



ACADEMIC ACHIEVEMENT, ATTENDANCE, AND DIGITAL LEARNING BEHAVIOUR AS PREDICTORS OF STUDENT ACADEMIC RISK IN HIGHER EDUCATION

Oinam Jimmy Singh¹, Dr. Neelam Tiwari², Dr. Tanay Vyas³

¹Bachelor of Arts, Regional College, Imphal, Manipur, India, Email id: jimmyoinam123@gmail.com BHMS

²MD (Hom), PhD (Practice of Medicine), Associate Professor, PG Department of Practice of Medicine, Rajasthan Vidyapeeth Homoeopathic Medical College & Hospital, Udaipur, Rajasthan, India, ORCID: 0009-0009-3403-3613, Email: drneelamtiwari0721@gmail.com

³BHMS, MD (Hom), PhD (Paediatrics), Associate Professor, PG Department of Paediatrics, Rajasthan Vidyapeeth Homoeopathic Medical College & Hospital, Udaipur, Rajasthan, India, ORCID: 0009-0008-6425-5399, Email: chayan4ever@gmail.com

Article History:

Article Type: **Research**

Received Date: 22/04/2026

Revised Date: 15/07/2026

Accepted Date: 11/08/2026

Published Date: 25/08/2026

Keywords: academic risk; higher education; learning analytics; digital learning behaviour; student performance prediction; educational data mining

ABSTRACT

One of the goals of learning analytics in higher education is to identify students at risk of low academic achievement early on. Whether academic achievement, attendance, and digital learning behaviour predicted academic risk and whether digital indicators provided additional information to grade point average (GPA) and attendance were studied. 1439 college students publicly available data was analyzed. Academic risk was determined by an average course grade in the lowest 25% ($n = 363$, 25.2%). The provided risk label was excluded, because it could be recovered predetermined from GPA and attendance, leading to target leakage. Bivariate tests, hierarchical logistic regression, and cross-validated logistic regression, Random Forest, and XGBoost models were used. Weak linear relationships were found between GPA, attendance, and digital indicators and the full logistic model accounted for little variance (McFadden pseudo- $R^2 = 0.010$). Random Forest had the best cross-validated discrimination (ROC-AUC=0.717), followed by XGBoost (0.681) and logistic regression, which was near chance (0.519). Adding digital behaviour to GPA and attendance increased Random Forest ROC-AUC from 0.589 to 0.678. It is important to note that the results are predictive and not causal and can be used to guide educational interventions. These results suggest that nonlinear combinations of these behavioural signals could enhance the ability to predict academic risk, though external and longitudinal validation must be completed before they can be used in institutions.

1. Introduction

Increasingly, universities are collecting data about student performance, attendance and engagement with digital learning environments. These data have become available to learning analytics, which is the measurement, collection, analysis, and reporting of data on learners and their contexts with the purpose of understanding and optimizing learning (Siemens & Long, 2011), and can now be used to create early-warning systems that identify students who are at risk of poor academic performance. Field reviews note an increase in the use of prediction and ongoing issues of generalizability, interpretability, and evidence of educational effect (Ifenthaler & Yau, 2020; Viberg et al., 2018).

Most early-warning strategies are based on a straightforward principle: that those who score poorly on prior achievement, attendance, and digital engagement are more likely to have trouble later on. There is evidence for each element in certain contexts (Crede et al., 2010; Macfadyen & Dawson, 2010; Richardson et al., 2012). However, the intensity of these links is different for each course, institution, and operationalisation (Conijn et al., 2017; Gašević et al., 2016). So it remains to be seen how much overlap there is between the conventional academic indicators and the LMS-based indicators and whether LMS based indicators can provide information that is not already offered by grades and attendance.

The second question is of a methodological nature. Flexible machine-learning algorithms have been used in many predictive studies, yet the implications of any advantage over traditional regression are rarely explored in the context of education research. A gain can be a result of nonlinear relationships or interactions that cannot be captured by additive models unless they are known a priori. It could also be due to data artefacts, where predictors provide information about the outcome (Kaufman et al., 2012), called target leakage. To distinguish between these possibilities, the outcome variable needs to be transparently audited, validation needs to be separated from model tuning and comparison of predictor domains are necessary.

The present study aims to answer both of the questions by employing a public multi-dimensional student data set (Hasan, 2026). It looks at academics achievement (GPA), attendance and digital learning behaviour as predictors of a relative indicator of low academic achievement, compares logistic regression with 2 tree-based ensemble methods, and quantifies the incremental predictive information of digital behaviour. Here machine learning is employed as a method. The contribution is intended to be educational – to clarify what different types of student information can and cannot contribute to the identification of students who may need support.

2. Literature Review and Conceptual Background

2.1 Academic Achievement and Academic Risk

GPA provides an overall picture of performance in courses and is commonly used as a measure of a student's academic status. Richardson et al. (2012) found that prior academic performance is one of the correlates of later university performance, along with self-regulatory and motivational constructs. Studies on predictive study in higher education also mention that previous and cumulative academic achievement is the most common predictor (Alyahyan & Düştegör, 2020).

Achievement-only warning systems have their limitations. Marks are awarded after assessment has taken place, influenced by assessment for the particular course and tell little about the behaviours they reflect. Academic success is also a multi-faceted phenomenon, not only grades but also satisfaction, learning of skills, persistence, and/or achievement of learning goals (York et al., 2015). This study therefore considers low performance in scores to be one aspect of operational academic difficulty, and not a measure of student success.

2.2 Attendance and Student Engagement

The notion of student engagement is multi-dimensional and includes behavioural, emotional and cognitive elements (Fredricks et al., 2004) and is treated differently in the context of higher education depending on the theorising of engagement that is used (Kahu, 2013). Attendance is a visible behaviourally active expression of engagement. A metaanalysis of college studies found a positive relationship between class attendance and grades; and that class attendance accounted for variance in student performance in college beyond several other predictors (Crede et al., 2010). There has also been a positive association between engagement measures and first year grades and persistence, albeit

modest (Kuh et al., 2008).

Attendance can give early-warning data as it can be observed early in a term, and it is not dependent on assessment outcomes. Its predictive power is still context dependent, however. Not only do attendance policies and course formats influence the actual meaning of an absence, students' circumstances do as well. The effect of attendance on outcomes could thus vary from context to context and may not be linear.

2.3 Digital Learning Behaviour and Learning Analytics

LMS data can provide a record of student activity, such as when they logged in, the amount of time spent in the LMS, when they submitted assignments, when they posted to forums, and when they watched videos. Initial research revealed that some LMS variables were associated with final grades and that these variables could be used to support a proof-of-concept early-warning model (Macfadyen & Dawson, 2010). Further studies have shown that the meaningfulness of trace data depends on the course design and the instructional context (Conijn et al., 2017; Gašević et al., 2016) and that trace data from various sources can complement each other (Tempelaar et al., 2015).

Digital traces should not be interpreted as learning measures. The number of logins does not mean how well a student engages and a long login could mean that the student is studying or not paying attention. However, when it is used outside of learning theory and context for learning, some learning analytics work has been discredited (Gašević et al., 2015). The indicators in this study are thus termed as LMS derived behavioural measures. They are interpreted as potential cues to patterns of engagement, NOT as evidence of learning quality.

2.4 Predictive Modeling of Academic Risk

The education data mining and learning analytics literature reviews reveal that many different predictive methods have been used with students' data (Baker & Siemens, 2014; Namoun & Alshamqiti, 2021; Romero & Ventura, 2020). Examples of empirical applications are LMS-based early-warning models (Akçapınar et al., 2019). Prediction is different from explanation. While explanatory modeling is based on the testing of theoretically specified relationships, predictive modeling is based on assessing the predictive power of the model to forecast new observations (Shmueli, 2010; Breiman, 2001b).

Logistic regression is an easily understood method widely used, but it models the log-odds of the outcome as an additive linear combination of the predictors (unless interaction or nonlinear terms are included) (Hosmer et al., 2013). Student behaviours can be complex combinations. For instance, the significance of a certain session duration might be contingent on the type of submission behavior or the type of course in which it is found. Some ensemble algorithms like decision trees or Random Forest (Breiman, 2001a) and gradient boosting algorithms like XGBoost (Chen & Guestrin, 2016) can capture interactions and thresholds without any upfront commitment. However, they are not necessarily better. They are more difficult to interpret (Rudin, 2019) and can give optimistic performance estimates if tuning and evaluation are not kept apart (Cawley & Talbot, 2010; Varma & Simon, 2006). When using in education, there is a need to consider calibration (Van Calster et al., 2019), algorithmic bias (Baker & Hawn, 2022), and ethical questions regarding student data (Slade & Prinsloo, 2013).

2.5 Research Gap, Objectives, and Research Questions

Two issues are still left unresolved from previous studies. First, incremental value of digital behaviour beyond GPA and attendance is often assumed and has not been tested by comparison of the predictor domains under the same validation scheme. Second, the integrity of the outcome label is rarely reported when it comes to comparisons of conventional and flexible models. This study does both by auditing the provided risk label, building an independent outcome and testing predictor domains using nested cross-validation.

The objectives were (RO1) to explore relationships of academic achievement, attendance and digital learning behaviour with academic risk in higher education; (RO2) to investigate whether and how digital learning behaviour contributes to a better prediction of academic risk than academic achievement and attendance alone; and (RO3) to assess the predictive power of nonlinear machine learning approaches compared to the conventional linear statistical modelling approach to identifying students at risk of low academic performance. These objectives are associated with three research

questions.

RQ1. How do academic achievement, attendance and digital learning behaviours relate to academic risk for students?

RQ2. Does adding digital learning behaviour to GPA and attendance contribute to prediction of academic risk?

RQ3. What are the performance of conventional logistic regression model in comparison with nonlinear machine-learning models in the identification of academic risk?

No directional hypotheses were formulated and the analyses are exploratory in terms of the direction of associations.

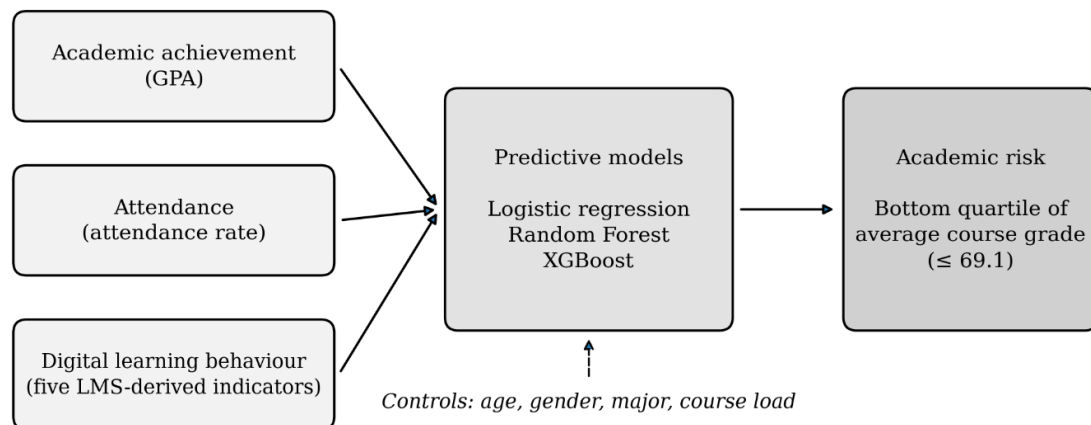


Figure 1: Conceptual and Analytical Framework

Note. Control variables: age, gender, major, and course load.

3. Methodology

3.1 Research Design

The study is quantitative secondary data analysis of a predictive educational study. The data available are cross-sectional, that is, they give one record for each student, and there is no indication of the time order of the variables. The design is not longitudinal or experimental, no causal or temporal inferences can be made based on the analysis. Conventional inferential statistics and machine-learning models, using cross validation, were used. The analytical framework is presented in Figure 1.

3.2 Data Source and Analytical Sample

Data were sourced from the publicly available repository, Mendeley Data, (Hasan, 2026). There are 1,439 records in the file for college (higher-education) students, including academic, attendance, demographic and digital-learning variables. The data audit identified no missing observations, no exact duplicate records and 1,439 unique student identifiers. Only identifiers were used to index records and were not included in any models. The information used for this study does not provide details about recruitment, sampling frame, institution, term, and data-collection dates, and none are assumed.

3.3 Data Auditing and Academic-Risk Operationalization

There is a risk categorical label in the dataset. A preliminary auditing of the data provided indicated that the risk label provided by the respondent was deterministically recoverable from GPA and attendance – a student was classified as high risk if he/she had a GPA of 2.75 or less or an attendance rate of 0.84 or less, low risk otherwise, with no exceptions. We were able to reconstruct all labels also with a shallow decision tree that only applied these two variables. The provided risk label was thus not used in the predictive analyses to prevent issue of circular prediction and target leakage.

Average course grade used to build an alternative outcome. Those students with average course grade below the bottom 25% of the analytical sample (≤ 69.1) were considered at risk ($n = 363$, 25.2%)

while the other students ($n = 1,076$, 74.8%) were not at risk. The average course grade was not included as a predictor as it is the outcome. The result is a relative measure of low academic achievement. None of the following were observed: institutional probation, failure, attrition, dropout, formal risk classification or disciplinary status. A quartile-based definition has the advantage of being analytically useful, and also being independent of institution-specific rules of grading. It is relative to this sample however, so that the same student might be above or below the cut-off if he or she were part of another cohort.

3.4 Variables

The variables and their functions are given in Table 1. GPA was used as an indicator of academic performance. The attendance rate was used to represent attendance. Digital learning behaviour was defined through five behavioural indicators derived from the LMS, including: number of logins per month, average session length, assignments submitted, number of forum entries and video completion rate. Covariates were age, gender, major, and course load. Enrollment status was not included in the primary models since it is not clear of its temporal relationship with academic performance.

Table 1: Variables, Analytical Roles, and Operational Definitions

Variable	Role	Definition (as supplied)
GPA	Predictor: academic achievement	Grade point average
attendance_rate	Predictor: attendance	Attendance rate (proportion)
lms_logins_past_month	Predictor: digital learning	Number of LMS logins in the past month
avg_session_duration_minutes	Predictor: digital learning	Average LMS session duration (minutes)
assignment_submission_rate	Predictor: digital learning	Assignment submission rate (proportion)
forum_participation_count	Predictor: digital learning	Count of forum participations
video_completion_rate	Predictor: digital learning	Video completion rate (proportion)
age, gender, major, course_load	Covariates	Demographic and enrollment characteristics
avg_course_grade	Outcome-defining	Average course grade; at risk if ≤ 69.1 (bottom quartile of sample)
risk_level (supplied)	Excluded	Categorical label recoverable from GPA and attendance (target leakage)
enrollment_status	Excluded	Temporal relationship with performance unclear
student_id	Excluded	Record identifier only

Note. LMS = learning management system. Excluded variables were not used in any predictive model.

3.5 Statistical Analysis

Continuous predictors were analyzed with descriptive statistics and risk group comparisons using Cohen's d (Cohen, 1988) as the effect-size index. Chi-square tests with Cramér's V (Cramér, 1946)

were performed to compare gender and majors. Pearson's correlation values were calculated between the predictors and the binary outcome, and variance inflation factors (VIF) were used to test for multicollinearity.

Predictors were entered in blocks to the hierarchical logistic regression: Model 0 (age, gender, major, course load); Model 1 (Model 0 + GPA); Model 2 (Model 1 + attendance); and Model 3 (Model 2 + the five digital-learning variables). The likelihood-ratio test was used for each block. Final model results were reported as adjusted odds ratios (ORs) with 95% confidence intervals and the model fit was summarized by the pseudo- R^2 of McFadden (1974). Results were considered statistically significant at $p < .05$ and p-values were analyzed with corresponding effect sizes.

3.6 Machine-Learning Validation

The models were tested by nested cross validation using strata of five folds and compared with each other by using logistic regression, Random Forest and XGBoost. Hyperparameter tuning was only applied to the training part of each fold, and the folds were evaluated on data that was not used for tuning (Cawley & Talbot, 2010; Varma & Simon, 2006). No external validation carried out.

ROC-AUC, PR-AUC, balanced accuracy, sensitivity, specificity, precision, F1-score and Brier score (Brier, 1950) were used to summarize performance. When the positive class is a minority, as is the case here (25.2%), precision-recall summaries respond directly to identification of positive cases, and the no-skill baseline for PR-AUC is equal to the positive-class prevalence (Saito & Rehmsmeier, 2015). Sensitivity, specificity, precision, and F1 were assessed for the Random Forest when using the default probability threshold, and the illustrative lower threshold of ~ 0.29 . The minimum level is only indicative and was not external validated.

3.7 Predictor-Domain, Importance, and Sensitivity Analyses

Random Forest models were examined for incremental predictive information by performing repeated 5-fold cross validation with six sets of predictors: using GPA only, attendance only, digital learning only, GPA plus attendance, GPA plus attendance plus digital learning, and the full model with control. Permutation importance, which provides the mean decrease in ROC-AUC on permutation of each predictor, was used to identify the contribution of each predictor to cross validated random forest discrimination. Permutation importance is used to measure the dependence of the model on a variable. It isn't a causal effect, it doesn't have direction, and it can be influenced by correlated predictors (Fisher et al., 2019; Strobl et al., 2007). Four different definitions of risk were used: bottom 20%, bottom 25% (primary), bottom 30% and an average grade of less than 70% (course grade).

4. Results

4.1 Sample and Risk Distribution

Of the 1,439 students, 363 (25.2%) met the academic-risk definition and 1,076 (74.8%) did not (Figure 2).

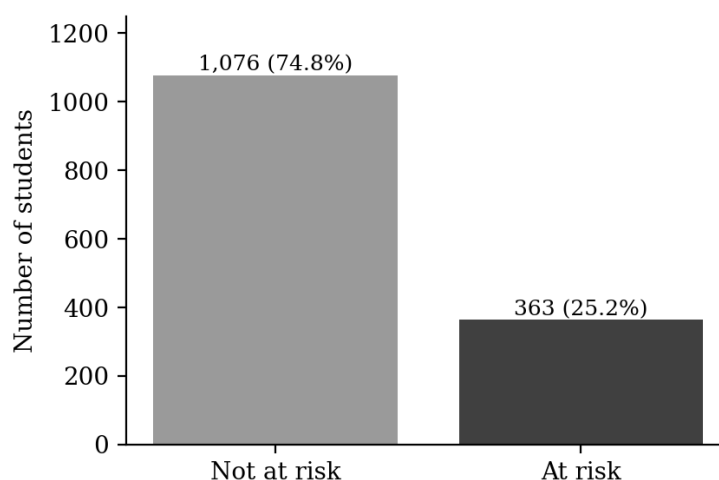


Figure 2: Distribution of Students by Academic-Risk Status

Note. Risk = average course grade ≤ 69.1 (bottom quartile of the sample).

4.2 Bivariate Findings

The comparisons are given in Table 2. There were no significant differences between at-risk and not-at-risk students with respect to GPA ($p = .087$, $d = 0.103$), attendance rate ($p = .979$), or any digital-learning indicator (all $p \geq .351$). All effect sizes were very small ($|d| \leq 0.103$). The point-biserial correlation values between GPA ($r = .045$), attendance ($r = -.001$), LMS logins ($r = .002$), session duration ($r = -.003$), assignment submission ($r = -.001$), forum participation ($r = -.008$), and video completion ($r = -.024$) and academic risk were all near zero. It was observed that the VIF values were not in the range of 1.05 to 9.00 which indicates absence of multicollinearity problem between the variables as all VIF values were approximately in the range of 1.00 to 1.04.

There was a small association between gender and risk status ($\chi^2 = 9.37$, $p = .009$, Cramér's $V = .081$). Major was not associated with risk status, $\chi^2 = 2.33$, $p = .507$, $V = .040$.

Table 2: Descriptive Statistics and Comparisons by Academic-Risk Status

Panel A. Continuous variables					
Variable	At risk ($n = 363$) M	Not at risk ($n = 1,076$) M	p	Cohen's d	r with risk
GPA	3.132 $\pm$ 0.585	3.071 $\pm$ 0.595	.087	0.103	.045
Attendance rate	0.811 $\pm$ 0.128	0.811 $\pm$ 0.128	.979	-0.002	-.001
LMS logins per month	19.15 $\pm$ 11.44	19.10 $\pm$ 11.57	.937	0.005	.002
Session duration (min)	49.75 $\pm$ 22.89	49.91 $\pm$ 23.22	.910	-0.007	-.003
Assignment submission rate	0.754 $\pm$ 0.142	0.755 $\pm$ 0.142	.956	-0.003	-.001
Forum participation	9.61 $\pm$ 5.74	9.72 $\pm$ 5.65	.754	-0.019	-.008
Video completion rate	0.747 $\pm$ 0.143	0.755 $\pm$ 0.147	.351	-0.056	-.024

Panel B. Categorical variables			
Variable	χ^2	p	Cramér's V
Gender	9.37	.009	.081
Major	2.33	.507	.040

Note. Positive d indicates a higher mean in the at-risk group. r is the correlation between the predictor and the binary risk indicator. VIF values ranged from approximately 1.00 to 1.04.

4.3 Hierarchical Logistic Regression

Fit of the covariate model was not significantly improved by the addition of GPA, LR $\chi^2(1) = 2.58$, $p = .108$. Adding attendance did not, LR $\chi^2(1) = 0.13$, $p = .718$. Adding the digital-learning block did not, LR $\chi^2(5) = 0.97$, $p = .965$ (Table 3, Panel A). The complete model accounted for very little of the variance (McFadden pseudo- $R^2 = 0.010$). All of the GPA, attendance, and digital-learning coefficient was non-statistically significant and contained 1 in the final model (Table 3, Panel B). Only one coefficient was statistically significant: gender, with male students having less chance of being classified as being at risk than female students (OR = 0.636, 95% CI [0.472, 0.856], $p = .003$). This

finding is not highlighted as the bivariate effect size is small.

Table 3: Hierarchical Logistic Regression Predicting Academic Risk

Panel A. Incremental likelihood-ratio tests

Block added	LR χ^2	df	p
GPA (Model 0 → Model 1)	2.58	1	.108
Attendance (Model 1 → Model 2)	0.13	1	.718
Five digital-learning variables (Model 2 → Model 3)	0.97	5	.965

Panel B. Final model (Model 3): adjusted odds ratios

Predictor	Adjusted OR	95% CI	p
GPA	1.190	[0.966, 1.465]	.103
Attendance rate	0.839	[0.323, 2.178]	.718
LMS logins per month	1.001	[0.991, 1.012]	.801
Session duration	1.000	[0.994, 1.005]	.872
Assignment submission rate	0.990	[0.424, 2.311]	.981
Forum participation	0.997	[0.976, 1.019]	.806
Video completion rate	0.684	[0.301, 1.558]	.366
Male (vs. female)	0.636	[0.472, 0.856]	.003

Note. Model 0 included age, gender, major, and course load. The final model had McFadden pseudo- $R^2 = 0.010$. Coefficients for age, major, and course load are not shown. OR = odds ratio; CI = confidence interval.

4.4 Predictive Model Performance

Nested cross-validation yielded an approximately chance level (ROC-AUC = 0.519 ± 0.029) for logistic regression. Random Forest had the highest cross-validated discrimination (ROC-AUC = 0.717 ± 0.038), followed by XGBoost (ROC-AUC = 0.681 ± 0.038 ; Table 4). PR-AUC followed the same order (0.285, 0.577, and 0.521), against a no-skill baseline of 0.252. The Brier scores were 0.190 for logistic regression and 0.167 for both the tree-based models. Logistic regression, Random Forest and XGBoost had balanced accuracy of 0.500, 0.590 and 0.640 respectively. No formal inferential comparison of the cross validated AUCs was performed and these differences are referred to as observed validation results.

The Random Forest was conservative at its default probability threshold with sensitivity ~19.3%, specificity ~98.6%, precision ~88.6%, and F1 ~0.297. When the threshold was set at ~0.29, sensitivity increased to ~49.6%, precision decreased to 51.0%, specificity was 84.0%, and balanced accuracy was ~66.8%. This threshold was not validated from external sources.

Table 4: Nested Cross-Validated Predictive Performance of Three Models

Model	ROC-AUC ($M \pm SD$)	PR-AUC	Balanced accuracy	Brier score
Logistic regression	0.519 ± 0.029	0.285	0.500	0.190
Random Forest	0.717 ± 0.038	0.577	0.590	0.167

Model	ROC-AUC ($M \pm SD$)	PR-AUC	Balanced accuracy	Brier score
XGBoost	0.681 ± 0.038	0.521	0.640	0.167

Note. Nested stratified five-fold cross-validation; no external validation. The PR-AUC no-skill baseline equals the positive-class prevalence (.252). No inferential comparison of AUCs was conducted. At the default threshold, the Random Forest showed sensitivity $\approx 19.3\%$, specificity $\approx 98.6\%$, precision $\approx 88.6\%$, and F1 ≈ 0.297 ; at an illustrative threshold of ≈ 0.29 , sensitivity $\approx 49.6\%$, specificity $\approx 84.0\%$, precision $\approx 51.0\%$, and balanced accuracy $\approx 66.8\%$.

4.5 Incremental Contribution of Digital Learning Behaviour

In repeated five fold Random Forest validation (Table 5; Figure 3), GPA alone failed to discriminate beyond chance (ROC-AUC = 0.495), and attendance alone showed limited discrimination (0.564). Five of the indicators addressed digital-learning only, and reached 0.670. GPA plus attendance yielded 0.589 ± 0.026 . Incorporating digital learning into the two predictors further boosted discrimination to 0.678 ± 0.034 , a net improvement of around 0.090. The complete model including controls was 0.727 ± 0.030 . The incremental predictive information of digital behaviour was thus provided in the nonlinear models, but not in the logistic model.

Table 5: Random Forest Predictive Performance by Predictor Domain

Predictor set	ROC-AUC
GPA only	0.495
Attendance only	0.564
Digital learning only	0.670
GPA + attendance	0.589 ± 0.026
GPA + attendance + digital learning	0.678 ± 0.034
Full model + controls	0.727 ± 0.030

Note. Repeated five-fold cross-validation. Standard deviations are shown where available. The increment from adding digital learning to GPA and attendance was approximately +0.090 ROC-AUC.

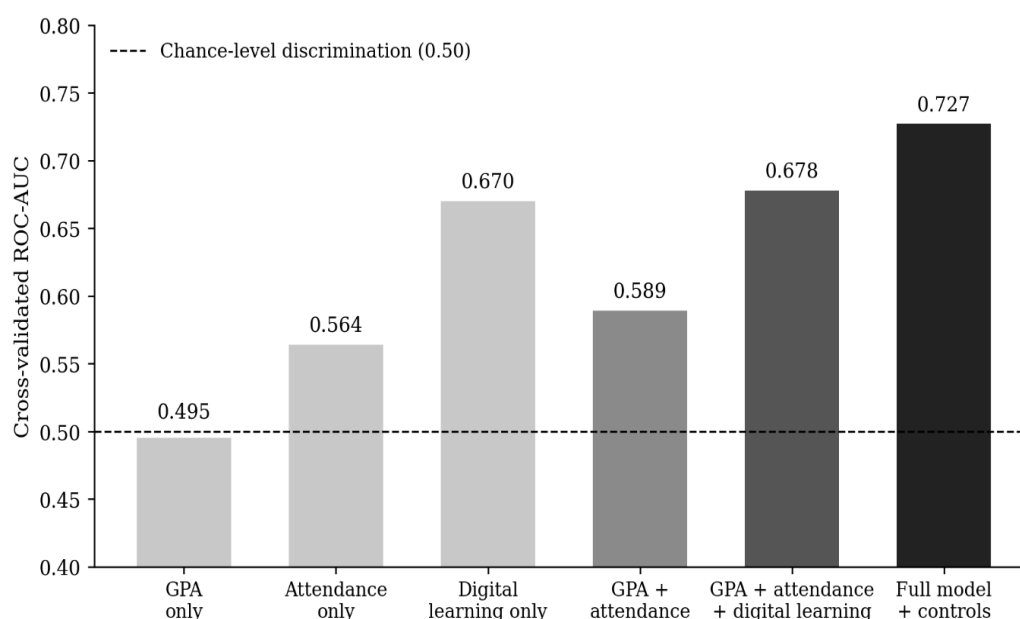


Figure 3: Random Forest ROC-AUC by Predictor Domain

Note. Dashed line indicates chance-level discrimination (0.50). Repeated five-fold cross-validation.

4.6 Variable Importance

The permutation importance was distributed across predictors (figure 4). The categories that saw the highest mean decrease in ROC-AUC were major (0.036), followed by average session duration (0.033) and forum participation (0.026), age (0.026) and GPA (0.026). Attendance (0.022), gender (0.021), assignment submission (0.020), course load (0.019), LMS logins (0.019), and video completion (0.018) followed. None of the variables were predominant. In the digital indicators, there was a higher relative value in session length and forum involvement. Permutation importance does not have a direction and so these values do not indicate whether the higher or lower values of any variable were associated with greater risk.

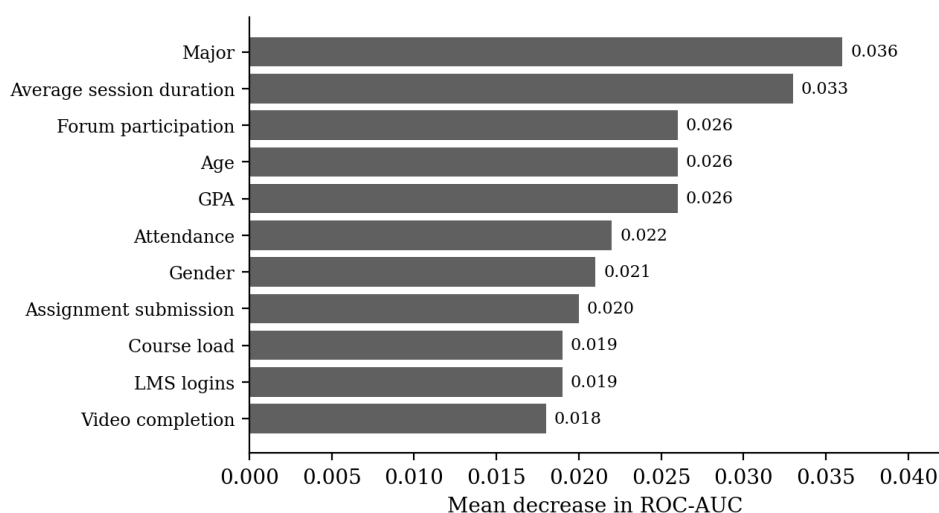


Figure 4: Random Forest Permutation Importance

Note. Mean decrease in ROC-AUC. Values indicate reliance of the model on each predictor and carry no causal meaning or direction.

4.7 Sensitivity Analysis

Random Forest discrimination was similar across outcome definitions (Table 6): ROC-AUC = 0.685 for the bottom 20% ($n = 290$), 0.719 for the primary bottom 25% ($n = 363$), 0.717 for the bottom 30% ($n = 433$), and 0.720 for average course grade below 70 ($n = 397$). Logistic regression was persistently around 0.50–0.52 no matter the definition. Note that the value for the primary definition (0.719) is slightly different from the nested cross validation estimate (0.717) as these are from different validation exercises.

Table 6: Sensitivity Analyses Under Alternative Academic-Risk Definitions

Risk definition	At-risk n	Random Forest ROC-AUC
Bottom 20%	290	0.685
Bottom 25% (primary)	363	0.719
Bottom 30%	433	0.717
Course grade < 70	397	0.720

Note. Logistic regression ROC-AUC remained approximately 0.50–0.52 across definitions.

5. Discussion

5.1 Principal Findings

Comparisons of results from additive and nonlinear analysis are shown. Bivariate relationships between individual predictors and the reconstructed outcome were negligible and the inclusion of GPA, attendance, or digital learning in hierarchical logistic models did not confer much added value. However, Random Forest and XGBoost had moderate cross-validated discrimination, and digital behaviour had better discrimination than GPA and attendance and predictive information was spread

out in many variables. The answer to RQ1 is that, as operationalized, linear associations were weak. The answer to RQ2 is yes in the case of the nonlinear model (between 0.589 and 0.678) and no in the case of the logistic model. For RQ3 the answer is that Random Forest and XGBoost demonstrated greater cross-validated discrimination than logistic regression on this data set.

5.2 Academic Achievement and Attendance

GPA and attendance were not revealed to be strong independent predictors of low average course grade. The result is specific to this data set, this definition of outcome, and this design of analysis. It is not to be interpreted as a reflection of the unimportance of achievement and attendance in higher education. Existing research shows that there are generally positive associations between attendance and grades, and between past and future performance (Crede et al., 2010; Richardson et al., 2012).

There are a number of methodological differences that might explain the divergence but the present data do not allow the adjudication. The result here is a relative quartile for the average course grade, while previous studies have generally considered course grades or GPA as a continuous outcome. The time of the predictors with respect to the criterion is not reported, and it is unclear whether GPA is prior to or concurrent with course grades. The population and the institutional context are also not documented. Finally, the risk label provided was completely dependent on the variables used to make it (GPA and attendance), and these variables demonstrated only weak and statistically insignificant association with an independent outcome. This pattern is a testament to the fact that a label's construction as well as its content must be examined.

5.3 Digital Learning Behaviour

The incremental findings are related to digital learning behaviour. However, none of the five indicators was significant in the logistic model and the digital block did not significantly improve the model. However, the combination of the block with GPA and attendance resulted in an increase of the Random Forest discrimination to 0.678, and the discrimination for the digital indicators was 0.670. The first is that digital behaviour is an interaction or threshold or conditional combination or a heterogeneous pattern, rather than a monotonic effect. This is an interpretation which is consistent with the results, but not a demonstrated mechanism; the interaction was not specifically tested in the analyses.

Results do not reveal any specific LMS behaviors that lead to improved or less successful outcomes. Section 2.3 mentioned that LMS tracks the activity of the platform, not of learning quality. Compared to those of other predictors, session duration and forum participation were more important in terms of permutation importance, but it remains unclear how these factors connected to risk and the shape of the relationship, and values also depend on the model and correlated predictors.

5.4 Linear and Nonlinear Academic-Risk Prediction

Logistic regression was able to achieve ROC-AUC of only 0.519, near chance, whereas Random Forest could achieve 0.717 and XGBoost 0.681. These are all numbers for middle of the road discrimination for the tree-based models, not great discrimination. One possible explanation is that the tree-based methods allow for nonlinearities and interactions without having to be predicted beforehand, whereas the logistic model had only main effects. The comparison is thus not between a particular additive specification and the flexible models; rather, it is not a comparison of the superiority of machine learning over statistical modelling in general. A logistic model with well selected interaction or spline terms may behave differently and was not investigated. Differences, if any, are not statistically determined but observed, based on the absence of a formal comparison of the cross-validated AUCs.

5.5 Implications for Higher-Education Practice

The results indicate that instructors, advisors, and student-support units should be wary of using any one of the indicators. GPA and attendance alone did not identify low performance under this outcome and combinations of behavioural signals contained more information than either alone. Digital behaviour can be used as an additional part of conventional academic records in an early-warning system, especially when used in combination with multiple weak signals. Many LMS administrators should be aware, however, that these indicators are activity traces and the interpretation of their meaning is dependent on the design of the course.

You can't just use model discrimination in practice. The Random Forest at the default threshold was

able to detect approximately one-fifth of the students at risk, which means that if this threshold were used in a system, the vast majority of students who the system was designed to serve would not be identified by the system. A lower threshold would increase sensitivity to about 49.6% and lower precision to 51.0%, resulting in about 50% of the students identified by the lower threshold being in the "at-risk" group. Hence, there is a trade-off in choosing a threshold, and institutions would choose it based on the cost of each error and the nature of the response. Responses should be helpful, for example, referring to contact or study-skills materials, and should not be the nature of a punishment. The score on a model should be used to inform, but not define, educational judgments, and a student should not be permanently labeled based on a model score.

5.6 Methodological Implications

The target-leakage discovery has implications for the broader field of educational data mining: any outcome data provided as input to the modeler should be checked first. A label that can be recomputed from candidate predictors would lead to almost perfect but circular prediction (Kaufman et al., 2012). The study also provides an example for the difference between explanatory and predictive goals (Shmueli, 2010). A predictor set may exhibit near zero coefficients in an additive model, and yet provide moderate discrimination out of sample in a flexible model. Nested cross-validation, predictor-domain comparisons and sensitivity analyses across outcome definitions guarded against over optimistic and definition dependent conclusions. Calibration was not evaluated beyond the Brier score, and calibration curves were not explored; calibration needs to be evaluated prior to using the probabilities to make decisions (Van Calster et al., 2019).

5.7 Limitations and Future Research

The conclusions are subject to several limitations. Data are secondary, cross-sectional and temporal ordering of predictor and outcome is not shown. The result is an analytically determined low-performance indicator (not an institutionally reported risk classification) and the bottom-quartile is defined in relation to this sample. External validity is unclear due to lack of documentation on sampling and data generation, and external validation was not conducted. Digital measures do not measure the quality of learning. Predictive associations do not show causal relationship. A moderate AUC is not indicative of readiness for deployment and the thresholds results demonstrate practical limitations of sensitivity. The analyses did not examine fairness of the assessments between student groups, an important factor to consider for operational use (Baker & Hawn, 2022).

These results should be validated in future studies on external data sets and on longitudinal data from institutions with ordered predictors. It is to be based on institutionally recorded outcomes (e.g., institutionally recorded course failure, probation, retention, or drop-out) and repeated-measures LMS traces. Models would then be useful in terms of additional practical value if demonstrated by fairness analyses, prospective validation and decision-curve / utility analysis (Vickers & Elkin, 2006). Research is required to establish whether or not an intervention study will make a difference in its outcomes for students.

6. Conclusion

Taking a provided risk label (recoverable from GPA and attendance), this study investigated whether academic achievement, attendance, and digital learning behaviour predicted a relative indicator of low academic performance among 1,439 higher-education students. As to RQ1, simple linear associations were weak in terms of the absence of any predictor being meaningfully different between the groups and the logistical hierarchical regression accounting for little variance. Digital learning behaviour was added to the predictive information in the nonlinear models, raising the Random Forest discrimination beyond GPA and attendance from 0.589 to 0.678, but not the logistic model, in answer to RQ2. Random Forest and XGBoost exhibited moderate levels of cross-validated discrimination, while logistic regression was near chance in response to RQ3. There was no one academic or digital variable that was the most prevalent.

In the context of higher education, the results indicate that perhaps it would be helpful to look at a set of behaviours to evaluate early-warning strategies, rather than just academic measures. They further demonstrate that discrimination is insufficient; threshold selection, sensitivity, calibration and supportive use of model outputs are all of importance for educational value. Results are predictive;

they are based on a single cross-sectional data collection; and they employ an analytically determined outcome. They need validation of the outside world, validation from the length of their stay, and validation from a fairness perspective before they can be used in any institutions.

Data Availability Statement

The dataset analyzed in this study is publicly available through Mendeley Data (Hasan, 2026; <https://doi.org/10.17632/fj67jtpfp8.1>).

Ethics Statement

This study involved secondary analysis of a publicly available dataset and did not involve new recruitment, intervention, or direct contact with human participants by the present authors. No additional personally identifiable information was collected. Ethical procedures relating to the original data collection are reported only to the extent documented by the dataset provider.

References

1. Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16, Article 40. <https://doi.org/10.1186/s41239-019-0172-z>
2. Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17, Article 3. <https://doi.org/10.1186/s41239-020-0177-7>
3. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
4. Baker, R. S. J. d., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253–272). Cambridge University Press.
5. Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
6. Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
7. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
8. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
10. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
11. Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17–29. <https://doi.org/10.1109/TLT.2016.2616312>
12. Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
13. Crede, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, 80(2), 272–295. <https://doi.org/10.3102/0034654310362998>
14. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
15. Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109.

- <https://doi.org/10.3102/00346543074001059>
16. Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84. <https://doi.org/10.1016/j.iheduc.2015.10.002>
 17. Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
 18. Hasan, F. (2026). A multi-dimensional college student behavioral dataset for educational data mining and risk assessment (Version 1) [Data set]. Mendeley Data. <https://doi.org/10.17632/fj67jtpfp8.1>
 19. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. <https://doi.org/10.1002/9781118548387>
 20. Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4), 1961–1990. <https://doi.org/10.1007/s11423-020-09788-z>
 21. Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education*, 38(5), 758–773. <https://doi.org/10.1080/03075079.2011.598505>
 22. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. <https://doi.org/10.1145/2382577.2382579>
 23. Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, 79(5), 540–563. <https://doi.org/10.1080/00221546.2008.11772116>
 24. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2), 588–599. <https://doi.org/10.1016/j.compedu.2009.09.008>
 25. McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.
 26. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), Article 237. <https://doi.org/10.3390/app11010237>
 27. Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138(2), 353–387. <https://doi.org/10.1037/a0026838>
 28. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
 29. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
 30. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
 31. Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
 32. Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 30–40.
 33. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
 34. Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, Article 25. <https://doi.org/10.1186/1471-2105-8-25>
 35. Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data

- for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167. <https://doi.org/10.1016/j.chb.2014.05.038>
36. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
37. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
38. Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
39. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574. <https://doi.org/10.1177/0272989X06295361>
40. York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation*, 20, Article 5. <https://doi.org/10.7275/hz5x-tx03>