their group. The solution is formulated as follows:

$$SSE = \sum_{j=1}^{k} \sum_{x \in Gi} \frac{p(X_i, C_j)}{n}$$

Where *p* is the proximity function, *k* is the number of groups, C_j is the number of centroids and *n* is the number of rows [16].

2.5 *K*-medoids

K-medoids is a clustering method, in which each cluster is represented by an observation present in the cluster. This observation, called a medoid, is an element within a cluster whose average distance between it and all other elements in the same cluster is as small as possible [17].

In this study, the Euclidean distance is chosen as the dissimilarity measure, although other measures can be adopted. The Euclidean distance between object i and object j is given by:

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (X_{ia} - X_{ja}\ )^2}$$

Where *i*=1, …, *n* ; *j*=1, … , n

2.6 Hierarchical cluster

Hierarchical clustering is a method of automatic data classification, which detects objects that can be considered similar and assign them to the same group, leaving different objects in another cluster. The similarity of objects is visualized in a dendrogram, which is a tree-like structure [18].

This classification method uses the Euclidean distance [19], which defines the distance between points *x* and *y* as:

$$d(x, y) = \left\|x - y\right\| = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2}$$

Where the point $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ of the n-dimensional Euclidean space.

2.7 Silhouette Index

The Silhouette index analyzes the intra-cluster and inter-cluster distances. The difference between the mean distance to points in the same cluster as itself, represents the cohesion, while the mean distance to points in other neighboring clusters is the separation [20]. The optimal number of clusters corresponds to the highest value of the Silhouette index, which is given by:

$$Sil(C) = \frac{1}{N}\sum_{C_k \in C}\sum_{X_i \in C_k}\frac{b\ (X_i, C_k) - \mathrm{a}\ (X_i, C_k)}{\max\{a(X_i, C_k), b(X_i, C_k)\}}$$

Where: $a(X_i, C_k) = \frac{1}{|C_k|}\sum_{X_j \in C_k} d_e(X_i, X_j)$

$$min_{C_l \in C \setminus C_k}\left\{\frac{1}{|C_l|}\sum_{X_j \in C_l} d_e(X_i, X_j)\right\}$$

Dataset $X = \{X_1, X_2, \ldots X_N\}$ of N objects, C_k - centroid of cluster, The partition X into K groups: $C = \{c_1, c_2, \ldots, c_k\}$, $d_e(X_i, X_j)$ – Euclidean distance.

2.8 Davis-Bouldin Index

David Bouldin Index considers the minimum distance of the points of a cluster to its centroid and the separation as a function of the distance between centroids [21]. The appropriate number of clusters according to the David Bouldin index corresponds to the smallest value. This index is defined as:

$$DB(C) = \frac{1}{K}\sum_{C_k \in C} max_{C_l \in C \setminus C_k}\left\{\frac{S\ (C_k) + S(C_l)}{d_e\ (C_k,\ C_l)}\right\}$$

Where:

$$S(c_k) = \frac{1}{|C_k|}\sum_{X_i \in C_k} d_e(X_i, C_k)$$

Dataset $X = \{X_1, X_2, \ldots X_N\}$ of N objects, C_k - centroid of cluster, The partition X into K groups: $C = \{c_1, c_2, \ldots, c_k\}$, $d_e(X_i, X_j)$ – Euclidean distance.

2.9 DUNN

Dunn index estimates the distance to the nearest neighbor and the maximum diameter of the cluster [22]. The higher Dunn's index is given by:

$$D(C) = \frac{min_{C_k \in C \setminus C_k}\left\{min_{C_l \in C \setminus C_k}\{\partial(C_k,\ C_l)\}\right\}}{max_{C_k \in C}\{\Delta\ (C_k)\}}$$

Where:

$$\partial(C_k, C_l) = min_{X_i \in C_k}\ min_{X_j \in C_l}\ \{d_e\ (X_i, X_j)\}$$

$$\Delta(C_k) = max_{X_i\, X_j \in C_k}\ \{d_e\ (X_i, X_j)\}$$

Dataset $X = \{X_1, X_2, \ldots X_N\}$ of N objects, C_k - centroid of cluster, The partition X into K groups: $C = \{c_1, c_2, \ldots, c_k\}$, $d_e(X_i, X_j)$ – Euclidean distance

2.10 Elbow method

The "Elbow method" is described in the document Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Technique [4].

3 Case Study

This paper presents a set of Soft Computing methods to analyze the operation of tourism companies in Ecuador. For this purpose, data from the country's tourism companies for the year 2015 are analyzed, from the Survey of Hotels, Restaurants, and Services, presented for the National Institute of Statistics and Census of Ecuador [23]. At the same time, data on the operation of companies in 2019 and 2020 are analyzed with the data presented by the National Institute of Statistics and Census of Ecuador, in the Structural Survey of Companies ENESEM, verifying the trend of the variables in these periods [24] [25]. The data sets are analyzed separately and together, with a total of 397 records.

The study considers the information compiled annually by the entity responsible for statistics and censuses in Ecuador. The variables considered for the analysis are:

1. PRODUCTION (V1) – USD. Degree of the utilization of all factors involved in the production process.
2. DEPRECIATION (V2) – USD. Asset value loss due to time or use.
3. STAFF (V3) – number of people. Including personnel who work in/or for the company with which they have a working relationship.
4. REMUNERATION (V4) – USD. Payments made by the company to executives, managers, employees, or workers, both in capital and other means.
5. EARNINGS (V5) – USD. Recognized percentage by the employer or company to its employees.
6. TAXES (V6) – USD. Sum of money made by taxpayers who by law are obliged to pay.
7. INTERMEDIATE CONSUMPTION (V7) – USD. Supplies that are used to produce other goods and services.

Data pre-processing is a fundamental step before the application of clustering methods. In the preprocessing stage of the present work, the data are normalized between 0 and 1, and outliers are eliminated, for which the values of each variable are analyzed by means of a Box Plot.

The preprocessing of the dataset of the present study discarded the highly correlated variables that were not going to influence the result, finally considering a data set of seven variables. In the analysis stage, case studies using seven variables' datasets were identified. Nasrin et *al.* [26] present a study comparing the usefulness of ant colony optimization (ACO), genetic algorithm (GA), and *K*-means methods for clustering climatic variables affecting rainfed wheat yield in northeastern Iran, this study uses a dataset with seven variables. In addition, the study conducted by Ngurah Krisnanda Putra et *al.* [27] uses *k*-means on a seven variables dataset for scholarship selection.

4 Experiments and Results

This work begins with the application of the PCA method in the data set of the year 2015. Table 1 corresponds to the correlation matrix and shows the correlation values of the variables to be analyzed.

Table 2. Correlation Matrix year 2015

	V1	V2	V3	V4	V5	V6	V7
V1	1.000	**.287**	.681	**.943**	**.813**	.718	**.976**
V2	.287	1.000	**.244**	**.167**	**.153**	.414	.270
V3	.681	.244	1.000	.765	.396	.527	.685
V4	-.943	.167	.765	1.000	**.842**	.662	**.882**
V5	.813	.153	.396	.842	1.000	.493	.690
V6	-.718	.414	-.527	.662	.493	1.000	.697
V7	-.976	270	-.685	.882	-.690	.697	1.000

a. Determinant = 1.793E-5

The correlation matrix shows variables with strong associations. The correlation between Production (V1) and Intermediate Consumption (V7) is 97.6%. Therefore, the higher the production, the higher the intermediate consumption. The same occurs for other variables in the correlation matrix. On the other hand, variables that do not have a strong association are marked in red.

To study the relationships between correlated variables, the set of original variables is transformed into a set of principal components. Fig. 1 corresponds to the diagram of principal components.

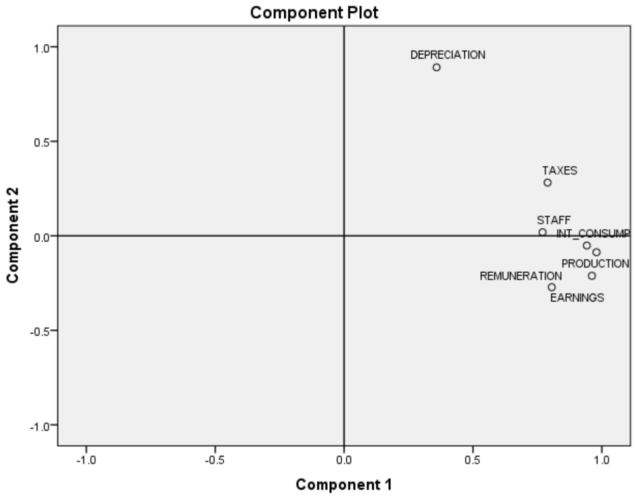

Fig. 1. Variables Diagram (PCA year 2015)

Figure 1 shows that the variables Production and Intermediate Consumption are very close, which means that they are highly correlated variables. The Remuneration variable has a high correlation with the Profits variable, indicating that as profits increase so do remunerations. The increase or decrease of the Depreciation variable does not provide information on the behavior of the rest of the variables.

The Principal Component Analysis – PCA, described above corresponds to the study carried out in the document Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Techniques [4]. The present paper searches for a possible data structure with PCA and uses ISOMAP and LLE, dimensionality reduction methods, which aim to facilitate data visualization, remove possible redundant features and improve data analysis.

Figure 2 shows two main components that project the entire dataset.

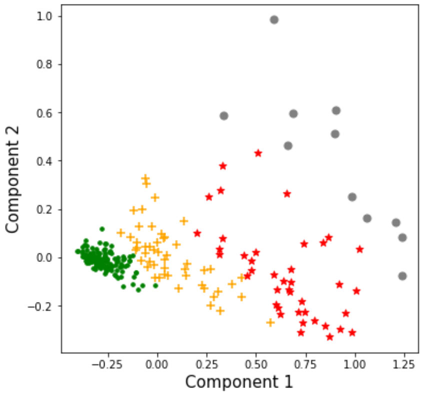

Fig. 2. PCA projection year 2015. Number of dimensions: 2. Label representation: ● micro-production, + small production, ★ medium production, and ● high production

Fig 2 shows the data according to the level of production. Distances are interpreted in terms of similarity. Therefore, micro-enterprises concentrate on the largest number of tourism enterprises with a fairly similar level of production. On the other hand, large tourism enterprises concentrate the smallest number of tourism enterprises in Ecuador, with a greater difference in their level of production. Small firms produce less volume and with similar values in quantity produced, while large firms produce much more and differ more in their numbers.

In Fig. 2, 95% of tourism enterprises correspond to micro, small and medium-sized tourism companies. This percentage should be considered by the authorities in the creation of specific strategies for the reactivation of the tourism industry, focusing on appropriate conditions for each group. The definition of this type of strategy should be based on studies such as the one presented by Conejero et al. [28], who implement a complete support system for public administrations in decision-making, based on Machine Learning techniques such as association rules.

Fig. 3 shows the PCA projection according to the number of employees.

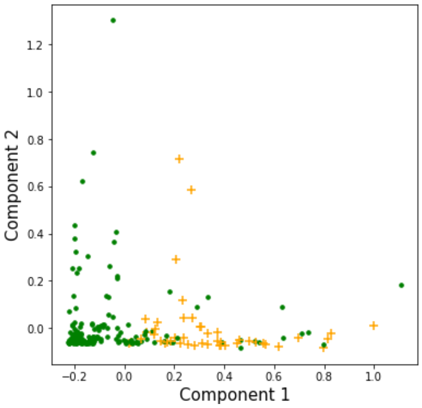

Fig. 3. PCA projection year 2015. Number of dimensions: 2. Representation of labels: ● low number of employees, + high number of employees

In Fig. 3, identified with the symbol '●' on the left side of the image, are the tourism companies with the lowest number of employees. Identified with the symbol '+ ', in the middle and right parts of the image are the companies with the highest number of employees. The first group of 50 to 99 employees and the second group of up to 49 employees. In Ecuador, most tourism companies have less than 49 employees.

The projection obtained is not as clear as the previous one (Fig. 2), which could be related to the labor reforms established in Ecuador since the year 2008. The objective of these reforms was to eliminate labor intermediation and hourly contracts to improve working conditions. However, these reforms generate problems in sectors such as tourism, which requires different employment modalities according to the demand for services.

Fig. 4 shows the ISOMAP projection, according to the level of production.

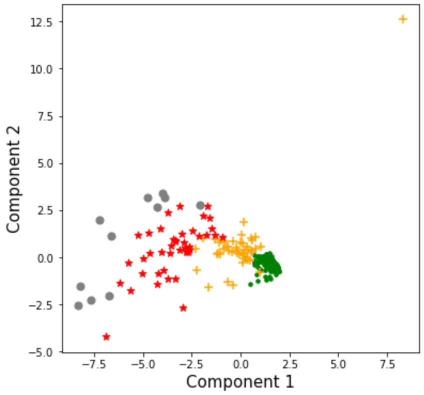

Fig. 4. ISOMAP projection year 2015. Number of dimensions: 2. Label representation: ● micro-production, + small production, ★ medium production, and ● high production

In Fig. 4, the dimensionality reduction method applied is ISOMAP and shows the data according to the level of production. The optimal number of neighbors is 12. and presents the reduction to 2 dimensions. When comparing the PCA and ISOMAP methods, it can be noted that PCA (Fig. 2) provides a better visualization of the groups, organized according to production level. PCA estimation is better, this may occur because the points do not form a uniform sample of the embedded space or there is noise in the data.

Figure 5 shows the projection using the LLE method.

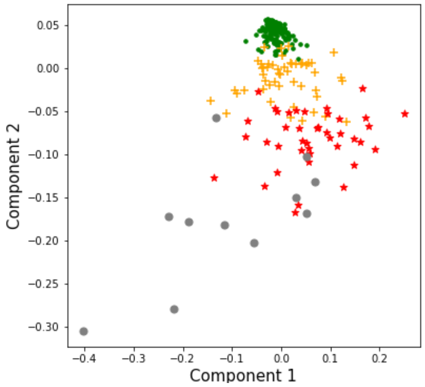

Fig. 5. LLE projection year 2015. Number of dimensions: 2. Label representation: ● micro-production, + small production, ★ medium production, and ● high production

In Fig. 5, the dimensionality reduction method applied is LLE and shows the data according to the level of production, with 12 nearest neighbor graphs, and presents the reduction to 2 dimensions. The LLE algorithm considers local information and its performance may be affected by choosing incorrectly the parameters, which influences the quality of the results.

The PCA projection (Fig. 2) achieves a better definition of the groups of tourism enterprises according to the level of production, compared to the LLE method (Fig. 5).

Table 2 corresponds to the correlation matrix and shows the correlation values of the variables to be analyzed for the 2019 year.

Table 2. Correlation Matrix year 2019

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**
V1	1.000	.475	.956	**.967**	**.754**	.609	**.984**
V2	.475	1.000	.509	.453	.410	.823	.434
V3	.956	.509	1.000	.920	.549	.701	.975
V4	.967	.453	.920	1.000	**.748**	.521	**.922**
V5	.754	.410	.549	.748	1.000	.317	.648
V6	.609	.823	.701	.521	.317	1.000	.639

V7	.984	.434	.975	.922	.648	.639	1.000

Determinant = 6.13E-009

In the 2019 data, as in the 2015 data, a strong association is observed. Table 2 shows a 98.4% correlation between Production (V1) and Intermediate Consumption (V7). The value of the determinant is practically 0, which means that the correlation matrix has highly correlated variables. In other words, the increase or decrease of one variable provides information about the behavior of the correlated variable [29]. Also, the personnel and production variables show a higher correlation in Table 2, this could reflect the increase in direct tourism employment generated from 2015 to 2019 [30].

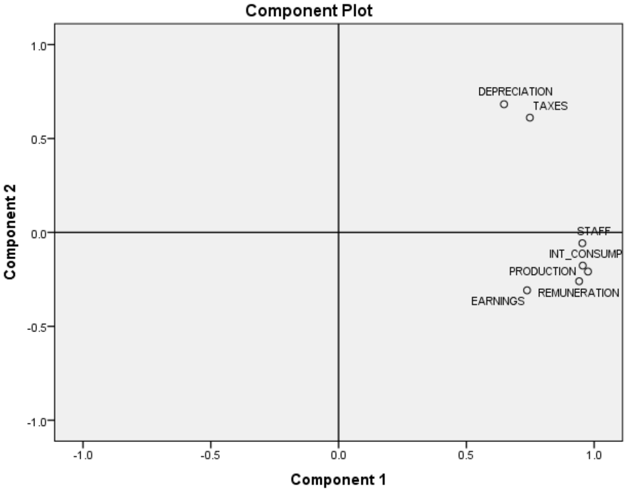

Fig. 6. Variables Diagram (PCA year 2019)

Table 3 corresponds to the correlation matrix and shows the correlation values of the variables to be analyzed for the 2020 year.

Table 3. Correlation Matrix year 2020

	V1	V2	V3	V4	V5	V6	V7
V1	1.000	.244	.965	.979	-0.29	**.868**	**.997**
V2	.244	1.000	.263	.289	-.290	.320	.243
V3	.965	.263	1.000	**.983**	-.149	**.886**	.978
V4	.979	.289	.983	1.000	-.109	.866	.969
V5	-0.29	-.290	-.149	-.109	1.000	-.319	.886
V6	.868	.320	.886	.866	-.319	1.000	.863
V7	.997	.243	.978	.969	.886	.863	1.000

a. Determinant = 4.488E-007

In the 2020 data, it is observed that the variable Earnings (v2) has negative values, which would reflect the economic problems generated in the tourism industry by Covid-19.

When applying dimensionality reduction techniques, 2, 3, and 4 groups are intuitively identified. The visual representation infers the most approximate values for the *k* parameter. However, it is important to use the methods that allow determining the optimal number of clusters. In the document Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Techniques [4], the "Elbow Method" is used to define the number of clusters. In addition, the Silhouette Coefficient, Davies-Bouldin Index, and Dunn Index are used to define the optimal number of clusters.

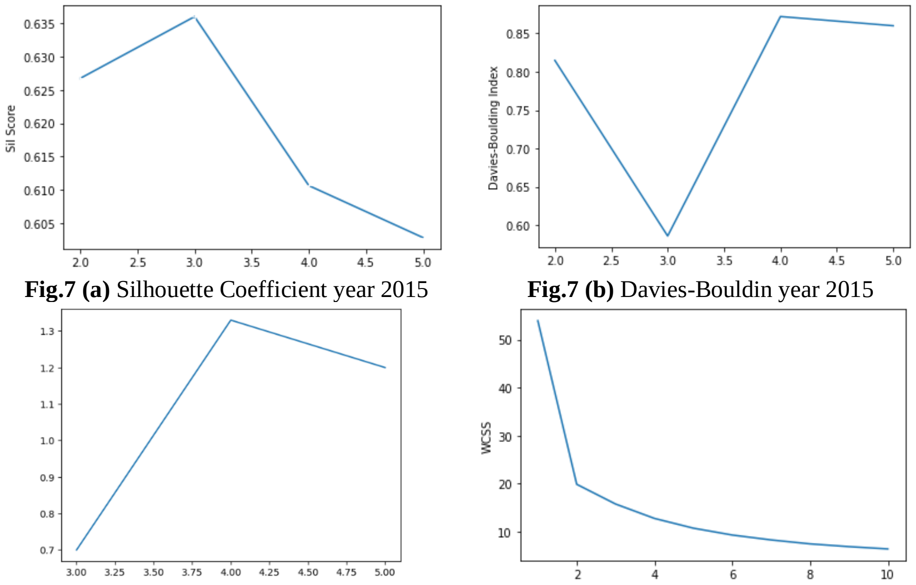

Fig.7 (a) Silhouette Coefficient year 2015

Fig.7 (b) Davies-Bouldin year 2015

Fig.7 (c) Dunn Index year 2015

Fig.7 (d) Elbow Method year 2015

According to the results of Figures 7(a), 7(b), and 7(d), the best value for *k* is 3, which coincides with the graphical representation presented below. In Figures 7(a), 7(b), 7(c), and 7(d), the *X-axis* represents the number of clusters (*k*), and the *Y-axis* represents the value of the index. The objective of the indices is to identify compact clusters, with little variation among cluster members, and well separated from other clusters.

In Fig 8. *k*-means applied to data from enterprises in Ecuador, in the year 2015, (a) parameter *k*=2, (b) parameter *k*=3, and (c) parameter *k*=4.

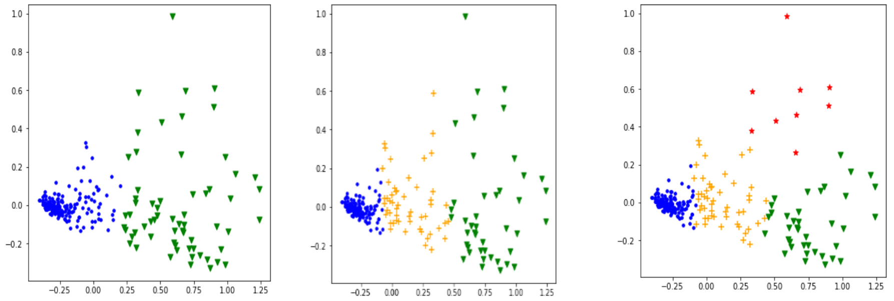

Fig.8 (a) *k*-means with *k* = 2 year 2015

Fig.8 (b) *k*-means with *k* = 3 year 2015

Fig.8 (c) *k*-means with *k* = 4 year 2015

The groups identified in Figures 8(a), 8(b), and 8(c) through the application of *k*-means, provide important information for the generation of strategies in the reactivation of the tourism industry. For example, this information could suggest the generation of national and international tourism promotion campaigns that consider the level of production of the tourism companies.

The results of the *k*-means technique are shown in Table 4.

Table 4. Clustering results for the operation of tourism companies' dataset year 2015

Method	***K***	**Micro companies**	**Small companies**	**Medium companies**	**Large companies**
K-means	2	[94 55]	[67 3]	[13 5]	[8 0]
K-means	3	[30 75 44]	[3 67 0]	[5 13 0]	[0 8 0]
K-means	4	[71 8 44 26]	[67 0 3 0]	[13 1 0 4]	[8 0 0 0]

In Table 4, column *k* represents the number of clusters defined for the algorithm. The columns: Micro, Small, Medium, and Large companies, represent the number of samples from this group that is assigned to each of the clusters. For example, [94 55] represents 94 of the samples assigned to the first cluster and 55 of the samples assigned to the second cluster.

Fig. 9 shows Hierarchical clustering dendrograms.

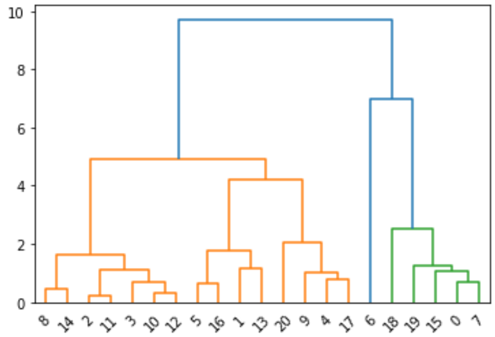

Fig. 9. Hierarchical clustering dendrogram year 2015

In this case study, the hierarchical clustering starts with 21 clusters, each one composed of a single element, corresponding to the tourism companies grouped according to their geographic location (province). Then, the two clusters with the greatest similarity are merged into the same cluster, this operation is repeated until a single cluster is formed.

Fig. 9 shows the dendrogram, the diagram with which the hierarchical clustering is represented. It presents three groups of data, the group on the left with the majority of the samples representing the locations with the highest tourism production, while the groups on the right have fewer samples and correspond to the locations with the highest tourism production.

The data in each group are closely related to each other and the optimal number of clusters corresponds to the branch with the greatest distance in the dendrogram.

Fig. 10 shows the *k*.medoids method for the 2015 data.

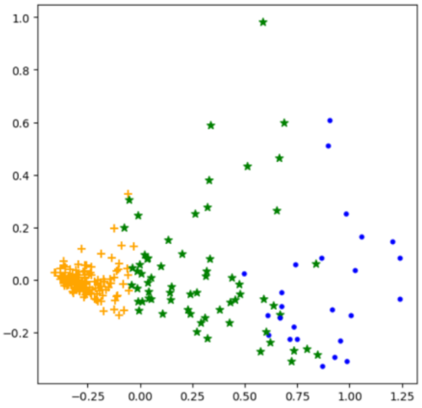

Fig. 10 *K*.medoids method year 2015

The clusters identified with the *k*-medoids method (Fig. 10) match with the results of the *k*-means method Fig. 8 (b).

The 2015 data will be compared with the 2019 and 2020 data, using the Soft Computing methods already described. Fig. 11 includes the clustering result for the year 2019 and Fig. 12 includes the clustering result for 2020, with the optimal number of clusters, determined with the metrics for evaluating clustering

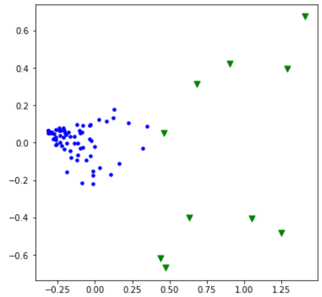

Fig.11 (a) *k*-means with *k* = 2, year 2019

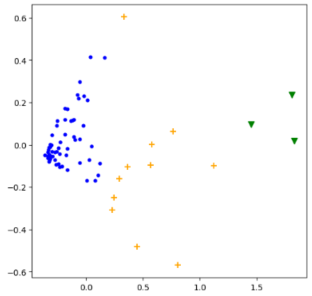

Fig.11 (b) *k*-means with *k* = 3, year 2020

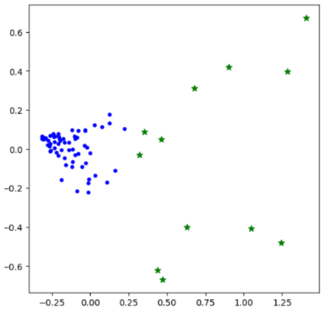

Fig.11 (c) *k*-medoids with *k* = 2, year 2019

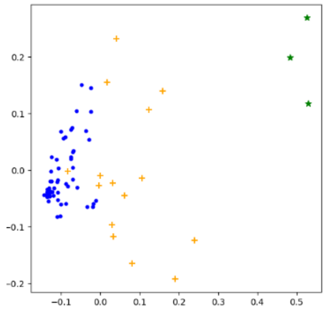

Fig.11 (d) *k*-medoids with *k* = 3, year 2020

To analyze the results obtained in this work, the treatment of the dataset formed by the 2015, 2019, and 2020 datasets is presented below. Figure 12 presents the dimensionality reduction performed with the PCA method, and Figure 13 the *k*-means clustering.

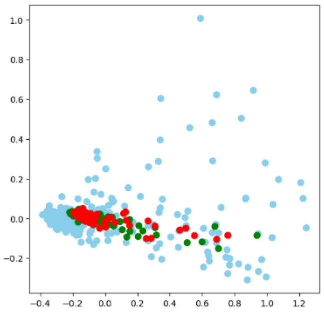

Fig.12 (a) PCA proyection year 2015, 2019, 2020. Number of dimensions: 2. Representation of labels: ● year 2015, ● year 2019, ● year 2020

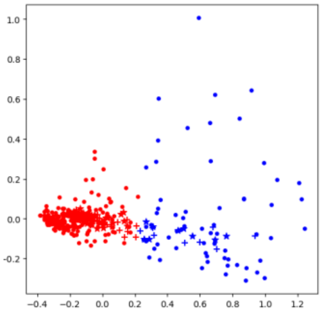

Fig.12 (b) *k*-means with *k*=2, year 2015, 2019, 2020

Figure 12(a) shows that the operation of tourism companies in 2019 decreased compared to 2015. This could be related to the strong social protests that occurred in 2019 in Ecuador, which generated cancellations and decreased the operation of tourism companies. Figure 12 also shows that the operation of tourism companies decreases even more in 2020, which would be related to the crisis in the tourism sector caused by the covid-19 pandemic.

Figure 12(b) shows the result of *k*-means with *k*=2. According to the cluster evaluation indexes, the most appropriate value of *k* is 2. Figure 12(b) identifies 2 groups according to the level of operation of tourism companies in Ecuador. It also identifies the level of operation of the industry in 2015, 2019 and 2020 ("o" year 2015, "+" year 2019, and "*" year 2020).

The first group, identified in red, includes tourism companies with a lower volume of operations. It can be seen that the elements of this group have less dispersion, that is, most of the tourism companies in Ecuador have a similar level of operations. On the other hand, the second group, identified with blue color, displays a much larger dispersion, indicating that these firms have a heterogeneous level of operations.

The groups identified in Figure 12(b) through the application of *k*-means provide important information for the creation of strategies to reactivate the tourism industry. For example, this information could suggest the generation of opportunities for the exchange of knowledge, such as technical visits, dialogue spaces, etc. for tourism companies with similar levels of operations in Ecuador.

5 Conclusions

The results obtained in this work show the importance of Soft Computing techniques in the generation of strategies to support the reactivation of the tourism industry.

The Soft Computing techniques applied in this study identify the relationship between the operating variables of tourism companies in Ecuador. This type of study should be considered for the creation of strategies and public policies that support the reactivation of the tourism industry, with a regulatory framework that adjusts to the new post-pandemic reality. The results obtained in the present study could suggest the convenience of creating lines of credit appropriate for each segment of tourism companies, to provide liquidity to avoid a greater number of personnel layoffs. Moreover, they also support the creation of tourist campaigns according to the production level of the tourist companies, as a strategy that seeks to motivate the travel of potential tourists, for the reactivation of the tourist industry.

The dimensionality reduction methods used in this study have facilitated the visualization of the data, obtaining better results in the PCA method according to the level of production of the tourism companies. The 2015, 2019, and 2020 variables used in the case study are highly correlated, this implies that the behavior of one variable depends on the other variables.

In turn, the clustering methods used have made it possible to group similar data, organizing them into an efficient representation that has characterized Ecuador's tourism companies. The clustering validation indices applied in this document, have allowed us to evaluate the goodness of clustering results and to define the number of appropriate clusters. It is possible to propose the identification of the services offered in each group, with the objective of creating national and international tourism promotion campaigns or creating strategies to attract foreign investment based on tourism services.

The unified analysis of the data sets corresponding to the years 2015, 2019 and 2020, allows visualizing the decrease in the operation levels of the tourism industry in Ecuador, in the years 2019 and 2020, which could be associated with external factors such as social protests and the covid-19 pandemic.

Finally, it is important to consider that the present work includes the study of the operating variables of tourism companies in Ecuador; however, future studies must analyze tourism information from other countries as part of market intelligence that provides valid information for the generation of marketing strategies.

References

[1] S. Richard, "Tourism and Development (2nd Ed)," in 1. Tourism: A Vehicle for Development?, S. Richard and J. T. David, Eds.: Channel View Publications, 2020, pp. 3-30.

[2] UNWTO, "2020: A year in review," eLibrary, [Online]. Available: https://www.unwto.org/UNWTO-academy

[3] (2020). Recovery Plan. [Online] Available: https://www.turismo.gob.ec/wp-content/uploads/2021/05/Plan-Reactivacion-Turistica-Red_compressed.pdf

[4] A. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero, "Analysis of the Tourism Industry in Ecuador by Means of Soft Computing Techniques," Cham, 2022: Springer International Publishing, pp. 811-820.

[5] G. I. Penagos-Londoño, C. Rodriguez–Sanchez, F. Ruiz-Moreno, and E. Torres, "A machine learning approach to segmentation of tourists based on perceived destination sustainability and trustworthiness," Journal of Destination Marketing & Management, vol. 19, p. 100532, 2021/03/01/ 2021.

[6] O. Anguera-Torrell, J. Vives-Perez, and J. P. Aznar-Alarcón, "Urban tourism performance index over the COVID-19 pandemic," International Journal of Tourism Cities, vol. 7, no. 3, pp. 622-639, 2021.

[7] L. Wang, "Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization," Applied Soft Computing, vol. 114, p. 108153, 2022/01/01/ 2022.

[8] L. A. Zadeh, "Soft computing and fuzzy logic," in Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi a Zadeh: World Scientific, 1996, pp. 796-804.

[9] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," J Mach Learn Res, vol. 10, no. 66-71, p. 13, 2009.

[10] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," science, vol. 290, no. 5500, pp. 2319-2323, 2000.

[11] L. Rokach and O. Maimon, "Clustering Methods," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2005, pp. 321-352.

[12] E. H. Ruspini, "A new approach to clustering," Information and Control, vol. 15, no. 1, pp. 22-32, 1969/07/01/ 1969.

[13] T. Gupta and S. P. Panda, "Clustering Validation of CLARA and *K*-Means Using Silhouette & DUNN Measures on Iris Dataset," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 10-13.

[14] H. Abdi and L. J. Williams, "Principal component analysis," WIREs Computational Statistics, https://doi.org/10.1002/wics.101 vol. 2, no. 4, pp. 433-459, 2010/07/01 2010.

[15] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press, 2018.

[16] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of *K* in *K*-means clustering," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 219, no. 1, pp. 103-119, 2005.

[17] E. Schubert and P. J. Rousseeuw, "Faster *k*-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms," Cham, 2019: Springer International Publishing, pp. 171-187.

[18] W. Revelle, "Hierarchical Cluster Analysis And The Internal Structure Of Tests," Multivariate Behavioral Research, vol. 14, no. 1, pp. 57-74, 1979/01/01 1979.

[19] P. Ryan, Euclidean and non-Euclidean geometry: an analytic approach. Cambridge university press, 1986.

[20] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient," in Knowledge and Systems Sciences, Singapore, 2019: Springer Singapore, pp. 1-17.

[21] D. L. Davies and D. W. Bouldin, "A cluster separation measure," (in eng), IEEE transactions on pattern analysis and machine intelligence, vol. 1, no. 2, pp. 224-7, Feb 1979.

[22] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," Journal of Cybernetics, vol. 3, no. 3, pp. 32-57, 1973/01/01 1973.

[23] INEC. (2015). Hotels, Restaurants and Services Survey [Online]. Available: https://www.ecuadorencifras.gob.ec/hoteles-restaurantes-y-servicios/.

[24] INEC. (2019). Business Structural Survey ENESEM [Online]. Available: https://www.ecuadorencifras.gob.ec/encuesta-estructural-empresarial-2019/.

[25] INEC, "Business Structural Survey ENESEM," 2020.

[26] N. Salehnia, N. Salehnia, H. Ansari, S. Kolsoumi, and M. Bannayan, "Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and *K*-means approaches," International Journal of Biometeorology, vol. 63, no. 7, pp. 861-872, 2019/07/01 2019.

[27] A. A. N. K. Putra, M. Nasucha, and H. Hermawan, "*K*-Means Clustering Algorithm in Web-Based Applications for Grouping Data on Scholarship Selection Results," in 2021 International Symposium on Electronics and Smart Devices (ISESD), 2021, pp. 1-6.

[28] J. M. Conejero, J. C. Preciado, A. J. Fernández-García, A. E. Prieto, and R. Rodríguez-Echeverría, "Towards the use of Data Engineering, Advanced Visualization techniques and Association Rules to support knowledge discovery for public policies," Expert Systems with Applications, vol. 170, p. 114509, 2021/05/15/ 2021.

[29] N. J. Gogtay and U. M. Thatte, "Principles of correlation analysis," Journal of the Association of Physicians of India, vol. 65, no. 3, pp. 78-81, 2017.

[30] Ministry-of-Tourism-Ecuador. (2022). Visualizer [Online]. Available: https://servicios.turismo.gob.ec/turismo-cifras.

Capítulo IV. Forecasting hotel cancellations through Machine Learning

Autores: Anita Herrera[1], Ángel Arroyo[1], Alfredo Jiménez[2] and Álvaro Herrero[1]

Afiliaciones:

[1] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela Politécnica Superior, Universidad de Burgos, Av.Cantabria s/n, 09006, Burgos, Spain, ahv1002@alu.ubu.es, {aarroyop, ahcosio}@ubu.es
[2] KEDGE Business School, 680 cours de la Liberation, Talence (Bordeaux) France. alfredo.jimenez@kedgebs.com

Resumen

Los pronósticos desempeñan un papel importante en la gestión del sector turístico, al anticipar la demanda, optimizar las operaciones y mejorar la experiencia del turista, lo que contribuye al éxito y competitividad en la industria turística. El presente estudio utiliza técnicas de clasificación para predecir cancelaciones en reservaciones de hotel, evaluando la precisión alcanzada y confirmando la eficacia de las técnicas de Machine Learning en los pronósticos del sector turístico.

Las técnicas utilizadas en este estudio incluyen Multilayer Perceptron Neural Network, Radial Basis Function Neural Network, Deep Neural Network, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier y XgBoost Classifier. La elección de las técnicas se fundamenta en su capacidad para capturar relaciones complejas entre variables, proporcionando un enfoque integral para abordar el desafío de generar mejores pronósticos.

Los datos analizados corresponden a las reservaciones realizadas en un hotel de Lisboa y un resort ubicado en la región Algarve de Portugal, durante el período comprendido entre el 1 de julio de 2015 y el 31 de agosto de 2017.

En relación al preprocesamiento del conjunto de datos, considerado como una etapa crítica e indispensable previa a la aplicación de las técnicas de Machine Learning, se realizó un Análisis Exploratorio de Datos para identificar posibles inconsistencias y definir estrategias adecuadas de corrección. A continuación, se realizó la limpieza de datos, la detección de outliers, transformación de datos y la selección de las variables significativas para el estudio. Esta fase de preprocesamiento garantiza una mayor efectividad en los modelos predictivos, al asegurar la calidad y coherencia de los datos utilizados en el análisis.

El nuevo conjunto de datos, resultado de la etapa de preprocesamiento, se divide en un conjunto de entrenamiento que está formado por el mayor porcentaje de muestras (70% - 80%) y el conjunto de prueba que abarca el porcentaje restante. La determinación del porcentaje dependerá de los resultados obtenidos en cada modelo predictivo.

Otro aspecto fundamental para mejorar la precisión de los pronósticos es la correcta configuración de los hiperparámetros de cada técnica. En el estudio se explora y selecciona los hiperparámetros más adecuados, y se emplea la Cross-Validation (CV) estratificada con *K*-Fold para evaluar la robustez y el rendimiento de los modelos.

Una vez entrenado el modelo, este se emplea para predecir las etiquetas de un nuevo conjunto de muestras pertenecientes al conjunto de prueba. A continuación, se procede a evaluar el desempeño del modelo mediante la construcción y análisis de la matriz de confusión. Esta herramienta permite determinar el número de muestras correctamente pronosticadas, lo que facilita la comprensión del comportamiento y la capacidad predictiva de cada modelo. Además, se han calculado las métricas accuracy, precision, recall, specificity y la puntuación F1, para evaluar el rendimiento de los modelos y comparar las diferentes técnicas utilizadas.

La evaluación de los modelos se complementa con un análisis visual mediante diferentes representaciones gráficas. Training and Validation Loss y Training and Validation Accuracy, proporcionan una perspectiva sobre la evolución de la pérdida y precisión del modelo durante el entrenamiento y validación. Las gráficas que ofrecen información sobre cómo varía la precisión del modelo en función de los centroides utilizados en la capa oculta. Asimismo, las gráficas que muestra la relación entre la precisión del modelo y el parámetro de dispersión (Sigma) en la Radial Base Function Neural Network. Por último, la curva ROC (Receiver Operating Characteristic) ofrece una evaluación visual de la tasa de verdaderos positivos frente a la tasa de falsos positivos en diferentes puntos de decisión. Estas representaciones gráficas ayudan en una evaluación completa de los modelos propuestos en la investigación.

Los modelos analizados en este estudio demuestran un alto nivel de precisión, con valores superiores al 90%, evidenciando su capacidad para generar pronósticos confiables para la gestión turística. Entre los modelos evaluados, se destaca XgBoost Classifier y Deep Neural Network, con los mejores resultados, representando una opción válida para los responsables de las tomas de decisiones del sector turístico.

Forecasting hotel cancellations through Machine Learning

Anita Herrera[1] [0000-0002-2655-412X], Ángel Arroyo[1][0000-0002-3561-6257], Alfredo Jiménez[2][0000-0001-7811-5113]
and Álvaro Herrero[1][0000-0002-2444-5384]

[1] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela
Politécnica Superior, Universidad de Burgos, Av.Cantabria s/n, 09006, Burgos, Spain,
ahv1002@alu.ubu.es, {aarroyop, ahcosio}@ubu.es
[2]KEDGE Business School, 680 cours de la Liberation, Talence (Bordeaux) France.
alfredo.jimenez@kedgebs.com

Abstract. Accurate and reliable forecasting of cancellations is important for successful revenue management in the tourism industry. The objective of this study is to develop classification models to predict hotel booking cancellations. The work involves a number of key steps, such as data preprocessing to properly prepare the data; feature engineering to identify relevant attributes to help improve the predictive ability of the models; hyperparameter settings of the models, including choice of optimizers and incorporation of dropout layers to avoid overfitting in the neural networks; potential overfitting is evaluated using *K*-fold cross-validation; and performance is analyzed using the confusion matrix and various performance metrics. The algorithms used are Multilayer Perceptron Neural Network, Radial Basis Function Neural Network, Deep Neural Network, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier and XgBoost Classifier. Finally, the results of all models are compared, visualizing Deep Neural Network and XgBoost as the most suitable models for predicting hotel reservation cancellations.

Keywords: Machine Learning, supervised learning, classification methods, neural networks, *k*-means, tourism industry, hotel booking.

1 Introduction

The tourism industry is on its way to pre-pandemic levels [1], a process that includes innovation, in a highly competitive industry that must adapt to the expectations and requirements of tourists. According to the World Tourism Organization, the tourism sector had shown high growth in the years before the pandemic declaration. In 2019, 1.5 billion international tourist arrivals were recorded in the world, versus 400 million in 2020. In 2022, more than 900 million tourists made international trips, which is equivalent to 63% of pre-pandemic levels [2].

The tourism industry's efforts for its revival have included creating strategies based on intelligent information. In this sense, Machine Learning (ML) models transform data into useful information for tourism companies, identifying customer demand, industry trends and future expectations. Techniques such as clustering, classification, and text mining have been used to analyze online tourism records to understand the new needs generated during the coronavirus pandemic [3]. ML models determine areas or tourism services to be improved, new demands and, in general, valid information for tourism management.

Managers or decision-makers can benefit from reliable forecasts generated from ML models that take into account the specific characteristics of the study area and management data sets [4]. A forecast of tourism demand provides information for the design and implementation of policies that provide a competitive advantage, reduce the uncertainty associated between supply and demand [5], and is an important element for government planning and management [6] and decision-makers in the tourism sector.

The tourism management of different establishments is reinforced with tourism demand forecasting models. Supervised learning methods are analyzed to identify the best-performing model according to the tourist service to be forecast [7]. Artificial neural networks (ANN) examine large amounts of data to create advanced forecasts that significantly improve management in the travel industry [8]. Likewise, ANN is used in hotel occupancy rate forecasting [9], Radial Basis Function (RBF) Neural Networks analyze online hotel bookings, forecasts that provide valid information for decision-making [10]. Other hybrid models formed with Backward Programming Neural Networks, General Regression Neural Networks, Least Squares Support Vector Regression (LSSVR), Random Forest (RF), Gaussian Process Regression (GPR), etc., reflect the validity of these techniques as valid mechanisms in hotel occupancy rate forecasting [11].

ML techniques allow to obtain highly accurate cancellation predictions, which contribute to the creation of better policies and means an important aspect due to the economic losses caused by an unsold tourist service. The research by Antonio et *al.* [12] is an experimental descriptive paper using ML models to forecast hotel reservation cancellations, with large volumes of data formed from multiple sources. The study concludes that it is possible to build ML models to predict hotel reservation cancellations with high accuracy, the best results are achieved by including variables that capture the characteristics and operating environment of each hotel. The study by Sánchez-Medina et *al.* [13] proposes to predict hotel booking cancellations using a smaller number of features, which coincide with the features most requested by hotel booking platforms. This is an experimental descriptive work using Random Forest, Support Vector Machine (SVM), C5.0, ANN methods optimized by genetic algorithms. The validation of the model performance is performed by repeated random subsampling. The model performance is presented in the ROC curve and the performance metrics Accuracy, Precision, F1 score, Specificity, Recall and AUC. The study concludes that it is possible to forecast cancellations with a smaller number of variables, obtaining an accuracy of 80% in models based on tree or SVM decisions, and 98% in ANN. The research by Sanchez et *al.* [14] analyzes personal name record (PNR) data to identify individuals who are likely to make cancellations in a short period of time before the check-in date. Different time horizons were considered in the study, ranging from 4 to 7 days before service time, and the methods used are C5.0, SVM, ANN and tree boosting ensemble. The study obtains the lowest accuracy percentages in the 4-day period (before the hotel presentation) and these correspond to 60% with ANNs and 73% with the ensemble method. This type of forecasting is very valuable for the hotel and lodging industry

In contrast to the works cited above, this study is distinguished by providing a detailed exposition of ML methods used to predict hotel reservation cancellations. Base classifiers, ensemble classifiers and neural networks are included. The importance of combining data processing techniques, model choice and performance evaluation to obtain more accurate predictions is highlighted. In contrast to the reviewed studies, this work implements a Deep Neural Network as a more powerful tool for forecasting hotel reservation cancellations. This work focuses on transparency and technical depth, including detailed data preprocessing, specific hyperparameters, confusion matrix, percentage of hits, and clear performance metrics. A thorough analysis of the results from a technical ML perspective is also performed, enriching the understanding of the application of these techniques in hotel cancellation prediction.

In summary, ML models are important for information generation, which is the basis of successful tourism management. In this paper we analyze supervised learning methods for hotel reservation prediction, using a reservation dataset from an urban hotel and a resort hotel, located in Portugal [15]. The work includes the description of data preprocessing, the identification of the most informative variables, the configuration of hyperparameters, the evaluation of *K*-fold cross-validation and the analysis of the performance of the methods from the confusion matrix, performance metrics and graphics.

The continuation of this paper includes Section 2 called *Techniques Applied*, in which different ML algorithms used for data classification are described, which will allow obtaining hotel reservation cancellation forecasts. In addition, the concept of unsupervised learning applied to the unlabeled dataset is discussed for the purpose of discovering possible patterns. A description of the indices used to evaluate the quality of the clusters is also described. Section 3 describes a *Case Study* with reservation data from two hotels in Portugal. Section 4 corresponds to *Experiments and Results*. Section 5 describes the *Practical Advantages and Limitations*. Section 6 corresponds to *Applications of Classification to Tourism Management*. Finally, Section 7 includes the *Conclusions* of the present work.

2 Techniques Applied

ML is the field of study that gives computers the ability to learn without being explicitly programmed. ML is an area focused on the use of data and algorithms to simulate human learning. Currently, several industries use ML algorithms to make classifications, predictions or discover strategic information.

Supervised learning is characterized by the use of labeled datasets that are used to train algorithms that classify data, or accurately predict outcomes. The goal of supervised learning is to predict or classify a specific result of interest [16].

Unsupervised learning uses algorithms for the treatment of unlabeled data. The goal of unsupervised learning is to analyze data to identify patterns, clusters, or relationships between features in the input data. In addition, unsupervised learning algorithms are suitable for creating labels in the data [17] and making predictions.

The algorithms used in the case study of the present work are described below:

2.1 Artificial Neural Networks (ANN)

Artificial intelligence tools whose functionality is based on biological neurons. ANNs consist of an input layer, one or more hidden layers and an output layer [18]. Training data is used to tune the model, which is then used to solve various problems accurately.

Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is an ANN, consisting of multiple hidden layers, which is characterized by the ability to learn more complicated functions. The MLP Neural Network generates a predictive model for one or more dependent variables based on the predictor variables. The input data, called features, correspond to the study variables.

The model multiplies each feature by its weight, then the weighted features are summed to obtain the scalar product, and the latter is added to the bias that feeds the activation function as the output of the neuron. MLP can approximate smooth and measurable functions by selecting the appropriate connection weights and transfer functions [19].

MLP network operation is expressed by

$$Z_k = \sum_{j=1}^{q} w_{kj}^{t}\, y_j - \theta_k^t = \sum_{j=1}^{q} w_{kj}^{t}\, f\left(\sum_{i=1}^{n} w_{ji}\, x_i - \theta_j\right) - \theta_k^t$$

Where:

x_i = input MLP, y_j = hidden layer outputs, Z_k = final layer results, w_{ij} = hidden layer weights, θ_j = hidden layer thresholds, w_{kj}^{t} = output layer weights, θ_k^t = output layer thresholds.

Radial Basis Function Neural Networks (RBF Neural Networks)

Three-layer neural network, an input layer receives the data and transmits it to the next layer without any processing, a hidden layer that performs the computations, and an output layer that performs prediction tasks such as classification or regression [20].

The operation of an RBF Neural Network is expressed by:

$$\emptyset_i(x) = \varphi((x - c_i)^T . R^{-1}(x - c_i))$$

Where:

$\emptyset$ = radial basis function used, c_i = set of radial basis function centers, $(x - c_i)^T . R^{-1}(x - c_i)$= distance from input x to center c.

Neurons in the hidden layer are activated by a radial function, in a different region of the input pattern space, there are different types of radial basis functions, but the most commonly used is the Gaussian function. The output layer neurons perform a linear combination of the activations of the hidden neurons. For the Gaussian function and the Euclidean metric, the output of the network is given by:

$$F(x) = \omega_o + \sum_{i=1}^{c} \omega_i \, . \, exp\left(-\frac{\|x - c_i\|^2}{r^2}\right)$$

Where:

C = number of radial basis functions used, c_i = synaptic weights, $\|.\|$ = standard Euclidean, r = radius of radial function, ω = weights.

Deep Neural Network (DNN)

Powerful and efficient tool for solving complex machine learning tasks. DNNs are identified as consisting of multiple hidden layers between the input layer and the output layer. DNNs use optimization algorithms to adjust the weights and biases of the connections between neurons to minimize a loss function [21].

Each neuron in a hidden layer performs a linear combination of the outputs of the previous layer, followed by the application of a nonlinear activation function. For a neuron in the hidden layer *j* of layer *l*:

Linear combination weighted: $Z_j^{(l)} = \sum_{(l)^{i=1}}^{n^{(l-1)}} W_{ji}^{(l)} * a_i^{(l-1)} + b_j^{(l)}$

Application of activation function: $a_j^{(l)} = f(z_j^{(l)})$

Where:

$Z_j^{(l)}$ = weighted input to neuron "j" in layer "l", $a_i^{(l-1)}$ = output of neuron "i" in the previous layer "*l-1*", $W_{ji}^{(l)}$ = weights associated with the connection between neuron "i" and "j" in layer "l", $b_j^{(l)}$ = bias of neuron "j" in layer "l", f(·)= nonlinear activation function.

The process is repeated in each neuron through the network, until it reaches the output layer. The output layer performs a weighted linear combination of the outputs of the previous layer and applies an appropriate activation function, to perform the prediction. DNN uses optimization algorithms to minimize the loss function.

2.2 Classification Algorithms

Classification is a supervised learning algorithm that aims to make predictions about the classifications of input data. A classification algorithm learns from training data and then evaluates the test data, and finally uses it to make predictions on new data [22].

Decision Tree Classifier

A decision tree is a supervised algorithm that divides the data into homogeneous groups according to the most significant input variable. This algorithm presents a hierarchical tree structure, consisting of a root node, branches, internal nodes, and leaf nodes. The classification starts at the root node of the tree and progresses recursively up to the leaf nodes. The leaf nodes represent all possible results within the dataset [23].

Random Forest Classifier

The Random Forest Classifier is based on decision trees, where each tree is trained with a subset of the dataset and gives a result, then the results of the decision trees are combined to obtain a more reliable result. Random Forest is essentially a collection of Decision Trees. The accuracy of Random Forest Classifier has generated the attention of researchers [24].

AdaBoost Classifier (Adaptive Boosting)

The AdaBoost algorithm trains single classifiers iteratively, cascading multiple classifiers so that each classifier focuses on data that were misclassified by the previous classifier. The algorithm achieves better results and more accurate classification [25].

XgBoost Classifier

XgBoost is the abbreviation for eXtreme Gradient Boosting, a supervised machine learning method for classification and regression based on Decision Trees and an efficient improvement over Random Forest and Gradient Boosting methods. XgBoost combines trees sequentially to learn from the result of the previous trees and correct the error until it can no longer be corrected. XgBoost is an adaptation of Gradient boosting that stands out for its efficiency and speed [26].

2.3 Clustering

Clustering is an unsupervised learning technique whose objective is to find homogeneous subgroups in the data set. The definition of the subgroups takes into account the distances between observations so that the elements of a cluster are more similar than the elements of other clusters [27].

K-means

K-means is an unsupervised learning algorithm that attempts to cluster unlabeled data into *k* well-defined clusters. The *k*-means algorithm requires setting the number of clusters and will assign each data to a cluster based on its characteristics. The goal of *k*-means is to define the centroid and members of each cluster by minimizing the distance from each point to the cluster centroid, a process that is repeated until the centroid does not change significantly [28].

The sum of the distance of each point to the centroid of its cluster is given by:

$$SSE = \sum_{j=1}^{k} \sum_{x\in\, Gi} \frac{p(X_i,C_j)}{n}$$

Where *p* is the proximity function, *k* is the number of groups, C_j is the number of centroids and *n* is the number of rows.

2.3.1 Cluster Validation Metrics

Metrics that assess the quality of clustering by examining how separate the clusters are and how compact they are, i.e., looking for clusters with a high similarity between their components and low similarity between clusters [29].

Silhouette Index

Silhouette is a metric that validates the coherence of data clusters by analyzing the distance between points in each cluster (cohesion) and the distance to other clusters (separation). The Silhouette index can take values in the interval [-1, 1] and the optimal number of clusters is given by the highest value of the Silhouette index, which is given by [30]:

$$Sil(C) = \frac{1}{N} \sum_{C_k \in C} \sum_{X_i \in C_k} \frac{b\ (X_i, C_k) - \mathrm{a}\ (X_i, C_k)}{\max\{a(X_i, C_k), b(X_i, C_k)\}}$$

Where:

$$a(X_i, C_k) = \frac{1}{|C_k|} \sum_{X_j \in C_k} d_e(X_i, X_j)$$

$$min_{C_l \in C \setminus C_k} \left\{ \frac{1}{|C_l|} \sum_{X_j \in C_l} d_e(X_i, X_j) \right\}$$

Dataset $X = \{X_1, X_2, \ldots X_N\}$ of N objects, C_k - centroid of cluster, The partition X into K groups: $C = \{c_1, c_2, \ldots, c_k\}$, $d_e(X_i, X_j)$ – Euclidean distance.

Davies Building Index

The Davies Building index evaluates clustering algorithms in terms of the minimum distance between the points of a cluster and its centroid, and the distance between the centroids. The appropriate number of clusters, according to the David Boulding index, corresponds to the smallest value [31]. The Davies Building index is defined as:

$$DB(C) = \frac{1}{K} \sum_{C_k \in C} max_{C_l \in C \setminus C_k} \left\{ \frac{S\,(C_k) + \ S(C_l)}{d_e\,(C_k,\ C_l)} \right\}$$

Where:

$$S(c_k) = \frac{1}{|C_k|} \sum_{X_i \in C_k} d_e(X_i, C_k)$$

Dataset $X = \{X_1, X_2, \ldots X_N\}$ of N objects, C_k - centroid of cluster, the partition X into K groups: $C = \{c_1, c_2, \ldots, c_k\}$, $d_e(X_i, X_j)$ – Euclidean distance.

Elbow Method

The elbow method evaluates the appropriate number of clusters (k) by calculating the sum of the squared distances from each cluster object to its centroid (WCSS). The graph of WCSS-versus-k, changes drastically at one point, creating the shape of an elbow. The optimal value of the cluster corresponds to the point in the graph from which it shifts parallel to the x-axis [32]. WCSS is given by:

$$WCSS = \sum_{i=1}^{N_c} \sum_{x \in C_i} d\left(x, \overline{x}_{C_i}\right)^2$$

Where: C_i = cluster, N_c = # clusters, $\overline{x}_{C_i}$= cluster centroid.

3 Case Study

This paper presents a set of ML methods to analyze the booking behavior of hotel reservations. For this purpose, a dataset of bookings from a resort hotel located in the tourist region of Algarve, and a hotel in the city of Lisbon (Portugal), is used the dataset includes variables related to the bookings that were due to arrive between July 1, 2015, and August 31, 2017. The analyzed dataset has a total of 119,210 records [15].

The variables considered for analysis are:

1. hotel: Categorical variable that takes the value of 1 for the city hotel, and 0 for the resort hotel.
2. is_canceled: Categorical variable that identifies canceled bookings as 1 and non-canceled bookings as 0.
3. lead_time: Numerical variable that represents the number of days between the booking date and the arrival date.
4. arrival_date_month: Numerical variable representing the month of arrival.
5. arrival_date_year: Numerical variable representing the year of arrival.
6. arrival_date_week_number: Numerical variable representing the week number of the arrival date.
7. arrival_date_day_of_month: Numerical variable representing the day of the month of the arrival date.
8. stays_in_weekend_nights: Numeric variable with the number of weekend nights the guest has stayed or booked.
9. stays_in_week_nights: Numerical variable representing the number of nights between Monday to Friday, the guest stayed or booked.

10. adults: Numeric variable of the number of adults in the booking.
11. children: Numeric variable of the number of children in the booking.
12. babies: Numeric variable of the number of babies in the booking.
13. meal: Categorical variable representing the type of food booked.
    0: BB – Bed & Breakfast; 1: FB – Full Board; 2: HB – Half Board; 3: SC – Self-Catering; 4: Undefined.
14. market_segment: Categorical variable representing market segment designation:
    0: Direct; 1: Corporate; 2: Online TA; 3: Offline Travel Agents/Tour Operators; 4: Complementary; 5: Groups; 6: Undefined; 7: Aviation.
15. distribution_channel: Categorical variable representing the booking distribution channel: 0: TA – Travel Agents; 1: TO – Tour Operators.
16. is_repeated_guest: Categorical variable with 1 if the booking name is of a repeated guest and 0 if it is not repeated.
17. previous_cancellations: Numerical variable representing the number of previous bookings canceled by the client before the current booking.
18. previous_bookings_not_canceled: Numerical variable representing the number of previous bookings not canceled by the client before the current booking.
19. reserved_room_type: Categorical variable of room type code booked.
20. deposit_type: Categorical variable indicating whether the client deposited to guarantee the booking.
    - 0: No Deposit – no deposit was made;
    - 1: Refundable – a deposit was made with a value under the total cost of the stay;
    - 2: Non Refund – a deposit was made in the value of the total stay cost.
21. agent: Numeric variable that corresponds to the ID of the travel agency that made the booking.
22. company: Numeric variable that corresponds to the ID of the company/entity that made the booking or is responsible for the payment of the booking.
23. customer_type: Categorical variable of booking type.
    - 0: Transcient – the booking is not part of a group or contract, and is not associated with another transient booking;
    - 1: Contract – the booking is associated with an award or other type of contract;
    - 2: Transient party – the booking is transient, but is associated with at least one other transient booking;
    - 3: Group – the booking is associated with a group.
24. adr: Numerical variable that corresponds to Average Daily Rate, calculated by dividing the sum of all lodging transactions by the total number of staying nights.
25. required_car_parking_spaces: Numerical variable related to the number of car parking spaces required by the customer.
26. total_of_special_requests: Numerical variable that corresponds to a number of special requests made by the customer.

Data preprocessing is a critical phase in building ML models. Dealing with data inconsistencies is an important process to ensure that models are robust and accurate [33]. The choice of specific methods depends on the nature of the data and the characteristics of the problem addressed. It is important to perform a detailed Exploratory Data Analysis (EDA) to identify inconsistencies before applying appropriate correction strategies. The following preprocessing techniques are employed in the present study:

- Data cleaning: identifies and corrects problems in the dataset, which may affect the quality and reliability of the analysis result. In the initial data set, composed of 32 variables, 4 variables with null values are identified and corrected. The null values of the children column are replaced by 0, because the average of the same column is almost 0. The null values of the country attribute are replaced with the text "Unknown". In addition, the null values of the attributes agent and company are replaced by 0.
- Outliers identified visually are eliminated. Also, business rules are validated, for example the rows in which null values were found for the fields adults, children and babies, at the same time, have been eliminated. Likewise, records that at the same time have a value of 0 in the field week_ends_nights_stays and in the field week_nights_stays, are eliminated since they are a situation that cannot happen.

- Data transformation: includes different operations to improve data quality. The categorical features in the dataset are transformed into numerical features, as indicated in the description of the variables. The data are then adjusted to the same specific scale or range by normalization.
- Feature selection: Evaluate the importance or relevance of each characteristic in relation to the target variable, using feature selection methods. Feature selection is one of the tasks necessary to achieve a robust and efficient classification model at training time [34].
- Data splitting: Division of the dataset into training set (largest percentage, commonly 70% or 80%) and testing set (with the remaining percentage). It is important to verify that the training dataset presents a balance of the classes of interest (non-canceled and canceled) [35].

Fig. 1 shows the treatment of data from raw data to ready to use data for training.

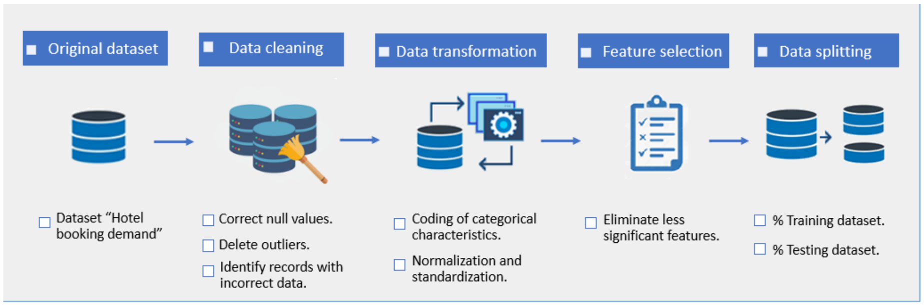
■ Original dataset
■ Data cleaning
■ Data transformation
■ Feature selection
■ Data splitting
□ Dataset "Hotel booking demand"
□ Correct null values.
□ Delete outliers.
□ Identify records with incorrect data.
□ Coding of categorical characteristics.
□ Normalization and standardization.
□ Eliminate less significant features.
□ % Training dataset.
□ % Testing dataset.

Fig. 1 Data Treatment

Effective data preprocessing is a key component in improving the performance of classification models. The combination of data processing techniques, hyperparameter definition, appropriate model choice and performance evaluation are combined in a comprehensive approach to obtain better forecasts.

4 Experiments and Results

The design of a classification model includes different steps and key decisions to achieve a model capable of making the most accurate predictions on the data. The present work begins with the application of the MLP Neural Network method on the dataset, which was previously preprocessed.

The model was trained with different hyperparameters until the model performance was optimized. The hyperparameters that generate the best performance for the MLP are as follows: the dataset is divided into 70% for training and 30% for testing, the network accepts 26 inputs, has a hidden layer with 100 nodes and Relu activation function, the second hidden layer with 100 neurons and Relu activation function, and the output layer with one neuron and sigmoid activation function. The backpropagation algorithm is used to train the MLP.

A representative and quality training data set ensures that classification models can generalize well to new data. A fragment of the input and output training data in MLP is presented in Table 1:

Table 1. Input and output training data - MLP Neural Network

is_canceled	**hotel**	**meal**	**…**	**required_car_parking _spaces**	**total_of_special_ requests**
0	1	0	…	0	0
0	1	0	…	0	1
1	1	0	…	0	0
0	0	0	…	0	2
…	…	…	…	…	…
0	0	2	…	0	0
0	1	0	…	0	1
1	1	0	…	0	0

To evaluate the robustness and performance of the classification methods, stratified cross-validation with *K*-Fold was used. The objective of cross-validation is to ensure that the results obtained in classification models are independent of the partition between training data and validation data. This concept is widely used in models generated in AI projects [36]. The MLP method has a stratified *K*-Fold Cross-Validation of 0.98061034, indicating high model performance.

The confusion matrix is obtained as follows. The confusion matrix of the MLP is presented in Table 2 to visualize the performance of the algorithm. The columns of the confusion matrix represent the number of predictions of each class, while the rows represent the instances in the actual class.

Table 2. Confusion Matrix – MLP Neural Network

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	21,940	577
	Canceled	138	13,108

According to the confusion matrix in Table 2:
21,940 reservations were not canceled and the model predicted that they would not be canceled.
13,108 reservations were canceled and the model predicted they would be canceled.
577 reservations were not canceled and the model predicted they would be canceled.
138 reservations were canceled and the model predicted that they would not be canceled.

In summary, MLP model correctly predicted 97.44% of non-canceled bookings and 98.96% of canceled bookings. The overall accuracy of the model is 98% [37].

Fig. 2 shows the Training and Validation Loss curve of the MLP model.

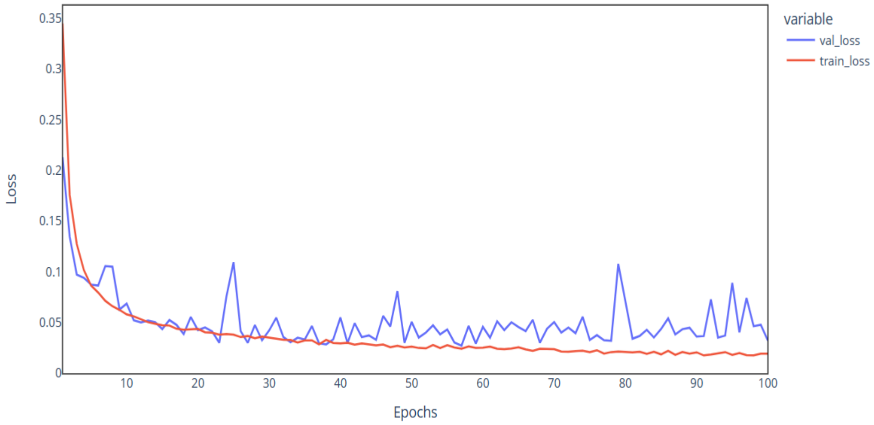

Fig. 2 Training and Validation Loss – MLP Neural Network

Fig. 2 presents the behavior of the MLP model during training and validation, and visualizes the loss of information, calculated from the training and validation dataset. The x-axis corresponds to the iterations and the y-axis to the losses. Fig. 2 shows that with each iteration, the losses decrease. This paper discusses the classification MLP, which learns from training data to model the relationships between input and output data and predict which hotel bookings will and will not be canceled.

Fig. 3 shows the Training and Validation Accuracy curve of the MLP model.

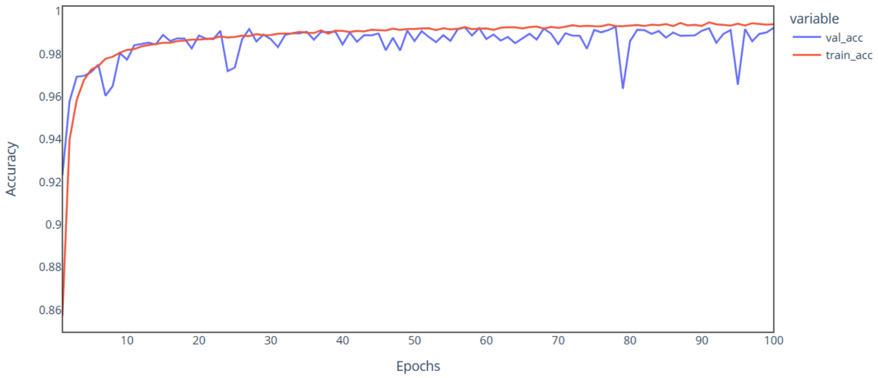

Fig. 3 Training and Validation Accuracy – MLP Neural Network

Fig. 3 shows the performance of the MLP model, it determines that the model is learning correctly and that there is no overfitting.

The next method to be analyzed is RBF Neural Network. The hyperparameters used in this model are: the dataset is divided into 80% for training and 20% for testing. The RBF includes an input layer with 26 neurons. A hidden layer consisting of 100 radial neurons, each of which applies a Gaussian radial basis function. In the output layer, the Random Forest classifier is used for final classification. The RBF model has a stratified *K*-Fold Cross-Validation of 0.94916532.

Table 3 presents the confusion matrix of the RBF Neural Network:

Table 3. Confusion Matrix – RBF Neural Network

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	21,501	972
	Canceled	3,664	9,626

The RBF model correctly predicted 95.67% of non-canceled bookings and 72.43% of canceled bookings. The overall accuracy of the model is 87.04%.

Fig. 4 shows the Accuracy as a Function of Centroid Number curve of the RBF model, which provides information for tuning and optimizing an RBF Network by varying the number of centroids used. It helps to find a balance between model performance and model complexity.

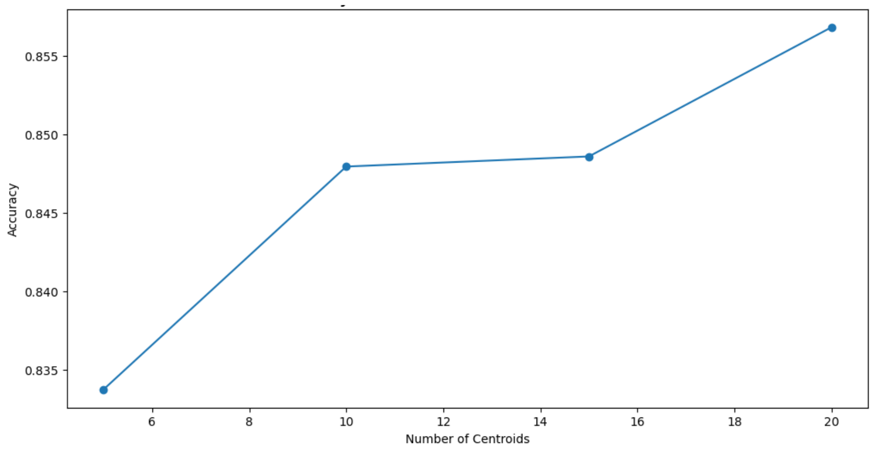

Fig. 4 Accuracy as a Function of Centroid Number

Fig. 4 shows that by setting the number of centroids to 20, the model achieves a performance higher than 85%.

Fig. 5 corresponds to the Precision as a Function of Sigma graph, which helps to find the optimal value of Sigma (parameter σ used in radial basis functions). The graph helps to understand how different values of Sigma affect the accuracy and generalizability of the model.

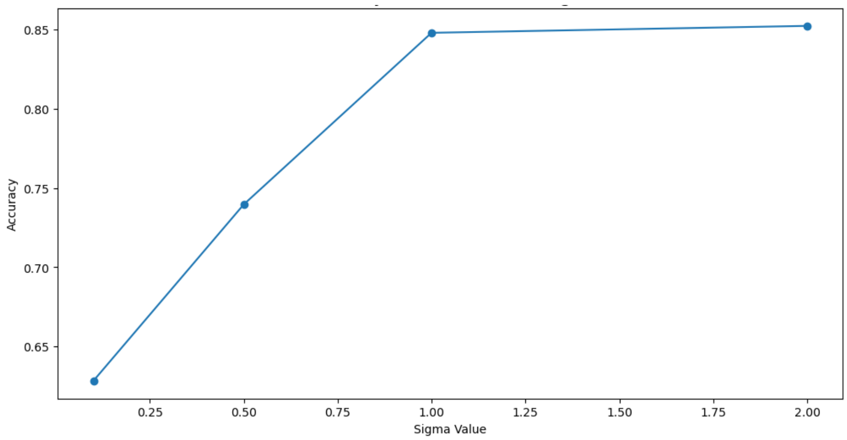

Fig. 5 Precision as a function of Sigma

Fig. 5 shows that the accuracy reaches a value higher than 85% when sigma takes the value 2.

The RBF network developed in this case study is a classification network, whose supervised phase training adjusts the weights and thresholds of the output layer to obtain the best possible forecast of hotel booking cancellations.

Next, we discuss the advantages of a DNN, which is compared to the neural network models described earlier in this paper [38]. DNN incorporates dropout layers to avoid overfitting in neural networks. The hyperparameters for configuring the DNN include a Sequential type network, with an input layer of 26 neurons according to the input features. The first hidden layer has 256 neurons and uses the Rectified Linear Unit (ReLU) activation function. In addition, a dropout layer with a rate of 0.5 is included, which means that randomly 50% of the neurons will be deactivated during training to avoid overfitting. The second hidden layer has 128 neurons and uses the ReLU activation function. As in the previous layer, a deactivation layer with a rate of 0.5 is included. The third hidden layer has 64 neurons and uses the ReLU activation function. Again, a dropout layer with a rate of 0.5 is applied. The output layer has one neuron and uses the sigmoid activation function for binary classification.

The ReLU activation function is used in ANN and deep learning models. The ReLU function generates an output equal to zero when the input is negative and an output equal to the input when the input is positive. This means that the ReLU function introduces nonlinearity into the neural network, which is important for the network to learn complex relationships and patterns in the data [39]

The DNN model has a stratified *K*-Fold Cross-Validation of 0.95292958. Table 4 shows the confusion matrix of the DNN model:

Table 4. Confusion Matrix – Deep Neural Network

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	14,944	14
	Canceled	412	8,472

DNN model correctly predicted 99.91% of non-canceled bookings and 95.36% of canceled bookings. The overall accuracy of the model is 98.21%.

Fig. 6 shows the Training and Validation Loss curve of the DNN model.

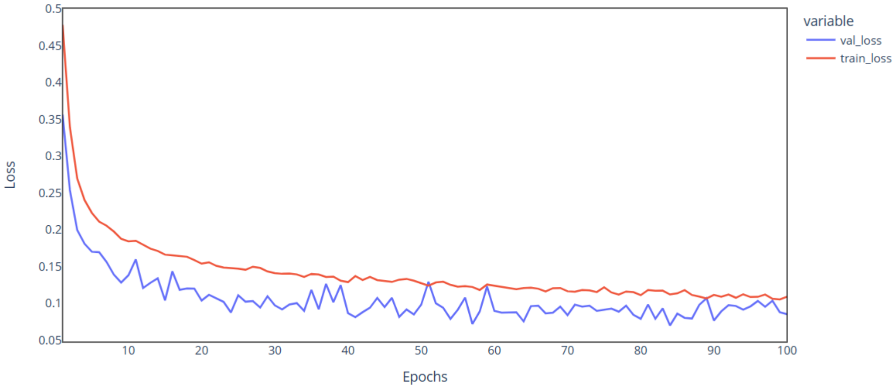

Fig. 6 Training and Validation Loss – Deep Neural Network

Fig. 6 presents the performance of the DNN model during training and validation. Fig. 6 shows that the model is learning well and generalizing appropriately to unseen data.

Fig. 7 shows the Training and Validation Accuracy curve of the DNN model:

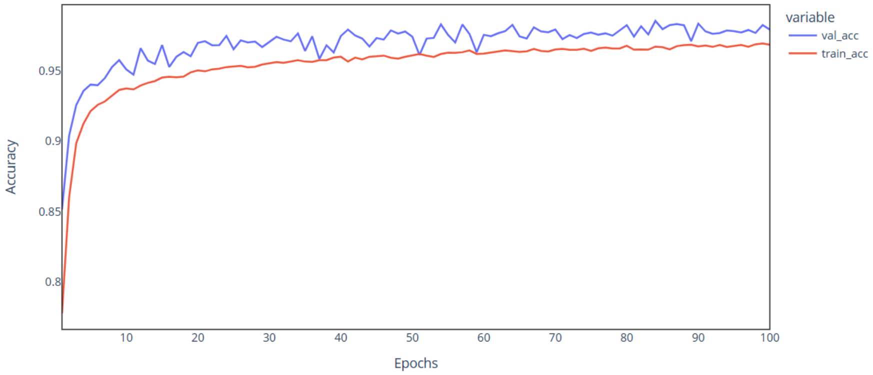

Fig. 7 Training and Validation Accuracy – Deep Neural Network

Fig. 7 shows that the model is learning effectively and is not experiencing significant overfitting.

This paper also analyzes the Decision Tree Classifier model. In this model, the data set is divided into 70% for training and 30% for testing. The hyperparameters of the Decision Tree classifier include the Gini index as the splitting criterion, the best splitter, a minimum number of samples to split a node with a value of 2, a minimum number of samples in a leaf with a value of 1. In addition, it considers all features for splitting (no feature restriction), and no random seed (random_state) is set. The Decision Tree Classifier model has a stratified *K*-Fold Cross-Validation of 0.94373667.

Table 5 corresponds to the confusion matrix and presents the performance of the Decision Tree Classifier.

Table 5. Confusion Matrix – Decision Tree Classifier

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	21,581	957
	Canceled	972	12,253

The Decision Tree model correctly predicted 95.75% of non-canceled reservations, and 92.65% of canceled reservations.

The next method to be analyzed is Random Forest. This model considers 70% for training and 30% for testing, and include the use of the Gini criterion for splits, the lack of a maximum depth in trees to allow full expansion, the requirement of at least 2 samples in a node to split it, and the guarantee that leaves contain at least 1 sample after a split. In addition, the maximum number of features considered in a split is automatically adjusted to suit the context. Parallel processing takes advantage of processor cores. No weights are assigned to classes to deal with imbalances and no iterative approach is used to aggregate trees. Detailed information during training is kept to a minimum.

Table 6 shows the confusion matrix of the Random Forest method:

Table 6. Confusion Matrix – Random Forest

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	22,335	203
	Canceled	1496	11,729

The Random Forest model correctly predicted 99.10% of non-canceled bookings and 88.69% of canceled bookings.

The method to be analyzed in the following is AdaBoost Classifier. The dataset is divided into 70% for training and 30% for testing. The AdaBoost classifier uses the Decision Tree as the base classifier, it automatically adjusts the weights of the instances during training. In addition, the learning rate for weighting the contribution of each classifier is 1. The SAMME.R algorithm is used to update the weights of the instances. The AdaBoost Classifier model has a stratified *K*-Fold Cross-Validation of 0.94207105.

Table 7 presents the confusion matrix of the AdaBoost Classifier method:

Table 7. Confusion Matrix – AdaBoost Classifier

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	21,598	940
	Canceled	965	12,260

The AdaBoost model correctly predicted 95.83% of non-canceled bookings and 92.70% of canceled bookings.

The last method considered for this analysis is XgBoost Classifier. In this model, the dataset is divided into 70% for training and 30% for testing. The hyperparameters of the XgBoost classifier include the use of the tree-based reinforcement type (gbtree), the learning rate that controls the contribution of each tree to the modeling is set to 0.1. In addition, the maximum depth of each tree is set to 5 and the number of trees to be constructed is set to 180. The XgBoost Xlassifier has a stratified *K*-Fold Cross-Validation of 0.98180881.

The confusion matrix of the XgBoost Classifier is presented in Table 8:

Table 8. Confusion Matrix – XgBoost Classifier

		Prediction	
		Non-canceled	**Canceled**
True class	Non-canceled	22,526	12
	Canceled	596	12,629

The XgBoost model correctly predicted 99.95% of non-canceled bookings and 95.49% of canceled bookings.

Fig. 8 presents the ROC curve (ratio between the true positive rate and the false positive rate) of all the models analyzed in this work:

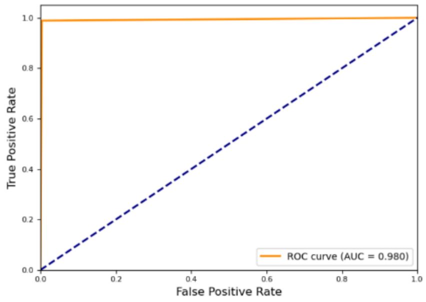

Fig. 8(a) ROC Curve – MLP

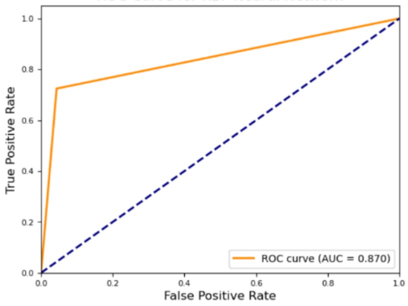

Fig. 8(b) ROC Curve – RBF

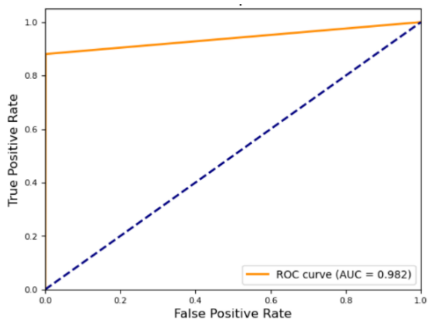

Fig. 8 (c) ROC Curve – DNN

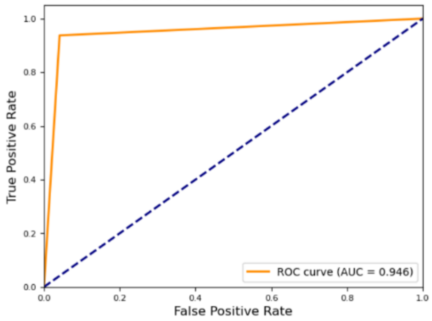

Fig. 8 (d) ROC Curve – Decision Tree

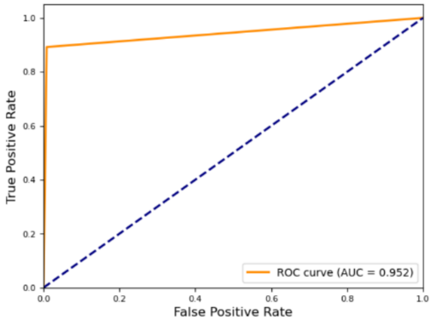

Fig. 8 (e) ROC Curve – Random Forest

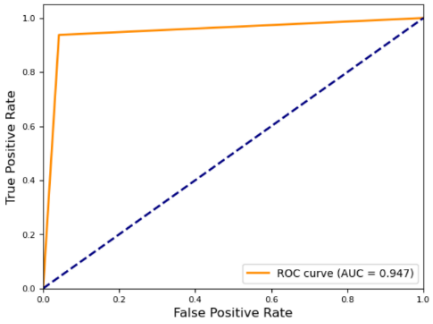

Fig. 8 (f) ROC Curve – Ada Boost Classifier

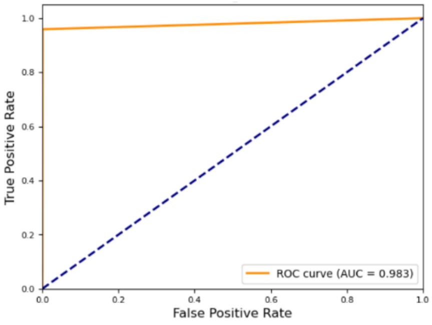

Fig. 8 (g) ROC Curve - XgBoost Classifier

Fig. 8 indicate that the models work very well for binary classification and are very effective in discriminating between the class of non-canceled and canceled reservations.

Table 9 summarizes the evaluation of the above models:

Table 9. Summary of model's evaluation

	Correct prediction	
Model	**Non-canceled**	**Canceled**
MLP	97.44%	**98.96%**
RBF	95.67%	72.43%
Deep	**99.91%**	95.36%
Decision Tree	95.75%	92.65%
Random Forest	**99.10%**	88.69%
AdaBoost	95.83%	92.70%
XgBoost	**99.95%**	95.49%

Table 9 details the correct forecast of the supervised learning models for each category (non-canceled and canceled), analyzed in this document. The DNN and XgBoost models show the highest percentage, provide the highest accuracy for this prediction task. The application of these models would improve the prediction of cancellations, providing important information for decision-making.

The performance of the classification models is obtained from the following metrics: Accuracy, Precision, Recall, Specificity, F1. Table 10 shows the performance metrics of the models analyzed in this work.

Table 10. Performance metrics

METHOD	ACCURACY	PRECISION	RECALL	SPECIFICITY	F1
MLP	**0.98000727**	0.95783704	**0.98958176**	0.97437492	0.97345067
RBF	0.87036881	0.90828458	0.72430398	0.95674810	0.80592766
Deep	**0.98213237**	**0.99835022**	0.95362449	**0.99906405**	0.97547496
Decision Tree	0.94606157	0.92755488	0.92650284	0.95753838	0.92702856
Random Forest	0.95249228	0.98298692	0.88688091	**0.99099298**	0.93246412
Ada Boost	0.94673266	0.92878788	0.92703214	0.95829266	0.92790918
XgBoost	**0.98299919**	**0.99905071**	0.95493384	**0.99946757**	0.97649424

Table 10 lists the performance metrics of the models analyzed in this paper. The DNN and XgBoost Classifier models present the highest overall metrics.

It is also interesting to analyze the reservation dataset with an unsupervised learning method to identify possible patterns, and it is part of classification approaches presented in different researches. For example, research by Onan *et al.* [40] describes a hybrid set pruning approach for sentiment classification in texts. This approach uses clustering and random search techniques. Similarly, Onan [41] presents an approach on topic extraction from scientific literature in bibliometric studies building on word embedding and cluster analysis.

By applying the *k*-means method it is possible to visually identify the appropriate number of clusters and, on the other hand, it is also possible to determine the optimal number of clusters, with methods such as the Silhouette Coefficient, Davies Building Index, Elbow Method, etc. The result of the application of the methods to determine the value of *k* in the study dataset is presented below.

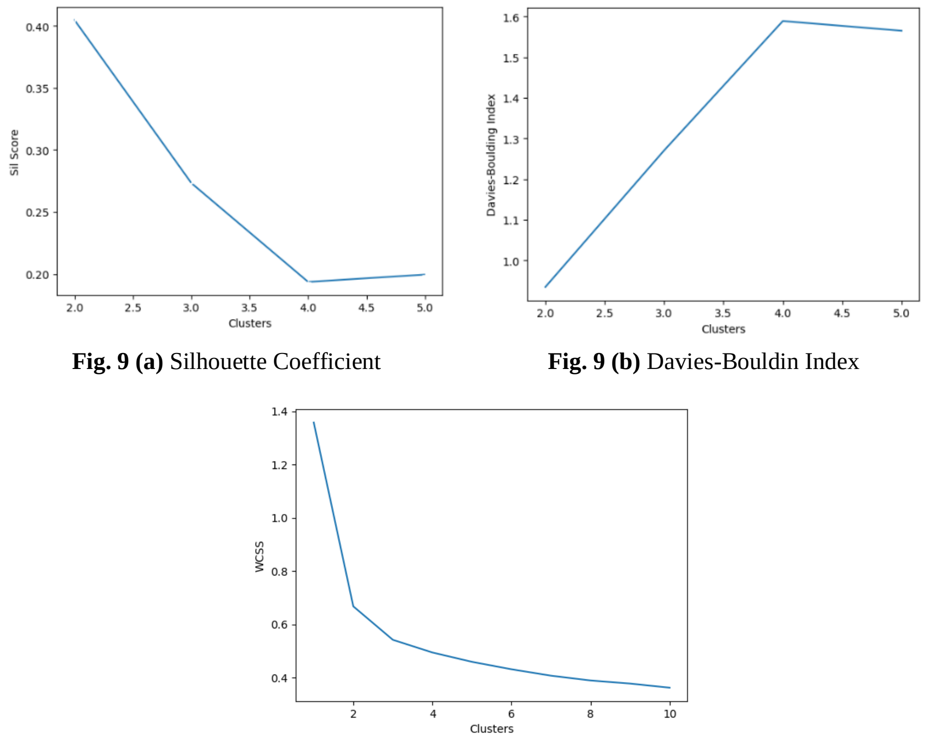

Fig. 9 (a) Silhouette Coefficient **Fig. 9 (b)** Davies-Bouldin Index

Fig. 9 (c) Elbow Method

In Fig. 9 (a) the Max Silhouette Score is reached at *k* = 2, so the optimal number of clusters is 2. The Davies-Boulding index (Fig. 9 (b)) takes the lowest value at *k*=2. Therefore, the optimal number of clusters according to the Davies-Boulding index is 2.

Finally, the optimal value of clusters according to the Elbow Method is 2 (Fig. 9 (c)).

The result of *k*-means (based on Euclidean distance) applied to the booking dataset is shown in Fig. 10.

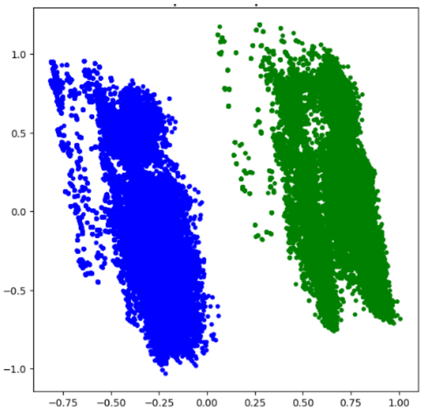

Fig. 10 *k*-means with *k* = 2

Fig. 10 presents two defined and separate clusters. The blue cluster is related to urban hotel bookings and includes the largest number of items, 79,163 bookings, with higher dispersion, implying lower similarity. The mean of the components is (-0.35; -0.06).

The green group is related to resort hotel bookings, with the lowest number of elements, 40,047 bookings, with the lowest dispersion, implying the highest similarity. The mean of the components is (0.68; 0.12).

The definition of clusters could be related to the difference between the characteristics of city hotel bookings and the characteristics of resort hotel bookings. This type of analysis is important in decision-making and valid for generating strategies. For example, this information could be useful for the creation of a cancellation policy according to the specific characteristics of the tourism business. An appropriate cancellation policy should attract the tourist's interest in booking the service, offer resale options for a canceled service and avoid economic losses.

5 Practical Advantages and Limitations

In the context of tourism management, forecasts provide important information for decision making, anticipation and effective management of resources in the face of possible cancellations. The application of classification methods for forecasting tourism cancellations has several advantages. Deep Learning models, with their ability to recognize hierarchical representations, can discover significant features to improve forecast accuracy. Decision Trees and Random Forests are easy to interpret and useful for understanding the factors influencing cancellations. In addition, AdaBoost and XGBoost combine models, improving robustness and performance.

Although the models considered in this work offer advantages in the generation of forecasts, it is necessary to take into account possible disadvantages when choosing and applying the models. RBFs are prone to overfitting, so hyperparameters and training data must be carefully defined. Deep learning models are more powerful, which implies higher computational resources. Decision trees are prone to instability in the face of noisy data. Random Forests can be computationally expensive in training. AdaBoost and XGBoost models can be sensitive to outliers.

In summary, the models analyzed in this work present important benefits for the generation of forecasts; however, it is essential to consider an effective preprocessing of the data and the correct definition of the hyperparameters for each model.

6 Applications of Classification to Tourism Management

Classification methods are fundamental tools in tourism management, as they allow the organization and analysis of large volumes of data, generating valuable information that contributes to strategic decision making. In this area, the processing of textual data sets, such as hotel reviews, travelers' comments or content on social networks, makes it possible to adapt the tourism offer to the needs and preferences of users, improving competitiveness and customer satisfaction in the tourism industry [42].

In social networks, people can express negative feelings using words with positive meanings and vice versa. Correct identification of messages will influence predictive performance. In this context, research presents deep learning-based schemes for sarcasm identification [43, 44].

For text analysis, the most common neural network architectures include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN). In study of [45], a hierarchical graph-based text classification framework with contextual node embedding is proposed. Other studies analyze schemes for classifying text according to genre, which can be useful for classifying comments according to the tourism service to which they correspond [46].

7 Conclusions

This paper attempts to review the efficiency of classification-oriented supervised and unsupervised learning models, demonstrating the importance of the ML methods in the processing of hotel bookings, due to the high accuracy they achieve. Therefore, the ML methods represent an opportunity to generate information for the creation of cancellation policies that attract more tourists and strengthen the management of the tourism sector.

The models are evaluated with the training data set (corresponding to 70% or 80% of the records) and the validation data set (corresponding to the remaining value of records). To properly evaluate the models' capabilities, the cross-validation technique is used.

When comparing the accuracy of the methods analyzed in this paper, it is determined that the XgBoost Classifier and DNN have the best Accuracy Score in Table 10, achieving better predictions of hotel reservation cancellations (Table 9). These results confirm the ability of these methods to capture complex relationships in the data.

DNN are well suited for predicting hotel reservation cancellations due to their ability to capture complex relationships in the data.

The methods based on Decision Trees enhance predictive models with high accuracy, stability, and ease of interpretation. The supervised learning models analyzed in this study are more than 98% accurate, generating reliable and useful forecasts to support decision-making.

RBF is one of the rare but extremely fast, effective, and intuitive machine learning algorithms.

The choice between DNN, MLP or RBF depends on the nature of the data, the complexity of the problem and the resources available. In general, DNNs are powerful and versatile, but they can also be computationally expensive and require large training datasets. On the other hand, MLP and RBF networks may be more appropriate in scenarios where simplicity and interpretation are priorities, or when little data is available.

The *k*-means method applied to the reservation data set (excluding the label), presents two compact and well-defined clusters, as shown in Fig. 10. The members of each cluster are fairly close to the other members of the same cluster, and the clusters are separated from each other. The clusters could be related to the characteristics of the reservations, depending on the type of hotel.

Fig. 10 presents the clusters identified with the *k*-means method, which are related to the booking characteristics of city hotels and resort hotels. Cluster analysis is useful for decision making and strategy generation. Cluster analysis provides valuable information on customer behavior, preferences and trends, enabling companies to adapt their strategies and services.

Forecasting is a relevant tool to reduce uncertainty in decision-making in tourism management, which is why ML methods have participated in the innovation and growth of this industry. This paper aims to highlight the high performance of the ML methods analyzed, which provide valid information for decision-making.

References

[1] UNWTO. (2023). Tourism set to return to pre-pandemic levels in some regions in 2023 [Online]. Available: https://www.unwto.org/taxonomy/term/347.

[2] UNWTO. (2023). Impact assessment of the covid-19 outbreak on international tourism [Online]. Available: https://www.unwto.org/impact-assessment-of-the-covid-19-outbreak-on-international-tourism.

[3] M. Zibarzani et al., "Customer satisfaction with Restaurants Service Quality during COVID-19 outbreak: A two-stage methodology," Technology in Society, vol. 70, p. 101977, 2022/08/01/ 2022.

[4] A. M. Fiori and I. Foroni, "Reservation Forecasting Models for Hospitality SMEs with a View to Enhance Their Economic Sustainability," (in English), Sustainability, Article vol. 11, no. 5, p. 24, Mar 2019, Art no. 1274.

[5] A. Ampountolas, "Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models," Forecasting, vol. 3, no. 3, pp. 580-595, 2021.

[6] S. L. Sun, M. C. Li, S. Y. Wang, and C. Y. Zhang, "Multi-step ahead tourism demand forecasting: The perspective of the learning using privileged information paradigm," (in English), Expert Systems with Applications, Article vol. 210, p. 12, Dec 2022, Art no. 118502.

[7] R. Khorsand, M. Rafiee, and V. Kayvanfar, "Insights into TripAdvisor's online reviews: The case of Tehran's hotels," (in English), Tourism Management Perspectives, Review vol. 34, p. 12, Apr 2020, Art no. 100673.

[8] M. Mamula, R. Folgieri, and K. Duvnjak, "Some considerations about Artificial Neural Networks in Hotel Industry: State of the art and future developments," in 5th International Scientific Conference on Tourism in Southern and Eastern Europe (ToSEE 2019), Opatija, CROATIA, 2019, vol. 5, OPATIJA: Univ Rijeka, Faculty Tourism & Hospitality Management, Opatija, 2019, pp. 431-440.

[9] A. G. Assaf and M. G. Tsionas, "Forecasting occupancy rate with Bayesian compression methods," (in English), Annals of Tourism Research, Article vol. 75, pp. 439-449, Mar 2019.

[10] M. Xiang, "Research on Quality Evaluation of Online Reservation Hotel APP Based on a RBF Neural Network and Support Vector Machine," (in English), Int. J. Inf. Syst. Serv. Sect., Article vol. 12, no. 2, pp. 50-64, Apr-Jun 2020.

[11] Y. M. Chang, C. H. Chen, J. P. Lai, Y. L. Lin, and P. F. Pai, "Forecasting Hotel Room Occupancy Using Long Short-Term Memory Networks with Sentiment Analysis and Scores of Customer Online Reviews," (in English), Appl. Sci.-Basel, Article vol. 11, no. 21, p. 14, Nov 2021, Art no. 10291.

[12] N. Antonio, A. de Almeida, and L. Nunes, "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior," Cornell Hospitality Quarterly, vol. 60, no. 4, pp. 298-319, 2019.

[13] A. J. Sanchez-Medina and E. C-Sanchez, "Using machine learning and big data for efficient forecasting of hotel booking cancellations," (in English), International Journal of Hospitality Management, Article vol. 89, p. 9, Aug 2020, Art no. 102546.

[14] E. C. Sánchez, A. J. Sánchez-Medina, and M. Pellejero, "Identifying critical hotel cancellations using artificial intelligence," Tourism Management Perspectives, vol. 35, p. 100718, 2020/07/01/ 2020.

[15] N. Antonio, A. de Almeida, and L. Nunes, "Hotel booking demand datasets," Data in Brief, vol. 22, pp. 41-49, 2019/02/01/ 2019.

[16] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 161-168.

[17] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, vol. 42, no. 1, pp. 177-196, 2001/01/01 2001.

[18] A. Onan, "Mining opinions from instructor evaluation reviews: A deep learning approach," Computer Applications in Engineering Education, vol. 28, no. 1, pp. 117-138, 2020.

[19] J. Singh and R. Banerjee, "A Study on Single and Multi-layer Perceptron Neural Network," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 35-40.

[20] X. Li and C. A. Micchelli, "Approximation by radial bases and neural networks," Numerical Algorithms, vol. 25, no. 1, pp. 241-262, 2000/09/01 2000.

[21] S. Mittal, "A survey on modeling and improving reliability of DNN algorithms and accelerators," Journal of Systems Architecture, vol. 104, p. 101689, 2020/03/01/ 2020.

[22] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised Classification Algorithms in Machine Learning: A Survey and Review," in Emerging Technology in Modelling and Graphics, Singapore, 2020: Springer Singapore, pp. 99-111.

[23] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 612-619, 2020.

[24] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," in International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, Cham, 2019: Springer International Publishing, pp. 758-763.

[25] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," Expert Systems with Applications, vol. 57, pp. 232-247, 2016/09/15/ 2016.

[26] T. Chen et al., "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1-4, 2015.

[27] X. Rui and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[28] M. Ahmed, R. Seraj, and S. M. S. Islam, "The *k*-means algorithm: A comprehensive survey and performance evaluation," Electronics, vol. 9, no. 8, p. 1295, 2020.

[29] A. Onan, "Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling," Computational and mathematical methods in medicine, vol. 2018, p. 2497471, 2018/07/22 2018.

[30] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient," in Knowledge and Systems Sciences, Singapore, 2019: Springer Singapore, pp. 1-17.

[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," (in eng), IEEE transactions on pattern analysis and machine intelligence, vol. 1, no. 2, pp. 224-7, Feb 1979.

[32] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration *K*-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," IOP Conference Series: Materials Science and Engineering, vol. 336, no. 1, p. 012017, 2018/04/01 2018.

[33] A. Onan, "SRL-ACO: A text augmentation framework based on semantic role labeling and ant colony optimization," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 7, p. 101611, 2023/07/01/ 2023.

[34] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," Journal of Information Science, vol. 43, no. 1, pp. 25-38, 2017.

[35] A. Onan, "Consensus Clustering-Based Undersampling Approach to Imbalanced Learning," Scientific Programming, vol. 2019, p. 5901087, 2019/03/03 2019.

[36] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified *K*-fold cross-validation on ML classifiers for predicting cervical cancer," Frontiers in Nanotechnology, vol. 4, p. 972421, 2022.

[37] A. ONAN, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," Computer Applications in Engineering Education, vol. 29, no. 3, pp. 572-589, 2021.

[38] A. Onan, "Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 2098-2117, 2022/05/01/ 2022.

[39] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, "A Review of Activation Function for Artificial Neural Network," in 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2020, pp. 281-286.

[40] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," Information Processing & Management, vol. 53, no. 4, pp. 814-833, 2017/07/01/ 2017.

[41] A. Onan, "Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering," IEEE Access, vol. 7, pp. 145614-145633, 2019.

[42] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," Concurrency and Computation: Practice and Experience, vol. 33, no. 23, p. e5909, 2021.

[43] A. Onan, "Topic-Enriched Word Embeddings for Sarcasm Identification," in Software Engineering Methods in Intelligent Algorithms, Cham, 2019: Springer International Publishing, pp. 293-304.

[44] A. Onan and M. A. Toçoğlu, "A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification," IEEE Access, vol. 9, pp. 7701-7722, 2021.

[45] A. Onan, "Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 7, p. 101610, 2023/07/01/ 2023.

[46] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," Journal of Information Science, vol. 44, no. 1, pp. 28-47, 2018.

Parte III. Discusión y Conclusiones

Conclusiones Finales

Este apartado destaca las principales conclusiones extraídas del análisis de los trabajos de investigación presentados en esta tesis doctoral, evidenciando la relevancia de estos temas en la definición de estrategias en la gestión turística.

Al iniciar esta descripción es pertinente señalar el cumplimiento del objetivo principal de la tesis doctoral, al demostrar la eficacia de las técnicas de Machine Learning para generar información relevante que ayude en la toma de decisiones en la gestión turística. El estudio ha sido abordado desde la revisión de la literatura existente acerca de la aplicación de la Inteligencia Artificial en el turismo, el análisis de las variables asociadas con la operación de las empresas turísticas y el estudio de modelos de pronósticos en el sector turístico.

La revisión de literatura ha revelado una amplia variedad de aplicaciones de Inteligencia Artificial dirigidas a mejorar la experiencia del viajero y optimizar la gestión de los servicios turísticos. Innovadoras aplicaciones basadas en el metaverso permiten a los usuarios experimentar los destinos turísticos sin viajar físicamente. Así también, la implementación de chatbots capaces de atender consultas reconociendo emociones en el texto o tono de voz consiguen brindar respuestas más personalizadas. Aplicaciones para la planificación de viajes y pronósticos de servicios turísticos han revolucionado la industria turística, mejorando la eficiencia operativa y la toma de decisiones estratégicas.

La revisión de literatura presentada en el Capítulo II – "Artificial Intelligence as Catalyst for the Tourism Sector: A Literature Review", muestra el potencial de la Inteligencia Artificial para impulsar la innovación y desarrollo sostenido en la industria turística.

En cuanto a la aplicación de las técnicas de reducción de dimensiones, se observa que permiten reducir el número de dimensiones y, por ende, disminuir la complejidad en el tratamiento de los datos. En este contexto, se ha priorizado el uso de PCA por su capacidad en el reconocimiento de las relaciones entre las variables del conjunto de datos y su robustez frente al ruido. Al evaluar el desempeño de PCA, LLE e ISOMAP se obtienen buenos resultados, asociados con la correcta identificación de los parámetros.

El uso de técnicas de agrupamiento revela su eficacia en el análisis con un enfoque numérico, facilitando la identificación de patrones en el conjunto de datos. En este estudio, se emplean Silhouette Coefficient, Davies-Bouldin Index y Dunn Index para determinar el número óptimo de grupos a formar. Los resultados de estas métricas coincidieron con la identificación visual lograda mediante técnicas de reducción de dimensionalidad. En cuanto a la medida de distancia, se encontró que la distancia euclídea ofreció los mejores resultados.

De las técnicas de agrupamiento utilizadas, *k*-means es más eficaz, al lograr una mejor identificación de cada grupo, en comparación con *k*-medoids. En cuanto a los resultados de la técnica jerárquica, estos son presentados en un dendograma, herramienta que brinda información sobre la subdivisión del conjunto de datos en grupos, y el nivel de muestras que se reúnen en una misma hoja.

El caso de estudio presentado en el Capítulo III – “Exploratory techniques to analyze Ecuador´s tourism industry”, aplica técnicas de reducción de dimensionalidad y técnicas de agrupamiento, sobre conjuntos de datos correspondientes a la operación de empresas turísticas en diferentes años, identificando el comportamiento y tendencia de la operación, durante los períodos analizados.

La selección de las técnicas de Machine Learning utilizadas en el pronóstico de la cancelación de reservaciones en el sector hotelero ha incluido la evaluación de las fortalezas y limitaciones de diversas técnicas, con el objeto de reconocer las más apropiadas. Durante el análisis se destacan las redes neuronales profundas como técnicas fuertes para el tratamiento de relaciones más complejas entre los datos. A continuación, el estudio se enfocó en la identificación de los hiperparámetros de cada técnica seleccionada, un paso crítico para maximizar la calidad y precisión de los pronósticos obtenidos, asegurando así la robustez y fiabilidad de los resultados.

El preprocesamiento de datos es uno de los factores más críticos, que contribuyen a la eficacia de los modelos de predicción. La elección de las técnicas de preprocesamiento depende de las características del conjunto de datos. Por lo tanto, es relevante realizar un Análisis Exploratorio de Datos para identificar inconsistencias y definir las estrategias más adecuadas para el tratamiento de los datos. En este estudio, el preprocesamiento ha incluido técnicas de limpieza de datos, identificación de outliers, transformación de datos y selección de las variables significativas para la investigación. Este proceso ha permitido obtener un conjunto de datos óptimo para la aplicación de los modelos de pronósticos.

A partir de las métricas y gráficos de desempeño es posible evaluar los resultados de los modelos de predicción para anticipar las cancelaciones de reservaciones de hotel. El caso de estudio presentado en el Capítulo IV – “Forecasting hotel cancellations through Machine Learning” destaca que las técnicas Deep Neural Network y XgBoost presentan la mayor precisión y eficacia en la predicción de cancelaciones.

Los estudios presentados en esta tesis doctoral evidencian el progreso del conocimiento en el campo de la Inteligencia Artificial aplicada a la industria turística. Desde el análisis de técnicas de reducción de dimensionalidad y agrupamiento en la operación de las empresas turísticas, la revisión de las aplicaciones de Inteligencia Artificial en el turismo y la aplicación de modelos predictivos para la gestión de reservaciones de hotel, se ofrece una visión de cómo estas técnicas pueden optimizar la gestión turística. Se espera que los casos de estudio presentados en esta tesis, abran nuevas oportunidades para futuras investigaciones, con el objetivo de contribuir a la innovación y al desarrollo continuo de esta industria.

Final Conclusions

This section highlights the main conclusions drawn from the analysis of the research presented in this doctoral thesis, demonstrating the relevance of these topics in the definition of strategies in tourism management.

At the beginning of this description, it is pertinent to point out the fulfillment of the main objective of the doctoral thesis, by demonstrating the effectiveness of Machine Learning techniques to generate relevant information to help decision-making in tourism management. The study has been approached from the review of the existing literature on the application of Artificial Intelligence in tourism, the analysis of the variables associated with the operation of tourism companies and the study of forecasting models in the tourism sector.

The literature review has revealed a wide variety of Artificial Intelligence applications aimed at improving the traveler experience and optimizing the management of tourism services. Innovative applications based on metaverses allow users to experience tourist destinations without the need to physically travel. Likewise, the implementation of chatbots capable of responding to queries by recognizing emotions in text or tone of voice can provide more personalized responses. Applications for travel planning and tourism service forecasting have revolutionized the tourism industry, improving operational efficiency and strategic decision making.

The literature review presented in Chapter II - "Artificial Intelligence as a Catalyst for the Tourism Sector: A Literature Review", shows the potential of Artificial Intelligence to drive innovation and sustained development in the tourism industry.

With regard to the application of dimensionality reduction techniques, it was observed that they make it possible to reduce the number of variables and, therefore, the complexity of data processing. In this context, the use of PCA has been prioritized for its ability to recognize the relationships between variables in the data set and its robustness to noise. When evaluating the performance of PCA, LLE and ISOMAP, good results are obtained, associated to the correct identification of the parameters; however.

The use of clustering techniques reveals their effectiveness in the analysis with a numerical approach, facilitating the identification of patterns in the data set. In this study, the Silhouette Coefficient, Davies-Bouldin Index and Dunn Index are used to determine the optimal number of clusters to form. The results of these metrics coincided with the visual identification achieved by dimensionality reduction techniques. As for the distance measure, Euclidean distance was found to provide the best results.

Of the clustering techniques used, *k*-means is more effective, as it achieves a better identification of each group, compared to *k*-medoids. As for the results of the hierarchical technique, these are presented in a dendogram, a tool that provides information on the subdivision of the data set into groups, and the level of samples that are collected on the same sheet.

The case study presented in Chapter III - "Exploratory technique for analyzing Ecuador's tourism industry", applies dimensionality reduction and clustering techniques on data sets corresponding to the operation of tourism companies in different years, identifying the behavior and trend of the operation during the periods analyzed.

The selection of machine learning techniques used in reservation cancellation forecasting in the hotel industry has included the evaluation of the strengths and limitations of various techniques, in order to recognize the most suitable ones. During the analysis, deep neural networks were highlighted as strong techniques for processing more complex relationships between data. Next, the study focused on identifying the hyperparameters of each selected technique, a critical step to maximize the quality and accuracy of the obtained forecasts, thus ensuring the robustness and reliability of the results.

Data preprocessing is one of the most critical factors contributing to the effectiveness of the prediction models. The choice of preprocessing techniques depends on the characteristics of the data set. Therefore, it is relevant to perform an Exploratory Data Analysis to identify inconsistencies and define the most appropriate strategies for data processing. In this study, preprocessing has included data cleaning techniques, identification of outliers, data transformation and selection of significant variables for the research. This process has allowed obtaining an optimal data set for the application of the forecasting models.

From performance metrics and graphs, it is possible to evaluate the results of prediction models to anticipate hotel cancellations. The case study presented in Chapter IV - "Forecasting hotel cancellations through Machine Learning" highlights that Deep Neural Networks and XgBoost techniques present the highest accuracy and efficiency in the prediction of cancellations.

The studies presented in this doctoral thesis show the progress of knowledge in the field of Artificial Intelligence applied to the tourism industry. From the analysis of dimensionality reduction and clustering techniques in the operation of tourism companies, the review of Artificial Intelligence applications in tourism and the application of predictive models for hotel reservation management, a vision of how these techniques can optimize tourism management is offered. It is hoped that the case studies presented in this thesis will open new opportunities for future research, with the aim of contributing to the innovation and continuous development of this industry.