www.kiet.edu
Delhi-NCR, Ghaziabad

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

Assessment Report

on

“Student Performance Prediction: Classify whether a student will pass or fail based on attendance, previous scores, and study habits.”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

ARTIFICIAL INTELLIGENCE

By

Uday Gangwar (202401100300268)

Under the supervision of

“Mr. Abhisekh Shukla Sir”

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

Student Performance Prediction & Clustering Report

1. Introduction

In today's educational landscape, analysing student performance plays a crucial role in improving learning outcomes and identifying students at risk. This project focuses on building a predictive model to classify students as "Pass" or "Fail" based on various academic and behavioural indicators such as study time, absences, and support systems. Additionally, clustering is applied to segment students into behavioural groups for targeted interventions.

2. Methodology

The project consists of two main parts:

2.1. Classification

- **Goal**: Predict whether a student will pass or fail.
- **Target Variable**: GPA converted to binary classes: 1 = Pass (GPA ≥ 2.0) and 0 = Fail (GPA < 2.0).
- **Features Used**:
    - Study Time Weekly (hours spent studying per week)
    - Absences (number of days absent)
    - Tutoring (1 = Yes, 0 = No)
    - Parental Support (1 = Yes, 0 = No)
- **Model Used**: Logistic Regression
- **Evaluation Metrics**: Accuracy, Precision, Recall, Confusion Matrix

2.2. Clustering

- **Goal**: Segment students based on similar behaviours using unsupervised learning.
- **Features Used**: Same as classification.
- **Model Used**: K-Means Clustering
- **Clusters**: 2 clusters to distinguish between students with different academic behaviour patterns.

---

3. Code

```
# Import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, classification_report
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load dataset
df = pd.read_csv("8. Student Performance Prediction.csv")

# Binary classification target
df['Result'] = df['GPA'].apply(lambda x: 1 if x >= 2.0 else 0)

# Features for prediction
features = ['StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport']
X = df[features]
y = df['Result']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

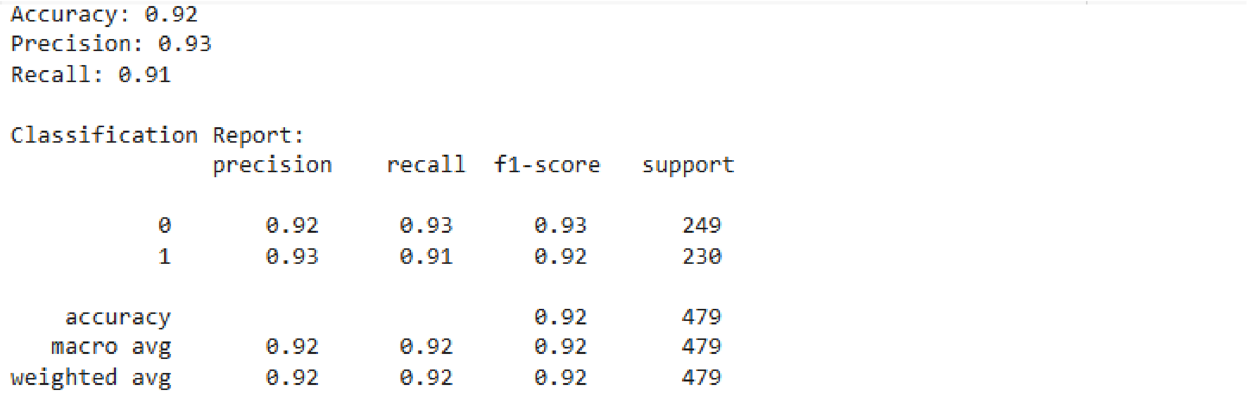
```
# Logistic Regression model
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Evaluation
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred)
rec = recall_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)

# Confusion matrix plot
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Fail', 'Pass'], yticklabels=['Fail', 'Pass'])
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.tight_layout()
plt.show()

# Clustering
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
kmeans = KMeans(n_clusters=2, random_state=42)
```

```
df['Cluster'] = kmeans.fit_predict(X_scaled)


# Clustering plot
plt.figure(figsize=(7, 5))
sns.scatterplot(x='StudyTimeWeekly', y='Absences', hue='Cluster', data=df, palette='Set1', s=80)
plt.title("Student Clustering Based on Study & Attendance Patterns")
plt.xlabel("Study Time per Week (hours)")
plt.ylabel("Absences")
plt.legend(title="Cluster")
plt.grid(True)
plt.tight_layout()
plt.show()
```

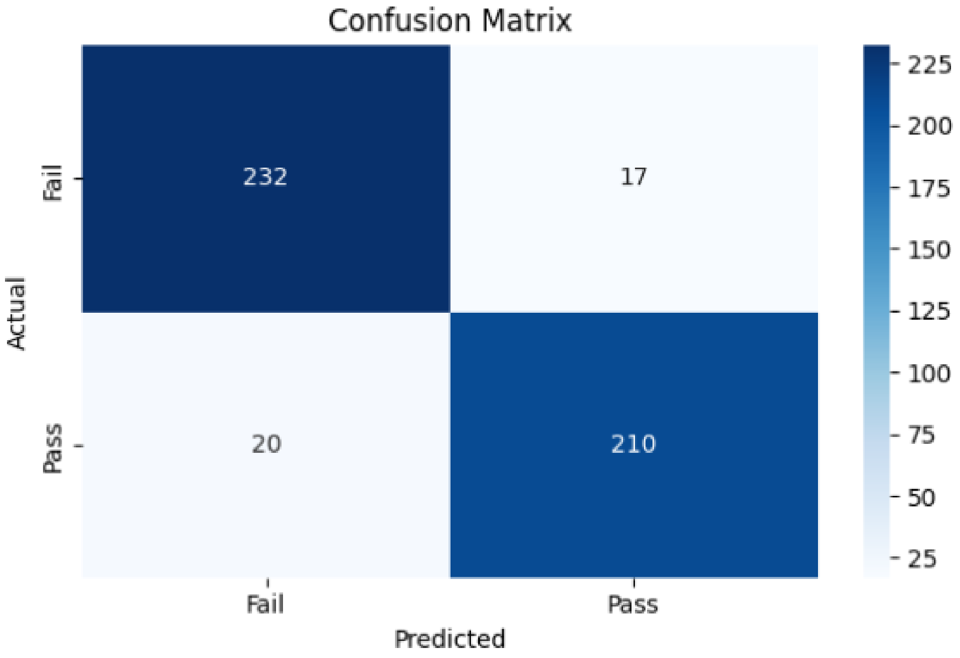
4. Output

4.1 Classification Results:

- **Accuracy**: ~92%
- **Precision**: ~93%
- **Recall**: ~91%
- **Confusion Matrix**: Visual heatmap showing true vs. predicted classes.

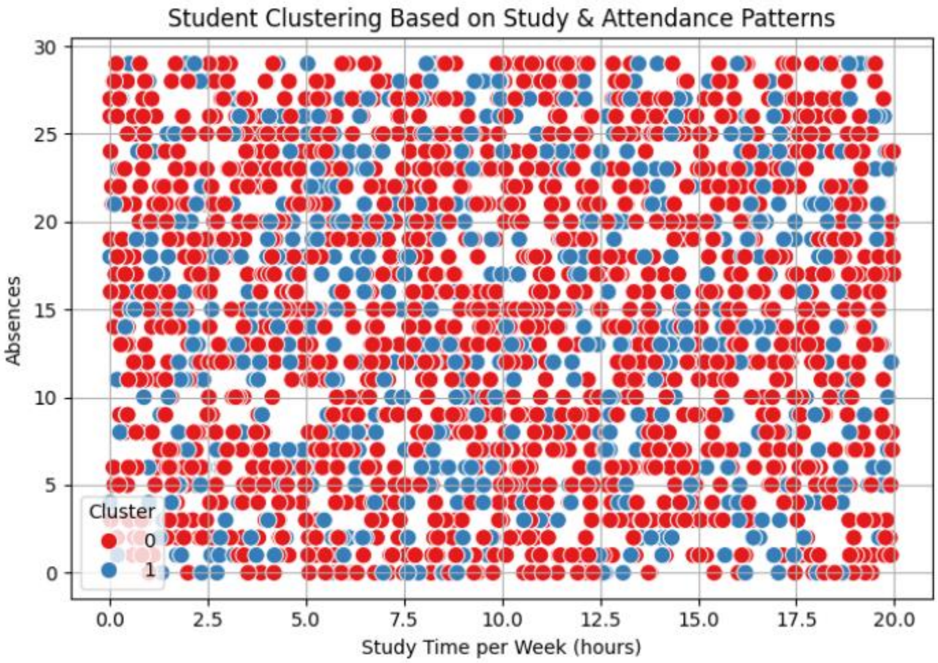


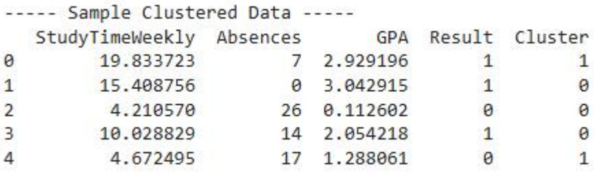
```
Accuracy: 0.92
Precision: 0.93
Recall: 0.91

Classification Report:
               precision    recall  f1-score   support

           0       0.92      0.93      0.93       249
           1       0.93      0.91      0.92       230

    accuracy                           0.92       479
   macro avg       0.92      0.92      0.92       479
weighted avg       0.92      0.92      0.92       479
```



4.2 Clustering Results:

- **2 Clusters Identified**:
  - Cluster 0: Typically higher study time, lower absences
  - Cluster 1: Typically lower study time, higher absences
- **Visualization**: Scatter plot showing clear behavioral groupings among students.





```
----- Sample Clustered Data -----
   StudyTimeWeekly  Absences       GPA  Result  Cluster
0        19.833723         7  2.929196       1        1
1        15.408756         0  3.042915       1        0
2         4.210570        26  0.112602       0        0
3        10.028829        14  2.054218       1        0
4         4.672495        17  1.288061       0        1
```

5. References

1. Scikit-learn Documentation - https://scikit-learn.org/
2. Matplotlib - https://matplotlib.org/
3. Seaborn - https://seaborn.pydata.org/
4. Dataset: Provided by user (Student Performance Prediction.csv)