Assessment Report

on

"Student Club Participation Prediction"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Name : Akeel Ahmad

Roll Number : 202401100400020

Section: A

Under the supervision of

"Vicky sir"

KIET Group of Institutions, Ghaziabad

May, 2025

1. Introduction

Extracurricular activities, such as student clubs, are vital to student development. The decision to join a club can depend on multiple factors, such as a student's level of interest and available time. This project aims to develop a predictive model that forecasts a student's club participation using data on interest levels and free hours. The insights from this model can assist universities in improving club engagement and resource planning

2. Problem Statement

Predict whether a student will join a club based on their interest levels and available free time.

3. Objectives

The objective of this project is to develop a machine learning model that accurately predicts a student's likelihood of joining a club based on their interest level and the number of free hours they have per week. By understanding these patterns, educational institutions can better engage students in extracurricular activities, optimize club offerings, and allocate resources more effectively.

4. Methodology

1. **Data Collection**: We used a dataset named club_participation.csv, which includes the following fields:

1. **Data Collection**: We used a dataset named club_participation.csv, which includes the following fields:
    - interest_level (1-10 scale)
    - free_hours_per_week

    - club_participation (Yes/No)
2. **Exploratory Data Analysis (EDA)**:
    - Count plots to observe participation distribution.
    - Boxplots to compare interest and free time against participation.
    - Scatter plots to visualize decision boundaries.
3. **Preprocessing**:
    - Converted participation labels into binary (Yes=1, No=0).
    - Split data into training and testing sets.
4. **Modeling**:
    - Applied logistic regression.
    - Evaluated performance with classification report and confusion matrix.
5. **Visualization**:
    - Probability distribution of predictions.
    - Decision boundary to visualize model separation.

---

5. Data Preprocessing

- Checked and handled any missing values (if present).
- Converted categorical club_participation values to binary: Yes = 1, No = 0.
- Verified and ensured numerical types for features.
- Split the data into training (80%) and testing (20%) sets for model evaluation.

---

6. Model Implementation

- Chose Logistic Regression due to its suitability for binary classification.
- Trained the model using training data with interest level and free hours as predictors.
- Predicted outcomes for test data to assess model performance.

7. Evaluation Metrics

- Used classification report metrics including:
  - **Accuracy**: Overall correctness of the model.
  - **Precision**: How many predicted positive cases were actually positive.
  - **Recall**: How many actual positive cases were correctly predicted.
  - **F1-score**: Harmonic mean of precision and recall.
- Confusion Matrix to visualize true positives, true negatives, false positives, and false negatives.

---

8. Results and Analysis

- The model showed an accuracy of approximately 55% which is slightly better than random guessing.
- The confusion matrix indicated class imbalance, with a bias toward predicting the majority class.
- Visualizations revealed some correlation between higher interest/free time and likelihood of participation.
- The model may benefit from additional features (e.g., past participation, peer influence) or more complex algorithms.

CODE:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression
```

```
from sklearn.metrics import classification_report, confusion_matrix,
ConfusionMatrixDisplay


# Load data
file_path = 'club_participation.csv'
df = pd.read_csv(file_path)


# Data Preprocessing
print("Initial Data Info:")
print(df.info())
print("\nChecking for missing values:\n", df.isnull().sum())
df['club_participation_binary'] = df['club_participation'].map({'yes': 1, 'no': 0})


# Visualizations
sns.countplot(x='club_participation', data=df); plt.title("Participation Count"); plt.show()
sns.boxplot(x='club_participation', y='interest_level', data=df); plt.title("Interest Level");
plt.show()
sns.boxplot(x='club_participation', y='free_hours_per_week', data=df); plt.title("Free
Hours"); plt.show()
sns.scatterplot(data=df, x='interest_level', y='free_hours_per_week',
hue='club_participation'); plt.title("Scatter Plot"); plt.show()


# Prepare data
X = df[['interest_level', 'free_hours_per_week']]
```

```python
y = df['club_participation_binary']

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


# Train model

model = LogisticRegression()

model.fit(X_train, y_train)

y_pred = model.predict(X_test)


# Evaluation

print(classification_report(y_test, y_pred))

disp = ConfusionMatrixDisplay(confusion_matrix(y_test, y_pred)); disp.plot(); plt.show()


# Probability Plot

y_proba = model.predict_proba(X_test)[:, 1]

sns.histplot(y_proba, bins=10, kde=True); plt.title("Prediction Probabilities"); plt.show()
```

9. Conclusion

---

10. References

- scikit-learn documentation
- pandas documentation
- Seaborn visualization library

- Research articles on credit risk prediction
