1. Configuración Inicial

- **Instalación de paquetes**: Se instalan librerías esenciales (optuna para optimización, torchvision para visión artificial, folium para mapas).
- **Importación de librerías**: Se cargan bibliotecas para manipulación de datos (pandas, numpy), visualización (matplotlib, seaborn, folium), aprendizaje profundo (torch), preprocesamiento (scikit-learn) y manejo de imágenes (PIL, OpenCV).
- **Comprobación de versiones**: Se verifican las versiones de las librerías clave para garantizar reproducibilidad.
- **Fijado de semilla**: Se establece una semilla (42) en numpy, random y torch para asegurar resultados consistentes en todos los experimentos.
- **Montaje de Google Drive**: Se accede al dataset almacenado en Google Drive (poi_dataset.csv), cargando 1,569 registros con 14 columnas, incluyendo ID, nombre, descripción, categorías, ubicación geográfica, interacciones (visitas, likes, dislikes, bookmarks) y rutas de imágenes.

---

2. Análisis Exploratorio de Datos (EDA)

2.1 Descripción general del dataset

- **Primeras filas**: Muestra POIs (Puntos de Interés) con datos como categorías (ej: *"Patrimonio"*, *"Cultura"*), ubicación (coordenadas), interacciones (visitas ~10,000, likes/dislikes variables) y rutas de imágenes.
- **Información general**: Confirma que no hay valores nulos y detalla tipos de datos: 6 numéricos (tier, xps, visitas, likes, dislikes, bookmarks), 6 de texto (ej: descripciones) y 2 geográficas (latitud, longitud).
- **Dimensiones**: 1,569 filas × 14 columnas.

2.2 Valores nulos

- No se detectan valores nulos, lo que indica un dataset limpio y listo para análisis sin requerir imputación.

2.3 Análisis de variables numéricas

- **Estadísticas descriptivas**:
    - **Visitas**: Muy homogéneas (~10,012 ± 5), sugiriendo que no son discriminativas para engagement.
    - **Likes/Dislikes/Bookmarks**: Alta dispersión (likes: 3,624 ± 4,818; dislikes: 2,526 ± 2,226). Bookmarks muestran asimetría positiva (media: 973, max: 8,157).
    - **tier**: Entero (1-4), nivel del POI.
    - **xps**: Puntos asociados al POI (0-1,000).
- **Histogramas**:
    - **Likes/Dislikes/Bookmarks**: Distribución asimétrica positiva. Mayoría de POIs tienen interacciones bajas, con pocos "superstars" (valores extremos altos).
    - **tier**: Frecuencia alta en tier 1 (POIs básicos).

    - **xps**: Concentración en 500, 700 y 1,000 puntos.
- **Interpretación**: Las interacciones son desiguales; unos pocos POIs concentran alta actividad. Esto justifica una métrica de engagement que capte estas diferencias.

2.4 Análisis de variables categóricas

- **Categorías**: 224 únicas, frecuentemente anidadas (ej: *["Patrimonio", "Cultura"]*). Las más comunes:
    1. **Patrimonio**: 400+ POIs.
    2. **Cultura**: ~350 POIs.
    3. **Historia/Escultura**: ~150-200 POIs.
- **Gráfico de frecuencias**: Confirmó dominio de categorías culturales y patrimoniales, reflejando el enfoque temático del dataset.

2.5 Análisis de ubicación geográfica

- **Mapa interactivo (Folium)**:
    - **Clusters**: Concentración en Madrid (40°N, -3.7°W) y Baleares (39.5°N, 2.6°E).
    - **Dispersión global**: Pocos POIs en América y Asia, mayoría en España.
- **Interpretación**: Distribución geográfica desigual, útil para contextualizar engagement (ej: POIs en zonas turísticas podrían tener más interacciones).

2.6 Análisis de texto (shortDescription)

- **Histograma de longitud**:
    - **Rango**: 500-1,000 caracteres.
    - **Pico**: ~750 caracteres.
- **Interpretación**: Descripciones consistentes en extensión, sin outliers extremos.

2.7 Análisis de interacciones

- **Ratios**:
    - **Likes/Dislikes**: Sesgado hacia valores altos (más likes), pero con outliers inversos (POIs controvertidos).
    - **Total interacciones (L+D+B)**: Asimetría positiva; mayoría con bajas interacciones, pocos con miles.
- **Histogramas**: Refuerzan que la distribución no es normal, validando la necesidad de una métrica robusta.

2.8 Análisis de la métrica de engagement

- **Fórmula**:

*Engagement = 0.4*Likes + 0.4*Bookmarks - 0.2*Dislikes*

- **Normalización**: Escalada a [0, 1] con MinMaxScaler.
- **Visualizaciones**:

    - **Histograma**: Gran densidad cerca de 0 (bajo engagement), cola larga hacia 1 (POIs destacados).
    - **Boxplot**: Múltiples outliers altos, confirmando la existencia de POIs excepcionalmente populares.
- **Interpretación de ponderaciones**:
    - **Likes & Bookmarks (40% cada uno)**: Señales positivas fuertes (aprobación e interés futuro).
    - **Dislikes (-20%)**: Penalización leve; reconoce que incluso POIs valiosos pueden tener críticas.
- **Exclusión de visitas**: Evita *data leakage* (visitas podrían ser causa o efecto del engagement, no una señal intrínseca).

---

3. Preprocesamiento de datos

3.1 Creación de engagement

- La métrica se calcula y normaliza. Reemplaza nulos por 0.

3.2 Tratamiento de categorías

- **Codificación con** MultiLabelBinarizer: Convierte listas de categorías en matriz binaria (224 columnas). Ej: *["Cultura"]* → [0, 1, 0, ..., 0].

3.3 División de datos

- **Partición estratificada**:
    - **Entrenamiento**: 70%.
    - **Validación**: 15%.
    - **Test**: 15%.
- **Estratificación**: Basada en engagement normalizado para mantener proporciones en todos los conjuntos.

3.4 Normalización de variables numéricas

- tier **y** xps: Escalados a [0, 1] con MinMaxScaler.

3.5 Procesamiento de imágenes

- **Transformaciones**:
    - Redimensionamiento (256x256).
    - Normalización (media=0.485, std=0.229).
    - Conversión a tensores de PyTorch.
- **Dataset personalizado**: Combina imágenes, características categóricas/numéricas y engagement.

---

4. Construcción del modelo

- **Arquitectura multimodal**:
    1. **Rama imagen**: ResNet50 preentrenada (extrae características visuales).
    2. **Rama tabular**: Capas densas para procesar categorías y datos numéricos.
    3. **Fusión**: Salidas de ambas ramas se concatenan y pasan a capas densas finales para predecir engagement.
- **Función de pérdida**: MSE (Error Cuadrático Medio), idónea para regresión.

---

5. Entrenamiento y validación

- **Hiperparámetros**:
    - Optimizador: Adam.
    - *Learning rate* inicial: 0.001 (ajustado con scheduler).
- **Curvas de aprendizaje**:
    - **Pérdida (train/val)**: Decrece establemente, sin sobreajuste (brecha mínima entre curvas).
    - **Estabilización**: Pérdida de validación se estanca tras ~10 épocas, sugiriendo convergencia.

Conclusiones

- **Engagement**: La métrica ponderada refleja interacciones reales de usuarios, evitando dominancia de una sola variable y fugas de información (visitas excluidas).
- **Modelo multimodal**: Combina imágenes y datos tabulares eficazmente, logrando error bajo en engagement.