YAYASAN AKRAB PEKANBARU
Akrab Juara : Jurnal Ilmu-ilmu Sosial
Volume 10 Nomor 2 Edisi Mei 2025 (477-485)
https://doi.org/10.58487/akrabjuara.v10i2

20

C4.5 ALGORITHM OPTIMIZATION WITH BACKWARD ELIMINATION SELECTION FEATURE FOR CREDITWORTHINESS ASSESSMENT

Untung Rohwadi, Amrin, Rudianto
Dosen Universitas Bina Sarana Informatika
(Naskah diterima: 12 April 2025, disetujui: 28 April 2025)

Abstract

Credit is now a trend in society. Credit problems are the history of incorrect use of credit cards. The impact can cause bad credit. If customers do not pay the debt that has been agreed with the bank, they can increase their credit risk. In this study, researchers applied the C4.5 algorithm without optimization and the C4.5 Algorithm with Backward Elimination Feature Selection Optimization to classify creditworthiness status. Researchers used 481 vehicle credit records with "bad" and "good" reviews. The independent variables used in the study were dependent status, age, last education, marital status, occupation, company status, income, employment status, house condition, length of stay and down payment. From the results of the study and testing, the performance of the C4.5 model without backward elimination for creditworthiness assessment provided a truth accuracy level of 91.90% with an area under the curve (AUC) value of 0.915. While the performance of the C4.5 model with backward elimination provided a truth accuracy level of 94.80% with an area under the curve (AUC) value of 0.973. This proves that optimization with backward elimination can improve the performance of the classification method used.

Keywords: *C4.5, backward elimination, confusion matrix, ROC curva.*

Abstrak

Kredit sekarang menjadi tren di masyarakat. Problem kredit adalah sejarah penggunaan kartu kredit yang salah. Dampak yang ditimbulkan dapat menyebabkan kredit macet. Jika pelanggan tidak membayar utang yang telah disepakati dengan bank, mereka dapat meningkatkan risiko kredit mereka. Dalam penelitian ini, peneliti menerapkan algoritma C4.5 tanpa optimasi dan Algorima C4.5 dengan Optimasi Fitur Seleksi Backward Elimination untuk mengklasifikasikan status kelayakan kredit. Peneliti menggunakan 481 catatan kredit kendaraan dengan ulasan ”bad” dan ”good”. Variabel independen digunakan dalam penelitinan adalah status tanggungan, usia, pendidikan terkahir, status pernikahan, pekerjaan, status perusahaan, pendapatan, status pekerjaan, kondisi rumah, lama tinggal dan uang muka. Dari hasil penelitian dan pengujian, performa model C4.5 tanpa backward elimination untuk penilaian kelayakan kredit memberikan tingkat akurasi kebenaran sebesar 91,90% dengan nilai area under the curva (AUC) sebesar 0,915. Sedangkan Performa model C4.5 dengan backward elimination memberikan tingkat akurasi kebenaran sebesar 94,80% dengan nilai area under the curve (AUC) sebesar 0,973. Hal ini membuktikan bahwa optimasi dengan backward elimination dapat meningkatkan kinerja metode klasifikasi yang digunakan.

Kata kunci: C4.5, backward elimination, confusion matrix, ROC curva.

I. INTRODUCTION

The process of getting a motorcycle loan at this time is very difficult. The finance company must conduct a creditworthiness analysis before providing credit to prospective debtors. The 5C principles, namely (ability), Character Capital (character), (capital), Capacity Collateral (collateral), and Condition (economic conditions), are used to conduct creditworthiness analysis (Aryanto & Widiatno, 2013). In addition, finance companies must also take into account quantitative factors, such as the income and expenses of potential debtors (Lasena & Ahmad, 2023). Most finance companies still conduct a manual creditworthiness analysis, which is prone to errors and subjectivity in decision-making. As a result, a system is needed that can assist finance companies in conducting creditworthiness analysis quickly and accurately (Amrin et al., 2024).

Credit evaluation is becoming increasingly difficult in the world of big data as new information technologies and networks allow for the collection of large amounts of data from various sources (Chern et al., 2021). Managing large, unorganized data sources and ever-changing credit requirements requires sophisticated processing. The credit analysis stage begins to be carried out during the loan or credit process. Due to the large number of documents that come in, it takes a long time to analyze customer data. The principles of Character, Capital, Capacity, Collateral, and Condition (5C) require professional field verification (Lasena & Ahmad, 2023).

Not all consumers or customers are able to pay their credit bills on time every month during credit activities, therefore developing billing problems can affect the company's ability to survive (Alfian & Nugroho, 2024). In addition, credit score data has a high dimension. In addition, redundant or irrelevant features often lead to overfitting the model, thereby interfering with performance (Jin et al., 2021). It is increasingly impossible to ignore the importance of a business's credit rating in the financial sector. This is important because it protects investors' money and reduces the knowledge imbalance that exists between businesses, financial institutions, and society (Song, 2023).

Today, many industries, such as business, banking, and credit unions, use data mining (Oktafriani et al., 2023). The classification model in data mining that functions as a decision model, is usually supported by feature dissection, data resampling, and feature selection methods. The above techniques can be used in a variety of publications. When a subset of relevant features is selected, the computational load can be reduced and the efficiency and understanding of the model can be improved (Ziemba et al., 2021). The stages of identifying and predicting customers properly and correctly can be done before

the loan process by checking the customer's loan historical data. This activity is an effort made by the banking industry at this time in dealing with credit risk problems (Amrin & Pahlevi, 2022).

The purpose of this study is to apply the classification algorithm C4.5 without optimization and the C4.5 Algorima with Optimization of the Backward Elimination Selection Feature in the assessment of creditworthiness. Another goal of this study is to create a classification model that can help very accurately in determining creditworthiness. The study also evaluates how the C4.5 algorithm without optimization and the C4.5 algorithm with optimization work in the creditworthiness assessment. As a result, this research is expected to help build a more efficient credit scoring system.

Several previous studies related to the theme of methods that have been used to solve creditworthiness predictions, including research conducted by (Religia et al., 2021), this study analyzed the Comparison of Optimization Algorithms in Random Forest for Bank Marketing Data Classification. The test results show that the use of GA or Bagging optimization has not been able to increase the accuracy of the RF algorithm for the classification of Bank Marketing datasets, where by using optimization or not the accuracy obtained is 88.30%. The optimization carried out has not been successful because the data used is too imbalanced, so that when the bagging or GA process is carried out, the distribution of the data produced is still quite large. Furthermore, the research conducted by (Putry et al., 2024), this research is about Decision Tree Optimization for Bank DKI's KMG Credit Risk. The result of this study is the optimization of Particle Swarm Optimization.

Next research by (Ubaedi & Djaksana, 2022), this study discusses C4.5 Algorithm Optimization Using Forward Selection and Stratified Sampling Methods for Credit Eligibility Prediction. The conclusion of this study is that the C4.5 Algorithm has been shown to be effective in predicting creditworthiness with an accuracy rate of 79.11%. The Forward Selection and Stratified Sampling methods have been proven to be successful in increasing the accuracy of the C4.5 algorithm by 9.2% in predicting creditworthiness. Then research by (Novichasari, 2021) which discusses Improving Credit Eligibility Accuracy Using Particle Swarm Optimization. The result of this study was that the accuracy of each classification technique increased when combined with PSO on attribute weighting. The SVM-PSO model has the best accuracy and precision among other techniques. The next previous research conducted by (Agustian & Bisri, 2019), this study discusses the Min Data.

II. THEORETICAL STUDIES

Decision Tree

Decision trees are analytical techniques applied to solve decision-making issues by describing various options and possible outcomes. This method breaks down the complicated process into simpler parts, making it easier for decision-makers to assess their options. Below is a thorough explanation of decision tree theory. Definition of Decision Tree. The decision tree is a visual representation of the various choices and the impact that arises from those choices. It shows the decision-making structure in a tree format with branches that reflect the options available, as well as the nodes that symbolize decisions or events (Alamanda, Cahyani, Novianti, Hilyah, & Haryono, 2024).

Feature Selection

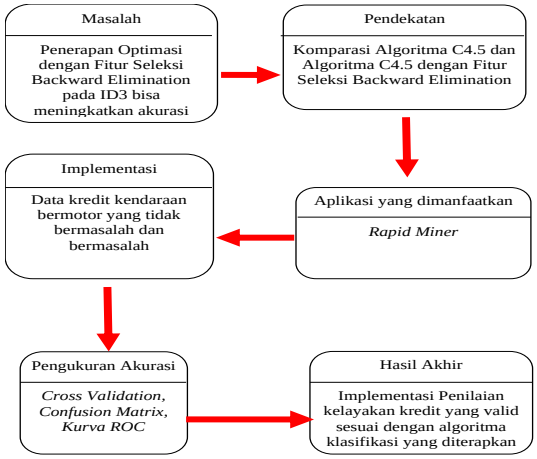
Feature selection is a step in data analysis that aims to select the most significant set of features from existing data, to improve predictive model performance and reduce complexity. This process also plays a role in reducing overfitting and improving understanding of models (Huizen, Ardima, & Idris, 2025).

Credit Eligibility

Creditworthiness is a method used to evaluate a borrower's ability to pay off debts. This concept pays attention to various elements, such as the nature of the borrower, ability, capital, situation, and collateral, known as the 5C principle (Widagda & Primantari, 2025).

III. RESEARCH METHODS

This research consists of several stages as seen in the frame of thought in Figure 1. The problem in this study is whether the application of optimization with the Backward Elimination Selection Feature in C4.5 can improve accuracy? For this reason, an approach (model) was made, namely the C4.5 algorithm and the C4.5 Algorithm with the Backward Elimination Selection Feature to solve the problem and then test the performance of the method. The test uses the Cross Validation, Confusion Matrix and ROC curve methods. To develop applications (development) based on the model created, Rapid Miner is used.



Source: Research Results (2025)

Figure 1 Research Framework

IV. RESULTS AND DISCUSSION

Data Analysis

The dataset used was 481 motor vehicle credit data, both problematic and non-problematic. The input variables in this study consisted of thirteen variables, namely: 1) Marital Status, 2) Number of Dependents, 3) Age 4) Residence Status, 5) Home Ownership, 6) Occupation, 7) Employment Status, 8) Company Status, 9) Income, 10) Down Payment, 11) Education, 12) Length of Residence, and 13) Home Condition. Figure 1 shows the sample dataset used to test the algorithm to be tested.

status perkawinan	jumlah tanggungan	pendidikan terakhir	usia	kepemilikan rumah	lama tinggal	kondisi rumah
menikah	2-3	SLTP	21-55	ortu	>5	permanen
belum menikah	tdk ada	SLTA	21-55	ortu	1-3	permanen
menikah	2-3	SLTA	21-55	KPR	3-5	permanen
belum menikah	tdk ada	SLTA	<21/>60	ortu	1-3	permanen
menikah	>3	SD	21-55	atas nama	>5	non permanen
menikah	2-3	SLTP	21-55	milik sendiri	>5	permanen
menikah	2-3	SLTA	21-55	milik sendiri	>5	permanen

Figure 2 Sample Dataset

Model Testing This research was carried out by testing experiments on the proposed model. Then the model was evaluated and validated to produce accuracy and AUC values. The test uses Rapidminer with a 10-fold cross-validation operator to obtain accuracy and AUC results on each algorithm tested. The evaluation carried out was with the Confusion Matrix and ROC Curve or Area Under Curve (AUC).

1. Confusion Matrix

A. C4.5 Algorithm

Table 1 is the confusion matrix for the C4.5 algorithm. It is known that 152 data classified as "bad" were predicted according to the actual data, then 27 data were predicted

to be "good" but turned out to be "bad". Then 290 data classified as "good" were predicted to be appropriate, and 12 data predicted "bad" turned out to be “good”.

Table 1 Model Confusion Matrix untuk Algoritma C4.5

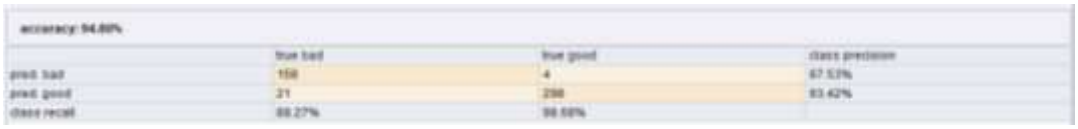
accuracy: 91.90% +/- 4.50% (mikro: 91.89%)			
	true bad	true good	class precision
pred. bad	152	12	92.68%
pred. good	27	290	91.48%
class recall	84.92%	96.03%	

Source: Processing Results Using RapidMiner (2025)

C4.5 Algorithm with Backward Elimination

Table 2 is the confusion matrix for the C4.5 algorithm with Backward Elimination. It is known that 298 data classified as "good" were predicted according to the actual data, then 4 data were predicted to be "bad" but turned out to be "good". Then 158 data classified as "bad" were predicted to be appropriate, and 21 data predicted "good" turned out to be "bad”.

Table 2 Model Confusion Matrix untuk Metode C4.5 dengan Backward Elimination



Source: Processing Results Using RapidMiner (2025)

2. Kurva ROC

a. Algoritma C4.5

ROC curve for the random forest algorithm as shown by figure 3 below.

Source: Processing Results Using RapidMiner (2025)
Figure 3 ROC Curve of C4.5 Algorithm

The ROC curve in figure 2 expresses the confusion matrix. Horizontal lines are false positives and vertical lines are true positives. C4.5 algorithm with Backward Elimination

The ROC curve for the C4.5 algorithm with Backward Elimination as shown by figure 4 below.

Source: Processing Results Using RapidMiner (2025) Figure 4 ROC Curve of the C4.5 Method with Backward Elimination.

Analyse des résultats

The comparison of accuracy, precision, recall and AUC values for the C4.5 and C4.5 algorithms with Bacward Elimination is shown by table 3 below.

Table 3 Model Evaluation and Validation

Algoritma	Accuracy	*Precision*	*Recall*	*AUC*
C4.5	91.90%	91.69%	96.03%	0.915
C4.5 dengan Backward Elimination	**94.80%**	93.42%	98.68%	**0.973**

Source: Processing Results Using RapidMiner (2023)

Table 3 compares the accuracy, precision, recall and AUC of the C4.5 algorithm without backward elimination and C4.5 with backward elimination. It can be seen that the accuracy value of the C4.5 algorithm with backward elimination has a higher accuracy compared to the C4.5 algorithm without backward elimination. Likewise with the AUC value, it is important to note that the AUC value of the C4.5 algorithm with backward elimination has a higher value compared to the C4.5 algorithm without backward elimination. For the classification of mining data, the AUC value can be divided into several groups [20].

a. 0.90-1.00 = Excellent classification
b. 0.80-0.90 = Good classification
c. 0.70-0.80 = La classification est suffisante
d. 0.60-0.70 = Bad classification
e. 0.50-0.60 = Misclassification

Based on the grouping above, it can be concluded that the C4.5 and C4.5 algorithm models with backward elimination are included in the classification category very well.

V. CONCLUSION

From the results of research and testing, the performance of the C4.5 model without backward elimination for creditworthiness assessment provides a correctness rate of 91.90% with an area value under the curve (AUC) of 0.915. Meanwhile, the performance of the C4.5 model with backward elimination provides a correctness rate of 94.80% with an area value under the curve (AUC) of 0.973. This proves that optimization with backward elimination can improve the performance of the classification method used.

REFERENCE

Agustian, A. A., & Bisri, A. (2019). Data Mining Optimization Using Sample Bootstrapping and Particle Swarm Optimization in the Credit Approval Classification. *Indonesian Journal of Artificial Intelligence and Data Mining*, *2*(1), 18–27. https://doi.org/10.24014/ijaidm.v2i1.6299

Alamanda, D. T., Cahyani, V., Novianti, D. N., Hilyah, A., & Haryono, Z. (2024). Aplikasi Game Theory Dalam Bisnis (Analisis Permainan Menggunakan Permainan). Kab. Sumedang: CV. Mega Press Nusantara.

Alfian, A. B., & Nugroho, A. H. D. (2024). Analisis Sistem Pengendalian Internal Terhadap Efektivitas Pemberian Kredit Kendaraan Bermotor di PT. Indomobil Finance Indonesia Cabang Semarang. *Journal of Economic, Bussines and Accounting (COSTING)*, *7*(4), 9071–9084. https://doi.org/10.31539/costing.v7i4.8902

Amrin, A.-, & Pahlevi, O.-. (2022). Implementation of Logistic Regression Classification Algorithm and Support Vector Machine for Credit Eligibility Prediction. *Journal of Informatics and Telecommunication Engineering*, *5*(2), 433–441. https://doi.org/10.31289/jite.v5i2.6220

Amrin, A., Rudianto, R., & Irfiani, E. (2024). Analisis Algoritma Iterative Dichotomiser 3 (ID3) untuk Penilaian Kelayakan Kredit Kendaraan Bermotor. *IMTechno: Journal of Industrial Management and Technology*, *5*(2), 36–40.

Aryanto, R., & Widiatno, A. (2013). Prioritas Alternatif Keputusan Pada Analisis Kredit Motor. *Binus Business Review*, *4*(9), 316–321.

Chern, C.-C., Lei, W.-U., Huang, K.-L., & Chen, S.-Y. (2021). A decision tree classifier for credit assessment problems in big data environments. *Information Systems and E-Business Management*, *19*(1), 363–386. https://doi.org/10.1007/s10257-021-00511-w

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Springer.

Huizen, L. M., Ardima, M. B., & Idris, M. (2025). Meningkatkan kinerja SVM: Dampak berbagai teknik seleksi fitur pada akurasi prediksi. AITI: Jurnal Teknologi Informasi, 1–14

Jin, Y., Zhang, W., Wu, X., Liu, Y., & Hu, Z. (2021). A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data. *IEEE Access*, *9*, 143593–143607. https://doi.org/10.1109/ACCESS.2021.3120086

Lasena, M., & Ahmad, S. R. (2023). Sistem Pendukung Keputusan Kelayakan Pemberian Kredit Nasabah Dengan Metode Electre. *Bulletin of Information Technology (BIT)*, *4*(2), 232–238. https://doi.org/10.47065/bit.v4i2.690

Novichasari, S. I. (2021). Peningkatan Akurasi Kelayakan Kredit Menggunakan Particle Swarm Optimization. *Jurnal Multimatrix*, *3*(1), 86–90. https://jurnal.unw.ac.id/index.php/mm/article/view/1533%0Ahttps://jurnal.unw.ac.id/index.php/mm/article/view/1533/992

Oktafriani, Y., Firmansyah, G., Tjahjono, B., & Widodo, A. M. (2023). Analysis of Data Mining Applications for Determining Credit Eligibility Using Classification Algorithms C4.5, Naïve Bayes, K-NN, and Random Forest. *Asian Journal of Social and Humanities*, *1*(12), 1139–1158. https://doi.org/10.59888/ajosh.v1i12.119

Putry, J. B. E., Sasongko, A. T., & Hadikristanto, W. (2024). Optimasi Decision Tree Menggunakan Particle Swarm Optimization (PSO) pada Risiko Kredit KMG Bank DKI. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1403–1410. https://doi.org/10.57152/malcom.v4i4.1521

Racmatl, R., & Likliwati, R. D. (2019). Optimasi k-Nearest Neighbor Dengan Particle Swarm Optimization Pada Klasifikasi Nasabah Kredit Kendaraan. *Jurnal JTRISTE*, *6*(1), 9–16. https://jurnal.kharisma.ac.id/jtriste/article/view/89

Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 187–192. https://doi.org/10.29207/resti.v5i1.2813

Song, Y. (2023). Enterprise Credit Rating Prediction Model Based on Data Mining Algorithm. In J. C. Hung, J.-W. Chang, & Y. Pei (Eds.), *Innovative Computing Vol 1 - Emerging Topics in Artificial Intelligence* (pp. 745–751). Springer Nature Singapore.

Widagda, I. N., & Primantari, A. A. (2025). Penyelesaian Wanprestasi Dalam Perjanjian Kredit Bank Tanpa Agunan Melalui Perspektif Hukum Positif Di Indonesia. JMA: Jurnal Media Akademik, 1-16

Ziemba, P., Becker, J., Becker, A., Radomska-Zalas, A., Pawluk, M., & Wierzba, D. (2021). Credit decision support based on real set of cash loans using integrated machine learning algorithms. *Electronics (Switzerland)*, *10*(17), 1–22. https://doi.org/10.3390/electronics10172099