# Analysis and Classification of Bollywood Movies (2005-2017)

## Introduction

This project analyzes a dataset of 1698 Bollywood movies released between 2005 and 2017. The objective is to predict whether a movie will be a hit or a flop based on various attributes using two classification models: Decision Tree and Naïve Bayes.

## Dataset Description

The dataset contains information on several attributes of Bollywood movies, including:

- Movie Name

- Release Period (Holiday/Normal)

- Whether Remake/Franchise (Yes/No)

- Genre

- New Actor, Director, Music Director (Yes/No)

- Lead Star

- Director

- Music Director

- Number of Screens

- Revenue (INR)

- Budget (INR)

## Data Preprocessing

### Creating the Target Variable

A new column 'hit\_or\_flop' was created based on the condition:

• Hit: If Revenue > Budget

• Flop: Otherwise

### Handling High Cardinality Categorical Variables

Some categorical features had a very high number of unique values:

• Genre: 14 unique values

• Lead Star: 764 unique values

• Director: 1048 unique values

• Music Director: 630 unique values

Due to their high cardinality, the columns Lead Star, Director, and Music Director were dropped to simplify the model.

### Encoding Categorical Variables

Binary categorical columns were label-encoded (Yes/No → 1/0).

### Checking for Data Imbalance

Class distribution:

• Hits: 981

• Flops: 377

Since the class distribution was not severely imbalanced, we proceeded without resampling.

## Model Training and Evaluation

We trained two classifiers:

1. Naïve Bayes Classifier

2. Decision Tree Classifier

Each model was evaluated using multiple metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC Score.

### Naïve Bayes Results

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.85 |
| Precision | 0.85 |
| Recall | 0.96 |
| F1-score | 0.90 |
| ROC-AUC | 0.76 |

### Decision Tree Results

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.96 |
| Precision | 0.96 |
| Recall | 0.98 |
| F1-score | 0.97 |
| ROC-AUC | 0.94 |

## Comparison of Classifiers

|  |  |  |
| --- | --- | --- |
| Metric | Decision Tree | Naïve Bayes |
| Accuracy | 0.96 | 0.85 |
| Precision | 0.96 | 0.85 |
| Recall | 0.98 | 0.96 |
| F1-score | 0.97 | 0.90 |
| ROC-AUC | 0.94 | 0.76 |

## Conclusion

• The Decision Tree classifier consistently outperformed Naïve Bayes in all evaluation metrics.

• The Naïve Bayes classifier performed relatively well but struggled with precision and recall.

• Future improvements could include:

- Feature engineering to extract more meaningful attributes

- Trying ensemble methods such as Random Forest

- Performing hyperparameter tuning for further model optimization.