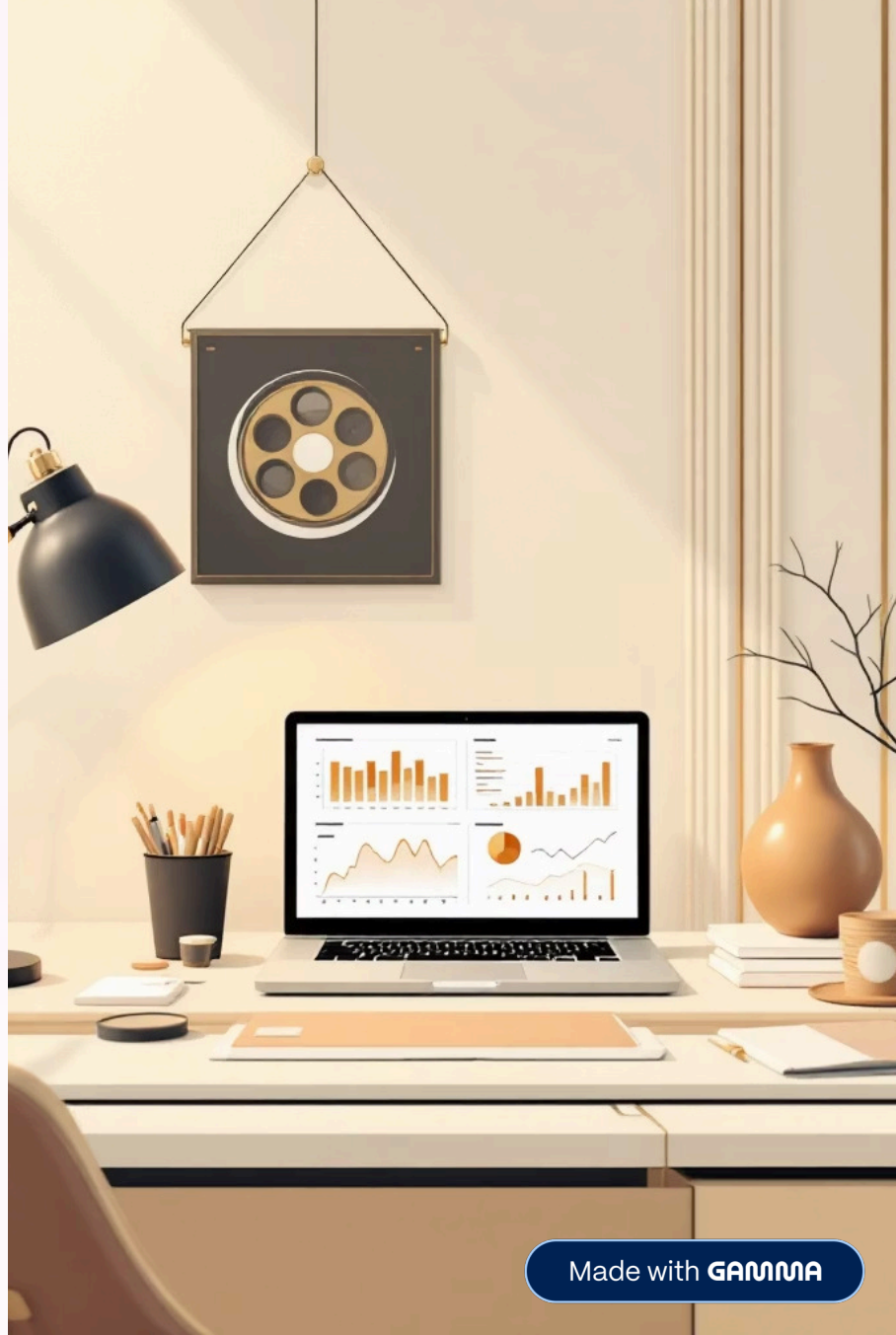
Movie Success Prediction Using Supervised Machine Learning

Submitted By: Utkarsha Patil

Course: Bachelor of Computer Applications (BCA)

Academic Year: 2025–2026



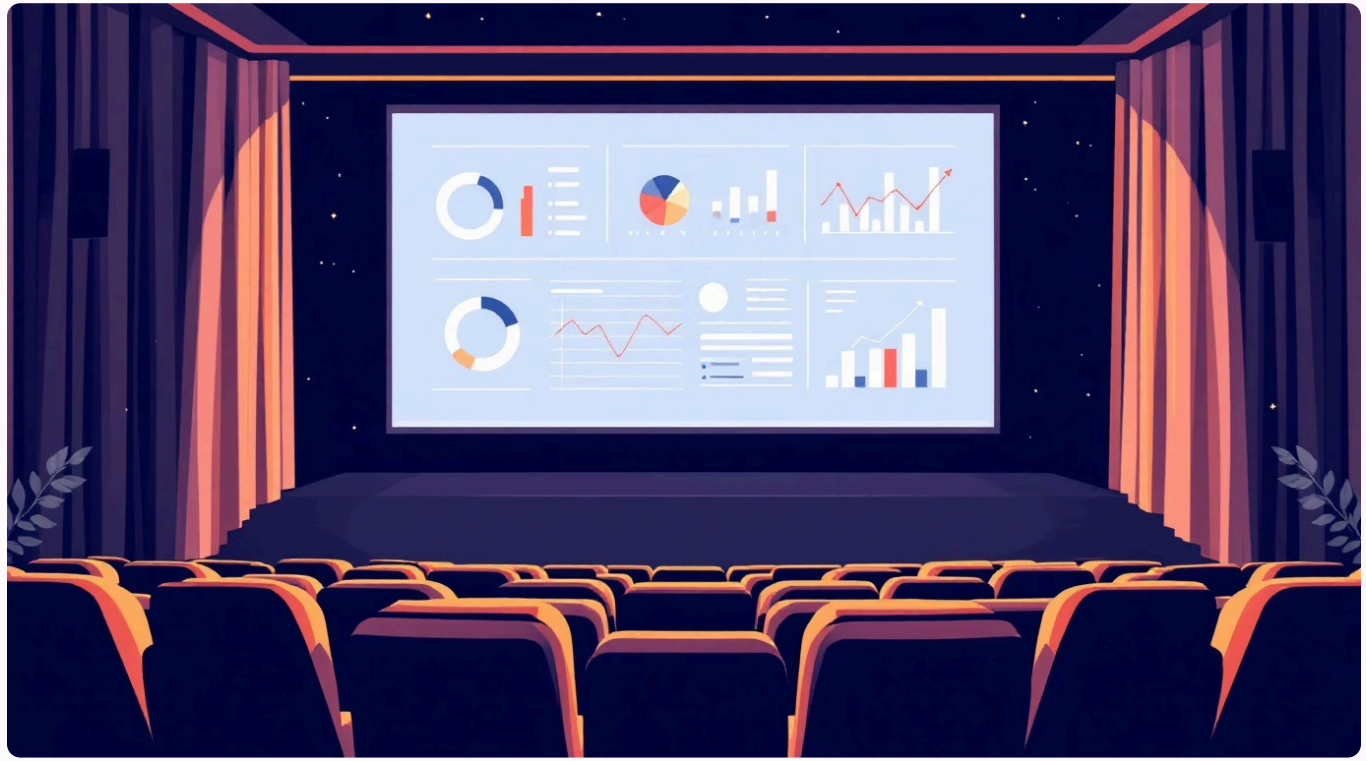
Abstract

This project aims to predict the success of a movie using supervised machine learning algorithms. The prediction is based on pre-release features such as budget, genre, cast popularity, director rating, marketing spend, release season, and runtime. Multiple classification algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting are implemented and compared. The model performance is evaluated using accuracy, confusion matrix, ROC curve, and precision-recall curve. The best performing model is selected and saved for future predictions.

Introduction

The film industry involves high financial risk, and predicting movie success before release is challenging. Machine learning provides an effective way to analyze historical movie data and identify patterns that influence box office success.

Supervised learning is used in this project because the dataset contains labeled outputs (Hit or Flop). The model learns from historical data and predicts the outcome for new movies.



Objectives

Build Model

To build a supervised learning model to predict movie success.

Compare Algorithms

To compare multiple classification algorithms.

Analyze Features

To analyze feature importance.

Evaluate Performance

To evaluate performance using different metrics.

Visualize Insights

To visualize insights using attractive graphs.

Save Model

To save the best performing model.

Literature Review (Short Version)

Previous studies show that budget, marketing spend, star power, and genre significantly influence movie success. Machine learning models such as Random Forest and Gradient Boosting often provide high accuracy in classification problems.

Methodology

The project follows the Machine Learning Pipeline:

01

Data Collection

Synthetic dataset created with movie features.

02

Data Preprocessing

- Handling categorical variables using Label Encoding
- Feature-target separation
- Train-test splitting

03

Model Training

Algorithms used:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- Gradient Boosting

04

Model Evaluation

Metrics used:

- Accuracy Score
- Confusion Matrix
- Classification Report
- ROC Curve
- Precision-Recall Curve
- Learning Curve

05

Model Selection

The best model is selected based on highest accuracy.

System Architecture

1. Input Features
2. Data Preprocessing
3. Model Training
4. Model Evaluation
5. Prediction Output

Algorithms Used

1. Logistic Regression

Used for binary classification problems.

2. Decision Tree

Tree-based model that splits data based on feature conditions.

3. Random Forest

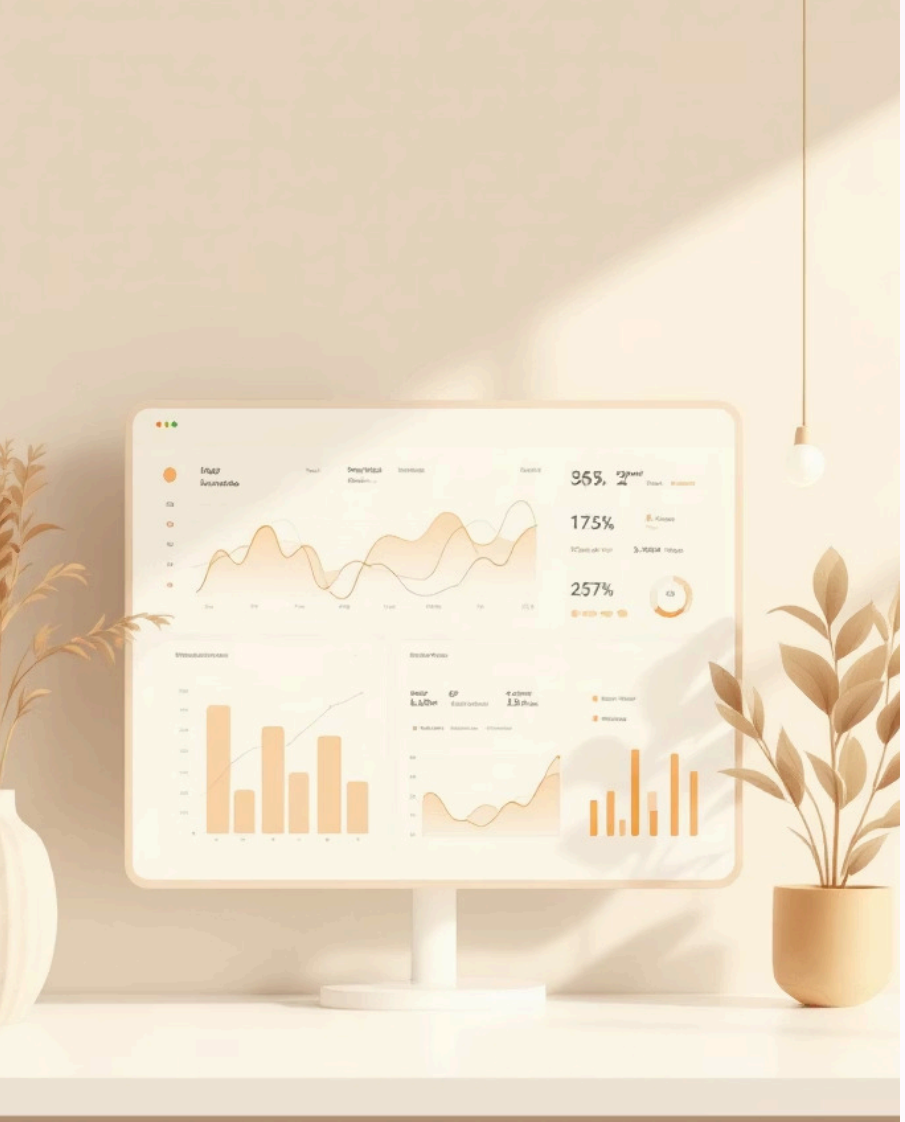
Ensemble method combining multiple decision trees.

4. Support Vector Machine

Finds optimal boundary to separate classes.

5. Gradient Boosting

Sequential boosting algorithm that improves weak learners.



Data Visualization

The following visualizations were used:

- Count Plot (Hit vs Flop distribution)
- Box Plot (Budget vs Status)
- Correlation Heatmap
- Pair Plot
- Feature Importance Graph
- Accuracy Comparison Graph
- ROC Curve
- Precision-Recall Curve
- Learning Curve
- Confusion Matrix

These graphs help in understanding relationships between features and model performance.

Results

Multiple models were trained and compared.

Best Models

Random Forest / Gradient Boosting achieved the highest accuracy.

Key Features

Budget and Marketing Spend were found to be the most important features.

Model Saved

The model was saved using joblib for future use.

Advantages

- Early prediction of movie success
- Helps producers in decision making
- Reduces financial risk
- Data-driven approach

Limitations

- Uses small synthetic dataset
- Real-world data may be more complex
- Accuracy depends on data quality

Future Enhancements

- Use real Kaggle dataset
- Add Deep Learning models
- Deploy as web application
- Integrate real-time data
- Add sentiment analysis from social media

Conclusion

This project successfully demonstrates the application of supervised machine learning in predicting movie success. By comparing multiple algorithms and analyzing feature importance, the system identifies key factors influencing movie performance. The model achieves good accuracy and can be extended for real-world deployment.

Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn
- Joblib

References

- Scikit-learn Documentation
- Kaggle Movie Dataset
- Machine Learning Research Papers
- Python Official Documentation