ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 252-256 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Utilization of Data Mining to Predict Non-Performing Loan

Yosaphat Catur Widiyono*[,1], Sani Muhamad Isa [2]

[1]*Computer Science Department, Binus Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia*

[2]*Computer Science Department, Binus Online Learning, Bina Nusantara University, Jakarta 11480, Indonesia*

ARTICLE INFO

Article history:
Received: 18 June, 2020
Accepted: 09 July, 2020
Online: 28 July, 2020

Keywords:
Data Mining
Non-Performing Loan
Machine Learning

ABSTRACT

In the banking industry, the existence of problem loans is inevitable. NPL (Non-Performing Loan) will certainly have an impact on the reduction in the capital of a bank. One good step in reducing the risk of credit default or the emergence of non-performing loans is to take proper care of debtors who begin to experience payment constraints. The main obstacle experienced in bank management, especially in the credit sector, is being unable to identify or detect potential debtors early due to a large amount of data and manual processing. In this study, the debtor payment history is presented as data to predict the existence of problem loans. History payment can be used to predict bad loans. The technic of data mining in this experiment is a new method. The results of research conducted using Naïve Bayes, Decision Tree, K-NN, Rule Induction, Logistic Regression, Random Forest, Generalized Linear Model, and Gradient Boosted Trees as a comparison then choose the method that has the highest accuracy to be implemented in making additional modules on the core banking system. Random Forest is the model that has the highest accuracy of 96.55%.

1. Introduction

The banks are one of the economic drivers having an important role in participating in advancing the people's economy. Bank is a financial institution that has the authority to collect funds from the public and channel it back to the community in the form of working capital loans to improve the standard of living of the general public. In its operations, banking is overseen by the Financial Services Authority (OJK). The Financial Services Authority is a state institution formed under the Act that functions to organize a system of regulation and supervision that is integrated into all activities in the financial services sector both in the banking sector, capital market, and financial services sector. Credit debtors are bank customers who receive loan funds that have been agreed through a credit agreement made between the bank and the debtor [1]. NPL or Non-Performing Loan is one of the financial ratios that reflects credit risk. NPL is defined as a loan that has problems/difficulties in repayment or is often called bad credit [2]. Credit collectibility has 5 groups, namely current, special attention, substandard, doubtful, and loss. If it is associated with credit ollectibility, then what is included in non-performing loans are credit that has a substandard, doubtful, and bad collectibility.

The challenge faced is how can the data mining process predict the emergence of problem loans? How are the results of trials conducted using several models for comparison?

The purpose of this study is to analyze the existing credit debtor payment data at XYZ Bank so that the payment patterns can be identified. This pattern can be used to predict non-performing loans in the next 3 months. The benefit of this research is that it helps companies to predict problem loans with active debtors so that they can be applied in credit management to make it more effective and efficient to use priority scale in determining actions to handle potential debtors.

2. Related Works

The research relates to the factors causing the occurrence of problem loans, states that from the 24 existing variables formed 8 factors that have a contribution and influence on problem loans. The factors that have the highest contribution weight are credit period (credit term), loan amount (ceiling), and loan interest rates to be aspects that need to be considered in granting credit. A study by building a multi-dimensional and multi-level credit risk

*Corresponding Author: Yosaphat Catur Widiyono, Email: yosaphat.widiyono@binus.ac.id

https://dx.doi.org/10.25046/aj050431

indicator system aims to find the most important credit risk characteristics that will cause serious default risks. With the existence of several algorithms that can be used in research related to the C4.5 algorithm it has been done [3], which is where they compare the performance of ID3 with C4.5. And the results are better using C4.5 [4]. The application of the C4.5 algorithm in evaluating scholarship granting to students has been conducted research [5]. In that study Wang et al. have the objective to analyze the relationship between student performance and the scholarships provided and by using the C4.5 decision tree algorithm in the scholarship evaluation system it is hoped that the scholarships will be efficient and fair and can be realized. In the study, testing was also carried out to compare the use of different algorithms, namely: C4.5, ID3, Fuzzy Mathematics and Set Pair Analysis. The comparison results of the 4 models C4.5 are the best. In [6], the author build and test classification models to predict student success in English exams. As in the study conducted in [7], they did an experiment using different values for the s parameter to execute C-C4.5. The results of this study indicate that the split criteria of the C-C4.5 algorithm is stronger against noise than the C4.5 criteria. Comparison with more models including Random Forest [8]. Shamshur and Weill conducted a study by examining the impact of bank efficiency on credit costs [9]. In this study combining company-level data with bank-level data so that it can identify the level of efficiency of banks that lend money to each company. Estimation was then carried out using a large data sample from 240,000 companies from nine countries in Europe. Their main finding is that higher bank efficiency is associated with lower credit costs. They, therefore, support the view that the effectiveness of banks in minimizing costs is transferred to the borrowing company through lower credit costs. Ye et al. conduct research on P2P lending loans [10]. The experiment resulted in contributions in the form of methods that made it possible to improve the quality of credit in P2P loans provided to borrowers. Moradi et al. proposed a dynamic model for credit risk assessment that outperforms the models currently used. Our model has a dynamic engine that assesses the behavior of bad customers on a monthly basis and a fuzzy inference system (FIS) that includes the factors of credit risk, especially in economic crises[11]. Previous research has used data: Gender, Age, Amount of Credit, Monthly Income, Monthly Expenditures, Current Payments Per Month, Savings (Income Payment), Type of Collateral, Collateral Value, Loan Period, Type of Business Activity, Sources of Funds, Credit Status previous. Data mining is used to suggest a decision tree model for credit assessment as it can indicate whether the request of lenders can be classified as performing or non-performing loans risk. Using C 5.0 methodology, a new decision tree model is generated [12].

3. Methodology

Experiments carried out using the CRISP-DM methodology that offers a structured approach to data mining. This research will be carried out in a 5-step process, namely business understanding, data understanding, data preparation, modeling, and evaluation. Tools or software used to process data are MS Excel 2016 and Rapid Miner 9.6.000. Several classification algorithms are used in modeling and testing. The study was conducted based on the CRISP-DM framework as shown in Figure 1.

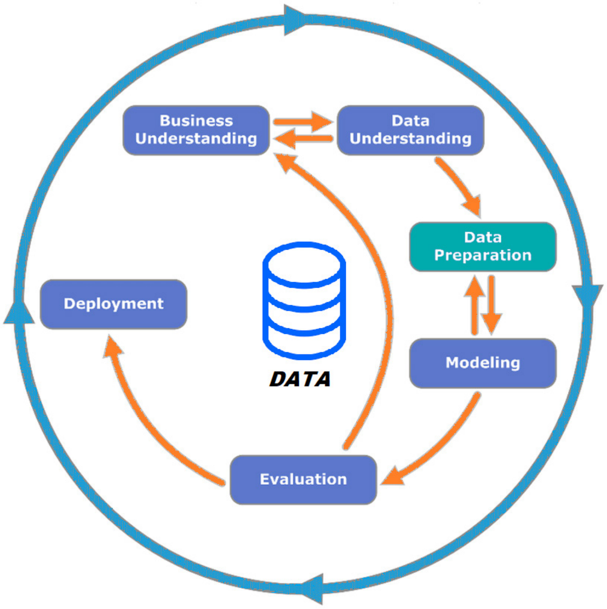


Figure 1: CRISP-DM Model

This study seeks to solve the challenge of how to predict the quality of debtors' credit based on payment history more precisely [13]. Some of the Algorithms we tested included the Decision Tree Algorithm which was chosen because it had many advantages. A trial was conducted to compare several methods in their ability to predict more accurately. The steps are taken in the collection of this data as shown in Figure 2.

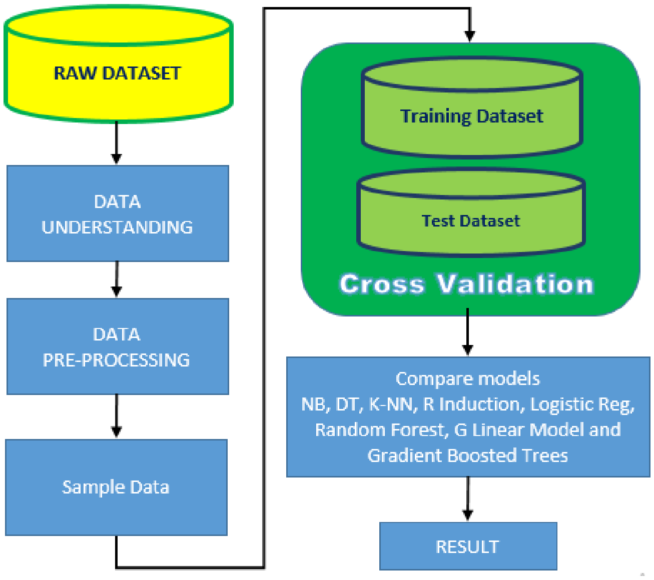


Figure 2: Experimental Design

4. Theory and Methods

Data mining is the activity of extracting to obtain important information that is implicit and previously unknown, from a data. The use of data mining for prediction has also been widely applied and continues to be studied. The algorithm used in this study, among others, the Decision Tree algorithm, is a series of algorithms for classification problems in a machine and data set. Next, the Naive Bayes algorithm is a classification algorithm based on the Bayesian theorem in statistics. The Naive Bayes algorithm can be used to predict the probability of class membership. The Bayesian theorem is a fundamental statistical approach to pattern recognition. Naive Bayes is based on the simplification assumption that attribute values are conditionally independent if output values are given [14]. K-Nearest Neighbor enters the classification algorithm so this algorithm can be used to predict new classes from datasets that have classes. K-Nearest Neighbor (K-NN) algorithm is a method for classifying a set of data based on learning data that has been classified previously [15]. This algorithm is also one of the lazy learning techniques. K-NN searches the k group of objects in the training data closest to or similar to the object in new data or test data.

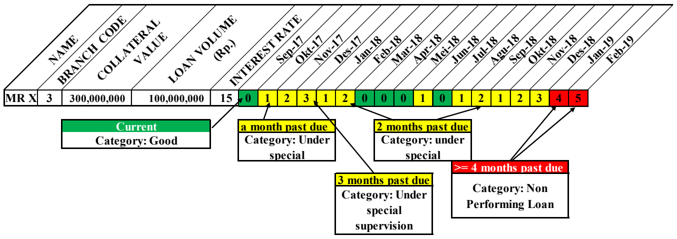
5. Proposed Method and Results

5.1. Business Understanding

XYZ Bank needs a way to do the Non-Performing Loan problem solving faster. Therefore, effective and early action is needed. With this effective action, the company can minimize losses that arise in the future due to Non-Performing Loans. Therefore, with debtor payment history data, machine learning will be made that can predict problem debtors in the future. By utilizing the results, the team that handles the problem of non-performing loans can take action earlier by using a priority.

5.2. Data Understanding

The data prepared is data on credit payment history every month from September 2017 to February 2020. Data collection is done by downloading arrears reports from XYZ Bank's core banking system. Initial data processing using MS Excel software. The arrears data is then processed into credit history data for each credit account each period. Within 1 period is 9 months of payment history. Debtor payment history data structure as shown in Figure 3.



NAME	BRANCH CODE	COLLATERAL VALUE	LOAN VOLUME (Rp.)	INTEREST RATE	Sep-17	Okt-17	Nov-17	Des-17	Jan-18	Feb-18	Mar-18	Apr-18	Mei-18	Jun-18	Jul-18	Agu-18	Sep-18	Okt-18	Nov-18	Des-18	Jan-19	Feb-19
MR X	3	300,000,000	100,000,000	15	0	1	2	3	1	2	0	0	0	1	0	1	2	1	2	3	4	5

Figure 3: Debtor payment history data structure

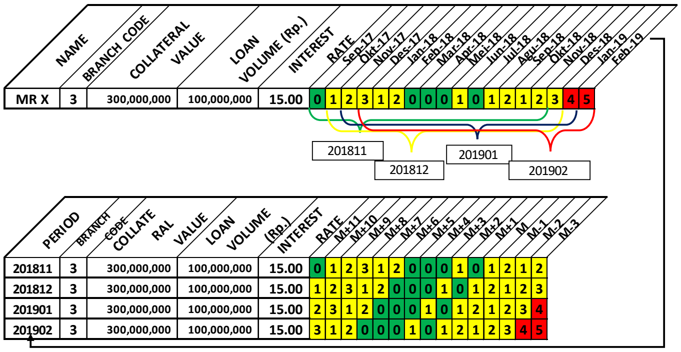
The long historical data is then divided into periods per month with a stipulation that the length of the month is 15.

5.3. Data Preparation

The data that has been collected will be identified, selected with the help of MS Excel application software. For data classification, it is necessary to transform data on several data that will be used as attributes. Next is the evaluation of data requirements, attribute determination, data type, and class. The attributes that will be used in this study are:

- Branch Code (Representation of the region)
- Collateral Value (Classified as BIG and SMALL)
- Loan Volume (Classified as BIG and SMALL)
- Interest rate (Classified as BIG and SMALL)
- History of late payment for 12 months (M+11, …, M+5, M+4, M+3, M+2, M+1, and M).

While the Label is M-3 (Prediction 3 months) with a value of {NOT and NPL}. NOT means it is not included in the NPL (Non-Performing Loan) category, which is a delay of <4 months. Whereas NPL means to be included in the category of Non Performing Loans, which is a delay of > 3 months. The process of collecting debtor delays data for each specified period can be seen in Figure 4.

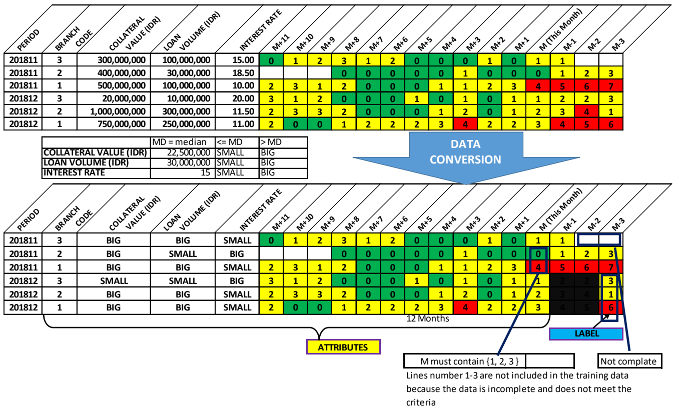


NAME	BRANCH CODE	COLLATERAL VALUE	LOAN VOLUME (Rp.)	INTEREST RATE	Sep-17	Okt-17	Nov-17	Des-17	Jan-18	Feb-18	Mar-18	Apr-18	Mei-18	Jun-18	Jul-18	Agu-18	Sep-18	Okt-18	Nov-18	Des-18	Jan-19	Feb-19
MR X	3	300,000,000	100,000,000	15.00	0	1	2	3	1	2	0	0	0	1	0	1	2	1	2	3	4	5

PERIOD	BRANCH ~~CODE~~	COLLATERAL VALUE	LOAN VOLUME (Rp.)	INTEREST RATE	M+11	M+10	M+9	M+8	M+7	M+6	M+5	M+4	M+3	M+2	M+1	M	M-1	M-2	M-3
201811	3	300,000,000	100,000,000	15.00	0	1	2	3	1	2	0	0	0	1	0	1	2	1	2
201812	3	300,000,000	100,000,000	15.00	1	2	3	1	2	0	0	0	1	0	1	2	1	2	3
201901	3	300,000,000	100,000,000	15.00	2	3	1	2	0	0	0	1	0	1	2	1	2	3	4
201902	3	300,000,000	100,000,000	15.00	3	1	2	0	0	0	1	0	1	2	1	2	3	4	5

Figure 4: Data Collection With Conversion

After the data is collected, data conversion is required . Data conversion methods as shown in Figure 5.



PERIOD	BRANCH CODE	COLLATERAL VALUE (IDR)	LOAN VOLUME (IDR)	INTEREST RATE	M+11	M+10	M+9	M+8	M+7	M+6	M+5	M+4	M+3	M+2	M+1	M (This Month)	M-1	M-2	M-3
201811	3	300,000,000	100,000,000	15.00	0	1	2	3	1	2	0	0	0	1	0	1	1		
201811	2	400,000,000	30,000,000	18.50				0	0	0	0	0	1	0	0	0	1	2	3
201811	1	500,000,000	100,000,000	10.00	2	3	1	2	0	0	0	1	1	2	3	4	5	6	7
201812	3	20,000,000	10,000,000	20.00	3	1	2	0	0	0	1	0	1	0	1	1	2	2	3
201812	2	1,000,000,000	300,000,000	11.50	2	3	3	2	0	0	0	1	2	0	1	2	3	4	1
201812	1	750,000,000	250,000,000	11.00	2	0	0	1	2	2	2	3	4	2	2	3	4	5	6

	MD = median	<= MD	> MD
COLLATERAL VALUE (IDR)	22,500,000	SMALL	BIG
LOAN VOLUME (IDR)	30,000,000	SMALL	BIG
INTEREST RATE	15	SMALL	BIG

PERIOD	BRANCH CODE	COLLATERAL VALUE (IDR)	LOAN VOLUME (IDR)	INTEREST RATE	M+11	M+10	M+9	M+8	M+7	M+6	M+5	M+4	M+3	M+2	M+1	M (This Month)	M-1	M-2	M-3
201811	3	BIG	BIG	SMALL	0	1	2	3	1	2	0	0	0	1	0	1	1		
201811	2	BIG	SMALL	BIG				0	0	0	0	0	1	0	0	0	1	2	3
201811	1	BIG	BIG	SMALL	2	3	1	2	0	0	0	1	1	2	3	4	5	6	7
201812	3	SMALL	SMALL	BIG	3	1	2	0	0	0	1	0	1	0	1	1	2	2	3
201812	2	BIG	BIG	SMALL	2	3	3	2	0	0	0	1	2	0	1	2	3	4	1
201812	1	BIG	BIG	SMALL	2	0	0	1	2	2	2	3	4	2	2	3	4	5	6

Lines number 1-3 are not included in the training data because the data is incomplete and does not meet the criteria

Figure 5: Data Conversion

From the data conversion, a total of 175,913 data were obtained, with the composition as shown in Table 1.

Table 1: Composition of Dataset

Class	Records	%
NOT	**153,856**	**87%**
NPL	**22,057**	**13%**

5.4. Modeling

By using the Rapidminer V 9.6 application, this study tested several models using the cross-validation method as shown in Figure 6. Because the composition of the data labeled NPL is far less than the data labeled NOT (not NPL), the sampling technique is done by taking all data labeled NPL and labeled NOT each as much as the amount of data labeled NPL. The implementation of the model in the Rapidminer application as shown in Figure 7.

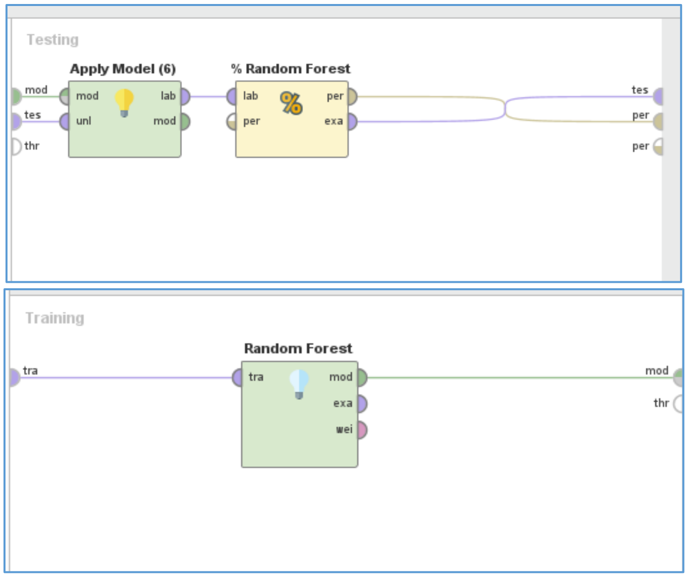


Figure 6: Cross-Validation Design

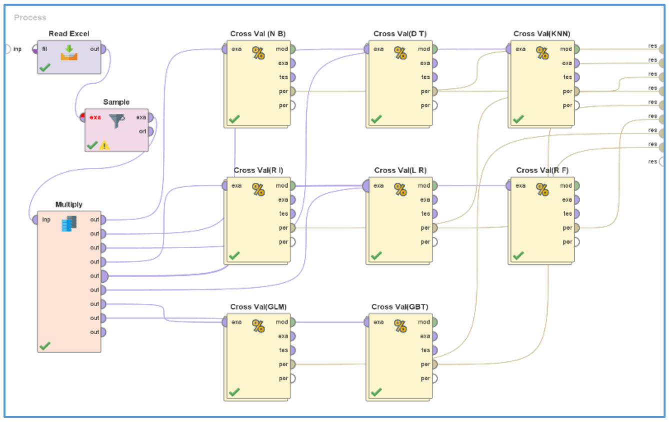


Figure 7: Model Implementation in the Rapidminer

Accuracy, AUC, F_measure, Sensitivity, Specificity between each model can be compared [15]. The models that have been tested are Naïve Bayes (NB), Decision Tree (DT) [16], K-NN, Rule Induction, Logistic Regression , Random Forest, Generalized Linear Model dan Gradient Boosted Trees.

5.5. Evaluation

To see the accuracy of the model to each class, in this study using a confusion matrix [17]. By calculating the accuracy of some test data, the effectiveness of classification can be seen. From the several models tested, the results as shown in Figure 8.

In testing, the technique used in drawing data is balanced in each class {NOT and NPL}. From the comparison of test results, it was found that Random Forest was the highest accuracy. Accuracy is calculated from the total number of two correct predictions (TP + TN) divided by the total number of datasets (P + N). An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever. Random Forest - AUC: 0.996 +/- 0.000 (micro average: 0.996) (positive class: NPL) as shown in Figure 9.

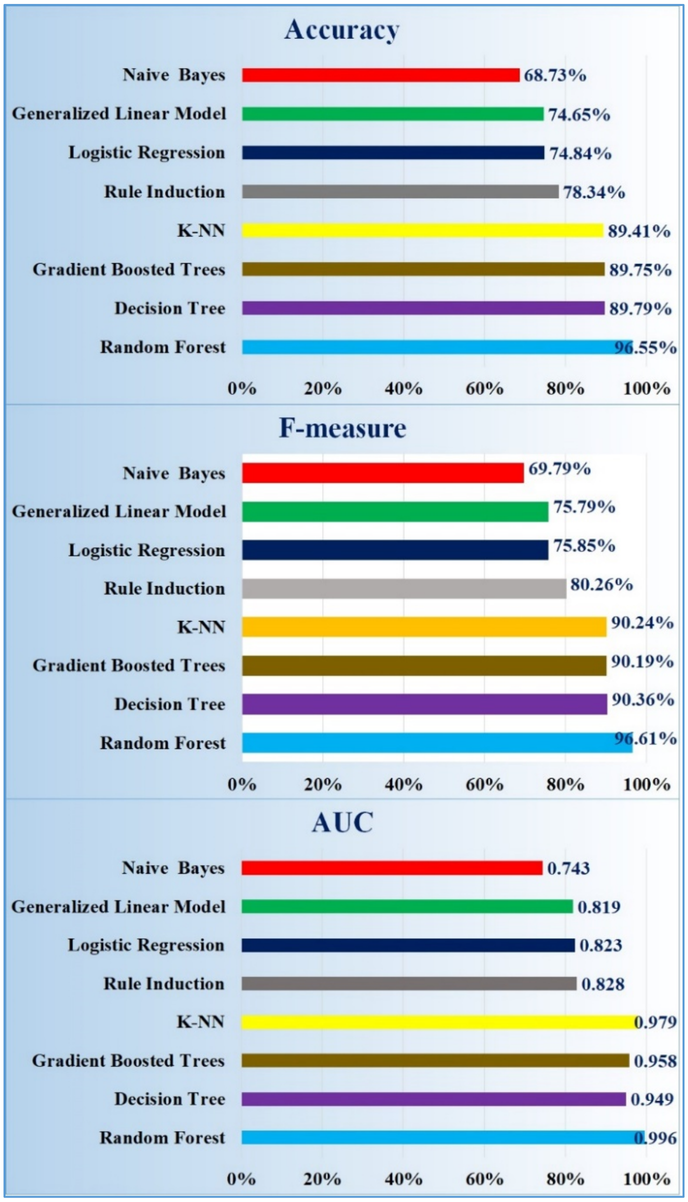


Figure 8: Graph for Accuracy, F-measure, and AUC on 8 models

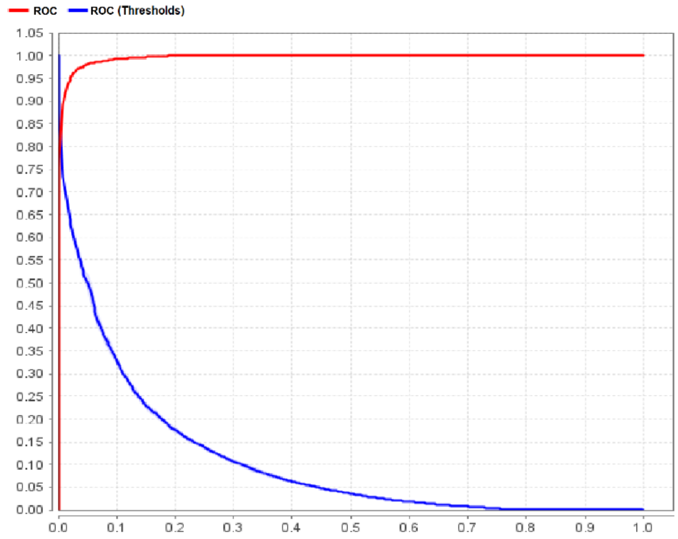


Figure 9: AUC – Random Forest

From the experiments conducted can show that the payment history can be used to predict the possibility of debtors becoming problematic in their payments. This is of course because the payment history is an orderly description of the debtor in fulfilling his obligations.

6. Conclusion

This research has discussed how historical data or loan repayment patterns can be studied through data mining and generate new knowledge to be able to predict future possibilities more accurately. In other studies it has been discussed that the dynamic behavior of the debtor during the last few years is more able to determine the condition of the debtor in the future which has an impact on the level of bad loans[11]. In this experiment, the data used are dynamic in the form of historical payments and payment patterns that can change from time to time, so that it can be better in terms of predicting uncertain conditions such as macroeconomic conditions, seasons, etc. Instead of using static data such as gender, marital status, employment, etc. less able to predict precisely [12]. The choice of attributes is very influential on the accuracy of the model made. One example is from this experiment found that the longer the payment history that we use, the more accurate the results obtained. In this experiment, it was found that factors such as branch codes also influence the determination of prediction results. The branch code represents the level of personal ability at the branch in handling this non-performing loan. Then the Collateral Value, Interest Rate and Loan Volume also have a correlation to the emergence of this NPL. Several models have been tested and compared to choose the best. Then it can be considered for implementation on the core banking system used by XYZ Bank. After predicting the NPL, we can use the data to be submitted to the relevant officers to follow up. Data mining in this study is still limited to credit loans with flat payment systems. From several models that have been tested in this experiment, it can be concluded that the Random Forest classification method is the most precise in predicting. It is expected that in subsequent studies it can predict types of credit models such as credit with seasonal payments, credit cards, and other models. Implementation of the model in the core banking system can be done by adding an NPL prediction module that processes data on the Database Server and provides reports and views to the user. The architecture of adding NPL prediction modules as shown in Figure 10.

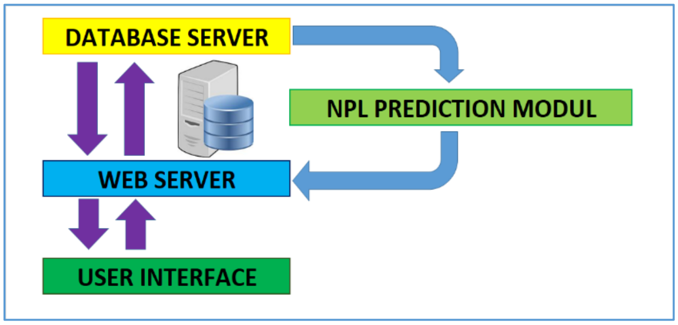


Figure 10: The architecture of adding NPL prediction modules

References

[1] M. B. Alexandri and T. I. Santoso, "Non Performing Loan: Impact of Internal and External Factor (Evidence in Indonesia)," *Int. J. Humanit. Soc. Sci. Invent.*, **4**(1), 87–91, 2015.

[2] R. Agustiningrum, "Analisis Pengaruh Car, Npl, Dan Ldr Terhadap Profitabilitas Pada Perusahaan Perbankan," *E-Jurnal Manaj. Univ. Udayana*, **2**(8), 255030, 2013.

[3] H. Elaidi, Z. Benabbou, and H. Abbar, "A comparative study of algorithms constructing decision trees: Id3 and c4.5," *ACM Int. Conf. Proceeding Ser.*, 3–7, 2018. doi: 10.1145/3230905.3230916.

[4] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' dropout indicators in public school using data mining approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, **9**(1), 774–778, 2020. doi: 10.30534/ijatcse/2020/110912020.

[5] X. Wang, C. Zhou, and X. Xu, "Application of C4.5 decision tree for scholarship evaluations," *Procedia Comput. Sci.*, **151**(2018) 179–184, 2019. doi: 10.1016/j.procs.2019.04.027.

[6] W. Puarungroj, N. Boonsirisumpun, P. Pongpatrakant, and S. Phromkhot, "Application of data mining techniques for predicting student success in English exit exam," *ACM Int. Conf. Proceeding Ser.*, 1–6, 2018. doi: 10.1145/3164541.3164638.

[7] C. J. Mantas, J. Abellán, and J. G. Castellano, "Analysis of Credal-C4.5 for classification in noisy domains," *Expert Syst. Appl.*, 2016. doi: 10.1016/j.eswa.2016.05.035.

[8] H. Byeon, "A Prediction Model for Mild Cognitive Impairment Using Random Forests," *Int. J. Adv. Comput. Sci. Appl.*, **6**(12), 8–12, 2015. doi: 10.14569/ijacsa.2015.061202.

[9] A. Shamshur and L. Weill, "Does bank efficiency influence the cost of credit?," *J. Bank. Financ.*, 2019. doi: 10.1016/j.jbankfin.2019.05.002.

[10] X. Ye, L. an Dong, and D. Ma, "Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score," *Electron. Commer. Res. Appl.*, 2018. doi: 10.1016/j.elerap.2018.10.004.

[11] S. Moradi and F. Mokhatab Rafiei, "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks," *Financ. Innov.*, **5**(1), 2019. doi: 10.1186/s40854-019-0121-9.

[12] I. G. N. N. Mandala, C. B. Nawangpalupi, and F. R. Praktikto, "Assessing Credit Risk: An Application of Data Mining in a Rural Bank," *Procedia Econ. Financ.*, **4**, Icsmed, 406–412, 2012. doi: 10.1016/s2212-5671(12)00355-3.

[13] O. R. Devi, "International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://www.warse.org/ijatcse/static/pdf/file/ijatcse02422015.pdf," **4**(2), 15–21, 2015.

[14] N. Sun, B. Sun, J. (Denny) Lin, and M. Y. C. Wu, "Lossless Pruned Naive Bayes for Big Data Classifications," *Big Data Res.*, 2018. doi: 10.1016/j.bdr.2018.05.007.

[15] A. Singh, M. N., and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, **8**(12), 1–10, 2017, doi: 10.14569/ijacsa.2017.081201.

[16] M. Hoechstoetter, A. Nazemi, and S. T. Rachev, "Recovery Rate Modelling of Non-performing Consumer Credit Using Data Mining Algorithms," 12, 2012.

[17] N. S. Buot, "Multiple intelligences and reading comprehension of senior high school students: A response evaluation through educational data mining technique," *Int. J. Adv. Trends Comput. Sci. Eng.*, **8**(6), 2871–2876, 2019, doi: 10.30534/ijatcse/2019/30862019.