7 (3) March 2025 147-154
JSIKTI: *Jurnal Sistem Informasi dan Komputer terapan Indonesia*
ISSN 2460-7258 (online) | ISSN 1978-1520 (print)

Contents lists available at www.infoteks.org

JSIKTI

Journal Page is available to https://infoteks.org/journals/index.php/jsikti

Research article

Comparison Of the Accuracy of Decision Tree Algorithms C4.5 And C5.0 In Predicting Tuition Payment Delays at Mts. Al-Jabar Bali

Ni Wayan Jeri Kusuma Dewi [a,*], *Nia Febriani,* [b], *I Gede Made Yudi Antara* [c]

[a, b] *Department Informatics Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
[c] *Department Computer Systems Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
email: [a,*] wayan.kusumadewi@instiki.ac.id niafebriani220202@gmail.com yudi.antara@instiki.ac.id
* *Correspondence*

ARTICLE INFO

Article history:
Received 7 December 2024
Revised 21 January 2025
Accepted 01 March 2025
Available online 27 March 2025

Keywords:
Education, SPP, Prediction, Decision Tree, C4.5, C5.0

Please cite this article in IEEE style as:
N. W. J. K. Dewi, N. Febriani, and I. G. M. Y. Antara, "Comparison Of the Accuracy of Decision Tree Algorithms C4.5 And C5.0 In Predicting Tuition Payment Delays at Mts. Al-Jabar Bali," *JSIKTI: Jurnal Sistem Informasi dan Komputer Terapan Indonesia*, vol. 7, no. 3, pp. 147–154, 2025.

ABSTRACT

Delays in the payment of Educational Development Contributions (SPP) have become a major issue impacting financial management at MTs. Al-Jabar Bali, with approximately 60% of students experiencing payment delays each year. This study aims to compare the accuracy of Decision Tree algorithms C4.5 and C5.0 in predicting SPP payment delays. The research method adopts the CRISP-DM approach and is implemented using Python on the Google Colaboratory platform. The data used includes students' payment histories, parents' occupations, and income. The models were evaluated using Accuracy, Precision, and Recall metrics. The results show that the C5.0 algorithm has higher accuracy (98%) compared to C4.5 (89%). The C5.0 algorithm is recommended as an effective predictive model to assist schools in making strategic financial management decisions.

1. Introduction

Education is one of the fundamental needs in human life. One form of support for the implementation of education, especially in private schools, is through the payment of Educational Development Contributions (SPP)[1],[2]. SPP serves as the main source of funding to support school operational activities such as the procurement of infrastructure, payment of teachers and staff salaries, as well as the development of education quality[3],[4]. MTs. Al-Jabar Bali, as a private educational institution, relies on SPP as its primary source of funding. However, in the past five years, the school has encountered serious problems regarding delays in SPP payments by students[5]. According to the school's finance department, more than 60% of students experience payment delays every year. This condition directly affects the school's financial stability and potentially hinders the learning process and school development[6],[7].

To address this challenge, a solution is needed that can accurately predict the potential for SPP payment delays[8],[9]. One approach that can be used is data mining methods with Decision Tree algorithms. Decision Tree is a classification method capable of mapping patterns from historical data and making decisions based on rules formed in a tree structure. Commonly used Decision Tree

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

algorithms include C4.5 and C5.0[10],[11]. The C4.5 algorithm is known for its ability to handle both categorical and numerical data and for generating fairly good decision trees[12],[13]. Meanwhile, the C5.0 algorithm is an enhanced version of C4.5, offering advantages in terms of computational speed, accuracy, and more optimal pruning capabilities[14].

Based on this background, this study was conducted to compare the accuracy performance of the C4.5 and C5.0 algorithms in predicting SPP payment delays at MTs[15],[16]. Al-Jabar Bali. The results of this study are expected to provide recommendations for schools to implement the appropriate predictive model in order to improve the efficiency of financial management[17],[18].

2. Research Method

2.1 Prediction

Prediction is the process of estimating future events based on historical data. In the context of this research, prediction is used to estimate whether a student will experience delays in paying tuition fees (SPP) based on previous data patterns. Prediction is similar to forecasting or estimation and is usually based on scientific or subjective methods. The benefits of making predictions, according to include:

1. Understanding future conditions.
2. Planning for production, marketing, finance, and other aspects.
3. Assisting in investment decision-making within a company.

2.2 Data Mining

Data mining is the process of discovering patterns or important information from large data sets. In this research, data mining is used to process student and financial data into accurate prediction models. The techniques used in data mining include classification, association, and clustering[19].

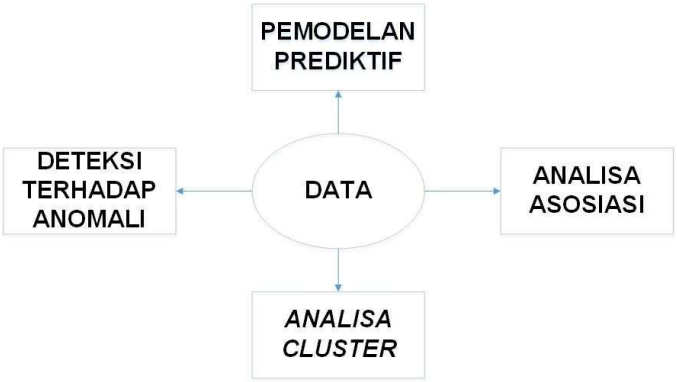

Figure 1. Four Core *Data Mining*

2.3 Decision Tree

A Decision Tree is a classification method that displays rules in a tree structure. Each node in the tree represents an attribute, branches represent the outcomes of the attributes, and the leaves represent the final decision or class.

2.4 C4.5 and C5.0Algorithms

1. C4.5 uses the concepts of entropy and information gain to determine the best attribute for splitting the data. This algorithm builds a decision tree with rules based on the dominant attribute values. This algorithm uses the calculation of Entropy, information gain, split info and gain ratio for selecting attributes to become nodes[20]. The formula for calculating Entropy is as follows:

$$Entropy(S) = \sum_{i=1}^{n} - \ pi * log_2 \ pi \quad (1)$$

Description:
S : Set of Classes
n : Number of partitions in S

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

pi : Proportion of Si relative to S

Where S represents the set of cases, n is the number of partitions in S, and pi is the proportion of the i-th case set relative to the entire case set. Meanwhile, the formula for calculating information gain is as follows:

$$InformationGain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|s_i|}{|s_i|} * Entropy(S) \quad (2)$$

Description:
S : Set of cases
A : Attribute
n : Number of partitions of attribute A
$|Si|$: Number of cases in the i-th partition
$|S|$: Total number of cases in S

2. C5.0 is an improvement over C4.5 that is more efficient, delivers higher accuracy, and supports better pruning and handling of missing values. C5.0 or See5 is an updated version of C4.5, which initially adopted the rules used by ID3[21]. C5 also follows the principles found in C4.5, thus having rules that are almost similar to the C4.5 algorithm. Like C4.5, the C5.0 algorithm also provides features such as attribute selection, cross-validation, and error pruning to reduce errors[22]. The first step in calculating the C5.0 algorithm is to calculate the entropy value using the following formula:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi \quad (3)$$

Description:
S : Set of Classes
n : Number of partitions in S
pi : Proportion of Si relative to S

After obtaining the entropy value for each attribute, the next step is to calculate the information gain using the following formula:

$$InformationGain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|s_i|}{|s_i|} * Entropy(S) \quad (4)$$

Description:
S : Set of cases
A : Attribute
n : Number of partitions of attribute A
$|Si|$: Number of cases in the i-th partition
$|S|$: Total number of cases in S

Finally, in the last stage, the gain ratio is calculated using the following formula:

$$Gain\ Ratio = \frac{InformationGain(S,A)}{\sum_{i=1}^{n} Entropy(s_i)} \quad (5)$$

2.5 Confusion Matrix

The Confusion Matrix is used to evaluate the performance of classification models. The values produced include True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). From these values, metrics such as Accuracy, Precision, and Recall are calculated.
Based on the values of TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative), the values of Accuracy, Precision, and Recall can be calculated.

Accuracy

Accuracy is the degree of closeness between the predicted value and the actual value[23]. The formula to calculate Accuracy is:

$$Accuracy \frac{TP+T}{TP+TN+FP+F} \text{ x } 100 \quad (6)$$

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

Precision

Precision is the level of accuracy between the information requested by the user and the response provided by the system[24]. The formula to calculate Precision is:

$$Precision \frac{TP}{TP+FP} \text{ x } 100 \quad (7)$$

Recall

Recall is the system's ability to successfully retrieve relevant information[25]. The formula to calculate Recall is:

$$recall \frac{TP}{TP+FN} \text{ x } 100 \quad (8)$$

TP : is a prediction result that is positive and correct.
TN : is a prediction result that is negative and correct.
FP : is a prediction result that is positive but incorrect.
FN : is a prediction result that is negative but incorrect.

2.6 Tools and Programming Languange

This research utilizes the Python programming language on the Google Colaboratory platform. Several libraries used include Pandas, Numpy, Matplotlib, and Sklearn for the data mining processes and model evaluation.

3. Results And Discussion

3.1 Data Decription

The data used in this study is the historical data of SPP (Educational Development Contribution) payments by students of MTs. Al-Jabar Bali during the period from 2019 to 2024. The data includes several important attributes that influence payment delays, namely: parents' occupation, parents' monthly income, number of family dependents, and payment delay status (problematic or non-problematic). The classification criteria for payment delays in this study are divided into two categories:

1. Non-Problematic: payment delay is less than 6 months
2. Problematic: payment delay is 6 months or more

3.2 Model Implementation

The implementation process was carried out using the Python programming language with the help of several libraries such as Pandas, NumPy, Sklearn, and Matplotlib on the Google Colaboratory platform. The processed data was then used for training and testing the models using the Decision Tree C4.5 and C5.0 algorithms.

C4.5 Algorithm

The C4.5 model was built using partitioning based on the information gain and entropy values of each attribute. The decision tree structure is formed with nodes representing the attributes and branches representing the outcomes of those attributes. After training and testing, the evaluation results obtained were:

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

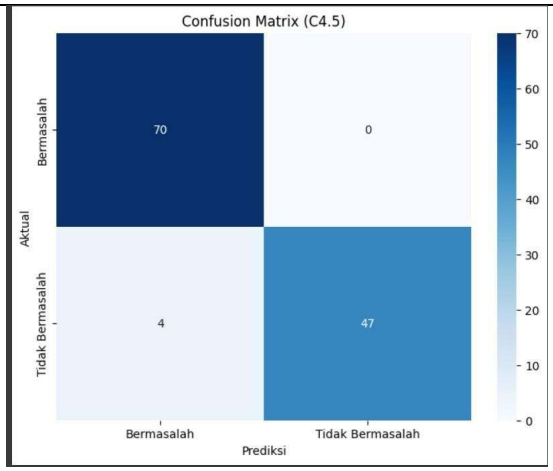

Figure 2. *Confusion Matrix* C4.5

Based on the obtained confusion matrix, it can be concluded that:

1. The correct predictions for problematic data (True Positive) amounted to 70 predictions, while the incorrect predictions (False Positive) totaled 4 predictions.
2. The correct predictions for non-problematic data (True Negative) amounted to 47 predictions, while the incorrect predictions (False Negative) totaled 0 predictions.

C5.0 Algorithm

The C5.0 algorithm is an enhancement of C4.5, employing more efficient pruning and rule generation processes. The C5.0 model produces a simpler and more accurate decision tree structure. The evaluation results of the C5.0 algorithm are:

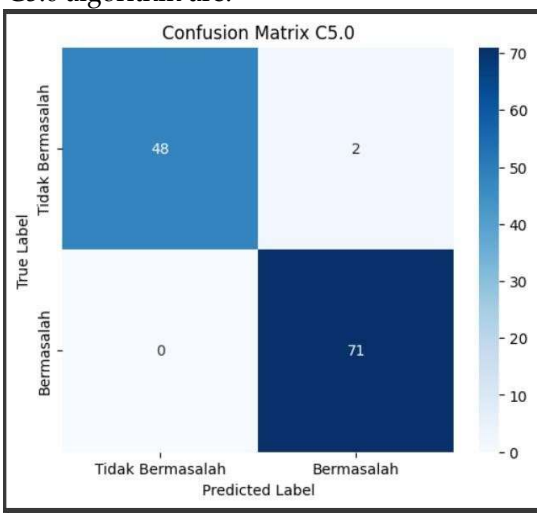

Figure 3. *Confusion Matrix* C5.0

Based on the obtained Confusion Matrix, it can be concluded that:

1. The correct predictions for problematic data (True Positive) amounted to 71 predictions, while the incorrect predictions (False Positive) totaled 2 predictions.
2. The correct predictions for non-problematic data (True Negative) amounted to 48 predictions, while the incorrect predictions (False Negative) totaled 0 predictions.

3.3 Result Comparasion

The analysis results show a significant performance difference between the C4.5 and C5.0 algorithms in predicting tuition payment delays (SPP) at MTs. Al-Jabar Bali. The C4.5 algorithm achieved an accuracy of 96.69%, a precision of 94.59%, and a recall of 100%. Meanwhile, the C5.0 algorithm achieved an accuracy of 98.34%, a precision of 97.26%, and also a recall of 100%. Based on this comparison, the C5.0 algorithm demonstrated a higher accuracy by 1.66% compared to C4.5. In terms

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

of precision, C5.0 also outperformed with a 2.67% higher value than C4.5. As for recall, both algorithms showed equally strong results, reaching 100%, indicating that both are highly effective in detecting problematic data without errors in classifying positive data. Overall, C5.0 is superior in terms of accuracy and precision.

The main factor contributing to these differences lies in the structure and working mechanism of each algorithm. The C5.0 algorithm is an improved version of C4.5 with various enhancements, especially in handling more complex data. C5.0 is capable of processing continuous attributes more effectively, generating more optimal decision trees, and producing more stable predictive models. This capability makes C5.0 more effective in processing historical student payment data, as it can produce models that are more accurate and responsive to data variations compared to C4.5. The improved performance in building decision trees also contributes to the increase in accuracy by 1.66% and precision by 2.67%, resulting in a more reliable prediction model.

The increase in the number of students from year to year can also affect the effectiveness of the prediction model; therefore, regular monitoring and evaluation are necessary to ensure the analysis results remain relevant. Thus, the utilization of the C5.0 algorithm is expected to support more accurate and efficient decision-making in addressing the issue of tuition payment delays (SPP).

Based on the evaluation using accuracy, precision, and recall metrics, it can be concluded that the C5.0 algorithm is more effective and accurate than the C4.5 algorithm. The advantages of C5.0 in producing more stable and precise prediction models make it more suitable for implementation as a prediction tool for SPP payment delays at MTs. Al-Jabar Bali. With higher performance—an increase of 1.66% in accuracy and 2.67% in precision—C5.0 is expected to assist the school in managing and predicting payment data more optimally, thereby supporting more accurate and efficient decision-making. The following table shows the comparison results of the C4.5 and C5.0 algorithms.

Table 1. Comparison Results

	Algoritma C4.5	Algoritma C5.0
Accuracy	96,69%	98,34%
Precision	94,59%	97,26%
Recall	100%	100%

3.4 Discussion

The application of this predictive method provides strategic benefits for the school. By identifying students who are likely to experience payment delays, the school can implement early interventions, such as issuing warnings, designing subsidy programs, or approaching parents directly. In addition, this study proves that data mining has a real contribution in the field of education, especially in terms of financial management. The predictive model built can also be used for other policy analysis needs, such as dropout prediction, learning performance, and so on.

4. Conclusion

Based on the accuracy comparison calculations of the Decision Tree Algorithms C4.5 and C5.0 in predicting tuition payment delays (SPP) at MTs. Al-Jabar Bali, several conclusions can be drawn as follows:

1. The analysis results show that the C4.5 algorithm achieved an accuracy of 96.69%, a precision of 94.59%, and a recall of 100%. Meanwhile, the C5.0 algorithm achieved an accuracy of 98.34%, a precision of 97.26%, and a recall of 100%.
2. The main factor that causes performance differences between the C4.5 and C5.0 algorithms in predicting tuition payment delays at MTs. Al-Jabar Bali lies in the structure and working mechanism of each algorithm. C5.0 is an improved version of C4.5 with several enhancements, including its ability to handle data with continuous attributes. C5.0 can process numerical data more effectively, generate more efficient decision trees, and produce more accurate predictive models. Therefore, C5.0 tends to provide more accurate prediction results, especially when applied to

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

students' historical SPP payment data, as it can produce a more stable and precise model compared to C4.5.

3. Based on the evaluation results using the Accuracy, Precision, and Recall metrics, the C5.0 algorithm is considered more effective and accurate than the C4.5 algorithm. Thus, the C5.0 algorithm is recommended as a more appropriate prediction model to be implemented at MTs. Al-Jabar Bali for predicting tuition payment delays.

Suggestions

Based on the analysis results of the accuracy comparison between the C4.5 and C5.0 Decision Tree Algorithms in predicting tuition payment delays at MTs. Al-Jabar Bali, there are still several improvements that can be considered for future research, such as:

1. Expanding and increasing the dataset used, as larger and more diverse data can improve prediction accuracy and provide more specific results.
2. To enhance the level of detail in the classification process, it is recommended to consider adding additional research attributes.
3. Conducting research on different objects using the C4.5 and C5.0 algorithms to evaluate the strengths and weaknesses of both algorithms in various contexts.

Author Contributions

From the results of the analysis of the calculation of the accuracy of the Decision Tree C4.5 and C5.0 Algorithms in predicting late payment of SPP at MTs. Al-Jabar Bali. For further research, it can be supplemented in the analysis of the calculation of the C4.5 and C5.0 Algorithms, namely: (1) Enlarging and expanding the dataset used, because more and more varied data can increase the accuracy of predictions and provide more specific results. (2) To increase the level of detail in the classification process, you can consider adding additional research attributes. (3) Conducting research with different objects using the C4.5 and C5.0 Algorithms. In order to be able to rotate the advantages and disadvantages of the two algorithms.

References

[1] P. Algoritma, C. Dan, R. Linear, and U. Prediksi, "Perbandingan algoritma c5.0 dan regresi linear untuk prediksi kelulusan mahasiswa," vol. 1, no. 2, pp. 52–59, 2024.

[2] D. Astuti, "Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 60–72, 2019, doi: 10.20895/inista.v1i2.71.

[3] K. A. Dharlie and S. Samanik, "IMAGERY ANALYSIS IN MATSUOKA'S CLOUD OF SPARROWS," *Linguist. Lit. J.*, vol. 2, no. 1, pp. 17–24, Jun. 2021, doi: 10.33365/llj.v2i1.514.

[4] M. Fithratullah, "Representation of Korean Values Sustainability in American Remake Movies," *TEKNOSASTIK*, vol. 19, no. 1, p. 60, Jan. 2021, doi: 10.33365/ts.v19i1.874.

[5] V. S. Ginting, K. Kusrini, and E. Taufiq, "Implementasi Algoritma C4.5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 1, Jun. 2020, doi: 10.35585/inspir.v10i1.2535.

[6] A. S. Huda, "Prediksi Penerimaan Pegawai Baru Dengan Metode Naive Bayes.," 2020.

[7] N. H., Sayekti, L., Studi, P., Informasi, T., Sains, F., Teknologi, D. A. N., Islam, U., & Walisongo, "Analisis sentimen berbasis leksikon terhadap opini mahasiswa tentang kinerja dosen.".

[8] R. P. Kurnia and Y. A. Atma, "ANALISIS REKOMENDASI FILM DARI DATA IMDB MENGGUNAKAN PYTHON," *DEVICE J. Inf. Syst. Comput. Sci. Inf. Technol.*, vol. 3, no. 2, pp. 23–28, Dec. 2022, doi: 10.46576/device.v3i2.2698.

[9] N. W. Wardani,2020. "Penerapan Data Mining Dalam Analytic CRM.

[10] Y. Mertania and D. Amelia, "Black Skin White Mask: Hybrid Identity of the Main Character as Depicted in Tagore's The Home and The World," *Linguist. Lit. J.*, vol. 1, no. 1, pp. 7–12, Jun. 2020, doi: 10.33365/llj.v1i1.233.

[11] S. W. Jannah, A. Muhtadi, M. D. Ridwan, and R. Kurlillah, "Design of an IoT-Based Automatic Weighing System for Catfish Farming to Support Smart Aquaculture," vol. 5, no. 2, pp. 111–118,

http://doi.org/10.26594/register.v8i2.2541

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
7 (3) March 2025 147-154

2024.

[12] Murjani, "Metodelogi Penelitian Kuantitatif,Kualitatif, Dan Ptk. Cross-Borde," 2022, doi: https://journal.iaisambas.ac.id/index.php/Cross- Border/article/view/1141.

[13] I. N. I. Wiradika, "Mobile-Based Decision Support System for Recommending Tourism Attraction Using MAGIQ-ARAS Methodology," vol. 5, no. 2, 2024.

[14] M. A. et al Muslim, "Data Mining Algoritma C4.5 Disertai Contoh Kasus Dan Penerapannya Dengan Program Komputer.," 2019.

[15] I. Mustika, Ardila, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Guntoro, Manuhutu, M. A., Ridwan, M., Hozairi, Wardhani, A. K., Alim, S., Romli, I., Religia, Y., Octafian, D. T., Sufandi, U. U., & Ernawati, "Data Mining dan Aplikasinya," 2021, [Online]. Available: https://repository.penerbitwidina.com/uk/publications/351768/data%0A-mining-dan-aplikasinya%0A

[16] F. U. Zaenal. A, Eka. N, Permata, "ANALISIS PERBANDINGAN ALGORITMA DECISION TREE," 2023.

[17] B. P. Pratiwi, "Pengukuran Kinerja Sistem Kualitas Udara," *J. Inform. UPGRIS, 6(2), 66–75.*, 2020.

[18] D. S. Wardhana, A. W., Patimah, E., Shafarindu, A. I., Siahaan, Y. M., Haekal, B. V., & Prasvita, "Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision Tree.," *Senamika, 2(1), 660–667.*, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/view/1389

[19] A. Kukuh Wahyudi, N. Azizah, and H. Saputro, "DATA MINING KLASIFIKASI KEPRIBADIAN SISWA SMP NEGERI 5 JEPARA MENGGUNAKAN METODE DECISION TREE ALGORITMA C4.5," *J. Inf. Syst. Comput.*, vol. 2, no. 2, pp. 8–13, Dec. 2022, doi: 10.34001/jister.v2i2.392.

[20] W. I. Rahayu, "Regresi Linier Untuk Jumlah Prediksi Jumlah Penjualan Terhadap Jumlah Permintaan.," 2020.

[21] M. Hakimi, M. S. Zarinkhail, H. Ghafory, and S. A. Hamidi, "Revolutionizing Technology Education with Artificial Intelligence and Machine Learning : A Comprehensive Systematic Literature Review," vol. 5, no. 2, pp. 94–110, 2024.

[22] M. Sari, M. S., & Zefri, "Pengaruh Akuntabilitas, Pengetahuan, dan Pengalaman Pegawai Negeri Sipil Beserta Kelompok Masyarakat (Pokmas) Terhadap Kualitas Pengelola Dana Kelurahan Di Lingkungan Kecamatan Langkapura," *J. Ekon. 21(3), 311.*, 2019.

[23] M. Hakimi, Zarinkhail, M.S., Ghafory, H., & Hamidi, S.H. "Revolutionizing Technology Education with Artificial Intelligence and Machine Learning: A Comprehensive Systematic Literature Review," *TIERS Information Technology Journal, 5(7), 94-110.*, 2024, doi: https://doi.org/10.38043/tiers.v6i2.5640.

[24] G.S. Mahendra & I.N.I Wiradika, "Mobile-Based Decision Support System for Recommending Tourism Attraction Using MAGIQ-ARAS Methodology.," *TIERS Information Technology Journal.* 5(2), *119-128.*, 2024. doi: https://doi.org/10.38043/tiers.v5i2.5655

[25] S. W. Janah, A. Muhtadi J., M.D Ridwan, & R. Kurlillah, "Design of an IoT-Based Automatic Weighing System for Catfish Farming to Support Smart Aquaculture" *TIERS Information Technology Journal, 2023. Indones.*, 2024, doi: https://doi.org/10.38043/tiers.v6i2.5633

http://doi.org/10.26594/register.v8i2.2541