PROJECT REPORT DATA SCIENCE FALL 2023

Movies Recommendation

Bilal Ahmad, Taha Nawaz, and Ali Ahmad

1

Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

**Abstract**

Movies are a form of storytelling through visual and auditory mediums, combining elements like narrative, acting, cinematography, sound, and special effects to evoke emotions and convey ideas. People tell stories about their lives or turn hard times into art, because it’s incredible that it’s even possible to do that. And when you see something that resonates with you and speaks to apart of you that you’ve never quite been able to put into words and makes you feel emotions.

A movie recommendation system suggests films to users based on their preferences, viewing history, or specific criteria. This project aimed to develop a movie recommendation system using machine learning techniques to predict movie success, identify trends, and provide actionable insights for stakeholders in the film industry.

For clustering, the K-Means algorithm grouped movies based on features like budget, popularity, and genre, with t-SNE used to visualize and confirm distinct clusters. Dimensionality reduction techniques, such as PCA and t-SNE, simplified the dataset and revealed relationships between movie attributes. Visualization tools like Matplotlib, Seaborn, and t-SNE were used to explore feature relationships, visualize clusters, and analyze the impact of various attributes on movie success.

In conclusion, this project demonstrates the potential of data science to transform the movie recommendations. By using clustering, dimensionality reduction, and predictive modeling, this work lays the groundwork for creating more effective and personalized movie recommendation systems in the future.

# Introduction

In the film industry, understanding what drives a movie's success is crucial for making data-driven decisions. With the rise of streaming platforms and personalized recommendations, movie recommendation systems have become integral tools for helping users discover content that suits their preferences.

This project aims to develop a comprehensive movie recommendation system using machine learning techniques, leveraging a variety of movie attributes such as genres, budget, revenue, popularity, and more.

*Project objective and goals:*

* **Data Preprocessing**: Clean and prepare the dataset for analysis.
* **Clustering**: Group movies based on shared attributes using K-Means.
* **Dimensionality Reduction**: Use PCA and t-SNE for better data visualization.
* **Revenue Prediction**: Predict movie revenue using Random Forest regression.
* **Recommendation System**: Build a system to suggest movies based on user preferences.
* **Model Evaluation**: Evaluate and optimize the models for accuracy.
* **Insights**: Provide actionable insights for film industry stakeholders.
* **Visualization**: Visualize results to better understand movie trends

# Methodology

The methodology of this project outlines the steps taken to process the movie dataset, apply machine learning models, and analyze the results. The project consists of several key phases: data collection, data preprocessing, clustering, dimensionality reduction, and predictive modeling. Below is a detailed breakdown of the approach used in each phase.

### ***Data Collection:***

* **Source**: The dataset was sourced from [Kaggle](https://www.kaggle.com/datasets/makray/tmdb-5000-movies), a popular platform for datasets and data science competitions.
* **Dataset Name**: "TMDB 5000 Movie Dataset," which includes information about over 4,803 movies.
* **Features**:
  + Includes **20 columns**, such as **budget**, **revenue**, **genres**, **runtime**, **release dates**, **popularity**, and **languages**.
  + Provides detailed attributes of movies like **cast**, **production companies**, and **overview**.
* **Origin**: The data was collected from **The Movie Database (TMDb)**, a community-driven movie database.
* **Relevance**: The dataset is ideal for analyzing various factors influencing movie success, such as production budgets, popularity, and audience reception.
* **Availability**: The dataset is publicly available on Kaggle, making it accessible for data science projects.
* **Purpose**: This rich dataset provided the foundation for building a movie recommendation system, performing clustering, and predicting movie revenue.

#### **Data Preprocessing:**

* The dataset was loaded using pandas, and its structure was inspected to identify numerical, categorical, and textual features.[1]
* Missing values in numerical columns (e.g., budget, revenue) were handled using mean imputation through the SimpleImputer from scikit-learn.
* Numerical features like budget, popularity, and runtime were scaled using StandardScaler to ensure all features had the same scale, improving the performance of clustering and regression models.
* Categorical variables, such as genres, were encoded using one-hot encoding. This transformed categorical variables into binary features, making them suitable for machine learning algorithms.

#### **Clustering:**

* The goal of clustering was to group movies into distinct categories based on their similarities in features such as budget, popularity, and genre.
* After applying K-Means, the results were visualized using t-SNE to confirm the distinctiveness of the clusters. t-SNE helped in understanding how well the movies were grouped based on their feature set.

#### **Dimensionality Reduction:**

The purpose of dimensionality reduction was to reduce the high-dimensional dataset into two dimensions while retaining as much of the variance as possible.

* Principal Component Analysis (PCA) was performed to reduce the dataset's dimensionality and identify key components while retaining most of the variance.
* t-Distributed Stochastic Neighbor embedding (t-SNE) was used to visualize high-dimensional data in a two-dimensional space.

#### **Predictive Modeling:**

The goal of predictive modeling was to forecast movie revenue based on key attributes such as budget, popularity, and genre.

* Recursive Feature Elimination (RFE) was employed to select the most important features for predicting revenue.
* A Random Forest regression model was trained to predict movie revenue, using the selected features from RFE.
* The dataset was split into training and test sets for model evaluation using metrics such as R² and Mean Absolute Error (MAE).

*Data Visualization*

* Visualizations were created using Matplotlib and Seaborn to display relationships between key features, clustering results, and regression predictions.

Each step was carefully designed to ensure replicability and clarity, enabling a comprehensive analysis of the dataset. These objectives provide a clear roadmap for building and evaluating a movie recommendation system, while also focusing on providing value to users and industry stakeholders.

# Experiments

The following experiments were conducted to analyze movie data, apply machine learning techniques, and build the recommendation system, it also mention about the various models been used in this project:

*Clustering:*

**Objective**: Group movies based on shared features (e.g., genres, budget, popularity).

**Method**: K-Means clustering was applied to the dataset.

**Details**: K-Means is an unsupervised learning algorithm that partitions the dataset into a predefined number of clusters, helping to identify distinct movie categories.

**Outcome**: Movies were categorized into distinct clusters, helping to identify patterns in movie types.

*Dimensionality Reduction:*

**Objective**: Reduce the high-dimensional dataset to two components for easier interpretation and visualization.

**Method**: Principal Component Analysis (PCA) reduced the features to two principal components, retaining most of the data's variance.

**Details**: PCA is a linear dimensionality reduction technique that helps visualize high-dimensional data in two or three dimensions for easier interpretation and analysis.

**Outcome**: Key dimensions influencing movie characteristics were identified.

**Details**: t-SNE is a non-linear technique used for visualizing clusters and relationships in data, especially effective for high-dimensional data like movie attributes.

*Predictive Modeling:*

**Objective**: Build a model to predict movie revenue based on attributes like budget, popularity, and genre.

**Method**: A Random Forest regression model was trained on the dataset.

**Details**: Random Forest is an ensemble learning model that builds multiple decision trees and combines their outputs. It is effective for regression tasks due to its robustness to overfitting and ability to handle complex, non-linear relationships.

**Outcome**: The model achieved high predictive accuracy, with R² and Mean Absolute Error (MAE) used as evaluation metrics.

**Feature Importance**: Budget, popularity, and genre were identified as the most significant predictors of movie revenue.

These experiments contributed to understanding movie characteristics, enabling the creation of a recommendation system that takes these factors into account and also extract valuable insights from the movie data, including clustering similar movies, reducing the complexity of the dataset, and predicting financial success based on key features

# Results & Discussion

In this section, we interpret the findings from the experiments conducted, explore interesting trends, and relate them to the initial hypotheses. We will also justify the choice of methods used and discuss potential sources of error.

#### **Clustering Results:**

* **K-Means Clustering** was applied to group movies into distinct categories based on their features like genres, budget, and popularity.
  + The clusters revealed distinct types of movies:
    - **Cluster 1**: High-budget action and adventure films.
    - **Cluster 2**: Low-budget independent films.
    - **Cluster 3**: Family-oriented and animated movies.
    - **Cluster 4**: Drama and documentary genres.
  + **t-SNE Visualization** It further clarified these clusters, showing that movies with similar attributes (e.g., high-budget vs. low-budget) were placed close together. This confirms that clustering effectively grouped movies by similar characteristics, making it a useful technique for identifying movie types for recommendations.

#### **Dimensionality Reduction Results:**

* **PCA** and **t-SNE** were used for dimensionality reduction to explore and visualize the relationships within the dataset.
  + **PCA** reduced the dataset’s dimensionality from 85 features to just two, retaining **49.35% of the variance**. The two principal components captured the most important factors influencing movie characteristics, such as budget and genre.
  + **t-SNE** provided an even clearer 2D visualization of the data, helping to identify movie groupings more accurately. The non-linear relationships uncovered by t-SNE showed that movies within the same cluster often had similar patterns, which further supports the clustering results.

#### **Unsupervised Learning:**

* **TF – IDF Vectorizer (Term Frequency - inverse document frequency)** was applied to predict movie revenue based on various features like budget, popularity, and genre.
* The overview column is vectorized using **TF-IDF** to convert textual data into numerical form for clustering.
* **Optimal Clusters Identified:** Clustering analysis determined an optimal number of clusters (optimal\_k), improving recommendations.
* **Silhouette Score Evaluation:** Silhouette scores validated the effectiveness of clustering, with a reported score like: ***Silhouette Score: 0.7521.***

#### **Discussion of Results:**

* The **K-Means clustering** and **dimensionality reduction** techniques confirmed that movies can be effectively grouped based on shared characteristics, such as genre and budget. This provides valuable insights for building personalized recommendation systems where users can be shown movies similar to those they have enjoyed in the past.

#### **Potential Sources of Error**

* **Data Quality**: One potential source of error is the quality of the data. Although the dataset is comprehensive, missing or inaccurate values in certain fields (such as revenue or genre classification) could skew the results, especially in the regression model.
* **Clustering Algorithm Limitations**: The K-Means algorithm assumes that clusters are spherical and equally sized, which may not always hold true in real-world data. Therefore, some of the clusters might not be perfectly representative of the underlying movie types.
* **Model Overfitting**: While the Random Forest model performed well, there is always a risk of overfitting, especially with a high number of features. Even though we used feature importance to identify key predictors, further fine-tuning or regularization techniques could improve the model's generalizability.

This project focuses more on **clustering for recommendation systems** rather than predictive modeling. However, some sections (like the use of RandomForestRegressor or LinearRegression, mentioned earlier) might involve predictive modeling for other tasks.

**5 Visualization Tools Used in the Project**

In this project, several visualization tools were utilized to explore the data, analyze trends, and interpret the results of machine learning models. Below are the key visualization tools, along with the columns or aspects of the dataset they were applied to:

#### **1. Matplotlib:**

**Feature Importance Plot**: Visualized the importance of features like budget, popularity, and runtime in predicting revenue.

**Revenue Distribution**: Histogram to understand the spread of revenue across movies.

#### **2. Seaborn**

**Correlation Heatmap**: Highlighted relationships between features like budget, popularity, runtime, and revenue.

**Pairplot**: Showed trends between numerical features such as budget and revenue.

**Boxplot**: Analyzed revenue distribution and outliers across different genres.

#### **3. t-SNE (t-Distributed Stochastic Neighbor Embedding)**

Visualized clusters of movies based on features like budget, genres, and popularity in 2D space.

#### **4. PCA (Principal Component Analysis)**

Reduced 85 features (e.g., budget, runtime, vote\_average) to two components, visualized via scatter plot to uncover patterns and groupings.

#### **5. Language Distribution Visualizations**

**Bar/Pie Charts**: Showed movie counts by original\_language, highlighting dominant languages.

**Stacked Bar Chart/Heatmap**: Displayed encoded language features across the dataset.

#### **6. Scatter Plots**

**Revenue vs. Budget**: Explored if higher budgets lead to higher revenue.

**Popularity vs. Revenue**: Examined the relationship between popularity and revenue.

These visualization tools provided both high-level overviews and in-depth analysis, helping to identify key trends, correlations, and relationships in the movie dataset. Each tool was used to highlight specific features, such as **budget**, **genres**, and **revenue**, and contributed to building a comprehensive understanding of the data.

# Conclusion and Future Work

This project provides valuable insights into the factors driving movie success, focusing on attributes like budget, popularity, and genres. By applying clustering, dimensionality reduction, and predictive modeling, we identified key factors that influence financial performance. The successful use of machine learning, including K-Means clustering and Random Forest regression, demonstrates the power of data science to predict movie revenue accurately (R² = 0.72). These findings suggest that data-driven approaches can revolutionize movie production, marketing, and recommendations, helping the film industry optimize resources, maximize revenue, and enhance audience engagement.

*Future Studies:*

* **Incorporating User Preferences**: Future research could focus on integrating user-specific data (e.g., viewing history, ratings) into the recommendation system to make even more personalized recommendations. Understanding how individual preferences impact movie success would provide more accurate predictions and enhance the user experience.
* **Deep Learning Models**: Another area for future work could involve exploring deep learning techniques, such as neural networks, to capture more complex patterns in the data. These models might provide better accuracy and handle larger datasets, allowing for more sophisticated recommendations.
* **Inclusion of Social Media and Reviews**: Future studies could incorporate data from social media platforms and user reviews to analyze public sentiment, which has become a key driver of movie success in the digital age. Combining this with the existing features could enhance the accuracy of revenue predictions and recommendations.

By building on these ideas, future research can further refine and expand upon the findings from this study, contributing to more effective strategies in the movie industry.