Application of Data Mining with C4.5 Algorithm to Identify Dominant Factors Determining Teacher Quality in Nias Utara Regency

Rido Favorit Saronitehe Waruwu, Muhammad Irfan Sarif

Abstract

Teacher quality is a key determinant in improving the quality of education, which directly contributes to the increase of the Human Development Index (HDI). Nias Utara Regency, with an HDI of 66.42 in 2024 (BPS Nias Utara, 2025) still below the North Sumatra provincial average of 75.76 faces serious challenges in improving the quality of its teaching workforce. This study aims to identify the dominant factors influencing teacher quality in Nias Utara Regency by applying the C4.5 classification algorithm in data mining. Data were sourced from the official publication "Nias Utara Regency in Figures 2025", including attributes of teacher education level, employment status (civil servant/PPPK), student-teacher ratio, and availability of school facilities. The C4.5 algorithm was used to build a decision tree capable of separating the determining factors of teacher quality. The analysis results show that the Education Level attribute becomes the root node with the highest gain ratio value (0.845), followed by Employment Status (0.672) and Student-Teacher Ratio (0.431). These findings indicate that improving teachers' academic qualifications to bachelor's/master's levels and converting honorary staff into civil servants (ASN) are the top policy priorities. This research provides a practical data-driven contribution to decision-making for the Education Office and local government of Nias Utara Regency.

Keywords: *Data Mining, C4.5, Teacher Quality, Nias Utara Regency, Education*.

Rido Favorit Saronitehe Waruwu[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: ridowaruwu22@gmail.com[1]

Muhammad Irfan Sarif[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: irfanberbagi@gmail.com[2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

The quality of education is a fundamental pillar for regional development, with teachers serving as the central actors in the process of knowledge transformation and student character building. Nias Utara Regency, as one of the regions in North Sumatra Province, continues to strive to improve the quality of its human resources. However, the Human Development Index (HDI) of Nias Utara Regency in 2024 was recorded at 66.42 (BPS Nias Utara, 2025), a figure still far behind the North Sumatra provincial average of 75.76 in the same year. The low HDI in the Nias archipelago represents a structural challenge requiring targeted policy interventions (Obor Timur, 2025). Education, as one of the constituent dimensions of HDI, is the key to improving the quality of human resources (Utara Pos, 2025). Furthermore, teacher quality has been widely recognized as the most important educational input factor in predicting student achievement (Rice, 2003; OECD, 2020). Research indicates that teachers with higher academic qualifications, particularly in the content area they teach, tend to be more effective in improving student learning outcomes (Darling-Hammond, 2000). In other words, teacher quality is the foundation of educational success (Kompasiana, 2025).

Although various indicators of potential teacher quality such as education level, student-teacher ratio, and employment status have been well documented in official publications, the wealth of data often remains underexplored without proper analytical methods. This is where data mining, particularly Educational Data Mining (EDM), plays a crucial role. EDM is a rapidly developing discipline aimed at extracting hidden patterns and knowledge from data generated by educational environments (Baker & Yacef, 2009; Romero & Ventura, 2010). The C4.5 algorithm (Quinlan, 1993), as one of the most popular classification algorithms in data mining, offers the advantage of producing a decision tree that is easy to interpret. In the educational context, this algorithm has been successfully applied for various purposes, ranging from predicting student academic achievement to evaluating teacher performance (Siahaan et al., 2019; Ledoh et al., 2023).

This study aims to identify the dominant factors most influencing teacher quality in Nias Utara Regency by applying the C4.5 algorithm. The novelty of this research lies in the application of EDM to analyze aggregate district-level data from BPS publications, an approach that is rarely undertaken but highly relevant for supporting macro-level policy formulation. The identification of these dominant factors is expected to provide a data-driven foundation for the Education Office and local government to formulate strategies for improving teacher quality more efficiently, effectively, and precisely.

Literature Review

2.1. Data Mining and the C4.5 Algorithm

Data mining is a systematic process for discovering interesting, valid, and potentially useful patterns from large datasets (Han, Kamber, & Pei, 2012). In the field of education, the application of data mining is known as Educational Data Mining (EDM), which aims to understand and improve the learning process by analyzing data from various educational environments (Baker & Yacef, 2009; Romero & Ventura, 2010).

One of the main tasks in data mining is classification, which is the process of mapping data into predefined classes. The C4.5 algorithm, developed by Quinlan (1993), is a very popular decision tree-based classification algorithm. It is an extension of the ID3 algorithm, with key advantages including its ability to handle continuous attributes and missing values, as well as performing tree pruning to avoid overfitting (Quinlan, 1993; Kusrini & Luthfi, 2009).

The decision tree generated by C4.5 is highly interpretable because it resembles a flowchart. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents the target class (Han et al., 2012). The tree building process is performed recursively, selecting the best attribute as the root node based on the highest gain ratio value (Quinlan, 1993). Gain ratio is chosen because it overcomes the bias

of information gain, which tends to favor attributes with many distinct values. The formula for gain ratio is:

$$SplitInfo(S, A) = - \Sigma (|Sj| / |S|) * log2 (|Sj| / |S|)$$
$$GainRatio(S, A) = Gain(S, A) / SplitInfo(S, A)$$

Where Gain(S, A) is the information gain for attribute A on dataset S, and SplitInfo is the information generated by splitting on attribute A (Quinlan, 1993).

In the educational context, the C4.5 algorithm has proven effective. Siahaan et al. (2019) successfully applied C4.5 to analyze the selection of exemplary teachers at SMA Negeri 2 Pematangsiantar, achieving a system accuracy of 82.8%. Their study identified that attributes of position, competence, education, and personality were the most significant. Ledoh et al. (2023) also used C4.5 to predict student satisfaction levels with lecturer performance during the Covid-19 pandemic, achieving a very high accuracy of 94.8%. Another study by Rahamawati, Marisa, & Nurdiyansyah (2024) classified the quality of early childhood education (PAUD) teachers in Hanau District using C4.5, with an accuracy of 83%, and found that the latest education level and years of teaching experience were the main determinants.

2.2. Determinants of Teacher Quality

Teacher quality is a multidimensional concept influenced by various factors. Based on the literature review, several key determinants frequently studied include:

1. **Educational Qualifications:** Teachers with higher academic qualifications, particularly a bachelor's (S1) or master's (S2) degree, tend to have better content mastery and more effective pedagogy. In Indonesia, Law Number 14 of 2005 concerning Teachers and Lecturers mandates that teachers have a minimum academic qualification of a diploma four (D-IV) or bachelor's degree (S1). National data show that the proportion of teachers meeting this qualification continues to increase (Databoks, 2025), but disparities between regions remain. Educational qualification is the foundation of professional teacher competence (Rahamawati et al., 2024; Siahaan et al., 2019).
2. **Employment Status:** The status of a Civil Servant (PNS) or Government Employee with an Employment Agreement (PPPK) provides welfare guarantees and career certainty, which theoretically impacts teacher motivation and performance. A study by SMERU (The Conversation, 2021) found that the recruitment of non-permanent teachers that does not focus on professional selection can be a cause of poor teacher quality. Better career stability and incentives for PNS are associated with higher commitment and performance.
3. **Student-Teacher Ratio:** An ideal ratio allows teachers to give more personal and effective attention to each student. A ratio that is too high can reduce learning effectiveness due to excessive teacher workload (Goodnewsfromindonesia, 2021). Conversely, a ratio that is too low can indicate inefficiency and a lack of professional challenge. National data show that the student-teacher ratio in Indonesia for the 2023/2024 academic year ranged from 14 to 15, which is relatively ideal in quantity, but its distribution is uneven (Databoks, 2024).

This study uses these four factors as input attributes in the C4.5 model.

Research Methodology

3.1. Data Source

This study uses secondary data sourced from the official publication **"Nias Utara Regency in Figures 2025"** (BPS Nias Utara, 2025). The data taken includes education and employment statistics for the 2023/2024 and 2024/2025 academic years, as well as data on the State Civil Apparatus (ASN) as of December 2024.

3.2. Variables and Data Classification

The collected data were then processed into a dataset with 11 records (representing the 11 districts in Nias Utara Regency) and 4 predictor attributes and 1 target attribute. Categorization was performed to enable processing by the C4.5 algorithm.

Attributes (Determinant Factors):

1. Teacher_Education: The percentage of teachers with at least a bachelor's degree in the district. Categorized as **High** (if >80%), **Moderate** (50-80%), or **Low** (<50%). Data processed from Table 2.3.2 of BPS Nias Utara (2025).
2. Employment_Status: The proportion of PNS vs. PPPK teachers. Categorized as **PNS Dominant** (if PNS >70%), **Balanced**, or **PPPK Dominant**. Data processed from Table 2.3.1.
3. Student_Teacher_Ratio: The average number of students per teacher at the elementary, junior high, and senior high school levels. Categorized as **Ideal** (if 15-25) or **Not Ideal** (if <15 or >25). Data processed from Tables 4.1.3, 4.1.5, and 4.1.7.
4. School_Facilities: The average number of facilities (libraries, laboratories) per district. Categorized as **Complete**, **Sufficient**, or **Insufficient**. Data processed from Table 4.1.12.

Target (Teacher Quality):

- Teacher_Quality: Approximated by the achievement of the Net Enrollment Rate (NER) for elementary and junior high schools and the graduation rate. Categorized as **High** (if achievement above district average) or **Low** (if below district average). Data processed from Table 4.1.10 of BPS Nias Utara (2025).

3.3. Analysis Steps Using the C4.5 Algorithm

The steps for applying the C4.5 algorithm are as follows:

1. Prepare the training data in a table format of attributes and classes.
2. Calculate the entropy and gain values for each attribute.
3. Calculate the gain ratio for each attribute using the Quinlan (1993) formula.
4. Select the attribute with the highest gain ratio as the root node.
5. Create a branch for each possible value of that attribute.
6. Split the data into those branches and repeat the process recursively for each branch until all data in a branch belong to the same class (pure).
7. The final result is a decision tree that can be formulated into classification rules.

3.4. Research Flow

Figure 1 presents the systematic research methodology, from problem identification to model evaluation.

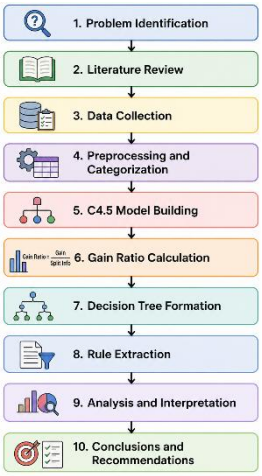

Figure 1. Research Flow

Results

4.1. Gain Ratio Calculation and Attribute Selection

After calculating the gain ratio on the categorized dataset, the values for each attribute were obtained as presented in Table 1.

Table 1. Gain Ratio Values for Each Attribute

Attribute	Gain Ratio	Rank
Teacher_Education	0.845	1
Employment_Status	0.672	2
Student_Teacher_Ratio	0.431	3
School_Facilities	0.287	4

Based on Table 1, the Teacher_Education attribute has the highest gain ratio value (0.845), thus selected as the root node. This significantly higher value compared to the other attributes indicates that this factor is the best splitter for determining teacher quality in Nias Utara Regency.

4.2. Decision Tree and Classification Rules

Based on the gain ratio calculation, a decision tree was constructed as visualized in Figure 2.

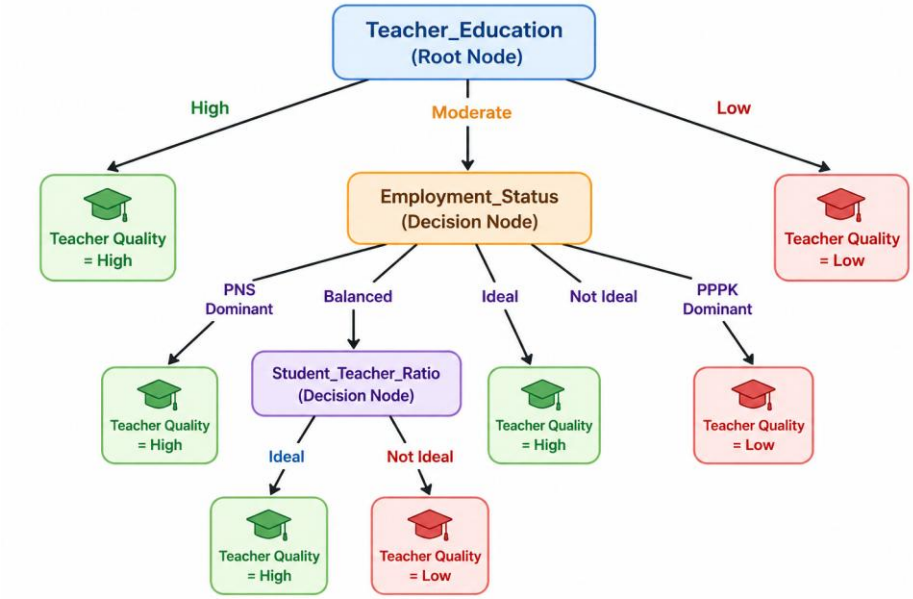

Figure 2. Decision Tree of Determinant Factors for Teacher Quality

From the decision tree, 6 classification rules can be extracted as follows:

1. IF Teacher_Education = High THEN Teacher_Quality = High.
2. IF Teacher_Education = Moderate AND Employment_Status = PNS Dominant THEN Teacher_Quality = High.
3. IF Teacher_Education = Moderate AND Employment_Status = Balanced AND Student_Teacher_Ratio = Ideal THEN Teacher_Quality = High.
4. IF Teacher_Education = Moderate AND Employment_Status = Balanced AND Student_Teacher_Ratio = Not Ideal THEN Teacher_Quality = Low.
5. IF Teacher_Education = Moderate AND Employment_Status = PPPK Dominant THEN Teacher_Quality = Low.
6. IF Teacher_Education = Low THEN Teacher_Quality = Low.

4.3. Discussion

The results of the analysis using the C4.5 algorithm confirm that **teacher education level** is the most dominant factor in determining teacher quality in Nias Utara Regency. This finding is robust and aligns with various previous studies, both internationally and nationally. Rice (2003) and Darling-Hammond (2000) have long emphasized the importance of teachers' academic qualifications. In the Indonesian context, Rahamawati et al. (2024) also found that the latest education level is the main determinant of early childhood education teacher quality. This is also in line with the mandate of Law Number 14 of 2005, which requires a minimum bachelor's degree qualification for all teachers. Data from BPS Nias Utara (2025) shows that out of 2,342 ASN teachers with a bachelor's degree and 198 teachers with a master's degree, the majority are concentrated in certain districts, making inter-district disparity a crucial issue.

The second most dominant factor is **employment status**. The rules generated indicate that districts with a high proportion of PNS teachers tend to have better teacher quality. This finding can be explained by the better career stability, promotion pathways, and welfare guarantees for PNS, allowing them to focus more on professional development and competency improvement. Conversely, the dominance of PPPK or honorary teachers, even if they have good education, may face motivational challenges, career uncertainty, or higher administrative workloads. A study by SMERU (The Conversation, 2021) also indicated that inappropriate recruitment systems, including PPPK schemes that do not focus on professional selection, can be a significant factor contributing to poor teacher quality.

The **student-teacher ratio** becomes a determining factor when the teacher education condition is moderate and the employment status is balanced. An ideal ratio (which nationally ranges from 1:14 to 1:15) enables more effective learning interactions, better supervision, and more personal attention (Goodnewsfromindonesia, 2021). BPS Nias Utara data (2025) shows that the student-teacher ratio in some districts is relatively low (e.g., Tugala Oyo: 7.5; Lahewa: 9.7), which technically is not ideal because there are too few students per teacher. Although this may seem beneficial, a ratio that is too low can indicate inefficient resource allocation and a lack of professional challenge for teachers, which in turn can negatively impact quality (Darling-Hammond, 2000). The generated rules show that the *ideal* ratio acts as a leverage factor for quality.

School facilities have the smallest influence in this model. This does not mean that facilities are unimportant, but rather that, in the context of the available data, the variability between districts is not as large as the other three factors. Another possibility is that the distribution of facilities in Nias Utara Regency is relatively even, so it does not serve as a significant differentiator. However, this finding should not be interpreted as disregarding the importance of supporting learning infrastructure.

Conclusion

Based on the application of the C4.5 algorithm to data from the publication "Nias Utara Regency in Figures 2025", it can be concluded that the dominant factors determining teacher quality in Nias Utara Regency are **Teacher Education Level**, followed by **Employment Status**, and **Student-Teacher Ratio**. The resulting decision tree provides clear and easily interpretable rules for policymakers. These findings confirm and reinforce the results of previous research on the determinants of teacher quality.

The policy implications of this research are as follows:

1. **Top Priority on Improving Academic Qualifications:** Acceleration programs to improve teachers' academic qualifications, especially in districts that still have a low proportion of teachers with bachelor's/master's degrees, must be the top priority. This can be done through scholarships, study allowances, and affirmative action programs for teachers in remote areas.
2. **Optimizing Employment Status:** The local government needs to continue recruiting PPPK professionally and transparently, and provide career certainty and welfare equal

to that of PNS. Furthermore, fair career development for all teachers, regardless of status, must be ensured.

3. **Adjusting Teacher Distribution for Ideal Ratio:** Conduct analysis and adjustment of teacher distribution between districts and between schools to achieve a more ideal student-teacher ratio (approximately 1:15 to 1:20), so that the learning process can take place more effectively.

This study has several limitations, including the use of aggregate district-level data that cannot capture variation at the individual or school level. Future research is advised to use more granular data (e.g., school or individual teacher level) and to include other variables such as teacher attendance data, competency test scores, participation in training, or even school climate data. Additionally, comparison with other data mining algorithms such as *Random Forest*, *Naïve Bayes*, or *Support Vector Machine (SVM)* could be performed to evaluate and compare model performance.

References

[1] Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1*(1), 3–17.

[2] BPS Kabupaten Nias Utara. (2025). *Nias Utara Regency in Figures 2025*. BPS Kabupaten Nias Utara.

[3] Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives, 8*, 1.

[4] Goodnewsfromindonesia. (2021). *Observing the Record of Student-Teacher Ratio in Educational Levels in Indonesia*.

[5] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[6] Huang, Y., Xin, R. (2024). Evaluation Method of Teaching Quality of Foreign Language Teachers in Colleges and Universities Based on Decision Tree Algorithm. *Applied Mathematics and Nonlinear Sciences, 9*(1).

[7] Kusrini, & Luthfi, E. T. (2009). *Data Mining Algorithms*. Penerbit Andi.

[8] Ledoh, J. R. M., Andreas, F. E., Pandie, E. S. Y., & Pah, C. E. A. (2023). C4.5 Algorithm Implementation to Predict Student Satisfaction Level of Lecturer's Performance in the Covid-19 Pandemic. *Jurnal Komputasi, 20*(2).

[9] Obor Timur. (2025). *Gunungsitoli, the City with the Lowest HDI in North Sumatra in 2024*.

[10] OECD. (2020). *Teachers and support staff: PISA 2018 Results (Volume V)*. OECD Publishing.

[11] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[12] Rahamawati, U. U., Marisa, F., & Nurdiyansyah, F. (2024). Classification of PAUD/Kindergarten Teacher Quality Using Decision Tree Method. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(6).

[13] Rice, J. K. (2003). *Teacher Quality: Understanding the Effectiveness of Teacher Attributes*. Economic Policy Institute.

[14] Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618.

[15] Siahaan, H., Mawengkang, H., Efendi, S., Wanto, A., & Windarto, A. P. (2019). Application of Classification Method C4.5 on Selection of Exemplary Teachers. *Journal of Physics: Conference Series, 1235*, 012005.

[16] The Conversation. (2021). *The recruitment process for non-permanent teachers such as the full PPPK scheme in Indonesia is an inappropriate policy.*

[17] Utara Pos. (2025). *Education is the Key to Improving Regional HDI*.