### Executive Summary

**Project Title:** Predicting Student Performance Using Machine Learning

### 1. Business Problem

Academic institutions often face challenges identifying students who are at risk of failing before it's too late. Traditional methods rely heavily on periodic assessments, which may come after critical learning gaps have already widened. Educators need early, data-driven signals to intervene and provide academic support proactively.

The primary goals of this project were to:

* Predict student outcomes (pass/fail) using early-term data
* Analyze the impact of academic, behavioral, and demographic features
* Design an interpretable and scalable ML model for educational insights
* Support data-informed strategies for student retention and success

### 2. Dataset Overview

We used the Student Performance Dataset from the UCI Machine Learning Repository, which includes data for 395 students enrolled in Math and Portuguese courses. The merged dataset includes:

* **Demographics**: Age, gender, family background, parental education
* **Academic & Behavioral**: Study time, past failures, absences, health, free time
* **Lifestyle Factors**: Alcohol use, travel time, romantic relationships
* **Target**: Binary classification – pass (1) or fail (0)

To avoid grade leakage, G1 and G2 scores were intentionally excluded from the features.

### 3. Methodology

Our approach followed a structured CRISP-DM framework:

**a. Data Preparation & Cleaning**

* Merged two course datasets for feature richness
* Removed duplicates, encoded categorical variables, handled imbalance using SMOTE

**b. Feature Engineering**

* Created new features such as studytime × failures, absences/age
* Normalized skewed distributions to improve model performance

**c. Model Development**

* Compared Logistic Regression, Decision Tree, and Random Forest
* Applied class weighting and threshold tuning to enhance minority class recall
* Evaluated models using F1 Score, Accuracy, and Recall

**d. Explainability**

* Implemented SHAP and LIME to make feature influence interpretable and actionable

### 4. Key Findings

* **Best Model**: Random Forest with tuned threshold = 0.60
  + Accuracy: 71.4%
  + Recall (Fail): 60%
  + F1 Score (Fail): 0.52
* **Top Influential Features**:
  + Past academic failures
  + Number of absences
  + Study time
  + Weekend alcohol consumption
  + Parental education level
* **Behavioral Patterns**: Students with consistent absences and low study time showed a higher likelihood of failing.

### 5. Recommendations

* **Early Warning System**: Deploy our model to flag at-risk students in the early phase of a semester
* **Targeted Interventions**: Guide counseling and tutoring based on key risk factors
* **Data-Driven Policy**: Inform school-wide strategies around attendance and behavioral health monitoring
* **Ethical Use**: Ensure predictions are used constructively and transparently, not to penalize students

### 6. Tools & Technologies

| **Component** | **Tools Used** |
| --- | --- |
| Data Handling | Pandas, NumPy |
| Modeling | scikit-learn, imbalanced-learn |
| Explainability | SHAP, LIME |
| Visualization | Matplotlib, Seaborn |
| Development Env. | Jupyter Notebook, Google Colab |

### 7. Conclusion

This project highlights how machine learning can be applied in education to improve outcomes. By analyzing non-grade indicators, our model offers a proactive tool for identifying struggling students before it's too late. The insights drawn can help shape support programs, reduce dropout rates, and ultimately contribute to student success.

In the future, the model can be expanded with time-series performance data, integrated into real-time academic dashboards, and generalized across different educational institutions.