Reporte Exploratorio del Dataset de Netflix

Equipo: Equipo 4

Integrantes:

- Luis Eduardo Diaz Valle - A00835871
- Isaac Yael Zaragoza Oldendorff - A00837859
- Luis Guzmán - A00837746
- Alan De Loa Larios - A00837006
- Leonardo Lopez Ruiz - A01254678

Unidad de formación: Inteligencia artificial avanzada para la ciencia de datos I (Gpo 102)

Profesor: Alder López Cerda

Fecha: 17 de agosto de 2025

Índice

Resumen

Este reporte presenta un análisis exploratorio del dataset de Netflix, con el propósito de comprender sus características y detectar patrones relevantes. Se documentan motivaciones, preguntas de investigación, la variable objetivo seleccionada y un diccionario de variables. También se identifican riesgos y limitaciones que impactan en el uso del dataset para futuros modelos de aprendizaje automático. Finalmente, se proponen conclusiones preliminares para guiar el modelado de la variable objetivo seleccionada.

Palabras clave:

Netflix, análisis exploratorio, machine learning, clasificación de audiencia, entretenimiento

1 Motivación

El consumo de contenido audiovisual a través de plataformas de *streaming* se ha consolidado como una de las principales formas de entretenimiento en la actualidad. Netflix, en particular, se ha posicionado como un referente mundial en la industria, ofreciendo una amplia variedad de títulos que abarcan múltiples géneros, países y formatos.

Analizar este dataset permite entender no solo la composición del catálogo, sino también las tendencias de producción, la distribución geográfica y las preferencias en clasificación de audiencia. Estos hallazgos tienen valor tanto académico como práctico, ya que pueden aplicarse en áreas como:

- **Modelos de recomendación:** Optimización de sugerencias de contenido a usuarios.
- **Estudios de mercado:** Identificación de géneros o regiones con mayor crecimiento.
- **Clasificación de audiencias:** Mejora en la predicción de a qué público va dirigido un nuevo título.

2 Preguntas de investigación

A partir del análisis del dataset, se plantean las siguientes preguntas de investigación:

1. ¿Cuáles son los géneros más comunes en películas y series y cómo varían según la región de producción?

2. ¿Qué patrones temporales (por décadas) se observan en la incorporación de títulos al catálogo de Netflix?

3. ¿Es posible predecir la clasificación de audiencia (**rating**) de un título a partir de sus características principales (tipo, género, duración, país, año de lanzamiento)?

3 Variable objetivo

La variable seleccionada como objetivo es **rating**, por las siguientes razones:

- Es categórica, relevante y tiene impacto directo en el consumo de contenido.
- Está presente en una gran parte de los registros.
- Permite desarrollar un problema de clasificación supervisada, común en machine learning.

4 Fuente

El dataset proviene de la plataforma *Kaggle*, titulado **Netflix Movies and TV Shows**.

- Registros: alrededor de 8,800.
- Variables: título, tipo, director, país, fecha de estreno, año de lanzamiento, duración, clasificación, géneros.
- Fuente: `https://www.kaggle.com/datasets/shivamb/netflix-shows`

5 Diccionario de variables

A continuación, se presenta un resumen de las principales variables utilizadas:

- **type** (categórica): Película o Serie.
- **title** (texto): Nombre del título.

- **director** (texto): Director (si está disponible).
- **country** (categórica): País de producción.
- **date_added** (fecha): Fecha de incorporación al catálogo.
- **release_year** (numérica): Año de lanzamiento.
- **rating** (categórica): Clasificación de audiencia (objetivo).
- **duration** (texto/numérica transformada): Minutos (películas) o temporadas (series).
- **listed_in** (categórica): Géneros asignados al título.

6 Riesgos y limitaciones

Durante el análisis exploratorio se detectaron diversos riesgos y limitaciones que deben considerarse antes de aplicar modelos de machine learning sobre este dataset:

- **Datos faltantes:** Variables como *director*, *country* y *rating* presentan una proporción significativa de valores nulos. Esto puede generar sesgos si no se aplican técnicas de imputación o eliminación adecuadas.
- **Desbalance de clases en rating:** Existen clasificaciones muy frecuentes (como *TV-MA* y *TV-14*), mientras que otras son poco representadas.
- **Inconsistencias en duración:** La variable *duration* mezcla minutos (para películas) con temporadas (para series). Si no se transforma correctamente, puede inducir errores en el entrenamiento de modelos.
- **Sesgo geográfico:** El dataset está dominado por títulos producidos en Estados Unidos e India, mientras que otros países aparecen de forma marginal. Esto puede limitar la capacidad del modelo para generalizar hacia producciones de regiones poco representadas.
- **Evolución temporal desigual:** El número de lanzamientos crece abruptamente en la última década, lo cual puede generar una sobre-representación de títulos recientes frente a décadas anteriores.

- **Campos textuales poco estructurados:** Variables como *listed_in* y *cast* requieren procesamiento de texto (tokenización, limpieza de caracteres, normalización) para ser útiles en un modelo predictivo. Su tratamiento inadecuado puede afectar el desempeño.

- **Posible obsolescencia del dataset:** La información se limita a un corte temporal fijo. Como el catálogo de Netflix cambia constantemente, algunos títulos pueden no estar disponibles actualmente, lo que reduce la vigencia del análisis.

7 Conclusiones iniciales

El análisis exploratorio permitió identificar patrones relevantes como la prevalencia de ciertos géneros y el crecimiento del catálogo en décadas recientes. Además, se determinó que la variable **rating** es adecuada como objetivo de un modelo de clasificación.

En fases posteriores se recomiendaria tomar las siguientes acciones:

- Aplicar limpieza y estandarización en variables textuales.

- Tratar el desbalance de clases en *rating*.

- Evaluar algoritmos de clasificación multiclase (árboles de decisión, random forest).