Análisis del Catálogo de Netflix

Contexto

El objetivo de este proyecto es analizar el catálogo de contenido de Netflix utilizando técnicas de Análisis Exploratorio de Datos (EDA) y modelos de Machine Learning (ML). El propósito es comprender los patrones de crecimiento del catálogo, la predominancia de títulos de Estados Unidos y la frecuencia del género 'Drama'. El análisis está guiado por tres hipótesis principales para obtener información sobre la distribución y evolución del contenido en Netflix.

Hipótesis

1. **Netflix ha aumentado significativamente la cantidad de contenido en los últimos años.**
    - Justificación: Dada la estrategia de Netflix de expandir su catálogo, se espera un aumento notable en el número de lanzamientos en la última década.
2. **La mayoría del contenido en Netflix proviene de Estados Unidos.**
    - Justificación: Al ser una empresa con sede en EE. UU., es posible que el catálogo tenga una proporción significativa de contenido proveniente de dicho país.
3. **El género 'Drama' es el más común en Netflix.**
    - Justificación: El drama es un género popular tanto en películas como en series, por lo que se espera que sea el más frecuente en la plataforma.

Análisis Exploratorio de Datos (EDA)

Análisis Univariante

- Análisis de la distribución del contenido por año para observar tendencias de crecimiento.
- Distribución de tipos de contenido (Películas vs. Series).
- Top 10 de géneros por frecuencia.
- Países más representados en el catálogo.

Análisis Bivariante

- Comparación entre películas y series a lo largo del tiempo para observar patrones de crecimiento.
- Distribución de géneros entre películas y series.
- Relación entre el año de lanzamiento y el género.

Análisis Multivariante

- Análisis de la distribución de países dentro de los diferentes géneros.
- Análisis cruzado de tipo de contenido, género y año de lanzamiento.

- Heatmaps para visualizar correlaciones entre variables numéricas y categóricas.

Visualizaciones

- Gráficos de líneas y barras para el crecimiento del contenido.
- Gráficos circulares para la distribución de géneros.
- Diagramas de caja para analizar la duración de películas y series.
- Heatmaps para correlaciones multivariantes.

Modelos de Machine Learning

Selección del Modelo

- Clasificador Random Forest para clasificación binaria.
- Tres modelos, uno para cada hipótesis.

Resultados

1. **Hipótesis 1 (Crecimiento del Contenido)**
    - Precisión: 62%
    - El modelo predice correctamente el aumento de contenido en los últimos años.
2. **Hipótesis 2 (Predominancia de Contenido de EE. UU.)**
    - Precisión: 64%
    - El modelo tiene un mejor rendimiento al identificar contenido no estadounidense, lo que refleja diversidad en el catálogo.
3. **Hipótesis 3 (Frecuencia del Género Drama)**
    - Precisión: 69%
    - El modelo confirma que el drama es el género más frecuente, especialmente en películas.

Conclusiones y Recomendaciones

- Netflix ha aumentado significativamente su volumen de contenido, especialmente entre 2012 y 2018, lo que refleja una estrategia de expansión.
- Aunque la mayoría del contenido proviene de Estados Unidos, la presencia de títulos de otros países muestra una creciente diversidad.
- El drama sigue siendo el género más común, destacando su atractivo global y preferencia del público.
- Se recomienda que Netflix continúe diversificando el origen del contenido mientras mantiene el enfoque en géneros que resuenan con la audiencia, como el drama.

Requerimientos

- Python 3.7+
- Bibliotecas: pandas, numpy, sklearn, matplotlib, seaborn