

**AIML Project Report**

**on**

Movie Recommendations System

Submitted by:

Hitesh Kumar 21CSU389

Utkarsh Sehrawat 21CSU398

Vishu 21CSU370

### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SCHOOL OF ENGINEERING AND TECHNOLOGY

### THE NORTHCAP UNIVERSITY GURUGRAM, HARYANA

**TABLE OF CONTENT**

[Introduction 5](#_TOC_250011)

[Problem Statement 5](#_TOC_250010)

[Implementation 6](#_TOC_250009)

[Data Collection 6](#_TOC_250008)

[Exploratory Data Analysis 8](#_TOC_250007)

[Machine Learning Model 8](#_TOC_250006)

[User Interface 9](#_TOC_250005)

[Data and Model Analytics 10](#_TOC_250004)

[Data Correlation 10](#_TOC_250003)

[Model Insights 11](#_TOC_250002)

[Conclusion 12](#_TOC_250001)

# INTRODUCTION

## PROBLEM STATEMENT

In the digital era, with an overwhelming amount of movie choices available across various streaming platforms, the need for an effective and personalized movie recommendation system has become crucial. The challenge lies in providing users with tailored movie suggestions that align with their individual preferences and viewing history. The existing movie recommendation systems often lack the ability to accurately predict users' tastes, leading to suboptimal user experiences and limited engagement.

This project aims to develop an advanced movie recommendation system that leverages cutting-edge machine learning techniques to address these shortcomings. The primary objective is to create a platform that can analyze users' historical movie preferences, genres, directors, actors, and viewing patterns to predict their future preferences accurately. The system will consider both explicit user feedback (ratings, reviews) and implicit behavior (viewing history, time spent on movies) to generate recommendations.

Key challenges include designing and implementing robust algorithms capable of handling sparse data, addressing the cold start problem for new users, and ensuring scalability to accommodate a growing user base. Ethical considerations such as user privacy and data security will also be prioritized throughout the development process.

The aim of our movie recommendation system is to provide personalized and engaging film suggestions to users. Leveraging advanced algorithms and user preferences, we strive to curate a tailored selection of movies that cater to individual tastes, enhancing their entertainment experience. By analyzing viewing history, genre preferences, and ratings, our system aims to offer diverse and relevant movie choices, making it easier for users to discover new favorites and enjoy a seamless cinematic journey.

# IMPLEMENTATION

## DATA COLLECTION

## The GroupLens dataset is a well-known resource in the field of recommendation systems. It contains various types of data collected from users of the MovieLens website, where users rate and review movies. This data is valuable for building movie recommendation systems. Here's an explanation of the data collected:

## **1.** **User Data**: This includes information about users such as their IDs, demographic details (age, gender, occupation), and potentially some personal characteristics. This data helps in understanding user preferences and tailoring recommendations to their profiles.

## **2.** **Movie Data**: This contains information about movies in the dataset, including their IDs, titles, release years, and genres. This helps in categorizing movies into genres and providing context for recommendations.

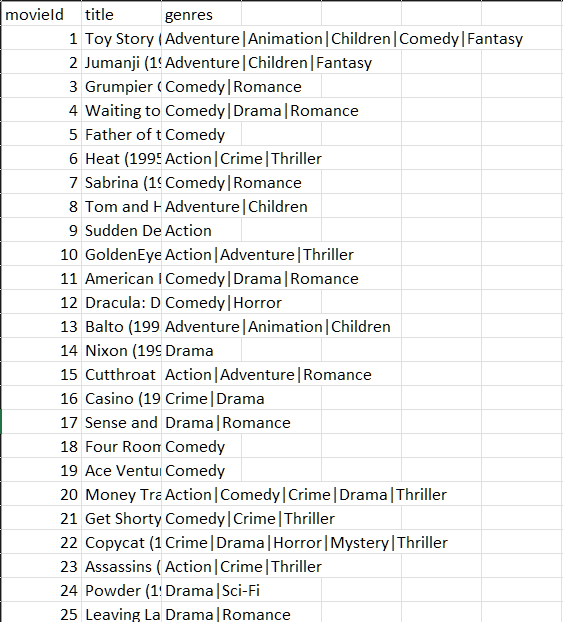
## **3.** **Ratings Data**: Users rate movies on a numerical scale, usually from 1 to 5. The ratings data links users to the movies they have rated and provides a quantitative measure of how much a user liked or disliked a movie. This forms the basis for personalized recommendations.

## **4. Timestamps**: Timestamps are often included to indicate when a user provided a rating or review. This temporal information can be useful for analyzing trends and potentially factoring in recency in recommendations.

## **5. Reviews and Textual Data**: Some datasets might include textual reviews or comments written by users about movies. This additional information can be used to understand user sentiment and preferences in more detail.

## By leveraging this data from GroupLens, recommendation systems can employ collaborative filtering techniques to find users with similar tastes and suggest movies that like-minded users enjoyed. Additionally, content-based filtering can be used by analyzing the characteristics and genres of movies that a user has shown interest in. Hybrid approaches combining both methods can provide even more accurate and diverse recommendations.

It's important to note that while the GroupLens dataset is widely used and has contributed to the development of recommendation systems, there are newer and more extensive datasets available now, which might lead to even better performance and insights for movie recommendations.



A screenshot of a table

Description automatically generated

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis or as abbreviated as EDA is essential to get a quick visual insight of the data. Maintaining its reputation, it will also be used by us to have a quick insight on our data. We have used various graphs, charts and algebraic functions provided by various python libraries to have a quick insight on our data.To read the data from CSV and convert it into a workable dataframe, we have used Pandas library. Further, we have applied various algebraic as well as aggregate methods on our data to get a quick insight about it. Now to get a visual representation of our data, we have used plotting libraries like Matplotlib.

This EDA has helped us to get a quick insight about our data and learn more about the user’s music taste. Moving forward, the EDA will also be used in the web page for users to have an insight about the model.

MACHINE LEARNING MODEL

As we can see, our extracted dataset is clearly classified and sorted. Hence moving forward, we will be applying “Content Analysis” in order to forecast whether the genre which user will be inputting has been watched by the user before or not. As we can see with our data - our prediction will have only 2 possible outcomes, wether the person have liked the genre before or not (1 or 0). This will be determined upon set of independent variables provided in the dataset. Hence, it is wise to use “Logistic Regression” for our model to get optimal accuracy.

We have used “Nearest neighbors” provided by Sklearn to use various methods in order to proceed with our program. It has helped us to train our model and make further predictions using the trained model seamlessly.

# DATA AND MODEL ANALYTICS

DATA CORRELATION

A graph of blue bars

Description automatically generated with medium confidence

This bar graph represents how many times a rating was used to rate movies.

A graph of different movies

Description automatically generated with medium confidence

This graph represents all genres and the movies in it

A graph with numbers and lines

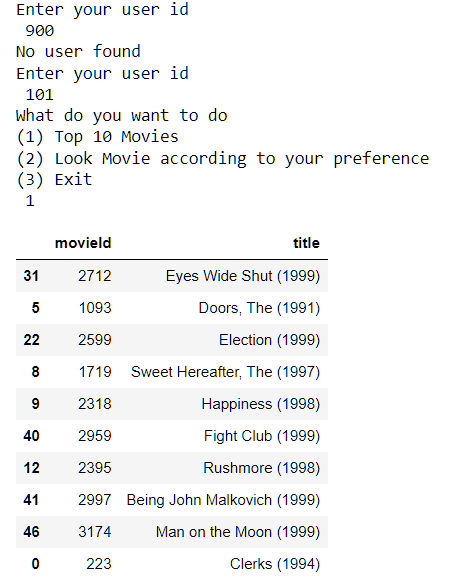
Description automatically generated.

This shows our top 10 users who have watched the maximum movies

# CONCLUSION

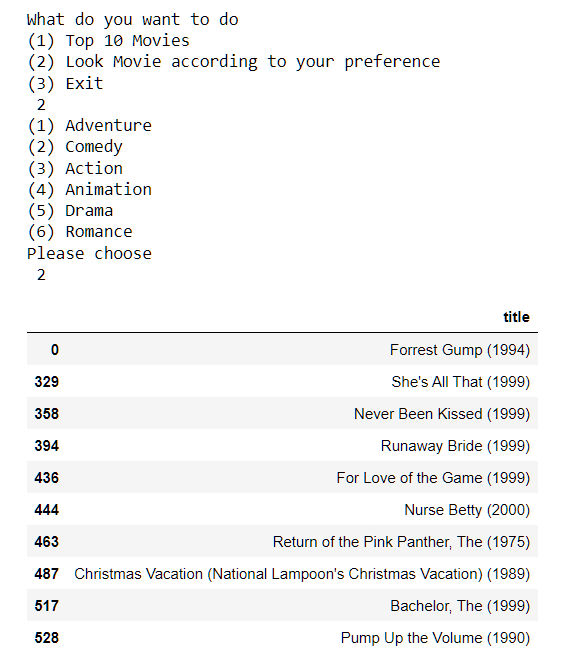
Top 10 Movies

When the user selects Top 10 Movies, he/she can see their top 10 movies rated by them. This helps the user to rewatch their old favorite movies.



Movie Search

Once the user selects to the watch movie according to specific genre then the program takes him/her to another area in which the program asks them which genre will they prefer. According to the users choice the program will provide them movies.



# 