Early Warning System for Identifying At-Risk Students Using Machine Learning

Name: Shreya Kirdat
Registration Number: 2024SEPVUGP0095

Abstract

Higher education institutions generate large volumes of academic data, including attendance records, internal assessments, assignment submissions, and prior academic performance. Despite the availability of such data, universities often rely on reactive approaches that identify struggling students only after examination results are declared.

This project proposes the design and deployment of an Early Warning System that leverages machine learning to categorize students into Low, Medium, and High academic risk levels before final evaluations. The system will utilize structured data stored in a SQL database, train and evaluate traditional machine learning models, and present actionable insights through an interactive dashboard. Additionally, periodic retraining will be incorporated to maintain predictive accuracy as new data becomes available.

The primary objective is to enable universities to shift from reactive academic monitoring to proactive, data-driven intervention strategies that support student success.

1. Problem Statement

Universities frequently face challenges in identifying students who are likely to experience academic difficulties before their performance declines significantly. Indicators such as low attendance, poor internal assessment scores, incomplete assignments, and reduced academic engagement are often present but remain underutilized due to the absence of automated predictive systems.

Most traditional monitoring methods rely on simple or binary classifications, which fail to capture the varying severity of academic risk. Institutions require a more structured approach that not only detects vulnerable students but also prioritizes them based on risk levels.

This project aims to develop a deployable machine learning-based Early Warning System that transforms institutional data into actionable insights by categorizing students into Low, Medium, and High risk groups. Such a system will assist academic administrators in making informed decisions and implementing timely intervention measures.

2. Data

The dataset for this project will either be sourced from publicly available educational datasets or synthetically generated to reflect realistic university conditions. A dedicated data generator script will be included in the project repository to ensure reproducibility and compliance with submission guidelines.

Proposed Dataset Attributes:

- Student ID
- Attendance Percentage
- Internal Assessment Scores
- Assignment Completion Rate
- Weekly Study Hours
- Previous Semester GPA
- Class Participation Level
- Number of Missed Submissions

Target Variable:
Academic Risk Level (Low / Medium / High)

All data will be stored in a SQL-based relational database such as MySQL or PostgreSQL to simulate real-world academic infrastructure and support scalable data management.

---

3. Trained and Tested Models

The project will implement and evaluate up to three traditional machine learning models:

- Logistic Regression
- Decision Tree
- Random Forest

Each model will be trained and tested using evaluation metrics such as accuracy, precision, recall, and F1-score. Comparative analysis will be conducted to identify the most suitable model for deployment.

The selected model will be serialized to enable seamless integration into the prediction system.

---

4. Dashboard

An interactive dashboard will be developed using Streamlit to effectively present analytical insights and predictive outcomes. The dashboard will include:

- Visualizations of academic performance trends
- Distribution of students across risk categories

- Model evaluation metrics
- Updates following model retraining

The dashboard will function as a decision-support interface, allowing academic administrators to quickly identify students who require attention.

---

5. Predictions

The deployed system will support real-time prediction capabilities through a dedicated interface within the dashboard. Users will be able to input student parameters, after which the model will generate a risk classification along with an associated probability score.

Additionally, the system will highlight key contributing factors influencing the prediction, improving interpretability and supporting data-driven academic decisions.

---

6. Updation and Maintenance Timelines

To ensure long-term effectiveness, the project will incorporate a structured model lifecycle strategy.

Weeks 1–2:
Dataset preparation, schema design, and SQL database setup

Week 3:
Data preprocessing, model training, testing, and evaluation

Week 4:
Model serialization and dashboard development

Week 5:
Integration of prediction functionality and system testing

Week 6:
Implementation of automated retraining workflow and project refinement

Final Phase:
Documentation, repository review, and deployment readiness

The model will be periodically retrained as new academic data is introduced, ensuring continued predictive relevance.

---

7. Git Repository

A public Git repository will be established to maintain version control and transparently document the development lifecycle. The repository will feature a structured folder hierarchy for data scripts, model pipelines, dashboard components, and documentation.

Large datasets and serialized model files will not be uploaded. Instead, the repository will contain a data generator script specifying either the download source or the synthesis methodology. Commit histories will clearly reflect incremental progress throughout the project timeline.

Expected Output

The proposed system is expected to deliver:

- A production-oriented machine learning pipeline
- Structured SQL-based data storage
- Trained and evaluated predictive models
- An interactive dashboard with real-time prediction capability
- Automated scripts for model training and retraining
- A publicly accessible Git repository demonstrating systematic development

Ultimately, the system will serve as a decision-support tool that enables early identification of academically at-risk students and promotes proactive institutional intervention