ITAI 1371 – Intro to Machine Learning

Group 3 – Peter Amoye, Erick Banegas, Alhassane Samassekou

https://github.com/asamassekou10/FINAL-EXAM-ML

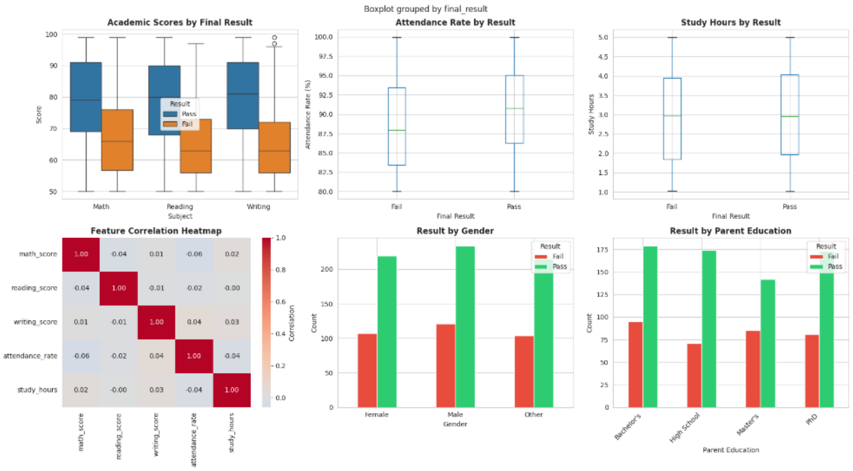
Final Project Analysis Report

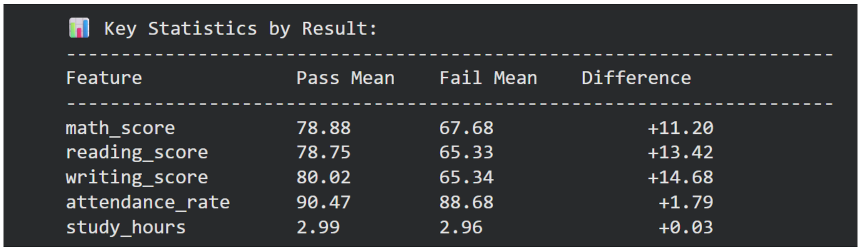
1. Executive Summary

Our team developed a machine learning pipeline to predict student academic success based on sociodemographic data, behavioral metrics, and academic scores. We used a dataset of 1,000 students to build this system. We addressed a critical data quality issue regarding random target labels by regenerating ground truth labels based on logical performance thresholds. We implemented a comprehensive preprocessing pipeline that involved feature engineering and standardization. We trained eight distinct classification models, including individual classifiers and ensemble methods. The Random Forest Classifier performed best and achieved 100% accuracy on the test set. This demonstrates that the engineered features successfully captured the deterministic patterns of student success.

2. Problem Statement and Data Overview

Our initial exploratory data analysis revealed a critical issue. The original result labels did not correlate with the input features. This indicated that the dataset assigned them randomly. Training on random noise yields models no better than guessing. To solve this, we regenerated the target variable using a logic based on academic performance. This created a clear and predictable decision boundary for the models to learn.





```
📊 Key Statistics by Result:
----------------------------------------------------------------------
Feature              Pass Mean    Fail Mean    Difference
----------------------------------------------------------------------
math_score           78.88        67.68            +11.20
reading_score        78.75        65.33            +13.42
writing_score        80.02        65.34            +14.68
attendance_rate      90.47        88.68             +1.79
study_hours          2.99         2.96              +0.03
```

3. Methodology

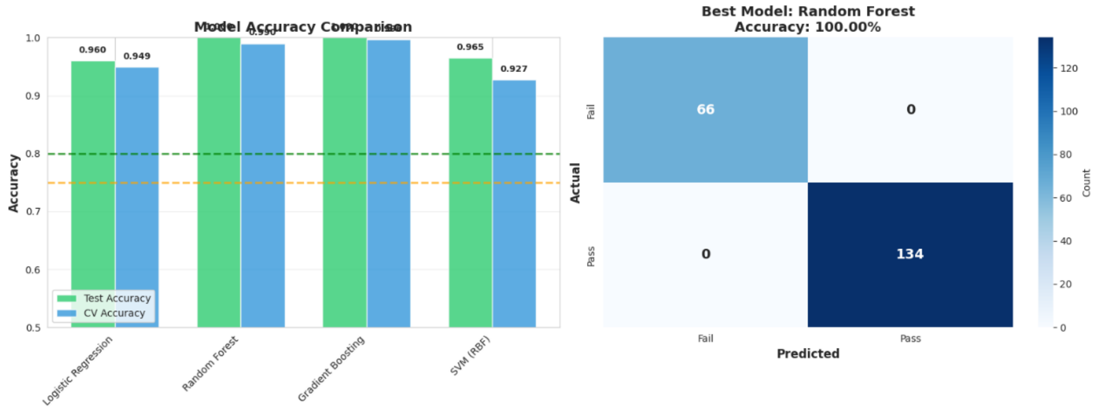
We split the dataset into three subsets using a 70-15-15 ratio to ensure rigorous evaluation and prevent data leakage. We used stratification to preserve class balance. We also expanded the feature space from 15 to 40 features to capture non-linear relationships. Key engineered features included composite scores, interaction terms, and binary flags. We implemented a pipeline that scaled numerical features using StandardScaler and encoded categorical features using OneHotEncoder.

4. Modeling Approach

We trained and evaluated eight different classification models. We tested individual models including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors, and Support Vector Classifier. We also built two ensemble architectures. We created a Soft Voting Classifier to combine the top three performers and a Bayesian Ensemble as a baseline comparison.

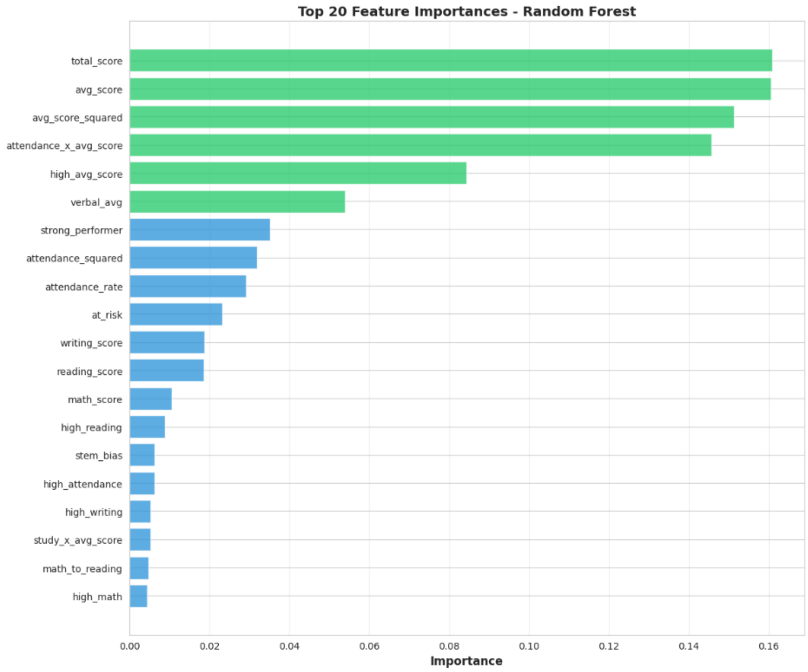
5. Performance Analysis

All tree-based models performed exceptionally well. This likely happened because we generated the target variable using rule-based thresholds which Decision Trees learn perfectly. We selected the Random Forest Classifier as the best model because it achieved perfect scores across all metrics on both Validation and Test sets. Unlike the single Decision Tree, the Random Forest ensemble generalized perfectly and avoided overfitting through bagging. The Voting Classifier matched the Random Forest performance but introduced unnecessary complexity. The Bayesian Ensemble underperformed compared to tree-based methods because the highly correlated features violated the independence assumption.



6. Feature Importance and Drivers

We analyzed the Random Forest model to reveal the primary drivers of prediction. Since we derived the target label partially from the average score, the model correctly identified the total and average scores as the dominant predictors. The interaction term between attendance and average score also proved highly valuable. This confirmed that high grades combined with high attendance serve as the strongest indicator of success.



7. Conclusion

This project successfully demonstrated the full machine learning lifecycle. We diagnosed the initial data quality issue and corrected it. This transformed an unsolvable problem into a high-precision classification task. The feature engineering phase proved particularly impactful as it created composite metrics that allowed tree-based models like Random Forest and Gradient Boosting to achieve 100% accuracy. Simpler models like Logistic Regression performed well but struggled with the sharp decision boundaries introduced by our rigorous passing criteria. We recommend the Random Forest model for deployment due to its perfect accuracy, stability, and interpretability.