Data Science Applications

TITLE: Student Performance Prediction

Name	ID
1. Yohannis Hailye	1501575
2. Eyerusalem Habtegebreal	1501186
3. Bantalem Mitiku	1501020
4. Firehywet Tesfaye	1501207
5. Addisu Guche	1500959

Submit to : Petros A.

Student Performance Prediction – Final Report

1. Introduction

The objective of this project is to predict whether a student will **Pass or Fail** based on academic, behavioral, and socio-demographic factors. Early identification of at-risk students enables timely academic support and better educational outcomes. This project applies a full machine learning pipeline—from data exploration to deployment—to build a reliable and interpretable classification model.

---

2. Data Description

The dataset contains **708 student records** with **10 features**, including:

- **Numerical:**
  Study_Hours_per_Week, Attendance_Rate, Past_Exam_Scores, Final_Exam_Score
- **Categorical:**
  Gender, Parental_Education_Level, Internet_Access_at_Home, Extracurricular_Activities
- **Target Variable:**
  `Pass_Fail` (binary classification)

There are **no missing values**, and the dataset is balanced (Pass ≈ Fail).

---

3. Exploratory Data Analysis (EDA)

Key observations from EDA:

- Higher **study hours**, **attendance rate**, and **past exam scores** strongly correlate with passing.

- Students with **internet access at home** and **extracurricular involvement** show slightly better outcomes.
- No extreme outliers requiring removal.
- The target variable is well-balanced, making standard classification metrics appropriate.

EDA confirmed that the dataset is clean, structured, and suitable for machine learning.

4. Data Preprocessing

The following steps were applied:

- Removed unnecessary columns: `Student_ID`, `Final_Exam_Score`
- Target defined as `Pass_Fail`
- Numerical features scaled using **StandardScaler**
- Categorical features encoded using **OneHotEncoding**
- Data split into **80% training** and **20% testing** with stratification
- Preprocessing pipeline saved for reuse

All preprocessing steps were modular and reusable.

5. Methodology

Three classification models were evaluated:

- **Logistic Regression** (baseline)
- **Support Vector Machine (SVM)**
- **Random Forest Classifier**

Models were compared using:

- Cross-validation accuracy
- Test accuracy
- Precision, Recall, F2-score
- ROC-AUC

Recall was emphasized to minimize false negatives (failing students incorrectly predicted as passing).

6. Model Results

Best Model: Random Forest

Performance on Test Set:

- Accuracy: **87.32%**
- Precision (Pass): **81%**
- Recall (Pass): **97%**
- ROC-AUC: **0.98**

Confusion Matrix:

```
[[55, 16],
 [ 2, 69]]
```

This shows excellent recall with very few failing students misclassified as passing.

Random Forest outperformed all other models across every key metric.

7. Discussion & Insights

- **Attendance rate** and **study hours** and **past exam scores** are the strongest predictors.
- Random Forest handles non-linear relationships effectively.
- Feature importance analysis confirms academic effort as the dominant factor.
- The model is robust but may not generalize to different educational systems without retraining.

Limitations:

- Dataset size is moderate
- No temporal or longitudinal data
- Socioeconomic factors are limited

---

8. Recommendations

- Use the model as an **early-warning system** for academic intervention.
- Encourage attendance and structured study habits.
- Collect additional features such as stress levels or course difficulty.
- Periodically retrain the model with new student data.

---

9. Conclusion

This project successfully developed a reliable student performance prediction system using a **Random Forest classifier**. The model demonstrates strong recall, high accuracy, and practical applicability. Future work can focus on expanding features, improving interpretability, and deploying the system at scale.