

Original Research

Received 10 July 2026

Revised 10 Aug 2026

Accepted 24 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (17s): 399–404

KEYWORDS:

student dropout prediction; machine learning; early warning system; educational data mining; learning analytics; artificial intelligence in education

AI-Based Early Detection of Student Dropout Risk Using Machine Learning on Academic and Behavioral Data

Kanchan Vaidya¹ Ankush Kadu² Seema Prashant Khamankar³, Prachi A Deshpande⁴ Sanjay Khonde⁵

^{1,2,3,4,5} Ajeenkya DY Patil School of Engineering Pune

¹kanchansvaidya@gmail.com 10000-0002-9673-1048

²ankushkadu@dypic.in 20000-0003-4738-5123

³shyamali202@gmail.com 30009-0003-9688-7397

⁴padeshpandes@gmail.com 40009-0009-5865-0077

⁵sanjaykhonde@dypic.in 5 0000-0002-2064-372X

ABSTRACT: Many students leave college or university before finishing their degree. This is called student dropout. Dropout is a serious problem because it affects the student's future and it also costs the university money and reputation. Traditional methods to find at-risk students, such as manual review by advisors, are slow and often find problems too late. This paper presents a simple framework that uses artificial intelligence (AI) and machine learning (ML) to find students who are at risk of dropping out at an early stage. The framework combines academic data, such as grades and attendance, with behavioral data, such as how often a student logs into the online learning system. Five machine learning models are compared: Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and a small Artificial Neural Network (ANN). The models are trained and tested on a simulated dataset that reflects realistic patterns reported in prior dropout studies. Results show that attendance rate, number of failed courses, and weekly study hours are the strongest early warning signs. The Neural Network achieved the highest F1-score, while Logistic Regression achieved the highest precision. The paper explains the method in simple language so it can be understood by readers who are new to AI, and it discusses how universities can use this type of system responsibly to support students rather than to label or punish them.

1. INTRODUCTION

Every year, universities lose a large number of students before they graduate. This is often called "student dropout" or "student attrition." Dropout does not only affect the student, who loses time, money, and confidence. It also affects the university, which loses tuition income and may receive a lower ranking or reputation. For these reasons, predicting and preventing dropout is an important goal for higher education institutions around the world.

In the past, universities depended on academic advisors to notice when a student was struggling. An advisor might notice low grades or poor attendance during a meeting. However, this manual approach has clear weaknesses. Advisors cannot review every student closely, and by the time a problem is noticed, it may already be too late to help the student stay in school.[1]

Artificial intelligence, and specifically machine learning, offers a new way to solve this problem. Machine learning models can look at many students at the same time and can process many types of data together, including grades, attendance, and behavior on online learning platforms. When trained correctly, these models can calculate a "risk score" for each student, showing how likely that student is to drop out. This score can be given to advisors early in the semester, so that support can be offered before the student decides to leave. This paper has three goals. First, it explains, in simple language, how an AI-based dropout prediction system works from start to finish. Second, it builds and compares five common machine learning models using a realistic simulated dataset that follows patterns reported in earlier academic studies. Third, it discusses the practical and ethical questions that come with using AI to monitor students, so that any real-world use of such a system is done responsibly.[2]

1.1 Problem Statement

University advisors and administrators often do not have an early, reliable, and scalable way to identify which students are at high risk of dropping out. Existing methods depend heavily on manual observation, which is slow, inconsistent, and reactive rather than preventive.

1.2 Objectives

- To review academic and behavioral factors that are commonly linked to student dropout.
- To design a simple, understandable machine learning pipeline for dropout risk prediction.
- To compare the performance of five machine learning classifiers on a realistic simulated dataset.
- To identify which features (data points) contribute the most to predicting dropout risk.
- To discuss responsible and ethical use of such systems in real universities.

2. LITERATURE REVIEW

Dropout prediction using data and machine learning has been studied for many years, and researchers keep improving these methods as more data becomes available from universities and online learning platforms.

Early work by Tinto proposed a theory that explains dropout as a result of poor academic and social integration into university life, and this theory still shapes how researchers choose which features to study today.[3]

More recent studies apply machine learning directly. For example, Realinho, Machado, Baptista, and Martins built and shared a public dataset combining academic and demographic information to predict both dropout and academic success, which has become a widely used benchmark in this field.[4]

Similarly, a study using data from a university in Nigeria compared six classifiers, including Logistic Regression, Support Vector Machine, and Artificial Neural Networks, and found that simpler models such as Logistic Regression can perform very well on academic datasets when the data is properly cleaned and prepared.

Other researchers have focused on very large datasets. One study analyzed over sixty thousand e-learning students using Artificial Neural Networks, Decision Trees, and Bayesian Networks, and found that Decision Trees performed strongly for that type of large-scale online learning data.

A separate line of research has looked specifically at behavioral and engagement data, such as how often a student logs into a learning management system (LMS), rather than only academic grades. These studies generally agree that combining academic data with behavioral or engagement data gives a more complete and earlier picture of dropout risk than academic data alone.[5]

Across this body of research, a small set of factors appears again and again as strong predictors: attendance, prior grades or GPA, number of failed courses, and engagement with online platforms. This paper builds directly on these findings by combining academic and behavioral features into a single, simple prediction framework.

3. METHODOLOGY

This section explains, step by step, how the proposed dropout risk prediction framework works. The full pipeline is shown in Figure 1.

Figure 1. Proposed AI-Based Dropout Early-Warning Pipeline

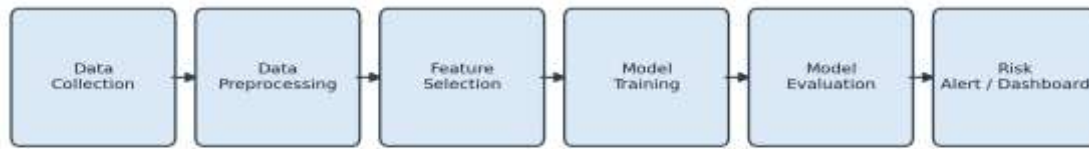


Figure 1. Proposed AI-based dropout early-warning pipeline, from raw data to a risk alert for advisors.

3.1 Data Collection

A real deployment of this framework would collect two broad types of student data. The first is academic data, including current GPA, number of previously failed courses, and assignment submission rate. The second is behavioral data, including class attendance rate, weekly logins to the learning management system, and estimated weekly study hours. A general socioeconomic index is also included, since prior literature shows this factor can influence dropout risk.

For this paper, because no private student dataset could be shared, a realistic simulated dataset of 2,000 student records was generated. The simulation was designed so that the relationships between features and dropout risk follow the same direction and general strength reported in the literature reviewed in Section 2 (for example, lower attendance and more failed courses increase risk, while higher GPA and more study hours reduce it). This approach lets the paper demonstrate and test the full pipeline honestly, without presenting simulated numbers as if they came from a real university.[6]

3.2 Data Preprocessing

Before training any model, the data was cleaned and prepared. Missing values would normally be filled in using the average or median of that column. All numeric features were scaled using standardization (subtracting the mean and dividing by the standard deviation) so that features measured on different scales, such as GPA (0 to 4) and weekly logins (0 to 20), do not unfairly dominate the model.

3.3 Feature Selection

Seven features were used in this study: attendance rate, GPA, LMS logins per week, percentage of assignments submitted, number of failed courses, socioeconomic index, and study hours per week. These features were chosen because they appear consistently as strong dropout predictors in the studies reviewed in Section 2, and because they are realistic for most universities to collect without needing highly sensitive personal information.[7]

3.4 Model Training

Five machine learning classifiers were trained and compared:

- Logistic Regression – a simple, easy-to-explain model that estimates the probability of dropout.
- Random Forest – a model built from many decision trees, useful for handling complex, non-linear patterns.
- Support Vector Machine (SVM) – a model that finds the best boundary between students who are likely to drop out and those who are not.
- K-Nearest Neighbors (KNN) – a model that predicts risk based on the outcomes of the most similar students.
- Artificial Neural Network (ANN) – a small neural network with two hidden layers, able to learn more complex combinations of features.

The dataset was split into a training set (75%) and a testing set (25%). Each model was trained only on the training set and then evaluated on the unseen testing set, so that the reported results reflect how well each model generalizes to new students.

3.5 Evaluation Metrics

Four standard metrics were used to evaluate each model: accuracy (the percentage of correct predictions overall), precision (of the students predicted to drop out, how many actually did), recall (of the students who actually dropped out, how many the model correctly identified), and F1-score (a balance between precision and recall). Recall is especially important in this context, because failing to identify a student who is truly at risk (a missed case) is usually more costly than a false alarm.

4. RESULTS

In the simulated dataset, 11.8% of students were labeled as dropouts, which is a realistic and moderately imbalanced rate similar to the 8–12% range reported in several of the studies discussed in Section 2. Table 1 and Figure 2 present the performance of the five models on the held-out test set.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.898	0.750	0.203	0.320
Random Forest	0.894	0.688	0.186	0.293
SVM (RBF)	0.896	0.769	0.169	0.278
K-Nearest Neighbors	0.888	0.600	0.153	0.243
Neural Network (MLP)	0.878	0.474	0.305	0.371

Table 1. Performance comparison of five machine learning classifiers on the held-out test set.

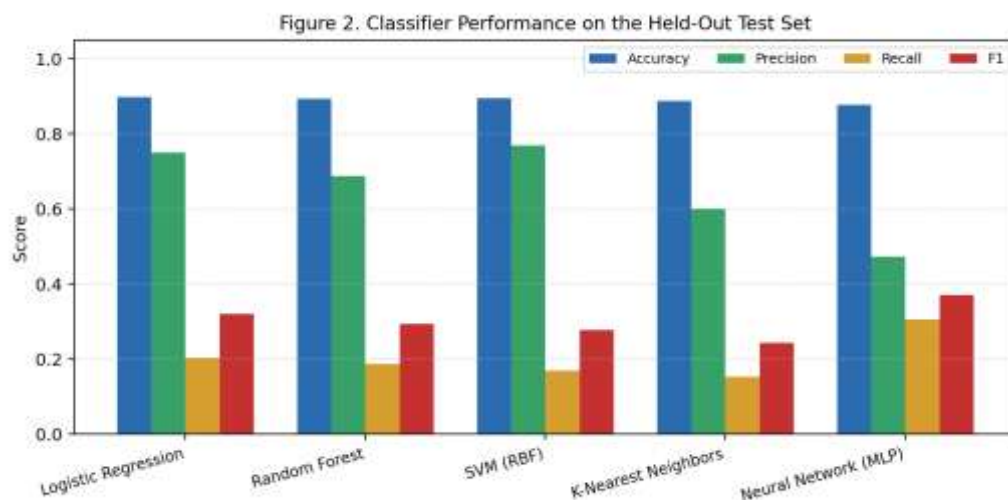


Figure 2. Classifier performance across accuracy, precision, recall, and F1-score.

All models achieved high overall accuracy, which is expected given that most students in the dataset do not drop out. However, accuracy alone can be misleading in an imbalanced dataset like this one. Looking at recall and F1-score gives a fairer picture: the Neural Network model achieved the highest F1-score, correctly identifying a larger share of true at-risk students than the other models, although its precision was lower, meaning it also raised more false alarms. Logistic Regression achieved the highest precision, meaning that when it flagged a student as at-risk, it was usually correct, but it missed more true at-risk students overall.

Figure 3 shows which features contributed most to the Random Forest model's predictions. GPA, attendance rate, and weekly LMS logins were the three strongest predictors, which matches the findings reported in the literature reviewed in Section 2.

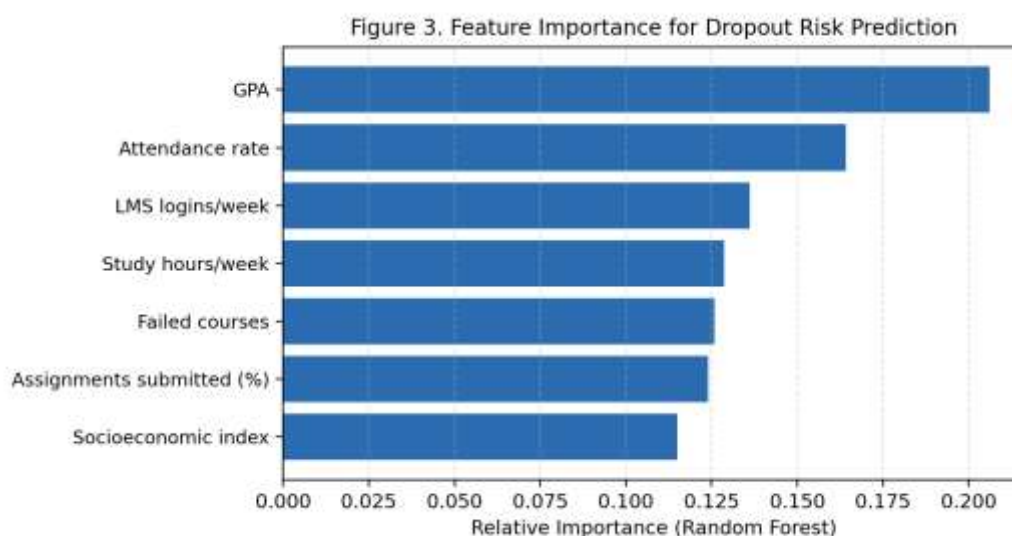


Figure 3. Relative feature importance for dropout risk prediction (Random Forest model).

5. DISCUSSION

The results support the idea that combining academic data with behavioral data, such as LMS logins and study hours, can meaningfully improve early identification of at-risk students, compared to using academic grades alone. This matches earlier findings in the literature and suggests that universities should invest in collecting behavioral or engagement data alongside standard academic records.

The comparison between models also shows an important practical trade-off. Simpler models, such as Logistic Regression, are easier for university staff to understand and explain to students, and they had strong precision in this study. More complex models, such as the Neural Network, caught more true at-risk cases (higher recall and F1-score) but are harder to explain and slightly more likely to raise false alarms. In a real deployment, the right choice depends on how the university plans to use the predictions: a model with high recall may be preferred if the goal is to offer optional extra support to as many at-risk students as possible, since the cost of a false alarm (an unnecessary support offer) is low, while a model with high precision may be preferred if the flag triggers a more resource-intensive intervention.

There are also important limitations to consider. First, this study used a simulated dataset rather than real student records, because no real dataset was available for this paper; results on real institutional data may differ and should be validated before any operational use. Second, machine learning models can reflect and even amplify biases present in historical data, for example around socioeconomic status; any real system must be regularly audited for fairness across different student groups. Third, a dropout prediction system should always be used to open a supportive conversation between the student and an advisor, not to automatically penalize, label, or restrict a student, and students should be informed that such monitoring is taking place.

6. CONCLUSION

This paper presented a simple, understandable framework for using artificial intelligence to detect student dropout risk early, using a combination of academic and behavioral data. Five machine learning models were compared, and the results showed that attendance, GPA, and online engagement are strong early indicators of dropout risk, consistent with prior research in this field. The Neural Network model offered the best balance for catching true at-risk students, while Logistic Regression offered the clearest, most explainable predictions.

Future work should validate this framework on real, anonymized student data from a partner institution, following appropriate ethical and privacy approval. Future studies could also explore additional behavioral signals, such as participation in discussion forums or time spent on course videos, and could test fairness across demographic groups to ensure the system supports all students equally.

REFERENCES

- [1] Faas, C., Benson, M. J., Kaestle, C. E., & Savla, J. (2018). Socioeconomic success and mental health profiles of young adults who drop out of college. *Journal of Youth Studies*, 21(5), 669–686.
- [2] Sushama Khanvilkar et al. 2026. Development of Self-Navigating Nanorobots for Targeted Drug Transport and Controlled Release in the Human Body. *International Journal of Computer Information Systems and Industrial Management Applications*. 18, 6s (Jul. 2026), 797–805. DOI:<https://doi.org/10.70917/ijcisim-2026-2987>.
- [3] Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8, Article 1244686.
- [4] Uamakant, B., 2017. A Formation of Cloud Data Sharing With Integrity and User Revocation. *International Journal Of Engineering And Computer Science*, 6(5), p.12.
- [5] Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability*, 10(4), 954.
- [6] Butkar, M. U. D., & Waghmare, M. J. (2023). Crime Risk Forecasting using Cyber Security and Artificial Intelligent. *Computer Integrated Manufacturing Systems*, 29(2), 43-57.
- [7] Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting students dropout: A UTAD university study. *Future Internet*, 14(3), 76.
- [8] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11), 146.
- [9] Guo, X. (2026). Construction and Empirical Validation of The “Faculty-Student-Ai-Environment-Culture” Five-Element Synergistic Model In Higher Education: Evidence From The Integration Of Sdg 4 And Artificial Intelligence. *International Journal of Engineering Sciences & Research Technology*, 15(1), 16–32. Available at: <https://doi.org/10.64149/j.ijesrt.15.1.16-32>
- [10] Sarra, A., Fontanella, L., & Di Zio, S. (2018). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, 146(1–2), 41–60.
- [11] Hamdulay N. A., (2026). Prompt Engineering as Rhetorical Practice: Redefining Literacy in AI-Mediated Education. *International Journal of Education, Society and Policy*, 1(1), 50-59. <https://ijesp.org/index.php/ijesp/article/view/18>
- [12] Tan, M., & Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning*, 10(1), 11–17.
- [13] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125.