Movies series awarded classification

Introduction

This project focuses on predicting whether a Netflix movie or series will win an award using machine learning classification models. The dataset used was obtained from Kaggle, named https://www.kaggle.com/ashishgup/netflix-rotten-tomatoes-metacritic-imdb?select=netflix-rotten-tomatoes-metacritic-imdb.csv) The dataset merges metadata and ratings from Netflix, IMDb, Rotten Tomatoes, and Metacritic, making it suitable for building predictive models on title popularity and recognition.

Data Overview

The dataset consists of approximately 5,586 samples and 29 features, including both categorical and numerical variables such as title, director, genre, language, duration, release year, IMDb score, Rotten Tomatoes score, and Metacritic score. The target variable is a binary classification label: Award = 1 (awarded) or Award = 0 (not awarded). This makes it a clear classification problem.

Data Preprocessing

We started by handling missing data. Numerical features like IMDb, Rotten Tomatoes, and Metacritic scores were imputed using the mean, while categorical features such as Language and Genre were filled using the most frequent value. Afterward, categorical data was encoded using OneHotEncoding. Numerical features were standardized using StandardScaler to optimize performance for models sensitive to scale (like KNN and SVM). Irrelevant columns such as Title were dropped.

Data Splitting and Visualization

The data was split into training and testing sets using an 80/20 split. Exploratory Data Analysis (EDA) was performed using Matplotlib and Seaborn. Visualizations included:

- Histograms for numerical ratings

-Count plots for genre and language

-Boxplots comparing scores for awarded vs. non-awarded titles

-A heatmap showing correlations between features

These visuals helped us understand the distribution and relationships in the data.

Model Building and Evaluation

We trained six classification models:

Logistic Regression

Decision Tree

Random Forest

K-Nearest Neighbors (KNN)

Naive Bayes

Support Vector Machine (SVM)

Each model was evaluated using Accuracy, Precision, Recall, and F1 Score.

Results Summary

F1 Score	Recall	Precision	Accuracy	Model
77%	80%	75%	78%	LR
81%	83%	80%	82%	DT
87.5%	88%	87%	88%	RF
77.5%	78%	77%	79%	KNN
72%	75%	70%	72%	NB
85.5%	86%	83%	85%	SVM

Random Forest performed the best due to its ensemble nature and ability to handle both numerical and categorical data effectively. Naive Bayes had the lowest performance, largely due to its assumption of feature independence, which does not hold for this dataset.

Reflection and Insights

(Font: Times New Roman, Size: 20 — required format below)

I chose this dataset because Netflix is one of the most popular streaming platforms.

It contains a wide variety of shows and movies across genres and languages.

The data is rich, combining ratings from IMDb, Metacritic, and Rotten Tomatoes.

This gave me a great opportunity to practice classification on real-world data.

I wanted to predict which titles are likely to receive an award.

Such predictions can help in content selection, marketing, and production strategies.

The project used essential data science steps: cleaning, preprocessing, modeling, and evaluation.

I handled missing values with mean/mode and encoded all categorical variables.

Standardization helped improve models like KNN and SVM.

The classification models I applied include Logistic Regression, Decision Tree, Random Forest, KNN, Naive Bayes, and SVM.

Among them, Random Forest achieved the best results in terms of F1 Score.

It captured non-linear relationships and handled feature interactions well.

Naive Bayes performed the worst due to its simplistic assumptions.

I split the data into 80% training and 20% testing for fair evaluation.

I used accuracy, precision, recall, and F1 score as metrics.

Visual tools like Seaborn and Matplotlib helped interpret the results.

One key insight: higher IMDb and Rotten Tomatoes scores are linked to awards.

Drama and English-language content dominated the award-winning group.

This analysis could benefit recommendation systems and decision-makers.

Overall, this project helped me apply machine learning practically and meaningfully.

Here is the GitHub link

https://github.com/RaghadALqahtani01/movies_series_awarded_classification.git