Report

Student Success Prediction using Machine Learning

Goutam Agarwal

1. Introduction

This project focuses on predicting whether a student will pass the final exam using machine learning models. The dataset comes from the UCI Machine Learning Repository, containing demographic, social, and academic factors.

2. Motivation

The COVID-19 pandemic highlighted declining student performance worldwide. Educational systems now require data-driven methods to understand learning difficulties early. With machine learning, historical student records can be analyzed to detect students at risk of failing and provide insights for teachers and parents to intervene proactively.

3. Problem Statement

This is a binary classification problem where the target variable is student status (Pass/Fail). The predictors include academic, family, lifestyle, and socio-economic features such as absences, study time, parent’s job, and health.

4. Data Processing

- Label Encoding: Converted categorical variables into numerical codes.
- Feature Scaling: Normalized continuous variables for efficient convergence.
- Binary Variables: Left unchanged (0/1 format).

5. Data Visualization

- Histograms: Showed distributions of features (e.g., internet access).
- Boxplots: Highlighted variation in performance across different factors.
- Correlation Heatmaps: Revealed strongest relationships with performance (study time, absences, parental education).

6. Models and Evaluation

Three classifiers were implemented and compared:

1. Logistic Regression

- Accuracy: ~78%, ROC AUC: 0.75
- Pros: Fast, interpretable. Cons: Limited for non-linear patterns.

2. K-Nearest Neighbors (KNN)
   - Accuracy: ~81%, ROC AUC: 0.77
   - Pros: Easy to understand. Cons: Expensive on large datasets.

3. Support Vector Machines (SVM)
   - Linear Kernel: Accuracy 84.3%, F1 = 0.82, ROC AUC = 0.80
   - Polynomial Kernel: Accuracy 78%, ROC AUC = 0.73
   - Gaussian Kernel: Accuracy 83%, ROC AUC = 0.74

Best Model: Linear SVM (balanced accuracy and F1 score).

7. Key Factors Extracted

Positive Impact (success predictors):
- Parental education, motivation for higher studies, study time, parental jobs.

Negative Impact (failure predictors):
- Absences, prior failures, excessive going out with friends, health issues, older age.

8. Conclusion

Among tested classifiers, SVM with Linear Kernel performed best (~84% accuracy). Strong predictive factors include study habits, parental education, and attendance. Recommendations include encouraging discipline, supporting weaker backgrounds, and early warning systems.

9. Future Work

- Use ensemble methods (Random Forest, Gradient Boosting).
- Apply feature selection to reduce noise.
- Expand dataset for better generalization.
- Explore deep learning if dataset size increases.