****

**Assessment Report**

on

**“Movie Watch Pattern Clustering”**

submitted as partial fulfilment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name: Aastha Srivastava

Roll Number: 202401100300004

Section: A

**KIET Group of Institutions, Ghaziabad**

**1. Introduction**

As streaming platforms continue to grow in popularity, understanding user behaviour is essential for enhancing user engagement. One key aspect of user behaviour is the amount of time they spend watching content. Predicting whether a user will watch more than a certain amount of time can help content providers make data-driven decisions, such as recommending content that keeps users engaged for longer periods. This project aims to build a machine learning model that predicts whether a user will watch more than 120 minutes of content based on their activity patterns and ratings.

**2. Problem Statement**

To predict whether a user will watch more than 120 minutes of content, using features such as the user's genre preferences, ratings, and the total time they watch in a given session. This classification will help in identifying users who are likely to engage more deeply with the platform, enabling targeted interventions to retain users.

**3. Objectives**

* Preprocess the dataset for training the machine learning model.
* Train a **Random Forest** classifier to classify whether a user will watch more than 120 minutes.
* Evaluate the model’s performance using accuracy, precision, recall, and classification report.
* Visualize the confusion matrix with a heatmap to better understand the model's predictions.

**4. Methodology**

**Data Collection: A dataset is generated, simulating user data such as time watched, genre preferences, and average ratings.**

**Data Preprocessing**

* One-hot encoding is applied to categorical features (e.g., genre preferences).
* Features are standardized to ensure they are on a similar scale.

**Model Building:** A Random Forest Classifier is used to classify users into two categories: those who watch more than 120 minutes and those who do not.

**Model Evaluation:** We evaluate the model using metrics like accuracy, precision, recall, and a classification report.

**5. Data Preprocessing**

The data is preprocessed as follows:

* One-Hot Encoding: Categorical variable genre\_preference is converted into numerical format using one-hot encoding.
* Feature Standardization: Features such as time\_watched and average\_rating are scaled using StandardScaler to normalize the data.
* Train-Test Split: The dataset is split into training and testing sets using a 80-20 ratio.

**6. Model Implementation**

The model implemented is a **Random Forest Classifier** due to its ability to handle high-dimensional data and perform well with minimal hyperparameter tuning. The classifier is trained on the preprocessed data and used to predict whether a user will watch more than 120 minutes.

**7. Evaluation Metrics**

The following evaluation metrics are used to assess the performance of the model:

* **Accuracy**: Measures the proportion of correct predictions.
* **Precision**: Indicates how many of the predicted high watch-time users were actually high watch-time users.
* **Recall**: Shows the proportion of actual high watch-time users correctly identified by the model.
* **F1-Score**: The harmonic mean of precision and recall.
* **Confusion Matrix**: A heatmap of the confusion matrix provides a visual representation of how well the model predicted the classes.

**8. Results and Analysis**

The model demonstrated reasonable performance, achieving satisfactory accuracy, precision, and recall. The **confusion matrix** heatmap helped to visually assess the model’s ability to correctly classify users into the Low Watch Time and High Watch Time categories.

* **Accuracy**: Measures how often the model predicts the correct class.
* **Precision**: Highlights how accurately the model identified users who watched more than 120 minutes.
* **Recall**: Measures how many of the actual high-watch-time users were successfully predicted.

The **classification report** provides a more detailed view of performance, showing precision, recall, and F1-score for both classes (Low Watch Time and High Watch Time).

**9. Conclusion**

The Random Forest model was successful in classifying users based on their likelihood to watch more than 120 minutes of content. The evaluation metrics and confusion matrix heatmap indicate that the model performs reasonably well in distinguishing between high and low watch-time users. However, further improvements can be made by exploring more advanced models and ensuring that the data is well-balanced for both classes.

This project demonstrates the potential for machine learning to enhance user engagement strategies on streaming platforms, offering content providers the ability to make data-driven decisions.

**10. References**

* **scikit-learn Documentation**: For machine learning algorithms and model evaluation metrics.
* **Pandas Documentation**: For data handling and manipulation.
* **Seaborn Documentation**: For data visualization (heatmaps and plots).

