Improved C45 performance with gain ratio for credit approval dataset

Ivandari [1], M. Adib Al Karomi [2], Much. Rifqi Maulana [3]

[1,2,3)] Computer Science, STMIK Widya Pratama Pekalongan, Indonesia

Abstract— People's shopping behavior has undergone many changes after the COVID-19 pandemic. Many people have switched to using the marketplace to make buying and selling transactions. The payment process in the marketplace is relatively easy, especially when using a credit card. The increase in demand for credit must be addressed better by financial providers to minimize bad loans. The best thing in minimizing bad credit is to be more selective in choosing credit customers. Data mining is a field that can study old data to become new knowledge in the future. In data mining, the classification of bad credit customers is mostly done. One of the algorithms that excels in handling credit approval datasets is C45. The C45 model is widely used because it has an output decision tree that is easier to understand in human language. The number of data attributes can affect the performance of the algorithm. Feature selection is a form of attribute reduction to improve data quality and improve classification algorithm performance. Gain ratio is the development of information gain and is the best feature selection model and is widely used by researchers. This study performs a classification using C45 and uses a gain ratio for the selection of credit approval data features. By using the gain ratio, the accuracy of the C45 classification algorithm increased from the previous 94.12% to 95.29%.

Keywords— Decision tree, information gain ratio, accuracy.

1. Introduction

The COVID-19 pandemic gave rise to many new behaviors in human life. In addition to changing social behavior, economic behavior has also changed a lot due to social restrictions in society. Some of the economic behavior that has changed due to the pandemic is people's spending habits. The existence of social restrictions in society has led to an increase in online buying and selling transactions [1]. The increase in the number of online transactions during 2020 has an impact on the trend of credit provider data. The credit provider banks must think of new ways to minimize bad loans. In fact, banks can use more than one model to decide on the credit approval of their customers [2]. In addition to using traditional methods, data mining is also widely used for credit approval classification [3].

Data mining is the study of data so that it can generate new knowledge. Data mining is widely used for classification [4] [5]. The classification process can use many algorithms, one of the best classification algorithms is C45 which can produce decision tree outputs [6]. C45 is proven to be able to handle numeric or nominal type [7]. One of the advantages of the C45 output is that it can be more easily understood in human language [3]. In the calculation process, the C45 algorithm uses the gain value to calculate the importance of each data attribute used. The attribute with the highest gain value will be used for the first node and so on until all data attributes are used up for the other nodes.

Credit approval classification using the C45 algorithm has been done [3]. In this study, a decision support system was made with an accuracy rate of 94.12%. In addition, several studies were also conducted to improve the accuracy of credit approval classification using the information gain method [8]. In its development, many improvements to the information gain method have been carried out. One of the most prominent improvements to the information gain method is the information gain ratio method [9]. In the information gain ratio method, the split information value is used to divide the information gain value. The information gain raio process is proven to improve the performance of the classification algorithm [10].

This study uses the information gain ratio method for the selection of credit customer data features. The data used is credit card customer data with a total of 14 regular attributes and 1 label attribute. This dataset has 766 records. After calculating the information gain ratio, the threshold value is set to be 0.21. The classification process is carried out using the C45 algorithm. Validation in the classification is carried out using 10 folds cross validation and using a confusion matrix for the evaluation process. The classification process was carried out and the results obtained an accuracy rate of 95.29%. Without using the information gain ratio, the accuracy rate is only 94.12%. In fact, the information gain ratio can improve the classification performance of the C45 algorithm by 1.17%.

2. Literature Review

2.1. Related Research

Research with the theme of credit approval has been carried out with various results. In the previous study [8] *used the credit approval dataset and information gain feature selection. This research uses the K-NN classification algorithm. In this study, the best accuracy was obtained using K-NN and information gain with an accuracy rate of 94.78%. Currently the development of information gain is found by dividing the split information and is known as the gain ratio. Classification research using C45 is also widely carried out. One of them is to classify the Covid 19 surveillance dataset [7]. The advantage of C45 is that it has an output decision tree that can be easily understood by human language. C45 is also one of the best classification algorithms and is widely recommended by international researchers [11].*

2.2. Data Mining

Data mining is a science that focuses on existing computerized datasets or records [12]. The current use of digital media clearly enriches existing digital data. The amount of data that has no meaning will only become digital waste that makes our data storage media full. The existence of data mining is very helpful in processing data so that it becomes a new knowledge. In data mining there are various main functions. Such as Estimation, Prediction, Clustering, Association and Classification. Classification is one of the main functions and is widely used because it can handle numeric and nominal data. Various algorithm models are the mainstay in the classification process. One of the most popular and proven to have good performance is the C45.

3. Research Methods

This study uses an experimental method. The experiment was carried out using existing datasets and using a rapid miner as a calculation tool. Figure 1 is the research method carried out. The stages of the research carried out are as follows:

3.1. Data collection

The data collection process is carried out using a credit approval dataset from banks. This dataset is classified as private data from the use of credit cards in one bank. The credit card usage data used has 766 records and 14 regular attributes and 1 label attribute. Table 1 is the metadata of the credit approval dataset.

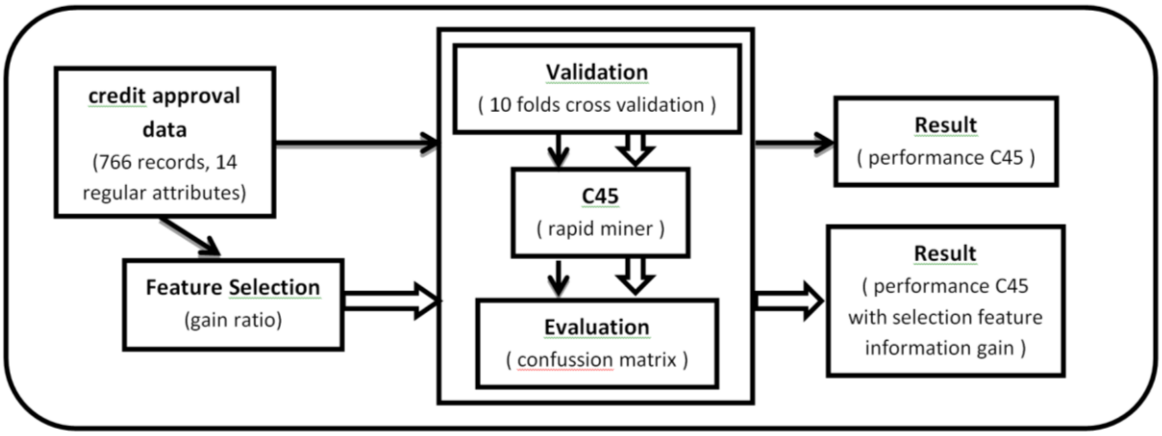


Figure 1. Research flow chart

Table 1. Credit approval metadata

Role	Name	Type	Statistics	Range	Missings
id	nama_nasabah	polynominal	mode = x1 (1), least = x1 (1)	Unique…	0.0
label	status kredit	binominal	mode = MACET (556), least = LANCAR (210)	MACET (556), LANCAR (210)	0.0
regular	jenis_kelamin	binominal	mode = P (462), least = L (304)	P (462), L (304)	0.0
regular	umur	integer	avg = 29.161 +/- 263.166	[-7162.000 ; 1043.000]	1.0
regular	jml_pinjaman	numeric	avg = 2712482.631 +/- 9995602.067	[83333.330 ; 228655000.000]	0.0
regular	jkw	integer	avg = 18.961 +/- 32.076	[1.000 ; 679.000]	0.0
regular	jml_angsuran_per_bulan	numeric	avg =	[0.000 ; 10350000.000]	0.0

			233391.702 +/- 548968.221		
regular	type_pinjaman	integer	avg = 100 +/- 0	[100.000 ; 100.000]	0.0
regular	jenis_pinjaman	integer	avg = 301.197 +/- 0.822	[301.000 ; 305.000]	0.0
regular	bi_sektor_ekonomi	integer	avg = 6013.046 +/- 216.196	[6000.000 ; 9990.000]	1.0
regular	col	integer	avg = 1.217 +/- 0.412	[1.000 ; 2.000]	0.0
regular	bi_golongan_debitur	polynominal	mode = 874 (757), least = 834 (1)	874 (757), 876 (8), 834 (1)	0.0
regular	bi_gol_penjamin	polynominal	mode = 000 (519), least = 835 (1)	875 (229), 000 (519), 800 (8), 874 (9), 835 (1)	0.0
regular	saldo_nominatif	numeric	avg = 2007385.712 +/- 8711282.360	[-4000000.000 ; 209404092.000]	0.0
regular	tunggakan_pokok	numeric	avg = 790085.298 +/- 4139216.644	[0.000 ; 91612122.240]	0.0
regular	tunggakan_bunga	numeric	avg = 87717.084 +/- 568231.776	[0.000 ; 11000000.000]	0.0

3. 2. *Feature selection*

The first process after data collection is feature selection. Feature selection in this study uses the gain ratio [9]. The gain ratio is proven to improve the performance of the classification algorithm [10]. The feature selection process is used to determine how high the influence of the attribute in the classification is according to the gain value of the attribute. Furthermore, some attributes that are considered to have no effect or have low gain will be set aside and not used in the classification process. The success of the feature selection process using the gain ratio is also influenced by the threshold used.

3. 3. *Validation*

The validation process is carried out using 10 folds cross validation. This process is used to ensure that all data records have participated in training and testing data. The process of this validation process uses a rapid miner application. Figure 2 is the running process of the rapid miner. Figure 3 is a validation display using cross validation in rapid miner.

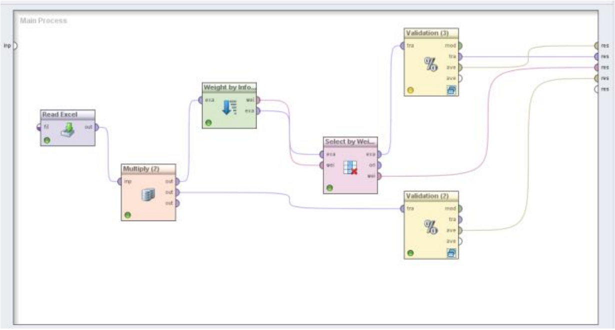


Figure 2. Research process in rapid miner

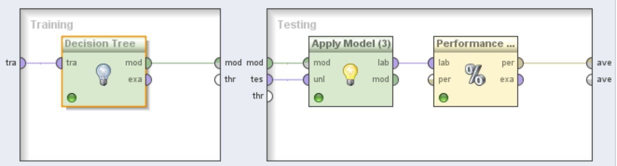


Figure 3. Validation process

3. 4. *Algorithm calculation and result evaluation*

Algorithm calculation is done by rapid miner. The process that is followed is as shown in Figure 2. In the image there is one data which is then separated into 2 using the multiply tools. One of them is the C45 calculation process using the original dataset. Next, the dataset is selected for features before C45 classification. Furthermore, the results of the accuracy of the two are compared. The comparison is done by using confusion matrix.

4. Results and Discussion

4. 1. *Results*

The gain ratio is used to determine the importance value of each attribute in the credit approval dataset. Table 2 is the result of calculating the gain ratio using a rapid miner. From these results, it is known that the ***tunggakan pokok*** attribute has the most influence in the classification with the highest importance value. While ***tipe pinjaman*** type attribute is the attribute with the lowest importance value.

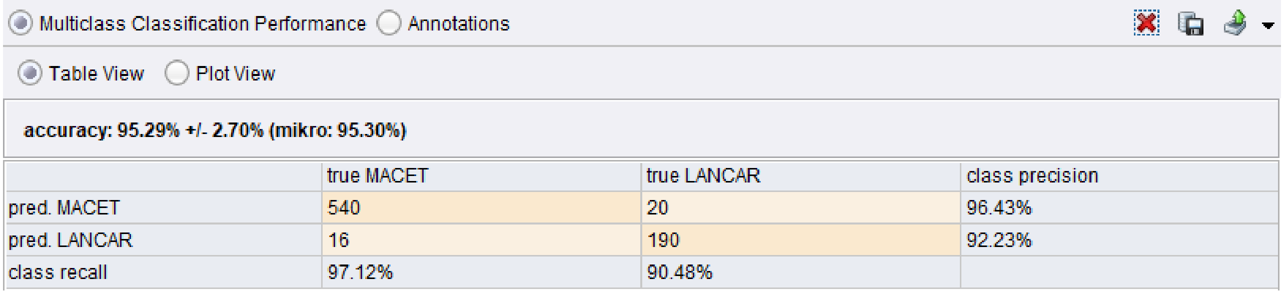
Table 2. Results of gain ratio

Atribut name	Weight by gain ratio
type_pinjaman	0.0
jenis_kelamin	0.004803862099373672
jenis_pinjaman	0.011990558155785886
bi_golongan_debitur	0.026122029885853118
jkw	0.07390805881361022
umur	0.07855777463435319
bi_sektor_ekonomi	0.08224417989421196
bi_gol_penjamin	0.1579360800243115
jml_pinjaman	0.16744927994578124
col	0.20724024788144652
tunggakan_bunga	0.2124907036888866
jml_angsuran_per_bulan	0.27133653191439594
saldo_nominatif	0.3329244527390554
tunggakan_pokok	1.0

The results showed an increase in the performance of the C45 algorithm by selecting features using the gain ratio. By using the gain ratio, the accuracy of the C45 algorithm increases to 95.29%. Previously, the accuracy of the C45 algorithm without using a gain ratio was 94.12%. This 1.17% increase in performance occurs by using a threshold value of 0.21 in the gain ratio. Figure 4 is the performance of the C45 algorithm with the gain ratio feature selection.

4.2. Discussion

This study uses an experimental method by trying every possibility that exists. From table 2 which has described the importance of all attributes using the gain ratio. From this gain ratio value, threshold can be taken for the classification process using C45. This threshold is used as a limit for attributes that will be used or left in the next classification process using C45. Table 3 is the overall result of the C45 classification using the threshold in accordance with the previous gain ratio results.



accuracy: 95.29% +/- 2.70% (mikro: 95.30%)

	true MACET	true LANCAR	class precision
pred. MACET	540	20	96.43%
pred. LANCAR	16	190	92.23%
class recall	97.12%	90.48%	

Figure 4. C45 accuracy results

Table 3. Experimental results of C45 and gain ratio

threshold	Attribute used	C45 accuracy	Increase/ decrease in accuracy
0	14	94,12 %	0
0,002	13	94,19 %	+ 0,07 %
0,006	12	94,25 %	+ 0,13 %
0,015	11	94,25 %	+ 0,13 %
0,03	10	94,25 %	+ 0,13 %
0,075	9	94,65 %	+ 0,53 %
0,08	8	94,39 %	+ 0,27 %
0,085	7	94,39 %	+ 0,27 %
0,16	6	94,91 %	+ 0,79 %
0,17	5	95,16 %	+ 1,04 %
0,21	4	95,29 %	**+ 1,17 %**
0,22	3	93,34 %	-0,78 %
0,28	2	93,08 %	-1,04 %
0,5	1	91,38 %	-2,74 %

Table 3 explains the increase in the accuracy of C45 by using the gain ratio. The greatest increase in accuracy occurs by using a threshold of 0.21. By using a threshold of 0.21 it means that only the best 4 attributes are used in the C45 classification process. The use of the gain ratio does not always improve the performance and accuracy of the algorithm. In fact, using a gain ratio with a threshold of 0.22 or more than 0.22 can reduce the accuracy of C45.

5. Conclusion

This study shows an increase in the accuracy of the C45 algorithm by adding a gain ratio feature selection for the credit approval dataset. The highest increase in the accuracy of the C45 algorithm, which is 1.17%, occurs using a threshold value of 0.21.

References

[1] B. J. G. Rozo, J. Crook, and G. Andreeva, “The role of web browsing in credit risk prediction,” *Decis. Support Syst.*, p. 113879, 2022, doi: 10.1016/j.dss.2022.113879.

[2] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, “Generating customer’s credit behavior with deep generative models,” *Knowledge-Based Syst.*, vol. 245, p. 108568, 2022, doi:

10.1016/j.knosys.2022.108568.

[3] M. R. Maulana and M. A. Al Karomi, “Sistem Pendukung Keputusan Persetujuan Kredit Menggunakan Algoritma C4.5,” *J. IC-Tech*, vol. Vol. XI No, no. 1, pp. 29–38, 2016, [Online]. Available: http://jurnal.stmik-wp.ac.id/gdl.php?mod=browse&op=read&id=ictech--muchrifqim-80.

[4] Ivandari and M. A. Al Karomi, “Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance,” *IC Tech*, vol. 16, no. 1, pp. 12–15, 2021, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137.

[5] M. A. Al Karomi, M. R. Maulana, S. J. Prasetiyono, Ivandari, and Arochman, “Strengthening campus finance by analyzing attribute attributes for student registration classifications.” p. 1, 2019, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/1431.

[6] V. K. Xindong Wu, *The Top Ten Algorithm in Data Mining*. 2009.

[7] Ivandari and M. A. Al Karomi, “Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm,” *Jaict*, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.

[8] Ivandari, T. T. Chasanah, S. W. Binabar, and M. A. Al Karomi, “Data Attribute Selection with Information Gain to Improve Credit Approval Classification Performance using K-Nearest Neighbor Algorithm,” *IJIBEC*, vol. I, pp. 15–24, 2017.

[9] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection using Gain Ratio and Correlation Based Feature Selection,” *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[10] I. Indrayanti, S. Devi, and M. A. Al Karomi, “Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus,” *IC-TECH*, vol. XIII, no. 2, pp. 1–6, 2017, [Online]. Available: ejournal.stmik-wp.ac.id.

[11] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2007.

[12] O. Maimoon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, vol. 40, no. 6. Springer, 2010.