APPLICATION DETERMINATION OF CREDIT FEASIBILITY IN SHARIA COOPERATIVE WITH C4.5 ALGORITHM

Siti Masripah
AMIK BSI Jakarta
Jl. RS. Fatmawati No. 24 Pondok Labu in South Jakarta
email: siti.stm@bsi.ac.id

Abstract - Credit is the provision of money or bills, based on the agreement between bank lending and other parties who require the borrower to pay off debts after a certain period of time with interest. Cooperative Financial Services Sharia (KJKS) is a cooperative business activities engaged in financing, investment, and savings according to the pattern of results (sharia). Like the banks, sharia cooperative funding as the process of granting financing from the filing of financing, analysis of the proposed financing, approval committee of sharia cooperative finance, the binding of financing, until disbursement stage. Each borrower (debtor) must perform the process. Analysis of the proposed financing is a process undertaken by the authorities to determine whether the borrower has a good value or not. If the borrower has a good value, it will reduce the credit risk that will be accepted by funders. This paper discusses how to predict credit worthiness in sharia cooperative with classification C4.5 algorithm. Tests performed with the Confusion Matrix produces an accuracy value of 88% at 0.898 and AUC values with Good Classification level diagnostics. Then the classification results are implemented in the web to know the status of the credit risk of a customer whether liquid or bad credit.

Keywords: Determination of credit feasibility, C4.5 algorithm

I. INTRODUCTION

In a broad sense, the credit risk is the uncertainty of earnings or fluctuations in credit activities (Yu, Chen, Koronios, Zhu, & Guo, 2007). To reduce the credit risk then credit analysis is important in the management of financial risks (Lai, Yu, Zhou, & Wang, 2006). Historical data is the training data or the data of experience, as with the data we're going to practice to gain knowledge. Classification algorithm will use training data so that it will produce knowledge to classify the credit risk of a customer in the future based on existing variables. As a customer benchmark material approved or rejected, can be seen from the data in the customer credit history cooperative sharia. Below is a chart that shows that customers are problematic in terms of loan installment payments greater than the liquid customer in the loan repayment, based on the data taken in 2010..

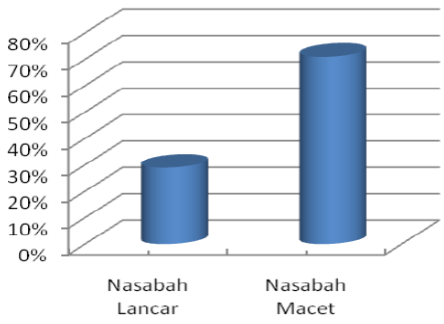


Figure 1.1 Chart of Customer Status (source :sharia cooperative)

This study aimed to apply the classification algorithm C4.5 on a web-based system to help the financing of the proposed analysis in determining the credit status of customers in sharia cooperative. The benefit of this research is divided into several benefits, namely, the practical benefits of the results of this study can be used by analysts a loan provider to do a better analysis. The benefits of the policy can be used as a material consideration in decision making on corporate credit analysis. And theoretical benefits, is expected to contribute to the data mining algorithm C4.5 in particular.

The framework of this study as follows:

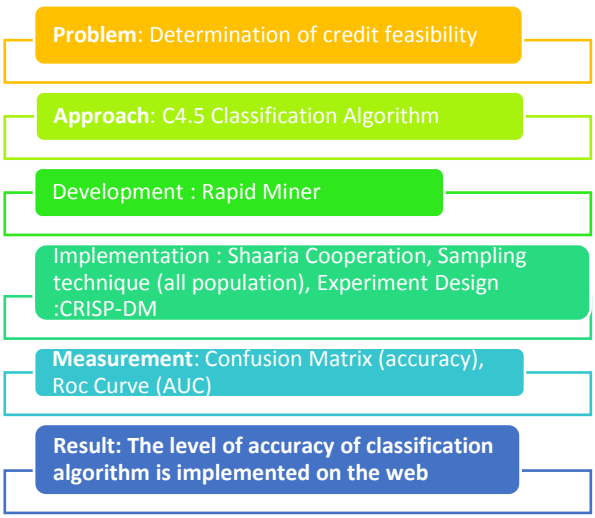


Figure 1. Framework

II. THEORY

2.1. Algorithm C4.5

One of the classification method that interesting involves the construction of a decision tree, a collection of decision nodes, connected by branches, extending down from the root node until it ends at a leaf node. Starting from the root node, which by convention is placed at the top of the decision tree diagram, the attributes are tested at decision nodes, with any outcome that may produces branch. Then each branch leads to another decision node or to a leaf node to end (Larose, 2005).

In

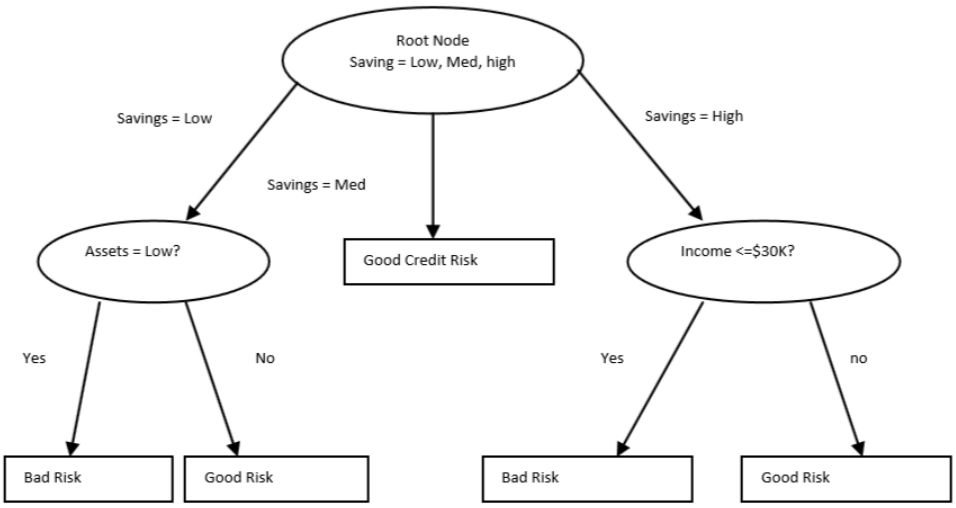


Figure 2.2 the target variable for decision trees are credit risk, with the potential customers who are classified as good or bad credit risk. Predictor variables are saving (low, med, high), Assets (low or non-low), and income (≤ $ 50,000 or> $ 50,000). Here, the root node is a decision node, test whether each has a saving rate (saving) a low, medium or high. C4.5 algorithm is part of a group of trees and a decision algorithm category 10 of the most popular algorithms. At the end of the 1970s until the early 1980s, Quinlan J.Rosss a researcher in the field of machine learning to develop a decision model, called ID3 (Interative Dichotomiser), although previously this project has been made by EB.

Hunt, J. Marin, and P.T. Stone. Quinlan then make the algorithm C4.5 (development of ID3) based supervised learning (Han & Kamber, 2006). Stages in making a decision tree algorithm C4.5 (Larose, 2005), namely:

1) Prepare the training data, the training data are usually taken from historical data that never happened before or referred to past data and are already grouped in certain classes.
2) Calculate the total entropy before look for each class Entropy

$$\mathrm{H}(T) = -\sum_j Pj \log 2\,(Pj) \qquad (2.1)$$

Remarks:
H = The set of cases
T = Attributes
Pj = proportion of Hj to H

3) Calculate the value of the information gain Gain averaging:

$$\text{Gain average} = \mathrm{H}(T) - \mathrm{Hsaving}(T) \qquad (2.2)$$

Remarks:
H (T) = Total Entropy
Hsaving (T) = Total Gain information for each Attribute

4) Repeat steps 2 and 3 until all tuples partitioned

Partitioning process stops when the decision tree:
a. All tuples in the N nodes get the same class
b. There is no attribute in the tuples are partitioned again
c. There is no branch in the empty tuple

2.2.1 Evaluation of Confusion Matrix and ROC Curve

1. Evaluation of Confusion Matrix

To evaluate the classification model based on the calculation of testing objects which are predicted correct and incorrect. These calculations are tabulated into a table called confusion matrix (Gorunescu, 2011). Form of confusion matrix is shown in Table 2.1 below:

Tabel 0.1 *Confusion matrix* (Gorunescu, 2011)

CLASSIFICATION		PREDICTED CLASS	
		Class = YES	Class = NO
OBSERVED CLASS	Class = YES	a (true positive-TP)	b (false negative-FN)
	Class = No	c (false positive-FP)	d (true negative-TN)

In Table 2.1, for True positive is a positive tuple in set data that classified positive, True negatives are the negative tuples in the data set were classified negative. False positives are positive tuples in the data set were classified negative False negatives is the number of negative tuples classified positive.

After subsequent confusion matrix will be calculated accuracy, sensitivity, specificity, PPV, NPV. Sensitivity is used to compare the number of true positives against the number of tuples that positives .

while specificity is the ratio of true negatives to the number of tuples that negatives. As for the PPV (positive predictive value) is the proportion of cases with a positive diagnosis, NPV (negative predictive value) is the proportion of cases with a negative diagnosis. Here's the calculation:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.5)$$

$$Sensitivity = \frac{numberof'TruePositives'}{Numberof'TruePositives'+Numberof'FalseNegatives'} \quad (2.6)$$

$$Specificity = \frac{numberof'TrueNegatives'}{Numberof'TrueNegatives'+Numberof'FalsePositives'} \quad (2.7)$$

$$PPV = \frac{numberof'TruePositives'}{Numberof'TruePositives'+Numberof'FalsePositives'} \quad (2.8)$$

$$NPV = \frac{numberof'TrueNegatives'}{Numberof'TrueNegatives'+Numberof'FalseNegatives'} \quad (2.9)$$

Sensitivity can also be said to be true positive rate (TP rate) or recall. A sensitivity of 100% means that the classification recognizes a positive observed cases. For example, all people have a malignant cancer is recognized as an illness.

2. Evaluation ROC Curve

ROC curve (Receiver Operating Characteristic) is a graphical illustration of the ability of the discriminant and is usually applied to the problem of binary classification (Yu, Chen, Koronios, Zhu, & Guo, 2007). Technically, the ROC curve is also called the ROC graphs, two-dimensional graphs, namely the TP rate is placed on the Y axis, while the FP rate is placed on the X axis ROC graph illustrates the trade-offs between benefits ('true positives') and costs ('false positives'). Below the display are two types of ROC curves (discrete and continuous).

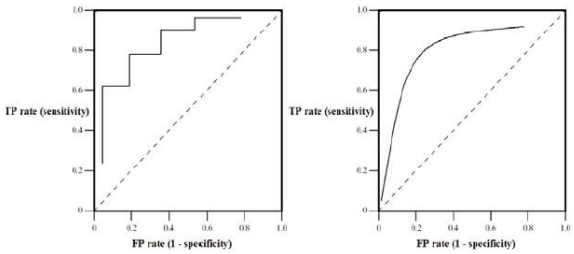


Figure 0.3 ROC graph (discrete and continuous).

III. THE RESEARCH METHOD

3.1. The Research Design

There are four commonly used research methods that is Action Reserch, Experiment, Case Study and Survey (Dawson, 2009). The research method used is a form of research Experiment. Experimental research is an investigation of causal relationships using controlled tests by researchers (Dawson, 2009). In experiments typically consist of:

1 Defining the theoretical hypothesis
2 Select a sample from a known population
3 Allocate samples to different experimental conditions
4 Introducing planned changes for one or more variables
5. Measuring a small number of variables
6 Controlling all the variables

Experimental studies are usually conducted in the project development, evaluation and problem solving (Dawson, 2009). In experimental studies used hardware and software specifications as a tool in the research contained in Table

Tabel 3.1 *Hardware* and *Software* Spesification

Hardware	*Software*
CPU : Intel Pentium Dual Core	Operating System : Windows 8
Memory : 1 GB	Data Mining : Rapid Miner 5.1
Hardisk : 120 GB	Application : Dreamweaver CS6
	Database : SQL

In experimental research methods, used process model CRISP-DM (Cross-Industry Standard Process for Data Mining), which consists of 6 stages (Larose, 2005):
1 Bussiness understanding
2 Data understanding
3 Data preparation
4 Modelling
5. Evaluation
6 Deployment

3.2. Data Understanding

The data obtained from the cooperative of sharia is a customer credit data in 2010, the amount of data as the data 866. Attributes or variables that have as many as 44 attributes (the data can be seen in appendix). After the data preparation process, attributes or variables used consists of 17 attributes of the data contained in the customer's credit status. These variables were classified as no predictor or predictor variables (predictor variable) is the variable that is used as a basic determinant of credit risk, and the goal variable is the variable that is used as credit risk (Susanto & Suryadi, 2010). Predictor variables ie customer name, gender, age, loan amount, term, monthly installment amount, loan type, loan type, bi economic sector, the debtor class bi, bi group guarantor, balance nominative, theoretical ceiling, principal arrears, and arrears interest. While the goal variable is the credit status..

3.2. Data Preparation

At this stage the data as much as 866 and attributes consisting of 44 attributes, some screening will be done to produces the required data, the stages are:

1) Data Cleaning to clean the empty value or an empty tuple. For example, attributes arrears penalties.
2) Data Integration with storage that serves to unite different places into one data. In this case there is only one data repository that customer credit status.
3) Data reduction used the number of attributes that may be too large, of the 44 attributes used only 17 of the required attributes, and attributes that are not required to be removed. The data in Table 3.2 below only as an example for the training data, for more on the attached appendices.

Nama nasabah	Jenis kelamin	umur	Jml Pinjaman	jkw	jml_angsuran per_bulan	Type Pinjaman	Jenis Pinjaman	bi_sektor ekonomi	Col	bi_gol debitur	bi_gol penjamin	Saldo nominatif	Plafond Teoritis	Tunggakan pokok	Tunggakan bunga	Status kredit
x1	P	40	345000	1	345000	100	301	6000	1	874	875	345000	0	345000	0	MACET
x2	L	31	350000	7	55716	100	301	6000	1	874	875	390000	278572	111428	0	MACET
x3	L	37	649926	6	108321	100	301	6000	1	874	875	649926	433284	216642	0	MACET
x4	P	25	459168	12	38264	100	301	6000	1	874	875	459168	76528	382640	0	MACET
x5	P	34	3055499	8	381937,41	100	301	6000	1	874	875	3055499	1527750	1527749,48	0	MACET
x6	L	49	2000000	16	0	100	301	6000	1	874	875	-85000	0	0	0	LANCAR
x7	L	31	8333334	10	833333,4	100	301	6000	1	874	875	8333334	5000000	3333333,6	0	MACET
x8	L	27	4435001	8	671098	100	301	6000	1	874	875	4435001	4635001	0	0	LANCAR
x9	L	42	560000	8	95221	100	301	6000	1	874	875	660800	560000	100800	0	MACET
x10	L	49	1443750	15	107800	100	301	6000	1	874	875	1617000	539000	1078000	0	MACET
x11	L	42	3066000	10	351670	100	301	6000	1	874	875	3066000	2452800	613200	90140	MACET
x12	P	26	4071669	20	203583,45	100	301	6000	1	874	875	3671669	2443001	1228667,6	0	MACET
x13	L	35	228655000	33	7495303,73	100	301	6000	1	874	875	209404092	1,2E+08	91612122,24	7362732	MACET
x14	L	55	840000	4	60000	100	301	6000	1	874	875	240000	0	240000	0	MACET
x15	L	38	3000000	24	147500	100	301	6000	1	874	000	2500000	1625000	875000	157500	MACET
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	
866	P	31	1000000	60	18167,06	100	301	6000	2	874	000	233333,64	150000	83333,3	7500	LANCAR

Based on Table 3.2 of all the attributes that exist in the table above are not all worth categorical, but there are valuable points. Based on Table 3.2 candidat tree then made the determination, the determination is done by inserting a tree candidat all the attributes then do attributes assessment resulting in a classification of attributes that affect credit risk, in Table 3.3 obtained candidat split the arrears in principal, the amount of the
loan, the amount of monthly installments, unpaid interest, balance nominative, so the value of the rule can be described as follows in Table 3.3:

Table 3.2 Candidate split and rule of attribute value C4.5 algorithm

Candidate split		Child nodes
1	Tunggakan pokok	Tunggakan pokok
	≤ 7055.330	>7055.330
	≤ 9000	> 9000
	≤ 166833.330	> 166833.330
	≤ 313750.005	> 313750.005
	≤ 606600	> 606600

2	Jml pinjaman ≤ 1831667 ≤ 505626 ≤ 925208.380 ≤ 14692499.500	Jml pinjaman > 1831667 > 505626 > 925208.380 > 14692499.500
3	Jml angsuran per bulan ≤ 12750 ≤ 239633.205	Jml angsuran per bulan > 12750 > 239633.205
4	Tunggakan bunga ≤ 1756 ≤ 9000 ≤ 15000 ≤ 41250 ≤ 112500	Tunggakan bunga > 1756 > 9000 >15000 > 41250 > 112500
5	Saldo nominatif ≤ 433750 ≤ 950000 ≤ 1669500.010	Saldo nominatif > 433750 > 950000 >1669500.010
6	Jkw ≤ 7.500 ≤ 22.500	Jkw > 7.500 > 22.500
7	Bi golongan penjamin = 000 Bi golongan penjamin = 875	

3.3. Modelling

At this stage, the data processing is done so that the training will result in some rules and will form a decision tree. The classification C4.5 algorithm, the following steps will be performed.

1. Counting the number of cases of class LIQUID and class BAD and Entropy of all cases and cases that are divided based on the attributes in Table 3.3. Total line of Entropy is calculated based on training data
2. Then calculate the gain of each attribute based on Table 3.3 above, as an example for arrears in principal.

And to information of Gain can be seen in Table 3.5 below:

Table 3.5 Information Gain for C4.5 algorithm

Kandidat Split	Child Nodes	Informasi Gain (Entropy Reduction)
1	Tunggakan pokok ≤ 7055.330 dan >7055.330	0.4358
2	Tunggakan pokok ≤ 166833.330 dan > 166833.330	0.4157
3	Tunggakan pokok ≤ 313750.005 dan > 313750.005	0.3039
4	Tunggakan pokok ≤ 606600 dan > 606600	0.15
5	Jumlah pinjaman ≤ 505626 dan > 505626	0.0256
6	Jumlah pinjaman ≤ 925208.380 dan > 925208.380	0.0876
7	Jumlah pinjaman ≤ 1831667 dan > 1831667	0.2908
8	Jumlah pinjaman ≤ 14692499.500 dan > 14692499.500	0.0014
9	Jumlah angsuran ≤ 12750 dan > 12750	0.0012
10	Jumlah angsuran ≤ 239633.205 dan > 239633.205	-0.0869
11	Tunggakan bunga ≤ 1756 dan > 1756	0.0302
12	Tunggakan bunga ≤ 9000 dan > 9000	-0.1917
13	Tunggakan bunga ≤ 15000 dan > 15000	0.0.2193

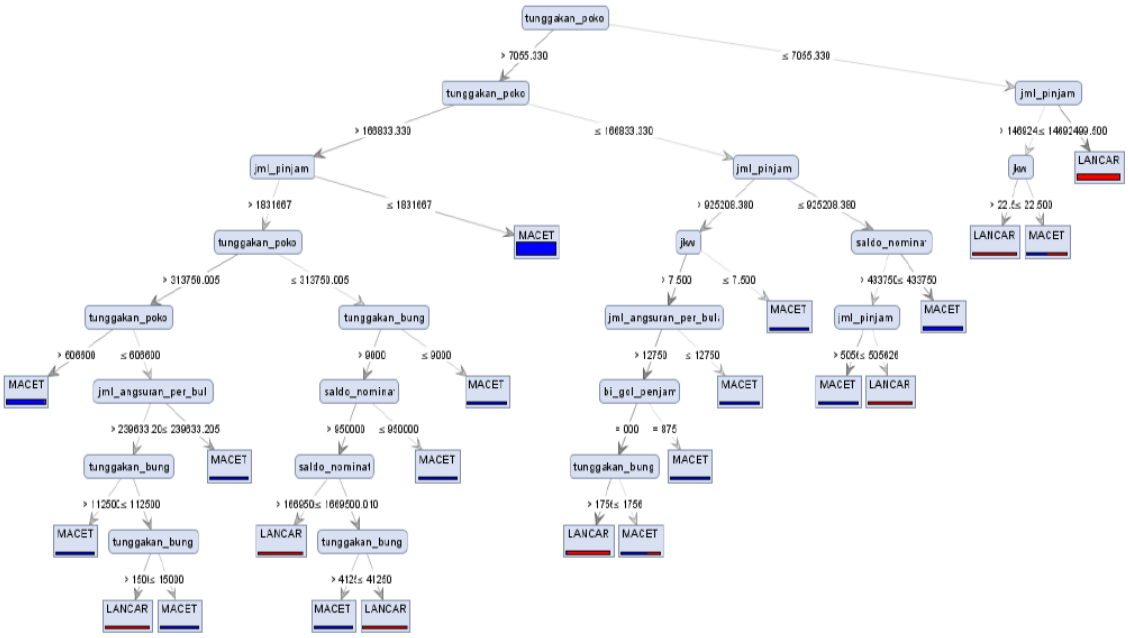


Figure 3.1. Decision tree of the customers classification to algorithm C4.5

IV. RESULTS AND DISCUSSION

4.1. Evaluation and Validation Model

The results of testing the model is for the credit worthiness with C4.5 classification algorithm to determine the value of accuracy, and AUC.

1. Testing Results Using C4.5 Algorithm
   The results of the experiments performed to produces value accuracy and AUC values (Area Under the Curve).
   a. Evaluation of the model with the Confusion Matrix
      Model confusion matrix to form a matrix consisting of true positive and true positive or negative tuple or tuples negative, then input the data into the testing that has been prepared so that the results obtained confusion matrix in Table 4.1 below:

Tabel 4.1 Konversi *confusion matrix* algoritma klasifikasi C4.5

	Predicate Class	
Observed Class	YES	NO
YES	50	3
NO	9	38

In Table 4.1 that for the number of True Positive (TP) is 50, for False Negative (FN) is 3, for False Positive (FP) is 9, and for True Negative (TN) is 38. Based on data contained in the confusion matrix above then can we count to find the value of accuracy, sensitivity, specifity, PPV, and NPV, outcome can be seen in Table 4.2 below:

Tabel 4.2 Nilai *sensitivity, specificity, ppv, npv,* dan *accuracy*

	Nilai (%)
Accuracy	88
Sensitivity	94,34
Specificity	80,85
PPV	84,75
NPV	92,68

Based on Table 4.2 show that, the accuracy of the C4.5 classification algorithm is used by 88%.

b. Evaluation of the ROC Curve

In Figure 4.1 shows a graph with the value of ROC AUC (Area Under the Curve) of 0898. Accuracy levels of diagnosis are (Gorunescu, 2011):

Accuracy is worth 0.90 - 1.00 = Excellent classification
Accuracy is worth 0.80 - 0.90 = Good classification
Accuracy is worth 0.70 - 0.80 = Fair classification
Accuracy is worth 0.60 - 0.70 = Poor classification
Accuracy is worth 0:50 - 0.60 = Failure

While the results obtained from the processing of ROC which can be seen in Figure 4.1 for 0898 with a diagnosis of Good classification level.

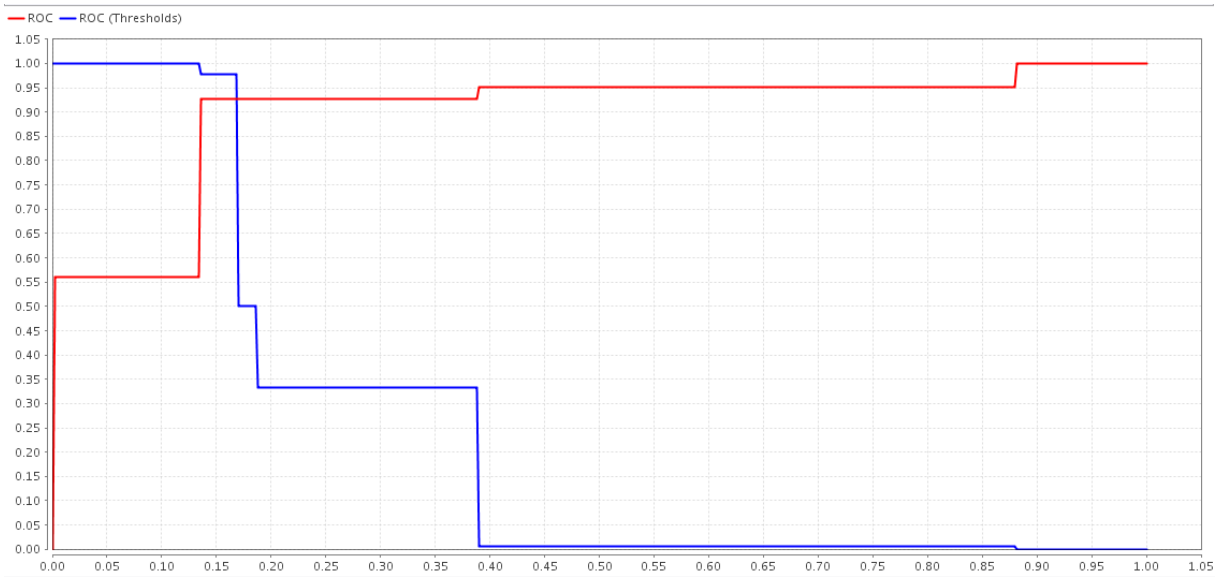
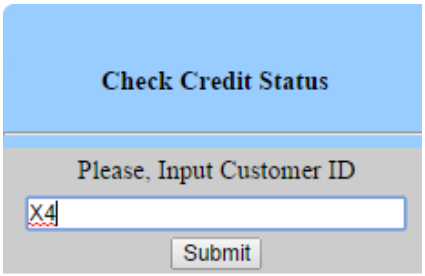


Figure 4.1 ROC AUC values in a graph algorithm C4.5

4.2. Implementation in web

The customers classification who have been tested with the confusion matrix and ROC curves is implemented into the new data for the next testing. The result testing on new data show the level of accuracy of classification results of customers by 88%. So the rule obtained from the customer classification can be applied to the applications of determination of credit feasibility web-based as follows:

1. Input customer ID for evaluation then submit



After input of customer data it will show customer data at the form clasification. Then fill the loan data. Click Submit. It will show classification results in the form of credit status

Figure 4.2. View of Credit evaluation

a. View of customer data reports

Data Nasabah

ID	Nama	Jenis Kelamin	Umur	Bi_Sektor_Ekonomi	Bi_Gol_Debitur	Col	Bi_Gol_Penjamin	Saldo Nominatif	Plafon Teoritis	ACTION
X1	X1	P	40	6000	874	1	875	345000	0	ADD EDIT
X2	X2	L	31	6000	874	1	875	390000	278572	ADD EDIT
X3	X3	L	37	6000	874	1	875	649926	433284	ADD EDIT
X4	X4	P	25	6000	874	1	875	459168	76528	ADD EDIT
X5	X5	P	34	6000	874	1	875	3055499	1527750	ADD EDIT
X6	X6	L	49	6000	874	1	875	85000	0	ADD EDIT
X7	X7	L	31	6000	874	1	875	8333334	5000000	ADD EDIT

Jumlah Nasabah : (7)

b. View of loan data report

LAPORAN DATA PINJAMAN

ID	Nasabah	Jumlah Pinjaman	JKW	Jumlah Angsuran Pinjaman	Type Pinjaman	Jenis Pinjaman	Tunggakan Pokok	Tunggakan Bunga	Status
10	X1	345000	1	0	100	301	345000	0	MACET
11	X2	350000	7	0	100	301	111428	0	MACET
12	X3	649926	6	108321	100	301	216642	0	MACET
13	X4	459168	12	38264	100	301	382640	0	MACET

Jumlah Dokumen : (4)

Created: Rahma Hanifah Wijaya

IV. CONCLUSION

The results of the study for accuracy classification algorithm C4.5 value by 88%. For AUC values based on ROC curve for C4.5 classification algorithm is worth 0898 with the diagnosis of Good classification level. So the rule obtained from the customer classification can be applied to the applications of determination of credit feasibility web-based.

As for the suggestion of this research are

1. Adding the amount of data that larger and more attributes, so the measurement results will be obtained even better.
2. Using optimization methods such as Ant Colony Optimization (ACO), Genetic Algorithm (GA), and others.
3. Development using selection methods other attributes such as chi-square, and so the index information for selecting the attribute accuracy.

REFERENCES

[1] Dawson, Chaterine., Introduction to RESEARCH METHODS: A practical guide for anyone undertaking a researdh project. Begbroke, Oxford OX5 IRX, United Kingdom: How to Bookd Ltd, 2009.

[2] Lai, K. K., Yu, L., Zhou, L., & Wang, S., Credit Risk Evaluation With Least Square Support Vector Machine, 2006.

[3] Larose, D. T., Discovering Knowledge In Data. Canada: Wiley- Interscience, 2005.

[4] Gorunescu, F., Data Mining Concepts, Model and Techniques. Berlin: Springer, 2011.

[5] Susanto, S., & Suryadi, D., Pengantar Data Mining menggali Pengetahuan dari Bongkahan Data. Yogyakarta: C.V ANDI OFFSET,2010

[6] Yu, L., Chen, G., Koronios, a., Zhu, S., & Guo, X. Application and Comparison of Classification Techniques in Controlling Credit Risk, World Scientific, 2007, p.111..

Siti Masripah, is currently a lecturer of the Study Program of Accounting Computerizzed, AMIK BSI. She received a Master Degree in Computer Science from STMIK Nusa Mandiri in 2010 on “Management Information System”. Siti Masripah, M. Kom research interests are in Data Mining. She is active involved as member in Consorsium of Accounting Computerized.