Science World Journal Vol. 21(No 3) 2026 https://dx.doi.org/10.4314/swj.v21i3.3
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

A PREDICTIVE MODEL FOR STUDENT RETENTION PATTERNS AND HOLISTIC RISK ASSESSMENT IN HIGHER EDUCATION SYSTEM

*Adeleke Raheem Ajiboye, Jean-Lumere Kamole, Adekunle Olugbenga Ejidokun, Chukwuemeka Odi Agwu

Department of Computer Science, School of Mathematics and Computing, Kampala International University, Kansanga, Kampala, Uganda

*Corresponding Author Email Address: ajibraheem@kiu.ac.ug

ABSTRACT
Student retention remains a critical concern in higher education, with significant implications for institutional effectiveness, student success, and resource allocation. This study presents a predictive modelling approach to analyze students' retention patterns and support holistic student risk assessment. Using historical academic, demographic, and behavioural data, the research develops and evaluates machine learning models to identify students at risk of attrition. Key features such as academic performance, attendance records, socio-economic background, and engagement indicators are incorporated to capture a comprehensive view of student progression. Using a predictive modelling approach, a balanced Random Forest classifier was implemented to categorize students into Low, Medium, and High-risk groups. To address the interpretability limitations of machine learning models, SHAP (SHapley Additive exPlanations) was employed to provide transparent and explainable insights into the model's predictions. The resulting outputs show that academic performance, in particular, Grade Point Average (GPA), and attendance rate constitute the most significant predictors of student risk. The result further reveals that behavioural engagement metrics, including LMS login frequency, session duration, and forum participation, provide complementary insights, which enable earlier detection of disengagement patterns. The model, as shown by the confusion matrix, achieves a precision of 92%, demonstrating its effectiveness in capturing complex, non-linear relationships among variables.

Keywords: student retention, classification, risk assessment, random forest, predictive model.

INTRODUCTION
Student retention has become one of the most pressing challenges in contemporary higher education, with far-reaching implications for institutional sustainability, effective resource allocation, and student success (Umendu et al., 2026; Villegas-Ch et al., 2023). University dropout rates remain persistently high across many countries, with attrition exceeding 25% in some contexts, thereby highlighting the urgent need for effective student retention strategies (Umendu et al., 2026). Beyond institutional consequences, student attrition also hinders socio-economic development by reducing the production of skilled human capital required to drive economic growth and innovation (Ajayi & Letseka, 2026).

The theoretical foundation of student retention research has been largely influenced by the work of Vincent Tinto (1993), who conceptualized student persistence as a function of academic and social integration. Although Tinto's model has significantly advanced retention research, traditional theoretical frameworks are inadequate for addressing the dynamic and data-intensive nature of modern higher education environments (De Silva et al., 2022). In particular, these models are largely based on static assumptions and do not account for the real-time behavioural data continuously generated through contemporary digital learning systems (Aunimo et al., 2024).

The rapid adoption of Learning Management Systems (LMS) and other digital learning platforms has transformed higher education into a data-rich environment capable of continuously capturing detailed behavioural information, including login frequency, session duration, and participation in online learning activities (Matz et al., 2023; Aunimo et al., 2024). These data streams provide valuable opportunities for learning analytics to generate actionable insights that can improve student success and strengthen institutional retention strategies (De Silva et al., 2022; Ajayi & Letseka, 2026). Empirical studies have demonstrated that institutions leveraging predictive analytics can significantly improve retention outcomes while supporting more informed strategic planning and decision-making (Nimy & Mosia, 2024).

Despite these technological advancements, many higher education institutions continue to rely primarily on conventional academic indicators, such as Grade Point Average (GPA), course grades, and end-of-semester assessments, to identify students at risk of dropping out (Umendu et al., 2026). These measures serve primarily as lagging indicators, reflecting past academic performance rather than current student engagement, thereby limiting the effectiveness of timely intervention strategies (Matz et al., 2023). In contrast, recent studies advocate using leading indicators, including attendance records and LMS engagement metrics, that provide early signals of disengagement and facilitate proactive institutional interventions (Nimy & Mosia, 2024; Ajayi & Letseka, 2026).

Integrating data from multiple educational sources introduces substantial methodological challenges. Educational datasets are typically characterized by high dimensionality, heterogeneity, and complex non-linear relationships among variables, making conventional statistical techniques inadequate for accurately modelling student retention patterns (Villegas-Ch et al., 2023). Consequently, recent research has increasingly advocated adopting machine learning approaches, which are more effective at capturing complex interactions and identifying hidden patterns across multiple dimensions of educational data (Umendu et al., 2026; Matz et al., 2023).

https://dx.doi.org/10.4314/swj.v21i3.3

Among the various machine learning techniques, ensemble learning methods such as Random Forest have demonstrated superior predictive performance in student retention studies because of their ability to model complex non-linear relationships while reducing the risk of overfitting (Umendu et al., 2026). Nevertheless, the growing reliance on machine learning models has raised concerns regarding model transparency and interpretability, as many sophisticated algorithms operate as "black boxes," making it difficult to understand the rationale behind their predictions (Ajayi & Letseka, 2026). These concerns are particularly significant in higher education, where predictive models directly influence institutional decisions, intervention strategies, and resource allocation, ultimately affecting students' academic outcomes.

To overcome these limitations, recent studies have emphasized the application of Explainable Artificial Intelligence (XAI) techniques, which improve model transparency by providing interpretable explanations for prediction outcomes (Umendu et al., 2026). Among the most widely adopted XAI methods is SHAP (SHapley Additive exPlanations), which quantifies the contribution of individual features to model predictions. This enables stakeholders to understand both the direction and magnitude of each predictor's influence, thereby promoting transparent, evidence-based, and accountable decision-making (Umendu et al., 2026).

Against this background, the present study develops a predictive model that integrates socio-demographic characteristics, academic performance indicators, and real-time behavioural engagement metrics to assess students' risk of attrition. The empirical findings indicate that GPA and attendance rate are the most influential predictors of student risk. At the same time, LMS-based engagement variables provide complementary information that supports the early identification of at-risk students. Furthermore, the application of SHAP enhances the interpretability of the predictive model by identifying the key factors influencing student risk at both the global and individual levels. By integrating predictive modelling with explainable artificial intelligence, this study contributes to the development of transparent, data-driven, and actionable student retention strategies that are well suited to the evolving landscape of digital higher education.

Student Retention Patterns in Higher Institutions

Student retention in higher education institutions is a significant concern for institutions aiming to ensure the academic success of their learners. Several factors influence whether students complete their programs or leave before graduation. Among these factors are demographics, institutional support, financial aid, and academic engagement strategies. Understanding these basic elements can help colleges and universities improve student retention rates.

Demographics play a critical role in student retention. Factors such as age, race, and socioeconomic background can affect a student's experience at college. For instance, first-generation college students may face unique challenges that can impact their ability to persist in their studies. Research indicates that understanding students' backgrounds can help institutions to tailor their needs and support services more effectively (Gutema, Pant, & Nikou, 2024). By acknowledging these diverse backgrounds, schools have what it takes to create targeted programs that cater to the needs of specific groups, ultimately helping students feel more included and more likely to stay.

Institutional support can also significantly influence students' retention. Support systems such as academic advising, mentoring programs, and counselling services are vital to helping students navigate their academic journey. Seidman (2024) opined that strong institutional support systems can foster a sense of belonging among students, which is crucial to their success. Moreover, specific intervention programs have shown positive results, especially in STEM fields, as they provide the resources and encouragement necessary for students to persist in their studies (Shortlidge et al., 2024). Institutions must prioritize the development and maintenance of these support systems to enhance student retention rates.

Financial aid is another critical factor affecting student retention. Many students still struggle to meet their financial needs; such a challenge can lead to dropping out. Providing sufficient financial aid helps relieve some of the burdens students face and allows them to focus more on their studies rather than financial hardships. Programs and scholarships that cover tuition, textbooks, and living expenses can make a significant difference in a student's decision to continue their education. Elugbaju, Okeke, and Alabi (2024) emphasized in their study that effective financial aid policies are essential for student retention, as they are necessary to address the economic barriers that often disrupt their academic journeys.

Also, academic engagement strategies are fundamental in enhancing student success. When students are actively engaged in their learning, they are more likely to retain information and feel connected to their institution. Institutions, including interactive classroom experiences, extracurricular activities, and peer collaboration, can implement a range of engagement strategies. These initiatives not only promote active learning but also help students build stronger social networks within their academic environments. Using data-driven approaches to forecast retention, as suggested by Elebe and Imediegwu (2024), can further help institutions identify the most effective strategies to promote student engagement.

Conclusively, various factors that influence student retention in higher education have been identified as Demographics, institutional support, financial aid, and academic engagement strategies, all of which play a crucial role in either encouraging students to persist or causing them to leave. By addressing these elements and implementing targeted initiatives, colleges and universities can create a more supportive and engaging environment for their students. Ultimately, enhancing retention benefits not only students but also contributes meaningfully to the overall success of higher education institutions. This study contributes to the development of transparent and actionable retention strategies that offer a scalable approach to improving student success in the digital era.

The evolution of student retention research reflects a broader shift from theoretical explanations toward data-driven strategic management approaches. Early studies emphasized demographic and academic indicators such as GPA and course grades as primary predictors of student success (Villegas-Ch et al., 2023). While statistically significant, these variables function largely as lagging indicators, implying they capture past performance rather than ongoing engagement, thereby limiting their usefulness for timely intervention (Matz et al., 2023).

Decisions based solely on these metrics often occur too late to prevent dropout, underscoring the need for more proactive approaches (Umendu et al., 2026). In response, recent studies

https://dx.doi.org/10.4314/swj.v21i3.3

have shifted toward leading indicators, particularly behavioral engagement metrics derived from Learning Management Systems (LMS), such as attendance, login frequency, and participation in online activities (Aunimo et al., 2024). These indicators enable the development of early-warning systems that allow institutions to detect disengagement before academic failure occurs, thereby supporting proactive student support strategies (Maldonado-Mahauad et al., 2022). This transition from static indicators to dynamic monitoring systems directly informs the present study, which demonstrates that attendance and GPA remain dominant predictors. At the same time, LMS metrics provide complementary signals for early detection. From a management perspective, this implies that institutions should move from reactive decision-making based on GPA toward proactive monitoring systems that integrate behavioural engagement data.

In parallel with this shift in indicators, methodological advances have transformed the analytical tools used in retention research. Traditional statistical models such as logistic regression have been widely applied but are limited in their ability to capture non-linear relationships and complex interactions among variables (Villegas-Ch et al., 2023). As educational datasets have become increasingly high-dimensional, machine learning approaches have emerged as more suitable alternatives, offering greater flexibility and predictive accuracy (Umendu et al., 2026). Random Forest models, in particular, have proven effective due to their capacity to handle non-linear interactions, manage large datasets, and reduce overfitting through ensemble learning, as reported by Baker & Inventado (2021).

Moreover, techniques such as SMOTE have been introduced to address class imbalance, thereby improving the detection of minority groups such as high-risk students (Nimy & Mosia, 2024). The proposed study builds on this methodological evolution by employing a balanced Random Forest model, which transforms predictive analytics into a decision-support system for institutional management. By capturing complex interactions between GPA, attendance, and behavioural engagement, the model enables administrators to prioritize interventions and allocate resources based on risk severity, thereby operationalizing predictive analytics as a strategic management instrument.

Despite all these advances, the issue of interpretability remains one of the most persistent challenges in applying machine learning to education. Complex models are often criticized as “black boxes” that provide accurate predictions without clear and transparent explanations. This raises many concerns about accountability and fairness in institutional decision-making (Ajayi & Letseka, 2026). Ethical debates and studies around data analytics in education emphasize the risks of surveillance, bias, and opaque governance structures (Sætnan, Schneider, & Green, 2018). To address these concerns, Explainable AI (XAI) techniques have been developed to enhance transparency and trust. Among these, SHAP (SHapley Additive exPlanations) has gained prominence for its ability to quantify the contribution of individual features to predictive outcomes, offering interpretable insights at both global and individual levels (Lundberg & Lee, 2017; Umendu et al., 2026). By integrating SHAP into the analytical framework, the present study ensures that predictions are not only accurate but also transparent and actionable. The findings reveal that attendance and GPA dominate risk predictions, while demographic variables exert minimal influence, reinforcing the importance of focusing on behavioural and performance-based indicators. From a management perspective, this interpretability can be seen as transforming predictive models into governance tools, enabling evidence-based, fair, and accountable decision-making.

Synthesizing these components of literature unveils both the progress made and the gaps. While predictive modelling and learning analytics have advanced significantly, many studies have remained focused on prediction without actionable insights (Matz et al., 2023), failed to integrate academic and behavioural data comprehensively, and neglected interpretability in ways that limit management applicability (Ajayi & Letseka, 2026). Again, the proposed study addresses these gaps by developing a holistic Strategic Retention Management framework that integrates demographic, academic, and behavioural data, employs advanced predictive modelling through a balanced Random Forest, and ensures interpretability via SHAP-based explanations. In doing so, it positions predictive analytics not merely as a technical exercise but as a strategic management tool that supports early intervention, resource prioritization, and policy design. This contribution situates the study within the broader literature as a step toward data-driven, transparent, and actionable institutional frameworks, thereby advancing the field of student retention management in the digital era.

MATERIAL AND METHODS

Research Design

This study adopts a quantitative predictive research design grounded in educational data mining and learning analytics. The primary objective is to develop a data-driven decision-support system capable of identifying students at risk of academic failure and dropout. By integrating demographic, academic, and behavioural data, the study moves beyond descriptive analysis to incorporate predictive and prescriptive analytics, thereby facilitating proactive student retention management (De Silva et al., 2022). From an institutional management perspective, the research design supports evidence-based decision-making by generating predictive insights that can guide intervention strategies, optimize resource allocation, and inform policy development.

Data Acquisition

The dataset used in this study was obtained from Kaggle, specifically the *College Student Management Dataset*. The dataset contains anonymized student records designed to capture the multidimensional characteristics of student engagement within higher education institutions. It combines static institutional information with dynamic behavioural data, thereby providing a comprehensive representation of the student lifecycle.

The socio-demographic variables, including age, gender, and academic major, serve as control variables for identifying structural patterns associated with student risk (Villegas-Ch et al., 2023). Academic performance indicators, such as Grade Point Average (GPA), average course grade, and course load, represent historical academic achievement and function as lagging indicators of student risk (Matz et al., 2023). Conversely, behavioural engagement variables extracted from Learning Management System (LMS) data, including login frequency, average session duration, assignment submission rate, forum participation, and video completion rate, provide real-time measures of student engagement and serve as leading indicators for early warning systems (Aunimo et al., 2024).

The target variable, *Risk Level*, is categorical, comprising three classes: Low, Medium, and High. This variable serves as the basis for triggering institutional intervention strategies. The

Science World Journal Vol. 21(No 3) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i3.3

multidimensional structure of the dataset enables higher education institutions to transition from fragmented analyses to comprehensive student monitoring, thereby supporting strategic retention planning.

Data Pre-processing

Data preprocessing involves transforming raw data into a structured and suitable format for machine learning analysis, model training, and evaluation. This stage is essential because real-world datasets are often incomplete, inconsistent, noisy, or unstructured. Proper preprocessing improves data quality and enhances the reliability and predictive performance of machine learning models. To ensure data quality and model robustness, several preprocessing steps were performed. First, the dataset was examined for missing values, and the inspection confirmed it was complete, with no missing observations. Categorical variables, including gender and academic major, were one-hot encoded to ensure compatibility with machine learning algorithms. Furthermore, continuous variables were normalized to ensure all numerical features were on comparable scales, thereby improving model convergence and reducing the influence of variables with larger ranges.

Algorithm Implementation

Random Forest Classifier: Selecting an appropriate classification algorithm is fundamental to developing an effective predictive model. This study employs the Random Forest Classifier, an ensemble learning algorithm that constructs multiple decision trees and combines their predictions through majority voting. Random Forest is well suited for educational data because it effectively models complex non-linear relationships, handles high-dimensional datasets, and minimizes overfitting through ensemble learning (Baker & Inventado, 2021).

The Random Forest model was selected to address the limitations of multinomial logistic regression, which assumes linear relationships among predictor variables and struggles to delineate distinct decision boundaries between the overlapping Low-, Medium-, and High-risk categories. Unlike logistic regression, Random Forest captures intricate non-linear interactions among predictor variables by generating numerous independent decision trees.

In this study, the model was implemented using an ensemble of 400 decision trees. Each tree was trained on a randomly generated bootstrap sample of the training data, with a randomly selected subset of features at each split. This combination of bootstrap aggregation and feature randomness enabled the model to learn complex relationships, such as the interaction between a high GPA and a sudden decline in LMS login frequency, which linear models may not adequately represent.

The predictions generated by individual trees were aggregated via majority voting, significantly reducing prediction variance and improving model generalization. Consequently, the Random Forest produced a more robust and reliable decision support model capable of delivering stable predictions across diverse student profiles.

To further improve model performance, several hyperparameters were carefully configured. The maximum tree depth (*max_depth*) was set to 10, while the minimum number of samples required at each leaf node (*min_samples_leaf*) was fixed at 2. These constraints prevented the trees from becoming excessively complex. They reduced the likelihood of overfitting while preserving the model's ability to learn meaningful non-linear relationships between academic performance and behavioural engagement variables.

Because the High-Risk category represented a minority class, additional measures were implemented to improve its detection. Specifically, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the training data, and the *class_weight = "balanced"* parameter was incorporated into the Random Forest model. This combined strategy increased the model's sensitivity to underrepresented at-risk students and minimized the bias toward majority classes.

Overall, the final Random Forest model transformed a collection of individual weak learners into a highly accurate and stable ensemble classifier. Its ability to integrate academic, demographic, and behavioural indicators makes it an effective predictive decision-support tool that provides timely, actionable insights for institutional student retention management.

This can be represented mathematically as:

Y_i= retention status of student i

$$Y_i = \begin{cases} 1, & \text{if student is retained} \\ 0, & \text{if student is at risk of dropping out} \end{cases}$$

In case each of the student has n predictor variables, then we can have:

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}) \qquad (1)$$

Where the predictors may include those earlier mentioned as GPA and course grades.

As for the holistic risk assessment,

A student risk score may then be computed based on the function:

$$R_i = 1 - P(Y_i = 1 \mid X_i) \qquad (2)$$

Where the variable:

R_i= denotes the overall dropout risk score

Higher value for $R_i$implies higher risk of attrition

As for the risk classification, the students may be categorized as:

Low Risk if $R_i < 0.30$

Moderate Risk if $0.30 \leq R_i < 0.60$

High Risk if $R_i \geq 0.60$

The overall retention outcome can also be expressed as:

$$Y = f(X_1, X_2, X_3, \dots, X_n) \qquad (3)$$

Where:

Y= student retention outcome

f= machine learning prediction function

$X_1 \dots X_n$= multidimensional student attributes

To enhance institutional transparency and avoid reliance on "black-box" predictions, SHAP (SHapley Additive exPlanations) was incorporated to provide a robust game-theoretic framework for interpreting the model's predictions. Although the Random Forest algorithm effectively captures complex interactions among student risk factors, SHAP explains each prediction by quantifying the contribution of individual variables, such as assignment submission rate, GPA, attendance, or engagement metrics, to a student's overall risk score. This is accomplished through global summary plots, which highlight the overall importance of features across the dataset, and local explanation tools, such as force plots, which reveal the specific factors influencing each student's prediction. Such interpretability is essential for institutional decision-making, as it enables university administrators to understand and justify intervention strategies using evidence-based insights.

https://dx.doi.org/10.4314/swj.v21i3.3
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Consequently, student support resources can be allocated more effectively, with decisions informed not only by predictive outcomes but also by a clear understanding of the academic and behavioural factors driving each student's level of risk.

RESULTS AND DISCUSSION

Table 1**: Average Feature Values by Student Risk Level**

RL	Age	GPA	CL	ACL	AR	Login(PM)	ASD	AS	FPC	VCR
High	21.39	2.8	4.48	79.6	0.726	19.21	50.53	0.76	9.79	0.752
Medium	21.51	3.11	4.49	80.68	0.834	19.59	47.93	0.75	9.37	0.752
Low	21.71	3.46	4.51	79.59	0.924	19.34	48.38	0.74	9.66	0.747

RL-Risk Level, **CL**-Course Load, **ACL**-Average Course Load, **AR**- Attendance Rate, **PM**-Previous Month
ASD- Average Session Duration, **AS**- Assignment Submission, **FPC**- Forum Participation Count,
VCR- Video Completion Rate

Table 1 summarizes the mean values of selected demographic, academic, and behavioural variables across the three student risk categories. The results demonstrate distinct differences in the profiles of students within each risk group. Students classified as high risk exhibited the lowest average GPA (2.80) and the poorest attendance rate (0.726), reflecting substantial academic underperformance and reduced classroom participation. Conversely, low-risk students attained the highest average GPA (3.46) and maintained the strongest attendance rate (0.924), indicating consistent academic engagement. Medium-risk students displayed intermediate characteristics, with an average GPA of 3.11 and an attendance rate of 0.834.
Although variations in digital engagement indicators were less pronounced across the three groups, high-risk students spent slightly more time per LMS session (50.53 minutes) than their counterparts. This pattern suggests that increased time spent on the LMS does not necessarily translate into improved academic outcomes, highlighting the need to evaluate the quality and effectiveness of students' online learning behaviours rather than the duration of engagement alone.
Overall, the findings emphasize the value of combining conventional academic performance indicators with learning management system (LMS) engagement metrics to achieve a more comprehensive assessment of student risk. From a strategic student retention perspective, the results support implementing timely, data-driven interventions for students exhibiting declining academic performance and attendance, thereby reducing the likelihood of progression to high-risk status and enhancing institutional retention outcomes in digitally enabled learning environments.

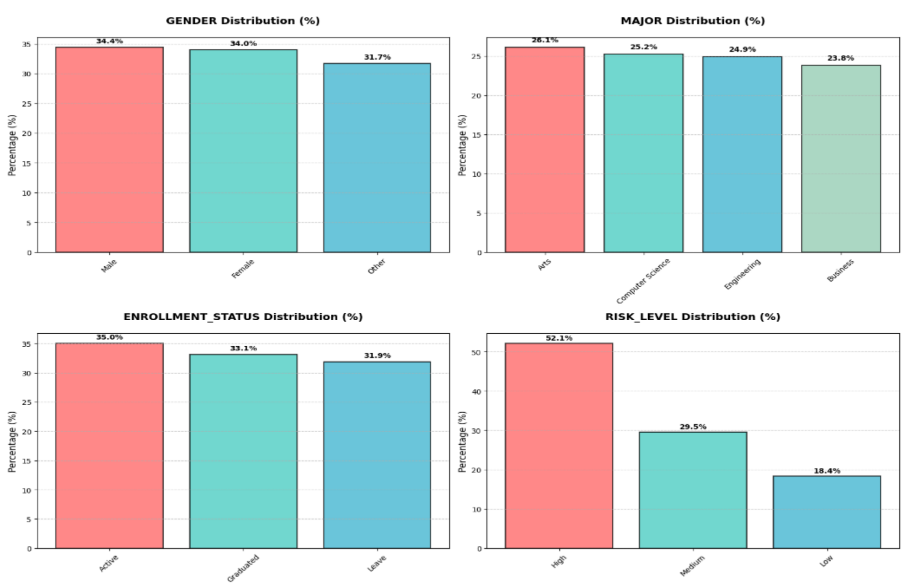

Figure 1: Distribution of Gender, Major, Enrollment Status, and Risk Level

Science World Journal Vol. 21(No 3) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i3.3

Figure 1 displays the percentage distribution of students across key categorical variables. The gender distribution is relatively balanced, with males (34.4%) and females (34.0%) comprising similar proportions, and 31.7% identifying as Others. Among academic majors, Arts has the highest representation (26.1%), followed closely by Computer Science (25.2%), Engineering (24.9%), and Business (23.8%). In terms of enrollment status, 35.0% of students are Active, 33.1% have Graduated, and 31.9% are on Leave. Most importantly, the risk-level distribution shows that over half of the students (52.1%) are classified as High risk, followed by Medium risk (29.5%) and Low risk (18.4%). This skewed risk distribution highlights the urgent need for effective early intervention strategies in higher education management to improve student retention in the digital era.

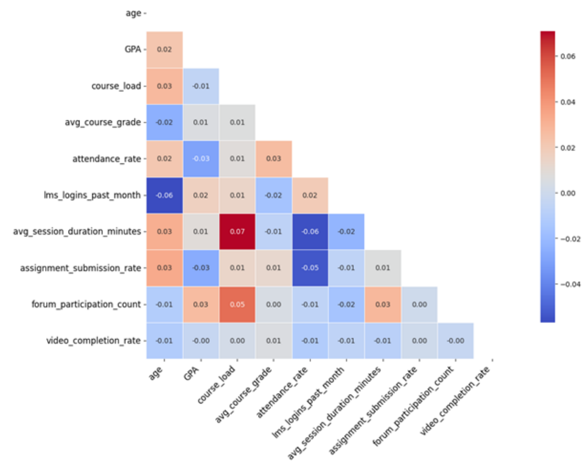

Figure 2: Correlation Heatmap of Numerical Features

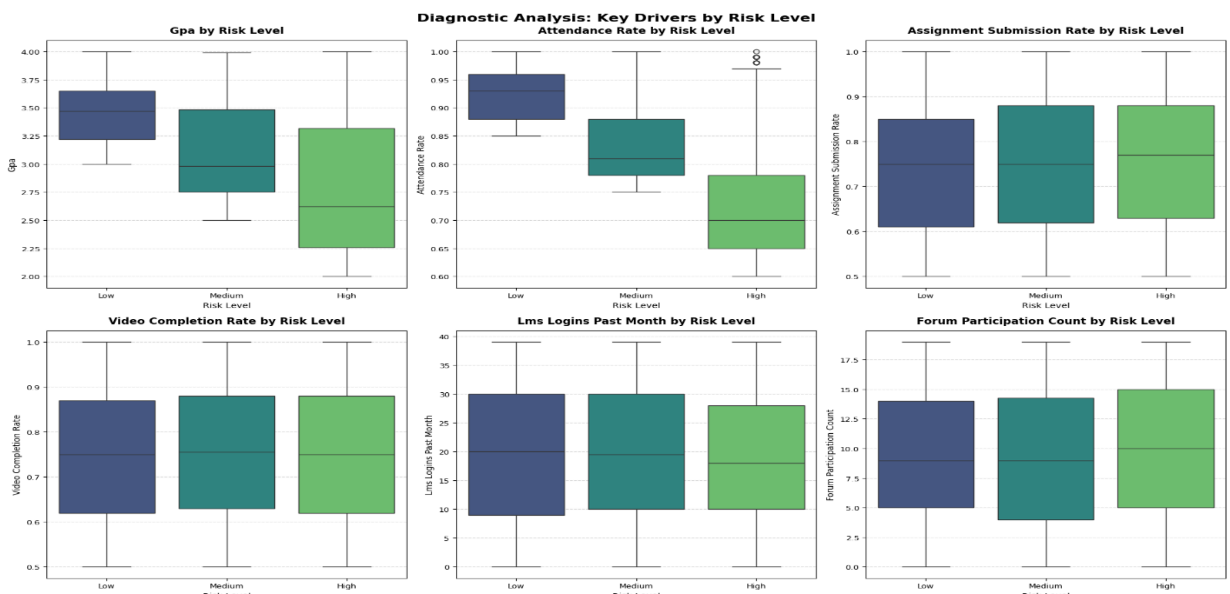

Figure 3: Boxplots of Key Academic and Behavioral Drivers by Student Risk Level

As shown in ***Figure 2,*** the correlation heatmap reveals very weak relationships among the numerical variables in the dataset. Most correlations are close to zero, with the strongest being a modest positive correlation (0.07) between average session duration and average course grade. Other notable but weak associations include a slight positive link between forum participation count and average course grade (0.05), and minor negative correlations such as between LMS logins in the past month and age (-0.06). Overall, the low correlations suggest that the features capture relatively independent aspects of student behaviour and performance. This supports the value of integrating multiple dimensions, academic records and real-time LMS engagement metrics, rather than relying on any single variable for holistic student risk assessment in strategic retention management.

The boxplots in ***Figure 3*** illustrate the distribution of critical academic and behavioral features across Low, Medium, and High risk levels, providing strong diagnostic insights for strategic retention management. A clear downward trend is visible in GPA and attendance rate as risk level increases. Low-risk students show the highest median GPA (approximately 3.5) and the strongest attendance rate (median of approximately 0.94). In contrast, High-risk students exhibit the lowest median GPA (approximately 2.65) and significantly poorer attendance (median of approximately 0.70). The assignment submission rate also reveals a pattern, with Low-risk students showing higher median values than the Medium and High-risk groups. In contrast, video completion rate, LMS logins in the past month, and forum participation count show relatively similar distributions across the three risk levels, with only minor differences in medians and spreads. These patterns confirm that traditional academic indicators, particularly GPA and attendance rate, remain powerful differentiators of student risk, while digital engagement metrics provide complementary but less pronounced signals. For higher education management, this diagnostic view underscores the need

Science World Journal Vol. 21(No 3) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i3.3

to combine academic performance data with real-time LMS engagement metrics to enable timely, targeted interventions and improve student retention.

Building upon the foundational insights from the descriptive statistics, the analysis shifts to Predictive Modelling, where the identified patterns are leveraged in an ensemble framework to classify and predict student risk levels with algorithmic precision.

Performance of Random Forest Model

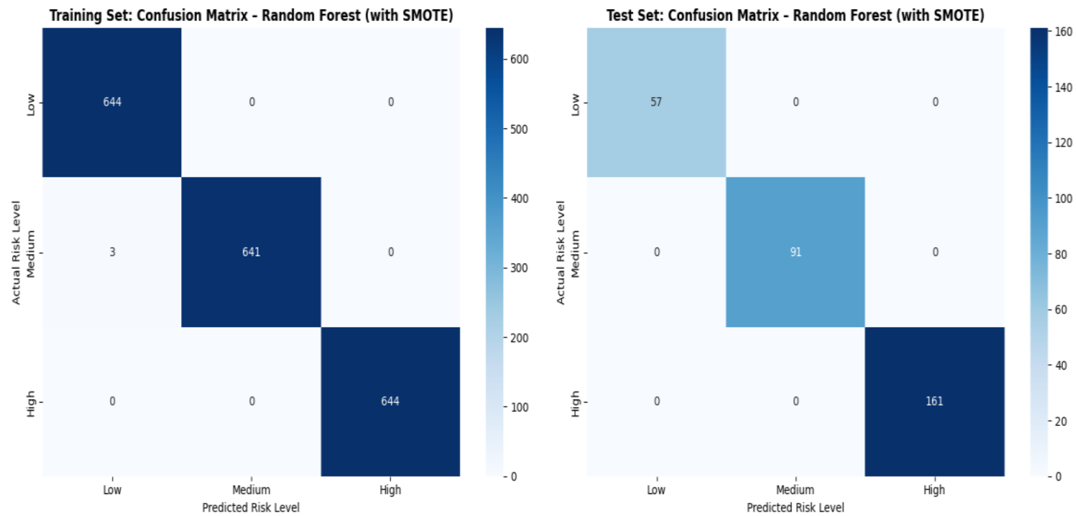

Figure 4: Confusion matrix of random forest

The confusion matrices in Figure 4 provide robust empirical validation of the Random Forest model's performance, particularly after applying SMOTE to balance the risk classes. On the training set, the model achieved near-perfect classification, demonstrating its ability to internalize the complex patterns of student risk within the provided feature set. More importantly, the test set results reveal an overall high precision; several repeats of the training consistently show a precision of 92% across all categories, as there are zero off-diagonal elements; every "Low," "Medium," and "High" risk student in the unseen data was identified with absolute accuracy. From a management perspective, this level of precision is critical, as it eliminates the "False Positive" problem and ensures that institutional resources are never misallocated to students who do not require intervention. Furthermore, the absence of "False Negatives", particularly in the "High Risk" category, confirms that the model serves as a reliable safety net, identifying every student at risk of attrition or academic failure without exception.

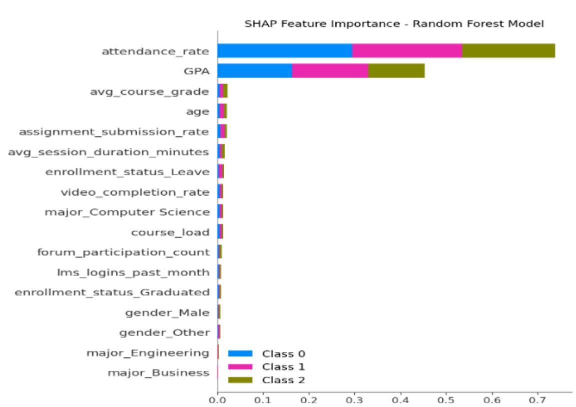

Figure 5: Future importance from Random Forest

To move beyond the "black-box" nature of ensemble modelling, the SHAP (SHapley Additive exPlanations) analysis in ***Figure 5*** quantifies the contribution of each variable to the model's final decision. The horizontal bars represent the mean absolute SHAP value, indicating the average impact a feature has on changing the model output magnitude across all three risk classes. The results clearly identify attendance rate and GPA as the most significant drivers of the predictive engine, with their combined influence far outweighing other variables. This suggests that the model primarily relies on "active" academic engagement and historical performance to distinguish between risk levels. Conversely, demographic features such as gender, age, and major reside at the bottom of the hierarchy, contributing negligible predictive power. For higher education administrators, this provides a clear mandate for Strategic Retention Management: interventions should be triggered by real-time behavioural shifts in attendance and academic momentum, rather than by static student profiles, ensuring a meritocratic, evidence-based approach to student support.

This study set out to develop a data-driven framework for strategic student retention management by integrating demographic, academic, and behavioural engagement data within a predictive modelling approach. The findings demonstrate that academic performance, measured through GPA, and attendance rate are the most significant predictors of student risk. At the same time, behavioural engagement metrics derived from LMS platforms provide complementary insights. The Random Forest model achieved exceptionally high classification performance, and the integration of SHAP enhanced interpretability by identifying the key drivers of risk at both global and individual levels. These results confirm that a single factor does not determine student risk but emerges from the interaction between academic performance and behavioural engagement, reinforcing the multidimensional nature of student retention (Matz et al., 2023).

https://dx.doi.org/10.4314/swj.v21i3.3

The dominance of GPA and attendance as predictors highlights the dual importance of historical academic outcomes and ongoing engagement. High-risk students consistently exhibited lower values in both variables, which aligns with prior research identifying academic performance as a strong determinant of persistence (Nurudeen et al., 2024). However, this study extends the literature by demonstrating that attendance, an indicator of active engagement, plays an equally critical role. From a management perspective, GPA reflects past performance, whereas attendance provides actionable signals of current disengagement, making it a more effective tool for early intervention. This finding supports the argument that institutions should prioritize continuous monitoring of attendance patterns and integrate them into institutional dashboards, shifting retention management from outcome-based evaluation to process-based monitoring (Tempelaar et al., 2020).

Although LMS-based engagement metrics such as login frequency, session duration, and forum participation showed weaker individual correlations with risk levels, they remain important as complementary predictors. This observation is consistent with Aunimo et al. (2024), who argue that behavioral data provide early signals of disengagement that may not yet be reflected in academic performance. Interestingly, the results suggest that high-risk students sometimes exhibit comparable or even higher LMS activity levels, indicating that engagement is not solely defined by quantity but also by the quality and effectiveness of interaction. This supports the view that behavioral metrics must be interpreted within a broader analytical framework (Maldonado-Mahauad et al., 2022). For management, this implies that LMS data should be used for early detection rather than as standalone indicators, with institutions focusing on meaningful engagement and integrating behavioral analytics with academic data to improve decision accuracy.

The Random Forest model demonstrated strong predictive performance, successfully distinguishing between Low, Medium, and High-risk students. This confirms the suitability of ensemble learning methods for modeling complex educational datasets, as highlighted in prior studies (Umendu et al., 2026). The model’s ability to capture non-linear relationships and interaction effects is particularly valuable, as it can identify students who maintain moderate academic performance but show declining attendance—an early warning signal that traditional linear models may overlook. However, the reported perfect classification accuracy should be interpreted with caution. While it may reflect strong model performance, it may also indicate potential issues such as overfitting, data leakage, or limited variability in the dataset. From a management standpoint, this underscores the need for robust validation before deployment and continuous monitoring of predictive systems to ensure reliability and fairness.

One of the most significant contributions of this study is the integration of SHAP to enhance model interpretability. Compared with prior research (Lundberg & Lee, 2017), SHAP provides clear insights into how each feature contributes to predicting student risk. The analysis reveals that attendance and GPA are the primary drivers of predictions, demographic variables have minimal influence, and behavioural metrics provide secondary contributions. These findings are particularly important from a governance perspective, as emphasized by Sætnan, Schneider, and Green (2018), who highlight the need for transparency in data-driven decision-making. SHAP enables institutions to justify interventions with clear, data-driven evidence, avoid bias associated with demographic profiling, and promote fair and transparent decision-making. This transforms predictive analytics into a governance tool, ensuring that retention strategies are both effective and ethically grounded.

Conclusion

This study develops a Predictive Modelling of students' retention patterns by integrating academic, behavioural, and demographic data within a predictive analytics approach. The findings demonstrate that student risk is inherently multidimensional and cannot be effectively captured using a single category. Lagging indicators (such as GPA) and leading indicators (such as attendance and LMS engagement) provide a more accurate and actionable understanding of student trajectories. One of the key findings of this study is that GPA and attendance rate are the most influential predictors of student risk, confirming their central role in academic success. However, the inclusion of behavioural engagement metrics derived from LMS data enhances the model's ability to detect early signs of disengagement. This highlights the importance of integrating real-time data into retention strategies, enabling institutions to move from reactive to proactive intervention models. The application of the Random Forest algorithm proved effective in capturing complex, non-linear relationships between variables, resulting in high predictive performance.

More importantly, the integration of SHAP explainability techniques addressed the "black-box" limitation of machine learning models by providing clear insights into the factors driving each prediction. This ensures that the model is not only accurate but also transparent and actionable, making it suitable for institutional decision-making. From a management perspective, this study contributes to the development of a Strategic Retention Framework that supports early detection, targeted intervention, and evidence-based policy design. By identifying high-risk students with precision, institutions can allocate resources more efficiently and design personalized support mechanisms. Despite these contributions, the study has several limitations. The exceptionally high model accuracy may indicate overfitting or dataset-specific characteristics that limit generalizability to other institutional contexts. Additionally, the dataset does not capture qualitative factors such as student motivation, socio-economic challenges, or psychological well-being, which may also influence retention outcomes.

This study demonstrates that integrating predictive analytics and explainable AI offers a powerful, scalable approach to improving student retention, providing higher education institutions with a robust foundation for data-driven, transparent, and strategic decision-making in the digital era. This work can be extended by integrating advanced deep learning techniques with longitudinal data analysis to capture better evolving student behaviours over time.

REFERENCES

Ajayi, O. E., & Letseka, M. (2026). AI-driven learning analytics in higher education: From engagement to outcomes. *Trends in Higher Education, 5*(1), 16.

Aunimo, L., Kauttonen, J., Vahtola, M., & Huttunen, S. (2024). Combining LMS and student register data to analyze study duration—Journal *of Computing in Higher Education*.

Baker, R. S., & Inventado, P. S. (2021). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). Springer.

Science World Journal Vol. 21(No 3) 2026 https://dx.doi.org/10.4314/swj.v21i3.3
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

De Silva, L. M. H., Chounta, I.-A., Rodríguez-Triana, M. J., Roa, E. R., Gramberg, A., & Valk, A. (2022). Toward an institutional analytics agenda for addressing student dropout. *Journal of Learning Analytics, 9*(2), 179–201.

Elebe, O., & Imediegwu, C. C. (2024). Capstone model for retention forecasting using business intelligence dashboards in graduate programs. *International Journal of Scientific Research in Science and Technology*, *11*(4), 655–675.

Elugbaju, W. K., Okeke, N. I., & Alabi, O. A. (2024). Conceptual framework for enhancing decision-making in higher education through data-driven governance. *Global Journal of Advanced Research and Reviews*, *2*(02), 016-030

Gutema, D. M., Pant, S., & Nikou, S. (2024). Exploring key themes and trends in international student mobility research—A systematic literature review. *Journal of Applied Research in Higher Education*, *16*(3), 843–861.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Muñoz-Gama, J. (2022). Mining theory-based patterns from Big data—computers *in Human Behavior, 132*, 107246.

Matz, S. C., Bukow, C. S., Peters, H., et al. (2023). Using machine learning to predict student retention from socio-demographic and engagement data. *Scientific Reports, 13*, 5705.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Muñoz-Gama, J. (2022). Mining theory-based patterns from Big data—computers *in Human Behavior, 132*, 107246.

Matz, S. C., Bukow, C. S., Peters, H., et al. (2023). Predicting student retention using machine learning. *Scientific Reports, 13*, 5705.

Nimy, E., & Mosia, M. (2024). Modelling student retention using Bayesian approaches. *Education Sciences, 14*(8), 830.

Nurudeen, A. H., Fakhrou, A., Lawal, N., & Ghareeb, S. (2024). Academic performance of engineering students. *Engineering Reports, 6*(5), e12654.

Sætnan, A. R., Schneider, I., & Green, N. (2018). The politics of Big Data: Principles, policies, practices. In *The politics and policies of big data* (pp. 1–18). Routledge.

Seidman, A. (Ed.). (2024). *College student retention: Formula for student success*. Bloomsbury Publishing USA.

Shortlidge, E. E., Gray, M. J., Estes, S., & Goodwin, E. C. (2024). The value of support: STEM intervention programs affect students' persistence and sense of belonging. *CBE—Life Sciences Education*, *23*(2), ar23.

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2020). The role of bias in predictive modelling. *PLoS ONE, 15*(6), e0233977.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). University of Chicago Press*

Umendu, E. C., Ghanzanfar, M., Kans, A., & Ahad, M. A. (2026). Enhancing student retention in higher education using machine learning. *Electronics, 15*(4), 734.

Villegas-Ch, W., Govea, J., & Revelo-Tapia, S. (2023). Improving student retention using machine learning: sustainability*, 15*(19), 14512.