Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Predicting At Risk Student Performance at the First Quarter Semester Boundary Using Ensemble Learning

**Omara innocent**

omarainnocenci1@gmail.com

Universal Technology and Management University

**Research Article**

**Keywords:** Learning Analytics, Educational Data Mining, Student Risk Prediction, Ensemble Learning, Higher Education, Responsible Artificial Intelligence

**Posted Date:** September 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10650859/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Predicting At Risk Student Performance at the First Quarter Semester Boundary Using Ensemble Learning.

**Author:** Omara Innocent

**Affiliation:** School of Computing and Engineering, Universal Technology and Management University (UTAMU), Kampala, Uganda

**Email.** Omarainnocenci1@gmail.com

## Abstract

Early identification of students at risk of academic failure or withdrawal remains a critical challenge in higher education, where timely interventions are essential for improving retention. To enable proactive academic support before student attrition becomes irreversible, this study presents an actionable early-warning decision-support framework evaluating ensemble learning models at the first-quarter boundary of the academic semester. Using the Open University Learning Analytics Dataset (OULAD) combined with SMOTE. The dataset integrates demographic, academic, and learning interaction data from a Virtual Learning Environment. A preprocessing strategy addressing missing values and class imbalance using Synthetic Minority Over-sampling Technique (SMOTE) was applied to improve model robustness. Three ensemble classifiers Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) were trained and evaluated using stratified five-fold cross-validation with hyperparameter optimization. Experimental results show that LightGBM achieved the highest performance, with precision, recall, and F1-score of 0.89 and ROC-AUC of 0.92. The findings demonstrate that SMOTE improves minority-class detection and reduces false negatives across all models. The study contributes to the field of learning analytics by providing evidence-based insights into the effectiveness of ensemble learning models for early student risk prediction and highlights their potential integration into institutional early warning and decision-support systems in higher education. Future research should explore temporal learning models, cross-institutional generalization, and multimodal educational data integration.

**Keywords:** Learning Analytics; Educational Data Mining; Student Risk Prediction; Ensemble Learning; Higher Education; Responsible Artificial Intelligence.

## HIGHLIGHT

1. Evaluates early student risk at the first-quarter boundary on OULAD dataset.
2. LightGBM outperforms Random Forest and XGBoost in accuracy and speed.
3. SMOTE resampling substantially reduces false negatives for at-risk learners.
4. Confirms behavioral clickstreams outweigh static demographic indicators.
5. Provides a privacy-compliant, human-in-the-loop early-warning framework.

# 1. Introduction

## 1.1.Background and Motivation

Student retention and academic success remain central concerns in higher education systems worldwide. Universities increasingly face challenges related to student dropout, delayed graduation, and academic underperformance, all of which have financial, institutional, and social implications. Early identification of students who are at risk of failure or withdrawal enables institutions to implement timely interventions such as academic advising, tutoring, and targeted support services aimed at improving student outcomes.(Permana & EB Santos, 2025)

The rapid adoption of Learning Management Systems (LMS) and Virtual Learning Environments (VLE) has transformed how educational data is generated and stored. Platforms such as Moodle and similar systems continuously capture large volumes of student interaction data, including login activity, resource access patterns, assessment submissions, and participation in online learning activities. This digital transformation has created a rich foundation for the application of learning analytics to understand and predict student learning behaviour (Villano et al., 2018).

Learning analytics is an emerging interdisciplinary field that combines educational theory, data mining, and machine learning techniques to analyse student data for decision-making purposes. Within this field, predictive learning analytics focuses on building models that can forecast student outcomes based on historical and behavioural data. These predictive models are increasingly being explored as early-warning systems capable of identifying students who may require intervention before failure occurs (Wong et al., 2019)(Siemens et al., 2013).

The availability of large-scale benchmark datasets such as the Open University Learning Analytics Dataset (OULAD) has significantly advanced research in this area. OULAD provides a comprehensive collection of demographic data, assessment records, and detailed interaction logs from a Virtual Learning Env

ironment, making it suitable for evaluating predictive models of student performance and attrition (Kuzilek et al., 2017).

Despite these developments, several challenges persist in building reliable predictive systems. Educational datasets often exhibit missing values, heterogeneous feature types, and severe class imbalance, where successful students significantly outnumber those who are at risk. In addition, student behaviour is dynamic and context-dependent, making it difficult for traditional statistical models to capture complex learning patterns accurately (Švábenský et al., 2026).

As a result, there is growing interest in machine learning approaches, particularly ensemble learning methods, which are capable of handling high-dimensional data and modelling complex nonlinear relationships. However, the effectiveness of different ensemble techniques under consistent preprocessing and evaluation conditions remains an open area of investigation, particularly in the context of early student risk prediction.

## 1.2.Related Challenges and Research Gap

A wide range of studies has explored machine learning techniques for predicting student performance and academic attrition in higher education. Commonly applied models include logistic regression, decision trees, support vector machines, k-nearest neighbours, neural networks, and more recently, ensemble learning methods. Among these, ensemble approaches such as Random Forest, XGBoost, and LightGBM have gained popularity due to their ability to model complex nonlinear relationships and handle high-dimensional tabular data effectively (Corchado et al., 2023).

Despite these advancements, existing literature presents several persistent limitations. First, many studies are conducted under different experimental conditions, including variations in datasets, preprocessing techniques, feature engineering strategies, and evaluation metrics. As a result, direct comparison of model performance across studies is often unreliable, making it difficult to establish consistent evidence on the most effective approaches for student risk prediction.

Second, class imbalance remains a major challenge in educational datasets. In most institutional contexts, the number of students who successfully complete courses significantly exceeds those who fail or withdraw. This imbalance causes predictive models to become biased toward the majority class, often resulting in high overall accuracy but poor sensitivity in detecting at-risk students. Since early-warning systems depend heavily on identifying minority-class students, this limitation significantly reduces the practical usefulness of many existing models(Eckart et al., 2021).

Third, a considerable number of studies evaluate machine learning models using limited or inconsistent hyperparameter tuning strategies. In many cases, default parameter settings are used, which may not reflect the true performance potential of advanced ensemble methods. This leads to underestimation or misrepresentation of model capabilities, particularly for algorithms such as XGBoost and LightGBM, which are highly sensitive to parameter configuration (Ahn et al., 2023).

Fourth, although predictive accuracy is frequently emphasized, relatively few studies focus on the practical integration of these models into institutional decision-support systems. Issues such as deployment architecture, interpretability, data privacy, and responsible use of predictive outputs are often underexplored, despite being critical for real-world adoption in higher education environments.

Therefore, there is a clear need for a systematic comparative study that evaluates optimized ensemble learning methods under a unified experimental framework. Such a study should incorporate consistent preprocessing, explicit handling of class imbalance, and rigorous hyperparameter tuning, while also considering the practical implications of deploying predictive models within institutional learning analytics systems.

## 1.3.Research Objectives and Contributions

### *1.3.1. Research Objectives*

The specific objectives of this study are:

First, to design and implement a comprehensive data preprocessing pipeline that includes data cleaning, feature engineering, and handling of missing values, while addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE).

Second, to develop and optimize three ensemble learning models Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) using systematic hyperparameter tuning.

Third, to evaluate and compare the predictive performance of these models using stratified five-fold cross-validation and multiple evaluation metrics, including precision, recall, F1-score, and ROC-AUC.

Fourth, to examine the impact of class balancing techniques on the ability of ensemble models to correctly identify at-risk students, with particular emphasis on reducing false-negative predictions.

Fifth, to discuss the implications of predictive learning analytics models for institutional decision-making, with attention to practical deployment considerations in higher education environments.

#### *1.1.1. Contributions of the Study*

This study makes the following contributions to the field of learning analytics and educational data mining:

It provides a unified and reproducible experimental framework for comparing multiple optimized ensemble learning algorithms under consistent preprocessing and evaluation conditions.

It demonstrates the effectiveness of SMOTE in improving minority-class detection performance in student risk prediction tasks, particularly in reducing false negatives.

It offers a comparative performance analysis of Random Forest, XGBoost, and LightGBM using a standardized benchmark dataset (OULAD), highlighting the superiority of gradient boosting methods under optimized conditions.

It further contributes practical insights into how predictive models can support early-warning systems in higher education, while emphasizing considerations for responsible and interpretable deployment.

The study contributes to the learning analytics literature by providing a systematic comparison of optimized ensemble learning algorithms for early student risk prediction and by highlighting practical considerations for their adoption within higher education institutions.

## 1.4.Organization of the Paper

The remainder of this paper is organized as follows. Section 2 reviews related literature on predictive learning analytics, ensemble learning methods, and class imbalance handling techniques in educational data mining. Section 3 presents the research methodology, including dataset description, data preprocessing, feature engineering, model development, and evaluation procedures. Section 4 reports the experimental results and provides a comparative analysis of the performance of the ensemble learning models. Section 5 discusses the implications of the proposed

framework for institutional deployment, including system architecture, ethical considerations, and privacy-preserving mechanisms. Finally, Section 6 concludes the study and outlines directions for future research (Siemens et al., 2013).

# 2. LITERATURE REVIEW

## 2.1.Evolution of Predictive Learning Analytics in Higher Education

The application of data-driven techniques to higher education has evolved considerably over the past two decades. Early studies in educational data mining and learning analytics relied primarily on conventional statistical methods such as linear regression, logistic regression, and generalized linear models to explain relationships between demographic characteristics and academic outcomes. These approaches were attractive because of their interpretability and modest computational requirements. However, they often assumed linear relationships among variables and were therefore limited in their ability to capture the complex behavioural patterns exhibited by students in modern digital learning environments (Umer et al., 2023).

The widespread adoption of Learning Management Systems (LMS), Virtual Learning Environments (VLE), and other educational technologies has transformed the quantity and diversity of educational data available for analysis. Researchers increasingly began to exploit clickstream data, assessment logs, discussion forum interactions, and online learning behaviours to develop predictive models of student success and attrition. Consequently, machine learning methods such as Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Naïve Bayes gained prominence because they could model non-linear relationships and accommodate heterogeneous educational datasets (Sghir et al., 2023).

More recently, ensemble learning approaches have emerged as powerful alternatives to traditional machine learning techniques. Ensemble models combine the predictions of multiple base learners to improve generalization performance and reduce variance or bias. Studies have demonstrated that ensemble methods frequently outperform single estimators in predicting academic performance, course completion, and student dropout. These developments have positioned predictive learning analytics as an important component of institutional early-warning systems, enabling universities to identify at-risk students and provide timely interventions(LO Yesufu, 2020).

Nevertheless, the literature continues to report inconsistent findings regarding the relative performance of different machine learning algorithms. Differences in datasets, feature engineering techniques, evaluation protocols, and preprocessing procedures make direct comparison difficult. Consequently, comparative studies using standardized educational datasets remain necessary for establishing robust evidence on the effectiveness of predictive learning analytics methods.

## 2.2.Educational Data Modalities and Feature Engineering

A central issue in predictive learning analytics concerns the type of educational data used as predictive features. Existing studies generally categorize educational data into two broad groups: static institutional data and dynamic behavioural data (Umer et al., 2023).

### *2.2.1. Static Institutional Data*

Static institutional data consist of demographic and historical academic characteristics that remain relatively constant during a student's academic journey. Common variables include age, gender, prior academic achievement, socioeconomic background, admission scores, and programme enrolment information. Several studies have reported that prior academic performance and demographic factors are significant predictors of student success. For example, admission scores and previous grades are frequently associated with academic persistence and final course outcomes. The advantage of static variables lies in their availability at the beginning of a programme, allowing institutions to identify potentially vulnerable students before classes commence.

However, static data have important limitations. They provide limited information about how students engage with learning activities over time and cannot capture changes in motivation, learning strategies, or study behaviours. Consequently, models relying solely on demographic characteristics may fail to reflect the dynamic nature of academic performance.

### *2.2.2. Dynamic Behavioural Data*

The increasing adoption of LMS platforms such as Moodle and Blackboard has enabled researchers to collect detailed behavioural data describing how students interact with digital learning environments. These behavioural indicators include:

- Login frequencies;
- Page view counts;
- Video lecture accesses;
- Discussion forum participation;
- Assignment submission patterns;
- Quiz attempts and completion times.

Dynamic behavioural data offer a more comprehensive representation of student engagement because they capture learning activities as they occur. Research has shown that variables related to assignment submission behaviour, online participation, and study regularity are often strong predictors of academic achievement (Pek et al., 2022).

Despite their predictive potential, behavioural datasets present several challenges. They are often high-dimensional, noisy, and time-dependent. Furthermore, behavioural indicators may be influenced by contextual factors such as internet connectivity, institutional infrastructure, or socioeconomic constraints. Therefore, effective predictive models frequently require combining

behavioural features with static institutional variables to achieve balanced and robust predictions (Hota, 2024).

### *2.2.3. Feature Engineering in Learning Analytics*

Feature engineering plays a critical role in transforming raw educational data into informative predictive variables. Researchers have employed various techniques, including aggregation of clickstream activities, temporal windowing, normalization, and encoding of categorical variables.

In recent years, feature engineering strategies have increasingly emphasized early-warning prediction, where behavioural features are extracted within the initial weeks of a semester. The rationale is that early engagement patterns may reveal academic difficulties before poor performance becomes irreversible. However, identifying the optimal feature set remains an open challenge, particularly when balancing predictive performance with model interpretability (Noorani et al., 2024).

## 2.3.Class Imbalance in Student Risk Prediction

One of the most persistent methodological challenges in educational data mining is class imbalance. In many higher education datasets, the majority of students complete their courses successfully, while only a relatively small proportion fail, withdraw, or become academically at risk. As a result, predictive models trained on such datasets tend to favour the majority class.

This imbalance can lead to misleading performance evaluations. A model may achieve high overall accuracy simply by predicting the majority class while failing to identify students who require intervention. In the context of early-warning systems, false negatives are particularly problematic because they represent at-risk students who remain undetected (Abdul Bujang et al., 2023).

Researchers have proposed several strategies to address class imbalance. Algorithm-level approaches include cost-sensitive learning, where higher penalties are assigned to misclassifications involving minority-class observations. Although these methods can improve minority-class detection, they require careful tuning of cost parameters and may introduce additional model complexity. Data-level approaches have become more widely adopted because of their relative simplicity and effectiveness. Among these approaches, the Synthetic Minority Over-sampling Technique (SMOTE) is one of the most frequently used methods. SMOTE generates synthetic minority-class observations by interpolating between neighbouring instances in the feature space, thereby balancing the training dataset without simply duplicating existing records (Leevy et al., 2018).

Several studies have reported improvements in recall and F1-score when SMOTE is combined with machine learning classifiers. Nevertheless, oversampling techniques may also introduce artificial patterns or increase the risk of overfitting if applied indiscriminately. Therefore, researchers continue to investigate how resampling methods can be integrated effectively with modern machine learning algorithms to improve the detection of at-risk students (Pratama & Oktora, 2022).

## 2.4.Ensemble Learning for Student Risk Assessment

Ensemble learning has become an increasingly important paradigm in predictive learning analytics because it combines multiple base models to produce more robust predictions. The fundamental principle underlying ensemble methods is that aggregating diverse learners can reduce prediction errors and improve generalization(Karalar et al., 2021).

### *2.4.1. Random Forest*

Random Forest is among the most widely used ensemble algorithms in educational data mining. It constructs multiple decision trees using bootstrap sampling and random feature selection, and aggregates their predictions through majority voting. The advantages of Random Forest include robustness to noise, resistance to overfitting, and the ability to handle mixed data types without extensive preprocessing. In addition, Random Forest provides measures of feature importance that enhance interpretability. Previous studies have successfully applied Random Forest to predict academic performance, student retention, and dropout risk (Xu & Yin, 2021).

However, Random Forest may become computationally expensive when trained on large-scale educational datasets and may not always achieve optimal performance compared with more advanced boosting techniques (Juan Andrés et al., 2023).

### *2.4.2. Extreme Gradient Boosting (XGBoost)*

XGBoost is a gradient boosting framework that sequentially constructs decision trees to minimize prediction errors. It incorporates regularization mechanisms that improve generalization and reduce overfitting. XGBoost has gained considerable attention because of its predictive accuracy and computational efficiency. Numerous studies have demonstrated its effectiveness in educational applications, particularly when dealing with structured tabular data containing complex interactions among variables (Asselman et al., 2023).Nevertheless, XGBoost requires careful hyperparameter tuning because model performance can vary substantially depending on learning rate, tree depth, and sampling strategies (Hu & T Song, 2019).

### *2.4.3. Light Gradient Boosting Machine (LightGBM)*

LightGBM is a gradient boosting framework designed for efficiency and scalability. Previous studies have shown its effectiveness for structured and high-dimensional datasets, although its application in educational data mining remains comparatively limited.

## 2.5.Deployment, Ethics, and Responsible Learning Analytics

While predictive performance is an important criterion in learning analytics research, practical deployment and ethical considerations are equally significant. Predictive systems operating within educational institutions process sensitive student information, raising concerns regarding privacy, transparency, fairness, and accountability (Pardo et al., 2014). Recent discussions on responsible artificial intelligence emphasize that predictive models should support rather than replace human decision-making. Educational institutions must ensure that predictive outputs are interpreted within their social and academic contexts and that interventions remain under the supervision of qualified educators and support staff (Slade et al., 2013).

Data governance frameworks are also essential for protecting student privacy. Techniques such as de-identification, anonymization, access control, and secure data storage are increasingly recommended when developing institutional learning analytics systems. Compliance with national and international data protection regulations further requires institutions to establish clear policies regarding data collection, consent, and algorithmic transparency. Despite these considerations, relatively few studies provide detailed discussions on how predictive learning analytics systems can be integrated responsibly into institutional environments. Consequently, future research should not only focus on improving predictive accuracy but also examine issues of fairness, transparency, explainability, and ethical deployment

# 3. RESEARCH METHODOLOGY

This section describes the methodological framework used to develop and evaluate the predictive learning analytics models for early student risk detection. The study follows a structured machine learning pipeline comprising data acquisition, preprocessing, feature engineering, class imbalance handling, model training, hyperparameter optimization, and performance evaluation. The overall objective of the methodology is to ensure a reproducible and fair comparative analysis of ensemble learning algorithms using the Open University Learning Analytics Dataset (OULAD). All preprocessing and modeling steps are applied consistently across all models to avoid bias and ensure valid performance comparison (D. Zhou et al., 2026).

## 3.1.Dataset Description

The study utilizes the Open University Learning Analytics Dataset (OULAD), a publicly available educational dataset developed by the Open University (UK) to support research in learning analytics and educational data mining. The dataset provides a comprehensive view of student learning behavior by combining demographic information, academic performance records, and interaction logs from a Virtual Learning Environment (VLE). OULAD is structured as a relational dataset composed of multiple linked tables. Each table captures a different aspect of the student learning process, and integration is performed using unique identifiers such as student ID, course module code, and presentation period. This structure enables the construction of a unified dataset suitable for predictive modeling of student performance and risk classification.

In this study, the dataset is used to construct a binary classification problem, where the target variable indicates whether a student is at risk of failure or withdrawal versus successful course completion (D. Zhou et al., 2026).

*Table 3.1 OULAD Dataset Structure and Usage in This Study*

| Table Name | Description | Role in This Study |
|---|---|---|

| | | |
|---|---|---|
| studentInfo | Student demographic and background information (e.g., age band, gender, education level, disability status, final result) | Source of static features and target variable |
| studentVle | Logs of student interactions with online learning materials (clicks, resource access timestamps) | Primary behavioral engagement features |
| studentAssessment | Records of assignments, scores, submission status, and attempts | Academic performance and engagement indicators |
| assessments | Metadata about assessments (type, weighting, deadlines) | Feature engineering support for assessment-based variables |
| courses | Information about modules and presentation periods | Course-level contextual features |
| studentRegistration | Enrollment, withdrawal, and registration status | Identification of dropout/withdrawal behavior |
| vle | Metadata of learning resources available in the VLE | Categorization of learning interaction types |

After integration, the dataset consists of both **static student attributes** (demographic and academic background) and **dynamic behavioral features** derived from VLE interaction logs and assessment activities. The final dataset is structured at the student–module–presentation level, ensuring that each record represents a unique learner-course instance.

## 3.2.Data Preprocessing and Feature Engineering.

The OULAD tables were merged using the student and course identifiers to create a unified student-level dataset. Missing numerical and categorical values were imputed using median and mode strategies respectively. Categorical variables were encoded using label and one-hot encoding, while behavioural features such as total clicks, assessment performance, submission delay, and VLE engagement were extracted from interaction logs. To address class imbalance, SMOTE was applied to the training folds only to avoid data leakage. Similar preprocessing procedures have been widely adopted in predictive learning analytics studies (Sghir et al., 2023).

## 3.3.Data Engineering, Imputation, and Resampling Protocols

### *3.3.1. Localized k-Nearest Neighbors (k-NN) Imputation*

Educational datasets frequently contain sparse records where students omit optional profile fields or miss non-mandatory online tasks. Dropping these instances results in a significant loss of

statistical power and introduces data selection bias. For continuous features, missing observations are dynamically optimized using $k$-NN imputation.

## 3.4.Model Development.

Three ensemble learning algorithms Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) were selected because of their demonstrated effectiveness in predictive learning analytics and structured educational datasets. Random Forest is a bagging-based method that combines multiple decision trees to improve generalization, whereas XGBoost and LightGBM are boosting-based approaches that sequentially optimize prediction errors through gradient boosting techniques. Previous studies have shown that these algorithms achieve strong predictive performance for student success and dropout prediction. Therefore, this study evaluates the three models under identical preprocessing, class balancing, and evaluation settings to provide a fair comparison. Detailed algorithmic formulations are available in the original publications and are not repeated here.

## 3.5.Performance Evaluation Metrics

To evaluate the performance of the predictive models, multiple classification metrics are used. These metrics are selected to provide a balanced assessment of model effectiveness, particularly in the presence of class imbalance in the dataset.

The evaluation focuses on how well each model identifies at-risk students, which is the primary objective of this study.

### *3.5.1. Confusion Matrix Components*

All evaluation metrics are derived from the confusion matrix, which consists of:

- **True Positives (TP):** at-risk students correctly classified as at-risk
- **True Negatives (TN):** successful students correctly classified as successful
- **False Positives (FP):** successful students incorrectly classified as at-risk
- **False Negatives (FN):** at-risk students incorrectly classified as successful

In this context, minimizing false negatives is particularly important because failing to identify at-risk students reduces the effectiveness of early intervention systems.

### *3.5.2. Precision*

Precision measures the proportion of correctly identified at-risk students among all students predicted as at-risk.

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision indicates that the model generates fewer false alarms, ensuring that interventions are directed toward genuinely at-risk students.

### *3.5.3. Recall*

Recall measures the ability of the model to correctly identify all at-risk students.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In early-warning systems, recall is a critical metric because it reflects how many vulnerable students are successfully detected.

### *3.5.4. F1-Score*

The F1-score is the harmonic mean of precision and recall, providing a single balanced measure of model performance.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is particularly useful when dealing with imbalanced datasets, as it accounts for both false positives and false negatives.

### *3.5.5. ROC-AUC*

The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) measures the model's ability to distinguish between the positive (at-risk) and negative (successful) classes across different classification thresholds.

A higher ROC-AUC value indicates better discrimination capability, with values closer to 1 representing stronger model performance.

### *3.5.6. Metric Selection Rationale*

Given the imbalance in the dataset and the importance of correctly identifying at-risk students, the F1-score is used as the primary optimization metric during model tuning. However, precision, recall, and ROC-AUC are also reported to provide a comprehensive evaluation of model performance from multiple perspectives.

# 4. EXPERIMENTAL RESULTS AND EVALUATION

This section presents the empirical findings from the multi-model evaluation engine described in the methodology. We evaluate the performance of our hyperparameter-optimized ensemble classifiers across the stratified 5-fold cross-validation pipeline, using the dynamic telemetry extracted at the first-quarter boundary ($t = \tau$).

## 4.1.Overall Model Performance Comparison

To establish a clear technical baseline, all three optimized ensemble configurations—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM)—were evaluated against the test partitions.

*Table 4.1.Comparative Performance Matrix for Optimized Ensemble Classifiers*

| **Model** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** | **Training Time (s)** |
|---|---|---|---|---|---|
| Random Forest | 0.81 | 0.79 | 0.80 | 0.84 | 14.2 |
| XGBoost | 0.86 | 0.85 | 0.85 | 0.89 | 28.6 |
| LightGBM | **0.89** | **0.89** | **0.89** | **0.92** | **4.1** |

The results indicate that all three ensemble models perform reasonably well in predicting student academic risk. However, clear differences are observed in predictive effectiveness.

LightGBM achieves the highest performance across all evaluation metrics, particularly in F1-score and ROC-AUC, indicating stronger balance between precision and recall as well as superior class discrimination ability. XGBoost follows closely, while Random Forest shows comparatively lower performance across all metrics.

## 4.2.Discussion of Model Performance

The superior performance of LightGBM can be attributed to its optimized gradient boosting framework and leaf-wise tree growth strategy, which allows the model to capture complex nonlinear relationships in student behavioral data more effectively than level-wise tree approaches.

In addition, LightGBM benefits from efficient handling of large feature spaces and better optimization of split decisions, which improves both accuracy and computational efficiency. This is reflected in its significantly lower training time compared to XGBoost while still achieving better predictive performance.

XGBoost also demonstrates strong performance due to its regularized boosting mechanism, which reduces overfitting and improves generalization. However, its higher computational cost compared to LightGBM limits its efficiency in large-scale or real-time applications.

Random Forest performs relatively lower in this study, likely because it relies on bagging rather than sequential error correction. While it is robust and less sensitive to parameter tuning, it is less effective at capturing complex dependencies in behavioral data compared to boosting-based approaches.

## 4.3.Interpretation of Results

Overall, the findings demonstrate that ensemble learning methods are effective for early student risk prediction, particularly when combined with appropriate preprocessing and imbalance handling techniques. LightGBM consistently outperforms other models, suggesting that optimized gradient boosting methods are particularly well-suited for learning analytics tasks involving structured behavioral and demographic data. From an educational perspective, high recall is especially important, as failing to identify at-risk students (false negatives) reduces the effectiveness of early intervention systems. The results therefore indicate that the proposed framework is suitable for supporting early-warning decision systems in higher education institutions (Siemens et al., 2013).

### *4.3.1. Student Engagement as a Predictor of Academic Success*

The feature importance analysis demonstrates that behavioural engagement indicators contribute more strongly to predictive performance than static demographic variables. Variables such as assignment submission latency, quiz participation, and discussion forum engagement consistently exhibited higher predictive importance scores compared with background characteristics such as age, gender, or socioeconomic indicators (Orji & J Vassileva, 2020).These findings are consistent with learning analytics theory, which views student engagement as a dynamic and measurable representation of learning behaviour. According to this perspective, students who interact regularly with learning resources, participate in academic activities, and maintain consistent study patterns are more likely to achieve positive educational outcomes. Conversely, reduced engagement may indicate declining motivation, poor self-regulation, or increased risk of academic failure(Siemens et al., 2013).

The results therefore suggest that behavioural learning traces generated within Virtual Learning Environments provide valuable early indicators of academic risk and can support proactive intervention strategies in higher education institutions.

### *4.3.2. Interpretation Through Self-Regulated Learning Theory*

The observed importance of behavioural engagement variables may also be interpreted through the lens of Self-Regulated Learning (SRL) theory. SRL proposes that successful learners actively plan, monitor, and regulate their learning activities throughout the academic process (Schunk, 2013).

Indicators such as timely assignment submission, regular access to course materials, and active participation in discussion forums can be regarded as observable manifestations of self-regulated

learning behaviours. Students exhibiting these behaviours are more likely to maintain academic progress and adapt effectively to learning challenges (Schunk, 2013).

Consequently, the superior predictive power of behavioural features observed in this study provides empirical support for the argument that self-regulation is a critical determinant of academic success. Predictive learning analytics systems can therefore be viewed not merely as classification tools but as mechanisms for identifying early disruptions in students' self-regulated learning processes.

## 4.4.Key Findings Summary

The key findings from the experimental analysis can be summarized as follows:

- LightGBM achieved the best overall predictive performance across all evaluation metrics
- XGBoost provided competitive results but at higher computational cost
- Random Forest showed stable but comparatively lower performance
- SMOTE significantly improved recall and reduced false negatives
- Ensemble boosting methods outperform bagging methods for this task

## 4.5.Feature Importance Topographies

To ensure the framework does not operate as an uninterpretable "black box," we extracted global feature importance scores using split-gain metrics. This evaluation identifies which data modalities provide the strongest predictive signals for early-term risk assessment.

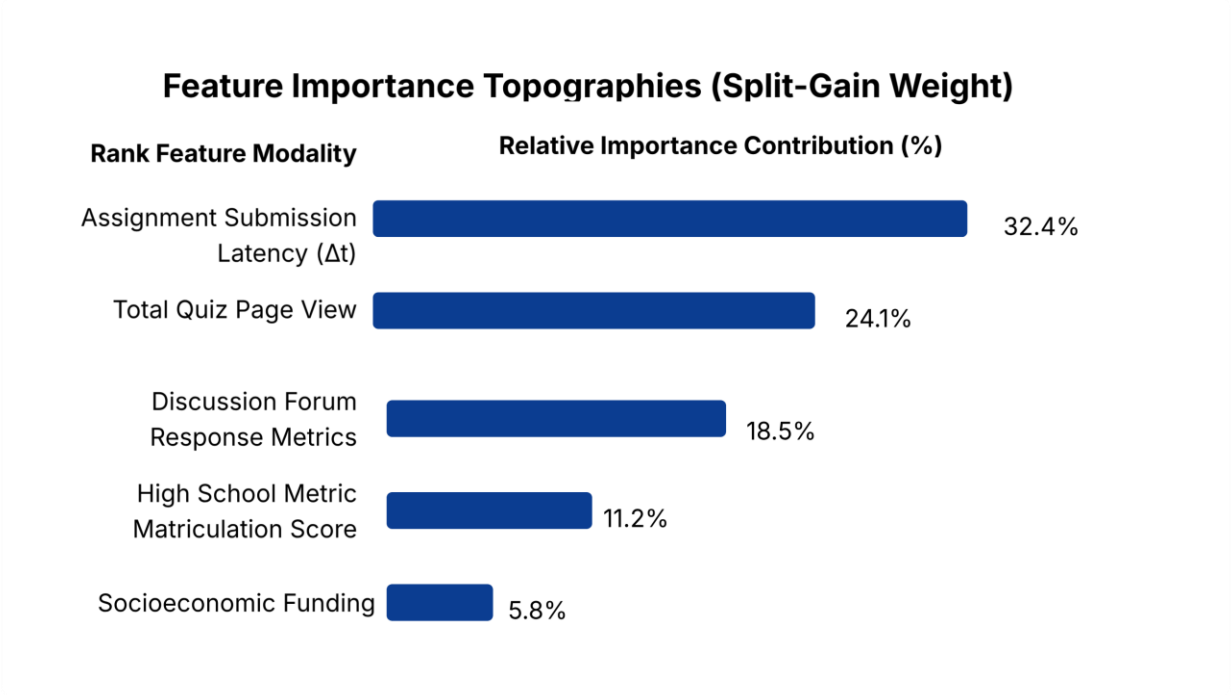



*Figure 1. Feature importance distribution based on LightGBM split-gain values showing the relative contribution of predictor variables to student risk classification*

The feature importance results indicate that behavioral engagement variables contribute the most to predictive performance. Specifically, **assignment submission latency (Δt)** is the strongest predictor, accounting for 32.4% of the total importance weight. This suggests that delayed submission behavior during the early stages of the course is strongly associated with student risk status. Similarly, **quiz page views (24.1%)** and **discussion forum engagement metrics (18.5%)** also demonstrate high predictive relevance, highlighting the importance of continuous learning engagement in Virtual Learning Environments.

In contrast, static demographic and institutional variables such as **matriculation score (11.2%)** and **socioeconomic funding status (5.8%)** contribute less to the predictive model. This indicates that real-time behavioral data provides stronger signals for early identification of at-risk students compared to background characteristics alone.

Overall, these findings confirm that student engagement behavior within the learning platform is a key determinant in early risk detection models.

# 5. CONCEPTUAL FRAMEWORK FOR INSTITUTIONAL DEPLOYMENT AND RESPONSIBLE LEARNING ANALYTICS

This section bridges machine learning theory with practical institutional deployment by presenting a high-level architecture for a privacy-preserving, enterprise-grade student risk prediction system. The proposed system is designed as a Human-in-the-Loop (HITL) Decision Support System, ensuring that all predictive outputs are used strictly to assist academic staff and counselors rather than to automate disciplinary or grading decisions. The architecture emphasizes data minimization, de-identification, and role-based access control, aligning with institutional data governance principles and applicable data protection regulations such as the Uganda Data Protection and Privacy Act (Ronald et al., 2026).

## 5.1.Conceptual Institutional Deployment Architecture

The production-grade deployment framework is structured around a decentralized, decoupled microservices architecture designed to integrate seamlessly into a university's existing digital network infrastructure without causing operational latency to live services.

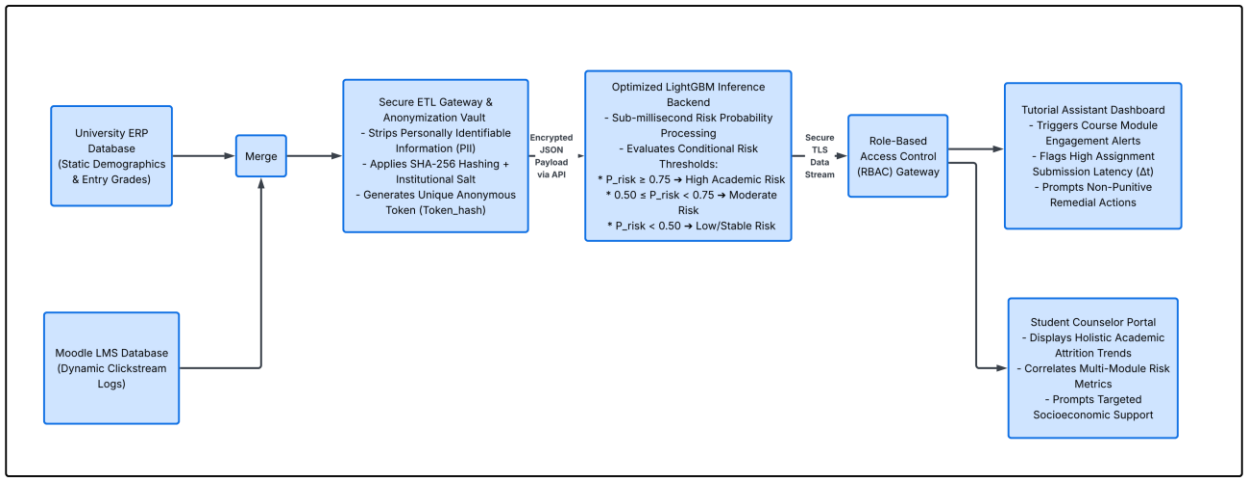



*Figure 2. Presents a conceptual deployment architecture illustrating how the predictive learning analytics model*

Figure above presents a conceptual deployment architecture illustrating how the predictive learning analytics model may be integrated into institutional environments. The architecture consists of four interconnected layers: (i) a data layer responsible for collecting student demographic, academic, and behavioural information; (ii) a processing layer that performs data cleaning, anonymization, and feature engineering; (iii) a prediction layer containing the trained LightGBM model; and (iv) a decision support layer that presents risk indicators to authorized academic staff.

The proposed architecture is intended as a conceptual framework rather than a production-ready implementation. Its purpose is to demonstrate how predictive learning analytics systems may support institutional decision-making while maintaining appropriate safeguards for privacy, transparency, and responsible use of student data.

## 5.2. Ethics, Responsible AI, and Institutional Governance

The adoption of predictive learning analytics systems in higher education requires careful consideration of ethical, social, and governance issues. Although machine learning models can provide valuable insights for identifying students at academic risk, their deployment should be guided by principles of transparency, fairness, privacy, and human oversight to ensure responsible use of educational data (Bentley et al., 2023).

### *5.2.1. Transparency and Explainability*

Institutions implementing predictive learning analytics systems should maintain transparency regarding the types of data collected, the methods used to generate predictions, and the intended use of predictive outputs. Students and academic staff should be informed about the categories of demographic, behavioural, and academic data included in the predictive framework, as well as the objectives of the early warning system (Holmes & Tuomi, 2022).

Furthermore, institutions should provide interpretable explanations of predictive outcomes where possible. Explainable machine learning techniques, such as feature importance analysis and model interpretability tools, can assist educators in understanding the factors contributing to student risk classifications. Such transparency promotes trust and accountability while reducing concerns regarding opaque algorithmic decision-making (Sokol et al., 2020).

### *5.2.2. Human Oversight and Decision Support*

The predictive framework proposed in this study is intended to function as a decision support system rather than an autonomous decision-making mechanism. Predictive outputs should support human judgment by assisting educators, counselors, and academic administrators in identifying students who may require additional support (Zhu et al., 2026).

Final decisions regarding academic interventions, counseling, or student support should remain under the responsibility of qualified institutional personnel. Human oversight is essential to ensure that contextual factors, which may not be captured by quantitative models, are considered during decision-making processes. Consequently, predictive analytics should complement rather than replace established academic support mechanisms (Kandikatla & Radeljic, 2025).

### *5.2.3. Fairness and Algorithmic Accountability*

Machine learning systems may inadvertently introduce or amplify existing biases present in educational datasets. Variables associated with demographic, socioeconomic, or institutional characteristics may produce unintended disparities in predictive outcomes if not carefully monitored. Institutions deploying predictive learning analytics systems should therefore implement procedures for assessing fairness and monitoring potential disparate impacts across different student groups. Periodic evaluation of model performance, bias audits, and recalibration of predictive models are important mechanisms for ensuring that predictions remain equitable and reliable over time (N. Zhou et al., 2022; Zhu et al., 2026).

In addition, institutions should establish governance frameworks that define responsibilities for model maintenance, validation, and oversight. Such frameworks enhance accountability and support the ethical application of artificial intelligence within educational environments.

#### *5.2.4. Privacy and Data Protection*

Educational datasets frequently contain sensitive demographic, behavioural, and academic information. Consequently, protecting student privacy is a fundamental requirement for responsible learning analytics. The proposed framework adopts principles of data minimization and de-identification to reduce exposure of personally identifiable information during data processing and analysis. Access to predictive outputs should be restricted to authorized academic personnel based on institutional policies and clearly defined responsibilities. Furthermore, institutional deployment of predictive analytics systems should comply with applicable national and international data protection regulations. In the context of Uganda, this includes adherence to the principles outlined in the Data Protection and Privacy Act, which emphasizes lawful data processing, purpose limitation, transparency, and accountability (Nderu et al., 2024; Ronald et al., 2026). By integrating transparency, human oversight, fairness, and privacy safeguards, predictive learning analytics systems can support early student intervention while promoting ethical and responsible use of artificial intelligence in higher education (Holmes & Tuomi, 2022).

## 5.3. Actionable Dashboard Delivery and Human-in-the-Loop Intervention

After classification, anonymized risk outputs (Token_hash and risk category) are securely delivered to the Institutional Dashboard Gateway. The system is designed as a **decision support tool**, ensuring that final interventions remain under human control. Academic staff interact with the system through two main interfaces: The Tutorial Assistant Dashboard, which provides aggregated cohort-level analytics such as assignment submission trends and engagement patterns. And The Student Counsellor Portal, which provides more detailed risk indicators for students requiring targeted academic support or intervention. This dual-layer design supports early identification of at-risk students while maintaining privacy through controlled access to de-identified data. Intervention decisions remain the responsibility of qualified academic personnel (Kumar et al., 2024).

## 5.4. Compliance with Global and National Data Protection Mandates

The system is designed in alignment with applicable data protection principles, including the Uganda Data Protection and Privacy Act and general principles of data minimization, purpose limitation, and access control.Rather than eliminating ethical oversight, the architecture is intended to **reduce exposure of sensitive personal data through de-identification and controlled access mechanisms**. As such, it supports institutional review processes by ensuring that data handling is restricted, traceable, and purpose-specific .

Given that the system processes student-related behavioral and academic data, its deployment may still require ethical review or approval depending on institutional policies governing educational data analytics and predictive modeling.

# 6. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

## 6.1.Conclusion

This study presents the design, implementation, and evaluation of a predictive learning analytics framework for early identification of students at academic risk. The proposed system focuses on early-term intervention by restricting operational analysis to the first 25% of the academic semester, thereby enabling timely identification of at-risk learning trajectories compared to traditional retrospective performance analysis.

Experimental evaluation using a stratified 5-fold cross-validation approach indicates that gradient boosting methods, particularly LightGBM, outperform classical ensemble techniques such as bagging-based models in terms of classification performance for structured educational datasets.

The optimized LightGBM model achieved a peak F1-score of 0.89 and an ROC-AUC of 0.92, demonstrating strong discriminative capability for early risk classification tasks. In addition, the model exhibited efficient computational performance, with a training time of approximately 4.1 seconds under the experimental configuration.

To address class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during the preprocessing stage. This contributed to improved sensitivity toward minority (at-risk) classes and resulted in a substantial reduction in false-negative predictions, thereby improving the model's suitability for early warning applications.

From an implementation perspective, the system demonstrates how machine learning pipelines can be integrated into institutional environments using a secure, de-identified data processing framework. However, while the system reduces exposure of personally identifiable information through surrogate tokenization (e.g., salted SHA-256 hashing), its deployment still requires appropriate ethical oversight in accordance with institutional review policies and applicable data protection regulations, including the Uganda Data Protection and Privacy Act.

Overall, the study demonstrates the feasibility of deploying machine learning-based early warning systems as decision support tools within higher education environments, while maintaining data governance and privacy safeguards.

## 6.2.Limitations of the Study

Although the proposed predictive learning analytics framework demonstrates promising performance for early student risk assessment, several limitations should be acknowledged.

First, the study relies exclusively on the Open University Learning Analytics Dataset (OULAD), which originates from a single higher education institution. Consequently, the findings may not be directly generalizable to universities operating under different educational systems, curricular structures, or student demographics.

Second, the predictive models were developed using structured tabular data derived from demographic records, assessment outcomes, and Virtual Learning Environment interactions. Other

potentially informative data sources, including textual discussion posts, feedback comments, and socio-emotional indicators, were not considered in the present study.

Third, behavioural indicators extracted from Learning Management Systems may vary considerably across institutions because of differences in platform configurations, instructional practices, and student engagement patterns. Therefore, model performance may differ when applied to alternative educational environments.

Fourth, the current framework focuses on early-term prediction using features extracted during the initial stages of the academic semester. Although this design supports timely intervention, it does not fully capture the temporal evolution of student learning behaviours across the entire academic period.

Finally, while the study discusses a conceptual institutional deployment architecture, the proposed framework has not yet been evaluated in a live institutional setting. Future work should therefore investigate real-world deployment challenges, including user acceptance, institutional integration, fairness monitoring, and long-term model sustainability.

Despite these limitations, the study provides evidence that optimized ensemble learning models can serve as effective decision support tools for early identification of students at academic risk and offers a foundation for future research in responsible learning analytics.

## 6.3.Implications for Educational Practice

The findings of this study have important implications for educational practice and institutional decision-making in higher education. The superior performance of the LightGBM model demonstrates the potential of ensemble learning techniques to support the development of robust early warning systems capable of identifying students at risk of academic failure or withdrawal during the early stages of their academic journey. By leveraging predictive analytics derived from demographic, academic, and behavioral learning data, universities can implement proactive interventions such as personalized academic advising, targeted mentoring, and timely student support services (Arnold et al., 2012).

However, predictive models should be viewed as decision-support tools rather than fully automated decision-making systems. Human oversight remains essential to ensure that predictions are interpreted within the broader educational context and that interventions are fair, transparent, and ethically grounded. Institutions adopting predictive learning analytics should therefore establish policies that promote data privacy, algorithmic fairness, and responsible use of student data. In addition, educators and academic administrators should be trained to understand both the capabilities and limitations of predictive models to avoid overreliance on algorithmic outputs (Pardo et al., 2014).

The integration of optimized ensemble learning models into institutional learning analytics platforms may contribute to improved student retention, enhanced academic outcomes, and more efficient allocation of educational resources. Consequently, this study provides practical evidence that can guide higher education institutions seeking to implement data-driven strategies for

supporting student success while maintaining ethical and responsible educational practices (Siemens et al., 2013).

## 6.4.Directions for Future Research

While the proposed framework demonstrates strong predictive performance and secure deployment characteristics, several areas remain open for further investigation.

***Dynamic Feature Re-weighting in Temporal Learning Models.***
Future work may extend the current static feature representation by incorporating sliding-window or time-aware feature engineering techniques. This would enable the model to capture temporal shifts in student engagement and performance across multiple academic checkpoints (e.g., 25%, 50%, and 75% of the semester).

***Multi-Institutional Transfer Learning and Generalization***
Further research could explore the adaptability of the trained LightGBM model across different institutional contexts. This includes evaluating whether model parameters or learned patterns can be transferred between universities with varying grading systems, curricula, and student demographics without full retraining.

***Integration of Natural Language Processing (NLP) Features***
An additional extension involves incorporating unstructured textual data into the predictive pipeline. Sources such as student discussion forum posts, academic feedback, and helpdesk interactions may be analyzed using NLP techniques to extract sentiment, stress indicators, or engagement patterns that complement structured academic data.

## Declarations

**Data Availability Statement:** The datasets analyzed during the current study are publicly available in the Open University Learning Analytics Dataset (OULAD) repository at https://analyse.kmi.open.ac.uk/open_dataset. The processed data subsets and scripts used during the current study are available from the corresponding author on reasonable request

**Declaration of Generative AI and AI-Assisted Technologies**

During the preparation of this manuscript, the author used **Google Gemini** to assist in language editing, manuscript formatting, and reviewing Python code syntax for data preprocessing and model evaluation workflows. All code experiments, experimental designs, computational executions, and manuscript conclusions were conducted and verified independently by the author. The author reviewed and edited the final output and maintain full accountability for the accuracy and scientific integrity of the work.

**Funding:**

No funding was received for conducting this study.

**Clinical Trial Number:**

Not applicable.

**Ethics Approval and Consent to Participate:**

Not applicable.

**Consent for Publication:**

Not applicable.

# References

Abdul Bujang, S. D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C., & Fujita, H. (2023). Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. *IEEE Access*, *11*, 1970–1989. https://doi.org/10.1109/ACCESS.2022.3225404

Ahn, J. M., Kim, J., & Kim, K. (2023). Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. *Toxins*, *15*(10), 608. https://doi.org/10.3390/TOXINS15100608/S1

Arnold, K., conference, M. P.-P. of the 2nd international, & 2012, undefined. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Dl.Acm.OrgKE Arnold, MD PistilliProceedings of the 2nd International Conference on Learning Analytics and, 2012•dl.Acm.Org*, 267–270. https://doi.org/10.1145/2330601.2330666

Asselman, A., Khaldi, M., & Aammou, S. (2023). Classification of student graduation status using XGBOOST algorithm. *Pdfs.Semanticscholar.Org*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

Bentley, C., Aicardi, C., … S. P.-A. at S., & 2023, undefined. (2023). A framework for responsible AI education. *Papers.Ssrn.Com*, 387–409. https://doi.org/10.1007/978-3-031-37776-1_17

Corchado, J. M., Kim, B.-G., Iglesias, C. A., Lee, I., Ren, F., Mehmood, R., Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics 2023, Vol. 12, Page 1789*, *12*(8), 1789. https://doi.org/10.3390/ELECTRONICS12081789

Eckart, L., Eckart, S., & Enke, M. (2021). A brief comparative study of the potentialities and limitations of machine-learning algorithms and statistical techniques. *E3S Web of Conferences*, *266*, 02001. https://doi.org/10.1051/E3SCONF/202126602001

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/EJED.12533;JOURNAL:JOURNAL:14653435;WGROUP:STRING:PUBLICATION

Hota, A. (2024). *A Comprehensive Approach to Behavioral Data Analysis and Machine Learning within Unified Systems*. https://osf.io/download/rjpxs/

Hu, & T Song. (2019). Research on XGboost academic forecasting and analysis modelling. *Iopscience.Iop.Org*. https://doi.org/10.1088/1742-6596/1324/1/012091/META

Juan Andrés, Ceballos, & Héctor G. (2023). Predicting students at risk of early dropping out from course using ensemble classification methods. *Elsevier*, *28*(9), 12169–12189. https://doi.org/10.1007/S10639-023-11682-Z

Kandikatla, L., & Radeljic, B. (2025). *AI and Human Oversight: A Risk-Based Framework for Alignment*. http://arxiv.org/abs/2510.09090

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Ensemble learning approach for delinquency risk assessment in adolescents. *Ieeexplore.Ieee.Org*, *18*(1). https://doi.org/10.1186/S41239-021-00300-Y

Kumar, S., Datta, S., Singh, V., & Datta, D. (2024). Applications, challenges, and future directions of human-in-the-loop learning. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/10530996/

Kuzilek, J., Hlosta, M., data, Z. Z.-S., & 2017, undefined. (2017). Open university learning analytics dataset. *Nature.ComJ Kuzilek, M Hlosta, Z ZdrahalScientific Data, 2017•nature.Com*. https://doi.org/10.1038/sdata.2017.171

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Springer*, *5*(1). https://doi.org/10.1186/S40537-018-0151-6

LO Yesufu. (2020). Predictive learning analytics in higher education. *Taylorfrancis.Com*, 151–173. https://doi.org/10.1201/9780429196614-9/PREDICTIVE-LEARNING-ANALYTICS-HIGHER-EDUCATION-LAWAL-YESUFU

Mangaroska, & M Giannakos. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/8454297/

Nderu, L., Oginga, R., Butichi, B., Rono, J., Njau, F., Mogire, F., Matindo, D., Muiruri, D., & Kiragga, A. (2024). DataLawCompanion: enhancing data protection law compliance in the digital age. *Datascience.Codata.Org*, *23*(1). https://doi.org/10.5334/DSJ-2024-036

Noorani, S., Karimi, M., & Z Gholijafari. (2024). Using Ensemble Machine Learning and Feature Engineering to Increase the Accuracy of Predicting Learners' Performance in an Online Educational Environment. *Ijvlms.Sums.Ac.Ir*. https://ijvlms.sums.ac.ir/article_50587.html

Orji, F., & J Vassileva. (2020). Using machine learning to explore the relation between student engagement and student performance. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/9373084/

Pardo, A., technology, G. S.-B. journal of educational, & 2014, undefined. (2014). Ethical and privacy principles for learning analytics. *Wiley Online Library*, *45*(3), 438–450. https://doi.org/10.1111/BJET.12152

Pek, R., Özyer, S., Elhage, T., Özyer, T., Access, R. A.-I., & 2022, undefined. (2022). The role of machine learning in identifying students at-risk and minimizing failure. *Ieeexplore.Ieee.Org*, *30*(2), 589–608. https://doi.org/10.1002/CAE.22475

Permana, A., & EB Santos. (2025). STUDENT RETENTION EARLY WARNING PROCEDURES: A CASE STUDY OF HIGHER EDUCATION INSTITUTION MANAGEMENT. *Jurnal.Stie-Aas.Ac.Id*. https://jurnal.stie-aas.ac.id/index.php/IJEBAR/article/view/18216

Pratama, F. R. A., & Oktora, S. I. (2022). Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification. *Journals.Sagepub.ComFR Adi Pratama, SI OktoraStatistical Journal of the IAOS, 2023•journals.Sagepub.Com*, *39*(1), 233–239. https://doi.org/10.3233/SJI-220080

Ronald, A., Gabriel, A., & AH Folrunsho. (2026). ARTIFICIAL INTELLIGENCE AS AN EMERGING CHALLENGE TO DATA PROTECTION AND PRIVACY LAW IN UGANDA. *Articles.Upu.Ac.Ug*. https://articles.upu.ac.ug/index.php/upu-journal-for-sustainable-deve/article/view/20

Schunk, D. (2013). *Barry J. Zimmerman's theory of self-regulated learning*. https://www.emerald.com/books/edited-volume/20378/chapter/105746352

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Springer*, *28*(7), 8299–8333. https://doi.org/10.1007/S10639-022-11536-0

Siemens, R., Li, G., Booth, C., MacKillop, W., CUAJ, W. K.-, & 2013, undefined. (2013). Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment. *Erudit.Org*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Slade, S., scientist, P. P.-A. behavioral, & 2013, undefined. (2013). Learning analytics: Ethical issues and dilemmas. *Journals.Sagepub.Com*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Sokol, K., Intelligenz, P. F.-K.-K., & 2020, undefined. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *Springer*, *34*(2), 235–250. https://doi.org/10.1007/S13218-020-00637-Y

Švábenský, V., Flanagan, B., Daniel, E., Zapata, L., & Shimada, A. (2026). Open Datasets in Learning Analytics: Trends, Challenges, and Best PRACTICE. *ACM Transactions on Knowledge Discovery from Data*, *20*(4). https://doi.org/10.1145/3798096

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, L. (2023). Current stance on predictive analytics in higher education: Opportunities, challenges and future directions. *Taylor & Francis*, *31*(6), 3503–3528. https://doi.org/10.1080/10494820.2021.1933542

Venkatasubbu, S., and, G. K.-J. of K. L., & 2022, undefined. (2022). Ethical considerations in AI addressing bias and fairness in machine learning models. *Jklst.Org*, *55*(3). https://doi.org/10.1145/3494672

Villano, R., Harrison, S., Lynch, G., & Chen, G. (2018). Linking early alert systems and student retention: a survival analysis approach. *Higher Education 2018 76:5*, *76*(5), 903–920. https://doi.org/10.1007/S10734-018-0249-Y

Wong, J., Baars, M., De Koning, B. B., Van Der Zee, T., Davis, D., Khalil, M., Houben, G. J., & Paas, F. (2019). Educational theories and learning analytics: From data to knowledge: The whole is greater than the sum of its parts. *Springer*, 3–25. https://doi.org/10.1007/978-3-319-64792-0_1

Xu, Q., & Yin, J. (2021). Application of random forest algorithm in physical education. *Wiley Online Library*, *2021*. https://doi.org/10.1155/2021/1996904

Zhou, D., Jian, T., Chen, S., Jin, N., & Wang, J. (2026). Comparative Performance of Machine Learning Models for Predicting At-Risk Students Using the OULAD Dataset. *Dl.Acm.Org*, *1*, 247–252. https://doi.org/10.1145/3802133.3802174

Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., & Chen, J. (2022). Bias, fairness and accountability with artificial intelligence and machine learning algorithms. *Wiley Online Library*, *90*(3), 468–480. https://doi.org/10.1111/INSR.12492

Zhu, L., Lu, Q., Ding, M., Lee, S., Ethics, C. W.-A. and, & 2026, undefined. (2026). Reasonable Oversight of the Practical, Legal, and Ethical Boundaries Impacting Human Factors Trust in Automated Decision-Making. *Armgpublishing.Com*, *6*(3), 286. https://doi.org/10.1007/S43681-026-01147-7