Comparative Analysis of Machine Learning Algorithm and Feature Selection on Student Grade Small Datasets

Agus Wantoro[1], Nur Aminudin[2]*, Dita Septasari[3], Ningsiah[4], Ikna Awaliyani[5], Anazwa Novia Zahara[6]

[1,6]Informatics Engineering, Faculty of Technology and Informatics, Aisyah University, A. Yani No.1A, Pringsewu, Indonesia
[2]Software Engineering, Faculty of Technology and Informatics, Aisyah University, A. Yani No.1A, Pringsewu, Indonesia
[3,5]Information Technology Education, Faculty of Teacher Training and Education, Aisyah University, A. Yani No.1A, Pringsewu, Indonesia
[4]Pharmacy, Faculty of Health, Aisyah University, A. Yani No.1A, Pringsewu, Indonesia

*Corresponding author: nuraminudin@aisyahuniversity.ac.id

Abstract. Assessment of student academic performance is an important aspect in improving the quality of the learning process in higher education. With the development of Machine Learning (ML) techniques, predictive analysis of student assessment data can be carried out more accurately and efficiently. This research aims to analyze the comparative performance of several ML algorithms and evaluate the effectiveness of various feature selection techniques on student assessment datasets. The dataset used is a limited dataset that contains attributes that represent components of student academic assessment. The research process includes determining cross-validation, applying ML algorithms, applying feature selection techniques, and performance evaluation. The ML algorithms compared are Naïve Bayes, K-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Tree, Random Forest, and AdaBoost. Several feature selection techniques used are Information Gain (IG), Gain Ratio (GR), Gini Dicrease (GD), ReliefF. Evaluation of the performance of the ML algorithm uses the Confusion Matrix in the form of accuracy, precision, recall and F1-score, while the performance evaluation of the feature selection technique uses Spearman Correlation. Experimental results show that the SVM algorithm shows the best performance followed by k-NN. However, based on computing time testing, the Tree algorithm shows the fastest time. Apart from that, the comparison results of feature selection techniques show that Gini Decrease has the best correlation. The findings of this research contribute to selecting a more effective educational data analysis method and can be a reference for developing ML algorithms in student academic evaluation.

Keywords: Comparative, Machine Learning, Feature Selection, Small Dataset, Student.

How to cite (in APA style):
Wantoro, A., Aminudin, N., Septasari, D., Ningsiah, Awaliyani, I., & Zahara, A. N. (2026). Comparative Analysis of Machine Learning Algorithm and Feature Selection on Student Grade Small Datasets. *Proceeding of IJC-MaSTEd*, 1(1), 1-10.

DOI: https://doi.org/10.31571/ijcmasted.v1i1.1010

1 Introduction

The development of information and computing technology has encouraged the use of Machine Learning (ML) in various fields, including the education sector (Korkmaz & Correia, 2019). Educational institutions currently produce academic data in ever-increasing amounts, such as data on course grades, attendance and student performance (Sorensen, 2019). This data has great potential to be analyzed to support academic decision making, such as predicting graduation, evaluating

learning performance, and identifying at-risk students. One technique that is widely used to carry out predictive analysis is a numerical-based Machine Learning (ML) algorithm.

In general, algorithms such as Naïve Bayes, k-NN, Tree, SVM, Random Forest, and AdaBoost are often used in academic data modeling (Boateng et al., 2020). However, most previous studies focused on large datasets with complex feature variations. In fact, in practice, we often find datasets with relatively small amounts of data and simple numerical types, such as data on course grades in one class or one study program.

Problems arise when ML algorithms are applied to small datasets. Some algorithms tend to experience overfitting, while other algorithms require quite large amounts of data to produce optimal performance (Sorensen, 2019). This condition raises the question of which algorithm is the most stable and effective when used on small numerical datasets. Until now, there is still limited research that specifically compares the performance of several ML algorithms in the scenario of small academic datasets with homogeneous numerical characteristics. Some studies only present implementation results without in-depth comparative analysis regarding performance stability, accuracy, precision, or generalization ability of the model. In addition, comparative studies of the performance of feature selection techniques have not been carried out on limited datasets.

Based on these research gaps, a systematic comparative analysis is needed to evaluate the performance of various ML algorithms and feature selection techniques on small category course grade datasets of the numerical type. This research aims to: (1) implement several ML algorithms on a dataset of course grades with a limited amount of data, (2) compare the performance of algorithms based on relevant evaluation metrics, such as accuracy, precision, recall, and F1-score, and (3) identify the most optimal and stable algorithm in handling small numerical datasets (4) comparative analysis of the performance of the feature selection technique using the Spearman Correlation technique

It is hoped that the results of this research can contribute to selecting appropriate methods for analyzing small-scale academic data, as well as becoming a reference for educational researchers and practitioners in implementing ML and feature selection techniques more effectively and efficiently.

2 Method

2.1 Type of research

This research uses a quantitative approach with comparative experimental methods. This approach aims to compare the performance of several ML algorithms in processing small category course grade datasets of numerical type. Experiments are carried out in a controlled manner with the same test scenarios for each algorithm so that the comparison results are objective and measurable

2.2 Research Design

This section describes the procedures and stages of the research. The research stage begins with data collection. The next step is to select the most optimal number of k-fold validations. The next step is testing algorithms such as Naïve Bayes, k-NN, Tree, SVM, Random Forest, and AdaBoost. Next, carry out comparative analysis in the form of Accuracy, Precision, Recall, and F1-Score. The research framework design is illustrated in Figure 1.

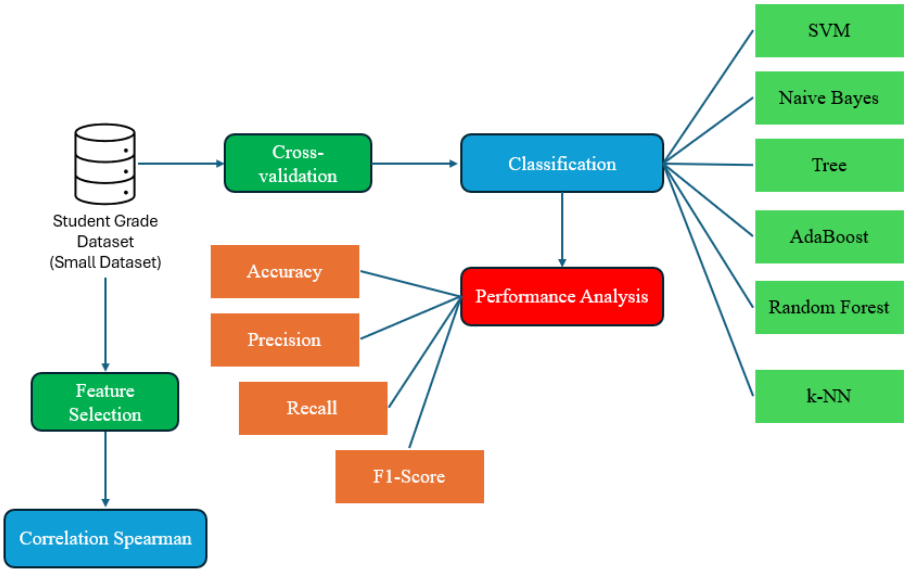

Figure 1. Research design

2.3 Dataset

The dataset used is student course grade data in the small category (306 data). All attributes are of numeric types, namely Assignment, Quiz Exam, Midterm Exam, Final Exam, and Grade. The variables in this research consist of two, are (a) independent variables (features): Assignment, Quizzes, Midterm Exam, Final Exam, and the dependent variable (target) is Grade. Dataset attributes are shown in Table 1.

Table 1. Atribute dataset

ID	Attribute	Type
f1	Assignment	Number
f2	Quiz exam	Number
f3	Midterm exam	Number
f4	Final exam	Number
f5	Grade	Text

2.4 Cross-validation (k-folds)

A classification is a form of data mining technique that is currently popular. It functions similarly to other methods like decision trees and neural networks. These strategies use various methods to assess the available data to produce their prediction (Sulistiani et al., 2024). After the feature selection stage and the features that affect the class label based on ranking are obtained, the next stage is classification. Furthermore, the accuracy, precision, recall, F1-Score, and time required in the classification process are compared using the method of Naïve Bayes, k-NN, Tree, SVM, Random Forest, and AdaBoost. The classification model will be validated using k-fold cross-validation. The cross-validation method is commonly used for training sets (Yan et al., 2022). Figure 2 shows the k-fold cross-validation.

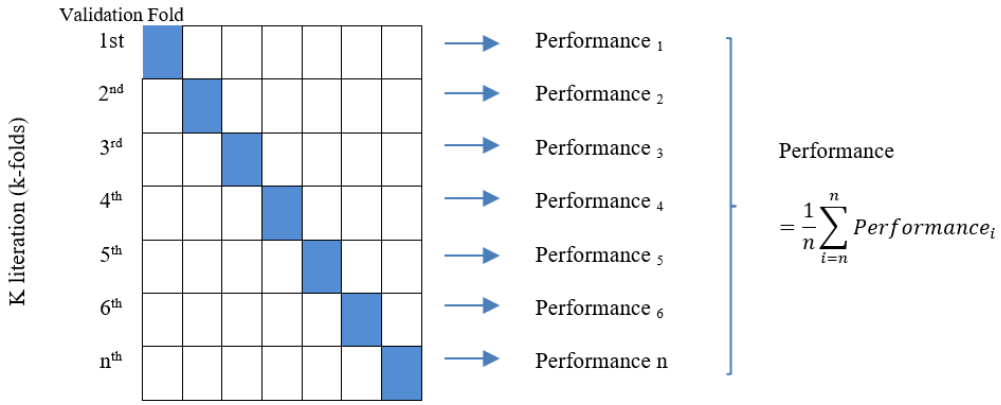

Figure 2. The procedure of k-fold validation

2.5 Performance Analysis

This study examines how the confusion matrix might be used to gauge accuracy and mistake rate. A confusion matrix of size $n \times n$ coupled to a classifier, where n is the total number of classes, displays the anticipated and actual categorization (Ohsaki et al., 2017). The confusion matrix for n=2 is shown in Table 2.

Table 2. Confusin matrix

Class	Predicted positives	Predicted negatives
Actual positives instances	Number of true positives instances (TP)	Number of false negatives instances (FN)
Actual negatives instances	Number of false positives (FP)	Number of true negatives instances (TN)

Calculating prediction accuracy, precision, recall, and f1-score is another method for evaluating and comparing classifiers. Both values can be obtained from the confusion matrix Table 1 and calculated using equation (1-4)

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

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

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

$$f1 - score = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

2.6 Feature Selection

One of the most important things in classification is determining features to get the best performance results (Sulistiani et al., 2024). Data sets used in ML processes usually contain redundant and irrelevant features (Wang et al., 2014), and has no effect on the learning model, and may even reduce the performance of the learning model (Bashir et al., 2019), therefore it is important to perform relevant feature analysis. In this case we apply the Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and ReliefF. These four techniques have good capabilities in selecting features by giving weight to each feature and displaying them in ranking (Casteren, 2017).

2.7 Spearman Correlations

This method is used to measure correlation with feature selection techniques such as Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and ReliefF. These four methods can rank features based on highest to lowest weight. A high correlation value indicates that a technique has good ranking consistency and is reliable (Jia et al., 2021). This testing provides validation of effectiveness in producing competitive and trustworthy rankings. The correlation range is shown in Table 3.

Table 3. Range correlation

Range	Correlation
0.00 – 0.25	Very weak
0.26 – 0.50	Weak
0.51 – 0.75	Moderate
0.76 – 0.99	Strong
1	Perfect

3 Results and Discussion

Tables We use orange software (version 3.3.8). This platform simplifies the construction of several technical analysis data. Orange has the ability to categorize, perform regression, classification, eliminate features, create association rules, and adjust classes in datasets (Popchev & Orozova, 2023). We used cross-validation (k-fold=5) because our findings indicate that this choice is the best for the classification of ML algorithms. The performance results of the algorithms are presented in Figure 3.

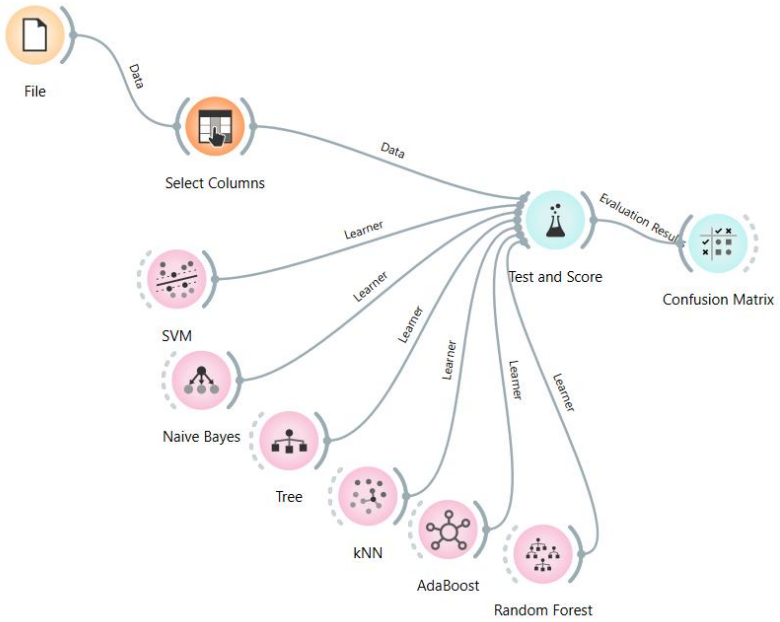

Figure 3. Model testing algorithms

Next, we tested six ML algorithms on a small dataset using the number feature. Testing is carried out by comparing actual data with predicted data using the Confusion Matrix calculation in Figure 4.

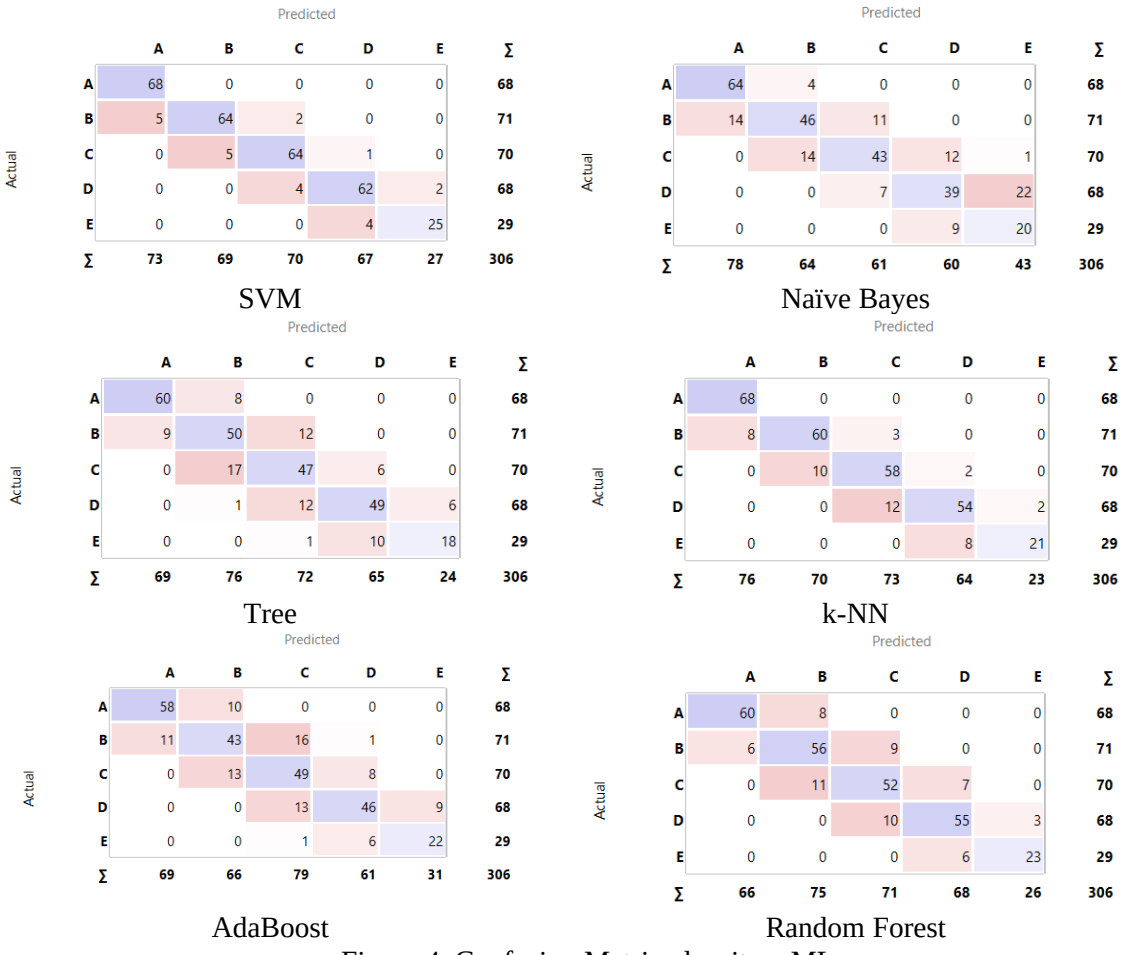

Figure 4. Confusion Matrix algoritma ML

Based on the values in the Figure, calculations are then carried out to obtain information in the form of performance based on accuracy, precision and recall which are shown in Table 4.

Table 4. ML algorithm performance

No	Algorithm	Accuracy	Precision	Recall
1	SVM	**92.5**	**92.5**	**92.5**
2	k-NN	**85.3**	**85.3**	**85.3**
3	Random Forest	78.8	79.5	78.8
4	Tree	73.2	73.2	73.2
5	AdaBoost	71.2	71.5	71.2
6	Naïve Bayes	69.3	69.1	69.3

Table 2 shows the performance of different ML algorithms. The ML algorithm that has the best performance is SVM. This is because this method is effective in both high and low dimensional spaces. SVM works very well even if the number of dimensions (features) is greater than the number of data samples. This makes it well suited for Natural Language Processing (NLP), text classification, and genomics. Additionally, SVM tend to be more robust against overfitting, especially on high-dimensional datasets, because they focus on maximizing the margin (distance) between classes.

Meanwhile, the worst performance of the ML algorithm is Naïve Bayes. This is because this method has an independence assumption. Naive Bayes assumes that all features (input variables) are independent of each other and contribute independently to the class probability. In the real world, features are often correlated which can reduce algorithm accuracy

Next, we carried out a performance comparison analysis using the F1-Score value. This value is used to evaluate the performance classification of ML algorithms which calculate the harmonic average of precision and recall. The F1 score provides a balance between the two metrics, especially useful for imbalanced datasets where accuracy is insufficient, with values ranging from 0 (complete failure) to 1 (perfect). Performance comparison results based on F1-Score values are displayed in Figure 5.

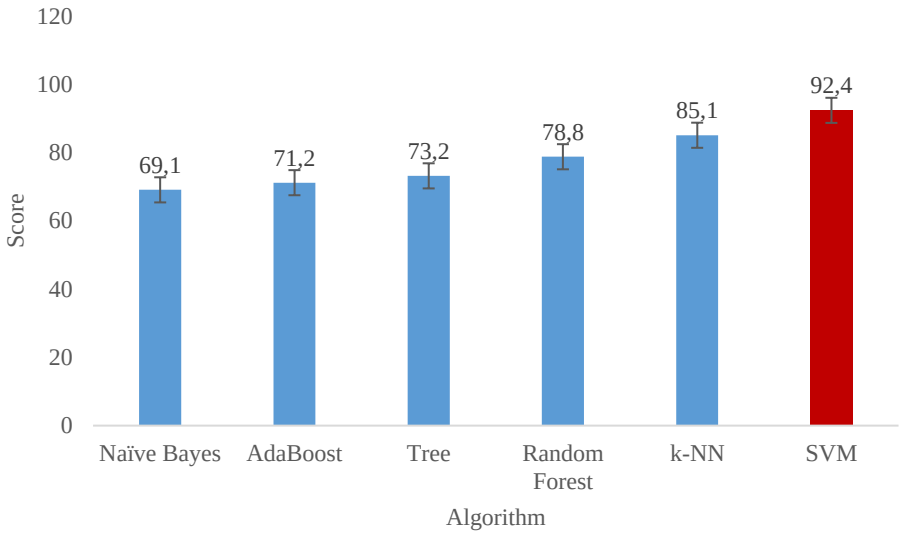

Figure 5. Performance comparison based on F1-Score

Performance comparison results based on F1-Score (Table 5) show different results. In this case, the SVM algorithm shows the best performance. This is because SVM focuses on maximizing the distance between the closest data points (margin), this algorithm is generally more resistant to outlier data compared to some other algorithms. Meanwhile, the Naïve Bayes algorithm shows the worst performance. This is because Naive Bayes assumes that all features or predictors are independent of each other (have no relationship). This is rarely fulfilled because features are often interdependent, which causes biased posterior probabilities and less accurate classification results in predicting grade.

Next, we compared the computing time of each algorithm shown in Figure 6. Figure 6 shows the different computing times. The Tree algorithm shows the fastest time, even though the accuracy of this algorithm is not significant, in terms of time it shows the best. Meanwhile, the k-NN algorithm shows the worst time.

In addition, we carried out a comparative analysis of feature selection techniques such as IG, GR, GD, and ReliefF with the original dataset. This is done to see the performance of each feature selection technique, to obtain information on new findings and references for further research. The results of the ranking comparison of feature selection techniques are shown in Table 5.

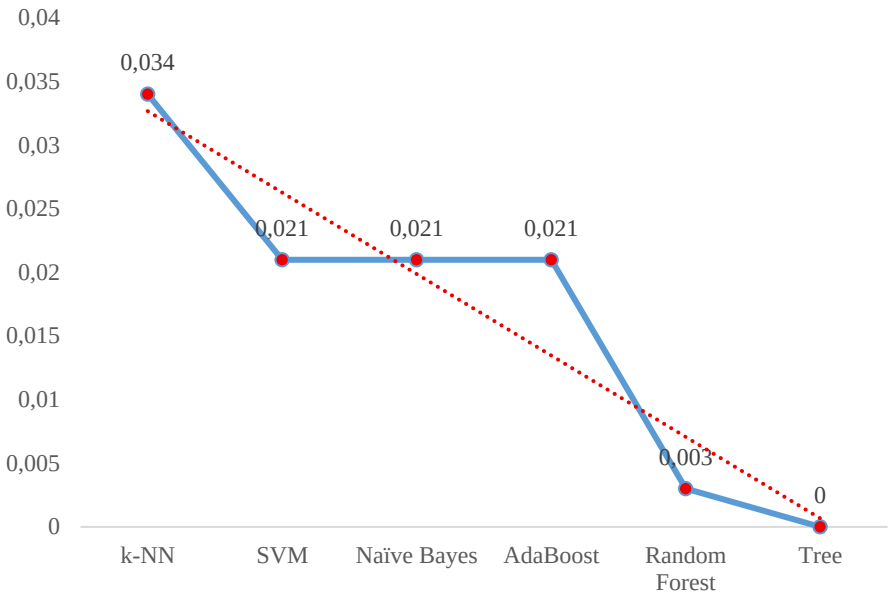

Figure 6. Compute time comparison

Table 5. Comparison of ranking feature selection techniques

No	Feature	Original	IG	GR	GD	ReliefF
1	Attendance (%)	**5**	3	3	4	4
2	Quiz	**4**	5	5	5	5
3	Midterm Exam	**2**	1	1	2	1
4	Final Exam	**1**	2	2	1	2
5	Assignments	**3**	4	4	3	3

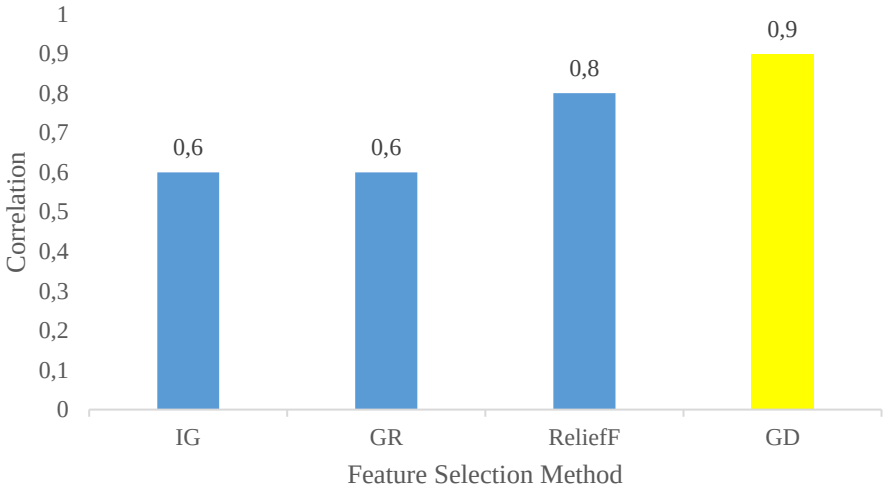

Figure 7. Comparison of correlation values Feature selection techniques

The comparison results (Table 5) then we carried out a ranking comparison using the Spearman correlation method to see the performance of the feature selection technique between the original

data and the ranking using the feature selection technique. The higher the correlation value, the better the performance of the feature selection technique. The comparison results are shown in Figure 7.

4 Conclusion

This research has carried out a comparative analysis of the performance of several Machine Learning (ML) algorithms on student assessment datasets. The ML algorithms used are Naïve Bayes, Tree, AdaBoost, Random Forest, and k-NN. Meanwhile, the feature selection techniques compared are Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and ReliefF. Algorithm performance evaluation uses the Confusion Matrix in the form of accuracy, precision, recall, and F1-score, while performance evaluation of the feature selection technique uses Spearman Correlation. Experimental results show that the SVM algorithm shows the best performance (92.5%) followed by k-NN (85.3%). However, based on computing time testing, the Tree algorithm shows the fastest time. Apart from that, the results of the comparison of feature selection techniques show that Gini Decrease (GD) has the best correlation (0.9) or the "Strong" category, while Information Gain (IG) and Gain Ratio (GR) show the worst correlation.

The findings of this research contribute to selecting a more effective educational data analysis method and can be a reference for developing ML algorithms in student academic evaluation. It is hoped that the results of this research can become the basis for further research in developing prediction models for student academic performance as well as implementing more adaptive and accurate educational analytical systems in higher education environments.

Reference

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving Heart Disease Prediction Using Feature Selection Approaches. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 619–623. https://doi.org/10.1109/IBCAST.2019.8667106

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K -Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network : A Review. *Journal of Data Analysis and Information Processing*, *8*(1), 341–357. https://doi.org/10.4236/jdaip.2020.84020

Casteren, W. Van. (2017). The Waterfall Model And The Agile Methodologies : A Comparison By Project Characteristics-Short The Waterfall Model and Agile Methodologies. *Academic Competences in the Bachelor*, *February*, 10–13. https://www.researchgate.net/publication/313768860

Jia, K., Yang, Z., Zheng, L., Zhu, Z., & Bi, T. (2021). Spearman Correlation-Based Pilot Protection for Transmission Line Connected to PMSGs and DFIGs. *IEEE Transactions on Industrial Informatics*, *17*(7), 4532–4544. https://doi.org/10.1109/TII.2020.3018499

Korkmaz, C., & Correia, A.-P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, *56*(3), 250–267. https://doi.org/10.1080/09523987.2019.1669875

Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H., & Ralescu, A. (2017). Confusion-matrix-based kernel logistic regression for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering*, *29*(9), 1806–1819. https://doi.org/10.1109/TKDE.2017.2682249

Popchev, I., & Orozova, D. (2023). Algorithms for Machine Learning with Orange System. *International Journal of Online and Biomedical Engineering*, *19*(4), 109–123. https://doi.org/10.3991/ijoe.v19i04.36897

Sorensen, Lucy C. (2019). "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. *Educational Administration Quarterly*, *55*(3), 404–446. https://doi.org/10.1177/0013161X18799439

Sulistiani, H., Syarif, A., Muludi, K., & Warsito. (2024). Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder. *Bulletin of Electrical Engineering and Informatics*, *13*(2), 1383–1391. https://doi.org/10.11591/eei.v13i2.6717

Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). An improved feature selection based on effective range for classification. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/972125

Yan, T., Shen, S.-L., Zhou, A., & Chen, X. (2022). Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(4), 1292–1303. https://doi.org/https://doi.org/10.1016/j.jrmge.2022.03.002