Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 2, May 2026, pp. 263-269
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v5i2.475

COMPARATIVE STUDY OF CLASSIFICATION MODELS IN PROCESSING STUDENT TEST SCORES DATASETS

Rico Pramestiawan[*1], Arry Verdian[2], Chindu Lintang Bhuana[3], Lilik Joko Susanto[4]

[1,2,3,4]Department of Education Informatics, STKIP Rosalia, Metro, Indonesia
Email: [1]ricosimple25@gmail.com, [2]vedian.2637@gmail.com, [3]chindu.lintangbhuana, [4]lilikjoko09@gmail.com

(Received: April 30, 2026; Revised: May 8, 2026; Published: May 11, 2026)

Abstract

The development of Machine Learning (ML) has contributed significantly to the field of education, particularly in analyzing student academic data to support data-driven decision-making. Predicting student exam results is important for identifying academic performance patterns, detecting potential failures, and improving learning interventions. However, variations in student characteristics and dataset complexity require the selection of appropriate classification models to achieve optimal prediction performance. This study aims to compare the effectiveness of several ML classification models in predicting student exam results using a student academic dataset. The dataset consists of 306 records, seven attributes, and five grade classes (A, B, C, D, and E), including attendance, quiz scores, midterm examination scores, final examination scores, and assignment scores. Data preprocessing was conducted to handle missing values, duplication, inconsistencies, and outliers. The dataset was split into training and testing data with a ratio of 75:25 and evaluated using 10-fold cross-validation. Several classification models were applied, including k-Nearest Neighbour (kNN), Decision Tree, Naive Bayes, Support Vector Machine (SVM), and Random Forest. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The experimental results showed that Random Forest achieved the best performance with an accuracy of 73.9%, precision of 74.0%, recall of 73.9%, and F1-score of 73.9%, followed by Naive Bayes and Decision Tree. Meanwhile, SVM produced the lowest performance among the tested models. The findings indicate that Random Forest is the most effective method for predicting student exam results and has strong potential to support educational decision-making systems.

Keywords: machine learning; classification; student test scores; model comparison; model evaluation.

1. INTRODUCTION

The development of information technology, especially in the fields of data mining and Machine Learning (ML), has made a significant contribution to the world of education [1]. One application is in analysing student academic data to support more accurate and data-based decision making. Student exam result datasets are an important source of information that can be used to predict academic performance, identify potential failures, and design more effective learning interventions[2]. However, the complexity of the data and variations in student characteristics require the use of appropriate analysis methods so that prediction results have a high level of accuracy.

Various classification models have been used in previous research to predict student learning outcomes, such as Decision Tree, Naive Bayes, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM)[3]. Recent studies show that there is no one model that consistently excels on various types of datasets, because model performance is very dependent on the characteristics of the data used, such as the number of features, data distribution, and the presence of noise and imbalanced data. Therefore, choosing the right model is a crucial aspect in producing an optimal prediction model.

Although various studies have discussed the application of classification models in educational contexts, most studies still focus on the use of one or two models without conducting comprehensive comparisons using diverse evaluation metrics [4]. Apart from that, there are still limitations in using specific and contextual student exam result datasets, especially in certain educational environments. This shows a research gap in the form of a lack of systematic comparative studies to evaluate the effectiveness of various classification models on student exam result datasets with a comprehensive evaluation approach.

Several ML studies on students have been conducted. Research by Rohayani and Umam, (2025) implemented the Backpropagation Algorithm to predict study program determination based on student grades. The results of the research using classification performance with Rapidminer software on targets which produced the greatest accuracy was 77.42% [5]. Furthermore, research by Putra et al, (2025) grouped student scores using the K-Means method. Tests carried out using the silhouette coefficient showed a value of 0.458. This value shows that the grouping results obtained are quite satisfactory, although they have not yet reached the optimal level [6]. Research by Surohman et al,

(2021) conducted a correlation analysis between students' profiles and academic scores using the k-means algorithm. The results of this research show that the elbow method for the most ideal number of clusters in this research is between 3 and 4 clusters, where the highest Silhoutte Coefficient value is 0.8103 for grouping 3 clusters [7]. Several studies apply ML algorithms to student-related datasets, but comparisons of several ML models on student grade datasets have not been found

Based on these problems, this research aims to compare the effectiveness of several ML classification methods in predicting student exam results. The model used in this research includes several popular methods that have different characteristics. Model performance evaluation is carried out using various metrics, such as accuracy, precision, recall, and F1-score, to provide a more comprehensive picture of the performance of each model. It is hoped that the results of this research can contribute to determining the most appropriate classification method for analysing educational data, as well as becoming a reference for further research in developing a prediction system for student learning outcomes.

2. RESEARCH METHODS

This research uses a quantitative approach with experimental methods to compare the performance of several ML classification models in predicting student exam results. The research stages were designed systematically starting from data collection to model evaluation. The framework design is illustrated in Figure 1.

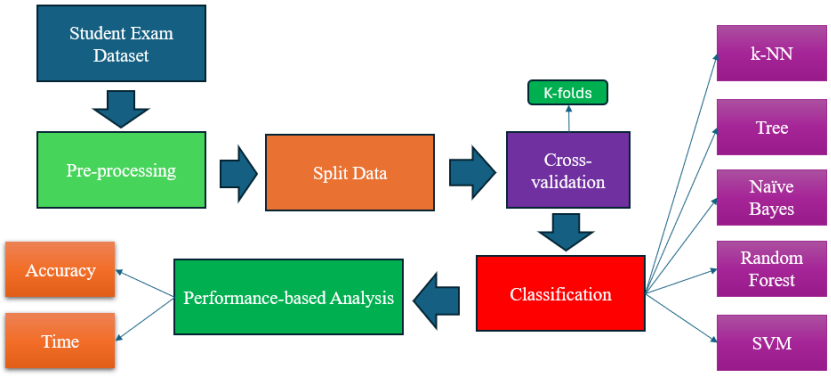


Figure 1. Framework design

After the diabetes classification is carried out, a comparative analysis is then carried out by calculating the accuracy value and the computing time required for the ML model to build the model

3.1. Dataset

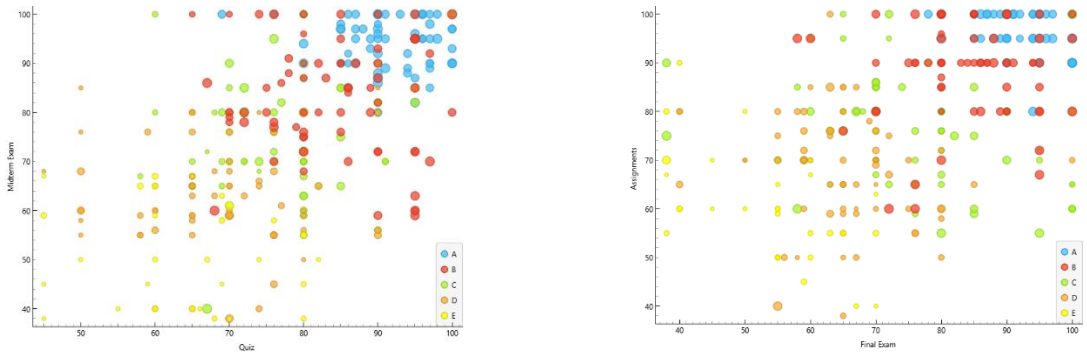
The data set used in this research is student exam score data which includes several attributes, such as assignment scores, midterm exam scores (UTS), final semester exam scores (UAS), attendance, and other supporting attributes. This dataset consists of 306 records, seven attributes, and 5 classes (A, B, C, D, dan E). Data can be obtained from educational institutions or relevant public datasets. The features used in this dataset are presented in Table 1.

Table 1. Features and attribute dataset

ID	Features	Information
f1	Name	Student name
f2	Attendance	Attendance value
f3	Quiz	Quiz exam scores
f4	Midterm Exam	Midterm exam scores
f5	Final Exam	Final exam or UAS scores
f6	Assignments	Assignment value
f7	Grade	Quality Letters

3.2. Pre-processing

The pre-processing stage is carried out to ensure data quality before it is used in the model training process. The steps taken include data cleaning: Addressing missing values, data duplication and inconsistencies (outliers). The results of the improved dataset are shown in Figure 2.



(a) Data inliers quiz and midterm exam (b) Data inliers final exam and assignment

Figure 2. Data inliers

Next is split data where the dataset is split into two parts, namely training data (training set) and test data (testing set) with a ratio of 75:25. It is based on the results of our comparison tests that in this case, splitting using this ratio has the best impact on model performance.

3.3. Cross-validation (k-fold)

Classification is a form of data mining technique that is currently popular [8]. This strategy uses various methods to assess available data to produce diabetes predictions [9]. The classification model will be validated using k-fold cross-validation. The cross-validation method is generally used for training sets [10]. Figure 3 displays the cross-validation procedure

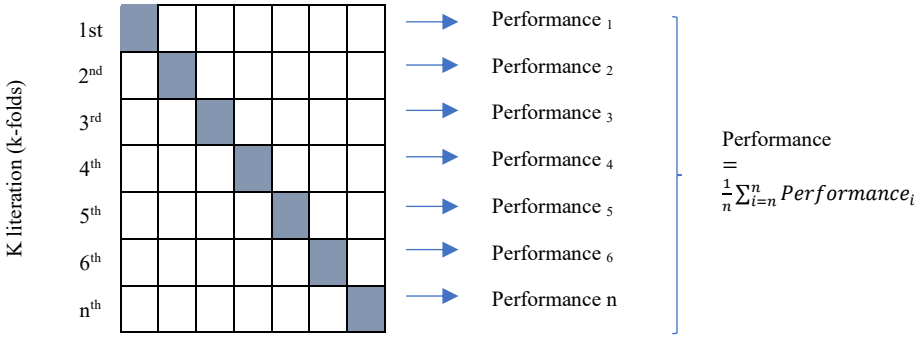


Figure 3. Procedure k-fold Validation

3.4. Classification

At this stage, several machine learning classification models are applied and their performance is compared [11]. The models used in this research include K-Nearest Neighbour (KNN), Decision Tree, Naive Bayes, Support Vector Machine (SVM), Random Forest [12]. Each model is trained using training data and tested using test data to get an idea of the performance of each model.

3.5. Model Performance Evaluation

This study examines how the Confusion Matrix can be used to measure accuracy and error rates. Confusion matrix is a table used to evaluate the performance of classification models in ML. This table compares the model's predicted results with the actual (actual) values from the data, so it can provide an in-depth picture of the model's strengths and weaknesses [13]. The confusion matrix variable is displayed in Table 2.

Table 2. Confusion Matrix

Class	**Positive Prediction**	**Negative Prediction**
Positive Actual	Number of True Positive (TP)	Number False Negative (FN)
Negative Actual	Number of False Positive (FP)	Number True Negative (TN)

Accuracy is a method for evaluating the performance of an ML model. These variables can be obtained from the Confusion Matrix in Table 3 and calculated using equation (1-4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2x \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

3. RESULTS AND DISCUSSION

3.1. Test Results

We use Orange software (version 3.39). This platform simplifies the construction of multiple data analysis techniques[14]. Orange has the ability to categorize, perform regression, classification, remove features, create association rules, and adjust classes to datasets [15]. We use k-fold=10, split data (training set 75%, and testing 25%) because our findings are this option is the best for ML algorithm classification.

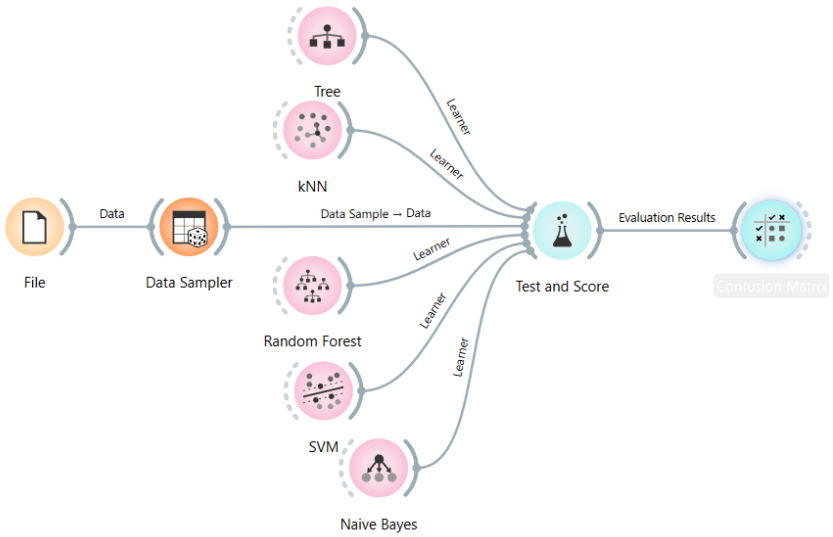


Figure 4. Model testing design in Orange tools

After carrying out the training and testing process on several ML models, the results of the model performance evaluation were obtained based on accuracy, precision, recall and F1-score metrics. The performance comparison results of each model are shown in Table 3.

Table 3. Classification Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.648	0.597	0.648	0.610
kNN	0.670	0.682	0.670	0.662
Tree	0.687	0.688	0.687	0.685
Naive Bayes	0.696	0.708	0.696	0.697
Random Forest	**0.739**	**0.740**	**0.739**	**0.739**
Average	0.688	0.683	0.688	0.678

Based on these results, the Random Forest model shows the best performance with an accuracy rate of 74%, followed by Naive Bayes with an accuracy of 69.6% followed by the Tree model. Meanwhile, SVM has the lowest performance among the models tested. Based on the average test value of the model, ranking is then carried out to obtain comprehensive information on the model performance shown in Figure 5.

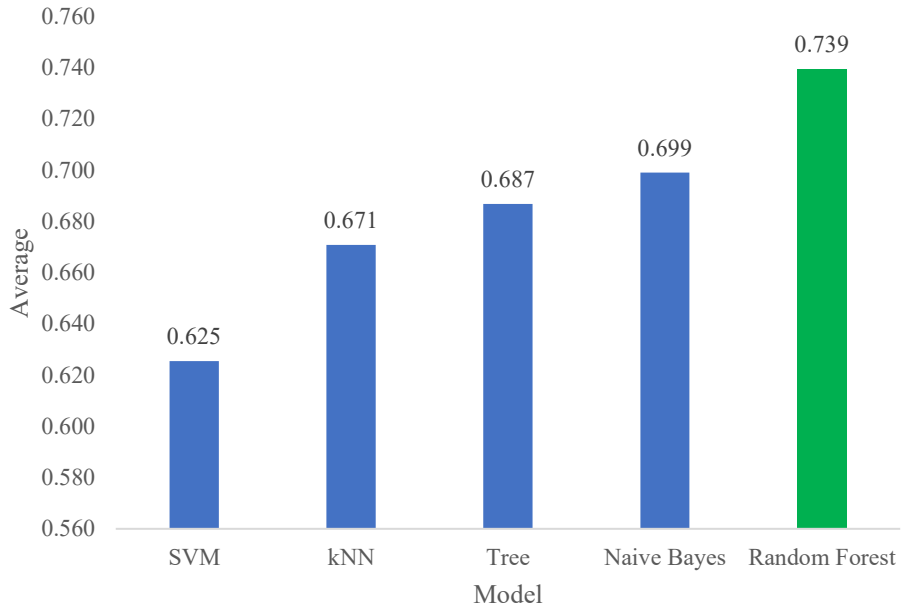


Figure 5. Machine learning model performance

Next, we compared the computing time required when building the model on testing data. The computing time for each model is shown in Figure 6.

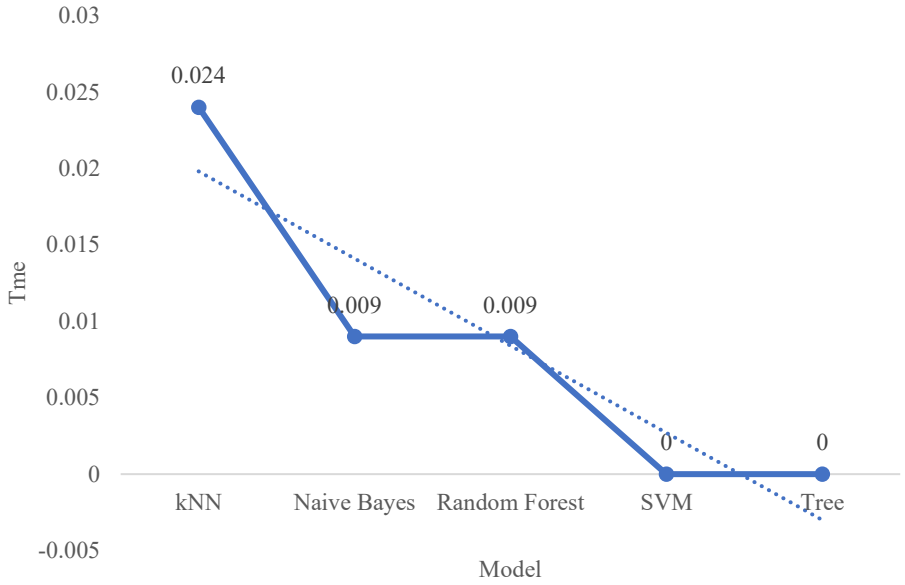


Figure 6. Computational time building the model

3.2. Model Performance Evaluation

The higher performance of Random Forest is due to its ability to combine several decision trees, thereby reducing the risk of overfitting and increasing model generalization. Support Vector Machine also shows good performance because of its ability to separate classes with optimal margins, especially on datasets with quite complex dimensions. On the other hand, the Naive Bayes and k-NN models show quite good performance but are sensitive to the choice of k parameters and the data scale. Decision Trees provide relatively stable results and are easy to interpret but tend to experience overfitting if no pruning is carried out. SVM and Tree, despite having advantages in computing speed, show lower performance. This is likely due to the independence between features not being fully fulfilled in the student test score dataset.

3.3. Research Implications

The results of this research can be used as a reference for educational institutions in selecting appropriate machine learning models for classifying student achievement levels. By choosing the right model, the academic data analysis process can be carried out more accurately and efficiently, thus supporting better decision making in the world of education.

3.4. Discussion

The results of this study demonstrate that different ML classification models exhibit varying performance in predicting student exam results, indicating that model selection plays an important role in educational data analysis. Based on the experimental results, the Random Forest model achieved the highest performance with an accuracy of 73.9%, precision of 74.0%, recall of 73.9%, and F1-score of 73.9%. These findings indicate that Random Forest is more effective in handling student exam datasets with multiple academic features, including attendance, quiz scores, midterm examinations, final examinations, and assignments. The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which combines multiple decision trees to improve prediction robustness and reduce the risk of overfitting. This capability enables the model to better generalize unseen data, particularly in educational datasets where student performance patterns may vary significantly.

The Naive Bayes model showed the second-best performance with an accuracy of 69.6%, indicating that probabilistic classification methods can still perform relatively well in student academic prediction tasks. The model assumes independence among features, which may partially align with student assessment components such as attendance, assignments, and examinations. However, in real educational settings, these features are often interrelated, which may explain why Naive Bayes did not outperform Random Forest. Similarly, the Decision Tree model achieved stable results with an accuracy of 68.7%, demonstrating its ability to model decision rules transparently and provide interpretable outcomes for educational stakeholders. However, the model may suffer from overfitting when handling limited datasets or when tree pruning is not optimally applied.

Meanwhile, k-Nearest Neighbour (kNN) produced moderate performance with an accuracy of 67.0%. This suggests that similarity-based learning methods can classify student performance reasonably well, although they are highly dependent on distance calculations and parameter selection, particularly the value of *k*. The performance of kNN may also be influenced by feature scaling and the diversity of student characteristics in the dataset. On the other hand, the Support Vector Machine (SVM) model demonstrated the lowest performance with an accuracy of 64.8%. Although SVM is generally effective in handling complex classification boundaries, its lower performance in this study may be caused by the characteristics of the dataset, which consists of relatively limited records (306 samples) and multiple grade categories (A–E). The complexity of multi-class classification and overlapping feature distributions may reduce SVM effectiveness in identifying clear decision boundaries.

The comparison of computational time also revealed important insights regarding model efficiency. Although Random Forest achieved the highest predictive performance, computational cost must also be considered in real-world educational implementations. Simpler models such as Decision Tree or Naive Bayes may still be preferable when computational efficiency and interpretability are prioritized. However, for predictive systems requiring higher accuracy, Random Forest appears to be the most suitable approach based on the results of this study.

Overall, the findings highlight that the effectiveness of ML models in educational datasets depends on data characteristics, feature interactions, and classification complexity. The study contributes to the growing body of educational data mining research by providing a systematic comparison of multiple classification models using comprehensive evaluation metrics. The results also suggest that integrating robust ML techniques into educational systems can support more accurate academic performance prediction, early identification of at-risk students, and data-driven educational decision-making.

4. CONCLUSION

This research has evaluated and analysed the performance of several ML models in classifying student exam score datasets. Based on the results of the experiments that have been carried out, it can be concluded that each model has different characteristics and levels of performance. The Random Forest model shows the best performance (73.9%) compared to other models with the highest accuracy, precision, recall and F1-score values. This shows that the ensemble learning approach is effective in increasing the generalization ability of the model on student test score datasets. Support Vector Machine also provides competitive performance, especially in handling complex class separations. Meanwhile, models such as Naive Bayes and Decision Tree provide quite good results but have limitations such as the potential for overfitting and sensitivity to parameters. The SVM model although superior in computational efficiency, shows relatively lower performance due to the assumption of feature independence which does not fully match the characteristics of the data. Overall, this research confirms that selecting the right machine learning model greatly influences classification results, and the use of multiple evaluation metrics is very important to get a more comprehensive picture of model performance.

In further research development, several things can be done, namely the use of larger and more diverse datasets. Future research is recommended to use a larger dataset and include more variables, such as students' social, economic and psychological factors. Exploration Other models can be tested against other more complex models, such as advanced ensemble methods such as Gradient Boosting, XGBoost or deep learning-based approaches. Deeper Hyperparameter Optimization uses techniques such as Grid Search or Random Search to get the best parameters from each model.

5. ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to the Faculty of Health, Aisyah Pringsewu University (UAP) and STMIK Rosalia Metro, for helping validate the anonymous breast cancer patient data used in this research. We also thank the Machine Learning (ML) Research Group at UAP and STKIP Metro for their valuable input during the model development and evaluation phase. Special awards are given to doctors whose clinical insight contributes significantly to the interpretation of breast cancer prediction results

REFERENCES

[1] Z. Syahputra and R. Kurniawan, "Journal of Computer Networks , Architecture and High Performance Computing Journal of Computer Networks , Architecture and High Performance Computing," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 7, no. 1, pp. 341–352, 2025.

[2] A. Wantoro, Zulkifli, P. Bintoro, T. H. Andika, F. Ardhy, and A. N. Al Aziz, "Performance Evaluation of Classification Multi Algorithms on Small Dataset: A Comparative-Based Analysis," in *2025 Tenth International Conference on Informatics and Computing (ICIC)*, 2025, pp. 1–6. doi: 10.1109/ICIC68054.2025.11309491.

[3] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, "Meta-iavp: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation," *Int. J. Mol. Sci.*, vol. 20, no. 22, 2019, doi: 10.3390/ijms20225743.

[4] C. Karima and W. Anggraeni, "Performance Analysis of the Ada-Boost Algorithm For Classification of Hypertension Risk With Clinical Imbalanced Dataset," *Procedia Comput. Sci.*, vol. 234, pp. 645–653, 2024, doi: https://doi.org/10.1016/j.procs.2024.03.050.

[5] H. Rohayani and M. C. Umam, "Prediksi Penentuan Program Studi Berdasarkan Nilai Siswa dengan Algoritma Backpropagation," *J. Inf. Syst. Res.*, vol. 3, no. 4, pp. 651–657, 2022, doi: 10.47065/josh.v3i4.1935.

[6] R. D. K. Putra, K. S. Palupi, and N. Wakhidah, "Pengelompokkan Data Nilai Mahasiswa Menggunakan Metode K-Means," *J. Algoritm.*, vol. 6, no. 1, pp. 88–99, 2025, doi: 10.35957/algoritme.v6i1.11313.

[7] Suraohman, L. Fabrianto, F. Riza, and N. M. Faizah, "Korelasi Antara Profil dan Nilai Akademis Siswa dengan Menggunakan Algoritma K-Means," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 845–852, 2021, doi: 10.25126/jtiik.202183034.

[8] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques*. London, 2019. doi: 10.1007/979-981-13-8798-2-12.

[9] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, "Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.

[10] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022, doi: https://doi.org/10.1016/j.jrmge.2022.03.002.

[11] A. Agliata, D. Giordano, F. Bardozzo, S. Bottiglieri, A. Facchiano, and R. Tagliaferri, "Machine Learning as a Support for the Diagnosis of Type 2 Diabetes," *International Journal of Molecular Sciences*, vol. 24, no. 7. 2023. doi: 10.3390/ijms24076775.

[12] A. Wantoro, A. F. Yuliana, D. Yana, A. Andini, and I. Awaliyani, "Optimizing Type 2 Diabetes Classification with Feature Selection and Class Balancing in Machine Learning," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2625–2637, 2025.

[13] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," *J. Phys. Conf. Ser.*, vol. 1229, no. 1, 2019, doi: 10.1088/1742-6596/1229/1/012055.

[14] B. Imran, H. Hambali, A. Subki, Z. Zaeniah, A. Yani, and M. R. Alfian, "Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer," *J. Pilar Nusa Mandiri*, vol. 18, no. 1, pp. 37–46, 2022, doi: 10.33480/pilar.v18i1.2912.

[15] E. Akkaya and S. Turgay, "Unveiling the Power: A Comparative Analysis of Data Mining Tools through Decision Tree Classification on the Bank Marketing Dataset," *Wseas Trans. Comput.*, vol. 23, pp. 95–105, 2024, doi: 10.37394/23205.2024.23.9.