MACHINE LEARNING

EXPERIENTIAL LEARNING PROJECT

Submitted by:

Arijit Debnath

Link to reference dataset and worksheets: https://drive.google.com/drive/folders/1-D3mS2q4zX8X7ZZnn8WxobGix_TDmFWD?usp=sharing

QUESTION 1: DATASET AND PROBLEM STATEMENT

The **dataset** contains comprehensive information about Netflix shows and movies, including attributes such as type (TV Show/Movie), title, director, cast, country, release year, rating, duration, and genre. This dataset serves as a valuable resource for analyzing patterns in content distribution and consumption trends on Netflix.

Problem Statement: Can we predict whether a Netflix title is a "Movie" or a "TV Show" based on its available attributes? This classification problem aims to leverage machine learning models to automate the categorization of Netflix content, which could be useful for content recommendation systems and metadata tagging.

Independent Variables:

- **Release Year:** The year the title was released. Older titles may have different trends compared to newer ones.
- **Rating:** The age rating assigned to the title, which indicates the suitability of the content for different age groups.
- **Duration:** The length of the title in minutes for movies or the number of seasons for TV shows, providing insight into the type of content.
- **Genre (Listed In):** The category of the content, such as Drama, Comedy, or Thriller, which plays a crucial role in differentiating TV Shows from Movies.
- **Country:** The country where the title was produced, as regional content trends can impact the type of content created.

Dependent Variable: Type: A binary variable (0 for Movie, 1 for TV Show) representing the classification to be predicted.

QUESTION 2: DATA CLEANING AND PRE-PROCESSING

To ensure accurate predictions, the dataset underwent extensive pre-processing. This included handling missing values by removing records with crucial missing data, ensuring the dataset retained its integrity without introducing biases. Categorical features such as ratings, genres, and countries were converted into numerical representations using label encoding, making them suitable for machine learning models. Additionally, duration values, which varied between movies (measured in minutes) and TV shows (measured in seasons), were standardized to ensure consistency and improve the accuracy of model predictions.

1. Logistic Regression: Logistic Regression, a linear classification model, was applied to predict the type of Netflix content. It showed exceptional performance with:

Logistic Regression:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1117
1	0.99	1.00	1.00	477
accuracy			1.00	1594
macro avg	1.00	1.00	1.00	1594
weighted avg	1.00	1.00	1.00	1594

This indicates that the model was able to correctly classify all instances without error.

2. k-Nearest Neighbours (kNN): kNN, a non-parametric model based on similarity measures, was used. It also performed well, achieving:

kNN:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1117
1	0.99	0.98	0.99	477
accuracy			0.99	1594
macro avg	0.99	0.99	0.99	1594
weighted avg	0.99	0.99	0.99	1594

This suggests that kNN was slightly less perfect than Logistic Regression but still highly reliable.

3. Naïve Bayes: Naïve Bayes, a probabilistic classifier based on Bayes' theorem, was implemented and produced the following results:

Naïve Bayes:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1117
1	1.00	0.97	0.99	477
accuracy			0.99	1594
macro avg	0.99	0.99	0.99	1594
weighted avg	0.99	0.99	0.99	1594

While it performed well, it had a minor trade-off in recall for TV Shows, indicating a slight difference in handling classification decisions.

QUESTION 3: MODEL SELECTION AND BUSINESS CONTEXT

After evaluating all models, **Logistic Regression** was selected as the best model for deployment. The rationale behind this decision is:

- It achieved a perfect accuracy score of **100%**, meaning there were no misclassifications.
- It has high interpretability, making it easier for businesses to understand the impact of different attributes on predictions.
- Computationally efficient, allowing for quick predictions at scale.

From a **business perspective**, deploying this model can help Netflix or other streaming services automate the classification of new content, reducing manual efforts in tagging and categorization. This can lead to enhanced content discovery, improving search and recommendation algorithms to offer more personalized suggestions to users based on their viewing preferences. Additionally, this model can assist content managers in making data-driven decisions regarding content curation, licensing, and regional content promotion by understanding trends in movies and TV shows. By efficiently categorizing content, streaming platforms can enhance user engagement, retention, and satisfaction, ultimately driving higher subscription rates and revenue growth.