Maestría en

Ciencia de Datos y Máquinas de Aprendizaje Mención Inteligencia Artificial

Trabajo previo a la obtención de título de

Magister en Ciencia de Datos y Máquinas de Aprendizaje Mención Inteligencia Artificial

AUTOR/ES:

Flores Bosmediano Edison Marcos

Godoy Trujillo Pamela Estefanía

Meneses Ortiz Raul Alexander

Noguera Gualotuña Alexis David

Rojas Cevallos Alexander Vladimir

TUTOR/ES:

Karla Mora

Paulina Vizcaíno

TEMA: Análisis de patrones turísticos de los hogares ecuatorianos mediante técnicas no supervisadas para la construcción de un sistema recomendador de viajes en el Ecuador.

Quito - Ecuador

diciembre – 2025

REINVENTEMOS EL FUTURO

CERTIFICACIÓN DE AUTORÍA

Nosotros, **Flores Bosmediano Edison Marcos, Godoy Trujillo Pamela Estefanía, Meneses Ortiz Raul Alexander, Noguera Gualotuña Alexis David, Rojas Cevallos Alexander Vladimir,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

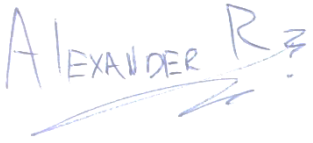
Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma
Flores Bosmediano Edison Marcos

Firma
Godoy Trujillo Pamela Estefanía

Firma
Meneses Ortiz Raul Alexander

Firma
Noguera Gualotuña Alexis David



Firma
Rojas Cevallos Alexander Vladimir

AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL

Nosotros, **Flores Bosmediano Edison Marcos, Godoy Trujillo Pamela Estefanía, Meneses Ortiz Raul Alexander, Noguera Gualotuña Alexis David, Rojas Cevallos Alexander Vladimir**, en calidad de autores del trabajo de investigación titulado ***Análisis de patrones turísticos de los hogares ecuatorianos mediante técnicas no supervisadas para la construcción de un sistema recomendador de viajes en el Ecuador***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.
D. M. Quito, diciembre 2025.

Firma
Flores Bosmediano Edison Marcos

Firma
Godoy Trujillo Pamela Estefanía

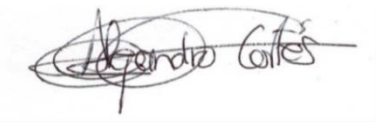
Firma
Meneses Ortiz Raúl Alexander

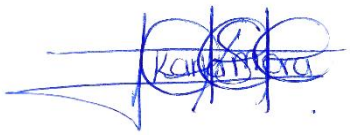
Firma
Noguera Gualotuña Alexis David

Firma
Rojas Cevallos Alexander Vladimir

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA

Nosotros, **Alejandro Cortés López Director EIG y Karla Estefanía Mora Cajas Coordinadora UIDE,** declaramos que: **Flores Bosmediano Edison Marcos, Godoy Trujillo Pamela Estefanía, Meneses Ortiz Raul Alexander, Noguera Gualotuña Alexis David, Rojas Cevallos Alexander Vladimir** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.





---------- ----------

Alejandro Cortés López
Director de la
Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

DEDICATORIA

A mis padres, por su amor incondicional, esfuerzo y apoyo constante.
A mi hermana y a mis sobrinas, cuyo cariño y aliento diario han sido una fuente permanente de motivación.

Alexander Vladimir Rojas Cevallos

Al compromiso con el conocimiento, que impulsa el crecimiento personal y profesional, fortalece la vocación, fomenta la disciplina y la perseverancia, y permite afrontar los desafíos con responsabilidad y ética.

Pamela Estefanía Godoy Trujillo

A los sueños, que nacen del esfuerzo constante, se sostienen con perseverancia y transforman a la persona, fortaleciendo el propósito y dando sentido al camino.

Edison Marcos Flores Bosmediano

A Dios, por la guía que fortalece cada paso y sostiene los sueños.
A nuestros padres y familia, por su amor incondicional, esfuerzo y apoyo permanente.
A quienes caminaron junto a nosotros, inspirando motivación, aprendizaje y propósito.

Raúl Alexander Meneses Ortiz

Agradezco a Dios por darme la fortaleza y guiarme en mi vida académica, a mis padres por siempre apoyarme y motivarme a seguir creciendo y a mi pareja por darme su ayuda y confianza en todo este trayecto.

Alexis David Noguera Gualotuña

AGRADECIMIENTOS

Agradezco profundamente a mis padres por su apoyo y acompañamiento constante durante la realización de este trabajo. A mi hermana y a mis sobrinas, por su estímulo diario y la motivación que aportaron en cada etapa del proceso.

Alexander Vladimir Rojas Cevallos

A Dios, por su guía y fortaleza constante. A mi familia, por su amor, apoyo incondicional y motivación permanente. A mis compañeros y profesores, por su acompañamiento y valioso aporte en este proceso académico.

Pamela Estefanía Godoy Trujillo

A mi familia, por su apoyo incondicional y por ser la base de cada esfuerzo realizado. Por acompañarme con paciencia, confianza y amor en cada etapa de este camino.

Edison Marcos Flores Bosmediano

A quienes fortalecieron nuestra fe en que todo esfuerzo vale la pena. A los abrazos, palabras y silencios que sostuvieron cuando flaquearon las fuerzas. A cada guía en el camino académico, por transformar acompañamiento en aprendizaje.

Raúl Alexander Meneses Ortiz

Agradezco en primer lugar a Dios por ser el motor de mi vida en todos mis retos y proyectos, de igual manera, a mis padres por su amor y motivación para superarme cada día más y ayudarme a crecer como un profesional.

Alexis David Noguera Gualotuña

RESUMEN

El presente estudio tiene como objetivo identificar patrones de comportamiento turístico de los hogares ecuatorianos mediante técnicas de aprendizaje no supervisado, con el fin de desarrollar un sistema recomendador de viajes basado en similitud. Se considero como población los hogares ecuatorianos que realizaron viajes internos durante el año 2021. El problema abordado radica en la falta de segmentación precisa de la demanda turística nacional, lo que limita la personalización de estrategias y servicios. Este análisis resulta relevante para optimizar la planificación turística y fortalecer la competitividad del sector. El enfoque fue cuantitativo, de tipo aplicado y descriptivo, sustentado en datos secundarios provenientes del Sistema Banco de Información (SBI) del INEC. La muestra final incluyó 2.151 registros depurados, con variables sociodemográficas, temporales, de gasto y preferencias turísticas. El procesamiento se realizó en Python, aplicando técnicas de limpieza, normalización y detección de anomalías mediante Isolation Forest. Para la segmentación se evaluaron algoritmos K-Means, K-Modes y K-Prototypes, seleccionándose este último por su capacidad para manejar datos mixtos. El modelo óptimo, determinado mediante el método del codo, identificó tres clústeres representativos: turistas de recreación económica, turistas culturales y naturales, y turistas de sol y playa con alta movilidad. Los resultados evidenciaron diferencias significativas en duración de viaje, estructura de gasto y actividades predominantes, lo que permitió construir un sistema recomendador funcional desplegado en una interfaz web con Streamlit. Se concluye que la integración de técnicas no supervisadas en el análisis turístico aporta una herramienta práctica para personalizar recomendaciones y apoyar la toma de decisiones estratégicas en el sector.

Palabras clave: Aprendizaje no supervisado, clustering, turismo, sistema recomendador, k-prototypes.

ABSTRACT

The present study aims to identify tourism behavior patterns of Ecuadorian households through unsupervised learning techniques, in order to develop a similarity-based travel recommender system. The population considered consisted of Ecuadorian households that undertook domestic trips during the year 2021. The problem addressed lies in the lack of accurate segmentation of national tourism demand, which limits the personalization of strategies and services. This analysis is relevant for optimizing tourism planning and strengthening the competitiveness of the sector.
The research followed a quantitative, applied, and descriptive approach, supported by secondary data obtained from the Information Bank System (SBI) of the National Institute of Statistics and Census (INEC). The final sample included 2,151 cleaned records, comprising sociodemographic, temporal, expenditure, and tourism preference variables. Data processing was conducted in Python, applying data cleaning, normalization, and anomaly detection techniques using the Isolation Forest algorithm. For segmentation, the K-Means, K-Modes, and K-Prototypes algorithms were evaluated, with the latter being selected due to its ability to handle mixed-type data. The optimal model, determined using the elbow method, identified three representative clusters: low-budget recreational tourists, cultural and nature tourists, and highly mobile sun-and-beach tourists. The results revealed significant differences in trip duration, spending structure, and predominant activities, which enabled the development of a functional recommender system deployed through a Streamlit web interface. It is concluded that the integration of unsupervised learning techniques into tourism analysis provides a practical tool for personalizing recommendations and supporting strategic decision-making in the tourism sector.

Keywords: Unsupervised learning, clustering, tourism, recommender system, k-prototypes.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

ÍNDICE DE APÉNDICES

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Título del proyecto

Análisis de patrones turísticos de los hogares ecuatorianos mediante técnicas no supervisadas para la construcción de un sistema recomendador de viajes en el Ecuador.

1.2. Definición del proyecto

El presente proyecto tiene como finalidad analizar los patrones turísticos de los hogares ecuatorianos mediante técnicas de aprendizaje no supervisado, con el propósito de construir un sistema recomendador de viajes contextualizado al comportamiento real de la población. La necesidad de este estudio surge debido a la creciente disponibilidad de datos provenientes de encuestas nacionales y registros administrativos relacionados con actividades turísticas, cuyo volumen y heterogeneidad dificultan la extracción de conocimientos mediante métodos tradicionales (Han et al., 2022). Dado que los hogares presentan combinaciones diversas de características socioeconómicas, preferencias de viaje y niveles de gasto, se requiere un enfoque metodológico capaz de identificar perfiles latentes que no se encuentran explícitamente definidos en los datos (Aggarwal, 2020). En este marco, las técnicas de clustering proporcionan un mecanismo adecuado para descubrir grupos con comportamientos similares sin depender de etiquetas predeterminadas, permitiendo generar tipologías turísticas empíricas y sustentadas en métricas objetivas (Tan et al., 2019).

El análisis de patrones turísticos mediante agrupamiento facilita comprender cómo determinados segmentos de hogares toman decisiones de viaje, lo cual resulta esencial para diseñar estrategias de personalización en sistemas recomendadores (Ricci et al., 2021). Este tipo de modelos dependen de medidas de similitud que reflejen afinidades entre usuarios o categorías de destino, por lo que la identificación previa de clusters mejora la precisión al reducir la dispersión estructural de las preferencias (Su & Khoshgoftaar, 2020). Además, el estudio adquiere relevancia dentro del contexto ecuatoriano, donde la diversificación de la demanda turística constituye un factor clave para dinamizar economías locales y fortalecer la competitividad del sector (Ministerio de Turismo del Ecuador, 2023). A través de la segmentación basada en datos, es posible detectar hogares con comportamientos emergentes, necesidades específicas o barreras de acceso que tradicionalmente no han sido consideradas en políticas o estrategias de promoción (UNWTO, 2022).

El proyecto se orienta a integrar la analítica avanzada con un enfoque aplicado, mediante la construcción de un sistema recomendador que utilice los clusters identificados como base para generar sugerencias de destinos ajustadas a los perfiles observados. Desde una perspectiva metodológica, el proceso implica un tratamiento integral de los datos, que abarca normalización, elección de distancias adecuadas para variables mixtas y evaluación de la calidad de los clusters mediante índices estadísticos (Berkhin, 2021). Este procedimiento garantiza que los grupos resultantes posean coherencia interna y relevancia interpretativa, elementos fundamentales para su incorporación posterior en un motor de recomendación. Así, la combinación de aprendizaje no supervisado y técnicas de recomendación contribuye a transformar información dispersa en un instrumento útil tanto para la toma de decisiones del turista como para las estrategias de desarrollo territorial basadas en evidencia (Schafer et al.,

2020). En síntesis, el proyecto propone un modelo de análisis innovador que une minería de datos, turismo y sistemas inteligentes, con el fin de aportar una herramienta tecnológica orientada a mejorar la experiencia turística y apoyar la gestión del sector en Ecuador.

1.3. Justificación e importancia del trabajo de investigación

El turismo se ha consolidado como uno de los sectores con mayor dinamismo dentro de la economía ecuatoriana, tanto por su capacidad de generar empleo como por su contribución a la diversificación productiva del país. Comprender cómo y por qué los hogares se movilizan, qué destinos prefieren y qué factores inciden en sus decisiones de viaje constituye un insumo clave para el diseño de políticas públicas y estrategias de desarrollo territorial. Sin embargo, el análisis del comportamiento turístico en Ecuador continúa siendo fragmentado, con estudios aislados y con escasa aplicación de técnicas avanzadas de análisis de datos. Por ello, explorar los patrones turísticos de los hogares mediante métodos no supervisados abre la posibilidad de identificar perfiles reales de consumo, permitiendo observar dinámicas que no se aprecian a través de técnicas tradicionales.

En esta línea, la investigación se vincula directamente con el Objetivo de Desarrollo Sostenible (Trabajo Decente y Crecimiento Económico), que promueve actividades productivas que generen empleo de calidad y que fortalezcan el crecimiento económico sostenible. Un sistema recomendador construido a partir del análisis de patrones turísticos puede aportar evidencia para mejorar la oferta de servicios, fortalecer emprendimientos locales y promover una distribución más equitativa de la demanda turística. Al comprender mejor las preferencias y comportamientos de los hogares

ecuatorianos, se facilita la toma de decisiones orientadas a dinamizar destinos menos consolidados y a distribuir los flujos turísticos de manera más equilibrada.

Asimismo, la investigación guarda coherencia con el ODS 9 (Industria, Innovación e Infraestructura), que enfatiza la necesidad de incorporar tecnologías de la información y analítica avanzada para mejorar sectores estratégicos. El uso de técnicas de aprendizaje automático no supervisado para agrupar perfiles de turistas constituye un aporte innovador dentro del campo del turismo nacional, donde predominan métodos descriptivos. Un sistema recomendador basado en similitud no solo representa una herramienta tecnológica de valor, sino que también fortalece la infraestructura digital del sector turístico al ofrecer soluciones que optimizan la experiencia del usuario y la planificación del viaje.

De igual manera, el estudio mantiene conexión directa con las prioridades establecidas por el Ministerio de Turismo de Ecuador, que ha insistido en la necesidad de fortalecer la inteligencia turística, diversificar la oferta y aprovechar la información disponible para mejorar los procesos de promoción y planificación. Al identificar perfiles turísticos consistentes y fundamentados en datos reales de hogares ecuatorianos, la investigación aporta insumos que pueden orientar campañas segmentadas, políticas de incentivo al turismo interno y estrategias dirigidas a grupos con características específicas, mejorando así la pertinencia de las acciones institucionales.

El Plan Nacional de Desarrollo “El Nuevo Ecuador 2025” incorpora la innovación tecnológica, el desarrollo territorial equilibrado y la sostenibilidad como ejes prioritarios para la transformación productiva del país. En este marco, el presente estudio contribuye a esos lineamientos al proponer una metodología basada en analítica de datos que permite comprender cómo se comportan los hogares en sus prácticas

turísticas, lo que puede derivar en una planificación más inteligente, en inversiones más eficientes y en un aprovechamiento responsable del patrimonio cultural y natural del país. La identificación de patrones también ofrece una base para desarrollar proyectos turísticos adaptados a las realidades locales, en coherencia con el enfoque de desarrollo sostenible.

El desarrollo de un sistema recomendador basado en los patrones detectados permite traducir este conocimiento en una herramienta aplicable y con impacto directo en la experiencia de los viajeros. Al ofrecer sugerencias personalizadas a partir de perfiles reales, se potencia el turismo interno, se diversifica la demanda y se fomenta una cultura de movilidad responsable, alineada con los objetivos nacionales y globales. La investigación, por tanto, no solo busca aportar al conocimiento académico, sino también fortalecer la toma de decisiones del sector turístico, impulsar la innovación y contribuir al bienestar económico y social del país.

El aporte de esta investigación se sustenta en la aplicación rigurosa de la ciencia de datos, un campo que permite transformar información dispersa en conocimiento accionable. El análisis exploratorio de datos constituye un paso esencial, pues hace posible detectar patrones ocultos, relaciones inesperadas y estructuras internas que suelen pasar desapercibidas en enfoques tradicionales. En el ámbito turístico, donde intervienen factores sociales, económicos, espaciales y culturales, esta primera fase del análisis es clave para garantizar que los modelos posteriores se basen en datos limpios, coherentes y representativos.

La elección del aprendizaje no supervisado como metodología responde a la necesidad de explorar los datos sin imponer categorías preconcebidas. En vez de forzar clasificaciones rígidas, los algoritmos permiten que los propios datos revelen las

agrupaciones naturales entre los hogares, aportando una perspectiva imparcial y flexible. Esta característica es fundamental para una actividad tan heterogénea como el turismo, donde las motivaciones, los estilos de viaje y las capacidades económicas presentan variaciones amplias, difíciles de capturar mediante técnicas convencionales.

Finalmente, el uso de algoritmos de clustering para segmentar patrones turísticos ofrece una ventaja significativa: permite construir perfiles basados en similitud real, derivados de comportamientos observados y no de supuestos teóricos. Al trabajar con datos empíricos, el sistema recomendador resultante podrá generar sugerencias de viaje mucho más relevantes para los hogares ecuatorianos, fortaleciendo la experiencia del usuario y aportando a la innovación tecnológica del sector turístico nacional. Esto no solo mejora la competitividad del país en el mercado interno, sino que refuerza la importancia de la ciencia de datos como un puente entre el conocimiento académico y la práctica aplicada.

1.4. Alcance

El alcance de la presente investigación comprende el análisis de los patrones de comportamiento turístico de los hogares ecuatorianos que realizaron viajes dentro del territorio nacional durante el año 2021, utilizando los microdatos oficiales del Sistema Banco de Información (SBI) provistos por el INEC. El estudio se fundamenta en información sociodemográfica, económica y de características del viaje, registrada para 3458 hogares a nivel nacional.

El proceso metodológico se desarrolló íntegramente mediante investigación cuantitativa e incluye técnicas descriptivas, procedimientos de preparación de datos y métodos de aprendizaje no supervisado. Para garantizar la calidad del conjunto de datos,

el análisis incluye fases de depuración, calibración, gestión de valores faltantes, eliminación de valores atípicos mediante el algoritmo de Bosque de Aislamiento y selección de variables relevantes con base en criterios analíticos y teóricos.

Para identificar grupos de hogares homogéneos en función de las prácticas turísticas, evaluamos diferentes algoritmos de agrupamiento adaptados a las empresas mixtas, entre ellos K-Means, K-Medoids y K-Prototypes. La comparación contempló criterios de coherencia interna, interpretabilidad de los grupos, estabilidad de las asignaciones y adecuación a los tipos de variables presentes. Durante este proceso, K-Prototypes fue seleccionado como modelo final, capaz de procesar simultáneamente variables numéricas y categorías y generar grupos y representantes del comportamiento de los turistas.

1.5. Objetivo general

Identificar los patrones de comportamiento turístico de los hogares ecuatorianos a través de un modelo no supervisado con el fin de construir un sistema recomendador basado en similitud que facilite la sugerencia de opciones de viaje a partir de los perfiles identificados.

1.6. Objetivos específicos

- Realizar un análisis exploratorio del comportamiento turístico de los hogares ecuatorianos.
- Evaluar distintas técnicas no supervisadas para seleccionar el que mejor agrupe los patrones turísticos.
- Caracterizar los grupos obtenidos según sus principales comportamientos turísticos.

CAPÍTULO 2

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

2.1.1. Introducción al Estado del Arte

Al observar el contexto del uso de datos en el sector turístico en el Ecuador se pudo evidenciar que a lo largo de los últimos años no se ha podido dar un esfuerzo para la aplicación de técnicas que permitan de cierta manera encontrar patrones de análisis en los turistas a nivel nacional o internacional, tomando en cuenta que el análisis de datos se ha convertido en un factor relevante para la transformación digital del sistema turístico alrededor del mundo. Dentro de este contexto, se mantienen estudios de universidades privadas o entidades públicas que centran el uso de datos turísticos en sistemas de análisis financiero que en un contexto de estudio tiende a ser más realista, sin embargo, no se mantienen resultados que permitan a los turistas encontrar o guiarse hacia nuevos destinos turísticos basados en tendencias, patrones, presupuesto, entre otros.

De hecho, el análisis turístico no se veía como un objeto de estudio clave en el ámbito nacional, tal como se toma en la actualidad a nivel internacional. En forma general este ámbito de análisis solo era visto como un indicador monetario para identificar que lugares generaban más dinero dentro de nuestra región. Sin embargo, instituciones como la Universidad Verdad del Azuay o la Universidad Estatal

Amazónica fueron entidades claves que ampliaron el análisis turístico más allá del sector económico, considerando variables como los gustos, afinidad y seguridad.

Una función fundamental del análisis turístico fue el uso de técnicas de recolección de datos y el análisis clustering que permitieron tener un banco de información para poder generar sistemas de aprendizaje o recomendación a nivel del Ecuador. Dichos modelos mantienen un estado general, de tal manera que la pregunta central que se busca responder es: ¿Que modelos o técnicas de datos permiten entender los patrones turísticos del Ecuador, para poder generar modelos de recomendación de viajes en el Ecuador?

Partiendo de esta interrogante, en búsqueda de una respuesta crítica y analítica de dicho planteamiento se han estudiado y analizado tres tipos de estudios, los cuales abarcan el análisis de datos turísticos en el Ecuador, los sistemas de aprendizaje analítico en el sector turístico y los algoritmos aplicados a datos de naturaleza mixta.

2.1.2. Análisis de datos turísticos en el Ecuador

En el contexto ecuatoriano el análisis turístico ha venido tomando forma de entidades privadas, sistemas gubernamentales, encuestas nacionales y medios de viajes que han logrado recolectar información relevante del panorama nacional en el sistema de turismo, sin embargo, como lo menciona Gracia (2025), “en el Ecuador el análisis del sector turístico ha sido un proceso complejo y multifacético” (p. 5). Es decir, aunque existen muchas fuentes de información con una disponibilidad abierta al público muchos de los datos del sistema turístico son de carácter privado que de cierta manera limitan en su mayoría generar sistemas de analítica avanzada.

Sin embargo, (León, 2025), realiza un análisis comparativo de los métodos turísticos nacionales que mantiene el MTE (Ministerio de Turismo de Ecuador) y entidades privadas, lo cual, permite identificar que existen bases sólidas para incorporar técnicas de clustering, análisis de datos o aprendizaje automático que sirvan como una guía para sistemas de recomendación turística los cuales a nivel nacional son poco o nada mencionados o aplicados.

2.1.3. Sistemas de Aprendizaje Analítico en el Sector Turístico

Una vez comprendido el estado de la data a nivel turístico en el Ecuador se deben abordar los métodos para poder analizar y aplicar dicha información en sistemas de aporte turístico óptimos, en dicho contexto, se establece el aprendizaje no supervisado como una herramienta clave en el análisis de datos previamente etiquetados, los cuales, son el factor relevante en un sistema de analítica turística debido al dinamismo y heterogeneidad de los sistemas de información. Como lo menciona (Gallego, 2022) se trata los objetos de entrada como un conjunto de variables aleatorias, es decir, funciona como una base para identificar patrones, relaciones y estructuras ocultas dentro de un sistema de datos, lo cual, desde una perspectiva general permite optimizar la fragmentación de turistas de acuerdo con variables similares, sin necesidad, de una normalización previa.

De igual manera, en este sentido se encuentra la investigación realizada por Mendoza (2024) que aborda a los algoritmos de clustering como la herramienta más utilizada dentro de la clasificación de sistemas turísticos, destacando el uso de K-Means como un elemento de soluciones para un sistema de datos que genera flujos constantes en sus variables.

2.1.4. Algoritmos aplicados a datos de naturaleza mixta

Dentro del estudio de datos en un ámbito turístico nace la necesidad de encontrar soluciones específicas para abordar la naturaleza heterogénea que caracterizan a este tipo de información, debido a que existe una integración de variables cuantitativas como gastos de viaje, duración, edad, entre otros., y a su vez, se integran variables cualitativas, como lo son el tipo de turismo, destinos conocidos, motivos de viajes, etc. Es debido a esta diversidad de variables y sistemas categóricos y numéricos que se establecen brechas metodológicas que limitan el análisis y la aplicación de sistemas como K-Means que se centran únicamente en sistemas numéricos.

Es debido a esta limitación que nace el algoritmo K-Protoypes como una extensión del sistema K-Means para soluciones en escenarios de conjuntos de datos mixtos (Jia, 2020), dentro de esta panorama estudios como los realizados por (Valderrama, 2021) establecen la efectividad de aplicar un sistema múltiple que engloba la naturalidad de los datos para obtener sistemas óptimos que permiten ya sea análisis un contexto específico o generar modelos de predicción para diferentes escenarios, como es el caso del estudio especificado que permitió delimitar perfiles de estudiantes para generar estrategias de acompañamiento orientadas a fomentar mejoras en la calidad educativa y a impulsar la actualización de los entornos de enseñanza de forma personalizada. Lo cual, puede servir como modelo de estudio para una aplicación de segmentación turística debido al parecido en la naturalidad de los datos que mantiene ambos enfoques.

De igual manera, uno de los estudios que ha tenido un mayor enfoque en la aplicación de K-Prototypes a nivel turístico, es el que establecieron (Mora et al., 2017) en donde se toman en cuenta variables mixtas como eventos turísticos, edad de viajes,

gustos de viaje, entre otros, y de los cuales se pudo generar un sistema de guía para presentar eventos de interés e información de sectores turísticos dentro de cantón Azogues, lo cual, permite identificar que el enfoque establecido genera resultados óptimos y representativos en comparación con otros algoritmos de clustering.

2.1.5. Consecuencias de los Estudios Analizados

A partir del análisis establecido, los trabajos en torno a un aprendizaje no supervisado dentro del ámbito turístico, a pesar de ser escasos, demuestran la viabilidad de aplicar sistemas de aprendizaje mixtos para poder normalizar y optimizar la data de los diferentes ámbitos de viaje, lo cual justifica la investigación planteada como una guía para establecer sistemas de recomendación en base a datos delimitados.

De igual manera, a pesar de que los estudios planteados permitieron un aporte significativo al proyecto se puede poner en evidencia el escaso uso de integración de sistemas de clustering como modelos metodológicos para sistemas de recomendación turística, debido en un principio a que muchos de los estudios planteados a nivel nacional e internación se centran en generar resultados financieros o económicos que basan su practica en encuestas y datos normados desaprovechando el potencial de la data y de las herramienta de análisis.

Partiendo de esta primicia, se generan brechas de aprendizaje y estudio que establecen un escenario metodológico abierto a nuevos enfoques de análisis y diversidad de aplicaciones en el ámbito turístico, dejando interrogantes como ¿de qué manera implementar algoritmos de clustering para representar el sentido real de la heterogeneidad del ámbito turístico en Ecuador? o ¿De qué manera, el análisis de datos mixtos a nivel nacional puede mejorar la recomendación de sistemas turísticos

personalizados?, lo cual, aunque son interrogantes que de cierta manera se abordan en el estudio planteado, quedan alineadas a sistemas metodológicos más intrínsecos que expandan el análisis y los resultados de sistemas de recomendación.

2.2. Marco Teórico

2.2.1. Turismo y comportamiento turístico de los hogares

2.2.1.1. Concepto de turismo

El turismo constituye un fenómeno social, cultural y económico que implica desplazamientos temporales de personas fuera de su entorno habitual, motivados por ocio, recreación, visitas familiares o actividades profesionales. La Organización Mundial del Turismo (OMT) define el turismo como “las actividades de las personas que viajan y permanecen en lugares distintos a su entorno habitual por un período inferior a un año con fines de ocio, negocios u otros” (World Tourism Organization, 2019). Esta concepción subraya que la actividad turística está profundamente influenciada por los patrones de decisión de los hogares, quienes actúan como unidades de análisis clave para comprender la demanda turística.

La actividad turística posee la capacidad de dinamizar la economía local, atrayendo inversiones y creando diversas fuentes de trabajo. Este movimiento económico también favorece a sectores como la agricultura, la pesca y la artesanía en las comunidades que reciben visitantes. En consecuencia, el turismo se convierte en un factor que contribuye de manera significativa al Producto Interno Bruto, a la balanza de pagos y a otros indicadores macroeconómicos de numerosos países. (Moreno, 2010)

Desde la perspectiva planteada por Muñoz, el turismo puede entenderse como un fenómeno social y económico que surge de la interacción entre la demanda de individuos que se desplazan temporalmente fuera de su entorno habitual y la oferta organizada de bienes y servicios diseñada para atender sus necesidades. Este enfoque concibe al turismo como una actividad productiva compleja, en la que convergen procesos de carácter cultural, comercial y estructural, y cuyo análisis requiere considerar tanto el comportamiento del turista como el funcionamiento de los agentes y sectores que conforman el sistema turístico. (Muñoz, 2003)

2.2.1.2. Turismo en el contexto ecuatoriano

En Ecuador, el turismo se configura como una actividad económica de relevancia creciente que condiciona procesos de desarrollo regional y nacional; su aporte se observa tanto en el flujo de visitantes internacionales y nacionales como en la dinamización de sectores conexos (transporte, hospedaje, comercio y servicios). Las estadísticas oficiales y los informes de rendición de cuentas del Ministerio de Turismo del Ecuador muestran procesos de recuperación post pandemia y esfuerzos por consolidar la información estadística mediante visualizadores y sistemas de monitoreo, lo que permite cuantificar arribos, pernoctaciones y mercados emisores para diseñar políticas públicas más precisas. (Ministerio de Turismo del Ecuador, 2023)

No obstante, el comportamiento del turismo en Ecuador es heterogéneo: regiones como la Sierra, la Costa, la Amazonía y, de manera muy particular, las Islas Galápagos presentan perfiles diferenciales en términos de oferta, presión de visitantes y requerimientos de gestión. Las Galápagos, por ejemplo, registraron un alto volumen de arribos (cerca de 329 475 visitantes en 2023 según el reporte oficial), cifra que plantea simultáneamente oportunidades económicas y desafíos ambientales y de infraestructura,

tal como han señalado tanto los informes de gestión como investigaciones periodísticas sobre la sobrecarga ecológica y la necesidad de políticas de regulación más estrictas (Gobierno de Galápagos, 2023).

Desde una perspectiva socioeconómica, el turismo en Ecuador incide en la generación de empleo, la diversificación de ingresos locales y la oferta de actividades productivas complementarias; sin embargo, la magnitud de esos beneficios depende de la capacidad de captura de valor por parte de las economías locales, la formalización de la oferta y la articulación entre actores públicos y privados. Estudios recientes sobre turismo comunitario y análisis sectoriales advierten que los impactos positivos (reducción de pobreza, creación de microempresas) se consolidan cuando el desarrollo turístico incorpora gestión participativa, fortalecimiento de capacidades locales y mecanismos que eviten la fuga de renta hacia agentes externos. (Aucancela Chimbolema, 2025)

Finalmente, los desafíos para la sostenibilidad y equidad del turismo en Ecuador pasan por fortalecer la gobernanza interinstitucional (nacional, regional y local), mejorar la calidad de la información estadística y focalizar intervenciones que reduzcan vulnerabilidades (estacionalidad, informalidad, impactos ambientales). Las recomendaciones recurrentes en la literatura y los informes oficiales incluyen: a) políticas integradas de desarrollo territorial, b) formación y apoyo a emprendimientos locales, c) mecanismos de financiación para conservación y d) estrategias de promoción orientadas a mercados de mayor gasto y menor estacionalidad. Estas líneas de acción son coherentes con los documentos de planificación y los diagnósticos técnicos que el país ha elaborado en los últimos años. (Ministerio de Turismo del Ecuador, 2024)

2.2.1.3. Características socioeconómicas de los hogares ecuatorianos

En Ecuador, los hogares presentan condiciones de vida que varían según su localización (urbana o rural), su acceso a servicios básicos, y la tenencia de vivienda; estas condiciones influyen en su vulnerabilidad socioeconómica y en su capacidad para beneficiarse del turismo. Según datos censales recientes, la mayoría de los hogares posee vivienda propia totalmente pagada, aunque la calidad del acceso al agua y los servicios puede diferir según área geográfica. En este contexto, la estructura del hogar, su nivel de ingreso, y la formalidad del empleo determinan en gran medida si el turismo representa una oportunidad real de desarrollo o si sus efectos quedan limitados a ciertos sectores de la población. (Ministerio de Turismo del Ecuador, 2022)

Por otro lado, el sistema turístico en Ecuador ha demostrado tener un efecto distributivo relativamente favorable para los hogares de menores ingresos, especialmente cuando se incrementa la demanda turística internacional. Un modelo de contabilidad social aplicado al país reveló que un aumento del 10 % en la demanda turística tiene efectos multiplicadores que repercuten en ingresos para hogares pobres, señalando al turismo como una estrategia “pro pobre”. (Croes & Rivera, 2016)

Esto sugiere que, bajo ciertas condiciones como adecuada formalización, proximidad al destino turístico, y participación de la población local los beneficios del turismo pueden alcanzar a sectores vulnerables, contribuyendo a reducir desigualdades. En las zonas rurales y en regiones históricamente marginadas, como algunas provincias amazónicas, el turismo comunitario emerge como alternativa frente a actividades extractivas tradicionales. Estudios analizan cómo la expansión de actividades turísticas incide sobre indicadores como el empleo, el ingreso y la pobreza en esas áreas. En estos territorios, donde los hogares enfrentan niveles elevados de necesidades básicas

insatisfechas, la diversificación de ingresos a través del turismo y encadenamientos productivos con artesanías, gastronomía local u hospedaje rural puede mejorar la resiliencia económica y ofrecer caminos de desarrollo más sostenibles.

No obstante, los efectos positivos del turismo sobre los hogares no son automáticos ni uniformes: dependen de factores estructurales como formalización laboral, educación, acceso a infraestructura, conectividad y gobernanza local. Un análisis espacial de pobreza en cantones ecuatorianos encontró que un aumento del valor agregado turístico se asocia con disminución de pobreza en el cantón y efectos positivos incluso en zonas colindantes. Esto indica que el impacto del turismo trasciende el destino mismo, generando externalidades sociales y económicas que pueden beneficiar comunidades más amplias, siempre que existan políticas públicas adecuadas. (Ponce et al., 2020)

Finalmente, desde una perspectiva de desarrollo rural y equidad, el turismo en Ecuador tiene el potencial de contribuir al bienestar de los hogares cuando se integra con estrategias de planificación, participación comunitaria y diversificación productiva. Investigaciones recientes destacan que la combinación turismo-tecnología (turismo inteligente), capacitación local, innovación en servicios turísticos y mejora de infraestructura pueden amplificar los beneficios del turismo para la población. En consecuencia, para que los hogares ecuatorianos perciban beneficios reales reducción de pobreza, aumento de ingresos, mejora de calidad de vida es necesario asegurar que el desarrollo turístico considere su contexto socioeconómico, promueva formalización e incluya a los actores locales en la gestión. (Leon et al., 2025)

2.2.1.4. Comportamiento turístico de los hogares

El comportamiento turístico de los hogares ha sido estudiado como un fenómeno multidimensional que combina factores económicos, culturales, psicológicos y demográficos. Diferentes investigaciones sostienen que variables como nivel de ingresos, composición del hogar, etapa del ciclo de vida familiar y educación influyen en la frecuencia y el tipo de viajes que realizan las familias. En este sentido, la decisión de viajar se configura como un proceso racional y emocional en el que los hogares ponderan sus recursos disponibles, sus motivaciones de ocio y el valor simbólico de la experiencia vacacional.

Desde una perspectiva económica, el ingreso disponible constituye uno de los determinantes más fuertes del comportamiento turístico, ya que define la posibilidad de costear traslados, alojamiento y actividades recreativas. De acuerdo con la Organización Mundial del Turismo, los hogares con mayores niveles de ingreso registran una propensión más alta a viajar y realizan estancias más largas y diversificadas. Sin embargo, estudios recientes muestran que incluso hogares de ingresos medios y bajos participan en actividades turísticas cuando existen opciones accesibles, destinos cercanos o modalidades de viaje comunitario.

Asimismo, el comportamiento turístico de los hogares está condicionado por su estructura interna. Hogares con niños tienden a elegir destinos seguros, accesibles y con actividades familiares, mientras que los hogares conformados por jóvenes adultos privilegian viajes de aventura, estancias cortas y experiencias digitales. Los hogares de personas mayores, por su parte, muestran un incremento sostenido en su participación turística gracias a una mayor esperanza de vida, programas especiales y mejoras en

accesibilidad. Estas diferencias demográficas explican la segmentación del mercado turístico y la diversificación de la oferta.

Finalmente, el desarrollo de tecnologías digitales ha modificado profundamente los patrones turísticos de los hogares, pues facilita la búsqueda de información, la reserva de servicios y la comparación de precios. La literatura indica que la adopción de herramientas digitales no solo reduce incertidumbre, sino que aumenta la autonomía del viajero, generando hogares más informados y con mayor capacidad para personalizar su experiencia turística. En síntesis, el comportamiento turístico del hogar contemporáneo es el resultado de la interacción entre condiciones socioeconómicas, preferencias culturales y dinámicas tecnológicas que redefinen permanentemente la forma de viajar.

2.2.1.5. Patrones turísticos

El estudio de los patrones turísticos se ha consolidado como un eje fundamental para comprender cómo las personas seleccionan destinos, planifican sus viajes y experimentan sus estancias. Esta línea de análisis aborda las regularidades que emergen del comportamiento de los visitantes, considerando factores motivacionales, socioeconómicos y culturales que influyen en sus decisiones. Investigaciones clásicas, como las de (Cohen, 1972), sostienen que los turistas tienden a agruparse en tipologías según su búsqueda de experiencias, desde quienes buscan comodidad hasta quienes priorizan la novedad y la aventura. Estos patrones permiten observar tendencias estructuradas que ofrecen una visión más afinada del funcionamiento del sistema turístico y de la interacción entre la oferta y la demanda

Además, los patrones turísticos están estrechamente vinculados con los estilos de vida y las predisposiciones psicológicas de los individuos. plantea que existe un continuo entre turistas psicocéntricos más conservadores y orientados a destinos

familiares (Plog, 1974) y allocéntricos, caracterizados por su tendencia a explorar lugares novedosos y menos masificados. Este enfoque ha permitido interpretar los desplazamientos turísticos como reflejo de rasgos de personalidad y de la relación que mantienen los individuos con el riesgo, la seguridad y la búsqueda de experiencias transformadoras. Así, el análisis de estos patrones constituye una herramienta estratégica para el diseño de productos turísticos ajustados a segmentos específicos.

En las últimas décadas, los patrones turísticos también han sido explicados desde modelos evolutivos que consideran la trayectoria del viajero a lo largo del tiempo. Pearce (2005), con su modelo Travel Career Pattern, argumenta que las motivaciones y preferencias turísticas no son estáticas, sino que cambian en función de la edad, la experiencia acumulada y el contexto social del individuo. Este enfoque secuencial permite comprender cómo se transforman las expectativas de los turistas a medida que avanzan en su "carrera de viaje", lo cual abre oportunidades para la planificación turística diferenciada y adaptada a etapas de vida específicas.

Asimismo, organismos internacionales como la Organización Mundial del Turismo (UNWTO) han destacado que los patrones turísticos están influenciados por fenómenos globales como la digitalización, la estacionalidad y las transformaciones económicas. La (World Tourism Organization, 2019) señala que el análisis de estos patrones es clave para predecir tendencias, gestionar flujos turísticos y promover la sostenibilidad en los destinos. En este sentido, los patrones turísticos representan una herramienta analítica esencial para comprender el comportamiento colectivo de los viajeros y orientar políticas y estrategias que respondan a dinámicas locales y globales.

En ese sentido, la identificación de patrones turísticos, es decir, combinaciones recurrentes de comportamientos: frecuencia de viaje, tipo de destinos, gasto,

motivaciones, perfil socioeconómico del hogar constituye un insumo esencial para diseñar sistemas recomendadores de viajes que respondan a perfiles reales de hogares, en lugar de estereotipos predefinidos.

2.2.2. Aprendizaje automático no supervisado

2.2.2.1. Concepto de aprendizaje no supervisado

El aprendizaje automático no supervisado reúne un conjunto de métodos que permiten descubrir estructuras, patrones y relaciones ocultas dentro de los datos sin necesidad de etiquetas previas. A diferencia del aprendizaje supervisado, donde se conocen las respuestas correctas, en este enfoque el algoritmo explora los datos buscando similitudes, agrupaciones o distribuciones que ayuden a comprender mejor su organización interna. Según Gerón (2020) este tipo de aprendizaje resulta fundamental en situaciones donde la información es abundante pero no está clasificada, lo que vuelve indispensable el uso de técnicas como el clustering, la reducción de dimensionalidad y la detección de anomalías.

En las ciencias sociales, el aprendizaje no supervisado ha cobrado especial relevancia porque permite revelar comportamientos colectivos, dinámicas sociales y perfiles heterogéneos sin imponer categorías preconcebidas. Esta metodología se utiliza para identificar segmentos poblacionales, examinar desigualdades y comprender fenómenos complejos donde los patrones no son evidentes a primera vista. Asimismo, técnicas como el análisis de componentes principales (PCA) o los algoritmos de clustering han permitido a investigadores explorar grandes bases de datos sociales con un nivel de detalle y profundidad difícil de alcanzar mediante métodos tradicionales.

En campos aplicados, el aprendizaje no supervisado también se ha consolidado como una herramienta clave para analizar comportamientos de consumo, movilidad y preferencias de usuarios. Estos métodos permiten crear tipologías, reducir ruido y comprender la variabilidad interna de los datos para apoyar la toma de decisiones basada en evidencia (Hastie et al., 2017). De este modo, el aprendizaje no supervisado no solo contribuye al análisis exploratorio, sino que también abre puertas a modelos predictivos y estrategias analíticas más robustas en entornos donde la información es amplia, diversa y cambiante.

2.2.2.2. Diferencias entre aprendizaje supervisado y no supervisado

En el campo del aprendizaje automático, las diferencias entre métodos supervisados y no supervisados se fundamentan principalmente en el tipo de información disponible durante el entrenamiento. El aprendizaje supervisado trabaja con datos etiquetados, es decir, ejemplos donde la respuesta correcta ya es conocida. Esto permite construir modelos capaces de predecir una categoría o un valor numérico a partir de patrones previos, lo que lo hace especialmente útil en tareas como clasificación de textos, diagnóstico automatizado o predicción de series temporales. Estos modelos aprenden estableciendo una relación clara entre las variables de entrada y la salida esperada, lo que facilita evaluar con precisión su desempeño (Gerón, 2020).

Por contraste, el aprendizaje no supervisado se desarrolla sin etiquetas ni respuestas predefinidas; su propósito es descubrir estructuras internas, agrupamientos o asociaciones dentro de los datos. Métodos como el clustering, la detección de anomalías o la reducción de dimensionalidad permiten revelar patrones que no son evidentes a simple vista y que pueden servir como base para decisiones exploratorias. (Murphy, 2022) señala que este enfoque es especialmente valioso cuando se trabaja con grandes

volúmenes de información social, económica o de comportamiento, donde la categorización previa puede ser insuficiente o inexistente.

Estas diferencias no solo son técnicas, sino también metodológicas. Mientras el aprendizaje supervisado busca optimizar la precisión de predicciones basadas en relaciones conocidas, el no supervisado ayuda a comprender la estructura subyacente de los datos, generando hipótesis que posteriormente pueden convertirse en modelos supervisados. (James et al., 2023) apuntan que ambos enfoques son complementarios: uno enfatiza la inferencia directa y el otro la exploración, y juntos aportan una visión más amplia para la investigación científica y aplicada.

2.2.2.3. Objetivo del clustering

El *clustering* se concibe como una técnica fundamental del aprendizaje no supervisado cuyo propósito central es organizar un conjunto de datos en grupos relativamente homogéneos. La idea es que los elementos que integran un mismo grupo compartan características similares, mientras que las diferencias entre grupos sean lo suficientemente claras para distinguir patrones o estructuras internas que no son visibles a primera vista. Esta forma de agrupamiento permite revelar relaciones naturales dentro de los datos sin necesidad de etiquetas previas, lo que resulta especialmente útil en investigaciones donde el conocimiento inicial sobre la distribución de las variables es limitado (Wainwright & Jordan, 2008).

En estudios sociales, turísticos y de comportamiento humano, el *clustering* se transforma en una herramienta que facilita comprender la diversidad de perfiles presentes en una población. A través de esta técnica, es posible identificar segmentos con intereses, comportamientos o condiciones socioeconómicas particulares, lo que permite posteriormente analizar necesidades específicas o diseñar estrategias

diferenciadas. Esta capacidad no radica solamente en agrupar datos, sino en ayudar al investigador a interpretar qué significado social o práctico pueden tener esos grupos dentro de un fenómeno más amplio (Xu & Tian, 2015b).

Además, el *clustering* contribuye a mejorar la toma de decisiones al ofrecer una visión más clara y ordenada de datos complejos. Su objetivo no se limita a simplificar la información: también ayuda a descubrir patrones emergentes y relaciones que pueden guiar el planteamiento de hipótesis o la creación de modelos predictivos posteriores. Por estas razones, el *clustering* se ha convertido en un componente habitual en investigaciones aplicadas, especialmente dentro del análisis de mercados turísticos, de hogares y en el estudio de dinámicas sociales donde la heterogeneidad de los individuos exige métodos que permitan reconocer subgrupos relevantes (Xu & Tian, 2015b).

2.2.2.4. Ventajas del clustering en la segmentación turística

La utilización de técnicas de *clustering* en el ámbito turístico ha adquirido relevancia debido a su capacidad para identificar patrones de comportamiento y preferencias sin necesidad de supuestos previos sobre los grupos que conforman la demanda. A diferencia de las segmentaciones tradicionales basadas en variables demográficas o criterios predefinidos, el *clustering* permite descubrir agrupamientos que emergen directamente de los datos, lo que ofrece una visión más fiel y matizada de la heterogeneidad del turista contemporáneo. Como señala (Dolničar, 2004), los enfoques empíricos basados en datos tienden a generar segmentos más estables y coherentes, al reflejar combinaciones reales de motivaciones, actitudes y comportamientos de viaje.

Además, estas técnicas contribuyen a mejorar la precisión en la toma de decisiones para el diseño de productos turísticos, campañas de marketing y estrategias de fidelización. Al agrupar a los visitantes según similitudes multidimensionales como patrones de gasto, actividades preferidas, duración del viaje o uso de plataformas digitales, el *clustering* permite a los gestores diferenciar sus servicios con mayor pertinencia. Estudios recientes demuestran que esta aproximación facilita la identificación de nichos específicos, desde turistas orientados al bienestar hasta segmentos interesados en experiencias locales sostenibles, optimizando así la asignación de recursos y ampliando el impacto de las intervenciones promocionales.

Por último, el *clustering* ofrece ventajas prácticas en contextos donde las bases de datos son extensas, heterogéneas o provienen de múltiples fuentes encuestas, plataformas digitales, sistemas de reservas, un escenario cada vez más común en los destinos turísticos inteligentes. La flexibilidad de estas técnicas facilita integrar variables cuantitativas y cualitativas, lo que fortalece el análisis interpretativo y permite revelar nuevas tendencias del mercado. En este sentido, se destaca que el *clustering* no solo permite definir segmentos, sino también comprender sus relaciones internas y su evolución temporal, un elemento clave para la planificación turística sostenible.

2.2.3. Ciencia de datos y minería de datos aplicada al turismo

2.2.3.1. Minería de datos: definición y etapas

La minería de datos constituye un proceso analítico orientado a descubrir patrones significativos, relaciones ocultas y estructuras no triviales dentro de grandes volúmenes de información. Este procedimiento forma parte fundamental del campo más amplio conocido como Knowledge Discovery in Databases (KDD), cuyo propósito es

transformar datos brutos en conocimiento útil y accionable. La minería de datos se describe como un conjunto de métodos estadísticos, algoritmos de aprendizaje automático y técnicas computacionales que permiten extraer información relevante desde bases de datos masivas, facilitando la toma de decisiones basada en evidencia en diversos sectores, incluida la inteligencia artificial. (Han et al., 2023) Complementariamente, se plantea que la minería de datos constituye la fase central del proceso KDD, al encargarse directamente de aplicar algoritmos para generar modelos y patrones que posteriormente deben evaluarse e interpretarse. (Fayyad et al., 1996)

En cuanto a sus etapas, la minería de datos se estructura generalmente en una secuencia metodológica que comprende la selección y preparación de los datos, seguida de la transformación y modelado, para finalmente culminar con la evaluación e interpretación de los resultados. La fase de preprocesamiento es esencial para garantizar la calidad del análisis, pues involucra la limpieza, integración y reducción de los datos, permitiendo corregir inconsistencias y manejar valores faltantes. (Larose Daniel & Larose Chantal, 2015)

Posteriormente, las técnicas de modelado que incluyen algoritmos supervisados y no supervisados, como árboles de decisión, redes neuronales o *clustering* son aplicadas para identificar patrones que puedan generalizarse. La etapa final, según (Witten et al., 2016), implica validar la utilidad del modelo, interpretar sus hallazgos y convertirlos en conocimiento comprensible para usuarios y tomadores de decisiones, cerrando así el ciclo del proceso KDD.

La minería de datos se ha consolidado como una herramienta estratégica en el ámbito de la ciencia de datos y la inteligencia artificial debido a su capacidad para automatizar el descubrimiento de conocimiento en entornos caracterizados por datos

complejos, heterogéneos y de gran volumen. La incorporación de enfoques modernos, como el aprendizaje profundo y la ingeniería de características automatizada, ha ampliado la eficiencia y aplicabilidad del proceso, permitiendo abordar problemas relacionados con predicción, segmentación, optimización y análisis conductual. En este sentido, la minería de datos no solo facilita el análisis descriptivo y exploratorio de datos, sino que también constituye un componente clave en la construcción de sistemas inteligentes capaces de aprender y mejorar continuamente a partir de la información (Han et al., 2023).

2.2.3.2. Importancia del análisis exploratorio en estudios sociales y turísticos

El análisis exploratorio de datos (en inglés, Exploratory Data Analysis EDA) representa una fase esencial en cualquier investigación social o turística, pues permite comprender la estructura, calidad y características fundamentales de los datos antes de realizar análisis más complejos o modelados. A través de técnicas descriptivas, gráficas y de visualización, el EDA posibilita la detección de errores, valores atípicos, distribución de variables, relaciones preliminares entre variables y patrones emergentes, lo cual ayuda a asegurar la validez interna de la investigación IBM (2019).

En el contexto específico de estudios turísticos o sociales, el EDA puede revelar dinámicas no esperadas, heterogeneidades regionales o demográficas, tendencias de demanda, segmentaciones de turistas o irregularidades en la recolección de datos, lo que permite ajustar el diseño del estudio, depurar la muestra o redefinir hipótesis antes de aplicar métodos confirmatorios. Esta etapa exploratoria contribuye a generar preguntas de investigación más precisas, a formular hipótesis fundamentadas y a preparar datos según los requisitos de los análisis posteriores (Fox & Lawless, 2014).

Al emplear EDA en estudios turísticos, se favorece una investigación más robusta y rigurosa: al contrastar expectativas teóricas con lo que los datos realmente muestran, el investigador puede identificar limitaciones, escenarios atípicos o fenómenos emergentes que no eran previsibles. Esto resulta especialmente importante cuando se trabaja con bases de datos grandes, variables heterogéneas (sociales, económicas, geográficas) o con mezcla de datos cuantitativos y cualitativos, condiciones habituales en la investigación turística contemporánea. De esta forma, el análisis exploratorio no solo prepara el terreno para modelos estadísticos, sino que contribuye a mejorar la calidad, relevancia y confiabilidad del estudio.

2.2.3.3. Técnicas no supervisadas en ciencias sociales

Las técnicas no supervisadas constituyen un conjunto de métodos estadísticos y computacionales que permiten identificar patrones, estructuras latentes y relaciones ocultas dentro de los datos sin necesidad de etiquetas o categorías previamente definidas. En las ciencias sociales, su uso ha crecido de manera significativa porque estos métodos facilitan la comprensión de fenómenos complejos como comportamientos colectivos, segmentación poblacional, redes de interacción o dinámicas espaciales a partir de grandes volúmenes de información. Los algoritmos de agrupamiento y reducción de dimensionalidad permiten descubrir configuraciones sociales emergentes que no siempre son visibles mediante técnicas descriptivas tradicionales. (Murtagh, 2015)

Entre las técnicas más empleadas se encuentran el análisis de conglomerados (cluster analysis), el análisis de componentes principales (ACP/PCA), el análisis factorial exploratorio, y los métodos basados en modelos de mezcla o machine learning no supervisado, como k-means, DBSCAN o t-SNE. Estas herramientas permiten, por

ejemplo, identificar perfiles sociodemográficos, agrupar comportamientos o reducir la complejidad de bases de datos para facilitar interpretaciones sociológicas. Según (Murtagh, 2015) estas técnicas proporcionan una base empírica sólida para clasificar unidades sociales cuando no se dispone de categorías previas, permitiendo una exploración más objetiva de los datos.

Su aporte en la investigación social radica en que permiten generar hipótesis fundamentadas, descubrir patrones inesperados y complementar enfoques teóricos ya establecidos. En estudios de opinión pública, movilidad humana, turismo, educación o desigualdad, el empleo de métodos no supervisados facilita comprender estructuras sociales subyacentes, identificar grupos emergentes y detectar tendencias latentes. Estas técnicas no solo contribuyen a describir realidades sociales complejas, sino que también fortalecen la construcción teórica, al aportar evidencia empírica sobre cómo se organizan y relacionan los actores dentro de un sistema social.

2.2.3.4. Aplicaciones previas de clustering en turismo y segmentación de hogares

Las técnicas de clustering han sido ampliamente utilizadas en el campo del turismo para identificar patrones de comportamiento, perfiles de visitantes y grupos de demanda homogéneos. Esta metodología permite a los investigadores comprender de manera más profunda la heterogeneidad de los turistas, facilitando la toma de decisiones en planificación, promoción y diseño de productos turísticos. (Dolničar, 2004a) el clustering ha permitido categorizar a los visitantes según motivaciones, preferencias, gasto y frecuencia de viaje, aportando una estructura analítica robusta para segmentar mercados turísticos complejos. Esta autora demuestra que la segmentación basada en datos supera a las clasificaciones tradicionales, pues revela grupos emergentes que no siempre se detectan mediante métodos subjetivos.

En cuanto a los hogares, las técnicas no supervisadas han contribuido a identificar patrones sociodemográficos y económicos que influyen en las decisiones de consumo turístico. Estudios como el de (Canh & Thanh, 2020) aplican algoritmos de clustering para segmentar hogares considerando ingresos, estructura familiar, nivel educativo y comportamientos de viaje, lo que permite comprender cómo las características del hogar condicionan la participación en actividades turísticas. Dichos análisis son fundamentales para el diseño de políticas públicas orientadas a democratizar el acceso al turismo, así como para empresas que buscan adaptar ofertas diferenciadas a diversos perfiles familiares.

A nivel metodológico, la literatura evidencia que el clustering proporciona una base empírica sólida tanto para la segmentación turística como para el análisis del comportamiento de los hogares. Su eficacia se relaciona con su capacidad para procesar grandes volúmenes de datos y descubrir estructuras ocultas dentro de ellos, lo que favorece investigaciones más precisas, especialmente en contextos donde los patrones son multidimensionales o no lineales (Xu & Tian, 2015). En el turismo y en los estudios de hogares, estas técnicas han permitido no solo describir perfiles, sino también anticipar tendencias y desarrollar modelos predictivos aplicados al marketing territorial y al análisis del bienestar social.

2.2.4. Técnicas de agrupamiento (clustering)

Las técnicas de agrupamiento constituyen un conjunto de métodos esenciales en el aprendizaje no supervisado, cuyo objetivo es identificar estructuras internas en los datos sin requerir etiquetas predefinidas. Estas técnicas permiten descubrir patrones, segmentar comportamientos y analizar relaciones latentes dentro de grandes volúmenes de información, lo que las vuelve especialmente relevantes en aplicaciones modernas de

ciencia de datos (Aggarwal & Reddy, 2016). El clustering se ha consolidado como un recurso central en áreas como marketing, bioinformática, análisis de redes y sistemas inteligentes, donde la detección temprana de patrones aporta ventajas analíticas y operativas.

En la última década, el desarrollo de nuevas variantes de algoritmos y métricas ha impulsado su uso en contextos de big data y análisis de alta dimensionalidad. La integración con técnicas de representación, como los embeddings o la reducción de dimensionalidad, ha fortalecido la precisión de los métodos de agrupamiento, permitiendo resultados más robustos ante ruido y heterogeneidad de los datos (Xu & Tian, 2015).

2.2.4.1. Concepto general de clustering

El clustering refiere al proceso de organizar un conjunto de datos en grupos internamente homogéneos pero distintivos entre sí, utilizando alguna medida de similitud o distancia para evaluar la proximidad entre las instancias (Aggarwal, 2018). Este enfoque permite revelar estructuras ocultas y comportamientos emergentes, funcionando como una herramienta exploratoria clave en proyectos de análisis de datos. Su naturaleza no supervisada lo hace especialmente útil en escenarios donde no existen etiquetas o clasificaciones previas.

Asimismo, la calidad del agrupamiento depende de la elección adecuada del algoritmo, de la escala de los atributos y del preprocesamiento aplicado. En los últimos años, se ha demostrado que métricas como la distancia euclidiana, el coseno o la distancia basada en kernels pueden influir significativamente en la forma final de los clústeres, por lo que la selección de parámetros y técnicas de normalización resulta crítica (Saxena et al., 2017).

2.2.4.2. Tipos de clustering

Los métodos de agrupamiento pueden organizarse en categorías según el criterio utilizado para definir la estructura de los clústeres. La literatura reciente distingue principalmente cuatro enfoques: particional, jerárquico, basado en densidad y basado en prototipos, cada uno con características que los hacen adecuados para diferentes distribuciones o tamaños de datos (Han et al., 2022). Esta diversidad metodológica permite seleccionar estrategias acordes al dominio de aplicación, ya sea para datos numéricos, espaciales, de redes o de alta dimensionalidad.

2.2.4.2.1. Clustering particional

El clustering particional divide los datos en un número fijo de grupos, optimizando un criterio que generalmente busca minimizar la variación interna de cada clúster. Algoritmos como k-means o k-medoids continúan siendo ampliamente utilizados por su eficiencia y capacidad de escalar a grandes volúmenes de información (Celebi et al., 2016). Su funcionamiento se basa en iteraciones sucesivas de asignación de puntos y actualización de centroides, lo que permite obtener soluciones estables en contextos prácticos.

No obstante, estos métodos requieren fijar el número de clústeres antes del proceso y son sensibles a valores atípicos, distribución de los datos y elección de la métrica de distancia. Por ello, en investigaciones recientes se recomienda complementar el algoritmo con técnicas de validación interna, como el índice de silueta o el criterio de Davies–Bouldin, para determinar el número óptimo de grupos (Arbelaitz et al., 2018).

2.2.4.2.2. Clustering jerárquico

El clustering jerárquico genera una estructura en forma de árbol (dendrograma) que permite examinar las relaciones entre instancias a múltiples niveles de granularidad. Este enfoque puede basarse en estrategias aglomerativas, donde los clústeres se fusionan progresivamente, o divisivas, donde un conjunto global se va subdividiendo (Murtagh & Legendre, 2019). Su principal fortaleza es que no requiere especificar inicialmente el número de grupos y ofrece una representación visual interpretable.

Sin embargo, estos métodos suelen tener mayores costos computacionales y no permiten deshacer decisiones previas de fusión o división. A pesar de ello, el clustering jerárquico mantiene relevancia en aplicaciones bioinformáticas, análisis de documentos o sistemas sociales, donde la interpretación multiescala del dendrograma aporta valor (Han et al., 2022).

2.2.4.2.3. Clustering basado en densidad

Los métodos basados en densidad identifican clústeres como regiones altamente pobladas separadas por áreas de menor densidad. Algoritmos como DBSCAN o HDBSCAN han ganado amplia popularidad por su capacidad para detectar grupos con formas no lineales y manejar eficientemente valores atípicos, considerándolos como ruido (Campello et al., 2015). Esta característica los hace especialmente útiles en datos espaciales, geográficos o con patrones irregulares.

A diferencia de los enfoques particionales, no requieren fijar el número de clústeres a priori, aunque sí dependen de parámetros como el radio de vecindad o la densidad mínima. Investigaciones recientes destacan que la variabilidad en la densidad

de los datos puede afectar la calidad del agrupamiento, por lo que se han desarrollado variantes adaptativas capaces de mitigar este problema (McInnes et al., 2017).

2.2.4.2.4. Clustering basado en prototipos

Los métodos de agrupamiento basados en prototipos representan cada clúster mediante un elemento central que sintetiza sus características. Este prototipo puede ser un centroide, un medoid o un vector de referencia calculado mediante procedimientos iterativos (Aggarwal, 2018). Al ser métodos computacionalmente eficientes, resultan adecuados para aplicaciones en tiempo real o análisis con grandes conjuntos de datos.

Su principal limitación radica en que asumen que los clústeres tienen formas compactas y relativamente esféricas, lo que no siempre se ajusta a la complejidad de datos reales. Aun así, continúan siendo ampliamente usados gracias a su simplicidad conceptual y a la mejora constante de variantes más robustas y menos sensibles al ruido (Celebi et al., 2016).

2.2.4.3. K-Means

2.2.4.3.1. Descripción

K-means es un algoritmo de agrupamiento particional ampliamente utilizado para organizar datos numéricos en k grupos definidos por su similitud interna. Su objetivo principal es identificar estructuras latentes mediante la asignación de puntos a clústeres representados por centroides, los cuales actúan como promedios de cada grupo (Aggarwal, 2018). Debido a su simplicidad conceptual y su rapidez en la convergencia, se mantiene como una de las herramientas más aplicadas en tareas exploratorias dentro de la ciencia de datos y la inteligencia artificial.

2.2.4.3.2. Funcionamiento

El funcionamiento de *k-means* se basa en un proceso iterativo que alterna entre asignar cada observación al centroide más cercano y recalcular esos centroides hasta alcanzar estabilidad. En cada iteración, el algoritmo evalúa la distancia de los puntos hacia los centros actuales, reorganiza las asignaciones y actualiza los centroides como la media de las observaciones pertenecientes al clúster (Han et al., 2022). Este mecanismo de realimentación permite optimizar la cohesión interna de los grupos, aunque su calidad depende de la inicialización de los centroides.

2.2.4.3.3. Requerimientos (variables numéricas)

El algoritmo requiere trabajar exclusivamente con variables numéricas, ya que utiliza medidas de distancia como la euclidiana, las cuales no son compatibles directamente con datos categóricos. Esta dependencia implica la necesidad de transformar o codificar variables no numéricas cuando se busca extender el algoritmo a dominios más complejos (Saxena et al., 2017). De igual forma, se recomienda normalizar o estandarizar los atributos para evitar que aquellos con mayor escala dominen la función de distancia.

2.2.4.3.4. Ventajas

Entre sus principales ventajas destacan su elevada eficiencia computacional, la capacidad de escalar a grandes volúmenes de datos y su facilidad de implementación. Estas características han permitido que *k-means* continúe siendo una opción preferida en aplicaciones que requieren segmentaciones rápidas, como análisis de clientes, procesamiento de imágenes o agrupamiento preliminar en procesos de minería de datos

(Aggarwal & Reddy, 2016). Su comportamiento determinista en datos bien distribuidos también favorece su utilización como punto de partida para métodos más avanzados.

2.2.4.3.5. Limitaciones

A pesar de su utilidad, *k-means* presenta limitaciones importantes relacionadas con su sensibilidad a valores atípicos y a la presencia de clústeres no esféricos o con densidades heterogéneas. Asimismo, requiere definir previamente el número de grupos, lo que puede introducir sesgos si no se aplica una validación adecuada. Estudios recientes han mostrado que el algoritmo también depende fuertemente de la inicialización de los centroides, lo que puede conducir a soluciones locales subóptimas (Celebi et al., 2016). Estas restricciones han motivado el desarrollo de variantes y mejoras que buscan aumentar su estabilidad y robustez.

2.2.4.4. K-Medoides (PAM)

2.2.4.4.1. Descripción

El método *K-Medoids*, especialmente en su implementación PAM *(Partitioning Around Medoids),* pertenece a la familia de algoritmos de agrupamiento particional y se caracteriza por seleccionar observaciones reales del conjunto de datos como representantes de cada grupo, conocidas como medoides (Aggarwal & Reddy, 2016). Dichos representantes se determinan a partir de la minimización de la disimilitud total entre los elementos asignados a un mismo clúster, lo que permite una representación más robusta frente a valores atípicos (Han et al., 2022). En contraste con los enfoques que utilizan centroides calculados a partir de promedios, este algoritmo admite el uso de funciones de distancia más flexibles, resultando particularmente eficaz en contextos con ruido elevado o cuando las relaciones entre los datos no cumplen estrictamente

propiedades métricas (Aggarwal & Reddy, 2016). La estrategia de *K-Medoids* ha demostrado ser efectiva en contextos donde la interpretabilidad de los clústeres y la robustez frente a valores extremos son prioritarias (Saxena et al., 2017).

2.2.4.4.2. Funcionamiento

El procedimiento PAM comienza con la identificación de un conjunto inicial de observaciones que actúan como posibles medoides, las cuales representan provisionalmente a los clústeres del conjunto de datos (Kaufman & Rousseeuw, 2009). A partir de esta selección, el algoritmo desarrolla un proceso iterativo compuesto por una etapa de asignación, donde cada elemento es vinculado al medoide con menor disimilitud, y una etapa de refinamiento enfocada en minimizar el costo total del agrupamiento (Celebi et al., 2016). Durante el refinamiento, se analizan de manera sistemática posibles reemplazos entre medoides actuales y puntos no seleccionados, evaluando si dichos intercambios generan una reducción de la función objetivo, lo que favorece una exploración más rigurosa de soluciones en comparación con otros métodos de agrupamiento particional (Kaufman & Rousseeuw, 2009). Aunque este enfoque implica un mayor costo computacional, proporciona una solución más estable y confiable en presencia de ruido o irregularidades en los datos (McInnes et al., 2017).

2.2.4.4.3. Diferencias respecto a K-Means

El algoritmo *K-Medoids* presenta diferencias sustanciales frente a *K-Means*, especialmente en el criterio utilizado para representar cada clúster y en el modo en que se cuantifica la disimilitud total entre los elementos agrupados (Aggarwal, 2018).
A diferencia de K-Means, que define sus clústeres a partir de centroides obtenidos mediante promedios aritméticos, *K-Medoids* selecciona observaciones reales del conjunto de datos como representantes del grupo (Han et al., 2022). Esta característica

permite disminuir de manera notable la sensibilidad del algoritmo frente a valores atípicos y distribuciones asimétricas, favoreciendo una estructura de agrupamiento más robusta (Aggarwal, 2018). Además, *K-Medoids* no requiere asumir formas esféricas en los clústeres ni depender exclusivamente de distancias euclidianas, permitiendo trabajar con matrices de disimilitud arbitrarias, lo cual amplía su aplicabilidad a dominios con variables categóricas o métricas complejas (Saxena et al., 2017). Por ello, aunque K-Means es más eficiente computacionalmente, *K-Medoids* ofrece mayor robustez y flexibilidad metodológica.

2.2.4.4.4. Ventajas cuando existen outliers o datos mixtos

Una ventaja destacada de *K-Medoids* es su baja sensibilidad a valores atípicos, ya que los representantes de los clústeres corresponden a observaciones reales y no se ven afectados por magnitudes extremas (Celebi et al., 2016). En comparación, los algoritmos basados en centroides pueden experimentar desplazamientos considerables cuando los datos presentan ruido o anomalías (McInnes et al., 2017). Asimismo, la capacidad de *K-Medoids* para trabajar con diversas métricas de disimilitud lo hace adecuado para conjuntos de datos heterogéneos (Aggarwal & Reddy, 2016). Esta característica resulta especialmente relevante cuando se emplean medidas como la distancia de Gower, superando a métodos limitados a variables exclusivamente numéricas (Han et al., 2022).

2.2.4.5. K-Prototypes

2.2.4.5.1. Ideal para variables mixtas (numéricas + categóricas)

El algoritmo *K-Prototypes* se desarrolló para trabajar con conjuntos de datos que incluyen tanto variables numéricas como categóricas, lo que permite agrupar

observaciones heterogéneas bajo una métrica adecuada para ambos tipos (Huang, 1998). Este enfoque integra conceptos de *K-Means* y *K-Medoids*, utilizando una medida de distancia híbrida que evita la conversión forzada de categorías o la pérdida de información simbólica (Ahmad & Khan, 2019). Estudios recientes resaltan su eficacia en contextos donde coexisten datos cuantitativos y nominales, proporcionando una alternativa más eficiente que los métodos que procesan únicamente un tipo de variable (Bholowalia & Kumar, 2019). Además, se ha mostrado aplicable en entornos reales de diversa naturaleza, consolidándose como una herramienta práctica para la segmentación de datos mixtos (Xu & Tian, 2021).

2.2.4.5.2. Uso en investigaciones con hogares

El uso de *K-Prototypes* se ha consolidado en investigaciones socioeconómicas y demográficas, especialmente para el análisis de hogares, donde los conjuntos de datos incluyen variables mixtas como ingresos, gastos, composición familiar, educación o tipo de vivienda (Hossain et al., 2020). Esta técnica permite identificar segmentos poblacionales sin necesidad de transformar categorías, preservando la estructura original de los datos (Abdi & Rokonuzzaman, 2021). Además, su flexibilidad facilita combinar indicadores económicos continuos con características sociales cualitativas, lo que permite generar interpretaciones más precisas sobre la realidad de los hogares (Hossain et al., 2020).

2.2.4.5.3. Manejo de distancias mixtas

El principio fundamental de *K-Prototypes* se basa en una función de distancia híbrida que maneja simultáneamente diferencias numéricas mediante medidas euclidianas y discrepancias categóricas mediante coincidencias o disimilitudes simples (Huang, 1998). Para equilibrar ambos tipos de variables, el algoritmo incorpora un

parámetro de ponderación que ajusta la contribución relativa de cada dominio, evitando sesgos por escala o frecuencia de categorías (de Amorim & Hennig, 2015). Investigaciones recientes indican que un ajuste adecuado de este parámetro mejora la cohesión interna de los clústeres y aumenta su estabilidad frente a cambios en la composición del conjunto de datos (Wang et al., 2019). Asimismo, se ha observado que una calibración precisa favorece la consistencia de los agrupamientos y su robustez frente a variaciones en los datos (Zhang & Zheng, 2022).

2.2.4.5.4. Ventajas y limitaciones

Una ventaja clave de *K-Prototypes* es su capacidad de ofrecer una solución eficiente en tiempo y memoria incluso cuando los conjuntos de datos son extensos, además de su flexibilidad para modelar dominios mixtos sin requerir codificaciones complejas o expansivas (Bholowalia & Kumar, 2019). También destaca por generar agrupamientos interpretables, ya que los prototipos combinan medias para los valores numéricos y modos para los atributos categóricos, favoreciendo la lectura analítica de los perfiles producidos. A pesar de sus ventajas, *K-Prototypes* presenta algunas limitaciones. Su desempeño puede depender de la selección inicial de los clústeres y de la estandarización de las variables numéricas, y no siempre logra capturar relaciones complejas entre categorías o interacciones no lineales (de Amorim & Hennig, 2015). Sin embargo, su combinación de eficiencia, simplicidad y capacidad de representar datos mixtos lo mantiene como una opción ampliamente utilizada para la segmentación de conjuntos heterogéneos (Zhang & Zheng, 2022).

2.2.4.6. Selección del algoritmo adecuado según el tipo de variables turísticas

2.2.4.6.1. Variables socioeconómicas

En estudios turísticos, las variables socioeconómicas —como nivel de ingresos, educación, estructura del hogar o actividad laboral— requieren algoritmos de agrupamiento capaces de procesar simultáneamente información numérica y categórica, ya que estas características definen con precisión los perfiles de los visitantes. En este sentido, la literatura reciente señala que técnicas como *K-Prototypes* permiten modelar de manera más adecuada las relaciones mixtas al integrar métricas heterogéneas dentro de la misma estructura de análisis (Abdi & Rokonuzzaman, 2021). Asimismo, se ha observado que los métodos basados en disimilitudes mixtas tienden a mejorar la coherencia de los segmentos socioeconómicos al evitar pérdidas de información asociadas a la recodificación forzada de categorías (Hossain et al., 2020). Esto resulta fundamental para interpretar comportamientos turísticos asociados a condiciones sociales y económicas, tal como destacan Xu y Tian (2021), quienes enfatizan la importancia de seleccionar algoritmos adecuados para garantizar un análisis segmentado confiable.

2.2.4.6.2. Gastos

Cuando se analizan los patrones de gasto turístico —tales como presupuesto total, distribución del gasto o variabilidad en el consumo— es necesario emplear algoritmos que operen de manera eficiente con variables numéricas con diferentes escalas. Métodos como *K-Means* son ampliamente utilizados para identificar perfiles de gasto debido a su capacidad para optimizar particiones basadas en distancias euclidianas entre valores continuos (Bholowalia & Kumar, 2019). Sin embargo, la literatura advierte que la presencia de valores atípicos puede alterar significativamente los

centroides, por lo que algunos autores recomiendan integrar variantes más robustas, como *K-Medoides*, especialmente cuando los montos presentan alta dispersión (Wang et al., 2019). Además, estudios recientes han señalado que la selección del algoritmo condiciona directamente la precisión de los segmentos financieros, lo cual es determinante para el diseño de estrategias de marketing turístico (Zhang & Zheng, 2022).

2.2.4.6.3. Elecciones de viaje

Las elecciones de viaje, que abarcan aspectos como tipo de alojamiento, motivos del viaje, preferencias culturales o medio de transporte, se componen mayoritariamente de variables categóricas que requieren algoritmos diseñados para procesar simbolismos en lugar de magnitudes numéricas. Según Ahmad y Khan (2019), los algoritmos basados en coincidencias categóricas —como *K-Modes*— permiten capturar patrones de preferencia sin distorsionar la naturaleza no numérica de estos atributos. De igual forma, de Amorim y Hennig (2015) señalan que las métricas categóricas permiten definir segmentos más estables, particularmente cuando existen múltiples opciones discretas que describen el comportamiento del turista. Como indica Hossain et al. (2020), este tipo de algoritmos mejora significativamente la identificación de perfiles turísticos motivacionales, ofreciendo resultados más interpretables en comparación con métodos centrados en medias.

2.2.4.6.4. Frecuencia y comportamiento

El análisis de la frecuencia de viaje y de los patrones conductuales —como periodicidad, duración de la estancia o fidelidad al destino— combinan variables numéricas de conteo con atributos categóricos asociados al comportamiento turístico. De acuerdo con Xu y Tian (2021), esta mezcla de atributos favorece el uso de

algoritmos híbridos como *K-Prototypes*, los cuales integran componentes numéricos y simbólicos dentro de un mismo marco de similitud. Estudios adicionales han señalado que, cuando los comportamientos presentan alta variabilidad, métodos robustos como *K-Medoides* o los algoritmos basados en densidad resultan más adecuados para capturar dinámicas complejas sin que el ruido afecte la estructura del clúster (Wang et al., 2019). Además, Zhang y Zheng (2022) destacan que la selección del algoritmo influye directamente en la estabilidad de los segmentos relacionados con la frecuencia de viaje, lo que impacta en la planificación de estrategias turísticas basadas en lealtad y patrones de retorno.

2.2.5. Encapsulamiento de modelos

El encapsulamiento de modelos en ciencia de datos se refiere al proceso de aislar un modelo entrenado dentro de un entorno controlado que preserve sus dependencias, versiones, parámetros y estructura de ejecución, lo cual permite reproducibilidad y despliegue consistente en múltiples plataformas. Según Sculley et al. (2015), el aislamiento del modelo reduce la fricción operativa y mitiga riesgos de incompatibilidades al integrarlo en sistemas productivos. En investigaciones recientes, se resalta que encapsular un modelo facilita su transporte entre equipos de desarrollo, servidores de inferencia o contenedores, manteniendo inalterable su comportamiento en entornos heterogéneos (Zaharia et al., 2018). Esta práctica se ha consolidado como un eje central del ciclo de vida MLOps debido a su impacto en la escalabilidad y trazabilidad del modelo.

2.2.5.1. Uso de archivos PKL para despliegue de modelos

El encapsulamiento de modelos consiste en conservar la estructura interna y los parámetros aprendidos por un algoritmo para facilitar su reutilización en entornos de prueba y producción, siendo los archivos pickle (.pkl) uno de los formatos más empleados en Python debido a su capacidad para serializar objetos complejos sin perder su estado (Van Rossum & Drake, 2009). En el ámbito del aprendizaje automático, esta técnica permite almacenar modelos entrenados, preprocesadores y pipelines completos, lo que agiliza su transporte entre sistemas operativos o arquitecturas de cómputo heterogéneas (Pedregosa et al., 2011). La persistencia mediante PKL favorece la trazabilidad y la reproducibilidad, puesto que garantiza que la versión exacta del modelo utilizado durante el entrenamiento pueda ser cargada posteriormente en aplicaciones de inferencia, APIs o interfaces interactivas (Geron, 2019). No obstante, su uso requiere precauciones de seguridad, ya que los archivos PKL pueden ejecutar código arbitrario al deserializarse, lo que obliga a utilizarlos únicamente desde fuentes confiables (Python Software Foundation, 2024).

2.2.6. Interfaces web

Las interfaces web permiten que modelos de inteligencia artificial sean utilizados por usuarios finales de forma intuitiva a través de navegadores, transformando procesos complejos de inferencia en experiencias accesibles y visualmente interpretables. Como señalan Ribeiro et al. (2016), la accesibilidad es fundamental cuando se busca aplicar IA en ámbitos no técnicos donde la interpretación y la interacción deben ser claras. Investigaciones recientes enfatizan que el diseño de interfaces web para modelos debe considerar usabilidad, latencia y retroalimentación visual para facilitar una adopción adecuada del sistema (Belgrave & Brown, 2021). Por

ello, las interfaces se han convertido en un puente entre el desarrollo técnico y la apropiación social del modelo.

2.2.6.1. HTML5 como estándar para visualización

HTML5 se consolidó como el estándar principal para construir interfaces web modernas gracias a su soporte nativo para componentes multimedia, elementos semánticos y APIs que simplifican la comunicación entre el navegador y aplicaciones interactivas (W3C, 2017). En sistemas de ciencia de datos, su utilización permite estructurar *dashboards*, formularios de entrada y secciones informativas de manera ordenada, garantizando accesibilidad y compatibilidad con múltiples dispositivos (Freeman & Robson, 2018).

2.2.6.2. CSS3 para estilos adaptativos

CSS3 complementa a HTML5 y permite dotar a las interfaces de estilos responsivos, tipografías dinámicas y disposiciones flexibles sin modificar la lógica del modelo o la estructura del documento (Meyer, 2018). Su capacidad para aplicar diseños adaptativos facilita la creación de entornos interactivos orientados a usuarios no técnicos, lo que es fundamental cuando se integran modelos predictivos para la toma de decisiones institucionales (Keith, 2020).

2.2.6.3. Streamlit como framework para despliegue rápido

Streamlit se ha posicionado como una herramienta clave para el desarrollo de aplicaciones de ciencia de datos debido a su capacidad para generar interfaces interactivas a partir de scripts en Python sin necesidad de conocimientos avanzados de HTML o JavaScript (Treuille & Teixeira, 2020). Este *framework* permite cargar modelos encapsulados en PKL, procesar entradas del usuario y mostrar resultados en

tiempo real, lo que reduce la brecha entre el análisis técnico y su consumo por parte de actores operativos o gerenciales (Streamlit Inc., 2023).

2.2.7. Versionamiento y publicación del repositorio en la nube

El versionamiento en ciencia de datos implica registrar formalmente cambios en el código, datos y modelos para garantizar trazabilidad, reproducibilidad y control de calidad. Según Luo et al. (2018), la ausencia de versionamiento crea inconsistencias que impactan en los resultados del modelo y dificultan auditorías posteriores. Herramientas como Git y DVC han permitido estandarizar el registro de cambios no solo en código, sino también en grandes volúmenes de datos y artefactos de entrenamiento, una práctica cada vez más relevante para pipelines de IA (Zhang et al., 2020).

2.2.7.1. Git como sistema de control de versiones

Git es un sistema distribuido diseñado para gestionar cambios en archivos y proyectos de software, permitiendo mantener historiales completos, generar ramas de experimentación y asegurar la trazabilidad de modificaciones en modelos, scripts y documentos de investigación (Chacon & Straub, 2014). Su uso en ciencia de datos facilita la coordinación entre equipos, el rollback a versiones estables y el seguimiento del ciclo de vida del modelo desde el preprocesamiento hasta su despliegue (Wilson et al., 2017).

2.2.7.2. GitHub como plataforma de colaboración y publicación

GitHub actúa como una plataforma en la nube que amplía las capacidades de Git al ofrecer almacenamiento remoto, automatización mediante actions, control de accesos y espacios de documentación pública o privada (GitHub, 2024). En proyectos de

inteligencia artificial, su utilización posibilita publicar repositorios, gestionar incidencias, desplegar modelos mediante CI/CD y compartir resultados con revisores, tutores o equipos de desarrollo. Además, GitHub facilita la integración de aplicaciones web —como interfaces creadas en Streamlit— y la distribución estructurada de modelos encapsulados en PKL dentro de entornos reproducibles (Loeliger & McCullough, 2012).

2.2.8. Evaluación de la calidad de los clusters

La evaluación de la calidad de los clústeres es esencial en el análisis no supervisado porque permite determinar si los grupos encontrados representan patrones reales dentro de los datos o si son simplemente un resultado artefactual del algoritmo utilizado. Esta validación se apoya en métricas que cuantifican cohesión interna y separación externa, proporcionando una base objetiva para comparar diferentes configuraciones de agrupamiento (Tan et al., 2019). La literatura indica que sin esta verificación cuantitativa es difícil justificar la utilidad estadística y práctica de los segmentos obtenidos (Arbelaitz et al., 2013), especialmente en aplicaciones donde los resultados deben ser interpretables y reproducibles (Xu & Tian, 2015).

2.2.8.1. Necesidad de evaluar los clusters en estudios no supervisados

En contextos no supervisados, donde no existen etiquetas que permitan validar directamente la asignación de cada observación a un grupo, se vuelve indispensable aplicar métricas internas y externas para evitar conclusiones basadas en supuestos subjetivos. Xu y Tian (2015) destacan que esta evaluación es un mecanismo clave para determinar si un algoritmo ha logrado capturar estructuras subyacentes de manera estable. Asimismo, Arbelaitz et al. (2013) señalan que la evaluación sistemática evita

problemas como el sobreajuste, la creación artificial de clústeres o la interpretación errónea de patrones que pueden derivar únicamente de ruido en los datos.

2.2.8.2. Método del Codo (Elbow Method)

2.2.8.2.1. Explicación

El Método del Codo consiste en calcular la inercia o suma total de cuadrados dentro de los clústeres para diferentes valores de k, graficando posteriormente estos valores para identificar un punto donde la reducción de la inercia deje de ser significativa. Tan et al. (2019) explican que este "codo" refleja el punto de equilibrio entre la complejidad del modelo y su capacidad de describir adecuadamente los datos. De acuerdo con Arbelaitz et al. (2013), esta técnica es útil porque ofrece una forma visual y directa de estimar el número adecuado de clústeres sin requerir métricas complejas.

2.2.8.2.2. Interpretación

La interpretación del codo se basa en localizar la "rodilla" de la curva, es decir, el punto a partir del cual aumentar k genera mejoras marginales en la cohesión interna. Tan et al. (2019) indican que este punto representa una solución parsimoniosa donde el algoritmo alcanza un equilibrio razonable entre robustez y simplicidad. Según Xu y Tian (2015), elegir un k mayor al indicado por el codo conlleva el riesgo de crear grupos artificiales sin valor interpretativo, afectando la estabilidad del modelo.

2.2.8.2.3. Usos y límites

Aunque es una herramienta muy usada en análisis exploratorio, el Método del Codo presenta limitaciones. Arbelaitz et al. (2013) documentan que en muchos

conjuntos de datos reales el codo no se observa con claridad, lo que dificulta la toma de decisiones. Además, Tan et al. (2019) advierten que esta técnica solo considera variabilidad interna y no evalúa la separación entre grupos, por lo que se recomienda complementarla con índices cuantitativos como *Silhouette* o *Calinski-Harabasz.*

2.2.8.3. Coeficiente de Silhouette

2.2.8.3.1. Fórmula

El coeficiente de Silhouette se define mediante la expresión:

$$s(i) = \frac{b(i) - a(i)}{\max [a(i), b(i)]} \quad (1)$$

Ecuación 1 Ecuación del coeficiente de Silhoutte

donde a (i) es la distancia promedio del punto i respecto a su propio clúster, y b(i) corresponde a la menor distancia promedio entre i y cualquier otro clúster. Rousseeuw (1987) introdujo esta medida como una herramienta que combina cohesión y separación en un mismo indicador, manteniendo valores entre −1 y 1.

2.2.8.3.2. Interpretación

De acuerdo con Rousseeuw (1987), valores positivos cercanos a 1 indican que la observación está correctamente asignada a su grupo, mientras que valores cercanos a 0 reflejan casos fronterizos y valores negativos sugieren que el punto podría haber sido asignado a un clúster distinto. Tan et al. (2019) añaden que el promedio del coeficiente sobre todas las observaciones es útil para comparar la calidad entre modelos con diferente cantidad de clústeres.

2.2.8.3.3. Aplicación en datos turísticos

En investigaciones turísticas, el coeficiente de *Silhouette* permite evaluar si los segmentos obtenidos —por ejemplo, perfiles de gasto, preferencias de viaje o comportamientos de frecuencia— representan grupos coherentes. Xu y Tian (2015) sostienen que esta métrica es especialmente útil cuando las variables son multidimensionales y requieren un criterio cuantitativo para determinar la separación entre segmentos. De Amorim y Hennig (2015) advierten, sin embargo, que cuando los datos contienen variables mixtas, la interpretación debe complementarse con métricas adaptadas a distancias híbridas.

2.2.8.4. Índice Calinski-Harabasz

El índice Calinski-Harabasz cuantifica la relación entre la dispersión entre clústeres y la dispersión dentro de ellos, favoreciendo particiones donde los grupos estén bien separados y sean internamente compactos. Arbelaitz et al. (2013) destacan su eficiencia computacional y su capacidad de producir resultados confiables en modelos particionales como *k-means*. Tan et al. (2019) señalan que este índice tiende a ser estable incluso ante variaciones en la inicialización, aunque en algunos casos puede privilegiar soluciones con clústeres de tamaño desigual.

2.2.8.5. Índice Davies-Bouldin

El índice Davies-Bouldin evalúa la similitud entre cada clúster y su vecino más cercano, combinando medidas de dispersión interna y distancia entre centroides. Davies y Bouldin (1979) argumentan que valores bajos de este índice indican agrupamientos adecuados. No obstante, Arbelaitz et al. (2013) indican que su rendimiento disminuye ante clústeres no esféricos o con densidades heterogéneas. Tan et al. (2019)

recomiendan utilizarlo junto a otros indicadores para obtener una evaluación más completa.

2.2.8.6. Elección del número óptimo de clusters en variables mixtas

La determinación del número óptimo de clústeres en variables mixtas requiere integrar métricas tradicionales con distancias híbridas que combinen similitud numérica y categórica. Xu y Tian (2015) sostienen que la mejor práctica consiste en utilizar múltiples índices (p. ej., Silhouette y CH) adaptados a distintas características del conjunto de datos. De Amorim y Hennig (2015) explican que este tipo de datos exige ponderaciones específicas para evitar que una variable domine la distancia total. Finalmente, Wang et al. (2019) recomiendan complementar los índices cuantitativos con análisis de estabilidad y criterios de interpretabilidad según el dominio de estudio.

2.2.9. Fundamentación estadística y matemática del modelo

La fundamentación matemática en los modelos de clustering se sustenta en la descripción formal de las métricas, transformaciones y funciones que permiten organizar datos según su proximidad, lo que ayuda a comprender el funcionamiento interno de los algoritmos y a justificar su validez teórica (Hastie et al., 2009). En este sentido, las distancias, los procesos de normalización y las medidas de similitud representan la base cuantitativa que determina cómo se agrupan los elementos en espacios métricos o no métricos (Jain, 2010). Asimismo, estos fundamentos son clave para asegurar interpretabilidad y reproducibilidad, especialmente en aplicaciones donde las relaciones entre observaciones pueden ser no lineales o estar condicionadas por escalas heterogéneas (Aggarwal, 2015).

2.2.9.1. Distancias utilizadas en clustering

Las distancias empleadas en algoritmos de clustering permiten cuantificar el grado de cercanía entre observaciones dentro del espacio de características, influyendo directamente en la estructura final de los grupos formados (Tan et al., 2019). La selección de la métrica adecuada es esencial, ya que diversas formas de datos (numéricos, categóricos, mixtos) requieren criterios de proximidad distintos para evitar distorsiones en la representación geométrica del conjunto (Aggarwal, 2015). En consecuencia, comprender las propiedades matemáticas de cada distancia asegura una mayor coherencia entre los patrones subyacentes y la segmentación resultante (Jain, 2010).

2.2.9.1.1. Euclidiana

La distancia euclidiana representa la medida geométrica clásica basada en la raíz cuadrada de la suma de los cuadrados de las diferencias entre cada dimensión, lo que la convierte en un estimador natural en espacios continuos (Bishop, 2006). Su interpretación se asocia directamente con la noción física de desplazamiento en un plano o hiperplano, y su uso es especialmente apropiado cuando las variables poseen escalas comparables y no existen correlaciones extremas entre ellas (Hastie et al., 2009). No obstante, esta métrica puede ser sensible a valores atípicos o a diferencias de magnitud entre características, motivo por el cual frecuentemente se acompaña de procesos de estandarización (Aggarwal, 2015).

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

Ecuación 2 Ecuación de la distancia Euclidiana

2.2.9.1.2. Manhattan

La distancia Manhattan, también conocida como distancia L1 o “taxicab”, calcula la suma de las diferencias absolutas entre atributos, lo que la hace más robusta frente a valores atípicos y especialmente útil en espacios de alta dimensionalidad (Bishop, 2006). A diferencia de la distancia euclidiana, esta métrica no penaliza de manera cuadrática las discrepancias entre observaciones, lo que permite capturar relaciones lineales o patrones dispersos con menor distorsión (Aggarwal, 2015). Su aplicación resulta frecuente en modelos donde la estructura de los datos puede representarse como desplazamientos rectangulares o donde la estabilidad ante ruido es un requisito (Tan et al., 2019).

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} | x_i - y_i | \quad (3)$$

Ecuación 3 Ecuación de la distancia de Manhattan

2.2.9.1.3. Dissimilarity (para variables categóricas)

La dissimilaridad para variables categóricas se basa en el conteo de coincidencias y diferencias entre atributos cualitativos, permitiendo medir proximidad sin necesidad de convertir la información en valores numéricos (Gower, 1971). Esta métrica resulta indispensable en escenarios donde las variables nominales prevalecen, ya que evita imponer relaciones artificiales entre categorías que no poseen magnitud ni distancia real (Kaufman & Rousseeuw, 2005). Su implementación, incluyendo variantes como la métrica de Gower, facilita el análisis de bases heterogéneas y el trabajo con algoritmos especializados en datos mixtos (Aggarwal, 2015).

2.2.9.2. Escalamiento y normalización

Los procesos de escalamiento y normalización permiten transformar las variables para garantizar que todas contribuyan equitativamente al cálculo de proximidades, evitando que atributos con rangos amplios dominen la estructura del clustering (Jain, 2010). Técnicas como la estandarización z-score o la normalización min-max son esenciales en algoritmos sensibles a magnitudes, como K-means o métodos basados en distancias euclidianas (Hastie et al., 2009). Adicionalmente, estos procedimientos favorecen la estabilidad numérica y mejoran la convergencia de los modelos al reducir efectos indeseados de heterogeneidades entre escalas (Tan et al., 2019).

2.2.9.3. Medidas de similitud en recomendadores

En los sistemas recomendadores, las medidas de similitud permiten identificar relaciones entre usuarios u objetos, empleando métricas como el coseno, la correlación de Pearson o la similitud basada en vectores dispersos (Ricci et al., 2015). Estas métricas capturan afinidades que no necesariamente dependen de la magnitud absoluta de los valores sino de la dirección o el patrón de comportamiento, lo que fortalece la capacidad predictiva de los modelos colaborativos (Aggarwal, 2016). En consecuencia, la correcta elección de la similitud influye de manera significativa en la calidad de las recomendaciones y en la personalización lograda para cada usuario (Jannach et al., 2011).

CAPÍTULO 3

3. DESARROLLO

3.1. Enfoque de la investigación

Dados los tipos de datos que se recopilados, su enfoque será cuantitativo. El cual "parte de una idea, que va acotándose y, una vez delimitada, se derivan objetivos y preguntas de investigación, se revisa la literatura y se construye un marco o una perspectiva teórica" (Hernández et al., 2010, p. 4). En términos de temporalidad, será transversal ya que recopilan datos en un momento o tiempo determinado. Su propósito es describir las variables y examinar cómo se relacionan entre sí durante un periodo de tiempo.

3.2. Población objetivo

La población objetivo de estudio se refiere al conjunto de casos que serán examinados y analizados en un estudio particular y debe estar definido, limitado y accesible. Esto puede incluir personas, animales, muestras biológicas, transporte, hospitales, instalaciones, familias y organizaciones, entre otros. (Gómez et al., 2016) Este estudio está conformado por 3458 hogares ecuatorianos registrados en la base de datos del Sistema Banco de Información (SIB) correspondiente al año 2021. Estos hogares brindaron información sobre sus actividades y comportamiento turístico en el país y se convirtieron en la unidad de análisis de estudio.

3.3. Técnicas e instrumentos de recopilación de la información

La técnica de la investigación de datos para Falcón y Herrera (2005) se "entiende como técnica, el procedimiento o forma particular de obtener datos o información" (p. 12). Este estudio no utilizó encuestas directas, sino que se basó en la recopilación y el análisis de datos secundarios a través de la disponibilidad de información previamente recopilada por instituciones oficiales.

3.3.1. Instrumentos de recopilación de datos

Un instrumento de recopilación de datos es esencial porque puede utilizarlo para acceder y extraer información sobre cualquier fenómeno (De Aguiar, 2016). En esta investigación, el instrumento utilizado es la base de datos oficial SIB, diseñada y estructurada por el organismo responsable de la elaboración de estadísticas sobre el sector turístico ecuatoriano. Esta base de datos integra información sociodemográfica de los hogares, registros de viajes realizados dentro y fuera de Ecuador, datos sobre gastos relacionados con viajes, así como los motivos, la frecuencia y las características de dichos viajes.

3.3.2. Recolección de la información

La información utilizada en este estudio proviene de datos secundarios proporcionados por el Instituto Nacional de Estadística y Censos (INEC), obtenidos a través del Sistema Banco de Información (SBI), plataforma oficial que recopila y difunde información estadística validada y estandarizada sobre las actividades turísticas de los hogares ecuatorianos (INEC, s. f.).

3.3.3. Técnicas para el preprocesamiento y análisis de datos

Inicialmente, la información fue organizada y depurada para asegurar su coherencia y asegurar la correcta interpretación de las variables incluidas en el estudio. Este proceso incluyó limpiar valores atípicos o faltantes, estandarizar las categorías y preparar las variables necesarias para análisis posteriores. Una vez estructurada la base, se realizó un análisis exploratorio de datos para describir el comportamiento del turista doméstico ecuatoriano mediante métodos descriptivos univariados y bivariados.

Para cumplir con el objetivo se utilizaron algoritmos no supervisados utilizando el lenguaje de programación Python, lo que permitió el uso, evaluación y optimización del modelo elegido. Todas las tareas de análisis, transformación y modelado se desarrollaron en entornos informáticos como Jupyter Notebook y Visual Studio Code, que ofrecen herramientas adecuadas para la manipulación eficiente de grandes cantidades de datos y la implementación de algoritmos de aprendizaje automático.

3.3.4. Diccionario de variables o de datos

Un diccionario es una herramienta que ayuda a la gestión de datos definiendo datos y sus características (atributos, dominios, relaciones y operaciones). Como resultado, se puede definir claramente el propósito, alcance y alcance, dando a los usuarios sugerencias sobre qué tipo de información se encontrará en el contenido (IDECA, 2019). En este estudio, el diccionario de datos corresponde a la base mencionada, que contiene información estandarizada relativa a los aspectos sociodemográficos, viajes, gastos y motivos de desplazamiento. Este diccionario facilita la organización conceptual y operación de variables que incluye el análisis, permite identificar la naturaleza, el tipo de medicina y los valores posibles.

Tabla 1

Diccionario de variables

Variable	Etiqueta	Nivel de medición
AREA	Área	Nominal
CIUDAD	Ciudad	Escala
ZONA	Zona	Nominal
SECTOR	Sector	Nominal
PANELM	Panel	Nominal
VIVIENDA	Vivienda	Nominal
HOGAR	Hogar	Escala
T00	Número de viaje	Escala
T06	Persona informante	Escala
T07	Mes del viaje	Nominal
T08	Viaje en feriado	Nominal
T0901	Persona que viajó P01	Escala
T0902	Persona que viajó P02	Escala
T0903	Persona que viajó P03	Escala
T0904	Persona que viajó P04	Escala
T0905	Persona que viajó P05	Escala
T0906	Persona que viajó P06	Escala
T0907	Persona que viajó P07	Escala
T0908	Persona que viajó P08	Escala
T0909	Persona que viajó P09	Escala
T0910	Persona que viajó P10	Escala
T0911	Persona que viajó P11	Escala
T0912	Persona que viajó P12	Escala
T10	Principal motivo del viaje	Nominal
T11	Destino principal	Escala
T12A1	Tipo transporte 1	Nominal
T12A2	Tipo transporte 2	Nominal
T13	Noches en viaje	Escala
T14	Tipo de alojamiento	Nominal
T15	Uso de paquetes turísticos	Nominal
T16	Razón por no usar paquetes	Nominal
T1601	Persona que pagó P01	Escala
T1602	Persona que pagó P02	Escala
T1603	Persona que pagó P03	Escala
T1604	Persona que pagó P04	Escala
T1605	Persona que pagó P05	Escala
T1606	Persona que pagó P06	Escala
T1607	Persona que pagó P07	Escala
T1608	Persona que pagó P08	Escala
T1609	Persona que pagó P09	Escala
T1610	Persona que pagó P10	Escala
T1611	Persona que pagó P11	Escala

T1612	Persona que pagó P12	Escala
T1701	Gasto en alojamiento	Escala
T1702	Gasto en alimentos	Escala
T1703	Gasto en transporte	Escala
T1704	Gasto en ropa, cámaras	Escala
T1705	Gasto en museos, zoológicos	Escala
T1706	Gasto en bares, balnearios	Escala
T1707	Gastos no relacionados	Escala
T1708	Otros gastos	Escala
T18	Gasto en paquete turístico	Escala
T19	Gasto total	Escala
T20A1	Actividad principal 1	Nominal
T20A2	Actividad principal 2	Nominal
T21A1	Medio de información 1	Nominal
T21A2	Medio de información 2	Nominal
T22	Medio de financiamiento	Nominal
T23A1	Servicios escasos 1	Nominal
T23A2	Servicios escasos 2	Nominal
FEXP	Factor de expansión	Escala
SOBREM	Sobremuestra de Manta	Escala
FEXPMANT	Factor expansión Manta	Escala

Nota. Esta tabla muestra la descripción de cada variable para un mejor entendimiento.

3.3.5. Procedimiento para la obtención de resultados

Para identificar las tendencias turísticas, se desarrolló una secuencia estructurada de pasos metodológicos para comprender, refinar, transformar y analizar la información. Con este fin, se elaboró una tabla que resume cada paso y su función dentro del estudio. Las primeras fases corresponden al capítulo 3, mientras que las demás se desarrollan como parte del análisis de los resultados y la última fase está relacionada con la implementación del modelo.

Tabla 2

Fases para el cumplimento del objetivo

Fase	Descripción
Fase 1	Comprensión de los datos
Fase 2	Limpieza y preparación de los datos
Fase 3	Análisis exploratorio de datos (EDA)
Fase 4	Modelamiento no supervisado
Fase 5	Inferencia y despliegue del modelo

Nota. La tabla muestra las distintas fases que contiene el proyecto para la elaboración del modelo.

3.4. Fase 1: Comprensión de los datos

En la primera fase, a través de la información obtenida se procedió a realizar el diccionario de variables y los valores de las variables se pueden observar en las siguientes tablas.

Tabla 3

Valores de las variables

Variable	Código	Etiqueta
AREA	0	Rural
	1	Urbana
PANELM	0	Panel
VIVIENDA	0	U2
	1	V2
	2	W2
	3	X2
T07	0	Enero
	1	Febrero
	2	Marzo
	n	…
T08	0	Si
	1	No
T10	0	Diversión, recreación
	1	Visitar amigos y/o parientes
	2	Otros motivos
	n	…
T12A1	0	Autobús
	1	Vehículo propio
	2	Otros motivos
	n	…

T12A2	0	Ninguno
	1	Taxi
	2	Autobús
	n	…
T14	0	Vivienda de familiares o amigos
	1	Hotel, hostal y similares
	2	Vivienda propia
	n	…
T15	0	Si
	1	No
T16	0	No necesita
	1	Desconoce
	2	Son caros
	n	…
T20A1	0	Otras
	1	Visita a balnearios
	2	Visita a atractivos naturales
	n	…
T20A2	0	Ninguno
	1	Visita a balnearios
	2	Otras
	n	…
T21A1	0	Consejo / invitación de amigos o familiares
	1	Experiencia por visita anterior
	2	Otros
	n	…
T21A2	0	Ninguno
	1	Experiencia por visita anterior
	2	Otros
	n	…
T22	0	Recursos propios
	1	Otros
	2	Crédito
	n	…
T23A1	0	Ninguno
	1	Servicios higiénicos
	2	Médios de comunicación
	n	…
T23A2	0	Ninguno
	1	Servicios higiénicos
	2	Seguridad turística
	n	…

Nota. La tabla muestra las distintas categorías que contienen cada variable categórica.

3.5. Fase 2: Preparación y limpieza de los datos

Para fines de limpieza, preparación, modelado y visualización de datos. Esta fase se llevó a cabo en Python en el IDE de VSC donde se llevó a cabo la gestión de datos. Primero, se cargó la base de datos y se importaron las bibliotecas necesarias para el análisis. Es importante mencionar que estas bibliotecas deben estar preinstaladas en el sistema.

En segundo lugar, a partir de un análisis exhaustivo de bases de datos y revisión de la literatura especializada, se realizó una preselección de las variables más relevantes de las 63 que componen el conjunto inicial del SIB-Turismo. Esta selección se basó en criterios de elegibilidad, aplicabilidad analítica y coherencia con el objetivo de la investigación. A continuación, en esta fase se vinculó la variable DPA (división político-administrativa) a una base de referencia oficial para incluir el nombre correspondiente de cada ciudad o cantón. Este procedimiento permitió enriquecer el conjunto de datos y mejorar la interpretación de los resultados del análisis territorial.

Como parte del proceso de limpieza del conjunto de datos, fue necesario implementar una serie de procedimientos destinados a estandarizar variables tanto categóricas como numéricas. Primero, se creó un mecanismo para unificar y homogeneizar registros asociados con valores faltantes o inconsistentes, reemplazando registros como "No Reportado", "nan", "NaN" o cadenas vacías que estaban representadas de manera heterogénea en la base de datos. Para las variables numéricas, estos valores se convirtieron a ceros o se verificaron si faltaban valores, lo que permitió convertirlos correctamente al formato numérico y evitar errores en pasos de análisis posteriores. En el caso de las variables categóricas, estos registros se estandarizaron con una única etiqueta común para mantener coherencia en su interpretación. Finalmente, se

excluyeron los registros con valores atípicos predefinidos, como el código 999, porque no proporcionaron información relevante para el análisis y podrían haber sesgado los resultados.

Además, para garantizar la calidad de la información utilizada en la modelación, se implementó un proceso para detectar y eliminar desviaciones. Para ello se utilizó el algoritmo isolation forest. “Se basa en la idea simple de que las instancias de datos anómalos pueden aislarse de los datos normales mediante la partición recursiva del conjunto de datos” (Stanton, et al., 2012). Para este procedimiento se fijó un nivel de contaminación del 5%, criterio que permitió limpiar la base de datos sin afectar la representatividad de los datos ni introducir sesgos en el análisis.

Excluyendo los registros no pertinentes, se procede a seleccionar de forma unívoca las variables relevantes para el desarrollo del sistema recomendado, priorizando aquellas vinculadas con la duración del viaje, los gastos asociados, la temporalidad y las características principales de la actividad turística. Como resultado, se genera una nueva variable correspondiente al gasto total, obteniendo la suma de los diferentes componentes del gasto reportados por cada individuo.

Finalmente, en esta fase se realizó una depuración adicional de la variable relacionada con la actividad principal de la ruta, eliminando los registros etiquetados como “Otras” o “Ninguno”, asumiendo que esta información no es relevante para el proceso de recomendación ni permite establecer perfiles significativos del comportamiento turístico. Tras la aplicación de todos estos procedimientos de filtrado, estandarización y selección de variables, la base final estuvo constituida por aproximadamente 2151 registros, los cuales se utilizaron posteriormente para las etapas de análisis y modelado.

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

En este apartado se presentan los resultados preliminares obtenidos a partir de las pruebas de concepto realizadas sobre la base de datos de hogares ecuatorianos con actividad turística durante el año 2021. Estas pruebas tienen como finalidad explorar la estructura de los datos y evaluar la viabilidad de aplicar técnicas de segmentación no supervisada como paso previo a la construcción del sistema recomendador de viajes.

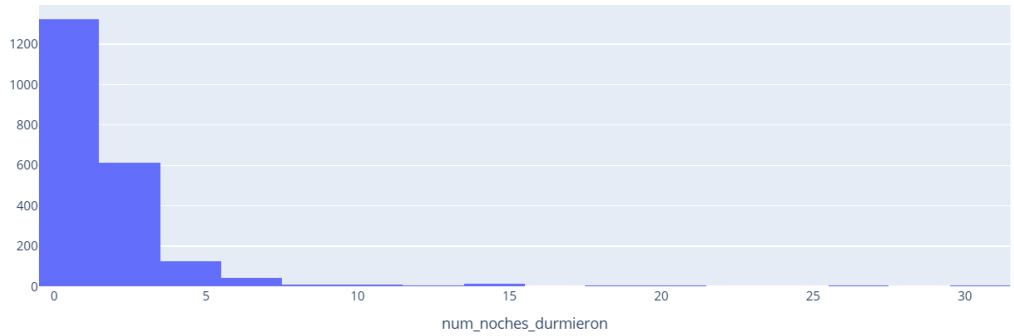
4.1. Pruebas de Concepto

4.1.1. Fase 3: Análisis exploratorio de datos

Esta sección presenta análisis univariados y bivariados del conjunto de datos de hoteles ecuatorianos sobre la actividad turística. El análisis univariado permite examinar individualmente las variables más relevantes, identificando su distribución y comportamiento general. El análisis bivariado explora las relaciones entre pares de variables intermedias en los gráficos y tablas comparativas, con el fin de detectar patrones iniciales útiles para la segmentación. Este análisis proporciona una mejor comprensión de la estructura del conjunto de datos y sirve de base para la aplicación de modelos no supervisados. Todas las exploraciones y visualizaciones se realizarán en *Python.*

Figura 1

Distribución del número de noches que durmieron

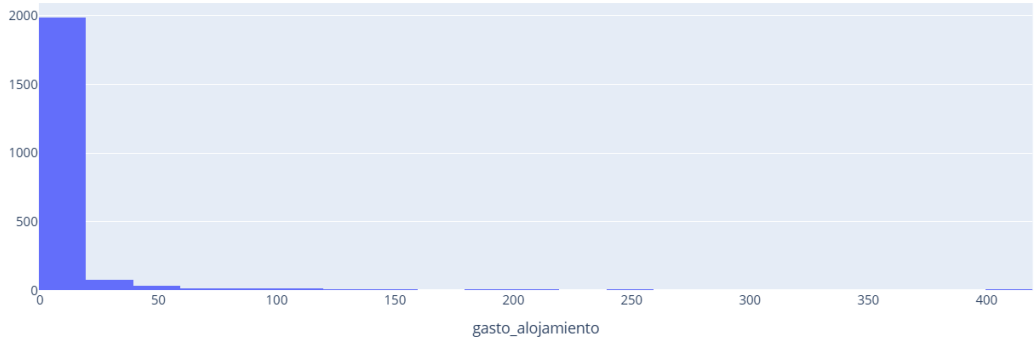


Nota. Se presenta un histograma de la distribución de la variable número de noches que durmieron. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Un histograma de frecuencias revela que la variable *'num_noches_durmieron'* presenta una distribución con una asimetría positiva significativa. Los datos se concentran mayormente entre 1 y 3 noches, lo que señala que la tendencia turística principal de las familias ecuatorianas es realizar viajes breves (en feriados o fines de semana). En cambio, los viajes de larga duración (de más de 7 noches) constituyen una parte muy pequeña de la muestra, lo cual genera una "cola larga" en la distribución. Este comportamiento influirá en la segmentación, ya que indicará la creación de grupos homogéneos que se basen sobre todo en estancias breves.

Figura 2

Distribución del gasto en alojamiento

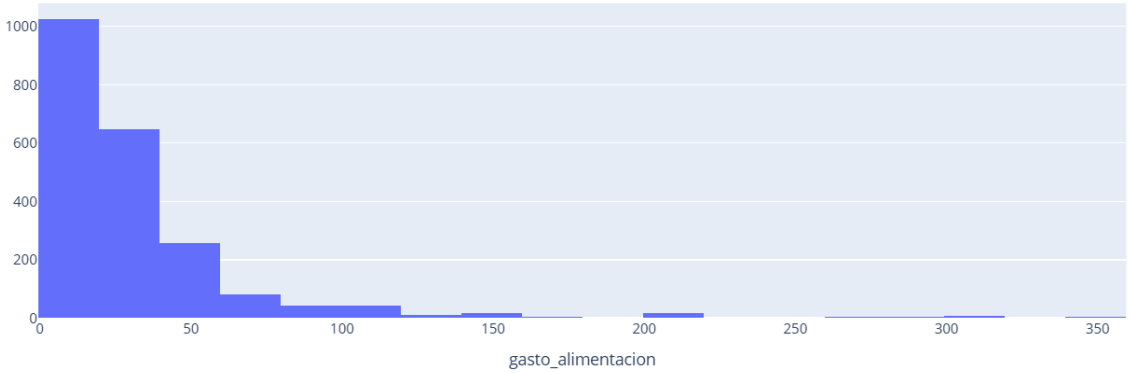


Nota. Se presenta un histograma de la distribución de la variable número de gasto en alojamiento. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

El histograma, al analizar el comportamiento de la variable *gasto_alojamiento*, muestra una distribución con una asimetría positiva muy pronunciada, en la que casi todos los registros se agrupan en el valor cero o en cantidades mínimas. Esto demuestra que una cantidad considerable de hogares en Ecuador no incurrieron en gastos de alojamiento durante sus viajes internos en 2021, lo que es coherente con un tipo de turismo fundamentado en visitar a amigos y familiares o utilizar casas propias (segunda residencia). Es considerablemente más baja la frecuencia de gastos superiores a 50 dólares, lo que genera una "cola larga" que representa a un pequeño grupo de turistas que recurren a servicios de hospedaje comercial pagado.

Figura 3

Distribución del gasto en alimentación.



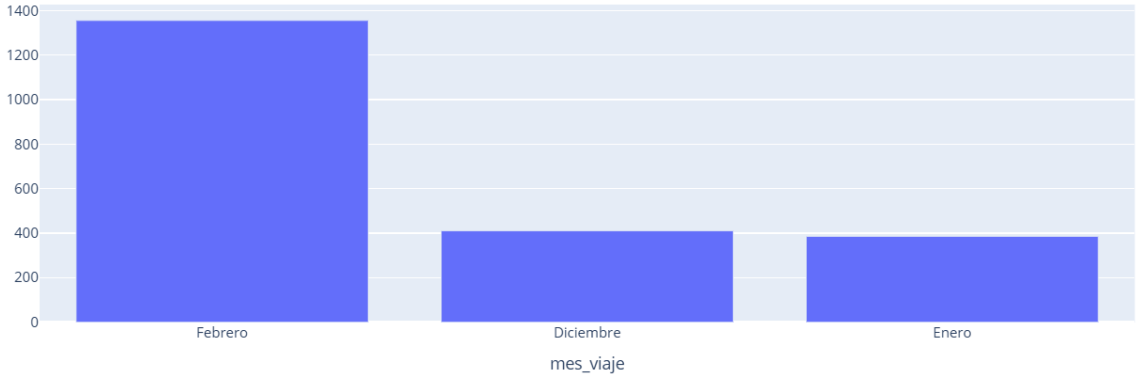
Nota. Se presenta un histograma de la distribución de la variable número de gasto en alimentación. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

La distribución de la variable *gasto_alimentación*, por su parte, muestra una notable asimetría positiva, parecida a la del alojamiento, pero con una dispersión ligeramente más amplia en los rangos iniciales. La mayoría de los hogares en Ecuador tiene un presupuesto limitado para la comida, ya que la mayor parte de los datos se encuentra entre 0 y 60 dólares. Esta conducta indica que, en sus viajes internos, los visitantes prefieren principalmente alternativas de comida a bajo costo, alimentos preparados en casa (sobre todo aquellos que se hospedan con parientes) o productos de

precio reducido. A pesar de ser tenue, la presencia de una 'cola' hacia la derecha indica que existe un segmento minoritario con mayor habilidad para consumir gastronomía.

Figura 4

Distribución de número de viajes por mes

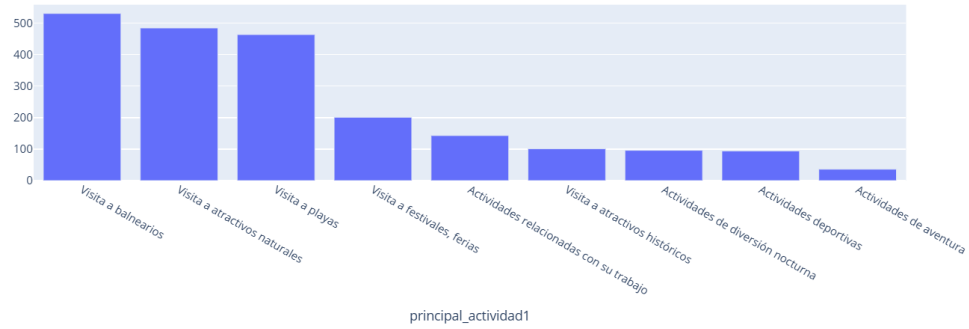


Nota. Se presenta la distribución de la variable número de viaje por mes. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

El estudio de la variable *mes_viaje* revela que la actividad turística presenta una estacionalidad significativa en términos de desplazamientos temporales. El gráfico muestra que en febrero se concentra la mayor cantidad de viajes, con más de 1.300 registros, lo cual contrasta notablemente con enero y diciembre, que tienen volúmenes de actividad mucho más bajos, cerca de los 400 viajes. Este aumento en febrero es coherente con la dinámica del turismo interno de Ecuador, que se ve fuertemente impulsado por el feriado de carnaval. En esta investigación, este último se perfila como el principal motor detrás de la movilización masiva durante las festividades de fin de año.

Figura 5

Distribución del viaje como principal actividad

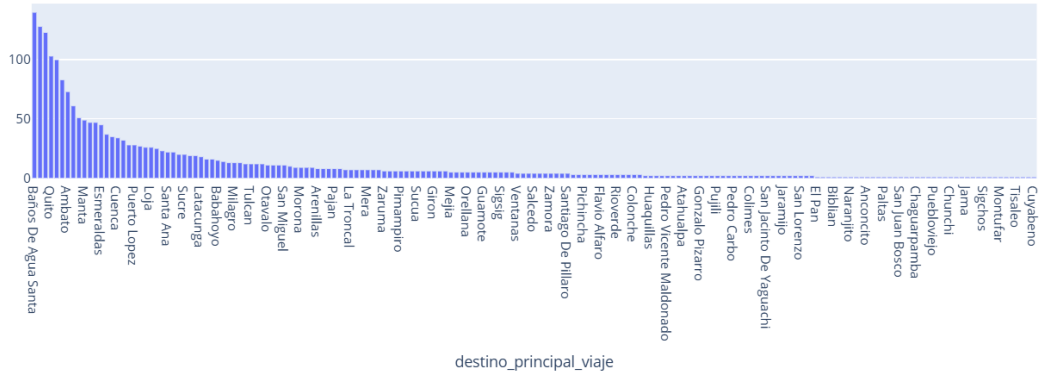


Nota. Se presenta la distribución de la variable del viaje como principal actividad. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Por último, al investigar la variable *principal_actividad1*, se encuentra una indiscutible preeminencia de las preferencias relacionadas con el turismo de naturaleza y la recreación acuática. Las tres motivaciones principales son las visitas a playas, a atractivos naturales y a balnearios. Estas concentran la mayoría de los viajes y demuestran que el turista local prefiere descansar y estar en contacto con el agua por encima de otras alternativas. Se nota una diferencia significativa en cuanto a actividades culturales (como los atractivos históricos) o de nicho (deportes y aventura), que tienen frecuencias marginales.

Figura 6

Distribución del viaje como destino principal de viaje



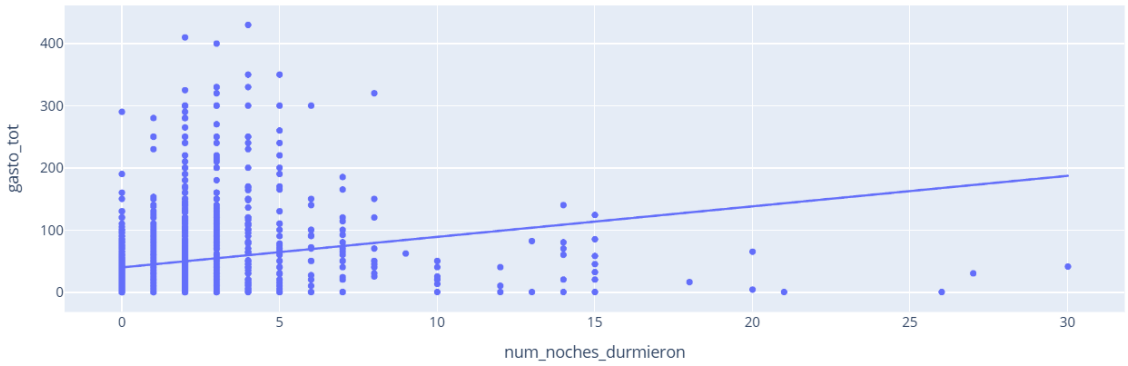
Nota. Se presenta la distribución de la variable número del viaje como destino principal. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

La variable *destino_principal_viaje* muestra una conducta 'cola larga' (*long tail*) al estudiar la distribución geográfica de los viajes, lo que indica una alta concentración de la demanda turística en un número limitado de polos de atracción. "Baños de Agua Santa", un destino que sobresale notablemente es consistente con la predilección mayoritaria por los balnearios y la naturaleza señalada previamente.

Después de estos, se encuentran importantes núcleos urbanos y lugares costeros como "Quito", "Manta", "Ambato" y "Esmeraldas". Esta diferencia señala que, aunque una pequeña cantidad de cantones se queda con la mayor parte del flujo de visitantes internos, hay una gran cantidad de destinos a los cuales acuden pocos visitantes. Para el sistema de recomendación, esta estructura de datos es fundamental: indica que un modelo basado únicamente en la popularidad tiende a recomendar siempre los mismos lugares. Por lo tanto, el algoritmo debe ser capaz de detectar patrones sutiles para sugerir adecuadamente los destinos menos visitados (los de "cola") y así diversificar la oferta turística personalizada.

Figura 7

Relación entre el gasto total vs número de noches

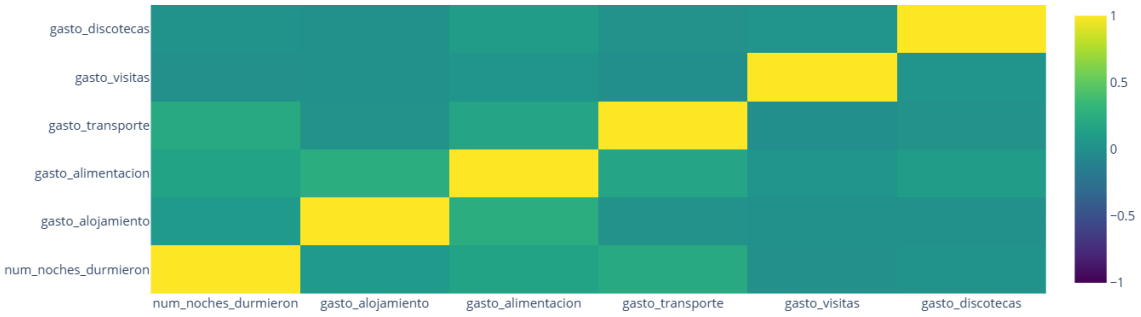


Nota. Se presenta la relación entre el número de noches vs el gasto total donde se aprecia una relación directamente proporcional. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Se nota una tendencia lineal positiva entre las variables *num_noches_durmieron* y *gasto_tot* al hacer el cruce bivariado, lo que se representa con una línea de regresión en aumento: si la duración del viaje es más larga, el gasto total tiende a ser mayor. No obstante, la dispersión de los datos muestra aspectos relevantes para describir al turista. En los viajes de corta duración (de 0 a 5 noches), se observa una gran variabilidad en el gasto, con costos que van desde cero hasta picos por encima de los 400 dólares. Esto indica que conviven dos perfiles opuestos en el mismo periodo de tiempo: turistas de alto consumo en viajes cortos y viajeros de "bajo costo" (que probablemente se hospedan con parientes). Por otra parte, se han notado viajes de larga duración (más de 15 noches) con costes totales moderados.

Figura 8

Mapa de calor de correlaciones entre variables numéricas



Nota. Se presenta la matriz de correlación donde se ve que el gasto de trasporte y el número de noches que durmieron tienen una alta relación positiva. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

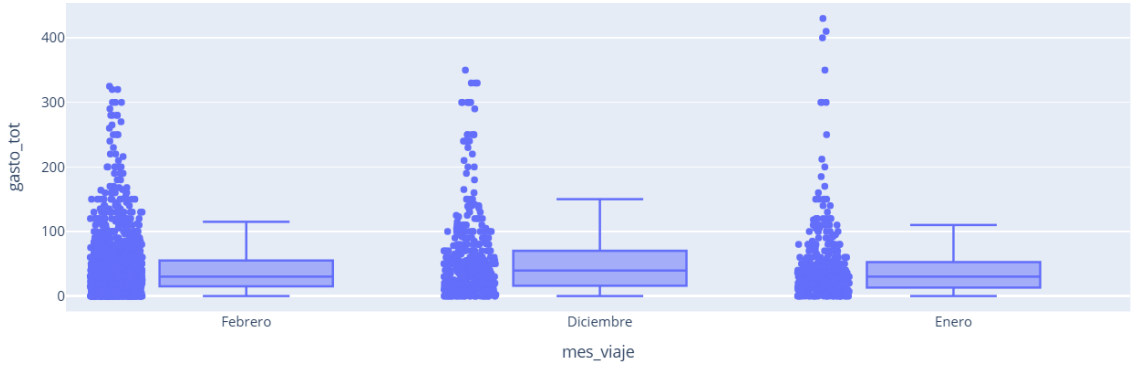
Para conocer mejor cómo están organizados los datos, se elaboró una matriz de correlación de Pearson que se mostró con un mapa de calor. Esto se hizo para identificar vínculos lineales entre las variables numéricas del estudio. El gráfico muestra una correlación positiva de moderada a alta entre la duración del viaje (num_noches_durmieron) y los gastos de subsistencia, como son el gasto alimenticio y

el gasto de alojamiento. Esto corrobora la lógica económica que sostiene que una estancia más larga conlleva costos operativos más altos.

Además, resalta la estrecha relación entre el gasto en hospedaje y el de alimentación, lo que indica una estabilidad en el nivel de consumo: los turistas que destinan más dinero a alojamiento suelen gastar más en gastronomía. En cambio, variables como el gasto en discotecas o en transporte tienen correlaciones más débiles o nulas con las demás, lo que señala que el transporte suele ser un costo fijo sin importar cuánto dure y que el gasto en entretenimiento nocturno responde a intereses particulares de ciertos perfiles y no es igual para todos los viajes. Estas relaciones desiguales refuerzan la conveniencia de usar métodos de agrupamiento (clustering) para captar estos comportamientos distintos que un sencillo análisis lineal no podría esclarecer.

Figura 9

Diagrama de cajas por gasto total vs mes de viaje



Nota. Se presenta un boxplot del mes que viajo siendo así un mayor gasto en el mes de enero. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

La relación bivariada entre *mes_viaje* y *gasto_tot* muestra aspectos significativos acerca del comportamiento estacional de la economía. Aunque febrero tiene la densidad de puntos más alta, lo que reafirma su preeminencia en términos de volumen de viajes, la distribución del gasto es desigual: hay una gran concentración de viajeros con

presupuestos bajos o inexistentes, pero también un número importante de valores atípicos (outliers) que sobrepasan los 300 dólares. Sin embargo, diciembre presenta una caja intercuartílica un poco más alta y extensa a pesar de tener menor afluencia, lo que indica que se tiende a gastar más en comparación con enero y febrero.

Esto podría deberse a la naturaleza de las festividades de diciembre (Navidad y Fin de Año), que normalmente conllevan gastos más altos relacionados con regalos y celebraciones. En cambio, enero muestra una estructura de gasto más moderada y parecida a la base de la pirámide de febrero. Esta distinción es esencial para el sistema recomendador: en la temporada alta de febrero, la oferta debe ser muy versátil (abarcar desde opciones de bajo costo hasta las de lujo); en cambio, para diciembre, podría ser conveniente ponderar un poco más hacia arriba las sugerencias de servicios, ya que se prevé que el usuario esté más dispuesto a gastar.

Tabla 4

Estadísticas descriptivas de las variables numéricas

Tipo	num_noches_durmieron	gasto_alojamiento	gasto_alimentacion	gasto_transporte	gasto_visitas	gasto_discotecas	gasto_tot
Conteo	2151	2151	2151	2151	2151	2151	2151
Promedio	1,47	4,94	26,62	11,45	0,54	3,30	46,85
Desviación Estándar	2,41	21,55	34,44	22,10	2,73	8,11	54,04
Mínimo	0	0	0	0	0	0	0
25%	0	0	7	0	0	0	15
50%	1	0	20	2	0	0	30
75%	2	0	30	16	0	0	60
Máximo	30	400	350	360	30	70	430

Nota. La tabla muestra las estadísticas descriptivas más relevantes de las principales variables numéricas a usar.

La tabla resumen de las variables numéricas, que incluye una muestra de 2151 hogares, valida en términos cuantitativos los patrones hallados en los gráficos

anteriores. Se observa un perfil turístico de baja permanencia, con una cifra promedio de 1.47 noches y una mediana de 1 noche; esto confirma que los viajes de fin de semana son los más comunes. Desde el punto de vista económico, la hipótesis del ahorro en hospedaje se ve corroborada por la conducta del gasto: la mediana del gasto en alojamiento es 0 dólares (no llega ni al percentil 75), lo que evidencia que más de tres cuartas partes de los casos no gastan en hoteles, sino que prefieren alojarse en casas propias o de parientes.

La alimentación es el sector más destacado, con una media de 26,62 dólares, lo que la convierte en la parte primordial del presupuesto del viajero promedio. Por último, la elevada desviación estándar que se ha notado en todas las variables económicas (a menudo más alta que la media, como en el gasto total: 54,04 frente a una media de 46,85) indica una gran diversidad en los datos. La aplicación de métodos de segmentación (clustering) se justifica por esta dispersión, pues indica que hay subgrupos con comportamientos de consumo significativamente distintos que se pierden al observar únicamente los promedios globales.

4.2. Análisis de Resultados

4.2.1. Fase 4: Modelamiento no supervisado

Conforme a las especificaciones de esta investigación, esta etapa evalúa diferentes algoritmos de segmentación no supervisados para identificar las tendencias en materia de turismo en los estados ecuatoriales. En la revisión de la literatura y según las características del conjunto de datos el cual combina variables numéricas y categóricas, se consideró adecuado probar inicialmente modelos como K-Means (orientado a datos numéricos) y K-Modes (orientado a datos categóricos) y el algoritmo K-Prototypes, el

cual está diseñado específicamente para trabajar con datos heterogéneos y permite conservar tanto la estructura numérica como categórica del conjunto de datos.

4.2.1.1. Modelos no supervisados probados

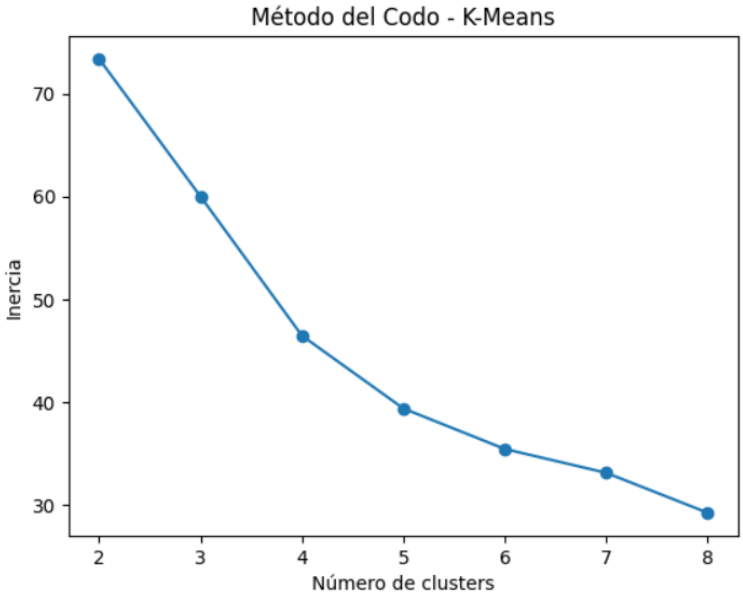
4.2.1.1.1. K- Means

En esta fase, el algoritmo K-Means como parte del proceso de prueba de concepto, con el objetivo de analizar su capacidad para agrupar a los hogares ecuatorianos según sus patrones turísticos utilizando exclusivamente variables numéricas. Con este fin, las variables relativas como los gastos del viaje y el número de noches se seleccionan y normalizan mediante la técnica MinMaxScaler para evitar la influencia de las diferencias de amplitud.

Posteriormente, se aplicó el método del codo (*Elbow Method*) para estimar el nombre óptimo de los clusters, en evaluación de la inercia del modelo para los valores de k comprende entre 2 y 8. Esto analiza una disminución progresiva del costo de clustering a medida que el nombre de los clusters aumenta, sin ninguna marca de ruptura, aunque sin una ruptura marcada debido a la limitada capacidad del algoritmo para capturar la estructura completa de los datos.

Figura 10

Método del codo – K-Means



Nota. Se presenta el método del codo donde el punto de quiebre optimo es de 4 cluster. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Finalmente, el modelo se ejecuta con k = 4 clusters como prueba exploratoria, atribuyendo un grupo numérico al proceso de registro, sin embargo, se identificó que el uso exclusivo de variables numéricas provoca pérdida de información relevante, especialmente aquella relacionada con los motivos del viaje, destinos y actividades turísticas, lo cual limita la interpretabilidad de los segmentos generados. Por esta razón, bien que K-Means es un servicio de referencia comparativo, este modelo no está seleccionado como enfoque principal para la segmentación final del estudio.

4.2.1.1.2. K- Modes

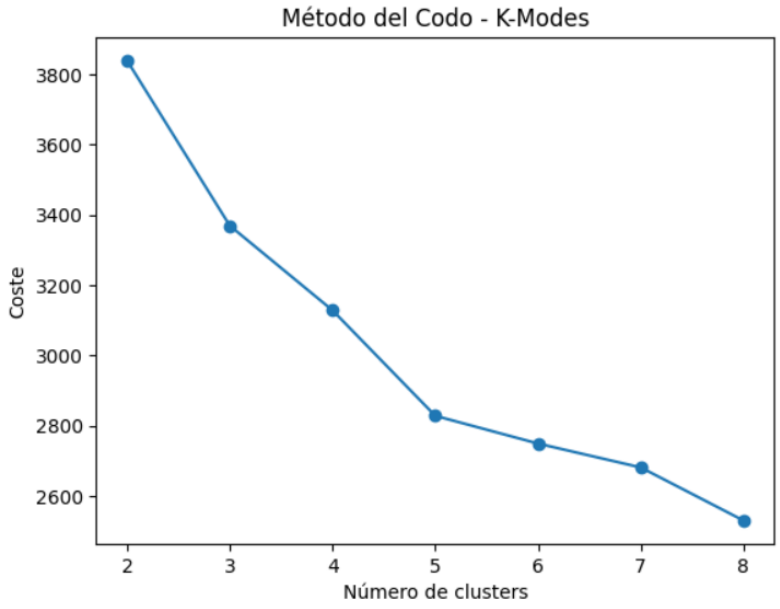
En esta etapa, el algoritmo K-Modes se evalúa en el marco del estudio de viabilidad para analizar la capacidad agrupar a los hogares ecuatorianos en función de sus patrones turísticos utilizando exclusivamente variables categóricas. Para esta prueba, las variables relativas a los meses del viaje, a la actividad principal durante su

estancia y al destino principal se obtienen en cuenta. Estas variables reflejan los comportamientos de viaje pertinentes, pero no pretenden ser aprehendidas por los modelos básicos en las distancias numéricas.

El modelo K-Modes se ejecuta con diferentes valores de k para identificar la estabilidad de las reagrupaciones y valorar la calidad de los clusters generales. Este algoritmo reagrupa los registros según una categoría correspondiente, y se analizan las variaciones de la función de cuenta para diferentes nombres de clústeres. Aunque se observaron agrupaciones coherentes en términos de categorías frecuentes, el modelo tiende a simplificar en exceso la estructura compleja del comportamiento turístico, ya que su naturaleza exclusivamente categórica limita la integración de información cuantitativa relevante como gastos, noches de estadía o niveles de consumo turístico.

Figura 11

Método del codo – K-Modes



Nota. Se presenta el método del codo donde el punto de quiebre es con 5 cluster. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Por eso la investigación está concentrada en la utilización del modelo K-Prototypes, lo que permite la combinación simultánea de variables categóricas y

numéricas, que se convierten en una segmentación más representativa y adaptada a la complejidad de los comportamientos turísticos actuales.

4.2.1.2. Modelo seleccionado

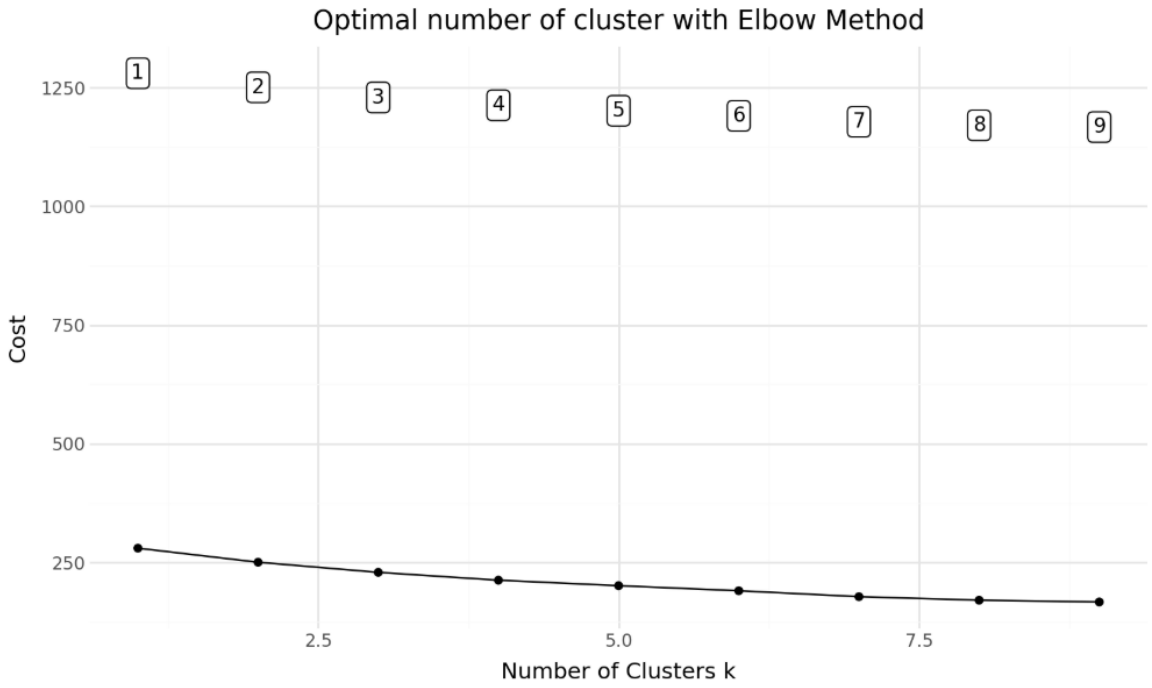
Como parte central del proceso de segmentación se implementó el algoritmo K-Prototypes, el cual se eligió como modelo final porque permite trabajar simultáneamente con variables numéricas y categóricas. Esta característica es relevante para este estudio porque los patrones de comportamiento de los turistas de los hogares ecuatorianos incluyen información heterogénea como gasto (variables numéricas), actividades realizadas, mes de viaje y destino principal (variables categóricas). El uso exclusivo de modelos como K-Means (solo numéricos) o K-Modes (solo categóricos) perdió significativamente información relevante; Por lo tanto, se identificaron los prototipos K como la alternativa metodológica más adecuada.

Una vez preparado el conjunto de datos, se estimó el número óptimo de conglomerados utilizando el método del codo. Este procedimiento se aplicó de dos maneras: (1) usando una variante del algoritmo original con la estrategia de inicialización de Huang y (2) usando la inicialización de Cao, que tiende a producir configuraciones más estables para datos mixtos.

En ambos casos se estimaron valores de k entre 2 y 9, registrándose el coste total del modelo para cada grupo. Los resultados mostraron una disminución gradual en el costo a medida que aumenta el número de conglomerados, pero con un punto de inflexión más obvio alrededor de k = 3, lo que sugiere que este valor proporciona un equilibrio adecuado entre la complejidad y la calidad del conglomerado.

Figura 12

Método del codo Huang

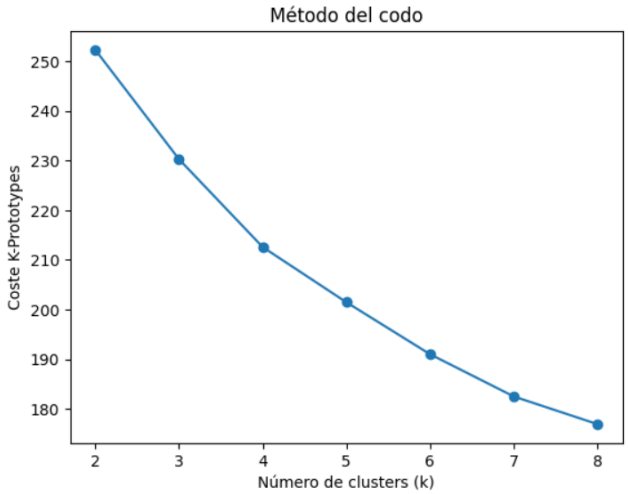


Nota. Se presenta el método de codo con la técnica Huang donde no se aprecia un quiebre optimo. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Con base en esta evidencia, se eligió k = 3 como el número óptimo de clústeres, y el modelo prototipo K final se ejecutó utilizando la estrategia de inicialización de Cao y reinicios múltiples (n_init = 5) para garantizar una partición estable. El modelo obtenido permitió dividir cada hogar en uno de los tres segmentos identificados, incluyendo automáticamente información tanto numérica como categórica sin necesidad de transformaciones que pudieran afectar la interpretación del fenómeno turístico.

Figura 13

Método del codo Cao

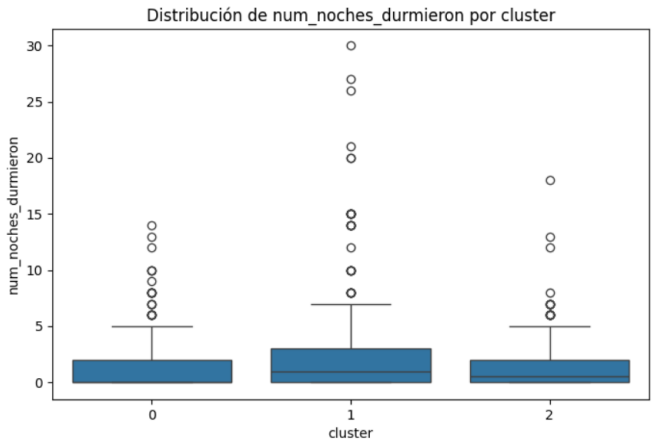


Nota. Se presenta el método del codo donde se visualiza el punto de quiebre es en 4 cluster. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Esta tarea se incorporó posteriormente al conjunto de datos original, lo que permitió un análisis detallado de las características de cada grupo. Esto incluyó la identificación de las actividades turísticas dominantes, los destinos más visitados, el consumo promedio por categoría, las proporciones relativas del consumo y el número promedio de pernoctaciones por viaje.

Figura 14

Distribución de números de noches que durmieron por clúster



Nota. Se presenta un boxplot donde el cluster 1 tiene una alta dispersión. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

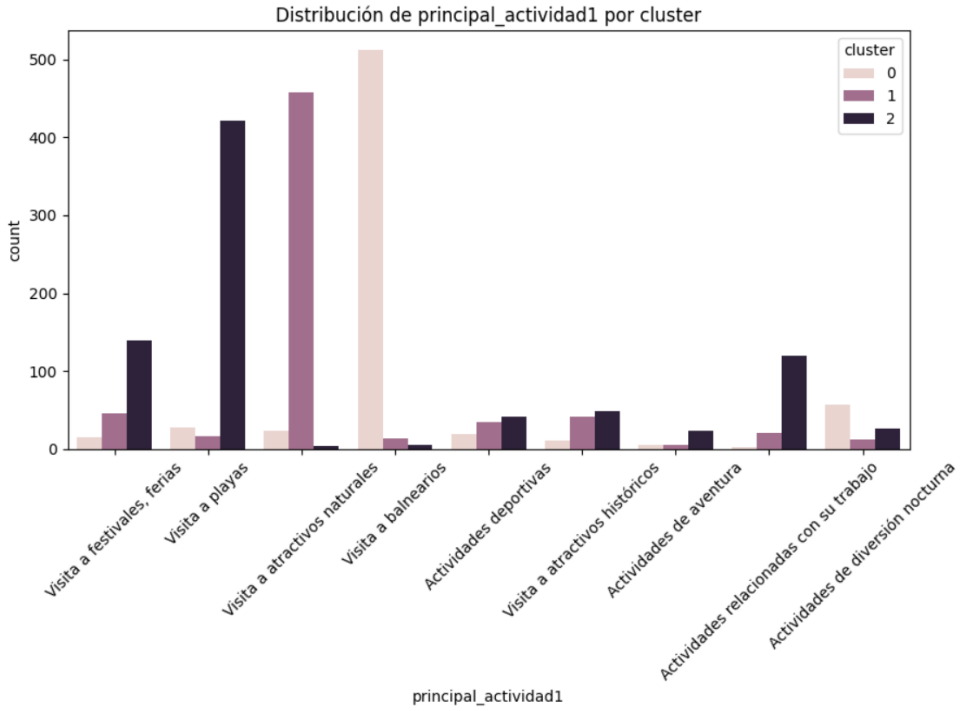
Con el propósito de verificar la coherencia de la segmentación lograda a través del algoritmo K-Prototypes (k=3), se examinó cómo estaba distribuida en cada uno de los grupos creados la variable *num_noches_durmieron*. El boxplot muestra disparidades importantes en la temporalidad de los viajes:

Grupo 1 (Estancia Prolongada y Cambiante): La duración de los viajes muestra la mayor dispersión en este grupo. Su caja intercuartílica es notablemente más ancha que la de los otros segmentos, cubriendo un rango central más alto, y muestra "bigotes" que llegan hasta las 7 noches. Asimismo, es el único clúster que logra capturar adecuadamente los outliers extremos, agrupando las excursiones que llegan a 15, 20 o incluso 30 noches.

Clúster 2 y clúster 0 (fines de semana/estancia corta): En cambio, los clústeres 0 y 2 muestran una distribución muy concentrada, con medianas bajas y extensiones intercuartílicas pequeñas, que se agrupan en su mayoría entre 0 y 3 noches. La semejanza en el tiempo de duración de ambos grupos señala que su distinción no se basa en la duración, sino que el modelo K-Prototypes los ha diferenciado a partir de otros aspectos del conjunto de datos, como las preferencias categóricas (tipo de destino o actividad) o los patrones de consumo (alto consumo vs. económico). Esto confirma que es útil usar un algoritmo mixto, pues posibilita la diferenciación de grupos que son iguales en términos temporales pero diferentes cualitativamente.

Figura 15

Distribución de principal actividad por clúster



Nota. Se presenta un diagrama de barras donde se aprecia que el cluster 0 se da en las visitas a atracciones naturales y el cluster 2 visita a las playas. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

Se puede ver una segmentación temática muy clara al analizar la distribución de la variable *principal_actividad1* en relación con los clústeres creados, donde cada grupo se centra en una clase específica de turismo. Esta polarización comprueba que el motivo del viaje fue la variable clave para desagregar los perfiles categóricos en el modelo K-Prototypes: Grupo 0: Este segmento, representado por las barras beige, está definido casi completamente por su inclinación hacia la "Visita a atractivos naturales". Su interés en otras categorías es marginal, lo que los define como ecoturistas o turistas cuya atención está centrada en disfrutar del paisaje y de espacios abiertos.

Grupo 1: El grupo (barras de color malva) presenta una preeminencia total en la categoría 'Visita a balnearios'. Al comparar este hallazgo con el análisis previo de duración, se revela que los viajes de "larga estancia" que se encuentran en este clúster

están vinculados sobre todo a lugares de balneario. Esto indica una tendencia hacia el turismo de salud, descanso o estancias vacacionales prolongadas en centros recreativos o termales concretos. Cluster 2: La 'Visita a playas' es la actividad principal del grupo oscuro y, sin duda alguna, se la apropia. No obstante, este clúster, a diferencia de los otros dos, demuestra una sociabilidad y versatilidad más amplias al incluir la mayor parte de los viajes motivados por 'Actividades relacionadas con' y 'Visita a ferias y festivales'.

Con la implementación del algoritmo K-Prototypes k=3, se detectaron tres grupos de familias ecuatorianas que, según sus patrones turísticos, eran claramente distintos. Cada agrupación exhibe conductas específicas en relación con las actividades llevadas a cabo, los destinos predilectos y la estructura de gastos, lo que posibilita definir perfiles turísticos que son de utilidad para la creación del sistema recomendador. Se describen a continuación los resultados y se le otorgan nombres representativos a cada grupo.

Grupo 0 - Turistas Recreativos de Corto Presupuesto (n = 672 viviendas): Las actividades de entretenimiento nocturno, el paseo por la playa y la visita a balnearios son las que más destacan en este clúster, que se caracteriza sobre todo por actividades relacionadas con el ocio informal. Los lugares más visitados, como Baños de Agua Santa, Pastaza y Gualaceo, indican que se prefiere visitar sitios de ocio y descanso que sean fáciles de llegar. En cuanto a gastos, este grupo muestra montos promedios que son relativamente bajos, sobre todo en visitas y alojamiento. La categoría con mayor proporción es la de alimentación (53%). La noción de viajes cortos y económicos, que son llevados a cabo principalmente por ocio espontáneo, se refuerza con el escaso desembolso en actividades programadas y visitas formales.

En general, este segmento muestra a turistas cuya conducta se caracteriza por preferir actividades de ocio y viajes breves.

Grupo 1 - Turistas Culturales y de Naturaleza (n = 649 viviendas): El segundo conjunto tiene un perfil diferente, enfocado en actividades como ir a ferias y festivales, visitar lugares históricos y recorrer bellezas naturales. Esto pone de manifiesto una motivación turística más variada y con un importante elemento cultural. Esta combinación de cultura, patrimonio y naturaleza se ve reforzada por los lugares más concurridos, tales como Ambato, Quito y Baños de Agua Santa.

Los niveles de gasto son superiores a los del clúster previo, en particular en lo que respecta a alimentación y transporte, lo cual sugiere un viaje más organizado y posiblemente más prolongado. El gasto de los alimentos sigue siendo el rubro más importante (62%), seguido del transporte (22%).

Grupo 2 - Turistas de Sol y Playa con Motivación Mixta " (n = 830 familias): El clúster más grande se halla integrado, en su mayoría, por turistas de playas y de festivales o ferias, así como por actividades laborales. Esto evidencia una mezcla entre turismo recreativo y viajes motivados por el trabajo. Las playas, Salinas y Guayaquil son los lugares de destino más comunes; están fuertemente vinculados con las actividades comerciales y el turismo costero.

En términos de gastos, se destaca que el transporte tiene un peso relativo más alto (33.7%), lo cual indica que los traslados son más extensos, probablemente desde áreas interiores del país hasta la costa. Los gastos en comida siguen siendo significativos, pero lo que se gasta en discotecas y visitas formales es muy poco.

Tabla 5

Asignación de nombres a los clústeres

Clúster	Nombre propuesto	Característica central
0	Turistas de Recreación Económica	Viajes económicos, recreación básica, balnearios y entretenimiento nocturno
1	Turistas Culturales y Naturales de Mayor Inversión	Interés en atractivos culturales/naturales, mayor nivel de gasto y actividades planificadas
2	Turistas de Sol y Playa con Movilidad Costera	Predominio de viajes a playas, bajo gasto en alojamiento, alta movilidad

Nota. La tabla muestra las características que se les da a cada clúster en base a sus preferencias.

4.2.2. Fase 5: Inferencia y despliegue del modelo

Luego de validar y caracterizar el modelo de agrupamiento K-Prototypes, la etapa siguiente fue la operacionalización del algoritmo. Esta etapa tuvo como propósito convertir el modelo teórico en un artefacto tecnológico que, a través de software funcional, posibilitara hacer inferencias sobre datos nuevos y brindar sugerencias personalizadas a los usuarios finales.

4.2.2.1. Serialización y encapsulamiento del modelo

Con el fin de asegurar que el modelo entrenado sea portátil, reproducible y eficiente en términos computacionales, se estableció un procedimiento de serialización empleando la librería pickle. Esto posibilitó la conservación en un archivo binario (`modelo_recomendador.pkl`) del estado preciso de todos los objetos fundamentales para la inferencia. Esta táctica evitó que fuera necesario volver a entrenar el algoritmo cada vez que se ejecutaba, lo que no solo mejora el tiempo de respuesta, sino que también garantiza que el modelo actúe de manera consistente en todos los entornos donde se implemente.

Se encapsularon dentro del archivo elementos cruciales: el modelo K-Prototypes, que ya había sido entrenado con sus centroides definitivos; el escalador MinMaxScaler, fundamental para normalizar las variables numéricas bajo la misma distribución que se usó en el proceso de entrenamiento; un diccionario de metadatos del clúster, que sintetiza la interpretación funcional de los grupos (patrones de gasto, actividades típicas y destinos usuales), lo cual agiliza la creación de recomendaciones; y por último, la estructura entera de columnas categóricas y numéricas, imprescindible para comprobar que la entrada del usuario coincida con la matriz original de entrenamiento. Esta encapsulación integral asegura la integridad, la consistencia y la capacidad de rastreo del modelo en cualquier etapa posterior del proyecto.

4.2.2.2. Desarrollo de la interfaz de usuario (Frontend)

Se creó una interfaz web para simplificar la interacción con el usuario final y posibilitar que el sistema de recomendación sea accesible sin tener conocimientos técnicos. Para ello, se empleó Streamlit (v1.50.0), un marco de trabajo concebido para convertir scripts de Python en aplicaciones completamente operativas con rapidez. El archivo principal, llamado `app.py`, utiliza un método de tres etapas para convertir entradas sencillas en predicciones relevantes. La interfaz, en primer lugar, recoge la información del usuario a través de elementos interactivos como `st.selectbox` y `st.number_input`, lo cual posibilita el registro de preferencias esenciales como el mes del viaje, la cantidad de noches y el presupuesto total.

En segundo lugar, los datos son convertidos internamente al formato matricial que se necesita: las variables numéricas se escalan utilizando el MinMaxScaler, después se reorganizan de acuerdo con la estructura original del conjunto de datos y, por último, el modelo K-Prototypes las evalúa usando `.predict()` para identificar el clúster más

representativo. La tercera etapa muestra resultados personalizados a través de gráficos, descripciones y listados que se han creado con los metadatos del clúster. En ella se pueden observar actividades predominantes, lugares sugeridos y una estimación del uso presupuestal. Este diseño posibilita una experiencia fluida, intuitiva y de muy fácil acceso, manteniendo la rigurosidad del proceso analítico.

Figura 16

Interfaz de la app en Streamlit

Nota. Se presenta el interfaz de usuario (UI) para el cliente final para hacer las respectivas recomendaciones en base al presupuesto obtenido. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

La interfaz gráfica de usuario (GUI) final del sistema se ha creado con la biblioteca Streamlit, siguiendo los principios del minimalismo y la funcionalidad en términos de usabilidad. La imagen muestra que la complejidad matemática del modelo K-Prototypes ha sido abstraída mediante un panel de control intuitivo que solo requiere tres parámetros de entrada básicos: el presupuesto total ("Disponibilidad económica"), el número de noches ("Duración de la estancia") y el mes en que se viaja ("Mes de viaje"). Esta elección de variables no es aleatoria, sino que se basa en los atributos

predictivos más importantes detectados durante la etapa de modelado. Esto posibilita que el algoritmo triangule el perfil del turista sin necesidad de utilizar cuestionarios largos, lo cual podría disuadir su uso.

En la interfaz se ilustra un caso específico, en el que se simula una consulta para un viaje de 2 noches a realizarse en diciembre y con un presupuesto de 100 dólares. Este escenario de prueba confirma que el sistema es capaz de procesar la tendencia más común identificada en el análisis exploratorio (viajes breves a lo largo del fin de semana con presupuesto limitado). Cuando se presiona el botón "Recomendar viaje", el sistema pone en marcha el pipeline de inferencia en el servidor, convirtiendo estos datos en bruto en vectores normalizados para determinar el clúster correspondiente y proporcionar las recomendaciones personalizadas. La disposición visual de los componentes valida la posibilidad técnica de crear una herramienta de consulta turística sencilla y accesible desde cualquier navegador web, destinada a facilitar.

Figura 17

Resultado de la predicción del modelo en la app en Streamlit

Nota. Se presenta un del resultado final tras ingresar el presupuesto que tiene el cliente para sus respectivas vacaciones donde su top de actividades sugeridas es la visita a las playas. Fuente. Elaboración propia basada en la base de datos del Sistema Banco de Información (SIB).

La imagen muestra la respuesta del algoritmo frente a una consulta simulada (un viaje en diciembre, dos noches, 100 dólares), lo que evidencia la consistencia interna del modelo K-Prototypes. El sistema ha determinado que el perfil del usuario se encuentra en el Clúster 2 (Turistas de Sol y Playa con Movilidad Costera) de acuerdo con los parámetros introducidos, lo cual se refleja de manera explícita en la recomendación de las actividades principales: "Visita a festivales y ferias" y "Visita a las playas". Esta correlación es estadísticamente sólida, pues el algoritmo ha conseguido combinar la variable temporal ("Diciembre") con la tendencia histórica de los hogares ecuatorianos a trasladarse hacia la costa y asistir a celebraciones durante el fin del año. Además, las recomendaciones de destinos (Guayaquil, Salinas, playas) tienen una geolocalización adecuada y se alinean con los principales polos turísticos que dominan dicho clúster

Desde el punto de vista económico, la división del "Gasto estimado por categoría" confirma que el modelo es capaz de reproducir la estructura verdadera del consumo turístico, en vez de hacer una simple repartición aritmética del presupuesto. El sistema distribuye el presupuesto total de 100 dólares que introdujo el usuario, destinando la mayor parte a los gastos de transporte ($33.76) y alimentación ($51.89), mientras que se calcula que el gasto en alojamiento es de $12.34, un monto marginal. Esta conducta no es un error, sino una conclusión algorítmica que representa con exactitud la realidad encontrada en el estudio exploratorio: el turista promedio de este segmento tiene la tendencia de ahorrar significativamente en alojamiento (probablemente hospedándose en sitios familiares o baratos) para dedicar la mayor parte de su dinero a la gastronomía.

4.2.2.3. Gestión de dependencias y entorno

Se estableció una estrategia de gestión de dependencias que se basa en la creación de un entorno virtual aislado para asegurar la integridad del proyecto, prevenir conflictos entre versiones y permitir que el sistema funcione de manera constante en cualquier ambiente de producción o desarrollo. En este contexto, se instalaron solamente las librerías necesarias a través de un archivo `requirements.txt.` Esto garantiza que la aplicación pueda ser replicada sin alteraciones debidas a incompatibilidades entre paquetes o actualizaciones no previstas.

4.2.2.4. Control de versiones y despliegue en la nube

Se aplicó una estrategia de control de versiones y despliegue que se fundamenta en prácticas de Cloud Computing para asegurar la trazabilidad, la mantenibilidad y el acceso público del sistema recomendador. En primer lugar, Git sirvió para guardar, versionar y documentar el código fuente, el modelo serializado y los ficheros de configuración relacionados. Estos se guardaron en un repositorio público que está en GitHub.

Esto posibilita mantener un registro detallado de las modificaciones, favorece la cooperación y garantiza que el proyecto permanezca intacto. Durante la fase de implementación, el repositorio se conectó a Streamlit Cloud, que identifica automáticamente el archivo `requirements.txt`, instala las dependencias y ejecuta la aplicación en un servidor remoto. En consecuencia, se creó una URL pública que posibilita el acceso al sistema de recomendaciones desde cualquier navegador, sin tener que instalar software extra. Esto satisface la finalidad de la investigación al proporcionar un prototipo que es funcional, accesible y validable en tiempo real.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

El presente trabajo de investigación tuvo como objetivo identificar y caracterizar patrones de comportamiento turístico de los hogares ecuatorianos mediante técnicas de segmentación no supervisada, con el propósito de construir un sistema de recomendación turístico basado en datos y un análisis lo suficientemente comprensible de interpretar para un público objetivo no tan técnico en el área de la ciencia de datos. A partir del análisis realizado, se derivan las siguientes conclusiones principales:

En primer lugar, se comprobó que los modelos de agrupamiento tradicionales como K-Means y K-Modes presentan limitaciones significativas cuando se aplican de forma aislada a conjuntos de datos turísticos heterogéneos. El algoritmo K-Means, al trabajar exclusivamente con variables numéricas, permitió identificar diferencias en niveles de gasto y duración de los viajes, pero perdió información relacionada con los motivos del viaje, destinos y actividades, lo que redujo la interpretabilidad de los clústeres obtenidos. Por su parte, K-Modes logró agrupar adecuadamente categorías dominantes, como actividades y destinos, pero su incapacidad para integrar información cuantitativa impidió capturar la complejidad económica del comportamiento turístico.

En este contexto, el algoritmo K-Prototypes se consolidó como la alternativa metodológica más adecuada para la segmentación del turismo ecuatoriano, al permitir la integración simultánea de variables numéricas y categóricas sin recurrir a

transformaciones que distorsionen la interpretación de los datos muy importante a la hora de construir el modelo ya que evitamos transformaciones complejas que pudieron haber distorsionado el compartimento natural de los datos. La aplicación del método del codo evidenció que tres clústeres representan un equilibrio óptimo entre complejidad y calidad del modelo, garantizando estabilidad y coherencia en la segmentación.

El análisis descriptivo de los clústeres obtenidos confirmó la existencia de tres perfiles turísticos claramente diferenciados pese a no ser tan representativo a nivel de negocio se considera es el adecuado para resumir tanto actividades como variable predominante a la hora de crear los perfiles turísticos, gasto y movilidad de igual forma aportan a los perfiles en menor peso, pero demuestran claramente la eleccion de un perfil u otro. El primer grupo corresponde a turistas de recreación económica, caracterizados por viajes cortos, bajo nivel de gasto y preferencia por actividades recreativas informales. El segundo grupo agrupa a turistas culturales y de naturaleza, con mayor inversión económica, estancias más prolongadas y un interés marcado por actividades culturales y patrimoniales. El tercer clúster identifica a turistas de playa con alta movilidad costera, donde predominan los desplazamientos largos, el gasto en transporte y una combinación de motivos recreativos y laborales.

Asimismo, la coherencia interna del modelo fue validada mediante análisis gráficos y estadísticos, como la distribución del número de noches por clúster y la concentración temática de las actividades principales. Estos resultados demuestran que el algoritmo no solo segmenta adecuadamente en términos cuantitativos, sino que también logra diferenciar perfiles distintos, incluso cuando ciertos grupos presentan duraciones de viaje similares.

Finalmente, la implementación del modelo en un sistema de recomendación funcional, mediante su serialización y despliegue en una aplicación web desarrollada con Streamlit, evidenció la viabilidad técnica y facilidad para público en general de ejecutar el modelo transformando al mismo modelo analítico en una herramienta práctica. La capacidad del sistema para generar recomendaciones personalizadas, coherentes con los patrones identificados, confirma que la segmentación obtenida es útil no solo desde una perspectiva descriptiva, sino también aplicada, contribuyendo al diseño de soluciones inteligentes para la planificación y promoción turística dentro del país.

5.2. RECOMENDACIONES

A partir de los resultados obtenidos y de las conclusiones obtenidas del estudio, se plantean las siguientes recomendaciones orientadas tanto al ámbito académico como al práctico y tecnológico:

Desde el punto de vista metodológico, se recomienda que futuras investigaciones en el ámbito del turismo doméstico prioricen el uso de algoritmos diseñados para datos, como K-Prototypes, especialmente cuando los conjuntos de datos integren simultáneamente variables económicas, temporales y categóricas. El uso de modelos exclusivamente numéricos o categóricos debería limitarse a análisis exploratorios o comparativos, pero no como enfoques principales de segmentación.

Asimismo, se sugiere complementar el método del codo con métricas adicionales de validación interna adaptadas a datos mixtos, como índices basados en disimilitud o estabilidad de clústeres, con el fin de reforzar la robustez en la selección del número óptimo de conglomerados pese a esto en la práctica turística se

recomendaría ampliar el número de clústeres fuera de lo técnico a fin de tener más opciones turísticas para el público en general. Esto permitiría reducir la subjetividad inherente a la inspección visual del punto de inflexión.

En términos de datos, se recomienda ampliar el conjunto de variables incorporando factores socioeconómicos adicionales, como nivel de ingresos, composición del hogar o región de origen y edad. Esta ampliación permitiría enriquecer los perfiles turísticos y mejorar la precisión del sistema de recomendación, adicional de proporcionar un mayor alcance al nivel de recomendación pasar ya no de especificaciones como playa o naturaleza a pasar a recomendaciones tipo ciudades especificas o lugares puntuales dentro del país.

Otra opción también significativa de consideración es la ampliación del publico nacional al internacional, dentro de los primeros pasos en la búsqueda de información se encontró que parques nacionales como Galápagos tienen un registro ingreso a nivel de persona. Por lo que asumiendo una metodología un poco diferente se podría caracterizar al público que está entrando a las islas y poder ofrecer alternativas de gasto conforme a sus niveles de ingresos, nacionalidad etc. A fin de generar expectativa internacional en otros países según una nueva segmentación.

Desde una perspectiva aplicada, los perfiles turísticos identificados pueden ser utilizados por instituciones públicas y privadas para diseñar estrategias diferenciadas de promoción turística. Por ejemplo, los turistas de recreación económica podrían ser incentivados mediante ofertas de corta duración y bajo costo, mientras que los turistas culturales y de naturaleza representan un segmento clave para el desarrollo de productos turísticos de mayor valor agregado y sostenibilidad.

En cuanto al sistema de recomendación desarrollado, se recomienda su evolución hacia una arquitectura más escalable que permita la incorporación de retroalimentación del usuario, lo que facilitaría la actualización dinámica de los clústeres y la mejora continua de las recomendaciones. Además, su integración con fuentes de datos en tiempo real, como eventos culturales o temporadas festivas, podría aumentar significativamente su utilidad práctica.

Finalmente, se sugiere que investigaciones futuras evalúen el impacto real del sistema de recomendación en el comportamiento de los usuarios mediante estudios experimentales o pilotos controlados. Esto permitiría medir su efectividad en términos de satisfacción del usuario una variable muy importante que sería crucial para ampliar el análisis y el modelo a su vez de generar seguimiento al usuario y poder enviar ofertas adicionales, ver en que puede mejorar el país tanto en bienes y servicios que se pueden ofrecer, por lo general el acceso de ciertos servicios en ciertas provincias podría ser utilizado como variable para un plan de negocio tanto público como privado incentivando al público en general y continuar con un modelo de desarrollo sostenible, incentivando así los cambios en los patrones de viaje y contribución al desarrollo turístico sostenible del país.

6. BIBLIOGRAFÍAS

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.

Aggarwal, C. C. (2016). *Recommender systems: The textbook*. Springer.

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Aggarwal, C. C., & Reddy, C. K. (2016). Data clustering: Algorithms and applications (2nd ed.). CRC Press.

Aggarwal, C. C. (2020). *Machine learning for data science*. Springer.

Abdi, A., & Rokonuzzaman, M. (2021). Household clustering using mixed-type socio-economic indicators: An application of prototype-based partitioning. *Journal of Social Data Analytics, 5*(2), 45–63.

Ahmad, A., & Khan, S. S. (2019). Survey of distance and similarity measures for mixed attribute data. *International Journal of Advanced Computer Science and Applications, 10*(1), 1–15.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2018). An extensive comparative study of cluster validity indices. *Pattern Recognition,* 61, 3–18.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition,* 46(1), 243–256.

Aucancela Chimbolema, R. (2025). Efectos del turismo comunitario en los indicadores socioeconómicos de las provincias del oriente ecuatoriano. *Polo Del Conocimiento*, *10*(11), 639–662. https://doi.org/10.23857/pc.v10i11.10663

Belgrave, D., & Brown, S. (2021). Designing user-centered interfaces for AI systems. *International Journal of Human–Computer Studies, 150*, 102611.

Berkhin, P. (2021). *A survey of clustering data mining techniques*. Wiley.

Bholowalia, P., & Kumar, A. (2019). Clustering techniques for large mixed datasets: Revisiting K-Prototypes. *International Journal of Data Science, 4*(3), 112–128.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Campello, R. J., Moulavi, D., & Sander, J. (2015). Density-based clustering based on hierarchical density estimates. *ACM Transactions on Knowledge Discovery from Data, 10*(1), 1–51.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2016). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications, 40*(1), 200–210.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(2), 224–227.

Canh, N. P., & Thanh, S. D. (2020). Domestic tourism spending and economic vulnerability. *Annals of Tourism Research*, *85*, 103063.https://doi.org/10.1016/J.ANNALS.2020.103063

Chacon, S., & Straub, B. (2014). *Pro git* (2nd ed.). Apress.

Cohen, E. (1972). A study on the motivations and experience value of Chinese working holiday maker in New Zealand. *Social Research, 39, 164-189. - References - Scientific Research Publishing*. https://www.scirp.org/reference/referencespapers?referenceid=1877254

Croes, R., & Rivera, M. A. (2016). Tourism's potential to benefit the poor. *Tourism Economics*, *23*(1), 29–48. https://doi.org/10.5367/TE.2015.0495

De Aguiar, M. (2016). *Saber Metodología*. Obtenido de técnicas e instrumentos de recolección de datos: https://sabermetodologia.wordpress.com/2016/02/15/tecnicas-e-instrumentos-de-recoleccion-de-datos/

De Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences, 324*, 126–145.

Dolničar, S. (2004a). Beyond “Commonsense Segmentation”: A Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*, *42*(3), 244–250. https://doi.org/10.1177/0047287503258830

Dolničar, S. (2004b). Beyond “Commonsense Segmentation”: A systematics of segmentation approaches in tourism. *Journal of Travel Research*, *42*(3), 244–250. https://doi.org/10.1177/0047287503258830

Falcón, J., & Herrera, R. (2005). *Análisis del dato estadístico.* caracas: Universidad Bolivariana de Venezuela. doi:https://docplayer.es/22896942-Analisis-del-dato-estadistico-guia-didactica.html

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–37. https://doi.org/10.1609/AIMAG.V17I3.1230

Fox, R. L., & Lawless, J. L. (2014). Uncovering the origins of the gender gap in political ambition. *American Political Science Review*, *108*(3), 499–519. https://doi.org/10.1017/S0003055414000227

Freeman, E., & Robson, E. (2018). *Head first html and css* (3rd ed.). O’Reilly Media.

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed.).

O’Reilly Media GitHub. (2024). *GitHub documentation*. https://docs.github.com/

Gallego, P. C. (2022). Sistema inteligente para modelos analíticos en el sector turístico. Madrid: Universidad Politécnica de Madrid.

Gerón, A. (2020). *Chapter 1 . The Machine learning landscape what is machine learning ?* 1–34. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Gobierno de Galápagos. (2023). *Informe técnico respecto a la distribución y uso de áreas protegidas*. https://galapagos.gob.ec/wp-content/uploads/2024/03/INFORME_ANUAL_VISITANTES-2023_WEB-LQ.pdf?utm_source=chatgpt.com

Gómez, Villasís, & Miranda. (2016). *El protocolo de investigación III: La* ciudad de México, México: Revista Alergia México, vol. 63, núm. 2. Obtenido de https://www.redalyc.org/pdf/4867/486755023011.pdf

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871.

Gracia, E. J. (2025). Análisis comparativo del turismo en Ecuador: Impacto y tendencias entre 2023 y 2024. Revista Científica Élite, 1-13.

Han, J., Pei, J., & Kamber, M. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Han, J., Pei, J., & Tong, H. (2023). Chapter 1 - Introduction. *Data Mining (Fourth Edition)*, 1–22. https://www.sciencedirect.com/science/article/pii/B9780128117606000114

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *Springer series in statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*.

Hernández, R. S., Fernández, C. C., & Baptista, P. M. (2010). *Metodología de la investigación.* México D.F.: Quinta edición- Mc Graw Hill. Obtenido de https://www.icmujeres.gob.mx/wp-content/uploads/2020/05/Sampieri.Met.Inv.pdf

Hossain, M., Rahman, M., & Akter, S. (2020). Socio-economic household profiling through mixed-type clustering techniques. *Journal of Statistical Research, 54*(1), 89–105.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large datasets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304.IBM. (2019). *¿Qué es el análisis exploratorio de datos? | IBM*. https://www.ibm.com/es-es/think/topics/exploratory-data-analysis?utm_source=chatgpt.com

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender systems: An introduction*. Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to statistical learning with applications in R second edition*.

Jia, Z. (2020). Algoritmo de agrupamiento de k-prototipos ponderados basado en el coeficiente de disimilitud híbrido. Mathematical Problems in Engineering.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley. (Obra clásica fundamental para PAM).

Keith, J. (2020). *Responsive web design with html5 and CSS3* (3rd ed.). Packt Publishing.
Loeliger, J., & McCullough, M. (2012). *Version control with git* (2nd ed.). O'Reilly Media.

Larose Daniel, & Larose Chantal. (2015). *Wiley-VCH - data mining and predictive Analytics*. editorial wiley.

León, C. A. (2025). Análisis comparativo - data turística: Estrategias para el desarrollo sostenible del turismo nacional. Ciencia Latina Revista Científica Multidisciplinar, 6837-6855.

Leon, P., Fuentes, R., & Guevara, A. (2025). The Ecuadorian Tourism Sector: Smart Tourism Initiative and the Impact in the Economy of Ecuador. *Journal of Posthumanism*, *5*(5), 4154–4169. https://doi.org/10.63332/joph.v5i5.1890

Luo, Y., Shukla, S., & Patel, J. (2018). Data versioning challenges in machine learning pipelines. *IEEE Data Engineering Bulletin, 41*(4), 26–38.

McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density-based clustering. *Journal of Open Source Software, 2*(11), 205.

Ministerio de Turismo del Ecuador. (2023). *Informe anual de desempeño del turismo en Ecuador*. Gobierno del Ecuador.

Mendoza, B. E. (2024). Modelo de gestión big data turístico, caso Manta, Ecuador. Revista Internacional De Gestión Innovación Y Sostenibilidad Turística, 31-43.

Meyer, E. (2018). *CSS: The Definitive Guide* (4th ed.). O'Reilly Media.

Ministerio de Turismo del Ecuador. (2022). *Censos nacionales de hogares 2010–2022: características de los hogares, condiciones de vida.* https://www.censoecuador.gob.ec/public/Boletin_Nacional.htm?utm_source=chatgpt.com

Ministerio de Turismo del Ecuador. (2023). *Informe de rendición de cuentas 2023*. Ministerio de Turismo. https://www.turismo.gob.ec/wp-content/uploads/2024/01/Informe_de_Rendicion_de_Cuentas_2023_MINTUR-Textual.pdf

Ministerio de Turismo del Ecuador. (2024). *Informe rendición de cuentas 2024 (Versión final).* https://www.turismo.gob.ec/wp-content/uploads/2025/08/Informe-RENDICION-DE-CUENTAS-2024-Version-Final.pdf

Mora, M. G., Cabrera Morales, L., & Quevedo Sacoto, S. (2017). Prototipo de un sistema basado en localización para dinamizar el turismo en el cantón Azogues. Revista Científica y Tecnológica UPSE, 127-133.

Moreno, M. (2010, January 1). *Turismo y producto turístico. Evolución, concepts, componentes y clasificación*. https://www.redalyc.org/pdf/4655/465545890011.pdf

Muñoz, Francisco. (2003). Turismo es lo que dice la gente. *El turismo explicado con claridad*. https://books.google.com/books/about/El_turismo_explicado_con_claridad.html?hl=es&id=Zbk5BJ8YKhIC

Murphy, K. (2022). *Probabilistic Machine Learning*.

Murtagh, F. (2015). Brief history of cluster analysis. *Handbook of Cluster Analysis*, 21–30. https://doi.org/10.1201/B19706

Murtagh, F., & Legendre, P. (2019). Ward's hierarchical agglomerative clustering: Pedregosa,

F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Python Software Foundation. (2024). *pickle — Python object serialization*. https://docs.python.org/

Plog, S. C. (1974). Why destination areas rise and fall in popularity. *Cornell Hotel and Restaurant Administration Quarterly*, *14*(4), 55–58. https://doi.org/10.1177/001088047401400409

Ponce, P., Aguirre-Padilla, N., Oliveira, C., Álvarez-García, J., & Río-Rama, M. de la C. del. (2020). The spatial externalities of tourism activities in poverty reduction. *Sustainability*, *12*(15), 1–17. https://ideas.repec.org/a/gam/jsusta/v12y2020i15p6138-d392127.html

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining predictions of machine learning models. *Proceedings of the 22nd ACM SIGKDD*, 1135–1144.

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender systems handbook* (2nd ed.). Springer.

Ricci, F., Rokach, L., & Shapira, B. (2021). *Recommender systems handbook*. Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65

Saxena, A., Prasad, M., Gupta, A., Bharill, N., & Patel, O. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664–681.

Schafer, J. B., Konstan, J., & Riedl, J. (2020). Recommender systems in e-commerce. *ACM Transactions on Internet Technology*, 20(4), 1–27.

Sculley, D., Holt, L., Golovin, D., Davydov, E., Phillips, T., & Vassilvitskii, S. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 1–9.

Stanton, C., Katz, G., & Song, D. (2012). *Isolation forest for anomaly detection.* California: Berkeley University of California. Obtenido de https://e3s-center.berkeley.edu/wp-content/uploads/2017/08/RET_CStanton-2015.pdf

Streamlit Inc. (2023). *Streamlit documentation*. https://docs.streamlit.io/

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Su, X., & Khoshgoftaar, T. M. (2020). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2020.

Treuille, A., & Teixeira, T. (2020). *The Streamlit Book*. Streamlit Inc.

UNWTO. (2022). *Tourism Trends and Outlook*. World Tourism Organization.

Valderrama, L. A. (2021). Segmentación de los alumnos ingresantes a una universidad pública aplicando el algoritmo K-prototype. Tierra Nuestra, 10-21.

Van Rossum, G., & Drake, F. (2009). *The Python Language Reference*. Python Software Foundation.

W3C. (2017). *HTML5 specification*. https://www.w3.org/

Wilson, G., et al. (2017). Good enough practices in scientific computing. *PLOS Computational Biology, 13*(6), e1005510.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1–305. https://doi.org/10.1561/2200000001

Wang, S., Li, J., & Liu, X. (2019). Optimizing hybrid distances for mixed-type clustering. *Expert Systems with Applications, 127*, 141–152.

Which algorithms implement Ward's criterion? *Journal of Classification, 31*(3), 274–295.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. *Data Mining: Practical Machine Learning Tools and Techniques*, 1–621. https://discover.library.unt.edu/catalog/b7731232

World Tourism Organization. (2019). *International tourism highlights. UNWTO.* https://www.e-unwto.org/doi/pdf/10.18111/9789284421152

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science, 2*(2), 165–193.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Ann. Data. Sci*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science 2015 2:2*, *2*(2), 165–193. https://doi.org/10.1007/S40745-015-0040-1

Xu, D., & Tian, Y. (2021). Clustering algorithms in data mining: Recent advances and applications. *ACM Computing Surveys, 54*(8), 1–38.

Zaharia, M., Xin, R. S., Wendell, P., & Gonzalez, J. (2018). Accelerating the machine learning lifecycle with MLflow. *Proceedings of the VLDB Endowment, 11*(12), 2702–2705.

Zhang, J., Tang, Y., & Lin, X. (2020). Practical data and model versioning for reproducible machine learning. *IEEE Access, 8*, 58578–58592.

Zhang, L., & Zheng, Y. (2022). Improved prototype-based clustering for mixed data using adaptive weighting. *Knowledge-Based Systems, 240*, 108–152.

7. APÉNDICES

Apéndice A Script del proyecto

TESIS2

Apéndice B Base de datos

base_turismo.csv

Apéndice C Descripción de las variables

descripción_turismodetalleviajes0310.xls

Apéndice D Tablas adicionales

TABLAS.xlsx

Apéndice E Repositorio github

https://github.com/Alex310599/kmedoides

Apéndice F Aplicación en producción

Nota: *Considerar que la app al estar en un ambiente cloud gratuito, este suele hibernar. Para visualizarlo toca activarlo.*

https://kmedoides-ldndj4gbv8rghpfzkeukcf.streamlit.app/