IMPLEMENTATION OF NAÏVE BAYES ALGORITHM TO DETERMINE CUSTOMER CREDIT STATUS IN PT. MULTINDO AUTO FINANCE SEMARANG

Muhammad Tosan Bingamawa[1], Heru Agus Santoso[2]

[1]Faculty of Computer Science, Dian Nuswantoro University
Email : 111201206940@mhs.dinus.ac.id
[2]Faculty of Computer Science, Dian Nuswantoro University
Email : herezadi@gmail.com

Abstract

As a finance company, PT. Multindo Auto Finance Semarang is giving fast, appropriate, and flexible finance solution for people to own cars. Finance solution offered by PT. Multindo Auto Finance Semarang is form as a credit loan. With the demand of the credit applicants, customer classification to provide information about customer credit status is needed for PT. Multindo Auto Finance Semarang. It is because credit risk will always be possible. An example of credit problems that frequently occur in the credit activity is loss credit. By using data mining classification approach that implemented in customer credit data, it would be possible to overcome the credit problems in PT. Multindo Auto Finance Semarang. In this study, Naïve Bayes Classification Algorithm is performed for customer credit status categorization in PT. Multindo Auto Finance Semarang. Moreover, Cross-Industry Standard Process for Data Mining (CRISP-DM) and Knowledge Discovery in Database (KDD) phase are also performed for data processing technique. Experimental result of customer classification using customer credit data in this study provides the result of 91.29% accuracy. From this experimental result, system prototyping is developed for the visualization that can help PT. Multindo Auto Finance Semarang to predict the status of new credit applicants and also control their credit customer from any credit problems.

Keywords: *data mining, classification, naïve bayes, credit, finance company*

1. INTRODUCTION

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. Previously, bank is the most common party that people usually use to apply any credit transaction. However, in this era there are a lot of relevant parties that are concerned to the credit activity.

One of relevant parties that concerned in credit activity is a financing company. Finance company is a financial institution that provides the supply of credit to the customer for the purchase of goods and services by granting loans directly to the customer with a contract period [1]. PT. Multindo Auto Finance is one example of financial company that works based on credit system in automotive field.

With the demand of the credit applicants, PT. Multindo Auto Finance has a credit scoring model to determine which applicants who will get the loan. Credit scoring is used to categorize the applicant's credit weather it will be accepted or rejected by the applicants' characteristics [2]. Whether credit scoring is very helpful for the company, the use of credit scoring evaluation model in PT. Multindo Auto Finance is not quite effective. Moreover, PT. Multindo Auto Finance prefers to do a real and direct survey to the credit applicants.

Although the real survey has been conducted, the credit risks will always be possible. An example of problems that frequently occur in the credit activity is loss credit. The applicants who have received the loans can become unpredictable. It will happen during the period of time on the credit repayment. This situation also occurs in PT. Multindo Auto Finance. That is why classification for the customers is required to determine the customer's credit status and help the company to take action on the customer.

Currently, the data mining approach is commonly used for data processing in data analysis activity. Data mining refers to a method in processing large amounts of data to find hidden patterns and new knowledge or useful information [3]. Data mining has several methods used to process data in the data analysis activity. One of the data mining method is Classification. Classification is a supervised learning approach to find rules and divide the data into specific groups [3]. By using data mining classification approach, it would be possible to overcome the problems that occurs in credit transaction activities such as loss credit.

2. THEORITICAL BACKGROUND

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 [4]. 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 4 categories. Those are fluent credit /current credit, substandard credit, doubtful credit, and bad credit / loss credit.

2.2. Data Mining

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 [5]. Based on the analysis task, data mining is divided into some of groups, which are description, estimation, prediction, classification, clustering, and association.

2.3. Data Mining Process

Data Mining and Knowledge Discovery in Database (KDD) are connected each other and both of them are relevant. KDD consist of 5 steps which are data selection, preprocessing / data cleaning, transformation, data mining, interpretation / evaluation.

2.4. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a standard data mining process as a general problem solving strategy of a business or research unit.

CRISP-DM divided into six phases. Those are business understanding, data understanding, data preparation, modeling, evaluation and deployment.

2.5. Naïve Bayes Algorithm

Bayesian classifier are statistical classifier that can predict class membership probabilities based on Bayes theorem [6]. Bayes theorem formulation can be described as follows:

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)}$$

Where:

- X = evidence / data tuple
- Ci = hypothesis such as the data tuple X belongs to a specific class C
- P(Ci|X) = the probability that the hypothesis Ci holds given the evidence or observed data tuple X
- P(Ci) = the prior probability of Ci
- P(X|Ci) = the posterior probability of X conditioned on Ci
- P(X) = the prior probability of X

2.6. Confusion Matrix

Confusion Matrix is a tool used for model evaluation classification to predict an object which is true or not [6]. 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 1: 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

2.7. 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 [7].

3. RESEARCH METHOD

3.1. Research Instrument

In this study, the author used observation method to collect and get the data that will be used and performed in this study. Direct observation is conducted to the financing company which is PT.

Multindo Auto Finance to ask and request the data. During the observation, the company agrees and gives a confirmation that the data can be exported and used in this study. The data itself was taken from PT. Multindo Auto Finance Semarang on April, 2015.

3.2. Data Collection Method

There are two types of data used in this study. Those are:

a. Primary Data

Data provided by the company which is PT. Multindo Auto Finance Semarang. Consist of two kinds of data, which are customer data and customer transaction data.

b. Secondary Data

Data that supporting the primary data and arranging this study which are Journal about credit and Data Mining Books.

3.3. Data Analysis Technique

This study follows the steps of Cross-Industry Standard Procedure for Data Mining (CRISP-DM). CRISP-DM provides a general data mining process as a problem solving strategy in business environment. These are the following steps:

1. Business Understanding

PT. Multindo Auto Finance, based on its business license, is carries on business in the area of consumer financing. Until now, PT. Multindo Auto Finance is giving fast, appropriate, and flexible finance solution for people to own cars. Finance solution offered by PT. Multindo Auto Finance to the customer is form as a credit loan. Credit activity in PT. Multindo Auto Finance is similar to any credit activity conducted by bank. It may also have a credit problems e.g. loss credit. In PT. Multindo Auto Finance Credit can be classified into three types which are Regular Credit (on time / within 0-29 days late), Problem Account (30-59 days late), and Loss Credit (more than 60 days late).

2. Data Understanding

The customer data given by PT. Multindo Auto Finance is the data about customer profile that has been registered as a credit customer in this company. This is the main data used in this study. This customer data consists of 10264 data records with 12 attributes.

Table 2: Detail Attributes in Customer Data

NopinAll	Customer credit ID
RealisasiDate	Date for the company to realize the credit loan
MerkName	The vendor of the vehicle manufacturer
CategoryName	Type of the vehicle
Occupation	Customer daily work
Grossincome	Customer income in a month
Tanggungan	Number of people that become the responsibility of the customer
LoanType	Type of loan given by the company
AngsuranReal	Amount of interest that customer need to repay every month to the company
Tenor	Credit repayment time
Total pinjaman	Total amount that customer need to repay including the credit interest
OS PINjaman PER 31 DES 14	Rest of total amount that customer need to repay to the company per December 31th, 2014

While, for the customer transaction data is the data of customer credit repayment report in period of 2014. This data is an additional data used as a selector in data preparation phase. This data consists of 62747 records data with 7 attributes.

Table 3: Detail Attributes in Customer Transaction Data

ID	Customer Credit ID
Apldate	Date of the transaction recorded
RealisasiDate	Date for the company to realize the credit loan
HARITGK	Amount of late day for the customer to repay the credit
BucketDueJT	Time of late classification
AngsuranReal	Amount of interest that customer need to repay every month to the company
LastPaidDate	Date of the last time to repay the credit each month

3. *Data Preparation*

In this study, not all of the data will be used in the process. In this phase, data preparation is conducted to prepare the data before it is ready to be processed. Data selection process is conducted to get the final data in this study. After the selection process, 3253 records of customer data is selected as a final data / training set that will be used in this study. Moreover, attributes selection is also performed in this phase.

Table 4: Detail Attributes in Training Set

Attribute	Utilization Detail		
NopinAll	×	No	-
Status	√	Yes	Target variable
RealisasiDate	×	No	-
MerkName	√	Yes	Predictor variable
CategoryName	√	Yes	Predictor variable
Occupation	√	Yes	Predictor variable
Grossincome	√	Yes	Predictor variable
Tanggungan	√	Yes	Predictor variable
LoanType	√	Yes	Predictor variable
AngsuranReal	√	Yes	Predictor variable
Tenor	√	Yes	Predictor variable
Total pinjaman	×	No	-
OS PINjaman PER 31 DES 14	×	No	-

4. *Modeling*

The model proposed to be used in this study is NBC model. This model will be applied in RapidMiner application. Thus, accuracy checking in this study performed by using RapidMiner Ver.5.3.013 framework.

5. *Evaluation*

In this phase, validation and accuracy measuring from the result model are performed. In this study, validation process is performed using split validation with relative split type and stratified sampling type. Moreover, for calculating the accuracy of the result, confusion matrix is performed.

6. *Deployment*

The result of this study is an analysis that can be proposed as a Decision Support System (DSS) for the company which is PT. Multindo Auto Finance Semarang. This result can be used by the company to classified credit customers based on credit status in the credit repayment. Moreover, this result can also be used as a predictor by predicting the credit applicant status in the future.

4. ANALYSIS AND DISCUSSION

4.1. Data Processing for Training Set

With large amount of data provided by PT. Multindo Auto Finance Semarang, data processing is needed for selecting the data that is suitable to be used in this study. To get the final data or training set used in this study, KDD phases are performed. In KDD, several steps need to be done for data preparation of training set. Those are data selection, data integration, data reduction, and data transformation. After this data preparation phase, 3253 records of customer data are selected as training set with 9 attributes.

4.2. Validation and Evaluation

In this phase, the measurement accuracy of Naïve Bayes Classification Algorithm using Split Validation method is implemented. The design of Naïve Bayes Classification using split validation method in RapidMiner Ver.5.3.013 framework.

4.3. Testing and Result

First testing in implementing NBC algorithm use customer data that has been processed in the data selection, data integration, and data reduction phase. The amount of the data are 3253 records of customer with 9 attributes. Those attributes are status, MerkName, CategoryName, occupation, grossincome, tanggungan, LoanType, AngsuranReal, and Tenor. From the first testing, the classification results are shown in confusion matrix as follows:

Table 5: Confusion Matrix using Final Data

Classification	True lancar	True macet	True bermasalah
Pred. lancar	730	20	12
Pred. macet	191	8	8
Pred. berasalah	7	0	0

From the first testing, the measurement accuracy shows the result of 75.61% of accuracy. This result can be categorize as fair classification.

Before the next testing is conducted, there are some parameters in performing split validation technique that must be concerned. Those are split type, split ratio, and sampling type. By changing the parameters in split validation technique, the second testing is conducted using the final data that has been transformed which is training set. The second testing result for measurement accuracy of NBC algorithm result is shown in confusion matrix below:

Table 6: Confusion Matrix using Training Set

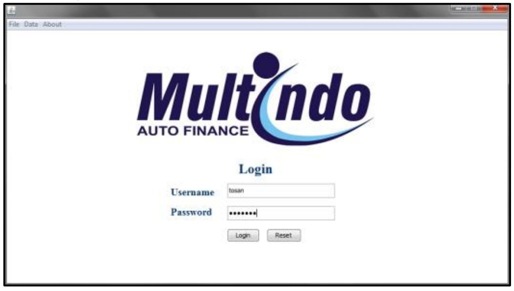
Classification	True lancar	True macet	True bermasalah
Pred. lancar	2665	65	89
Pred. bermasalah	32	3	5
Pred. macet	61	2	5

By changing some parameters in split validation technique and used training data set, the second testing for the measurement accuracy of NBC algorithm provides a result of 91.29% accuracy that is categorized as Excellent classification.

4.4. Implementation

Based on the testing result and evaluation performance in this study, the system development to implementing Naïve Bayes Classification technique is conducted using Netbeans IDE 7.4. The system is functioned as a visualization that can help the company which is PT. Multindo Auto Finance Semarang to determine customer credit status using Naïve Bayes Classification algorithm. In this application, there are some interfaces that users can access. They are Login, Home, Classification, Applicants, Customer, and About.

1. *Login*

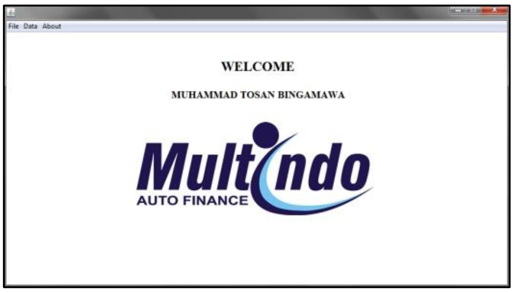


Username: tosan
Password: •••••••

Figure 1. Interface Login

In login page, users are required to input their username and password that have been registered in the database system.

2. *Home*



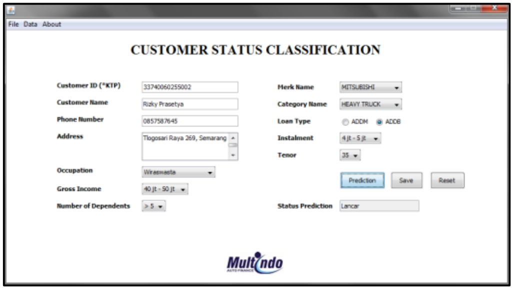
WELCOME

MUHAMMAD TOSAN BINGAMAWA

Figure 2. Interface Home

In this application, home page is a transition page after the users successfully login to the system. This page shows the user’s name from the database. This page provides some menu that can be used by the users.

3. *Classification*



CUSTOMER STATUS CLASSIFICATION

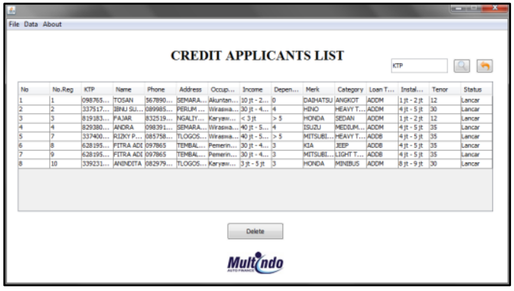
Customer ID (*KTP): 33740060255002
Customer Name: Rizky Prasetya
Phone Number: 0857587645
Address: Tlogosari Raya 269, Semarang
Occupation: Wiraswasta
Gross Income: 40 jt - 50 jt
Number of Dependents: > 5
Merk Name: MITSUBISHI
Category Name: HEAVY TRUCK
Loan Type: ○ ADDM ● ADDB
Instalment: 4 jt - 5 jt
Tenor: 35
Status Prediction: Lancar

Figure 3. Interface Classification

In classification user interface, the application is required some inputs to perform status prediction. Those inputs have been appropriated with the attributes needs for classification technique in this application. From those inputs, Naïve Bayes Classification that has been implemented in this application will calculate and determine the customer status whether it is categorized as regular credit, problem account, or loss credit.

The classification technique performed in this application is the implementation of Naïve Bayes performance that has been tested before with the result of 91.29% of accuracy. Based on that accuracy result, the customer classification perform in this application is quite accurate to determine new credit applicants status.

4. *Applicants*



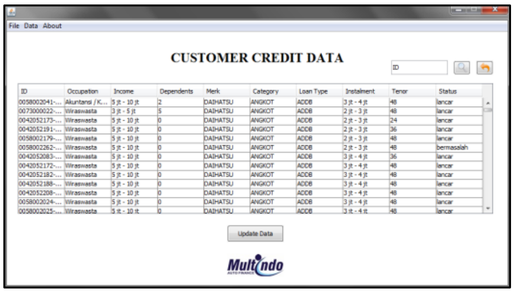
CREDIT APPLICANTS LIST

No	No.Reg	KTP	Name	Phone	Address	Occup...	Income	Depen...	Merk	Category	Loan T...	Instal...	Tenor	Status
1	1	098765...	TOSAN	567890...	SEMARA...	Akuntan...	10 jt - 2...	0	DAIHATSU	ANGKOT	ADDM	1 jt - 2 jt	12	Lancar
2	2	337517...	IBNU SU...	089985...	PERUM ...	Wiraswa...	30 jt - 4...	4	HINO	HEAVY T...	ADDM	4 jt - 5 jt	30	Lancar
3	3	819183...	FAJAR	832519...	NGALIY...	Karyaw...	< 3 jt	> 5	HONDA	SEDAN	ADDM	1 jt - 2 jt	12	Lancar
4	4	829380...	ANDRA	098391...	SEMARA...	Wiraswa...	40 jt - 5...	4	ISUZU	MEDIUM...	ADDM	4 jt - 5 jt	35	Lancar
5	7	337400...	RIZKY P...	085758...	TLOGOS...	Wiraswa...	40 jt - 5...	> 5	MITSUBI...	HEAVY T...	ADDB	4 jt - 5 jt	35	Lancar
6	8	628195...	FITRA ADI	097865	TEMBAL...	Pemerin...	30 jt - 4...	3	KIA	JEEP	ADDB	4 jt - 5 jt	35	Lancar
7	9	628195...	FITRA ADI	097865	TEMBAL...	Pemerin...	30 jt - 4...	3	MITSUBI...	LIGHT T...	ADDB	4 jt - 5 jt	35	Lancar
8	10	339231...	ANINDITA	082979...	TLOGOS...	Karyaw...	3 jt - 5 jt	3	HONDA	MINIBUS	ADDM	8 jt - 9 jt	30	Lancar

Figure 4. Interface Applicants

New credit applicants that have been classified from classification page can be saved into database. Applicants' page is used to show list of credit applicants who want to apply new credit loan to the company.

5. *Customer*



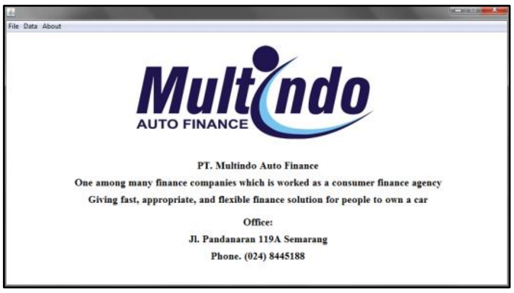
CUSTOMER CREDIT DATA

ID	Occupation	Income	Dependents	Merk	Category	Loan Type	Instalment	Tenor	Status
0058002041-...	Akuntansi / K...	5 jt - 10 jt	2	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0073000022-...	Wiraswasta	3 jt - 5 jt	5	DAIHATSU	ANGKOT	ADDB	2 jt - 3 jt	48	lancar
0042052173-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	2 jt - 3 jt	24	lancar
0042052191-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	2 jt - 3 jt	36	lancar
0058002179-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	2 jt - 3 jt	48	lancar
0058002262-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	2 jt - 3 jt	48	bermasalah
0042052083-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	36	lancar
0042052172-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0042052182-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0042052188-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0042052208-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0058002024-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar
0058002025-...	Wiraswasta	5 jt - 10 jt	0	DAIHATSU	ANGKOT	ADDB	3 jt - 4 jt	48	lancar

Figure 5. Interface Customer

Customer page shows customer credit list in the company. The customers are still on the credit repayment process. That is why in this page, the users can change the status of the customer based on credit transaction system recorded by the company. It may change frequently each month or once a year based on the company regulation. Through this page, the company can also monitor the customer status. It can help the company to control their customer and reduce customer credit problems.

6. *About*



PT. Multindo Auto Finance

One among many finance companies which is worked as a consumer finance agency

Giving fast, appropriate, and flexible finance solution for people to own a car

Office:
Jl. Pandanaran 119A Semarang
Phone. (024) 8445188

Figure 6. Interface About

About page is used to show the information about the company.

5. CONCLUSION

In this study, the implementation of Naïve Bayes algorithm for customer credit status classification in PT. Multindo Auto Finance Semarang can be performed well. It is proved by the result of the measurement accuracy that can be categorized as excellent classification. In the first testing, the evaluation result using confusion matrix provides 75.61% of accuracy. Furthermore, the second testing that is conducted after data preprocessing provides the result of 91.29% of accuracy. It can be concluded that the data preprocessing have a great impact on the calculation of measurement accuracy. Moreover, based on the result of this study, the customer classification using Naïve Bayes algorithm to determine customer credit status in PT. Multindo Auto Finance Semarang can be used to predict the status of new credit applicants and also control their customers from getting credit problems.

In further studies, more data records are recommended. Meanwhile, the amount of data itself must be balance between each labels. Different classification algorithm technique besides Naïve Bayes Classification Algorithm is recommended. Moreover, the used of optimization algorithm like genetic algorithm or fuzzy-logic is also recommended. More additional attributes as predictor variable are recommended besides the attributes that has been used in this study to support the classification technique.

REFERENCES

[1] T. E. o. E. Britannica, "Encyclopædia Britannica," 2015. [Online]. Available: htttp://www.britannica.com/EBchecked/topic/207149/finance-company. [Accessed 9 April 2015].

[2] C.-L. Huang, M.-C. Chen and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications,* vol. 33, p. 847–856, 2007.

[3] P.Marikkannu and K.Shanmugapriya, "Classification Of Customer Credit Data For Intelligent Credit Scoring System Using Fuzzy Set and MC2 – Domain Driven Approach," *Electronics Computer Technology (ICECT),* vol. 3, pp. 410-414, 2011.

[4] M. Sinungan, Dasar-Dasar dan Teknik Manajemen Kredit, Jakarta: Bumi Aksara, 1993.

[5] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, New Jersey: John Wiley & Sons, Inc., 2005.

[6] J. Han, M. Kamber and J. Pei, Data Mining Concept and Technique Third Edition, San Fransisco: Morgan Kaufmann, 2001.

[7] Kursini and E. T. Lutfi, Algoritma Data Mining, Yogyakarta: Andi Offset, 2009.