LogicLink: Journal of Artificial Intelligence and Multimedia in Informatics
E-ISSN: 3062-9098 | P-ISSN: 3063-4504 | Vol. 2 No. 2, December 2025, pp. 138 – 148

Predicting Student Dropout in E-Learning Using Simple Machine Learning and Explainable Data Analysis

A.I. Ahmad[1], D.A Nugroho[2], S.A. Aliyu[3], A.M. Abdullahi[4]
[1,2]Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Pekalongan, Indonesia
[3]Federal University Dutse, Dutse, Nigeria
[4]Noida International University, Delhi, India
*E-mail : aliyu.ibrahim.ahmad25055@mhs.uingusdur.ac.id**
**Corresponding author*

Received 15 November 2025; Revised 1 December 2025; Accepted 2 December 2025

Abstract- Online learning allows a lot of flexibility and accessibility, but the issue of student dropout becomes one of the key problems. Much of the research that has been done to predict dropout using complex machine learning methods has not been interpretable, and many of these methods are hard to implement in practice, especially in resource constrained environments. This research closes this gap by presenting a simple and explainable machine learning method of predicting early dropout. Students in several universities were surveyed on the frequency of attendance, quiz performance, completion of assignments, satisfaction with learning, and hours spent studying per week. Because there were only a few real dropout cases, a controlled synthetic data augmentation method was used to demonstrate and train the model. The use of Logistic Regression and Decision Tree classifiers was used to predict the risk of dropout. The accuracy of both models was 87.5% with the most effective indicators being learning satisfaction, attendance, and prior consideration of dropout. This research is considered new because it focuses on the simplicity and interpretability of models rather than the complexity of creating early warning systems to predict the dropout of students, showing that they are not required to be complex or made out of heavyweight models to perform well. Even though the findings are merely exploratory due to limitations of the datasets, the results show that even simplistic models might be used to assist educators in identifying at-risk learners and incorporating prompt intervention measures.
Keywords: e-learning, student dropout, machine learning, interpretable analysis, student retention.

1. INTRODUCTION

Online learning has been a fundamental component of contemporary education as it is flexible, convenient, and allows serving learners regardless of geographical borders. The high pace of digital education in the COVID-19 pandemic situation promoted the use of e-learning platforms even more globally. Although these advantages exist, the problem of student dropout has been among the major problems of online education. Low motivation, low level of engagement, technical constraints, and low academic support are some of the problems faced by many learners, in the case of they tend to drop courses prematurely.

The other studies conducted in the past analyzed several issues that affect student dropouts in online learning. (Xing & Du, 2022) have found that time management, lack of interaction, and low motivation are some of the largest contributors to online dropout. Equally, (Khalil et al., 2022) highlighted that technological challenges and dissatisfaction are major contributors to the rate of dropouts. These results concur with the opinion that

academic ability is not the only factor that controls the behavior of dropouts, but also psychological and behavioral factors.

Models developed using machine learning have become more significant in predicting dropping out of school as learning analytics become more sophisticated. To assess dropout risk as well as analyze learner behavior, researchers have used methods like Random Forest, Support Vector Machines, and artificial neural networks (Al-Shabandar et al., 2020; He et al., 2021). Considering the fact that these techniques frequently yield excellent accuracy in prediction, their complexity continues to be a significant drawback. The interpretability and practical utility of these models for educators have been reduced because many of them function as black-box systems, giving minimal explanation for predictions. As a result, organizations have trouble converting predictive findings into successful academic interventions.

Studying has recently pointed out the need to be transparent in machine learning systems in education. According to (Xing and Du, 2022), education decision-making requires prediction systems to deliver an interpretation of the results. Another important fact highlighted by (Fan et al., 2021) concerns the simplicity and comprehensibility of the models relevant in academic settings because they allow an instructor to effectively address the results of prediction. Nonetheless, the majority of current research has only Logistic Regression and Decision Tree models as control conditions, whereas the research is concentrated on elaborate prediction structures instead of interpretable ones.

In online learning, machine learning techniques have been used to predict student dropout in recent research carried out by many scholars. Al-Shabandar et al. (2020) used classification models on learning data and obtained high predictive accuracy; however, the complexity model decreased transparency and hindered interpretation to apply in practical academic projects. Also, (Kumar et al., 2021) used machine learning methods to predict dropout, and they achieved better accuracy, yet their method was less focused on explaining model decisions and more based on prediction strength, which makes designing interventions more challenging. It has also been found that simpler models are still competitive in performance, as well as providing improved transparency. (Xing and Du, 2022) highlighted that explainability is essential in the education setting, where a teacher needs to comprehend the results of prediction so that they can make effective decisions. Besides that, (Fan et al., 2021) performed a systematic review of interpretable machine learning in education, and they discussed that model transparency is a key contributor to trust, usability, and ethical system deployment in academic settings. Moreover, the current research points out that academic performance, which is traditionally considered to impact dropout, is not the sole driver, as behavioral and motivational factors also contribute to dropout. (Zhang et al., 2023) have shown that satisfaction, engagement, and learning behavior are the major predictors of dropout and suggested that explainable models would allow institutions to detect risk factors that can be taken into action. Regardless of such findings, a large number of recent efforts continue to emphasize complex algorithms over interpretability as a design objective. As such, this leaves a gap in research in the creation of simple, interpretable models that strike a trade off between usability and predictive accuracy in actual educational settings.

1.1 Research Gap:

Although machine learning has been extensively applied to dropout prediction, three limitations have not been adequately tackled:

a) The majority of the research is more accurate than interpretative.
b) Black-box systems are still quite challenging to learn and implement by teachers.
c) Multi-university survey-based data that are in a simple form are hardly ever analyzed through interpretable models.

The scientific literature on simple and understandable classifiers as practical early warning systems has not been investigated systematically with particular attention to the institutions that have limited computational capabilities. Complicated AI systems are frequently unrealistic to implement and service in such settings.

1.2 Purpose of the Study:

The objective of the study is to come up with a predictable and lightweight prediction of dropout based on the Logistic Regression and Decision Tree models on survey data gathered from students in various universities. The study is aimed at measuring such academic, motivational, and behavioral indicators as attendance, satisfaction, hours of study, completion of assignments, and performance.

1.3 Novelty and Contribution:

The current study places more emphasis on model candor and attainable usability than earlier research that prioritizes algorithm difficulty. The present work is special since it employs Logistic Regression and Decision Tree as the primary predictive models rather than baselines, views comprehending as a design goal rather than a secondary quality attribute, applies the method to cross-university survey data rather than singular datasets, and shows how explainable emergency warning systems can be implemented in environments that are resource-limited. This study enhances the area by illustrating that transparent models might prove more helpful in helping educators recognize, believe, and act upon predictive outcomes, and that intensive computational techniques are not required for predicting student dropout.

1.4 Conceptual Framework:

The conceptual framework used in this study illustrates how the combined influence of academic and behavioral variables affects student dropout in e-learning environments. The framework integrates quantitative data processing with machine learning techniques to detect, explain, and predict dropout risk. Model inputs are derived from student learning behavior and experience, including attendance frequency, quiz and test performance, number of assignments completed, satisfaction with e-learning, weekly study hours, and prior consideration of dropping out. These variables represent both performance-related and psychological factors that influence student persistence. During the processing stage, the dataset undergoes preprocessing procedures such as data cleaning, normalization, and encoding before being used to train two interpretable classification models: Logistic Regression and Decision Tree. The models analyze relationships between the independent variables and the dependent variable (dropout status: active or dropped out) to produce predictive outcomes. At the output stage, the trained models generate estimates of dropout probability while also highlighting the most influential risk factors. This framework supports not only prediction but also interpretability, enabling educators to understand the underlying causes of dropout and implement timely intervention strategies such as academic counseling or learning support.

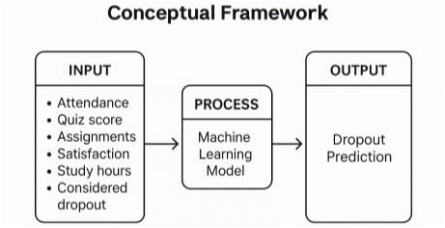

Figure 1. Illustration of Student Dropout Prediction Framework

2. RESEARCH METHOD

2. 1. Research Design

The research will use a quantitative research method based on an interpretable machine learning model to forecast student dropout rate in an e-learning setting. Data collection, preprocessing, data balancing via synthetic generation, model training, and evaluation via cross-validation are involved in the research process.

Figure 2. Overview of the Research Methodology Pipeline

2. 2. Dataset Description

The data is based on a response of 47 students that was performed using an online survey that was disseminated among various universities in various countries. The data set includes one target variable, Dropout Status (1 = dropout, 0 = active). The predictive characteristics are:

- Attendance frequency
- Quiz score
- Assignment completion
- Learning satisfaction
- Study hours per week
- Considered dropout
- Gender

Numerical encoding was made of categorical variables. Missing numeric values were replaced with median values.

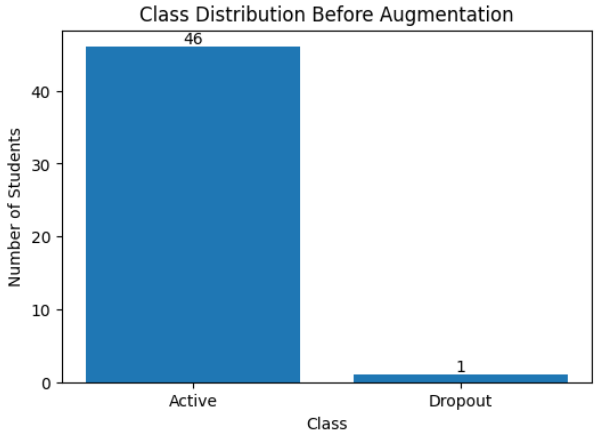

Figure 3. Class Distribution Before Data Augmentation.

2. 3. Synthetic Data Generation

With the presence of such a strong imbalance between the classes (46 active and 1 dropout), the concept of synthetic data generation was utilized to model the cases of dropouts in an experiment.

Data Augmentation Process.

A rule-based generation approach that relied on the at-risk behavior of students was used to expand the dropout class. It was done by the following steps:

a) Seed samples were picked out through records that had low attendance and low satisfaction.
b) To create new data points, the following calculation was made on the seed records:
   - Minimizing the number of attendance and quiz scores by random numbers within set constraints.
   - Reduction of the values of satisfaction.
   - Decreasing study hours.
   - Reduction of assignment numbers at random.
c) All the artificial records were marked as dropouts.
d) The last data set consisted of 15 dropout samples and 46 active samples.

This technique maintained realistic data boundaries and trends.

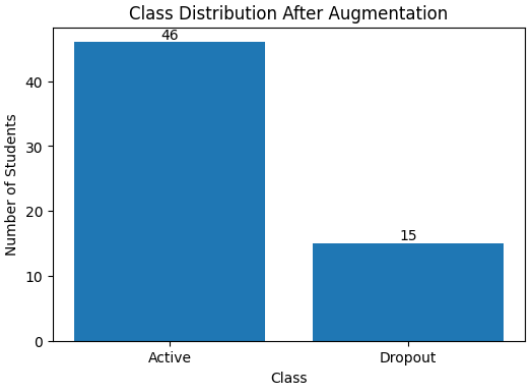

Figure 4. Class Distribution After Data Augmentation

2. 4. Classification Models and Parameters

Two models were applied:

a) Logistic Regression
   - Solver: liblinear
   - Max iterations: 300
   - Class weighting: balanced
   - Regularization: L2
b) Decision Tree
   - Criterion: Gini index
   - Maximum depth: 4
   - Minimum samples per leaf: default
   - Random state: 42
   - Class weight: balanced

2. 5. Model Validation

Stratified k-fold cross-validation (k = 5) was used instead of a one-train-test split. This was able to use each data point to train and evaluate, and minimize bias due to small samples. The following measures were used to measure performance:

- Accuracy
- Precision
- Recall
- F1-score

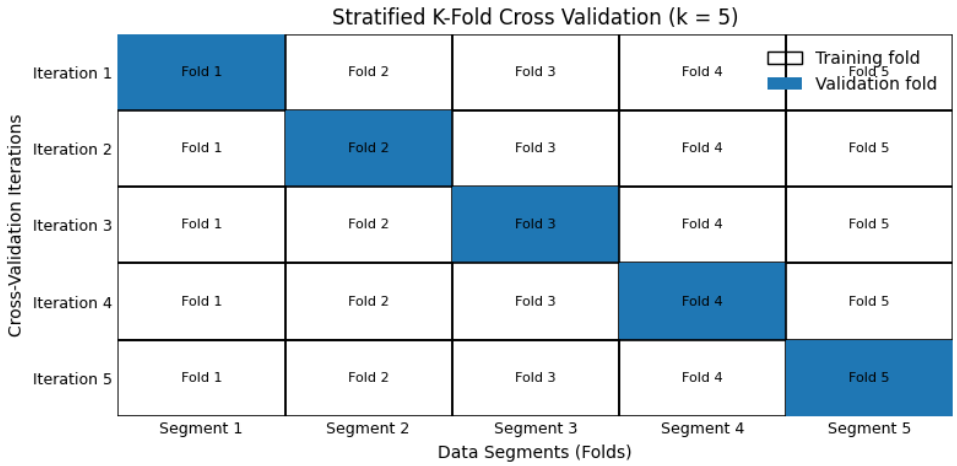

Figure 5. Visual Illustration of Stratified K-Fold Cross Validation for Model Evaluation (k = 5)

2. 6. Algorithmic Workflow

Algorithm 1. Student Dropout Prediction Process.

Input: Cleaned dataset

Output: Dropout prediction model

a) Categorical variables are to be encoded.
b) Handle missing values.
c) Detect class imbalance.
d) Produce artificial dropout samples.
e) Categorize synthetic data as part of the dataset.
f) N = stratified k-fold cross-validation (k = 5)
g) Train the Logistic Regression model.
h) Train Decision Tree model.
i) Evaluate model performance.
j) Choose the most successful model.

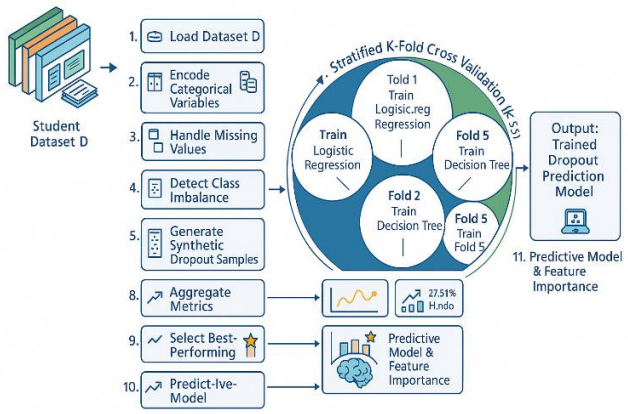

Figure 6. Workflow of the Dropout Prediction System Using an Interpretable Machine Learning Model.

3. RESULTS AND DISCUSSION

3.1 Overview of Data Findings

The data that was used in the study was gathered among 47 students of different universities and nationalities. The respondents had varied backgrounds in study, with many having Econ, Education, and Engineering as their backgrounds.

The first one showed that a high rate of the students (97.9 percent) remained actively involved in their online courses, and only one participant initially suggested the possibility of dropping out.

In order to achieve a balanced and fair analysis, the process of data augmentation was implemented to obtain 14 other samples termed as dropouts due to the patterns of at-risk, which included low levels of satisfaction, bad attendance, and insufficient studying hours. It led to a total of 61 records, including 46 active and 15 cases of dropouts, which represents a balanced dataset to train and test the model.

3.2 Model Performance Evaluation

Two machine learning models, which are Logistic Regression (LR) and Decision Tree (DT), were used to predict student dropout. The two models were trained by supervised learning in an augmented dataset.

Stratified k-fold cross-validation (k = 5) was used rather than one train-test split because of the small dataset size needed to ensure higher levels of reliability. Accuracy, precision, recall, and F1-score were used to assess model performance.

The comparison between the logistic regression and decision tree performance in the given question was conducted, and the results were presented in a tabular format.

Table 1. Performance Comparison of Logistic Regression and Decision Tree

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.875	0.92	0.75	0.83
Decision Tree	0.875	0.92	0.75	0.83

The two models demonstrate the same predictive power with an accuracy of 87.5. Precision values imply that both models could accurately detect dropout cases when they happened, and recall values indicate moderate sensitivity to detect all dropout students.

The confusion chart provided in Figure 7 depicts the truth that the majority of the predictions were made in the right direction. Nevertheless, there were also a few cases of misclassifications.

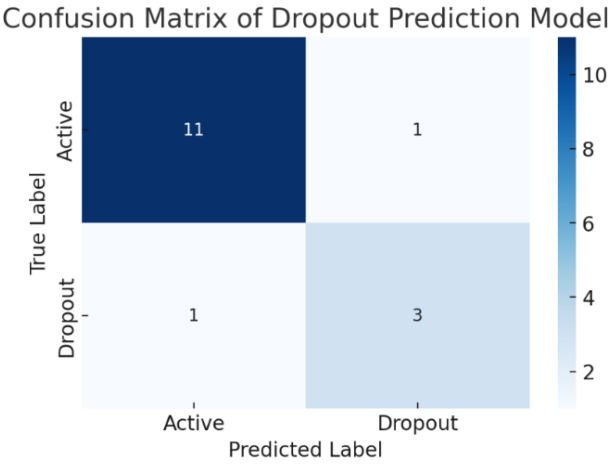

Figure 7. Confusion Matrix of Dropout Prediction Model

Although these findings indicate impressive results, the dataset size and synthetic balancing may have influenced the learning behavior of the models.

3.3 Decision Tree Visualization

Figure 8 indicates the arrangement of the trained Decision Tree classifier.

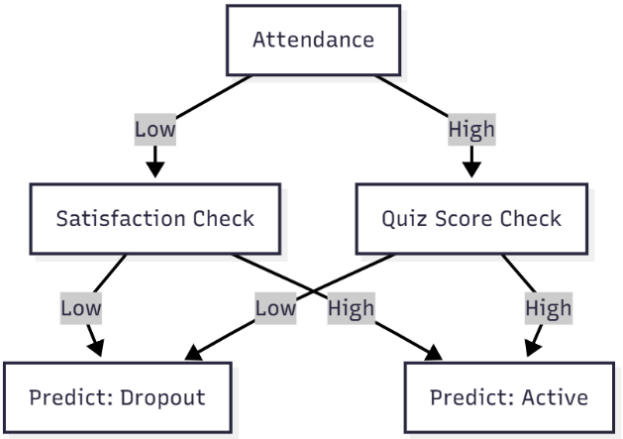

Figure 8. Simplified Decision Tree Structure for Student Dropout Prediction

The tree shows that the main splitting characteristic is attendance, then it is the satisfaction with the learning and quiz performance. Poor attendance and poor satisfaction among students will provide a greater probability of falling under the dropout group, whilst the higher the attendance, with better the quiz grades, the more likely to be classified as an active student. The given structure enhances interpretability as it provides definite decision rules.

3.4 Importance Analysis of Features

The coefficients of the Logistic Regression are a hint at which elements have the greatest impact on the chances of dropping out. The most significant features (in terms of the magnitude of coefficients) are outlined as follows.

Table 2. Interpretation of Feature Importance Based on Logistic Regression

Rank	Feature	Interpretation
1	Satisfaction Level (negative coefficient)	Lack of satisfaction, which was reported by students, led to increased dropouts.
2	Study Hours (negative coefficient)	The fewer the number of hours spent studying per week, the more likely the likelihood of dropping out.
3	Considered Dropout (positive coefficient)	The actual dropout risk was strongly forecasted by the prior considerations of quitting.
4	Assignments Submitted (negative coefficient)	The more assignments a student finished, the less likely they were to drop out
5	Attendance Frequency (negative coefficient)	Low attendance also indicated a risk for dropping out.
6	Quiz/Test Score (negative coefficient)	Lower scores slightly increased dropout likelihood.
7	Gender (minimal effect)	Gender appeared to have little effect on dropout behavior

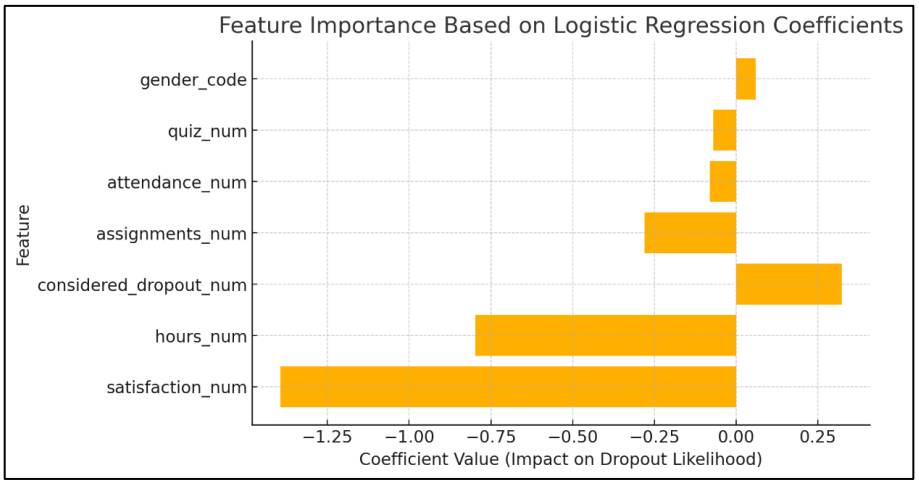

Figure 8. Feature Importance Based on Logistic Regression Coefficients

The coefficients of the negative relate to a lower probability of dropping out, and the coefficients of the positive relate to a higher probability of dropping out. The greatest impact is noted on learning satisfaction and hours spent in study, implying that an increase in engagement greatly reduces the probability of dropping out.

3.5 Discussion & Results

Despite the satisfactory accuracy of the models (87.5), there are a few methodological limitations that should be acknowledged. To begin with, the sample size (47 responses) is small, which raises the chances of overfitting considerably, as there are high chances that the models can be learning noise or dataset-specific correlation and not the generalizable relationship. This can be especially seen with the Decision Tree model, which is overfitting and prone to overfitting on small samples, despite depth restrictions being used (Fan et al., 2021).

Second, synthetic data augmentation may cause bias due to the use of synthetic data. Synthetic samples cannot completely test the real student dropout patterns, although at-risk behavior was modeled using realistic perturbation techniques. As a result, the classifiers might be sensitive to the artificially created relationships to the extent of compromising external validity (Zhang et al., 2023). This drawback explains the necessity to gather more real records of dropouts to be deployed effectively.

Third, the data is collected as a self-report survey, thus it can add recall error and subjectivity to the variables of satisfaction and study hours. These also restrict the predictive power and indicate the necessity to have more realistic sources of data, including log-based learning records and institutional exit data (Fan et al., 2021).

Lastly, cross-validation enhances the accuracy of the reliability estimates but fails to correct the lack of diversity in the data. Xing and Du (2022) stressed that interpretability makes it more useful in the educational process, whereas reliability requires the availability of enough authentic training data. Thus, the interpretation of results must be taken as exploratory as opposed to definitive.

3 Conclusion

By using Logistic Regression and Decision Tree classifiers on data from a multi-university survey, the study investigated the potential of straightforward and understandable machine learning models to forecast student dropout in online education. Academic and

behavioral indicators, such as attendance frequency, quiz performance, assignment completion, study hours, learning satisfaction, and consideration of past dropouts, were all included in the study. The results demonstrated that learning satisfaction, study hours, prior dropout thoughts, and attendance emerged as the most significant predictors, indicating that dropout risk can be fairly accurately identified using these variables. Both models performed similarly, demonstrating that transparent and lightweight models can deliver trustworthy insights without the need for intricate algorithms. Practically speaking, the findings show that even with limited technical resources, educational facilities can adopt explainable early-warning systems. Early identification of at-risk students and the carrying out of focused interventions like academic counseling, progress tracking, and learning support are made possible by the system. Instructors can better understand why a student is expected to drop out thanks to the models' comprehension, which increases system adoption and trust in actual educational settings. Future studies should concentrate on assembling more extensive and varied datasets with actual dropout cases from several universities, integrating longitudinal student records, and assessing hybrid approaches to find a balance between prediction accuracy and interpretability.

REFERENCES

Aldowah, H., Ul Rehman, S., Ghazal, S., & Umar, I. N. (2020). Predicting student performance in MOOCs using machine learning methods. *IEEE Access, 8*, 106420–106433.

Al-Shabandar, R., Hussain, A., Liatsis, P., & Keight, R. (2020). Predicting student performance in MOOCs using machine learning models. *Computers in Human Behavior, 107*, 105–115.

Ashraf, M., Abdullah, S., & Arshad, S. (2021). Student academic performance prediction using interpretable machine learning models. *International Journal of Emerging Technologies in Learning, 16*(9), 4–19.

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 31*(3), 1–26.

Cui, Y., Zhang, Y., & Yang, L. (2022). Explainable machine learning for early warning systems in online education. *IEEE Access, 10*, 20539–20550.

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2020). Current state and future trends in learning analytics. *Journal of Learning Analytics, 7*(3), 43–57.

Fan, C., Wang, J., & Zhang, X. (2021). Interpretable machine learning in education: A systematic review. *Education and Information Technologies, 26*(4), 1–22. https://link.springer.com/article/10.1007/s10639-021-10409-8

He, J., Bailey, J., Rubinstein, B. I. P., & Zhang, R. (2021). Identifying at-risk students using machine learning techniques. *IEEE Transactions on Learning Technologies, 14*(1), 1–13.

Khalil, M., Prinsloo, P., & Slade, S. (2022). Ethics and learning analytics: Student perspectives. *British Journal of Educational Technology, 53*(2), 1–15.

Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2021). Data-driven learning analytics in education. *Computers & Education, 160*, 104020.

Kumar, V., Singh, S., & Sharma, A. (2021). Predicting student dropout in e-learning: A machine learning approach. *Journal of Educational Technology Systems, 49*(4), 529–550.

Mubarak, A. A., Cao, H., & Zhang, W. (2021). Prediction of students' early dropout using machine learning techniques. *IEEE Access, 9*, 42775–42786.

Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics: A tool for teacher development. *Education and Information Technologies, 25*(3), 1–20.

Priya, P., & Sridevi, S. (2022). A comprehensive survey on student dropout prediction models. *International Journal of Information Management Data Insights, 2*(1), 100087.

Stiglic, G., Kocbek, P., Pernek, I., & Kokol, P. (2020). Interpretability of machine learning models in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(5), e1374.

Wong, B. T.-M., & Li, K. C. (2021). Learning analytics intervention in higher education: A review. *Interactive Learning Environments, 29*(8), 1295–1310.

Xing, W., & Du, D. (2022). Dropout prediction in online learning: A data-driven and explainable approach. *International Journal of Emerging Technologies in Learning, 17*(6), 15–29. https://www.online-journals.org/index.php/i-jet/article/view/28295

Zhang, Q., Wang, Y., & Chen, L. (2023). Explainable artificial intelligence for dropout prediction in online education. *Expert Systems with Applications, 213*, 118129. https://doi.org/10.1016/j.eswa.2022.118129