**Using Machine Learning for Early Prediction of Student Academic Performance**

**1. Introduction**

In education, early identification of struggling students is important, but most efforts come too late, often after a midterm exam or a failing final grade. By that time, a student's performance trajectory may be difficult or impossible to reverse. This project proposes an applied machine learning system that predicts student academic performance early in the term, using features like attendance, assignment submissions, and prior academic data. The goal is to help educators intervene earlier and more effectively, improving outcomes while reducing dropout and failure rates.

**2. Problem Statement**

Educational institutions often lack the tools to proactively support at-risk students. Traditional academic reporting systems are reactive, identifying issues only after performance has declined significantly. As a result, institutions miss the window for timely support, and students fall through the cracks.

This project addresses that gap by developing a predictive system that can detect early warning signs of poor academic performance — using machine learning models trained on historical data and early-term indicators.

**3. Objectives**

* Build a machine learning pipeline to predict student outcomes (pass/fail/risk level) using early-course data.
* Evaluate and compare models such as Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN).
* Use feature importance and model explainability (e.g., SHAP) to provide transparency into predictions.
* Deliver results in a user-friendly format suitable for educators, advisors, or academic dashboards.

**4. Methodology**

* **Data Collection**: Use publicly available datasets (e.g., UCI Student Performance) and/or anonymized institutional data.
* **Preprocessing**: Clean and encode data, manage missing values, engineer relevant features (e.g., assignment trends, attendance rates).
* **Modeling**: Train and test several classification algorithms (SVM, Decision Tree, Naïve Bayes, KNN) using k-fold cross-validation.
* **Evaluation**: Measure performance using accuracy, precision, recall, and confusion matrices.
* **Explainability**: Integrate tools to explain risk predictions for each student.
* **Output**: Create a lightweight dashboard or reporting template that lists at-risk students with suggested intervention triggers.