CHAPTER II

THEORITICAL BACKGROUND

2.1. Related Study

To prove that this research area is quite important in the business activity field and also for academic purpose, these are some of related study that has been conducted. It would be a references to conducting this study about "Implementation of Naïve Bayes Algorithm to Determine Customer Credit Status in PT. Multindo Auto Finance Semarang".

First, is the research conducted by Angga Ginanjar Mabrur and Riani Lubis about Implementation of Data Mining to Predict Credit Customer Criteria. The purpose of this research is to help and solve the problem in investigating customer data by developing data mining application to predict customer credit criteria that potentially do a new application for credit in bank. In this research, the authors choose bank XY located in Bandung to applying the research.

Data mining technique that the author used for this research is classification technique. Moreover, for the method, the authors choose to use Decision Tree with C4.5 algorithm to create the decision tree. Then, for the data used in this research is the credit instalment payment data from bank XY on June, 2009. This data is provided in Microsoft excel format. The original data set given by the bank consist 27 fields. Although, in this research the author select only 8 parameters to be used which are Gender, Age, Village, *Kecamatan*, Marital_Status, Loan_Amount, Installment, Integration_Code.

The result of this research is an application that can help financial division in bank XY to fulfill marketing target in the future. The data mining application produced is described in the research using use case diagram, scenario use case, activity diagram, sequence diagram, and class diagram.

Second, the research conducted by Henny Leidiyana about Implementation of K-Nearest Neighbor Algorithm to Determine Credit Risk in Motorcycle. In this research, the author applied k-Nearest Neighbor (kNN) Algorithm on customer data who used motorcycle credit finance. Then, to measure the performance of the algorithm, the author used cross validation, confusion matrix and ROC curve from the test result.

The data used in this research are provided by a leasing company in Cikarang. The original data set provided consists of 558 record and 14 attributes. The attributes consists of 13 predictor and 1 label. To calculate the distance from each attributes, each attribute has its own integrity. The integrity distance given is a value between 0 and 1. 0 means the attribute is not really affected to the result of classification, while, 1 means the attribute is really affected on the result of classification. After calculating the distance value, then the test result will be tested again to check the performance of the algorithm using cross validation, confusion matrix, and ROC curve. In this research, it is known that the accuracy value provided is 81.46% and the AUC value of 0.984. It is include as excellent classification because the AUC value is between 0.90 – 1.00.

Third, the research conducted by Claudia Clarentia Ciptohartono about Naïve Bayes Classification Algorithm for Credit Assessment. In this research, the author used Naïve Bayes model to identify useful information from a large size of data. The author chooses Naïve Bayes model because this model holds the assumption that the relationship between the features or attributes are independent, which make it simple and efficient.

The result of this research proves that Naïve Bayes algorithm can be applied to assess the credit data. The data used in this research is provided by bank, which is BCA Finance Jakarta. In this research, the author used 14 attributes to perform Naive Bayes algorithm. These are Gender, Customer_job, Marital_Status, City, Merk, Year, Tenor, OTR, DP, Insurance, Income, Interest, disposal, and payment.

After the experiment conducted, it provided the result that the credit assessment accuracy on BCA Finance Jakarta using the initial data set with preprocessing provide 85.57% accuracy. Moreover, after initial data processing and preprocessing it provides 92.53% accuracy. From the result provide that credit assessment using Naïve Bayes algorithm for BCA Finance Jakarta is superior if initial data are preprocessed. It is shown that the preprocessing is a step that greatly affects the final result to get an excellent category for its accuracy.

Fourth, the research conducted by I Gusti Ngurah Narindra Mandala, Catharina Badra Nawangpalupi, and Fransiscus Rian Praktikto about Assessing Credit Risk: an Application of Data Mining in a Rural Bank. The purpose from this research is to identify factors which are necessary for a rural bank to assess credit application. A decision tree model is proposed by applying data mining methodology in this research.

This credit assessment model is applied to PT. BPR X in Bali. This company has alarming performance where the NPL is much higher than 5% (11.99%). This research is carried to evaluate the credit assessment criteria to suggest better model to lower NPL rate. This research used CRISP-DM as a standard procedure of data mining in business field. For the algorithm, it used CART and C5.0 algorithm.

The data used in this research consists of gender, age, credit amount, monthly income, expenses each month, current payment per month, saving, collateral types, collateral value, loan period, type of business activities, source of funding, and previous credit status / rating. As the result, proposed model using decision tree provide lower NPL rate which is 3%.

Fifth, research conducted by Weimin Chen, Guocheng Xiang, Youjin Liu, and Kexi Wang about Credit risk Evaluation by hybrid data mining technique. This research is proposed a hybrid data mining technique with two processing stages. It is combining clustering and classification technique of data mining.

The algorithm used for clustering is k-means cluster, while for classification it used support vector machines classification. For the data used in this research, it is provided by a local bank in China. The data contains fifteen predictor variable which are country / state, certificate, gender, native place, age, marital status, number of dependents to support, educational level, telephone, residential status, occupation, ownership of employment establishment, annual income, a client of the bank or not, and credit limits.

As the result of this research, in the clustering stage, accepted credit and new applicant credit are grouped into homogeneous cluster. Then, in classification stages, support vector machines use the sample with new labels for building the scoring model. Credit scoring models proposed in this research is different with the other credit scoring because the samples were classified into three or four classes rather than two classes which is god and bad credit.

Table 2.1 State-of-The-Art

Title	Year	Attribute	Method	Result
Implementation of Data Mining to Predict Credit Customer's Criteria [7]	2012	Gender, Age, Marital_Status, Loan_Amount, Installment, Integration_Code, Village, *Kecamatan*	Decision Tree C4.5	Experiment and analysis in this research provide an application that can help fund section in bank or finance company to fulfill marketing target in the future

Implementation of K-Nearest Neighbor Algorithm to Determine Credit Risk in Motorcycle [4]	2013	Marital status, j*umlah tanggungan, pendidikan terakhir*, age, *kepemilikan rumah, lama tinggal, kondisi rumah*, kind of work, company status, employment status, work period, income, first payment	K-Nearest Neighbor	In this research, the kNN algorithm is used in customer credit data. To providing a good result preprocessing data is implemented. To measure the performance of the algorithm, cross validation, confusion matrix, and ROC curve is implemented and providing 81.64% of accuracy
Naïve Bayes Classification Algorithm for Credit Assessment [8]	2014	Gender, Customer_job, Marital_Status, City, Merk, Year, Tenor, OTR, DP, Insurance, Income, Interest, disposal, payment	Naïve Bayes Classifier	Credit assessment accuracy in BCA Finance Jakarta using the initial data set with preprocessing provide 85.57% accuracy.

				Moreover, after initial data processing and preprocessing it provides 92.53% accuracy
Assessing Credit Risk: an Application of Data Mining in a Rural Bank [9]	2012	gender, age, credit amount, monthly income, expenses each month, current payment per month, saving, collateral types, collateral value, loan period, type of business activities, source of funding, and previous credit status / rating	CART, Decision Tree, C5.0	Proposed model using decision tree in this study provide lower NPL rate which is 3%. It is much lower than the current rate.
Credit risk Evaluation by Hybrid Data Mining Technique [10]	2012	country / state, certificate, gender, native place, age, marital status, number of dependents to support, educational level, telephone, residential status,	K-means clustering, Support Vector Machines	Credit scoring models proposed in this research is different with the other credit scoring because the samples were classified into three or four classes rather

		occupation, ownership of employment establishment, annual income, a client of the bank or not, and credit limits		than two classes which is god and bad credit

2.2. Literature Review

2.2.1. Credit

The term of Credit is derived from Greek word which is *"Credere"* that means trusty / faith. Because of that, the most fundamental thing in credit is the trust from the one who give a credit, both personal or company, to the credit applicants [11]. In credit activity, there are two different side who involved, which are the supplier and the borrower (client / customer). The borrower is the side who will ask the supplier to give the loan to purchase their goods. The supplier is the side who will provide the money as a loan for the borrower (client / customer).

Based on the Indonesian Banking Regulation No. 10 in 1998, the definition of credit is an activity for supplying money or resources based on the agreement from the borrower (customer / client) and the supplier that require the borrower to repay the money and the amount of interest to the supplier in a period of time. Moreover, in Banking Principle Statute No. 14/67, Credit is a money supplying based on the loan agreement between the bank and the borrower then the borrower has a responsibility to repay the loan in the period of time and also the amount of interest that has been decided.

In credit activity there are some aspects to be concerned before the supplier give a credit to the borrower, these are [11]:

1. Trust

   Trust is a belief of the supplier (bank / company) that the credit that has been given to the borrower will be fully repaid in period of time and also the interest in the future.

2. Time Period

   In every credit activity both the supplier and the borrower in giving and returning are limited by period of time based on the agreement.

3. Degree of Risk

   Giving a credit means that it may create a problem and risk. Risk may appears for the supplier because the supplier has been given a money / service / good to the borrower.

4. Achievement / Performance

   Previously, achievement in credit can be represent as goods, services or money. But, by the improvement of credit activity in modern era, achievement gives in the credit means giving money to borrower.

Credit can be categorized to many categories. Based on decision letter from direction of Bank Indonesia (BI) No. 32/268/KEP/DIR in February 27th, 1998, credit can be classified to be:

1. Fluent Credit / Current Credit

   Means that the repayment of the loan and also the interest of credit is on time. Fluent credit can be indicates by the improvement of bank account, there is no instalment remainders and it appropriate with the credit requirement.

2. Substandard Credit

   It is a credit that have a remainders on the instalment within 90 days until 180 days from the due date that has been agreed for the repayment of the loan and also the credit interest.

3. Doubtful Credit

   It is a credit that have a remainders on the instalment within 180 days until 270 days from the due date that has been agreed for the repayment of the loan and also the credit interest.

4. Bad Credit / Loss Credit

   It is a credit that have a remainders on the instalment pass over 270 days from the due date that has been agreed for the repayment of the loan and also the credit interest.

2.2.2. Data Mining

There are a lot of understanding about data mining definition. Data mining is the analysis of observational data sets to find data relationship and summarize the data to make it understandable and useful to the data owner [12]. In other definition, data mining is the process of discovering new correlations, patterns and trends through large amounts of data that stored in repositories using pattern recognition technologies like statistical and mathematical techniques [13].

Based on some the definition that has been mentioned above, data mining in general is the technique to discover knowledge from data. Data mining is used to retrieve or extract an interesting (non-trivial, implicit, previously unknown and potentially useful) patterns from huge amount of data. Data mining also has some alternative names that like Knowledge Discovery in Database (KDD), knowledge extraction, data / patterns analysis, data archeology, data dredging, information harvesting, etc.

There are some aspects that make data mining technique is growing rapidly in current technology era's [13]:

1. Fast improvement of data collection
2. The storing of data in data warehouse, so all of the entire enterprise has an access into reliable database.
3. The improvement of data access through web navigation and intranet.
4. Business competition for increase market share in a global economic.
5. The improvement of software tools to implement data mining techniques.
6. Rapid development of computational skill and storage.

Based on the analysis task, data mining is divided into some of groups, which are [13]:

1. Description

   Researcher and analyst are plainly to find a way to describe pattern and trend analysis from the data. Description is the method used to perform this kind of data mining approach and give the description about how to explain certain pattern or trend.

2. Estimation

   Estimation is similar to classification. The different between estimation and classification is located on the use of variable. In estimation, target variable used is more likely numerical variable rather than categorical variable. The model in estimation are built from complete record which provide the value of the target variable as a predictor.

3. Prediction

   Prediction is similar to classification and estimation. In prediction, the value of the result lie in the future. One example of prediction tasks in business and research is predicting the price of rice based on the stock for the next three month.

4. Classification

   In classification, there is a target categorical variable. The example of classification is income bracket which is can be classified into high income, middle income and low income.

5. Clustering

   Clustering refers to the grouping of records, observation or creating classes of similar objects that has the same characteristic. A cluster means collection of records that has the same characteristic each other, but it is different from the other records in other cluster. Clustering differs from classification in that there is no target variable for clustering.

6. Association

   Association in data mining means finding the relation based on the appearance of attributes at the same time. In business field, association is commonly known as market based analysis. The task of association is to seeks uncover rules for quantifying the relationship between two or more attributes.

2.2.3. Data Mining Process

Data mining and Knowledge Discovery in Database (KDD) have a different concept, they are connected each other and both of them are relevant. One of KDD process is data mining. These are the following KDD process in general [14]:

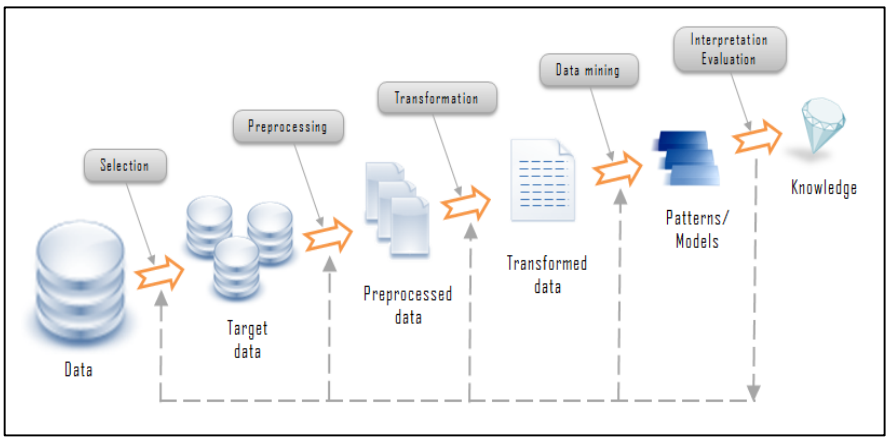


Figure 2.1 Knowledge Discovery in Database (KDD) Phase

1. Data Selection

   Data selection means creating a target data set. Data selection process is needed to get potential data from all of operational data to be retrieved as target data set in KDD phase. Selected data from data selection process will be used in data mining process and it should be saved in one directory that is different from the operational basis data.

2. Pre-processing / Data Cleaning

   Data cleaning process is needed to be done in selected data / target data set before performing data mining process. Data cleaning process are including remove redundant data, check inconsistent data, data reparation e.g. typography, and data enrichment. In data enrichment, it is possible to add the target data set with any others relevant information from external information.

3. Transformation

   Coding is data transformation process to be implemented in target data set, so this target data set can be appropriate for data mining process. In KDD phase, coding is a creativity process and it is strongly dependent to the kind of information pattern that going to be search.

4. Data Mining

   Data mining is a process to find a pattern or information from target data set that has been selected using specific technique or method. There are many various technique, method, and algorithm used in data mining. Choosing the appropriate method or algorithm in data mining phase is depend on the purpose and overall KDD phase.

5. Interpretation / Evaluation

   Interpretation is KDD phase to show information pattern as the result from data mining process in understandable form. This phase including checking for the result pattern or information retrieved from data mining process whether it is relevant or not with the previous hypothesis.

2.2.4. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed by industrial analysts in 1996. CRISP-DM provides a standard data mining process as a general problem solving strategy of a business or research unit. According to CRISP-DM, data mining life cycle is divided into six phases [13]:

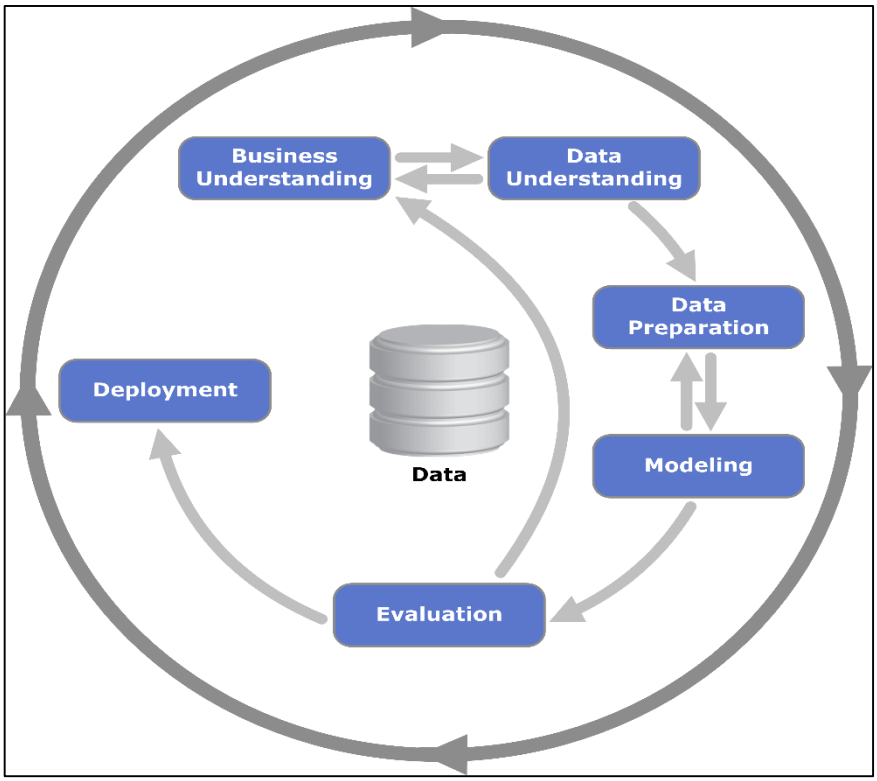


Figure 2.2 CRISP-DM

1. Business Understanding Phase

   In this phase, the understanding about the purpose of data mining project that would be done is needed. The activities conducted in this phase includes finding the purpose and the target of business, finding the detail requirement in business field, and preparing the strategies to reach the goal.

2. Data Understanding Phase

   Phase for collecting the data and analyze the data to be recognizing the data that will be used. This phase is used for doing data analysis and discover initial insights about the data.

3. Data Preparation Phase

   Preparing a collection of data that would be used in the next phase. This phase is a difficult work and need to be done intensively. This phase is used for preparing the variable to be transformed into database and perform a changing variable if needed.

4. Modeling Phase

   This phase is used to selecting and applying the appropriate modeling technique and calibrate the model setting to optimize the result.

5. Evaluation Phase

   Deep evaluation is needed for adjusting the model to be appropriate with the target in the first phase. Evaluation phase used for evaluating the model to provide the quality and the effectiveness before deployment.

6. Deployment Phase

   Creating a report of the result model and the implementation of data mining.

2.2.5. Naïve Bayes Algorithm

Bayesian classifier are statistical classifier that can predict class membership probabilities based on Bayes theorem [12]. Naïve Bayes Classification algorithm perform the classification function similar as decision tree and neural network. The Naïve Bayes Classifier works as follows [12]:

1. Let D is a training set of tuple and associated with the class label. Each tuple is represented by an *n*-dimensional attributes vector, $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$, where *n* are illustrated and measured on the tuple from *n* attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. There are *m* classes which is set of categories, $C_1, C_2, \ldots, C_m$. Given a tuple $\boldsymbol{X}$, the classifier will predict $\boldsymbol{X}$ belongs to the class with the highest posterior probability in $\boldsymbol{X}$. The Naïve Bayes Classifier predicts that tuple $\boldsymbol{X}$ belongs to the class C_i if and only if

   $P(C_i \mid X) > P(C_j \mid X)$ for $1 \leq j \leq m, j \neq i$

   Maximize $P(C_i|X)$. Maximized class C_i for $P(C_i|X)$ is called the maximum posteriori hypothesis

   $$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)}$$
3. $P(X)$ is a constant for all classes, only $P(X|Ci)$. $P(Ci)$ needs to be maximized. If the class prior probability are not known, then it is commonly assumed that the classes are equal with other categories like $P(C_1) = P(C_2) = \ldots = P(C_m)$, because of that, $P(X|Ci)$ needs to be maximized. Need to be noted that class prior probability can be predicted by calculate $P(C_i) = \frac{S_i}{S}$, where S_i is amount of data training S from C_i categories and S is total amount of data training.
4. Data sets with many attributes would be extremely computationally to compute $P(X|C_i)$. To reduce computation in evaluating $P(X|Ci)$, it can be used by this calculation:

   $$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

   $$= P(X_1|C_i) \times P(X_2|C_i) \times \ldots\ldots \times P(X_n|C_i)$$

   We can easily estimate the probabilities $P(X_1|C_i), P(X_2|C_i), \ldots\ldots, P(X_n|C_i)$ from the training tuples.

5. To predict class label of ***X***, *P(X|Ci). P(Ci)* is evaluated for each class *Ci*. The classifier predicts that the class label of tuple ***X*** is class *Ci* if and only if
   $P(X \mid C_i).P(C_i) > P(X \mid C_j).P(C_j)$ for $1 \leq j \leq m, j \neq i$
   In other word, the predicted class label is the class *Ci* for which *P(X|Ci). P(Ci)* is the maximum.

The main advantages of using Naïve Bayes Classification are [15]:

1. Robust to isolated noise in data
2. Robust to irrelevant attributes
3. In case of missing values, it will be ignores the corresponding objects during the process of computing probabilities.

2.2.6. Confusion Matrix

Confusion Matrix is a tool used for model evaluation classification to predict an object which is true or not [12]. A prediction matrix will be compared with the original input class. In other word, confusion matrix consists of actual information and prediction in classification.

Table 2.2 Confusion Matrix Table with 2 classes

Classification	Predicted Class	
	Class = Yes	Class = No
Class = Yes	a (true positive-TP)	b (true negative-TN)
Class = No	c (false positive-FP)	d (false negative-FN)

Formula to calculate the accuracy level in confusion matrix is:

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}$$

Classification level can be divided into some categories which are:

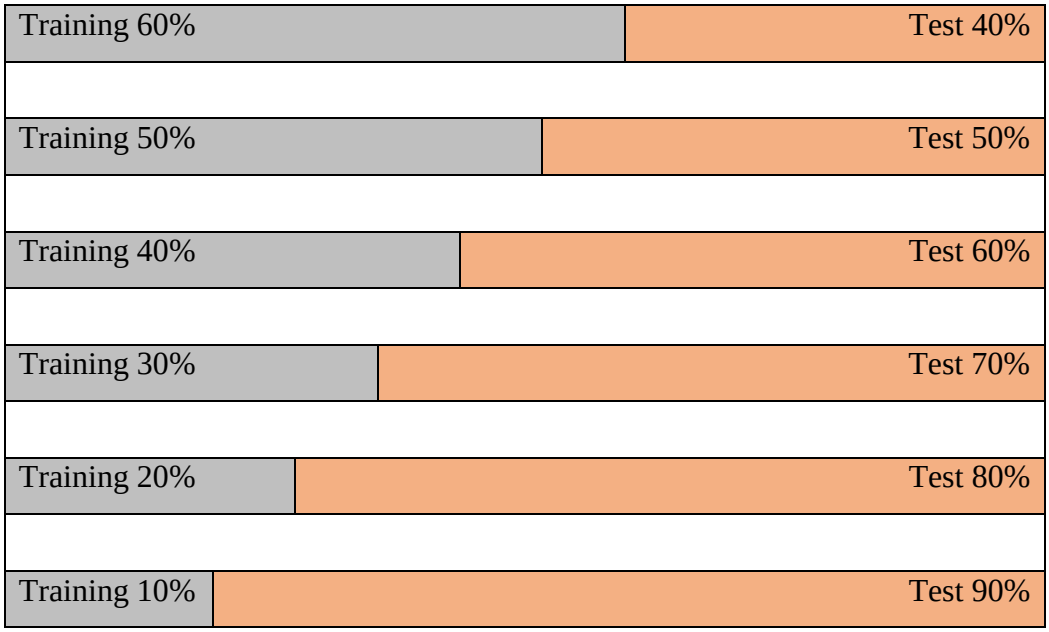
- 0.90 – 1.00 accuracy = Excellent classification
- 0.80 – 0.90 accuracy = Good classification
- 0.70 – 0.80 accuracy = Fair classification
- 0.60 – 0.70 accuracy = Poor classification
- 0.50 – 0.60 accuracy = Failure

2.2.7. Split Validation

Split validation is a validation technique by dividing the data set into two different part randomly, which are data training and data testing. By using split validation, it will conduct a training based on the split ratio that has been decided before. Then, the rest of data from split ratio in data training will used as a data testing. Data training is a set of data used to learning process. Moreover, data testing is a set of data that have not been used in learning and it will be used as a data testing in providing accuracy result [16].

Table 2.3 Split Validation Illustration

Training 90%	Test 10%
Training 80%	Test 20%
Training 70%	Test 30%



2.3. Framework of Study

In this study, framework of study is needed for keeping the study to following each phases in order and consistent. The problem happened in PT. Multindo Auto Finance is the demand of credit applicants but there is no classification for credit customer itself, so the need of credit customer classification is becoming main focus on this study to determine customer credit status.

In this study, Data Mining Approach is used by implementing Naïve Bayes Algorithm on customer data to do a classification. Moreover, CRISP-DM model is used as the design in this study and RapidMiner tools as application model. After performing Naïve Bayes Classifier (NBC) Algorithm, split validation and confusion matrix are used as testing result evaluation from the work of Naïve Bayes Classification Algorithm (NBC).

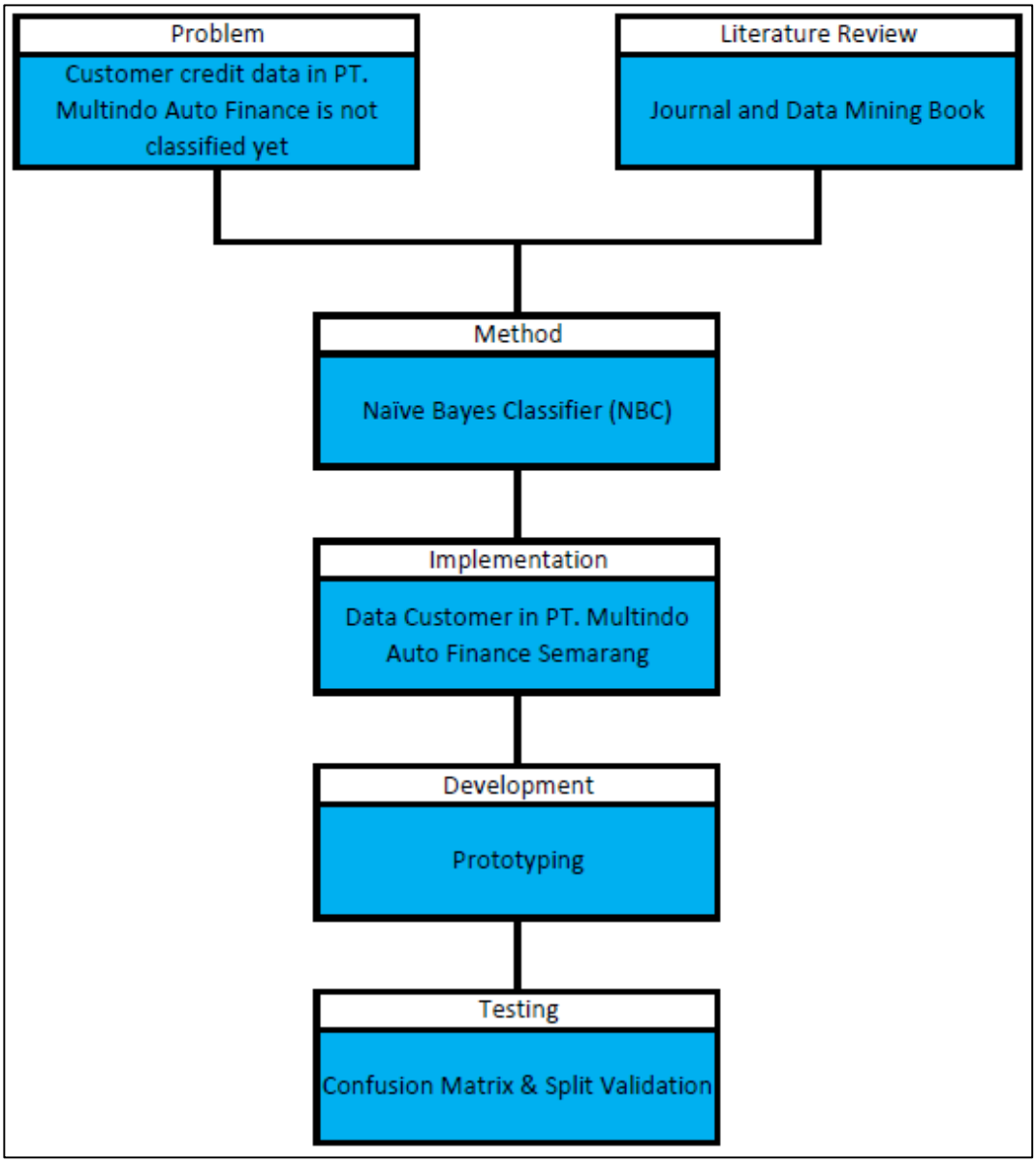


Figure 2.3 Framework of Study