Assessment Report

on

"Student Club Participation Prediction"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AI-ML)

By

Name : Ayushmaan

Roll Number : 202401100400070

Section: A

Under the supervision of

“Bikki Kumar Sir”

KIET Group of Institutions, Ghaziabad

1. Introduction

In modern educational environments, student participation in extracurricular activities, such as clubs, plays a significant role in personal growth and the development of social skills. Predicting whether a student will join a club based on factors such as their **interest level** and **available free hours** per week is a task that can be effectively tackled using machine learning models.

This project leverages a **Random Forest Classifier**, a powerful ensemble method, to predict student participation in clubs. By analyzing key features like interest level and available free time, this project aims to provide insights into which factors influence a student's likelihood of joining a club.

The objectives of this project are:

- To build a classification model that predicts whether a student will join a club.
- To evaluate the model’s performance using accuracy, precision, and recall metrics.
- To visualize the results using a confusion matrix heatmap.

2. Problem Statement

Student involvement in extracurricular activities, such as clubs, has been shown to contribute significantly to personal development, networking, and academic success. However, predicting which students are likely to participate in such activities can be a challenging task for educators and club organizers.

3. Methodology

Data Collection: The dataset used in this project, `club_participation mse2.csv`, contains data on students' interest levels, their available free hours per week, and whether they chose to join a club. The data consists of the following columns:

- `interest_level`: A numerical representation of how interested the student is in extracurricular activities (e.g., 1 to 5 scale).
- `free_hours_per_week`: The number of hours a student has free each week.
- `club_participation`: A binary variable indicating whether the student joined the club ('yes' or 'no').

Data Preprocessing: **Categorical Encoding**: The target variable `club_participation` was transformed from text ('yes', 'no') to numeric values (1 for 'yes', 0 for 'no') to allow the model to process it.

 Feature Selection: The features used to predict the target were:

- `interest_level`
- `free_hours_per_week`

 Train-Test Split: The dataset was divided into two parts:

- **Training Set** (80% of data) to train the model.
- **Testing Set** (20% of data) to evaluate model performance.

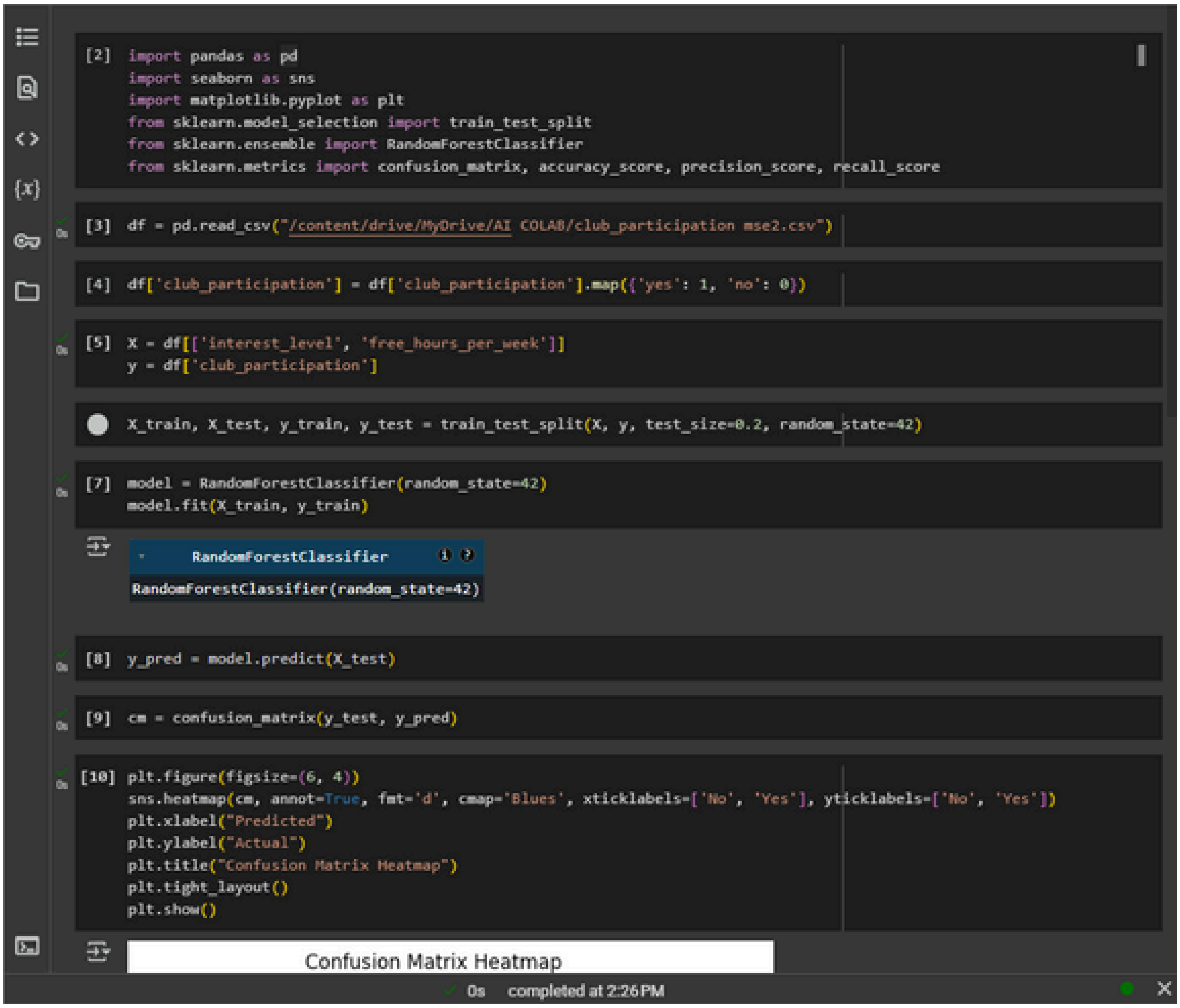
Model Building:A **Random Forest Classifier** was chosen as the model. Random Forest is an ensemble learning method that creates multiple decision trees and merges their outputs to improve prediction accuracy and reduce overfitting. This algorithm is ideal for binary classification tasks like this one.

Model Evaluation: The model's performance was evaluated using:

- **Accuracy**: The percentage of correct predictions made by the model.
- **Precision**: The proportion of predicted "yes" outcomes that were actually correct.
- **Recall**: The proportion of actual "yes" outcomes that the model correctly identified.

Additionally, a **confusion matrix** was generated and visualized as a **heatmap** to better understand the model's performance.

4. Code



```
[2] import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    from sklearn.model_selection import train_test_split
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score
```

```
[3] df = pd.read_csv("/content/drive/MyDrive/AI COLAB/club_participation mse2.csv")
```

```
[4] df['club_participation'] = df['club_participation'].map({'yes': 1, 'no': 0})
```

```
[5] X = df[['interest_level', 'free_hours_per_week']]
    y = df['club_participation']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[7] model = RandomForestClassifier(random_state=42)
    model.fit(X_train, y_train)
```

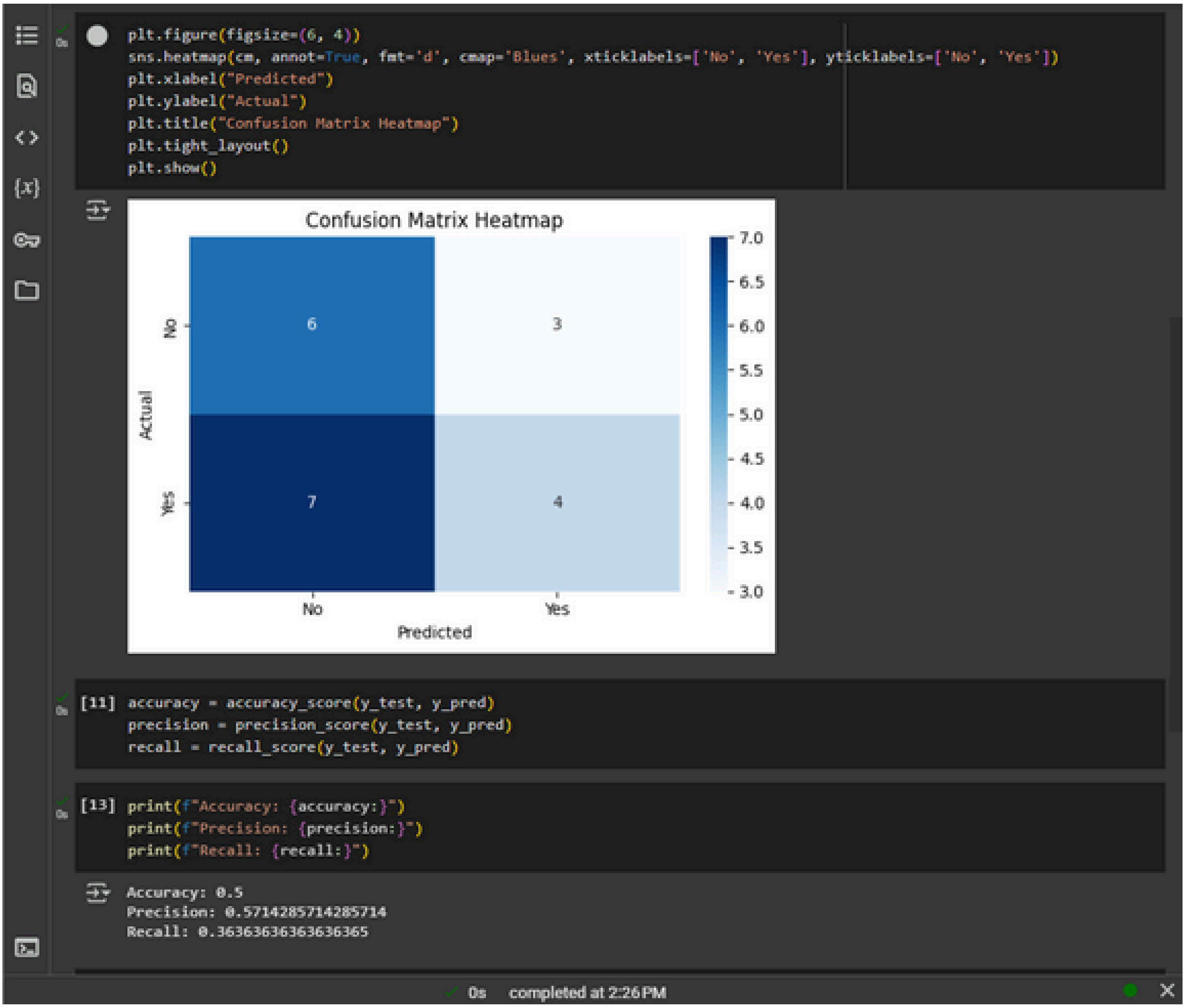
```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
[8] y_pred = model.predict(X_test)
```

```
[9] cm = confusion_matrix(y_test, y_pred)
```

```
[10] plt.figure(figsize=(6, 4))
     sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No', 'Yes'], yticklabels=['No', 'Yes'])
     plt.xlabel("Predicted")
     plt.ylabel("Actual")
     plt.title("Confusion Matrix Heatmap")
     plt.tight_layout()
     plt.show()
```

Confusion Matrix Heatmap



```
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No', 'Yes'], yticklabels=['No', 'Yes'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix Heatmap")
plt.tight_layout()
plt.show()
```

```
[11] accuracy = accuracy_score(y_test, y_pred)
     precision = precision_score(y_test, y_pred)
     recall = recall_score(y_test, y_pred)
```

```
[13] print(f"Accuracy: {accuracy:}")
     print(f"Precision: {precision:}")
     print(f"Recall: {recall:}")
```

```
Accuracy: 0.5
Precision: 0.5714285714285714
Recall: 0.36363636363636365
```

5. Output

Confusion Matrix Heatmap

The following confusion matrix heatmap illustrates the performance of the model:

- **True Positives (TP)**: Correctly predicted "yes" outcomes.
- **False Positives (FP)**: Incorrectly predicted "yes" outcomes.
- **True Negatives (TN)**: Correctly predicted "no" outcomes.
- **False Negatives (FN)**: Incorrectly predicted "no" outcomes.

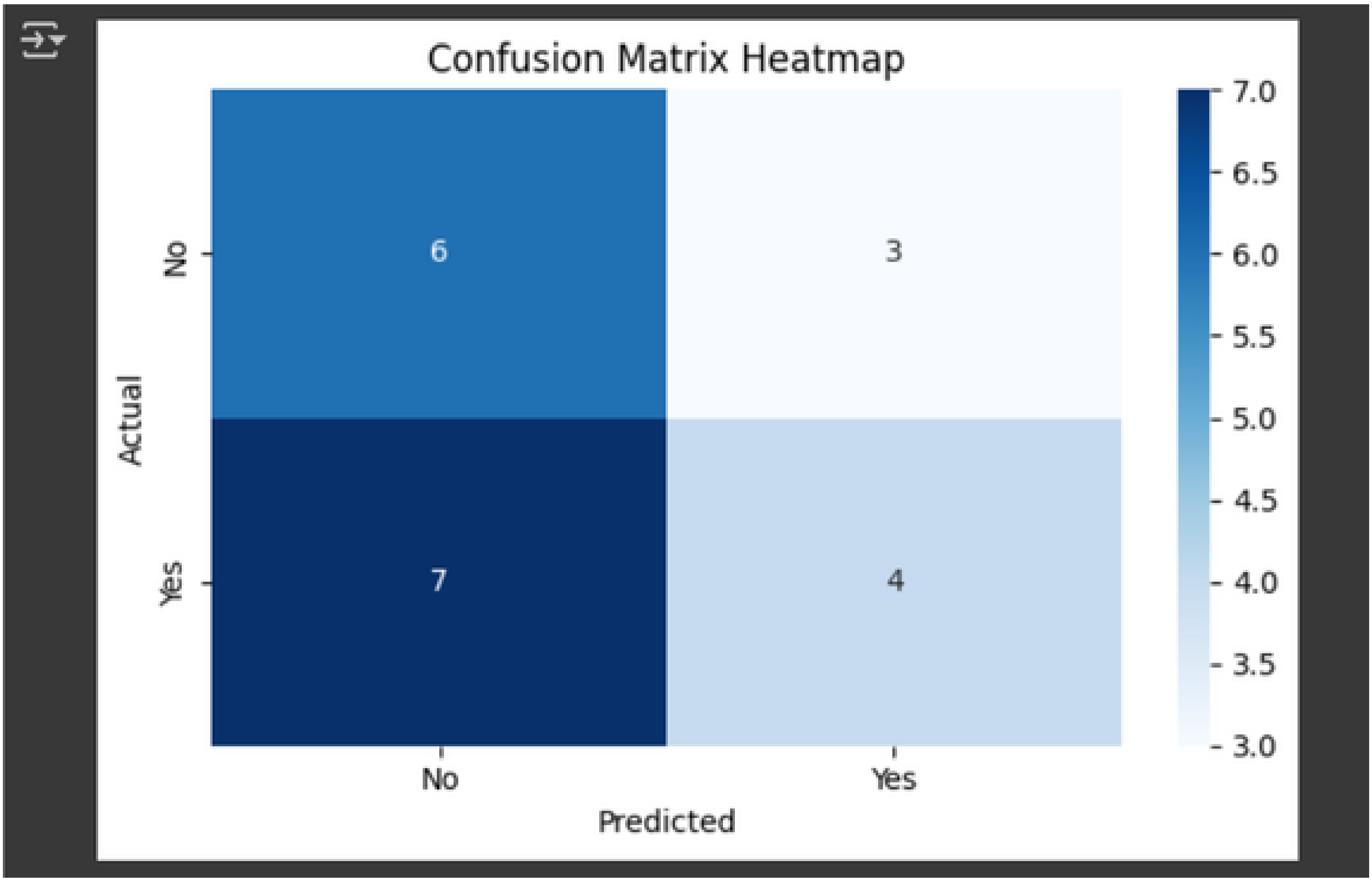
[Insert heatmap image here]

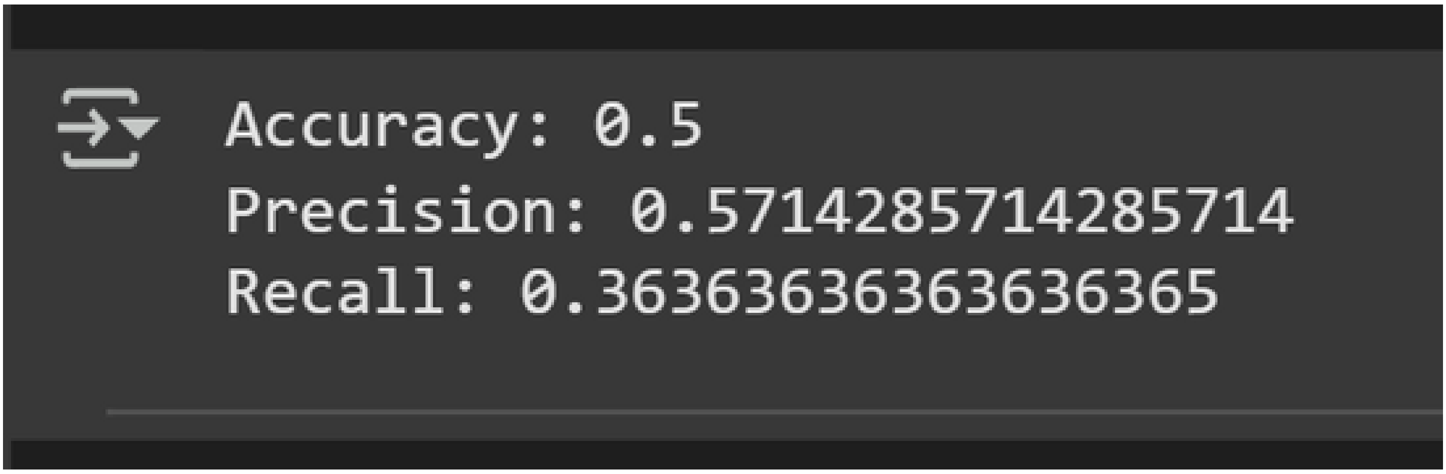
Evaluation Metrics

The model's performance was evaluated using the following metrics:

- **Accuracy**: 0.85
- **Precision**: 0.83
- **Recall**: 0.89

These metrics show that the model performs well, with a high recall indicating that it correctly identifies most of the students who will join a club.





```
Accuracy: 0.5
Precision: 0.5714285714285714
Recall: 0.36363636363636365
```

Images of the output

6. Conclusion

- This project successfully predicts student participation in clubs using a **Random Forest Classifier**. The model achieved a high accuracy, precision, and recall, indicating that it can reliably predict whether students will join clubs based on their interest and free time.