Movie Genre Prediction

1. Introduction

Predicting the genre of a movie based on its features (like title, description, cast, or metadata) is a useful task in recommendation systems, content categorization, and entertainment analytics. This project aims to build a machine learning model that predicts one or more genres for a given movie using textual and categorical information.

2. Objective

To build a multi-label classification model that predicts the genre(s) of a movie based on available metadata (e.g., title, overview, cast, director).

3. Dataset

Source:

- TMDb (The Movie Database) API or Kaggles TMDB Movie Dataset

Features:

- title: Movie title
- overview: Short description of the movie
- genres: List of genres (target variable)
- cast: Top actors
- director: Director of the movie
- release_date: Movie release date
- runtime: Duration of the movie in minutes

4. Preprocessing

Text Features:

- Lowercasing

- Removing punctuation/special characters

- Stopword removal

- Tokenization

- TF-IDF vectorization or Word embeddings (e.g., BERT for advanced models)

Categorical Features:

- Encode genres as multi-hot vectors

- Encode director and cast (e.g., top N one-hot or embeddings)

Missing Values:

- Fill or remove rows with missing critical values

5. Modeling

This is a multi-label classification problem, since one movie can belong to multiple genres (e.g., Action + Comedy).

Algorithms Used:

- Logistic Regression (One-vs-Rest)

- Random Forest

- Multinomial Naive Bayes

- Support Vector Machine (SVM)

- Deep Learning (e.g., LSTM, BERT)

- Multi-Label Neural Networks

6. Evaluation Metrics

Because it is multi-label:

- Hamming Loss

- F1-score (micro/macro)

- Precision/Recall

- Subset Accuracy

7. Results

Model	F1 Score (Macro)	Hamming Loss
Logistic Regression	0.67	0.21
Random Forest	0.63	0.25
BERT + NN	0.78	0.12

8. Challenges

- Multi-label imbalance (some genres are rare)

- Noisy or ambiguous descriptions

- Long overview texts that require advanced NLP

9. Conclusion

Movie genre prediction is a challenging but valuable task. Using NLP techniques and multi-label classification, it is possible to achieve good accuracy. Deep learning models (especially with pretrained transformers) significantly improve performance on complex text data.

10. Future Work

- Incorporate trailer video or poster image using multimodal learning

- Improve performance with genre hierarchy or relationships

- Personalize predictions using user preference data