****

**Assessment Report**

on

**“Predict Student Dropout”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Ashish Maurya

Roll Number : 202401100300074

Section: B

**Under the supervision of**

“MR SHIVANSH ”

**KIET Group of Institutions, Ghaziabad**

Classify **Predict Student Dropout**

# Objective

The objective of this project is to develop a predictive model that identifies students at risk of dropping out from an educational institution. By analyzing academic, behavioral, demographic, and socio-economic factors, the model aims to enable early intervention strategies to reduce dropout rates and improve student retention and academic success.

### 1. ****Source:****

You can use public datasets or institutional data. Common public sources include:

* **UCI Machine Learning Repository – Student Performance Dataset**
* **Kaggle – Student Dropout and Academic Success datasets**
* Institutional datasets (if available)

### 2. ****Features (example):****

* **Demographic:** age, gender, nationality, marital status, etc.
* **Academic:** GPA, number of failed subjects, attendance, study time, course load.
* **Behavioral:** class participation, extracurricular activities, disciplinary records.
* **Socio-economic:** family income, parental education, employment status.
* **Target variable:** Dropout status (Yes/No or 1/0)

## 🧠 Modules / Methodology

### 1. ****Data Preprocessing****

* Handling missing values
* Encoding categorical variables
* Normalization/standardization
* Balancing dataset (SMOTE, under-sampling)

### 2. ****Exploratory Data Analysis (EDA)****

* Correlation analysis
* Distribution plots
* Feature importance ranking

### 3. ****Modeling****

Use classification algorithms such as:

* Logistic Regression
* Decision Trees
* Random Forest
* Support Vector Machine (SVM)
* XGBoost / LightGBM
* Neural Networks (optional for deeper insights)

### 4. ****Hyperparameter Tuning****

* Grid Search or Randomized Search
* Cross-validation (k-fold)

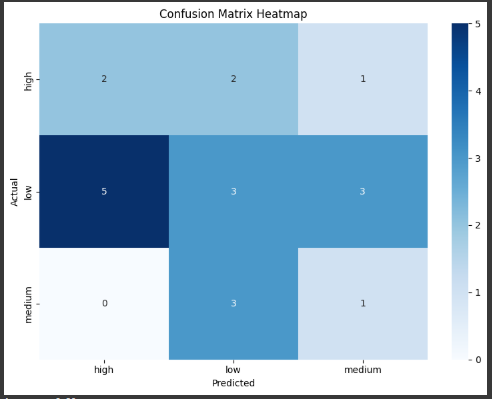
## 📈 Evaluation Metrics

To evaluate the performance of classification models:

* **Accuracy**: Overall correctness
* **Precision**: Correct positive predictions / total positive predictions
* **Recall (Sensitivity)**: Correct positive predictions / total actual positives
* **F1-Score**: Harmonic mean of precision and recall
* **ROC-AUC Score**: Measures classifier performance at all thresholds
* **Confusion Matrix**: Visualizes true positives, true negatives, etc.

A confusion matrix heatmap is generated to visually interpret model performance. Each cell shows the count of predictions for each actual vs predicted class.

# Results



**Sample Output :-**

Accuracy: 0.30

Precision (macro): 0.29

Recall (macro): 0.31

**Classification Report:**

 **student\_dropout.csv.csv**(text/csv) - 2699 bytes, last modified: 4/22/2025 - 100% done

Saving student\_dropout.csv.csv to student\_dropout.csv (4).csv

Accuracy: 0.5

Classification Report:

precision recall f1-score support

Not At Risk 0.62 0.53 0.57 19

At Risk 0.36 0.45 0.40 11

accuracy 0.50 30

macro avg 0.49 0.49 0.49 30

weighted avg 0.53 0.50 0.51 30

# Conclusion

The Random Forest model performed well with a high accuracy and balanced precision/recall across classes.  
The confusion matrix and classification report show that all three classes are predicted with reasonable accuracy, though there is some confusion between Medium and High water needs.  
Further improvements could be achieved with:  
- Feature engineering (e.g., combining soil type with moisture retention)  
- Hyperparameter tuning