



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85254>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

AI-Driven Early Warning System for Predicting Student Dropouts

Illapu Sankara Srinivasa Rao¹, Pedumajji Kusuma Priya², Urlapu Anitha Sri³, Kurmana Sasank⁴, Pothala Anil Kumar⁵
Department of Computer Science & Engineering – Data Science, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

Abstract: Student dropout is a significant educational issue in India, influenced by regional differences and unequal resources. This paper introduces an AI-driven early warning system for predicting dropout risk at the state level using multi-year UDISE+ data. The proposed framework combines Logistic Regression, XGBoost, and CatBoost through a weighted ensemble strategy to enhance predictive accuracy. It includes automated preprocessing, feature engineering, classification, and time-series forecasting to classify states into Low, Medium, and High risk levels. An interactive Streamlit dashboard allows for visualization and supports decision-making. To connect prediction with action, a Community Learning Hub offers tailored learning and mentoring support for high-risk areas. The modular design allows for scalability and fosters data-driven educational planning.

Keywords: Student dropout prediction, Early warning system, Machine learning, UDISE+, Time-series forecasting.

I. INTRODUCTION

Education is an essential factor for the socio-economic development of a nation. Though school enrolment levels have shown improvement in India, dropout levels among students are a major problem that needs to be addressed for better employability and economic development. Hence, it is important to identify student dropout risks for better educational planning and intervention strategies.

The traditional methods of dropout risk analysis are based on statistical reports of past data. But these methods are not helpful in predicting new risks. Student dropout is a complicated issue that is affected by many factors, including a student's economic background, living conditions, and the availability of proper facilities and resources. Hence, data analysis is required for better decision-making for predicting student dropout risks.

The Unified District Information System for Education Plus (UDISE+) offers detailed data that helps in planning and analysing the education system in India. Though the data is comprehensive for analysis and decision-making, its predictive capabilities are not fully explored. Recent advances in artificial intelligence and machine learning have made it possible to analyse large amounts of educational data more effectively.

The main aim of this research is to build an AI-based early warning system that can predict the risk of student dropouts across different states in India, using data from UDISE+, to support better decision-making. The research aims to develop a predictive framework based on a weighted ensemble of Logistic Regression, XGBoost, and CatBoost machine learning algorithms for predicting student dropout risks for different states in India [6], [7]. In addition, time-series analysis is also carried out for better analysis of student dropout risks [8].

II. RELATED WORK

Student dropout has been examined through socio-economic analysis, statistical methods, and machine learning techniques. Early research pointed out the socio-economic effects of dropout. Sarker et al. demonstrated that high dropout rates hurt economic growth and the development of human capital [3]. Kumar et al. analysed adolescent dropout in India using UDAYA survey data and identified factors such as poverty, gender inequality, and parental education [2].

Recent studies have applied machine learning for dropout prediction. Selim and Rezk developed a classification-based early warning system [4], while Aulck et al. demonstrated that demographic and academic indicators can predict student attrition in higher education [5]. However, many approaches rely on single models and limited datasets.

Fig. 1 illustrates a typical dropout prediction framework based on historical data analysis. Despite existing work, limited research uses large-scale national datasets, ensemble learning, and integrated intervention systems. Therefore, this study proposes an AI-driven early warning system using multi-year UDISE+ data and ensemble machine learning models.

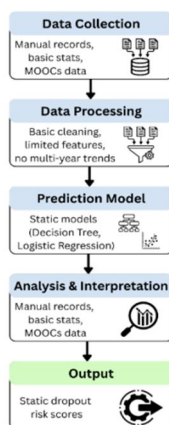


Fig. 1. Existing Dropout Prediction System

III. METHODOLOGY

This section presents an AI-driven system for predicting state-level student dropout risk using multi-year UDISE+ data [1], integrating preprocessing, classification, forecasting, and visualization (Fig. 2 and Fig. 3).

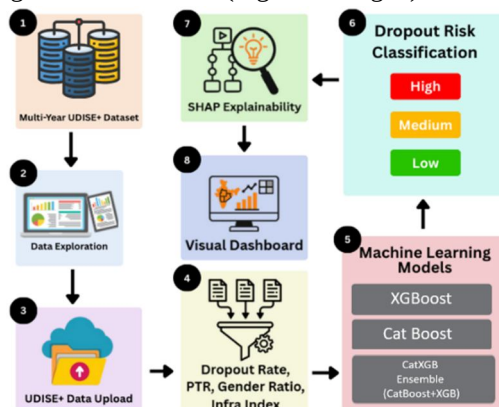


Fig. 2. Proposed Dropout Prediction System

System Architecture. The architecture is based on the three-tier architecture, which includes the data layer, processing layer, and presentation layer.

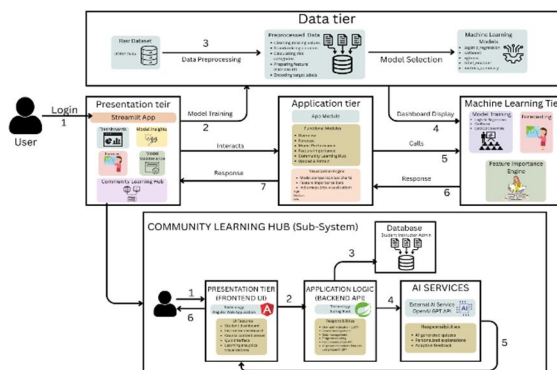


Fig. 3. Proposed System Architecture

Data Collection and Preprocessing. The data set is collected using the multi-year data sets using the UDISE+ data set. It includes data related to enrolment, dropout rates, availability of teachers, and infrastructure. Preprocessing is done using techniques such as normalization, feature engineering, and the calculation of the dropout rates, pupil-teacher ratio, and gender parity index.

Machine Learning Models. Different machine learning models can be used to predict dropout rates. In this study, logistic regression, XGBoost, and CatBoost are used for this purpose.

Ensemble Strategy. The results from each model are combined to make the predictions more accurate and improve overall performance. The ensemble strategy used in this work is the weighted ensemble strategy. The final prediction made by the ensemble strategy can be calculated as follows:

$$P_{final} = w_1 P_{LR} + w_2 P_{XGB} + w_3 P_{CB}$$

where, w_1, w_2, w_3 are the weights assigned to each model, and the weights should satisfy the following constraint:

$$w_1 + w_2 + w_3 = 1$$

This strategy improves the classification accuracy and reduces the variance of the predictions.

Forecasting and Trend Analysis. The system uses time series forecasting techniques like Prophet to perform the analysis of the dropout trends and forecast the risks.

Community Learning Hub. The proposed framework integrates the Community Learning Hub, i.e., Vidya Setu, which can be accessed via the Streamlit dashboard.

IV. EXPERIMENTAL RESULTS

This section shows how well the proposed AI-based early warning system predicts the risk of school dropouts. The performance of the models used in the system is measured using classification metrics.

Quantitative Results. Logistic Regression, XGBoost, and the proposed CatXGB ensemble model were tested using the processed UDISE+ dataset, and the results are presented in Table 1.

Table 1. Performance comparison of dropout prediction models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.6486	0.7221	0.6486	0.6730
XGBoost (Standalone)	0.8649	0.8688	0.8649	0.8506
CatXGB Ensemble (Proposed)	0.8919	0.8962	0.8919	0.8795

The proposed CatXGB ensemble model performs better than the other models, reaching an accuracy of 89.19% and an F1-score of 0.8795.

Performance Analysis. Fig. 4 shows the performance analysis of the proposed model. It compares the CatXGB ensemble model with the individual models, and the results show that the ensemble model performs more effectively than each individual model.

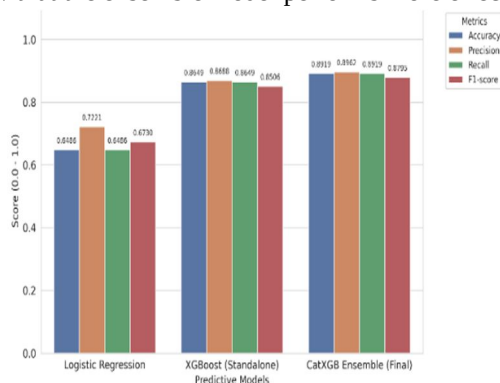


Fig. 4. Comparative analysis of model performance metrics for Logistic Regression, XGBoost, and the proposed CatXGB ensemble model.

The confusion matrix of the proposed model (Fig. 5) shows strong values along the main diagonal, indicating that the model correctly classifies most cases across the Low, Medium, and High dropout risk categories.

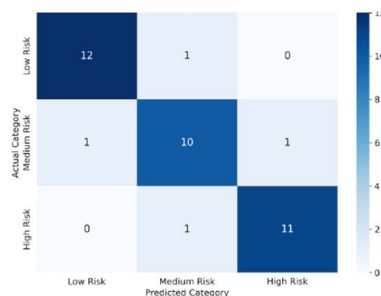


Fig. 5. Confusion matrix of the proposed CatXGB ensemble model for multi-class state-level dropout risk classification.

The feature importance analysis in Fig. 6 identifies key factors affecting dropout risk prediction results. It is observed that Total Enrolment and Pupil Teacher Ratio are significant factors in dropout risk prediction, followed by infrastructure indicators like Schools with Electricity and Total Schools.

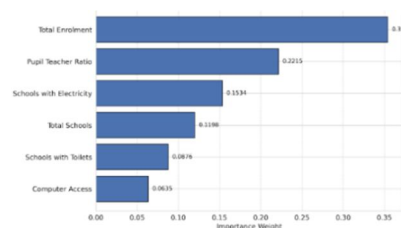


Fig. 6. Normalized feature importance of educational indicators in the CatXGB ensemble model.

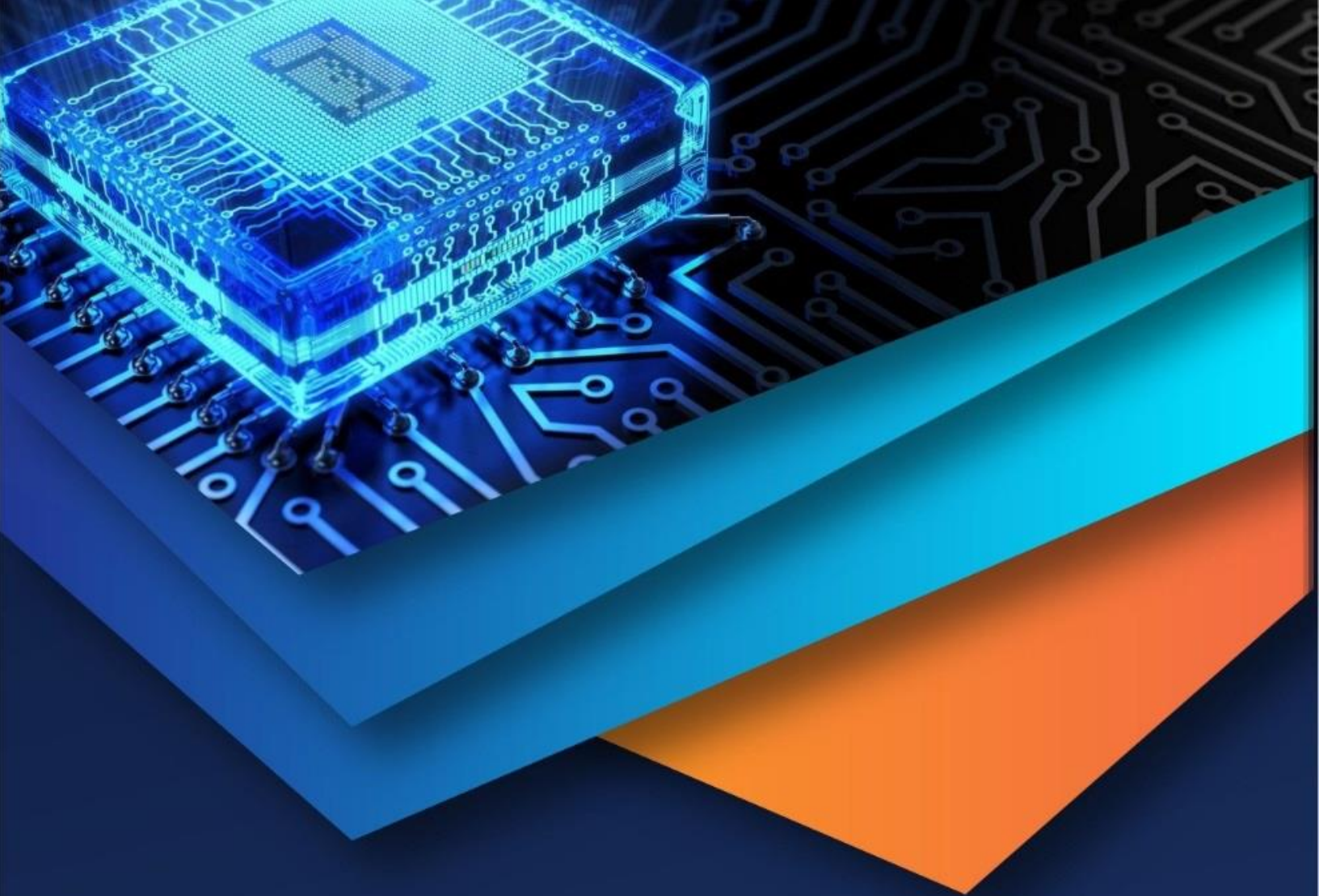
Overall, the results confirm that the proposed CatXGB ensemble framework is effective in improving both the accuracy and reliability of dropout risk predictions at the state level.'

V. CONCLUSION

This paper presents an AI-driven early warning system for predicting state-level school dropout risk in India using multi-year UDISE+ data [1]. The framework combines Logistic Regression, XGBoost, and CatBoost through a weighted ensemble to improve prediction accuracy, along with time-series forecasting for trend analysis [8]. The system is implemented using a Streamlit dashboard and integrated with the VidyaSetu Community Learning Hub to support targeted interventions. The experimental results show that the ensemble model performs better than the individual models, proving that it is effective for analysing large-scale educational data and supporting decision-making.

REFERENCES

- [1] Ministry of Education, Government of India provides the Unified District Information System for Education Plus (UDISE+), which can be accessed online at the official website.
- [2] P. Kumar, S.K. Patel, S. Debbarma, and N. Saggurti (2023) studied the factors that influence school dropout among adolescents in India in their paper published in PLOS ONE.
- [3] M.N.I. Sarker, M. Wu, and M.A. Hossin (2019) discussed the economic impact of school dropout in Bangladesh in the Journal of Social and Economic Development.
- [4] K.S. Selim and S.S. Rezk (2023) explored how machine learning can be used to predict school dropout in Egypt, published in Education and Information Technologies.
- [5] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West (2016) worked on predicting student dropout in higher education, presented at the ICML Workshop on Data for Education.
- [6] T. Chen and C. Guestrin (2016) introduced XGBoost, a scalable and efficient boosting system, at the ACM SIGKDD Conference.
- [7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin (2018) developed CatBoost, a boosting method designed to handle categorical features, presented at NeurIPS.
- [8] S.J. Taylor and B. Letham (2018) discussed large-scale forecasting methods in their work published in The American Statistician.
- [9] Streamlit Inc. (2023) provides Streamlit, a framework used for building interactive data applications.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)