University of Illinois Springfield

Final Project Report

Predicting Netflix Movie Genres with Machine Learning

Table of Contents

Abstract

This project focuses on using machine learning techniques to predict the genre of Netflix movies based on their plot descriptions. The goal is to classify movies into one of six genres: drama, comedy, horror, action, thriller, and romance. By using supervised learning algorithms like Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Networks, we aim to identify patterns in the textual data that can help determine the genre of a movie. We preprocess the text data using techniques like TF-IDF vectorization and evaluate the models using metrics like accuracy, precision, recall, and F1-score. The project not only predicts the genres but also offers insights into how well these models perform. This work contributes to improving content recommendation systems by utilizing machine learning to automate genre classification, making it easier for users to discover movies they will enjoy.

Problem Definition and Project Goals

With the growing volume of content on streaming platforms like Netflix, manually assigning genres to each title is time-consuming and prone to inconsistencies. Titles often span multiple genres, and their classification can vary based on interpretation. This creates challenges in organising content and delivering accurate recommendations to users. The problem, therefore, is to develop an automated, reliable method for predicting genres based on a title's description and metadata, using machine learning techniques. Originally, the data had the following variables :

Variable Name	Data Type	Description
Release_year	Integer	The year the movie or show was released.
Title	String	The name of the Netflix movie or series.
Origin	String	The country or region where the content was produced.
Director	String	Name of the director responsible for the film or series.
Cast	String	A list or comma-separated string of actors featured in the title.

Variable Name	Data Type	Description
Genre	String / List	Genre or multiple genres assigned to the title (e.g., Drama, Comedy).
Plot	String	A brief description or summary of the storyline or content.

The dataset used in this project includes around 30000 entries, each representing a Netflix movie or TV show. For every title, details such as the year of release, country of origin, director, cast members, genre, and a short plot description are provided. With a substantial number of records, it provides enough data to train and evaluate machine learning models effectively. The diversity within the dataset also helps in building a more generalizable model for predicting genres based on textual features.

Related Work

Previous research has shown the effectiveness of machine learning in text classification, including genre and sentiment prediction. Approaches often use TF-IDF vectorization or word embeddings, followed by classifiers such as Naive Bayes, SVM, and neural networks. Prior studies on IMDb and similar datasets have faced issues like multi-label classification and class imbalance, addressed through label filtering and balanced sampling. This project builds on these strategies, using similar preprocessing and modelling steps tailored for Netflix content.

Data Exploration and Preprocessing

1. Understanding the Dataset

The dataset includes around 30000 records, each representing a Netflix title. Key fields like Title, Plot, Genre, Cast, Director, and Release_year offer both textual and categorical information. Since the primary goal is to predict genre using descriptions, the focus was placed on the Plot column.

2. Missing Value Handling

- The dataset may contain movies labeled with an unknown genre or missing genre information. These entries could distort the model's training and testing, so they are removed.

- We only worked with rows which had single genre entries. Movies with multiple genres were not considered.

3. Text Cleaning and Tokenization

- Text Cleaning: The raw plot descriptions of movies are cleaned to remove noise such as punctuation, numbers, and unnecessary whitespace. Text preprocessing techniques are applied to ensure that the data is in a consistent format for feature extraction.

- Tokenization: The plot descriptions are tokenized into individual words using custom tokenization. This breaks down the text into manageable units for analysis.

- Stopword Removal: Common, non-informative words such as "the," "and," "is," etc., are removed from the text to prevent them from affecting the classification process.

4. Converting Text to Features

Since machine learning models work with numbers, the cleaned plot summaries were transformed into numeric vectors using **TF-IDF (Term Frequency–Inverse Document Frequency)**. This method gives more weight to unique words in each plot, which helps models focus on important terms rather than frequent but uninformative ones.

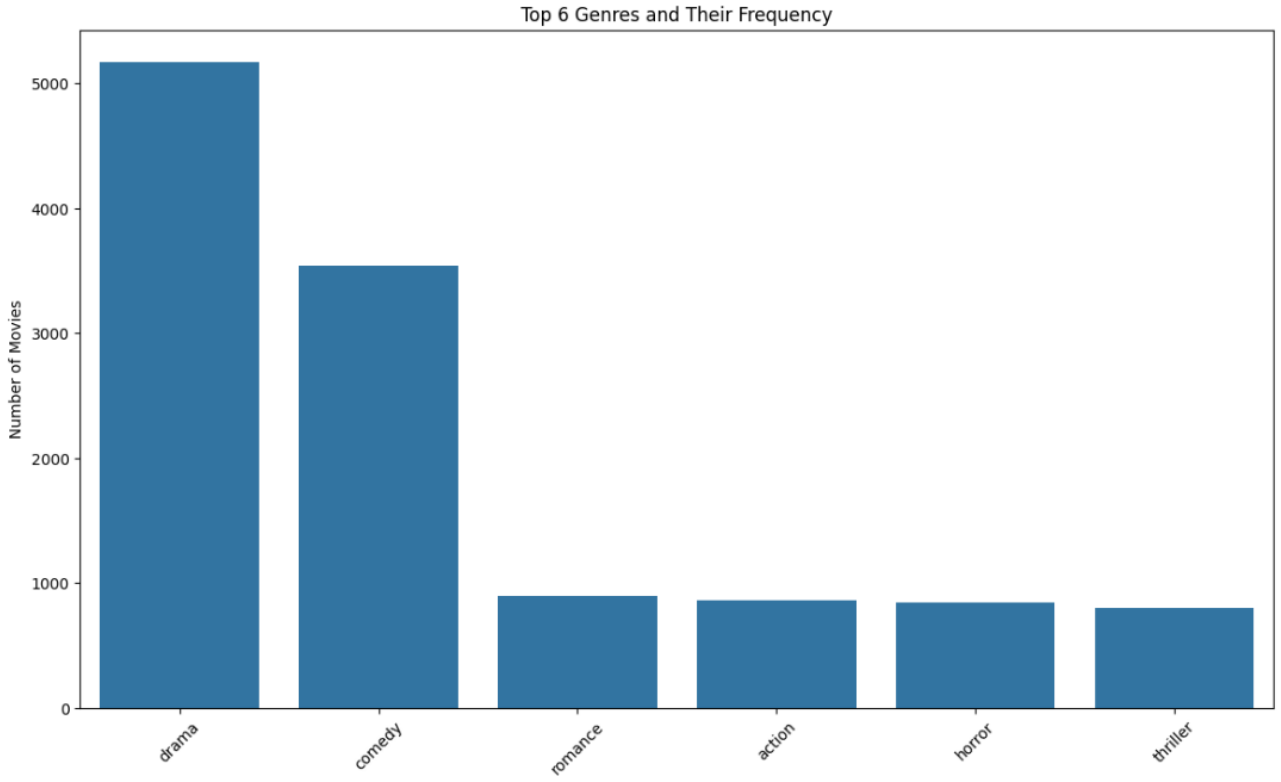
6. Genre Label Encoding

To train classification models, genre labels were converted from text to numbers using a label encoder. This allowed the models to understand the genres as target variables and also we worked on 6 genres {drama, comedy, romance, action, horror, thriller}, each were labeled 1-6 respectively.

Data Analysis and Experimental Results

To understand the linguistic patterns behind different movie genres, frequency distribution graphs were created for the top 10 most used words in plot summaries, represented in Figures F1 to F6. In the figure 0, we can see the top 6 most frewuently occurring genres and we can say the dataset is imbalance.

Fig 0 : Genres and their frequency



We work on the plot description to find the genre of the movie. So it is essential to find what kind of words are frequently occurring in the plot description for a specific genre. We tried to analyse the top occurring words for each genre as graphs. Drama genres often involve strong emotional connections and family-oriented themes, as evidenced by frequent words like "father", "home", and "love". The high frequency of "one" may reflect a focus on individual journeys or personal stories. This aligns with typical themes in drama films, such as family dynamics, emotional struggles, and personal growth.

Common words like "death", "killed", and "body" show a focus on dark, intense, and often violent themes. Horror films often revolve around places (like houses) where something sinister or supernatural is discovered. The action genre includes words like "kill", "police", and "father", which are typical of high-energy, conflict-driven plots. Action films often revolve around violence, confrontation, and dramatic conflict, such as battles between police and criminals, or personal vendettas.

Words like "police", "one", and "man" point to intense, suspenseful narratives. Thrillers often focus on tension, investigations, or survival. The use of "house" and

"finds" suggests that many thrillers center around mysterious locations or discoveries, while "police" highlights crime-related or investigative themes. Romance films are centered around relationships, with words like "love", "family", and "get" showing a focus on emotional connections, familial involvement, and romantic pursuits. The use of "father" implies themes involving family approval, relationships with parents, or conflicts related to love and family.

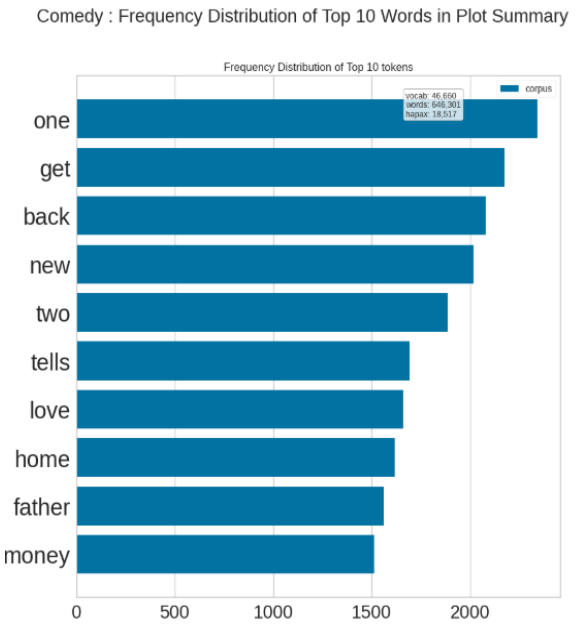


Figure 1

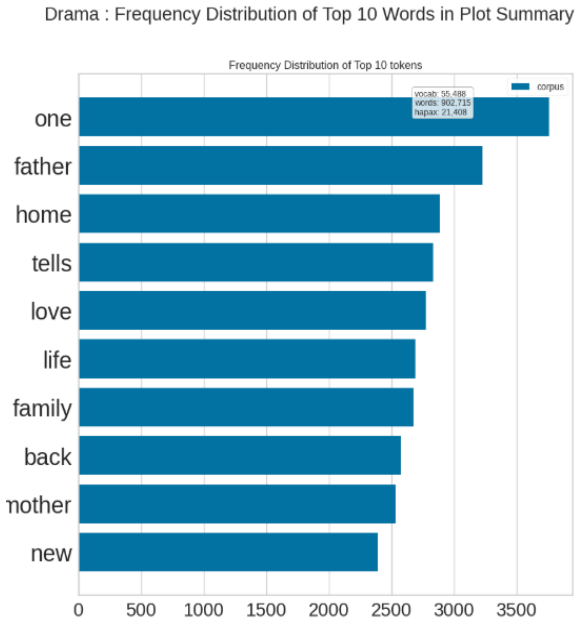


Figure 2

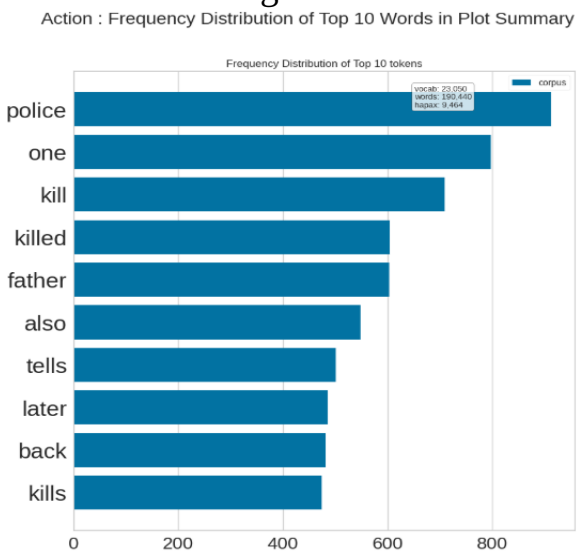


Figure 3

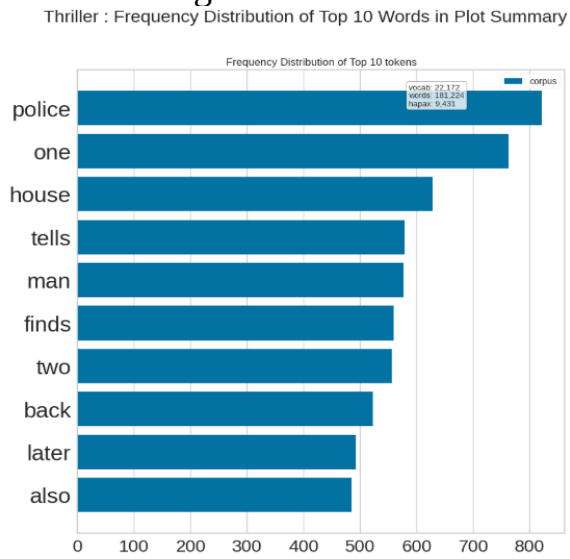


Figure 4

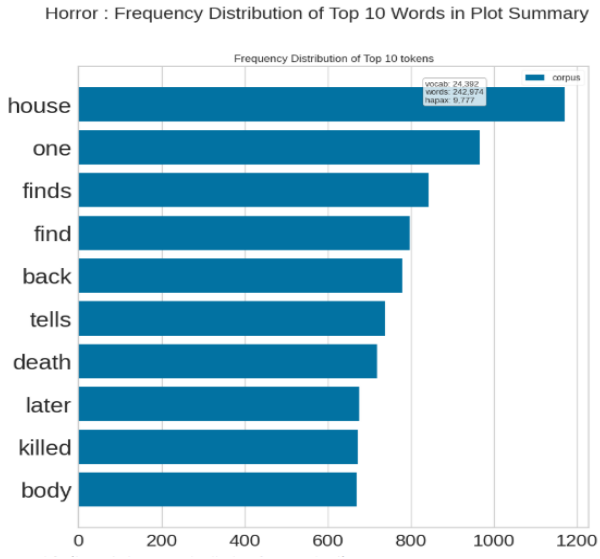


Figure 5

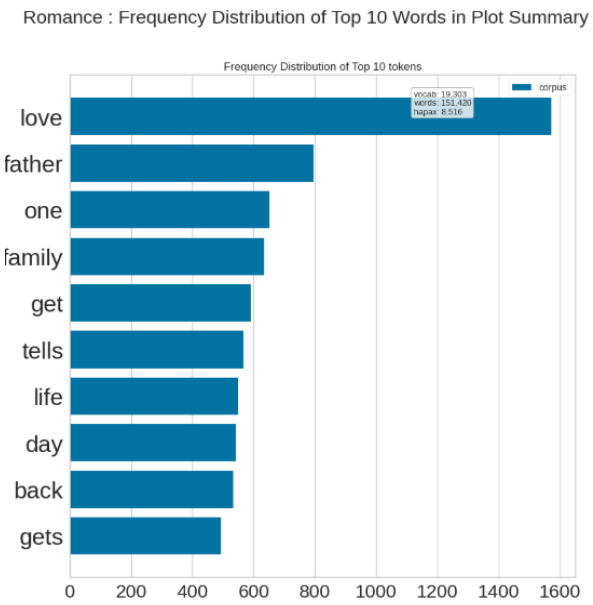


Figure 6

Data Modelling and Results

TF-IDF

To convert the movie plot descriptions into a form that machine learning models can understand, we used TF-IDF (Term Frequency–Inverse Document Frequency). This technique transforms text into numerical values by measuring how important a word is in a specific document compared to all documents in the dataset. Words that appear frequently in a particular plot but not in others are given higher importance. This way, the model focuses more on unique, meaningful words instead of common ones like “the” or “and.”

Once the text was converted into TF-IDF vectors, we trained and tested tfour different machine learning models:

- **Neural Networks:** These models mimic how the human brain learns. In our case, the network took the TF-IDF features as input and learned patterns through multiple layers of interconnected nodes (neurons). It was especially good at capturing complex relationships in the text, leading to the highest accuracy among all models tested.
- **Support Vector Machines (SVM):** SVMs work by finding the best boundary (or hyperplane) that separates data points of different genres. With the TF-IDF features, the SVM learned how to distinguish between genres by identifying which word combinations best defined each class. It performed very well, especially in handling high-dimensional TF-IDF data.

- **Random Forest:** This model is made up of multiple decision trees. Each tree makes a prediction based on a subset of the data, and the final result is based on majority voting. Random Forest was used to learn decision rules from the word patterns in the plots. While slightly less accurate than the neural network and SVM, it was still effective and easy to interpret.

- **Neural Network:** Neural Networks are used to classify movies into genres based on plot descriptions. By leveraging deep learning, the neural network can model complex, non-linear relationships between the plot text and the genre labels. The text is transformed into numerical features using TF-IDF vectorization, and the network learns to predict genres from these features.

After classifying the data using each model, based on the performance metrics, we evaluated each model. The performance of the four machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Networks—was evaluated across the top 6 genres: Drama, Comedy, Horror, Action, Thriller, and Romance.

Model	Precision	F1 score	Recall	Support
LogisticRegression	0.59	0.61	0.58	2424
Random Forest	0.58	0.56	0.5	2424
Support Vector Machine	0.60	0.60	0.57	2424
Neural Network	0.55	0.55	0.55	2424

Logistic Regression achieved the highest average F1-score, precision, and recall, making it the best-performing model overall. The model performs particularly well on common genres like Drama, with high recall indicating its ability to identify these genres reliably. However, its performance on rarer genres like Horror and Romance could be improved.

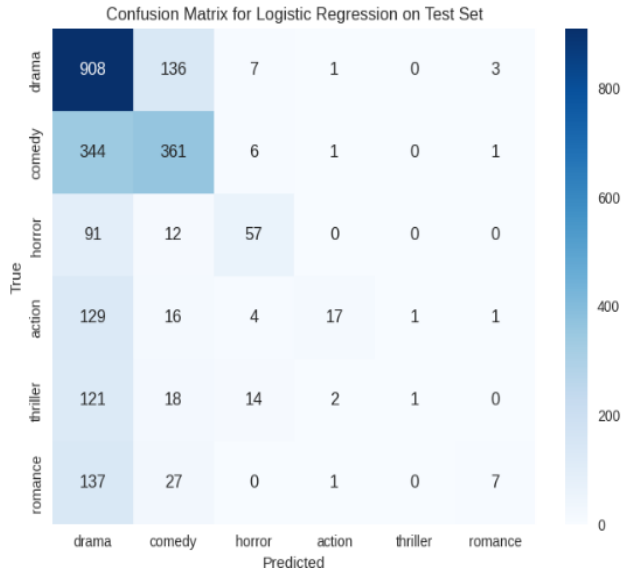
Random Forest exhibits similar performance to Logistic Regression but slightly lags behind in terms of the F1-score. The model handles the Drama genre well, as indicated by the high recall, but its performance drops significantly on Thriller and Romance, as evidenced by the low recall and precision for those genres.

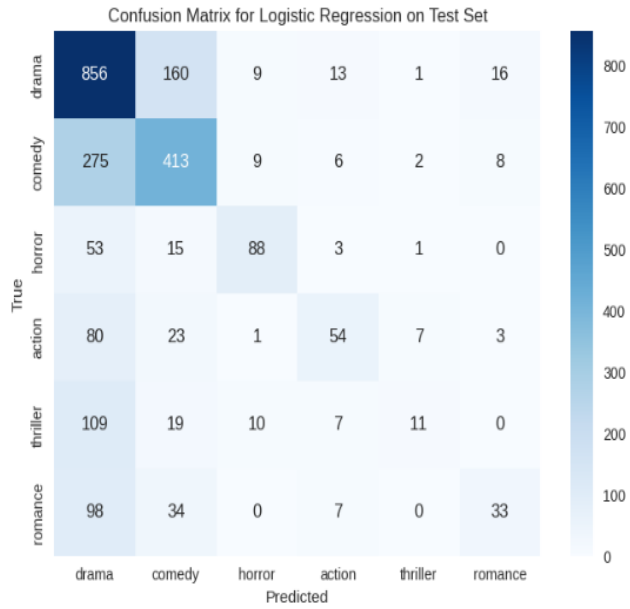
SVM shows competitive performance with Logistic Regression, with precision and recall values slightly higher than Random Forest. Its average recall for Drama and Comedy genres is good, but like Random Forest, it struggles with Horror and Romance genres, particularly in recall. SVM is generally effective for this classification task, but

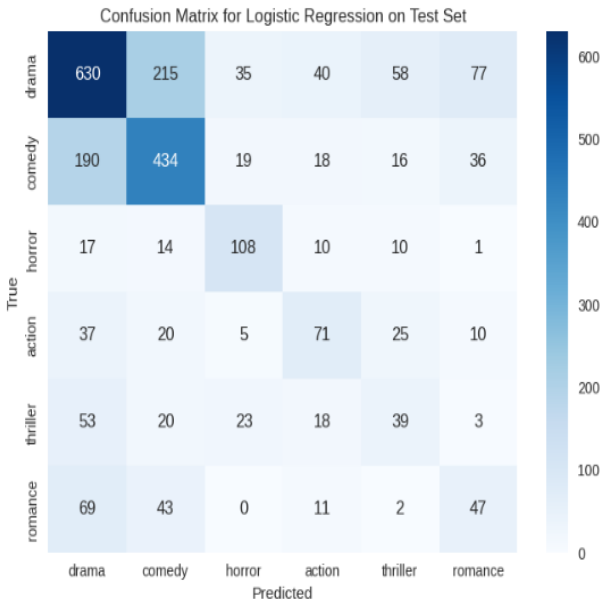
it may benefit from parameter tuning or additional features to boost performance on certain genres.

Neural Networks have the lowest precision, recall, and F1-score across all models. While the model performed decently on Drama and Comedy, it struggled significantly with Romance and Action genres.

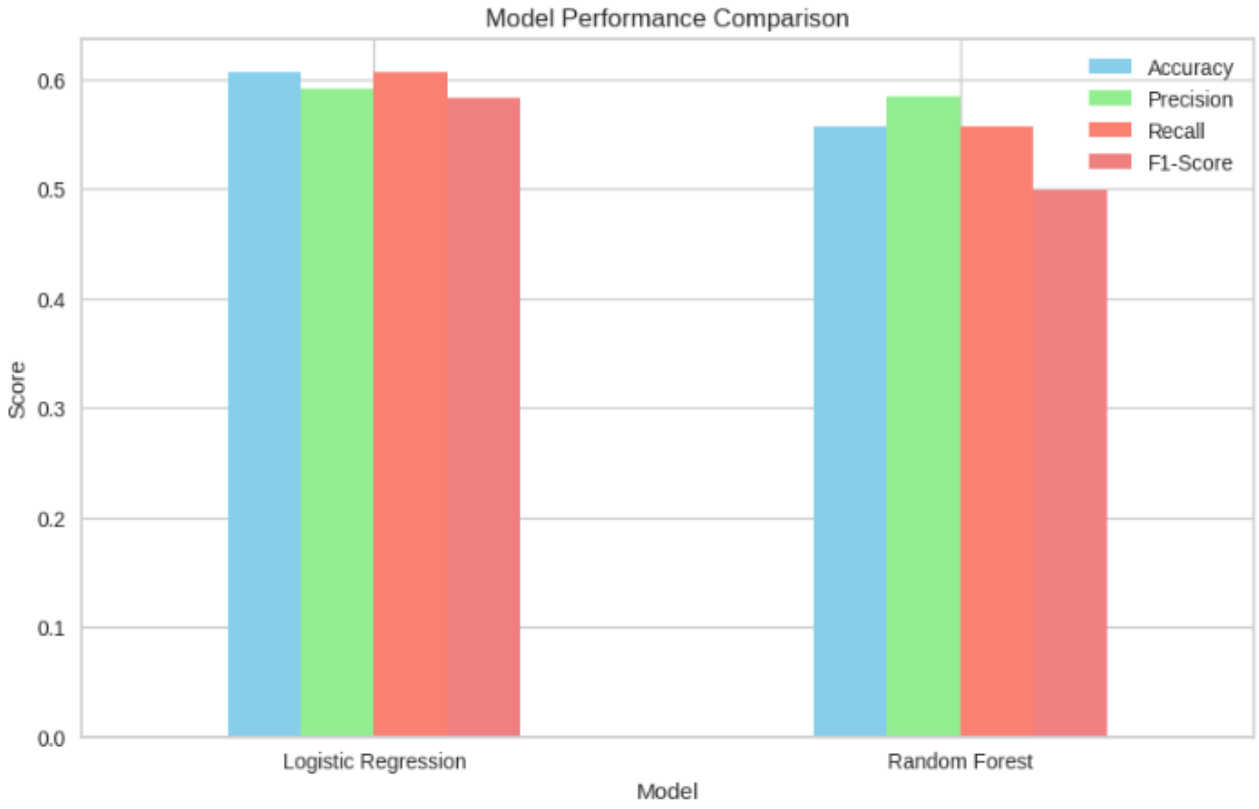
We also plotted confusion matrix for all models on the test set, most of the models performed well on the drama and comedy due to the dominance of these genres in dataset.







Logistic Regression and Random Forest perform similarly, with small differences between them across all four metrics.



Conclusion

The Netflix Movie Genre Prediction project aimed to classify movies into six genres based on plot descriptions using machine learning models. This project demonstrated that plot descriptions contain meaningful signals for predicting genres using machine learning. Through careful preprocessing and model evaluation, we achieved good performance across all classifiers. Among the models tested, Logistic Regression performed the best, with the highest accuracy and recall, particularly for common genres like Drama. Random Forest showed slightly better F1-score due to its balance between precision and recall, but struggled with rare genres. SVM and Neural Networks showed decent performance but were outperformed by Logistic Regression and Random Forest.

References

- https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots
- Netflix public datasets and Kaggle resources on movie metadata
- https://scikit-learn.org/1.4/tutorial/text_analytics/working_with_text_data.html

- https://scikit-learn.org/stable/modules/model_evaluation.html
-

Contributions

Data collection and cleaning: Sanjana Siluveru and Neeraj Giramoni

Text preprocessing and feature engineering: Sanjana Siluveru

Model building and evaluation: Neeraj Giramoni and Sanjana Siluveru

Visualization and analysis: Neeraj Giramoni