Student Performance Classification — Project Report

Project: Student-Performance-Classification

Link: Student-Performance-Classification
Author: Mayuresh Thorve

1. Introduction

Academic performance prediction plays a crucial role in identifying at-risk students and enabling early intervention strategies. With the growing availability of educational data, machine learning techniques can be effectively used to analyze patterns in student behavior, demographics, and academic history.

This project focuses on building a **classification system** to predict whether a student will **Pass or Fail** using **Decision Tree** and **Random Forest** algorithms. The work demonstrates a complete machine learning workflow, including data exploration, preprocessing, model training, evaluation, and comparison.

2. Objective

The primary objectives of this project are:

- To analyze student-related academic and socio-economic features.
- To convert student grades into a binary Pass/Fail target variable.
- To train and evaluate Decision Tree and Random Forest classifiers.
- To compare model performance and identify important contributing features.

3. Dataset Overview

The dataset contains student-related attributes that influence academic outcomes. These include:

- Academic features: Study time, past failures, grades
- Socio-economic features: Parental education, family support
- Behavioral features: Absences, extracurricular involvement

Key Observations

- The dataset contains minimal or no missing values, reducing the need for heavy imputation.
- The distribution of grades is suitable for binary classification.
- Initial exploration shows that study time, parental education, and previous grades have a strong influence on final performance

4. Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to understand feature distributions and relationships:

- **Correlation analysis revealed that absences negatively impact performance.**
- **Study time and parental education (Medu, Fedu) show a positive correlation with passing outcomes.**
- **Visualization techniques were used to identify trends and outliers in the data.**

These insights guided feature selection and model training

5. Data Preprocessing

Data preprocessing steps included:

- Encoding categorical variables where required.
- Feature-target separation.
- Converting final grades into a binary target variable:
    - **Pass (1):** Final grade ≥ 10
    - **Fail (0):** Final grade < 10
- Splitting the dataset into **training (80%)** and **testing (20%)** sets using stratification to preserve class balance.

6. Model Performance — Comparison

Model	**Accuracy**	**F1-score**	**Precision**	**Recall**
Decision Tree	0.86923	0.92237	0.92660	0.91818
Random Forest	0.9000	0.94170	0.92920	0.95454

7. Performance Summary

- The **Random Forest model outperformed the Decision Tree**, achieving higher accuracy and F1-score.
- Decision Tree showed good interpretability but was more prone to overfitting.
- Random Forest demonstrated robust and consistent performance on unseen data

8. Feature Importance Analysis

Feature importance analysis highlighted the most influential factors affecting student performance:

- **Previous grades / grade class**
- **Absences**
- **Weekly study time**
- **Parental education level**

These insights align well with real-world educational understanding and validate the model's predictions.

9. Conclusion

This project successfully demonstrates the application of machine learning techniques to predict student academic outcomes. By leveraging Decision Tree and Random Forest models, meaningful patterns were identified that can help educators and institutions take proactive measures to support students.