Analysis of Civil Servant Teacher Certification Patterns in Asahan Regency Using the C4.5 and Random Forest Algorithm

Nurhawa Nizar Siagian, Muhammad Iqbal, Andysah Putera Utama Siahaan, Darmeli Nasution, Zulham Sitorus

Abstract

The determination of civil servant teacher certification patterns has so far been conducted administratively based on minimum requirement fulfillment, without systematically uncovering the relationships among teacher attributes, particularly for a population of thousands of teachers with varying characteristics. This study aims to analyze the certification pattern of teacher elementary school teachers in Asahan Regency using the C4.5 and Random Forest algorithms, to identify the most dominant attributes associated with certification status, and to compare the classification performance of both algorithms. A quantitative approach was applied using secondary data from the Basic Education Data of the Asahan Educational Office for the 2024/2025 academic year, covering 2,330 of elementary school teachers (civil servants and government contract employees) with eight predictor attributes (age, sub-district, position, education level, employment status, rank/grade, years of service, and weekly teaching hours) and one target attribute (certification status). The data were split using an 80:20 train–test scheme and evaluated using confusion matrix, accuracy, precision, recall, F1-score, and the ROC-AUC curve. Results show that C4.5 achieved an accuracy of 82.90%, a precision of 84.42% and 80.65%, and an AUC of 0.8984, while Random Forest achieved a higher accuracy of 83.98%, with a more balanced precision of 87.36% and 79.60% across classes. Feature importance analysis identified rank as the most dominant predictor of certification status for both algorithms, followed by age and years of service. Overall, the two algorithms achieved comparable performance, with C4.5 offering stronger interpretability and Random Forest offering more stable feature importance estimates, making both valuable tools for education authorities in evaluating teacher certification patterns.

Keywords: *Teacher certification; C4.5 Algorithm; Random Forest*

Nurhawa Nizar Siagian
[1] Master of Information Technology, Panca Budi University, Indonesia
Email: hawanisiagian@gmail.com

Muhammad Iqbal, Andysah Putera Utama Siahaan, Darmeli Nasution, Zulham Sitorus
[2,3,4,5]Master of Information Technology, Panca Budi University, Indonesia
Email: muhammadiqbal@dosen.pancabudi.ac.id, andiesiahaan@gmail.com, darmelinasution@gmail.com, zulhamsitorus@dosen.pancabudi.ac.id

Introduction

Improving the quality of education through greater teacher professionalism is a national priority in Indonesia. The teacher certification program, mandated under Constitution Number. 14 of 2005 on Teachers and Lecturers, is one of the primary instruments for ensuring that teachers possess the pedagogical competence and professional accountability required to meet national standards [1].

In practice, however, certification patterns are still determined administratively based on minimum requirement fulfillment, without systematically analyzing how attributes such as rank, age, years of service, and educational background interact. This limitation becomes more pronounced when thousands of teachers with varying characteristics must be evaluated, motivating the use of analytical techniques capable of processing large volumes of administrative data and revealing hidden patterns.

Machine learning offers a powerful approach for such evaluation. The C4.5 decision tree is a widely used method due to its interpretability and rule-based classification output [2], while Random Forest, an ensemble learning technique, has demonstrated improved predictive accuracy and reduced overfitting by combining multiple decision trees [3]. Prior studies applying C4.5 to teacher certification data have identified rank as the most decisive root attribute [2], yet few studies have directly compared C4.5 and Random Forest for this task in the Indonesian context using both classification-status targets and an expanded predictor set.
This study addresses that gap by comparing C4.5 and Random Forest for predicting teacher certification status in Asahan Regency, using eight predictor attributes. The objectives are to: (1) evaluate and contrast the predictive performance of C4.5 and Random Forest, (2) determine the most dominant attributes influencing certification status, and (3) provide evidence-based insights to support the Asahan Educational Office in evaluating certification patterns.

Literature Review

A decision tree maps data attributes into a tree structure based on decision rules. C4.5 builds this tree by selecting the splitting attribute with the highest gain ratio, computed from the entropy of each attribute, producing rules that are directly interpretable [2]. Random Forest extends this idea through ensemble learning: it builds many decision trees from bootstrap samples and random feature subsets, then aggregates predictions through majority voting [3], which increases robustness against overfitting compared with a single tree .

Several studies have compared the two algorithms across classification domains. Ernawati and Karim compared Random Forest and C4.5 for car-price classification [4], while Sudarto et al. compared their accuracy for classifying student performance [5]; both studies reported Random Forest slightly outperforming C4.5 in accuracy, while C4.5 retained an advantage in interpretability. In the health domain, Barutu et al. found C4.5 slightly outperforming Random Forest when predicting diarrheal disease in toddlers [6], indicating that the relative performance of the two algorithms is dataset-dependent.

Within the same institutional context, several studies applied C4.5 or Random Forest to administrative classification problems similar to this study. Hamzah and Sitorus applied C4.5 to classify diabetes indications with an accuracy of 85.78% [7]. Ismail et al. applied Random Forest and Gradient Boosting Machines to analyze the ASN Professionalism Index in Asahan Regency the same research location as this study—identifying the implementing-staff position as the most dominant attribute [8]. Ernawati et al. compared Random Forest and K-Nearest Neighbors for classifying poverty status in Labuhanbatu Regency [9], and Siahaan et al. applied C4.5 to predict user age groups from demographic attributes in a public satisfaction survey [10]. These studies collectively support the relevance of C4.5 and Random

Forest for classifying administrative status based on demographic and employment attributes, as undertaken in this study.

Beyond direct algorithm comparisons, several studies have applied machine learning classification more broadly to administrative and decision-support contexts related to this study. Rasyid et al. applied machine learning to improve the efficiency of the student admission decision-making process at Pembangunan Panca Budi University[11] . Siregar et al. compared Naive Bayes and C4.5 for selecting UMKM product types based on vocational training data [12], and Sunjaya et al. applied Naive Bayes and Support Vector Machine to classify online shopping ratings [13]. In a directly analogous classification task, Fitriyanti et al. applied C4.5 to classify student graduation predicates based on GPA and study-period requirements, demonstrating the algorithm's suitability for predicting categorical academic and administrative outcomes similar to certification status [14].

Research Methodology

This study follows a quantitative approach, which emphasizes the analysis of numerical data processed using statistical and computational methods [15]. The research procedure comprises seven interrelated stages, as illustrated in Fig. 1: literature study, data collection, data processing, exploratory data analysis, model implementation, model evaluation, and results and reporting.

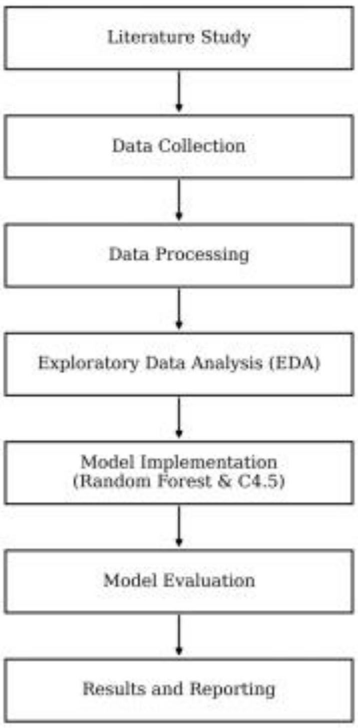

Fig. 1. Research methodology framework [17].

In the literature study stage, relevant prior research on the C4.5 and Random Forest algorithms, teacher certification, and related classification studies was reviewed to establish the theoretical basis and identify the research gap addressed in this study.

The data collection stage involved gathering secondary data from the Basic Education Data (Dapodik) of the Asahan Regency Education Office for the 2024/2025 academic year, covering teacher records relevant to the certification program.

During data processing, the collected records were cleaned and standardized through data selection, encoding of categorical attributes, and removal of incomplete or duplicate entries, followed by an 80:20 split into training and test sets.

Exploratory Data Analysis (EDA) was then performed using descriptive statistics and visualization to examine attribute distributions, relationships, and the balance between certified and not-yet-certified classes, informing subsequent modeling decisions.

In the model implementation stage, classification models were built using the C4.5 and Random Forest algorithms on the training data, as detailed in Sections III-B and III-C. The model evaluation stage assessed both models on the held-out test set using confusion matrix, accuracy, precision, recall, F1-score, and ROC-AUC, enabling a direct comparison of their classification performance, as detailed in Section III-D.Finally, in the results and reporting stage, the evaluation outcomes and feature importance findings were compiled into statistical summaries and interpretive discussion to support decision-making by the Asahan Educational Office, as presented in Section IV.

3.1 Data Acquisition and Preparation

The dataset used in this study represents elementary school civil servant teachers participating in the certification program, obtained from the Basic Education Data of the Asahan Educational Office for the 2024/2025 academic year. A total of 2,310 records were retained for analysis out of 9,101 initial records after data cleaning and selection.

The dataset includes eight predictor attributes and one target attribute, certification status, as shown in Table I. Age, sub-district, position, education, employment status, rank, years of service, and weekly teaching hours are the predictors used to build the classification models.

A hold-out method with an 80:20 train–test split was employed, using total sampling for the population [16]. This resulted in 1,868 training records and 462 test records, on which model performance was evaluated.

Table I. Attribute Descritions

Variable	**Data Type**
Age	Int64
Region Name	Object
Position (PTK)	Object
Education	Object
Employment Status	Object
Rank	Object
Years of Service	Int64
Weekly Teaching Hours	Int64
Certification Status	Object (Target)

3.2 *Implementation of C4.5 for Teacher Certification*

A decision tree for predicting teacher certification status was built using the C4.5 algorithm. The algorithm splits the dataset based on the attribute that produces the greatest reduction in entropy, defined as [16]:

$$Entropy = -\sum_{i=1}^{k} P_i * log_2 P_i \quad (16)$$

where P_i denotes the proportion of class i in dataset S. Because information gain can favor attributes with many distinct values, C4.5 uses the Gain Ratio for attribute selection [17]:

$$GainRatio\ (S, A) = \frac{Gain(S,A)}{Split\ (S,A)} \quad (19)$$

The attribute with the highest gain ratio is chosen as the decision node, and the process repeats recursively until a stopping criterion is met. The resulting tree produces rules that explain how attributes such as rank, age, and years of service relate to certification status. [18]

3.3 *Implementation of Random Forest for Teacher Certification*

Random Forest builds multiple decision trees from bootstrap samples of the dataset, considering only a random subset of attributes at each split. This reduces correlation between trees and helps prevent overfitting. The prediction for an input instance is obtained through majority voting across all trees: [19]

$$\hat{y} = mode\ \{h_1(x), h_2(x), \ldots\ldots, h_T(x)\} \quad [20]$$

where $h_1(x)$denotes the prediction of the $t-th$ tree and T is the total number of trees. In addition to improving predictive stability, Random Forest produces feature importance scores that highlight the attributes most strongly associated with certification status, which is valuable for policymakers seeking to understand certification patterns.

3.4 *Model Evaluation*

Both models were evaluated on the same 462-record test set using confusion matrix, accuracy, precision, recall, F1-score, and the ROC-AUC curve, allowing for a direct and consistent comparison of classification performance.

Results

Testing was carried out on 462 records (20% of the dataset), separate from the 1,868 training records used to build the models.

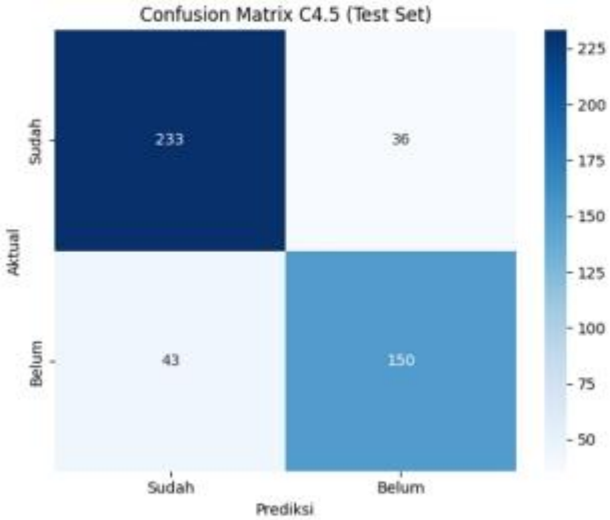

Fig. 2. C4.5 confusion matrix.

The confusion matrix of the C4.5 model (Fig. 2) shows that the model correctly classified 233 certified teachers and 150 not-yet-certified teachers, while 36 certified teachers were misclassified as not-yet-certified and 43 not-yet-certified teachers were misclassified as certified.

Table II presents the C4.5 classification report, showing an overall accuracy of 82.90% on 462 test records, with precision of 84.42% and recall of 86.62% for the certified class, and precision of 80.65% and recall of 77.72% for the not-yet-certified class. Compared with the imbalanced-recall behavior often reported for C4.5 on skewed datasets, this model shows a relatively balanced recall across both classes, reflecting a moderately balanced class distribution in the dataset (269 certified vs. 193 not-yet-certified in the test set); this contrasts with more heavily skewed datasets, where SMOTE-based resampling has been shown necessary to improve minority-class recall for Random Forest and Gradient Boosting classifiers [22].

Table II. C4.5 Classification Report

Class	Precision	Recall	F1-Score	Support
Certified	84.42%	86.62%	85.50%	269
Not Cert.	80.65%	77.72%	79.16%	193
Accuracy			82.90%	462
Macro Avg	82.53%	82.17%	82.33%	462
Weighted Avg	82.84%	82.90%	82.85%	462

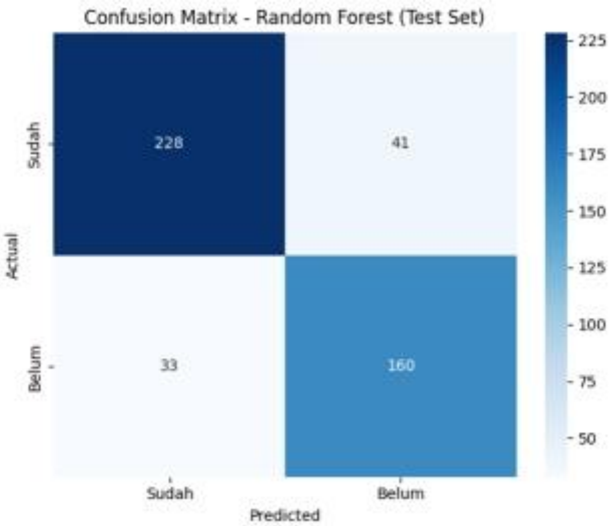

Fig. 3. Random Forest Confusion Matrix.

The confusion matrix of the Random Forest model (Fig. 3) shows that the model correctly classified 228 certified teachers and 160 not-yet-certified teachers, with 41 certified teachers misclassified as not-yet-certified and 33 not-yet-certified teachers misclassified as certified.

As shown in Table III, Random Forest achieved an overall accuracy of 83.98%, with precision of 87.36% and recall of 84.76% for the certified class, and precision of 79.60% and recall of 82.90% for the not-yet-certified class. Compared with C4.5, Random Forest produced higher precision for the certified class and higher recall for the not-yet-certified class, resulting in a higher weighted-average F1-score (84.02% versus 82.85%).

Table III. Random Forest Classification Report

Class	Precision	Recall	F1-Score	Support
Certified	87.36%	84.76%	86.04%	269
Not Cert.	79.60%	82.90%	81.22%	193
Accuracy			83.98%	462
Macro Avg	83.48%	83.83%	83.63%	462
Weighted Avg	84.12%	83.98%	84.02%	462

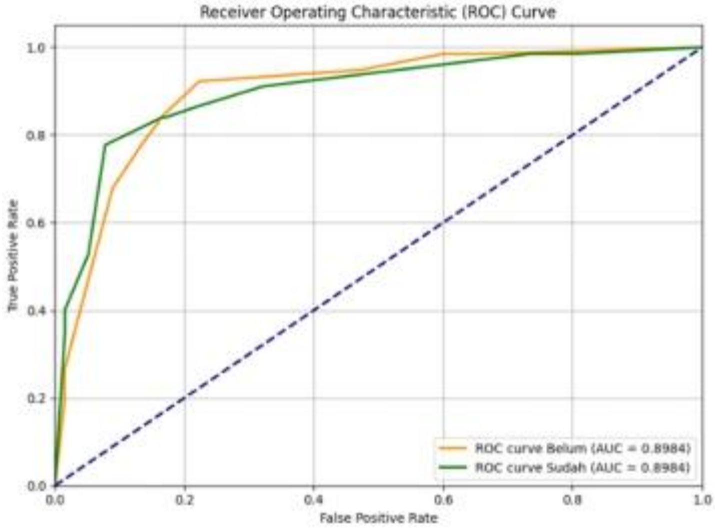

Fig. 4. ROC curve of the C4.5 decision tree.

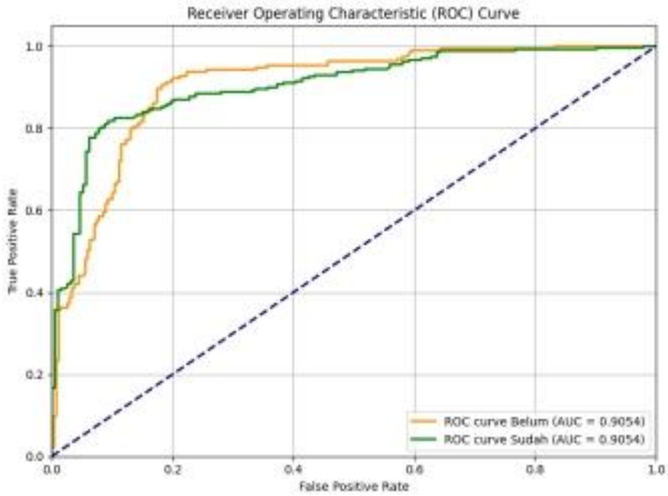

Fig. 5. ROC curve of the Random Forest.

Receiver Operating Characteristic (ROC) analysis was used to further evaluate the discriminatory ability of the C4.5 classifier. As shown in Fig. 4, the ROC curve for both the certified and not-yet-certified classes sits well above the diagonal reference line, with an Area Under the Curve (AUC) of 0.8984 for both classes, indicating good discriminatory capability. An equivalent ROC-AUC evaluation for Random Forest as shown in Fig. 5, the ROC curve for both the certified and not-yet-certified classes sits well above the diagonal reference line, with an Area Under the Curve (AUC) of 0.9054 for both classes, indicating good discriminatory capability

Both algorithms consistently identified rank as the most dominant attribute associated with certification status, with a gain ratio of 0.7515 in C4.5 and a feature importance of 0.395 in Random Forest, consistent with methodological guidance on interpreting random forest variable importance scores in behavioral and social science research [21]. Teachers in lower ranks (II/a–III/b) were predominantly not yet certified, while teachers in higher ranks (III/c and above) were predominantly certified. In C4.5, subsequent splits occurred on age and years of service, whereas in Random Forest, years of service (0.155) and employment status (0.127) showed a more pronounced contribution, reflecting the ensemble's ability to capture attribute contributions that may be masked by rank in a single tree. This pattern parallels findings in other administrative classification contexts, such as employee promotion prediction using Random Forest, where similarly formal hierarchical attributes have been shown to be strong predictors of status outcomes [22]. Because holding a teaching certificate is itself often a prerequisite for rank promotion among civil servants, this relationship should be interpreted as associative rather than causal.

Overall, the accuracy difference between the two algorithms (1.08 percentage points) indicates no substantial practical gap in performance for this dataset. C4.5 offers stronger interpretability through its explicit decision rules, while Random Forest offers more balanced classification performance and more stable feature importance estimates, making both algorithms valuable tools for the Asahan Regency Education Office in evaluating teacher certification patterns.

Conclusion

This study successfully built C4.5 and Random Forest classification models to analyze the certification pattern of ASN elementary school teachers in Asahan Regency based on eight predictor attributes. The C4.5 algorithm identified rank/grade as the root node with the highest information gain, followed by subsequent splits on age and years of service, achieving an accuracy of 82.90% on 462 test data points. The Random Forest algorithm produced a similar pattern through feature importance analysis across 100 decision trees, achieving an accuracy of 83.98%, with rank/grade remaining the top attribute while also revealing a more pronounced contribution from years of service and employment status compared to C4.5.

The performance comparison shows that the two algorithms have relatively comparable classification capability (an accuracy difference of 1.08 percentage points), so both are worth considering as the basis for an evaluation support system for the Asahan Regency Education Office. Because the dominance of the rank/grade attribute is potentially associative due to administrative promotion requirements, further qualitative verification is recommended before this pattern is used as a basis for causal policy. Future research is recommended to consider non-demographic factors such as teacher performance, competency test results, and motivation, and to maintain regular updates of Dapodik data to improve the accuracy and reliability of predictive models in the future.

Refrences

[1] UU No. 14, "Undang-undang (UU) Nomor 14 Tahun 2005 tentang Guru dan Dosen," Jakarta, Dec. 2005.

[2] S. Jailani, Defit and G. W. Nurcahyo, "Penerapan Algoritma C4 . 5 Pada NUPTK Untuk Menentukan Pola Sertifikasi Guru Dengan Menggunakan Metode Klasifikasi Decision Tree (Studi Kasus : Pada Dinas Pendidikan Pemuda dan Olahraga Kabupaten Rokan Hulu) Of Computer Science," *J. Comput. Sci.*, vol. 1, no. 1, pp. 69–83, 2015.

[3] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[4] Ernawati and A. Karim, "Komparasi Kinerja Algoritma Random Forest dan C4.5 untuk Klasifikasi Harga Mobil," *Bul. Ilm. Inform. Teknol.*, vol. 3, no. 1, pp. 25–32, Sep. 2024, doi: 10.58369/biit.v3i1.95.

[5] Sudarto, E. Sufarnap, and M. Ilhami, "Komparasi Tingkat Akurasi Random Forest dan C4.5 Pada Klasifikasi Performa Mahasiswa," *Prem. Informatics Eng. Yayasan Literasi Emas Nusant.*, vol. 2, no. 1, Apr. 2025.

[6] Barutu, Sipra, Iqbal, Muhammad, Khairu, Khairul, Nasution, "Comparative Analysis of the C4.5 and Random Forest Algorithms for the Prediction of Diarrheal Disease,"

Sean Inst., vol. Vol. 4 No., 2025.

[7] Hamzah, Zulham Sitorus, and Khairul, “Analisa Classification Decision Tree C45 dan Naïve Bayes Pada Indikasi Penyakit Diabetes Menggunakan Rapid Miner,” *J. Nas. Teknol. Komput.*, vol. 4, no. 1, pp. 25–33, Jan. 2024, doi: 10.61306/jnastek.v4i1.126.

[8] R. F. Ismail, Syahlan, Iqbal, Muhammad, Wijaya, “Analisis Pola Peningkatan Indeks Profesionalitas ASN Kabupaten Asahan Menggunakan Random Forest Dan Gradient Boosting Machines,” *Cattleya Darmaya Fortuna*, vol. Vol. 7 No., 2025.

[9] D. Ernawati, Andi, Khairul, Khairul, Sitorus, Zulham, Iqbal, Muhammad, Nasution, “Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan K-Nearest Neighbors,” *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 22–35, 2025, doi: https://doi.org/10.47065/bit.v6i1.1783.

[10] S. Utama, Putera, Andysah, Jabal, Abdul, Ami, Parhusip, Nelviony, Indrayani, Maida, Barutu, “Analysis of User Age Predictions in Public Satisfaction Surveys at Public Service Malls Using Decision Tree C4.5,” *JITCSE*, vol. 1, pp. 192–133, 2024, doi: https://doi.org/10.61306/jitcse.v1i2.58.

[11] P. U. S. Andi. Siregar, L. Marlina, Khairul, M. Iqbal, “Machine Learning Analysis in Improving the Efficiency of the Student Admission Decision Making Process New At Panca Budi Medan Development University,” *Bull. Eng. Sci. Technol. Ind.*, vol. 2, no. 3, pp. 239–225, 2024.

[12] P. U. Siregar, Alex, Marlina, Leni, Khairul, Khairul, Iqbal, Muhammad, Siahaan, “COMPARATIVE ANALYSIS OF NAIVE BAYES ALGORITHM AND C4.5 ALGORITHM IN SELECTING TYPES OF UMKM PRODUCTS IN BBPSDMP KOMINFO MEDAN TRAINING,” *PT. Radja Intercont. Publ.*, vol. Vol. 2 No., 2024.

[13] Muit Sunjaya, Zulham Sitorus, Khairul, Muhammad Iqbal, and A.P.U Siahaan, “Analysis of machine learning approaches to determine online shopping ratings using naïve bayes and svm,” *Int. J. Comput. Sci. Math. Eng.*, vol. 3, no. 1, pp. 7–16, May 2024, doi: 10.61306/ijecom.v3i1.60.

[14] V. Fitriyanti, G. Testiana, and C. Eri Gunawan, “Klasifikasi Predikat Kelulusan Mahasiswa Menggunakan Algoritma C4.5,” *J. SAINTEKOM*, vol. 14, no. 2, pp. 217–232, Sep. 2024, doi: 10.33020/saintekom.v14i2.626.

[15] V. W. Sujarweni, “Metodologi Penelitian Bisnis dan Ekonomi V. Wiratna Sujarweni,” *Metodol. Penelit.*, 2015.

[16] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, 2011.

[17] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How Many Trees in a Random Forest?,” 2012, pp. 154–168. doi: 10.1007/978-3-642-31537-4_13.

[18] C. B. Sucahyo *et al.*, “Performance Analysis of Random Forest on Quartile Classification Journal,” *Appl. Eng. Technol.*, vol. 3, no. 1, pp. 1–17, Mar. 2024, doi: 10.31763/aet.v3i1.1189.

[19] S. Hong and H. S. Lynn, “Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction,” *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 199, Dec. 2020, doi: 10.1186/s12874-020-01080-1.

[20] I. J. Rifka Alkhilyatul Ma’rifat, I Made Suraharta,” *World Resources Institute Indonesia*, vol. 2, Indonesia, pp. 306–312, Apr. 22, 2024. [Online]. Available: https://wri-indonesia.org/en/insights/advancing-strong-and-sustainable-downstreaming-indonesias-fisheries-sector

[21] Y. Rothacher and C. Strobl, “Identifying Informative Predictor Variables With Random Forests,” *J. Educ. Behav. Stat.*, vol. 49, no. 4, pp. 595–629, Aug. 2024, doi: 10.3102/10769986231193327.

[22] J. Tuhairwe, J. K. Mukisa, and A. A. Al-Absi, “Web-Based Employee Promotion Prediction Using Random Forest Classifier,” 2024, pp. 459–475. doi: 10.1007/978-981-97-0573-3_37.