APPLIED OF CLASSIFICATION TECHNIQUE IN DATA MINING FOR CREDIT SCORING

Heriyanto[1], Ika Kurniawati[2*], Fachri Amsury[3], M. Rizki Fahdia[4], Irwansyah Saputra[5],
Nanang Ruhyana[6], Asrul[7*]
Teknologi Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia[1,2,3,4,5,6]
Bisnis Digital, Universitas Teknologi Akba Makassar, Indonesia[7]
E-mail address: heriyanto.hio@nusamandiri.ac.id[1], ika.iki@nusamandiri.ac.id[2*],
fachri.fcy@nusamandiri.ac.id[3], rizki.muz@nusamandiri.ac.id[4],
irwansyah.iys@nusamandiri.ac.id[5], nanang.ngy@nusamandiri.ac.id[6], asrul@akba.ac.id[7*]

Received: 28, October, 2022 Revised: 24, December, 2022 Accepted: 24, December, 2022

ABSTRACT

In the development of the banking business, credit issues remain interesting to study and uncover. Most of the problems occur not in the system implemented by the bank, but the problem occurs precisely in the human resources who manage credit, either in their relationship with consumers or in errors on the part of the bank which mispredicts in assessing consumers who apply for credit. Several studies in the computer field have been carried out to reduce credit risk which causes losses to the company. In this study, a comparison of the Naive Bayes, C4.5 and KNN algorithms was carried out which was applied to consumer data that received credit eligibility for good and bad customers. The best prediction results are nave Bayes with an accuracy of 95.95% and an AUC of 0.974. The results of this classification are implemented in the form of a website-based application that can be used to facilitate related parties in the credit scoring system.

Keywords: C4.5; Credit Scoring; Data Mining; K-NN, Naïve Bayes.

1. INTRODUCTION

In distributing credit funds, the bank finds that there are several loans that are said to be substandard or bad credit which will then affect the provision of further credit or can also affect the bank's ability to channel credit. One way that the Bank Jabar can do to prevent bad credit is by knowing the quality of credit early on. Classification is a data mining technique that examines the behavior and attributes of a group of data. This method can classify new data with data that has been classified and generate a number of rules/patterns. The decision tree is an example of an easy and popular classification because it is easy to implement. Using classification method on credit customer data, it is expected to be able to predict the type of credit and reduce the number of bad loans. So that the results of this study are expected to be used by Bank Jabar to determine the classification of customer data that is classified as current or not and reduce the number of bad loans, and implementation into the application.

In banking, credit scoring uses machine learning technology primarily to predict the credit status of prospective customers to reduce losses caused by credit defaults. Credit assessment with machine learning is carried out by helping credit or banking institutions find important attributes that affect credit status, potential customers, then conduct an assessment of potential customers based on important attributes so that the level of bad loans can be reduced (Zhang, Yang, & Zhou, 2018). In this study, we conducted a credit analysis by digging up existing data on credit customer data based on their attributes using data mining techniques with the C4.5, Naive Bayes and K-NN. Application of the C4.5 Agorithm to analyze the feasibility of providing customer credit, in this case, the examiner compares the test with 8 attributes and 9 attributes to find the highest level of accuracy. Prediction of bad credit through customer behavior in savings and loan cooperatives by using the C4.5 classification technique.

This research was conducted by reviewing several previous studies related to research methods and objects. Research conducted by (Sucipto, 2015) predicts the occurrence of bad loans through customer behavior using classifier C4.5, using sample data from 1312 cooperative customers resulting in an accuracy value of 91.05%. Comparison of C4.5 Data Mining Algorithms with Naive Bayes for Evaluation of Crediting. In research on lending decision analysis, we compare the C4.5 algorithm with Naive Bayes to find out which results from processing the data sets of the two algorithms are more accurate and better. The results obtained for the same data set, the accuracy obtained in Algorithm C4.5 was 88.90% while Naive Bayes obtained an accuracy of 80.00% (Siti, 2016). Research by (Li, 2009) the classification method with KNN is an effective method for classifying credit. This method has been used in various real-world applications with success.

2. THEORY

2.1. Data Mining

Data mining is the process of analyzing data with an emphasis on finding hidden information in large amounts of data stored when running a company's business. Data mining technique is an information extraction process to explore knowledge (knowledge discovery) and find patterns (pattern recognition) in piles of data in databases which are usually large-scale. (Larose, 2007). An information extraction process to explore knowledge (knowledge discovery) and find patterns (pattern recognition) in large-scale databases is a data mining technique (Larose, 2007). In knowledge mining and data analysis, methods are needed to find patterns or patterns that have meaning. The methods used in data mining are description, estimation, prediction, classification, clustering, and association (Kusrini, 2015).

2.2. Credit Scoring

An important part of the credit risk management system is the process of assessing potential customers, this is done at financial institutions to predict the risk of loan applications. This process often relies on statistics that estimate account information from applications and customers also consider the possibility of default. In this time, the approach to customer scores was replaced or combined with an automated approach using machine learning (Jung, Thomas, & So, 2015) (Kennedy, Namee, & Delany, 2013). Borrowing process, bad credit often occurs, which until now has not been resolved. The cause of this problem could be due to weak analysis of potential customers who will borrow money. Because an analysis accuracy is important in determining the feasibility level of a prospective customer to be approved or not to borrow the funds. Here the author tries to do research on credit classification at Bank Jabar, in order to help the management of Bank Jabar Parung Panjang in determining the feasibility level of customers in borrowing loan money.

Data mining techniques are used to provide a model so that the bank is quick in making decisions for customers who are entitled to be given credit or rejected. With several classification algorithms tested on the training data, a C4.5 algorithm classification model is given which has the highest accuracy value. After being implemented on the test data, it is obtained the customer's decision which is rejected and the credit is accepted. To determine quickly and reduce the risk in "bad credit" in lending, it is necessary to analyze the training data patterns from existing customers to extract knowledge in the form of decision trees or rules that are easy to understand, so that the bank is easy to determine which credit is rejected or credit. the application received is based on the processed data.

2.3. Algorithm C4.5

One of the algorithms used for solving classification problems in data mining. C4.5 is an algorithm used to build a decision tree from data. C4.5 is a development of the ID3 algorithm which is also an algorithm for building a decision tree. Algorithm C4.5 recursively visits each decision node, selecting the optimal branch, until no more branches are possible (Larose, 2007). C4.5 is one of the algorithms in the decision tree. The building of tree is conducted by gain values of every feature. C4.5 is the development of ID3, using modifications of the information gain known as the gain ratio, which is used to overcome bias (Karegowda, Manjunath, Ratio, & Evaluation, 2010). In C4.5 algorithm, selecting the node as root is conducted based on the highest gain values from all of the features. The gain value is obtained by calculating the entropy (Muslim, Nurzahputra, & Prasetiyo, 2018).

$$Entropy(S) = \sum_{i=1}^{n} - p_i \, log_2(p_i) \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \; (1)$$

After calculating the entropy, calculate the gain value by using the result of entropy (2). The calculating is conducted until all the features became node in tree.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad \ldots\ldots \; (2)$$

2.4. Naive Bayes

The Naive Bayes (NB) is a The probabilistic classification method works with the assumption that all variables used are independent except for the target variable. This algorithm uses a parametric and probabilistic approach. This classification technique has Bayes rule to calculate the probability of class label Ci with all Aj attributes and predicts the class with the highest probability (Twala, 2010). Bayesian classification is also known as Naïve Bayes, which has comparable capabilities to decision trees and neural networks (Han & Kamber, 2006). Bayes classification is a statistical classification that can be used to predict the probability of membership of a class (Kusrini, 2009). Naive Bayes is a simple probability classifier that applies Bayes' Theorem with a high independence assumption (Telaumbanua & Kurniawati, 2022).

2.5. K-Nearest Neighbor

K-Nearest Neighboring (KNN) is a nonparametric classifier and has been applied in various classification problems (Peterson, Doom, & Raymer, 2005). Distance measurements are used by KNN classifiers to make predictions without building a model. The prediction for a new instance is given by most of the environment class labels in the training data. Find the nearest neighbor in the training data and use adjacent categories to determine the class of the given input. The training phase of the KNN classifier consists of storing the feature vectors and class labels of the training patterns. During the actual classification phase, the same properties as before were calculated for the specimens of the unknown class. The distance from the new vector to all stored vectors is calculated and the next sample is selected. The predicted new point includes the most numerous classes in the set (Feng-Chia Li, 2009). This algorithm is also a lazy learning technique. KNN is done by looking for groups of objects in the training data that are closest (similar) to objects in the new data or data testing (Wu, 2009).

3. METHOD

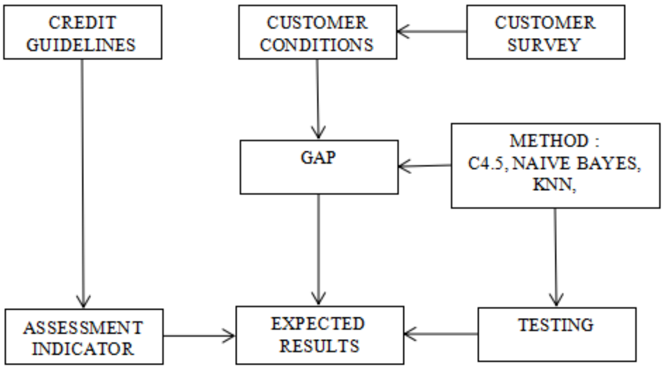


Figure 1. Proposed Method

Guidelines for creditworthiness and assessment indicators are used as the basis for granting credit from the head of the Bank Jabar branch. Determination of creditworthiness is currently still using conventional feasibility analysis based on the decision of the head of the Bank Jabar branch. Bank Jabar team conducted a survey to customers to see the customer's financial condition and analyze whether the customer had a problem or gap which would later be classified by comparing the C4.5, Naive Bayes and KNN methods. After testing using rapidminer tools, the results of the greatest accuracy will later be used to build a decision support system that is obtained is creditworthiness.

4. RESULTS AND DISCUSSION

In this work, the 866 data is divided into two parts, 766 is used for training data and 100 is used for testing data. Testing data as many as 866 and attributes consisting of 9 attributes. To verify the feasibility and feasibility of the credit model using C 4.5, Naive Bayes, and KNN, using a customer data set provided by the Bank Jabar Banten. Each bank customer in the data set contains: nine predictor variables, namely, id, gender, age, number of credits, term, number of installments, type of credit, nominative balance, and credit status (good or bad) as labels. Our data was obtained from the Bank Jabar Banten, several selections are made to produce the required data, the stages are:

1) Data Cleanup : to clean empty value or empty tuples. For example the attribute of arrears of fines.
2) Data Integration : which functions to unite different storage places into one data. In this case, there is only one place to store data, namely the customer's credit status.
3) Data reduction : the number of attributes used may be too large, from 9 attributes used only 5 required attributes consisting of 4 predictor attributes and 1 objective attribute, and attributes that are not needed will be deleted.
4) Modeling : the computing approach in this study was chosen based on a literature study on algorithm, namely C 4.5 which is able to classify current and credit credit status congested.
5) Analysis and evaluation pattern : the algorithm that has been developed in this research will be applied to the creditworthiness data through a simulation model. 80% of the data will be used as training data and 20% of the data will be used as checking or testing data.

Table 1. List of variables in dataset.

Variable Description	Rows	Measurement Level
Name	Insert	Text
Id	Insert	Number
Gender	Insert	Choice
Age	Insert	Number
Number Of Credits	Insert	Number
Term	Insert	Text
Number Of Installments	Insert	Number
Type Of Credit	Insert	Number
Nominative Balance	Insert	Number
Credit Status	Output	Binary

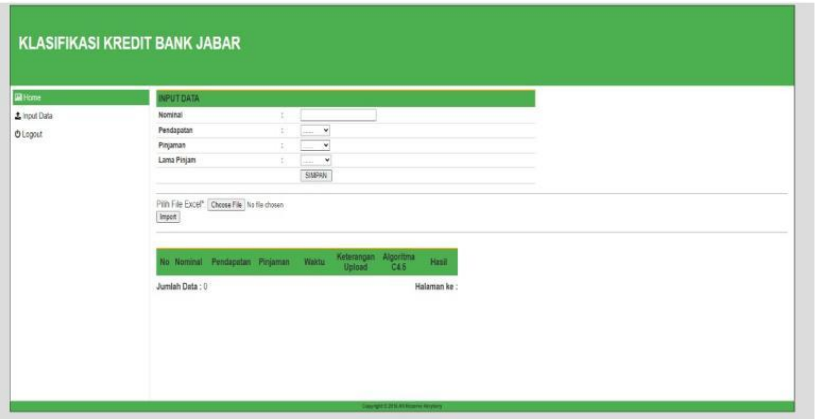
The cleaned training data will be processed using Rapidminer and modeled with C4.5, Naive Bayes and KNN to obtain a classification of customer creditworthiness data. From the results of testing data using Rapidminer, the highest results obtained from the Accuracy of testing the Naive Bayes 95.95% with an AUC of 0.974 which is categorized as very good. The accuracy obtained from KNN is 87.34% with an AUC of 0.958. For the C4.5, the accuracy results are still low at 71.36% with an AUC of 0.645 which is categorized as poor. And the following is a detailed comparison table of the Naive Bayes and C4.5 algorithms in determining creditworthiness. After the modeling has been carried out, it will be implemented into the development of a credit decision support system using a web programming.

Table 2. Model Comparison

Creterion	Naive Bayes	C4.5	K-NN
Accuracy	**95.95%**	71.13%	87.34%
Precision	89.08%	100.00%	100.00%
Recall	97.14%	0.40%	53.81%
AUC(optimistic)	0.974	1.000	1.000
AUC	0.974	0.500	0.958
AUC(pessimistic)	0.974	0.004	0.916

Implementation

After we tested the dataset using the classification algorithm, namely C4.5, Naive Bayes and K-NN and it was found that the highest accuracy was Naive Bayes with 95, 95%. We implemented the algorithm into an application called Bank Jabar credit classification. The implementation of the program uses a web programming framework. This information system can be used by Bank Jabar to predict customers who will apply for credit loans whether these customers are creditworthy or not and can be used as a decision support system for Bank Jabar in determining lending. The image below is a dashboard display in figure 2 where users can input customer data one by one or if the user wants to predict large amounts of customer data, they can use the import data feature in excel format.



KLASIFIKASI KREDIT BANK JABAR

Home
Input Data
Logout

INPUT DATA

Nominal	:	
Pendapatan	:	......
Pinjaman	:	......
Lama Pinjam	:	......

SIMPAN

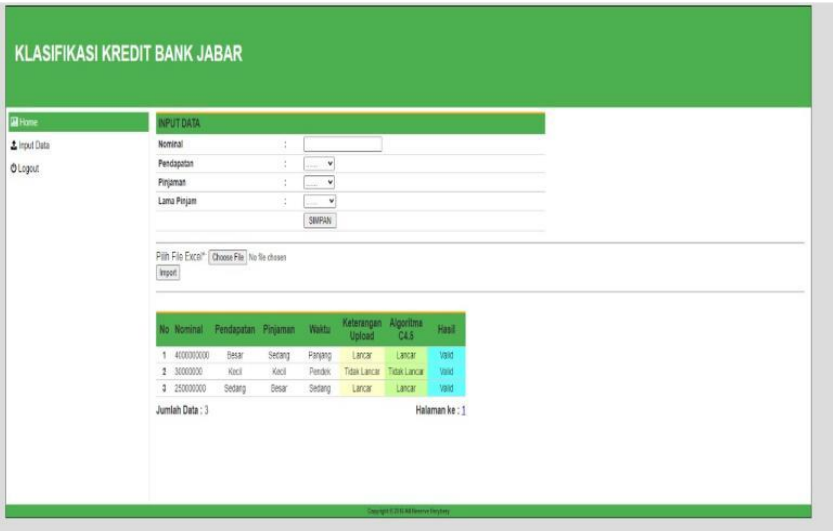
Pilih File Excel* Choose File No file chosen
Import

No	Nominal	Pendapatan	Pinjaman	Waktu	Keterangan Upload	Algoritma C4.5	Hasil

Jumlah Data : 0 Halaman ke :

Figure 2. Dashboard Data Input

In figure 3, based on the inputted data, the Bank Jabarcredit classification system is able to display the prediction results of customer data with good and bad status. This program is very useful for Bank Jabar to predict which customers will apply for credit.



KLASIFIKASI KREDIT BANK JABAR

Home
Input Data
Logout

INPUT DATA

Nominal	:	
Pendapatan	:	......
Pinjaman	:	......
Lama Pinjam	:	......

SIMPAN

Pilih File Excel* Choose File No file chosen
Import

No	Nominal	Pendapatan	Pinjaman	Waktu	Keterangan Upload	Algoritma C4.5	Hasil
1	4000000000	Besar	Sedang	Panjang	Lancar	Lancar	Valid
2	30000000	Kecil	Kecil	Pendek	Tidak Lancar	Tidak Lancar	Valid
3	250000000	Sedang	Besar	Sedang	Lancar	Lancar	Valid

Jumlah Data : 3 Halaman ke : 1

Figure 2. Display of Prediction

5. CONCLUSIONS AND SUGGESTIONS

In this study, a classification test was carried out using the C4.5, Naive Bayes and K-NN algorithms on customer data to assess creditworthiness. Then the results are compared to find out the best algorithm in determining credit risk. To measure the performance of the three algorithms, the Cross Validation, Confusion Matrix and ROC Curve testing methods are used. It is known that the Naive Bayes algorithm has the highest accuracy and AUC values, followed by the K-NN method, and the lowest. Method C4.5 Naive Bayes implementation is a fairly good method for classifying data with an accuracy of 95.95% and an AUC of 0.974. The Naive Bayes algorithm can provide a solution to the problem of determining a customer's eligibility to receive credit or not. We implemented the results of testing the three algorithms into Bank Jabar's creditworthiness classification information system which can be used to help support Bank Jabar's decision to provide customer credit.

REFERENCES

Feng-Chia Li, (2009). The hybrid credit scoring strategies based on KNN classifier. *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*, *1*, 330–334. https://doi.org/10.1109/FSKD.2009.261

Han J & Kamber. (2006), "Data Mining: Concepts and Techniques second edition". Simon Fraser University. USA: Morgan Kaufman Publisher.

Jung, K. M., Thomas, L. C., & So, M. C. (2015). When to rebuild or when to adjust scorecards. *Journal of the Operational Research Society*, *66*(10), 1656–1668. https://doi.org/10.1057/jors.2015.43

Karegowda, A. G., Manjunath, A. S., Ratio, G., & Evaluation, C. F. (2010). 3.Comparative study of Attribute Selection Using Gain Ratio. *International Journal of Information Technology and Knowledge and Knowledge Management*, *2*(2), 271–277. Retrieved from https://pdfs.semanticscholar.org/3555/1bc9ec8b6ee3c97c524f9c9ceee798c2026e.pdf%0Ahttp://csjournals.com/IJITKM/PDF 3-1/19.pdf

Kennedy, K., Namee, B. Mac, & Delany, S. J. (2013). Using semi-supervised classifiers for credit scoring. *Journal of the Operational Research Society*, *64*(4), 513–529. https://doi.org/10.1057/jors.2011.30

Kusrini Dan Taufig Lutfia(2009). Algoritma Data Mining. Yogyakarta : Andi Offset.

Kusrini and T. Emha, (2015). "Definisi Data Mining," Data Mining.,

Larose DT. (2007). *Discovering Knowledge in Databases. New Jersey: John Willey & Sons Inc.Myatt, Glenn J. Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data mining*. New Jersey: John Wiley & Sons, Inc.

Muslim, M. A., Nurzahputra, A., & Prasetiyo, B. (2018). Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction. *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, *2018-Janua*(1996), 141–145. https://doi.org/10.1109/ICOIACT.2018.8350753

Peterson, M. R., Doom, T. E., & Raymer, M. L. (2005). GA-facilitated classifier optimization with varying similarity measures. *GECCO 2005 - Genetic and Evolutionary Computation Conference*, 1549–1550. https://doi.org/10.1145/1068009.1068253

Siti, M. (2016). Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit. *Bina Insani Ict Journal*, *3*(1), 187–193. Retrieved from http://ejournal-binainsani.ac.id/index.php/BIICT/article/view/815/658

Sucipto, A. (2015). Prediksi Kredit Macet Melalui Perilaku Nasabah Pada Koperasi Simpan Pinjam Dengan Menggunakan Metode Alogaritma Klasifikasi C4.5. *Jurnal DISPROTEK*, *6*(1), 75–87.

Telaumbanua, D., & Kurniawati, I. (2022). Penerapan Algoritma C4. 5 Untuk Klasifikasi Kepuasan Pelanggan Pada Jasa Layanan Pengiriman. *JoMMiT: Jurnal Multi Media dan IT*, *6*(1).

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, *37*(4), 3326–3336. https://doi.org/10.1016/j.eswa.2009.10.018

Wu, X., Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Boca Raton (US): CRC Press

Zhang, X., Yang, Y., & Zhou, Z. (2018). A novel credit scoring model based on optimized random forest. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, *2018-Janua*, 60–65. https://doi.org/10.1109/CCWC.2018.8301707.