EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol3.iss2.art2

Research Article

Comparison of the Naïve Bayes Classifier and Decision Tree J48 for Credit Classification of Bank Customers

Alifia Tanza[1], Dina Tri Utari[1,*]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Jl. Kaliurang Km 14.5, Yogyakarta

[*] *Corresponding author: dina.t.utari@uii.ac.id*

Received: 13 July 2022; Accepted: 22 August 2022; Published: 29 August 2022

Abstract: The bank conducts an analysis or survey in the credit system to determine whether the customer is eligible to receive credit. With a case study of Bank BJB debtor data in December 2021, credit classification analysis was carried out by forming a model using the Naïve Bayes Classifier and Decision Tree J48. Thus it is expected to minimize the occurrence of bad loans. The data are divided into several categories: debtors with good, substandard, doubtful, and bad credit. The analysis was carried out using a 10-fold cross-validation model, where the results obtained from both tests, the highest accuracy value was the Decision Tree J48 of 78.26%. While the Naïve Bayes Classifier has a lower level of accuracy, the prediction results tend to be better than the Decision Tree J48. The prediction results with the Naïve Bayes Classifier can predict all classes and the most influential variable in classifying credit is the loan term.

Keywords: Credit classification, Naïve Bayes Classifier, Decision Tree J48

Introduction

Credit has a meaning in the form of a trust where the meaning of the word credit comes from the word credere. Faith means that if someone gets credit, they have earned trust. Meanwhile, for lenders, it means giving confidence to creditors that the money given with the aim of lending will return [1].

To make credit, banks apply the principle when conducting credit analysis to avoid problems in credit activities such as bad loans. In general, credit analysis is carried out based on the 5C code. The values included in the 5C principle have the customer's ability (Capacity), capital owned (Capital), the customer's economic condition (Condition), and guarantees provided (Collateral) [2].

The 5C principle does not necessarily make banks avoid problems in credit activities. One of the things that will happen to credit activities is bad credit. Credit can be harmful if the debtor does not meet the pre-agreed conditions [3].

To avoid bad credit, the bank must analyze credit data first. From these problems, the authors want to research credit classification analysis with a case study on BJB Bank debtor data in December 2021 using the Naive Bayes Classifier and Decision Tree J48. According to [4] is considered to have good potential in terms of accuracy and efficiency in computation compared to other classification methods. Meanwhile, the Decision Tree J48 was chosen as a comparison method because the classification with the Decision Tree J48 can take a complex decision into a simpler one, making it easy for the decision-making process.

In 2020, Ketjie et al. compares the Nave Bayes method and the C4.5 algorithm to predict credit card submission. Age, gender, recent education, marital status, number of dependents, type of company, monthly income, and salary slip are among the criteria used to make decisions. The final result discovered that both methods are relatively similar [5].

Materials and Methods

Materials

The data used in this study is secondary data obtained directly from the Bank BJB. The data were from Bank BJB's nominative loan debtors in December 2021. This study uses quantitative data, which is

then processed and analyzed to obtain each method's classification and prediction classes. The variables used in this study are age, loan amount, loan term, and collectibility, with a sample of 575 records divided into train and test data. Data distribution uses 80% for train and 20% for test data. This study will classify data by comparing two methods, Naïve Bayes Classifier and Decision Tree J48.

Naïve Bayes Classifier

The Naïve Bayes classifier is a fairly simple probabilistic classification method. The calculation carried out in this method is on a set of opportunities by adding up the frequency and combination of values from the dataset owned. The Naïve Bayes Classifier method assumes that each variable in each category is independent of the other [6].

Classification capabilities in Naïve Bayes are based on the Bayes theorem. In addition, the Naïve Bayes Classifier can also be said to be a method for making predictions in the future based on previous experience [7]. Based the Bayes theorem formula can be written as follows:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

where:

X : unknown class data
H : hypothesis from data X, whose class specifications are known
$P(H|X)$: the probability of conjecture H based on state X (posterior probability)
$P(H)$: the probability of conjecture H (prior probability)
$P(X|H)$: the probability of X based on the state of the conjecture H
$P(X)$: the probability of X

Decision Tree

The decision tree is a method with the basic concept of turning data into a decision tree and its rules. The selected variable will result in a constraint with the same data, and a simple decision tree can be generated with fewer iterations. This decision tree is based on rules that aim to divide the number of populations with heterogeneous properties into more detailed and homogeneous characteristics [8].

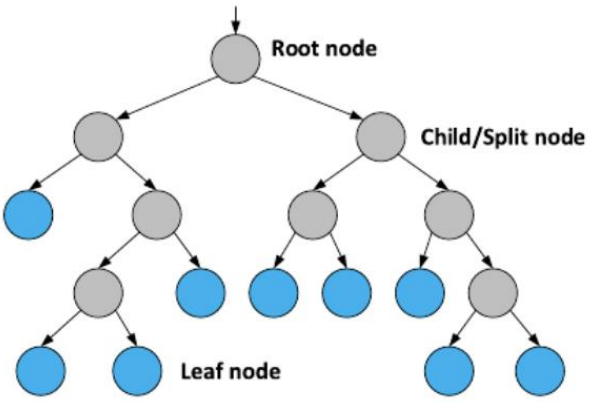

Figure 1. Illustration of decision tree [9]

Generally, the decision tree method is a top-down solution search strategy. In classifying unknown data, a test value will be carried out on the attribute value by tracking the track from the root (root node) to the leaf (end node) then class predictions will be made on the new data. In short, the decision tree is a classification method used in text mining [10].

One of the Decision Tree algorithms is C4.5, developed by Quinlan. Quinlan's C4.5 algorithm transforms J48 into a trimmed C4.5 decision tree. Every aspect of the information must be divided into minor subsets to decide. J48 examines the standardized data gain that indeed the results of splitting the information by selecting an attribute. To summarize, the acquired attribute extreme standardized data is used. The algorithm finds the minor subsets. If a subgroup has a place with a similar class in all instances, the split strategies end. J48 creates a decision node based on the class's expected estimations. It can handle

specific characteristics, lost or missing attribute estimations, and varying attribute costs. Pruning can improve its accuracy in this case [11].

The following are the steps in the J48 algorithm [12]:

Step 1: If the instances belong to the same class, the leaf is labeled with a similar category.

Step 2: For each attribute, the potential data will be calculated, and the data gained from the attribute test will be calculated.

Step 3: The best attribute will be chosen based on the result in Step 2.

J48 algorithms focus on some essential factors of a decision tree, such as entropy, information gain, and information gain ratio [13].

1. Entropy

Entropy is a quantitative measure of system disorder. It calculates a dataset's homogeneity to divide it into several classes. If the resulting category contains similar data, the entropy is zero; if the resulting class is equally divided into two datasets, the entropy is one. On the other hand, it measures the impurity of the dataset, which means that the higher the entropy value, the more information content.

$$Entropy = \sum_{i=1}^{c} p_i \ \log p_i \qquad (2)$$

2. Information Gain

The information gain metric is used to assess the purity or homogeneity of a dataset. The concept of information gain describes how data can be distributed concerning a response variable. If the information gain is high, the variable is more informative and should be considered for the root node. Information gain divides the dataset based on an attribute with lower entropy. Find the highest information gain attribute in the decision tree that divides the dataset into more homogeneous classes or sub-datasets.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \qquad (3)$$

where:

$\{S_1, \dots, S_i, \dots, S_N\}$: partition of S according to the value of attribute A

n : number of attribute A

$|S_i|$: the number of cases in the section S_i

$|S|$: total number of cases in S

3. Information Gain Ratio

Information gain does not apply to high branch attributes. The info gain ratio reduces the bias of information gain. When data is evenly divided, the resulting gain ratio value is significant and small when data belongs to only one class. The number and size of an attribute's branches are also considered when calculating the gain ratio. It removes the bias of information gain by using the intrinsic information of an attribute split.

$$IntrinsicInfo(S, A) = -\sum \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \qquad (4)$$

If an attribute's intrinsic value is high, it is less informative or essential. Less intrinsic values for an attribute increase its importance.

$$GainRatio(S, A) = \frac{Gain(S,A)}{IntrinsicInfo(S,A)} \qquad (5)$$

Results and Discussions

Data

The variables used are age, loan amount, loan term, and collectibility. Furthermore, the transformation is carried out by changing the data type from numeric to the interval. The description of the data set is defined in Table 1.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Table 1. Transformation Result

Attribute	Category	Transformation
Loans (rupiah)	0 – 150,000,000	P1
	150,000,001 – 300,000,000	P2
	300,000,001 – 450,000,000	P3
	450,000,001 – 600,000,000	P4
	>600,000,000	P5
Age	28 – 36	U1
	37 – 45	U2
	46 – 54	U3
	55 – 63	U4
	63 – 73	U5
Loans term (year)	1 – 5	W1
	6 – 10	W2
	11 – 15	W3
	16 – 20	W4
Collectibility	1 (good)	K1
	2 (substandard)	K2
	3 (doubtful)	K3
	4 (bad)	K4

Descriptive Statistics

By conducting a descriptive analysis, a general description of the collectibility of debtors at Bank BJB was obtained in December 2021.

Table 2. Descriptive Statistics

Variables	Minimum	Average	Maximum
Loans	2,183,554	135,268,893	671,626,121
Age	28	50	73
Loans term	2	11	20

Based on Table 2. it is found that the debtor with the largest loan is Rp. 671,626,121, while the lowest loan is Rp. 2,183,554. The longest period is 20 years, and the minimum is two years. Meanwhile, the explanation of the collectibility variable can be seen in the following graph.

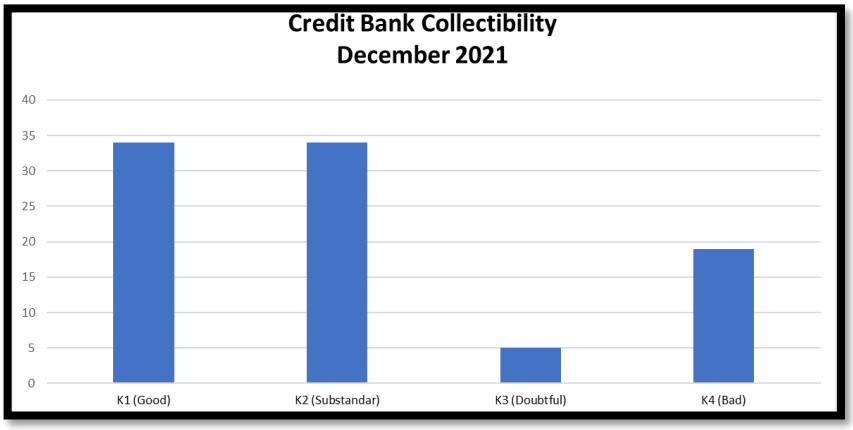

Figure 1. BJB Bank Debtor Collectibility

Figure 1. shows that in December 2021, the credit at Bank BJB was very healthy and stable, as seen from the number of debtors who were in class K1 (good) compared to class K3 (doubtful) and K4 (bad).

Naïve Bayes Classifier Analysis

Table 3. provides precise accuracy by class of NBC. Furthermore, Table 4. illustrates the classification based on NBC.

Table 3. Detailed Accuracy by Class of NBC

	TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area	Class
	0.714	0.137	0.750	0.714	0.732	0.584	0.946	0.908	K1
	0.744	0.306	0.593	0.744	0.660	0.425	0.835	0.729	K2
	0.000	0.009	0.000	0.000	0.000	-0.022	0.802	0.149	K3
	0.583	0.066	0.700	0.583	0.636	0.555	0.943	0.838	K4
Weighted average	0.661	0.179	0.642	0.661	0.647	0.487	0.896	0.787	

Table 4. Confusion Matrix of NBC

		Prediction			
		K1	K2	K3	K4
Actual	K1	30	12	0	0
	K2	8	32	0	3
	K3	0	3	0	3
	K4	2	7	1	14

Table 3. shows the model evaluation values for each collectibility class. It can be seen that the accuracy value for each class is quite good, except for K3. In addition, Table 4. issues the characteristics of prediction results.

1. The decision tree has classified 30 K1 objects as K1, and 12 as K2, leading to 12 misclassifications.
2. The decision tree has classified 32 K2 objects as K2 and eight as K1, leading to 8 misclassifications.
3. The decision tree has classified 0 K3 objects as K3 and three as K2 and K4, leading to 6 misclassifications.
4. The decision tree has classified 14 K4 objects as K4, two as K1, seven as K2, and one as K3, leading to 10 misclassifications.

From the prediction results obtained, the accuracy rate of NBC in the classification success is 66.09%.

Decision Tree J48 Analysis

The decision tree using the J48 algorithm gives the result that from the model formed, the tree size is 10 and 8 decision rules, which can be written as follows:

1. IF loans term = W1 THEN class = K1
2. IF loans term = W2 AND loans = P1 THEN class = K1
3. IF loans term = W2 AND loans = P2 THEN class = K2
4. IF loans term = W2 AND loans = P5 THEN class = K1
5. IF loans term = W2 AND loans = P3 THEN class = K1
6. IF loans term = W2 AND loans = p4 THEN class = K1
7. IF loans term = W3 THEN class K2
8. IF loans term = W4 THEN class K4

In addition, the prediction results of the classification using the J48 algorithm can be described in the Table 5 and 6.

Table 5. Detailed Accuracy by Class of Decision Tree J48

	TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area	Class
	1.000	0.151	0.792	1.000	0.884	0.820	0.957	0.923	K1
	0.744	0.194	0.696	0.744	0.719	0.543	0.827	0.683	K2
	0.000	0.000	?	0.000	?	?	0.768	0.113	K3
	0.667	0.000	1.000	0.667	0.800	0.783	0.900	0.806	K4
Weighted average	0.783	0.128	?	0.783	?	?	0.886	0.767	

Note: "?" indicates that both of TP and FP has 0 values

Table 6. Confusion Matrix of Decision Tree J48

		Prediction			
		K1	K2	K3	K4
Actual	K1	42	0	0	0
	K2	11	32	0	0
	K3	0	6	0	0
	K4	0	8	0	16

Table 5. shows that each evaluation value of the model for the collectibility of K1, K2, and K4 is quite good. Nevertheless, in K3, there is a "?" result. This condition is because, for this same class, the TP and FP values are 0. It seems that J48 did not assign any observations to this class. Meanwhile, the confusion matrix in Table 6. provides the following information:

1. The decision tree has classified 42 K1 objects as K1.
2. The decision tree has classified 32 K2 objects as K2 and 11 as K1, leading to 11 misclassifications.
3. The decision tree has classified 0 K3 objects as K3 and six as K2, leading to 6 misclassifications.
4. The decision tree has classified 16 K4 objects as K4 and eight as K2, leading to 8 misclassifications.

Based on the classification results for each class, it is obtained that the overall value of the Decision Tree classification accuracy using the J48 algorithm is 78.26%.

Comparison of the Accuracy

The test model for the two methods divides the determined data based on iterations carried out using the K-Fold Cross Validation model, where the iterations used are ten times [14]. With this data sharing method, the existing data will be divided into ten folds, with each part having the same size. Thus, ten subsets of data will be used to evaluate the performance of the model or algorithm. The results of the tests are concluded and described in Figure 2.

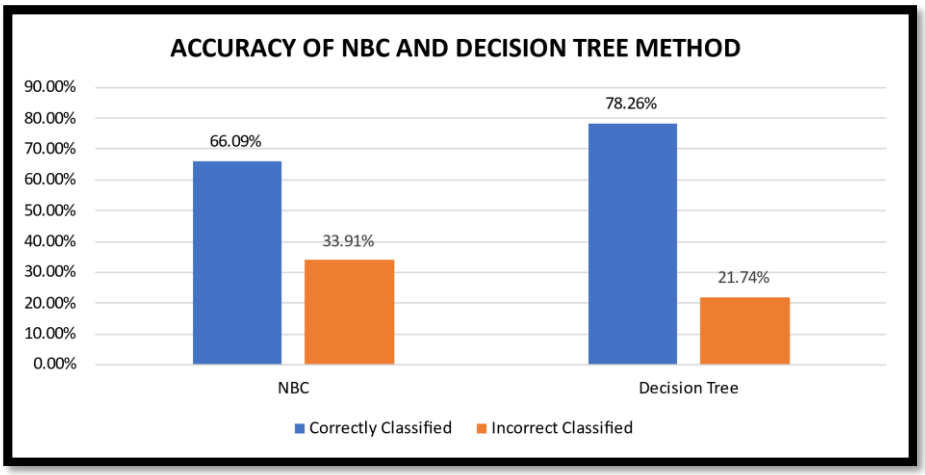

Figure 2. Comparison of Accuracy

From testing the two methods using the confusion matrix, it was obtained that the highest accuracy value was classification using the Decision Tree J48, which was 78.26%. For classification that is done using the Naïve Bayes Classifier, it has a lower accuracy rate of 66.09%. From the results of the comparison of accuracy obtained, it can be concluded that the best method according to the level of accuracy is the Decision Tree J48. However, as seen from the prediction results in Table 5., the Naïve Bayes Classifier method is the best. The Naïve Bayes Classifier is said to be good because, based on the prediction results made in both ways, the Naive Bayes Classifier can predict all classes, compared to the Decision Tree J48, which can only indicate three categories. Therefore, the Naïve Bayes Classifier predicts better than the Decision Tree J48.

Feature Selection

After obtaining the best model, the feature selection step is carried out using the Correlation Attribute Evaluation method to determine the percentage level of the most influential variable. Correlation Attribute Evaluation is a feature selection method that uses a ranking search method. Correlation Attribute Evaluation pays attention to the target class. Pearson's Correlation Method measured the correlation between each attribute and the target class[15]. Each value acts as an indicator by considering the nominal characteristics in the value base.

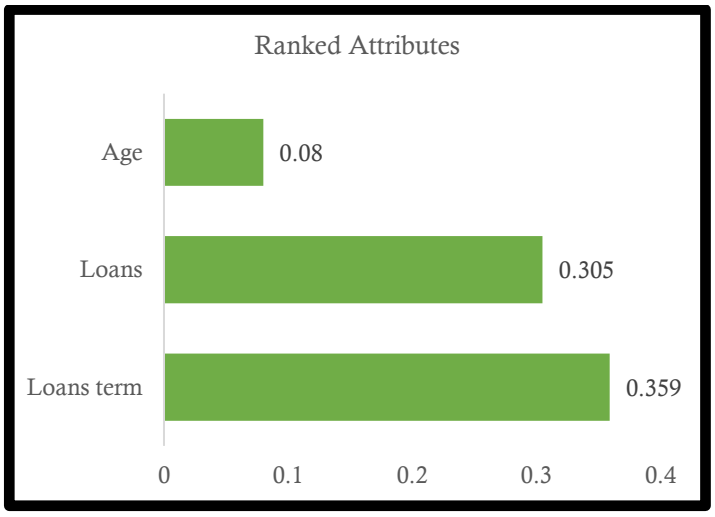

Figure 3. Ranked of Attributes

Figure 3. presents the result of feature selections where the purpose of feature selection in this study is to find the most influential variables. The output displayed is a ranking of features or variables. A variable is said to be the most significant if the ranked level obtained is close to 1. The output shows that the most influential variable in the Naïve Bayes Classifier is the loans term variable with a weight of 0.359.

Conclusion

The tests were performed by dividing the data using a 10-fold cross-validation model and the confusion matrix accuracy method. The results obtained from both tests using the Naïve Bayes Classifier and Decision Tree J48, namely the tendency of the highest accuracy value of the test is to use the Decision Tree J48, which is 78.26%. However, the Naïve Bayes Classifier tends to be better than the Decision Tree J48. In the prediction results with the Naïve Bayes Classifier, all classes can be predicted, while only three categories can be expected in the Decision Tree J48. Therefore, Naïve Bayes Classifier is the best method, and the weight order of each variable or feature is loans term, loans, and age.

Acknowledgment

The authors thank the Department of Statistics, Universitas Islam Indonesia, and Bank BJB Cabang Banjar for their valuable support.

References

[1] Kasmir, Manajemen Perbankan, PT. Raja Grafindo Persada, Jakarta, 2011.
[2] A. N. Kholifah and N. Insani, Analisis Klasifikasi Pada Nasabah Kredit Koperasi X Menggunakan Decision Tree C4.5 dan Naive Bayes, Jurnal Pendidikan Matematika dan Sains (2016).
[3] Rahmadeni, Susandi, R. Yendra, and A. P. Desvina, Analisis Diskriminan Fisher Untuk Klasifikasi Risiko Kredit, Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 11 (2019).
[4] D. S. L. Ting, P. W. H. Ip, and D. H. C. Tsang, Is Naïve Bayes a Good Classifier for Document Classification, Int. J. Softw. Eng. its Appl., 5(3) (2011) 37–46.
[5] Ketjie, V. C. Mawardi, and N. J. Perdana, Prediction of Credit Card Using the Naïve Bayes Method and C4.5 Algorithm, IOP Conf. Ser.: Mater. Sci. Eng., (2020).
[6] A. Nafalski and A. P. Wibawa, Machine translation with Javanese speech levels' classification, Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, 6(1) (2016) 21–25.
[7] Bustami, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, Jurnal Informatika, 8(1) (2014) 884–898.
[8] D. Yusuf and E. Sestri, Metode Decision Tree Dalam Klasifikasi Kredit Pada Nasabah PT Bank Perkreditan Rakyat (Studi Kasus : PT BPR Lubuk Raya Mandiri), Jurnal Sistem Informasi (JUSIN), 1(1) (2020) 21–28.
[9] M. R. Camana, S. Ahmed, C. E. Garcia, and I. Koo, Extremely Randomized Trees-BasedScheme for Stealthy Cyber-Attack Detection in Smart Grid Networks, IEEE Access 4(2016) (2020) 1–13.
[10] Rusito and M. T. Firmansyah, Implementasi Metode Decision Tree dan Algoritma C4.5 Untuk Klasifikasi Data Nasabah, INFOKAM, 1(12) (2016) 1–12.
[11] E. v. Venkatesan, Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, Indian Journal of Science and Technology (2015).
[12] N. Saravanan and V. Gayathri, Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48), Int. j. comput. intell. inform., 7(4) (2018) 188–198.
[13] P. Gulati, A. Sharma, and M. Gupta, Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review, Int. J. Comput. Appl., 141(14) (2016) 19–25.
[14] T.-T. Wong and P.-Y. Yeh, Reliable Accuracy Estimates from k-Fold Cross Validation, IEEE Trans. Knowl. Data Eng., 32(8) (2020).
[15] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, and R. Damaševičius, Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training, sensors, 20(6793) (2020).