Student Success Prediction Project

# Objective

This project focuses on building AI models to support academic decision-making by predicting two critical outcomes:  
1. First Year Persistence: Whether a student continues to the second year (Classification: Yes/No)  
2. Second Term GPA: The expected GPA score in the second term (Regression: Continuous Value)  
  
We use the same student dataset for both tasks, enriched with academic, demographic, and preparatory features.

# Problem 1 – First Year Persistence (Classification)

## Target Variable

First Year Persistence: Binary value (0 = Did not continue, 1 = Continued)

## Recommended AI Models

We aim to capture both linear and non-linear patterns, testing models of increasing complexity:  
- Logistic Regression: Baseline model to understand linear decision boundaries.  
- Support Vector Machine (SVM): Strong for high-dimensional spaces, sensitive to feature scaling.  
- Neural Network (Sequential API – TensorFlow): For deeper learning of complex interactions.  
- Recommended: Random Forest Classifier – Robust to overfitting, handles mixed feature types well, interpretable feature importance, and provides strong baseline performance on educational datasets.

## Evaluation Methods

- Accuracy: Proportion of correct predictions overall.  
- Precision: Measures how many predicted positives are truly positive.  
- Recall: Measures how many actual positives were correctly predicted.  
- F1-Score: Harmonic mean of precision and recall, useful for imbalanced classes.  
- ROC AUC: Measures model’s ability to distinguish between classes across different thresholds.

# Problem 2 – Second Term GPA (Regression)

## Target Variable

Second Term GPA: Numeric value between 0 and 4.5

## Recommended AI Models

We aim to capture both linear and complex non-linear relationships:  
- Linear Regression: Benchmark model for linear trends.  
- Support Vector Regression (SVR): Effective for non-linear relationships using kernel methods.  
- Neural Network (Sequential API – TensorFlow): Suitable for capturing complex score interactions.  
- Recommended: Gradient Boosting Regressor (e.g., XGBoost or LightGBM) – Offers superior performance on structured data, handles missing values, and provides model explainability.

## Evaluation Methods

- Mean Absolute Error (MAE): Average of absolute differences between predicted and actual values. Lower values indicate better performance.  
- Mean Squared Error (MSE): Penalizes larger errors more than MAE, sensitive to outliers.  
- Root Mean Squared Error (RMSE): Square root of MSE, in the same unit as the target variable.  
- R² Score: Proportion of variance in the dependent variable explained by the model (1.0 is perfect).