

Early Identification of At-Risk Students Using Attendance and Quiz Data: An Explainable Learning Analytics Model for Low-Infrastructure Philippine Higher Education Institutions

Joshua C. Cervantes

2025-00739-CM-0 · Section MSIT_A
Master of Science in Information Technology
Polytechnic University of the Philippines, Open University System / CCIS
Manila, Philippines
joshuacruz cervantes@gmail.com

ABSTRACT

Students who fail a course usually show warning signs weeks before the final grade: missed classes, skipped quizzes and falling scores. Instructors in many Philippine higher education institutions (HEIs) see these signs too late because the records sit in separate attendance sheets, gradebooks and learning management systems (LMS). Most learning analytics research predicts at-risk students from rich LMS clickstream data collected in online or well-resourced settings, which does not match blended, low-infrastructure classrooms where LMS use is uneven. This proposal describes a quantitative, predictive study that will build and evaluate an explainable early warning model using only the two data sources such classrooms reliably have: attendance check-ins and quiz records. Logistic regression, decision tree, Random Forest and gradient boosting classifiers will be compared at three checkpoints in the term, with recall as the primary metric and SHAP values used to explain each prediction. The expected output is a validated model and a weekly, printable risk list that tells the instructor which students to contact and why.

CCS CONCEPTS

- Applied computing → Education; E-learning
- Computing methodologies → Machine learning; Supervised learning by classification

KEYWORDS

learning analytics; early warning system; at-risk students; explainable AI; attendance; higher education; Philippines

ACM Reference Format:

Joshua C. Cervantes. 2026. Early Identification of At-Risk Students Using Attendance and Quiz Data: An Explainable Learning Analytics Model for Low-Infrastructure Philippine Higher Education Institutions. MSIT 632 Technology and Project Management, Individual Research Proposal. Polytechnic University of the Philippines, Open University System, Manila, Philippines, 7 pages.

I. INTRODUCTION

A. Background of the Study

Learning analytics is the measurement, collection and analysis of data about learners in order to understand and improve learning. One of its most studied applications is the early identification of students at risk of failing, so that instructors can intervene while there is still time to change the outcome. Systematic reviews show that this area has matured quickly: classification models, especially tree-based ensembles, predict student performance with useful accuracy, and assessment grades are consistently among the strongest predictors [2].

Recent work has pushed the prediction earlier in the term. Santos and Henriques showed that useful predictions can be made from LMS logs when only 25% of a course has elapsed, using a small feature set that works across different courses [8]. Marcolino et al. found that access patterns and consecutive days without platform access were among the strongest signals of dropout [6]. Dai et al. went a step further and showed that a model trained on one term can flag students in the next, and that a well-designed feedback message leads many flagged students to re-engage [4].

These results come almost entirely from settings where students interact heavily with an LMS: fully online courses, massive open online courses and universities with stable connectivity. Many Philippine HEIs operate differently. Classes are mostly face-to-face or blended, laboratories and internet access are not always reliable, and LMS use varies widely from one instructor to the next. In such settings the richest clickstream features in the literature simply do not exist. What does exist in almost every class is attendance, often captured through a paper list or a simple online check-in form, and quiz results, whether taken on paper or on an LMS.

This study asks whether those two modest data sources are enough to identify at-risk students early and reliably, and whether the resulting predictions can be explained in terms an instructor can act on.

The question matters because failure and attrition carry real costs in the Philippine setting. A failed course adds a semester of tuition and delays graduation, and students who stop attending often do so quietly, without formally dropping. Instructors who handle several sections across more than one institution, a common arrangement for part-time and newly hired faculty, may teach well over a hundred students in a term. At that scale, noticing a single student's slide from the records alone is unrealistic without some form of automated support.

B. Statement of the Problem

Instructors in blended, low-infrastructure HEIs lack a practical way to identify at-risk students early in the term. Warning signs are scattered across attendance records, gradebooks and LMS reports, and they are usually reviewed only when grades are encoded at the end of a grading period. By then the student may already have missed too much to recover.

Specifically, this study seeks to answer the following questions:

1. Which attendance and quiz features are most strongly associated with failing a course in blended, low-infrastructure HEI classes?
2. How accurately can logistic regression, decision tree, Random Forest and gradient boosting classifiers identify at-risk students at week 3, at the end of the prelim period and at the end of the midterm period?
3. Which model offers the best balance of recall, precision and interpretability for use by instructors?
4. How can the model's predictions be explained and presented so that an instructor can decide on an intervention in a few minutes each week?

The study tests the following working hypothesis: at the end of the prelim period, the best machine learning model will identify at-risk students with higher recall than the simple rule instructors commonly use, which is to flag a student with three or more absences or a quiz average below the passing mark.

C. Objectives of the Study

The general objective is to develop and evaluate an explainable early warning model that identifies at-risk students from attendance and quiz data in blended, low-infrastructure HEI classes. The specific objectives are to:

1. Build an anonymized dataset of attendance and quiz features from one academic term of undergraduate IT courses.
2. Train and compare four classification models at three prediction checkpoints.
3. Evaluate the models using recall, precision, F1 score and area under the ROC curve (AUC), with recall for the at-risk class as the primary metric.
4. Explain individual and overall predictions using SHAP values.
5. Design a one-page weekly risk list, usable on paper or on screen, that states each flagged student's main risk factors and a suggested action.
6. Validate the selected model on a new term and assess the usability of the risk list with instructors.

D. Scope and Delimitation

The study covers undergraduate IT courses in two private HEIs over one term, so results may not generalize to other programs or regions. It uses only attendance and quiz data; factors such as motivation, finances and family situation are outside its scope, although they clearly affect performance. The model identifies risk; it does not measure the effect of interventions, which is left for future work.

E. Significance of the Study

The study is expected to produce (1) a list of the attendance and quiz features most associated with failure in blended, low-infrastructure classes; (2) a comparison of four models at three checkpoints; (3) a selected,

explainable model; and (4) a weekly risk list template.

For **instructors**, the output turns records they already keep into a short weekly action list. For **students**, it increases the chance that help arrives before a failing grade is final. For **institutions**, it offers an early warning approach that does not depend on expensive infrastructure. For **researchers**, it adds evidence from a setting that the learning analytics literature has largely overlooked.

Beyond the immediate study, the approach supports the broader goal of Philippine higher education to improve student retention and completion. Because the model uses only attendance and quiz data, it can be repeated in other programs and institutions with little adaptation, and the feature definitions in Table 3 can serve as a common starting point for comparison across schools.

II. REVIEW OF RELATED LITERATURE

A. Predicting Student Performance

Alwarthan, Aslam and Khan reviewed 45 studies on predicting student performance in higher education published in the five years up to April 2022 [2]. Classification was used in 87% of the studies, and Random Forest and ensemble models were the most accurate in many comparisons. University features such as assessment grades and cumulative grade point average were the most frequently used predictors. The review identified three gaps that this study addresses directly: most studies used small single-institution samples, few predicted performance early enough for intervention, and many models lacked interpretability.

Deep learning has also been applied to LMS activity. Al-Sulami, Al-Masre and Al-Malki compared recurrent neural network, LSTM and GRU models on 3,529 students in a foundation-year course and reported accuracy of up to 98.90% with the GRU model, with all three models reaching about 81% recall for at-risk students [1]. Wang and Yu, working with 300 students in 15 online courses, found that only three of eleven behavioral indicators were strongly associated with performance, and that an enhanced logistic regression outperformed several baseline methods [10]. Together these studies suggest that a small, well-chosen feature set can perform well, and that simple models remain competitive.

B. Early Prediction and Early Warning Systems

The value of a prediction depends on its timing. Santos and Henriques used Moodle logs from 2,140 students across 138 courses and built datasets at five points in the term [8]. Random Forest reached an AUC of 0.756 for at-risk prediction with full-course data, and useful prediction was already possible at 25% of the course. Importantly, a common set of 14 features worked across courses of different lengths and grading schemes, which suggests that a single model can serve many subjects.

Marcolino et al. applied CatBoost with multi-objective hyperparameter optimization and the ADASYN oversampling method to 567 student-course instances from 23 courses [6]. A single model trained on data from all weeks reached about 0.80 F1, 0.85 recall and 0.95 AUC. Boudjehem and Lafifi proposed an algorithm that monitors learners' performance-based behavior to forecast those likely to struggle or drop out, framing the result as an operational early warning system rather than a model alone [3].

C. Explainability

An instructor cannot act on a risk score without knowing why a student was flagged. Ujkani, Minkovska and Hinov applied several machine learning and deep learning models to the Open University Learning Analytics Dataset of 32,593 students and used SHapley Additive exPlanations (SHAP) to show that engagement and registration timing were the most important factors [9]. Mustofa et al. combined logistic regression with a neural network on a public dataset of 4,424 students, handled class imbalance through upsampling, and used SHAP and LIME to explain the contributing factors [7]. Both studies show that explainability can be added to a predictive model without giving up accuracy.

D. From Prediction to Intervention

Dai et al. applied models trained on a previous course offering to a new semester, where they kept an AUC above 0.8 [4]. Flagged students received feedback emails written with a relational feedback framework, and more than 30% of them opened previously unused learning activities within two weeks. Most surveyed students wanted more frequent interventions. The lesson for this study is that the output must be a regular, concrete action for the instructor, not a report.

E. Privacy and Data Protection

Liu and Khalil reviewed 47 papers and identified eight privacy and data protection issues across the learning analytics cycle, noting a lack of applied evidence that proposed solutions work [5]. In the Philippines, the Data Privacy Act of 2012 (Republic Act No. 10173) governs the processing of personal information, including student records. Any early warning model must therefore limit data collection to what is necessary, anonymize records for analysis and restrict access to the risk list.

F. Synthesis and Research Gap

Table 1 summarizes the data used in the reviewed studies.

Table 1: Data sources in the reviewed studies

Study	Main data source	Setting
[1]	LMS activity	Single course, Saudi Arabia
[3]	E-learning traces	Online environment
[4]	Learning traces, grades	Single course, Australia
[6]	Moodle logs	23 courses, Brazil
[7]	Demographic, academic	Public dataset
[8]	Moodle logs	138 courses, Portugal
[9]	OULAD (clicks)	Online, United Kingdom
[10]	MOOC behavior logs	15 online courses

The pattern is consistent: the literature depends on LMS clickstream data from online or well-resourced institutions. No reviewed study models students in blended, low-infrastructure classrooms where attendance and quiz records are the most reliable signals. This study addresses that gap.

The reviewed studies also point to three design choices that this study adopts. First, predict early, since interventions after the midterm have little room to work [4, 8]. Second, prefer explainable models or add explanations, since instructors must justify why they contacted a student [7, 9]. Third, treat privacy as part of the design rather than an afterthought [5].

G. Conceptual Framework

The study follows an Input-Process-Output model, shown in Table 2.

Table 2: Conceptual framework (Input-Process-Output)

Input	Process	Output
Attendance check-ins	Anonymize and clean	Ranked risk features
Quiz scores and completion	Build features per checkpoint	Model comparison
Prelim examination scores	Train and compare four models	Explainable selected model
Final grades (labels)	Evaluate; explain with SHAP	Weekly risk list

The input consists of records that instructors already keep. The process transforms these into features, trains and evaluates the models, and explains their predictions. The output is both a research result, namely the features and model that best identify risk, and a practical tool for instructors.

H. Definition of Terms

At-risk student. A student whose final grade falls below the passing mark of 75, or who drops or stops attending the course.

Checkpoint. A fixed point in the term at which a prediction is made using only data available up to that point.

Early warning model. A predictive model that flags at-risk students while there is still time to intervene.

Low-infrastructure setting. A class where laboratory access, internet connectivity or LMS use cannot be relied on every session.

Recall. The proportion of truly at-risk students that the model correctly flags.

SHAP. SHapley Additive exPlanations, a method that assigns each feature a contribution to an individual prediction.

III. METHODOLOGY

A. Research Design

The study uses a quantitative, predictive research design and follows the Cross-Industry Standard Process for Data Mining (CRISP-DM): business understanding, data understanding, data preparation, modeling, evaluation and deployment.

B. Setting and Data

Data will come from undergraduate IT courses handled by the proponent in two private HEIs in Bulacan, Philippines, over one academic term. Both institutions run blended

classes where attendance is captured through a check-in form or paper list, and quizzes are given on paper or on an LMS. Data collection will begin only after written approval from each institution. All records will be anonymized before analysis by replacing student names and numbers with random codes; the mapping file will stay with the proponent and will not be shared.

Data will be collected in the following steps:

1. Secure written approval from the dean or program head of each institution, stating the data to be used and the purpose.
2. Inform students at the start of the term, through a short written notice, that anonymized attendance and quiz records will be used for research and will not affect their grades.
3. Export attendance check-ins, quiz scores and period grades at the end of each grading period.
4. Replace identifiers with random codes on export and store the key file offline.
5. Merge the exports into one dataset per course section, then into one pooled dataset.
6. Record final grades and drop or withdrawal status at the end of the term to create the labels.

C. Features and Target

Features will be computed at each checkpoint using only data available up to that point:

- **Attendance:** attendance rate, number of absences, longest run of consecutive absences, and number of late check-ins.
- **Quizzes:** quiz average, lowest quiz score, quiz completion rate, and the trend in scores between the first and latest quiz.
- **Assessment:** prelim examination score (for the second and third checkpoints only).

Table 3 defines each feature and its source.

Table 3: Candidate features

Feature	Definition	Source
att_rate	Sessions attended / sessions held	Attendance
absences	Count of missed sessions	Attendance
max_consec_abs	Longest run of missed sessions	Attendance
late_count	Check-ins after the cut-off time	Attendance
quiz_avg	Mean quiz score, percent	Quiz records
quiz_min	Lowest quiz score,	Quiz

	percent	records
quiz_completion	Quizzes taken / quizzes given	Quiz records
quiz_trend	Latest minus first quiz score	Quiz records
prelim_score	Prelim examination, percent	Gradebook

The target variable is binary. A student is labeled at-risk if the final grade is below the passing mark of 75, or if the student dropped or stopped attending the course.

Preprocessing rules will be fixed before modeling. Sessions declared as no-class days will be excluded from the attendance denominator. Students who enrolled late will have their attendance rate computed only from their enrollment date. A missed quiz will count as zero for the quiz average but will also be captured separately through quiz completion, so the model can distinguish a low score from a missing one. Quiz scores with different maximums will be converted to percentages. Every excluded record will be logged with a reason.

D. Checkpoints

Models will be trained and evaluated at three checkpoints: the end of week 3 (about 20% of the term), the end of the prelim period and the end of the midterm period. The first checkpoint tests how early a useful prediction is possible; the later checkpoints test how much accuracy improves as more data arrive.

E. Models and Evaluation

Four classifiers will be compared: logistic regression, decision tree, Random Forest and gradient boosting. Logistic regression and decision trees are included because instructors can read them directly; the ensembles are included because the literature reports them as the most accurate [2, 8]. Because failing students are a minority, class weighting and the SMOTE oversampling method will be applied to the training folds only. Models will be evaluated with stratified 5-fold cross-validation. Recall for the at-risk class is the primary metric, because missing a student who needs help costs more than contacting one who does not. Precision, F1 score and AUC will also be reported.

Hyperparameters will be tuned by grid search inside the cross-validation loop to avoid optimistic estimates. Every model will be compared against a rule-based baseline that mirrors current practice: flag a student with three or more absences or a quiz average below

75. A model is only worth deploying if it clearly beats this rule, and this comparison tests the working hypothesis. Differences in recall between the best model and the baseline will be tested with McNemar's test on the same students.

To check that the model generalizes over time, as Dai et al. did [4], the selected model will be trained on the first term and then applied, without retraining, to the first weeks of the following term. Its recall at the week-3 and prelim checkpoints on the new cohort will be reported separately from the cross-validation results.

F. Explainability

SHAP values will be computed for the selected model. Global SHAP summaries will show which features matter most across all students, answering research question 1. Local SHAP values will show the top two or three reasons each individual student was flagged, for example "four consecutive absences" or "two missed quizzes".

G. Statistical Treatment

Descriptive statistics (mean, standard deviation, minimum and maximum) will summarize each feature for passing and at-risk students separately. The association between each feature and the at-risk label will be measured with the point-biserial correlation for continuous features and the chi-square test for counts grouped into bands. These results, together with the global SHAP ranking, answer research question 1. Model metrics will be reported as the mean and standard deviation across the five folds, which answers research question 2. Research question 3 is answered by weighing recall, precision and the interpretability of each model, and research question 4 by the usability results described below.

H. Output: The Weekly Risk List

The deployment output is a one-page risk list generated each week. For each flagged student it shows the anonymized code or name (visible only to the instructor), the risk level, the main reasons from SHAP and a suggested action such as a check-in message, a consultation or a make-up quiz.

Weekly Risk List | IT 101 Section A | Week 3

Model: Random Forest | Flagged: 4 of 38 students

Code	Risk	Main reasons (SHAP)	Suggested action
S-014	High	4 consecutive absences; 2 missed quizzes	Message today; consult
S-031	High	Quiz average 58%; falling trend	Consult; make-up quiz
S-007	Medium	3 late check-ins; quiz average 71%	Check-in message
S-022	Medium	Missed Quiz 3; attendance 80%	Make-up quiz reminder

Figure 1: Sample weekly risk list (illustrative data)

The list is designed to work on paper as well as on screen, so it remains usable when a laboratory or internet connection is not available.

The usability of the risk list will be assessed with at least five instructors who will use it for two weeks and then complete the System Usability Scale (SUS), a ten-item questionnaire scored from 0 to 100, followed by short interviews on whether the listed reasons were clear and the suggested actions practical. A mean SUS score of 68 or higher, the commonly used benchmark for average usability, will be treated as acceptable.

I. Ethical Considerations

The study will comply with Republic Act No. 10173. Only attendance and quiz data will be collected; no demographic, financial or health information will be used. The model will support, not replace, the instructor's judgment: no grade, sanction or record will be based on a prediction. Risk lists will not be shown to other students or posted publicly.

IV. PROJECT PLAN

The study is managed as an IT project with defined phases, a schedule, resources and risks.

Table 4: Work breakdown and schedule

Phase	Main activities	Weeks
1 Initiation	Approvals, data request letters	1-2
2 Data collection	Export attendance and quiz records	3-16
3 Preparation	Anonymize, clean, build features	5-17
4 Modeling	Train and tune four models	17-19
5 Evaluation	Metrics, SHAP, model selection	19-20
6 Deployment	Risk list, next-term test, SUS survey	21-22

7 Closure	Final paper and presentation	23-24
-----------	------------------------------	-------

Resources. The study uses free and open-source tools: Python with pandas, scikit-learn, imbalanced-learn and SHAP, run in Google Colab, with records exported from Google Sheets and the LMS. Direct costs are limited to printing and binding, estimated at PHP 1,500.

Table 5: Key risks and responses

Risk	Response
Institution denies data access	Request early; fall back to one institution
Too few failing students	Pool sections and subjects; use class weighting
Missing or inconsistent records	Define cleaning rules in advance; log exclusions
Privacy breach	Anonymize on export; keep the key offline

REFERENCES

- [1] Amnah Al-Sulami, Miada Al-Masre, and Norah Al-Malki. 2023. Predicting At-Risk Students' Performance Based on LMS Activity using Deep Learning. *International Journal of Advanced Computer Science and Applications* 14, 6 (2023). <https://doi.org/10.14569/IJACSA.2023.01406129>
- [2] Sarah A. Alwarthan, Nida Aslam, and Irfan Ullah Khan. 2022. Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review. *Applied Computational Intelligence and Soft Computing* 2022 (2022), Article 8924028, 26 pages. <https://doi.org/10.1155/2022/8924028>
- [3] Rochdi Boudjehem and Yacine Lafifi. 2024. An early warning system to predict dropouts inside e-learning environments. *Education and Information Technologies* 29 (2024), 16365-16385. <https://doi.org/10.1007/s10639-024-12498-1>
- [4] Wei Dai, Jionghao Lin, Flora Ji-Yoon Jin, Yi-Shan Tsai, Namrata Srivastava, Pierre Le Bodic, Dragan Gašević, and Guanliang Chen. 2025. Learning Analytics for Early Identification of At-Risk Students and Feedback Intervention. *Journal of Learning Analytics* 12, 3 (2025), 102-125. <https://doi.org/10.18608/jla.2025.8735>
- [5] Qinyi Liu and Mohammad Khalil. 2023. Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology* 54, 6 (2023), 1715-1747. <https://doi.org/10.1111/bjet.13388>
- [6] Markson Rebelo Marcolino, Thiago Reis Porto, Tiago Thompsen Primo, Rafael Targino, Vinicius Ramos, Emanuel Marques Queiroga, Roberto Munoz, and Cristian Cechinel. 2025. Student dropout prediction through machine learning optimization: insights from moodle log data. *Scientific Reports* 15 (2025), Article 9840. <https://doi.org/10.1038/s41598-025-93918-1>
- [7] Sumaya Mustofa, Yousuf Rayhan Emon, Sajib Bin Mamun, Shabnur Anonna Akhy, and Md Taimur Ahad. 2025. A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence* 8 (2025), Article 100352. <https://doi.org/10.1016/j.caeai.2024.100352>
- [8] Ricardo Miguel Santos and Roberto Henriques. 2023. Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs. *Computers and Education: Artificial Intelligence* 5 (2023), Article 100175. <https://doi.org/10.1016/j.caeai.2023.100175>
- [9] Berat Ujkani, Daniela Minkovska, and Nikolay Hinov. 2024. Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence. *Electronics* 13, 21 (2024), Article 4157. <https://doi.org/10.3390/electronics13214157>
- [10] Jing Wang and Yun Yu. 2025. Machine learning approach to student performance prediction of online learning. *PLOS ONE* 20, 1 (2025), e0299018. <https://doi.org/10.1371/journal.pone.0299018>