International Journal of Scientific Research in Engineering & Technology
Volume5, Issue4 (July-August 2025), PP: 11-14.
https://www.doi.org/10.59256/ijsreat.20250504002
www.ijsreat.com
ISSN No: 2583-1240

AI In Academia: Forecasting Student Dropouts

Dr. Deepali Y. Kirange[1], Dr. Yogesh N. Chaudhari[2]
[1,2]*Assistant Professor, KCES's Institute of Management and Research, Jalgaon, Maharashtra, India.*

To Cite this Article: *Dr. Deepali Y. Kirange[1], Dr. Yogesh N. Chaudhari[2], "AI In Academia: Forecasting Student Dropouts", International Journal of Scientific Research in Engineering & Technology, Volume 05, Issue 04, July-August 2025, PP:11-14.*

Abstract: *Student dropout is a major challenge in higher education, often linked to academic, demographic, and socioeconomic factors. This study uses machine learning techniques to predict student dropout and support early intervention. A dataset containing attributes like attendance, CGPA, internet access, parental education, and socioeconomic status was analyzed. Data preprocessing, feature selection, and model evaluation were performed using algorithms such as Random Forest, Logistic Regression, and XGBoost. Among these, XGBoost achieved the highest accuracy of 87%. The findings show that data-driven models can effectively identify at-risk students, helping institutions make informed decisions to improve retention*

Key Word: *Student Dropout, Machine Learning, Predictive Analytics, Educational Data Mining, Data-Driven Decision Making.*

I.INTRODUCTION

Predicting which students may drop out allows schools and colleges to intervene early—offering academic help, counseling, or financial aid before issues escalate. Early action improves student retention, ensures more students complete their studies, and helps institutions use their support resources more effectively. This is supported by research showing how machine learning models can accurately flag at-risk students and guide targeted interventions [1].

Using real educational data—like test scores, attendance, engagement metrics, and demographic information—helps teachers and administrators make informed decisions instead of relying on guesses or intuition. This empowers personalized learning, supports timely interventions for struggling students, improves curriculum and resource planning, and leads to better overall student outcomes [2].

Student dropout negatively affects both individuals and educational institutions—causing wasted resources, lower completion rates, and broader social challenges. Early detection remains a challenge due to diverse factors like academic performance, socio-economic background, and activity metrics. Using data-driven methods can help accurately identify students at risk, enabling timely support and better retention outcomes [3].

II.LITERATURE REVIEW

Researchers have applied a variety of machine learning methods—such as logistic regression, decision trees, neural networks, SVM, and ensemble models—to predict student dropout both in schools and universities. For instance, a study using XGBoost on Portuguese higher-education data achieved over 90% accuracy in identifying at-risk students [4].

Another comparative analysis in Finland revealed that features like failed courses, LMS activity, and accumulated credits are among the strongest predictors, with models achieving high AUC scores [5].

Across multiple contexts, academic performance has consistently emerged as the most influential factor in predicting dropout, often outperforming demographic or socio-economic features [6]. Researchers typically use supervised learning methods like logistic regression, decision trees, random forests, support vector machines (SVM), K-nearest neighbors (KNN), naive Bayes, and neural networks (NN) to predict student dropout. More recently, boosting algorithms (e.g. XGBoost, LightGBM, CatBoost) have become popular due to their high accuracy on structured educational datasets [6].

Despite progress, many studies rely on small or homogeneous datasets—often from one institution or region—making it hard to generalize results. There is also a lack of robust evaluation: some models use unrealistic data splits that artificially inflate accuracy, and few studies link predictions to actionable interventions. Finally, issues like imbalanced dropout data, limited use of school-level contextual variables, and insufficient explain ability remain under-addressed [7].

III.METHODOLOGY

3.1. Dataset Description

The dataset consists of academic and demographic data from enrolled students, including attributes such as student ID, gender, age, attendance percentage, CGPA, parental education, financial background, and dropout status (Yes / No). This dataset was synthesized or collected from institutional records while ensuring anonymity and privacy. The following table 3.1 shows the sample of the dataset.

Table 3.1 Sample Dataset

Student ID	Age	Gender	Attendance (%)	CGPA	Socio economic Status	Internet Access	Parental Education	Extracurricular Participation	Dropout (Target)
1001	18	F	85	7.8	Medium	Yes	Graduate	Yes	No
1002	20	M	60	5.2	Low	No	High School	No	Yes
1003	19	F	92	8.5	High	Yes	Postgraduate	Yes	No
1004	21	M	45	4.9	Low	No	High School	No	Yes
1005	18	F	78	6.3	Medium	Yes	Graduate	Yes	No
1006	22	M	50	5	Low	No	None	No	Yes
1007	19	F	88	7.1	Medium	Yes	Graduate	Yes	No
1008	20	M	58	5.5	Low	No	High School	No	Yes

The attributes used in dataset are

- **Age:** Student's age (numeric)
- **Gender:** Male/Female
- **Attendance (%):** Overall class attendance percentage
- **CGPA:** Current Grade Point Average (scale 0 to 10)
- **Socioeconomic Status:** Low / Medium / High
- **Internet Access:** Yes / No
- **Parental Education:** Highest education level of parents
- **Extracurricular Participation:** Yes / No
- **Dropout: Target label** (Yes = student dropped out; No = student continued)

3.2. Data Preprocessing

The following techniques are used to preprocess the data.

- **Handling Missing Values:** Missing values in features like attendance or CGPA were imputed using mean or median strategies. Records with critical missing labels were removed.

- **Encoding Categorical Variables:** Variables such as gender, parental education, and course type were encoded using One-Hot Encoding or Label Encoding.

- **Normalization/Scaling:** Numerical features such as age, CGPA, and attendance were normalized using Min-Max Scaling to fit the 0-1 range, especially for distance-based algorithms like SVM and neural networks.

- **Balancing Classes:** Since dropout cases were underrepresented, SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the target classes.

3.3 Feature Selection

- **Correlation Analysis** was performed to eliminate highly correlated features.
- **Recursive Feature Elimination (RFE)** and **feature importance from Random Forests** were used to select the top features contributing to dropout prediction.

3.4 Model Training

There are several machine learning models were applied and compared on the dataset.

3.4.1 Logistic Regression

Logistic Regression is a statistical model used for binary classification. It predicts the probability that a student will drop out based on input features. It's simple, interpretable, and useful as a baseline model.

3.4.2. Decision Tree Classifier

A decision tree splits the dataset based on feature values to create a tree-like structure. It helps understand how different student attributes (e.g., attendance, CGPA) lead to a dropout decision. It is prone to overfitting but interpretable.

3.4.3. Random Forest

Random Forest is an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce over fitting. It performs well in handling missing data and provides feature importance scores for dropout prediction.

3.4.4. Support Vector Machine (SVM)

SVM finds the optimal boundary (hyperplane) between students who drop out and those who don't. It works well in high-dimensional spaces, especially when data is not linearly separable, though it's computationally intensive.

3.4.5. XG Boost (Extreme Gradient Boosting)

XG Boost is a powerful ensemble technique that uses gradient boosting. It builds decision trees sequentially and optimizes the error function. XG Boost achieved the highest accuracy in this study due to its robustness, speed, and ability to handle class imbalance.

3.4.6. Neural Network (Multi-Layer Perceptron - MLP)

MLP is a feedforward artificial neural network. It learns complex patterns by adjusting weights through backpropagation. Although more complex, MLP showed good performance in capturing nonlinear relationships among features.

IV.RESULT AND DISCUSSION

Models were evaluated using parameters Accuracy, Precision, Recall and F1-Score. A **5-fold cross-validation** technique was used to ensure reliability of the results.

Model	Accuracy	Precision	Recall	F1-Score	Strength
Logistic Regression	78%	75%	72%	73%	Simple and interpretable
Decision Tree	80%	77%	75%	76%	Easy to visualize
Random Forest	85%	83%	81%	82%	Robust and handles feature noise
SVM	81%	78%	80%	79%	Effective in high-dimensional data
XG Boost	87%	85%	84%	84.50%	Fast and high-performing
Neural Network	86%	83%	85%	84%	Captures complex relationships

In this study, six machine learning models were evaluated for predicting student dropout using academic, demographic, and socioeconomic data. Among the models, XG Boost achieved the highest performance with an accuracy of 87%, precision of 85%, recall of 84%, and an F1-score of 84.5%, demonstrating its ability to handle complex data patterns and imbalanced classes effectively. The Neural Network (MLP) closely followed, achieving 86% accuracy and strong generalization capability due to its capacity to learn non-linear relationships.

The Random Forest model also performed well, with 85% accuracy and robust feature handling, making it a reliable choice for dropout prediction. Traditional classifiers such as Logistic Regression and Decision Trees showed moderate performance, with accuracies of 78% and 80%, respectively. Although these models are easier to interpret, they struggled with capturing more complex interactions among features. The SVM classifier achieved 81% accuracy, showing a balance between precision and recall and proving effective in high-dimensional feature spaces.

Overall, the results indicate that ensemble methods like XGBoost and Random Forest outperform simpler models in predicting dropout risk, especially when the dataset includes diverse and non-linear attributes. These findings support the use of advanced machine learning techniques for informed decision-making in educational environments.

V.CONCLUSION

This study demonstrates the potential of machine learning models in accurately predicting student dropout rates using diverse academic and socioeconomic features. By evaluating multiple algorithms, it was found that XGBoost provided the highest predictive performance, followed closely by Neural Networks and Random Forest, highlighting their suitability for handling complex and imbalanced educational data. Simpler models like Logistic Regression and Decision Trees, while easier to interpret, were less effective in capturing intricate data relationships.

The outcomes of this research can assist educational institutions in developing early intervention strategies to support at-risk students and improve overall retention rates. Future work could explore the integration of real-time data, explainable AI techniques, and cross-institutional validation to further enhance model reliability and impact in real-world educational settings.

References

1. Lee, J., Kim, M., Kim, D., & Gil, J.-M. (2021). Evaluation of predictive models for early identification of dropout students. Journal of Information Processing Systems, 17(3), 630–644. https://doi.org/10.3745/JIPS.04.0218
2. Fernandes, J. (2023). The role of data-driven decision-making in effective educational leadership. Academy of Educational Leadership Journal, 27(S2), 1–3.
3. Kim, S., Yoo, E., & Kim, S. (2023). Why do students drop out? University dropout prediction and associated factor analysis using machine learning techniques (arXiv: 2310.10987). arXiv. https://doi.org/10.48550/arXiv.2310.10987
4. Alhardi, A., & Alan, S. (2024), Predicting Student Dropout in Higher Education Using Machine Learning Techniques: A Predictive Model Using XGBoost Algorithm. International Conference on Engineering Technologies (ICENTE'24).
5. Vaarma, M., & Li, H. (2024), Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, 76, Article 102474. https://doi.org/10.1016/j.techsoc.2024.102474
6. Albugami, S., Almaghrabi, H., & Wali, A. (2024), from data to decision: Machine learning and explainable AI in student dropout prediction. Journal of e-Learning and Higher Education, 2024, Article 246301. https://doi.org/10.5171/2024.246301
7. Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. Data Science Journal, 18, *1–23.* https://doi.org/10.5334/dsj-2019-014