# Movie Analysis and Classification Report

## 1. Introduction

This project focuses on analyzing and classifying movies using the TMDB Movies Dataset. With over 350,000 movies, the dataset contains various attributes such as ratings, revenues, runtimes, original language, and genres. The goal of this project is to explore trends in the dataset, build a classification model, and deploy it for real-world testing.

## 2. Environment and Tools

The project was executed on Kaggle, leveraging the following Python libraries:  
- pandas for data manipulation  
- NumPy for numerical operations  
- Matplotlib & Seaborn for visualization  
- Scikit-learn for machine learning model development

## 3. Data Preprocessing

### 3.1 Handling Missing Data

The dataset was examined for missing values using `df.isna().sum()`. It was found that the 'genres' column contained 59,600 missing values, which were replaced with 'unknown'.

### 3.2 Data Cleaning

To improve data quality, movies with a vote average of 0 were removed. The release\_date column was converted into a datetime format for better analysis.

## 4. Exploratory Data Analysis (EDA)

### 4.1 Distribution of Ratings and Popularity

The distribution of vote averages was analyzed, showing that most movies had ratings between 5 and 8. A correlation analysis between runtime and ratings was also performed, revealing no strong relationship.

### 4.2 Genre Analysis

The dataset contained multiple genres per movie. The most common genres were:  
- Action  
- Drama  
- Comedy  
- Thriller  
- Science Fiction

## 5. Movie Classification Model

### 5.1 Model Selection

A classification model was developed to predict movie genres based on the overview text. TF-IDF was used for text vectorization, and the following models were tested:  
- Logistic Regression  
- Random Forest  
- Naïve Bayes

### 5.2 Model Performance

The models were evaluated using Accuracy and F1-score. Logistic Regression with TF-IDF achieved the highest accuracy, surpassing 80%.

## 6. Deployment

The final model was saved and tested on new data. The project was documented and uploaded to GitHub for further improvements and collaboration.

## 7. Conclusion and Recommendations

### 7.1 Key Findings

- Action, Drama, and Comedy were the most common genres.  
- There was no strong correlation between runtime and ratings.  
- Logistic Regression with TF-IDF performed best for genre classification.

### 7.2 Future Improvements

- Implement deep learning models like BERT for better text classification.  
- Integrate director and actor information to enhance predictions.  
- Deploy the model as an API using Flask or FastAPI for real-world applications.