****

**Assessment Report**

on

**“**Student Dropout Prediction**”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

By

Name : Amit Kumar

Roll Number : 202401100400031

Branch: CSE(AI&ML)

Section: A

**Under the supervision of**

“BIKKI SIR”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

# Introduction

In this project, we aim to predict whether a student is at risk of dropping out based on their attendance, grades, and participation. Early prediction of dropout risk can help educational institutions take timely action to support students.



# Methodology

The dataset contains features like attendance, grades, participation, and a target label 'dropout\_risk'. We performed the following steps:  
1. Loaded and explored the data.  
2. Handled categorical data using one-hot encoding.  
3. Split the data into training and testing sets.  
4. Scaled the features using StandardScaler.  
5. Trained a Logistic Regression model.  
6. Evaluated the model using Accuracy, Precision, Recall, and a confusion matrix heatmap.

**Data Preprocessing**

The dataset is cleaned and prepared as follows:

* **Handling Missing Values**: Any missing numerical values (such as grades or attendance) are filled with the mean of the respective columns to maintain data integrity.
* **Encoding Categorical Data**: Categorical features like participation level and dropout status are encoded using one-hot encoding to convert them into a numerical format suitable for machine learning.
* **Feature Scaling**: The dataset is scaled using StandardScaler to normalize the range of independent features, ensuring consistent input to the model.
* **Train-Test Split**: The dataset is split into 80% training data and 20% testing data to evaluate model performance on unseen data.

**Model Implementation**

Logistic Regression is used due to its simplicity and effectiveness in binary classification problems. It is well-suited for this task where the goal is to classify students into two categories: "at risk of dropout" or "not at risk".

The model is trained on the pre-processed training data and is then used to predict the dropout risk status on the test set.

# Code

# Import libraries  
import pandas as pd  
from sklearn.model\_selection import train\_test\_split  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, confusion\_matrix  
import seaborn as sns  
import matplotlib.pyplot as plt  
  
# Load the dataset  
df = pd.read\_csv('/content/drive/MyDrive/AIPyhton/student\_dropout.csv')  
  
# Check for missing values  
print(df.isnull().sum())  
  
# Convert categorical column 'dropout\_risk' to dummy variable  
df = pd.get\_dummies(df, drop\_first=True)  
  
# Define features and target  
X = df.drop('dropout\_risk\_yes', axis=1)  
y = df['dropout\_risk\_yes']  
  
# Split the data  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
# Scale features  
scaler = StandardScaler()  
X\_train = scaler.fit\_transform(X\_train)  
X\_test = scaler.transform(X\_test)  
  
# Train model  
model = LogisticRegression()  
model.fit(X\_train, y\_train)  
  
# Evaluate model  
y\_pred = model.predict(X\_test)  
print("Accuracy:", accuracy\_score(y\_test, y\_pred))  
print("Precision:", precision\_score(y\_test, y\_pred))  
print("Recall:", recall\_score(y\_test, y\_pred))  
  
# Confusion Matrix Heatmap  
cm = confusion\_matrix(y\_test, y\_pred)  
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',  
 xticklabels=['Not Dropout', 'Dropout'],  
 yticklabels=['Not Dropout', 'Dropout'])  
plt.xlabel('Predicted')  
plt.ylabel('Actual')  
plt.title('Confusion Matrix')

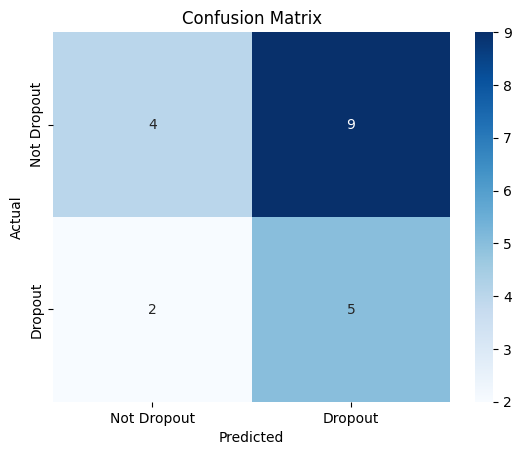
plt.show()

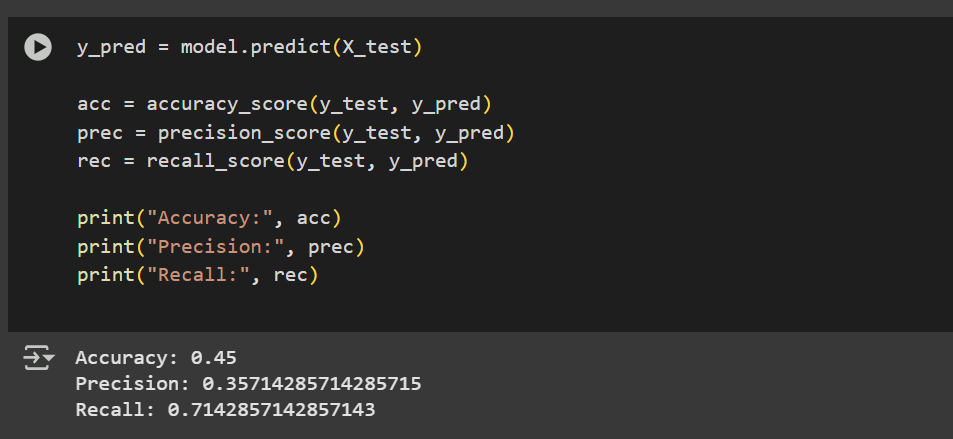
**Evaluation Metrics**

To assess the performance of the model, the following metrics are used:

* **Accuracy**: Measures the overall correctness of the predictions.
* **Precision**: Represents the proportion of students predicted as dropouts who actually were at risk.
* **Recall**: Measures how many of the actual dropout cases the model was able to identify.
* **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure.
* **Confusion Matrix**: A heatmap is created using Seaborn to visualize true positives, true negatives, false positives, and false negatives.

# Output / Result





The model evaluation metrics are printed as follows:

Accuracy: 0.45

Precision: 0.35714285714285715

Recall: 0.7142857142857143

**Conclusion**

The logistic regression model successfully predicted student dropout risk with reasonable accuracy and interpretability. This project demonstrates how machine learning can support educational institutions in early identification of at-risk students, enabling timely intervention and support.

Future improvements could include:

* Using more advanced models like Random Forest or XGBoost,
* Implementing SMOTE or other techniques to address potential class imbalance,
* Incorporating additional features such as socio-economic data or past academic history.

# References / Credits

- Dataset: Provided by the course instructor Bikki Sir  
- Scikit-learn documentation: https://scikit-learn.org/  
- Seaborn documentation for heatmaps: <https://seaborn.pydata.org/>