Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 717 – 727
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Enhancing Student Performance Classification Through Dimensionality Reduction and Feature Selection in Machine Learning

[1]Mustakim, [2*]Windy Junita Sari, [3]Fara Ulfa
[1,2]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Department of Psychology, Faculty of Psychology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
Email: [1]mustakim@uin-suska.ac.id, [2]12150324759@students.uin-suska.ac.id, [3]farapsi@uin-suska.ac.id

Article Info

Article history:
Received Jul 12th, 2025
Revised Oct 09th, 2025
Accepted Nov 27th, 2025

Keyword:
Dimensional Reduction
Education
Feature Selection
Machine Learning
Student Performance

ABSTRACT

Education plays an important role in shaping the intellectual and character of the nation's next generation. However, poor student academic performance poses a significant challenge, particularly in terms of student retention and dropout risk. This study aims to evaluate the performance of machine learning algorithms, namely K-Nearest Neighbor (K-NN), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGB), and analyze the effect of dimensionality reduction using Principal Component Analysis (PCA) and feature selection with Recursive Feature Elimination (RFE) on student performance prediction accuracy. The research dataset consists of 395 student samples with demographic, social, and academic attributes. The results show that XGB has the best performance with 98.32% accuracy and can predict all classes with perfect 100% accuracy. LightGBM and K-NN achieved 94.87% and 93.88% accuracy, respectively. The best attributes affecting student performance were found in the "Highly Prioritized" category, including study time, family support, family, and health. Although PCA slightly degraded the model performance, feature selection with RFE significantly improved accuracy. This study concludes that proper algorithm selection and focus on relevant attributes can improve prediction accuracy and efficiency, making an important contribution to the development of more effective education prediction systems.

Corresponding Author:
Windy Junita Sari,
Department of Information Systems, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau
HR. Soebrantas Street, Tuah Madani - Pekanbaru, Riau, Indonesia.
Email: 12150324759@students.uin-suska.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.37783

1. INTRODUCTION

Education is a national development priority, as mandated in Article 31 of the 1945 Constitution paragraphs (3) and (4), emphasizing the government's responsibility to enhance the nation's quality of life through the education system [1]. Education not only improves intellectual capacity but also shapes the character and competence of future generations [2], [3]. Student academic performance is a key indicator of learning success, impacting dropout rates and prospects [4], [5], [6]. According to Permendikbud Number 66 of 2013, student performance assessment must be planned, implemented, and reported professionally, objectively, and informatively [7], [8]. However, assessments often rely solely on academic grades, neglecting internal factors such as cognitive ability, attitude, and motivation, and external factors like family

support, teachers, peers, and facilities [6]. With technological advancements, data-driven approaches, and machine learning can help identify key factors affecting student performance [9], [10].

Algorithms such as KNN, LightGBM, and XGB have been widely used in student performance prediction with high accuracy [11], [12]. These algorithms have been proven effective in academic performance prediction. XGBoost enhances Tree Gradient Boosting Regressor (GBRT) for effective prediction [13], [14] LightGBM, a tree-based ensemble model, addresses overfitting in traditional methods [15], [16]. KNN, a simple yet effective classification technique, predicts outcomes based on data similarity [17], [18]. However, due to the complexity of student data, it is essential to apply methods for reducing dimensionality, such as Principal Component Analysis (PCA), along with feature selection using Recursive Feature Elimination (RFE), to enhance model efficiency and accuracy [19], [20], [21], [22].

This study differs from previous research by integrating expert-based prioritization of attributes with machine learning models. Unlike prior studies that rely solely on statistical feature selection or full feature sets, this research incorporates validation from psychology experts to classify attributes into priority and highly prioritized categories before model training. In addition, this study combines dimensionality reduction (PCA) and feature selection (RFE) with three classification algorithms (K-NN, LightGBM, and XGBoost) using multiple data-splitting scenarios. This integrated approach provides a more comprehensive analysis of feature importance and produces a more robust prediction model for student performance [23], [24], [25].

This study introduces a novel approach by selecting key attributes based on expert-validated importance levels. This study combines three machine learning algorithms with dimensionality reduction and feature selection approaches. Model performance will be tested using accuracy, precision, recall, F1-score, and a confusion matrix to identify significant factors impacting student performance and evaluate model efficacy. To guide the readers, this paper is structured as follows: Section 2 describes the materials and methods used in this study, including data preprocessing, feature selection, and model implementation. Section 3 presents the experimental results and analysis of model performance. Section 4 discusses the findings, limitations, and implications of this research. Finally, Section 5 provides conclusions and recommendations for future studies.

2. RESEARCH METHOD

This research followed systematically designed stages to achieve its objectives, ensuring proper design, method selection, data processing, and result interpretation, as seen in Figure 1.

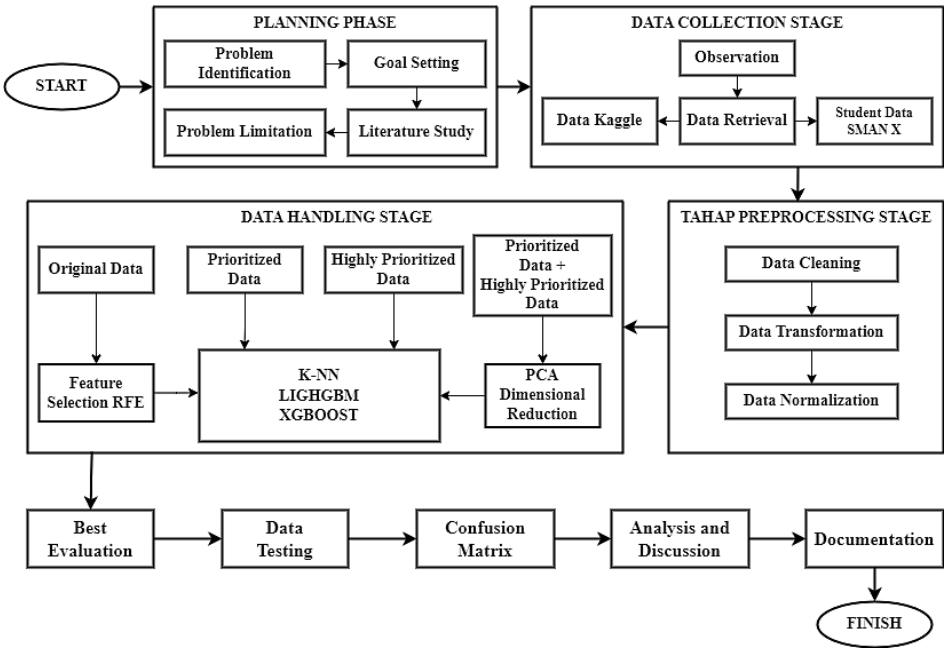

Figure 1. Research Methodology

2.1. Data Collection

The data collection aimed to gather relevant information from primary and second-ary sources. Secondary data, comprising 395 student records, was obtained from Kaggle [26], [27], [28] . With the help of a teacher and field observations, primary data was collected directly from SMAN X 12 students having 32

demographic, social, and academic metrics. This primary data was used to validate the research results. Table 1 displays the full details of the dataset attributes.

Table 1. Dataset Attribute Information

Attribute	Description
School	Student's School
Sex	Gender (binary: F-female or M-male)
Age	Student Age (numeric: 15-22)
Address	Type of residence (binary: U for urban, R for rural).
Famsize	Family Size (binary: LE3 for ≤3 members, GT3 for >3 members)
Pstatus	Parental living arrangement (binary: T for together, A for separated).
Medu	Mother's education level (numeric: 0 - None, 1 - Primary, 3 - Middle school, 4 - Higher education).
Fedu	Father's education level (numeric: 0 - None, 1 - Primary, 3 - Middle school, 4 - Higher education).
Mjob	Mother's occupation (nominal: teacher, healthcare, administration, police, homemaker, or other).
Fjob	Father's occupation (nominal: teacher, healthcare, administration, police, homemaker, or other).
Reason	Reason for school choice (nominal: proximity, reputation, preference, or other).
Guardian	Primary guardian (nominal: mother, father, or other).
Traveltime	Time taken to reach school (numeric: 1 - <15 min, 2 - 15-30 min, 3 - 30 min-1 hour, 4 - >1 hour).
Study time	Weekly study duration (numeric: 1 - <2 hours, 2 - 2-5 hours, 3 - 5-10 hours, 4 - >10 hours).
Failures	Number of failed subjects (numeric: n if $1 \leq n < 3$, otherwise 4).
Schools	Additional academic support from school (binary: yes or no).
Famsup	Family-provided educational support (binary: yes or no).
Paid	Enrollment in extra paid classes (binary: yes or no).
Activities	Participation in extracurricular activities (binary: yes or no).
Nursery	Attending kindergarten (binary: yes or no)
Higher	Aspiration for higher education (binary: yes or no).
Internet	Availability of internet at home (binary: yes or no).
Romantic	Involvement in a romantic relationship (binary: yes or no).
Famrel	Family relationship quality (numeric: 1 - very poor to 5 - excellent).
Freetime	Free time after school (numeric: 1 - very low to 5 - very high).
Go out	Frequency of socializing with friends (numeric: 1 - very low to 5 - very high).
Dalc	Alcohol consumption on weekdays (numeric: 1 - very low to 5 - very high)
Walc	Weekend alcohol consumption (numeric: 1-very low to 5-very high)
Health	Health status (numerical: from 1-very poor to 5-very good)
Absences	Number of school absences (numeric: 0-93)
Ipa	Final academic grade (numeric: 0-20).
Prediction	Student performance category (Excellent, good, fair, and poor)

2.2. Data Preprocessing

Before classification, data preprocessing is conducted using Google Colab. This involves data cleaning (removing irrelevant, empty, or duplicate data), data transformation (converting categorical data to numerical via label coding), and data normalization (using min-max scaling for uniform attribute ranges and efficient processing).

2.3. Data Modelling

The ensemble-based algorithms K-NN, LightGBM, and XGBoost were selected in this study due to their proven effectiveness in classification and high-performance prediction tasks. K-NN was chosen for its simplicity and strong performance in pattern recognition problems involving numerical and categorical data [17], [18]. LightGBM was selected because of its fast training speed, low memory usage, and strong capability in handling large and complex datasets with high accuracy [15], [16]. XGBoost was chosen due to its superior predictive power, integrated regularization mechanism, and robustness against overfitting, as demonstrated in various recent prediction studies [14], [26], [27].

2.4. K-NN Algorithm

K-NN is a learning method that relies on instance-based comparisons. This algorithm is also known as a lazy learning technique. K-NN classifies new data is deriving from the distance to the previous training data, this is done by comparing the nearest neighbors [28], [17] and the algorithm uses supervised learning over the lifetime of the queried instances based on the majority of their categories to make predictions [29], [30]. This neighbor distance calculation uses the Euclidean algorithm, as shown in equation 1.

$$\text{Euc (d)} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots (X_n - Y_n)^2} \quad (1)$$

2.5. LightGBM Algorithm

Light Gradient Boosting Machine (LightGBM) is an ensemble learning algorithm created by Microsoft, and it is an optimized implementation of the gradient boosting framework [31], which has been used for various data mining tasks, such as ranking and classification, and is particularly effective in relying on decision tree algorithms. This algorithm combines two novel approaches: gradient-based one-sided sampling and exclusive feature bundling [26].

By using the boosting algorithm, LightGBM is built as a strong regression tree by integrating several weak regression trees. Equation 2 is a description of the tree integration model [15]:

$$F(x) = \sum_{k=1}^{k} f_k(\text{x}) \quad (2)$$

where F(x) is the sample prediction, f_k (x) is the output of the weak regression tree.

2.6. XGBoost Algorithm

A decision tree-based method called Extreme Gradient Boosting (XGBoost) optimizes the loss function and is an algorithm that can find optimal solutions to regression, classification, and ranking problems [32]. The unique feature of XGBoost is its integrated regularization, which reduces overfitting and improves the model's capacity to generalize to new data. Equation 3 illustrates how the XGBoost prediction model F(x) is constructed as the sum of the outputs from separate decision trees H [27].

$$F(x) = \sum_{h=1}^{H} f_{h(x)} \quad (3)$$

Here, H represents the total number of trees in the model, while fh(x) denotes the influence of each tree. More trees are added to the program over time.

2.7. Literature Review

In this stage, the model evaluation results will be analyzed to compare the performance of K-NN, LightGBM, and XGBoost algorithms before and after the application of PCA and RFE. The focus of the discussion will be on examining the impact of RFE feature selection and dimensionality reduction on model performance, as well as the factors that influence the experimental results and their practical implications. Additionally, the findings will be used to inform suggestions for future research.

K-Nearest Neighbor (K-NN) is a simple yet powerful instance-based learning algorithm that classifies data based on the majority vote of the nearest neighbors using distance measurements such as Euclidean distance, with advantages in simplicity and adaptability to complex classification boundaries, but limitations in high computational cost and sensitivity to irrelevant features [17], [18], [28]. Light Gradient Boosting Machine (LightGBM) is a tree-based ensemble algorithm that applies gradient boosting with Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), enabling fast training speed, low memory usage, and high accuracy, although it requires careful hyperparameter tuning to avoid model bias [15], [16], [31]. Extreme Gradient Boosting (XGBoost) is an advanced boosting algorithm that integrates regularization, parallel processing, and optimized objective functions, offering superior predictive accuracy, robustness, and strong generalization capability, but with higher computational complexity and sensitivity to parameter tuning [14], [31], [32].

3. RESULTS AND ANALYSIS

Data preprocessing to enhance data quality for further analysis is a crucial step. To ensure that the dataset is ready for classification, this process involves data cleaning, which removes irrelevant, empty, missing, or duplicate data. Table 2. shows the results of this preprocessing stage.

Table 2. Data Preprocessing Results

No	School	Sex	Age	...	Absences	IPA	Prediction
1	0.0	0.0	0.428571	...	0.080000	0.1250	0.666667
2	0.0	0.0	0.285714	...	0.053333	0.1250	0.666667
3	0.0	0.0	0.000000	...	0.133333	0.2500	0.333333
4	0.0	0.0	0.000000	...	0.026667	0.7500	0.000000
5	0.0	0.0	0.142857	...	0.053333	0.1875	0.333333

No	School	Sex	Age	...	Absences	IPA	Prediction
...	...	...	...	...	...	...	...
391	1.0	1.0	0.714286	...	0.146667	0.3750	0.333333
392	1.0	1.0	0.285714	...	0.040000	0.6875	0.000000
393	1.0	1.0	0.857143	...	0.040000	0.4375	0.333333
394	1.0	1.0	0.428571	...	0.000000	0.5000	0.000000
395	1.0	1.0	0.571429	...	0.066667	0.3125	0.333333

3.1. PCA Dimensionality Reduction

To ensure optimal performance in data dimensionality reduction, the use of parameters in the PCA and StandardScaler processes is critical. The values of these parameters are configured with default or customized values to achieve the desired goals, such as maintaining the scale of the data and maximizing the variance retained by PCA. Table 3 provides detailed information on parameter usage at each stage.

Table 3. PCA Parameters

Stage	Parameter	Value	Description
PCA	N_components	0.95	PCA will select the number of components that retain 95% of the total variation in the original data.
	Svd_solver	Auto (default)	The SVD solver automatically selects the best method based on the data size to calculate the singular value decomposition.
	Whiten	False (default)	If true, the output components will be whitened, which may change the scale of the features but make them uncorrelated.

The results of applying PCA are presented in Table 4.

Table 4. PCA Result

Most Influential Features	Contribution Value
IPA	0.6850
School	0.6269
Health	0.6435
Higher	0.6169
Pstatus	0.5922
Walc	0.4991
Higher	0.4891
Medu	0.4745

PCA analysis shows that there are 8 components with the highest contribution, as shown in Table 4. The most influential component is the IPA attribute with a contribution value of 0.6850, while the lowest contribution is in the Medu attribute with a value of 0.4745.

3.2. RFE Feature Selection

This research uses the RFE feature selection technique to evaluate the performance of each algorithm. Feature selection aims to find the most relevant features, reduce model complexity, and improve prediction accuracy. Table 5 shows the process parameters used.

Table 5. RFE Parameter

Stage	Parameter	Value	Description
Model	Logistic Regression	Default	In RFE, the logistic regression model is used as the estimator.
Solver	Solver	lbfgs	The logistic regression parameters are calculated by an optimization algorithm.
Iterasi maks	Max_iter	1000	The highest number of iterations is required for convergence during the model training process.

Table 6 shows the feature selection results after the parameters are applied and the feature selection process is run. From the feature selection process at various data-sharing ratios, the 15 most relevant features are obtained, which are characterized by a ranking of 1. From each ratio, 11 attributes always appear with a ranking of 1, as shown in Table 6.

Table 6. RFE Feature Selection Results

Most Relevant Features	Ranking
Sex, Mjob, Reason, Failures, Famsup, Paid, Activities, Higher, Internet, Romantic, Famrel, Freetime, Goout, IPA, Dalc, School.	1
School, Address, Mjob	2

Most Relevant Features	Ranking
Famsize, Freetime	3
Famsup	4
Paid	5
Romantic	6
Famrel	7
Freetime	8
Go out	9
Dalc	10

3.3. Original Data Modeling Process Using Machine Learning Algorithms with RFE Feature Selection Technique

This section describes the stages of processing the original data using machine learning algorithms such as K-NN, LightGBM, and XGBoost, combined with the RFE feature selection technique. This technique is used to select the most relevant features, thereby improving the accuracy and efficiency of the model. The process involves selecting an algorithm, applying RFE, evaluating the feature selection results, and testing the model using simplified data.

Tabel 7. Original Data Result

Algorithm	Data Splitting	Value			
		Accuracy	Precision	Recall	F1-Score
K-NN	80:20	96.20%	96.65%	96.20%	96.28%
	70:30	94.96%	94.83%	94.96%	94.80%
	60:40	93.04%	93.25%	93.04%	92.77%
LightGBM	80:20	96.20%	92.70%	96.20%	94.38%
	70:30	96.64%	93.50%	96.64%	95.01%
	60:40	96.84%	93.87%	96.84%	95.31%
XGBoost	80:20	97.47%	97.60%	97.47%	96.87%
	70:30	97.48%	97.61%	97.48%	96.76%
	60:40	98.10%	98.18%	98.10%	97.71%

The performance of the three algorithms K-NN, LightGBM, and XGBoost is shown in Table 7 based on different data sharing ratios of 80:20, 70:30, and 60:40. Accuracy, Precision, Recall, and F1-Score metrics are used to measure the performance of the algorithms. The results show that XGBoost has the best performance across all sharing ratios, with accuracy reaching 98.10% at a 60:40 ratio, followed by LightGBM and K-NN. This result shows that XGBoost has higher accuracy and classification ability.

3.4. Modeling Process of Priority Attribute Importance Data by Applying Machine Learning

This stage describes the data processing procedure related to the prioritized attributes obtained through validation with psychology experts. Selected to ensure relevance and accuracy, these attributes are then analyzed using machine learning algorithms to find patterns and prioritize the most relevant attributes. This technique yields a more accurate and efficient analysis by focusing the model's performance on the components that have the greatest impact.

Table 8. Prioritization Data Results

Algorithm	Data Splitting	Value			
		Accuracy	Precision	Recall	F1-Score
K-NN	80:20	88.61%	85.23%	88.61%	86.82%
	70:30	89.92%	87.30%	89.92%	88.19%
	60:40	89.24%	86.91%	89.24%	87.61%
LightGBM	80:20	96.20%	92.70%	96.20%	96.20%
	70:30	96.64%	93.50%	96.64%	95.01%
	60:40	96.84%	93.87%	96.84%	95.31%
XGBoost	80:20	97.47%	97.60%	97.47%	96.87%
	70:30	97.48%	97.61%	97.48%	96.76%
	60:40	98.10%	98.18%	98.10%	97.71%

Table 8 shows the performance evaluation results of the three algorithms K-NN, LightGBM, and XGBoost at various data sharing ratios of 80:20, 70:30, and 60:40. Based on the data, XGBoost shows the best performance with an Accuracy value of 98.10% at a ratio of 60:40.

Overall, the results presented in Table 8 indicate that ensemble-based machine learning algorithms, particularly XGBoost, consistently outperform K-NN and LightGBM across all data splitting scenarios. The superior performance of XGBoost can be attributed to its ability to handle complex nonlinear relationships and reduce overfitting through gradient boosting optimization. Meanwhile, LightGBM also demonstrates

strong and stable performance, suggesting its effectiveness in modeling prioritized attribute importance. In contrast, K-NN shows comparatively lower performance, which may be influenced by its sensitivity to data distribution and distance metrics. These findings confirm that advanced boosting algorithms are more suitable for modeling prioritized attribute importance data, as they provide higher accuracy and robustness in capturing critical attribute patterns.

3.5. Modeling Process of Highly Prioritized Attribute Importance Data by Applying Mchine Learning

This stage explains the use of machine learning to process data with very important features. The purpose of applying machine learning is to improve the accuracy of decision-making based on very important data.

Table 9. Highly Prioritized Data Results

Algorithm	Data Splitting	Nilai			
		Accuracy	Precision	Recall	F1-Score
K-NN	80:20	97.47%	97.63%	97.47%	97.47%
	70:30	97.48%	97.56%	97.48%	97.37%
	60:40	94.94%	95.20%	94.94%	94.07%
LightGBM	80:20	96.20%	92.70%	96.20%	94.38%
	70:30	96.64%	93.50%	96.64%	95.01%
	60:40	96.84%	93.87%	96.84%	95.31%
XGBoost	80:20	96.20%	92.70%	96.20%	94.38%
	70:30	98.32%	98.38%	98.32%	98.05%
	60:40	97.47%	97.60%	97.47%	96.66%

Table 9 presents the performance evaluation results of the three algorithms, K-NN, LightGBM, and XGBoost, at data sharing ratios of 80:20, 70:30, and 60:40. The results indicate that XGBoost achieves the best performance at a ratio of 70:30, with 98.32%, followed by K-NN and LightGBM. Although not better than XGBoost, LightGBM performed similarly and best at a 60:40 ratio. Overall, XGBoost showed more stable and better results for most metrics and data-sharing ratios.

3.6. Data Modeling Process of Combination of Priority and Highly Priority Attributes by Applying PCA Dimensionality Reduction

In this section, highly important attributes are combined to create a broader dataset. Then, to reduce the complexity of the data without losing important information, PCA is used. The purpose of PCA is to enhance the efficiency and accuracy of the classification model that will be applied to this larger and more complex dataset.

Table 10. PCA Data Results

Algorithm	Data Splitting	Value			
		Accuracy	Precision	Recall	F1-Score
K-NN	80:20	86.08%	86.13%	86.08%	85.27%
	70:30	89.08%	89.19%	89.08%	88.28%
	60:40	88.61%	88.87%	88.61%	88.26%
LightGBM	80:20	74.68%	63.14%	74.68%	68.36%
	70:30	76.47%	75.83%	76.47%	73.20%
	60:40	76.58%	66.26%	76.58%	71.01%
XGBoost	80:20	78.48%	77.37%	78.48%	74.67%
	70:30	79.83%	78.39%	79.83%	78.91%
	60:40	79.11%	68.52%	79.11%	73.43%

Table 10 shows the performance evaluation results of the three algorithms K-NN, LightGBM, and XGBoost at data sharing ratios of 80:20, 70:30, and 60:40. K-NN shows the best performance in all ratios, with the highest accuracy reaching 89.08%.

The results also suggest that the application of PCA influences each algorithm differently. K-NN benefits the most from dimensionality reduction, likely due to improved distance calculations in a reduced feature space. Conversely, the decreased performance of LightGBM and XGBoost indicates that some discriminative information may be lost during PCA transformation, affecting tree-based learning effectiveness.

Figure 5. shows the comparison of model accuracy based on the type of data used in the analysis process. Prioritized Data and RFE feature selection methods have the same accuracy of 98.10%, indicating that the features selected through these two methods are quite relevant to the model. Meanwhile, the feature subset with the highest significance level, Highly Prioritized Data, has the highest accuracy of 98.32%, suggesting that more focused feature selection can improve the performance of the model.

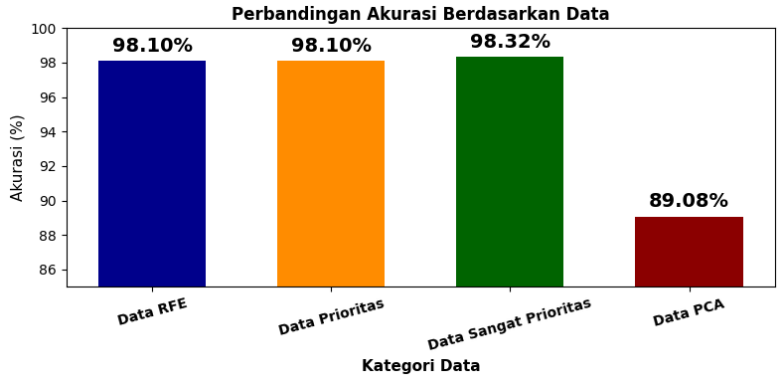

Figure 2. Highest Comparison Visualization

3.7. Data Testing

After all data processing was completed using the three models, a thorough evaluation of the best model was conducted. Based on the evaluation results, the highest-performing XGBoost algorithm was selected for the prediction process on the validation data. The prediction targets to be tested are shown in Table 11, with a rating scale based on predetermined categories of Riau Province SMAN Plus student data. This prediction target is obtained from the range of student scores, namely Poor = 0-79, Fair = 80-86, Good = 87-93, and Very Good = 94-100.

Table 11. Target Category

No	Target Category	Description
1	0	Less
2	1	Fair
3	2	Good
4	3	Very Good

Table 12. Prediction Result

No	Data Class	Predicted
1	Good	Good
2	Enough	Enough
3	Good	Good
4	Enough	Enough
5	Good	Good
6	Enough	Enough
7	Good	Good
8	Enough	Enough
9	Good	Good
10	Good	Good
11	Good	Good
12	Good	Good

As shown in Table 12, the model's prediction results on the 12 validation data reach 100%. All data were correctly predicted according to their original classes, indicating that the algorithm can learn the data patterns and classify them accurately. Additionally, data with “Good” and “Fair” classes were predicted consistently, demonstrating the model's effectiveness and accuracy. This success ensures that the research objectives are met and strengthens confidence in the model's ability to inform decision-making.

3.8. Discussion

The findings of this study are consistent with previous research that demonstrated the superiority of XGBoost in classification and prediction tasks. Hossen and Uddin (2023) reported that XGBoost achieved high accuracy in student attention detection, while Ahmed et al. (2020) confirmed the improved performance of modified K-NN in predicting student performance. The results of this study further strengthen these findings, where XGBoost achieved the highest accuracy of 98.32%, outperforming both K-NN and LightGBM. Furthermore, the effectiveness of RFE feature selection in this study aligns with prior studies [20], [21], which reported that feature selection improves model stability and accuracy by removing irrelevant features..

Model performance is significantly influenced by the application of dimensionality reduction techniques, including PCA, and feature selection using RFE. In particular, feature selection with RFE makes

the model more efficient and effective by reducing the number of irrelevant features, improving prediction accuracy, and increasing processing time. The feature selection results also showed correlations with features that were highly influential on student performance. At this stage, the highest accuracy achieved was 98.32%, indicating that an approach that focuses on more important attributes can improve accuracy. While the application of PCA to the data reduced performance, it may be that the application of PCA dimension reduction did not result in better model performance or even greater accuracy when compared to the RFE technique, as PCA transforms the feature representation into a new combination. In addition, the interpretation of the model becomes more complicated, which may affect the ability of the model to identify patterns in the original data.

Therefore, the RFE feature selection technique is more recommended for use in predicting student academic success. Overall, this research shows that choosing the right features and using machine learning algorithms can significantly improve the accuracy of predicting student performance. Future research can further explore research in this area by looking at other machine learning algorithms or incorporating additional data sources to prove the analysis.

In summary, this study reinforces existing literature by confirming that both algorithm selection and feature optimization play a crucial role in educational data modeling. Consistent with prior research, XGBoost demonstrates superior predictive capability, particularly when combined with effective feature selection methods such as RFE. The comparative analysis between RFE and PCA highlights that preserving meaningful original feature representations is essential for maintaining model interpretability and performance. These results collectively support previous findings that emphasize targeted feature selection over dimensionality reduction for improving prediction accuracy in student performance studies.

4. CONCLUSION

The use of machine learning algorithms such as XGBoost, K-NN, and LightGBM can provide significant results in predicting student performance. Of the three algorithms tested, XGBoost showed the highest accuracy, at 98.32%, and managed to predict all classes perfectly, achieving 100% accuracy. This demonstrates that XGBoost can effectively recognize patterns in the data. In addition, LightGBM and K-NN also yielded good results with accuracies of 94.87% and 93.88%, respectively. Feature selection techniques, such as RFE, and dimensionality reduction using PCA contributed to improving the model's performance, with proper feature selection enhancing prediction accuracy. In this case, applying the XGBoost model with prioritized and highly prioritized data provided highly accurate predictions, resulting in more optimized performance compared to using all features. Although PCA reduced the dimensionality of the data, its application to the combined priority and high-priority data slightly degraded performance, indicating the importance of further customization in model evaluation. The findings of this study have significant consequences for current educational policies and practices. Educational institutions can utilize the findings of this analysis to develop more effective strategies for enhancing student learning outcomes. For example, a data-driven personalized learning approach can help identify student needs more accurately, allowing for more targeted interventions. This research can also help policymakers create teacher training programs that better suit the needs of students in the modern era.

This study has several limitations that should be taken into consideration. First, the dataset used was relatively limited in size and derived from a specific educational context, which may affect the generalizability of the results to other regions or education levels. Second, this study only applied three machine learning algorithms and two feature selection techniques, which may not fully represent the potential performance of other advanced models, such as deep learning or hybrid ensemble approaches. Third, the validation data were limited to a single school, which may introduce selection bias. Future research is recommended to use larger and more diverse datasets from multiple institutions to improve model robustness. Additional machine learning algorithms, including deep neural networks and hybrid ensemble models, should be explored. Further optimization using hyperparameter tuning and cross-validation techniques is also suggested to enhance model reliability. Moreover, integrating socio-economic and behavioral data in greater detail could provide deeper insights into the factors affecting student academic performance.

REFERENCES

[1] F. N. A. Kurniawati, "Meninjau Permasalahan Rendahnya Kualitas Pendidikan Di Indonesia Dan Solusi," *Acad. Educ. J.*, vol. 13, no. 1, pp. 1–13, 2022, doi: 10.47200/aoej.v13i1.765.

[2] Imamah, U. L. Yuhana, A. Djunaidy, and M. H. Purnomo, "Enhancing students performance through dynamic personalized learning path using ant colony and item response theory (ACOIRT)," *Comput. Educ. Artif. Intell.*, vol. 7, no. April, p. 100280, 2024, doi: 10.1016/j.caeai.2024.100280.

[3] S. M. Saadullah, S. Ammar, and A. Alazzani, "Exploring verbal, interpersonal, and visual intelligences in accounting education: Effects on student learning and performance," *J. Account. Educ.*, vol. 68, no. August 2023,

p. 100917, 2024, doi: 10.1016/j.jaccedu.2024.100917.
[4] I. Costa, M. Angˆ, and M. Angˆ, "ScienceDirect Student Student Performance Performance Prediction Prediction on on Primary Primary and and Secondary Secondary Schools Schools Systematic Literature Review - A Systematic Literature Review," vol. 00, 2022, doi: 10.1016/j.procs.2022.11.229.
[5] F. A. Al-azazi and M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction in MOOC," *Heliyon*, vol. 9, no. 4, p. e15382, 2023, doi: 10.1016/j.heliyon.2023.e15382.
[6] T. Gori, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.
[7] T. Tadhkiroh, B. Akbar, and T. I. Hartini, "Pengembangan Instrumen Penilaian Kinerja pada Muatan IPA Kurikulum 2013 Tingkat Sekolah Dasar," *J. Basicedu*, vol. 7, no. 1, pp. 631–644, 2023, doi: 10.31004/basicedu.v7i1.4720.
[8] R. Nuraini, F. Fadllurrohman, and N. Norfaizah, "Implementasi Penilaian Hasil Belajar Siswa Berbasis Rapor Digital Madrasah Di MI Mathla'ul Anwar HSU," *Al-Madrasah J. Pendidik. Madrasah Ibtidaiyah*, vol. 6, no. 4, p. 1053, 2022, doi: 10.35931/am.v6i4.1174.
[9] W. J. Sari *et al.*, "Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 34–43, 2024, doi: 10.57152/predatecs.v2i1.1119.
[10] M. Chen and Z. Liu, "Heliyon Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: 10.1016/j.heliyon.2024.e32570.
[11] S. A. Priyambada, T. Usagawa, and M. ER, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge," *Comput. Educ. Artif. Intell.*, vol. 5, no. June, p. 100149, 2023, doi: 10.1016/j.caeai.2023.100149.
[12] C. I. Hatleberg *et al.*, "Predictors of Ischemic and Hemorrhagic Strokes Among People Living With HIV: The D:A:D International Prospective Multicohort Study," *EClinicalMedicine*, vol. 13, pp. 91–100, 2019, doi: 10.1016/j.eclinm.2019.07.008.
[13] N. B. Shaik, K. Jongkittinarukorn, and K. Bingi, "Jo ur na l P of," *Case Stud. Chem. Environ. Eng.*, p. 100775, 2024, doi: 10.1016/j.cscee.2024.100775.
[14] M. Shehab, R. Taherdangkoo, and C. Butscher, "Computers and Geotechnics Towards Reliable Barrier Systems : A Constrained XGBoost Model Coupled with Gray Wolf Optimization for Maximum Swelling Pressure of Bentonite," *Comput. Geotech.*, vol. 168, no. February, p. 106132, 2024, doi: 10.1016/j.compgeo.2024.106132.
[15] X. Mao *et al.*, "A variable weight combination prediction model for climate in a greenhouse based on BiGRU-Attention and LightGBM," *Comput. Electron. Agric.*, vol. 219, no. July 2023, p. 108818, 2024, doi: 10.1016/j.compag.2024.108818.
[16] D. Zhang, "The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure," vol. 8, 2020, doi: 10.1109/ACCESS.2020.3042848.
[17] A. Yasar, "Intelligent Systems And Applications In Engineering Data Classification of Early-Stage Diabetes Risk Prediction Datasets and Analysis of Algorithm Performance Using Feature Extraction Methods and Machine Learning Techniques," vol. 9, no. 4, pp. 273–281, 2021, doi: 10.1039/b000000x.
[18] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 26–33, 2024, doi: 10.57152/predatecs.v2i1.1110.
[19] A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," vol. 8, no. Ml, 2020, doi: 10.1109/ACCESS.2020.3001149.
[20] M. Qaraad, A. K. Kelany, and X. Chen, "An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021, doi: 10.1109/ACCESS.2021.3123090.
[21] D. P. Utomo, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," vol. 4, no. April, pp. 437–444, 2020, doi: 10.30865/mib.v4i2.2080.
[22] F. Ardiansyah and I. W. Siadi, "Klasifikasi Customer Relationship Management Menggunakan Dataset KDD Cup 2009 dengan Teknik Reduksi Dimensi Classification of Customer Relationship Management using KDD Cup 2009 Dataset with Dimension Reduction Technique," vol. 11, no. 28, 2022, doi: 10.34010/komputika.v11i2.6498.
[23] M. Kamal and M. Shorif, "Computers and Education : Artificial Intelligence Attention monitoring of students during online classes using XGBoost classifier," *Comput. Educ. Artif. Intell.*, vol. 5, no. November, p. 100191, 2023, doi: 10.1016/j.caeai.2023.100191.
[24] W. Alia *et al.*, "Factors Affecting Student ' s Academic Performance," vol. 7, no. 1, pp. 99–107, 2021.
[25] S. T. Ahmed, R. Al-hamdani, and M. S. Croock, "Enhancement of student performance prediction using modified K-nearest neighbor," vol. 18, no. 4, 2020, doi: 10.12928/TELKOMNIKA.v18i4.13849.
[26] S. Hussain, H. F. Öztop, A. Madhi, and F. Ertam, "Mixed bioconvection of nanofluid of oxytactic bacteria through a porous cavity with inlet and outlet under periodic magnetic field using artificial intelligence based on LightGBM algorithm," *Therm. Sci. Eng. Prog.*, vol. 50, no. April, p. 102589, 2024, doi: 10.1016/j.tsep.2024.102589.
[27] S. Jafari, J. H. Yang, and Y. C. Byun, "Optimized XGBoost modeling for accurate battery capacity degradation prediction," *Results Eng.*, vol. 24, no. July, p. 102786, 2024, doi: 10.1016/j.rineng.2024.102786.
[28] A. F. Lubis *et al.*, "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors,

Naïve Bayes and Decission Tree," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 44–51, 2024, doi: 10.57152/predatecs.v2i1.1103.
[29] A. R. Lubis and M. Lubis, "Optimization of distance formula in K-Nearest Neighbor method," vol. 9, no. 1, pp. 326–338, 2020, doi: 10.11591/eei.v9i1.1464.
[30] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor , Genetic , Support Vector Machine , Decision Tree , and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. May, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.
[31] D. Sarkasme, P. Dataset, N. Headline, D. Metode, and E. Deep, "Sarcasm Detection in News Headline Dataset with Ensemble Deep Learning Method," vol. 6, no. 2, pp. 47–52, 2023, doi: https://doi.org/10.21070/joincs.v6i2.1628.
[32] W. Liu, W. D. Liu, and J. Gu, "Predictive model for water absorption in sublayers using a Joint Distribution Adaption based XGBoost transfer learning method," *J. Pet. Sci. Eng.*, vol. 188, no. August 2019, p. 106937, 2020, doi: 10.1016/j.petrol.2020.106937.

BIBLIOGRAPHY OF AUTHORS

Mustakim, is a lecturer at Department of Information System Faculty of Science and Technology Universitas Islam Negeri Sultan Syarif Kasim Riau. Subject expert at Data Mining, Artificial Intelligence and Big Data. Most of his research a nd publication are integrated study of Data Mining with Moslem, Data Mining with Education and Data Mining with other sciences. He is a founder of Puzzle Research Data Technology, research gorup of national level at Universitas Islam Negeri Sultan Syarif Kasim Riau.

Windy Junita Sari is a graduate of the Information Systems Program, Faculty of Science and Technology, Sultan Syarif Kasim Riau State Islamic University. She earned a Bachelor of Computer Science degree with a focus on data science and machine learning. Her academic interests include data analysis, artificial intelligence, and administrative information systems. Windy is also an active member of Puzzle Research Data Technology, a national-scale research group focused on integrating data mining with various other disciplines, based at the State Islamic University Sultan Syarif Kasim Riau.

Fara Ulfa, is a lecturer at the Faculty of Psychology, Universitas Islam Negeri Sultan Syarif Kasim Riau. She holds a Master's degree in Professional Psychology from Universitas Padjadjaran. She has over six years of clinical experience as a psychologist working with children and adolescents. Her research interests include health psychology, computer-based interventions for ADHD, and child development.