ISSN 2278-3091

Volume 9 No.2, March -April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse27922020.pdf

https://doi.org/10.30534/ijatcse/2020/27922020

Optimization of Debtor Credit Quality Determining Prediction using Decision Tree

Abba Suganda Girsang[1], **Abner** [2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, agirsang@binus.edu

[2]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, abner@binus.ac.id

ABSTRACT

Credit is the one of the biggest contributor to the bank profit, other than products and services, but credit can also become the biggest loss contributor to the bank if credit quality, which is called as collectability not maintained properly. Good credit growth and expansion is needed to be in line with credit quality improvement so the maximum profit could be generated. This research seek to answer these question, how quantitative and qualitative data inside system that use decision tree can optimalized credit quality determination on PT. XYZ. Algorithm that is used in the Decision Tree method is Iterative Dichotomizer Three (ID3). Some of the advantages of ID3 Algorithm which are it could create understandable predicition rules, more faster, and only need some attribute test until all data is classified.

Key words : Bank , Prediction, Credit Scoring, Decision Tree

1. INTRODUCTION

In today's banking business competition requires companies to provide better quality, as well as faster information that improves business processes, so that the company's existence can excel in very tight competition. Use of optimal technology by processing data sets and relate between understandable data collected into new information that can become trends so as to provide and support a decision that considered as strategic.

Cedit is one of the biggest contributors to profit in Bank [1] [2], in addition to other products and services, but credit can also be the biggest loss contributor if quality credit, called collectability[3][4], is not well maintained. Improvement of good credit is have to be in line with improvement of quality of credit so the profit can be maximized to the maximum extent possible[5]. The community's need for good credit for consumer credit as well as productive credit with the target of ever-growing business growth push the company to make business process improvements in increasing speed and facilitating the process of analyzing and selecting prospective borrowers in order to maintain credit quality[6].

As one of the largest banking companies in Indonesia, PT Bank XYZ, in its various businesses, one of which is lending has a big challenge in maintaining credit quality, because many credit in arrears, the Company must provide deferred funds to Bank Indonesia that require funds that must be in circulation, which become Dead Funds or Unmoving Funds that cause loss for Bank because they have to pay interest to the customer that save their money in PT Bank XYZ.

In credit approval with regard to creditquality, the assessment of repaymentcapacity must meet predetermined standards, as defined [7], Credit approval is a technique that helps creditors, providing credit assistance to prospective debtors based on predetermined standards. A scoring system used to overcome credit problems, provide credit assessment and assistance in making decisions based on scoring of repayment capacity. Credit scoring is useful for reducing the cost of credit analysis, enabling faster credit decisions, closer monitoringto debtors and prioritizing credit collectibility [8]. Scoring system is part of the company's risk management with the aim to avoid the risk of accounts receivable loss by using, analyzing, and borrowing credit risk of the debtor [9][6].

The credit process starts with a pre-screening process with provisions determined by the Bank and Ministry of Economy regulations [10]. In the Bank's regulation, the credit analysis process is carried out with quantitative data that is done manually, while the qualitative data is carried out with adjustments, which causes the percentage of accuracy of bad loans prior to write offs in 2016 amounting to Rp.1,324,157,332,870, or 6.35 % of the Debit Tray and in 2017 increased to Rp3,357,776,685,202 or 8.23% of the Debit Tray, it is expected that using qualitative data will provide a faster prediction of credit quality.

2. RELATED WORK

The initial decision tree development is to analyze the value of risk and the value of information contained in an alternative problem solving[11]. The role of this decision tree as a tool in making decisions (decision support tool). The decision tree also shows the likelihood / probability factors that will affect the alternatives of the decision, accompanied by an estimate of the final results that will be obtained if we take the alternative decision. Refers to the ID3 algorithm (Iterative Dichotomizes 3) developed by J. Ross Quinlan[12][13]. The ID3 algorithm can be implemented using a recursive function (a function that calls itself). The ID3 algorithm attempts to build a top-down decision tree, starting with the question: "which attribute must first be checked and placed at the root?" This question is answered by evaluating all the attributes available using a statistical measure (which is widely used is information gain) to measure the effectiveness of an attribute in classifying a sample of data[6]. As in the case of credit assessment in multi finance[14][15], the determination of credit worthiness using the ID3 method produces 100% accuracy for 10 testing data from 100 training data and 30 testing data from 60 training data.

In summary, the workings of the ID3 Algorithm can be described as follows, the selection of attributes using Information Gain [9]. The steps taken are (a). Select the attribute where the information gain value is greatest. (b). Create a node that contains these attributes. (c)[16]. The information gain calculation process will continue until all data have been included in the same class. The selected attribute is not included anymore in the calculation of the information gain value. Whereas the selection of attributes using the ID3 algorithm is done by statistical properties, which are called information gain. Gain measures how well an attribute separates the training example into the target class. The attribute with the highest information will be selected. In order to define gain, the idea of information theory called entropy is first used. Entropy measures the amount of information contained in an attribute using the formula:

$$Entropy\ (S) = -\ p_{+}\ \log_2\ p_{+}\ -\ p_{-}\ \log_2\ p_{-} \qquad [1]$$

Where:

S is sample set.

Log2 is log base 2.

p is the proportion of S.

In ID3 algorithm entropy substraction is called as information gain and can be calculated with formula as follows:

$$Gain\ (S, A) = Entropy\ (S)\ -\sum_{v E\ value\ (A)} \frac{|\,S_v\,|}{|S|}\ Entropy\ (\,S_v\,) \qquad [2]$$

Where:

S is each value v of all possible values of attribute A.

S_v = subset of S for which attribute A has value v.

$|S_v|$ = number of elements in $S_{v.}$

$|S|$ = number of elements in S.

3. PROPOSED METHOD

Based on literature study and research on previous page, for debtor quality assessment which will be used, the phases are conducted which are data collecting, testing data with algorithm application, evaluation, and test results form data that will be involved as part of the system result[17][18]. In the research that are conducted by the researcher, data collection that will be used is using quantitative data and qualitative data. Quantitative data has a value or sum that has been stated in number. Quantitative data consists of various numerical variable that show a quantity. On the other hand, qualitative data is a data that shows the object types and it can be represented by name, symbol, and numeric code. Qualitative data consists of various categorical variables that shows quality. And then the collected data which already in csv form will be processed with a open source software that called Rapid Miner, which is an environement of machine learning, data mining, text mining, and data analysis [19]. Proposed method that will use for analysis as show in figure 1.

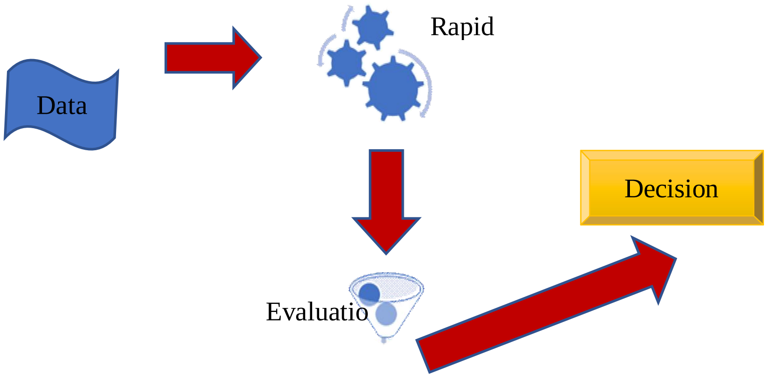

Figure 1: Concept

For data process that conducted in Rapidminer software, K-Fold Cross Validation will be usedinID3 algorithm and for comparison will be using C4.5 algorithm. For the design that will be processed in Rapidminer as show in figure 2.

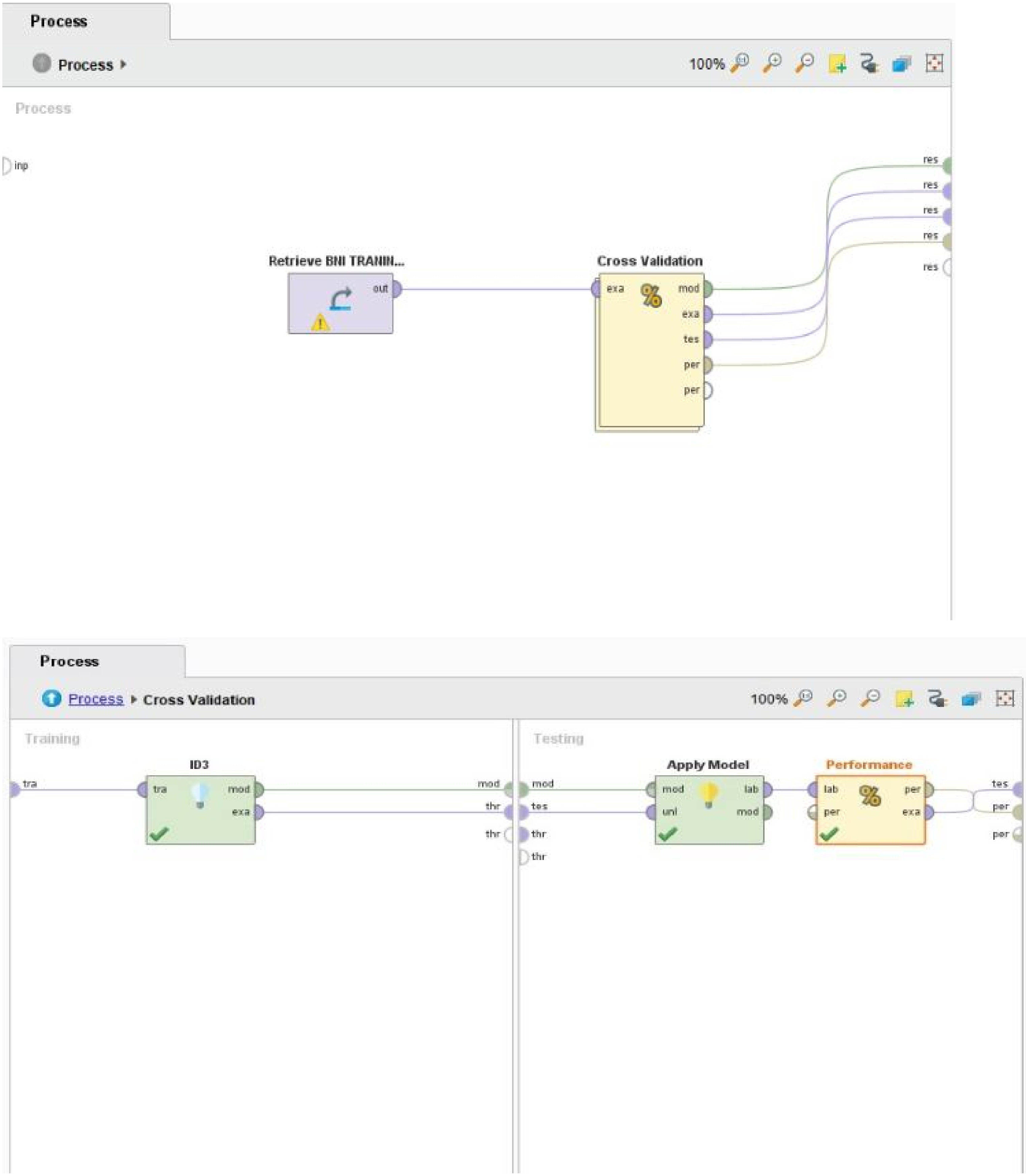

Figure 2: Cross Validation Design

4. ANALYSIS RESULTS

Based with Cross Validation process results using ID3and C.45 algorithm it creates a prediction value about debtor credit quality represented with "REJECTED" or "ACCEPTED" and accuracy value and precision value. For result with ID3 as show in table 1 and table 2. For result with C4.5 as show in table 3 and table 4.

Process is using ID3 algorithm:

Table 1: Accuracy 10 Fold Cross Validation Results ID3

accuracy: 75.31% +/- 1.68% (micro average: 75.31%)(positive: LANCAR)			
	true MACET	true LANCAR	class precision
pred. MACET	180	153	54,05%
pred. LANCAR	464	1702	78,58%
class recall	27,95%	91,75%	

Table 2: Precision 10 Fold Cross Validation Results ID3

precision: 78.59% +/- 1.09% (micro average: 78.58%)(positive: LANCAR)			
	true MACET	true LANCAR	class precision
pred. MACET	180	153	54,05%
pred. LANCAR	464	1702	78,58%
class recall	27,95%	91,75%	

Process is using C4.5 algorithm:

Table 3: Accuracy 10 Fold Cross ValidationResults C4.5

accuracy: 74.51% +/- 0.87% (micro average: 74.51%)(positive: LANCAR)			
	true MACET	true LANCAR	class precision
pred. MACET	18	11	62,07%
pred. LANCAR	626	1844	74,66%
class recall	2,80%	99,41%	

Table 4: Precision 10 Fold Cross Validation Results C4.5

precision: 74.66% +/- 0.55% (micro average: 74.66%)(positive: LANCAR)			
	true MACET	true LANCAR	class precision
pred. MACET	18	11	62,07%
pred. LANCAR	626	1844	74,66%
class recall	2,80%	99,41%	

After the accuracy and precision value seen from the data that has been processed with Rapidminer, the ROC curve will appear to show the relationship between sensitivity test and specificity test. These are used to explain the test in many cut-point stage in reading appropriate specificity with current sensitivity. Precision of the test can be explained from area below ROC curve. More bigger the area, more better the test results.

Evaluation with Receiver Operating Character Curve (ROC Curve), technically illustrates a two-dimensional graph, where the level of True Positive (TP) lies on the Y axis, while for False Positive (FP) lies on the X axis[20]. Thus the ROC illustrates the tradeoff between TP and FP. Recording in the ROC is stated in a clause, ie the lower the left point (0.0), then it is stated as a prediction classification approaching / becoming negative, while the higher the right point (1.1), then stated as a prediction classification approaching / becoming positive. A point with a value of 1 is expressed as a True Positive (TP) level, while a point with a value of 0 is expressed as a False Positive (FP) level. At point (0.1) the prediction classification is perfect because all cases both positive and negative are correctly said. Whereas for (1.0) prediction classification everything is stated as false (False). The closer to 1 AUC value, the better the prediction model. Graphic curve for ID3 and C4.5 algorithm as show in figure 3 and figure 4.

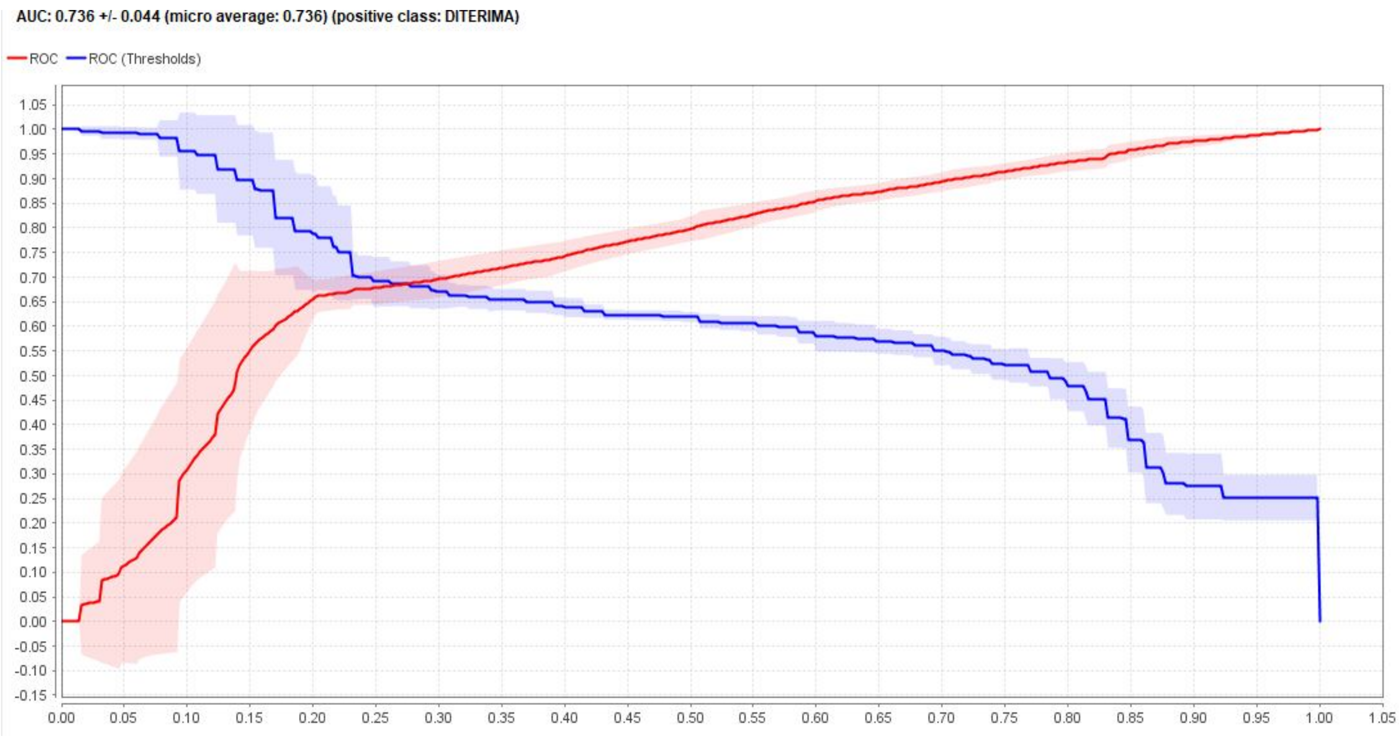

Figure 3:AUC from ID3 algorithm

Results accuracy is 74.63% from ID3 algorithm. In the ID3 algorithm trial test result for 2500 record. Data which has the most biggest impact is SektorEK, SektorEK attribute has big impact to the acceptance or rejection of loan application. This attribute is business type attribute that has been owned by customer. In fact the business type attribute also has a the biggest impact followed with customer age, business tenure, and guarantee.

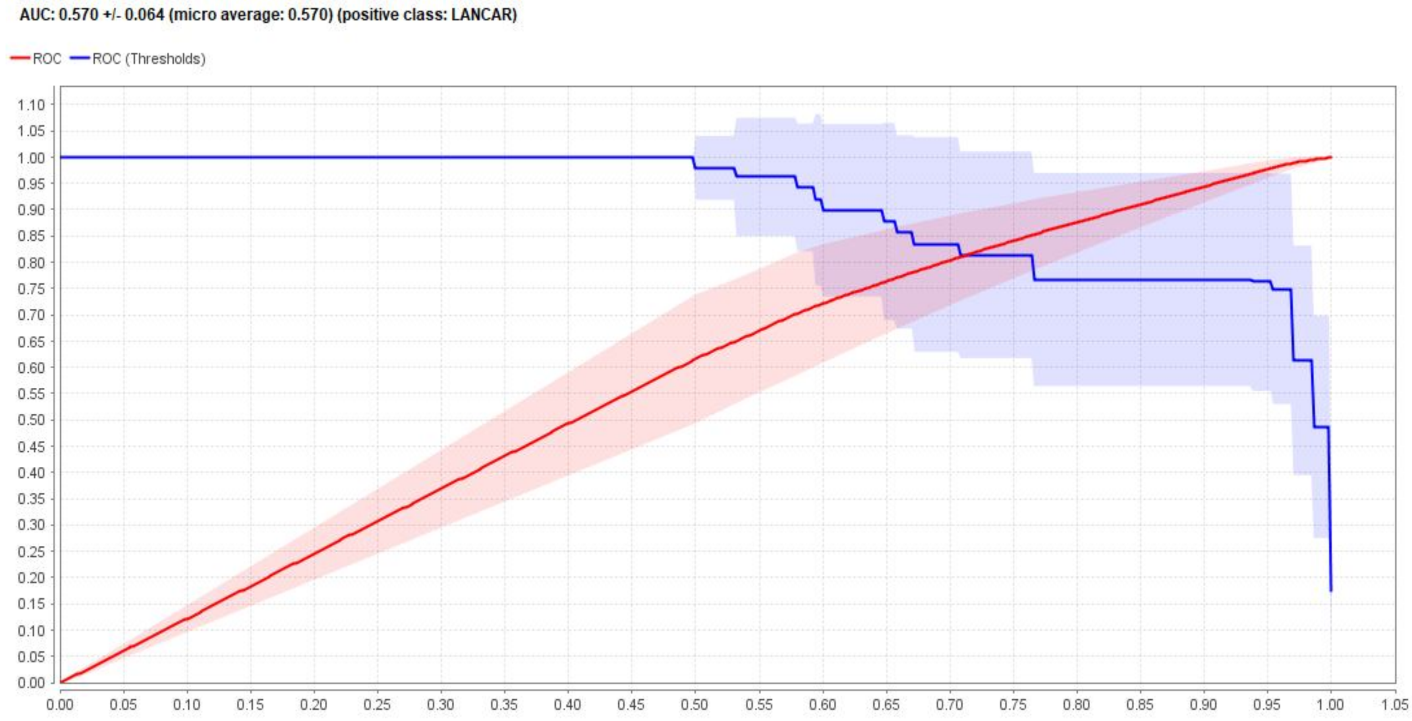

Figure 4: AUC from C4.5 algorithm

Results accuracy is 74.51% from C4.5 algorithm. In the C4.5 algorithm trial test result for 2500 record.

5. CONCLUSION

After conducting the research and testing to predict credit worthiness of PT Bank XYZ using ID3 algorithm, it can be concluded that:

a) After doing data mining process to qualitative data it give accuracy value 75.31% using ID3 algorithm and when using C4.5 algorithm it give accuracy value 74.51%

b) With this research it can help PT. Bank XYZ to be more careful on selection with adding qualitative data as supporting data in determining the customer credit worthiness

REFERENCES

[1] Australian Bureau of Statistics. (2013, July 4). Retrieved from https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+quantitative+and+qualitative+data

[2] Astiko. (1995). Manajemen Perkreditan. Yogyakarta: Andi Offset.

[3] Leong, K. C. (2015). Credit Risk Scoring With Bayesian Network Models . Journal Computational, 47 (3), 423-446. https://doi.org/10.1007/s10614-015-9505-8

[4] Luo, C., Wu, D., & Wu, D. (2016). A Deep Learning Approach For Credit Scoring Using Credit Default Swaps. Applications Of Artificial Intelegence, 465-470.

[5] Hanley, E. W., & Hand, G. J. (1997). Construction of a k-nearest neighbour credit scoring system. IMA Journal of Management Mathematics, 305-321. https://doi.org/10.1093/imaman/8.4.305

[6] Johnson, W. R., & Kallberg, G. J. (1986). Management of accounts receivable and payble. New Yyork: Wiley.

[7] Thomas, C. L. (2000). A Survey of Credit and Behavioral Scoring : forecasting financial risk of lending to consumers. International Journal Of Forecasting, 16, 149-172.

[8] Brill, J. (1998). The importance of credit scoring models in improving cash flow and collection. Business Credit.

[9] Brigham, F. E. (1992). Fundamentals of Financial management (6th ed). Forth Worth : Dryden.

[10] Departemen Koperasi dan UMKM. (2013). Retrieved from Departemen Koperasi: http://www.depkom.go.id/index.php?option=com_ (Departemen Koperasi dan UMKM , 2013)phocadownload&view=cat egory&id =126; rencana-kerja-pemerintah&Itemid =93

[11] Bouliceul, J. F., Espoito, F., Giannotti, F., & Pedreschi, D. (2004). Machine Learning ECML. Spinger, Science & Business Media.

[12] Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Boston: John Wiley & Sons. https://doi.org/10.1002/9781118029145

[13] Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Andi Publisher.

[14] Kristiyani, N. (2011). Sistem Pendukung Keputusan dengan Menggunakan Algoritma Iterative Dichotomizer Three (Studi Kasus Sistem PT.Warna Agung Semarang). AITI Jurnal Teknologi Informasi, 8 (1).

[15] Nuraeni, N. (2017). Penentuan Kelayakan Kredit Dengan Algoritma Naïve Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC. Teknik Komputer AMIK BSI, Voll III.

[16] maimon, o. z., & rokach, l. (2005). Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications. World Scientific,. https://doi.org/10.1142/5686

[17] Tan, N. P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Pearson Education Inc.

[18] Turban, E., Sharda, R., Delen , D., & King, D. (2011). Business Intelegence A Managerial Approach. Pearson Education.

[19] Hamid, A. J., & Agmed, T. M. (2016). Developing Prediction Model of Loan RiskIn Banks Using Data Mining. Machine Learning & Application.

[20] Afeni, B. o., Aruleba, T. I., & Oloyede, A. I. (2017). Hypertension Prediction System Using Naive Bayes Classifier. Journal of Advances in Mathematics and Computer Science, 10. https://doi.org/10.9734/JAMCS/2017/35610