Entrega 1. Proyecto de aula Aprendizaje de Máquina Aplicado: EDA + Data Card + Baseline

[Kevin Genez / Juan Luis Amaya]

October 2, 2025

Contents

1 Introducción

Objetivo del proyecto

El objetivo principal de este proyecto es desarrollar un modelo de aprendizaje automático que, a partir del texto de las reseñas publicadas en la plataforma TripAdvisor, sea capaz de predecir la calificación otorgada por los usuarios. Para ello, se realizará un análisis exploratorio de los datos con el fin de comprender su estructura, identificar patrones en el lenguaje empleado y detectar posibles sesgos o limitaciones. Posteriormente, se construirá un modelo base que sirva como punto de referencia inicial para evaluar el desempeño de técnicas más avanzadas en etapas futuras. De esta manera, el proyecto busca sentar las bases para el diseño de sistemas de análisis de opiniones que permitan extraer información útil del lenguaje natural en contextos turísticos y de hospitalidad.

Contexto de los datos

El dataset utilizado en este proyecto corresponde a reseñas de usuarios publicadas en la plataforma TripAdvisor, una de las principales fuentes de información turística a nivel mundial. Estas reseñas incluyen tanto el texto libre escrito por los usuarios como la calificación numérica (*rating*) otorgada a hoteles, restaurantes y atracciones turísticas.

La fuente de los datos proviene de un repositorio público disponible en **Kaggle: Trip Advisor Hotel Reviews Dataset**, que recopila más de 20.000 reseñas asociadas a distintos servicios de hospitalidad. Al ser un dataset abierto, permite explorar técnicas de análisis de texto y modelado predictivo sin comprometer la privacidad de los usuarios originales.

La relevancia de este tipo de datos radica en que las reseñas en línea se han convertido en un factor decisivo en la industria del turismo y la hospitalidad. La percepción expresada por los usuarios influye directamente en la reputación de los establecimientos y en la toma de decisiones de futuros clientes. Desde la perspectiva de la ciencia de datos, el valor de este dataset se encuentra en la combinación de información estructurada (ratings) e información no estructurada (texto libre), lo que ofrece la oportunidad de aplicar técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático para generar conocimiento útil y apoyar la toma de decisiones estratégicas.

Alcance de la entrega

En esta primera entrega se desarrollarán tres componentes fundamentales que constituyen la base del proyecto:

1. **Análisis Exploratorio de Datos (EDA):** se realizará un estudio inicial de la estructura y calidad del dataset, identificando patrones, distribuciones, valores atípicos y posibles sesgos en las reseñas y calificaciones. Este análisis permitirá comprender las características principales de los datos y orientar las decisiones posteriores de modelado.

2. **Ficha de Datos (Data Card):** se elaborará un documento descriptivo que resuma los elementos esenciales del dataset, incluyendo su origen, características de las variables, limitaciones, consideraciones éticas y posibles fuentes de sesgo. Este componente busca garantizar la transparencia y trazabilidad en el uso de los datos.

3. **Modelo Base (Baseline):** se construirá un primer modelo predictivo sencillo que permita establecer una línea de referencia frente a la cual se compararán futuros modelos más complejos. Este baseline servirá como punto de partida para evaluar el potencial de aplicar técnicas avanzadas de aprendizaje automático en el análisis de las reseñas de TripAdvisor.

Con este alcance, la entrega busca proporcionar una comprensión integral del dataset y sentar las bases para el desarrollo de modelos más sofisticados en etapas posteriores del proyecto.

2 Data Card (Ficha de Datos)

2.1 Descripción general

- **Nombre del dataset:** Trip Advisor Hotel Reviews.
- **Fuente:**: Kaggle, https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews
- **Tamaño:** 20491 filas y 2 columnas.
- **Periodo de recolección:** 2016–2018.

2.2 Diccionario de variables

Variable	Tipo	Descripción
Review	Texto	Texto libre escrito por los clientes de los hoteles en la plataforma TripAdvisor
Rating	Numérica	1 a 5 estrellas

2.3 Consideraciones

- Posibles sesgos: El dataset de reseñas de TripAdvisor puede presentar diferentes tipos de sesgos que es importante tener en cuenta:
  - **Sesgo de selección:** las reseñas provienen únicamente de usuarios que decidieron compartir voluntariamente su experiencia en la plataforma, lo cual puede no ser representativo de la población total de clientes. Generalmente, los usuarios tienden a dejar comentarios cuando su experiencia fue muy positiva o muy negativa, lo que puede generar una distribución polarizada de las calificaciones.
  - **Sesgo geográfico y cultural:** dado que TripAdvisor es más popular en ciertas regiones del mundo, es posible que el dataset no refleje de manera equitativa las percepciones de usuarios de diferentes países o contextos culturales. Esto puede influir en el lenguaje empleado, la valoración de los servicios y la interpretación de las experiencias.

  - **Sesgo lingüístico:** las reseñas están escritas en un idioma específico (por ejemplo, inglés), lo que limita el alcance del análisis y puede excluir a usuarios que se expresan en otras lenguas. Además, el uso de expresiones coloquiales, abreviaciones o errores ortográficos puede afectar el desempeño de los modelos de NLP.
  - **Sesgo temporal:** las reseñas corresponden a un periodo determinado y podrían no reflejar cambios en la calidad de los servicios o en el comportamiento de los usuarios a lo largo del tiempo. Esto limita la capacidad de generalizar los resultados a contextos actuales o futuros.
  - **Sesgo de representación en categorías:** algunas categorías como hoteles, restaurantes o atracciones turísticas pueden estar sobrerrepresentadas respecto a otras, lo cual puede sesgar el análisis hacia ciertos tipos de servicios.

  Reconocer estos sesgos es clave para interpretar los resultados con cautela y considerar estrategias de mitigación, como el balanceo de clases, la inclusión de metadatos contextuales o la validación cruzada en diferentes subconjuntos del dataset.

- Limitaciones: Aunque el dataset de reseñas de TripAdvisor ofrece un recurso valioso para el análisis de opiniones en el sector turístico y de hospitalidad, también presenta ciertas limitaciones que deben ser consideradas:

  - **Representatividad limitada:** el conjunto de datos se centra en usuarios de TripAdvisor, por lo que no necesariamente refleja las percepciones de clientes que utilizan otras plataformas o que no publican reseñas en línea.
  - **Ausencia de metadatos adicionales:** el dataset solo incluye el texto de la reseña y la calificación numérica, sin información complementaria como la fecha de publicación, país de origen del usuario o características específicas del establecimiento. Esto restringe los análisis contextuales más detallados.
  - **Desbalance en la distribución de calificaciones:** en muchos casos, las calificaciones pueden estar concentradas en ciertos valores (por ejemplo, 4 o 5 estrellas), lo que dificulta la predicción equilibrada de todas las categorías de rating.
  - **Posible ruido en los datos:** el lenguaje natural empleado en las reseñas puede contener errores ortográficos, abreviaciones, expresiones coloquiales o repeticiones, lo que complica el preprocesamiento y el modelado.
  - **Limitación temporal:** las reseñas corresponden a un periodo específico que no siempre se encuentra claramente delimitado en el dataset, lo que dificulta evaluar cambios recientes en tendencias de satisfacción o en la calidad de los servicios.

  Estas limitaciones deben tenerse en cuenta al interpretar los resultados del análisis y al momento de extrapolar conclusiones hacia contextos más amplios o distintos del propio dataset.

- Consideraciones éticas: El uso del dataset de reseñas de TripAdvisor implica tener en cuenta diversos aspectos éticos relacionados con la privacidad, la equidad y el impacto de los modelos desarrollados:

- **Privacidad y anonimato:** aunque el dataset disponible en Kaggle no contiene información personal identificable, es importante recordar que las reseñas originales fueron generadas por individuos en un contexto público. El análisis debe enfocarse en los patrones lingüísticos y no en intentar inferir la identidad de los autores.
- **Sesgos en los datos:** el contenido refleja las opiniones de un grupo específico de usuarios que deciden voluntariamente publicar reseñas. Esto puede excluir a ciertos segmentos de la población y generar modelos que no representen de manera justa a todos los clientes.
- **Uso responsable de los resultados:** los modelos predictivos derivados de este proyecto tienen fines académicos y experimentales. No deben ser utilizados directamente en entornos productivos para la toma de decisiones comerciales sin antes realizar validaciones adicionales, debido a posibles errores de predicción o sesgos.
- **Respeto al contexto de aplicación:** dado que las reseñas impactan la reputación de establecimientos y la toma de decisiones de otros usuarios, se debe evitar el uso de estos análisis para manipular artificialmente percepciones en el sector turístico.
- **Limitaciones en la interpretación:** los resultados obtenidos no deben interpretarse como verdades absolutas, sino como aproximaciones estadísticas basadas en un conjunto de datos específico, con restricciones en su representatividad y alcance.

Estas consideraciones buscan garantizar que el uso del dataset y de los modelos derivados sea coherente con principios de transparencia, responsabilidad y equidad en el ámbito de la ciencia de datos.

3 Exploratory Data Analysis (EDA)

3.1 Carga y vista general

- Número de registros: 20491 filas
- Variables disponibles: Review y Rating
- Primeras observaciones: se presentan en la Figura 1.

3.2 Calidad de los datos

- Valores faltantes: Completitud: 100.0
- Consistencia: Todas las reseñas tienen rating entre 1-5
- Variabilidad: Ratings distribuidos de 1 a 5 estrellas
- Duplicados: Ninguno.
- Tipos de datos: Texto y numéricos.

Figure 1: Vista de las primeras filas del dataset de TripAdvisor.

3.3 Estadísticas descriptivas

- Variables numéricas: media, mediana, desviación estándar de la columna de Ratings. Las estadísticas descriptivas de la columna de Ratings se presentan en la Figura 2.

3.4 Visualizaciones

- Histogramas para variables numéricas: En la figura 3 se presenta el histograma de la columna Ratings del dataset de TripsAdvisor.
- Distribución porcentual de ratings: En la figura 4 se presenta el porcentaje por cada uno de los Ratings que se encuentran en el dataset de TripsAdvisor.
- Estadísticas de la longitud de los reviews:
- Heatmap de correlación.
- Gráficas resumen del EDA.

3.5 Insights relevantes

- Alrededor del 70 por ciento de los reviews corresponden a valoraciones positivas (ratings 4 y 5 estrellas).
- En términos generales, se puede esperar que los reviews más extensos correspondan a ratings bajos; cuando un cliente no está satisfecho, lo normal es que se extienda en las explicaciones de los motivos.

4 Baseline Model (Modelo Base)

4.1 Definición de la tarea

- Tipo de problema: El reto planteado corresponde a un problema de clasificación supervisada multiclase. El objetivo es predecir la calificación otorgada por un usuario (Rating) a una reseña de hotel en TripAdvisor, a partir del contenido textual de la

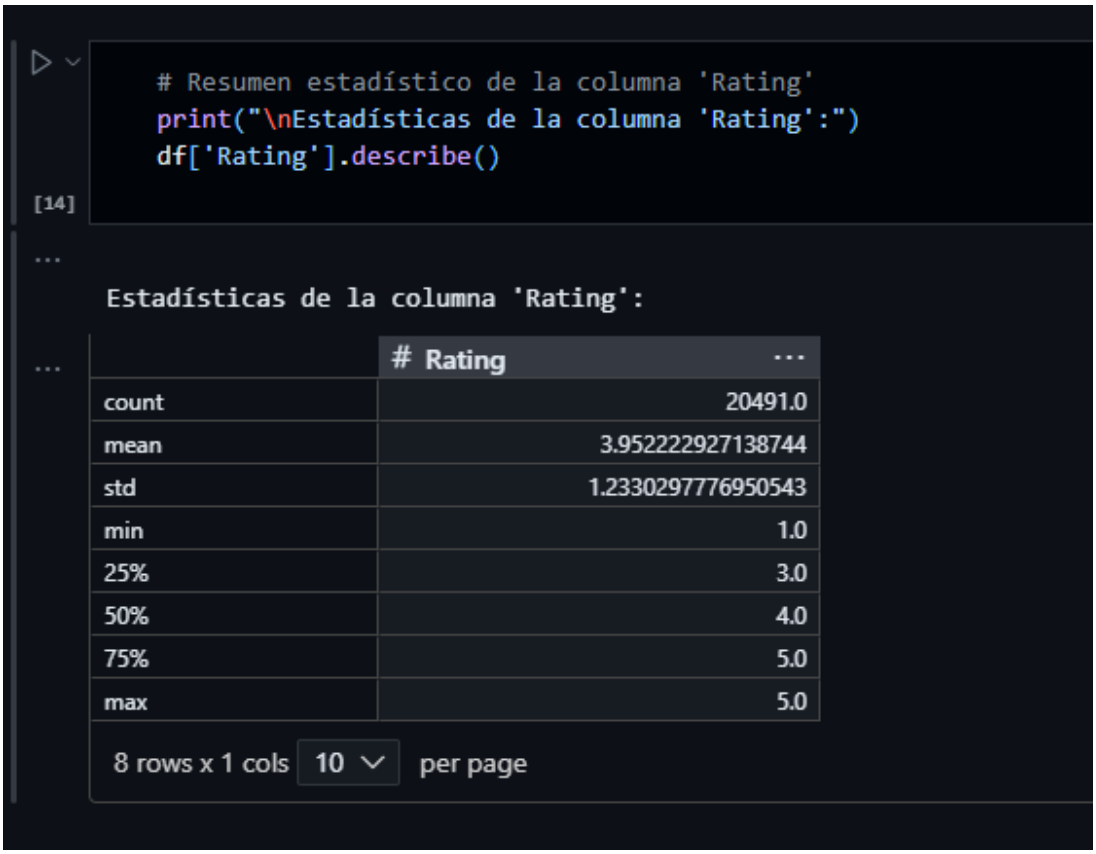

Figure 2: Vista de las estadísticas descriptivas de la columna Ratings del dataset de TripAdvisor.

misma (Review). Dado que las calificaciones se encuentran en una escala discreta (por ejemplo, de 1 a 5), el modelo debe asignar a cada texto la clase correspondiente, lo cual lo sitúa en el dominio de la clasificación de texto.

- Variable objetivo (target): La variable dependiente u objetivo (target) es el Rating, correspondiente a la valoración numérica asignada por los usuarios a los hoteles en el rango de 1 a 5 estrellas. Esta variable representa el nivel de satisfacción expresado en la reseña y constituye la categoría que se busca predecir a partir del contenido textual.

4.2 Estrategia de baseline

- Clasificación: La estrategia de baseline consiste en implementar un enfoque inicial de clasificación de texto empleando técnicas de representación vectorial simple y algoritmos clásicos de aprendizaje supervisado. En particular, se construyó un pipeline de preprocesamiento que transforma los textos mediante TF-IDF (Term Frequency-Inverse Document Frequency), seguido de la aplicación de dos modelos de referencia: Naive Bayes Multinomial, adecuado para datos textuales por su simplicidad y eficiencia, y Regresión Logística, reconocida por su robustez en problemas de clasificación multiclase. Los resultados obtenidos en términos de métricas de desempeño, como la exactitud (accuracy) y el F1-score, constituyen un punto de comparación inicial que permitirá valorar el impacto de modelos más sofisticados en

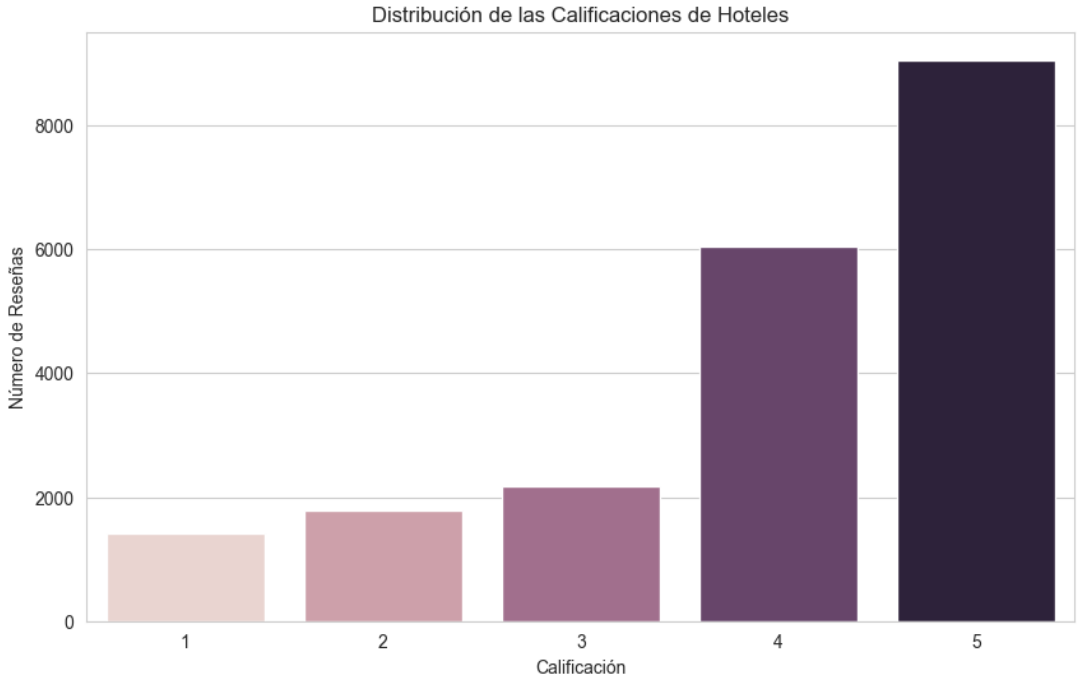

Figure 3: Histograma de la columna Ratings del dataset de TripAdvisor.

etapas posteriores del trabajo, incluyendo representaciones semánticas avanzadas y técnicas de aprendizaje profundo.

4.3 Implementación

- Definición del problema y variable objetivo: Se identifica que la tarea es un problema de clasificación supervisada multiclase, donde se busca predecir el rating (1 a 5 estrellas) a partir del texto de la reseña. Definir claramente el tipo de problema y la variable dependiente es esencial para seleccionar las técnicas de modelado y las métricas de evaluación adecuadas.

- Exploración y análisis inicial de los datos: Se revisa el número de registros, la distribución de las calificaciones y se observan ejemplos de reseñas. Esto permite detectar posibles problemas de desbalance de clases, comprender el comportamiento de los datos y orientar el preprocesamiento.

- Preprocesamiento básico de texto: Se convierten los textos a minúsculas, se eliminan caracteres especiales, números y puntuación, y se conservan únicamente las palabras relevantes. Estos pasos reducen el ruido en los datos y normalizan las reseñas para que el modelo se enfoque en la información semántica más relevante.

- Transformación de texto en representaciones numéricas (TF-IDF): Se utiliza la técnica TF-IDF (Term Frequency - Inverse Document Frequency) para convertir las reseñas en vectores numéricos que reflejan la importancia de cada palabra en el corpus. Los algoritmos de aprendizaje supervisado requieren entradas numéricas; TF-IDF es una representación eficiente y ampliamente utilizada en tareas de clasificación de texto.

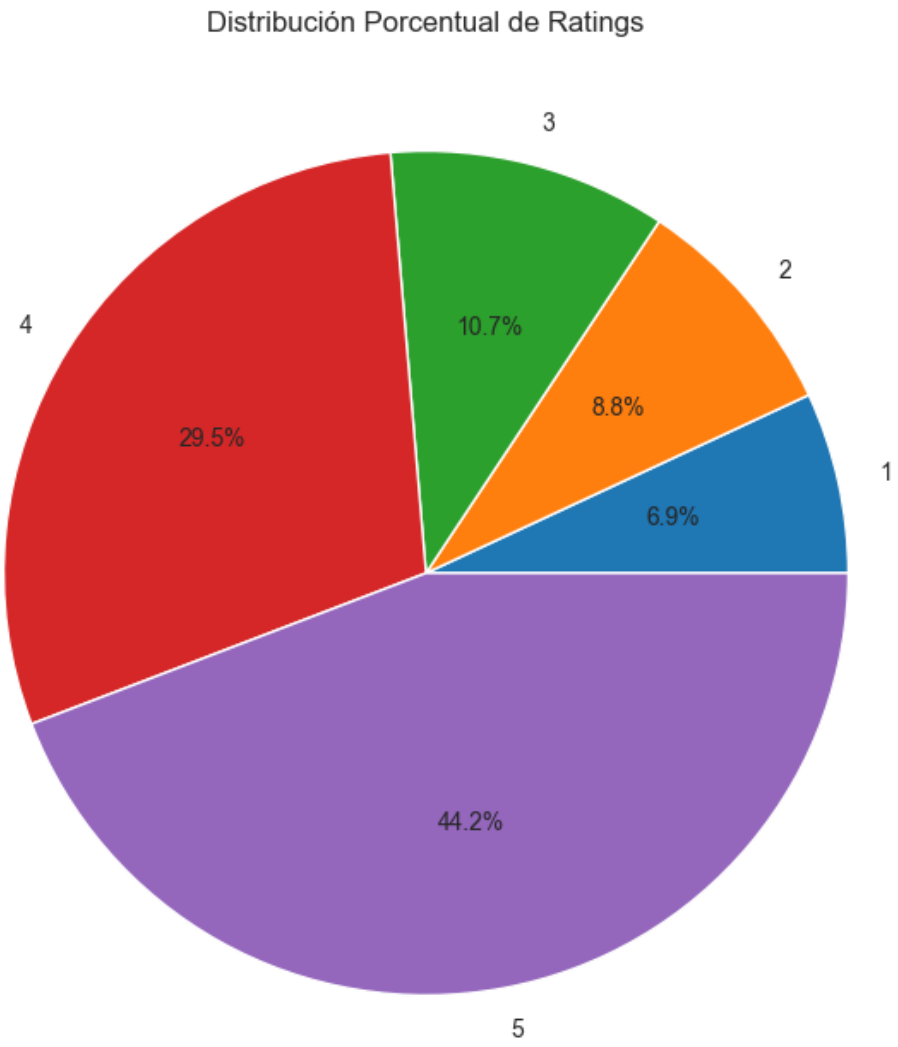

Figure 4: Distribución porcentual de los Ratings.

- División del dataset en entrenamiento y prueba.: Se separan los datos en conjuntos de entrenamiento (para ajustar el modelo) y prueba (para evaluar su desempeño). Esta práctica garantiza una evaluación imparcial del modelo, evitando el sobreajuste y asegurando que las métricas reflejen la capacidad de generalización.

- Entrenamiento de modelos baseline: Se entrenan dos modelos sencillos: Naive Bayes multinomial y regresión logística. Ambos algoritmos son eficientes y constituyen referentes clásicos en clasificación de texto. Permiten establecer un nivel de desempeño inicial contra el cual se podrán comparar modelos más sofisticados en el futuro.

- Evaluación del desempeño: Se calculan métricas como accuracy, precisión, recall y F1-score sobre el conjunto de prueba. Estas métricas permiten evaluar el rendimiento global del modelo y su capacidad para clasificar correctamente cada clase, brindando una visión integral del desempeño.

- Definición de la línea base: Se toman los resultados obtenidos como punto de comparación inicial. La línea base sirve para medir el valor agregado de futuros modelos

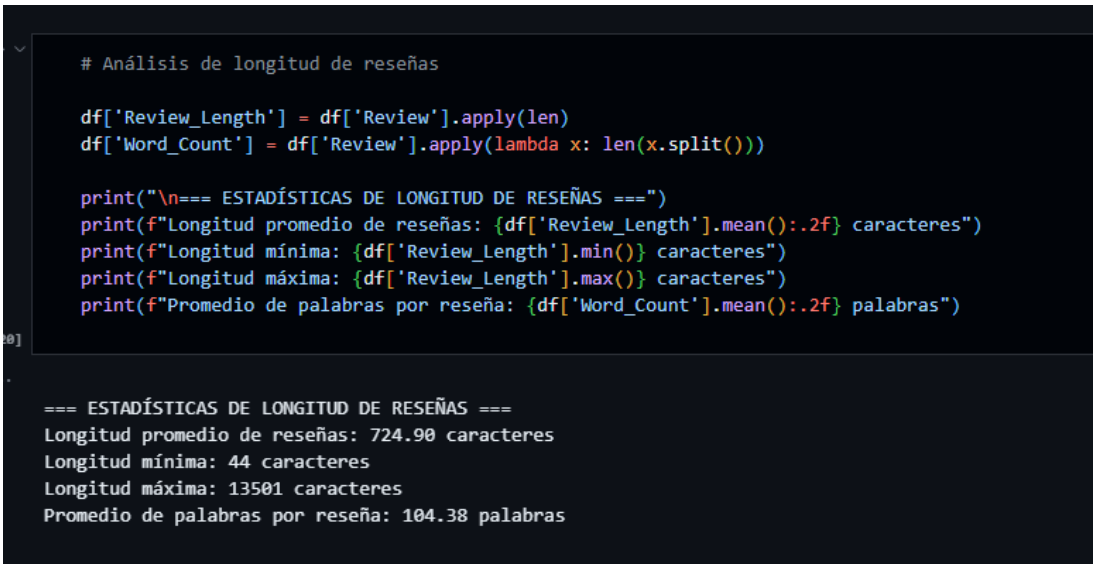

Figure 5: Estadísticas de la longitud de los reviews.

más avanzados, como embeddings semánticos o arquitecturas de deep learning, validando si efectivamente aportan una mejora sustancial.

4.4 Resultados

- El modelo 1: Nieve Bayes alcanzó un accuracy de 55%.
- El modelo 2: Regresión logística alcanzó un accuracy de 61%.
- Este resultado servirá como referencia para modelos futuros.

5 Conclusiones y próximos pasos

- Resumen de hallazgos principales del EDA: El análisis exploratorio de datos (EDA) permitió obtener una caracterización inicial del dataset de reseñas de hoteles en TripAdvisor. Se identificó que el conjunto contiene un volumen considerable de observaciones, lo cual resulta adecuado para entrenar y validar modelos de aprendizaje automático. La distribución de la variable objetivo (Rating) mostró cierta concentración en los valores más altos (4 y 5 estrellas), lo que sugiere la existencia de un desbalance de clases que podría influir en el desempeño del modelo si no se implementan estrategias de compensación. Asimismo, se observó una variabilidad significativa en la longitud de las reseñas, con textos que van desde comentarios muy breves hasta descripciones extensas, lo que refleja la heterogeneidad de los usuarios en la forma de expresar sus opiniones. Finalmente, el análisis de las palabras más frecuentes indicó la presencia de términos asociados a experiencias positivas (“excellent”, “location”, “friendly”) y negativas (“dirty”, “bad”, “poor”), lo cual confirma que la información contenida en el texto es potencialmente predictiva del puntaje asignado.
- Reflexión sobre limitaciones del dataset: Pese a su utilidad, el dataset presenta algunas limitaciones que deben ser consideradas en el proceso de modelado. En

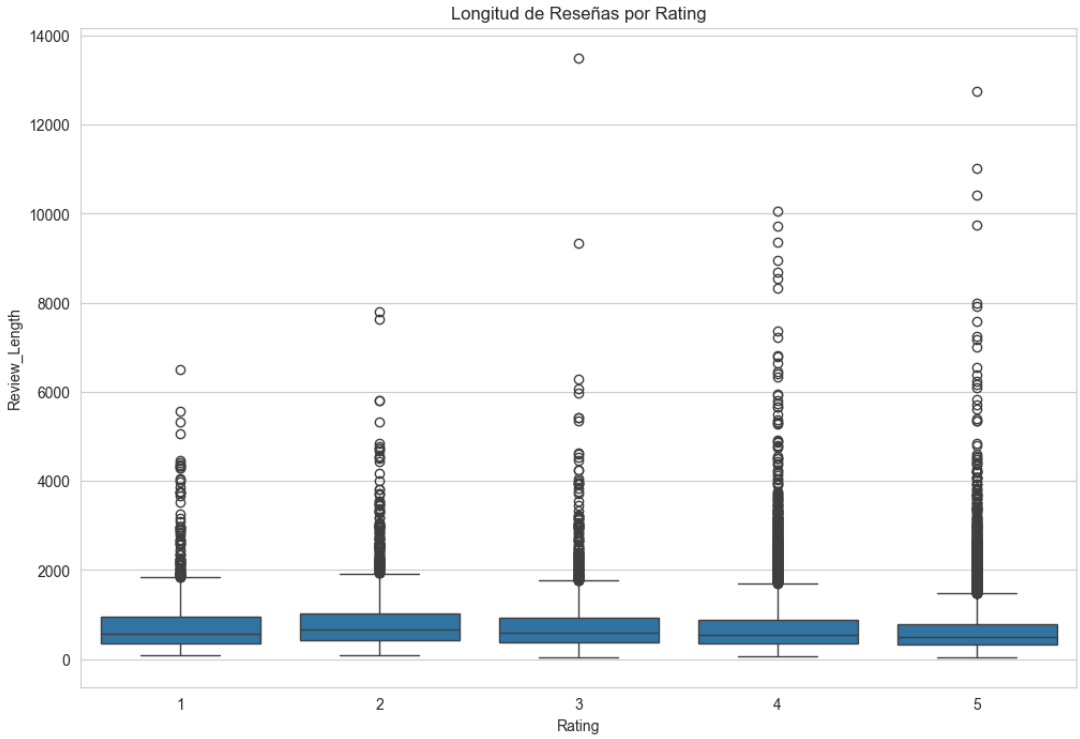

Figure 6: Longitud de los reviews.

primer lugar, la distribución desigual de las clases puede sesgar el aprendizaje hacia las categorías mayoritarias, reduciendo la capacidad de predicción en las calificaciones minoritarias (particularmente 1 y 2 estrellas). En segundo lugar, la ausencia de metadatos adicionales (como información sobre el usuario, el hotel o la fecha de la reseña) limita la posibilidad de incorporar variables contextuales que podrían enriquecer el modelo. Además, la naturaleza subjetiva de las reseñas implica la existencia de sesgos personales y lingüísticos, lo que añade complejidad a la interpretación de los resultados. Finalmente, no se dispone de una validación externa que confirme la veracidad de las opiniones, lo que introduce un posible ruido en los datos que debe ser reconocido como limitación inherente.

- Acciones futuras: Con base en los hallazgos y limitaciones identificados, se plantean las siguientes acciones para fortalecer el proceso de modelado:
  - Ingeniería de características (Feature Engineering): Incorporar variables derivadas del texto, tales como la longitud de la reseña, la polaridad y subjetividad mediante análisis de sentimiento, así como el uso de n-grams para capturar relaciones entre palabras.
  - Modelos más complejos: Evaluar el desempeño de representaciones semánticas avanzadas como Word2Vec, GloVe o embeddings contextuales derivados de modelos preentrenados como BERT. Asimismo, explorar arquitecturas de aprendizaje profundo (redes recurrentes, transformers) que han mostrado un rendimiento superior en tareas de clasificación de texto.
  - Validación más robusta: Implementar esquemas de validación cruzada estratificada para garantizar la estabilidad y generalización de los resultados, especialmente en un escenario con clases desbalanceadas.

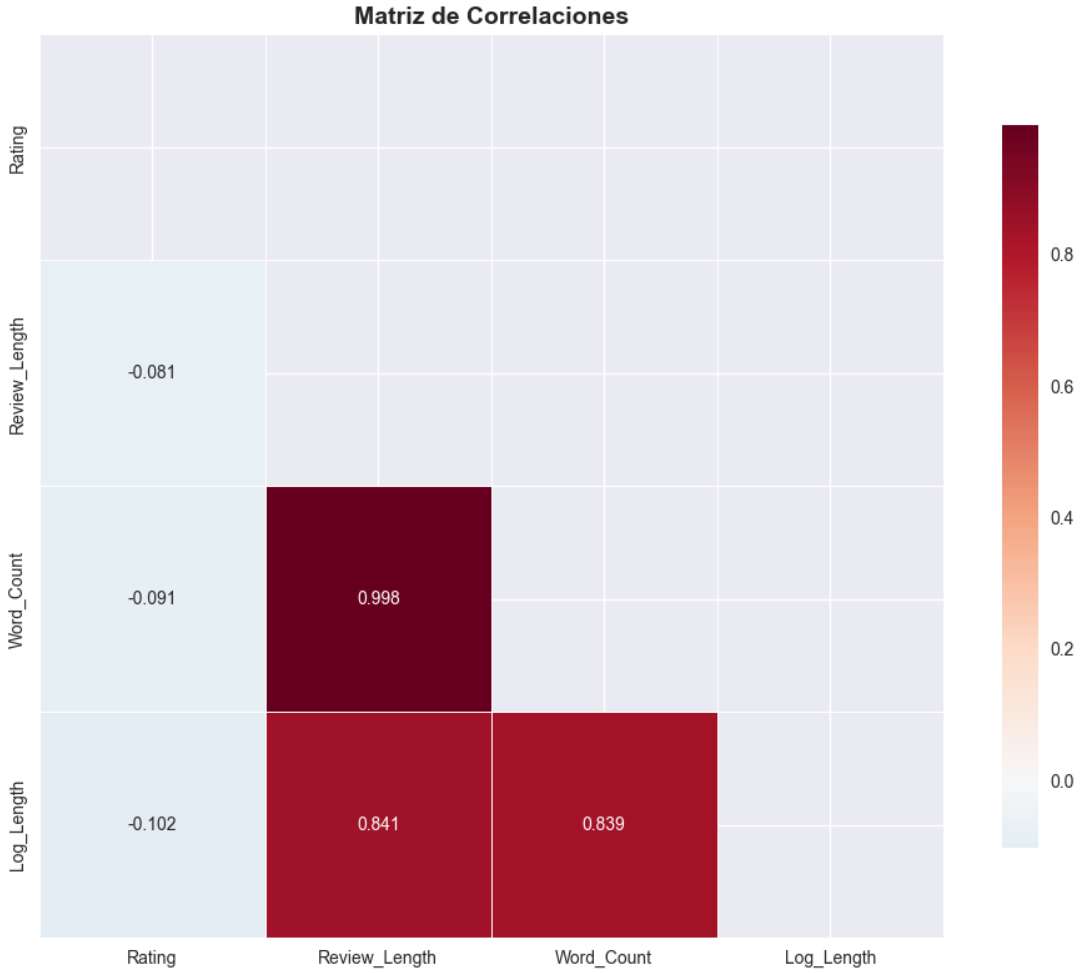

Figure 7: Matriz de Correlación.

- Manejo de desbalance de clases: Considerar estrategias como oversampling, undersampling o el uso de métricas ajustadas (F1 macro, balanced accuracy) que permitan una evaluación más justa del modelo en todas las categorías.
- Explicabilidad del modelo: Integrar técnicas de interpretación (por ejemplo, SHAP o LIME) que permitan comprender qué términos o características del texto influyen en la predicción, lo cual resulta valioso para la validación académica y práctica del modelo.

6 Enlace al repositorio de GitHub

- Enlace a repositorio de Github

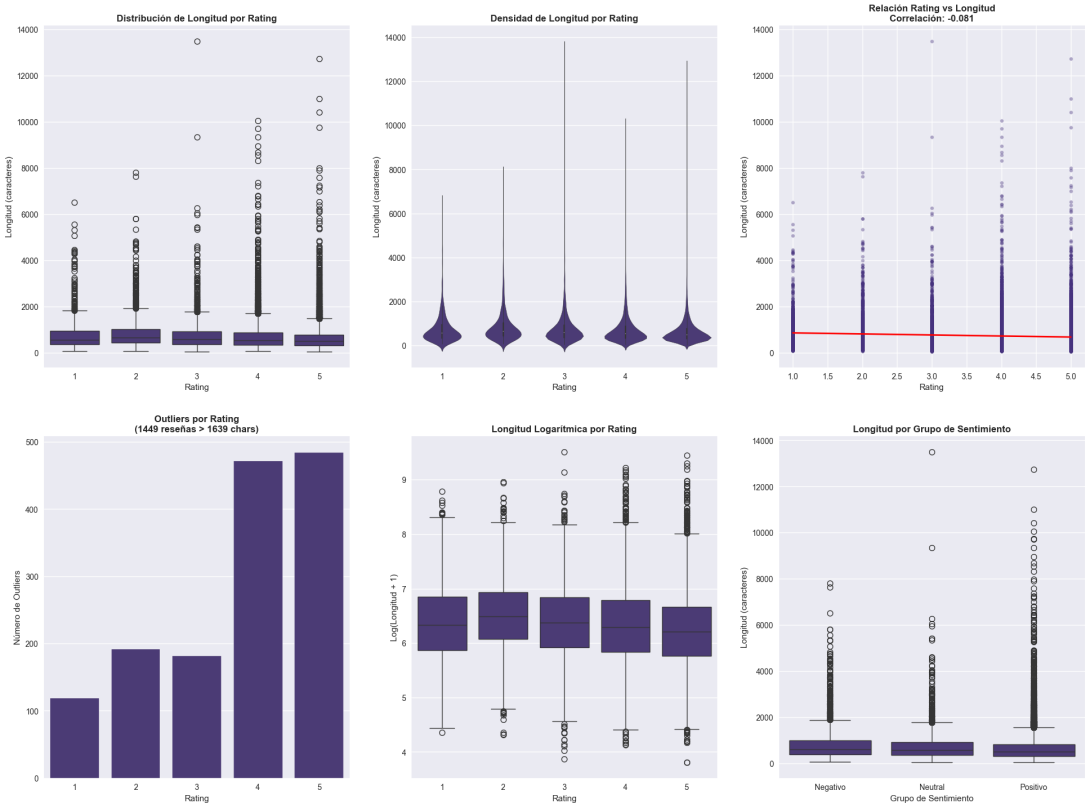

Figure 8: Gráficas resumen EDA.