Exploratory machine-learning classification of student performance in physical education using fitness, engagement, and academic indicators

Princes Lloyda Peralta[1*,A,B,D-F], Shekiro Raposas[2,A,C,D-F], Reinalyn Bucsit[1,A,D,E,F]

[1] General Education Department, College of Inclusive Education, Lorma Colleges, Urbiztondo, San Juan, La Union 2514, Philippines
[2] College of Computer Studies and Engineering, Lorma Colleges, Urbiztondo, San Juan, La Union 2514, Philippines

***Corresponding author:** Princes Lloyda Peralta; General Education Department, College of Inclusive Education, Lorma Colleges, Urbiztondo, San Juan, La Union 2514, Philippines; email: princeslloyda.peralta@lorma.edu

Received: 2026-08-12
Accepted: 2026-09-30
Published: 2026-10-01

A – Research concept and design
B – Collection and/or assembly of data
C – Data analysis and interpretation
D – Writing the article
E – Critical revision of the article
F – Final approval of article

Check for updates

This is an Open Access article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License

ABSTRACT

Background: Physical Education (PE) performance is multidimensional and may be associated with fitness, engagement, attendance, motivation, and prior achievement. Machine learning can identify classification patterns across these indicators, but predictive claims require clear temporal ordering and rigorous validation.

Objective: This study compared four machine-learning classifiers for categorizing student PE performance while minimizing target leakage and addressing class imbalance.

Methods: An exploratory, leakage-controlled classification analysis used a public Kaggle dataset of uncertain provenance. The dataset contained 500 records, 19 substantive variables, and one identifier, with Performance categorized as Low, Average, or High Performer. Because the label-construction rule was undocumented, Performance was treated as an inherited categorical outcome. Overall PE Performance Score, Improvement Rate, and Final Grade were excluded because they could contain current, change-based, or outcome-related information. Categorical predictors were one-hot encoded, numeric predictors were standardized, and preprocessing was performed within training folds. Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression were evaluated using stratified five-fold cross-validation with class-weighted learning.

Results: Random Forest achieved the highest accuracy (0.856), whereas Logistic Regression achieved the highest balanced accuracy (0.749) and macro-F1 (0.645). Its out-of-fold recall was 0.759 for Low, 0.720 for Average, and 0.750 for High Performers, indicating that accuracy alone may inadequately reflect performance in an imbalanced multiclass setting.

Conclusion: Leakage-controlled, class-aware evaluation is essential in machine-learning classification of student performance. These findings are exploratory and should not be interpreted as validated prospective predictions. Future studies should use documented data, explicit temporal ordering, larger balanced samples, and independent external validation.

Keywords: class imbalance, data leakage, exploratory classification, machine learning, physical education, student performance.

How to cite this article: Peralta, P. L., Raposas, S., & Bucsit, R. (2026). Exploratory machine-learning classification of student performance in physical education using fitness, engagement, and academic indicators. *Physical Education and Sports: Studies and Research*, *5*(3), 276-289. https://doi.org/10.56003/pessr.v5i3.790

INTRODUCTION

Physical Education (PE) contributes to students' physical competence, participation, health-related knowledge, and engagement. Performance assessment in PE is multidimensional and may encompass technical skills, decision-making, knowledge, fitness, participation, and other indicators, depending on the assessment framework and instructional context (Lieberman et al., 2024; Manso-Lorenzo et al., 2024). Recent systematic reviews have also shown that artificial intelligence and other data-driven technologies are increasingly being explored in PE for assessment, feedback, performance analysis, and personalized learning (Zhou et al., 2024). These developments suggest that data-driven approaches can complement, rather than replace, teacher judgment by enabling the simultaneous examination of patterns across multiple indicators.

Educational data mining has increasingly been used to analyze and classify student performance, with machine-learning approaches applied to identify patterns associated with academic outcomes (Barbeiro et al., 2024; Bellaj et al., 2024; Ersozlu et al., 2024; Xiao et al., 2022; Yağcı, 2022). Reviews of the field indicate that supervised learning remains widely used, with decision trees, support vector machines, random forests, and logistic regression among the commonly applied approaches (Xiao et al., 2022; Ersozlu et al., 2024). However, student-performance modelling requires careful attention to data preparation, outcome definition, class imbalance, validation, and interpretation (Xiao et al., 2022). These issues are particularly important when the outcome is represented by performance categories rather than a continuous score and when one category is substantially underrepresented. In PE, the availability of domain-specific indicators, such as strength, endurance, flexibility, agility, body mass index, physical-activity behavior, attendance, participation, motivation, and health-fitness knowledge, provides a potentially useful basis for exploratory classification.

The dataset used in the present study was published on Kaggle as Student Physical Education Performance (Ziya07, 2024). The data card describes 500 records containing demographic, physical-fitness, engagement, academic, and performance variables and identifies the dataset as CC0/Public Domain. A recent study also reported using a 500-instance Kaggle PE dataset containing similar types of variables and applied a substantially more complex analytical pipeline involving label encoding, SMOTE, feature selection, and hyperparameter optimization (Li & Cao, 2025). However, the public data card does not provide information about the school or sampling frame, data-collection protocol, collection dates, measurement instruments, or independent evidence confirming that the records were empirically collected from students. It also does not disclose the exact rule, cut-points, or component weights used to derive the Performance categories from the underlying measures. Consequently, the provenance of the dataset and the target-generation process cannot be independently verified.

Because educational performance modelling can be sensitive to outcome definition, class imbalance, predictor timing, and target leakage, the present study adopts a deliberately conservative classification framework. Leakage is a recognized source of overly optimistic model evaluation when information related to the target is incorporated into predictors or model-development procedures (Kaufman et al., 2012). Therefore, the analysis focuses on whether the supplied variables can

distinguish among the three observed performance categories after potentially outcome-overlapping variables are excluded and all preprocessing procedures are confined to the training folds. Accordingly, the findings are interpreted as exploratory classification results rather than as evidence of a deployable early-warning system or a validated prospective prediction model.

The study therefore aimed to: (1) establish the dataset size and class distribution; (2) identify and exclude predictors that may contain outcome or post-outcome information; (3) perform transparent data-integrity and outlier screening; (4) implement preprocessing within a reproducible machine-learning pipeline; (5) compare four classification algorithms using stratified five-fold cross-validation and class-aware metrics; (6) quantify uncertainty in cross-validated performance; and (7) identify predictors that remain influential after leakage-prone variables are removed.

METHODS

Study Design and Participants

A quantitative, non-experimental secondary-data analysis was conducted using the publicly available Kaggle dataset Student Physical Education Performance (Ziya07, 2024). The downloaded dataset contained 500 records and 20 columns, comprising one identifier, 18 candidate predictor variables, and the categorical performance outcome. The dataset contained no missing values; therefore, all 500 records were retained for analysis.

The public data card describes the variables and their intended meanings but does not document the original population, school, sampling procedure, measurement instruments, data-collection period, or validation procedures. It also does not explicitly indicate whether the records represent empirical observations or synthetic/simulated data. Performance is described as a final performance category, but the documentation does not provide the exact coding rule, threshold values, or weighting scheme used to generate the label. Accordingly, the target variable was treated as an inherited categorical outcome, and no attempt was made to reconstruct or infer its underlying generation formula. This uncertainty regarding dataset provenance and target construction was considered an important constraint on external validity and interpretation.

Ethical approval statement

This study involved secondary analysis of a publicly accessible, de-identified dataset obtained from Kaggle. The researchers did not recruit, contact, intervene with, or directly collect data from human participants and had no access to directly identifying information. In accordance with the institutional policy of Lorma Colleges, secondary analysis of publicly available, de-identified data does not require formal ethical review or ethics committee approval. The original data-collection procedures, informed-consent procedures, and ethical review status associated with the creation of the publicly available dataset could not be independently verified from the available documentation and are therefore acknowledged as a limitation of the study.

Target Variable and Class Distribution

The target variable, performance, consisted of three observed categories: low performer, average performer, and high performer. The supplied labels were retained without modification or reclassification. Because the public documentation does not disclose the thresholds, component variables, or weighting formula used to construct these categories, the analysis cannot establish whether the target was completely independent of the retained predictors during the original data-generation process.

This distinction is important because controlling for predictor leakage during model development cannot eliminate structural dependence that may already be embedded in a dataset-generated target.

Leakage Audit and Predictor Selection

A systematic audit of the candidate predictors was conducted before model development to identify variables that could introduce target leakage. A predictor was retained only when it could reasonably be interpreted as information available independently of the current performance outcome.

Overall PE performance score (Overall_PE_Performance_Score) was excluded because it was described as a composite measure of overall PE performance. Its conceptual overlap with the categorical performance outcome created a substantial risk that it contained information directly related to the target.

Improvement rate (Improvement_Rate) was also excluded because it represents change in PE performance over time. The available documentation does not provide sufficient temporal information to determine whether this variable was calculated independently of, or after, the Performance outcome. Its inclusion could therefore introduce information that would not necessarily be available at the intended prediction point.

Final grade (Final_Grade) was excluded because it represents a final academic or course outcome and may contain information available only concurrently with or after the performance outcome. Its temporal relationship with Performance could not be independently established from the available documentation.

The identifier (ID) was excluded because it had no substantive educational or PE meaning. Previous semester PE grade (Previous_Semester_PE_Grade) was retained because it was explicitly described as a prior-semester measure and could therefore reasonably be interpreted as historical academic information preceding the current performance outcome. The remaining fitness, engagement, demographic, physical-activity, and health-related variables were retained as candidate predictors because the available documentation did not identify them as direct components of the performance target. Nevertheless, because detailed collection dates and a documented prediction timeline were unavailable, the temporal availability of every retained predictor could not be independently verified. The resulting models were therefore interpreted as exploratory classifiers rather than as validated prospective prediction models.

Data Preprocessing

Preprocessing and machine-learning pipelines were implemented using scikit-learn (Pedregosa et al., 2011). Numeric predictors were standardized using StandardScaler, whereas categorical predictors were transformed using one-hot encoding. One-hot encoding was used to avoid imposing artificial numerical

distances on nominal or undocumented ordinal categories. All preprocessing parameters were estimated exclusively from the training portion of each cross-validation fold and subsequently applied to the corresponding validation fold.

Based on the data types supplied in the CSV file, Gender, Grade Level (Grade_Level), Class Participation Level (Class_Participation_Level), Motivation Level (Motivation_Level), and Previous Semester PE Grade (Previous_Semester_PE_Grade) were treated as categorical predictors and one-hot encoded. Age, fitness scores, body mass index (BMI), Attendance Rate (Attendance_Rate), and Hours of Physical Activity per Week (Hours_Physical_Activity_Per_Week) were treated as numeric predictors and standardized. No undocumented ordinal distances between categorical labels were assumed because the public data card did not provide a validated coding dictionary.

Class Imbalance and Model Validation

The performance outcome was substantially imbalanced, with the high performer category representing only a small proportion of the dataset. Rather than applying oversampling or undersampling to the complete dataset, class-weighted learning was implemented during model fitting. This approach increased the relative contribution of underrepresented classes without generating synthetic observations outside the training folds.

Model performance was estimated using stratified five-fold cross-validation with shuffle=True and random_state=42. Stratification was used to preserve the relative distribution of the three performance categories across folds. Given the small number of high performer observations, each validation fold contained only a small number of cases from this category, and this limitation was considered when interpreting class-specific performance and variability. No resampling was applied. Class weighting was implemented only during model fitting so that validation observations did not contribute to the construction or weighting of the training data.

Model Specification and Reproducibility

Four classifiers were evaluated without post hoc hyperparameter optimization: Random Forest, Decision Tree, Support Vector Machine (SVM) with a radial basis function (RBF) kernel, and Logistic Regression. Random Forest was specified with 500 trees, class_weight='balanced', random_state=42, and n_jobs=-1. Decision Tree used class_weight='balanced' and random_state=42. SVM used an RBF kernel, class_weight='balanced', and random_state=42. Logistic Regression used class_weight='balanced', random_state=42, and max_iter=5000.

Hyperparameter optimization was not performed because the study was designed as a transparent exploratory comparison of baseline classification algorithms rather than as a model-optimization study. Avoiding an additional model-selection layer was also considered appropriate given the very small minority class.

The original analysis record did not preserve the exact historical Python and package patch versions. For final verification of the manuscript, the complete analytical workflow was rerun using Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, and Matplotlib 3.10.8. These versions document the final verification environment and are not presented as the exact software versions used in any earlier analysis.

Evaluation Metrics

Model performance was evaluated using accuracy, balanced accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. Because macro-averaging assigns equal weight to each outcome category, it was considered particularly informative in the presence of severe class imbalance.

Class-specific precision, recall, and F1-scores were also calculated from out-of-fold predictions. A confusion matrix was generated for the model achieving the highest macro-F1 score. Approximate 95% confidence intervals for cross-validated mean performance metrics were calculated from the five fold-level estimates using a t distribution with four degrees of freedom. These intervals represent variability across the five validation folds and should not be interpreted as population-level confidence intervals.

Dataset Description

The candidate variables were organized into demographic, physical-fitness, engagement/behavioral, academic/history, and outcome domains. Gender, grade level, class participation level, motivation level, and previous semester PE Grade were treated as categorical variables. Age, strength score, endurance score, flexibility score, speed and agility score, BMI, health and fitness knowledge score, skills score, attendance rate, and hours of physical activity per week were treated as numeric variables.

Overall PE performance score, improvement rate, and final grade were excluded during the leakage audit, while ID was excluded as a non-substantive identifier. Performance was retained as the three-category outcome variable. Because the public data card does not provide complete coding dictionaries for all variables, the original category coding and the thresholds used to generate the performance outcome could not be independently reconstructed.

Data Integrity and Outlier Screening

Before model fitting, the dataset was screened for missing values, exact duplicate rows, duplicate identifiers, duplicate substantive records, constant columns, categorical consistency, numeric data-type consistency, and broad range plausibility. Univariate outliers were screened using Tukey's rule, whereby observations below (Q1 - 1.5 times IQR) or above (Q3 + 1.5 times IQR) were flagged for further inspection. The interquartile range (IQR) procedure was used as a screening tool rather than as an automatic exclusion criterion because statistical extremity alone does not establish that an observation is erroneous or invalid.

Observations identified by the outlier-screening procedure were therefore reviewed for potential data-integrity problems and were excluded only if there was evidence of data-entry error or implausibility. The results of these data-integrity and outlier checks are reported in the Results section.

Data Analysis

The analysis proceeded sequentially through data-integrity screening, predictor selection, preprocessing, model fitting, cross-validation, and performance evaluation. Following the leakage audit, the retained categorical and numeric predictors were processed within the cross-validation pipeline to prevent information from the

validation folds from influencing training-data preprocessing. Class imbalance was addressed through class-weighted learning within each classifier.

The four classifiers were evaluated using stratified five-fold cross-validation. No hyperparameter optimization was performed because the objective was to compare transparent baseline classification approaches rather than to identify an optimized predictive model. Performance was assessed using accuracy, balanced accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and class-specific performance measures. Out-of-fold predictions were used to construct the confusion matrix and evaluate classification behavior across the three performance categories.

Cross-validated performance estimates were summarized using fold-level means, standard deviations, and approximate 95% confidence intervals. In addition to classification performance, impurity-based feature importance was examined using the random forest model fitted to the full dataset with the leakage-controlled predictor set. For categorical predictors represented by multiple one-hot-encoded variables, the Gini importance values of the corresponding dummy variables were summed to obtain an aggregate importance value for the original predictor. Feature importance was interpreted descriptively and was not considered evidence of causal relationships.

All analyses were conducted using Python and the machine-learning workflow described above.

RESULTS

Verified Dataset Characteristics

The supplied dataset contained 500 records with no missing values. The Performance distribution comprised 245 Low Performers (49.0%), 243 Average Performers (48.6%), and 12 High Performers (2.4%). Data-integrity screening identified no exact duplicate rows, duplicate IDs, or duplicate substantive records. The univariate interquartile range (IQR) screening flagged 35 records across the numeric variables; none were excluded because the flagged values did not indicate evident data-entry errors or implausible observations. The 500 records should not, however, be interpreted as a verified empirical sample of students because the provenance of the publicly available dataset could not be independently established.

Cross-Validated Model Performance

Across five stratified cross-validation folds, Random Forest achieved the highest mean accuracy (0.856), whereas Logistic Regression achieved the highest mean macro-F1 (0.645) and balanced accuracy (0.749). Despite achieving the highest overall accuracy, Random Forest did not correctly identify any of the 12 High Performer records in the out-of-fold predictions, indicating poor sensitivity to the minority class. Decision Tree achieved a macro-F1 of 0.600, while Support Vector Machine (SVM) achieved a macro-F1 of 0.580. Approximate 95% fold-level t-based confidence intervals for macro-F1 were 0.553–0.601 for Random Forest, 0.498–0.702 for Decision Tree, 0.477–0.683 for SVM, and 0.541–0.749 for Logistic Regression. For Logistic Regression, the corresponding confidence interval for balanced accuracy was 0.650–0.848. These intervals indicate considerable variability across folds, particularly given the very small number of High Performer cases. Therefore, the

model with the highest overall accuracy should not automatically be interpreted as the strongest classifier when class-balanced performance is considered.

The comparative performance of the four classifiers is presented in Figure 1, while the corresponding mean values and approximate 95% fold-level t-based confidence intervals are summarized in Table 1.

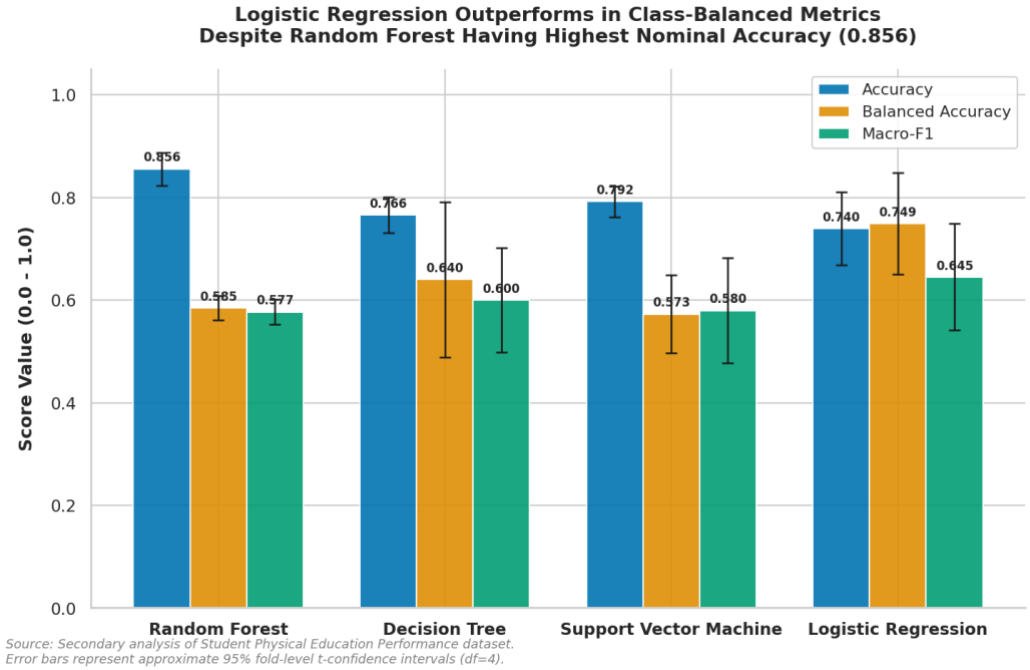


Figure 1. Five-fold cross-validated accuracy, balanced accuracy, and macro-F1 across the four classifiers. Error bars represent approximate 95% fold-level t-based confidence intervals (df = 4).

Table 1. Mean cross-validated performance with approximate 95% fold-level t confidence intervals in parentheses

Model	Accuracy	Balanced accuracy	Macro-F1
Random Forest	0.856 (0.824–0.888)	0.585 (0.561–0.609)	0.577 (0.553–0.601)
Decision Tree	0.766 (0.731–0.801)	0.640 (0.489–0.791)	0.600 (0.498–0.702)
Support Vector Machine	0.792 (0.762–0.822)	0.573 (0.497–0.649)	0.580 (0.477–0.683)
Logistic Regression	0.740 (0.669–0.811)	0.749 (0.650–0.848)	0.645 (0.541–0.749)

Class-Specific Performance of the Selected Model

Logistic Regression was selected for primary interpretation because it achieved the highest macro-F1 and balanced accuracy among the four evaluated classifiers. Its out-of-fold class-specific performance was as follows: for Low Performers, precision was 0.791, recall was 0.759, and F1-score was 0.775; for Average Performers, precision was 0.761, recall was 0.720, and F1-score was 0.740; and for High Performers, precision was 0.257, recall was 0.750, and F1-score was 0.383.

The model correctly classified nine of the 12 High Performer records. However, the relatively low precision for the High Performer class indicates that several observations classified as High Performer actually belonged to other performance categories. Given that this minority class contained only 12 records, these class-specific estimates should be interpreted cautiously.

The out-of-fold confusion matrix for the selected Logistic Regression model is presented in Figure 2. The confusion matrix further illustrates the model's class-

specific classification behavior. Although nine of the 12 High Performer records were correctly identified, several records from the other categories were incorrectly classified as High Performers, resulting in a relatively high number of false-positive classifications for this minority class. Thus, class weighting improved the model's ability to identify High Performer cases but did not eliminate misclassification into this category.

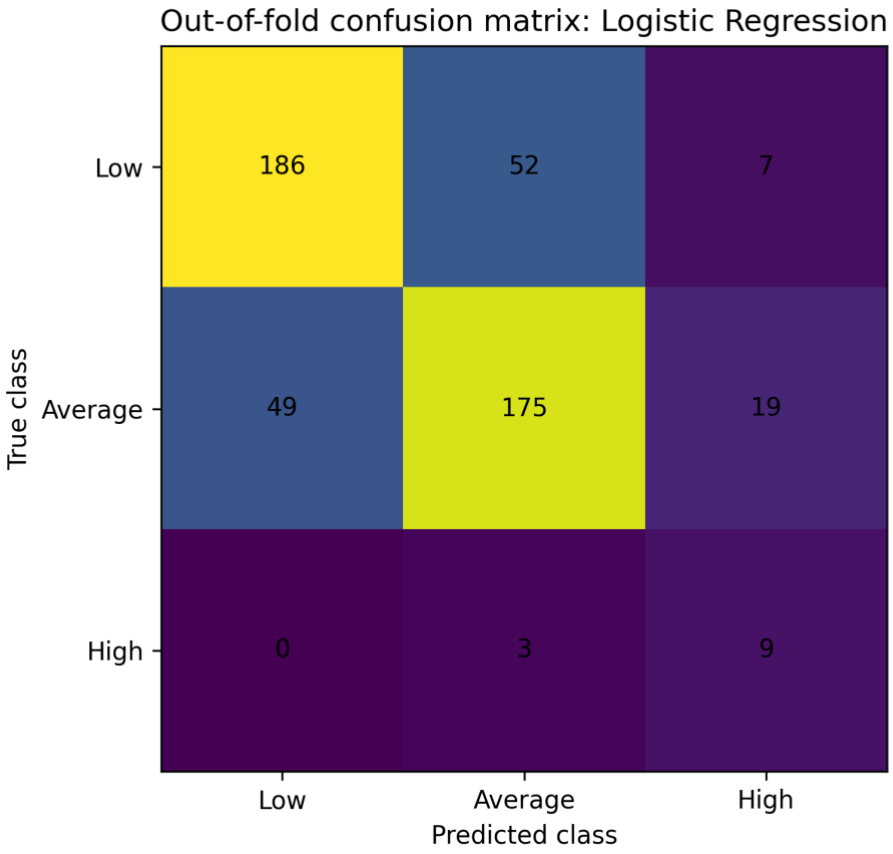


Figure 2. Out-of-fold confusion matrix for Logistic Regression.

Random Forest Feature Importance

The Random Forest impurity-based feature-importance ranking is presented in Table 2. After leakage-prone variables were excluded, Attendance Rate and Motivation Level showed the highest impurity-based importance values. Fitness-related variables, including Flexibility Score, Endurance Score, Skills Score, Speed and Agility Score, and Strength Score, also contributed to the model, although no single fitness indicator dominated the classification.

Table 2. Random Forest Feature Importance

Rank	Predictor	Gini importance
1	Attendance Rate	0.219
2	Motivation Level	0.197
3	Flexibility Score	0.075
4	Endurance Score	0.062
5	Skills Score	0.057
6	Health Fitness Knowledge Score	0.057
7	Speed Agility Score	0.051
8	Strength Score	0.050
9	Hours Physical Activity per Week	0.046
10	BMI	0.046

For categorical predictors, the Gini importance values of the corresponding one-hot-encoded dummy variables were summed to obtain an aggregate importance value for each original predictor. Motivation Level was the only categorical predictor among the ten highest-ranked predictors. The remaining predictors in the top-10 ranking were numeric variables.

These impurity-based feature-importance values represent descriptive associations within the fitted Random Forest model and should not be interpreted as causal effects. Because permutation importance was not performed in the recorded analytical workflow, the reported Gini importance rankings should be interpreted as model-specific descriptive evidence rather than as definitive measures of predictor contribution.

DISCUSSION

The leakage-controlled analysis provides a deliberately conservative interpretation of machine-learning classification within the supplied dataset. Excluding overall PE performance score, improvement rate, and final grade reduced the potential for direct or temporal overlap between the predictors and the performance outcome. This approach is consistent with established guidance indicating that target leakage can produce overly optimistic estimates when information unavailable at a legitimate prediction point is incorporated into model development (Kaufman et al., 2012). The use of stratified five-fold cross-validation and class-aware evaluation metrics is also consistent with the broader educational data-mining literature, which emphasizes careful data preparation, validation, and interpretation when modelling student performance (Xiao et al., 2022; Ersozlu et al., 2024).

The contrast between overall accuracy and class-balanced performance is particularly important in this dataset. Random forest achieved the highest overall accuracy (0.856), yet its out-of-fold recall for the high performer class was 0.000, indicating that none of the 12 high performer records were correctly identified. In contrast, logistic regression correctly identified nine high performer cases, yielding a recall of 0.750 for this minority class. However, its corresponding precision was only 0.257, indicating that several records classified as high performers actually belonged to other categories. This pattern illustrates why overall accuracy can be misleading in severely imbalanced classification problems and why class-specific measures, macro-F1, and balanced accuracy provide important complementary information (He & Garcia, 2009). In the present analysis, logistic regression therefore provided a more balanced representation of performance across the three classes despite having lower overall accuracy than random forest.

Attendance rate and motivation level showed the highest impurity-based feature-importance values in the random forest model after leakage-prone variables were excluded. This pattern is broadly consistent with educational data-mining research showing that behavioral and engagement-related indicators can contribute to the classification of student performance (Wong et al., 2024; Xiao et al., 2022; Ersozlu et al., 2024). Nevertheless, these findings should be interpreted as descriptive model associations rather than evidence of causal relationships. The impurity-based importance values are specific to the fitted random forest model, and the uncertain provenance and cross-sectional structure of the dataset preclude causal inference.

Moreover, because permutation importance was not performed, the relative ranking of predictors should not be interpreted as a definitive measure of their substantive importance.

The findings also impose important boundaries on the practical interpretation of the models. Although the evaluated algorithms may be useful for exploratory classification within the supplied dataset, the data do not provide documented longitudinal timing, a verified student population, or an independently validated target-generation process. A recent study using a matching 500-instance Kaggle PE dataset reported substantially higher model performance after applying label encoding, synthetic minority oversampling, feature selection, and hyperparameter optimization (Li & Cao, 2025). Those findings are not directly comparable with the present results because the analytical pipelines differ substantially. In particular, the present study deliberately excluded variables with potential outcome overlap and used fixed baseline model configurations to minimize leakage-related inflation and maintain analytical transparency. This contrast highlights the importance of documenting preprocessing decisions, class-imbalance handling, and leakage-control procedures when evaluating machine-learning models (Lamba, Rani, & Shabaz, 2026).

Consequently, the present findings should not be interpreted as evidence that these models can reliably predict future student performance early in a semester or are ready for deployment in educational decision-making. Establishing prospective predictive utility would require primary or independently verified educational data, clearly documented target construction, explicit temporal ordering between predictors and outcomes, prespecified prediction windows, and external validation in independent student populations.

Limitations of the study

Several limitations should be considered when interpreting the findings. First, the provenance of the dataset is incompletely documented, and the authors cannot verify whether the records represent empirical, synthetic, or simulated observations. Second, the exact rules, thresholds, and component weights used to construct the Performance target are not disclosed in the public data card. Consequently, structural relationships between the target and retained predictors cannot be completely ruled out.

Third, the dataset contains only 12 high performer records, representing 2.4% of the total sample. This severe class imbalance limits the precision and stability of minority-class performance estimates, even under stratified cross-validation and class-weighted learning. Fourth, the available documentation does not provide sufficient temporal information to confirm that every retained predictor was measured before the performance outcome. Therefore, the present models cannot be interpreted as validated prospective prediction models.

Fifth, the univariate IQR screening identified statistically unusual observations, but such screening cannot determine whether these observations are substantively valid or representative of the source population. Sixth, the original analytical record did not preserve the exact historical Python and package patch versions. The software versions reported in this manuscript therefore document the final verification environment rather than the precise environment of any earlier analytical run.

Seventh, no hyperparameter optimization was performed; therefore, the reported models should be interpreted as transparent baseline configurations rather than optimized classifiers. In addition, feature importance was assessed using Random Forest impurity-based importance without independent permutation-importance analysis, and the resulting rankings are therefore model-specific. Finally, the analysis was based on a single publicly available dataset of uncertain provenance, which substantially limits the generalizability of the findings to other schools, age groups, countries, educational systems, or PE curricula.

CONCLUSION

This leakage-controlled exploratory classification study demonstrates that machine-learning algorithms produced different performance profiles when classifying PE performance categories within the supplied dataset. Random forest achieved the highest cross-validated accuracy (0.856), whereas logistic regression achieved the highest macro-F1 (0.645) and balanced accuracy (0.749). Importantly, random forest failed to correctly identify any high performer records, whereas logistic regression achieved a high performer recall of 0.750, illustrating the importance of class-aware evaluation when outcome categories are severely imbalanced.

Data-integrity screening identified no missing values or duplicate records, while 35 records were flagged as statistical outliers using the univariate IQR criterion and were retained because no evident data-entry inconsistencies were identified. The analysis further demonstrates the importance of excluding variables with potential outcome overlap, restricting preprocessing to training folds, accounting for class imbalance, and reporting class-specific performance alongside overall accuracy.

The findings should be interpreted only as exploratory evidence from the supplied dataset and not as evidence of a deployable early-warning or prospective prediction system. Future research should use documented and ethically collected student data, explicit target-construction procedures, verified temporal ordering, larger and more balanced samples, prespecified prediction windows, independent external validation, and archived analytical environments to support reproducibility and stronger claims of predictive validity.

ACKNOWLEDGMENTS

The authors would like to thank the creators of the open Kaggle dataset used in this study for making the data publicly available for research purposes.

AI DISCLOSURE STATEMENT

During the preparation of this work, the authors used ChatGPT (OpenAI) and Grammarly to assist with language editing, grammar improvement, and manuscript organization. After using these tools, the authors carefully reviewed, revised, and verified the content and take full responsibility for the accuracy, integrity, and final content of the published article.

DATA AVAILABILITY

The dataset analyzed in this study is publicly available through the Kaggle dataset Student Physical Education Performance by Ziya07 (2024): https://www.kaggle.com/datasets/ziya07/student-physical-education-performance. The analysis was conducted using the 500-record CSV file obtained from this dataset. The SHA-256 checksum of the analyzed file is a46d3d78ec3c563788a988c7e16b9bb8be8f43a04dd38cafffe79dce4bd158a1. The public data card was last accessed on September 7, 2026. Because the available documentation does not describe the original data-collection process or the procedure used to construct the Performance target, the authors cannot independently verify the provenance of the records or the target-generation process. Accordingly, the present study reports a reproducible secondary analysis of the supplied dataset rather than a validated analysis of a known student population.

The analytical workflow is described in sufficient detail to reproduce the predictor-selection process, leakage exclusions, preprocessing procedures, class-weighting strategy, cross-validation design, data-integrity checks, model specifications, and evaluation metrics. The original analytical record did not preserve the exact historical software patch versions used in the initial analysis. Therefore, the software versions reported here represent the final verification environment: Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, and Matplotlib 3.10.8. The complete workflow was rerun in this environment and reproduced the reported cross-validated model-performance estimates and class-specific Logistic Regression results from the analyzed CSV file.

FUNDING

This research did not receive any external funding.

CONFLICT OF INTEREST

The authors declare that they have no financial or personal relationships that could have influenced the conduct or reporting of this study..

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or a claim its manufacturer may make, is not guaranteed or endorsed by the publisher.

REFERENCES

Barbeiro, L., Gomes, A., Correia, F. B., & Bernardino, J. (2024). A review of educational data mining trends. *Procedia Computer Science, 237*, 88–95. https://doi.org/10.1016/j.procs.2024.05.083

Bellaj, M., Ben Dahmane, A., Boudra, S., & Lamarti Sefian, M. (2024). Educational data mining: Employing machine learning techniques and hyperparameter optimization to improve students' academic performance. *International Journal*

of Online and Biomedical Engineering, 20(3), 55–74. https://doi.org/10.3991/ijoe.v20i03.46287

Ersozlu, Z., Taheri, S., & Koch, I. (2024). A review of machine learning methods used for educational data. *Education and Information Technologies, 29*, 22125–22145. https://doi.org/10.1007/s10639-024-12704-0

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15, 1–21. https://doi.org/10.1145/2382577.2382579

Lamba, K., Rani, S., & Shabaz, M. (2026). Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability. *Discover Mental Health*, *6*(1), 25. https://doi.org/10.1007/s44192-026-00373-z

Li, C., & Cao, Z. (2025). Deep learning-based AI model for predicting academic success and engagement among physical higher education students. *Scientific Reports, 15*, 45471. https://doi.org/10.1038/s41598-025-29000-7

Lieberman, L. J., Houston-Wilson, C., & Grenier, M. (2024). *Strategies for inclusion: Physical education for everyone*. Human Kinetics. https://doi.org/10.5040/9781718260023

Manso-Lorenzo, V., Evangelio, C., Ruiz-Tendero, G., & González-Víllora, S. (2024). Game performance assessment instrument in physical education: a systematic review from 2015 to 2024. *Physical Education and Sport Pedagogy*, 1-16. https://doi.org/10.1080/17408989.2024.2432313

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of educational psychology*, *116*(1), 48. https://doi.org/10.1037/edu0000833

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports, 4*(5), e12482. https://doi.org/10.1002/eng2.12482

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Zhou, T., Wu, X., Wang, Y., Wang, Y., & Zhang, S. (2024). Application of artificial intelligence in physical education: A systematic review. *Education and Information Technologies, 29*(7), 8203–8220. https://doi.org/10.1007/s10639-023-12128-2

Ziya07. (2024). Student Physical Education Performance [Data set]. Kaggle. https://www.kaggle.com/datasets/ziya07/student-physical-education-performance