Comparison of C4.5 & Random Forest Based on AdaBoost For Determining Loan Eligibility Customer Funds

Lenny[1)*], Violyn[2)], Achmad Ridwan[3)], Yennimar[4)]
[1,2,3,4)]Universitas Prima Indonesia, Indonesia
[1)]lenny6768@gmail.com, [2)]violynnk@gmail.com, [3)] achmadridwan@unprimdn.ac.id,
[4)]yennimar@unprimdn.ac.id

Submitted :Jan 20, 2025 | **Accepted** : Feb 2, 2025 | **Published** : Feb 5, 2025

Abstract: This research discusses the comparison between two data mining algorithms, namely Decision Tree C4.5 and Random Forest based on AdaBoost, in determining the creditworthiness of customer funds. The main objective of this research is to evaluate and compare the performance of the two algorithms in predicting loan eligibility based on customer data. Algorithm performance is measured using accuracy, precision, recall, and misclassification error metrics. The research results show that the AdaBoost-based Random Forest is superior with an accuracy of 78.86%, recall of 98.75%, and the lowest misclassification error of 21.14%. Meanwhile, Decision Tree C4.5 provides lower performance than AdaBoost-based Random Forest. This research recommends further exploration of other algorithms, such as Support Vector Machine (SVM) and Neural Networks, to obtain more optimal results in determining customer loan eligibility.

Keywords: Data mining, Decision Tree C4.5, Random Forest, AdaBoost, Eligibility of Loan Funds, Customers.

INTRODUCTION

Credit financing, also known as loan provision, is the activity of providing loans to individuals based on an agreement between the borrower and the lending institution, with the condition that the borrower must repay the loan within a predetermined period. When selecting potential borrowers, each lending institution conducts several analyses, such as assessing trustworthiness and the risk of delayed loan repayment. This is done to minimize the risk of delays or postponements in loan payments that have been previously agreed upon (Ardiyansyah et al., 2023).

Over time, the number of loan applications from potential borrowers continues to increase, considering that many entrepreneurs require funding to support their businesses. The large number of loan applicants increases the possibility of errors in determining the eligibility of loan recipients by the lending institution. Therefore, in the process of analyzing credit applications, decision-making techniques based on information technology are needed to accelerate the process of determining loan eligibility. Additionally, computational methods are required to analyze the eligibility of potential borrowers in accordance with applicable requirements, ensuring greater accuracy (Sholihaningtias, 2023).

The utilization of information technology in determining loan eligibility can be carried out using data mining techniques (Andi et al., 2023). Data mining is a series of processes or actions aimed at discovering meaningful relationships through patterns and trends in large datasets stored using specific methods or algorithms (Mulyana et al., 2019). The advantage of using information systems and data mining techniques is the ability to determine loan eligibility quickly and accurately. Moreover, data mining allows for predictions involving multiple parameters rather than relying on a single parameter, making the evaluation more comprehensive (Ramayu et al., 2022).

This study conducts a comparative analysis of previously implemented data mining algorithms, namely the Decision Tree C4.5 and Random Forest, in determining loan eligibility. These two algorithms were chosen because, based on a comparative study from previous research, it was concluded that they have higher accuracy compared to other data mining algorithms such as Naïve Bayes and K-Nearest Neighbor. The Random Forest algorithm ranked first with an accuracy rate of 85.67%, while the Decision Tree C4.5 algorithm ranked second with an accuracy rate of 80.33% (Wibisono & Fahrurozi, 2019).

*name of corresponding author

LITERATURE REVIEW

The following section outlines the novelty of this research through a literature review of previous studies related to the research topic. A study titled "Classification of Customer Creditworthiness at Bank XYZ Using C4.5 and Naïve Bayes Algorithm Methods" obtained accuracy results from the C4.5 algorithm across three consecutive tests of 65.75%, 67.70%, and 64.95%, while Naïve Bayes produced accuracies of 64.72%, 66.67%, and 63.40% (Igo et al., 2022). These two algorithms were used to classify the creditworthiness of Bank XYZ customers. This comparison provides an initial understanding of the expected performance of classification algorithms in the context of credit decision-making. A study titled "Evaluation of Decision Tree Models in Creditworthiness Decisions" evaluated the performance of the Decision Tree algorithm in determining loan eligibility, achieving an accuracy of 98%. This evaluation demonstrates the significant potential of using Decision Tree as a tool for predicting loan eligibility with high accuracy (Arnomo et al., 2023). Another study, "Implementation of the Random Forest Classification Algorithm for Creditworthiness Assessment," evaluated the performance of the Random Forest algorithm in determining loan eligibility, obtaining an accuracy of 78.60% (Pahlevi et al., 2023). This study provides an alternative choice in selecting algorithms for predicting creditworthiness by considering the accuracy obtained from Random Forest. Additionally, research titled "Prediction of Non-Performing Loans Using the C4.5 Algorithm at Bank BRI Wonodadi" assessed the performance of the C4.5 algorithm in determining loan eligibility, achieving an accuracy of 89% (Putra, 2024). This result highlights the strong potential of the C4.5 algorithm in credit risk evaluation.

The novelty of this research makes a significant contribution to the development of understanding regarding the application of Decision Tree C4.5 and Random Forest algorithms combined with AdaBoost in determining customer loan eligibility. Unlike previous studies that focused on the performance of each algorithm separately, this research presents a direct comparison that integrates boosting techniques to improve prediction accuracy.

METHOD

In this study, the researcher employs a quantitative research approach, which involves the collection of numerical data and statistical analysis to understand phenomena or answer research questions (Andi et al., 2019). This method is often used to measure relationships between variables and identify patterns or trends in data (Ghodang & Hantono, 2020). In this context, the research focuses on comparing the Decision Tree C4.5 and Random Forest algorithms with AdaBoost in determining customer loan eligibility. In this research, the research methods that will be carried out can be seen in Figure 1.

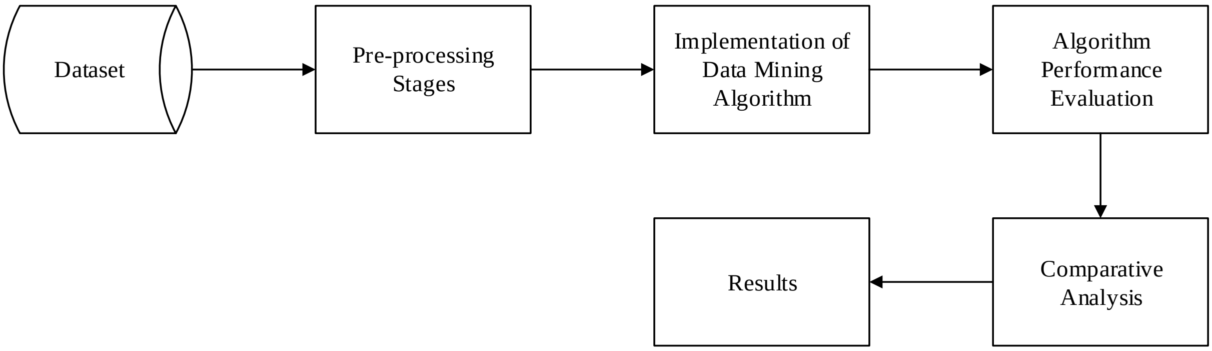

Dataset
Pre-processing Stages
Implementation of Data Mining Algorithm
Algorithm Performance Evaluation
Comparative Analysis
Results

Fig. 1 Research Method Stages

RESULT

The test results obtained in this study were processed using the Python programming language with Google Colab. The aim of this research is to determine the performance of the Decision Tree C4.5 and Random Forest algorithms, based on AdaBoost, in assessing the eligibility of customer loan applications.

Dataset

The dataset used in this study was obtained from the Kaggle website, specifically the Loan Eligibility Dataset, with a total of 614 data entries. This dataset includes various attributes relevant to creditworthiness assessment, such as applicant income, loan amount, credit history, marital status, education level, employment status, and property location. These features serve as key indicators in determining loan eligibility, providing a comprehensive foundation for evaluating the performance of classification algorithms in predicting credit approval outcomes. Table 1 presents the dataset used in this study.

Table 1. Research Dataset

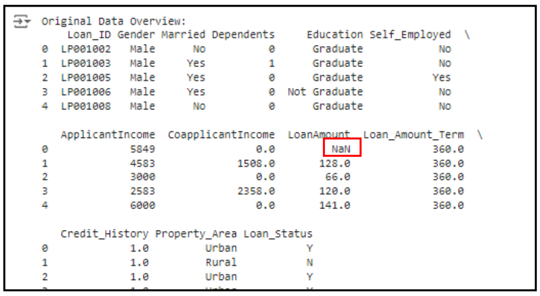
*name of corresponding author

No	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
2	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
3	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
4	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
5	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
…	…	…	…	…	…	…	…	…	…	…	…	…	…
614	LP002990	Female	No	0	Graduate	Yes	4583	0	133	360	0	Semiurban	N

Data Pre-processing Results

The data pre-processing stages in this study include several methods: imputation and normalization. The imputation method is used to handle missing data in the dataset (Karrar, 2022). In this process, missing values in an attribute (column) are replaced using the mean, median, mode, or imputation with a new category. This method helps maintain data consistency and integrity in the dataset (Prasetya & Priyatno, 2023). In this study, categorical attributes were imputed using a new category, while numerical columns were handled using mean, median, or mode imputation. Meanwhile, the normalization method transforms the dataset into a numerical format to facilitate data processing for algorithms. The normalization method applied to numerical data in this study is Z-score Normalization, while categorical data was transformed using One-Hot Encoding, a commonly used technique to convert categorical variables into a format understandable by machine learning algorithms (Sholeh et al., 2022).

Previously, data mining algorithms were applied, and the preprocessing stage was carried out as shown in Figure 2, which represents the dataset before preprocessing, and Figure 3, which shows the dataset after preprocessing.

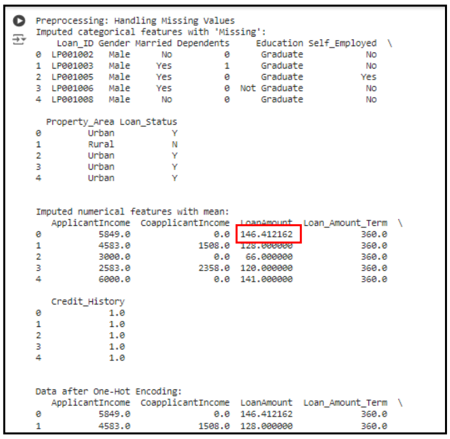


```
Original Data Overview:
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001002   Male      No          0      Graduate            No
1  LP001003   Male     Yes          1      Graduate            No
2  LP001005   Male     Yes          0      Graduate           Yes
3  LP001006   Male     Yes          0  Not Graduate            No
4  LP001008   Male      No          0      Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5849                0.0         NaN             360.0
1             4583             1508.0       128.0             360.0
2             3000                0.0        66.0             360.0
3             2583             2358.0       120.0             360.0
4             6000                0.0       141.0             360.0

   Credit_History Property_Area Loan_Status
0             1.0         Urban           Y
1             1.0         Rural           N
2             1.0         Urban           Y
```

Fig. 2 Dataset Before Pre-processing



```
Preprocessing: Handling Missing Values
Imputed categorical features with 'Missing':
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001002   Male      No          0      Graduate            No
1  LP001003   Male     Yes          1      Graduate            No
2  LP001005   Male     Yes          0      Graduate           Yes
3  LP001006   Male     Yes          0  Not Graduate            No
4  LP001008   Male      No          0      Graduate            No

  Property_Area Loan_Status
0         Urban           Y
1         Rural           N
2         Urban           Y
3         Urban           Y
4         Urban           Y


Imputed numerical features with mean:
   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0           5849.0                0.0  146.412162             360.0
1           4583.0             1508.0  128.000000             360.0
2           3000.0                0.0   66.000000             360.0
3           2583.0             2358.0  120.000000             360.0
4           6000.0                0.0  141.000000             360.0

   Credit_History
0             1.0
1             1.0
2             1.0
3             1.0
4             1.0


Data after One-Hot Encoding:
   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0           5849.0                0.0  146.412162             360.0
1           4583.0             1508.0  128.000000             360.0
```

*name of corresponding author

Fig. 3 Dataset After Pre-processing

In Figure 2, the results show that the "loan amount" attribute contains missing values, highlighted in red. These missing values are then handled through imputation, where they are replaced with the average value, as shown in Figure 3 after preprocessing.

Results of Classification Algorithm Implementation

In this study, the implementation of classification algorithms was carried out to evaluate the performance of the C4.5 Decision Tree and Random Forest algorithms based on AdaBoost in classifying the eligibility of customer loan applications. The data mining algorithm implementation process was carried out using the Google Colab tool with the Python programming language.

Results of Algorithm Performance Evaluation

The evaluation of algorithm performance in this study is conducted using a Confusion Matrix, which is a cross-tabulation of positive and negative class data categorized into predicted and actual classes (Andi et al., 2021). The data is split into training and testing sets, with 80% used for training and 20% for testing. The following Figure 4 shows the results of the performance evaluation of the Decision Tree C4.5 algorithm with AdaBoost and without AdaBoost.

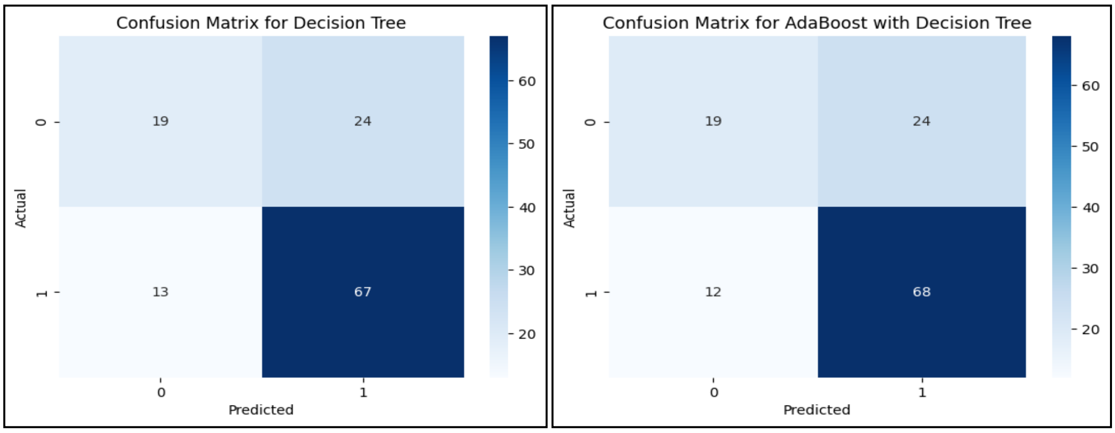


Fig. 4 Confusion Matrix Plot of the Decision Tree C4.5 Algorithm With AdaBoost and Without AdaBoost

Based on Figure 4, it can be seen that the performance of the Decision Tree C4.5 algorithm without AdaBoost achieves a maximum accuracy of 69.92%, a maximum precision of 73.63%, a maximum recall of 83.75%, and a maximum misclassification error of 30.08%. Next, it can be seen that the performance of the Decision Tree C4.5 algorithm based on AdaBoost produces an accuracy of 70.73%, precision of 73.91%, recall of 85.00%, and a misclassification error of 29.27%. The following Figure 5 shows the results of the performance evaluation of the Random Forest algorithm with AdaBoost and without AdaBoost.

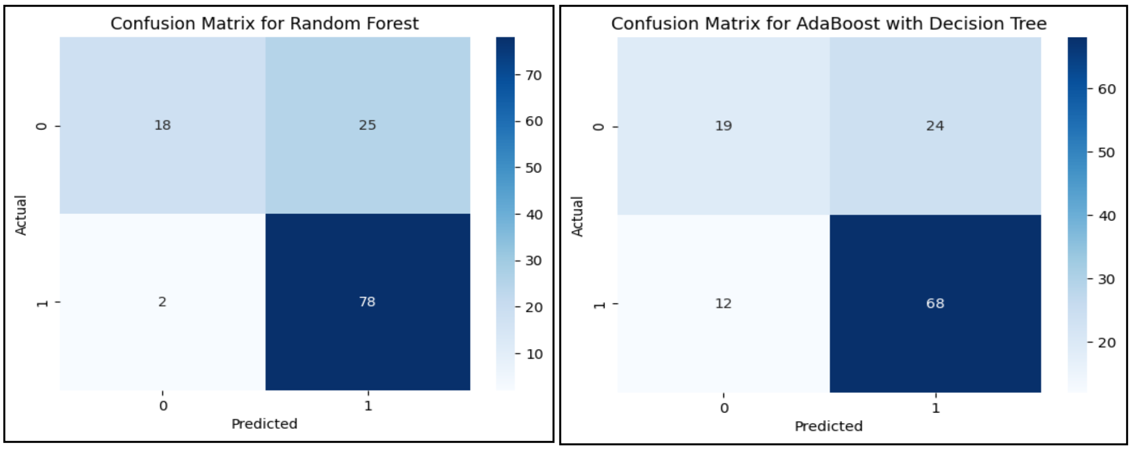


Fig. 5 Confusion Matrix Plot of the Random Forest Algorithm With AdaBoost and Without AdaBoost

*name of corresponding author

Based on Figure 5, it can be seen that the performance of the Random Forest algorithm without AdaBoost achieves a maximum accuracy of 78.05%, a maximum precision of 75.73%, a maximum recall of 97.50%, and a maximum misclassification error of 21.95%. Next, it can be seen that the performance of the Random Forest algorithm based on AdaBoost produces an accuracy of 78.86%, precision of 75.96%, recall of 98.75%, and a misclassification error of 21.14%.

Results of Comparative Analysis

After the performance evaluation process is carried out, the next step is to perform a comparison between the three algorithms tested in this research to determine which algorithm is more effective in predicting loan eligibility. The results of the comparative analysis are presented in Table 2 below.

Table 2. Results of Comparative Analysis of Algorithms

Algorithm	Accuracy		Precision		Recall		Missclassification Error	
	*	**	*	**	*	**	*	**
Decision Tree C4.5	69.92	70.73	73.63	73.91	83.75	85.00	30.08	29.27
Random Forest	78.05	78.86	75.73	75.96	97.50	98.75	21.95	21.14

Information:
* = Without AdaBoost
** = With AdaBoost

Next, the Relative Operating Characteristics (ROC) curve is shown in Figure 6 to provide a more comprehensive illustration of the performance of the algorithm model.

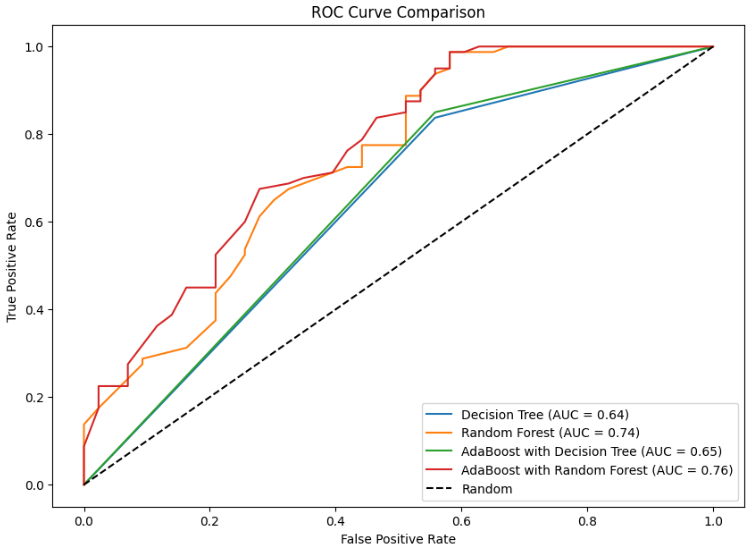


Fig. 6 Receiver Operating Characteristics (ROC) Graphic Results

From the results of the testing in Figure 6, the evaluation using the ROC curve indicates that the Random Forest algorithm based on AdaBoost has a higher accuracy compared to the prediction results of the Decision Tree algorithm based on AdaBoost. The performance of the Random Forest model is significantly better, with an Area Under the Curve (AUC) value of 76.00%.

DISCUSSIONS

The research results show that the Decision Tree C4.5 model without AdaBoost achieved an accuracy of 69.92%, which indicates good but suboptimal performance. The precision was 70.73%, meaning the model effectively predicted positive outcomes, but the recall was 73.63%, indicating that some important positive cases were missed. The misclassification error was 30.08%. When AdaBoost was applied to C4.5, the accuracy increased to 70.73%, indicating a slight improvement, while precision rose to 73.91%. More notably, the recall improved significantly to 85.00%, meaning the model became more effective at detecting positive cases, though the misclassification error dropped to 29.27%.

On the other hand, Random Forest demonstrated better performance. Without AdaBoost, it achieved an accuracy of 78.05%, indicating better classification performance. The precision was 78.86%, showing that it was better at predicting positive cases. The recall was 75.73%, meaning it still missed some positive cases. The misclassification error was 21.95%, reflecting a relatively low proportion of incorrect predictions. With AdaBoost, Random Forest's accuracy increased to 78.86%, while precision slightly decreased to 75.96%. However, the recall dramatically improved to 98.75%, showing the model's excellent ability to detect positive cases. The misclassification error reduced to 21.14%.

In summary, the analysis shows that Random Forest with AdaBoost performed the best, with the highest accuracy, excellent recall, and the lowest misclassification error. This combination was more effective and reliable in predicting loan feasibility compared to Decision Tree C4.5 with or without AdaBoost.

Furthermore, the AUC analysis indicated differences in algorithm performance. The Decision Tree without AdaBoost had an AUC of 0.64, which is satisfactory. However, when AdaBoost was applied, the AUC slightly improved to 0.65. On the other hand, Random Forest without AdaBoost had a better AUC of 0.74, classified as good. With AdaBoost, the AUC increased to 0.76, indicating a significant improvement in performance.

In conclusion, Random Forest with AdaBoost provided the best performance in predicting loan feasibility, with the highest accuracy, excellent recall, and lowest misclassification error. Decision Tree C4.5 with AdaBoost showed some improvement in recall and accuracy but remained in the satisfactory category. Random Forest consistently delivered better results, especially with the addition of AdaBoost, confirming its efficiency and effectiveness in predicting loan feasibility.

CONCLUSION

Based on the research conducted, several conclusions can be drawn. The Decision Tree C4.5 model with AdaBoost achieved an accuracy of 70.73%, precision of 73.91%, recall of 85.00%, and a misclassification error of 29.27%. The increase in recall indicates better capability in detecting positive cases compared to the model without AdaBoost. However, despite the improvement in recall, the misclassification error remained relatively high, showing a significant proportion of incorrect predictions. On the other hand, Random Forest with AdaBoost demonstrated superior performance with an accuracy of 78.86%, precision of 75.96%, a very high recall of 98.75%, and a low misclassification error of 21.14%. The exceptionally high recall reflects its excellent effectiveness in detecting almost all positive cases. The lower classification error indicates that this model is more reliable in predicting loan feasibility with fewer mistakes. Overall, Random Forest with AdaBoost outperformed Decision Tree C4.5 with AdaBoost, proving to be more consistent and effective in predicting loan feasibility.

REFERENCES

Andi, A., Juliandy, C., & David, D. (2023). Clustering Analysis of Tweets About COVID-19 Using the K-Means Algorithm. *Sinkron*, *8*(1), 543–533. https://doi.org/10.33395/sinkron.v8i1.12145

Andi, Juliandy, C., Robet, R., Pribadi, O., & Wijaya, R. (2021). Image Authentication Application with Blockchain to Prevent and Detect Image Plagiarism. *2021 6th International Conference on Informatics and Computing, ICIC 2021*, *December*. https://doi.org/10.1109/ICIC54025.2021.9632966

Andi, Purba, R., & Yunis, R. (2019). Application of Blockchain Technology to Prevent The Potential Of Plagiarism in Scientific Publication. *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*. https://doi.org/10.1109/ICIC47613.2019.8985920

Ardiyansyah, Sa’adah, R., Lisnawanty, & Purwaningtias, D. (2023). Peningkatan Akurasi Metode C4 . 5 Untuk Memprediksi Kelayakan Kredit Berbasis Stratified Sampling Dan Optimize Selection. *Kumpulan Jurnal Ilmu Komputer (KLIK)*, *10*(2), 239–249.

Arnomo, S. A., Fajrin, A. A., Siyamto, Y., Fairuz, S., & Sadikin, N. (2023). Evaluasi Model Decision Tree Pada Keputusan Kelayakan Kredit. *Jurnal Desain Dan Aanlisis Teknologi (JDDAT)*, *2*(2), 200–206.

Ghodang, H., & Hantono. (2020). *Metode Penelitian Kuantitatif Konsep Dasar & Aplikasi Analisis Regresi Dengan Jalur SPSS*. PT. Penerbit Mitra Grup.

Igo, Y. S. R., Aziz, A., & Ahsan, M. (2022). Klasifikasi Kelayakan Pemberian Kredit Nasabah Bank Xyz Menggunakan Metode Algoritma C4.5 Dan Naive Bayes. *Kurawal - Jurnal Teknologi, Informasi Dan Industri*, *5*(1), 57–64. https://doi.org/10.33479/kurawal.v5i1.551

Karrar, A. E. (2022). The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values. *Indonesian Journal of Electrical Engineering and Informatics*, *10*(2), 375–384. https://doi.org/10.52549/ijeei.v10i2.3730

Mulyana, M. A., Religia, Y., & Suwarno, A. (2019). Optimasi Berbasis Particle Swarm Optimization Pada Algoritma Naive Bayes Dalam Memprediksi Penyakit Stroke. *Pelita Teknologi: Jurnal Ilmiah Informatika, Arsitektur Dan Lingkungan*, *14*(1), 1–15.

Pahlevi, O., Amrin, A., & Handrianto, Y. (2023). Implementasi Algoritma Klasifikasi Random Forest Untuk Penilaian Kelayakan Kredit. *Jurnal Infortech*, *5*(1), 71–76. https://doi.org/10.31294/infortech.v5i1.15829

Prasetya, M. R. A., & Priyatno, A. M. (2023). Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining. *Jurnal Informasi Dan Teknologi*, *5*(2), 56–62. https://doi.org/10.37034/jidt.v5i1.324

Putra, S. M. A. (2024). Prediksi Kredit Macet Menggunakan Algoritma C4.5 Di Bank BRI Wonodadi. *Journal Zetroem*, *6*(1), 10–12. https://doi.org/10.36526/ztr.v6i1.3130

Ramayu, I. M. S., Susanto, F., & Mahendra, G. S. (2022). Penerapan Data Mining Dengan Algoritma C4.5 Dalam Pemesanan Obat Guna Meningkatkan Keuntungan Apotek. *Prosiding Seminar Nasional Manajemen, Desain & Aplikasi Bisnis Teknologi (SENADA)*, *5*, 237–245. http://senada.idbbali.ac.id

Sholeh, M., Andayati, D., & Rachmawati, R. Y. (2022). Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes. *TeIKa*, *12*(02), 77–87. https://doi.org/10.36342/teika.v12i02.2911

Sholihaningtias, D. N. (2023). Rekomendasi Kelayakan Penerima Kredit Menggunakan Metode TOPSIS dengan Pembobotan ROC. *Jurnal SAINTEKOM*, *13*(1), 88–99. https://doi.org/10.33020/saintekom.v13i1.376

Wibisono, A. B., & Fahrurozi, A. (2019). Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner. *Jurnal Ilmiah Teknologi Dan Rekayasa*, *24*(3), 161–170. https://doi.org/10.35760/tr.2019.v24i3.2393