International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 341-351 March 2026-Sept 2026 www.csjournals.com

Predicting Students' Academic Progress and Dropout: A Comparative Study of Random Forest and XGBoost Models

Neha Rani, Dr. Anupam Bhatia, Dr. Naveen
MCA Student, Associate Professor, Asst. Prof.
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

In today's world, Student dropout rest a critical problem in higher education. Many students leave school before finishing their degree, which affects their future jobs and life opportunities. At the same time, teachers and school managers do not know which student is at risk until it is too late for help. This research uses machine learning to predict whether a student will dropout or not. Two significant ML algorithms, Random Forest, and XGBoost, were employed in this study. This research is based on a dataset called Predict Student's Academic Progress and Dropout which is composed of 4,424 rows and 36 attributes. The findings indicate that both models are excellent in predicting students' dropout and academic progression. According to the results, XGBoost was capable to categorize the visuals with the accuracy of 88.31%, while the Random Forest was capable to categorize the visuals with the accuracy of 86%. The highest predictive factors for student failure were the attendance rate, admission grade, grade point average in the first semester of college attendance, age at enrollment, previous qualification grade, and number of subjects that were failed. The results of these findings can inform the schools to develop their early warning systems which can detect struggling students early and offer support in time. XGBoost produces the best prediction accuracy by comparing the aforementioned algorithms.

Keywords- Student dropout, Academic progress, Machine learning, Random Forest, XGBoost, Educational Data Mining, Early Warning System.

1. Introduction

Dropout of higher education has been a major issue all over the world both in individuals and institutions. Students who do not complete their educational program may have economic and social issues. High school dropout rates can affect the institutions' reputation and efficiency. This study shows that the OECD 23 countries' average drop

out rate for undergraduates is 33 percent, highlighting the need to address the issue [1].

Understanding the behaviour of successful students is essential at universities. One of the major concerns for them is the need to predict dropout characteristics [2].

The research is based on Educational Data Mining and Machine Learning. The major aim is to predict whether a student will continue their studies or drop out.

Educational Data Mining helps us analyse student data and identify useful patterns, while Machine Learning enables us to make predictions based on that data.

A dataset that contains various types of student behavior such as Academic performance behaviour, Attendance & Participation behaviour, Financial behaviour, Family & social background behaviour, Learning capability behaviour, Dropout risk behaviour, Socio-Economic behaviour, Demographic behaviour. In this dataset, various preprocessing steps, including duplicate removal, feature engineering, encoding, and scaling, were applied.

With the rapid growth of student dropout, understanding student behavioral intentions has become very difficult. Whether a student will dropout or not, the educational institutions wants to predict it.

Traditionally, schools are identify struggling students only after they have already failed in exams or stopped attending classes. By then, it is too late to help them effectively. What we need is a system that can predict problems before they happen — an early warning system.

Here the machine learning comes in. ML is a field of computer science where computers learn patterns from data and make predictions. In this, we use two popular ML algorithms — XGBoost and Random Forest— to predict two things:

- Which students are seems to drop out of school
- How well a student will perform academically (their academic progress)

By comparing these two models, we aim to find out which one works better for predicting student outcomes, and which factors (like attendance, grades, or financial situation) are most important in making these predictions.

By comparing these algorithms, XGBoost achieved highest accuracy (~88%).

1.1 Motivation

The main ambition of this study is to understand the student behavior and to predict whether a student will dropout or not. At present, most of the students take admissions but they leave their studies midway. This is the main challenge for the educational institutions, machine learning offers an effective way to instantly analyze the data and improve prediction accuracy.

2. Literature Review

- A. Ridwan and A. M. Priyatno (2024), This paper focuses on the problem of student dropout in higher education and its poor effect on students as well as educational organizations. This work applies the XGBoost machine learning algorithm to the prediction of student's academic progress and dropout risk using academic, demographic, and socio-economic factors. The model showed the best performance with an 80:20 training and testing ratio, helping institutions identify at-risk students early for better support and retention[1].

- S. P. Patel and H. N. Patel (2025), This paper discuss the importance of predicting student academic performance to provide early support and improve educational results. It explains how machine learning techniques in EDM are used to predict outcomes such as graduation, dropout, and continued enrollment. The study also highlights the role of feature selection and class imbalance methods like SMOTE in improving prediction accuracy [2].

- J. A. Esquivel (2023), This paper highlights the importance of predicting students' future behavior to improve curriculum planning and provide academic support at the right timeIt aims at providing the explanation of how the educational data can be analyzed through Data Mining and ML models to detect the students at risk of poor performance or dropout. Also mentioned in the study is the influence of Virtual Learning Environments (VLE) to manage online learning and prediction of dropout rates in distance education [3].

- V. Realinho, J. A. Machado, L. M. T. Bapista, and M. Martins (2022), This paper will be discussing the issue of student dropouts and academic failures in higher education and its effect on the society and economic growth, which are the major problem in higher education.Issues of student dropouts and academics failures in higher studies have been the major problem and their effects on the society and economic growth are discussed in this paper. It merges data from three categories – demographics, socio-economic, academic, and economic – to form a rich dataset suitable for analysis. The study emphasizes the significance of machine learning algorithms to forecast student performance and early identification of students at risk. It also conveys the importance of socioeconomic conditions on dropout rates and discusses issues such as imbalanced data and multicollinearity in educational studies [4].

- A. O. Ameen, M. A. Alarape, K. S. Adewole (2019), This paper reviews in detail the existing research on student study related progress and dropout prediction. It categorizes and summarizes various approaches and aspects of past research, and summarizes the main issues and limitations in forecasting student outcomes. The study also sheds light on the limitations of current research and proposes future directions for the improvement of prediction models and educational analytics [5].

3. Methodology

3.1 Data Collection:

The dataset in this work is obtained from kaggle, which includes information about students and their academic performance. In our dataset there are 4,424 records and 37 features. The dataset is publicly available and was originally published by Realinho (2022) et al. on the UCI Machine Learning Repository. This dataset is ideal for our study because it contains a broad range of student information including academic performance, demographic data, and socioeconomic factors.

The dataset includes the following types of information about each student:

Table 1: Summary of Features Used in the Dataset

Category	Features	Example
Academic Performance	Grades in 1st and 2nd semester, number of approved subjects, number of failed subjects	GPA = 14.5 / 20
Attendance & Engagement	Attendance rate, number of classes attended, assignments submitted on time	85% attendance
Demographic Info	Age, gender, nationality, marital status	Age = 19, Female
Socioeconomic Factors	Parents' education level, parents' occupation, scholarship holder, tuition fees paid on time	Mother: university graduate
Enrollment Info	Course enrolled, application mode, daytime or evening attendance	Nursing, evening class

3.2 Data Preprocessing:

This is the most crucial step in machine learning. It is used to clean our data and improve the performance of our model.

Steps include: Handling missing values, remove duplicates convert the categorical data into numerical data and feature scaling.

3.3 Feature Selection:

In this step, select the important feature that impact student behavior.

Most important features in the dataset are:

- Semester grades
- Approved subjects
- Tuition fee status
- Debtor status

- Scholarship status
- Admission grades
- Previous qualification
- Age at enrollment
- Target

...

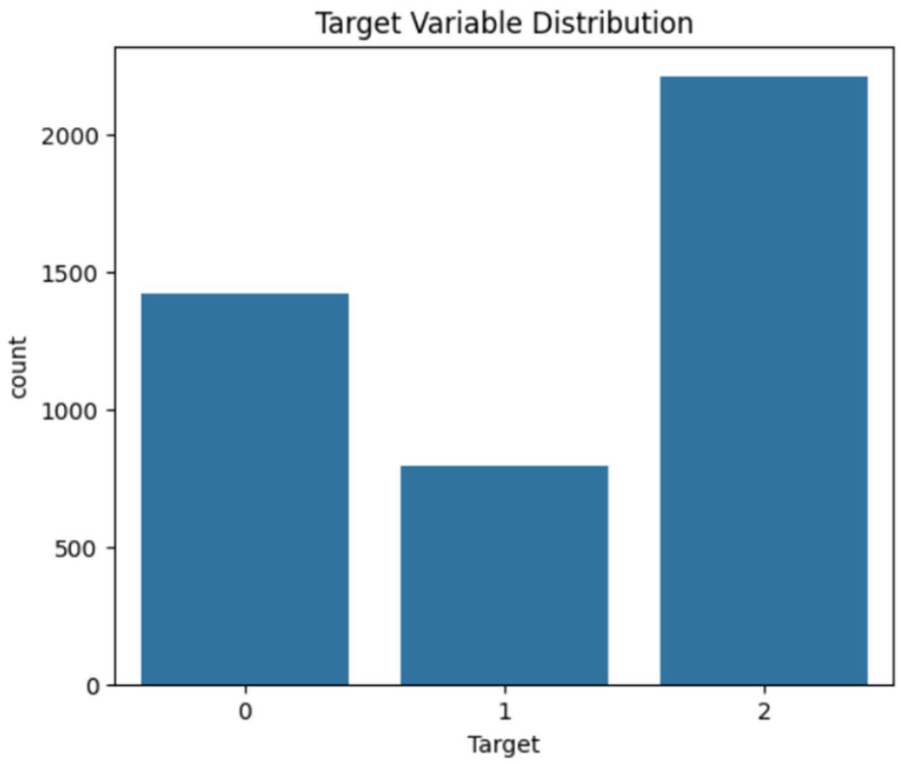

Figure1. Target Class Distribution (Graduate, Dropout and Enrolled)

3.4 Splitting:

Splitting data is an important stage in the design of a ML model. This study includes splits the dataset into two major parts: the training set and the testing set. The sections assigned for training limits from 50% - 90%. Whereas those for testing differ from 50% - 10% [1]. This phase is executed to ensure that the model can efficiently study from the most of the present records, Meanwhile delivering some separate data to assess the results of the model trained.

The dataset is splitting by the use of scikit-learn library, with the train test split function. The function unsystematically divide the dataset into two sections. The data is methodically splits into two sections, with 50 to 90% assigned for training and 50 to 10% designed to testing.

We split our dataset into two parts:

- Training set (80%): To train the ml model.
- Testing set (20%): To evaluate how effectively the ml model learned.

3.5 Machine Learning Models:

We used two ML models in this study. Let us explain each one in simple terms.

3.5.1 Random Forest

It is the most strong and widely used ML algorithm, that can used for both regression and classification. Random Forest builds many decision trees (each tree asks simple yes/no questions about the data) then it combines their results to make final prediction. Because it uses many trees instead of just one, it is much more accurate and makes less mistakes.

Key parameters we tuned for Random Forest:

- Number of trees: 200 (we found this gave the best results after testing different values)
- Maximum tree depth: 15 levels
- Minimum samples per leaf: 5

3.5.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGboost) is ML algorithm. which is based on decision trees, especially designed to improve the boosting process that assuring high performance computation and strong accuracy prediction[1]. XGBoost is another most powerful machine learning algorithm, but it works very different from the Random Forest. While Random Forest builds all the trees separately and in parallel, XGBoost builds trees one after another, where every new tree tries to valid the previous tree mistakes.

This process is called 'boosting' where each step boosts (improves) the performance. XGBoost is an ‘extreme’ version, which does this in a very improved and fast fashion. We fine-tuned some of the parameters for XGBoost.

3.6 Evaluation

The evaluation of model is an important step in the ML process, where the model is tested on data that has not been used for building the model, known as the test data. The model performance is assessed by four main metrics: precision, recall, F1-score and accuracy [1].

We used the following metrics as measures of the efficacy of our models: What percentage of the predictions were correct?

- The accuracy (90%, for example) indicates how often the prediction was correct (out of 100). Precision: What percentage of the students that were predicted to drop out, actually dropped out?
- Recall: What percentage of the students did the model correctly identify out of those who have dropped out?
- F1-Score: A combined score that balances recall and precision. It is beneficial in the event of an imbalance between classes.

4. Results & Discussions

In this section, we deliver the results of our research. We compare the outputs of Random Forest and XGBoost on our student dataset and identify the most important factors that predict dropout and academic progress.

4.1 Model Performance Metrics

The below table shows how both the models performed on the dataset:

Table 2: Performance Comparison of Random Forest vs XGBoost

Metric	Random Forest	XGBoost	Difference	Better Model
Accuracy	86.3%	88%	+1.7%	XGBoost
Precision	84%	87%	+3%	XGBoost
Recall	86%	88%	+2%	XGBoost
F1-Score	82%	86%	+4%	XGBoost

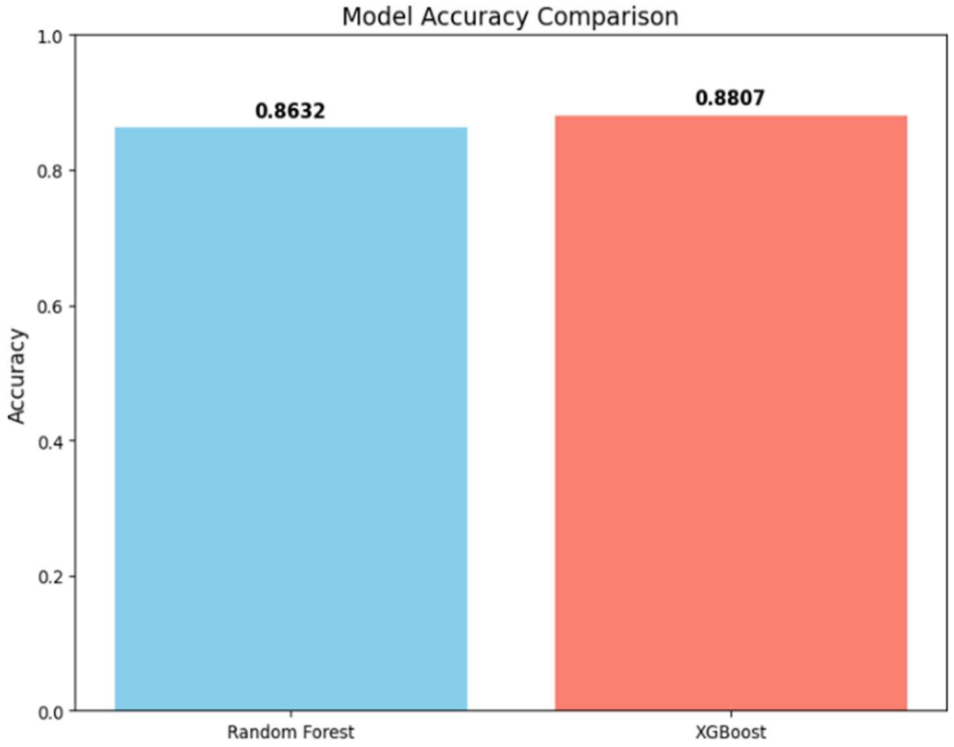

Figure 2. Model Accuracy Comparison

4.2 Most Important Features

One of the key element of machine learning is that it can tell us which factors are most important in making predictions. The figure below shows the top 12 most important feature identified by models:

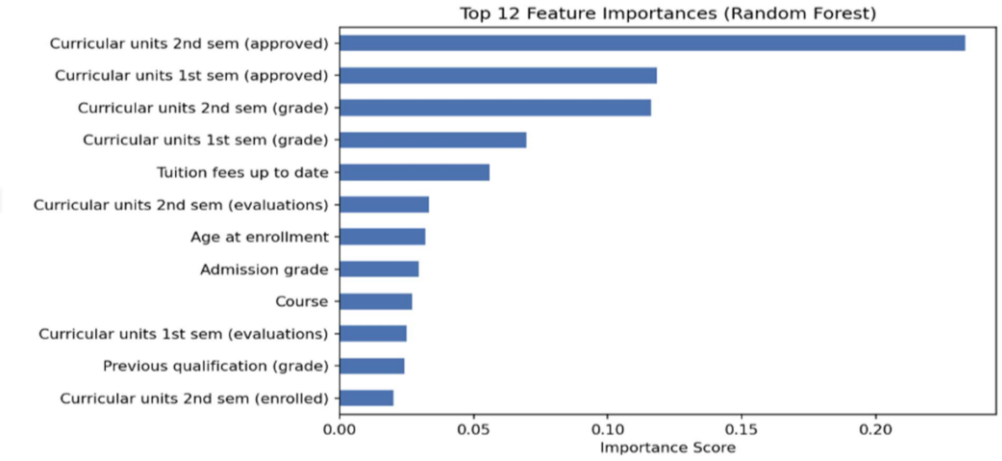

Figure 3. Top 12 Most important features for Dropout Prediction

4.3 Discussion

The learning obtained by using the XGBoost and Random Forest ml models can be a useful tool for higher studies teachers and educational professionals. A on time projection of students educational performance allows universities to carry out actions (e.g., assigning the session programs to student at risk of failure) [6].

Our results verify that the machine learning (especially XGBoost) can efficiently predict student dropout and academic progress with high accuracy.

5. Conclusion and Future Work

5.1 Conclusion

In this work, a model was created using some of the selected input variables. Among all the input variables, some of most important were chosen and then used to predict the student's educational performance. The student dataset was used to build a ML model to predict the students performance using two models: XGboost nd Random Forest . The data for this study comes from the Students' Dropout and Academic Progress Prediction dataset, consisting of 4,424 rows and 36 attributes. From the above analysis, we have concluded that the Random Forest achieves the accuracy 86% and XGboost achieves the accuracy 88%. Best results are achieved for the 80:20 split in the data, with a recall of 88%, precision of 87% and F1-score of 86%. The proposed methodology can be used to anticipate student performances and assist the teacher and student to improve the standard of learning and performance of students by making the right selection on the proper time.

5.2 Future Scope

Data could be further improved with additional details on the students and higher quality data that may improve the performance of the current model and also lead to more precise student achievements and determination of student response. Also, the work can be performed with other advanced techniques to obtain a general strategy and more trustworthy results. Another method to imbalanced class is SMOTE.

References:

[1] A. Ridwan and A. M. Priyatno, “Predict Students’ Dropout and Academic Success with XGBoost,” *Journal of Education and Computer Applications*, vol. 1, no. 2, pp. 1–8, Dec. 2024, doi: 10.69693/jeca.v1i2.13.
[2] S. P. Patel and H. N. Patel, “Student Success Through Algorithms: An Analytical Study of Machine Learning Models,” Jun. 2025, doi: 10.5281/zenodo.15623328.

[3] J. A. Esquivel, “Towards Predicting Student’s Dropout in Higher Education Using Supervised Machine Learning Techniques,” Mar. 2023, doi: 10.46254/an13.20230654.
[4] V. Realinho, J. A. D. Machado, L. M. T. Baptista, and M. Martins, “Predicting Student Dropout and Academic Success,” *Data*, vol. 7, no. 11, p. 146, Oct. 2022, doi: 10.3390/data7110146.
[5] A. O. Ameen, M. A. Alarape, and K. S. Adewole, “Students’ academic performance and dropout prediction,” vol. 4, no. 2, pp. 278–303, Nov. 2019, doi: 10.24191/MJOC.V4I2.6701.
[6] B. Perez, C. Castellanos, and D. Correal, "Applying data mining techniques to predict student dropout: A case study," in *Proc. 1st Colombian Conf. Appl. Comput. Intell. (ColCACI)*, Bogota, Colombia, IEEE, 2018.
[7] S. Bhutto, Q. A. Arain, I. F. Siddiqui, and M. Anwar, "Predicting students' academic performance through supervised machine learning," in *Proc. Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, Karachi, Pakistan, IEEE, 2020.
[8] V. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Coimbatore, India, IEEE, 2018.
[9] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, "Perspectives to predict dropout in university students with machine learning," in *Proc. IEEE Int. Conf.*, 2018.
[10] B. Sravani and M. M. Bala, "Prediction of student performance using linear regression," in *Proc. Int. Conf. Emerging Technol. (INCET)*, Belgaum, India, IEEE, 2020.
[11] H. P. Singh and H. N. Alhulail, "Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach," *IEEE Access*, vol. 10, pp. 6470–6482, 2022.
[12] C. E. L. Guarin, E. L. Guzman, and F. A. Gonzalez, "A model to predict low academic performance at a specific enrollment using data mining," *IEEE Access*, vol. 10, pp. 119–125, 2015.