# Academic Success Prediction Project Report

## 1. Introduction

This project focuses on predicting student academic success using machine learning. The dataset includes demographic, academic, and socio-economic variables that influence student outcomes. The goal is to build a reliable model to identify students who may need additional academic support.

## 2. Dataset Overview

The dataset contains multiple features such as marital status, application mode, course, previous qualification, curricular unit performance, and economic indicators. These features help predict whether a student is likely to succeed or face challenges in their academic journey.

## 3. Data Preprocessing

Preprocessing steps included:

- Handling missing values

- Encoding categorical features

- Feature scaling using StandardScaler

- Removing highly correlated features based on correlation threshold (0.8)

- Selecting top features based on variance and SelectKBest

## 4. Exploratory Data Analysis (EDA)

EDA involved visualizing feature distributions, detecting skewness, and identifying outliers. Correlation analysis was performed to understand relationships among features.

## 5. Feature Selection

Highly correlated features were removed using correlation matrices. Top 10 numerical features were selected using variance ranking, and SelectKBest was applied using ANOVA F-test.

## 6. Model Building

Multiple machine learning models were trained, including:

- Logistic Regression

- Random Forest

- Gradient Boosting

Random Forest and Gradient Boosting performed the best among all trained models.

## 7. Hyperparameter Tuning

Hyperparameter tuning was performed using GridSearchCV. The best models were identified as:

- Random Forest (Best Accuracy: ~0.752)

- Gradient Boosting (Best Accuracy: ~0.751)

Random Forest achieved the highest accuracy after tuning.

## 8. Machine Learning Pipeline

Pipelines were implemented for automation of preprocessing, feature selection, scaling, and model training. Separate pipelines were created for Random Forest and Gradient Boosting.

Final tuned pipeline models were saved for reuse.

## 9. Testing with Unseen Data

An unseen dataset was preprocessed by aligning its columns with the training data. Predictions from Random Forest and Gradient Boosting models were generated.

Sample Predictions (RF): [0, 1, 0, 1, 1, 0, 1]

Interpretation:

- 4 students predicted to succeed

- 3 students predicted to face challenges

## 10. Conclusion

The prediction model provides valuable insights that can help educators identify at-risk students. Random Forest performed the best overall. The system can be improved by adding more data, enhancing features, and periodically retraining the model to maintain accuracy.