Data Mining Classification Of Filing Credit Customers Without Collateral With K-Nearest Neighbor Algorithm (Case study: PT. BPR Diori Double)

Jeprianto Sinaga¹, Bosker Sinaga²

¹²Informatics Engineering Study Program, STMIK Pelita Nusantara, Jl. Iskandar Muda No. 1 Medan, North Sumatra, Indonesia 20154

E-mail: Jeffreysinaga1903@gmail.com, sinagab8@gmail.com

Abstract- Unsecured loans are the community's choice for lending to banks that provide these services. PT. RB Diori Ganda is a regional private banking company that serves savings and loans and loans without collateral for the community. Submission of unsecured loans must go through an assessor team to process the analysis of the attributes that affect the customer's classification so that credit can be approved, which is then submitted to the commissioner for credit approval. But what if those who apply for credit on the same day in large amounts, of course this will make the process of credit analysis and approval take a long time. If it is seen from the many needs of the community to apply for loans without collateral, a classification application is needed, in order to facilitate the work of the assessor team in the process of analyzing the attributes that affect customer classification. To find out the classification of customers who apply for unsecured loans for using data mining with the K-Nearest Neighbor algorithm. The result of this research is the classification of problematic or non-performing customers for credit applications without collateral.

Keywords: Data Mining, Customer Classification, K-Nearest Neighbor Algorithm

1. Introduction

Classification is one of the tasks in data mining to predict a future state through state testing in the past with a series of processes or analysis for the calculation of distance closeness that state. Data mining is a process of extracting data or filtering of data by using a data set with a large enough size through a series of processes to obtain valuable information from these data.

Diori PT. BPR Ganda is a private banking company that serves the community savings and loans for medium and supervised by the FSA. Register and problematic customer data already collected by PT. RB Dual Diori stored in the database. Pengelolah companies must be able to predict if any existing customers apply for credit without collateral then must predict whether the customer's classification is problematic or not problematic future in the mortgage payment, the classification prediction aims to avoid the problem of bad loans in the future. The process of data analysis conducted tanpaagunan submitting credit assessor teams are still using manual processes are observed and to sort out the customer's data, so that this process takes a long time or can be called inefficient time.

K-Nearest Neighbor algorithm is a classification method that classifies test data (testing) of the training data (training) based on the test data within the data to several neighbors (neighbor) nearby. K-Nearest Neighbor algorithm often used in the completion of data mining for predictive job classification. This algorithm uses training data (training) who have been there and its output has been known.

2. Theory

2.1 Data Mining

Data mining search containing the desired trends or patterns in large databases to support decision-making at a time when that will come. These patterns are recognized by specific devices that can provide a useful analysis and insightful data that can then be studied more closely, which may use other decision support tools. In the real world, data mining is more than just apply a of the algorithms are simple. Often the data is too crowded and incomplete, if it is justified, slow lorises diginkan great many patterns will be lost and any pattern detection capabilities will go down. Furthermore, the analysis should decide what kind of mining algorithms are used, apply it into subsets of the data sample, variable, digest the results and the process iterates (Fajar Astuti Hermawati, 2013: 2-3). Employment in data mining can be divided into four groups: the model prediction
(prediction modeling), cluster analysis (cluster analysis), analysis of association (association analysis), and detection of anomalies (anomaly detection).

The prediction model (prediction modeling) This work deals with the creation of a model that can melaukan mapping of any set of variables for each target, and then use that model to give a target on him [newly acquired unan. There are 2 Jenin prediction model that is klasosfokasi and regression. Classification is used for discrete target variable, whereas regression is used for continuous target variable. (Eko Prasetyo, 2014: 5).

Data mining actually part of a series of information on the database search process. Pattern information generated from the data mining process needs to be presented in a form easily understood by the parties concerned. This stage is part of the prose Knowledge Discovery in Databases (KDD) called interpretation. This phase also includes an examination of whether a pattern or information found contradicted by the facts or hypotheses that existed before. (Muchlisin Riadi, 2017). to take more details can be seen in Figure 2.1.

![Fig 1. Stages KDD](Source: Muchlisin Riadi, 2017)

2.2 Algorithm K-Nearest Neighbor

*K-Nearest Neighbors* one of the Nearest Neighbor method of classification based on the oldest and most popular. Rated K used herein express the number of nearest neighbors were involved in the determination of the class label prediction on the test data. (Eko Prasetyo, 2014: 150).

Jodi Irjaya Kartika, Eddy Santoso, and Sutrisno (2017) found a method or algorithm K-nearest Neighbor (K-NN) is a method that uses supervised algorithm in which the results of the new query/instance diklasifikaikan by a majority of the category in the K-nearest Neighbor. The purpose of this algorithm is to classify the new objects based on attributes and sample training. In the algorithm K-Nearest Neighbor (KNN) there are five (5) ways to search for the nearest neighbors are: [1] The Euclidean, [2] Distance Manhattan, [3] The Cosine, [4] The correlation, [5] Hamming distance.

In this study the means used to search for the nearest neighbors only use the Euclidean distance with the following formula:

\[ \text{Similarity } p,q = \frac{\sum_{i=1}^{n} f p_i q_i w_i}{w_i} \]

Keterangan:
P = The new cases
q = cases exist in storage
n = number of attributes in each case
i = attribute of individuals between 1 and n
f = function f attribute similarity between the case p and q case
w = The weight given to the attributes of the i-th

Characteristics of K-Nearest Neighbor algorithm (KNN) is a supervised learning algorithm that means this algorithm uses existing data and output are known. KNN widely used in applications of data mining, pattern recognition, image processing, etc. The steps of the algorithm calculation K-Nearest Neighbor (K-NN) in resolving a problem is as follows:

a) Menentukan value of k. N Determining the value of k used in klasiifikasi not have standard rules, but in this study the value of k used are 3, 5, 7.

b) Calculating the distance between the object / new data on all objects / data yang telah in training.

c) Sort the results of these calculations.

d) Determine the nearest neighbors by the minimum distance to K.

e) Use categories majority as object classification / new data,
2.1 Unsecured Loan

The word credit comes from the Latin is credere means trust. The owner of money or goods (creditor) gave keperayaan to the borrower (debtor) for the use of money or goods for a certain time. Such borrowing accompanied by the belief that the debtor can repay the money or goods lent. Thus, the use of a credit is giving money or goods to another person in a specific period of time with collateral or unsecured, with the provision of services of interest or no interest.

According to the Act. No.10 of 1998 on banking, noted that credit is a provision of cash or which can be likened to it, based on the approval or kesepeakatan lending - borrowing between banks and other parties who require the pemijam to repay the debt after a certain period of time with interest. From this it can be concluded uaraian take credit can be given to a person or institution on account of the trust, risk, time, and achievement. Risks include a reference to the fact that the net present value will vary with the value for money when it is returned.

3. Analysis

3.1 Data Analysis Research

The data in this study come from PT. RB Diori Simpang Dual Wire, namely customer data were ever filed and do credit (loan) on the marketing of the region 7. The data obtained and which would be predicted classification has been organizing (performed encoding the data prior to the data the original). Namely the training data and testing data.

1. Training data is customer data that has been applying for credit and classification known. Kolektibilas also decisive klasifikikasinya. If the customer where collectibility between 0 to 2, the customer dklasifikasikan not problematic, and if the customer kolektibilatas between 3 or more, the customer in classifying problematic.

| Marital status | The number of dependents | Last education | Age         |
|----------------|-------------------------|----------------|-------------|
| Married        | There is no             | SLTA           | 21-35 years old |
| Married        | 5 people                | SD             | 35-50 years old |

Table 2

| Work ownership Ruma | monthly income | credit Status |
|---------------------|----------------|---------------|
| employee            | <= 3 jt        | No problem    |
| farmer              | > 3 jt <= 6 jt | problematic   |

2. Data testing is the customer data that would apply for credit without collateral to be classified.

| Marital status | The number of dependents | Last education | Age         |
|----------------|-------------------------|----------------|-------------|
| Married        | 3 people                | SLTA           | 21-35 years old |

Table 4

| Work ownership | Income Per month | Credit Status |
|----------------|------------------|---------------|
| employee       | Parents          | <= 3 jt       | -            |

The data obtained in this study were 116 credit transaction record whether or not bermaslah bemasalah, obtained from a banking company PT. RB Diori Simpang Dual Wire. All the Attributes in the category valued credit data, as shown in Table 4. Training data consists of eight attributes, in which 7-value attributes (attributes influential), and 1 attribute label. Results obtained data preprocessing is necessary to obtain high quality data,