

Predicting Student Dropout in Undergraduate Programmes: A Comparative Study of Machine Learning Models

Megha Srivastava

Dept. of Computer Applications,
Your University, Lucknow
megha@youremail.edu

Faizan Ahmed

Dept. of Statistics, Your University,
Lucknow
faizan@youremail.edu

Pooja Tiwari

Dept. of Computer Applications,
Your University, Lucknow
pooja@youremail.edu

Abstract— Student dropout is a serious concern for colleges, affecting students, families and institutions. This paper compares five machine-learning models for predicting which first-year undergraduate students are at risk of leaving their programme. Using anonymised academic, attendance and demographic records, we evaluate logistic regression, decision trees, random forests, gradient boosting and a small neural network. Gradient boosting gives the best balance of recall and precision, and attendance in the first eight weeks emerges as the strongest early indicator. The paper also discusses how such models can be used responsibly to offer support rather than to label students. All results in this template are sample values.

Index Terms— *educational data mining, dropout prediction, machine learning, early warning systems, higher education*

This is a template: all data, results and references are placeholders for the author to replace.

I. INTRODUCTION

Many students who enrol in undergraduate programmes do not complete them. Reasons include academic difficulty, financial stress, family responsibilities and lack of a sense of belonging. If institutions can identify students at risk early, mentors and counsellors can offer timely support.

Institutions already hold data that may signal risk, such as attendance, internal assessment marks and fee payment status. This paper asks which machine-learning models make the best use of such data for early prediction, and which features matter most.

The main contributions of this paper are as follows:

- A comparison of five common models on an anonymised dataset of first-year students.
- An analysis of the most important early indicators of dropout risk.
- Practical guidelines for using predictions ethically in student support programmes.

The rest of the paper is organised as follows. Section II reviews related work, Section III describes the proposed prediction framework, Section IV presents the experimental setup, Section V reports and discusses the results, and Section VI concludes the paper.

II. RELATED WORK

Statistical studies have linked dropout to prior academic performance and socio-economic factors [1].

Decision-tree models have been used to produce interpretable rules for advisers [2].

Ensemble methods have improved accuracy in several studies of online and distance learning [3].

Researchers have raised concerns about fairness and the risk of reinforcing bias in automated predictions [4].

Ref.	Approach	Data	Limitation
[1]	Logistic regression	Survey + marks	Linear assumptions
[2]	Decision tree	College records	Overfits small data
[3]	Random forest	Online courses	Less interpretable
[4]	Fairness audit	Mixed	No prediction model

TABLE I. Summary of related work

III. PROPOSED PREDICTION FRAMEWORK

The framework has four stages: data preparation, feature engineering, model training and interpretation. Records were anonymised before analysis, and students' names and identifiers were removed.

Features include attendance in the first eight weeks, internal assessment marks, entrance score, distance from home, scholarship status and participation in clubs.

A. Overview

Raw records from the academic, attendance and accounts systems are merged by an anonymous key, cleaned, encoded and split into training and test sets. Five classifiers are trained with cross-validation, and the best model is explained with feature-importance analysis.

Figure placeholder — block diagram of the proposed prediction framework

Fig. 1. Overview of the proposed prediction framework

B. Algorithm

1. Merge and anonymise records from different systems.
2. Handle missing values and encode categorical variables.
3. Balance the classes using oversampling of the minority class.
4. Train each model with five-fold cross-validation.
5. Evaluate on a held-out test set.
6. Explain predictions using feature importance.

IV. EXPERIMENTAL SETUP

The dataset contains records of first-year students over three admission years (sample size for this template: 2,400 records, of which 14% dropped out). Models were implemented in Python with scikit-learn. Hyper-parameters were tuned with grid search on the training folds.

Parameter	Value
Records	2,400 (sample)
Dropout rate	14% (sample)
Train/test split	80 / 20
Cross-validation	5-fold
Class balancing	SMOTE
Tool	Python, scikit-learn

TABLE II. Experimental parameters

A. Evaluation Metrics

- Recall for the dropout class (most important, to avoid missing at-risk students)
- Precision and F1-score
- Area under the ROC curve (AUC)

V. RESULTS AND DISCUSSION

Gradient boosting achieved the highest recall and AUC, followed closely by random forest. Logistic regression performed reasonably and remains attractive because it is easy to explain.

Model	Recall	Precision	F1	AUC
Logistic regression	0.71	0.62	0.66	0.81
Decision tree	0.68	0.58	0.63	0.76
Random forest	0.76	0.67	0.71	0.86
Gradient boosting	0.79	0.69	0.74	0.88
Neural network	0.74	0.64	0.69	0.85

TABLE III. Comparison with baseline methods (sample data)

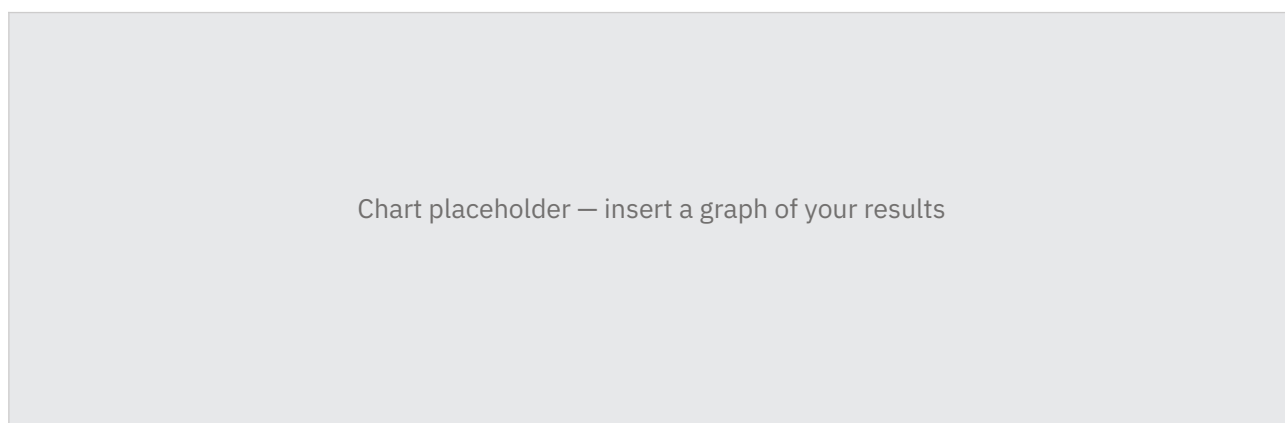


Fig. 2. Feature importance for the gradient boosting model (sample data)

Attendance in the first eight weeks was the most important feature, followed by first internal assessment marks and distance from home. This suggests that an early-warning check after the first two months could be effective.

Predictions should be used to offer help — mentoring, counselling, fee guidance — and never to penalise students. Mentors reviewed a sample of flagged cases and found most of them reasonable.

A. Ablation Study

To understand the contribution of each component, we removed one component at a time and repeated the evaluation. Table IV shows that every component contributes to the final result, with the largest drop observed when the attendance features is removed.

Variant	Recall
Full model	0.79
Without attendance features	0.66
Without pre-processing	0.74

Variant	Recall
Baseline only	0.61

TABLE IV. Ablation results (sample data)

B. Limitations

- Data came from a single institution.
- Reasons for dropout were not recorded, so causes cannot be inferred.
- Some students leave for positive reasons, such as admission elsewhere.

VI. CONCLUSION AND FUTURE WORK

This study compared five machine-learning models for early dropout prediction and found gradient boosting to perform best, with early attendance as the strongest signal.

Future work will test the models across institutions, include student well-being surveys, and evaluate the effect of support programmes triggered by the predictions.

Acknowledgment

The authors thank the department and the participants who helped with data collection. Replace this paragraph with your own acknowledgment, including any funding support.

References

[1] A. Author, B. Author, and C. Author, “Title of the journal article,” Journal Name, vol. 10, no. 2, pp. 101–112, 2023.

[2] D. Author and E. Author, “Title of the conference paper,” in Proc. Int. Conf. on Topic, City, 2022, pp. 45–50.

[3] F. Author, Title of the Book, 3rd ed. City: Publisher, 2021.

[4] G. Author et al., “Title of the article,” Transactions on Topic, vol. 7, pp. 220–231, 2024.

[5] H. Author, “Title of the thesis,” Ph.D. dissertation, Dept. Name, University Name, City, 2020.

[6] I. Author and J. Author, “Title of the preprint,” arXiv preprint, 2024.

[7] K. Author, “Title of the dataset or software,” Version 2.1, 2023. [Online]. Available: URL

[8] L. Author, M. Author, “Title of the survey paper,” ACM Computing Surveys, vol. 55, 2023.