Student Performance Prediction Using Machine Learning

Team VectorSpace
Uday Choudhary
B Sai Thrishul
Hariksh
R Sai Praneeth
AI Study Coach

***Abstract*—This project presents a structured machine learning study for predicting student final grades using progressively expanded feature sets. The dataset initially contained 25,000 records and was reduced to 15,000 after duplicate removal. Four controlled experiments were conducted using different feature configurations. Experiment 1 used academic and behavioral features. Experiment 2 added contextual features. Experiment 3 used all available features. Experiment 4 extended the evaluation to multiple additional classifiers. Performance was evaluated using accuracy, weighted F1-score, confusion matrices, cross-validation, learning curves, and feature importance analysis.**

***Index Terms*—Machine Learning, Classification, Educational Data Mining, Feature Engineering.**

I. Introduction

Predicting student academic performance enables early identification of at-risk students and supports personalized learning strategies. This project evaluates multiple machine learning models across four structured experiments to analyze the impact of feature expansion on predictive performance.

II. Methodology

A. Dataset Description

The dataset contained 25,000 student records with 16 attributes. After removing 10,000 duplicate rows, 15,000 records remained. The target variable `final_grade` was encoded from 0 (F) to 5 (A).

B. Preprocessing

- Duplicate removal
- IQR-based outlier capping
- Binary encoding
- Ordinal encoding
- One-hot encoding
- Standard scaling
- Stratified 80-20 train-test split

C. Experiments

Experiment 1: Academic + Behavioral Features
Models: Logistic Regression, Decision Tree

Experiment 2: + Contextual Features
Models: Logistic Regression, Decision Tree

Experiment 3: All Features
Models: Logistic Regression, Decision Tree

Experiment 4: Extended Model Comparison
Models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, Multi-Layer Perceptron

III. Results

A. Experiment 1 — Academic + Behavioral Features

TABLE I
Experiment 1 Performance

Model	Test Accuracy	Weighted F1
Logistic Regression	0.7693	0.7694
Decision Tree	0.6643	0.6644

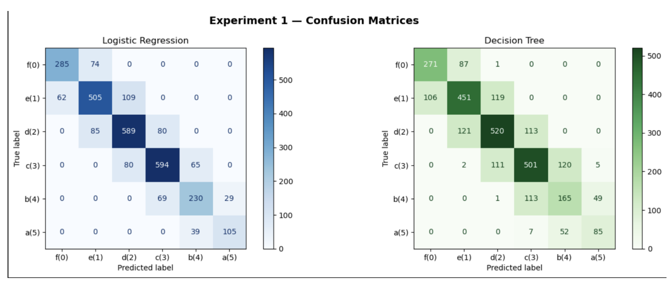

Fig. 1. Experiment 1 Confusion Matrices

B. Experiment 2 — + Contextual Features

TABLE II
Experiment 2 Performance

Model	Test Accuracy	Weighted F1
Logistic Regression	0.7723	0.7725
Decision Tree	0.6590	0.6590

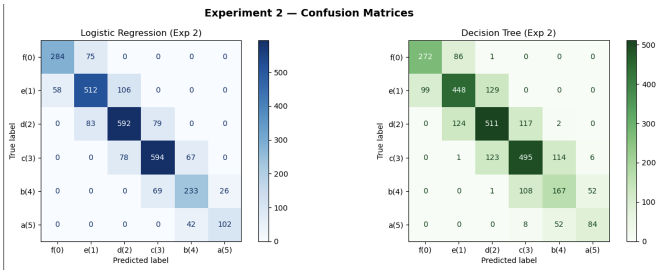

Fig. 2. Experiment 2 Confusion Matrices

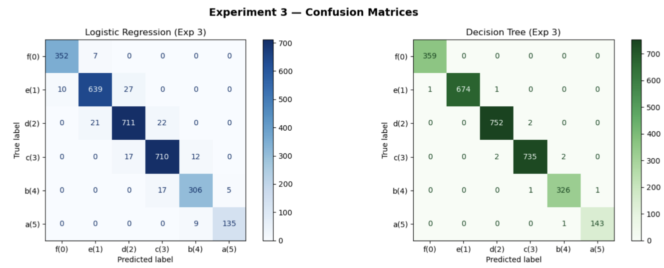

Fig. 3. Experiment 3 Confusion Matrices

TABLE III
EXPERIMENT 3 PERFORMANCE

Model	Test Accuracy	Weighted F1
Logistic Regression	0.9510	0.9510
Decision Tree	0.9963	0.9963

TABLE IV
EXPERIMENT 4 MODEL PERFORMANCE

Model	Test Accuracy	Weighted F1
Random Forest	0.9973	0.9973
Decision Tree	0.9963	0.9963
Gradient Boosting	0.9960	0.9960
MLP	0.9787	0.9787
Logistic Regression	0.9510	0.9510
SVM	0.9367	0.9367
KNN	0.7277	0.7257

C. Experiment 3 — All Features

D. Experiment 4 — Extended Model Comparison

E. Feature Importance Analysis

F. Learning Curve Analysis

IV. DISCUSSION

Model performance improved as additional feature groups were included. Experiments 3 and 4 achieved the highest accuracy when all features were used. Cross-validation showed minimal training-validation gap for tree-based ensemble models.

V. CONCLUSION

This study implemented four structured experiments to evaluate feature expansion and model comparison for student grade prediction. Ensemble methods achieved the highest predictive performance among evaluated models.

REFERENCES

APPENDIX

https://github.com/crazylogic03/VectoSpace

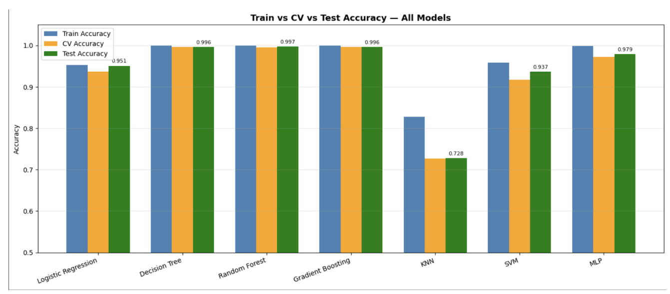

Fig. 4. Train vs CV vs Test Accuracy (All Models)

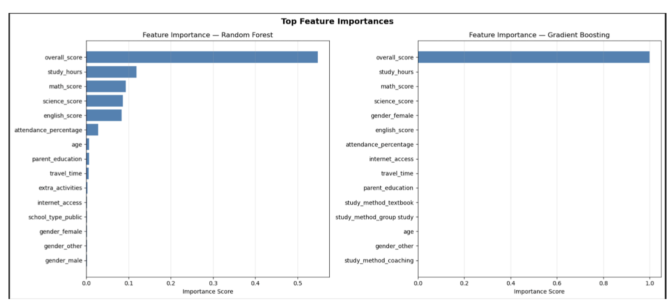

Fig. 5. Top Feature Importances (Random Forest Gradient Boosting)

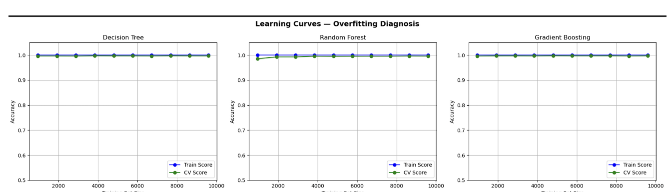

Fig. 6. Learning Curves for Tree-Based Models