Predicting the loan risk towards new customer applying data mining using nearest neighbor algorithm

S Samsir*, S Suparno and M Giatman
Department of Engineering, Technology and Vocational Education, Padang State University, Padang, Indonesia

*samsirst111@gmail.com

Abstract. Unstable economic conditions require Bank must be careful in deciding towards lending customers. Banks should not take the risk of giving loans to customers who cannot afford to pay. This study aims to assist bank in predicting lending. The study was conducted at the Bank Perkreditan Rakyat in Medan. The study was conducted applying data mining using the nearest neighbor algorithm. This algorithm was chosen because the nearest neighbor can calculate the closeness between new cases and old cases based on matching weights from a number of existing features. This algorithm will calculate the closeness with predetermined criteria. Hoped bank will be helped in making predictions.

1. Introduction
Bank Rakyat Indonesia is a state-owned bank, one of the services of which provides loan assistance to customers. The purpose of lending is usually many factors, including aims to help the customer's business and problems in accordance with the wishes of the customer [1]. Bank Rakyat Indonesia Medan branch received many loan applications. The request cannot be ignored because it is the bank's obligation to accommodate customers who need financial assistance. But the bank also cannot arbitrarily approve the loan application. It must be well researched and ascertained before deciding to grant a customer's loan application. Do not let the customer receive the loan but cannot return the loan [2]. In the end the bank experienced a problem. So far, the Bank Rakyat Indonesia field branch has experienced difficulty in terms of time and problems many customers who are in arrears on loans. This happens because of an error in predicting the customer's ability to make payments. This problem certainly must be resolved. Do not let the assessment of Bank Rakyat Indonesia field branch becomes bad because of the many customers who are overdue [3].

The rapid development of information technology today can be a solution for Bank Rakyat Indonesia Medan branch. These developments made it possible to build an information system that could be a tool to help predict customers' ability to repay loans. So this information system can be used to decide whether a customer is worthy of being given a loan or not. The information system in question combines a computerized programming language with one of the methods, the Nearest Neighbor method [4]. Nearest Neighbor is an approach to look for cases by calculating the closeness between new cases and old cases, which is based on matching the weights of a number of existing features. So the Nearest Neighbor can be predicted. Prediction is used to estimate the future value, for example predicting the stock of goods one year in the future. This function includes Neural Network, Decision Tree, and Nearest
Neighbor methods. Prediction uses several variables or database fields to predict the values of future variables that are needed, which are unknown at this time [5-7].

The combination of Nearest Neighbor with an information technology-based programming language will be supported by criteria that become Bank Rakyat Indonesia's rules in deciding whether a customer can receive a loan or not. These criteria include: Customer has a history of bad loans, Customer collateral is considered less feasible, Customer income is less able to compensate for the loan, Customer does not provide their own capital, Business age is still too young, Risk factors in a profession, Customer reputation is considered not good, and Data invalid customer. Thus the information system that is built is in accordance with the needs of Bank Rakyat Indonesia, so that efforts to improve the quality of lending data can also be done better [8,9].

2. Method

2.1. Preliminary research
Preliminary research is the first step to see the level of risk experienced by Bank Rakyat Indonesia in providing loan funds to customers that may result in losses to Bank Rakyat Indonesia [10].

2.2. Data collection
The data needed includes data on the amount of the loan, the purpose of the loan, the time period, the debtor's condition, the debtor's income and the collateral. The data used in research is quantitative data by collecting as much information as possible and then presented as best as possible so that it will produce a quality study.

2.3. Analysis

2.3.1. Data analysis. The data analysis process is carried out to get accurate data based on the data mining algorithm that will be used.

2.3.2. Analysis process. The method used in the research process is the classification method will use the nearest neighbor algorithm which will classify objects based on the data closest to the previous data.

2.3.3. Design. The design is done by using a design tool that is the Unified Modeling Language (UML). This study uses a use case diagram to see the relationship of actors with the system.

2.3.4. Implementation. Implementation is the process of translating the design model into an application that will be used by using the PHP programming language and supported by a MySQL database.

2.3.5. Testing. Testing is done to determine whether the application produced matches the results of the algorithm used.

3. Results and discussion

3.1. Results

3.1.1. Data analysis. The study was conducted at Bank Rakyat Indonesia Medan branch. This study uses 83 fund lending transaction records at Bank Rakyat Indonesia Medan branch. Before conducting the data mining process, it is necessary to clean the data to eliminate data duplication and data inconsistencies. After cleaning the data, clean data will be obtained and ready to be used for the data mining process. In this study, data that has been cleared into 65 records to be used as training data [11,12].
Data on the determination of the risk of loan funds at Bank Rakyat Indonesia Medan branch consists of 8 attributes, of which 7 are attribute predictors and 1 attribute are labels, as shown in the following table:

Table 1. List of attributes and their attribute values.

| Attribute               | Attribute Values                              |
|-------------------------|----------------------------------------------|
| Loan Amount             | <= 20.000.000                               |
|                         | > 20.000.000 <= 50.000.000                  |
|                         | > 50.000.000 <= 75.000.000                  |
|                         | > 75.000.000                                |
| Loan Objectives         | Investation                                  |
|                         | Consumptive                                   |
|                         | Working capital                               |
| Time period             | 12 Month                                      |
|                         | 24 Month                                      |
|                         | 36 Month                                      |
|                         | 48 Month                                      |
|                         | 60 Month                                      |
|                         | 72 Month                                      |
| Debtor Conditions       | Pretty good                                   |
|                         | Good                                          |
|                         | Very Good                                     |
| Debtor Work             | Housewife                                     |
|                         | Student                                       |
|                         | PNS                                           |
|                         | Employee                                      |
|                         | Private employees                             |
|                         | Tax Employee                                  |
|                         | Entrepreneur                                  |
| Debtor Income per Month | < 2.500.000                                  |
|                         | 2.500.000 <= 6.500.000                      |
|                         | > 6.500.000                                  |
| Guarantee               | BPKB two-wheeled vehicle                      |
|                         | BPKB four-wheeled vehicle                     |
|                         | Land certificate                              |
|                         | Home certificate                              |
|                         | SK PNS                                        |
| Loan Risk               | ?                                             |

3.1.2. Process analysis. From the grouping of attributes above, the next step is to analyze the process in data mining using the Nearest Neighbor Algorithm. To measure the distance of each attribute, we need to give a weight. The weight given is between 0 and 1, where 0 indicates the attribute has no effect and 1 indicates the attribute is very influential. The weighting of each attribute can be seen in the following table:

Table 2. Weight of each predictor.

| No. | Attribute               | Attribute |
|-----|-------------------------|-----------|
| 1.  | Loan Amount             | 0.8       |
| 2.  | Loan Objectives         | 0.8       |
| 3.  | Time period             | 0.8       |
| 4.  | Debtor Conditions       | 0.6       |
| 5.  | Debtor Work             | 0.6       |
| 6.  | Debtor Income per Month | 0.6       |
| 7.  | Guarantee               | 0.5       |
3.2. Discussion

3.2.1. Proximity of attribute values. The proximity value of the Loan Amount attribute is shown in the following table:

**Loan Amount:**

\- \( A1 = <= 20.000.000 \)
\- \( A2 = > 20.000.000 <= 50.000.000 \)
\- \( A3 = > 50.000.000 s/d <= 75.000.000 \)
\- \( A4 = > 75.000.000 \)

**Table 3. Proximity of the attribute value of Loan Amount.**

| No. | Loan Amount | Attribute Value | Weight |
|-----|-------------|-----------------|--------|
| 1.  | A1          | <= 20.000.000   | 0.8    |
| 2.  | A2          | > 20.000.000 <= 50.000.000 | 0.8 |
| 3.  | A3          | > 50.000.000 s/d <= 75.000.000 | 0.6 |
| 4.  | A4          | > 75.000.000    | 0.5    |

3.2.2. Sample training data. The next process is to do the calculation process based on training data that will be used. Used 3 training data sample records from 65 training data records.

**Table 4. Classification data.**

| .     | G1  | G2  | G3  | G4  | G5  | G5  |
|-------|-----|-----|-----|-----|-----|-----|
| G1    | 1   | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 |
| G2    | 0.9 | 1   | 0.899 | 0.778 | 0.667 | 0.556 |
| G3    | 0.8 | 0.889 | 1   | 0.875 | 0.75 | 0.625 |
| G4    | 0.7 | 0.778 | 0.875 | 1   | 0.857 | 0.714 |
| G5    | 0.6 | 0.667 | 0.75 | 0.857 | 1   | 0.833 |
| G6    | 0.5 | 0.556 | 0.625 | 0.714 | 0.833 | 1   |

3.2.3. Proximity of cases. Calculation of the proximity of new cases to the old case data in the training data in table:

**Table 5. Proximity of new cases to case.**

| No. | Attribute            | Case Attribute Value number 1 | New Case Attribute Value | Proximity (a) | Weight (b) |
|-----|----------------------|-------------------------------|--------------------------|---------------|------------|
| 1.  | Loan Amount          | > 20.000.000 <= 50.000.000   | > 20.000.000             | 1             | 0.8        |
| 2.  | Loan Objectives      | Working capital               | Consumptive              | 0.571         | 0.8        |
| 3.  | Time period          | 36 Month                      | 60 Month                 | 0.75          | 0.8        |
| 4.  | Debtor Conditions    | Pretty good                   | Good                     | 0.7           | 0.6        |
| 5.  | Debtor Work          | Private employees             | Private employees        | 1             | 0.6        |
| 6.  | Debtor Income per Month | 2.500.000                   | 2.500.000                | 1             | 0.6        |
| 7.  | Guarantee            | Land certificate              | BPKB two-wheeled vehicle | 0.8           | 0.5        |

From the table above, you can calculate the proximity of new cases to cases, by:
Similarity =
\[
((a1*b1)+(a2*b2)+(a3*b3)+(a4*b4)+(a5*b5)+(a6*b6)+(a7*b7)) / \\
(b1+b2+b3+b4+b5+b6+b7)
\]

Similarity =
\[
((1*0.8)+(0.571*0.8)+(0.75*0.8)+(0.7*0.6) \\
+(1*0.6)+(1*0.6)+(0.8*0.5)) / \\
(0.8+0.8+0.8+0.6+0.6+0.6+0.5)
\]

Similarity = 0.825

| No. | Attribute          | Case Attribute Value number 1 | New Case Attribute Value | Proximity (a) | Weight (b) |
|-----|--------------------|-------------------------------|--------------------------|---------------|-----------|
| 1.  | Loan Amount        | <= 20.000.000                 | > 20.000.000             | 0.8           | 0.8       |
|     |                    | <= 50.000.000                 |                          |               |           |
| 2.  | Loan Objectives    | Consumptive                   | Consumptive              | 1             | 0.8       |
| 3.  | Time period        | 36 Month                      | 60 Month                 | 0.75          | 0.8       |
| 4.  | Debtor Conditions  | Good                          | Good                     | 1             | 0.6       |
| 5.  | Debtor Work        | Private employees             | Private employees        | 1             | 0.6       |
|     |                    |                               |                          |               |           |
| 6.  | Debtor Income per  | 2.500.000                     | 2.500.000                | 1             | 0.6       |
|     | Month              | 6.500.000                     | 6.500.000                |               |           |
| 7.  | Guarantee          | BPKB four-wheeled vehicle     | BPKB two-wheeled vehicle | 0.9           | 0.5       |

From the table above, you can calculate the proximity of new cases to cases, by:

Similarity =
\[
((a1*b1)+(a2*b2)+(a3*b3)+(a4*b4)+(a5*b5)+(a6*b6)+(a7*b7)) / (b1+b2+b3+b4+b5+b6+b7)
\]

Similarity =
\[
((1*0.8)+(0.571*0.8)+(0.75*0.8)+(0.7*0.6) \\
+(1*0.6)+(1*0.6)+(0.8*0.5)) / \\
(0.8+0.8+0.8+0.6+0.6+0.6+0.5)
\]

Similarity = 0.913

4. Conclusions
Based on testing of 3 cases, the highest score is the third case, the case which is the closest case to the new case. The risk classification of loan funds in new cases is Low. The Nearest Neighbor that is generated has been able to show the relationship between one attribute with another attribute. The Nearest Neighbor method is able to divide each attribute into each class. Splitting credit history attributes into several small groups helps the user in understanding the information generated by the application used. Credit history attributes are the root nodes of the six existing attributes, namely gender, savings, work, credit history, income and collateral. Followed by the Income Attribute as a branch node. Thus, a person's Credit and Income History is a major factor for banks to determine the level of credit risk to prospective creditors.

References
[1] Leo M, Sharma S and Maddulety K 2019 Machine learning in banking risk management: A literature review Risks 7(1) 29
[2] Basel Committee on Banking Supervision 2011 *Principles for the Sound Management of Operational Risk* (Basel: Bank for International Settlements) pp 1–27

[3] Torvekar N and Game P S 2019 Predictive Analysis of Credit Score for Credit Card Defaulters *International Journal of Recent Technology and Engineering* **7**(5S2) 282-286

[4] Koklu M and Sabanci K 2016 Estimation of Credit Card Customers Payment Status by Using kNN and MLP *International Journal of Intelligent Systems and Applications in Engineering* **4** 249-251

[5] Venkatesh A, Gracia A and Shomona 2016 Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers *International Journal of Computer Applications* **145** 36-41

[6] Yeh I C and Lien C H 2009 The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients *Expert Systems with Applications* **36**(2) 2473-2480

[7] Wah Y B and Ibrahim I R 2010 Using data mining predictive models to classify credit card applicants *In 2010 6th International Conference on Advanced Information Management and Service (IMS)*

[8] Keramati A and Yousefi N 2011 A proposed classification of data mining techniques in credit scoring *In International Conference on Industrial Engineering and Operations Management* pp 416-424

[9] Lin S L 2009 A new two-stage hybrid approach of credit risk in banking industry *Expert Systems with Applications* **36**(4) 8333–8341

[10] Yu L, Wang S, Wen F, Lai K K and He S 2008 Designing A Hybrid Intelligent Mining System for Credit Risk Evaluation *Journal of Systems Science and Complexity* **21**(4) 527-539

[11] Hsieh N C and Hung L P 2010 A data driven ensemble classifier for credit scoring analysis *Expert Systems with Applications* **37**(1) 534–545

[12] Chen F L and Li F C 2010 Combination of feature selection approaches with SVM in credit scoring *Expert Systems with Applications* **37**(7) 4902-4909