MovieLens Dataset Analysis Report

A Report submitted for COGNITIVE COMPUTING-UCS420 by

Yuvraj 102317204

Shivansh 102317207

Garvish 102317190

1. Introduction & Problem Statement

In the modern digital era, personalized recommendations play a crucial role in enhancing user experience. Movie recommendation systems help users discover films that match their preferences based on past interactions. This project aims to analyze the MovieLens dataset to extract meaningful insights and implement a basic machine learning model to predict whether a user will like a movie based on its rating and release year.

2. Dataset Overview

The dataset used in this project is the **MovieLens dataset**, which contains:

- `movies.csv` – Movie IDs, titles, and genres.
- `ratings.csv` – User ratings for movies (scale: 0.5 to 5).
- `tags.csv` – User-assigned movie tags.

Key characteristics:

- Large collection of movies with metadata.

- User-generated ratings allowing sentiment-based analysis.
- Genres available for classification and filtering.

3. Technology Stack

The project utilizes the following technologies and libraries:

- **Programming Language:** Python
- **Libraries Used:**
    - `pandas`, `numpy` – Data manipulation
    - `matplotlib`, `seaborn`, `wordcloud` – Data visualization
    - `scikit-learn` – Machine learning (Random Forest Classifier)

4. ML Model Implementation & Evaluation

A **Random Forest Classifier** was used to predict whether a user liked a movie (rating > 3). The implementation steps included:

- **Feature Engineering:** Extracting the movie release year and rating as input features.
- **Data Preprocessing:** Handling missing values, converting text to numeric, and splitting data into training/testing sets.
- **Hyperparameter Tuning:** Using `RandomizedSearchCV` to optimize the model parameters.
- **Evaluation Metrics:** Accuracy, classification report, and confusion matrix were used to assess model performance.

5. Results & Insights (Visualizations & Metrics)

- **Most common movie genres:** Visualized using a word cloud and bar chart.
- **High and low-rated movies:** Identified based on rating distribution.
- **ML Model Performance:**
    - Confusion matrix showed model performance in classifying liked vs. not liked movies.

    - Feature importance analysis indicated **rating** had the highest impact on predictions.

6. Challenges & Future Improvements

- **Challenges Faced:**
    - Handling missing values in the dataset.
    - Balancing the dataset for better classification.
- **Future Enhancements:**
    - Incorporating additional features (e.g., genre, user interactions).
    - Testing different ML models (e.g., Neural Networks, SVM).

7. Conclusion & Learnings

This project provided insights into data analysis, visualization, and machine learning model development. The MovieLens dataset offers valuable information for recommendation systems, and further improvements can be made by incorporating more features and refining model accuracy.

8. References

- MovieLens Dataset: https://grouplens.org/datasets/movielens/
- Scikit-learn Documentation: https://scikit-learn.org/