Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

Phase-2 Submission

Student Name: Balaji G.

Register Number: 712523205011

Institution: PPG institute of technology

Department: B tech information technology

Date of Submission: 08/05/2025

Github Repository Link:https://github.com/balaji712523/NM_BALAJI_DS

1. Problem Statement

In an era dominated by digital streaming, users face an overwhelming number of movie choices. Navigating through this content without assistance often leads to choice fatigue and suboptimal viewing experiences. This project aims to solve this by creating an AI-powered movie recommendation system that uses collaborative and content-based filtering to provide personalized movie suggestions tailored to individual user preferences.

- ***Problem Type**: Recommendation (ranking), with elements of regression (rating prediction) and clustering (user segmentation).*
- ***Impact**: Enhances user satisfaction, engagement, and retention on movie streaming platforms by helping users discover content they are most likely to enjoy.*

2. Project Objectives

- *To build a machine learning-based movie recommendation engine.*
- *To apply collaborative filtering and content-based filtering algorithms.*
- *To compare performance across models such as KNN, SVD, and Matrix Factorization.*

Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

- *To improve accuracy through user and movie feature engineering.*
- *To produce explainable and real-time movie suggestions.*

3. Flowchart of the Project Workflow

4. Data Description

- ***Dataset**: MovieLens 100K*
- ***Source**: https://grouplens.org/datasets/movielens/*
- ***Type**: Structured (ratings.csv, movies.csv, users.csv)*
- ***Records**: 100,000 ratings from 943 users on 1682 movies*
- ***Target Variable**: rating (1–5 scale)*
- ***Static/Dynamic**: Static*

5. Data Preprocessing

- *Merged ratings, movies, and users data.*
- *Removed duplicates and handled missing values.*
- *Converted genres into multi-hot encoded features.*
- *Encoded user IDs and movie IDs.*
- *Normalized rating scale if needed.*
- *Transformed data for input into matrix factorization models.*

userId

- ***Type**: Integer*
- ***Description**: A unique identifier for each user in the dataset.*
- ***Properties**:*
  - *Represents an **anonymous user**; personal data is not included.*

Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

- *IDs are consistent across the dataset and used to **link** ratings with other user-specific data.*
- *Important for **collaborative filtering** models that rely on user similarity.*

movieId

- ***Type**: Integer*
- ***Description**: A unique identifier for each movie.*
- ***Properties**:*
  - *Used to join with movies.csv to retrieve movie titles and genres.*
  - *Helps define the **item space** in the recommendation system.*
  - *Supports **content-based filtering** (by connecting with movie metadata).*

rating

- ***Type**: Float (usually in 0.5 increments, from 0.5 to 5.0)*
- ***Description**: The score given by a user to a movie.*
- ***Properties**:*
  - ***Target variable** in rating prediction tasks.*
  - *Reflects user **preference** (higher ratings indicate greater liking).*
  - *Can be used to:*
    - *Predict future ratings (regression task),*
    - *Rank movies for recommendation (top-N),*
    - *Segment users (clustering).*

timestamp

- ***Type**: Integer (Unix time format — seconds since Jan 1, 1970)*
- ***Description**: The time when the rating was recorded.*
- ***Properties**:*

Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

- *Useful for **time-based analysis**, like:*
  - *Detecting trends over time,*
  - *Implementing **time decay** (recent ratings get more weight),*
  - *Splitting data chronologically for **train/test**.*
- *Can be converted into readable formats (e.g., using Python's datetime module).*

6. Exploratory Data Analysis (EDA)

***Univariate Analysis**:*

- *The document details the distribution of ratings (with a note that most ratings are either 4 or 5).*
- *It identifies popular genres and frequently rated movies.*
- *Top genres and most-rated movies*

***Bivariate/Multivariate**:*

- *Correlation heatmaps between user preferences and genres*
- *Average rating per user and per genre*
- *The report includes correlation analyses (heatmaps) to study the relationship between user preferences and movie genres.*
- *It also examines average ratings per user and per genre.*

Key Insights:

- *A small set of "power users" contribute the majority of ratings.*
- *Genres such as romance and action tend to have higher ratings.*
- *User preferences vary, suggesting segmentation could improve recommendation accuracy.*
- *Genre preferences vary significantly by user clusters.*

7. Feature Engineering

- *Genre embeddings are created to capture movie characteristics.*
- *User profile vectors are developed based on historical movie preferences.*

Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

- *Principal Component Analysis (PCA) is applied to the user-genre matrix for dimensionality reduction.*
- *Temporal features (such as time decay on ratings) are engineered to capture time-related behaviors in ratings.*
- *Created genre embeddings.*
- *Built user profile vectors based on past preferences.*
- *Applied PCA on user-genre matrix for dimensionality reduction.*
- *Engineered temporal features (e.g., rating time decay).*

8. Model Building

- ***Models Implemented:***
- ***KNN (Collaborative Filtering):*** *This neighborhood-based approach leverages similarities between users.*
- ***SVD (Matrix Factorization):*** *A latent factor model to handle sparse data and extract underlying patterns.*
- ***Content-Based Filtering:*** *Uses movie genres, tags, and historical data to recommend similar movies.*
- ***Training Process:***
- *Data is split into training (80%) and testing (20%) sets.*
- *Performance is evaluated using metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), as well as precision and recall at top-K recommendations.*

9. Visualization of Results & Model Insights

Performance Charts:

- *A comparison chart of RMSE across different models is mentioned.*
- *Bar plots are used to display Precision/Recall at K, illustrating the quality of top-N recommendations.*

Confusion Matrix*:*

Created by Xodo PDF Studio. Learn more at xodo.com/pdf-studio

- ***RMSE (Root Mean Squared Error)*** *and* ***MAE (Mean Absolute Error)*** *— for evaluating rating predictions.*
- ***Precision@k*** *and* ***Recall@k*** *— for evaluating the quality of top-N recommendations.*
- ***Ranking correlation metrics*** *(like NDCG or MAP) — for assessing the relevance order of recommendations.*
- *Not applicable for rating regression*

Feature Importance*:*

- *Genre influence on user preferences*
- *Visualizations (such as plots) highlight which features (especially genres) most impact user preferences.*

10. Tools and Technologies Used

Programming and Development:

- *Language: Python.*
- *IDE: Jupyter Notebook.*
- *Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, Surprise (for recommender systems), and LightFM.*

Data Visualization:

- *Tools like seaborn, matplotlib, and Plotly are used for creating insightful visualizations.*

11. Team Members and Contributions

S NO	***TEAM MEMBERS***	***CONTRIBUTIONS***
1.	***Kaviya Bharathi K.***	*Data preprocessing*
2.	Manjima M.	*EDA*, Evaluation
3.	Balaji G.	*Feature engineering*
4.	Vishnu Kumar K.	*modeling*
5.	Dhinesh Kumar B.	*documentation*