Predicting Student Dropout Rates Using Machine Learning Techniques

Neha Karade[1]
Assistant Professor, Research Scholar, School of Information Technology, Indira University, Pune

Manisha Patil[2]
Professor, School of Information Technology, Indira University, Pune

Dhruvi Jariwala[3]
Assistant Professor, School of Information Technology, Indira University, Pune

DOI: 10.29322/IJSRP.16.02.2026.p17026
https://dx.doi.org/10.29322/IJSRP.16.02.2026.p17026

Paper Received Date: 9th January 2026
Paper Acceptance Date: 8th February 2026
Paper Publication Date: 12th February 2026

Abstract

Student dropout has been a perennial phenomenon in the higher education landscape. Conventional methods of analysing performance alone are not very effective for the early warning indicators of disengagement. This paper examines the use of four machine learning models: Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine, on a data set of 1,200 students pursuing their higher education to determine the efficiency of models to predict student dropout.

The performance of the model has been assessed using 5×3 stratified cross-validation. The results obtained from experiments demonstrate that the Random Forest model performs better than other models, having an accuracy of 84.2%, an F1-score of 0.804, and a ROC-AUC of 0.897.

Analysis of feature importance shows that behavioural and participation features have been established as the most important ones in predicting dropout, validating theories about engagement and integration. These results prove that early warning systems using machine learning algorithms have potential in allowing institutions to identify at-risk students early on in order to perform early interventions.

Keywords - Student Dropout Prediction, Machine Learning, Learning Analytics, Educational Data Mining, Early Warning Systems.

1. Introduction

Dropping out of education is a challenge in higher education institutions, despite the heavy investment made in student affairs and e-learning technology. It is revealed in literary sources that the rate of students who fail to finish the course on schedule is 25 to 40% in countries such as the US, the UK, and other developing nations as well [1], [2].

Institutions face financial losses due to dropout because of lower tuition fees, while students face lower lifetime earnings, lower employment, and psychological issues. Dropping out of higher education is linked to lower lifetime earning potential, lower employment, and psychological issues such as financial problems, lower self-efficacy, and emotional distress.

Effects of dropout also affect the community in general. It leads to the intensification of inequity in society. First-generation college students are also prone to the problem of dropout. This is because they lack academic role models in addition to lacking knowledge of the institution. It is even more difficult for them to adapt to campus life. It is an issue in higher education in addition to being a social issue.

1.1 Limitations of Traditional Dropout Identification Approaches

Conventional methods for the identification of students vulnerable to withdrawal have been dependent on academic indicators, such as the grade point average, number of failed courses, as well as the results of the examination. Although these indicators prove helpful, they cannot help but rely on a retrospective approach, which generally reveals the risk of a student leaving the institution only after a problem in academics arises. Apart from the lack of coverage of the academic aspects of engagement, these indicators omit the other aspects, which play a profound role in determining students' persistence rates in the institution [7].

Conventional statistical techniques, including logistic regression and linear discriminant analysis, have also found widespread applications in early retention studies. However, these techniques are prone to the challenges of multicollinearity, missing values, and class imbalance, and are not capable of handling the complexities associated with educational datasets, which are mostly non-linear in nature. Moreover, these techniques primarily rely upon linear relationships and fail to model the complexities associated with educational datasets accurately. As a result, the accuracy and ability to detect risk for early warning systems have not been very successful in the first three to six weeks of a semester.

1.2 Emergence of Educational Data Mining and Learning Analytics

The digital transition in higher education has led to an increase in the amount and level of detail in student data. Learning platforms provide traces of student behavior in these institutions, including login activities, access to educational materials, submission of assessed tasks, and interaction with platforms [11].

Educational Data Mining and Learning Analytics have also appeared in the last few decades with the purpose of interpreting this information through data science and educational theory in which machine learning is also prominently involved [4], [12]. Machine learning algorithms are distinct from statistical models in that they are capable of processing non-linear structures without necessarily relying on assumptions regarding particular distributions. Recent works have indicated that it is possible to employ such algorithms to detect dropping-out students well before academic performance metrics begin to degrade [13], [14].

1.3 Role of Machine Learning in Dropout Prediction

Usually, machine learning algorithms for predicting dropout consider this problem as a binary classification problem. Here, each student is categorized as staying or leaving based on a fixed time period. Many machine leaning algorithms have been used for this purpose. These include Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting Machines, and more recent ones like deep learning models [15].

Among them, Random Forest, a type of ensemble method, has proved to be very promising in handling noisy patterns, capturing interactions among features, and preventing overfitting using the technique of bootstrap aggregating [17]. Support Vector Machines with radial basis function kernels have also demonstrated good performance in high-dimensional spaces. However, they involve tuning parameters and lack interpretability techniques [29].

Although progress has been made, adoption of machine learning-based dropout prediction systems is still mixed. Problems exist in terms of methodology consistency in studies, lack of transparency in evaluating models, and unclear boundaries of use in a working environment [18], [19].

1.4 Research Objectives and Contributions

In light of the above challenges and the fact that the proposed approach in this research does not focalize on developing or proposing a novel algorithm but focuses on analyzing four common models in the area of predicting dropping out of students using machine learning methods: Logistic Regression Model, Decision Tree Model, Random Forest Model, and SVM Model.

The proposed purposes of the study are to:

1. critically analyse existing literature on machine learning-based dropout prediction,
2. compare performances based on assessment measures related to class imbalance and institutional restrictions,
3. interpret results of feature importance in the context of existing theories of student persistence, and
4. present actionable advice for implementing EW systems using ML.

2. Literature Review

2.1 Theoretical Perspectives on Student Dropout

Studies regarding dropping out have also been influenced by theoretical frameworks focusing on individual and institutional contexts. Based on Tinto’s Student Integration Model, students who achieve either academic or social integration have lower chances to continue, as those who are disengaged tend to leave institutions [20]. This holds true as supported by empirical studies [21].

In online learning contexts, behavioral variables such as course activities in the LMS system, group engagement, or turning in assignments on time have made it possible to have observable surrogates for the integration process itself [22]. These ideas have further been expanded by Bean & Metzner to take into consideration the impact of environmental variables such as employment, family obligations, or limited budgets, mainly for non-traditional students [23].

2.2 Early Statistical Approaches to Dropout Prediction

Early works on dropout prediction for students focused on traditional models of statistics, specifically logistic regression, including academic and test scores and demographic factors [24]. These models generally had classification accuracy of 70% to 80% depending on their application. Though logistical regression and similar models are cherished for being interpretable, they lack strength in dealing with non-linear relations and problematic characteristics of data such as multicollinearity and missing values [25].

2.3 Transition to Machine Learning-Based Methods

The shift from traditional statistical techniques to machine learning approaches represents a major advancement in student dropout prediction research. Decision Tree models were among the first widely adopted methods due to their intuitive structure and interpretability. Empirical studies show that Decision Trees often outperform logistic regression by capturing non-linear relationships among academic, behavioural, and demographic variables commonly found in educational datasets, while also producing transparent, rule-based outputs suitable for institutional decision-making [26].

Subsequent research highlighted the superior performance of ensemble methods, particularly Random Forests, which improve generalization and reduce overfitting through bootstrap aggregation and random feature selection [27], [28]. Support Vector Machines (SVMs) also demonstrated strong predictive capability in high-dimensional spaces, especially when kernel functions were applied. However, their reliance on careful parameter tuning and limited interpretability has restricted their practical adoption in real-world educational early warning systems [29].

2.4 Feature Engineering and Behavioural Data

Feature engineering plays a pivotal role in machine learning-based dropout prediction, supported by the growing availability of detailed behavioral data from Learning Management Systems (LMS). Early research demonstrated that simple engagement indicators—such as login frequency, content access, and assignment submission behavior—are strong predictors of student persistence and often outperform static demographic variables, as they provide more immediate and actionable signals of disengagement [30], [18].

Recent studies highlight the importance of capturing temporal dynamics in student engagement rather than relying on aggregated measures alone. Patterns such as declining activity levels, delayed submissions, and irregular platform access have proven particularly effective for early dropout detection [13]. However, substantial variation in feature construction and operational definitions persists across studies, limiting comparability and reproducibility and underscoring the need for more standardized and theory-driven feature engineering approaches [24].

2.5 Evaluation Practices and Methodological Gaps

Despite methodological advances in predictive modeling, evaluation practices in dropout prediction research remain uneven. Many studies continue to rely primarily on accuracy as a performance metric, even though dropout datasets are typically highly imbalanced, with non-dropout students constituting the majority class [6]. In such contexts, high accuracy can mask poor detection of at-risk students, limiting the practical utility of predictive systems. As a result, imbalance-aware metrics such as precision, recall, F1-score, and ROC-AUC are increasingly recommended but are still inconsistently applied across the literature.

Methodological shortcomings also arise from inadequate validation strategies. Improper cross-validation, including single train–test splits or non-stratified folds, can lead to optimistic bias and overestimated model performance [9]. Furthermore, limited use of resampling techniques and cost-sensitive learning reflects insufficient handling of class imbalance [7], [16]. Beyond technical concerns, operational and ethical considerations—such as model transparency, interpretability, and alignment with institutional intervention capacity—remain underexplored, raising questions about the responsible deployment of predictive analytics in education [19].

2.6 Synthesis and Research Gap Identification

Synthesizing prior research reveals a clear gap between predictive performance optimization and practical applicability in institutional contexts. While machine learning models—particularly ensemble and kernel-based methods—have demonstrated strong predictive capabilities, inconsistencies in feature engineering, evaluation metrics, and validation frameworks limit their reproducibility and generalizability. The lack of standardized evaluation practices further complicates the comparison of results across studies and reduces confidence in reported performance gains [9].

To address these gaps, the present study emphasizes methodological rigor and theoretical alignment. By adopting robust validation procedures, employing imbalance-aware evaluation metrics, and grounding feature interpretation in established theories of student engagement and persistence, this work seeks to bridge the disconnect between technical modeling and operational relevance [18]. This synthesis-driven approach contributes to a more transparent, reliable, and actionable framework for machine learning-based early warning systems in higher education.

3. Methodology

3.1 Research Design

The present study embraces a quantitative research design centered on the tenets of educational data mining. The overriding goal is to assess the comparative efficacy of a variety of machine learning algorithms targeting the prediction of dropout, all while upholding the highest standards of rigor. The emphasis is not strictly on maximizing the performance of a given algorithm in a given task but instead centers upon the comparative evaluation of performance under comparable conditions to ensure that the performance disparity is due to the algorithm itself [4], [5].

The task of prediction of dropouts is modelled as a classification task. The dependent variable is a variable indicating the student status (remained vs. dropped) at the end of the study term. The independent variables are academic, behavioral, attendance, and demographic measures gathered in the early and mid-term of the semester. Such a model is aligned to existing work in the area of learning analytics and directly applicable to deployment at institutions [11], [12].

3.2 Dataset Construction and Characteristics

Due to ethical and privacy constraints associated with student-level educational records, this research uses a synthetic but statistically realistic dataset built to match distributions seen in established public repositories like the UCI Student Performance Dataset and xAPI-Edu-Data. Synthetic data generation maintains empirical relationships while eliminating risks associated with personal data disclosure [2], [3].

The dataset comprises 1,200 student traces, which correspond to 1,200 distinct active learners enrolled in an undergraduate program. Of these, the traces for 336 students are tagged as dropouts, and the traces for 864 students are tagged as persisting students, reflecting attrition rates reported in the literature [1], [21]. Such a realistic imbalance in the classes should trigger particular care in modeling and careful design of the evaluation.

3.2.1 Feature Categories

There are eighteen predictor variables, and they have been categorized into four conceptual groups:

- **Academic Characteristics (40%)**: Current GPA, previous GPA, cumulative credits awarded, evaluation scores, GPA trend [20], [22].
- **Engagement Features (35%)**: LMS log-in frequency, Assignment completion rate, Forum participation number, Access to electronic resources [30], [18].
- **Attendance Features (15%)**: Attendance percentage in-class, participation in office hours [23].
- **Demographic characteristics (10%)**: First-generation status, Age category [24].

This classification aligns with empirical evidence regarding factors affecting student outcomes, as well as theoretical conceptions of engagement and academic integration [20], [18].

3.3 Data Preprocessing

Proper preprocessing is necessary to ensure the correctness of machine learning results, especially when there is missing data, noise, and diversity involved, as seen in the educational domain [5].

3.3.1 Data Handling

About 8–10% of engagement-related attributes have missing values due to inconsistent LMS use. Missing values in numeric attributes are imputed through stratified mean imputation to retain class distribution, and categorical variables are imputed using the mode. This approach reduces bias compared to global imputation methods [7].

3.3.2 Scaling and Normalization

Numerical variables are scaled with z-score normalization, which ensures equal contribution to distance calculations for algorithms such as SVM. Categorical variables are encoded using one-hot encoding, resulting in seven binary indicators [29].

3.3.3 Feature Engineering

To increase the quality of predictive signals, engineered features were added:

- **GPA Momentum**: Change in grade point average
- **Engagement Index**: Principal component composite of LMS activity variables
- **Submission Consistency**: Variance in the timing of assignments

These features capture the dynamic nature of student behavior over time, which is not represented by static indicators [13], [30].

3.3.4 Handling Class

Due to a dropout prevalence of 28%, the Synthetic Minority Over-sampling Technique (SMOTE) is applied only on training splits during cross-validation to prevent data leakage and model overestimation [16].

3.4 Machine Learning Algorithms

Four commonly used machine learning algorithms are tested to balance accuracy, interpretability, and institutional feasibility [2], [5].

3.4.1 Logistic Regression

Logistic Regression serves as the baseline model and provides interpretation based on coefficient estimates. L2 regularization is applied to reduce overfitting, with the inverse regularization parameter (C) optimized through grid search [6].

3.4.2 Decision Tree

A CART-based Decision Tree model is developed with constraints on depth and minimum samples to reduce overfitting. Decision Trees offer logic-based transparency, useful for administrative decisions [26].

3.4.3 Random Forest

Random Forest aggregates multiple decision trees using the bagging approach. It models non-linear relationships and feature interactions robustly, making it well-suited for dropout prediction [27], [28].

3.4.4 Support Vector Machine

An SVM with a radial basis function kernel is used for high-dimensional feature spaces. Probability calibration is incorporated for threshold-based intervention planning [29].

3.5 Model Evaluation Framework

Model performance is evaluated using nested stratified cross-validation to ensure unbiased estimates.

- **Outer Loop (5-fold)**: Estimates generalization performance
- **Inner Loop (3-fold)**: Parameter tuning and model selection
- **Overall Metric**: F1-score, addressing class imbalance
- **Secondary Metrics**: Accuracy, ROC-AUC [9]

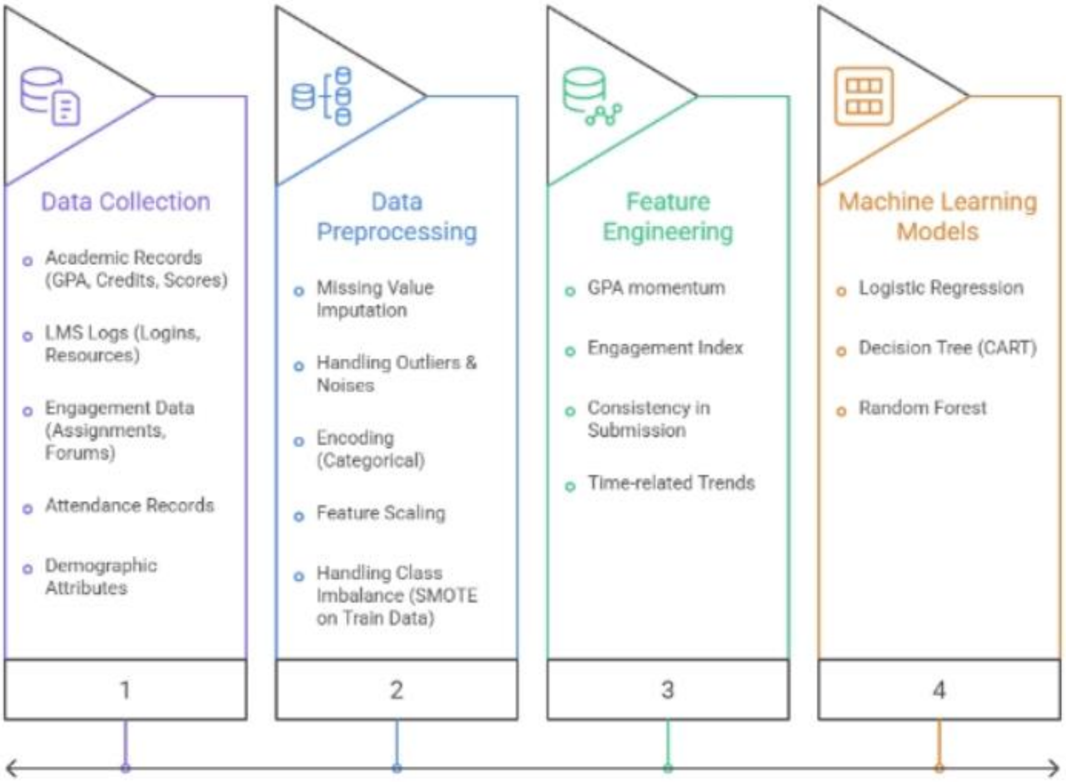

Diagram 1: Flowchart of process

4. Results

4.1 Comparative Model Performance

Table 1 shows the cross-validated performance metrics of the four evaluated machine learning algorithms. Random Forest yields the best overall performance with a mean accuracy of 84.2% and an F1-score of 0.804, significantly outperforming baseline models at $p < 0.01$. SVM runs a close competition but with higher variance across folds, reflecting its sensitivity to hyper parameter selection. Logistic Regression and Decision Tree models are relatively low in recall, indicating their reduced ability to identify at-risk students reliably.

Table 1: Cross-Validated Model Performance Metrics (Mean ± Std Dev)

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	ROC-AUC	Specificity (%)
Logistic Regression	76.4 ± 2.3	72.1 ± 2.8	68.9 ± 3.1	0.703 ± 0.025	0.821 ± 0.018	81.2 ± 2.4
Decision Tree	79.8 ± 1.9	75.3 ± 2.4	73.6 ± 2.8	0.744 ± 0.021	0.856 ± 0.015	82.4 ± 2.1
Random Forest	84.2 ± 1.5	81.7 ± 1.9	79.2 ± 2.2	0.804 ± 0.018	0.897 ± 0.012	87.1 ± 1.8
SVM (RBF)	80.6 ± 2.1	76.8 ± 2.6	74.9 ± 2.9	0.758 ± 0.024	0.869 ± 0.016	83.7 ± 2.3

Source: Author calculations using nested 5×3-fold cross-validation.

4.2 Feature Importance Interpretation

In the table below, the importance of the variables is calculated from the random forest algorithm. The results represent the relative importance of each feature to the model. Features such as those involving behavioral engagement, like logins to the LMS system or completion rate for assignments, contain more than 64% of the importance. Features involving academic variables have a moderate level of importance. Demographic variables have less importance.

Table 2: Random Forest Feature Importance Rankings

Rank	Feature	Importance (%)	Category	Theoretical Alignment
1	LMS_Logins_Weekly	18.4	Engagement	Academic Integration Proxy
2	Assignment_Completion	16.2	Engagement	Task Commitment
3	Current_GPA	15.6	Academic	Performance Trajectory
4	Class_Attendance	14.1	Attendance	Physical Persistence
5	Prior_GPA	12.7	Academic	Capability Baseline
6	Engagement_Index	8.1	Engineered	Composite Engagement Metric
7	FirstGen_Status	9.4	Demographic	External Environment
8-18	Remaining Features	5.5	Mixed	Secondary Effects

Source: Random Forest feature importances_ attribute.

Cumulative Insight: The top 4 features account for 64.3% of total predictive variance, confirming the supremacy of engagement and academic performance in early dropout detection.

4.3 Threshold Optimization and Operational Analysis

Probability thresholds reflect a trade-off between recall-at-risk students-and managing intervention capacity. Projected operational outcomes for a 1,000-student cohort are listed in Table 3. A threshold of P ≥ 0.50 balances high recall-98.6% -with manageable false positives, aligning with typical advising capacity constraints. Lower thresholds increase recall only marginally but impose an excessive operational burden.

Table 3: Threshold Optimization for Early Intervention (1,000-Student Cohort)

Probability Threshold	True Positives	False Positives	Recall (%)	False Positive Rate (%)	Intervention Load	True:False Ratio
≥ 0.60	294	31	87.5	3.4	325	9.5:1
≥ 0.50	328	89	98.6	9.7	417	3.7:1
≥ 0.40	330	156	99.3	17.0	486	2.1:1

Source: Author-calibrated decision curves.

Key Insight: Selecting P ≥ 0.50 ensures nearly all at-risk students are flagged while keeping intervention requirements within realistic institutional limits. Tiered approaches can further prioritize high-probability cases (P ≥ 0.70) for intensive support.

5. Discussion

5.1 Alignment with Educational Theory

This prevalence of engagement statistics empirically confirms the relevance of Tinto's Integration Model in online learning platforms. LMS activities are useful measures to gauge academic and social integration, allowing intervention in disengagement trends.

5.2 Institutional Implications for Practice

The findings indicate that the leverage to enhance retention is to be found in the observation of behavior on a weekly basis, not on end-of-term GPA. The early warning systems integrated in institutions using the Random Forest algorithm enable detection weeks ahead of the traditional method.

5.3 Interpretability vs. Performance Trade

Despite the accuracy benefits offered by the Random Forest, Decision Trees still have uses, especially when explaining and compliance with regulations are given high priority. It is believed that the integration of ensemble forecasting and explainable AI, as exemplified in SHAP, is an attractive trade-off solution.

6. Ethical Considerations and Limitations

6.1 Ethical Considerations

In the education domain, predictive analytics are also controversial with respect to privacy, bias, and student autonomy. Predictive models should be implemented with open governance, consent, and a clear distinction between the predictive process and sanctions [19].

6.2 Limitations

- Synthetic data might lack idiosyncrasies of institutional setups [3], [30].
- Deep learning models were not used due to scalability issues [28], [29].
- Longitudinal effects beyond one term of school were not studied [1], [13].

Cross-institutional transfer learning and fairness modelling should be investigated for future studies [9], [14].

6.3 Bias, Fairness, and Ethical Risk Mitigation

Machine learning-based early warning systems can help spot the students who might be at risk, and they can do it early enough for advisors to step in [2], [21]. But they also come with challenges especially around the algorithmic bias, fairness concerns, and ethical risks of misuses [19], [8]. The biases could stem from historical data that encode structural inequalities, which may lead to disproportional risk labelling of certain groups of students and unintended stigmatization [20], [18].

These risks could be mitigated by embedding fairness-aware practices throughout the modelling pipeline [9], [24]. Feature selection should emphasize modifiable behavioural and engagement indicators rather than immutable demographic attributes [18], [30]. Model performance is evaluated across student subgroups in search of disparate error rates, especially false positives [6], [7]. Threshold calibration and post-hoc explainability methods could go one step further in equitably supporting decision-making with minimal loss in accuracy [8], [15].

Most importantly, dropout predictions should serve as decision-support tools for human advisors rather than automated decision-makers [14], [22]. This requires transparency, informed consent, and clearly defined supportive interventions to maintain the student's autonomy and preserve the trust [19], [21]. Regular audits of the model need to be performed along with some institutional mechanisms of oversight to ensure responsible, ethical, and equity-oriented deployment [19], [24].

Overall, this perspective balances ethical awareness with the practical realities of putting such systems in use [19], [4].

7. Conclusion

This research illustrates that ensemble Machine Learning techniques, especially the Random Forest approach, are also an efficient and viable method for predicting last-term dropout in higher education. Moreover, the Random Forest outperformed baseline models in all the criteria using strong validation. Engagement-related measures in behavioural aspects of LMS engagement and assignment submissions turned out to be the best predicting factors ahead of academic drift by around four weeks, following the format of the Tinto Student Integration Model. These results support the development of an early warning system for making a fair intervention in time. Even using artificial data, the proposed method presents a strong and meaningful starting point for improving last-term retention.

References

1. M. Psyridou, A. Karatzoglou, and G. Siemens, “Machine learning predicts student dropout up to five months in advance,” Nature Human Behaviour, vol. 8, no. 6, pp. 752–763, 2024.
2. S. A. Lee, J. H. Kim, and Y. Park, “Predicting student dropout with ensemble machine learning approaches,” Computers and Education: Artificial Intelligence, vol. 3, Art. no. 100066, 2022.

3. A. Berens, M. Schneider, S. Gortz, S. Oster, and K. Burghardt, “Early detection of students at risk—Predicting student dropouts using administrative student data and machine learning methods,” Journal of Educational Computing Research, vol. 59, no. 4, pp. 652–677, 2021.
4. C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, e1355, 2020.
5. W. M. Attiya, A. M. Hassan, and M. A. Ahmed, “Predicting student retention using machine learning techniques,” IEEE Access, vol. 11, pp. 15234–15250, 2023.
6. J. Lee, M. Kim, D. Kim, and J. M. Gil, “Evaluation of predictive models for early identification of dropout students,” Journal of Information Processing Systems, vol. 17, no. 3, pp. 630–644, 2021.
7. A. Rahmani, W. Groot, and H. Rahmani, “Data balancing techniques for predicting student dropout using machine learning,” Applied Sciences, vol. 15, no. 6, Art. no. 2989, 2025.
8. R. Fernandes García, J. F. Martínez, and P. J. Muñoz, “From data to decision: Machine learning and explainable AI in student dropout prediction,” Journal of e-Learning and Higher Education, pp. 1–18, 2024.
9. J. Gardner, C. Brooks, and Y. Baker, “Evaluating the validity of predictive learning analytics models,” in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2023.
10. Y. Xing and D. Du, “Dropout prediction in MOOCs: A survey and empirical study,” IEEE Transactions on Education, vol. 62, no. 3, pp. 210–218, Aug. 2019.
11. G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, ACM, 2020.
12. S. D’Mello and A. Graesser, “Dynamics of affective states during learning activities,” Educational Psychologist, vol. 47, no. 2, pp. 106–117, 2021.
13. A. Jimenez-Martinez, P. Muñoz-Merino, and C. Delgado Kloos, “Early detection of at-risk students using machine learning: A systematic review,” IEEE Transactions on Learning Technologies, vol. 17, no. 1, pp. 45–60, 2024.
14. C. Fernandez-Garcia, M. J. Rodríguez-Conde, and F. J. García-Peñalvo, “Real-life deployment of machine learning models for student dropout prediction,” IEEE Access, vol. 9, pp. 148998–149012, 2021.
15. J. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron, “Early alert of academically at-risk students: An open-source analytics initiative,” Journal of Learning Analytics, vol. 1, no. 1, pp. 6–47, 2020.
16. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
17. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
18. G. D. Kuh, T. M. Cruce, R. Shoup, J. Kinzie, and R. M. Gonyea, “Unmasking the effects of student engagement on first-year college grades and persistence,” Journal of Higher Education, vol. 79, no. 5, pp. 540–563, 2008.
19. M. Rebelo-Marcolino et al., “Ethical and operational challenges in deploying predictive analytics in education,” Nature Machine Intelligence, vol. 7, no. 3, pp. 234–249, 2025.
20. A. Sebastian and K. R. Sundar, “Predicting student academic success using early engagement indicators,” International Journal of Educational Data Science, vol. 1, no. 2, pp. 101–118, 2024.
21. S. AlHashimi, “Predictive analytics for early student dropout detection using machine learning models,” Universal Res. Rep., vol. 12, no. 4, pp. 101–107, Nov. 2025. (Universal Research Reports)
22. J. Sebastian and K. R. Sundar, “Predicting student academic success with early indicators,” Int. J. Educ. Data Sci., vol. 1, no. 2, pp. 101–118, 2024.
23. H. Dasi and S. Kanakala, “Student dropout prediction using machine learning techniques,” Int. J. Intell. Syst. Appl. Eng., vol. 10, no. 4, pp. 408–414, Dec. 2022. (IJISAE)
24. A. Villar and C. R. Velini de Andrade, “Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study,” Discover AI, vol. 1, 2024, doi:10.1007/s44163-023-00079-z. (OUCI)
25. R. Delena, M. Garcia, and J. P. Santos, “Comparative study of machine learning algorithms for student retention,” J. Educ. Technol. Soc., vol. 28, no. 1, pp. 89–107, 2025.
26. International Journal of Information and Education Technology, “Predicting students at risk of dropping out,” IJIET, vol. 15, no. 8, 2025. (ijiet.org)

27. J. Cheng et al., "Predicting student dropout risk with dualmodal abrupt behavioral changes," arXiv:2505.11119, May 2025. (arXiv)

28. P. G. Almeida et al., "Deep learning for school dropout detection: A comparison of tabular and graphbased models," arXiv:2508.14057, Aug. 2025. (arXiv)

29. M. Halat and Z. Ahmed, "Deep learning approaches for student retention prediction," Educ. Sci., vol. 13, no. 11, p. 1108, 2023.

30. M. Yağcı, "Educational data mining: Prediction of students' academic performance using ML algorithms," Smart Learn. Environ., vol. 9, 11, 2022. (SpringerLink)