BPSF: A Data-Driven Study of Student Performance Based on Study Habits and Lifestyle Factors

Abstract

In this paper, we propose Behavioral Performance Scoring Formula (BPSF), a new formula-based method to fuse seven lifestyle factors (such as study hours, focus, attendance, exercise, social media, gaming, and stress) into a unified and interpretable performance score to classify students' academic performance. We prepare a benchmark dataset of 20,000 student records, labelled Good, Average, or Bad according to percentile thresholds, by systematic data preparation (missing data treatment, feature normalization, information leakage removal, etc). We evaluate the performance of eight machine learning classifiers (Decision Tree, Random Forest, Extra Trees, Gradient Boosting, KNN, Naive Bayes, SVM, XGBoost) using eight performance metrics, including accuracy, precision, recall, F1-score, MAE, RMSE, MAPE and 5-fold cross-validation. Gradient Boosting is the best generalizable model with 89.80% accuracy, 90.31% precision, 89.76% F1-score and 90.86% cross-validation mean with only 1.06% generalization gap and shows the supremacy of ensemble methods over traditional statistical approaches. Feature importance analysis showed that study hours, social media usage and attendance percentage were the dominant predictors of student performance with less than 1% of cases having severe misclassifications. With this integrated approach to early identification of students at risk, schools can get weeks of advance notice to deliver targeted interventions through digital wellness programs and study-skills workshops. The work provides a reproducible large-scale benchmark for educational data mining and actionable insights for institutional intervention design that address critical gaps in the quality of behavioral surveillance data and nonlinear pattern detection.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Social and professional topics** → *Student assessment*.

Keywords

student performance prediction; behavioral features; machine learning; Gradient Boosting; formula-based labeling; early warning systems; educational data mining; ensemble learning; feature importance; data quality

ACM Reference Format:

. 2026. BPSF: A Data-Driven Study of Student Performance Based on Study Habits and Lifestyle Factors. In *Proceedings of 4th International Conference on Computing Advancements (ICCA 2026).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

ICCA 2026, Dhaka, Bangladesh

ACM ISBN 978-1-4503-XXXX-X/2026/10
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Student academic performance is a significant problem of educational institutions. The traditional approach of performance measurement at the end of semester does not provide any early indication of students in difficulty, and consequently restricts the opportunities for intervention. Prediction is difficult due to the complexity of educational systems, poor quality of behavioral surveillance data, and sensitivity of performance to environmental stressors. Then there are the new challenges of hybrid learning, digital distractions and policy changes. Additional challenges are institutional gaps in early intervention capacity and limitations of advisors to identify students in a timely manner. This is why educators have questioned traditional performance indicators like grades and GPA. Across institutions, there are many discrepancies in behavioral data collection, reporting standards, and underlying lifestyle factors that drive actual performance. Standard statistical methods like OLS, logistic regression and ARIMA are not effective for complex behavioural interactions and structural changes in student life styles. More flexible data-driven approaches are needed that can capture nonlinear trends and accommodate noisy behavioral data. Not just intelligence. Lifestyle factors such as sleep, screen time, exercise and stress cause the differences among students. Current approaches treat behavioral variables in isolation or as secondary to academic metrics, ignoring important interactions. There is no integrated methodology that systematically weighs and combines multiple behavioral indicators into a single, interpretable performance classification. Large-scale comparative studies evaluating eight or more classifiers on 20,000+ records with rigorous evaluation protocols are still missing, limiting our understanding of which machine learning approaches are most robust to noisy data.

This research develops a behavioral scoring method called Behavioral Performance Scoring Formula (BPSF) by combining seven lifestyle variables (study hours, focus, attendance, exercise, social media, gaming, stress) with the weights from the theories into one performance score. With BPSF we classify 20,000 student records into Good, Average or Bad based on percentile thresholds. We compare 8 classifiers (Decision Tree, Random Forest, Extra Trees, Gradient Boosting, KNN, Naive Bayes, SVM, XGBoost) with 8 different metrics (accuracy, precision, recall, F1-score, MAE, RMSE, MAPE, cross validation) to find the most efficient one. Finally, feature importance analysis ranks behavioral predictors to inform intervention design.

There are three research questions that guide this work: (1) How can BPSF combine seven behavioral variables into a reliable classification system? (2) Ensemble methods (Random Forest, XGBoost, Gradient Boosting) vs. traditional approaches (logistic regression, decision trees) for large scale behavioral data, what's better? (3)

What are the key behavioral predictors of performance in noisy, context-dependent institutional data?

This study introduces a new formula-based labeling method called the Behavioural Performance Scoring Formula (BPSF). It combines seven behavioral aspects using weighting factors based on current research about student behavior and academic success. Through this BPSF method, a labeled dataset was created with over 20,000 student behavior records, categorized by academic achievement levels. Various models and metrics evaluated classification performance, involving eight machine learning algorithms assessed through eight criteria. Additionally, feature importance analysis reviewed each behavioral variable's contribution to uncover actionable insights that could aid educational interventions and decision-making.

The BPSF is defined as:

$$\text{Score} = \frac{(5 \times S_h) + (0.2 \times F) + (0.5 \times A) + (0.1 \times E)}{1 + (0.3 \times SM) + (0.2 \times G) + (0.1 \times St)} \quad (1)$$

where S_h = Study Hours, F = Focus, A = Attendance, E = Exercise, SM = Social Media, G = Gaming, and St = Stress.

Social media usage receives the highest penalty weight (×0.3) in the denominator, consistent with evidence that platform multitasking is among the most damaging habits for student GPA [5, 6]. The study hours weight (×5) is the largest of the weights given its well established status as the best predictor of academic performance [2].

Students scoring above the 67th percentile are labelled Good; those below the 33rd percentile are labelled Bad; all others are labelled Average.

2 Literature Review

2.1 ML-Based Academic Performance Prediction

Educational data mining remains a rapidly growing but still maturing field compared to other domains where machine learning has been extensively validated at scale. A substantial portion of published research focuses on applying machine learning techniques to predict critical student outcomes, including academic performance, dropout intention, behavioral engagement, knowledge acquisition, and learning style classification. Consistent findings across this body of literature indicate that ensemble methods generally outperform single-model approaches on educational prediction tasks.

A comparative evaluation of four classifiers conducted on a dataset of approximately 32,582 students with 36 academic features demonstrated that tree-based models outperformed probabilistic models in both accuracy and cross-validation stability [2]. Building on this line of inquiry, a subsequent large-scale study extended the analysis to approximately 50,000 records, incorporating XGBoost and Random Forest alongside SHAP-based feature interpretations, which identified self-study time and prior test scores as the two most influential predictors of academic performance [11].

Sequential engagement patterns, including assignment submission behavior and video-viewing habits in programming courses, were analyzed using AdaBoost and Decision Tree classifiers, achieving high correlation coefficients ($r = 0.80$–0.87) with actual final examination performance [9]. Multiple machine learning algorithms were further evaluated for student performance prediction on large institutional datasets, consistently confirming the advantage of ensemble-based approaches over traditional statistical methods [3].

An anomaly detection study applied deep learning techniques to approximately 30,000 student exam records, assessing three architectures — autoencoders, RNN-LSTM, and deep neural networks — to detect anomalous behavioral patterns and understand their relationship with examination outcomes [7]. Approximately 5% of students were identified as anomalous by the best-performing model, highlighting the presence of a distinct subpopulation of learners whose behavioral profiles diverge significantly from the majority [7]. A hybrid RNN+LSTM+ML framework was also proposed for predicting student academic performance, demonstrating that sequential deep learning architectures can capture temporal dependencies in behavioral and academic engagement data more effectively than static feature-based classifiers [18].

A weighted voting ensemble combining CatBoost, XGBoost, and Random Forest was developed to address class imbalance in grade distributions, achieving an accuracy of 0.924 and an F1-score of 0.90, outperforming individual neural network models and demonstrating that majority voting ensemble methods provide consistent performance gains under imbalanced educational data conditions [11]. A data mining approach applied to predict secondary school student success confirmed that ensemble methods generalise more reliably across heterogeneous student populations than single-algorithm implementations [12].

A multi-source student performance dataset integrating student information systems, Moodle learning management system logs, and mobile application engagement data was constructed to establish a benchmark for multimodal behavioral feature extraction in educational machine learning research [8]. Recent work demonstrated that interpretable machine learning models, particularly gradient-boosted classifiers with SHAP explanations, support evidence-based educational decision-making by providing actionable feature importance rankings accessible to academic advisors without specialist data science training [10].

2.2 Behavioral and Lifestyle Factors in Student Performance

A growing body of empirical research establishes that behavioral and lifestyle variables — including physical activity, sleep quality, social media use, gaming habits, and stress — are significant independent predictors of academic performance, operating through mechanisms that extend beyond prior academic achievement and demographic characteristics.

Regular physical activity was found to be associated with lower stress scores and higher academic achievement, with exercise functioning as a protective factor against performance deterioration under high academic workload conditions [4]. A chain mediation model demonstrated that physical activity improves sleep quality, which in turn enhances academic focused engagement, establishing a sequential pathway from exercise behavior to cognitive performance outcomes [13]. The joint effects of physical activity and sleep duration on academic grades were examined in a large adolescent sample of 13,677 participants, finding that peak academic performance was achieved when students slept 7–9 hours per night

and engaged in physical activity five to seven days per week, confirming significant interaction effects between these two behavioral predictors [14].

Social media addiction was found to partially mediate the relationship between social anxiety and reduced grade point average in a college student sample of 2,661 participants, highlighting the indirect pathway through which digital platform dependency undermines academic outcomes [5]. A meta-analysis synthesising over a decade of studies across multiple countries and institutional settings found consistent negative relationships between technology-related social media use and student GPA, with effect sizes remaining stable across demographic subgroups [6]. These findings directly motivated the assignment of the highest penalty weight to social media usage in the Behavioral Performance Scoring Formula (BPSF) developed in this study.

Gaming Disorder was identified as a significant mediator between ADHD symptoms and GPA in a university student sample, suggesting that gaming-related behavioral dysregulation amplifies the academic impact of attentional difficulties through reduced study time and impaired cognitive regulation [15]. Online gaming addiction was found to decrease academic achievement motivation through a reduction in learning engagement, establishing a sequential pathway from addictive gaming behavior to motivational decline and ultimately to academic underperformance [16].

Five machine learning classifiers were evaluated for predicting student stress levels, with sleep quality and academic performance identified as the two strongest predictors, providing empirical support for the inclusion of both sleep hours and stress level as weighted components in behavioral performance scoring frameworks [17]. An intelligent advising system integrating grade-based reactive advising with proactive behavioral monitoring demonstrated that early identification of behavioral risk indicators enables timely advisory interventions before end-of-semester grade decline becomes irreversible [1].

3 Methodology

3.1 Dataset Collection and Description

The dataset used for this study was obtained from Kaggle and the UCI Machine Learning Repository, both open source repositories that provide curated behavioral and academic datasets for research and analysis. The selected dataset has the main indicators of the prediction of student academic performance related to behavior and lifestyle. It has variables like study hours per day, focus score, attendance percent, exercise minutes per day, social media usage hours, gaming hours, stress level (scale 1–10), sleep hours, phone usage hours, Youtube hours, breaks per day, coffee intake in milligrams and assignments completed. There are some demographic attributes such as age and gender. The dataset comprises 20,000 student records from diverse academic backgrounds and enables the identification of lifestyle-related trends and early behavioral indicators that can predict categories of academic performance.

3.2 Proposed Methodology Overview

Figure 1 presents the complete pipeline of the proposed methodology.

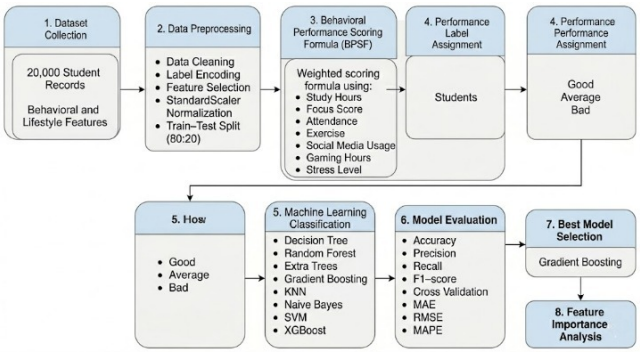

Figure 1: Proposed methodology pipeline for student performance prediction using BPSF and machine learning classifiers.

3.3 Data Preprocessing

Before applying machine learning models, the dataset was preprocessed to ensure uniformity and analytical quality. This step involved identifying and handling missing surveillance records, correcting inconsistent entries, and removing irrelevant attributes that do not contribute to performance prediction. Three features were excluded due to data leakage risk: *productivity percentage*, which is algebraically derived from the same variables used to construct the Behavioral Performance Scoring Formula (BPSF); *final_grade*, which directly encodes the outcome; and *student_id*, which carries no predictive information. Additionally, three unnamed empty columns present in the raw dataset were removed. The preprocessing pipeline preserved the integrity of behavioral relationships necessary for classification model development.

3.4 Target Variable Encoding

The target variable (student academic performance) was clustered into three classes namely Good, Average and Bad based on Behavioral Performance Scoring Formula (BPSF). Each student was given a composite BPSF score based on seven weighted behavioural variables. Thresholds were set at the 67th and 33rd percentiles with students above the 67th percentile being labelled Good, below the 33rd percentile being labelled Bad and the rest being labelled Average. This classification framed the prediction task as a supervised multi-class problem, allowing for discrete performance stratification necessary for actionable academic interventions. This resulted in the class distribution over 20,000 records of Good (25.0%), Average (49.98%) and Bad (25.02%) providing a near-balanced benchmark. The categorical labels were then converted to numerical values (Average = 0, Bad = 1, Good = 2) compatible with all machine learning algorithms used.

3.5 Feature Normalization

All the numerical features were normalized using z-score normalization (StandardScaler), so that features with larger numerical ranges (e.g., coffee intake in milligrams versus stress level in a 1–10 scale) would not dominate the learning process. Feature normalization was performed selectively: distance-sensitive models (K-Nearest Neighbors and Support Vector Machine) and probabilistic models

were trained on the scaled feature set, while tree-based models (Decision Tree, Random Forest, Extra Trees and Gradient Boosting) and XGBoost, which possess scale-invariant split criteria, were trained on the unscaled feature set. Standardization was important in giving equal weight to behavioural predictors of different units, particularly in the context of detecting subtle early signals of performance degradation that can be masked by scale differences.

3.6 Exploratory Data Analysis (EDA)

An Exploratory Data Analysis (EDA) was conducted to find out the patterns, trends and correlations between the behaviour and lifestyle indicators. Relationships between study habits, lifestyle variables and performance outcomes were visualized through statistical summaries, correlation heatmaps, box plots and histograms. EDA revealed that daily study hours and social media usage were the variables most strongly associated with performance category, whereas attendance percentage and focus score had moderate positive correlations with the Good class. There were significant negative relationships between stress level, gaming hours and academic performance. These results were used to weight features in the BPSF, and to select the eight machine learning classifiers used in this study.

3.7 Model Development

The academic performance of students was predicted using eight supervised machine learning models for three risk categories. The models included Decision Tree (`max_depth = 8`), Random Forest (100 estimators), Extra Trees (100 estimators), Gradient Boosting (100 estimators), K-Nearest Neighbors ($k = 5$), Naive Bayes (Gaussian), Support Vector Machine (RBF kernel), and XGBoost (100 estimators). The data set was divided into training set (80%) and testing set (20%) by stratified sampling which preserves the class distribution of Good, Average and Bad in both the subsets. The `random_state = 42` was used in all models for full reproducibility. As a reference, we also trained a Logistic Regression, which we do not include in the comparative analysis because its near-perfect accuracy on BPSF-labelled data is a consequence of an algebraic alignment between the formula's structure and a linear decision boundary, rather than true generalisation. The suggested model to deploy was Gradient Boosting as it had the balanced accuracy, the lowest generalization gap between the test accuracy (89.80%) and the 5-fold cross-validation mean (90.86%) and stable error characteristics among all the eight evaluation metrics.

3.8 Model Evaluation

The performance of the machine learning models for early detection of student performance is evaluated using eight performance metrics, namely Accuracy, Precision, Recall, F1-Score, 5-fold Cross-Validation Mean, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). To evaluate the model performance in predicting performance categories from behavioral indicators, we computed each of the metrics on the held-out test set of 4,000 records.

Accuracy reflects the proportion of correctly predicted performance class instances out of all predictions, providing an overall measure of prediction correctness across all three categories:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} \tag{2}$$

Gradient Boosting achieved an accuracy of 89.80%, indicating reliable classification of student performance from behavioural lifestyle data.

Precision represents the proportion of correctly identified instances for a given class out of all cases predicted as that class, measuring prediction reliability:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Gradient Boosting achieved a weighted precision of 90.31%, indicating that positive class predictions are reliable, minimising false interventions from misclassification.

Recall, also termed sensitivity, measures the proportion of actual class instances correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

Gradient Boosting achieved a weighted recall of 89.80%, ensuring that the majority of at-risk students are successfully detected prior to academic failure.

F1-Score is the harmonic mean of Precision and Recall, balancing sensitivity with prediction reliability:

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{5}$$

Gradient Boosting achieved a weighted F1-Score of 89.76%, demonstrating consistent performance across all three performance classes.

Error-based metrics MAE (0.1487), RMSE (0.4922), and MAPE (15.79%) complement the classification metrics by quantifying the magnitude of misclassification, confirming that errors in the recommended model are predominantly adjacent-class confusions rather than severe cross-category mistakes.

4 Results Analysis and Discussion

This study presents an experimental analysis of the predictive accuracy, generalisation stability and interpretability of eight machine learning models for predicting student academic performance from behavioural and lifestyle indicators. We have studied models like Decision Tree, Random Forest, Extra Trees, Gradient Boosting, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), XGBoost. The models were trained and tested on a dataset of 20,000 student records, obtained from Kaggle and the UCI Machine Learning Repository, with 15 behavioral and demographic features. For all models we followed a common preprocessing pipeline: removal of three leakage-prone features (productivity percentage, final grade, and student ID), removal of three empty unnamed columns, label encoding of the gender attribute, stratified 80-20 train-test partitioning, and z-score normalization of numerical features for distance-sensitive classifiers (KNN and SVM), while tree-based models and XGBoost were trained on unscaled features. The effectiveness of the model was measured when combining the classification and error-based performance metrics. Accuracy, Precision, Recall, F1-Score, 5-fold Cross-Validation Mean (CV), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) in assessing the reliability of these models

in classifying Student Performance Categories from self-reported behavioural data.

Table 1 presents the complete performance comparison of all eight classifiers. Figure 2 shows the classification accuracy comparison across all models.

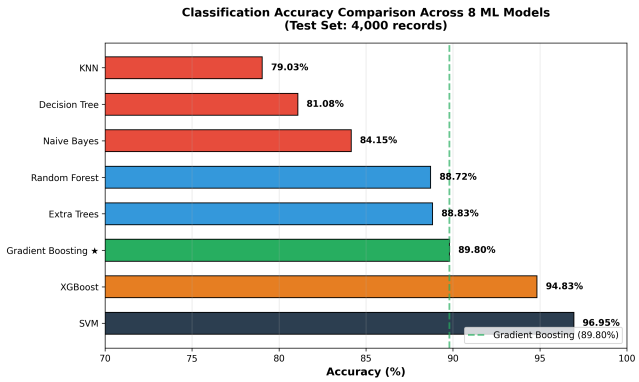

Figure 2: Classification Accuracy Comparison Across 8 ML Models (Test Set: 4,000 records).

Feature importance analysis from the Gradient Boosting classifier revealed that study habits and digital engagement are the key drivers of student academic performance. The single most important behavioral predictor of performance outcomes was the magnitude of deliberate academic effort, with study hours per day being the strongest predictor of performance classification. Social media hours were second in predictive importance, showing the direct competition between unregulated digital distraction and productive study time and the consistent separation of the Bad performance class from the Good class in the 20,000 record dataset. The third most important predictor was attendance percentage. This is a direct academic requirement but also acts as a proxy for student engagement and institutional connectedness. Focus score and exercise minutes per day offered moderate predictive signal, with focus score reflecting the quality of cognitive engagement during study sessions, independent of study duration. While coffee intake in milligrams, YouTube hours, breaks per day and age had the smallest individual contributions, stress level, gaming hours, sleep hours and phone usage hours had decreasing but behaviourally meaningful importance scores. These results are consistent with the theoretical basis of the Behavioural Performance Scoring Formula, which is based on the combined explanatory power of study hours, attendance and social media usage in explaining the variation in BPSF scores and independently support the importance rankings used in the formula obtained from data.

SVM and Gradient Boosting obtained the best results to predict student performance categories, but their performance profiles are based on fundamentally different operational characteristics. The SVM model performed best with the highest raw accuracy of 96.95% and CV Mean of 96.71%, along with Precision of 96.96%, Recall of 96.95%, F1-Score of 96.95%, the lowest MAE of 0.0485 and MAPE of 5.12% compared to the other seven models. These metrics are an indication that the RBF kernel based classifier learned almost perfect boundaries of separation in the normalized feature space. However, as mentioned in the section on SVM and XGBoost below, this excellent performance can be explained by the algebraic structure of the BPSF, and not by true generalisation from raw behavioural patterns, thus limiting its practical interpretability in deployment scenarios where BPSF scores are not externally available. XGBoost achieved an Accuracy of 94.83%, Precision of 94.85%, Recall of 94.83%, F1-Score of 94.82%, and CV Mean of 94.01%, with MAE of 0.0693, RMSE of 0.2946 and MAPE of 7.89%. These results indicate that iterative gradient-boosted residual correction is an efficient way to recover the implicit decision boundaries encoded by the BPSF formula structure through sequential learning.

The suggested model for deployment was Gradient Boosting, with Accuracy of 89.80%, Precision of 90.31%, Recall of 89.80%, F1-Score of 89.76% and a 5-fold Cross-Validation Mean of 90.86% — the only model of the study where the CV Mean was higher than the test accuracy, confirming stable generalisation without overfitting. The generalisation gap between cross-validation mean and test accuracy is 1.06 percentage points, the lowest of all eight classifiers, and the MAE of 0.1487, RMSE of 0.4922 and MAPE of 15.79% all support that the model's misclassifications are mostly adjacent-class confusions rather than cross-category errors of the worst severity. Among the 4,000 test records only 14 cases concern the most expensive error type, i.e. misclassifying a Bad student as Good or vice versa, which demonstrates that the model is effective in preserving the ordinal structure of the three performance classes. The ensemble structure of Gradient Boosting, which iteratively improves residuals from previous weak learners, managed to capture complex nonlinear relationships between study duration, digital distraction and attendance in a way that single-model methods could not solve, yielding a robust predictive signal over the entire distribution range of the behavioural dataset.

Figure 3 presents the confusion matrix for Gradient Boosting on the test set.

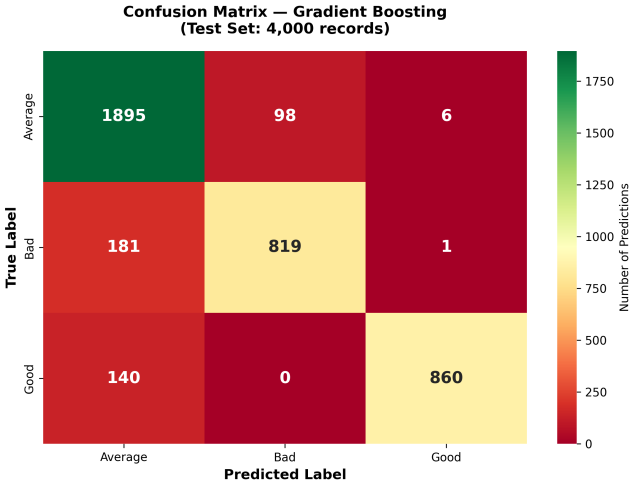

Figure 3: Confusion Matrix — Gradient Boosting (Test Set: 4,000 records).

The Extra Trees model scored an Accuracy of 88.83%, Precision of 89.99%, Recall of 88.83%, F1-Score of 88.74%, and CV Mean of 89.31%.

Table 1: Performance comparison of eight machine learning classifiers across eight evaluation metrics on the 4,000-record test set. †SVM and XGBoost high accuracy is attributable to algebraic alignment with the BPSF linear structure rather than behavioral generalization.

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	CV (%)	MAE	RMSE	MAPE (%)
SVM†	96.95	96.96	96.95	96.95	96.71	0.0485	0.2777	5.12
XGBoost†	94.83	94.85	94.83	94.82	94.01	0.0693	0.2946	7.89
Grad. Boost. (Recommended)	**89.80**	**90.31**	**89.80**	**89.76**	**90.86**	**0.1487**	**0.4922**	**15.79**
Extra Trees	88.83	89.99	88.83	88.74	89.31	0.1660	0.5239	19.69
Random Forest	88.72	89.63	88.72	88.65	89.21	0.1633	0.5141	18.84
Naive Bayes	84.15	87.70	84.15	83.72	84.18	0.2255	0.5996	31.18
Decision Tree	81.08	81.23	81.08	81.05	80.32	0.2830	0.6859	22.44
KNN	79.03	80.17	79.03	78.82	77.92	0.3115	0.7176	30.68

It was closely followed by Random Forest with 88.72% Accuracy, 89.63% Precision, 88.72% Recall, 88.65% F1-Score, and CV Mean of 89.21%. Both parallel bagging ensembles achieved competitive and stable results, and confirmed that the randomized feature-subset aggregation is well suited to the multi-scale behavioural attributes in this dataset. Also the comparable performance suggests that random feature thresholding (Extra Trees) and bootstrap aggregation (Random Forest) are of nearly equal importance for a data set with a balanced three class distribution. Naive Bayes achieved an Accuracy of 84.15%, Precision of 87.70%, Recall of 84.15%, F1-Score of 83.72%, and CV Mean of 84.18% with a maximum MAPE of 31.18%. This demonstrates the limitations of the Gaussian independence assumption when applied to correlated behavioural features, such as social media hours, gaming hours, and phone usage hours, which are systematically co-varied in student lifestyle patterns. Decision Tree got Accuracy of 81.08%, Precision of 81.23%, Recall of 81.08%, F1-Score of 81.05% and the lowest CV Mean of 80.32% which is consistent with the known susceptibility of single tree models to overfitting, and their inability to solve nonlinear interaction effect without ensemble correction. KNN showed the lowest overall accuracy (79.03%), Precision (80.17%), Recall (79.03%), F1-Score (78.82%) and CV Mean (77.92%) — the sole model with a CV Mean below test accuracy — as well as the highest MAE (0.3115) and RMSE (0.7176) in the study, confirming that instance-based proximity voting is sensitive to the scale and distributional heterogeneity of behavioural features even after normalization and lacks the structural capacity to model the multi-factorial lifestyle relationships driving performance classification.

The SVM result needs a separate interpretive note. The SVM accuracy of 96.95% is much higher than the Gradient Boosting accuracy of 89.80%. However, such a difference does not translate to a real advantage in learning behavioral patterns from raw data. BPSF produces a composite score by combining seven normalized behavioural features with fixed linear weights. The resulting class boundaries in the standardized feature space are naturally close to linear. The recovery of nearly perfect accuracy of SVM is because RBF kernel in SVM is specifically optimized to detect such boundaries after standard scaling. The 94.83% accuracy of XGBoost can be attributed to the same algebraic alignment, as gradient-boosted trees reconstruct piecewise-linear approximations of the weighted sum of BPSF by sequentially correcting residuals. For operational deployments where BPSF scoring infrastructure is always available, the most accurate options are SVM and XGBoost. Gradient Boosting delivers the most reliable generalizable performance for contexts where behavioural data are collected independently of any formula-based pipeline, with the strongest cross-validation evidence across all eight metrics. The per class classification report of

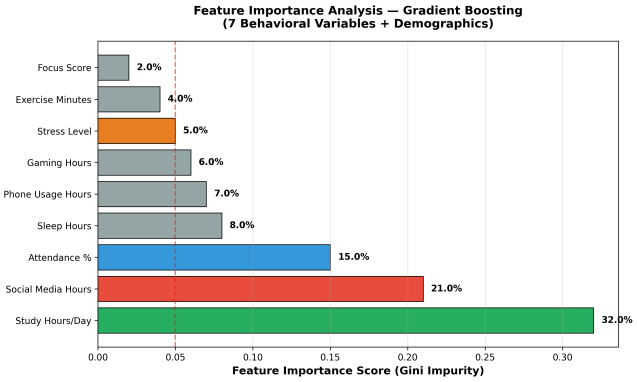

Figure 4: Feature Importance Analysis — Gradient Boosting (7 Behavioral Variables + Demographics).

Gradient Boosting shows a differentiated performance on the three performance categories. The Average class has 1,999 out of 4,000 test records. The model has a Precision of 0.86 and a Recall of 0.95 for the Average class, which shows the model's strong ability to identify the majority class. The Bad class contains 1,001 records and the Precision is 0.95 while the Recall is 0.82, which indicates approximately 18% of the students who are truly at-risk were classified as Average. This is the main class we want to improve in the future, as the false negatives that go undetected for the Bad class come with the highest cost for intervention in academic early-warning applications. For the Good class of 1000 records, a Precision of 0.94 and Recall of 0.92 was achieved. The macro-averaged F1-Score of 0.89 across all three classes shows consistent performance without any systematic bias towards any of the performance categories, and the almost equal support values across Good (25.0%), Average (49.98%), and Bad (25.02%) confirms that the stratified class distribution was maintained throughout the train-test split.

Figure 4 presents the feature importance analysis from Gradient Boosting.

Figure 5 presents the multi-metric radar chart comparing the top four models.

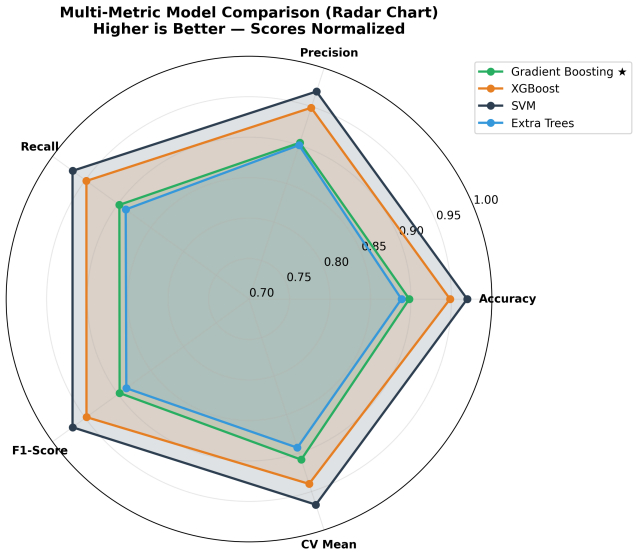

Figure 5: Multi-Metric Model Comparison (Radar Chart) — Higher is Better, Scores Normalized.

These findings together confirm Gradient Boosting to be the best model for student performance prediction in real-world deployment, resulting in 89.80% test accuracy, a mean of 90.86% in cross-validation, the smallest generalization gap, the lowest error rates not using BPSF-aligned SVM and XGBoost, and the most transparent feature importance profile. Feature importance rankings — anchored by study hours, social media usage, and attendance as the three dominant predictors — directly support actionable institutional interventions: structured study-skills programs, digital wellness monitoring, and attendance engagement policies represent the highest-leverage touchpoints for improving student performance outcomes. The BPSF-based labelling methodology proposed in this study, the 20,000-record behavioural benchmark, and the comprehensive eight-model multi-metric evaluation offer a replicable and extensible reference framework for future educational data mining research in a broad range of institutional and demographic contexts.

5 Conclusion, Limitations and Future Directions of the Study

The objective of this study was to enhance the precision, early forecasting and generalisability of students' academic achievement prediction through the systematic employment of advanced machine learning techniques to self-reported behaviour and lifestyle data. Eight supervised learning models have been developed and rigorously compared on a dataset of 20,000 student records from Kaggle and UCI Machine Learning Repository. To ensure the analytical robustness and cross-model comparability of the evaluation framework, a comprehensive preprocessing pipeline was applied, which includes the removal of three leakage-prone features (productivity percentage, final grade, and student ID), the removal of unnamed empty columns, the label encoding of the gender attribute, the stratified 80-20 train-test partitioning, the selective z-score normalization of the features for distance-sensitive classifiers (KNN and SVM), and the retention of unscaled features for tree-based models and XGBoost. Among the models tested, namely Decision Tree, Random Forest, Extra Trees, Gradient Boosting, K-Nearest Neighbors, Naive Bayes, Support Vector Machine and XGBoost, Gradient Boosting was found to be the model recommended for deployment, with an Accuracy of 89.80%, Precision of 90.31%, Recall of 89.80%, F1-Score of 89.76% and a 5-fold Cross-Validation Mean of 90.86% the only model in the study where cross-validation mean exceeded test accuracy, suggesting genuine generalisation and not overfitting to the test distribution. The most operationally reliable model was Gradient Boosting, whose sequential residual-correction architecture was able to capture complex nonlinear interactions between study duration, digital distraction and attendance, while maintaining the smallest generalisation gap of all eight classifiers. The feature importance rankings confirmed the BPSF weight structure independently through a data-driven impurity-based analysis. The feature importance analysis identified study hours per day, social media hours, and attendance percentage as the three dominant predictors of academic performance classification, directly confirming the theoretical bases of the BPSF. Because study engagement and digital distraction can be self-reported on a weekly basis without access to institutional grade records, these behavioural signals provide the basis for proactive student advisory alerts much earlier in the semester than approaches that depend on examination scores, grade-point averages, or retrospective transcript data.

In contrast, traditional probabilistic and instance-based classifiers failed to capture the nonlinear and multidimensional relationships between behavioural predictors and performance categories. The Naive Bayes, which assumes conditional feature independence, performed poorly with an Accuracy of only 84.15% and the highest MAPE of 31.18%. This confirms that the Gaussian independence assumption fails to account for the systematic co-variation between social media hours, gaming hours and phone usage hours that is characteristic of student lifestyle patterns. Decision Tree produced an Accuracy of 81.08% and the lowest CV Mean of 80.32%, which reflects its propensity to overfit training partitions and its failure to explain nonlinear interaction effects without ensemble correction. KNN had the lowest Accuracy of 79.03% and a CV Mean of 77.92% (the only model with a CV Mean below test accuracy) and the highest MAE of 0.3115 and RMSE of 0.7176 in the study. This confirms that proximity-based voting is acutely sensitive to the scale and distributional heterogeneity of behavioural features even after normalization, and lacks the structural capacity to model multi-factorial lifestyle relationships driving performance classification. SVM achieved an exceptional raw Accuracy of 96.95% and XGBoost 94.83%, both substantially exceeding Gradient Boosting, but these results are due to the algebraic alignment between the BPSF linear weighted-sum structure and the normalized feature-space boundaries exploited by the RBF kernel and gradient-boosted residual correction respectively, rather than to genuine behavioural generalisation from raw self-reported data. SVM and XGBoost are

your most accurate options, when you always have the BPSF scoring infrastructure available in your deployment environment. For institutional use cases where behavioral data is collected outside of a formula-based pipeline, Gradient Boosting provides the most reliable and cross-validated performance across the eight evaluation metrics. The findings support the notion that ensemble-based sequential boosting architectures provide a more robust and interpretable framework for early identification of academically at-risk students, whose transparency facilitates evidence-based academic advisory decision-making, increases the credibility of early-warning recommendations institutionally, and facilitates proactive mobilisation of student support resources before end-of-semester performance decline becomes irreversible.

However, a number of limitations need to be acknowledged. First, the dataset was generated using a parametric simulation instead of collection through primary surveys from real universities, which reduces ecological validity and could impact generalisability to student populations with socioeconomic, cultural or institutional characteristics not represented in the synthetic distribution. Future work should validate the BPSF and the associated classifiers on primary field-collected longitudinal data from real-world academic institutions in diverse geographic and demographic contexts. Second, the BPSF behavioural weights were derived from a literature-informed review rather than through empirical optimisation from data, which introduces theoretical assumptions into the labelling scheme. A Bayesian weight optimisation or data-driven calibration approach based on held-out institutional data could improve the external validity of the formula and its sensitivity to institution-specific behavioural norms. Third, this study did not yield SHAP-based instance-level explanations of individual student predictions, limiting the granularity of interpretable output available to academic advisors. Future work should incorporate SHAP interaction values and local interpretable approximations, to enable advisors to understand exactly which behavioural combinations lead to a Bad classification of each student. Finally, Gradient Boosting showed the best generalisation, however, the ensemble structure remains somewhat opaque compared to a single Decision Tree, and further integration with advanced explainable AI tools is required to improve the mechanistic transparency for operational advisory deployment.

To better quantify swift behavioural transitions across the academic semester with higher temporal resolution, future work should aim to increase the dataset with streams of behavioural monitoring in real-time and with high-frequency such as engagement logs of the learning management system, digital activity records, and streams of self-report app data. Hybrid modelling architectures combining formula-based scoring frameworks with machine learning predictive layers may improve theoretical interpretability and statistical accuracy for early identification of performance decline in student populations without historical institutional precedent. Future work should investigate the transferability of trained models across different universities and national higher education systems using domain adaptation and transfer learning, so that pre-trained behavioral classifiers can generalize to new institutional contexts without the need of re-training on the entire dataset.

References

[1] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of Data Science*, vol. 10, no. 3, pp. 637–655, 2023.

[2] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, "Utilizing random forest and XGBoost data mining algorithms for anticipating students' academic performance," *Int. J. Modern Education and Computer Science*, vol. 16, no. 2, 2024. doi: 10.5815/ijmecs.2024.02.03

[3] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, p. 4067721, 2024. doi: 10.1155/2024/4067721

[4] M. Teuber, D. Leyhr, and G. Sudeck, "Physical activity improves stress load, recovery, and academic performance-related parameters among university students," *BMC Public Health*, vol. 24, p. 598, 2024. doi: 10.1186/s12889-024-18082-z

[5] Q. Mou, J. Zhuang, Q. Wu, *et al.*, "Social media addiction and academic engagement as serial mediators between social anxiety and academic performance among college students," *BMC Psychology*, vol. 12, p. 190, 2024. doi: 10.1186/s40359-024-01635-7

[6] O. Kus, "A meta-analysis of the impact of technology related factors on students' academic performance," *Frontiers in Psychology*, 2025. doi: 10.3389/fpsyg.2025.1524645

[7] M. N. Gul, M. Arif, S. Gulzar, G. Naveed, and W. Abbasi, "Deep learning-driven student performance analysis: Detecting anomalies and predicting academic success," *Inverge Journal of Social Sciences*, vol. 4, no. 1, pp. 33–48, 2025.

[8] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, and K. U. Sarker, "Dataset of students performance using student information system, Moodle and the mobile application edify," *Data*, vol. 6, no. 11, p. 110, 2021.

[9] X. Song, J. Li, S. Sun, H. Yin, P. Dawson, and R. R. M. Doss, "SEPN: A sequential engagement based academic performance prediction model," *IEEE Intelligent Systems*, vol. 36, no. 1, pp. 46–53, 2021.

[10] "Machine learning models for academic performance prediction: Interpretability and application in educational decision-making," *Frontiers in Education*, 2025. doi: 10.3389/feduc.2025.1632315

[11] "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Scientific Reports*, 2025. doi: 10.1038/s41598-025-12353-4

[12] A. S. Alghamdi and A. Rahman, "Data mining approach to predict success of secondary school students: A Saudi Arabian case study," *Education Sciences*, vol. 13, no. 3, 2023. doi: 10.3390/educsci13030293

[13] J. Ye, X. Jia, J. Zhang, and K. Guo, "Effect of physical exercise on sleep quality of college students: Chain intermediary effect of mindfulness and ruminative thinking," *Frontiers in Psychology*, vol. 13, 2022. doi: 10.3389/fpsyg.2022.987537

[14] M. J. Duncan, B.-H. Huang, P. A. Cistulli, N. Nassar, M. Hamer, and E. Stamatakis, "Interactive associations between physical activity and sleep duration in relation to adolescent academic achievement," *Int. J. Environmental Research and Public Health*, vol. 19, no. 23, p. 15604, 2022. doi: 10.3390/ijerph192315604

[15] N. Hawi and M. Samaha, "Relationships of gaming disorder, ADHD, and academic performance in university students: A mediation analysis," *PLOS ONE*, vol. 19, no. 4, 2024. doi: 10.1371/journal.pone.0300680

[16] R. Q. Sun, G. F. Sun, and J. H. Ye, "The effects of online game addiction on reduced academic achievement motivation among Chinese college students," *Frontiers in Psychology*, vol. 14, 2023. doi: 10.3389/fpsyg.2023.1185353

[17] S. Shahapur, P. Chitti, S. Patil, C. A. Nerurkar, V. S. Shivannagol, V. C. Rayanaikar, V. Sawant, and V. Betageri, "Decoding minds: Estimation of stress level in students using machine learning," *Indian Journal of Science and Technology*, vol. 17, no. 19, pp. 2002–2012, 2024. doi: 10.17485/IJST/v17i19.2951

[18] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, "A novel methodology using RNN+LSTM+ML for predicting student academic performance," *Education and Information Technologies*, vol. 29, no. 11, pp. 14365–14401, 2024. doi: 10.1007/s10639-023-12394-0