Project Title:

Movie Watch Pattern Clustering.

Cluster users based on time of watching, genre preference, and rating behavior.

Project number - 4

Submitted By – Akash

CSE(AI) Sec -A

202401100300024

Submitted to :

DATE – 22/04/2025

Introduction:

In today's world of streaming platforms like Netflix, Amazon Prime, and Disney+, understanding the viewing habits of users is essential for improving recommendations and personalized experiences. By analyzing and clustering movie watch patterns, streaming services can provide better content suggestions and optimize user engagement. This report focuses on clustering users based on three major factors:

1. **Time of Watching**: When users typically watch movies.
2. **Genre Preference**: The types of genres users prefer to watch.
3. **Rating Behavior**: How users rate movies after watching them.

Objective

The goal of this clustering is to group users based on common patterns in the three criteria: **watching time, genre preferences, and rating behaviors**. By doing so, we can identify distinct user groups and tailor recommendations accordingly.

Clustering Approach

To perform clustering, we need to analyze and group users who share similar viewing behaviors. Below is a breakdown of the main factors that will be used for clustering.

1. **Time of Watching**
    - Users have different schedules and might watch movies at different times of the day. For example, some might prefer watching movies at night, while others may prefer morning or afternoon viewing.

- **Data Collection**: The time of the day when each user watches movies can be recorded and categorized into time slots (e.g., morning, afternoon, evening, night).

2. **Genre Preference**
    - People have varying tastes in movie genres. Some may prefer action films, while others may lean toward romantic comedies, thrillers, or documentaries.
    - **Data Collection**: A list of genres watched by each user can be compiled. This data can be represented as a frequency count of genres watched over a specific period.
3. **Rating Behavior**
    - After watching a movie, users often rate it, providing insight into their opinions about the content. Some users rate movies generously, while others may rate more conservatively.
    - **Data Collection**: The average rating each user gives to movies or the frequency of high ratings (e.g., 4-5 stars) vs. low ratings (e.g., 1-2 stars) can be tracked.

Clustering Algorithm

The clustering process involves grouping similar users based on the three attributes: time of watching, genre preferences, and rating behavior. A **K-means clustering algorithm** is commonly used for such tasks because it is efficient and works well with continuous data.

1. **Preprocessing**: Data must be normalized to ensure each feature (time, genre preference, and rating) has an equal influence on the clustering process.
2. **Choosing K**: The number of clusters (K) needs to be determined. Methods like the **elbow method** or **silhouette analysis** can be used to find the optimal number of clusters.
3. **Clustering Execution**: The K-means algorithm will group users into K clusters, where each user is assigned to a group that represents their similar viewing patterns.

Cluster Analysis

Once the clustering is done, we can analyze the resulting clusters to understand the common viewing behaviors in each group.

1. **Cluster 1: Night Owls**

- Users in this group prefer watching movies late at night. They tend to have more relaxed viewing habits, often binge-watching several movies in a row.
- **Genre Preference**: This group may have a diverse range of genre preferences, but they may lean toward genres that are less mainstream, such as horror or drama.
- **Rating Behavior**: Users in this group may rate movies more moderately, giving higher ratings to niche films that are not as widely appreciated.

2. **Cluster 2: Early Birds**
    - Users in this group tend to watch movies in the early morning or during the day.
    - **Genre Preference**: They may prefer lighter genres, such as family films, comedies, or documentaries.
    - **Rating Behavior**: These users may rate movies positively, especially feel-good or light-hearted films.
3. **Cluster 3: Genre-Focused**
    - These users are highly specific in their genre preferences. For example, they might only watch action movies or strictly romantic comedies.
    - **Time of Watching**: They may not have a specific time preference, but their movie selection is strongly influenced by their favorite genre.
    - **Rating Behavior**: This group tends to give ratings based on their genre interest. They might give high ratings to films within their preferred genre and lower ratings to others.
4. **Cluster 4: Casual Watchers**
    - These users watch movies sporadically and do not have a fixed time for watching movies.
    - **Genre Preference**: Their genre preferences may be quite broad, watching a variety of genres without sticking to any one type.
    - **Rating Behavior**: Their rating behavior is unpredictable, often giving a mix of ratings, but generally conservative in their scoring.

Visualization and Insights

Using visualizations such as scatter plots, heat maps, and bar graphs, we can observe the clustering of users. For example, we can plot the average time of watching on the x-

axis, genre preference on the y-axis, and the average rating on the z-axis to visualize the clusters.

1. **Heatmap of Genre Preferences**: A heatmap can show which genres are most popular among each cluster.
2. **Cluster Distribution by Time of Day**: A bar graph can show how the users in different clusters are distributed throughout the day.
3. **Rating Distribution**: A line graph can be used to show the distribution of ratings across the clusters, helping to identify patterns in rating behaviors.

Applications of Movie Watch Pattern Clustering

The insights derived from clustering can be used for various purposes:

1. **Personalized Recommendations**: Streaming platforms can recommend content based on the user's cluster. For instance, a "Night Owl" user could receive recommendations for late-night thrillers or horror films, while an "Early Bird" user might get suggestions for feel-good movies.
2. **Targeted Marketing**: Streaming services can create targeted ads or promotions for specific clusters. For example, a "Genre-Focused" group can be shown trailers for new releases in their favorite genre.
3. **Content Strategy**: Understanding user clusters can help content creators or streaming platforms make decisions about which types of movies to produce or license, based on popular genres or viewing times.
4. **User Engagement**: Streaming platforms can craft engaging campaigns that appeal to users' habits. For instance, special movie marathons for "Night Owls" or exclusive early access screenings for "Early Birds."

Conclusion

Clustering movie watch patterns based on time of watching, genre preference, and rating behavior helps create a more personalized and enjoyable user experience. By understanding the different groups of users, streaming platforms can offer better recommendations, improve user satisfaction, and increase engagement. Moreover, this clustering approach can be applied to optimize content strategies, marketing campaigns, and even content creation, paving the way for a more customized streaming experience for everyone.

CODE:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Sample Data: user_id, timestamp (time of watching), genre, rating
data = {
    'user_id': [1, 2, 1, 3, 2, 3, 1, 2, 3],
    'timestamp': [
        '2025-04-01 18:30', '2025-04-01 20:15', '2025-04-02 19:00',
        '2025-04-02 21:00', '2025-04-02 22:30', '2025-04-03 17:00',
        '2025-04-03 18:45', '2025-04-03 20:00', '2025-04-03 19:30'
    ],
    'genre': ['Action', 'Comedy', 'Action', 'Drama', 'Comedy', 'Action', 'Drama', 'Action',
'Comedy'],
    'rating': [5, 4, 3, 4, 5, 3, 4, 4, 5]
}

# Create DataFrame
df = pd.DataFrame(data)

# Convert timestamp to datetime and extract hour (time of day) and day of week
df['timestamp'] = pd.to_datetime(df['timestamp'])
df['hour'] = df['timestamp'].dt.hour
df['day_of_week'] = df['timestamp'].dt.dayofweek

# For simplicity, we will convert genres into numeric values
```

```
df['genre_numeric'] = df['genre'].map({'Action': 0, 'Comedy': 1, 'Drama': 2})

# Prepare the data for clustering (features: hour, day_of_week, genre_numeric, rating)
features = df[['hour', 'day_of_week', 'genre_numeric', 'rating']]

# Normalize the features using StandardScaler
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Apply KMeans clustering
kmeans = KMeans(n_clusters=3, random_state=42)
df['cluster'] = kmeans.fit_predict(scaled_features)

# Plot the results
plt.scatter(df['hour'], df['rating'], c=df['cluster'], cmap='viridis', s=100)
plt.title('Movie Watch Patterns Clustering')
plt.xlabel('Hour of Day')
plt.ylabel('Rating')
plt.colorbar(label='Cluster')
plt.show()

# Print the data with cluster information
print(df[['user_id', 'hour', 'day_of_week', 'genre', 'rating', 'cluster']])
```