International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 5, No. 1, March 2026, pp. 43∼54
ISSN: 2828-5611
DOI: 10.30812/IJECSA.v5i1.6032

Evaluation of Classification Methods for Predicting Junior High School Accreditation Ranks in Indonesia

Miftahul Jannah, Prajna Pramita Izati
Universitas Diponegoro, Semarang, Indonesia

Article Info

Article history:

Received December 16, 2025
Revised January 22, 2026
Accepted February 8, 2026

Keywords:

Area Under Curve (AUC)
Boosting
Classification
Random Forest
Support Vector Machine

ABSTRACT

In Indonesia, school accreditation is a crucial process for assessing educational institutions' eligibility to meet national education standards. However, this process is resource-intensive and requires significant time, manpower, and financial resources. This study aimed to explore the application of machine learning classification methods: Random Forest, Boosting, and Support Vector Machine (SVM) to predict the accreditation ranks of Junior High Schools in Indonesia. The goal was to create an efficient, automated model to predict school accreditation status, improve the accreditation process, and facilitate better resource allocation. Data preparation included handling missing values, reducing dimensionality, and addressing data imbalance. The dataset consisted of 23,954 Junior Schools from 34 provinces, with 37 variables, including 36 predictors and one target variable (accreditation status). The study found that Random Forest outperformed Boosting and SVM, with the highest Area Under Curve (AUC) of 0.8133. Random Forest also demonstrated the lowest average classification error of 19.32%, indicating its superior performance in predicting junior high school accreditation ranks. The results suggest that machine learning models, particularly Random Forest, can provide a more efficient and reliable alternative to manual accreditation evaluations. This approach can optimize educational assessments, improve resource allocation, and provide policymakers with valuable insights to enhance school performance, particularly in underserved regions.

Corresponding Author:

Miftahul Jannah,
Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia.
Email: miftamj@lecturer.undip.ac.id

How to Cite: M. Jannah, and P. P. Izati, "Evaluation of Classification Methods for Predicting Junior High School Accreditation Ranks in Indonesia", *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 5, no. 1, pp. 43–54, Mar. 2026. doi: 10.30812/ijecsa.v5i1.6032.

1. INTRODUCTION

The Indonesian government implements school accreditation to evaluate educational institutions' compliance with national education standards. This process involves evaluating various aspects of a school's performance, including curriculum, teaching quality, and facilities, to ensure a baseline quality of education across the archipelago [1]. This accreditation process, led by the National Accreditation Board for Schools/Madrasahs (BAN-S/M), is essential for maintaining educational quality across the country. However, this procedure requires considerable time, labor, and financial investment. Moreover, the numerous variables involved in the accreditation assessment process intensify the evaluation challenges. Given the significant disparities in data quality and educational resource distribution across regions [2, 3]. There is a critical need for efficient, data-driven alternatives to traditional manual evaluation methods. To address these challenges, previous studies have explored a variety of modeling techniques, ranging from predictive analytics for student performance [3] to specific classification tasks for educational data, including Random Forest and Support Vector Machines (SVM) for predicting Madrasah data accuracy [2], KNN versus MARS for accreditation prediction [4], and boosting for classifying school quality status [5].

For instance, one study utilized K-Nearest Neighbors and Multivariate Adaptive Regression Splines to predict accreditation values in public elementary schools, where evaluation metrics such as specificity and sensitivity indicated that the MARS method outperformed K-NN [4]. Another research has demonstrated the effectiveness of combining Latent Semantic Indexing with SVM for classifying accreditation documents [6], while other investigations have compared Boosting against Random Forest and decision tree algorithms to identify the most influential factors affecting school accreditation rank and educational quality [5]. Furthermore, comparative analyses of Random Forest and SVM have been conducted to predict the accuracy of Madrasah data, highlighting the applicability of these algorithms in educational settings [2].

The methodological robustness of Random Forest has been extensively documented in broader literature, where it excels at handling high-dimensional data, capturing non-linear relationships, and mitigating overfitting through ensemble learning, making it well-suited for complex classification tasks [7–9]. Its feature selection capabilities have been validated across diverse domains, including forest biomass modeling [10] and land cover assessment [8, 11]. Similarly, Support Vector Machines are renowned for identifying optimal hyperplanes to maximize classification margins in high-dimensional spaces [12]. Meanwhile, boosting algorithms, such as Adaptive Boosting, enhance predictive performance by iteratively correcting errors from weak learners, which has proven particularly effective in addressing class imbalance problems within educational datasets [5]. Building upon this theoretical and practical foundation, this study evaluates the application of these advanced classification methods, namely Random Forest, Boosting, and SVM, to predict school accreditation ranks with high accuracy. This study aimed to explore the application of machine learning classification methods: Random Forest, Boosting, and Support Vector Machine (SVM) to predict the accreditation ranks of Junior High Schools in Indonesia.

2. RESEARCH METHOD

2.1. Dataset and Methods

The dataset used in this study comprises data from 23,954 junior high schools across 34 provinces in Indonesia, totaling 86 variables. This data is sourced from the DAPODIK (Education Data and Information) of the Ministry of Education and Culture. The dataset includes various attributes, such as school type (public or private), school facilities (e.g., classrooms, toilets, libraries), student enrollment figures, teacher information, and accreditation rankings. The analysis in this study was conducted using RStudio version 2025.09.1. To predict the accreditation ranks of the schools, three classification algorithms are used in this study:

1. Random Forest

This ensemble learning method operates by constructing multiple decision trees during training and outputting the class with the highest frequency (classification) or the mean prediction (regression) across the individual trees, making it robust against overfitting [7]. The inherent stability and generalization capabilities of Random Forest classifiers arise from their ability to decorrelate individual trees through random feature selection at each node, thereby significantly enhancing predictive performance [13, 14]. Specifically, Random Forests achieve this by training each tree on a bootstrapped sample of the data and considering only a random subset of features at each split, thereby reducing the variance and improving the accuracy [8, 15]. Two critical hyperparameters, the number of trees (ntree) and the number of features considered at each split (mtry), require careful tuning to optimize the Random Forest model's performance [8]. Grid search and K-fold cross-validation are commonly employed for hyperparameter optimization, systematically evaluating various combinations of these parameters to identify the configuration that yields the best model performance and further reduces bias in the estimates [14]. The algorithm of Random Forest is as follows: (1) The algorithm begins by selecting a random

sample from the training data with replacement, a process known as bootstrapping, to grow each decision tree [8], as follows; (2) For each node in the decision tree, a random subset of features is considered for splitting, preventing any single feature from dominating the tree construction [13, 16]. This randomization across features contributes to the decorrelation of individual trees, which enhances the overall stability and predictive power of the ensemble [14]; (3) Each tree is grown to its maximum possible depth without pruning, ensuring that it captures as much information as possible from its bootstrapped sample [9]. (4) For classification tasks, the final prediction is determined by aggregating the predictions of all individual trees using majority voting [13, 17]. This aggregation process effectively mitigates the high variance often observed in individual decision trees, leading to a more robust and accurate classification [18].

2. Boosting

In contrast, boosting algorithms build an ensemble sequentially, where each new model corrects the errors of the previous model, progressively reducing bias and improving overall predictive accuracy [19]. This iterative refinement focuses on misclassified instances from prior iterations by assigning them higher weights, thereby compelling subsequent models to focus on these challenging cases [20]. The boosting algorithm is as follows: (1) Boosting algorithms are initiated by training a weak learner on the original dataset, where all instances are initially assigned equal weights; (2) Subsequently, the performance of this initial learner is evaluated, and the weights of the misclassified instances are increased to emphasize their importance in subsequent training rounds. This iterative process continues, with new weak learners being added and trained to correct the errors of the combined ensemble until a predefined stopping criterion is met or a satisfactory level of accuracy is achieved [20].

3. Support Vector Machine

Support Vector Machines operate on the principle of finding an optimal hyperplane that distinctly separates different classes in a high-dimensional feature space, maximizing the margin between them. This maximization of the margin ensures robust classification performance and generalization to unseen data, even in complex, nonlinearly separable scenarios [21]. The core idea involves transforming the data into a higher-dimensional space where a linear separating hyperplane can be found, and utilizing kernel functions to efficiently compute the dot products in this transformed space without explicitly performing the transformation [22]. The SVM algorithm begins with the identification of support vectors, which are the data points closest to the hyperplane and play a crucial role in defining its orientation and position. The primary objective of SVM is to maximize the margin between these support vectors, thereby enhancing the model's ability to generalize effectively to novel and unseen data [12].

2.2. Research Flow

The research process involves several key steps to ensure a robust classification model for predicting junior high school accreditation ranks. These steps include data preparation, feature selection, model training, and evaluation. The research flow is outlined as follows:

1. Data preparation:

   The raw dataset undergoes several preprocessing steps to ensure its quality:

   (a) Data cleaning: Irrelevant variables, such as school ID, district name, and elementary school-related attributes, are removed to reduce the dimensionality of the data.
   (b) Handling missing data: Missing values are addressed using imputation methods, with some missing data handled by using related variables.
   (c) Feature selection: Only relevant features for classification are retained.

2. Pre-processing:

   (a) Discretization: Several continuous variables, such as the number of classrooms and rooms for specific purposes, are discretized using the Minimum Description Length Principle (MDLP). This process reduces complexity while maintaining essential information for model training.
   (b) Training and testing set partitioning: The dataset is partitioned into training and testing sets using the createDataPartition function in R, ensuring that the training set is used to train the model. In contrast, the testing set is used for model evaluation.

3. Handling imbalance data:

The dataset exhibits an imbalance in accreditation ranks, with most schools having non-A accreditation. To address this issue, undersampling is applied to balance the dataset and ensure equal numbers of observations for both accreditation categories (A and non-A). This step ensures that the classification model is not biased towards the majority class.

4. Model training and evaluation:

   Three classification algorithms—Random Forest, Boosting, and Support Vector Machine (SVM)—are trained on the prepared dataset. Cross-validation (5-fold) is used to consistently evaluate model performance. The following metrics are used to assess the models:

   (a) Accuracy: Measures the overall correctness of the model's predictions.
   (b) Sensitivity: Measures the model's ability to correctly identify positive instances (A accreditation).
   (c) Specificity: Measures the model's ability to correctly identify negative instances (non-A accreditation).
   (d) AUC (Area Under the Curve): Indicates the model's ability to distinguish between the classes, with a higher AUC signifying better performance.

5. Model selection: The model with the best performance (lowest classification error and highest AUC) is selected for final predictions.
6. Feature importance analysis: Feature importance analysis is conducted to identify which variables have the greatest influence on the model's predictions, helping to guide potential educational policy decisions.

3. RESULT AND ANALYSIS

The dataset consisted of 23,954 junior high schools from 34 provinces in Indonesia and 86 variables. Before conducting the analysis, the data underwent a crucial preprocessing step: data cleaning and partitioning into training and test sets to ensure robust model evaluation.

3.1. Data Preparation

The data used in this analysis were derived from Junior High School data across Indonesia, adapted from the DAPODIK (Education Data and Information) of the Ministry of Education and Culture. These raw data require an initial preparation step to ensure high quality. The quality of the input data significantly affects the classification results. High-quality analysis results are crucial for achieving efficiency and effectiveness in the workplace. One of the processes in data preparation is data cleaning, which involves several tasks, such as reducing data dimensions, identifying and handling missing values, detecting and addressing outliers, and correcting inconsistencies.

The first step in the data preparation process was to reduce the data's dimensionality. The Junior High School accreditation data include variables that are not suitable for analysis and should be excluded. This data reduction yielded 47 variables, including one target variable and 46 explanatory variables. The excluded variables included school ID, district/city name, type of education, number of classes for grades 1-6, number of classrooms in good, lightly damaged, heavily damaged, and totally damaged conditions, number of students by age groups, and number of teachers, both permanent civil servants and non-permanent staff. The number of classes and students at the elementary school level is irrelevant to the Junior High School accreditation assessment; therefore, these variables are excluded as predictor variables. Additionally, variables related to the number of permanent teachers and other personnel were excluded to avoid bias in the classification process, as private schools tend to have non-permanent staff.

Furthermore, the number of classrooms in different conditions (good, lightly damaged, etc.) was excluded, as inconsistencies were observed in these data. For instance, the total number of classrooms in the dataset did not match the sum of the categories for damaged classrooms, and many of these variables contained missing values.

3.2. Detection and Handling of Missing Data

Missing data is a serious issue that must be addressed appropriately. As shown in Figure 1, each variable in the dataset had varying amounts of missing data. For variables with 0% missing data, some contained artificially zero values owing to data entry errors. For example, the number of students graduating and the number of classes for grades 7-9 may be zero, which are treated as missing data.

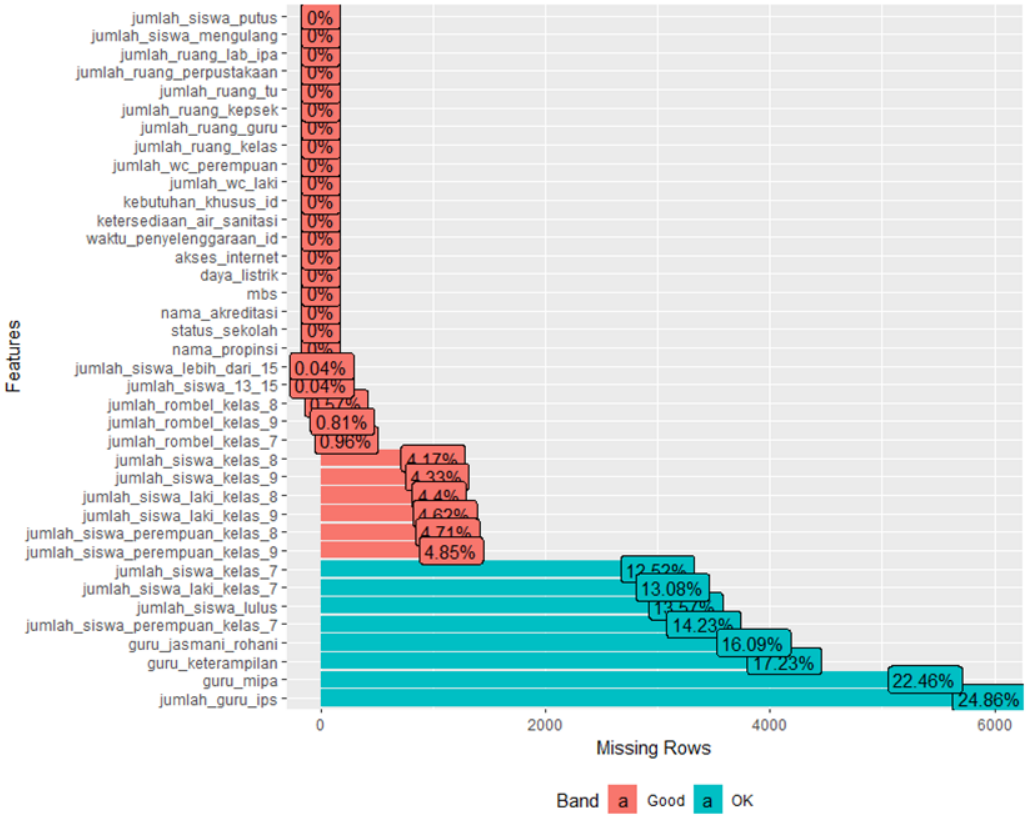

Figure 1. Number of missing values in each variable

Imputation was used for some variables with minor missing data. However, before imputation, some missing values were handled using related variables. For example, missing values in the ”number of classrooms” variable are addressed by using the ”number of classes” variable. Based on the accreditation instrument, the minimum number of classrooms was equal to the total number of classes across all grades. Therefore, for missing values in the ”number of classrooms” variable, the total class count was used to estimate them. When missing data occurred across all grade levels (7, 8, and 9), missing values were estimated by assuming a minimum of 1 class per grade level. Conversely, variables with a substantial proportion of missing data, particularly those exceeding 50% across the dataset, necessitated a different approach, often leading to their exclusion from the analysis to prevent significant bias or a reduction in model reliability.

Serious missing data issues arise in the ”number of teachers” variable, prompting the creation of new variables that combine relevant data. New variables, such as ”`guru_mipa`” (teachers in science subjects) and ”`guru keterampilan`” (teachers in skill subjects), were created by summing the relevant teacher counts. After these steps, the missing data were no longer present.

3.3. Pre-processing

In this step, discretization was performed on several variables. This process is based on the accreditation instrument, which evaluates spaces such as the teacher’s room, the principal’s room, and other facilities based on their size and equipment, not the number of rooms. Discretization also helps in handling outliers. The data were split into training and testing sets using the createDataPartition function in R, and discretization was performed on the training data. The Minimum Description Length Principle (MDLP) was used for discretization. This method aims to determine the optimal number of bins for each continuous variable by minimizing the total description length of the data, thereby preserving essential information while reducing complexity.

Discretization is then applied to the testing data using the same cut points derived from the training data. After discretization, descriptive statistics for the variables related to the number of classrooms were calculated. This ensures that both datasets are processed consistently, maintaining the integrity of the data distribution for subsequent classification modeling.

3.4. Handling Imbalance Data

The dataset suffers from an imbalance, with most schools having non-A accreditation, as shown in Figure 2. This imbalance can lead to skewed analyses and biased outcomes, as classification models may favor the majority class. To address this, we applied undersampling to balance the dataset, ensuring both accreditation categories (A and non-A) had equal numbers of observations (5,628 each). Undersampling was chosen over other techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), for several reasons. First, the dataset was large enough that undersampling could effectively reduce the majority class without significant data loss. In contrast, SMOTE might have introduced synthetic data that could distort the true data distribution. Additionally, undersampling prevents overfitting, ensures computational efficiency, and maintains the natural distribution of features within each class. This method ensures that both classes are represented equally, resulting in a more accurate, balanced classification model. After balancing the data, we applied normalization and feature scaling to ensure that all features contributed equally to the model's learning process [23].

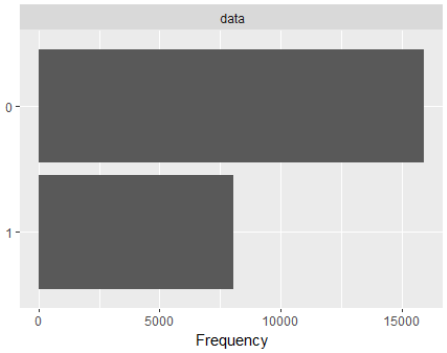

Figure 2. Bar chart of accreditation status

After completing these steps, the dataset was prepared for further analysis, with cleaner and more balanced data ready for modeling.

3.5. Data Exploration

Following the data preprocessing procedures outlined earlier, 37 variables were retained for subsequent analysis, consisting of one target variable and 36 predictors. The spatial distribution of junior high schools across Indonesia is heavily skewed toward Java Island, as shown in Figure 3. The provinces of West Java, East Java, and Central Java have the largest concentrations of Junior High Schools, substantially outnumbering those in other regions. In contrast, eastern Indonesia features markedly fewer schools than its western counterparts, a pattern largely attributable to the nation's uneven population distribution, which is densest in Java and Sumatra.

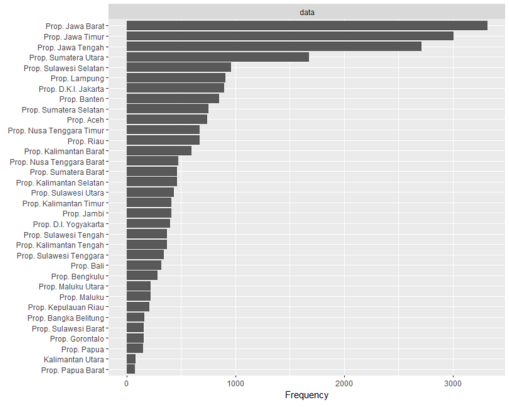

Figure 3. Junior High School Distribution by Province in Indonesia

Public (state-owned) schools attaining A accreditation are far more prevalent than private counterparts. Conversely, the proportions of public and private schools holding non-A accreditation are roughly equivalent, implying that institutional ownership may exert a discernible effect on accreditation. Table 1 presents summary statistics for the discretized predictors, including male and female toilet counts, total classrooms, and class sizes per grade level. This discretization was calibrated to accreditation rubrics that prioritize facility sufficiency over raw volumetric measures.

Table 1. Descriptive statistics of discrete variables

Variables	Min	Q1	Median	Q3	Max
Number of male toilets	0	1	1	1	34
Number of female toilets	0	0	1	1	31
Number of classrooms	1	5	9	15	113
Number of learning groups in Grade 7	1	2	3	5	32
Number of learning groups in Grade 8	1	2	3	5	35
Number of learning groups in Grade 9	1	1	3	5	28
Number of students repeating a grade	0	0	0	0	11
Number of students dropping out	0	0	0	1	177
Number of students graduating	1	29	62	135	794
Number of students aged 13-15	0	81	174	373	1942
Number of students aged >15	0	22	45	83	896
Number of students in Grade 7	1	36	75	149	707
Number of students in Grade 8	1	39	80	160	1057
Number of students in Grade 9	1	35	70	141	759
Number of male students in Grade 7	1	19	39	77	446
Number of male students in Grade 8	1	20	41	81	503
Number of male students in Grade 9	1	18	36	72	388
Number of female students in Grade 7	1	16	35	72	374
Number of female students in Grade 8	1	18	38	79	554
Number of female students in Grade 9	1	16	34	71	403
Number of Social Studies teachers	1	1	2	3	15
Number of Science teachers	1	1	2	3	20
Number of Skills teachers	1	2	3	5	30
Number of Physical Education teachers	1	2	4	6	31

3.6. Best Classification Methods

For the classification modeling of junior high school accreditation data in Indonesia, three classification methods were employed: Random Forest, Boosting, and Support Vector Machine (SVM). The Receiver Operating Characteristic (ROC) curve was used to evaluate the sensitivity and specificity of these methods. A higher area under the curve (AUC) indicates a better classification performance. These methodologies are widely recognized for their robustness in handling complex, high-dimensional datasets and their capacity to mitigate issues such as overfitting, which are common in educational classification problems [24, 25]. The ROC curve is shown in Figure 4.

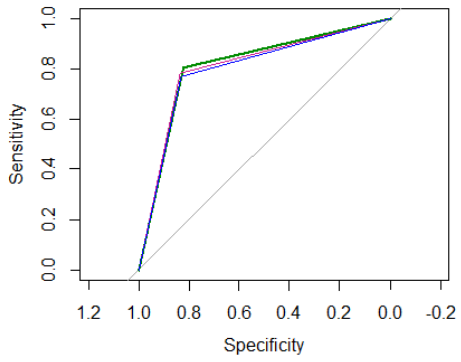

Figure 4. ROC curve of Random Forest, Boosting, and SVM

Based on the ROC curve in Figure 4, the Random Forest method had the highest AUC of 0.8133 (green line), followed by Boosting (0.8073, red line) and SVM (0.7998, blue line). These results suggest that Random Forest is likely the best method for classifying Junior School accreditation in Indonesia. This outcome aligns with previous research highlighting the superior performance of Random Forest in various classification tasks, including land cover classification and educational assessment, owing to its ability to effectively process a wide range of input features [7, 11, 26]. While Random Forest demonstrated robust performance, the slight differences in AUC values among the methods suggest that the optimal algorithm can be case-specific, aligning with findings from other classification studies [11]. The accuracy, sensitivity, and specificity of each method are presented in Table 2.

Table 2. Model Accuracy, Sensitivity, and Specificity of Random Forest, Boosting, and SVM

Classification Method	Accuracy	Sensitivity	Specificity
Random Forest	0.8161959	0.8935542	0.6936354
Boosting	0.8170307	0.8827677	0.7044168
SVM	0.8102129	0.8776185	0.694972

3.7. Cross-Validation Results

To ensure consistent results, cross-validation was used to identify the method with the smallest classification error. The average classification errors from 5-fold cross-validation were analyzed, revealing that the Random Forest consistently exhibited the lowest error rate. This reinforces its suitability for predicting junior high school accreditation ranks, echoing similar findings in which Random Forest models demonstrate superior accuracy in educational and land classification contexts [7, 27]. The results of the cross-validation are shown in Figure 5, and the exact values are listed in Table 3.

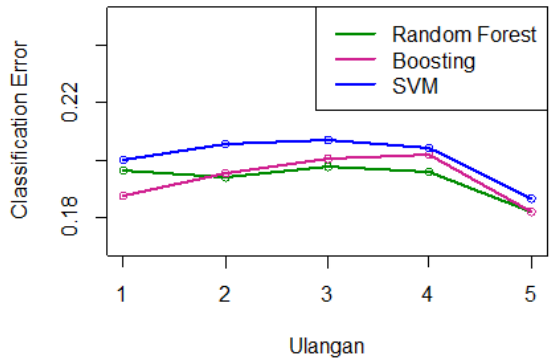

Figure 5. Classification Error Plot with 5-Fold Cross-validation

Table 3. Average Classification Error of 5-Fold Cross-validation

Classification Method	Classification Error (CE)
Random forest	0.1932469
Boosting	0.1934961
SVM	0.2007227

Among the three methods, Random Forest produced the lowest average classification error of 19.32%, making it the best method for predicting Junior School accreditation ratings. This outcome is further supported by observations in other studies, where Random Forest consistently outperformed SVM and Boosting techniques, particularly when resampling methods were applied to address data imbalances [2]. Furthermore, the ability of Random Forest to handle high-dimensional data and model complex interactions between variables without extensive parameter tuning contributes to its robust performance [28].

3.8. Feature Importance Analysis

The feature importance analysis of the Random Forest model, conducted using the MeanDecreaseGini method, reveals that the number of students aged 13-15 (”`jumlah_siswa_13_15`”) is the most influential feature, indicating that the student population size plays a critical role in determining school accreditation. The province name (”`nama_propinsi`”) ranks second,

closely behind the first factor, suggesting that regional differences in educational quality, infrastructure, and resources significantly affect accreditation outcomes. Other important features include the number of school staff ("`jumlah_staff`"), school facilities ("`jumlah_ruang_perpustakaan`"), and the percentage of female students ("`persentase_siswa_perempuan`"), which also contribute meaningfully to predicting accreditation. School-based management ("`mbs`"), however, was found to be the least important among the features, indicating that, despite its relevance to school governance, it has minimal impact on accreditation outcomes compared to demographic and regional factors. These results, shown in Figure 6, suggest that policy interventions focusing on increasing student enrollment, addressing regional educational disparities, and improving school facilities and staff may have a more direct impact on school accreditation.

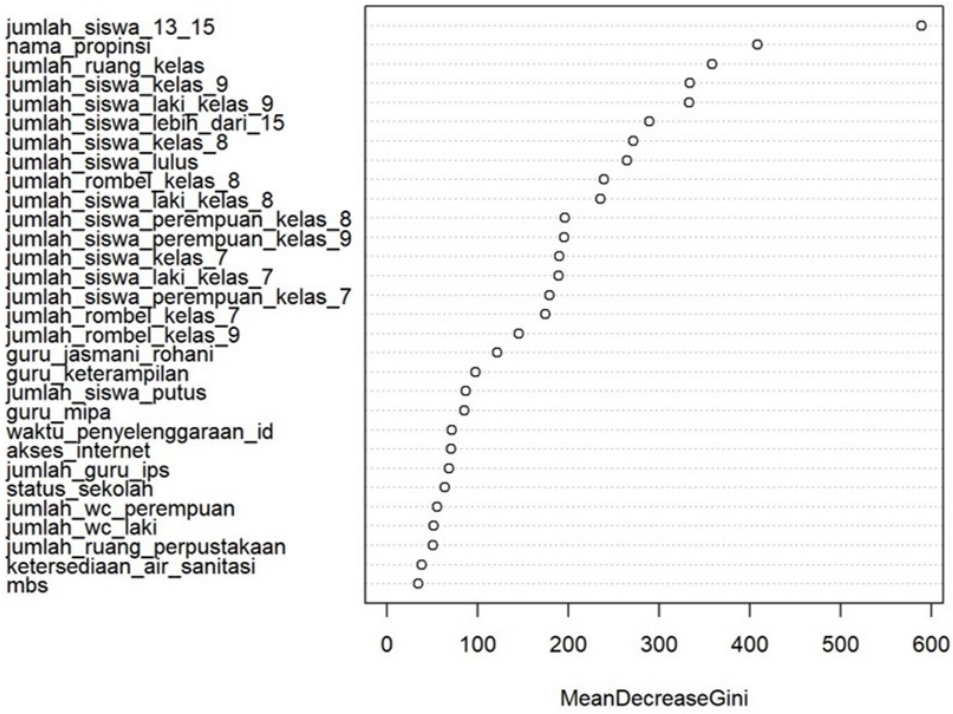

Figure 6. Feature Importance Analysis of Random Forest

The findings of this research are consistent with previous studies that have applied machine learning techniques to classify school accreditation data. In this study, Random Forest outperformed Boosting and SVM, achieving the highest accuracy and AUC, which aligns with findings from [2] and [5]. Both studies highlighted Random Forest's strength in handling complex educational data. Additionally, Random Forest has been shown to excel across various applications, including land cover assessment and forest biomass modeling, further supporting its use in classification tasks. These results reinforce the suitability of Random Forest for predicting school accreditation ranks, as also demonstrated in prior research.

4. CONCLUSION

This study aimed to evaluate the effectiveness of three classification methods–Random Forest, Boosting, and Support Vector Machine (SVM)–in predicting the accreditation status of Junior High Schools in Indonesia. The results indicated that Random Forest was the most effective classification method, achieving the highest Area Under the Curve (AUC) of 0.8133, followed by Boosting (AUC = 0.8073) and SVM (AUC = 0.7998). These findings suggest that Random Forest is particularly well-suited for classifying school accreditation ranks owing to its ability to handle high-dimensional data, capture non-linear relationships, and reduce overfitting through ensemble learning. Additionally, the Random Forest method had the lowest average classification error (19.32%) among Boosting and SVM, further reinforcing its superiority in this context. The high accuracy, sensitivity, and specificity values achieved by Random Forest indicate its strong performance in predicting school accreditation ranks, which is crucial for

improving the efficiency of the accreditation process in Indonesia.

This study also highlights the importance of data preprocessing, including handling missing values, reducing dimensionality, and addressing data imbalances, which significantly impacted the quality of the predictive model. By employing techniques such as imputation and undersampling, the dataset was cleaned and balanced, ensuring the models operated on consistent, complete data. In conclusion, machine learning models, specifically Random Forest, can provide a robust and efficient alternative to traditional manual methods of school accreditation assessment. This approach can help streamline the accreditation process, facilitate better resource allocation, and provide actionable insights for educational policymakers to improve school performance, especially in underserved regions.

ACKNOWLEDGEMENTS

The author would like to acknowledge that no external assistance was received during the research and writing of this article.

REFERENCES

[1] T. A. Y. Siswa and N. A. Verdikha, "Komparasi Algoritma Klasifikasi untuk Menentukan Evaluasi Kinerja Terbaik pada Status Akreditasi Sekolah/Madrasah Kalimantan Timur Berdasarkan LASP 2020," *Jurnal Informatika, Teknologi dan Sains*, vol. 4, no. 3, pp. 185–192, Aug. 2, 2022. DOI: 10.51401/jinteks.v4i3.1807

[2] D. I. Syarip, K. A. Notodiputro, and B. Sartono, "Comparison of Random Forest and Support Vector Machine Classification Methods for Predicting the Accuracy Level of Madrasah Data," *Media Statistika*, vol. 18, no. 1, pp. 37–48, Oct. 14, 2025. DOI: 10.14710/medstat.18.1.37-48

[3] Y. N. Sukmaningtyas, R. M. Akbar, and G. R. U. Asyafiiyah, "Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa," *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 2, pp. 127–145, Dec. 30, 2024. DOI: 10.29240/arcitech.v4i2.12048

[4] M. B. Musthafa et al., "Evaluation of university accreditation prediction system," *IOP Conference Series: Materials Science and Engineering*, vol. 732, no. 1, p. 012 041, Jan. 1, 2020. DOI: 10.1088/1757-899X/732/1/012041

[5] S. Wibowo, "Building a Classification Model to Predict School Quality in Indonesia:," presented at the International Conference on Educational Assessment and Policy (ICEAP 2020), Jakarta, Indonesia, 2021. DOI: 10.2991/assehr.k.210423.074

[6] A. Warjaya et al., "Kombinasi Latent Semantic Indexing dan Support Vector Machine pada Klasifikasi Dokumen Akreditasi: Studi Kasus: Pascasarjana Universitas Negeri Medan," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, pp. 6400–6407, May 25, 2025. DOI: 10.36040/jati.v9i4.14102

[7] M. I. Habibie et al., "Integrating Sentinel-2 and ESA world cover for effective land use and land cover assessment using machine learning," *Advances in Space Research*, vol. 76, no. 9, pp. 4925–4958, Nov. 2025. DOI: 10.1016/j.asr.2025.07.083

[8] V. Nasiri et al., "Modeling Forest Canopy Cover: A Synergistic Use of Sentinel-2, Aerial Photogrammetry Data, and Machine Learning," *Remote Sensing*, vol. 14, no. 6, p. 1453, Mar. 17, 2022. DOI: 10.3390/rs14061453

[9] M. Sipper and J. H. Moore, "Conservation machine learning: A case study of random forests," *Scientific Reports*, vol. 11, no. 1, p. 3629, Feb. 11, 2021. DOI: 10.1038/s41598-021-83247-4

[10] A. Fadila et al., "Classification of Dropout Rates in West Sumatra Using the Random Forest Algorithm with Synthetic Minority Oversampling Technique," *UNP Journal of Statistics and Data Science*, vol. 2, no. 3, pp. 279–286, Aug. 24, 2024. DOI: 10.24036/ujsds/vol2-iss3/183

[11] Y. Xi et al., "Mapping tree species in natural and planted forests using Sentinel-2 images," *Remote Sensing Letters*, vol. 13, no. 6, pp. 544–555, Jun. 3, 2022. DOI: 10.1080/2150704X.2022.2051636

[12] M. S. Başarslan and F. Bal, "The effect of text representation and model selection on classification performance: A comprehensive comparison of tf-idf, bow and transformer-based methods on the covid19-fnir dataset," *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 14, no. 4, pp. 1447–1461, Oct. 15, 2025. DOI: 10.28948/ngumuh.1694988

[13] R. Saini, "Integrating Vegetation Indices and Spectral Features for Vegetation Mapping from Multispectral Satellite Imagery Using AdaBoost and Random Forest Machine Learning Classifiers," *Geomatics and Environmental Engineering*, vol. 17, no. 1, pp. 57–74, Dec. 11, 2022. DOI: 10.7494/geom.2023.17.1.57

[14] K. Mohammed et al., "Integrating participatory GIS, remote sensing, and explainable machine learning to assess forest provisioning services," *Environmental Impact Assessment Review*, vol. 117, p. 108 245, Mar. 2026. DOI: 10.1016/j.eiar.2025.108245

[15] P. K. Rajput, "Machine learning approach for forest biomass modelling with in-situ and remote sensing data in Narmadapuram central India," *Modeling Earth Systems and Environment*, vol. 11, no. 5, p. 350, Oct. 2025. DOI: 10.1007/s40808-025-02527-4

[16] K. S. Bjerreskov, T. Nord-Larsen, and R. Fensholt, "Classification of Nemoral Forests with Fusion of Multi-Temporal Sentinel-1 and 2 Data," *Remote Sensing*, vol. 13, no. 5, p. 950, Mar. 3, 2021. DOI: 10.3390/rs13050950

[17] L. Yang et al., "Mapping above-ground biomass and canopy mean height in high mountainous forest areas with Sentinel-2 multi-spectral image based on machine learning algorithms," *International Journal of Digital Earth*, vol. 18, no. 2, p. 2 558 924, Dec. 31, 2025. DOI: 10.1080/17538947.2025.2558924

[18] M. Berriri et al., "Multi-Class Assessment Based on Random Forests," *Education Sciences*, vol. 11, no. 3, p. 92, Feb. 26, 2021. DOI: 10.3390/educsci11030092

[19] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 8, 2024. DOI: 10.58496/BJML/2024/007

[20] T. Chernenkova et al., "Spatiotemporal Modeling of Coniferous Forests Dynamics along the Southern Edge of Their Range in the Central Russian Plain," *Remote Sensing*, vol. 13, no. 10, p. 1886, May 11, 2021. DOI: 10.3390/rs13101886

[21] S. W. Akram et al., "Comparative Analysis Using Machine Learning Algorithms to Detect Parkinson's Disease using Voice Dataset," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 3, pp. 1933–1942, Mar. 31, 2024. DOI: 10.22214/ijraset.2024.59252

[22] Q. Fan et al., "Forest Carbon Storage Dynamics and Influencing Factors in Southeastern Tibet: GEE and Machine Learning Analysis," *Forests*, vol. 16, no. 5, p. 825, May 15, 2025. DOI: 10.3390/f16050825

[23] A. Indryani, U. Khaira, and M. F. Putri, "Rainfall Prediction Using Long Short-Term Memory Method (Case Study: Jambi City)," *Jurnal Pepadun*, vol. 6, no. 1, pp. 57–70, Apr. 15, 2025. DOI: 10.23960/pepadun.v6i1.256

[24] B. Alsubhi et al., "Effective Feature Prediction Models for Student Performance," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11 937–11 944, Oct. 13, 2023. DOI: 10.48084/etasr.6345

[25] O. Iparraguirre-Villanueva et al., "Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, 2023. DOI: 10.14569/IJACSA.2023.0140455

[26] A. B. I. Bernardo et al., "Contrasting Profiles of Low-Performing Mathematics Students in Public and Private Schools in the Philippines: Insights from Machine Learning," *Journal of Intelligence*, vol. 10, no. 3, p. 61, Aug. 30, 2022. DOI: 10.3390/jintelligence10030061

[27] M. Wang and S. Liu. "Machine Learning-Based Research on the Adaptability of Adolescents to Online Education." version 1, pre-published.

[28] K. Huang, I. Zimmerman, and D. Bein, "Study on the Use of Random Forest Classifier model and Multi Output Classifier model for Predicting Student Academic Performance and Identifying Area of Concern," in *2025 ASEE Annual Conference & Exposition Proceedings*, Montreal, Quebec, Canada: ASEE Conferences, Jun. 2025, p. 57 159. DOI: 10.18260/1-2--57159

[This page intentionally left blank.]