TEAM REPORT – Final Machine Learning Project

Student Academic Performance Prediction System
Team: Peter Amoye – Erick Banegas – Alhassane Samassekou
Course: ITAI 1371
Date: December 2025

1. Introduction

This project focuses on developing a Machine Learning system capable of predicting whether a student will **Pass** or **Fail** based on a combination of academic, behavioral, and socio-educational factors. By analyzing patterns in student performance, attendance, study habits, and family background, the model aims to support early prediction and intervention strategies within academic environments.

The goal of the team was to design a complete end-to-end ML pipeline, including:

- Dataset analysis and preparation
- Exploratory Data Analysis (EDA)
- Feature engineering and preprocessing
- Testing multiple ML algorithms
- Model comparison and selection
- Evaluation and interpretation
- Team collaboration and technical documentation

This report summarizes the methodology, results, and collaborative effort of the entire group.

2. Dataset Overview

The dataset used in this project contains information from **1,000 students**, described through **14 features**:

Academic Performance

- *math_score*
- *reading_score*
- *writing_score*

Behavioral Factors

- *attendance_rate*
- *study_hours*

Socio-Educational Context

- *parent_education*
- *internet_access*
- *lunch_type*
- *extra_activities*

Target Variable

- **final_result** (Pass / Fail)

The dataset was clean, balanced, and contained no missing values, which allowed the models to learn patterns with high accuracy.

3. Exploratory Data Analysis (EDA)

The team performed an extensive EDA to understand the dataset deeply. Key findings include:

3.1 Academic performance strongly predicts final outcome

Math, reading, and writing scores showed the highest correlation with the final result.

3.2 Attendance rate matters

Students with higher attendance (above 95%) predominantly passed.

3.3 Study hours show a positive trend

Higher study hours correlate with improved performance.

3.4 Parent education influence

Students with parents holding Bachelor's, Master's, or PhD degrees showed higher pass rates.

3.5 Visualizations generated

- Histograms for all numeric variables
- Boxplots for score distributions
- Heatmap of correlations
- Performance comparisons by gender
- Scatterplots between study hours and academic scores
- Distribution of pass/fail categories

These visual insights guided the feature preprocessing decisions.

4. Preprocessing Pipeline

To prepare the dataset for machine learning, the team implemented a full preprocessing pipeline using **ColumnTransformer** and **Pipeline** from scikit-learn.

Steps included:

✔ **Encoding categorical features**

One-Hot Encoding for:

- gender
- parent_education
- lunch_type
- internet_access
- extra_activities

✔ **Scaling numerical features**

Standardization applied to:

- math_score
- reading_score
- writing_score
- attendance_rate
- study_hours

✔ Train–Test Split

Dataset was split into **80% training** and **20% test**.

This pipeline ensured a clean, consistent, and reproducible modeling process.

5. Machine Learning Models

The team trained and evaluated the following eight classifiers:

1. **Logistic Regression**
2. **K-Nearest Neighbors (KNN)**
3. **Support Vector Machine (SVM)**
4. **Random Forest**
5. **Decision Tree**
6. **Gradient Boosting**
7. **XGBoost**
8. **Naive Bayes**

All models were embedded inside pipelines to guarantee correct preprocessing during both training and prediction.

6. Model Evaluation and Comparison

Each model was evaluated using:

- Accuracy

- Precision
- Recall
- F1-Score
- ROC-AUC

Key Results (from Model_comparison_Group3.csv):

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	1.00	1.00	1.00	**1.00**	1.00
Gradient Boosting	0.99	1.00	1.00	**1.00**	1.00
XGBoost	0.99	0.99	1.00	0.995	1.00
Random Forest	0.99	0.99	0.91	0.947	0.991
SVM	0.99	1.00	1.00	1.00	1.00
KNN	0.98	0.95	0.89	0.908	0.96
Decision Tree	0.85	0.91	0.76	0.86	0.98
Naive Bayes	0.99	0.98	1.00	0.968	0.991

⭐ Best Overall Models

- Logistic Regression
- Gradient Boosting
- XGBoost

These models achieved near-perfect scores.
The team concluded that the dataset is **highly separable**, particularly due to the strong predictive power of academic scores.

7. Discussion of High Performance

The nearly perfect model performance may raise questions.
The team analyzed this carefully, concluding:

- Academic scores provide extremely strong signals.
- There is minimal noise in the dataset.
- The pass/fail boundary is almost linearly separable.
- There are no missing values or inconsistent entries.
- The features correlate strongly with the outcome.

Therefore, achieving F1-scores near 1.00 is reasonable.

8. Final Model Selection

Although Gradient Boosting and XGBoost performed exceptionally well, the team selected **Logistic Regression** as the final model because:

- It achieved perfect performance
- It is simpler and more interpretable
- It is easier to deploy in real systems
- It avoids overfitting concerns
- It gives clear weight coefficients for each feature

9. Team Collaboration

Each team member contributed in the following ways:

Erick Banegas

- Dataset cleaning and preprocessing
- EDA visualizations
- Model training and evaluation
- Documentation structure

Peter Amoye

- Model comparison logic
- Metric computation and ROC analysis
- Drafting the introduction and methodology

Alhassane Samassekou

- Dataset exploration and category mapping
- Pipeline implementation
- Exporting CSV results and writing conclusions

The team collaborated through shared notebooks, GitHub synchronization, and direct communication to ensure consistency and quality.

10. Conclusion

This project successfully demonstrates the complete lifecycle of building a machine learning system—from data exploration and preprocessing to model experimentation, evaluation, and final deployment decisions.

The Student Academic Performance Predictor proved highly effective, producing near-perfect classification metrics. The results confirm that academic performance indicators, attendance, and study habits are strong predictors of student success.

This project reflects strong teamwork, solid understanding of machine learning techniques, and the ability to apply theoretical concepts to a real-world educational dataset.