Early Warning System for Student Academic Risk Prediction

Abstract

This project proposes a machine learning-based Early Warning System to identify students at academic risk. Using performance indicators such as attendance, internal scores, assignment completion, study hours, participation, previous GPA, and missed submissions, the system predicts risk levels to support timely academic intervention. Traditional models including Logistic Regression, Decision Tree, and Random Forest were evaluated, with Random Forest selected for deployment due to its robustness and strong generalization capability. The solution integrates model training, periodic retraining, and an interactive dashboard to enable real-time predictions and data-driven decision-making.

Problem Statement

Universities often struggle to detect academically at-risk students early enough for effective intervention. Manual monitoring is reactive and time-consuming. This project addresses the problem by developing a predictive analytics system that automatically classifies students into risk categories, enabling proactive academic support.

Proposed Solution

The system analyzes student performance data using machine learning to generate risk predictions. A synthetic data generator simulates realistic academic records, ensuring continuous data availability while avoiding sensitive institutional data. Multiple models were trained and evaluated, and Random Forest was selected as the final model because ensemble methods reduce variance and improve predictive stability.

The solution includes a reusable training pipeline, model evaluation, feature importance analysis for interpretability, a retraining mechanism, and an interactive dashboard for analytics and prediction. This architecture ensures adaptability to evolving academic patterns while maintaining consistent performance.

Data Description

The dataset consists of synthetically generated student records reflecting realistic academic behavior. Key attributes include attendance, internal assessment scores, assignment completion rate, weekly study hours, previous GPA, participation level, missed submissions, and academic risk classification. The structured data enables effective training and testing of predictive models.

Model Training and Evaluation

Logistic Regression, Decision Tree, and Random Forest models were trained and evaluated using accuracy, precision, recall, and F1-score to ensure balanced predictive performance. Although the models demonstrated similar accuracy due to the structured dataset, Random Forest was chosen for deployment because of its reliability and superior generalization. Feature importance analysis identified previous GPA and attendance as the strongest predictors, improving model interpretability.

Dashboard and Predictions

An interactive dashboard was developed within the notebook environment to present risk analytics and enable prediction functionality. Users can input student parameters to receive immediate risk classifications along with confidence scores. This supports real-time inference and facilitates early identification of students requiring academic assistance.

Updation and Maintenance Timeline

The system follows a structured model lifecycle to maintain predictive effectiveness. As new student data becomes available, the training pipeline can be re-executed to retrain the model. Retraining is scheduled at the end of each academic term or triggered when performance declines, helping mitigate model drift. Regular maintenance, including database checks and dashboard validation, ensures reliable system operation.

Project Timeline

• Problem Definition & Proposal – Week 1

• Data Generation & Storage – Week 2

• Model Training & Evaluation – Weeks 3–4

• Retraining Pipeline Development – Week 5

• Dashboard Implementation – Week 6

• Testing & Final Optimization – Week 7

Expected Outcome

The project delivers a deployment-ready predictive system capable of reliably identifying students at academic risk. By integrating training, retraining, analytics, and inference into a unified architecture, the solution demonstrates the practical application of machine learning in education and supports proactive, data-driven academic decision-making.