**Name:** Md. Masud Rana
**ID:** 212-15-14760
**Training Name:** Enhancing Digital Government and Economy(EDGE)

# Project Report: Student Performance Analysis Using Machine Learning

## 1. Introduction

Machine learning plays a crucial role in analyzing educational data to predict student performance. This project focuses on examining student-related factors that influence academic success using machine learning techniques. The dataset contains demographic, academic, and behavioral attributes that help in performance prediction. The study involves data preprocessing, feature selection, model implementation, and evaluation to identify the most effective predictive approach.

## 2. Objectives

- To preprocess and clean the student performance dataset for better model accuracy.
- To handle missing values and encode categorical features effectively.
- To implement and compare multiple machine learning models.
- To evaluate models using accuracy, precision, recall, and F1-score.
- To identify the best predictive model for student performance.

## 3. Dataset Description

The dataset used in this study contains details about students' demographics, academic history, and lifestyle choices. The key attributes include:

**Demographic Information:**

- Age: Student's age.
- Gender: Male or Female.
- Family Size: Size of the student's family.
- Parental Education Level: Education level of mother and father.

**Academic Details:**

- Study Time: Hours spent studying per week.
- Past Failures: Number of past class failures.

- Absences: Number of school absences.
- School Support: Whether the student receives extra academic support.

**Behavioral and Lifestyle Attributes:**

- Internet Access: Availability of internet at home.
- Extra Activities: Participation in extracurricular activities.
- Free Time and Going Out: Leisure activities and social behavior.
- Alcohol Consumption: Weekly alcohol consumption (workday and weekend).

**Target Variable:**

- Final Grade: Student's final exam performance (categorical classification).

# 4. Data Preprocessing

Data preprocessing is essential to improve model accuracy and efficiency. The following steps were applied:

**Handling Missing Values:**

- Missing values in critical academic features were replaced using mean imputation.
- Categorical missing values were handled using mode imputation.

**Feature Encoding:**

- Categorical variables like parental education and internet access were label encoded.
- One-hot encoding was applied to nominal categorical features.

**Feature Scaling:**

- Min-max scaling was applied to numerical attributes like study time and absences to normalize data.

# 5. Machine Learning Models Implemented

To evaluate different predictive techniques, the following models were implemented:

- **K-Nearest Neighbors (KNN):** A distance-based classification model where performance depends on the choice of k-neighbors.
- **Random Forest Classifier:** An ensemble learning method that constructs multiple decision trees and enhances accuracy.
- **Support Vector Machine (SVM):** A model that maximizes decision boundaries for classification.

# 6. Model Evaluation

To assess model performance, the following evaluation metrics were used:

- **Accuracy:** Measures the overall correctness of predictions.
- **Precision, Recall, and F1-score:** Evaluates model performance considering false positives and false negatives.
- **Confusion Matrix:** Provides a detailed breakdown of correct and incorrect predictions.

# 7. Results and Discussion

- **KNN:** Achieved moderate accuracy but was highly dependent on feature scaling and k-value optimization.
- **Random Forest:** The best-performing model with the highest accuracy due to its robustness and feature importance assessment.
- **SVM:** Showed strong performance but required careful hyperparameter tuning to avoid overfitting.

# 8. Conclusion

This study successfully predicted student performance using different machine-learning models. Among the models tested, the Random Forest classifier delivered the best results due to its ability to handle complex features and provide high accuracy. This model can be effectively used for early academic performance prediction, allowing educators to provide targeted support to students.

# 9. Future Work

To further improve performance and extend the study, the following aspects can be explored:

- **Hyperparameter Tuning:** Further optimization of machine learning models for better accuracy.
- **Advanced Feature Engineering:** Exploring additional factors affecting student performance.
- **Deep Learning Models:** Utilizing neural networks for improved predictive capabilities.
- **Early Dropout Prediction:** Extending the study to identify students at risk of dropping out based on academic and behavioral trends.