Movie Watch Pattern Clustering

Author: Amit Yadav

Roll No: 202401100300034

Branch: CSE AI Section: A

Submitted to : Mr. Bikki Gupta Sir

College: KIET Group of Institutions

Date: 22-04-2025

Tools Used: Python, Google Colab, Scikit-learn, Pandas, Matplotlib

Introduction

This project focuses on identifying patterns in movie-watching behavior using clustering algorithms. The aim is to group users based on their time of watching movies, genre preferences, and rating behavior. With the proliferation of online streaming platforms, users have developed distinct viewing patterns. Analyzing and clustering these patterns enables the development of personalized content delivery, improved user engagement, and targeted marketing strategies. The outcome of this analysis not only helps businesses optimize their services but also enhances user satisfaction.

Methodology

1. **Dataset:** A movie rating dataset containing information such as user ID, timestamp, genre, and user ratings. A publicly available dataset such as MovieLens was used due to its comprehensive and well-documented nature.
2. **Preprocessing:**
    - Converted timestamps into time-of-day categories (morning, afternoon, evening, night).
    - One-hot encoded movie genres for each user, allowing multiple genres to be represented accurately.
    - Normalized rating scores using StandardScaler to bring all features to a comparable scale.
3. **Feature Engineering:**
    - Grouped and aggregated user data to calculate average rating behavior.
    - Determined the most-watched genres per user.
    - Calculated the proportion of movies watched in each time segment.
4. **Clustering Algorithm:**
    - Used the K-Means clustering algorithm to group users based on their feature vectors.
    - Evaluated the optimal number of clusters using the Elbow Method and silhouette score.
5. **Visualization:**
    - Reduced dimensionality with PCA for 2D plotting of clusters.
    - Applied Seaborn's scatterplot for visual representation of user segments

```
#importing libraries for data manipulation, k-means used for clustering

import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import LabelEncoder



# 1. Data Preparation
df = pd.read_csv("/content/movie_watch.csv")  # Replace with your actual file path
le = LabelEncoder()
df['genre_encoded'] = le.fit_transform(df['genre_preference'])
features = df[['watch_time_hour', 'genre_encoded', 'avg_rating_given']]
# 2. Clustering
kmeans = KMeans(n_clusters=3)  # You can change the number of clusters here
kmeans.fit(features)
labels = kmeans.labels_

# Assign cluster labels to the DataFrame
df['cluster'] = labels
# Assign cluster labels to the DataFrame
df['cluster'] = labels
```

Output/Result

The analysis revealed four distinct user clusters:

1. **Cluster A:** Early morning watchers, with a balanced preference across genres and consistent high ratings.
2. **Cluster B:** Afternoon viewers who prefer documentaries and dramas with moderate rating variance.
3. **Cluster C:** Evening watchers who frequently watch thrillers, sci-fi, and action genres.
4. **Cluster D:** Night-time users who binge-watch comedy and romance with varying rating behaviors.

These insights can be instrumental for streaming platforms to:

- Offer time-based personalized content
- Optimize recommendation engines
- Design user-centric promotional strategies

References/Credits

- MovieLens Dataset: https://grouplens.org/datasets/movielens/
- Scikit-learn documentation: https://scikit-learn.org/
- Matplotlib & Seaborn for visualization
- Google Colab for implementation and testing