Netflix Titles Analysis and AI-Based Classification

Miriyala Dhanush Kumar
B.Tech, Artificial Intelligence & Data Science, IITH

Abstract

This report presents a comprehensive data analysis and machine learning application on the Netflix Titles dataset. The project focuses on understanding content distribution trends and applying a classification model to predict whether a title is a Movie or a TV Show using features such as duration, release year, rating, and genre. The model achieves over 99% accuracy, making this project a strong demonstration of both exploratory data analysis and AI techniques.

1. Introduction

The Netflix Titles dataset from Kaggle contains information about films and series available on Netflix, including attributes such as title, type, director, cast, release year, rating, duration, and genres. The goal of this project is twofold:

1. Perform exploratory data analysis (EDA) to discover insights and trends.
2. Develop a supervised machine learning model to classify content type.

2. Data Cleaning

The dataset required several preprocessing steps:

- Removed columns with excessive missing values (e.g., `cast`, `director`).
- Converted the `duration` field into numerical format.
- Extracted the primary genre from the `listed_in` column.
- Filled or dropped rows with missing values in critical columns like `rating`, `country`, and `date_added`.

3. Exploratory Data Analysis

Key findings from the data:

- Netflix has a higher number of **Movies** compared to TV Shows.
- The **United States** leads in content production.
- Most content was added between **2018 and 2020**.
- **TV-MA** is the most frequent rating.

Visualizations included bar plots, line charts, and pie charts to highlight the distribution of content types, top countries, release year trends, and rating frequencies.

4. AI Component: Classification Model

To make this project AI-driven, a classification model was developed to predict the content type (Movie or TV Show) using the following features:

- `duration_int`
- `release_year`
- `main_genre`
- `rating`

Model Used

A **Random Forest Classifier** was chosen for its robustness and ability to handle categorical and numerical data without extensive parameter tuning.

Performance

- Accuracy: **99.88%**
- F1-score: **1.00** for both Movies and TV Shows
- Evaluation techniques included confusion matrix and classification report

5. Technologies Used

- Python
- Jupyter Notebook
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

6. Conclusion

This project demonstrates how basic metadata from a streaming platform can be used not only for visualization but also for predictive modeling. With minimal preprocessing and a few selected features, the AI model achieved near-perfect classification of content types.

Dataset Source

Netflix Titles dataset by Shivam Bansal on Kaggle:
https://www.kaggle.com/datasets/shivamb/netflix-shows

Author: Miriyala Dhanush Kumar
B.Tech in Artificial Intelligence & Data Science, IITH
LinkedIn: https://www.linkedin.com/in/miriyala-dhanush-kumar-7b1482334/