# Poyecto de Gen AI - Data Science Coding Bootcamps

Aarón Villacis Arroyave, Darwin Pacheco

17 de noviembre de 2025

## 1. Contexto del problema

Han sido contratados por la reconocida **consultora DARAVA**, especializada en estrategia de negocios y análisis de mercado. Esta firma integra servicios de web mining, analítica de datos y aplicación de algoritmos de inteligencia artificial para evaluar la reputación en línea de marcas e identificar oportunidades de inversión para empresas y corporaciones.
En esta ocasión, un destacado promotor turístico ha solicitado los servicios de la consultora con el objetivo de determinar en qué playas del país debería invertir. Su propósito es organizar una serie de eventos musicales y culturales en el año 2026, para lo cual requiere comprender con mayor profundidad cuáles son los destinos costeros más populares del Ecuador.
Entre los principales requerimientos del estudio se encuentra el análisis de la **capacidad hospitalaria** y **nivel de hospitalidad** de la oferta de alojamientos en los destinos seleccionados: **Villamil Playas, Salinas, Montañita, Puerto López, Ayampe, Manta y Atacames.**

## 2. Estrategia de recolección de datos

Para la generación de un dataset con los datos procedentes de los alojamientos disponibles en los destinos requeridos, se desarrolló un *scraper* para **Booking.com** utilizando Selenium y `BeautifulSoup`. Ambas librerías de Python permiten una extracción de datos casi homogénea, dentro de un rango de fechas específico y con condiciones de búsqueda coincidentes. El script se divide en las siguientes etapas:

1. **Aspectos básicos del alojamiento:** nombre, precio, puntuación media, número de reseñas.
2. **Información de la ubicación:** distancia promedio al centro o a la playa.
3. **Características adicionales del alojamiento:** lista de servicios destacados.
4. **Reseñas** fragmentos de retroalimentación positiva y negativa por parte de huéspedes.

La totalidad de los datos se almacenó en archivos `JSON` para cada destino, lo que sirvió de base para el procesamiento y el análisis posteriores.

El detalle de la codificación se encuentra disponible en los scripts utilizados para la extracción de los datos (*booking_scraper_refactor.py*) y obtener resultados del proyecto (*genai_darava_analysis.py*).

## 3. Diseño metodológico

La metodología aplicada para clasificar y jerarquizar los destinos turísticos consideró tres criterios clave: **capacidad hospitalaria *(CH)*, nivel de hospitalidad *(NH)* y relación calidad-precio *(RCP)***.

La **capacidad hospitalaria (*CH*)** es un aspecto fundamental para la oportunidad de inversión descrita en el presente proyecto. Este criterio se subcataloga de acuerdo a los siguientes parámetros:

1. **Total de alojamientos (*TA*)**: El ***TA*** fue calculado como el número total de alojamientos listados por cada destino en Booking.com.

2. **Distribución por categoría (*DC*)**: El ***DC*** fue definido como un equilibrio en la distribución de alojamientos de acuerdo a nuevas categorías conforme al precio del hospedaje (accesible, estándar, premium). Una oferta más diversificada se logra de obtenerse un valor mayor de ***DC***.

3. **Proximidad media hasta el centro (*PC*)**: El cálculo de la distancia media reportada al centro, utilizando una normalización y escalamiento simultaneos a un rango de 0–1, donde valores más altos de ***PC*** indican alojamientos, en promedio, más cercanos al punto de interés (centro).

4. **Proporción de alojamientos con servicios críticos (*SC*)**: La identificación de la presencia de tres servicios considerados críticos para el atractivo turístico: piscina, desayuno incluido y ubicación frente al mar. El ***SC*** corresponde a la proporción promedio de alojamientos en cada destino que ofrezca estos servicios.

Cada componente, para apreciarse con un grado de importancia equitativo y preservar la interpretabilidad, se escala a un rango de 0–1, empleando *min–max scaling*, la **capacidad hospitalaria** del destino se definió como una combinación ponderada:

$$CH = 0{,}4 \cdot TA + 0{,}25 \cdot DC + 0{,}25 \cdot PC + 0{,}1 \cdot SC$$

Adicionalmente, el **nivel de hospitalidad (*NH*)** es otro criterio clave para evaluar con imparcialidad la reputación en línea de los alojamientos presentes en Booking.com. Se integraron tanto métricas cuantitativas (puntuaciones numéricas) como cualitativas (comentarios/reseñas).

Los comentarios brindados por huéspedes previos pueden analizarse con el fin de detectar la bilateralidad en las perspectivas de los usuarios, como sugieren Pinar et al. [E+24], facilitando la aprehensión de lo que el viajero tiene como expectativa de su destino. El uso de algoritmos/modelos de Machine Learning simplifica el proceso de abstracción de dicha información valiosa para interpretar de forma más completa la experiencia del usuario en cada alojamiento y destino analizado, aparte de poder utilizarse en diversidad de redes sociales y páginas web dinámicas, indicado en el estudio de [AG24]. El componente emocional subyacente en cadenas de texto, como las presentes en los comentarios de los huéspedes, puede llegar a tener la mayor influencia para la decisión de una compra/reserva de un alojamiento para futuros usuarios interesados en un destino en particular, así como lo planteó Guo et al. [GWW20].

Para cada destino, se consideraron los siguientes criterios:

1. **Rating promedio del destino (*RP*)** La media de las puntuaciones asignadas al alojamiento en la plataforma.

2. **Mediana del Rating (*MR*)** La mediana de las puntuaciones, reduciendo la sensibilidad a valores atípicos o outliers.

3. **Análisis de sentimiento (*HF*)** Las reseñas textuales (comentarios positivos y negativos) se procesaron con un modelo de análisis de sentimiento multilenguaje basado en la biblioteca open-source Transformers (*Hugging Face*), que asigna a cada reseña una puntuación en una escala equivalente a 1–5 estrellas. El valor ***HF*** corresponde al promedio de dichas puntuaciones para todas las reseñas asociadas al destino.

4. **Hospitality Experience Score (*HES*)** Con el objetivo de determinar la consistencia temática de la experiencia, se generaron vectores (*embeddings*) a partir de las reseñas y se agruparon mediante técnicas de *clustering*, específicamente K-means. Para cada destino se calculó el promedio del sentimiento por grupo temático (por ejemplo, limpieza, atención del personal, ruido, comodidad), y ***HES*** se definió como el promedio de estos valores por destino.

Como en el primer criterio, todos los componentes se reescalaron en 0–1 y se combinaron de acuerdo con las siguientes ponderaciones, priorizando tanto a la puntuación numérica como al contenido textual:

$$NH = 0{,}25 \cdot HF + 0{,}2 \cdot HES + 0{,}2 \cdot MR + 0{,}35 \cdot RP$$

Por otro lado, la **relación calidad-precio (*RCP*)** es un factor adicional que permite evaluar la eficiencia económica percibida de cada destino, es decir, cuánto valor recibe el visitante en función del precio pagado. Se definieron los siguientes parámetros para evaluar cada destino:

1. **Precio promedio por categoría (*PPC*)** Se calculó el precio medio por el período de fechas en búsqueda para cada categoría (accesible, estándar y premium) y se obtuvo un promedio general ***PPC*** para el destino.

2. **Precio vs. Rating (*PVR*)** Se creó un indicador de eficiencia calidad–precio a nivel de alojamiento, relacionando rating y precio, y se promedió a nivel de destino, de manera que valores más altos de ***PVR*** indican alojamientos con una mejor evaluación relativa al costo.

3. **Dispersión de precios (*DP*)** Se estimó la desviación estándar de los precios por destino. Una mayor dispersión puede interpretarse como una mayor variedad de oferta para distintos presupuestos.

Tras aplicar *min–max scaling* en los componentes, la relación calidad–precio se definió como:

$$RCP = 0{,}4 \cdot PPC + 0{,}35 \cdot PVR + 0{,}25 \cdot DP$$

Finalmente, se definio una ecuacion para determinar *Score Final* por destino, resultado de la combinación ponderada de los tres criterios anteriores. Dado el contexto de eventos turísticos y potencial uso para actividades musicales o culturales, se priorizó ligeramente la **capacidad hospitalaria** y el **nivel de hospitalidad** sobre la **relación calidad–precio**:

$$Score\ Final = 0{,}45 \cdot CH + 0{,}35 \cdot NH + 0{,}2 \cdot RCP$$

El *Score Final* fue usado para clasificar y jerarquizar los destinos, convirtiéndose en la base cuantitativa para tomar decisiones estratégicas sobre dónde concentrar esfuerzos de inversión para eventos masivos.

# 4. Análisis comparativo de resultados

La Tabla 1 presenta los resultados obtenidos para los tres criterios de la metodología planteada (CH, NH y RCP) y el Score Final para cada uno de los siete destinos evaluados. A partir de estos resultados, se identificaron patrones diferenciados que orientaron a la toma de decisiones.

Primero, Ayampe se posiciona como el destino con el **mayor Score Final** (0,586 aproximadamente), con una muy alta capacidad hospitalaria ($CH \approx 0{,}82$) y valores intermedios en nivel de hospitalidad y relación calidad–precio. La elevada puntuación en CH indica que, a pesar de ser percibido tradicionalmente como un destino más tranquilo, el número y las características de sus alojamientos lo convierten en una alternativa robusta para mayor demanda turística con disponibilidad de servicios críticos.

Salinas y Puerto López ocupan la segunda y tercera posición en el ranking, con Scores Finales de (0,536 y 0,480, respectivamente). Para ambas opciones, su **Nivel de Hospitalidad** les caracteriza, con valores de $NH$ superiores a 0.88. Esto refleja una percepción positiva por parte de los huéspedes en la atención, la limpieza y la experiencia general del servicio, reforzada por el análisis de sentimiento de las reseñas. Además, Salinas presenta una de las mejores relaciones calidad–precio ($RCP \approx 0{,}74$), lo cual la sitúa como un destino especialmente atractivo para visitantes que buscan servicio de calidad a un costo competitivo.

Montañita y General Villamil conforman un grupo intermedio, con Scores Finales moderados. Montañita muestra una capacidad hospitalaria y una relación calidad–precio razonables, pero un nivel de hospitalidad regular, percibiéndose una experiencia más heterogénea. General Villamil, por su parte, se sitúa en la mitad de la tabla, con valores equilibrados pero sin destacar en ninguno de los tres criterios, interpretándose como un comportamiento intermedio, con margen de mejora tanto en oferta como en percepción del servicio.

En la parte baja del ranking están Manta y Atacames. En el caso de Manta, a pesar de contar con una capacidad hospitalaria alta ($CH \approx 0{,}62$), el **Nivel de Hospitalidad** es significativamente bajo ($NH \approx 0{,}04$). Esta combinación sugiere que, si bien existe infraestructura de alojamiento, la experiencia reportada por los usuarios no es favorable, representando un riesgo desde la perspectiva de reputación y satisfacción. Atacames presenta el Score Final más bajo, con valores moderados en

CH y NH pero una relación calidad–precio particularmente reducida ($RCP \approx 0{,}03$), indicando que, de acuerdo con los datos analizados, la percepción de valor recibido frente al costo es menor que en el resto de destinos.

Cuadro 1: Resumen de indicadores por destino

| Destino | CH | NH | RCP | Score Final |
|---|---|---|---|---|
| Ayampe | 0.824247 | 0.574705 | 0.417981 | **0.585706** |
| Salinas | 0.299171 | 0.889851 | 0.744639 | 0.535552 |
| Puerto López | 0.302724 | 0.903112 | 0.437990 | 0.479621 |
| Montañita | 0.504422 | 0.435706 | 0.428658 | 0.418212 |
| General Villamil | 0.345763 | 0.499955 | 0.408739 | 0.370039 |
| Manta | 0.620757 | 0.038742 | 0.283548 | 0.316635 |
| Atacames | 0.337281 | 0.315785 | 0.034199 | 0.236487 |

# 5. Conclusiones

Para una eficaz planificación de eventos turísticos, los resultados obtenidos sugieren estrategias diferenciadas. Ayampe y Salinas surgen como candidatos naturales para iniciativas que requieran buena infraestructura, alta satisfacción del usuario y equilibrio costo–beneficio. Puerto López, con su excelente nivel de hospitalidad, podría ser atractivo para eventos de menor escala o experiencias más orientadas al ecoturismo. En contraste, Manta y Atacames requerirían intervenciones focalizadas orientadas a mejorar la experiencia del huésped y revisar estrategias de fijación de precios y servicios ofrecidos, antes de considerarlos como sedes prioritarias para eventos masivos.

Finalmente, es importante reconocer algunas limitaciones del análisis: los resultados se basan exclusivamente en información disponible en Booking.com para un rango de fechas específico, y la percepción de los usuarios puede variar en función de la temporada, tipo de visitante y otros factores externos. No obstante, la metodología propuesta ofrece un marco cuantitativo coherente y reproducible para comparar destinos y apoyar decisiones estratégicas en el ámbito del turismo y la hospitalidad.

# Referencias


[AG24] Modupe Agagu and Samuel Damilare Gbadebo. A Web-based Sentiment Analyzer for Tweets Using Machine Learning Techniques. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, pages 1–7, April 2024.

[GWW20] Junpeng Guo, Xiaopan Wang, and Yi Wu. Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer Services*, 52:101891, January 2020.

[E+24] Pınar Çelik Çaylak, Mehmet Kayakuş, Nisa Eksili, Fatma Yiğit Açikgöz, Artuğ Eren Coşkun, Mirona Ana Maria Ichimov, and Georgiana Moiceanu. Analysing Online Reviews Consumers' Experiences of Mobile Travel Applications with Sentiment Analysis and Topic Modelling: The Example of Booking and Expedia. *Applied Sciences*, 14(24):11800, December 2024. Publisher: Multidisciplinary Digital Publishing Institute.