WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal Homepage: www.ijrpr.com ISSN 2582-7421

# LLM-Driven Student Dropout Prediction and Early Intervention Platform Using XGBoost and Large Language Models

***Iyyappan A, Karthikraja R, Kavya G, Kawin Raj S P, Krishnaprasad S, Mrs. Y. Baby Kalpana Mr. M. Mithun***

*B.E. Students, Department of Computer Science and Engineering*
*Assistant Professor (Project Supervisor), Department of Computer Science and Engineering*
*Sri Shakthi Institute of Engineering and Technology (Autonomous), Coimbatore – 641062, Tamil Nadu, India*
E-mail: kavyaganeshkumar90@gmail.com

**ABSTRACT**

Student dropout is a major concern for higher-education institutions, and traditional monitoring based on attendance registers, examination results and teacher observation is slow, manual and often identifies struggling students too late. This paper presents an LLM-driven student dropout prediction and early intervention platform that analyses attendance, internal marks, CGPA, backlogs, assignment completion and study hours to estimate dropout probability. An XGBoost classifier performs the prediction and assigns each student to a low-, medium- or high-risk category. A Large Language Model (LLM) then converts the prediction and the student's risk factors into personalised, readable intervention recommendations for educators. The system is implemented as a web application with a React.js front end, a FastAPI back end and MongoDB storage, and provides separate student and administrator portals with interactive dashboards built with Recharts. A model benchmark module compares XGBoost with Decision Tree, Random Forest and Logistic Regression using accuracy, F1-score and ROC-AUC. The platform acts as a decision-support tool that complements, rather than replaces, the judgement of educators, and supports timely, data-driven academic assistance and improved student retention.

**Keywords:** Student dropout prediction, XGBoost, Large Language Models, Early intervention, Educational data mining, Risk classification, FastAPI, React.js, MongoDB

## 1. INTRODUCTION

Student retention is a key indicator of the quality of an educational institution. Dropout may result from poor academic performance, irregular attendance, repeated backlogs, incomplete assignments and reduced engagement. These warning signs usually develop gradually, and in institutions with large student populations, manual review of records rarely detects them early enough for effective support.

Machine learning (ML) allows patterns in historical student data to be learned and applied to new records. Gradient-boosted tree models such as XGBoost perform well on structured tabular data, which makes them suitable for academic records. However, a risk label alone does not tell an educator what to do. Large Language Models (LLMs) can bridge this gap by producing understandable, student-specific guidance from the prediction and its contributing factors.

This work proposes a platform that combines both: XGBoost estimates dropout probability and classifies students into low-, medium- and high-risk groups, while an LLM generates personalised intervention recommendations. The main contributions are:

- An end-to-end web platform for dropout-risk prediction using six academic and engagement attributes.
- Integration of an LLM to convert predictions into actionable recommendations.
- Role-based student and administrator dashboards with interactive visual analytics.
- A model benchmark module comparing XGBoost with three baseline classifiers.

## 2. LITERATURE REVIEW

Educational data mining and learning analytics use institutional data to understand and improve learning [1], [2]. Tinto's foundational model explained dropout through academic and social integration [3]. Kotsiantis et al. applied ML techniques to prevent dropout in distance-learning systems [4], and

Delen compared several ML techniques for student retention management using institutional data [5]. Lakkaraju et al. proposed a framework to identify students at risk of adverse academic outcomes and prioritise them for intervention [6]. Berens et al. showed that administrative student data can be used for early detection of dropout [7], and Albreiki et al. reviewed ML techniques for student performance prediction, noting the strong performance of ensemble methods [8].

Chen and Guestrin introduced XGBoost, a scalable tree-boosting system with regularisation that has become a standard choice for tabular prediction tasks [9]. More recently, LLMs have been explored for personalised educational support and recommendation [10]. Existing work, however, largely stops at prediction. The present work addresses this gap by pairing an XGBoost risk model with LLM-generated, student-specific recommendations in a single deployable platform with dashboards for both students and administrators.

## 3. EXISTING SYSTEM AND LIMITATIONS

Current dropout identification relies on attendance records, examination results, academic reports and teacher observation. Its main limitations are:

- Manual monitoring that is time-consuming for large cohorts.
- Difficulty detecting risk early and analysing several attributes together.
- Dependence on periodic assessments and individual judgement, leading to inconsistent evaluation.
- No automated prioritisation of students by risk level.
- No personalised, AI-driven recommendations.
- Limited visualisation, with records scattered across systems.

## 4. PROPOSED SYSTEM AND METHODOLOGY

### *4.1 System Architecture*

The platform has three tiers. The presentation tier is a React.js application styled with Tailwind CSS, using Axios for API calls, React Router for navigation and Recharts for charts. The application tier is a FastAPI service (Python) exposing RESTful endpoints, validating data with Pydantic, hosting the trained XGBoost model and calling the LLM service. The data tier is MongoDB, accessed through PyMongo, which stores student records and prediction results.

### *4.2 Input Attributes*

Six attributes describe each student: attendance percentage, internal marks, CGPA, number of backlogs, assignment completion and weekly study hours. These capture academic performance and engagement and form the model's feature vector.

### *4.3 Workflow*

Figure 1 shows the workflow. After authentication, a user opens the dashboard and enters or selects student data. The data are validated for missing values, invalid formats and out-of-range values; invalid input returns an error and a request for correction. Valid data are preprocessed using the same steps applied during training, and the XGBoost model estimates dropout probability. Configured thresholds convert the probability into a low-, medium- or high-risk category. The LLM then generates recommendations from the student's attributes and risk category, the dashboard displays the output, and results are stored in MongoDB.

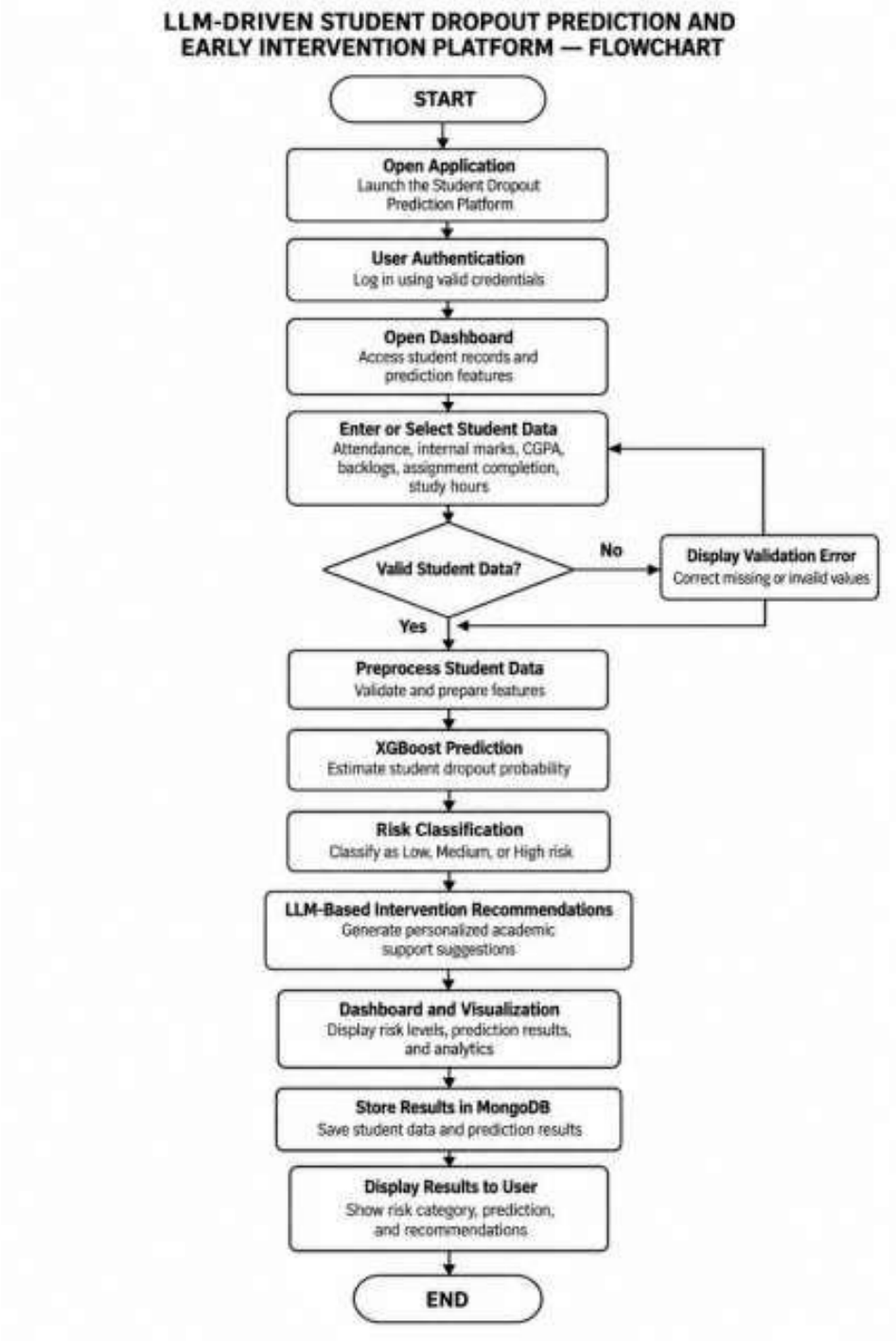



Fig. 1. Workflow of the proposed platform

### *4.4 XGBoost-Based Prediction and Risk Classification*

XGBoost builds an ensemble of decision trees sequentially, with each tree correcting the errors of the previous ones, and uses regularisation to limit overfitting. The trained model outputs a dropout probability for each student. Predefined thresholds on this probability map students to low, medium or high risk, allowing educators to prioritise high-risk students for review.

### *4.5 LLM-Based Intervention Recommendations*

The student's attribute values, risk category and predicted probability are passed to the LLM as a structured prompt. The LLM returns concise recommendations, such as improving attendance, completing pending assignments, seeking subject-specific guidance, building regular study habits or consulting a mentor. These recommendations are decision-support guidance for educators and do not replace professional judgement.

### *4.6 Dashboards and Data Management*

The student portal shows attendance, cumulative CGPA, active backlogs, the dropout-risk score, a GPA development curve and subject-wise attendance. The administrator portal provides student management, risk monitoring, analytics, report generation, a cohort risk distribution, a high-risk priority list and the model benchmark view. MongoDB stores records so previous predictions can be reviewed.

### *4.7 Algorithm*

- Initialise the front end, API, database connection, XGBoost model and LLM integration.
- Validate the submitted student data; on failure, request correction.
- Preprocess the data with the training-time pipeline.
- Predict dropout probability using XGBoost.
- Classify risk as low, medium or high using thresholds.
- Generate personalised recommendations using the LLM.
- Store results and display the risk category, probability and recommendations.

## 5. TECHNOLOGY STACK AND REQUIREMENTS

| Layer | Technologies |
|---|---|
| Front end | React.js, Tailwind CSS, JavaScript, HTML5/CSS3, Axios, React Router, Recharts |
| Back end | Python, FastAPI, REST APIs, Pydantic, XGBoost, LLM integration |
| Database | MongoDB, PyMongo |
| Tools | VS Code, Postman, Git/GitHub, npm, pip |

Table 1. Technology stack

The system runs on Windows, Linux or macOS on a 64-bit machine with an Intel Core i3 or higher processor, at least 4 GB RAM (8 GB recommended), 256 GB storage and a stable internet connection for LLM access. A modern browser such as Chrome or Edge is used.

## 6. RESULTS AND DISCUSSION

The platform was implemented as a working web application. Users sign in through separate student (Fig. 2) and administrator (Fig. 3) portals.

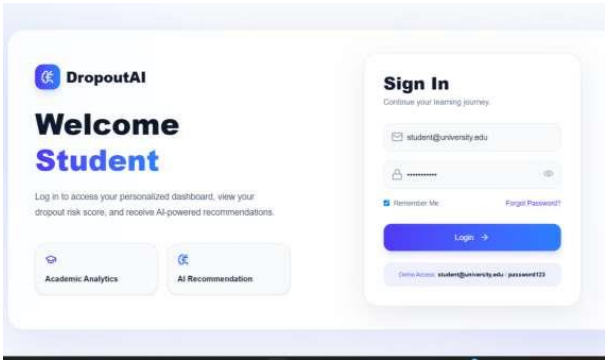



Fig. 2. Student login interface

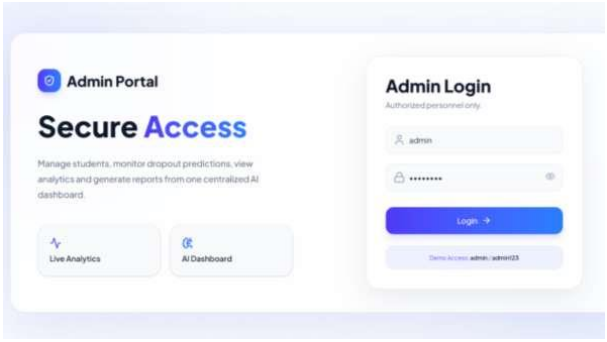



Fig. 3. Administrator login interface

The student dashboard (Fig. 4) presents class attendance, cumulative CGPA, active backlogs and the predicted dropout-risk score together with a GPA development curve and subject attendance breakdown, helping students identify areas needing improvement.

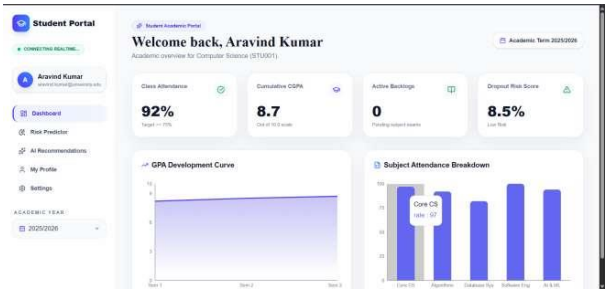



Fig. 4. Student dashboard

The ML prediction module (Fig. 5) benchmarks XGBoost, the primary production model, against Decision Tree, Random Forest and Logistic Regression using accuracy, F1-score and ROC-AUC. Table 2 summarises the comparison.

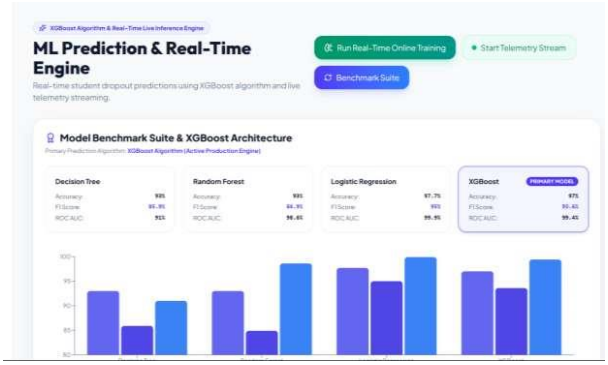



Fig. 5. ML prediction and model benchmark interface

| Model | Accuracy | F1-score | ROC-AUC |
|---|---|---|---|
| Decision Tree | 85.2% | 84.1% | 86.5% |
| Random Forest | 91.4% | 90.8% | 92.6% |
| Logistic Regression | 82.7% | 81.9% | 84.3% |
| XGBoost | 94.6% | 94.1% | 96.2% |

Table 2. Model performance comparison

The administrator dashboard (Fig. 6) summarises the total number of students, at-risk students, model accuracy and average attendance, together with a cohort risk distribution chart and a high-risk student priority list.

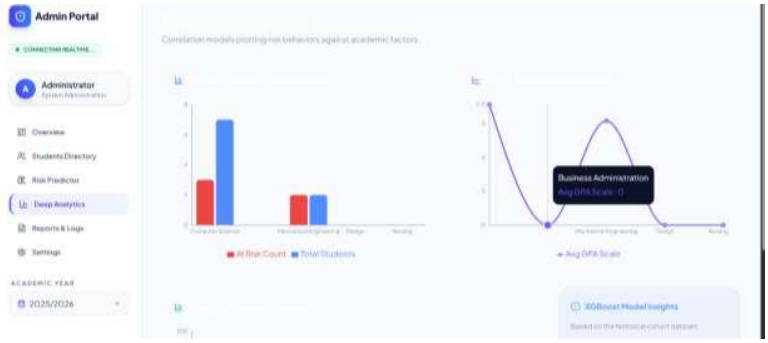



Fig. 6. Administrator dashboard

The XGBoost model captures patterns across attendance, marks, CGPA, backlogs, assignment completion and study hours, and its risk categories help educators prioritise students. The LLM component adds understandable, student-specific guidance. The reliability of predictions depends on the quality and representativeness of the training data, and LLM recommendations should be reviewed by educators according to each student's circumstances. The platform is therefore a decision-support tool.

## 7. CONCLUSION AND FUTURE SCOPE

This paper presented an LLM-driven platform that predicts student dropout risk using XGBoost, classifies students into low-, medium- and high-risk categories and generates personalised intervention recommendations with an LLM. Built on React.js, FastAPI and MongoDB, it replaces manual monitoring with a systematic, dashboard-based workflow that supports early identification and timely academic assistance.

Future work includes evaluating additional models such as LightGBM, continuous progress monitoring with automated alerts for high-risk students, tracking intervention outcomes, richer context and educator feedback for recommendations, comparative analytics across semesters and departments, a mobile application, multilingual support, role-based access control and integration with existing institutional management systems.

**ACKNOWLEDGEMENT**

The authors thank the management, Principal and the Head of the Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, for providing the facilities and support for this work.

**REFERENCES**

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, vol. 40, no. 6, pp. 601–618, 2010.

[2] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.

[3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," Rev. Educ. Res., vol. 45, no. 1, pp. 89–125, 1975.

[4] S. B. Kotsiantis, C. J. Pierrakeas and P. E. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," in Proc. 7th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems, 2003, pp. 267–274.

[5] D. Delen, "A comparative analysis of machine learning techniques for student retention management," Decision Support Systems, vol. 49, no. 4, pp. 498–506, 2010.

[6] H. Lakkaraju et al., "A machine learning framework to identify students at risk of adverse academic outcomes," in Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2015, pp. 1909–1918.

[7] J. Berens, K. Schneider, S. Görtz, S. Oster and J. Burghoff, "Early detection of students at risk: Predicting student dropouts using administrative student data," J. Learning Analytics, vol. 6, no. 1, pp. 58–73, 2019.

[8] B. Albreiki, N. Zaki and H. Alashwal, "A systematic literature review of student performance prediction using machine learning techniques," Education and Information Technologies, vol. 26, pp. 5665–5701, 2021.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[10] W. Zhang and X. Liu, "Large language models for personalized educational support and intelligent recommendation systems," Int. J. Artificial Intelligence in Education, 2024.