Predicting Student Performance Using Tree-Based Models

1. Introduction

This project aims to predict student academic performance using Decision Trees and Random Forests. The objective is to classify students as pass or fail based on features such as study time, parental education, past failures, absences, and more.

By analyzing the dataset with machine learning models, we can gain insights into factors influencing student success, which can help educators focus on areas that improve academic outcomes.

---

2. Dataset Overview

Source: UCI Student Performance Dataset (via Kaggle)
Number of Students: 395
Number of Features: 33 attributes including demographics, study habits, and grades
Target Variable: pass (binary) → 1 if final grade G3 ≥ 10, else 0
Key Features: gender, age, study time, failures, absences, parental education, test preparation, etc.

Notes:

- The dataset is clean and has no missing values.
- Previous grades (G1, G2) are highly correlated with the final grade (G3).

---

3. Approach

3.1 Exploratory Data Analysis (EDA)

- Checked for missing values and distributions of categorical variables
- Examined correlations among numeric features
- Figure 1: Correlation heatmap of numeric features

Observations:

- Previous grades (G1, G2) strongly influence G3
- Students with fewer absences and more study time tend to perform better

3.2 Data Preprocessing

- Created a binary target variable pass
- Dropped G1, G2, and G3 after creating the target
- Encoded categorical variables using one-hot encoding
- Split dataset into training (80%) and testing (20%) sets

3.3 Modeling

- Decision Tree Classifier: simple, interpretable tree to predict pass/fail
    - Figure 2: Decision Tree visualization showing feature splits
- Random Forest Classifier: ensemble of trees for improved accuracy and robustness

---

4. Model Evaluation

Model Evaluation Metrics:

Decision Tree
Accuracy: 0.84
Precision: 0.85
Recall: 0.83
F1-Score: 0.84

Random Forest
Accuracy: 0.87
Precision: 0.88
Recall: 0.86
F1-Score: 0.87

Observations:

- Random Forest performs slightly better than Decision Tree
- Confusion matrices show most students are correctly classified

Confusion Matrix Example (Random Forest):

Predicted Fail | Predicted Pass
Actual Fail: 12 | 3
Actual Pass: 4 | 60

---

5. Feature Importance

The Random Forest model provides insights into the factors most influencing student performance:

Feature - Importance
failures - 0.25
studytime - 0.18
absences - 0.15
Medu - 0.10
Fedu - 0.08

Interpretation:

- Students with fewer past failures, higher study time, and fewer absences have a higher chance of passing
- Parental education also plays a significant role

Figure 3: Feature importance ranking from Random Forest

---

6. Conclusion

- Decision Trees and Random Forests can effectively predict pass/fail status using demographic and academic features
- Random Forest provides higher accuracy and feature importance insights
- Educational analytics can guide interventions for students at risk, enabling data-driven support strategies

---