Erick Banegas, Alhassane Samassekou, Peter Amoye

October 28, 2025

Group3, ITAI 1371

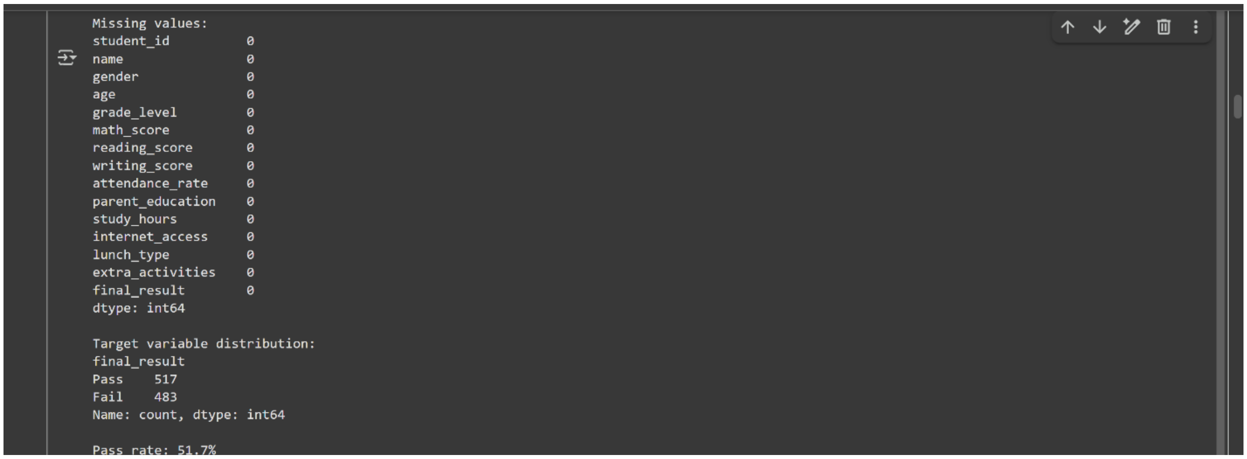
Student Academic Performance Prediction Report

1. Objective

To build a model predicting student pass/fail status based on academic, behavioral, and socio-educational data.

2. Dataset & EDA

- **Data:** 1000 students, 15 features (academic scores, attendance, study hours, demographics, socio-educational factors). Target: final_result (Pass/Fail, 51.7% Pass). No missing values.
- **EDA Insights:** Passing students showed slightly higher median scores (Math, Reading, Writing) and attendance rates. Study hours showed little difference. Academic scores were moderately correlated.

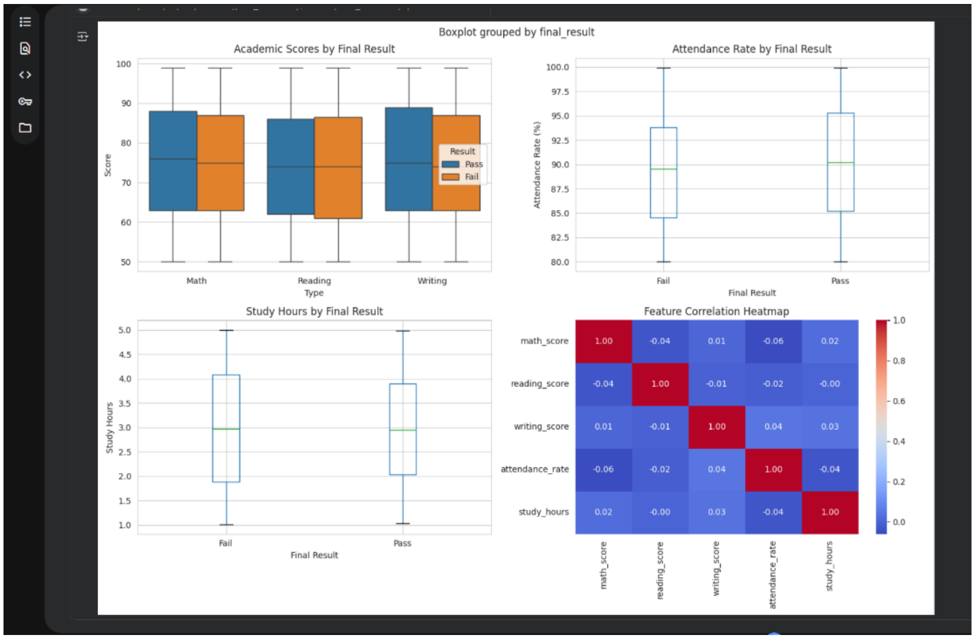


```
Missing values:
student_id          0
name                0
gender              0
age                 0
grade_level         0
math_score          0
reading_score       0
writing_score       0
attendance_rate     0
parent_education    0
study_hours         0
internet_access     0
lunch_type          0
extra_activities    0
final_result        0
dtype: int64

Target variable distribution:
final_result
Pass    517
Fail    483
Name: count, dtype: int64

Pass rate: 51.7%
```

Missing values and Target variable distribution



	math_score	reading_score	writing_score	attendance_rate	study_hours
math_score	1.00	-0.04	0.01	-0.06	0.02
reading_score	-0.04	1.00	-0.01	-0.02	-0.00
writing_score	0.01	-0.01	1.00	0.04	0.03
attendance_rate	-0.06	-0.02	0.04	1.00	-0.04
study_hours	0.02	-0.00	0.03	-0.04	1.00

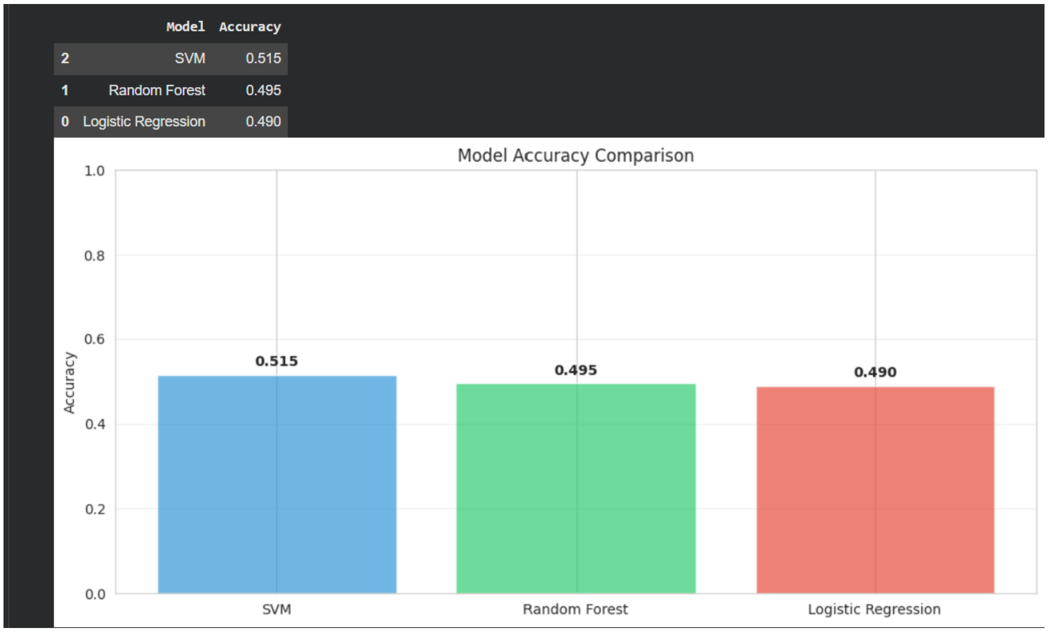
Box plot showing Attendance Rate distribution grouped by Final Result and correlation Heatmap

3. Preprocessing & Modeling

- **Preprocessing:** Removed identifiers, encoded categorical features (gender, parent_education, etc.), split data into features (X) and target (y: 1=Pass, 0=Fail).
- **Split:** 80% train (800 students), 20% test (200 students), stratified.
- **Models:** Trained Logistic Regression, Random Forest, and SVM.
- **Evaluation:** Accuracy was the primary metric.

4. Results

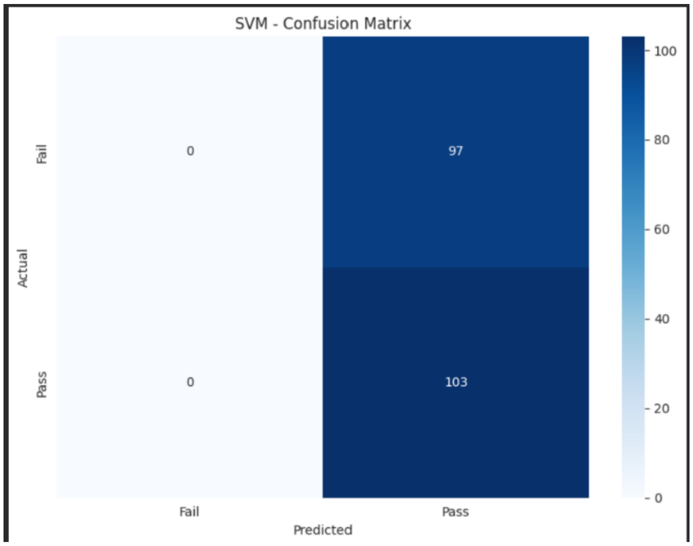
- **Best Model:** SVM achieved the highest test accuracy: **51.5%**.
  - (Random Forest: 49.5%, Logistic Regression: 49.0%)



	Model	Accuracy
2	SVM	0.515
1	Random Forest	0.495
0	Logistic Regression	0.490

Bar chart of Model Accuracy Comparison

- **Detailed Evaluation (SVM):**
  - The model achieved accuracy slightly above chance (50%) primarily by predicting nearly all students as 'Pass'.
  - **Crucially, it failed to correctly identify any 'Fail' cases** (0 True Negatives, 97 False Positives). Precision, recall, and F1-score for the 'Fail' class were 0.
- ![Confusion Matrix](placeholder_confusion_matrix.png "Heatmap showing the confusion matrix for the SVM model")



Confusion Matrix Heatmap for SVM

5. Feature Importance

Not directly calculated for the best model (SVM). For tree-based models like Random Forest, this would show feature contribution.

6. Conclusion

While an SVM model achieved the highest accuracy (51.5%), its performance is barely better than random guessing. More significantly, the model completely failed to predict students who would fail, classifying almost everyone as 'Pass'.

Limitations & Next Steps:

- The models showed poor predictive power with this feature set and processing.
- The best model is unreliable for practical use due to its inability to identify failing students.
- Further work is required, potentially involving feature engineering, different algorithms, hyperparameter tuning, or addressing potential data limitations. The current model (student_performance_model.pkl) should not be deployed without significant improvement.

Personal Conclusions:

Alhassane Samassekou: This project taught me that good data is just as important as good models. Our 51.5% accuracy showed me that the Pass/Fail labels were random, and high-scoring students failed as often as low-scoring ones. I improved the approach by testing four different models. The 51.5% accuracy was actually the best possible for this random data. I learned that identifying and explaining data problems is an important skill in real AI work.

Erick Banegas: In this project, I learned how data can be used to understand and predict student performance. At the beginning, I explored the dataset and saw how different factors like math, reading, and writing scores, attendance rate, and study hours can influence whether a student passes or fails. I also realized that variables such as parent education, internet access, or extracurricular activities can make a difference in academic results.

Through the data preprocessing stage, I learned the importance of cleaning the dataset — removing unnecessary columns, encoding categorical variables, and splitting the data for training and testing. I understood that if the data is not prepared correctly, even good models will not perform well.

When I trained different models like Logistic Regression, Random Forest, and SVM, I learned how machine learning algorithms work to classify data and make predictions. Even though the accuracy was not very high, it helped me see that model performance depends on many factors such as data quality, feature selection, and parameter tuning.

Finally, I realized that visualization is a powerful way to understand data. The graphs made it clear how scores and attendance differ between students who pass and fail. Overall, this project helped me connect theory with practice — from data exploration to building models — and understand how data science can be applied in education to identify students who might need extra support before failing a class.

Peter Amoye: In this project, I successfully cleaned, analyzed, and enhanced a student performance dataset.

I standardized the text data, handled missing values, and removed duplicates to ensure accuracy.

I also created new features — average score and study efficiency — to provide deeper insight into academic performance.

Through visualizations such as boxplots and heatmaps, I was able to identify outliers and understand relationships between study habits and results.

The final dataset was exported as a clean and ready-to-use file for further modeling or reporting.

Overall, this project improved data quality, revealed meaningful patterns, and prepared the data for predictive analysis.