STUDENT PERFORMANCE ANALYSIS & PASS/FAIL PREDICTION USING MACHINE LEARNING

Prepared by: Simanto Podder

1. Introduction

This report presents a comprehensive analysis of student academic performance data using data science and machine learning. The primary objective is to predict whether a student will pass or fail based on exam scores and demographic features.

2. Dataset Overview

The dataset used, 'StudentsPerformance.csv', contains exam scores in math, reading, and writing, as well as attributes such as gender, lunch type, and test preparation course for 1000 students.

3. Data Preprocessing

- All categorical values were cleaned (e.g., converted to lowercase, whitespace removed).
- Categorical features were encoded numerically:
  * Gender: female = 0, male = 1
  * Lunch: free/reduced = 0, standard = 1
  * Test Preparation: none = 0, completed = 1
- A new feature 'average_score' was created using the mean of all scores.
- A target variable 'result' was created: Pass (1) if average_score >= 60, otherwise Fail (0).

4. Exploratory Data Analysis

- Visualization and statistical analysis were performed to discover relationships.
- Reading and writing scores were strongly correlated.
- Students with completed test preparation and standard lunch scored higher.

- Females showed slightly higher average scores overall.

5. Model Building

Two classification models were trained:

- Logistic Regression
- Random Forest Classifier

Random Forest achieved superior performance with an accuracy of 99%, while Logistic Regression gave 96%.

6. Evaluation Metrics

- Accuracy: Overall correctness of model predictions.
- Precision, Recall, and F1-score: Analyzed using a classification report.
- Confusion Matrix: Visual tool to assess classification correctness.

7. Feature Importance

Top predictors of student performance:

1. Reading Score
2. Writing Score
3. Math Score
4. Test Preparation Course
5. Lunch Type
6. Gender

8. Conclusion

The Random Forest model accurately predicts student pass/fail outcomes. The model highlights that reading and writing performance are critical indicators of academic success. Test preparation courses and standard lunch programs also show positive influence.

9. Tools & Technologies Used

- Python
- Pandas, Matplotlib, Seaborn
- Scikit-learn
- Jupyter Notebook

10. Future Work

- Add more features like parental education or study hours
- Use hyperparameter tuning
- Deploy the model using Streamlit
- Apply similar analysis to other educational datasets

11. Author

Simanto Podder

Department of Software Engineering

AIUB