Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

CLASSIFICATION AGORITHM ANALYSIS FOR PREDICTING THE TYPE OF SENIOR HIGH SCHOOL ON ALUMNI SMP 2 BALONG PONOROGO

Nabiilah Winda Kurnia Putri[1], Wiyli Yustanti[2]

[1,2]*Information System, Faculty of Engineering, State University of Surabaya, Surabaya, Indonesia*

nabiilah.22156@mhs.unesa.ac.id, wiyliyustanti@unesa.ac.id

ABSTRACT

This study aims to analyze the performance of various classification algorithms in predicting the type of Senior High School (SLTA) that students choose based on academic scores and achievements. The study was conducted at SMPN 2 Balong Ponorogo using the SEMMA (Sample, Explore, Modify, Model, Assess) approach. Secondary data from 1,113 students were used and processed through the stages of data exploration, normalization, feature selection (using Pearson Correlation, Mutual Information, Random Forest, and Lasso Logistic Regression), and dimension reduction using Principal Component Analysis (PCA). Eight classification algorithms were tested, namely Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, LightGBM, CatBoost, and Naïve Bayes. Model evaluation is done using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results show that the Random Forest and KNN models with the Hybrid Feature Selection approach provide the best performance, with the F1-score value reaching 84%. This research contributes to data-based decision making for student guidance in choosing the right further education pathway.

Keyword: Classification, Secondary School, SEMMA, Feature Selection, Machine Learning

Article Info:
Article history:
Received July 19, 2025
Revised August 26, 2025
Accepted September 11, 2025

Corresponding Author
Nabiilah Winda Kurnia Putri
State University of Surabaya, Surabaya, Indonesia
Nabiilah.22156@mhs.unesa.ac.id

1. INTRODUCTION

The development of information technology has had a significant impact in the world of education, especially in the utilization of machine learning for data-based decision making. One important application is the prediction of student education pathway selection, such as the selection of the type of Senior High School (SLTA). The problem often faced by ninth grade students is the ignorance in choosing a school that suits their academic abilities and non-academic achievements. This requires a predictive system that can help students and schools in the career guidance process objectively and systematically.

Various studies have proven the reliability of classification algorithms such as Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) in predicting student performance. However, research that examines the selection of high school types based on a combination of academic grades and extracurricular achievements is still limited. Therefore, this research was conducted using the SEMMA (Sample, Explore, Modify, Model, Assess) approach, as well as the integration of feature selection and dimension reduction techniques to improve prediction performance.

The data used came from 1,113 students of SMPN 2 Balong Ponorogo in five batches (2019-2023). The research process includes normalization, label encoding, data balancing with SMOTE, and feature selection using Pearson Correlation, Mutual Information, Random Forest, and Lasso Logistic Regression. Selected features were then reduced using Principal Component Analysis (PCA). Modeling was performed with eight classification algorithms, and performance evaluation using accuracy, precision, recall, and F1-score metrics. Results showed that the combination of Random Forest and KNN algorithms with Hybrid Feature Selection approach gave the best performance with F1-score reaching 83% and 82%. This finding proves that data-based predictive models are reliable in helping students make the right high school choices. In addition, this research provides a practical contribution in the form of a framework for developing an educational recommendation system that can be utilized by teachers, counselors, and school policy makers. In the future, the model development can be improved by adding variables such as demographic background or career interests. Integration of the system into a digital platform will also expand the usefulness of the prediction results in educational practice.

2. METHODS

This research is a predictive quantitative study that applies a machine learning approach in a data mining framework using the SEMMA (Sample, Explore, Modify, Model, Assess) method [1]. This approach was chosen because it is able to facilitate data analysis systematically from the pre- processing stage to model evaluation. The following is a depiction of the flow of this research method:

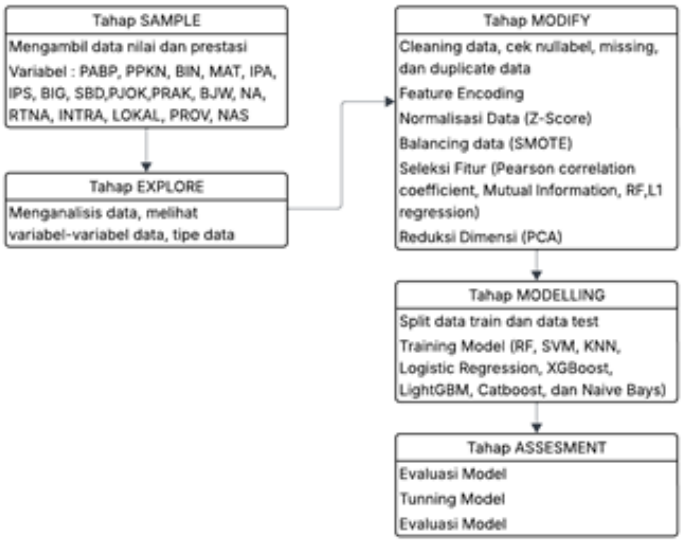

Figure 1 Research Flow Design

This approach was chosen because it is able to facilitate systematic data analysis from the pre-processing stage to model evaluation. The data used is secondary, obtained from school documentation in the form of academic This approach was chosen because it is able to facilitate systematic data analysis from the pre-processing stage to model evaluation. The data used is secondary, obtained from school documentation in the form of academic grades and achievement data for students in grade IX of SMPN 2 Balong Ponorogo for five school years (2019-2023), with a total of 1,113 student data. Data collection techniques are carried out through the documentation method, while data processing starts from the data cleaning stage, normalization using the Z-score method, and transformation of categorical variables through label encoding. Next, class distribution balancing was performed on the target variable using Synthetic Minority Oversampling Technique (SMOTE) to avoid model bias towards the majority class. The feature selection process uses a combination of filter methods (Pearson Correlation and Mutual Information) and embedded methods (Random Forest and Lasso Logistic Regression), the results of which are combined with a Hybrid Feature Selection approach to improve the relevance of features to predictions. Dimensionality reduction was performed using Principal Component Analysis (PCA) to simplify the data without reducing the meaning of the information. Eight classification algorithms were used in the modeling: Logistic Regression, K- Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, XGBoost, LightGBM, CatBoost, and Naïve Bayes, which were then evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. The best parameter search was conducted using the Grid Search method to optimize model performance. This method is believed to produce predictive models that are robust, interpretive, and can be applied in an educational context to support data-driven decision- making.

3. RESULTS AND DISCUSSION

This section presents the results and analysis of the entire classification process. The process includes feature selection, dimensionality reduction, modeling, and performance evaluation of the eight classification algorithms. The results are organized into the following subsections:

3.1 Feature Selection Results

The feature selection process was conducted through four approaches: Pearson Correlation, Mutual Information, Random Forest, and Lasso Logistic Regression (LS). Each resulted in a ranking of the importance of the feature to the target (Y).

a. Pearson Correlation Feature Result

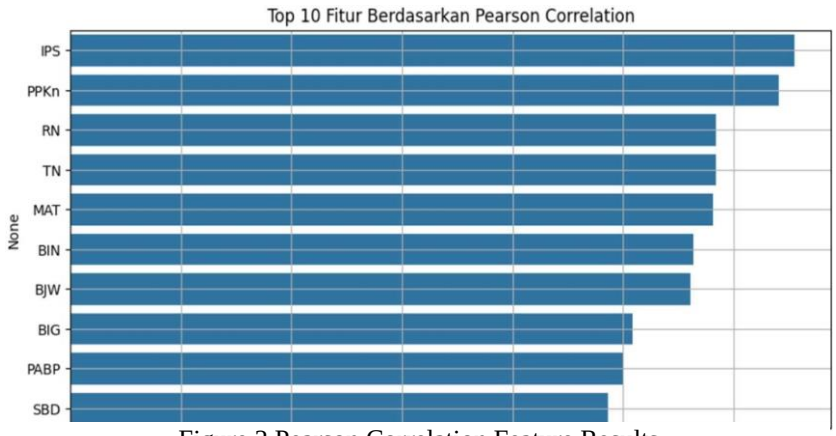

Figure 2 Pearson Correlation Feature Results

The results of the Pearson Correlation feature selection are shown in Figure 2 which produces the main features: Social Studies, Civics, RN, MAT, BIN, BJW, BIG, PABP, and SBD.

b. Mutual Information Feature Selection Results

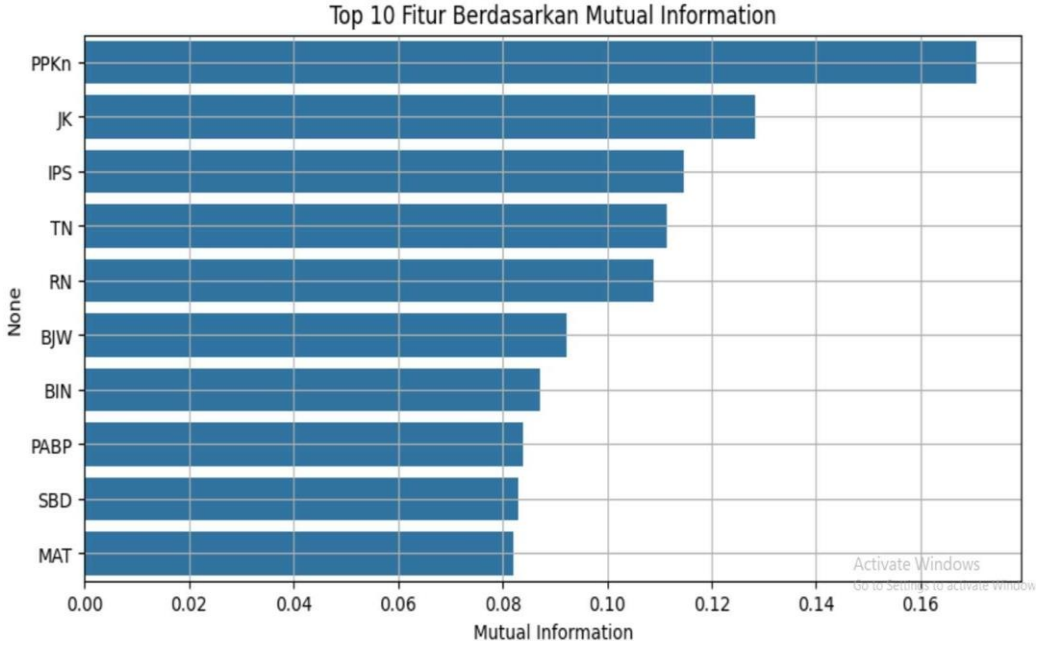

Figure 3 Mutual Information Feature Selection Results

The result of Mutual Information feature selection is shown in Figure 3, which produces the main features: PPkn, JK, IPS, TN, RN, BJW, BIN, PABP, SBD and MAT.

c. Random Forest Feature Selection Results

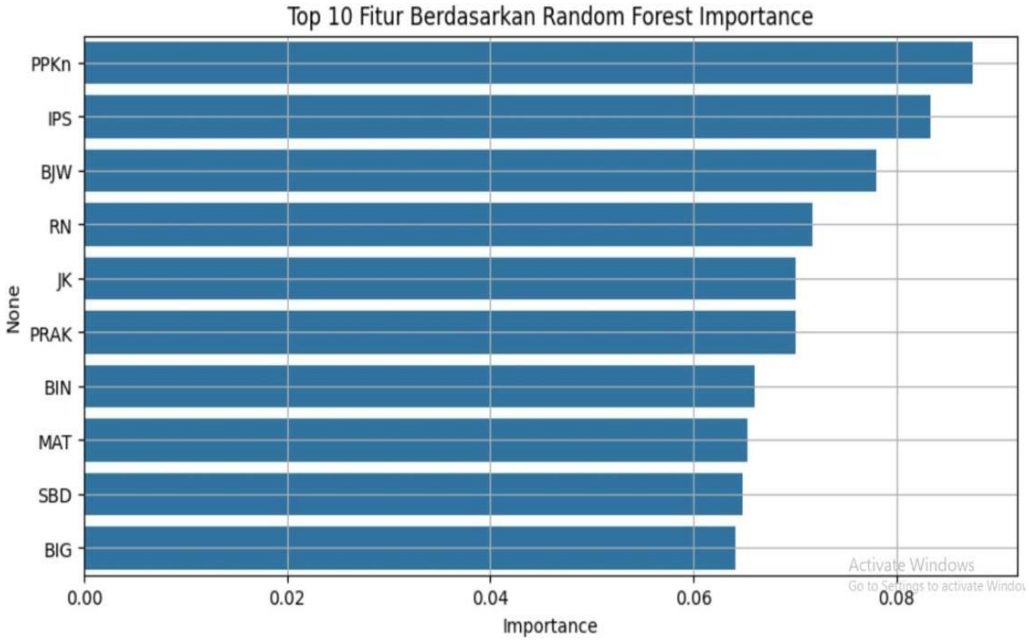

Figure 4 Random Forest Feature Selection Results

The results of the Random Forest feature selection are shown in Figure 4, which produces the main features: PPkn, IPS, BJW, RN, JK, PRAK, BIN, MAT, SDB, and BIG.

d. Lasso Regression Feature Selection Result

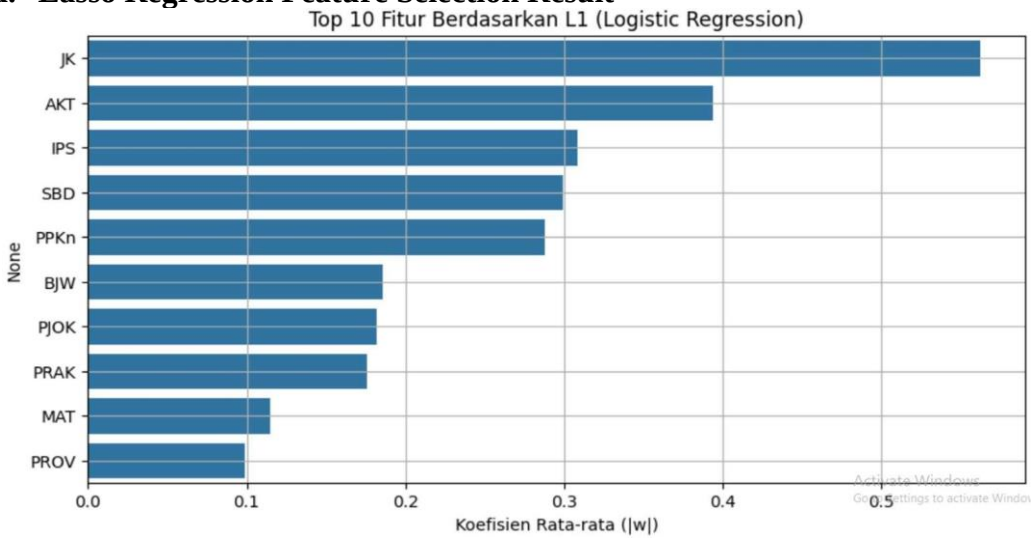

Figure 5 Lasso Regression Feature Results

The results of the Random Forest feature selection are shown in Figure 5, which produces the main features: JK, AKT, IPS, SBD, PPKn, BJW, PJOK, PRAK, MAT, and PROV.

e. Hybrid Feature Selection Results

Next, the selection results are combined using the Hybrid Feature Selection approach, which retains features that appear at least twice from the four approaches.

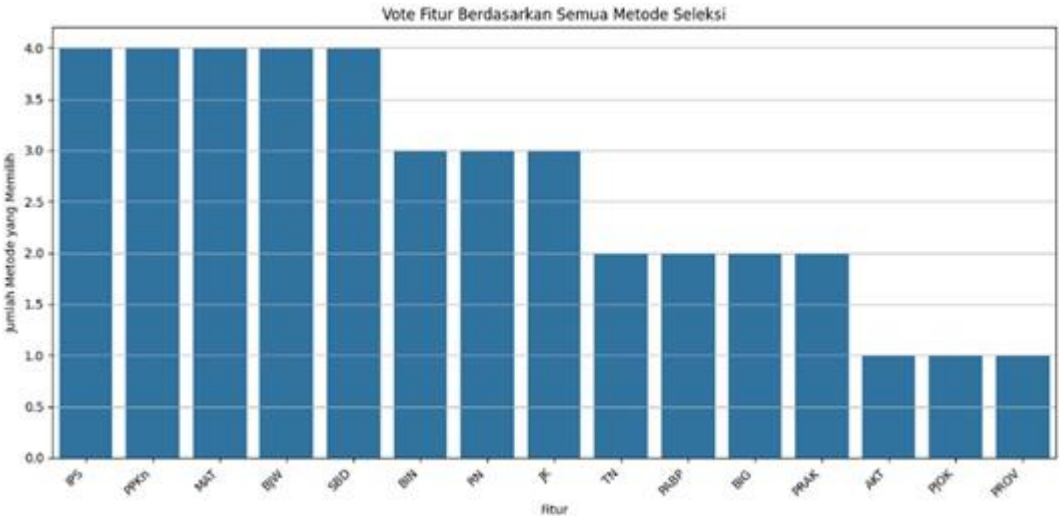

Figure 6 Hybrid Feature Selection Results

The final result of the combined feature selection is shown in Figure 6, which produces 14 main features: IPS, PPKn, BJW, SBD, RN, BIN, MAT, PRAK, JK, TN, PABP, PJOK, PROV,and AKT.

3.2 Dimension Reduction Using PCA

Dimensionality reduction was performed using Principal Component Analysis (PCA) technique on 14 selected features. The goal is to reduce complexity without losing important information.

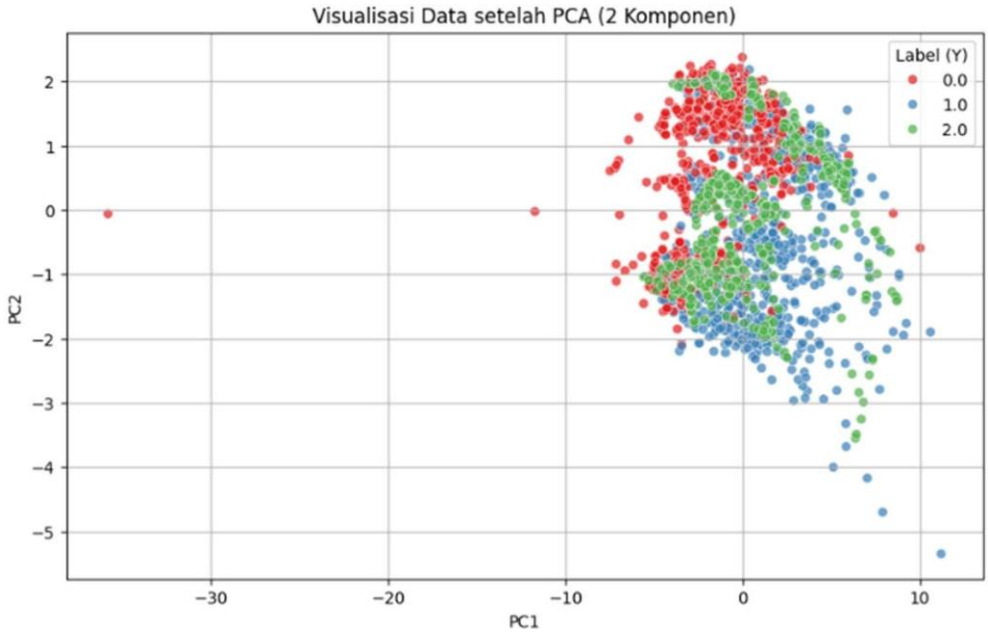

Figure 7 Visualization of PCA Plot

The PCA results are visualized in the form of a two-dimensional scatter plot in Figure 7, where the student data points are grouped based on the class label Y. PCA reveals the overlap between classes. Although class 2 (blue) tends to accumulate at the bottom right and class 0 (red) at the top left, class 1 (green) is widely spread in the center. This shows that the separation between classes is not completely linear.

3.3 Model Evaluation

Modeling was performed by applying eight classification algorithms to five feature scenarios: All Features, PCA, Mutual Information, Random Forest, Lasso (LS), and Hybrid Feature Selection. The evaluation was based on F1-score, accuracy, precision, and recall, as shown in Table 1.

Table 1 F1 Score Accuracy

Fitur	LS	KNN	SVM	RF	XGB	LGBM	CB	NV
All	63%	60%	63%	66%	65%	65%	67%	15%
PC	58%	54%	60%	63%	61%	61%	62%	56%
MI	62%	59%	60%	62%	61%	61%	62%	60%
RF	62%	59%	60%	65%	65%	64%	66%	61%
LS	63%	62%	65%	67%	66%	64%	67%	13%
HF	57%	82%	73%	83%	82%	83%	79%	40%

The Random Forest and KNN models gave the best performance with F1-score of 83% and 82% on the Hybrid Selection features. This shows that careful feature selection determines classification accuracy.

3.4 Model Comparison Analysis

The following is a description of the model performance accuracy comparison which can be seen in Table 2.

Table 2 Model Performance Comparison

Model	Feature	Accuracy	F1-Score Macro	F1-Score Weighted	Recall
RF	All	81%	78%	81%	84%
	PC	83%	79%	83%	86%
	MI	82%	79%	82%	85%
	RF	84%	81%	83%	88%
	L1	83%	80%	82%	87%
	HF	84%	82%	84%	92%
KNN	All	80%	76%	80%	82%
	PC	83%	79%	82%	85%
	MI	82%	78%	81%	84%
	RF	83%	80%	82%	87%
	L1	82%	79%	81%	86%
	HF	84%	81%	84%	89%
XGB	HF	83%	80%	83%	85%
SVM	HF	81%	79%	81%	78%
LR	HF	81%	77%	80%	76%

In addition to F1-score, other metrics such as recall on minority classes are the main focus due to unbalanced data. The Random Forest model had the highest recall of 92%, outperforming other models in detecting the minority class (Class 2). In contrast, Naïve Bayes performed poorly in most scenarios, except when using all features for Class 0 with a precision of 84%. Figure 3.7 below shows the performance comparison graph between models.

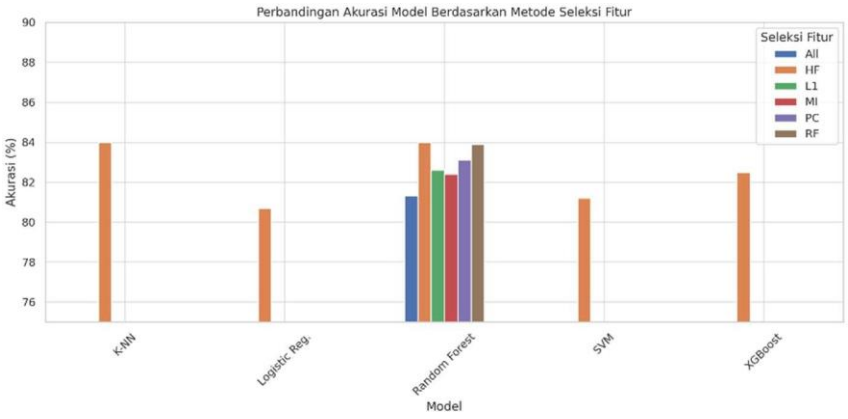

Figure 8 Model Performance Comparison

3.5 Interpretation and Implementation

The results of this study show that the combination of a tree-based algorithm, such as Random Forest, with a Hybrid Feature Selection approach provides the most optimal performance in predicting the type of secondary school. This finding provides practical implications that schools can use this approach as a tool to map student potential and provide

recommendations for educational pathways based on grades and achievement data.

CONCLUSION

This study successfully built a classification model to predict the type of senior high school (SLTA) that students choose based on academic grades and non-academic achievements. By applying the SEMMA approach, data analysis was carried out through the stages of exploration, feature selection, dimension reduction, modeling, and performance evaluation.

The results of feature selection using a hybrid approach resulted in 14 main features that were most relevant. Dimensionality reduction using PCA helped simplify the data structure and showed an initial visualization of the distribution between classes. Eight classification algorithms were tested on various combinations of features. Random Forest and K-Nearest Neighbors (KNN) models proved to give the best performance with an F1-score of 83% and 82% respectively on the Hybrid Feature Selection results.

From these results, it can be concluded that proper feature selection greatly affects the quality of classification. A hybrid approach that combines the strengths of several feature selection methods can significantly improve model accuracy and generalization. In addition, tree-based models such as Random Forest, XGBoost, and LightGBM also proved reliable in handling complex and imbalanced educational data.

Practically, the resulting model can be used as a decision-making tool in schools, especially in providing further education recommendations to students. This system can assist counseling teachers and homeroom teachers in mapping student potential more objectively and purposefully.

For future research, it is recommended to expand the features used, including career interests, family background, or psychological aspects of students. The development of a web-based system interface or mobile application can also support the implementation of the model results in daily educational practice.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to all those who have provided support in the process of preparing this research. Special thanks go to SMPN 2 Balong Ponorogo for the permission and support in providing data, which is an important part of this research.

REFERENCES

[1] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," Smart Learn. Environ., vol. 9, art. no. 11, pp. 1-17, 2022.

[2] Y. Badrani, A. E. Moustapha, and H. O. El Ghazi, "Classification for educational track prediction in Morocco using decision trees and random forest," Int. J. Educ. Dev. Using Inf. Commun. Technol., vol. 18, no. 1, pp. 87-101, 2023.

[3] M. Psyridou, K. Papamitsiou, and A. Economides, "Comparative study on machine learning algorithms for academic dropout prediction," Comput. Educ., vol. 177, art. no. 104370, 2024.

[4] S. Malik, H. A. Khattak, A. S. Latif, A. Javed, and S. Ahmad, "Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms with machine learning models," Sci. Rep., vol. 13, no. 1, art. no. 21298, 2023.

[5] X. Ren, "A hybrid model combining environmental analysis and machine learning for predicting AI education quality," Sci. Rep., vol. 13, no. 1, art. no. 12577, 2023.

[6] J. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 14, no. 2, art. e1477, 2024.

[7] D. A. Garcia, R. García-Sánchez, and M. S. Pons, "Machine learning in education: a bibliometric review," Educ. Inf. Technol., vol. 29, pp. 1081-1106, 2024.

[8] A. S. Gurusamy and G. Marimuthu, "Prediction of student academic performance using hybrid ensemble models," J. Intell. Fuzzy Syst., vol. 45, no. 2, pp. 1733-1746, 2023.

[9] R. García-Sánchez, D. A. Garcia, and A. R. Silla, "Application of machine learning in dropout prediction: a systematic review," Educ. Sci., vol. 13, no. 2, art. no. 180, 2023.

[10] R. H. Devadiga and P. M. Iyer, "Automated student performance prediction using feature engineering and machine learning," Int. J. Educ. Technol. High. Educ., vol. 20, no. 1, art. no. 47, 2023.

[11] A. G. M. M. Kandil and E. F. Elkhatib, "Enhancing student academic performance prediction using optimized hybrid feature selection," Expert Syst. Appl., vol. 213, art. no. 119193, 2023.

[12] N. A. Al-Khateeb, "Using machine learning algorithms to predict students' performance and identify at-risk students," Educ. Inf. Technol., vol. 28, pp. 2099-2119, 2023.

[13] N. A. Salim and Z. B. Abdul-Rahman, "A comparative analysis of feature selection techniques for educational data classification," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 3, pp. 11-18, 2023.

[14] H. H. Zhu, Y. J. Liu, and J. Y. Wei, "Predicting student performance based on learning behavior data with LSTM," IEEE Access, vol. 11, pp. 4526-4535, 2023.

[15] C. Bentéjac, A. Csörgo, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artif. Intell. Rev., vol. 54, pp. 1937-1967, 2021.