



Early Prediction of Students' Academic Performance using Machine Learning

Karishma R. Mujavar

Department of Computer Science & Engineering, KIT College, Kolhapur, Maharashtra, India

Samrudhi M. Bongale

Department of Computer Science & Engineering, KIT College, Kolhapur, Maharashtra, India

Siddhi S. Shah

Department of Computer Science & Engineering, KIT College, Kolhapur, Maharashtra, India

Anisha J. Shelake

Department of Computer Science & Engineering, KIT College, Kolhapur, Maharashtra, India

Dr. Deepali K. Jadhav

Associate Professor, Department of Computer Science & Engineering, KIT College, Kolhapur, Maharashtra, India

Abstract:

In most colleges, a student's academic performance issues only come to light after the damage is already done. The general thing is to wait till end semester results, but this leaves very little time for improvement. The research done shows that there are 3 different types of data concerning how students' grades have changed, how their behavioural factors are, and how the socio-economic factors affect their learning. We use this to flag students who may be heading toward academic trouble, early enough to actually do something about it.

The prediction model used is Logistic Regression, which produces a probability score for each student. This score is then classified into one of the four risk levels — Low, Medium, High, or Critical — using cutoff values of 0.25, 0.50, and 0.75. Four new features were created — a CGPI Progression Index tracking whether grades are going up or down over the years, a Study Focus Distribution checking how a student splits time across theory, practicals, assignments and revision, a Digital Accessibility Index based on whether the student has a phone, computer and internet at home, and a Socio-Economic Vulnerability Score built from family income and parental education levels. When a risk level is assigned, a built-in recommendation engine suggests specific actions the student or faculty should take. An AI layer using Groq's Llama 3.1 70B model is used to generate written explanations and study advice tailored to each student.

The whole system runs as a web application with separate views for students, teachers, and admins, built using FastAPI, MongoDB, and React. On training data, the model hit 93.54% accuracy, 99.58% ROC-AUC, and 98.32% recall. On unseen test data, accuracy held at 93.12% with ROC-AUC of 99.38%, showing the model generalises well without overfitting.

The reason for choosing Logistic Regression here is that it does not just perform well, it also explains itself. Faculty and admins can look at which factors drove a particular student's risk score, which builds trust in the system.

Keywords:

Academic Performance Prediction, Machine Learning, Early Intervention, Student Analytics, Risk Classification, Logistic Regression, Socio-Economic Risk Analysis.