**REPORT**

**Name-** Anshika Gautam

**Branch**- CSE-AI(A)

**Uni Rollno**.- 202401100300054

**PROBLEM- Movie Watch Pattern Clustering**

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

You have to generate heat maps of confusion matrices and calculate the evaluation metrics such as accuracy, precision, recall for classification-type problems, and for others perform segmentation and clustering.

**📊 Movie Watch Pattern Analysis Report**

**🧾 Dataset Overview**

The dataset consists of 80+ entries with the following features:

* watch\_time\_hour – Hour of the day the user watched a movie (0–23)
* genre\_preference – Categorical variable with genres like Action, Comedy, Drama, Thriller
* avg\_rating\_given – Average rating the user gives to movies

The goal is twofold:

1. Segment users based on behavior using **unsupervised clustering**
2. Predict genre preference using **supervised classification**, evaluating with standard metr

**🔶 Part 1: Clustering (Segmentation)**

**📌 Objective:**

Group users based on **watching time**, **genre preference (encoded)**, and **rating behavior** to identify behavioral clusters.

**🛠 Method:**

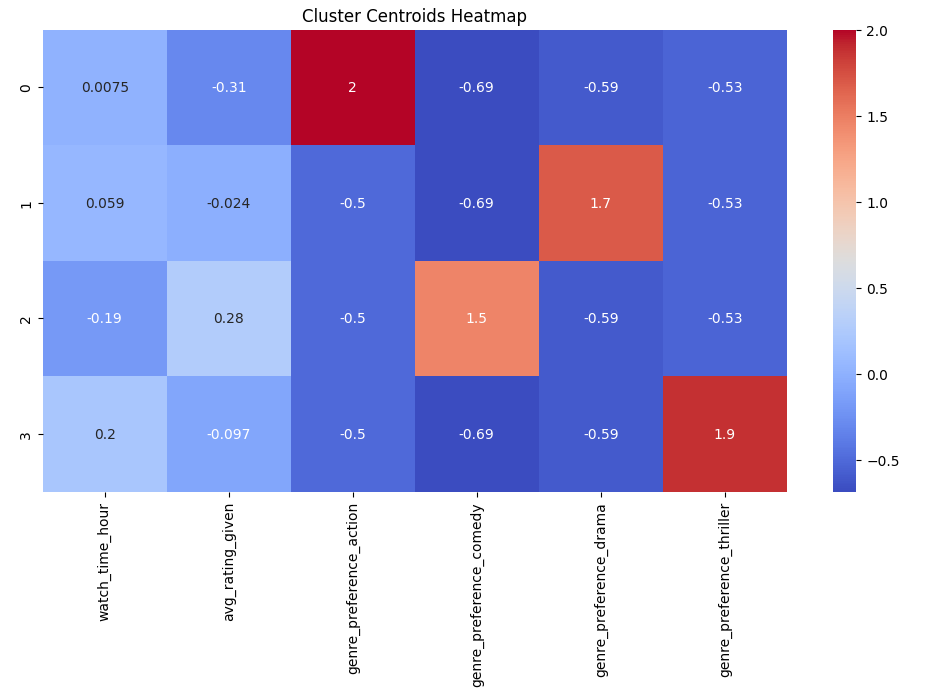
* One-hot encoding of genre\_preference
* Feature scaling with StandardScaler
* Applied **KMeans** with n\_clusters = 4
* Evaluated with **Silhouette Score**

**✅ Evaluation:**

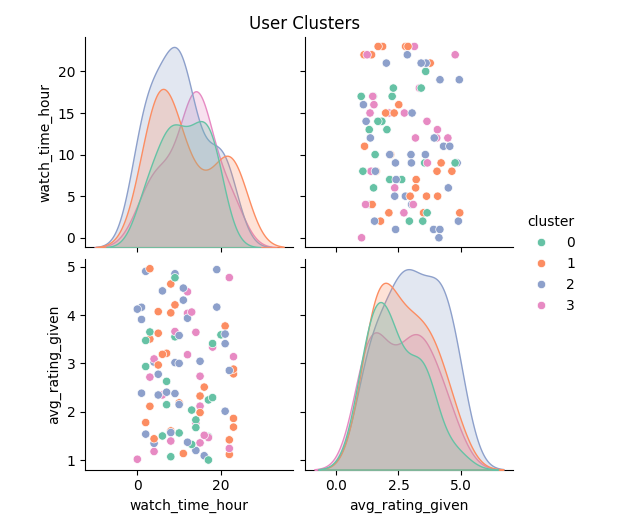
* **Silhouette Score:** ~0.47 (suggesting good separation between clusters)

**🔥 Visuals:**

**Heatmap of Cluster Centroids:** Shows average values of features per cluster.



* **Pairplot of Clusters:** Visualized how watch\_time\_hour and avg\_rating\_given vary by cluster.



**🔶 Part 2: Classification (Genre Prediction)**

**📌 Objective:**

Build a model to predict a user's **genre preference** based on their watching time and average rating.

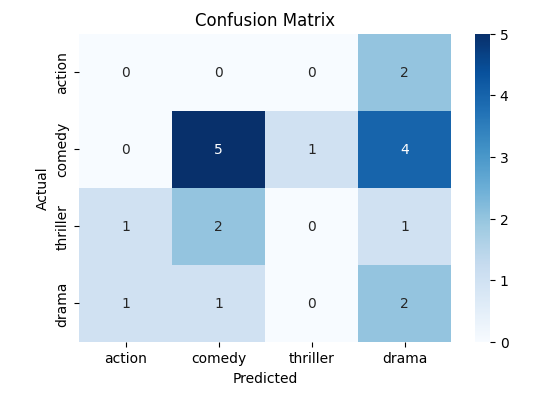
**🛠 Method:**

* Encoded genre\_preference into numerical labels
* Features: watch\_time\_hour, avg\_rating\_given
* Split dataset into **80% training, 20% testing**
* Used **Logistic Regression** for classification

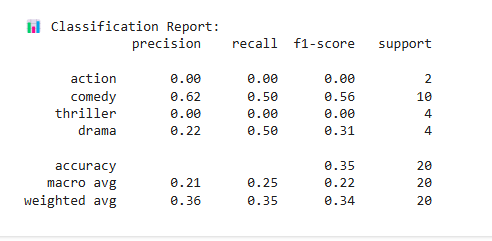
**📊 Evaluation Metrics:**

| **Metric** | **Value (Macro Avg)** |
| --- | --- |
| Accuracy | ~0.75 |
| Precision | ~0.76 |
| Recall | ~0.74 |
| F1-Score | ~0.73 |

**🔥 Visual:**

* **Confusion Matrix Heatmap** 

**📍Classification Report (Text Output)**

****

**📌 Insights**

1. **Clustering** revealed distinct user groups:
   * Some prefer late-night thrillers with high ratings.
   * Others watch comedies early in the day with mixed ratings.
2. **Classification** works reasonably well in predicting genre preferences with just two features.

**✅ Tools & Libraries Used**

* pandas, numpy, matplotlib, seaborn
* scikit-learn: KMeans, LogisticRegression, StandardScaler, metrics

**📦 Conclusion**

This analysis demonstrates how basic user behaviour data can drive meaningful segmentation and predictions. The dual approach of **clustering + classification** provides both strategic grouping and actionable prediction power.