Informe académico

Trabajo Final Lenguajes

Integrantes:

Llanos Gisel,

Rios Maria Eugenia

12-12-2025

INTRODUCCIÓN

El presente trabajo aborda un análisis descriptivo del dataset TMDB 5000 Movies, con el objetivo de explorar patrones en la industria cinematográfica a partir de variables como rentabilidad, duración, presupuesto y rating.

El análisis se desarrolló íntegramente en Python mediante Jupyter Notebook, complementado con gráficos y con la construcción de una mini-API en FastAPI que expone los resultados obtenidos.

OBJETIVOS

Este trabajo busca cumplir los siguientes propósitos establecidos en el Trabajo Práctico Final:

- Realizar la lectura, limpieza y preparación del dataset.
- Aplicar estadística descriptiva y gráficos para responder preguntas de análisis.
- Desarrollar un análisis basado en tres ejes obligatorios:
    1. Rentabilidad por género y país
    2. Relación entre presupuesto y rating
    3. Evolución de la duración de películas en los últimos 50 años
- Exportar resultados resumidos en formato CSV/JSON.
- Implementar una mininAPI capaz de entregar estos resultados a cualquier usuario o aplicación.

METODOLOGÍA

El análisis se desarrolló en un entorno Python utilizando Jupyter Notebook, empleando librerías como Pandas, NumPy, Matplotlib, Seaborn y NetworkX.

La metodología aplicada se organizó en cuatro etapas principales:

Lectura y revisión del dataset

Se cargaron los archivos tmdb_5000_movies.csv y tmdb_5000_credits.csv, revisando su estructura, cantidad de registros, tipos de datos y consistencia general.

Limpieza y preparación

Las principales tareas realizadas fueron:

- Conversión de release_date a formato fecha.
- Creación de columnas derivadas: release_year y release_decade.
- Conversión de columnas tipo JSON (géneros, países, actores).
- Eliminación de películas sin valores válidos de presupuesto o revenue para el análisis de rentabilidad.
- Tratamiento de valores nulos en la duración y variables numéricas.

Desarrollo de los ejes de análisis

Se trabajó sobre los tres ejes descriptivos obligatorios:

1. Rentabilidad por género y país
2. Relación entre presupuesto y rating
3. Evolución de la duración en los últimos 50 años.

Exportación de resultados para la API

Los resultados obtenidos se resumieron y exportaron en formato JSON/CSV dentro del directorio /artifacts/:

- roi_by_genre.json
- roi_by_country.json
- runtime_evolution.json
- budget_rating_corr.json

Construcción de la API

Se implementó una miniAPI con FastAPI, organizada en servicios independientes: **roi_service.py**, **runtime_service.py** y **correlation_service.py**.

RESULTADOS Y DISCUSIÓN

A continuación, se resumen los hallazgos principales de cada uno de los ejes analizados.

➔ **Rentabilidad por género**

El ROI promedio presenta grandes variaciones entre géneros. Comedia, Drama y Horror muestran los valores más altos debido a presupuestos relativamente bajos en comparación con la recaudación. Este resultado confirma que géneros populares pueden generar alta rentabilidad sin inversiones extraordinarias.

➔ **Relación entre presupuesto y rating**

La correlación es muy baja y el gráfico de dispersión no evidencia ninguna tendencia clara. Esto indica que invertir más dinero no implica obtener mejores evaluaciones; películas costosas y económicas pueden recibir ratings similares.

➔ **Evolución del runtime en los últimos 50 años**

Respecto a la duración de las películas, el análisis de las últimas cinco décadas mostró variaciones moderadas. Las películas de los años 70 a 90 tendían a ser ligeramente más largas, mientras que en los años 2000 en adelante la duración se estabiliza en torno a valores similares, tanto en la media como en la mediana.

CONCLUSIONES

El trabajo permitió integrar diversas etapas del análisis de datos: limpieza, procesamiento, exploración estadística y visualización.

De los resultados obtenidos se concluye que:

- La rentabilidad depende más del costo de producción que del ingreso total: películas de bajo presupuesto pueden lograr altos retornos relativos.
- No existe relación significativa entre presupuesto y rating; películas costosas y económicas pueden tener evaluaciones similares.
- La duración promedio de las películas se ha mantenido relativamente estable durante las últimas décadas.

Respecto a la mini-API

Finalmente, la mini-API se integra al proyecto leyendo exclusivamente los archivos resumidos generados en el notebook. Toda la lógica de procesamiento ocurre en el análisis, y la API solo cumple el rol de exponer esos resultados como endpoints, tal como lo exige el Trabajo Final.

BIBLIOGRAFÍA Y FUENTES

- Kaggle — TMDB 5000 Movie Dataset. https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- Documentación Pandas. https://pandas.pydata.org/docs/
- Documentación NumPy. https://numpy.org/doc/
- Documentación Matplotlib. https://matplotlib.org/stable/contents.html
- Documentación Seaborn. https://seaborn.pydata.org/
- Documentación NetworkX. https://networkx.org/documentation/stable/
- Documentación FastAPI. https://fastapi.tiangolo.com/