Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 18(1) 2025: 37-48
http://ejournal.undip.ac.id/index.php/media_statistika

COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE CLASSIFICATION METHODS FOR PREDICTING THE ACCURACY LEVEL OF MADRASAH DATA

Dodi Irawan Syarip[1, 2], Khairil Anwar Notodiputro[1], Bagus Sartono[1]
[1] Study Program in Statistics and Data Science - School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia
[2] Ministry of Religious Affairs, Republic of Indonesia

E-mail: *dodi.irawan.syarip@apps.ipb.ac.id*

DOI: 10.14710/medstat.18.1.37-48

Article Info:
Received: 17 May 2025
Accepted: 10 October 2025
Available Online: 14 October 2025

Keywords:
Accuracy, AUC, Random Forest, ROSE, SVM

Abstract: This study aims to identify the most effective classification method for predicting the accuracy level of madrasah data with class imbalance. Two machine learning approaches were employed: Random Forest (RF) and Support Vector Machine (SVM). Based on the AUC values, it was concluded that the RF model had a slightly better performance in predicting the accuracy level of the madrasah data, with an average AUC of 62.82, compared to the SVM model, which had an average AUC of 62.33. Among all models, the highest and consistent performance was achieved by the RF model using ROSE techniques. The results of measuring variable importance showed that the predictor variables with the greatest influence in predicting the accuracy level of the madrasah data are the number of students and the student-to-teacher and staff ratio. This finding suggests that school principals and madrasah administrative staff should prioritize ensuring the completeness of student, teacher, and staff data to improve the overall reliability of madrasah data.

1. INTRODUCTION

The rapid advancement of information technology has driven public organizations, including government agencies, to adopt digital systems that enhance the accuracy, efficiency, and reliability of data management. High-quality data is essential for evidence-based decision-making, strategic planning, and effective implementation of public programs, as unreliable data risks undermining policy outcomes (Batini et al., 2009).

The Directorate General of Islamic Education at the Ministry of Religious Affairs (MORA) has developed the Education Management Information System (EMIS) as the backbone for managing Islamic education data nationwide, enabling systematic collection, storage, analysis, and dissemination (Syarip & Rosidin, 2003). To ensure madrasah data quality, MORA initiated an annual data accuracy audit survey in 2020, with the 2023 survey of 1,210 madrasahs reporting a 76% national data accuracy rate. This relatively good achievement indicates that EMIS data are generally reliable for policymaking, yet further improvements are required to enhance accuracy and strengthen its role in supporting evidence-based decision-making. Despite its importance, empirical studies on measuring

and predicting data accuracy in government education information systems, particularly in Islamic education, remain limited, highlighting a significant research gap.

This study aims to model and predict the accuracy level of madrasah data using two machine learning–based classification methods: Random Forest (RF) and Support Vector Machine (SVM). These methods have been proven effective in handling complex classification problems and improving predictive performance compared to traditional statistical approaches (Breiman, 2001; Cortes & Vapnik, 1995). Several studies have compared RF and SVM in classification tasks across different domains. For example, Caruana & Niculescu-Mizil (2006) found that both RF and SVM consistently achieved strong predictive accuracy in large-scale empirical evaluations, with RF often performing better in handling noisy datasets, while SVM demonstrated superior performance in high-dimensional spaces. Similarly, Fernández-Delgado et al. (2014) conducted a comprehensive benchmark analysis involving 179 classifiers across 121 datasets and reported that both RF and SVM ranked among the most accurate classifiers, though RF generally showed higher robustness and lower variance across diverse data conditions.

The novelty of this research lies in integrating data quality auditing with advanced classification models to generate actionable insights for improving the reliability of madrasah data. By bridging the gap between data quality assessment and machine learning applications, this study contributes theoretically to the literature on public sector data governance and practically to policy recommendations for strengthening data management in Islamic education in Indonesia.

2. LITERATURE REVIEW

2.1. Random Forest

Random Forest is a technique employed for both classification and regression. It operates as an ensemble of learning methods, utilizing decision trees as fundamental classifiers, which are constructed and amalgamated (Kulkarni & Sinha, 2014). In simple terms, the algorithm of Random Forest can be explained as follows: if the training dataset has n magnitude size and consists of p predictor variables, next come the steps of preparation and estimation using the Random Forest method (Breiman & Cutler, 2003):

1) Bootstrap stage: performs random sampling with n-size recovery from the training dataset.
2) Random sub-setting stage: a tree is constructed using the bootstrap dataset, extending to its fullest extent (without pruning). At every node, the sorter (*mtry*) is chosen by randomly selecting up to m predictor variables, where $m < p$.
3) The process is reiterated k times, duplicating steps (1) and (2), resulting in a forest composed of k random trees.

In Random Forest, the number of predictor variables randomly selected at each node, denoted as m (or *mtry*), is commonly determined using practical rules of thumb, where m is set to the square root of the total number of predictors (p) for classification tasks, and approximately one-third of p for regression tasks (Liaw & Wiener, 2002). By aggregating the predictions of k trees, Random Forest estimates an observation's response, while its accuracy is evaluated through out-of-bag (OOB) error. OOB data, excluded from bootstrap samples, provide an unbiased estimate of misclassification and are also used to assess variable importance (Breiman, 2001). Each observation is predicted by about one-third of the trees, and the OOB error is calculated as the proportion of misclassified predictions across all observations in the dataset.

Breiman & Cutler (2003) suggest observing OOB errors when the number of trees (k) is small and selecting the number of predictors (m) that minimizes OOB error. For generating variable importance, it is advisable to use many trees—typically 1000 or more—especially when many predictors are analyzed, to ensure more stable results.

2.2. Support Vector Machine

Support Vector Machine (SVM) stands out among supervised learning algorithms, commonly employed for both regression and classification (Awad & Khanna, 2015; Cortes & Vapnik, 1995). In classification modeling, SVM offers a well-established and conceptually clear framework compared to other methods. The objective of the SVM algorithm is to establish an optimal decision line or boundary, termed the hyperplane, that effectively partitions n-dimensional spaces into distinct classes. This hyperplane facilitates the accurate categorization of new data points in the future. The SVM algorithm identifies extreme points or vectors, referred to as support vectors, that are crucial for the hyperplane construction. Therefore, it is referred to as a Support Vector Machine (Noble, 2006).

In real-world applications, many problems are inherently non-linear; therefore, kernel functions are applied to address the issue of nonlinearity. According to Hsu et al. (2003), four basic kernel functions are commonly used, namely:

1) Linear: $K(x_i, x_j) = x_i^T x_j$
2) Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + R)^d,\ \gamma > 0$
3) Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2),\ \gamma > 0$
4) Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

where γ,r, and d are kernel parameters.

2.3. Classification Accuracy

The performance of a classification algorithm is commonly evaluated using a confusion matrix, which summarizes the number of correctly and incorrectly predicted observations (Gorunescu, 2011). The matrix distinguishes four outcomes: True Positive (TP) when a positive case is correctly classified, True Negative (TN) when a negative case is correctly classified, False Positive (FP) when a negative case is incorrectly classified as positive, and False Negative (FN) when a positive case is misclassified as negative. Higher TP and TN values generally correspond to better accuracy, precision, and recall. Based on these components, various performance metrics can be derived, including accuracy, sensitivity/recall, specificity, precision, and F1-Score. The formulas for these measures are expressed as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (1)$$

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (4)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (5)$$

2.4. Area Under Curve (AUC)

The Receiver Operating Characteristic (ROC) curve is one of the most used tools for evaluating classification systems because it can evaluate algorithms very well. The threshold for sensitivity and specificity determines the construction of the ROC curve and the derivation of the Area Under Curve (AUC) value (Kuhn, 2008). The use of the AUC as a performance metric is essential, particularly when evaluating classification models on imbalanced datasets. Unlike accuracy, which can be misleading in the presence of class imbalance, AUC provides a threshold-independent measure of the model's overall discriminative ability between positive and negative classes (Kuhn & Johnson, 2013).

In addition to AUC, the Matthews Correlation Coefficient (MCC) is often used for imbalanced datasets as it incorporates all elements of the confusion matrix into a single value (Chicco & Jurman, 2020). However, MCC is threshold-dependent, while AUC evaluates performance across all thresholds, offering a more comprehensive measure of classification power. Given its broad adoption in machine learning, this study employs AUC as the primary evaluation metric, with its formula expressed as follows:

$$\text{AUC} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \times 100\% \tag{6}$$

2.5. Class Imbalance

Class imbalance is one of the problems that often occurs in data mining. This happens when there is a significant difference in size or frequency between the minority and majority classes (Ren et al., 2017). Models built with imbalanced data will have poor accuracy for minority predictions. There are biased decision boundaries in the categorization system because the dominant class has richer knowledge than the minority class (Jian et al., 2016). To overcome class imbalance, resampling techniques are applied at the preprocessing stage to rebalance the data distribution, thereby reducing bias in the learning process. These techniques include undersampling, which reduces the majority class; oversampling, which replicates the minority class; and hybrid approaches that combine both (Jian et al., 2016). Among various oversampling methods, Random Over-Sampling Examples (ROSE) is widely used to effectively mitigate imbalance and improve model performance (He et al., 2018).

2.6. Variable Importance

One commonly used approach to assessing variable importance is by evaluating the frequency with which a variable is used for splitting in a group of decision trees. The more frequently a variable is used and the greater its contribution to reducing impurity, the more important it is considered. A widely recognized method for quantifying variable importance in the Random Forest is the Mean Decrease Gini (MDG), as proposed by Breiman (2001). The MDG index measures the importance of the predictor variable x_h, suppose there are p predictor variables with $h = (1, 2, \dots, p)$. This measure reflects the average decrease in Gini impurity attributed to splits using x_h across all trees in the forest.

$$\text{MDG}(x_h) = \frac{1}{k}[1 - \sum_k \text{Gini}(h)^k] \tag{7}$$

where $\text{Gini}(h)^k$: Gini index for predictor variables x_h in the k^{th} tree; k: the number of trees in Random Forest

Unlike Random Forest, SVM does not inherently provide measures of variable importance, particularly when using nonlinear kernels such as the Radial Basis Function

(RBF). Therefore, external techniques are required to assess which features contribute most to the model's predictive performance. A widely adopted method is Permutation Feature Importance (PFI), which is model-agnostic and applicable to SVMs. The principle is simple: by permuting the values of a given feature and observing the drop in model performance, one can quantify the importance of that feature (Fisher et al., 2019). Features causing substantial performance degradation when shuffled are considered highly relevant, while those with negligible or negative effects may be irrelevant or noisy. In practice, PFI results provide ranked importance scores with uncertainty intervals, allowing researchers to better understand the role of each predictor in classification tasks.

3. MATERIAL AND METHOD

3.1. Source of Data

The data used in this study consisted of one response variable and nine predictor variables, as detailed in Table 1.

Table 1. List of Variables in This Study

No.	Label	Variable Name	Information		Data Type
1	Y	Data Accuracy Level	1: Good	2: Less	Ordinal
2	X1	Madrasah Level	1: RA	3: MTs	Nominal
			2: MI	4: MA	
3	X2	Accreditation Status	1: Not Accredited	3: B	Ordinal
			2: C	4: A	
4	X3	Operator Education Qualification	1: ≤ High School	3: S1	Ordinal
			2: Diploma	4: > S1	
5	X4	Headmaster's Account Authority	1: Handed over to the Operator		Nominal
			2: Held by Headmaster Himself		
6	X5	Number of Students	Number of students in each madrasah		Numeric
7	X6	Ratio of Students to Teachers and Staff	Number of students divided by the number of teachers and staff		Ratio
8	X7	Gender of Operator	1: Male	2: Female	Nominal
9	X8	Gender of Headmaster	1: Male	2: Female	Nominal
10	X9	Madrasah Location	1: Lowland	3: Mountains	Nominal
			2: Coastline		

The data for this study were obtained from the 2023 Islamic Education Data Accuracy Audit survey conducted by the Directorate General of Islamic Education, Ministry of Religious Affairs, through observations and direct interviews with school principals and data operators. The survey covered 1,210 madrasahs at the level of MA (Madrasah Aliyah), MTs (Madrasah Tsanawiyah), MI (Madrasah Ibtidaiyah), and RA (Raudhatul Athfal) in 102 regencies/cities. Given the large, geographically dispersed, and hierarchically structured population of 86,343 madrasahs across 34 provinces and 514 districts/cities in Indonesia as of 2023, a multistage random sampling technique was employed. In the first stage, a random selection of districts/cities was conducted, followed by the random selection of madrasahs within the selected areas. This sampling method was chosen to ensure adequate representation of the target population while optimizing the use of time, financial resources, and logistical capacity. Such an approach is particularly appropriate for educational surveys involving clustered populations, where full population enumeration is impractical (Babbie, 2020). Based on the survey results, the percentage of data accuracy (PDA) for each sample madrasah was obtained, which was calculated using the formula:

$$\text{PDA} = \frac{\text{the number of correct EMIS data attributes according to survey results}}{\text{the number of all surveyed data attributes}} \times 100\% \quad (7)$$

Furthermore, the PDA obtained is converted into a Data Accuracy Level (DAL) category consisting of two classifications: “Less” for values below 71% and “Good“ for values equal to or above 71%.

3.2. Research Methodology

As mentioned above, this study applied two classification methods: Random Forest (RF) and Support Vector Machine (SVM), to predict the accuracy level of madrasah data. These two ensemble methods were evaluated and compared to determine which method showed better performance in classifying and predicting the accuracy level of madrasah data.

The methodology in this study consists of several steps (see Figure 1), including: (1) data collection; (2) pre-processing of data, consisting of: (a) check for the condition of class imbalance and perform resampling; (b) for each sampling method, split the observation data into two (training and testing); (3) create the model with the original training data and the resampled training data; (4) testing the trained RF and SVM models using the test dataset to generate predictions; (5) create confusion matrix and evaluation the model; (6) measurement of variable importance.

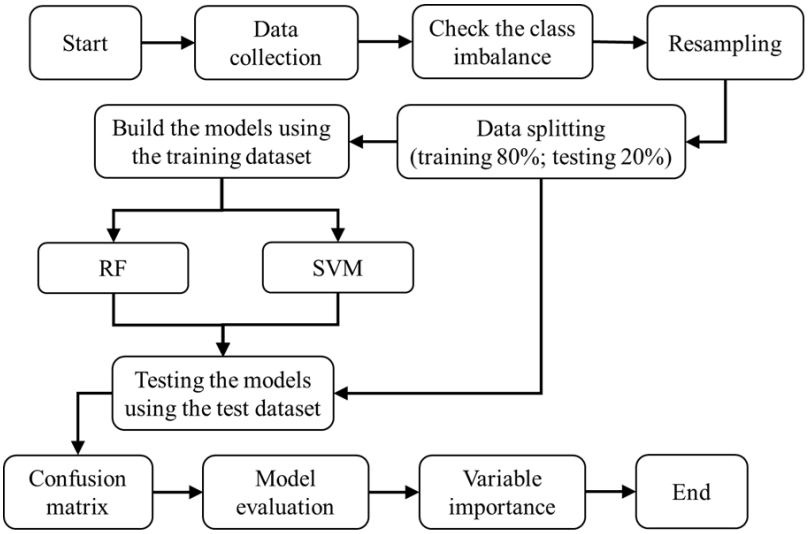

Figure 1. Research Methodology

The RF model was developed by partitioning the observation data into two subsets, with 80% allocated for training and 20% for testing. Model tuning was performed by varying the number of trees (*ntree*) at 300, 500, and 1000, and the number of predictor variables used at each split (*mtry*) at 2, 3, and 6. The results across these parameter configurations were compared to identify the optimal model for predicting the accuracy level of madrasah data, with the smallest out-of-bag (OOB) error serving as the selection criterion.

Similarly, the SVM model was trained using an 80/20 split of the data, with the Radial Basis Function (RBF) kernel employed to capture nonlinear relationships. Two key hyperparameters were considered: cost (C), which regulates the trade-off between margin width and classification error, and sigma (σ), which controls the influence of each data point in shaping the decision boundary. The tuning process of the SVM model was carried out by adjusting the cost (C) parameter at values of 0.25, 0.50, and 1.00.

To address the issue of class imbalance, three resampling methods were applied: undersampling, oversampling, and ROSE. Undersampling reduces the size of the majority class through random selection, thereby balancing it with the minority class. Oversampling increases the representation of the minority class by randomly replicating its instances until class proportions are balanced. The ROSE method generates synthetic observations using a smoothed bootstrap approach, creating a more representative and balanced training dataset.

4. RESULTS AND DISCUSSION

4.1. Overview of Observation Data

The research dataset comprised 1,210 observations with an imbalanced response variable. The "Good" class has a percentage of 80.08% or 969 data, much higher than the "Less" class with a percentage of 19.92% or 241 data. Furthermore, the 1,210 observational data were divided into two data groups with a ratio of 80% or 968 training data, and 20% or 242 testing data. Of the 968 training data, it consisted of 766 or 79.13% of the "Good" class data and 202 or 20.87% of the "Less" class data. A comparison between observational data and training dataset based on class is presented in Figure 2.

Figure 2. Classes Comparison of Response Variable (Data Accuracy Level)

4.2. Resampling

The issue of class imbalance in observational data is addressed through resampling, aiming to procure a training dataset with a more equitable class distribution. A comparison between the data volume of the majority and minority classes in both the original training data and the resampled outcomes is presented in Table 2.

Table 2. Comparison of Training Data for each Sampling Method

Sampling Methods	Amount of Data		Ratio	
	Good	Less	Good	Less
Original	766	202	79.13%	20.87%
Undersampling	184	202	47.67%	52.33%
Oversampling	766	796	49.04%	50.96%
ROSE	480	488	49.59%	50.41%

The comparison of the original training data shows a very unbalanced ratio between the majority class ("Good") of 79.13% and the minority class ("Less") of 20.87%. Meanwhile, training data that has been resampled using the under-sampling method shows a relatively balanced ratio, with the "Good" class of 47.67% and the "Less" class of 52.33%. Similarly, for the oversampling method, the "Good" class has a ratio of 49.04% and the "Less" class of 50.96%. As for the ROSE method, the ratio of the "Good" class is 49.59%, and the "Less" class is 50.41%.

4.3. Random Forest Classification Analysis

To obtain the optimal Random Forest model with the lowest OOB error, the algorithm was trained repeatedly using combinations of ntree values (300, 500, and 1000) and mtry values (2, 3, and 6). The smallest OOB error, 21.19%, was achieved with ntree = 1000 and mtry = 2. Table 3 reports the performance of this optimal model, evaluated on the original training data as well as three resampled datasets (undersampling, oversampling, and ROSE). The evaluation includes confusion matrices and performance metrics—accuracy, precision, recall, F1-Score, and AUC—providing a comprehensive assessment of the model's ability to classify the accuracy level of madrasah data.

The model trained on the original dataset achieved the highest accuracy (82.64) and F1-Score (90.37), showing strong performance in identifying the majority class ("Good") with balanced precision (84.55) and recall (97.04). However, its low AUC (52.37) indicates limited ability to distinguish the minority class ("Less"). In contrast, undersampling resulted in the weakest performance, with accuracy dropping to 59.92 and recall to 58.13. Although its AUC improved to 63.68, the overall predictive power declined due to loss of information from reducing the majority class. This finding supports previous research highlighting the drawbacks of undersampling and its risk of information loss (Haibo He & Garcia, 2009).

Table 3. Confusion Matrix of Random Forest Model

Sampling Methods	Prediction	Actual		Value				
		Good	Less	Accuracy	Precision	Recall	F1-Score	AUC
Original	Good	197	36	82.64	84.55	97.04	90.37	52.37
	Less	6	3					
Undersampling	Good	18	12	59.92	90.77	58.13	70.87	63.68
	Less	85	27					
Oversampling	Good	147	15	70.66	90.74	72.41	80.55	66.98
	Less	56	24					
ROSE	Good	147	14	71.07	91.30	72.41	80.77	68.26
	Less	56	25					

The application of oversampling produced more balanced classification outcomes, raising the AUC to 66.98, with a precision of 90.74 and an accuracy of 70.66. This demonstrates its effectiveness in improving the detection of minority class instances while maintaining predictive performance for the majority class. Similarly, the ROSE method achieved the highest AUC (68.26) and precision (91.30), with an accuracy of 71.07, showing comparable overall performance. By generating synthetic minority class examples, ROSE effectively mitigates class imbalance and enhances the model's ability to distinguish between classes (Lunardon et al., 2014).

Overall, the findings demonstrate that the choice of sampling method greatly affects the performance of Random Forest on imbalanced data. While the original dataset yielded the highest accuracy and recall, its low AUC reflects poor performance to identify the minority class. In contrast, the ROSE technique provides the best trade-off and most consistent balance across evaluation metrics, making it the most effective resampling approach for addressing class imbalance in predicting the accuracy level of madrasah data.

4.4. Support Vector Machine Classification Analysis

Hyperparameter tuning for the SVM model with a radial basis kernel identified the optimal configuration at sigma (σ) = 0.0449 and cost (C) = 0.25, achieving the highest accuracy (79.6%) compared to other C values (0.50 and 1.00). Table 4 shows the model's

performance (σ = 0.0449, C = 0.25) across different sampling methods: original data, undersampling, oversampling, and ROSE. The model trained on the original data yielded the highest accuracy (83.88), recall (100.00), and F1-score (91.24), indicating perfect classification of the majority class ("Good"). However, its AUC score was only 50.00, suggesting poor discrimination between majority and minority classes and a bias toward the majority class, thus limiting its effectiveness on imbalanced data.

Resampling techniques enhanced the model's ability to detect minority classes. Oversampling produced the most balanced performance, achieving the highest AUC (67.08), precision (91.33), and accuracy (67.36), indicating improved discriminative power and stability of SVM. ROSE delivered comparable results with slightly lower AUC (65.85), accuracy (65.29), and precision (91.03). Undersampling increased AUC (66.39) but reduced accuracy (64.46) and F1-Score (75.00) due to the loss of data volume. Overall, oversampling proved the most effective approach for improving SVM performance on imbalanced madrasah data, ensuring fairer classification between majority and minority classes.

Table 4. Confusion Matrix of SVM Model

Sampling Methods	Prediction	Actual		Value				
		Good	Less	Accuracy	Precision	Recall	F1-Score	AUC
Original	Good	203	39	83.88	83.88	100.00	91.24	50.00
	Less	-	-					
Undersampling	Good	129	12	64.46	91.49	63.35	75.00	66.39
	Less	74	27					
Oversampling	Good	137	13	67.36	91.33	67.49	77.62	67.08
	Less	66	26					
ROSE	Good	132	13	6.29	91.03	65.02	75.86	65.85
	Less	71	26					

4.5. Model Evaluation

The comparison between Random Forest (RF) and Support Vector Machine (SVM) reveals notable differences in handling class imbalance in madrasah data. Trained on the original dataset, both models achieved high accuracy (RF = 82.64, SVM = 83.88) and recall (RF = 97.04, SVM = 100.00), demonstrating strong performance for the majority class ("Good"). However, their very low AUC values (RF = 52.37, SVM = 50.00) highlight poor discrimination of the minority class ("Less"), confirming that models without resampling are biased toward the dominant class and fail to generalize effectively.

When resampling techniques were applied, both models exhibited improvements in their ability to recognize minority cases. For Random Forest, the ROSE method produced the most consistent performance, with the highest AUC (68.26), the highest accuracy (71.07), and a balanced recall (72.41). For SVM, oversampling yielded the best results, achieving the highest AUC (67.08) along with the highest accuracy (67.36) and an acceptable recall (67.49). While both algorithms benefited from resampling, Random Forest consistently demonstrated slightly better balance across recall, accuracy, and AUC compared to SVM.

These results align with previous studies conducted by Caruana & Niculescu-Mizil (2006) and Fernández-Delgado et al. (2014), which show the superiority of the Random Forest over SVM. In the context of madrasah data, Random Forest with ROSE technique provides the most reliable performance by addressing class imbalance and improving the representation of minority data.

4.6. Variable Importance

Variable importance was calculated for the model that was considered the best classification model in this study: Random Forest model with the ROSE sampling and SVM with oversampling technique. The importance of predictor variables in the Random Forest model is measured using the Mean Decrease Gini (MDG). The greater the MDG value of a variable, the more important the variable is in the model. Table 5 displays the MDG value for each predictor variable in the Random Forest model using the ROSE.

Table 5 shows that the number of students (X5) and the student-to-teacher and staff ratio (X6) are the strongest predictors of madrasah performance in data management, with MDG indices of 79.44366 and 74.27680, respectively. These results highlight the need for accurate data on student enrollment and teacher allocation to support effective resource planning and prevent issues such as overcrowded classrooms or underutilized staff. The madrasah level (X1), with MDG index of 48.93449, further underscores the role of valid institutional data in strategic planning and quality assurance. Strengthening data reporting systems is thus essential for policymakers and madrasah administrators to design interventions that target the most critical performance factors.

Table 5. Variable Importance of the ROSE Random Forest Model

No.		Variables	MDG
1	X5	Number of Students	79.44366
2	X6	Ratio of Students to Teachers and Staff	74.27680
3	X1	Madrasah Level	48.93449
4	X2	Accreditation Status	29.43915
5	X7	Gender of Operator	23.31659
6	X3	Operator Education Qualification	20.62864
7	X9	Madrasah Location	20.10117
8	X8	Gender of Headmaster	13.02222
9	X4	Headmaster's Account Authority	12.02313

Table 6. Variable Importance of the Oversampling SVM Model

No	Variables	PFI (0.05)	PFI (Average)	PFI (0.95)
1	X5 Number of Students	0.03606	0.04667	0.05163
2	X6 Ratio of Students to Teachers and Staff	0.03737	0.04439	0.05437
3	X1 Madrasah Level	0.03445	0.03753	0.04749
4	X3 Operator Education Qualification	0.01975	0.02350	0.03006
5	X7 Gender Operator	0.01074	0.02121	0.02719
6	X4 Headmaster's Account Authority	0.01465	0.01958	0.02513
7	X8 Gender of Headmaster	0.00783	0.02108	0.01407
8	X9 Madrasah Location	0.00271	0.02242	0.01609
9	X2 Accreditation Status	-0.00013	0.00587	0.01217

Table 6 presents the variable importance values of the oversampling SVM model based on Permutation Feature Importance (PFI). The results show the relative contribution of each predictor to the model's performance. Among the variables, the number of students (X5), the student-to-teacher and staff ratio (X6), and madrasah level (X1) emerge as the most influential, with average PFI scores of 0.04667, 0.04439, and 0.03753, respectively. This suggests that school size, the proportional distribution of students relative to staff, and madrasah level substantially affect the SVM model's classification accuracy. Conversely, variables such as accreditation status (X2) and madrasah location (X9) record the lowest PFI values. Notably, accreditation status even shows a slightly negative value at the 0.05 confidence interval, indicating that its contribution may be negligible or potentially introduce noise into the predictive process.

The ranking of these features highlights the practical implication that quantitative and structural characteristics of schools (such as student numbers, ratios, and institutional levels) are more critical for accurate predictions than demographic or administrative attributes. This finding suggests that educational data analysis using SVM models may yield stronger predictive accuracy when emphasizing measurable institutional capacity indicators rather than identity-based variables.

5. CONCLUSION

This study demonstrates that the Random Forest model consistently outperforms the SVM model in predicting madrasah data accuracy, with the best performance achieved by Random Forest using the ROSE sampling. Variable importance analysis reveals that the number of students and the student-to-teacher and staff ratio are the most influential predictors. These findings highlight the importance of maintaining accurate and complete records of students, teachers, and staff, verified through official documentation such as family cards, birth certificates, and assignment letters. Achieving this goal requires effective coordination among principals, teachers, data operators, students, and parents. Strengthening data governance will enhance the quality and reliability of educational data, support evidence-based decision-making, and enhance management and accountability in madrasahs.

ACKNOWLEDGMENTS

The author expresses gratitude to the Ministry of Religious Affairs and Lembaga Pengelola Dana Pendidikan (LPDP) for supporting the author in continuing his studies through the Beasiswa Indonesia Bangkit (BIB) Program.

REFERENCES

Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines*, pp. 67–80. Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Babbie, E. R. (2020). *The Practice of Social Research* (15th ed.). USA: Cengage Learning.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(5), 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (2003). *Manual on Setting Up, Using, and Understanding Random Forests V4.0*. https://www.Stat.Berkeley.Edu/~breiman/Using_random_forests_v4.0.Pdf.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 161–168. https://doi.org/10.1145/1143844.1143865

Chicco, D., & Jurman, G. (2020). The Advantages of The Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Cortes, C., & Vapnik, V. (1995). Support-vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15, 3133-3181. http://www.mathworks.es/products/neural-network.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Gorunescu, F. (2011). *Classification Performance Evaluation* (Vol. 12, pp. 319–330). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-19721-5_6

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, H., Zhang, W., & Zhang, S. (2018). A Novel Ensemble Method for Credit Scoring: Adaption of Different Imbalance Ratios. *Expert Systems with Applications*, *98*, 105–117. https://doi.org/10.1016/j.eswa.2018.01.012

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Jian, C., Gao, J., & Ao, Y. (2016). A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. *Neurocomputing*, *193*, 115–122. https://doi.org/10.1016/j.neucom.2016.02.006

Kuhn, M. (2008). Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (Vol. 26). USA: Springer.

Kulkarni, V. Y., & Sinha, P. K. (2014). Effective Learning and Classification using Random Forest Algorithm. *International Journal of Engineering and Innovative Technology (IJEIT)*, *3*(11), 267–273.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. http://www.stat.berkeley.edu/

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, *6*(1), 79–89.

Noble, W. S. (2006). What is a Support Vector Machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Ren, F., Cao, P., Li, W., Zhao, D., & Zaiane, O. (2017). Ensemble Based Adaptive Over-Sampling Method for Imbalanced Data Learning in Computer Aided Detection of Microaneurysm. *Computerized Medical Imaging and Graphics*, *55*, 54–67. https://doi.org/10.1016/j.compmedimag.2016.07.011.

Syarip, D. I., & Rosidin. (2003). *Education Data and Information Management System within the Directorate General of Islamic Institutions*. Directorate General of Islamic Institutions, Ministry of Religious Affairs of the Republic of Indonesia.