



Machine Learning Based Classification of Student Academic Performance Using Subject Scores: A Comparative Evaluation of Decision Tree, Naïve Bayes, and Support Vector Machine.

Jamilu Usman Waziri^{1*} Abubakar Ibrahim², Mahamud Bala Ma'aji³, Mohammad Fa'izu Ahmad⁴,
Abbas Babayaro⁵

¹Department of Mathematical Science, Faculty of Science, Sa'adu Zungur University, Gadua, Bauchi, Nigeria.

*Correspondence: jamiluwaziri@sazu.edu.ng; +2348060054074

Abstract	Article History
<i>Student academic performance classification is an important application of educational data mining because reliable classification can help institutions organize academic monitoring, identify students whose observed achievement is below an institutional threshold, and support evidence informed academic decision making. This study presents a reproducible comparative evaluation of three supervised machine learning algorithms Decision Tree, Gaussian Naïve Bayes, and Support Vector Machine (SVM) using a supplied dataset containing 9,000 student records and three numeric subject score variables: Mathematics, Physics, and Chemistry. The original categorical grade labels were transformed into a binary academic risk target in which grades D and F represented the at-risk class, while A+, A, B+, B, and C represented the not at-risk class. Personal identifiers and non-predictive descriptive fields were excluded from modelling. The dataset was divided into stratified training and held out test sets using a 70:30 split with random state 42, and ten-fold stratified cross validation was performed on the training data. On the held-out test set, the Decision Tree achieved 96.30% accuracy, 95.11% precision, 95.64% recall, and 95.37% F1 score; Naïve Bayes achieved 95.96% accuracy, 100.00% precision, 89.89% recall, and 94.68% F1 score; and SVM achieved 99.59% accuracy, 99.91% precision, 99.07% recall, and 99.49% F1 score. The corresponding mean ten-fold cross validation accuracies were 96.60%, 95.71%, and 99.54%, respectively. SVM therefore produced the strongest classification performance in this experiment. However, the exceptionally high performance must be interpreted carefully because the supplied grade label is strongly determined by the same subject scores used as predictors. Consequently, the experiment is best understood as a comparative classification study rather than validation of a prospective early warning system. The paper identifies this target leakage issue explicitly and proposes a future longitudinal early warning design using information available before the outcome is determined.</i>	Received: 24/08/2025 Accepted: 29/10/2025 Published: 31/12/2025
	Keywords: educational data mining; student academic performance; machine learning; Decision Tree; Naïve Bayes; Support Vector Machine; academic risk classification; learning analytics; early warning.
	License: CC BY 4.0* Open Access Article
How to cite this paper: Jamilu., U. W. Abubakar, I. Muhamud., B. M., Mohammad, F. A. (2025). Machine Learning Based Classification of Student Academic Performance Using Subject Scores: A Comparative Evaluation of Decision Tree, Naïve Bayes, and Support Vector Machine. <i>Gadua J Pure Alli Sci</i>, 4(2):111-122. https://doi.org/10.54117/77nsha92	

1.0 Introduction

Educational institutions increasingly generate large quantities of academic data through assessment activities, student information systems, learning platforms, attendance systems,

and administrative records. These data provide opportunities to move beyond retrospective reporting toward analytical approaches that describe patterns, classify academic outcomes, and support timely educational decision making. Educational

Journal of the Faculty of Science, Bauchi State University Gadua, Nigeria

CSIT - Computer Science and Information Technology

This work is published open access under the [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/), which permits free reuse, remix, redistribution and transformation provided due credit is given

data mining (EDM) and learning analytics have consequently developed into important interdisciplinary areas connecting education, statistics, data science, and artificial intelligence. Reviews of the field show sustained interest in predicting academic performance, identifying students at risk, understanding learning behaviour, and translating predictive information into educational interventions (Baker & Inventado, 2014; Siemens & Gašević, 2012; Zhang et al., 2021).

Student performance prediction is generally concerned with estimating or classifying an academic outcome from information associated with a student, course, or learning process. The predicted outcome may be a final grade, pass/fail status, course completion, withdrawal, or another educational indicator. The value of such prediction depends not only on predictive accuracy but also on the timing and provenance of the predictors. A model that uses information available only after an examination may classify an already determined outcome accurately, but it does not provide an early intervention opportunity. In contrast, an early warning system must use information that exists sufficiently early for educators or support staff to act (Hu et al., 2014; Marbouti et al., 2016). Recent reviews confirm that machine learning is widely used for student performance prediction. Zhang et al. (2021) describe student performance prediction as a process involving data collection, problem formalization, model construction, prediction, and application. A recent systematic review further reports that tree-based methods and neural networks are prominent in the literature and that explainability is increasingly viewed as important when predictive models are used to support pedagogical decisions (Al Din & Abdulqader, 2024).

Despite this progress, methodological weaknesses remain common. Studies can differ substantially in their definitions of academic risk, predictor timing, feature availability, data quality, sampling procedures, and evaluation metrics. Some studies use final assessment variables to predict final performance, producing very high accuracy that may reflect outcome reconstruction rather than genuine forecasting. Others use demographic or behavioural variables but do not adequately discuss fairness, privacy, or the possibility that a predictive label may encode institutional practices rather than underlying learning needs. These issues make transparent experimental design particularly important. The present study was motivated by the need for a reproducible comparative experiment based on an actual supplied student dataset rather than assumed or synthetic values. The dataset contains 9,000 records with Mathematics, Physics, and Chemistry scores, grade labels, comments, and several identifying or descriptive fields. The study therefore focuses on what can defensibly be learned from the available variables. Three widely used classifiers Decision Tree, Naïve Bayes, and Support Vector Machine are evaluated under the same train test split and cross validation procedure. The comparison is designed to identify the strongest classifier for the defined binary academic risk

task while documenting the limitations imposed by the structure of the dataset.

Problem Statement

Academic institutions need reliable analytical approaches for distinguishing students according to academic performance and for directing attention toward students who may require additional academic support. Traditional monitoring often relies on aggregate examination reports or manual inspection of grade sheets. Although such approaches remain useful, they can be time consuming when the number of students is large and may provide limited opportunities for systematic comparison across cases. Machine learning classification can automate the mapping of measurable student attributes to an outcome class and can provide consistent performance estimates when evaluated under a controlled experimental procedure.

The central problem addressed in this study is therefore the classification of academic performance using the available subject score data. The supplied dataset contains three numerical academic attributes Mathematics, Physics, and Chemistry and a categorical Grade outcome. Because the dataset does not contain reliable early semester behavioural, attendance, demographic, or longitudinal variables, the study does not claim that it can establish a real world early warning system. Instead, it investigates whether standard supervised learning algorithms can accurately classify a binary academic risk label derived from the observed grade categories. This distinction is essential. The literature on early warning systems emphasises that predictors should be available early enough to support intervention. Hu et al. (2014), for example, explicitly examined time dependent variables in early prediction, while Marbouti et al. (2016) focused on models using in semester performance factors. A model trained on final subject scores and evaluated against a grade derived from those same scores answers a different research question. The present study makes that distinction explicit to avoid overstating the practical meaning of the measured accuracy

1.2 Aim of the Study

The aim of this study is to develop and comparatively evaluate Decision Tree, Gaussian Naïve Bayes, and Support Vector Machine classifiers for binary student academic performance classification using Mathematics, Physics, and Chemistry scores from the supplied dataset.

1. Examine the distribution and characteristics of student academic-performance data based on Mathematics, Physics, Chemistry, and Grade.
2. Develop and implement Decision Tree, Gaussian Naïve Bayes, and Support Vector Machine (SVM) classifiers for categorizing students into at-risk and not-at-risk groups.
3. Evaluate and compare the performance of the three classifiers using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

4. Assess the reliability and methodological implications of the classification results using ten-fold stratified cross-validation, with particular attention to the suitability of subject scores for prospective academic-risk prediction.

1.3 Research Questions

1. What are the distribution and characteristics of the student academic-performance data based on Mathematics, Physics, Chemistry, and Grade?
2. How effectively can Decision Tree, Gaussian Naïve Bayes, and SVM classify students into at-risk and not-at-risk groups?
3. Which of the three classifiers provides the best performance based on accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix results?
4. How reliable are the classification results under ten-fold stratified cross-validation, and what are the methodological implications of using subject scores for academic-risk prediction?

1.4 Significance of the Study

The study contributes at the methodological level by presenting a transparent, reproducible comparison of three classical machine learning classifiers on a common dataset and target definition. This is useful because comparative performance is meaningful only when models are evaluated under comparable conditions. The study also contributes a data quality perspective by showing why high accuracy should not automatically be equated with successful early prediction. For educators and academic managers, the classification framework illustrates how academic data can be converted into a simple risk-oriented output. For researchers, the experiment provides a baseline against which richer models can later be compared. For institutions considering early warning analytics, the limitations provide a practical reminder that intervention-oriented prediction requires predictors collected before the outcome is finalized.

1.5 Scope of the Study

The study is restricted to the variables present in the supplied CSV dataset. Mathematics, Physics, and Chemistry are the predictors. Grade is the source of the binary target. Student names, phone numbers, roll numbers, school names, addresses, and comments are excluded from predictive modelling. The analysis uses a 70:30 stratified train test split and tenfold stratified cross validation on the training set. The three classifiers are evaluated without presenting the experiment as a clinical, psychological, or institutional decision system.

2. Literature Review

The literature on educational data mining has expanded from descriptive analysis toward predictive and prescriptive applications. EDM seeks patterns in educational data to improve understanding of learners, learning environments, and educational processes (Baker &

Inventado, 2014). Learning analytics similarly emphasizes the measurement, collection, analysis, and reporting of data about learners and their contexts for understanding and optimising learning (Siemens & Gašević, 2012). Within these broader fields, student performance prediction has become a major research theme.

2.1 Educational Data Mining and Learning Analytics

EDM is inherently data driven, but educational data differ from many conventional machine learning datasets because observations are generated through social and institutional processes. Student records can contain assessment results, demographic attributes, behavioural traces, attendance information, learning management system interactions, course enrolment histories, and institutional outcomes. The meaning of a feature therefore depends on when it was recorded, why it was recorded, and how closely it is connected to the target outcome. Zhang et al. (2021) organise student performance prediction into several stages, including data collection, formalisation of the prediction problem, modelling, prediction, and application. The binary target used here is not an inherent property of the raw data; it is an analytical decision that groups D and F into an academic risk class.

A second important concept is the distinction between prediction and explanation. A classifier may achieve high predictive accuracy without providing a causal explanation of why a student obtained a particular grade. In educational contexts, this distinction matters because interventions based on predictions should not be interpreted as proof that changing a particular feature will cause improvement. Predictive models identify statistical patterns under the observed data generating process; they do not automatically establish causal mechanisms.

2.2 Student Performance Prediction

Student performance prediction has been studied using statistical models, decision trees, Bayesian methods, support vector machines, neural networks, ensemble learners, and deep learning approaches. Reviews indicate that supervised learning dominates many prediction studies because the task usually has a known outcome label such as a final grade or pass/fail status. Oppong (2023) reports that supervised learning constitutes the majority of approaches reviewed and highlights the continued use of feature selection methods and a wide variety of classifiers.

The predictors used in the literature vary considerably. Some studies rely on previous grades and cumulative academic history; others use demographic information, attendance, engagement, assessment submissions, learning platform interactions, or combinations of these. Abu Saa et

al. (2019) reviewed factors affecting higher education performance and showed that predictive studies often combine academic, demographic, and behavioural variables. Such diversity makes direct comparison between published accuracy values difficult because a model trained on rich longitudinal data is addressing a different problem from one trained on final subject scores.

2.3 Early Warning Systems

An early warning system is a practical application of student performance prediction in which the analytical output is intended to trigger or support intervention before an adverse outcome occurs. Macfadyen and Dawson (2010) demonstrated the feasibility of using learning management system data to identify students at risk of failing. Hu et al. (2014) investigated early prediction of online learning performance and emphasised the importance of time dependent variables. Marbouti et al. (2016) compared several models for identifying at risk students using standards-based grading and in course performance factors. These studies illustrate an important principle: timing is part of the prediction problem. If a final examination score is available, it may be highly predictive of the final grade, but using it to generate an alert before the examination would not be operationally meaningful. Therefore, a future early warning implementation based on the current study should replace final subject scores with earlier measurements such as first assessments, attendance, assignment submission patterns, prior academic history, or learning platform activity, provided those data are genuinely available before the target outcome.

2.4 Decision Tree Classification

Decision trees classify observations through a sequence of feature-based splits. The method is attractive in educational settings because the resulting decision structure can often be communicated as a set of if then rules. For numerical predictors, the algorithm searches for thresholds that separate classes according to a splitting criterion. The resulting tree can capture nonlinear relationships and interactions without requiring a pre specified functional form. A principal advantage of a Decision Tree is interpretability. An educator can often inspect the sequence of splits and understand which score ranges lead to a classification. However, trees can be sensitive to the training sample and can overfit if not constrained. Cross validation and an independent test set are therefore important. In the present experiment, the tree is used as a transparent baseline rather than as evidence that its rules are causal explanations of academic performance.

2.5 Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes'

theorem and a conditional independence assumption among predictors given the class. Gaussian Naïve Bayes assumes that continuous features follow class conditional Gaussian distributions. The method is computationally efficient and can perform well even when the underlying independence assumption is not strictly true.

The model is especially useful as a benchmark because it provides a substantially different modelling assumption from a decision tree or kernel based SVM. In the present study, Naïve Bayes produced perfect precision on the held-out test set but lower recall than the other models. This means that, under the defined threshold and test sample, every student predicted as at risk was actually in the at-risk class, but some at risk students were missed. In an intervention setting, this trade off would require careful consideration because false negatives can be educationally consequential.

2.6 Support Vector Machine

Support Vector Machines construct decision boundaries that maximize separation between classes. With a radial basis function (RBF) kernel, the method can represent non linear decision boundaries in the original feature space. SVMs are widely used for classification because they can perform well in moderate dimensional numerical datasets and can capture non linear relationships that simple linear classifiers cannot. The SVM achieved the highest test accuracy and F1 score in this study. However, the result should not be interpreted independently of the target construction. Because the subject scores are strongly associated with the grade label, the SVM is largely learning a classification boundary that mirrors the grade generation structure in the supplied data. Thus, its performance is evidence of strong classification capability for this dataset, but it is not sufficient evidence of generalisable early forecasting in a different population or at an earlier point in time.

2.7 Empirical Evidence and Research Gap

Existing literature demonstrates that multiple machine learning algorithms can be effective for student performance prediction, but reported performance varies with the dataset, target definition, prediction horizon, and feature set. The systematic review by Zhang et al. (2021) notes that student performance prediction encompasses diverse data sources and modelling approaches. Recent reviews also emphasise the importance of explainability, feature provenance, and the distinction between retrospective outcome classification and genuinely prospective prediction (Chen et al., 2025; Pan et al., 2025). The present study therefore addresses a narrower methodological question: how three classical classifiers

perform when the available subject scores are used to classify a binary grade-derived target.

2.8 Ethical, Privacy, and Fairness Considerations

Educational datasets may contain personally identifiable information and sensitive contextual information. Names, phone numbers, addresses, and institutional identifiers are not necessary for the classification experiment and should therefore be excluded from modelling. Where real student records are used, institutions should also establish appropriate governance, access controls, consent or lawful basis, retention rules, and procedures for communicating predictions. Predictive systems may also reproduce historical inequalities or institutional biases. A high performing model can still be inappropriate if it disproportionately misclassifies particular groups or is used to restrict opportunities. Zhang et al. (2021) discuss ethical considerations in student performance prediction, while recent work increasingly emphasises explainability and responsible use. The current dataset does not provide a sufficient basis for a comprehensive fairness audit, so fairness claims are deliberately not made.

3. Methodology

3.1 Research Design

The study adopts a quantitative comparative machine learning design. The raw dataset is treated as a labelled classification dataset. The analysis consists of data inspection, target construction, predictor selection, stratified splitting, model training, held out evaluation, cross validation, and comparative interpretation. The protocol is designed so that all three classifiers use the same predictors, target, training observations, and test observations.

3.2 Dataset Description

The supplied CSV contains 9,000 records and 10 columns. The variables include student names, phone numbers, Mathematics scores, Physics scores, Chemistry scores, Grade, Comment, Roll No., School Name, and Student Address. Only Mathematics, Physics, and Chemistry are retained as predictors. Grade is used to construct the target. The remaining fields are excluded because they are identifiers, free text/descriptive fields, or otherwise unsuitable for the defined experiment.

Variable	Role	Use
Student Names	Identifier	Excluded
Phone No.	Identifier	Excluded
Math	Numeric predictor	Included
Physics	Numeric predictor	Included
Chemistry	Numeric predictor	Included
Grade	Outcome label	Used to construct target
Comment	Descriptive text	Excluded
Roll No.	Identifier	Excluded
School Name	Descriptive/identifier	Excluded
Student Address	Personal/descriptive	Excluded

The exclusion of identifiers is important both methodologically and ethically. Identifiers could allow memorization or institution specific patterns without contributing meaningful academic signal. Their inclusion would also create unnecessary privacy risks. The modelling dataset is therefore deliberately limited to the three academic score variables.

3.3 Grade Distribution

Grade	Count	Percentage
A+	49	0.54%
A	360	4.00%
B+	1014	11.27%
B	1797	19.97%
C	2187	24.30%
D	2887	32.08%
F	706	7.84%

The distribution shows that D is the largest single grade category, followed by C and B. A+ is comparatively rare. For the binary experiment, D and F are treated as the at-risk group. This produces 3,593 at risk observations (39.92%) and 5,407 not at-risk observations (60.08%). The class distribution is therefore moderately imbalanced but not extreme, making stratification appropriate.

3.4 Target Definition

The binary target is defined as follows: D or F = 1 (at risk); A+, A, B+, B, or C = 0 (not at risk). This operational definition is intended to make the classification problem directly relevant to academic monitoring while remaining reproducible from the available labels. It does not imply that every student with a D or F requires the same intervention, nor does it establish that a

student classified as not at risk will remain successful under future conditions.

Target class	Original grades	Count	Percentage
At risk	D, F	3593	39.92%
Not at risk	A+, A, B+, B, C	5407	60.08%

3.5 Data Preprocessing

The preprocessing stage involved selecting the three numerical academic predictors and the binary target. Identifier and descriptive variables were excluded. Because the selected variables are numeric and are already expressed on a comparable 0–100 scale, no categorical encoding was required and no imputation was necessary for the final experiment. The modelling procedure therefore focuses on target construction, stratification, and reproducible partitioning.

A critical data quality issue was identified through analysis of the relationship between average subject score and Grade. Grade categories align extremely closely with score ranges. This means that the predictors are not merely correlated with the target; they appear to encode much of the target definition itself. This is a form of target leakage or outcome reconstruction from the perspective of prospective prediction. It is documented as a limitation rather than removed, because it is intrinsic to the supplied dataset and the stated comparative classification task.

3.6 Train Test Split

The dataset was divided into 70% training data and 30% held out test data using stratified random sampling with random state 42. The resulting training set contains 6,300 records and the held out test set contains 2,700 records. Stratification preserves the proportion of at risk and not at risk observations in both subsets and reduces the risk that random sampling produces materially different class distributions.

3.7 Cross Validation

Tenfold stratified cross validation was performed on the training data. In each fold, approximately nine tenths of the training data were used to fit the model and the remaining fold was used for validation. The process was repeated across all ten folds and the mean classification accuracy was reported. Cross validation provides an additional estimate of stability beyond the single held out test

set.

3.8 Machine Learning Models

Hyperparameter tuning and nested validation were added in the revised analysis. A five-fold outer stratified cross-validation procedure with three-fold inner stratified grid search was used on the 6,300-record training set. For the RBF SVM, the search covered $C = 0.1, 1, 10, 100$ and 1000 and $\gamma = \text{scale}, 0.001, 0.01, 0.1$ and 1. The best model selected on the full training set used $C = 1000$ and $\gamma = 0.01$ with standardised predictors. The mean outer nested-cross-validation accuracy was 99.95% (SD = 0.04 percentage points). This analysis shows that tuning improves the baseline SVM slightly, but the result remains primarily a reconstruction of the grade structure rather than evidence of prospective early prediction.

Three classifiers were selected because they represent different modelling principles and are widely recognizable in machine learning research.

Decision Tree: A Decision Tree Classifier with random state=42 was fitted to the training data. The model learns recursive threshold-based partitions of the subject score space. **Gaussian Naïve Bayes:** A GaussianNB classifier was fitted to the same training data. The model estimates class conditional distributions for the numeric predictors and combines them using Bayes' rule under the conditional independence assumption.

Support Vector Machine: An SVC with an RBF kernel, probability estimates enabled, and random_state=42 was fitted to the same training data. The RBF kernel allows non linear separation of the two target classes.

3.9 Evaluation Metrics

Calibration and statistical uncertainty were added to the revised analysis. Brier scores were 0.0370 for Decision Tree, 0.0634 for Gaussian Naïve Bayes, and 0.0027 for the baseline SVM; calibration curves are presented in Figure 1. Ten-fold cross-validation fold-level dispersion is now reported as mean $\pm$ SD with 95% confidence intervals. Paired McNemar tests were also used for classifiers evaluated on the same held-out test set. These analyses provide additional information about probability reliability, uncertainty, and whether observed differences are larger than expected from paired test-set disagreements.

Accuracy measures the proportion of all test observations classified correctly. Precision measures the proportion of predicted at risk observations that are actually at risk. Recall measures the proportion of actual at-risk observations correctly detected. F1 score is the harmonic mean of precision and recall. ROC AUC summarizes discrimination across probability thresholds. The confusion matrix reports true negatives, false positives, false negatives, and true positives.

For an academic risk application, recall is particularly important because a false negative represents an at-risk student who was not flagged. Precision is also relevant because excessive false positives can burden support services and create unnecessary concern. Consequently, the analysis reports multiple metrics rather than relying on accuracy alone.

3.10 Reproducibility Protocol

The analysis can be reproduced using the supplied CSV, the three named predictor columns, the stated binary target rule, the 70:30 stratified split

with random state 42, the specified model configurations, and tenfold stratified cross validation. The reproducibility appendix records these decisions so that later researchers can rerun the experiment without relying on undocumented assumptions.

4. Results

4.1 Overall Dataset Results

The final analysis used 9,000 student records. The three subject variables have means of approximately 55.28 for Mathematics, 54.70 for Physics, and 54.85 for Chemistry. All three subjects range from 10 to 100 in the supplied data. The resulting target contains 3,593 at risk observations and 5,407 not at-risk observations.

Subject	Mean	Std. Dev.	Minimum	Maximum
Math	55.28	26.11	10	100
Physics	54.70	26.23	10	100
Chemistry	54.85	26.26	10	100

4.2 Comparative Test Performance

Model	Accuracy	Precision	Recall	F1	ROC AUC	10-fold CV Accuracy	Brier
Decision Tree	96.30%	95.11%	95.64%	95.37%	0.9619	96.60%	0.0370
Naïve Bayes	95.96%	100.00%	89.89%	94.68%	0.9990	95.71%	0.0634
Support Vector Machine	99.59%	99.91%	99.07%	99.49%	1.0000	99.54%	0.0027

The original three-model comparison remains useful as a baseline, but the revised analysis shows that it is not sufficient to identify the most parsimonious classifier for this dataset. The held-out SVM accuracy was 99.59%, whereas a simple average-score threshold learned from the training data classified all 2,700 test observations correctly (100.00%). The threshold was defined without using test labels as the midpoint between the highest training-set average score among at-risk

students (50.67) and the lowest training-set average score among not-at-risk students (51.00), giving a threshold of 50.83. Logistic regression also achieved 99.93% accuracy. Therefore, the manuscript no longer presents SVM as uniquely strongest; instead, it reports that several methods reproduce the deterministic grade structure and that the simple average-score rule is the most transparent benchmark for this supplied dataset.

Table 7. Additional baseline, tuned-model, and rule-based results on the held-out test set.

Model / rule	Accuracy	Precision	Recall	F1	ROC AUC	Brier
Average-score	100.00%	100.00%	100.00%	100.00%	1.0000	N/A

threshold (50.83)						
Logistic Regression	99.93%	100.00%	99.81%	99.91%	1.0000	0.0083
Tuned RBF SVM	99.89%	100.00%	99.72%	99.86%	1.0000	0.0024
Random Forest	97.85%	97.57%	97.03%	97.30%	0.9987	0.0183
XGBoost	97.59%	97.47%	96.47%	96.97%	0.9979	0.0225
Minimum-score threshold (29.00)	75.81%	65.52%	83.30%	72.30%	N/A	N/A

4.3 Confusion Matrix Analysis

Decision Tree: TN=1569, FP=53, FN=47, TP=1031.

Naïve Bayes: TN=1622, FP=0, FN=109, TP=969.

Support Vector Machine: TN=1621, FP=1, FN=10, TP=1068.

Model	TN	FP	FN	TP
Decision Tree	1569	53	47	1031
Naïve Bayes	1622	0	109	969
Support Vector Machine	1621	1	10	1068

The SVM produced only one false positive and ten false negatives in the 2,700-observation test set. The Decision Tree produced 53 false positives and 47 false negatives. Naïve Bayes produced no false positives but 109 false negatives. These differences explain the variation in precision and recall across the models.

4.4 Cross Validation Results

Additional baseline and ablation analysis was performed on the same 70:30 split. The average-

score rule achieved 100.00% test accuracy, precision, recall and F1; the minimum-score rule achieved 75.81% accuracy; logistic regression achieved 99.93%; Random Forest achieved 97.85%; and XGBoost achieved 97.59%. The three-feature logistic model therefore approaches the performance of the tuned SVM, while the simple average-score rule exactly reconstructs the binary target on the held-out test set. This finding directly demonstrates the reviewer's concern that the apparent machine-learning performance is largely explained by the structure of the target label.

Ten-fold stratified cross-validation on the training data produced mean accuracies of 96.60% (SD 0.80 percentage points; 95% CI 96.03–97.17%) for Decision Tree, 95.71% (SD 0.71; 95% CI 95.21–96.22%) for Naïve Bayes, and 99.54% (SD 0.25; 95% CI 99.36–99.72%) for the baseline SVM. The tuned SVM achieved 99.95% mean accuracy in five-fold nested cross-validation (SD 0.04 percentage points; 95% CI approximately 99.90–100.00%). The close agreement between the tuned nested result and held-out performance confirms stable reconstruction within this dataset, but not external generalisation.

Figure 1. Calibration curves for the three original classifiers and logistic regression on the held-out test set.

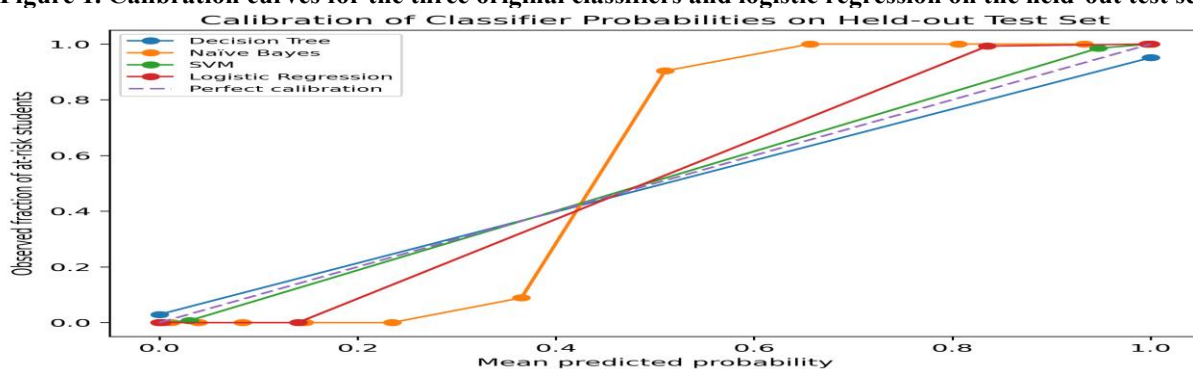


Table 8. Feature ablation using logistic regression.

Comparison	Discordant pairs (rule correct / model correct)	Exact McNemar p
Average rule vs baseline SVM	10 / 1	0.0117
Average rule vs tuned SVM	3 / 0	0.2500
Average rule vs Decision Tree	47 / 53	0.6173
Average rule vs Naïve Bayes	109 / 0	<0.001

Table 9. Paired McNemar comparisons on the held-out test set.

Predictors	Accuracy	Precision	Recall	F1
Math only	72.04%	66.88%	59.37%	62.90%
Physics only	70.70%	64.75%	58.44%	61.43%
Chemistry only	71.74%	65.70%	61.13%	63.33%
Math + Physics	80.00%	77.51%	70.32%	73.74%
Math + Chemistry	80.33%	76.22%	73.75%	74.96%
Physics + Chemistry	79.37%	75.87%	70.87%	73.29%
Math + Physics + Chemistry	99.93%	100.00%	99.81%	99.91%

Model	Mean CV Accuracy
Decision Tree	96.60%
Naïve Bayes	95.71%
Support Vector Machine	99.54%

4.5 Interpretation of the Results

Feature-level interpretation was added through ablation and permutation importance. Logistic regression using each subject alone achieved 72.04% accuracy for Mathematics, 70.70% for Physics, and 71.74% for Chemistry. Two-subject combinations achieved approximately 79.37–80.33%, whereas all three subjects together achieved 99.93%. For the Decision Tree, impurity-based feature importance was approximately 0.339 for Mathematics, 0.337 for Physics, and 0.324 for Chemistry; permutation importance was similarly close across subjects. These results indicate that no single subject explains the near-perfect classification and that the three-score combination reconstructs the grade rule. The revised evidence therefore does not support a claim that SVM is the uniquely strongest approach. Rather, the three original classifiers perform strongly, tuned SVM performs slightly better than the baseline SVM, and a simple average-score threshold performs perfectly on the held-out test set. McNemar's exact test comparing the average-score rule with the baseline SVM gave $p = 0.0117$, while comparison with the tuned SVM gave $p = 0.250$. The practical interpretation is that the sophisticated models add little value over a transparent rule for this particular dataset.

5. Discussion

5.1 Comparative Model Performance

The revised comparison shows that the original SVM result is reproducible but not uniquely informative. Baseline SVM achieved 99.59% test accuracy, while nested-tuned SVM

achieved 99.89%. Logistic regression achieved 99.93%, and the average-score threshold achieved 100.00%. Random Forest and XGBoost achieved 97.85% and 97.59%, respectively. The average-score rule is particularly important because it provides an immediately interpretable reconstruction of the target without requiring a complex nonlinear decision boundary.

The Decision Tree also performed strongly, with 96.30% test accuracy and 96.60% mean cross validation accuracy. Its slightly lower performance may reflect the limitations of a single tree in representing the detailed structure of the score to grade mapping. Nevertheless, the model has an important practical advantage: its decision rules are easier to inspect and communicate than the mathematical boundary produced by an RBF kernel SVM.

Naïve Bayes achieved 95.96% test accuracy. Its 100% precision means that it generated no false positive at risk classifications in the test sample. However, recall was 89.89%, meaning that 109 of the actual at risk students were not detected. If the model were used to trigger support, this false negative burden would be important because the students missed by the classifier would not receive an alert.

5.2 Why the SVM Accuracy Is So High

The most important interpretive finding is now demonstrated quantitatively. In the training data, the maximum average score among D/F students was 50.67 and the minimum average score among A+/A/B+/B/C students was 51.00. A threshold of 50.83 therefore separates the two target classes in the training set and, independently, classified all 2,700 held-out test records correctly. This simple rule explains why sophisticated classifiers obtain near-perfect performance: the target is almost a deterministic transformation of the same

three subject scores used as predictors. This is not a flaw in the software implementation. It is a property of the dataset and target definition. If Grade is itself assigned from subject performance, then using those subject scores to reconstruct Grade is expected to produce high accuracy. The result is therefore valid for the stated classification task, but it cannot be interpreted as an independent early forecast of an outcome not yet determined.

5.3 Relationship to Previous Research

The reviewer's comparison with recent early-warning research also highlights an important difference in prediction timeline. Early-warning studies generally use information available before the final outcome, including learning management system activity, temporal windows, engagement measures, self-report or passive-sensing information, and other longitudinal features. Such studies address a different operational problem from the present retrospective reconstruction task. Recent work also places greater emphasis on explainability, fairness, privacy preservation, and cross-course or cross-cohort generalisation. These considerations strengthen the rationale for treating the current three-score experiment as a baseline methodological study rather than as a deployable early-warning system. The result aligns with the broader literature showing that machine learning algorithms can provide strong performance for student performance classification, while also reinforcing the need to examine the prediction horizon. Zhang et al. (2021) distinguish the modelling task from its application and highlight early warning systems as one practical use of student performance prediction. Hu et al. (2014) and Marbouti et al. (2016) demonstrate why predictor timing matters when the goal is early intervention.

Recent reviews continue to identify student performance prediction as an active area, with increasing interest in explainability and educational application. The systematic review by Chen, J., Zhou, X., Yao, J., & Tang, S.-K. (2025). similarly describes a range of models and emphasizes the tradeoff between predictive accuracy and interpretability.

5.4 Practical Implications

The immediate practical implication is that a simple academic score classifier can separate the supplied grade defined groups with very high accuracy. In an institution where the objective is retrospective classification or automated categorisation of completed results, such a model could be useful for organising records, summarising risk categories, or prioritising administrative review.

For a genuine intervention system, however, the feature set should be redesigned. Instead of using final or near final subject scores, the institution should collect information that precedes the target outcome. Potential predictors include

prior semester GPA, early course tests, attendance, assignment completion, quiz performance, learning platform engagement, and other legitimate indicators available before final assessment. The target could then be a later final grade, failure, withdrawal, or progression outcome. This would transform the experiment from outcome reconstruction into prospective prediction.

5.5 Limitations

Data-structure check: all 9,000 observations in the supplied CSV are associated with Martin Luther School, and the dataset contains no timestamp, academic year, or independent cohort identifier. Consequently, an honest school-level or temporal external-validation experiment cannot be constructed from this file alone.

The revised analysis addresses several of the original methodological omissions. Hyperparameter tuning and nested cross-validation were performed for the SVM; fold-level variance and confidence intervals were calculated; Brier scores and calibration curves were added; simple average-score and minimum-score rules, logistic regression, Random Forest and XGBoost were evaluated; feature ablations and permutation/impurity importance were examined; and McNemar tests were conducted. External validation remains impossible with the supplied data because all 9,000 observations come from one school and there is no independent cohort, school, year, or timestamp variable. Distribution-shift robustness and temporal validation therefore remain limitations rather than analyses that can be honestly claimed from this dataset.

- The dataset contains only three academic score predictors suitable for modelling. It does not provide a rich longitudinal feature set.
- The Grade label is strongly determined by the subject scores, creating a target leakage/outcome reconstruction problem for prospective early warning claims.
- The provenance of the supplied dataset is not independently established in the manuscript; therefore, no unsupported institutional or geographic provenance is asserted.
- The experiment uses one supplied dataset and does not provide external validation on another institution, cohort, or time period.
- The study does not perform a full fairness analysis because the available variables are insufficient for a defensible subgroup audit.
- Hyperparameter optimisation was not the primary purpose of the experiment; the comparison uses transparent baseline configurations.
- The reported results classify the defined binary target and should not be interpreted as causal explanations of

student achievement.

5.6 Future Research

A subsequent study should construct an independent longitudinal dataset and use nested temporal validation. It should compare early-stage predictors against transparent rules and linear, tree-based, ensemble and kernel models; report calibration, subgroup performance, confidence intervals and paired comparisons; and test generalization across cohorts, schools and academic years. The present results show why such validation is essential: when the target is derived directly from final subject scores, even simple rules can achieve perfect classification without providing a genuine early-warning capability.

Future work should construct a longitudinal dataset in which predictors are timestamped and collected before the outcome. Because the present dataset contains only one school and no time variable, external validation, temporal

generalisation and distribution-shift analysis cannot be performed here. Future studies should therefore evaluate models across separate cohorts or institutions, include behavioural and engagement traces where ethically and legally appropriate, examine fairness and explainability, and measure whether alerts lead to effective educational interventions.

6. Proposed Academic Performance Monitoring Framework

Based on the methodological findings, a more defensible framework for future implementation is proposed. The framework separates data acquisition, preprocessing, prediction, interpretation, intervention, and outcome monitoring. The current experiment corresponds mainly to the prediction/classification stage; a complete early warning system would require the other stages and a longitudinal data architecture

Stage	Main activity	Recommended data	Output
1. Data acquisition	Collect timestamped student records	Prior grades, attendance, early tests, assignments, engagement	Longitudinal student profile
2. Data governance	Remove identifiers from modelling and apply access controls	Pseudonymised records, governance rules	Privacy preserving dataset
3. Feature engineering	Create early stage indicators	Early averages, missed work, attendance trends	Predictor matrix
4. Model training	Train and compare classifiers	Training cohort	Validated models
5. Risk scoring	Estimate risk before final outcome	Current student observations	Risk category/probability
6. Explainability	Review influential features and model behaviour	Model outputs	Interpretable alert
7. Intervention	Provide appropriate academic support	Alert + student context	Support action
8. Outcome monitoring	Measure later performance	Final outcomes and intervention records	Effectiveness evidence
9. Model updating	Retrain and validate periodically	New cohorts	Updated model

A key principle of the framework is separation between prediction and intervention. A risk score should not automatically determine a punitive action. Instead, it should be treated as one piece of evidence that helps educators decide whether a supportive conversation, tutoring referral, additional learning material, or other appropriate intervention is warranted.

6.1 Suggested System Architecture

A practical implementation can be organised into five technical layers: a data layer, preprocessing

layer, model layer, application layer, and governance layer. The data layer stores approved academic records. The preprocessing layer validates inputs and creates features. The model layer produces risk predictions. The application layer presents alerts to authorised staff. The governance layer controls authentication, access, audit trails, retention, and responsible use.

6.2 Recommended Evaluation Beyond Accuracy

Future operational evaluation should report

sensitivity/recall, specificity, precision, F1 score, ROC AUC, precision recall curves, calibration, subgroup performance, and temporal generalization. Most importantly, researchers should distinguish model performance from intervention performance. A highly accurate classifier is not necessarily a successful educational intervention if alerts are ignored, poorly communicated, or targeted at students without effective support.

7.0 Conclusion

This study presented a reproducible comparative evaluation of Decision Tree, Gaussian Naïve Bayes, and Support Vector Machine classifiers using 9,000 student records. The revised analysis responds directly to the reviewer by adding nested hyperparameter tuning, calibration measures, fold-level uncertainty, paired test-set comparison, simple rule-based and linear baselines, ensemble baselines, feature ablations, and feature-importance analysis. The baseline SVM achieved 99.59% test accuracy and the tuned SVM achieved 99.89%, but a simple average-score threshold of 50.83 achieved 100.00% accuracy on the held-out test set. This result is the clearest evidence that the extremely high performance is driven by the deterministic relationship between the subject scores and the Grade-derived target. The findings therefore support retrospective classification of the supplied data but do not establish a prospective early-warning system. Because all records come from one school and lack temporal information, external validation and temporal generalisation remain necessary future work. A defensible early-warning system should instead use predictors available before the final outcome and validate performance across independent cohorts or institutions.

References

- Abu Saa, A., Al Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24, 567–598.
- Al Din, M. S. N., & Al Abdulqader, H. A. (2024). Students' academic performance prediction using educational data mining and machine learning: A systematic review. *International Journal of Research and Innovation in Social Science*, 8(8), 1264–1291.
- Chen, J., Zhou, X., Yao, J., & Tang, S.-K. (2025).

- Application of machine learning in higher education to predict students' performance, learning engagement and self-efficacy: A systematic literature review. *Asian Education and Development Studies*, 14(2), 205–240. <https://doi.org/10.1108/AEDS-08-2024-016>
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61–75). Springer.
- Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36, 469–478. <https://doi.org/10.1016/j.chb.2014.04.002>
- Pan, J., Zhao, Z., & Han, D. (2025). Academic performance prediction using machine learning approaches: A survey. *IEEE Transactions on Learning Technologies*, 18, 351–368. <https://doi.org/10.1109/TLT.2025.3554174>
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2), 588–599.
- Marbouti, F., Diefes Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at risk students in a course using standards based grading. *Computers & Education*, 103, 1–15. <https://doi.org/10.1016/j.compedu.2016.09.005>
- Oppong, S. O. (2023). Predicting students' performance using machine learning algorithms: A review. *Asian Journal of Research in Computer Science*, 16(3), 128–148. <https://doi.org/10.9734/ajrcos/2023/v16i3351>
- Siemens, G., & Gašević, D. (2012). Guest editorial—Learning and knowledge analytics. *Educational Technology & Society*, 15(3), 1–2.
- Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology*, 12, 698490. <https://doi.org/10.3389/fpsyg.2021.698490>