International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-5, Issue-6, November-December 2026
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

SPARK: Student Performance Analysis and Risk Knowledge Based on Machine Learning for Early Academic Risk Warning

Muhammad Aiman bin Mat Yacob[1], Maged Nasser[1*]

Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
muhammad_22007254@utp.edu.my, maged.nasser@utp.edu.my

Received: 05 Oct 2026, Received in revised form: 02 Nov 2026, Accepted: 07 Nov 2026, Available online: 12 Nov 2026
©2026 The Author(s). Published by IJTLE. This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Abstract

Lecturers in large classes often identify struggling students late because monitoring relies on manual observation. This paper presents SPARK (Student Performance Analysis and Risk Knowledge), a machine learning early warning system that classifies students as High, Medium or Low academic risk from the grades available during a semester and delivers the results through an interactive Streamlit dashboard. Risk is defined from the final grade and estimated at two points: after the first period grade (Stage 1) and after the second (Stage 2). Six classifiers (Decision Tree, Random Forest, KNN, Logistic Regression, SVM and XGBoost) and a rule baseline were evaluated on 316 development students of the UCI Student Performance (Mathematics) data, a sealed hold-out set of 79 students, and 275 non-overlapping Portuguese-language students for cross-subject testing. The High-risk threshold was set to reach a recall of 0.90. The Random Forest, chosen by a rule fixed before testing, caught 24 of 26 High-risk hold-out students without labelling any as Low risk; on Portuguese students its recall fell to 0.83 and false alarms rose. A rule that applies the risk bands to the latest grade performed almost as well as every model, showing that the first two grades carry most of the usable signal. Thresholds did not transfer across subjects and should be calibrated per course.

Keywords*— *early warning system, student risk prediction, learning analytics, machine learning, class imbalance, dashboard.

I. INTRODUCTION

Universities record large volumes of academic data, such as attendance and assessment results, yet struggling students are still identified mainly through lecturers' manual observation. This does not scale to large classes and tends to surface problems too late for support to change the outcome. Machine learning can learn from historical records to estimate which students are likely to fail and which will do well, but accurate models alone rarely reach lecturers in a form they can act on.

Three connected problems motivate this work. First, institutions collect a growing volume of student data without a matching ability to act on it. Second, manual monitoring does not scale with class size and depends on each lecturer's intuition. Third, both problems delay early intervention, so support arrives when it is least able to influence results and retention.

This paper presents SPARK, an early warning system that classifies each student as High, Medium or Low academic risk from the grades available during a semester and presents the result through an interactive dashboard. The work had three objectives: (1) to design

a machine learning model that predicts academic performance, (2) to identify at-risk students from its predictions, and (3) to provide an interactive dashboard for the results.

The contributions are as follows. First, six classifiers are compared against a transparent rule baseline under a protocol that fixes the main-model selection rule before testing and keeps a sealed hold-out set and a cross-subject set. Second, the High-risk decision threshold is tied directly to a target recall of 0.90 at two prediction stages. Third, the same frozen models are delivered through a Streamlit dashboard with filtering, ranking and rule-based suggested actions. Fourth, the paper documents where the final system departs from the original proposal and why. Experiments use the public UCI Student Performance dataset [1], with grades as the model inputs and the final grade defining the risk category; using open data avoids the ethical and time costs of primary collection. Validation on university students, behavioural data, explanations of why a student is at risk, and production deployment are outside the scope of this study.

The rest of the paper is organised as follows. Section II reviews related work, Section III describes the method and the system, Section IV reports results, Section V discusses them against the objectives and the baseline, and Section VI concludes with recommendations and future work.

II. RELATED WORK

A. Predicting academic performance. Ensemble models achieve strong accuracy on academic datasets [2], although that evaluation used static, batch-processed data rather than a system intended for ongoing use. SVM and Decision Tree models have reached 93–95% accuracy on real student data [3], but were validated only in standalone scripts with no deployment or monitoring component. Cortez and Silva [1], whose dataset underpins this work, modelled secondary school achievement from period grades and demographic, social and school attributes. They found that predicting the final grade becomes markedly harder, but more useful, when earlier period grades are excluded, which supports treating the final outcome strictly as ground truth rather than an input. Across these studies, accuracy has been shown repeatedly, while turning it into a usable early warning tool has received far less attention.

B. Identifying at-risk students. Behavioural indicators such as attendance and participation correlate strongly with later performance [4], although that study used static statistical analysis rather than a predictive model. Hamsiah *et al.* [5] used minimal early data with SMOTE balancing and reported an F1 of 0.770 for CatBoost on the Open University dataset, showing the potential of early prediction and the importance of addressing class imbalance. El Jihaoui *et al.* [6] note that accuracy and AUC-ROC can mislead for a minority class and recommend recall, precision, precision–recall curves and explicit threshold rules for separating the classes. Together these studies imply that a bare prediction is not enough: the score must be linked to a risk category a lecturer can act on.

C. Dashboards. LearningViz [7] helps instructors find and close performance gaps and was judged effective in expert usability tests, but it has no predictive risk model. A Moodle dashboard plugin [8-11] helps lecturers track student progress but is descriptive, intended for general tracking rather than prediction. Graf *et al.* [12] observe that many dashboards act as visualisations with little effect on outcomes and advocate interactive features to raise engagement. None of these dashboards is driven by a predictive model, so instructors must draw their own conclusions about risk. Table I summarises all nine studies.

Table I Summary of related studies

Study	Method	Key finding	Limitation
Gul *et al.* [2]	Ensemble models on academic datasets	High accuracy	Static, batch data; no at-risk identification
Wiyono and Abidin [3]	SVM, Decision Tree on real student data	93–95% accuracy	Standalone scripts; no monitoring interface

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

Cortez and Silva [1]	Data mining on grades and background attributes	Prediction much harder without earlier grades	Secondary level; no risk labels or dashboard
Wang and Yu [4]	Eleven behavioural indicators	Attendance and participation strongly linked to performance	Static statistics; no ML system
Hamsiah *et al.* [5]	CatBoost with SMOTE on minimal data	F1 of 0.770	External dataset; no lecturer-facing layer
El Jihaoui *et al.* [6]	Review of thresholding and fairness practice	Accuracy and AUC can mislead; use recall and PR curves	Evaluation focus; no end-user system
Pei *et al.* [7]	Interactive gap-finding dashboard	Helps detect performance gaps	No predictive risk model
Santoso *et al.* [8]	Moodle lecturer dashboard	Better activity monitoring	Descriptive only
Graf *et al.* [12]	Static versus interactive dashboards	Interactivity improves engagement	Student-facing, not lecturer early intervention

D. Research gap and SPARK's response. Three gaps follow from the literature. Few works build an end-to-end prediction system; few provide an interactive lecturer-facing dashboard; and few translate predictions into explicit High, Medium and Low categories. SPARK addresses the first by training and comparing six models on the grades available during the semester, the second with a Streamlit dashboard that offers single-student scoring, cohort filtering and suggested actions, and the third by converting model output into risk categories with a threshold chosen to reach a recall of 0.90 for the High class, so that lecturers can intervene without missing at-risk students. Unlike most prior work, it also benchmarks the models against a simple baseline.

III. METHODOLOGY

A. System Architecture

SPARK has three layers (Fig. 1). The data layer holds the Mathematics and Portuguese files of the UCI dataset. The processing layer builds the risk label and the Stage 1 and Stage 2 inputs, trains and tunes the models, sets thresholds and exposes a scoring function. The presentation layer is a Streamlit dashboard that applies the frozen models to new grades. The layers correspond to the objectives: model training addresses prediction, the threshold and risk categories address identification of at-risk students, and the dashboard addresses visualisation. The work was done in three phases (data preparation, modelling and risk classification, dashboard) and in two passes: an exploratory pass on all 395 Mathematics students, and briefly on both subject files together, to decide which inputs, models and settings to keep, and a final pass with a fixed protocol, a sealed hold-out set and a cross-subject test. All results below come from the final pass unless stated.

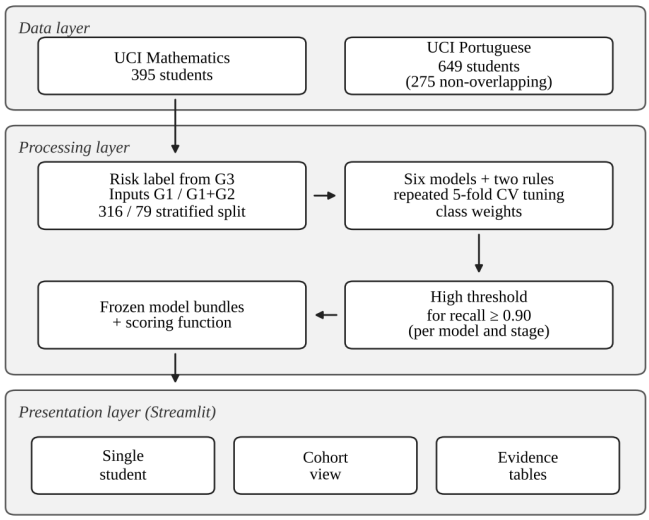


Fig. 1. SPARK architecture as implemented: data, processing and presentation layers.

B. Phase 1: Data Acquisition and Preprocessing

The originally planned Google Form survey was replaced by the public UCI dataset on the examiner's advice, because primary collection needed a defined target population, posed a risk to the timeline and meant collecting personal academic records. The two files were downloaded from the UCI repository as semicolon-separated tables. The Mathematics file had the expected 395 students and 33 attributes and no missing values or duplicate rows. Because the retained inputs are complete numeric grades on a fixed 0–20 scale, no

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

imputation or encoding was needed. Outliers were not removed, since grades are bounded and extreme values such as a final grade of zero are genuine outcomes, and scaling was learned inside each model pipeline from training data only.

Four preparation steps followed. (i) Each student was labelled from the final grade G_3: High (0–9, a failing grade since the pass mark is 10), Medium (10–13) or Low (14–20). (ii) Inputs were limited to G_1 alone (Stage 1) or G_1 and G_2 (Stage 2); G_3 was never an input, because the system must warn before the final result is known. (iii) The Mathematics file was split with stratification into 316 development and 79 hold-out students, and the hold-out set was stored unopened until the models were frozen. (iv) For the cross-subject set, Portuguese records matching Mathematics students on thirteen demographic attributes were removed, leaving 275 of 649 students. This phase gave a clean basis for training and testing in which the final outcome cannot leak into the inputs and test students are kept apart from training students.

C. Dataset

The dataset [1] was compiled from two Portuguese secondary schools and is widely used in the literature, which supports comparison with earlier work. It has two files, Mathematics (395 students) and Portuguese language (649), each with 33 demographic, social and school attributes, grades G_1, G_2 and G_3 on a 0–20 scale, and absences. The Mathematics file was used for training and testing and the Portuguese file for the cross-subject test. Table II gives the class distribution after the risk bands were applied.

Table II: Class distribution of the datasets

Dataset	Students	High	Medium	Low
Mathematics, all	395	130	165	100
Mathematics, development	316	104	132	80
Mathematics, hold-out	79	26	33	20
Portuguese, non-overlapping	275	70	141	64
Portuguese, full file	649	100	355	194

Two features of the data affected handling. Thirty-eight Mathematics students have $G_3 = 0$ and all fall in the High class; thirteen of them already had $G_2 = 0$, so the other twenty-five had to be predicted from earlier grades. All thirty-eight also have zero recorded absences, which probably means absences were not recorded for students who left, so absences was excluded as an input. These students were kept, because a student who leaves a course is exactly the kind of student an early warning system should flag. The main limitation of the dataset is that it describes secondary school students in Portugal, not university students in Malaysia, and offers only two informative inputs and a small number of students, which limits what any model can learn.

D. Phase 2: Model Development and Risk Classification

Six classifiers were developed: Decision Tree (DT), Random Forest (RF) and KNN, which were named in the proposal, plus Logistic Regression (LR) as a linear baseline and SVM and XGBoost (XGB) to cover the main method families in the literature. Each sits in a pipeline with a standard scaler, and class weights handle the mild class imbalance for models that support them (SMOTE gave no measurable gain in the exploratory pass and was dropped). Table III lists the settings searched. All models were tuned on the development set with the same repeated stratified five-fold cross-validation (three repeats), scored by average precision (AP) for the High class, and then re-scored on five new repeats with different folds because the tuning score is optimistic. Two non-learning baselines apply the risk bands directly to the latest grade (G_1 or G_2); they are called the *rule*.

Table III: Models and tuning ranges

Model	Settings searched
Decision Tree	Max depth 2–6; min leaf 3, 5, 10, 20; balanced class weights
Random Forest	300 trees; max depth 3, 5, 8, unlimited; min leaf 1, 3, 5, 10; balanced
KNN	Neighbours 5, 9, 15, 25, 35; uniform or distance weighting
Logistic Regression	C = 0.1, 0.3, 1, 3, 10, 30, 100; balanced
SVM	RBF kernel; C = 0.1, 1, 10, 100; gamma = scale, 0.01, 0.1, 1; balanced
XGBoost	Depth 2, 3, 4; 100 or 200 trees; learning rate 0.03 or 0.1

The main model was chosen by a rule written before any test set was opened: among the three proposed models, the one with the highest Stage 2 AP on the new

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

folds, with the Random Forest kept whenever it was within 0.01 of the best. SVM and XGBoost were added after the hold-out set was first scored, were tuned on development data only, and could not alter this choice. A paired bootstrap of development predictions estimated each model's gain over the rule.

The High threshold was set separately for each model and stage. Out-of-fold probabilities of High were collected over three repeats of five-fold cross-validation, and the threshold was the highest probability at which at least 90% of High students were flagged, which ties the operating point directly to the aim of not missing at-risk students. Probabilities were rounded to four decimals, because many students share identical grades and tiny numerical differences would otherwise move whole groups across the threshold. Each model was then refitted on all development students and stored with its thresholds. A student is classified High if P(High) reaches the threshold, otherwise Medium if P(Medium) ≥ P(Low), and Low if not. Stage 2 is used when both grades are available and Stage 1 when only G_1 is given; the scoring function rejects a missing G_1 and grades outside 0–20.

E. Phase 3: Dashboard and Early Intervention

The dashboard was built with Streamlit instead of the proposed Power BI report, because it had to score new grades immediately, which Power BI cannot do without extra setup. It loads the saved model bundles and uses the same scoring code as the experiments, so a student receives the same result in the dashboard as in the notebooks. It has three tabs. *Single student* takes G_1 and optionally G_2 and shows the risk label, class probabilities, stage, threshold and a note when a grade is next to a band boundary. *Cohort* accepts an uploaded file or a demonstration cohort, filters by risk level, stage, borderline status and background attributes, ranks flagged students by probability (ties broken by lower grades), exports the table and, when final grades are present, compares predictions with actual outcomes. *Evidence* shows the saved result tables.

Both the Single student and Cohort tabs show rule-based suggested actions for High and Medium students. They are chosen from the risk level, stage, G_1–G_2 change and borderline flag, for example arranging a meeting and catch-up plan for a High-risk student, or asking what changed when a grade has fallen. They are general guidance, not model output, were not tested for effectiveness, and the dashboard says so. The dashboard is decision support, not an automatic decision.

F. Evaluation Protocol

Development used repeated cross-validation on the 316 development students. The hold-out set and the Portuguese data were then scored once with the frozen models, with the 275-student version as the main cross-subject result and the full 649-student file as a secondary result. Because the Portuguese data were examined in the exploratory pass, this is a cross-subject evaluation rather than a blind test. The settings of the four original models were not changed after the first scoring of the test sets; SVM and XGBoost were added afterwards and tuned on development data only. Table IV lists the metrics and their priority, and recall is reported with 95% Wilson intervals. With only 26 High hold-out students, one student shifts recall by almost four points, so differences of a few points are not treated as meaningful and the test results are read as a check on the frozen system rather than a model ranking.

Table IV: Evaluation metrics and their priority

Metric	Purpose	Priority
High-class recall and precision	Share of failing students caught; share of flagged students truly High	Primary
High predicted as Low	Failing students given the lowest label, the costliest error for a warning tool	Primary
Confusion matrix	Every type of mistake among the three classes	Primary
Accuracy, macro precision, recall, F1	Overall performance across the three classes	Secondary
ROC-AUC (one-vs-rest), AP for High	Ability to rank students by risk, independent of threshold	Supporting

G. Tools and Reproducibility

The system was written in Python with pandas, NumPy, scikit-learn, XGBoost, matplotlib and seaborn, using Jupyter notebooks in Visual Studio Code and Streamlit for the dashboard. Shared functions sit in a single module used by both the notebooks and the dashboard, all random settings were fixed, and trained models were saved as files. Progress was recorded in Git, with tagged versions marking the end of the exploratory pass and the main stages of the final pass.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

H. Changes from the Proposed Methodology

Table V lists where the final implementation differs from the methodology proposed at the start of the project. The changes were made as the experiments showed what the data could support, and their effect on results is discussed in Sections IV and V.

Table V: Changes from the proposed methodology

Proposed	**Implemented**	**Reason**
Primary survey data	UCI Mathematics file; Portuguese file for cross-subject test	Examiner advice, time, avoiding personal records
Attendance, quiz, assignment and test features	G_1 and G_2	Only assessment measures with predictive value; other attributes added nothing
Risk probability cut at 0.3 and 0.7	Three grade bands; High threshold for recall 0.90	Ties threshold to the aim of catching at-risk students
SMOTE balancing	Class weights	Imbalance was mild; oversampling gave no gain
Composite risk score, wide feature engineering	Not used	Added no information beyond G_1 and G_2
Power BI dashboard	Streamlit dashboard	Needed to score new input and test the interface
Three models	Six models	LR, SVM and XGBoost added for comparison with the literature

IV. RESULTS

A. Exploratory Findings

On the 316 development students, four findings shaped the final design. (i) The first two grades carry almost all the information: a Random Forest on G_1 and G_2 alone reached High recall of 0.885, against 0.884 with all other attributes added, 0.737 without G_2 and 0.478 with no grades (Fig. 2). (ii) SMOTE, class weights and no balancing gave recalls of 0.862, 0.884 and 0.856, within one standard deviation of each other. (iii) Removing absences left the best result at 0.884. (iv) A depth-two Decision Tree split on G_2 at 9.5 and 13.5, the band edges, and its predictions were identical to the G_2 rule on all 395 Mathematics students.

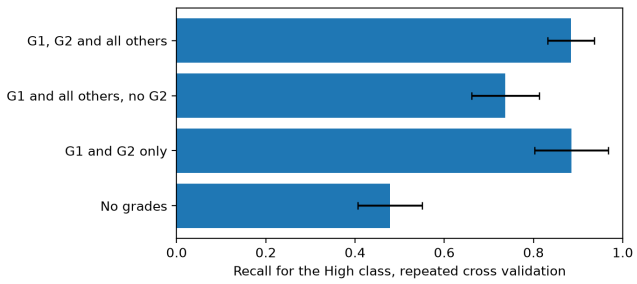


Fig. 2. Random Forest recall for the High class with different input sets (repeated cross-validation; error bars show variation across folds).

The exploratory pass also trained on both subject files together with a subject indicator. This gave a small but consistent gain over the rule (0.016 AP for Mathematics and 0.037 for Portuguese students), and the model weighted the first grade more heavily for Portuguese students. Because Portuguese data were examined in this step, and the final pass returned to the planned design of training on Mathematics only with a separate cross-subject test, the pooled design was not carried forward and is reported only as exploratory evidence.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

B. Development Results

Table VI gives development results on the 316 development students. Each model is scored by AP for the High class on five repeats of five-fold cross-validation with new fold assignments, and the gain columns show the improvement over the rule with a 95% bootstrap interval for the gap.

Table VI: Development results: AP for the High class and gain over the rule

Model	**S1 AP**	**S1 gain (95% CI)**	**S2 AP**	**S2 gain (95% CI)**
Rule (latest grade)	0.810	n/a	0.929	n/a
Decision Tree	0.798	−0.020 (−0.050, 0.011)	0.935	0.012 (0.002, 0.024)
Random Forest	**0.793**	**−0.025 (−0.054, 0.000)**	**0.949**	**0.020 (0.006, 0.037)**
KNN	0.797	−0.018 (−0.050, 0.009)	0.939	0.014 (0.004, 0.027)
Logistic Reg.	0.823	0.008 (−0.008, 0.022)	0.946	0.016 (0.007, 0.028)
SVM	0.715	−0.056 (−0.110, −0.005)	0.944	0.011 (0.003, 0.020)
XGBoost	0.797	−0.030 (−0.062, −0.001)	0.949	0.019 (0.005, 0.035)

S1/S2: Stage 1/Stage 2; five repeats of five-fold CV.

At Stage 2 every model gained 0.011–0.020 AP over the rule and all six intervals lie above zero, so the gain is consistent but small (Fig. 3). The Random Forest and XGBoost were highest (0.949), followed by Logistic Regression (0.946). At Stage 1 no model gained over the rule in a way distinguishable from zero: Logistic Regression gained 0.008, while the Random Forest, XGBoost and SVM fell slightly below the rule. With one input a model can do little beyond reordering students by that grade, which the rule already does. The SVM was weakest at Stage 1 (AP 0.715, precision 0.461 at 90% recall), most likely because its built-in probability estimates were unreliable with a single whole-number input.

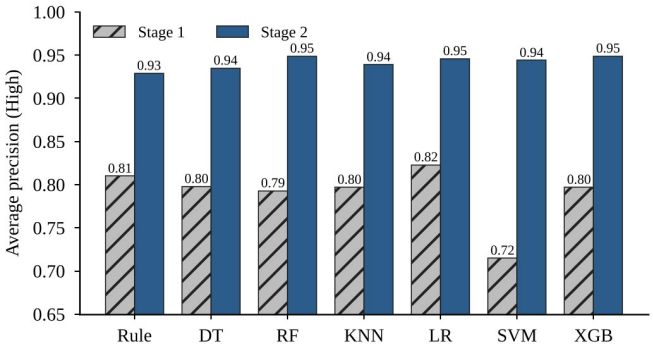


Fig. 3. Development AP for the High class by model and stage (axis starts at 0.65).

Setting the threshold for 0.90 recall gave the Random Forest thresholds of 0.496 (Stage 1) and 0.562 (Stage 2). On development data Stage 1 flagged 44.5% of students at precision 0.666 and Stage 2 flagged 35.0% at precision 0.849, with recall 0.901 and 0.904 (Fig. 4). The other three original models flagged 44.0–45.8% at Stage 1 and 35.1–36.2% at Stage 2, so the cost of an earlier warning was similar for all of them: about ten more students flagged per hundred, and about one in three flagged students not at risk, against about one in seven after the second grade. These figures are slightly optimistic because the thresholds were chosen on the same data.

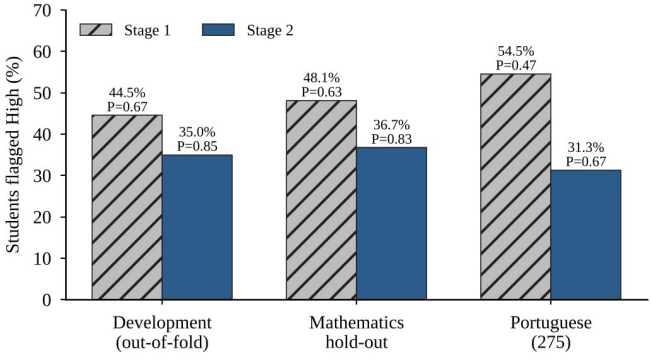


Fig. 4. Random Forest: share of students flagged High, with precision (P), on development, hold-out and Portuguese data.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

C. Mathematics Hold-Out Results

The frozen models were scored once on the 79 hold-out students (26 High). Tables VII and VIII give Stage 2 and Stage 1 results; each lists how many High students were caught, how many students were flagged, how many flags were false alarms, and recall with its 95% interval. No model classified any High student as Low risk.

Table VII: Mathematics hold-out, Stage 2 (26 of 79 High)

Approach	Caught	Flagged	False al.	Recall (95% CI)	Prec.
Rule (G_2)	24	30	6	0.923 (0.759–0.979)	0.800
Decision Tree	24	30	6	0.923 (0.759–0.979)	0.800
Random Forest	**24**	**29**	**5**	**0.923 (0.759–0.979)**	**0.828**
KNN	23	27	4	0.885 (0.710–0.960)	0.852
Logistic Reg.	23	27	4	0.885 (0.710–0.960)	0.852
SVM	24	30	6	0.923 (0.759–0.979)	0.800
XGBoost	24	30	6	0.923 (0.759–0.979)	0.800

Caught: High students flagged; Flagged: all students flagged High; False al.: false alarms; Prec.: precision for High; CI: Wilson interval. Bold: main model.

Table VIII: Mathematics hold-out, Stage 1 (26 of 79 High)

Approach	Caught	Flagged	False al.	Recall (95% CI)	Prec.
Rule (G_1)	21	28	7	0.808 (0.621–0.915)	0.750
Decision Tree	24	38	14	0.923 (0.759–0.979)	0.632
Random Forest	**24**	**38**	**14**	**0.923 (0.759–0.979)**	**0.632**
KNN	21	28	7	0.808 (0.621–0.915)	0.750
Logistic Reg.	21	28	7	0.808 (0.621–0.915)	0.750
SVM	24	38	14	0.923 (0.759–0.979)	0.632
XGBoost	24	38	14	0.923 (0.759–0.979)	0.632

At Stage 2 the Random Forest caught 24 of 26 High students (recall 0.923), flagged 29 and raised five false alarms; the two it missed were labelled Medium, not Low. The G_2 rule caught the same number and flagged one more student. Three of the six models matched the rule exactly, KNN and LR caught one student fewer, and accuracy, macro-F1 and ROC-AUC differed by no more than 0.013, 0.016 and 0.008 across models, which one or two students can change.

At Stage 1 the Random Forest again caught 24 of 26 but flagged 38 students, 14 of them false alarms, so precision fell from 0.828 to 0.632. The rule, KNN and LR flagged 28 and caught 21 (recall 0.808, below target). This split most likely reflects where the threshold fell on whole-number G_1 values, so that the RF, DT, SVM and XGBoost flagged one more group of students sharing the same grade, not a difference in how well the models learned. The confusion matrices (Fig. 5) show the cost of the earlier warning: 14 Medium students were flagged High at Stage 1 against 5 at Stage 2. The SVM at Stage 1 had accuracy 0.544 and macro-F1 0.421 although its High decisions matched the Random Forest's, which points to unreliable probabilities for the Medium and Low classes.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

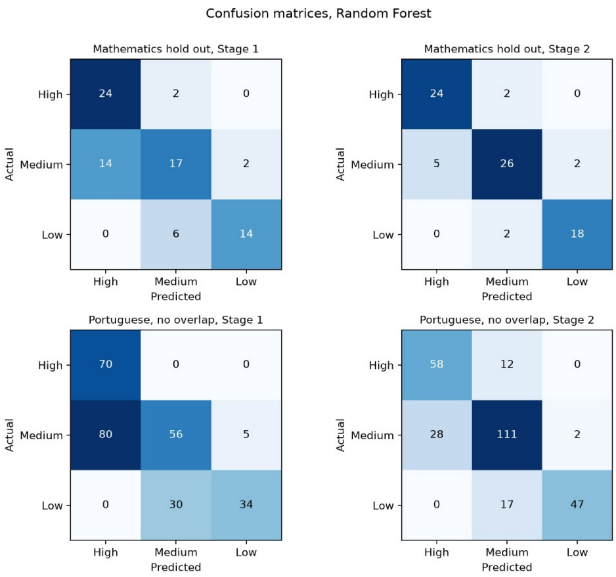


Fig. 5. Confusion matrices of the Random Forest on the Mathematics hold-out set and the non-overlapping Portuguese set, Stages 1 and 2.

D. Cross-Subject Results

The Mathematics models were scored without retuning on the 275 non-overlapping Portuguese students (70 High). Tables IX and X give the results. As on the hold-out set, no model labelled a High student as Low.

Table IX: Portuguese students, Stage 2 (70 of 275 High)

Approach	Caught	Flagged	False al.	Recall (95% CI)	Prec.
Rule (G_2)	63	94	31	0.900 (0.808–0.951)	0.670
Decision Tree	63	94	31	0.900 (0.808–0.951)	0.670
Random Forest	**58**	**86**	**28**	**0.829 (0.724–0.899)**	**0.674**
KNN	63	89	26	0.900 (0.808–0.951)	0.708
Logistic Reg.	63	89	26	0.900 (0.808–0.951)	0.708
SVM	63	94	31	0.900 (0.808–0.951)	0.670
XGBoost	60	100	40	0.857 (0.757–0.921)	0.600

Table X: Portuguese students, Stage 1 (70 of 275 High)

Approach	Caught	Flagged	False al.	Recall (95% CI)	Prec.
Rule (G_1)	63	105	42	0.900 (0.808–0.951)	0.600
Decision Tree	70	150	80	1.000 (0.948–1.000)	0.467
Random Forest	**70**	**150**	**80**	**1.000 (0.948–1.000)**	**0.467**
KNN	63	105	42	0.900 (0.808–0.951)	0.600
Logistic Reg.	63	105	42	0.900 (0.808–0.951)	0.600
SVM	70	150	80	1.000 (0.948–1.000)	0.467
XGBoost	70	150	80	1.000 (0.948–1.000)	0.467

At Stage 2, recall stayed at or near 0.90 for most models, but precision fell from 0.80–0.85 on the hold-out set to 0.60–0.71 (Fig. 6). About one in three flagged students was not at risk, for example 28 of the 86 flagged by the Random Forest. Among the proposed models the Random Forest fell furthest from target, catching 58 of 70 (recall 0.829), against 63 for the rule, and all twelve it missed were labelled Medium. XGBoost caught 60 and raised the most false alarms (40). LR and KNN matched the rule's recall with five fewer false alarms, and LR had the highest Portuguese ROC-AUC at both stages (0.950 and 0.886). The full 649-student file (100 High) gave the same picture: at Stage 2 the Random Forest caught 82, the rule 89, and LR and KNN 88 with 45 and 46 false alarms against 56 for the rule (Appendix Table A1).

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

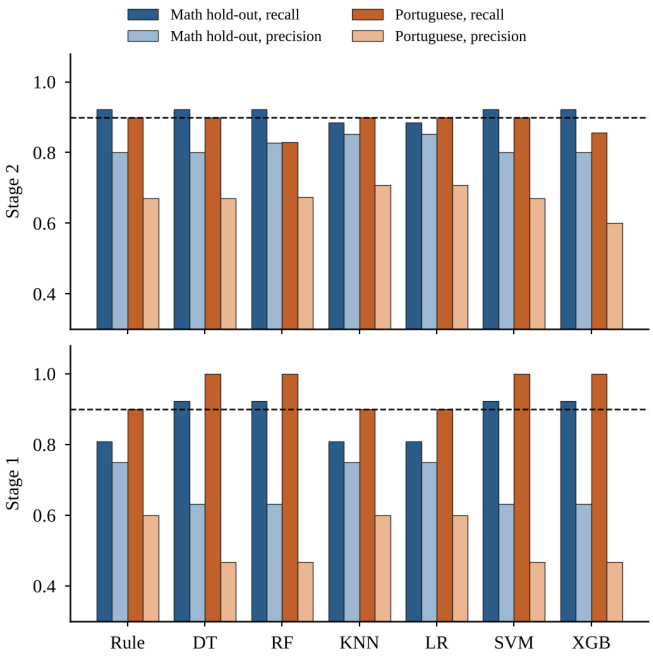


Fig. 6. High-class recall and precision on the Mathematics hold-out and Portuguese sets at Stage 2 (top) and Stage 1 (bottom). Dashed line: 0.90 recall target.

At Stage 1 the Mathematics thresholds did not transfer: the Random Forest, DT, SVM and XGBoost flagged 150 of 275 students (55%) to catch all 70 High students, giving recall 1.000 but precision 0.467 and 80 false alarms, while KNN, LR and the rule flagged 105 and caught 63. Flagging over half a cohort is not a useful warning, and the result shows that a threshold chosen on one subject should be recalibrated before use on another.

A likely cause is that Portuguese students tend to improve after the second period while Mathematics students slip slightly. The mean change from G_2 to G_3 was −0.30 for Mathematics and +0.34 and +0.41 for the full and non-overlapping Portuguese sets (Table XI). Among students with G_2 of 8 or 9, just below the pass mark, 29% of Mathematics students reached a final grade of 10 or more against 50% and 47% of Portuguese students (Fig. 7). A cut placed at the Mathematics boundary therefore flags many Portuguese students who go on to pass.

Table XI: Change from the second period grade to the final grade

Dataset	n	Mean change	G_2 of 8–9	Reached ≥10	Share
Mathematics, all	395	−0.299	82	24	0.293
Portuguese, full file	649	0.336	112	56	0.500
Portuguese, non-overlap	275	0.411	66	31	0.470

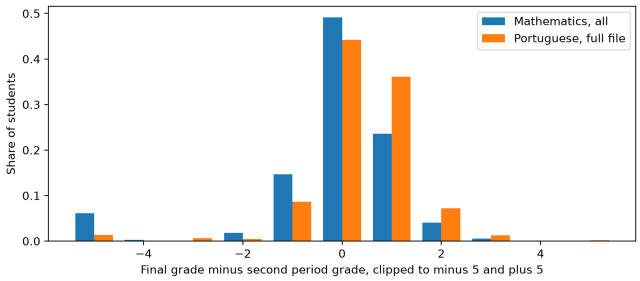


Fig. 7. Final grade minus second period grade, Mathematics and Portuguese (clipped to −5 and +5).

E. Dashboard

The dashboard was completed as a Streamlit application with the three tabs of Section III-E and a sidebar selector for any of the six frozen models, with the Random Forest as default. Fig. 8 shows two views. For a student with $G_1 = 10$ and $G_2 = 8$ the Single student tab returns High risk with the class probabilities, the threshold of 0.562 and four suggested actions, including one prompted by the fall between G_1 and G_2. On the demonstration cohort (the 275 Portuguese students, whom the models never saw in training) the Cohort tab at Stage 2 shows 86 students flagged High (31%) and 126 borderline.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1



Fig. 8. Dashboard: (a) Single student tab with a High-risk result and suggested actions; (b) Cohort tab with filters, summary counts and risk distribution for the demonstration cohort (Stage 2, Random Forest).

Two further views support checking and traceability (Fig. 9). When final grades are present, the Cohort tab compares predictions with actual outcomes and reports 58 of 70 High students caught, matching Table IX. The Evidence tab lists the saved result tables, including the hold-out and cross-subject results and the cross-validation results on fresh folds. Testing found one fault: a downloaded table could not be uploaded again because it contained the dashboard's own output columns; these are now discarded before scoring.

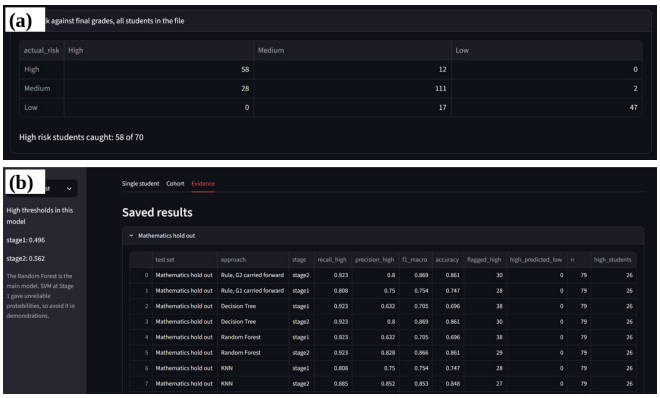


Fig. 9. Dashboard: (a) comparison of predictions with final grades for the demonstration cohort; (b) Evidence tab with saved results.

V. DISCUSSION

Objective 1: prediction. Six models were trained, tuned and compared, and all classified students into the three risk bands well above chance, with ROC-AUC of 0.96–0.97 at Stage 2 on the hold-out set. The Random Forest was the main model under the pre-set rule, having the highest Stage 2 AP among the three proposed models (0.949 against 0.939 for KNN and 0.935 for the Decision Tree). The models did not, however, separate clearly, and on the hold-out set differences of one or two students decided every comparison. The objective was met in that a working, evaluated model exists, but the evidence does not support calling one model better than another.

Objective 2: identifying at-risk students. Setting the High threshold for a recall of 0.90 on development data gave the Random Forest a hold-out recall of 0.923 at both stages with no failing student labelled Low risk, and the same absence of High-to-Low errors held for every model on Portuguese students. The cost was false alarms: about one in six flagged students after the second grade and more than one in three after the first on Mathematics students, and one in three after the second grade on Portuguese students. This is an acceptable trade for a tool whose purpose is not to miss students, but a flag should prompt a conversation, not a decision. *Objective 3* was met by the dashboard, which applies the same frozen models to new grades and turns the output into a ranked, filterable list with suggested actions.

The baseline. The most important finding is that a rule with nothing to learn performs almost as well as the trained models. On the hold-out set it caught the same number of High students as the Random Forest and was within one student of any model at Stage 2, and on development data the models gained only 0.011–0.020 AP, although consistently. This follows from the problem definition: the risk bands are cut on the same 0–20 scale as the grades, so a student's band after the second period is already a strong guess for the final band, as the depth-two tree showed directly. Models add a probability that ranks students within a band, a threshold tunable to a recall target and a small ranking gain, but little predictive information beyond G_2. In practical terms the first two grades are the signal, and none of the other attributes added anything measurable.

Stages and transfer. After the first grade the system caught as many High hold-out students but flagged nine more and precision fell to 0.63, and no model beat the

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

rule, because one input leaves little for a model to add. Across subjects the ordering of risk carried over but the operating point did not: recall stayed near 0.90 at Stage 2 for most models while precision fell, and at Stage 1 the Mathematics thresholds flagged over half the Portuguese students. The difference in how grades change after the second period is a likely cause, so thresholds should be set per subject, or per course in a university, before use. The Random Forest was less robust than the smoother models (58 of 70 High caught against 63 for LR, KNN and the rule), and LR had the best Portuguese ROC-AUC. Because the main model was fixed in advance it was not changed, and LR is noted as the most robust candidate if the work continues with a new hold-out set.

Relation to prior work. The studies in Section II reported high accuracy on different datasets and tasks, so their figures cannot be compared directly with these results. What this study adds is a comparison with a simple baseline, a sealed hold-out set, a threshold tied to a recall target, a second subject for testing and a dashboard that applies the same models to new input. It also shows why a baseline matters: an accuracy figure alone would have made the models look better than they are.

A. Limitations

The data describe secondary school students in Portugal, and the results cannot be assumed to hold for a university cohort. The hold-out set has 79 students (26 High) and the Portuguese test has 70 High students, so recall intervals are often 0.15–0.25 wide and differences between models are not reliable. Thresholds were chosen on development data and tested once, and recall on new students ranged from 0.808 to 1.000 around the 0.90 target. Many students share identical grades, so the models see only a few dozen distinct inputs and their probabilities come in jumps, which makes the threshold coarse.

The Portuguese data were seen in the exploratory pass, so the cross-subject test is not fully independent. SVM and XGBoost were added after the first look at the hold-out set, although tuned on development data only and unable to change the choice of main model. Two grades cannot explain why a student is at risk, and the suggested actions are general guidance not tested for effectiveness. Finally, the 38 Mathematics students with a final grade of zero, almost a tenth of the file, probably include students who left the course and are only partly predictable from earlier grades.

VI. CONCLUSION

SPARK classifies students as High, Medium or Low risk from the first two grades and presents the result through a Streamlit dashboard that scores single students and whole cohorts. All three objectives were met, with an important qualification about what the models add. With the High threshold set for 0.90 recall, the Random Forest caught 24 of 26 failing students in a sealed hold-out set with none labelled Low risk, and no model made a High-to-Low error on 275 Portuguese students. Three findings stand out: an early warning system built on two grades is useful, because it flagged nearly every failing student; a transparent baseline is essential, since the six models could not be separated from one another or clearly from a rule on the latest grade; and the operating point does not transfer between subjects, because Mathematics thresholds produced more false alarms on Portuguese students and, after the first grade, flagged more than half of them. These conclusions are limited by the small test sets, the use of secondary school data and the use of only two grades.

A. Recommendations

Three recommendations follow. First, treat a flag as a reason to talk to the student, not as a decision, and give extra attention to students whose grade sits next to a band boundary, where the label is least certain. Second, set thresholds per course on that course's own past data before use, because the Mathematics thresholds did not carry over. Third, keep the latest-grade rule as a baseline in later work, so that added complexity has to show a clear gain over it.

B. Future Work

Four directions would extend this work. The first is validation on university data from a defined year of study, the setting the system is intended for. The second is richer inputs: because the first two grades carry almost all the signal, a model can add value only if it receives information the grades lack, such as attendance, engagement with learning materials or coursework submission patterns, as in the Open University dataset used in one of the reviewed studies [5]. The third is a fairer model comparison on a new hold-out set with per-subject thresholds, with Logistic Regression as a candidate main model because it was the most robust across subjects. The fourth is testing the dashboard with lecturers, to learn whether the suggested actions are useful and whether flags lead to earlier help for students.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

REFERENCES

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Future Bus. Technol. Conf. (FUBUTEC)*, A. Brito and J. Teixeira, Eds. EUROSIS, 2008, pp. 5–12. [Online]. Available: https://archive.ics.uci.edu/dataset/320/student+performance

[2] M. N. Gul, W. Abbasi, M. Z. Babar *et al.*, "Data driven decisions in education using a comprehensive machine learning framework for student performance prediction," *Discover Comput.*, vol. 28, Art. no. 153, 2025, doi: 10.1007/s10791-025-09585-3.

[3] S. Wiyono and T. Abidin, "Comparative study of machine learning KNN, SVM, and Decision Tree algorithm to predict student's performance," *Int. J. Res. GRANTHAALAYAH*, vol. 7, no. 1, pp. 190–196, 2019, doi: 10.29121/granthaalayah.v7.i1.2019.1048.

[4] J. Wang and Y. Yu, "Machine learning approach to student performance prediction of online learning," *PLOS ONE*, vol. 20, no. 1, Art. no. e0299018, 2025, doi: 10.1371/journal.pone.0299018.

[5] Hamsiah, N. Adiyati, and R. Subekti, "Early prediction of at risk students using minimal data: A machine learning framework for higher education," *Digitus: J. Comput. Sci. Appl.*, vol. 3, no. 2, 2025, doi: 10.61978/digitus.v3i2.953.

[6] M. El Jihaoui, O. E. K. Abra, and K. Mansouri, "Towards a fair and transparent early warning system for identifying students at risk of academic failure," *Front. Educ.*, vol. 11, 2026, doi: 10.3389/feduc.2026.1878686.

[7] B. Pei, Y. Cheng, A. Ambrose, E. Dziadula, W. Xing, and J. Lu, "LearningViz: A dashboard for visualizing, analyzing and closing learning performance gaps, a case study approach," *Smart Learn. Environ.*, vol. 11, Art. no. 56, 2024, doi: 10.1186/s40561-024-00346-1.

[8] H. B. Santoso, A. K. Batuparan, R. Y. K. Isal, and W. H. Goodridge, "The development of a learning dashboard for lecturers: A case study on a student-centered e-learning environment," *J. Educators Online*, vol. 15, no. 1, 2018. [Online]. Available: https://eric.ed.gov/?id=EJ1168957

[9] Danyaro, K. U., Nasser, M., Zakari, A., Abdullahi, S., Khanzada, A., Yakubu, M. M., & Shoaib, S. (2025). LLM-based code generation: a systematic literature review with technical and demographic insights. IEEE Access, 13, 194915-194939.

[10] Nasser, M. (2024). Personalized learning through AI: enhancing student engagement and teacher effectiveness. International Journal of Teaching, Learning and Education, 3(6), 23-26.

[11] Nasser, M., Yusof, U. K., & Salim, N. (2023). Deep learning based methods for molecular similarity searching: a systematic review. Processes, 11(5), 1340.

[12] L. Graf, P. Bassner, M. Anzinger, F. Dietrich, S. Krusche, and O. Poquet, "Interactive learning dashboards: Rethinking learning visualisations as engagement tools," *Educ. Inf. Technol.*, 2026, doi: 10.1007/s10639-026-14082-1.

Appendix A: Supplementary Tables

Table A1: Full Portuguese file, High-class detection (100 High of 649)

Approach	S1 caught	S1 flagged	S1 FA	S2 caught	S2 flagged	S2 FA
Rule (latest grade)	88	157	69	89	145	56
Decision Tree	98	252	154	89	145	56
Random Forest	**98**	**252**	**154**	**82**	**131**	**49**
KNN	88	157	69	88	134	46
Logistic Reg.	88	157	69	88	133	45
SVM	98	252	154	89	145	56
XGBoost	98	252	154	85	156	71

S1/S2: Stage 1/2; FA: false alarms.

Table A2: Secondary metrics, Mathematics hold-out

Approach	S1 Acc.	S1 mF1	S1 AUC	S2 Acc.	S2 mF1	S2 AUC
Rule (latest grade)	.747	.754	n/a	.861	.869	n/a
Decision Tree	.696	.705	.890	.861	.869	.959
Random Forest	**.696**	**.705**	**.894**	**.861**	**.866**	**.962**
KNN	.747	.754	.889	.848	.853	.964
Logistic Reg.	.747	.754	.898	.861	.866	.966
SVM	.544	.421	.684	.861	.869	.962
XGBoost	.696	.705	.890	.861	.869	.967

Table A3: Secondary metrics, Portuguese students (275)

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1

Approach	S1 Acc.	S1 mF1	S1 AUC	S2 Acc.	S2 mF1	S2 AUC
Rule (latest grade)	.695	.690	n/a	.793	.797	n/a
Decision Tree	.582	.597	.859	.793	.797	.920
Random Forest	**.582**	**.597**	**.881**	**.785**	**.788**	**.930**
KNN	.695	.690	.858	.807	.808	.937
Logistic Reg.	.709	.717	.886	.807	.810	.950
SVM	.476	.365	.696	.793	.797	.942
XGBoost	.582	.597	.858	.749	.760	.932

Table A4: Secondary metrics, full Portuguese file (649)

Approach	S1 Acc.	S1 mF1	S1 AUC	S2 Acc.	S2 mF1	S2 AUC
Rule (latest grade)	.752	.741	n/a	.823	.808	n/a
Decision Tree	.641	.647	.881	.823	.808	.935
Random Forest	**.641**	**.647**	**.898**	**.821**	**.803**	**.942**
KNN	.752	.741	.880	.835	.821	.941
Logistic Reg.	.730	.726	.902	.832	.819	.953
SVM	.465	.366	.682	.823	.808	.949
XGBoost	.641	.647	.880	.794	.778	.940

Acc.: accuracy; mF1: macro-F1; AUC: one-vs-rest ROC-AUC. Macro precision and recall are in the full report.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.1