Assessment Report

on

“MOVIE WATCH CLUSTERING”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AI)

By

Name : Mayan Prajapati

Roll Number : 202401100300151

Section: C

Under the supervision of

“MAYANK LAKHOTIA”

KIET Group of Institutions, Ghaziabad

Introduction

The goal of this problem is to identify patterns in user behavior by clustering them based on when they watch movies, the genres they prefer, and how they rate the content. Clustering techniques can help in personalizing content delivery, designing recommendation systems, and understanding viewer trends. This unsupervised learning task uses a real dataset containing user behavior to segment users into meaningful groups.

Methodology

1. Data Preprocessing:
- watch_time_hour: The hour of the day the user watched movies.
- genre_preference: The genre of movie preferred by the user (categorical).
- avg_rating_given: The average rating given by the user.

2. Encoding and Scaling:
- The categorical feature genre_preference was encoded using LabelEncoder.
- All features were standardized using StandardScaler for better clustering performance.

3. Clustering:
- KMeans clustering with n_clusters=3 was applied to segment the users.
- PCA (Principal Component Analysis) was used to reduce the feature space to two dimensions for visualization.

Code

```
import pandas as pd
from sklearn.preprocessing import
StandardScaler, LabelEncoder
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("movie_watch.csv")

label_encoder = LabelEncoder()
df['genre_encoded'] =
label_encoder.fit_transform(df['genre_prefere
nce'])

features = df[['watch_time_hour',
'genre_encoded', 'avg_rating_given']]

scaler = StandardScaler()
scaled_features =
scaler.fit_transform(features)
```

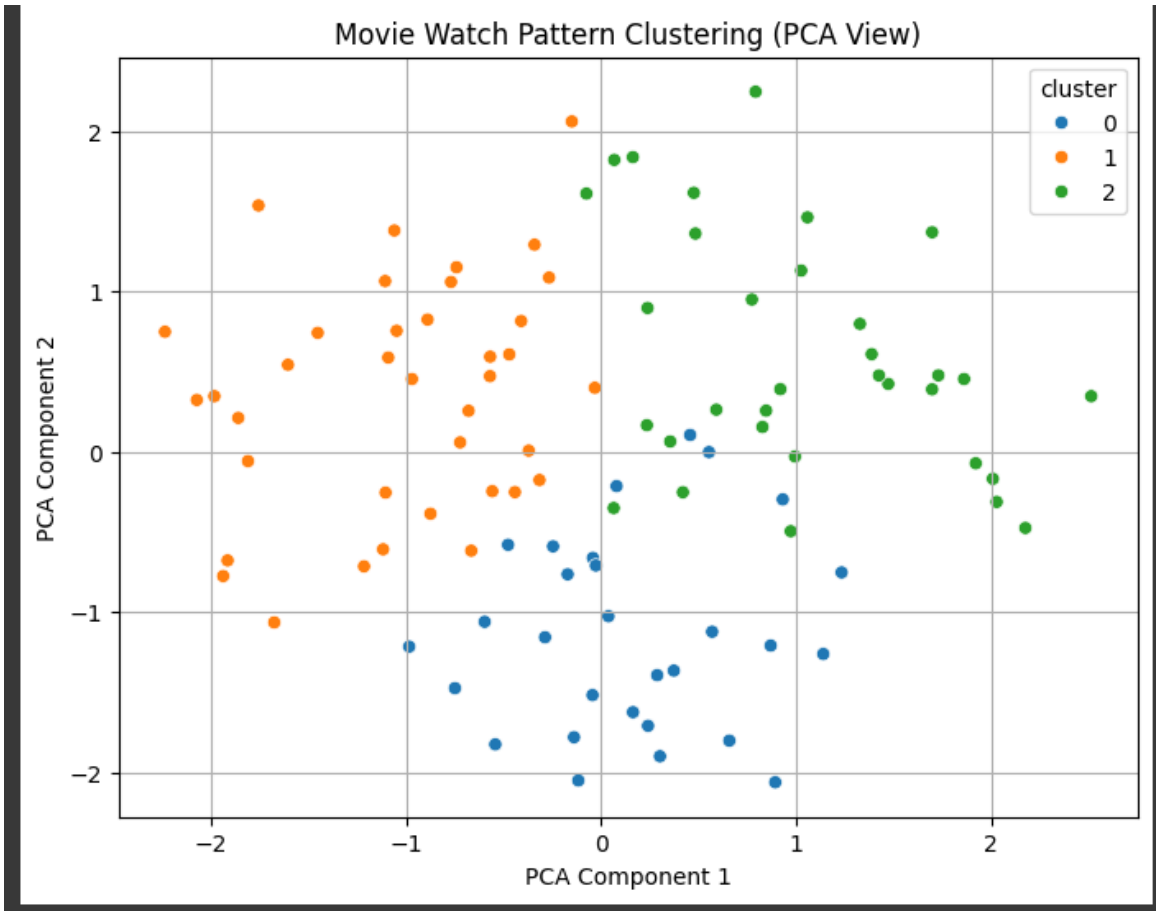
```
kmeans = KMeans(n_clusters=3,
random_state=42)
df['cluster'] =
kmeans.fit_predict(scaled_features)

pca = PCA(n_components=2)
pca_result = pca.fit_transform(scaled_features)
df['pca1'] = pca_result[:, 0]
df['pca2'] = pca_result[:, 1]

plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='pca1', y='pca2',
hue='cluster', palette='tab10')
plt.title('Movie Watch Pattern Clustering (PCA
View)')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Output/Result

A PCA scatter plot showing users clustered into 3 groups based on watch time, genre preference, and rating behavior.



References/Credits

1 **Scikit-learn Documentation**
https://scikit-learn.org/stable/documentation.html
(Used for clustering algorithms, preprocessing, and PCA)

2 **Pandas Documentation**
https://pandas.pydata.org/docs/
(Used for loading, cleaning, and manipulating the dataset)

- Libraries: pandas, scikit-learn, matplotlib, seaborn
- All work done using Google Colab