ISSN 2347 - 3983

Volume 8. No. 6, June 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter28862020.pdf

https://doi.org/10.30534/ijeter/2020/28862020

Credit Classification Using CRISP-DM Method On Bank ABC Customers

Toni Darmawan[1], Anggih Surya Birawa[2], Ery Eryanto[3], Tuga Mauritsius[4]

[1234]Information Systems Management Departement, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University Jakarta, Indonesia 11480

[1]toni.darmawan@binus.ac.id, [2]anggih.suryabirawa@binus.ac.id, [3]ery.eryanto@binus.ac.id, [4]tuga.mauritsius@binus.ac.id

ABSTRACT

Information systems are very important in the banking industry. This is because to simplify the decision-making process, one of which is the determination of prospective credit customers at Bank ABC. Bank ABC is one of the financial institutions in the growth phase needed to improve the quality of safe, high-quality and accurate credit. The purpose of this study is to accelerate the decision making process of credit customers, namely whether credit is accepted or rejected. One method used is CRISP-DM as standart processes for data mining. Election algorithm using the C4.5 decision tree algorithm. The final result of this study is the prediction of the pattern of credit results for credit customers. This pattern is generated from decision tree is expexted to be a reference for credit customer take a credit at Bank ABC.

Key words : Information systems, CRISP-DM, C4.5 algoritm, pattern of credit, data mining.

1. INTRODUCTION

Bank is financial institution with main activity to collect funding and redistribute to society in term of credit. In order to running its business activities, Bank must distribute their funding with precautionary principle to ensure that lender able to return it in timely manner.

Credit scoring or credit classification is a method of predicting potential risk corresponding to a credit portfolio. These tools can be used by financial institution to evaluate portfolios in term of risk. In evaluating credit, of course data is something that is very necessary. Data is growing rapidly, there is a need for advanced analytic techniques that operates on such data and extracts effective information, unknown patterns, and relationships that help in making decisions [11].

Each credit customer must have their own patterns such as the amount of income and the purpose of loaning funds to the bank so that the duration of the loan will be known and the amount of loan needed. An analysis of the pattern of prospective credit customers needs to be done more deeply isn order to reduce the risk of default on the bank.In this research we are using CRISP-DM. this method provide standard for data mining that can be implemented in the strategy for general problem solving. CRISP DM compare other data mining methodology more fully and well documented. Every phase structured and define well and easy in implementation.

Besides using right method, selecting correct method is important. C.4.5 decision tree algorithm is the most accurate algorithm to be used to produce classification rules. The data mining classification was indeed a very good way to avoid the subjectivity in decision making because it use the historical data [12]. The results of this study are to predict customer credit classification so that it can help and facilitate Bank ABC in making decisions on prospective credit customers so as to minimize bank risk, namely the failure to pay credit customers in the future using the decision tree algorithm C4.5 method.

2. LITERATUR REVIEW

2.1 Data Mining

Data mining is a process that employs one or more machine learning techniques to automatically analyze and extract knowledge. Other definitions include induction-based learning is the process of forming general concept definitions which are carried out by observing specific examples of the concepts to be learned. Knowledge Discovery in Databases (KDD) is the application of scientific methods to data mining. In this context data mining is one step in the KDD process [1]

Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequent knowledge from large databases [2].

Most of the data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures [3]. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases [4]. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain [3]. Several attempts have been made to design and develop the generic data mining system but no system found completely generic [3].

Data mining can be divided into several groups with tasks that can be carried out as follows [5] :

a. Description: used to describe patterns and trends contained in the data [5]
b. Estimation: the model uses a complete record that provides the value of the target variable as a predictive value [5]
c. Prediction: the predicted value of the results will be in the future [5]

d. Classification: a case in point is the classification of income can be separated into three categories, namely high income, medium income, and low income [5]
e. Clustering: is a grouping of records, observations, or pay attention and form classes of objects that have similarities [5]
f. Association: The task of association in data mining is to find attributes that appear at one time. In the business world it is more commonly called shopping basket analysis [5]

In data mining there are several stages, that are [6] :

1. Data Selection: Selection of data from a set of operational data. The selected data will be used for the data mining process, and stored in a separate file and operational database [6]
2. Pre-processing / cleaning process is done by removing noise, removing duplicate data, checking inconsistent data, and correcting errors in the data, such as typographical errors. [6]
3. Data Transformation This stage is the process of transforming data that has been selected, so that the data is suitable for the data mining process [6]
4. Data mining is the process of finding very interesting patterns or information in selected data using certain techniques or methods [6]
5. Interpretation / Evalution This stage involves checking whether the pattern or information found is contrary to the facts or hypotheses that existed before [6]

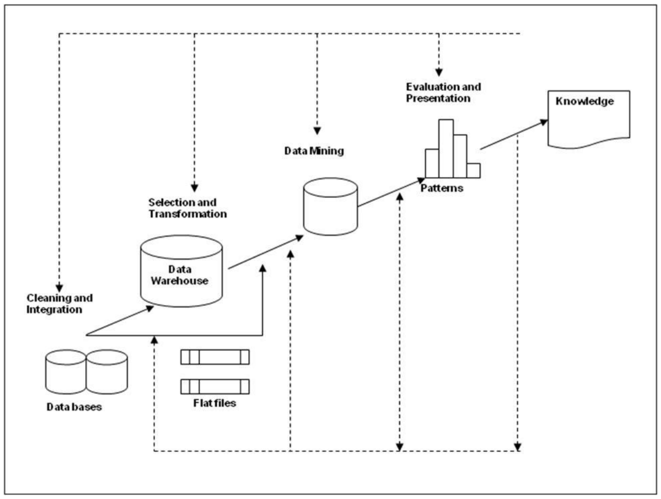

Figure 1 : Data Mining Steps

2.2 CRISP-DM Methodology

Cross-Industry standart Process for data mining or CRISP-DM, which is a sequence of six steps that starts with a good understanding of the business and th need for the data mining project (i.e, the application domain) and ends with the deployment of the solution that satisfied the specific business need. CRISP-DM is a standardization of data mining processes as a general problem solving strategy of a business or research unit [7]

In general the CRISP-DM life cycle is as shown below (figure 2) :

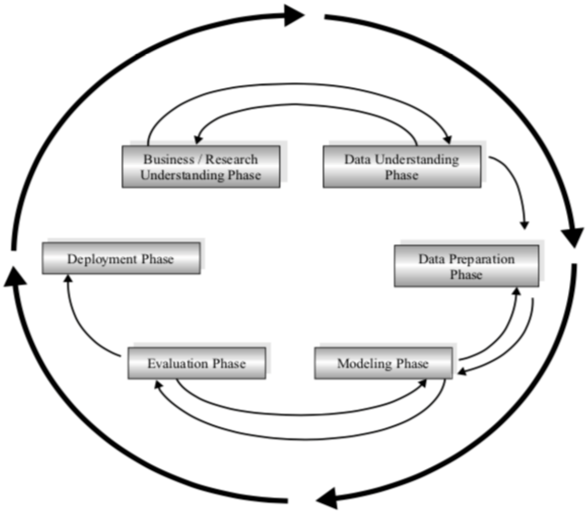

Figure 2 : CRISP-DM Lifecycle [4]

Steps of CRISP-DM Methodology [4] :

1. Business understanding phase
   a. Detailed project objectives and needs in the overall scope of the business or research unit.
   b. Translating goals and boundaries into formulas and definitions of data mining problems.
   c. Prepare an initial strategy for achieving goals.

2. Data understanding phase
   a. Collecting data.
   b. Using data analysis investigations to further identify data and search for the initial knowledge in it.
   c. Evaluating data quality.
   d. If desired, select a small group of data that might contain patterns of problems.
3. Data preparation phase
   a. Prepare data that is available from the beginning, because it is a data set that will be used for the entire next phase. This phase is heavy work that needs to be carried out intensively.
   b. Select the cases and variables that you want to analyze and those that are in accordance with the analysis you want to do.
   c. Make changes to several variables if needed.
   d. Prepare initial data so that it is ready for the modeling device.
4. Modelling phase
   a. Select and apply appropriate modeling techniques.
   b. Calibrate model rules to optimize results.
   c. It should be noted that several techniques may be used for the same data mining problem.
   d. If needed, the process can return to the data processing phase to make the data into a form in accordance with the requirements specifications of certain data mining techniques.
5. Evaluation phase
   a. Evaluate one or more models used in the modeling phase to get quality and effectiveness before they are deployed for use.
   b. Determine whether there is a model that meets the objectives in the initial phase.
   c. Determine whether there are important issues from business or research that are not handled properly.
   d. Making decisions relating to the use of results from data mining.
6. Deployment phase
   a. Using the resulting model. The formation of the model does not indicate the completion of the project.
   b. A simple example of deployment is report generation.
   c. A complex example of deployment is the application of data mining processes in parallel to other departments.

2.3 C4.5 Algorithm

C4.5 algorithm is an algorithm developed by J. Ross Quinlan in 1993 [8]. The C4.5 algorithm is a continued development of the previous algorithm, namely the ID3 algorithm. Therefore, the actual ID3 and C4.5 algorithms have the same basic principles. Some developments are done on the algorithm C4.5 algorithm that makes the C4.5 different from its predecessor, that is [8] :

- Ability to handle attributes with discrete or continuous type [8]
- Ability to handle empty attribute (missing value) [8].
- Can do pruning on branches [8].
- The selection is done using a calculation attribute Gain Ratio [8]

Here are the three principles of the work done by the C4.5 algorithm according to [8] :

- First, perform decision tree construction. The purpose of this decision tree construction algorithm is to create a model of a set of training data that will be used to predict the class of a new data [8]
- Second, the decision tree pruning. Since the results of decision tree construction can be bulky and not easy to "read", the C4.5 algorithm can simplify the decision tree with pruning based on the value of the level of confidence. Pruning also aims to reduce the prediction error rate on new data [8]
- Third, making the rules for the decision tree that has been constructed. The rules are in if-then form that derived from the decision tree by tracing from the root node to the leaf node [8]

Basic algorithms used by the C4.5 algorithm for decision tree induction is a greedy algorithm that builds decision tree from top to bottom (topdown) recursively by divide and conquer [8]

2.4 Decision Tree

A tree is a data structure that consists of nodes and edges. Nodes in a tree can be divided into three, namely the root node, branching node / internal branch / internal node and leaf node [8].

Decision trees (figure 3) are simple representations of classification techniques for a finite number of classes, where both internal nodes and root nodes are marked with attribute names, their ribs are labeled as possible attribute values and leaf nodes are marked with different classes [9].

Decision tree classification technique is a supervised learning. The class labels or categories are already defined in the beginning and in the process of making a model using the training data to classify new data. Decision tree itself consists of several parts of the node [9] :

- Root node, a node that is at the top of the tree, this node has no incoming branches and has more than one branch; sometimes it does not have a branch at all. This node is usually the most attributes that have the greatest influence on a particular class [9].
- Internal node, a branching node that only has one incoming branch, and has more than one branch coming out [9].
- Leaf node, an end node that only has one incoming branch, and has no branches at all. It also marks the node as a class label [9]

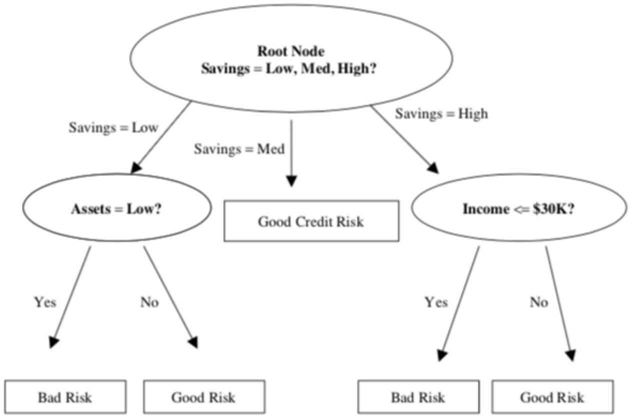

Figure 3 : Simple Decision Tree [10]

3. RESEARCH METHODOLOGY

The flow in this study, adjusts to the stages of the CRISP-DM method. Understanding at each stage refers to the research conducted, but of course there are some differences in the objects and research variables.The following is attached to the research flow which is used as a guide and direction of the research conducted, based on the object under study, and the needs of each.

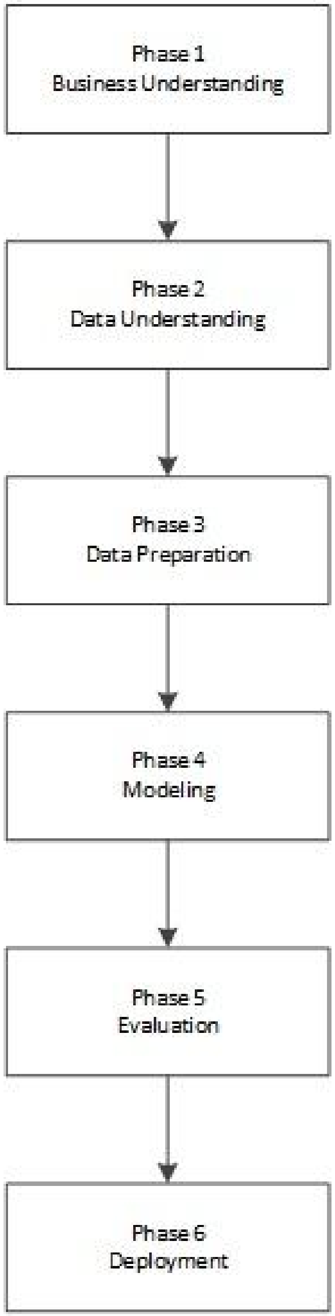

Figure 4 : Steps CRISP-DM Methodology

4. RESULT AND ANALYSIS

This research uses the process contained in the CRISP-DM method, but does not carry out the final process, namely the implementation process.

4.1 Discussion

4.1.1 Business Understanding

Stages of business understanding focuses on understanding the purpose of the needs based on business valuation. Then the understanding is transformed into an initial data mining plan designed to achieve the objectives. Business understanding refers to the rules for determining credit collectability in bank ABC. At this stage an understanding of the background and objectives of the business process is required in relation to the assessment of credit quality:

1. Determine Business Objectives. The business objective of conducting research is to recognize customer quality patterns to find out predictions or analyze customer credit quality
2. Assess the Situation. Bank ABC core banking deals with credit customers Bank ABC. Core banking is a system developed to carry out transactions and customer data processing that is customer data, savings data, credit data, etc.
3. Determine the Data Mining Goals. The purpose of data mining or the purpose of this research is discovering knowledge about pattern credit classification related to customer quality through current or non-performing collectibility.

4.1.2 Data Understanding

At this stage of understanding this data begins with initial data collection to determine the first insight into the data. Initial data collection was carried out by literature study and observation. In the observations obtained random sample data (table 1) that is :

Table 1 : Sample data credit customer Bank ABC

No	Income	Loan	Time	Collectibility
1	Small	Medium	Short	Bad
2	Medium	Large	Short	Current
3	Small	Small	Short	Current
4	Small	Medium	Medium	Bad
5	Small	Small	Short	Bad
6	Small	Small	Short	Bad
7	Large	Medium	Short	Current
8	Small	Medium	Short	Bad
9	Medium	Medium	Short	Current
10	Small	Medium	Long	Current

4.1.3 Data Preparation

The data preparation stage includes all activities that build the final dataset (data to be included in the modeling) from the initial raw data. Data preparation includes all activities to build a data set that will be included in the modeling tool from the initial raw data or create a new database for data mining setups. Data preparation includes all activities to build a data set that will be processed in the modeling process using the C4.5 algorithm to build a decision tree.

4.1.4 Modeling

The choice of data mining techniques, algorithms and determining parameters with optimal values. At the modeling stage, there are several things to do that is select modeling techniques, building models and Asses models.

1. Select Modelling Technique. The data mining technique chosen is the decision tree using the C4.5 algorithm. Decision tree and C4.5 algorithm are very appropriate to be used to achieve the initial objective of this study, which is to gain knowledge about the classification of active credit customers. Data mining modeling begins with making rules for the formation of decision trees.
2. Building Model which will be used as a benchmark in the classification of customers who are smooth or non-current. Customer assessment criteria become one of the benchmarks in classifying credit customers, namely the amount of income earned each month, the amount of credit proposed and the time / duration of the credit collection schedule by the customer.
3. Asses Model, modeling is done by forming a decision tree using the C.45 algorithm with predetermined rules Of the three conditions, 30 random customer data will be used.

4.1.5 Evaluation

The evaluation in this study is more focused on the model or pattern produced by the C4.5 algorithm. The resulting model is analyzed to determine whether the resulting pattern is in accordance with the standards contained in the bank ABC.

4.1.6 Deployment

Evaluation is focused on the patterns generated by the C4.5 algorithm. The resulting model is analyzed to determine whether the resulting pattern is in accordance with the standards contained in Bank ABC. If the resulting pattern is not appropriate, then further analysis of the resulting pattern can result in recommendations for improvements expected in determining whether credit customers are rejected or accepted, in order to avoid the risk of default by credit customers. For the stage of spreading data mining in this study was not done.

4.2 Research Result

Formation of a decision tree using the C4.5 algorithm to solve problems. The first thing that needs to be done is to calculate the number of quality of credit customers who experience traffic jams and smoothly from all cases and cases that are divided based on the total customer income attribute, the amount of the loan amount and the length of credit period.
Decision Tree

After testing the C4.5 algorithm method on weka, a decision tree is formed like Figure 5

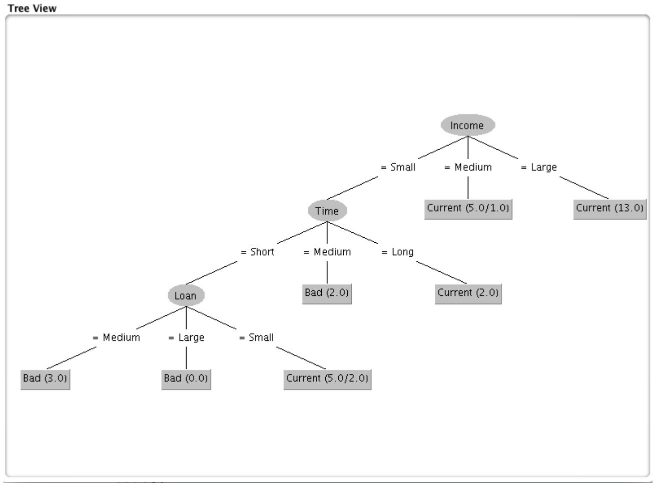

Figure 5 : The root Of Decision Tree

In Figure 5 the root of the decision tree is Revenue. The decision tree above produces rules that will be implemented in the program. The following is an explanation of the rules formed by modeling the C4.5 algorithm :

1. If income "medium" and income "large", then collectibility = Lancar
2. If income "small", time "long" then collectibillity = current.
3. If income "small" and time "medium" then collectibillity = bad.
4. If income "small", time "short", loan "medium" and "large" then collectibillity = bad.
5. If income "small", time "short" and loan "small" then collectibillity = current.

5. CONCLUSION

Based on the results of the research conducted the following conclusions can be drawn :

1. The CRISP-DM methodology can be applied to predict the credit classification classification.
2. C4.5 algorithm is used to make decision tree formation
3. Bank ABC will find out more quickly about the classification of prospective credit customers so that it will have a good impact on the company, that is reducing the risk of default by credit customers in the future.

REFERENCES

[1] Turban , Efraim & Aronson, Jay E. **Decision Support Systems and Intelligent Systems**. 6th edition. Prentice Hall: Upper Saddle River, NJ, 2001

[2] Sharda, Rames , Delen, Dursun , Turban, Efraim, **Business Intelligence, Analytics, and Data Science**, 4th ed. New York, Pearson , 2013

[3] Deshpande, Shrinivas & Thakare, V. M. **Intelligence Ingrained Data Mining Engine Architecture.** International Journal of Computer Technology and Applications. 02, 2011

[4] Larose, D. T, ***Discovering Knowledge in Data*** New Jersey: John Willey & Sons, Inc, 2014 https://doi.org/10.1002/9781118874059

[5] Fayyad, U, **Advances in Knowledge Discovery and Data Minin,. MIT Press**, 1996

[6] Larose, D. T, **Discovering Knowledge in Data: An Introduction to Data mining**. John Willey and Sons, Inc, 2005

[7] J. R. Quinlan, C4.5: **Programs for Machine Learning**, USA: Morgan Kaufmann, 1993

[8] Ananda, David & Wibisono, Ari, **C4.5 Decision Tree Implementation In Sistem Informasi Zakat (SIZAKAT) To Automatically Determining The Amount Of Zakat Received By Mustahik**. Jurnal Sistem Informasi. 10. 28. 10.21609/jsi.v10i1.375, 2014

[9] Hermawati. F. Astuti, **Data Mining**. Yogyakarta: Andi Offset, 2013

[10] P.-N. Tan, M. Steinbach and V. Kumar, I**ntroduction to Data Mining, 1st ed**., Boston: Pearson Addison-Wesley, 2006.

[11] Dhankhar Amita, Solanki Kamna. **A Comprehensive Review of Tools & Techniques for Big Data Analytics**. *International Jurnal Emerging Trends in Enginering Research* 7. 556-562. 10.30534/ijeter/2019/06852020, 2019

[12] Amos Pah, Clarissa & Utama, Ditdit. **Decision Support Model for Employee Recruitment Using Data Mining Classification**. *International Jurnal Emerging Trends in Engineering Research* 8. 1511-1516. 10.30534/ijeter/2020/06852020, 2020