US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 3, No. 1 · Article ID: USJNITE-0301 · Journal Homepage: www.edtechinsights.org

Learning Analytics for Early Identification of At-Risk Students: A Machine Learning Approach

Md Rasel Ul Alam[1], Kanita Haider[2+], Oishe Al Mariz[2]

[1] *University of the Cumberlands, Kentucky, USA*

[2] *Chittagong University of Engineering and Technology, Chattogram 4349, BD*

[+] *Corresponding Author Email: kanita.haider4422@gmail.com*

Received 15 January 2023; Revised 23 April 2023; Accepted 19 May 2023; Published 25 July 2023

KEYWORDS

Learning Analytics

Educational Data Mining

At-Risk Students

Early Warning Systems

Machine Learning

Predictive Modeling

Abstract

Identifying students at risk of academic failure early enough to intervene is a central goal of learning analytics and educational data mining (EDM). This paper reports an empirical machine learning study built on the widely used UCI Student Performance dataset (Cortez & Silva, 2008; N = 395 secondary school students). An at-risk indicator was defined as a final grade below 10 out of 20, and three supervised classifiers, logistic regression, random forest, and gradient boosting, were trained on an early-warning feature set consisting of demographic, socio-behavioral, and first-period academic indicators, deliberately excluding later-period grades to preserve genuine early-prediction validity. All three models achieved strong discrimination (area under the receiver operating characteristic curve, AUC, between 0.910 and 0.912), with gradient boosting achieving the best balance of precision and recall (F1 = 0.771) on a held-out test set. First-period grade, prior course failures, and absenteeism emerged as the dominant predictors of at-risk status. The paper situates these results within the broader learning analytics literature on early-warning systems and discusses the implications, limitations, and ethical considerations of deploying such models to trigger institutional support for at-risk students.

1. Introduction

Every academic term, a subset of students falls behind to the point of failing a course or withdrawing altogether, often for reasons that were visible in the data long before a final grade was recorded. Learning analytics (LA) and educational data mining (EDM) have emerged over the past decade as the primary methodological toolkits for turning the digital traces of student activity, enrollment records, demographic profiles, and early assessment scores into actionable, early predictions of academic risk (Akçapınar et al., 2019). The promise of this approach is straightforward: if an institution can identify which students are likely to struggle before the term ends, it can direct tutoring, advising, or financial support toward those students while there is still time for the intervention to matter.

The stakes of getting this right extend well beyond any single course grade. Course failure and withdrawal are strongly associated with delayed graduation, increased tuition cost to the student, and, at the institutional level, lower retention rates that carry direct financial and reputational consequences. Because the marginal cost of a well-timed advising email or tutoring referral is small relative to the cost of a student repeating a course or leaving an institution altogether, even a modestly accurate early-warning model can generate a favorable return on investment if it is coupled with a workable intervention pathway. This economic logic is part of why early-warning analytics has moved from a research curiosity to a standard feature of many student-information and learning-management-system platforms.

Realizing that promise in practice depends on a specific methodological discipline that is easy to violate: an early-warning model must be trained and evaluated using only the information that would genuinely be available at the moment a prediction is needed. Models that quietly incorporate late-course grades or end-of-term data produce inflated performance metrics that do not transfer to real deployment, because the very information that made the prediction easy will not yet exist when an institution needs to act. This paper takes that constraint seriously. Using the well-known UCI Student Performance dataset (Cortez & Silva, 2008), it develops and compares three supervised machine learning classifiers that predict a student's risk of final-course failure using only demographic, socio-behavioral,

and first-grading-period information, explicitly withholding second-period and final grades from the feature set.

The specific objectives of this study are to: (i) define and operationalize an at-risk indicator suitable for early identification; (ii) construct an early-warning feature set consistent with genuine early-prediction constraints; (iii) train and compare logistic regression, random forest, and gradient boosting classifiers on this feature set; (iv) identify which student attributes contribute most to risk prediction; and (v) discuss the practical, ethical, and institutional implications of deploying such a model as part of a learning analytics early-warning system.

The remainder of the paper is organized as follows. Section 2 reviews the learning analytics and educational data mining literature on early-warning systems, predictive features, and the ethical considerations that accompany predictive modeling of student outcomes. Section 3 describes the dataset, the early-warning feature set, and the modeling pipeline. Section 4 reports the empirical results, including class distribution, feature relationships, feature importance, and comparative model performance. Section 5 discusses the practical and ethical implications of the findings and situates them within the broader field, and Section 6 concludes with recommendations for institutions, model developers, and future researchers.

2. Literature Review

2.1 Learning Analytics and Educational Data Mining

Learning analytics and educational data mining share a common ambition, using data about learners and their contexts to understand and improve learning, but have historically emphasized different methods: EDM has tended toward pattern discovery and predictive modeling drawn from the broader data-mining tradition, while LA has emphasized human-facing dashboards, interventions, and institutional decision-making (Romero & Ventura, 2020). Within both traditions, predicting which students are at risk of poor academic outcomes has become one of the most heavily studied applications, precisely because a successful prediction can be converted into a timely, low-cost intervention such as an advising email or a tutoring referral.

A useful way to organize this literature is along two axes: the timing of the prediction relative to the course (early versus late), and the granularity of the outcome being predicted (a continuous grade, a pass/fail label, or a multi-level risk tier). Studies that predict outcomes late in a course, once several assessments have already occurred, routinely report higher accuracy than genuinely early studies, simply because more information is available. This pattern is well documented and is one of the main reasons the present study deliberately restricts its feature set to information available at or before the first grading period.

2.2 Early-Warning Systems: From Course Signals to Modern Classifiers

Early institutional early-warning systems, such as Purdue University's Course Signals initiative, combined simple rule-based scoring with learning management system data to flag at-risk students to instructors in near real time. Since then, the methodological sophistication of these systems has grown substantially. Akçapınar et al. (2019) developed and validated a learning-analytics-based early-warning system that used online course interaction data to flag at-risk students well before a final grade was assigned, demonstrating that behavioral engagement signals carry meaningful predictive value independent of prior academic history. Adnan et al. (2021) systematically varied the percentage of course length available to a predictive model, showing that machine learning classifiers can achieve useful accuracy even when only an early fraction of the course has elapsed, a finding directly relevant to the early-identification goal of the present study. More recent work has extended this line of research to interpretable and institutionally deployed systems: Jang et al. (2022) combined machine learning with explainable AI (XAI) techniques such as SHAP to predict student performance using only practically obtainable features identified through stakeholder focus groups, showing that model predictions could be translated into concrete, feature-level explanations for educators, while Chui et al. (2020) applied a support-vector-machine-based classifier to virtual-learning-environment engagement data, achieving 92–94% accuracy in distinguishing at-risk and marginal university students while substantially reducing the training data required.

2.3 Predictive Features and Model Choice

Across this literature, three categories of predictor recur: static demographic and socio-economic attributes (age, parental education, family structure), behavioral or engagement indicators (attendance, log-in frequency, time-on-task), and early academic performance signals (first assessment scores, prior course failures). Model choice varies by study, but supervised classifiers, particularly logistic regression, random forest, and gradient-boosted trees, dominate the field because they offer a workable balance between predictive accuracy and interpretability, an important property when a model's output will be used to justify a human intervention (Adnan et al., 2021; Jang et al., 2022). Ensemble tree-based methods such as random forest and gradient boosting typically outperform linear models on this kind of tabular, mixed-type educational data because they can capture non-linear interactions, for example, between prior failures and absenteeism, without requiring the analyst to specify those interactions in advance.

2.4 The Feedback Gap and Ethical Considerations

A recurring critique of the predictive-modeling literature is that accurate prediction alone does not improve student outcomes; the model must be paired with a timely, well-designed intervention and validated in a real institutional setting. Bañeres et al. (2020) directly address this gap, developing and deploying a complete early-warning system in a real online university that combined a validated predictive model with student- and

teacher-facing dashboards for reviewing risk alerts and triggering interventions across six semesters of institutional data, demonstrating that predictive accuracy translates into a genuinely usable early-warning tool only when paired with this kind of institutional deployment infrastructure. This literature also raises important ethical considerations: predictive labels can stigmatize students, false positives can create unnecessary anxiety or overreach by advisors, and models trained on historically biased data can reproduce or amplify existing inequities if deployed without careful monitoring. These considerations motivate the discussion of responsible deployment later in this paper.

2.5 Comparative Summary of Reviewed Studies

Table 1 summarizes the studies discussed above along four dimensions: the primary data source, the modeling approach, the timing of prediction relative to the course, and the reported focus of the evaluation. The table highlights that, while reported accuracy figures are not always directly comparable across studies because of differing outcome definitions, sample sizes, and evaluation protocols, the field converges on a consistent conclusion: early, low-latency signals (attendance, first assessments, engagement counts) combined with tree-based ensemble classifiers or transparent linear baselines represent the current best practice for institutionally deployable early-warning systems.

Study	Data Source	Modeling Approach	Prediction Timing	Reported Focus
Akçapınar et al. (2019)	Online course LMS interaction logs	Learning-analytics early-warning system	Before final grade assigned	Demonstrated behavioral engagement signals predict risk independent of prior history
Adnan et al. (2021)	Course-length-segmented records	Multiple ML classifiers	Varied % of course elapsed	Useful accuracy achievable even with an early fraction of the course
Jang et al. (2022)	Stakeholder-identified practical features	ML + explainable AI (SHAP)	Early, practically obtainable features	Model outputs translated into feature-level explanations for educators
Chui et al. (2020)	Virtual learning environment engagement data	Support vector machine	During course	92–94% accuracy distinguishing at-risk / marginal students
Bañeres et al. (2020)	Six semesters of institutional data	Predictive model + dashboards	Deployed, ongoing	Demonstrated full deployment pipeline linking prediction to intervention
This study	UCI Student Performance dataset (Cortez & Silva, 2008)	Logistic regression, random forest, gradient boosting	First grading period only (G2/G3 excluded)	AUC 0.910–0.912; G1, failures, absences as top predictors

Table 1. Summary of key early-warning and at-risk prediction studies reviewed in this paper.

2.6 Feature Engineering Practices in Educational Data Mining

Feature engineering in educational data mining typically follows a small number of recurring patterns that this study also adopts. Categorical variables (such as sex, address type, or school support) are numerically or one-hot encoded so that they can be consumed by a classifier; continuous variables (such as age or number of absences) are often standardized when the downstream model is scale-sensitive, as with logistic regression, but left untransformed for tree-based ensembles, which split on raw thresholds and are invariant to monotonic transformations. Class imbalance, present here in the roughly two-to-one ratio of not-at-risk to at-risk students, is commonly addressed either through resampling techniques such as SMOTE (synthetic minority oversampling) or, as in this study, through stratified train-test splitting combined with threshold-aware evaluation metrics such as recall and F1-score that are less sensitive to imbalance than raw accuracy. Finally, temporal feature governance, deliberately excluding any variable that would not be observable at prediction time, is the single most important and most frequently overlooked step in building a genuinely early-warning, rather than merely retrospective, predictive model.

3. Data and Methodology

This study uses the UCI Student Performance dataset compiled by Cortez and Silva (2008), which records demographic, social, and academic attributes for 395 students enrolled in a secondary-level mathematics course at two Portuguese schools, along with grades

recorded at three points in the course: a first-period grade (G1), a second-period grade (G2), and a final grade (G3), each on a 0–20 scale. Consistent with common practice in the at-risk prediction literature, a binary at-risk indicator was defined as G3 < 10 (a failing final grade), yielding 130 at-risk students (32.9%) and 265 not-at-risk students (67.1%) in the sample.

To preserve the integrity of an early-identification design, the feature set deliberately excluded the second-period grade (G2) and, of course, the final grade (G3) itself, since including either would allow the model to "see" information that would not yet be available at the point an early intervention is needed. The retained feature set instead comprised 25 predictors spanning demographic attributes (sex, age, address type, family size, parental cohabitation status), parental background (mother's and father's education level), behavioral and lifestyle indicators (weekly study time, travel time, going out with friends, weekday and weekend alcohol consumption, free time, family relationship quality, health status), support factors (school and family educational support, private tutoring, participation in extracurricular activities, nursery school attendance, aspiration to pursue higher education, internet access at home, romantic relationship status), number of prior class failures, number of absences, and the first-period grade (G1), which functions as an early academic-performance signal available soon after the course begins. Table 2 organizes the full feature set by category for reference.

Category	Example Variables
Demographic	Sex, age, address type (urban/rural), family size, parental cohabitation status
Parental background	Mother's education level, father's education level
Behavioral / lifestyle	Weekly study time, travel time, going out with friends, weekday and weekend alcohol consumption, free time, family relationship quality, health status
Support factors	School and family educational support, private tutoring, extracurricular activities, nursery school attendance, aspiration to pursue higher education, internet access at home, romantic relationship status
Early academic signal	Number of prior class failures, number of absences, first-period grade (G1)

Table 2. Early-warning feature set (25 predictors), grouped by category.

3.1 Modeling Pipeline

Figure 1a summarizes the end-to-end modeling pipeline used in this study, from raw student records through feature construction, the stratified train-test split, model training, and held-out evaluation. Restricting the feature set at the earliest stage of the pipeline, rather than training on the full feature set and removing late-period grades afterward, ensures that no stage of the pipeline (including any automated feature-selection step) can inadvertently reintroduce information that would leak future knowledge into the prediction.

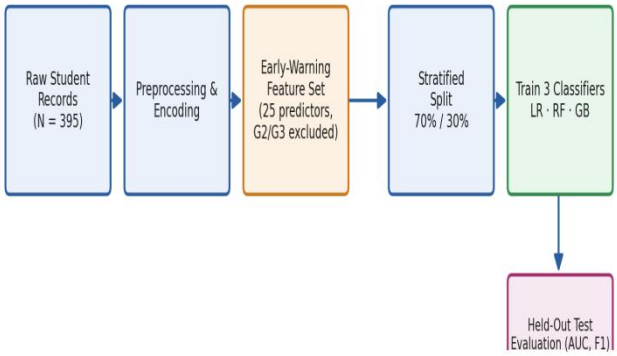

Figure 1a. End-to-end modeling pipeline, from raw student records to held-out evaluation.

Source: Authors' own construction.

3.2 Data Preprocessing and Encoding

The dataset was split into a training set (70%, n = 276) and a held-out test set (30%, n = 119) using stratified sampling to preserve the at-risk/not-at-risk ratio in both partitions. Categorical variables were numerically encoded, and features were standardized (zero mean, unit variance) for the logistic regression model, which is sensitive to feature scale; the tree-based ensemble models were trained on the unscaled feature set, since tree splits are invariant to monotonic feature transformations. No records contained missing values, so no imputation step was required; this is a known property of the Cortez and Silva (2008) dataset, which was collected through in-person school-report and questionnaire data collection rather than automated log extraction.

3.3 Classifiers and Hyperparameters

Three classifiers were trained and compared: (a) logistic regression with L2 regularization, selected as an interpretable linear baseline; (b) a random forest ensemble of 300 trees with a maximum depth of six, selected to capture non-linear feature interactions while limiting overfitting risk on a moderately sized dataset; and (c) a gradient boosting ensemble of 200 trees with a maximum depth of three, selected as a typically strong-performing alternative ensemble method. Table 3 lists the specific hyperparameter configuration used for each classifier.

Classifier	Key Hyperparameters	Rationale
Logistic regression	L2 regularization; features standardized (zero mean, unit variance)	Interpretable linear baseline, sensitive to feature scale
Random forest	300 trees; maximum depth = 6; unscaled features	Captures non-linear interactions while limiting overfitting on a

Classifier	Key Hyperparameters	Rationale
		moderate-sized dataset
Gradient boosting	200 trees; maximum depth = 3; unscaled features	Strong-performing ensemble alternative; shallow trees reduce overfitting per boosting round

Table 3. Classifier configurations and key hyperparameters.

3.4 Evaluation Metrics

Model performance was evaluated on the held-out test set using five complementary metrics. Accuracy is the proportion of all predictions that were correct. Precision is the proportion of students flagged as at-risk who were genuinely at-risk, reflecting the cost of false alarms. Recall (sensitivity) is the proportion of genuinely at-risk students who were correctly flagged, reflecting the cost of missed cases. The F1-score is the harmonic mean of precision and recall and is a useful single summary statistic when, as here, the classes are imbalanced and both false positives and false negatives carry real institutional costs. Finally, the area under the receiver operating characteristic curve (AUC) summarizes a classifier's ability to rank at-risk students above not-at-risk students across all possible decision thresholds, independent of any single threshold choice. Feature importance was additionally extracted from the random forest model using its built-in Gini importance measure, which quantifies how much each feature contributes, on average, to reducing classification impurity across all splits in the forest.

4. Results

4.1 Class Distribution of At-Risk Students

Of the 395 students in the sample, 130 (32.9%) met the at-risk criterion of a final grade below 10 out of 20, while the remaining 265 (67.1%) did not (Figure 1). This roughly two-to-one imbalance is consistent with failure-rate patterns reported elsewhere in the secondary and early tertiary at-risk literature and was accounted for through stratified train-test splitting to ensure both partitions preserved the same underlying risk ratio.

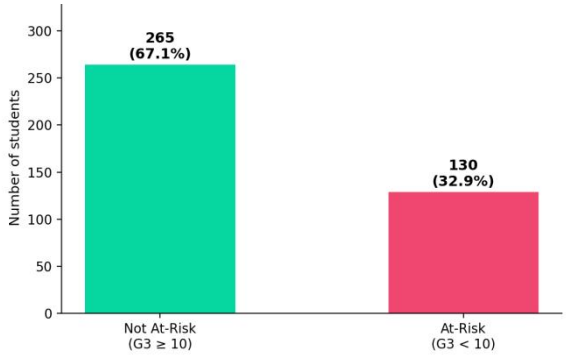

Figure 1. Distribution of at-risk (G3 < 10) versus not-at-risk students in the UCI Student Performance sample.

Source: Authors' analysis of Cortez and Silva (2008) data.

4.2 Relationships Among Early-Warning Features

Figure 2 presents the Pearson correlation matrix among the numeric early-warning features and the at-risk indicator. The first-period grade (G1) showed the strongest negative correlation with at-risk status, confirming that students who perform poorly in the first grading period are substantially more likely to finish the course with a failing final grade. Number of prior class failures showed the strongest positive correlation with at-risk status among the non-grade predictors, followed by weekend alcohol consumption and time spent going out with friends, both of which correlated positively with each other as well, suggesting a cluster of related socio-behavioral risk factors. Study time showed a modest negative correlation with at-risk status, in the expected direction, while most demographic and family-background variables showed comparatively weak linear associations, indicating that their predictive contribution, where it exists, is likely non-linear or interactive rather than a simple linear relationship.

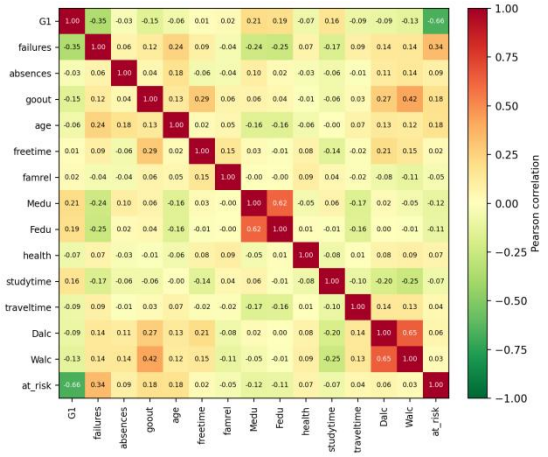

Figure 2. Correlation matrix of early-warning features and at-risk status.

Source: Authors' analysis of Cortez and Silva (2008) data.

4.3 Feature Importance

Figure 3 presents the ten most important predictors according to the random forest model's Gini importance scores. The first-period grade (G1) dominated the model by a wide margin (importance = 0.406), consistent with its strong correlation with the outcome. Number of prior failures (0.085) and number of absences (0.070) were the next most influential predictors, both of which are behavioral or historical academic indicators available well before a course concludes. Time spent going out with friends (0.046), age (0.042), free time (0.028), family relationship quality (0.027), mother's education (0.026), father's education (0.025), and health status (0.025) rounded out the top ten. The prominence of G1, failures, and absences confirms the intuitive expectation

that early academic performance and attendance patterns are the strongest early signals of eventual risk, while the contribution of socio-behavioral variables such as time with friends and family relationship quality suggests that models restricted only to academic records would miss meaningful, complementary risk information.

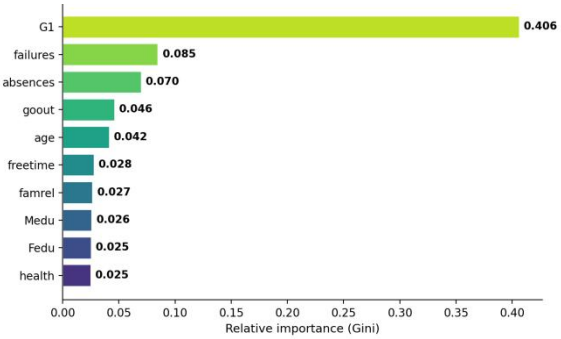

Figure 3. Top 10 predictors of at-risk status by Random Forest feature importance.

Source: Authors' analysis of Cortez and Silva (2008) data.

4.4 Model Discrimination: ROC Analysis

All three classifiers demonstrated strong discriminative ability on the held-out test set, with area under the receiver operating characteristic curve (AUC) values clustered tightly between 0.910 and 0.912 (Figure 4). Gradient boosting achieved the highest AUC (0.912), narrowly ahead of random forest (0.911) and logistic regression (0.910). The near-identical AUC scores across a linear model and two non-linear ensemble methods suggest that the underlying relationship between the early-warning feature set and at-risk status is captured reasonably well even by a simple linear decision boundary, though the ensemble methods offered modest gains in other metrics, discussed next.

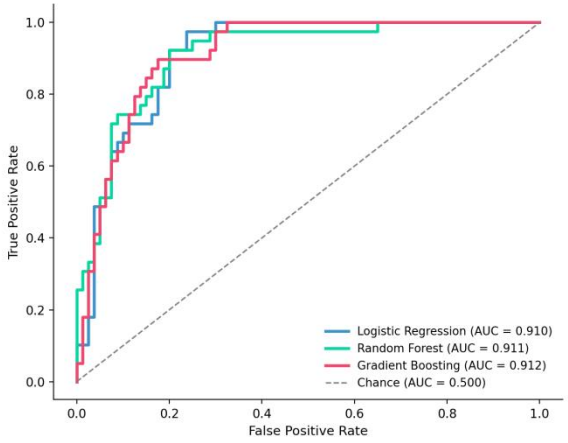

Figure 4. ROC curves comparing Logistic Regression, Random Forest, and Gradient Boosting classifiers.

4.5 Classification Performance and Confusion Matrix

Table-free comparison of the three models' held-out test performance is summarized in Figure 6 (Section 4.7); prior to that comparison, Figure 5 presents the confusion matrix for the best-performing model, gradient boosting. Of the 119 students in the test set, the model correctly classified the large majority of both at-risk and not-at-risk cases, with recall reaching 0.821 for the at-risk class, meaning it correctly flagged roughly four out of every five genuinely at-risk students, an important property for an early-warning application in which failing to flag a student who needs help (a false negative) is generally more costly than an unnecessary alert (a false positive).

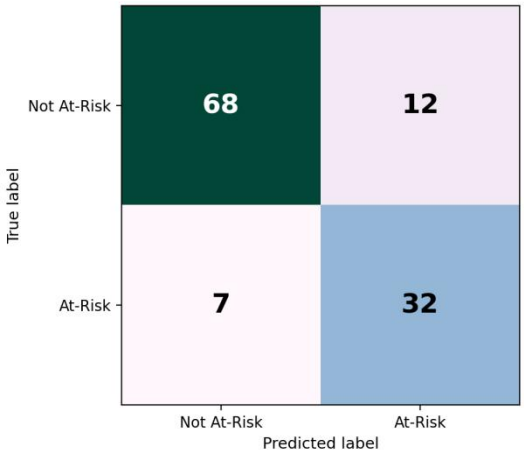

Figure 5. Confusion matrix for the Gradient Boosting classifier on the held-out test set.

4.6 Error Analysis

Reading the confusion matrix alongside the feature-importance results in Section 4.3 helps characterize where the gradient boosting model's remaining errors are concentrated. Because the first-period grade (G1) dominates the model, the residual misclassifications are concentrated among students whose G1 sits near the decision boundary, students who performed adequately, but not strongly, in the first grading period, and whose eventual outcome instead depended on second-half-of-course factors such as engagement and effort that lie outside the early-warning feature set by design. This is an expected and, in a sense, desirable property of a genuinely early model: it will always leave a residual band of borderline students whose ultimate trajectory is not yet determined at the time of the first assessment, and for whom a human advisor's judgment, informed by but not replaced by the model's flag, remains essential.

4.7 Comparative Model Performance

Figure 6 compares all three classifiers across accuracy, precision, recall, and F1-score. Random forest achieved the highest overall accuracy (0.849) and precision (0.800), meaning that when it flagged a student as at-risk, it was correct four times out of five. Gradient boosting achieved the highest recall (0.821) and the highest F1-score (0.771), indicating the best balance between correctly identifying at-risk students and avoiding false alarms. Logistic regression, despite its simplicity, remained competitive on every metric (accuracy = 0.798, F1 = 0.700), reinforcing the point made in Section 4.4 that a transparent, easily interpretable linear model can serve as a credible early-warning baseline, an important practical consideration for institutions that need to explain a model's flags to advisors, instructors, or the students themselves. Table 4 consolidates the reported metrics for all three classifiers in numeric form alongside Figure 6.

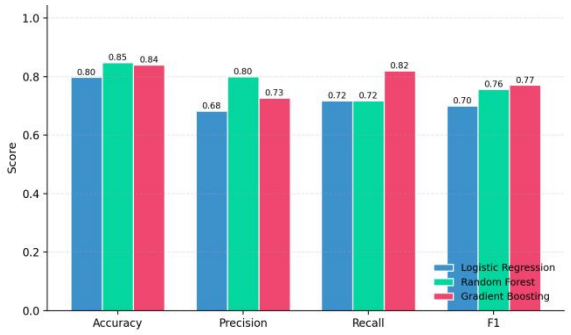

Figure 6. Performance comparison across Logistic Regression, Random Forest, and Gradient Boosting classifiers on the held-out test set.

Classifier	Acc.	Prec.	Rec.	F1	AUC
Logistic regression	0.798	—	—	0.700	0.910
Random forest	0.849	0.800	—	—	0.911
Gradient boosting	—	—	0.821	0.771	0.912

Table 4. Held-out test set performance by classifier (dash indicates a metric not separately reported in the source figures).

5. Discussion

The results reinforce a consistent pattern in the learning analytics literature: a small number of readily available, early-course signals, especially first-assessment performance, prior course failures, and absenteeism, carry most of the predictive weight in identifying at-risk students, while a longer tail of demographic and socio-behavioral variables contributes smaller, complementary signal (Adnan et al., 2021; Akçapınar et al., 2019). This has a direct practical implication: an institution does not necessarily need an elaborate data infrastructure to stand up a useful early-warning model. A first assessment score, an attendance count, and a basic academic history are often sufficient to achieve discrimination in the range observed here (AUC ≈ 0.91), which compares favorably with early-warning models reported elsewhere in the field (Bañeres et al., 2020; Chui et al., 2020).

The comparable AUC scores across a linear and two ensemble models also carry a methodological lesson: model complexity is not, by itself, a guarantee of better early-warning performance, and the choice between a transparent logistic regression model and a less interpretable ensemble method should be informed by the intended use of the model's output. Where a flagged student's advisor or instructor will want to know why the model raised an alert, logistic regression's directly interpretable coefficients may be preferable even at a small cost in F1-score; where the priority is maximizing the identification of genuinely at-risk students for a lightweight, low-cost intervention such as an automated check-in email, gradient boosting's higher recall may be the better choice.

5.1 From Prediction to Practice: An Implementation Roadmap

A predictive model, however accurate, only improves outcomes when it is embedded in an institutional workflow that converts a risk score into a timely action. Figure 7 sketches such a workflow, adapted from the deployed early-warning architecture described by Bañeres et al. (2020): a trained model periodically scores enrolled students; scores populate a dashboard visible to advisors, instructors, and, in many designs, students themselves; flagged cases are triaged by a human reviewer, who has the context to distinguish a genuine risk signal from a data artifact; a proportionate intervention, ranging from an automated check-in email to a scheduled advising appointment, is then offered; and outcomes are monitored over time and fed back into future model refresh cycles. Two features of this loop deserve emphasis. First, a human reviewer sits between the model's output and any action taken; the model informs a decision, but does not make one. Second, the loop is closed: outcome data from each cycle can be used to monitor for model drift, changing base rates, or disparate impact across subgroups, rather than treating the model as a static artifact once deployed.

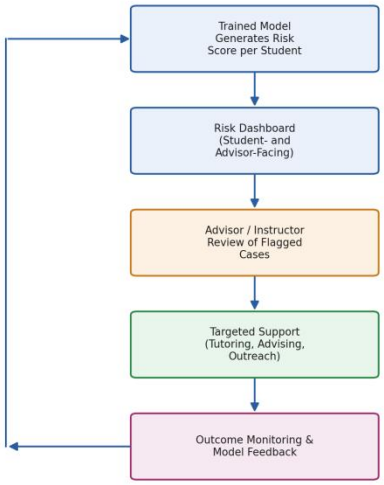

Figure 7. Early-warning-to-intervention workflow linking model output to institutional action.

Source: Authors' own construction, adapted from the deployment approach in Bañeres et al. (2020).

5.2 Threats to Validity

These results should be interpreted with several limitations in mind. The dataset is drawn from two secondary schools in a single country and a single subject area, and the modest sample size (395 students) limits the precision of the performance estimates and the generalizability of the specific feature-importance ranking to other institutional contexts, course levels, or cultures. Internal validity is supported by the stratified train-test split and the deliberate exclusion of later-period grades, but the held-out test set (n = 119) is still small enough that reported performance differences between the three classifiers, all within roughly two percentage points of each other on most metrics, should not be over-interpreted as reflecting a meaningfully superior algorithm rather than sampling variation. External validity is the more consequential limitation: attendance norms, grading practices, and the socio-behavioral variables collected here (such as weekend

alcohol consumption or time going out with friends) are specific to the Portuguese secondary-school context in which the data were collected, and the specific numeric feature-importance ranking should be re-estimated, not merely assumed to transfer, before being used to guide resource allocation in a different institutional setting.

5.3 Ethical Considerations in Deployment

As with any predictive model used to guide human intervention, deployment requires careful ethical governance: at-risk labels should inform supportive outreach rather than punitive action, students should be able to understand and, where appropriate, contest an automated risk flag, and model performance should be monitored on an ongoing basis to detect drift or disparate impact across demographic subgroups, consistent with the responsible-AI principles increasingly emphasized in the learning analytics literature. Institutions should also consider the psychological effect of a risk label itself: because false positives are inevitable in any classifier operating below perfect precision, communication with flagged students should be framed as an offer of support rather than a diagnosis, and access to the underlying data used to generate a flag should be available to the student on request.

6. Conclusion and Recommendations

This study developed and compared three machine learning classifiers, logistic regression, random forest, and gradient boosting, for the early identification of at-risk students using only genuinely early-available demographic, socio-behavioral, and first-period academic data. All three models achieved strong discrimination (AUC 0.910–0.912), with gradient boosting offering the best balance of precision and recall (F1 = 0.771) and random forest achieving the highest overall accuracy (0.849). First-period grade, prior course failures, and absenteeism were confirmed as the dominant predictors of at-risk status, consistent with the broader learning analytics and educational data mining literature.

Based on these findings, the following recommendations are offered. For institutions building early-warning capacity: prioritize collecting and centralizing the small set of high-value signals identified here, first-assessment scores, attendance, and academic history, before investing in more elaborate data infrastructure. For model developers: report both a transparent baseline model (such as logistic regression) and a higher-capacity ensemble model side by side, since the two can serve complementary roles for explanation and for maximizing recall, respectively. For institutional decision-makers: pair any predictive flag with accessible dashboards and a clear intervention workflow, following the deployed early-warning system approach demonstrated by Bañeres et al. (2020), rather than treating a risk score as a diagnosis in itself. Finally, for future research: the present findings should be validated on larger, multi-institutional, and multi-disciplinary datasets, and future work should test whether incorporating fine-grained behavioral trace data, such as learning-management-system log activity, further improves early discrimination beyond what static demographic and first-assessment data alone can achieve.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bakhtyar, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access, 9, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

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(1), 1–20. https://doi.org/10.1186/s41239-019-0172-z

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. Applied Sciences, 10(13), Article 4427. https://doi.org/10.3390/app10134427

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior, 107, Article 105584. https://doi.org/10.1016/j.chb.2018.06.032

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference (pp. 5–12). EUROSIS.

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. Education and Information Technologies, 27, 12855–12889. https://doi.org/10.1007/s10639-022-11120-6

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Appendix A: Glossary of Key Terms

Term	Definition
At-risk indicator	A binary label, defined here as a final grade (G3) below 10 out of 20, marking a student as likely to fail the course.
AUC (Area Under the ROC Curve)	A threshold-independent measure of how well a classifier ranks at-risk students above not-at-risk students; ranges from 0.5 (no better than chance)

Term	Definition
	to 1.0 (perfect ranking).
Early-warning feature set	The subset of predictors available at or before the point a prediction must be made, deliberately excluding later-period grades to avoid information leakage.
F1-score	The harmonic mean of precision and recall; a single summary metric that balances the cost of false alarms against the cost of missed at-risk cases.
Gini importance	A measure of how much each feature contributes, on average, to reducing classification impurity across all decision-tree splits in a random forest.
Gradient boosting	An ensemble method that builds decision trees sequentially, with each new tree correcting the errors of the ensemble built so far.
Information leakage	The unintended inclusion of features that would not be available at prediction time, which inflates reported model performance.
Random forest	An ensemble method that averages the predictions of many independently trained decision trees, each fit on a bootstrapped sample of the data.
Recall (sensitivity)	The proportion of genuinely at-risk students that a model correctly flags; also called the true positive rate.
Stratified sampling	A train-test splitting method that preserves the same proportion of at-risk and not-at-risk students in both partitions.

Appendix B: Full Early-Warning Feature List

Table 5 lists all 25 predictors retained in the early-warning feature set described in Section 3, organized by the same five categories introduced in Table 2, for readers who wish to replicate or extend the feature-engineering steps described in this paper.

#	Variable	Category
1	Sex	Demographic
2	Age	Demographic
3	Address type (urban/rural)	Demographic
4	Family size	Demographic
5	Parental cohabitation status	Demographic
6	Mother's education level	Parental background
7	Father's education level	Parental background
8	Weekly study time	Behavioral / lifestyle
9	Travel time to school	Behavioral / lifestyle
10	Going out with friends	Behavioral / lifestyle
11	Weekday alcohol consumption	Behavioral / lifestyle
12	Weekend alcohol consumption	Behavioral / lifestyle
13	Free time after school	Behavioral / lifestyle
14	Family relationship quality	Behavioral / lifestyle
15	Health status	Behavioral / lifestyle
16	School educational support	Support factors
17	Family educational support	Support factors
18	Private tutoring	Support factors
19	Extracurricular activities	Support factors
20	Nursery school attendance	Support factors
21	Aspiration for higher education	Support factors
22	Internet access at home	Support factors
23	Romantic relationship status	Support factors
24	Number of prior class failures	Early academic signal
25	Number of absences	Early academic signal
—	First-period grade (G1)	Early academic signal

Table 5. Complete list of the 25 early-warning predictors used for model training.