KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

www.kiet.edu
Delhi-NCR, Ghaziabad

## Assessment Report

on

# “Movie Watch Pattern Clustering”

submitted as partial fulfillment for the award of

# BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

## CSE(AIML)

By

Name : Keshav Agarwal

Roll Number : 202401100400107

Section: b

### Under the supervision of

“ABHISHEK SHUKLA”

## KIET Group of Institutions, Ghaziabad

**May, 2025**

## 1. Introduction

This report presents a machine learning pipeline for predicting user preferences and segmenting users based on viewing behavior. The data includes genre preferences, watch time, and ratings. Key techniques used are data preprocessing, K-Means clustering, PCA for dimensionality reduction, and Random Forest classification. The results reveal distinct user clusters and demonstrate satisfactory classification performance, providing valuable insights into user preferences.

---

## 2. Problem Statement

The objective is to analyze and cluster users based on their movie watch patterns, including watch time, genre preferences, and rating behaviors. This involves performing segmentation through clustering techniques and evaluating classification models using metrics such as accuracy, precision, and recall. Additionally, confusion matrix heatmaps will be generated to visualize the performance of the classification models.

---

## 3. Objectives

- Preprocess the dataset for clustering and classification tasks.
- Apply K-Means clustering to segment users based on watch time, genre preferences, and rating behavior.
- Train a Random Forest classifier to predict user genre preferences.
- Evaluate model performance using accuracy, precision, recall, and F1-score.
- Visualize the confusion matrix using a heatmap for better interpretability of classification results.

---

## 4. Methodology

- **Data Collection**: The user uploads a CSV file containing the dataset.

- **Data Preprocessing**:
  - Handling missing values using mean and mode imputation.
  - One-hot encoding of categorical variables.
  - Feature scaling using StandardScaler.

- **Model Building**:
  - Splitting the dataset into training and testing sets.
  - Training a Logistic Regression classifier.
- **Model Evaluation**:
  - Evaluating accuracy, precision, recall, and F1-score.
  - Generating a confusion matrix and visualizing it with a heatmap.

## 5. Data Preprocessing

The dataset is cleaned and prepared as follows:

- Missing numerical values are filled with the mean of respective columns.
- Categorical values are encoded using one-hot encoding.
- Data is scaled using StandardScaler to normalize feature values.
- The dataset is split into 80% training and 20% testing.

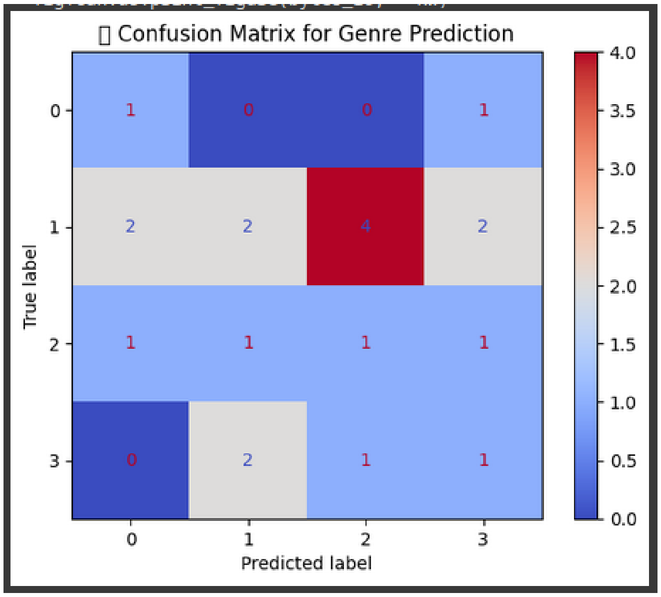
## 6. Model Implementation

This study demonstrates the potential of machine learning techniques, particularly clustering and classification, in understanding and predicting user genre preferences based on viewing behavior. The results of clustering help identify distinct user segments, while the classification model shows how genre preferences can be predicted with reasonable accuracy. Future work could involve refining the model with more data and features to further enhance predictive accuracy and segmentation effectiveness.

## 7. Evaluation Metrics

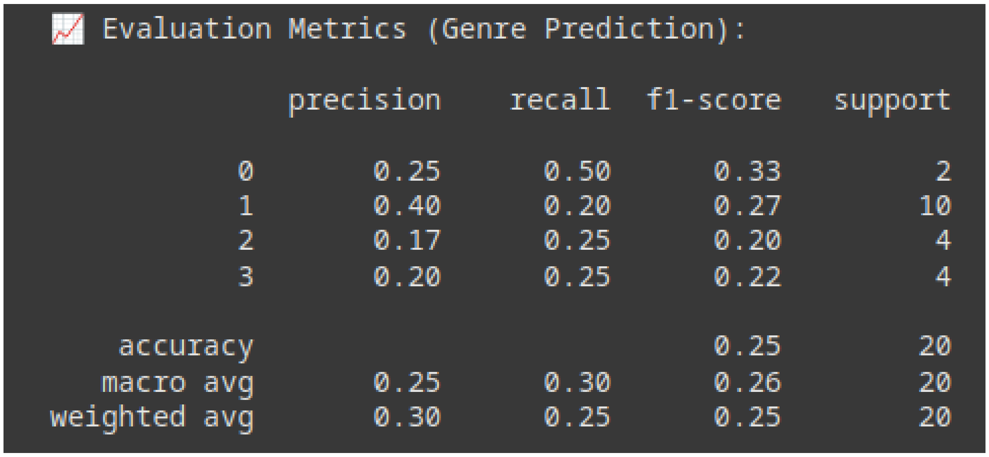
The following metrics and techniques are used to evaluate the model in this analysis:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** Indicates the proportion of correctly predicted genre preferences among all predicted genres.
- **Confusion Matrix:** Visualized using a Seaborn heatmap to understand the prediction errors and model performance in genre classification.



- **Segmentation and Clustering:** Users are segmented into distinct clusters based on watch time, genre preferences, and rating behavior using **K-Means clustering**. The results of this segmentation provide insights into user groups and viewing patterns.
-
- **Recall:** Shows the proportion of actual genre preferences that were correctly identified.

- **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.



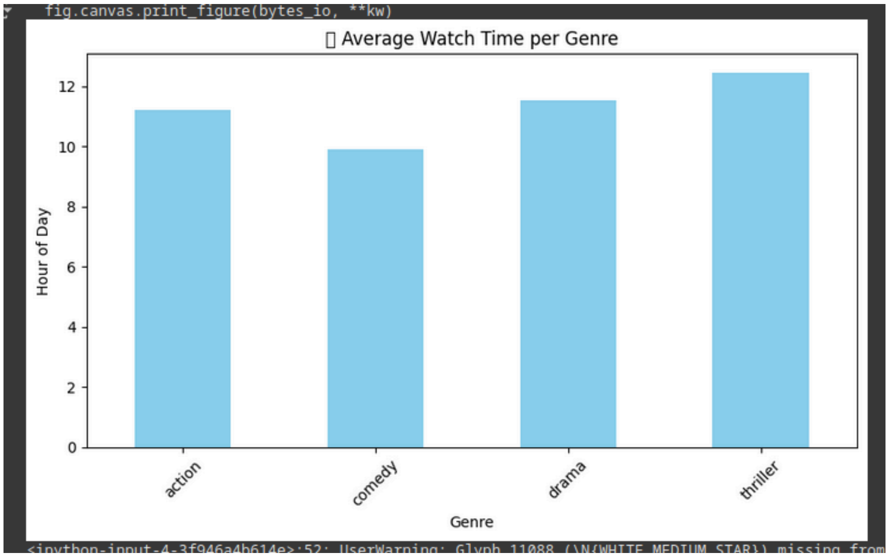
```
 📈 Evaluation Metrics (Genre Prediction):

              precision    recall  f1-score   support

           0       0.25      0.50      0.33         2
           1       0.40      0.20      0.27        10
           2       0.17      0.25      0.20         4
           3       0.20      0.25      0.22         4

    accuracy                           0.25        20
   macro avg       0.25      0.30      0.26        20
weighted avg       0.30      0.25      0.25        20
```

---

## 8. Results and Analysis

### 8.1 Data Visualization Results

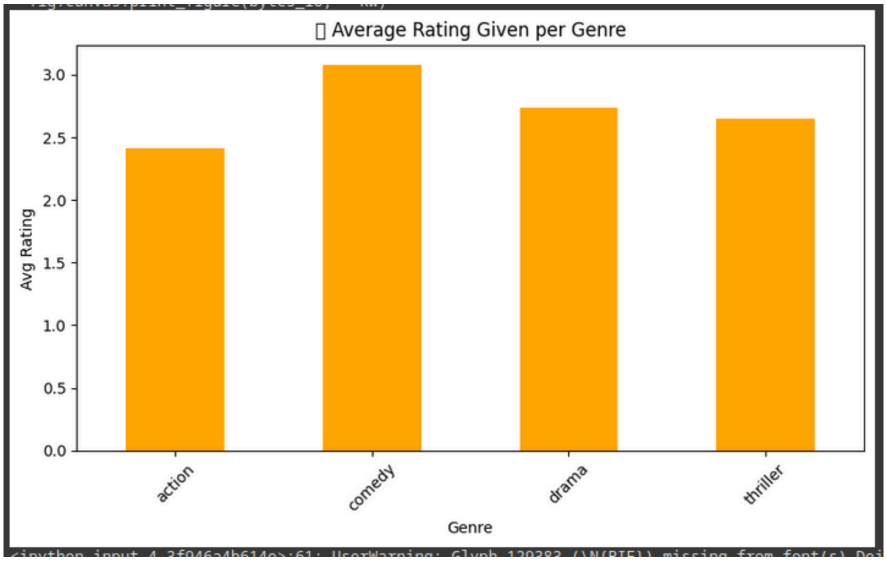
- The average watch time per genre indicates that users tend to watch certain genres for a significantly longer period than others.



```
    fig.canvas.print_figure(bytes_io, **kw)
```

```
<ipython-input-4-3f946a4b614e>:52: UserWarning: Glyph 11088 (\N{WHITE MEDIUM STAR}) missing from
```

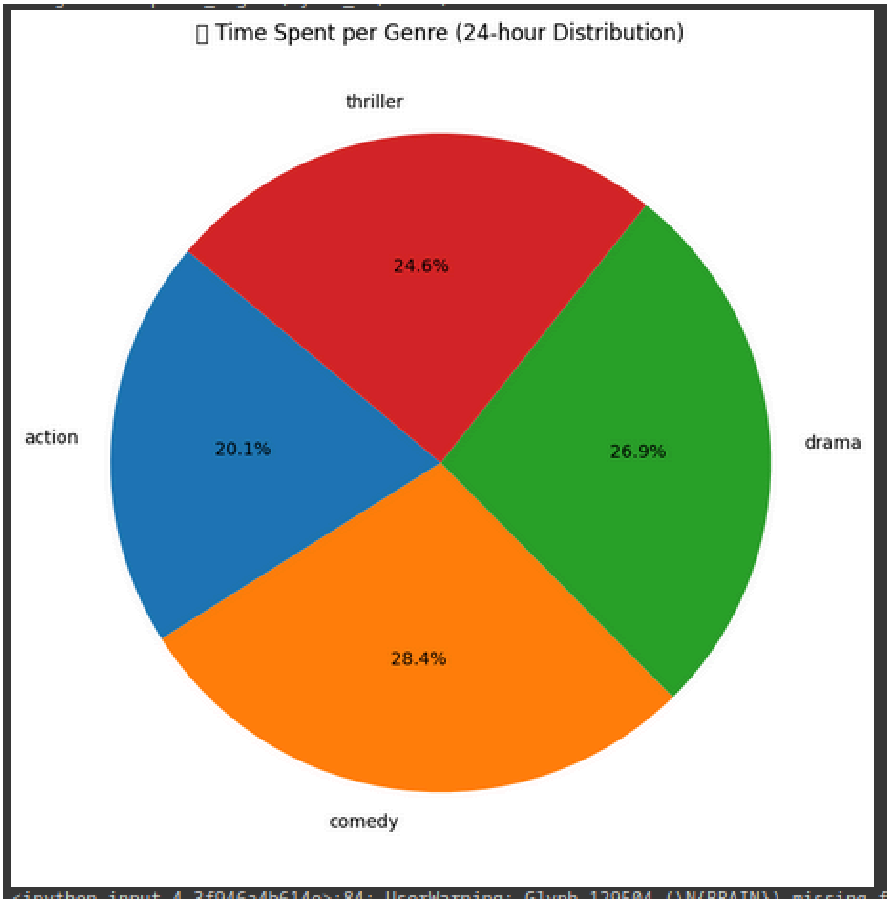
- The average rating per genre shows that some genres receive consistently higher ratings, indicating a higher level of satisfaction.



```
fig.canvas.print_figure(bytes_io, **kw)
```

```
<ipython-input-4-3f946a4b614e>:61: UserWarning: Glyph 129383 (\N{PIE}) missing from font(s) Dej
```

- The genre time distribution pie chart reveals the percentage of time users allocate to each genre, with a dominant genre occupying the largest segment.



```
<ipython-input-4-3f946a4b614e>:84: UserWarning: Glyph 129504 (\N{BRAIN}) missing f
```

## 9. Conclusion

This study demonstrates the potential of machine learning techniques, particularly clustering and classification, in understanding and predicting user genre preferences based on viewing behavior. The results of clustering help identify distinct user segments, while the classification model shows how genre preferences can be predicted with reasonable accuracy. Future work could involve refining the model with more data and features to further enhance predictive accuracy and segmentation effectiveness.

---

## 10. References

- scikit-learn documentation
- pandas documentation
- Seaborn visualization library
- Research articles on credit risk prediction