JOIN (Jurnal Online Informatika)
Volume 7 No. 1 | June 2022: 7-16
DOI: 10.15575/join.v7i1.762

p-ISSN: 2527-1682
e-ISSN: 2527-9165

Classification of the Fluency Multipurpose of Bank Mandiri Credit Payments Based on Debtor Preferences Using Naive Bayes and Neural Network Method

Putri Armilia Prayesy[1], Edi Surya Negara[2]
[1,2]Department of Informatics, Data Science Interdisciplinary Research Center, Bina Darma University

Article Info

Article history:

Received June 11, 2021
Revised June 21, 2021
Accepted June 28, 2021
Published June 30, 2022

Keywords:

Classification
Credit Debtors
Naive Bayes
Neural Network

ABSTRACT

One that has an important role in generating bank profits is providing credit to customers, but credit also carries a very high risk. For this reason, in providing credit to debtors, of course the bank will utilize the personal data of prospective debtors in detail to avoid the risk of problems that will arise in the future. One of the appropriate risks for banks in providing credit is the behavior of customers who do not pay installments at the time which causes bad loans. To overcome and overcome the many bad events, there is an algorithmic calculation method with an intelligent computing system that helps banks in selecting prospective debtors who will be given credit. There are many algorithmic methods that can be used in this kind of research. This study analyzes the classification of staffing credit based on the criteria that become the Bank's standard.The data used by the author in this study uses existing debtor credit data from 2017 to 2020, the modeling process is carried out using split validation with the Naive Bayes algorithm and Neural Network, with this algorithm the 1,314 datasets is divided into 2 parts, namely 80% used as training data and 20% used as testing data.

The results showed that the Neural Network algorithm has better results with a correct value of 84.13%, while the Naive Bayes algorithm only produces a value of 72.62%.

Corresponding Author:

Edi Surya Negara
Department of Informatics,
Data Science Interdisciplinary Research Center,
Bina Darma University
Jl. Jendral Ahmad Yani No 3 Palembang, Indonesia
Email: e.s.negara@binadarma.ac.id

1. INTRODUCTION

A bank is a financial institution that carries out operational activities to serve customers in the form of deposit and withdrawal transactions, investment, customer complaint services, and lending. Bank Mandiri is a state-owned company (BUMN) that operates as a provider of financial services in Indonesia, including cash processing, payment processing, foreign exchange, vehicle, and residential loan financing and provides loan solutions for companies, retail entrepreneurs, and employees.

One of the risks experienced by banks in providing credit is the behavior of customers who do not pay installments on time or delay payments, causing bad loans. To avoid the occurrence of bad credit, banks set requirements in the credit application process as a mitigation of the risk of bad credit.

To overcome and reduce the number of bad credit, it is necessary to have a calculation algorithm method with an intelligent computing system that helps banks in selecting prospective debtors to be given credit.

This study analyzes the risk of providing credit from various aspects that are taken into consideration in the process of providing credit for personnel based on the criteria that become the Bank's standards. Researchers use the *Naive Bayes* algorithm and *Neural Network* as a comparison of the accuracy of customer payments and Rapid Miner as the tools.

Maryam Hasan, In this research Predicting the smoothness of credit payments using the Naive Bayes algorithm based on Forward Selection which is able to predict the smoothness of credit payments in the future,

e-ISSN 2527-1682
p-ISSN 2527-9165

this is proven by the acquisition of the *Naive Bayes* accuracy value based on Forward Selection, which is able to achieve an accuracy value of 71.97%[5].

Yuli Murdianingsih conducting research on the classification of good and problem customers using the Naive Bayes method Obtained a web-based system using PHP and MySQL that can determine the problematic and good customers using the Naive Bayes method. The magnitude of the probability value of the class value resulting from manual calculations is the same as the magnitude of the probability value of the system result [13].

Amrin conducted Based on the results of the study, the performance of the backpropagation *neural network model* for providing car loans formed from training data and validated with an accuracy level of 89% with an area under the curve (AUC) value of 0.831[2].

From several studies can be seen the difference in the method of classification and accuracy results. The purpose of this study is to determine the classification process and the results of the accuracy of the comparison between the *Naive Bayes* and *Neural Network* methods so that the best model used in this study could be found.

2. METHOD

Data mining can be applied to explore the value of a knowledge data set that has not been known manually [17]. Data mining and Knowledge Discovery in Databases (KDD) are often used interchangeably in describing the process of extracting information in very large databases but are related to each other, the schema of the KDD process [4], can be seen in the following image:

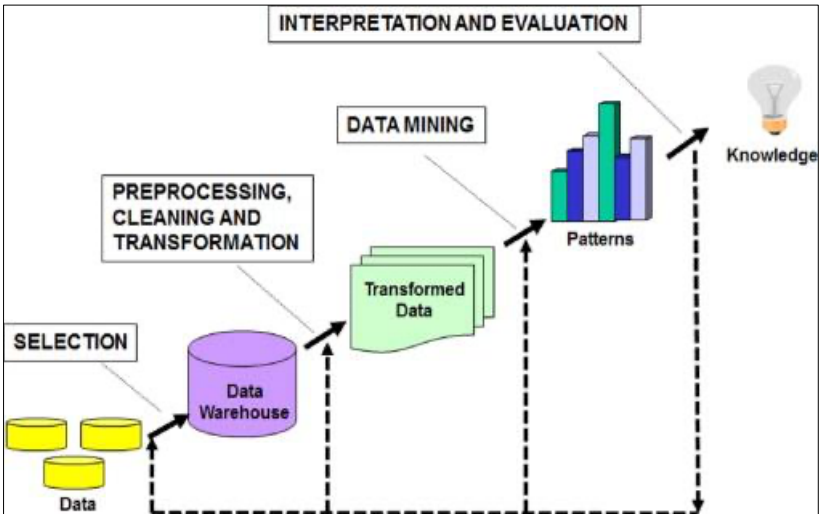

Figure 1. Schematic KDD

From the stages of the KDD process is data mining, the stages are as follows [3]:

1. Data Selection, where the data relevant to the analysis is decided and taken from the data location.
2. Transformed Data, at this stage the data cleaning process is carried out and checks if there are gaps or incompleteness, duplication of data, or irrelevant data. so that the data is in accordance with the data mining process.
3. Data Mining, the process of finding the appropriate pattern from the data you want to display using certain techniques and methods.
4. Interpretation/Evaluation, this stage is translating the patterns and methods that have been generated from data mining and also how the information can be easily understood.

In this study, we will classify the Smoothness of Mandiri Multipurpose Credit Payments using the *Naive Bayes* Method and *Neural Network* as the algorithm. In classifying the smoothness of debtor credit payments, several variables are used to test the accrual rate such as type of work, application limit, application period (tenor), credit installments, income, years of service, number of dependents, types of payroll, years of service and credit status.

In order for this research to be directed and in accordance with the expected goals, the research methodology uses a research framework. The research flowchart can be seen in Figure 2:

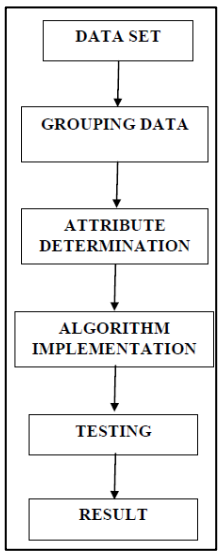

Figure 2. Research Flowchart

The flowchart Classification of the smooth rate of independent multipurpose credit payments using the *Naive Bayes* algorithm and *Neural Network* is shown in Figure 3 below:

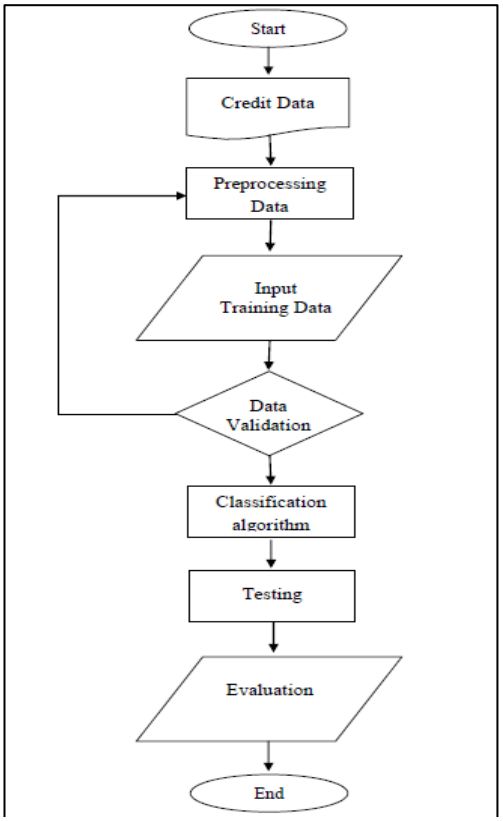

Figure 3. *Flowchart* Classification

Figure 3 describes the flow of the classification process, starting with credit data preprocessing, then inputting the data into rapidminer for the classification process, after the data has been successfully input, data validation is carried out to check for missing data or inconsistent data if there is missing data. or inconsistent, then re-preprocessing is carried out. After validation of credit data, modeling is carried out using the *Naive Bayes* algorithm and *Neural Network*, modeling is used to test accuracy, precision, and recall on credit data.

e-ISSN 2527-1682
p-ISSN 2527-9165

The results of the *Naive Bayes* algorithm and *Neural Network* can be used by the bank in making decisions to provide credit to prospective debtors.

2.1. Data Research

To support this research, data collection methods are needed. The research data consists of sources, data collection techniques, types, and sources of data along with data analysis.

Table 1. Sample Data

No	Debtors Name	Job	Limit	Time Period	Installment	Income	Length Of Work	Dependents	Payroll	Collateral	Credit Status
1	A Fathoni Arief	BUMN	Rp. 800.000.000	72	Rp. 15..953.920	> 25	15	3	Payroll	No	Good
2	A Frans Saputra	SWASTA	Rp. 50.000.000	60	Rp. 1.352.668	> 5	9	2	Payroll	Yes	Good
3	A Rachman Rachim	BUMN	Rp. 200.000.000	90	Rp. 4.004.812	> 5	10	3	Payroll	No	Good
4	Aan Irawan	SWASTA	Rp. 74.000.000	60	Rp. 1.947.300	> 5	6	1	Payroll	Yes	Good
5	Abdoel Gani	SWASTA	Rp. 93.000.000	76	Rp. 2.167.003	> 5	7	2	Payroll	Yes	Arrears
6	Abdul Rachman	BUMN	Rp. 200,000.000	90	Rp. 4.004.812	> 5	13	3	Payroll	No	Good
7	Abdul Rohim	SWASTA	Rp. 71.000.000	60	Rp. 1.881.066	> 5	8	2	Payroll	Yes	Good
8	Abdul Somad Fauzi	SWASTA	Rp. 45.000.000	45	Rp. 1.429.376	> 5	8	2	Payroll	Yes	Good
9	Abdur Roni Rachman	PNS	Rp. 50.000.000	60	Rp. 1.320.833	> 5	7	4	Payroll	Yes	Good
10	Abi Alhadi	BUMN	Rp. 130.00.000	60	Rp. 2.924.732	> 5	8	1	Payroll	No	Good
11	Abu Hasan Asyari	BUMN	Rp. 340.000.000	72	Rp. 7.497.437	> 10	11	2	Payroll	No	Arrears
12	Achiruddin	SWASTA	Rp. 100.000.000	60	Rp. 2.325.788	> 5	11	3	Payroll	Yes	Good
13	Achmad Azhari	BUMN	Rp. 210.000.000	170	Rp. 2.991.620	> 5	9	0	Payroll	No	Good
14	Achmad Djailani	SWASTA	Rp. 53.000.000	36	Rp. 1.995.146	> 5	9	2	Payroll	Yes	Good
15	Achmad Fauzi	BUMN	Rp. 180.000.000	60	Rp. 4.049.629	> 5	12	1	Payroll	No	Good
16	Achmad Firmansyah Hidayatulah	BUMN	Rp. 150.000.000	96	Rp. 2.913.182	> 5	7	2	Payroll	No	Good
17	Achmad Komarudin	SWASTA	Rp. 70.000.000	44	Rp. 1.991.375	> 5	8	1	Payroll	Yes	Good
18	Achmad Rio Fizriyan	BUMN	Rp. 70.000.000	60	Rp. 1.610.689	> 5	7	0	Payroll	No	Good
19	Achmad Rusli	SWASTA	Rp. 45.000.000	36	Rp. 1.672.361	> 5	7	2	Payroll	Yes	Good
20	Achmad Salleh	SWASTA	Rp. 50.000.000	60	Rp. 1.235.918	> 5	7	2	Payroll	Yes	Good
21	Achmad Sentosa	BUMN	Rp. 900.000.000	91	Rp. 15.478.661	> 25	17	2	Payroll	No	Good
22	Achmad Sepriansyah	SWASTA	Rp. 50.000.000	48	Rp. 1.488.418	> 5	9	1	Payroll	Yes	Good
23	Ade Afriansyah	BUMN	Rp. 500.000.000	180	Rp. 6.326.211	> 10	13	2	Payroll	No	Good
24	Ade Arrosyid	BUMN	Rp. 130.000.000	96	Rp. 2.524.757	> 5	9	1	Payroll	No	Good

2.2. Data Source

The source of the data used in this research is based on data related to credit given to debtors, while the source of data used is secondary data. secondary data is something on existing data.

2.3. Algorithm Analysis

Classification requires a number of instructions to determine the class that is analyzed. In this study it was carried out by calculating the prior probability and posterior probability using 1,314 records.

2.3.1 *Naive Bayes*

The current Naïve Bayes Classifier method has been developed to calculate the probabilistic size of each word and provide an assessment for each class [16]. Naïve Bayes is a classification method that is often used to estimate a value whose label is unknown. Naïve Bayes is the fastest and simplest Bayesian Learning method. This is derived from the Bayes theorem and the freedom hypothesis, resulting in a statistical classifier based on probability [10]. The following is the equation formula for Bayes' theorem:

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \quad (1)$$

X : Data with unknown class
H : Hypothesis data is a specific class
P(H|X : Probability of hypothesis H based on condition X (posteriori probability)

P(H) : Hypothesis probability H (prior probability)
P(X|H) : Probability of X based on the conditions on the hypothesis H
P(X) : Probability of X

2.3.2 *Neural Network*

Neural Network is a concept of knowledge engineering in the field of artificial intelligence designed by adopting the human nervous system, whose main processing is in the brain. The smallest part of the human brain is a nerve cell which is called the basic information processing unit or neuron. There are about 10 billion neurons in the human brain and about 60 trillion connections (Synapses) between neurons in the human brain. By using these neuro-neurons simultaneously the human brain can process information in parallel and fast [2].

2.4 Evaluation

The last stage is to evaluate the success rate of predictions made on data processing using the Naive Bayes method so that the results of this study are useful for banks in order to reduce the level of bad loans. The models that have been applied will be compared using a confusion matrix.

Confusion Matrix is a method used to perform accuracy calculations on the concept of data mining. The Confusion Matrix model will form a matrix consisting of true positives, true negatives, false positives and false negatives.

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

3. RESULTS AND DISCUSSION

3.1. Data Analysis

This chapter describes the description of the debtor's credit analysis process using the Naive Bayes algorithm and *Neural Network* to determine the level of accuracy of debtor credit payments in the lending process based on the variables and attributes that have been determined to classify the level of credit payments. The data is first processed by determining training data, target data and testing data, to conduct training data the author takes a sample of credit data that already exists in 2017 to 2020 credit data which will be the target of research data based on current loans and bad loans for 4 years The last one consists of 1,314 data records consisting of 1,089 current credit data and 226 bad credit data using 10 attributes.

Before carrying out the data mining process, the data is cleaned first by removing duplicate data, correcting inconsistent data errors and incomplete data.

All attributes in the dataset will be selected to obtain attributes that contain relevant values, not redundant, and not missing values. Data is classified as missing value if the data contained in the attributes has no value or is empty, while the data is redundant if there is the same data in one record.

The results of the classification will produce an accuracy value that can be considered by the bank in making decisions to provide credit to prospective debtors.

Based on the framework described in chapter II, the stages of the debtor credit classification process are carried out by processing the raw data first, then grouping the data based on each variable, then determining the attributes to be used. The data is processed by applying *Naive Bayes* and *Neural Network* after testing the debtor data, which is the object of this research, the tests that have been carried out have resulted in a quality assessment of the debtor's payment level.

3.2. Classification

Classification is used to predict future credit opportunities based on credit data that has existed in previous years [15]. The *Naive Bayes* classification algorithm and *Neural Network* are used as a comparison of accuracy levels to assist in making credit decisions. The classification process is carried out in split validation.

e-ISSN 2527-1682
p-ISSN 2527-9165

3.2.1 *Naive Bayes*

At this stage, it displays a process to determine the accuracy value using rapidminer tools. The stages of implementation are as follows:

1. Select the data to be used in both excel and csv formats. This study used data with excel format (.xls). then drag and drop the read excel operator.
2. Import the data that is the object of research, then specify the label that will be the classification target.
3. Drag and drop operator Split validation to run the data mining process.

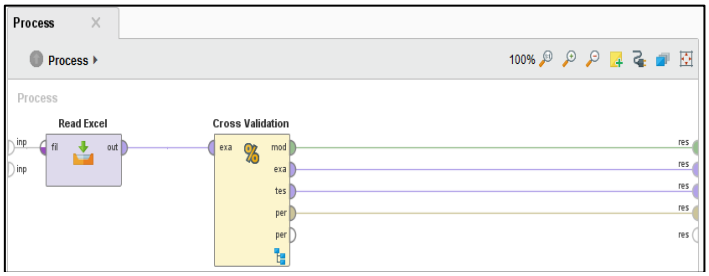

Figure 4. *Naive Bayes* Model Process

Figure 4 validation after reading the data. The read excel operator is connected to the Split validation operator. In the Split validation operator, there are training and testing sub-processes.

Drag and drop *Naive Bayes* operators in the training block, this block is used as a data training process with the Naive Bayes algorithm, while in the testing block, drag and drop applies to operators, apply models and performance operators, this block is used for testing the data and producing accuracy values, precision, recall, AUC, graph or classification pattern.

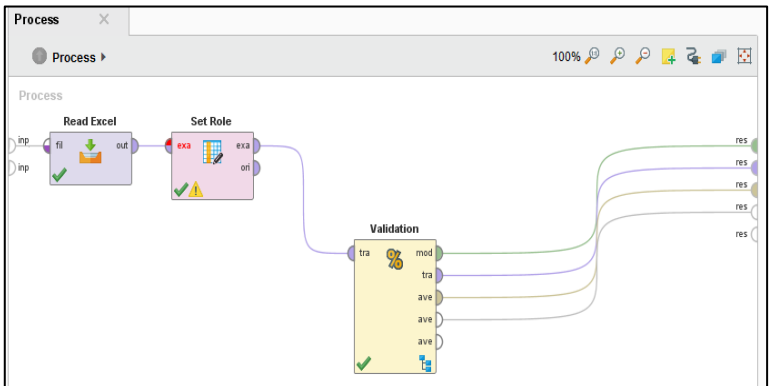

Figure 5. *Split Validation* Sub Process Model

Figure 5 shows the sub-process of split validation, the training block used for the *Naive Bayes* model is connected by a connecting line in the testing block of the apply model operator and operator performance.

Table View Plot View

accuracy: 72.62%

	true Lancar	true Macet	class precision
pred. Lancar	161	15	91.48%
pred. Macet	57	30	34.48%
class recall	73.85%	66.67%	

Figure 6. Results of Confusion Matrix of *Naive Bayes* Algorithm

In Figure 5.4, it can be seen that the accuracy value generated from the yahoo Naive Bayes process is an accuracy value of 72.62%, class recall is 91.48%, while for class precision the current prediction result is 73.85%. Of the 1,314 data that were predicted to be true positive as many as 161 data, as many as 15 false negative data, as many as 57 false positive data predicted true negative as many as 30 data.

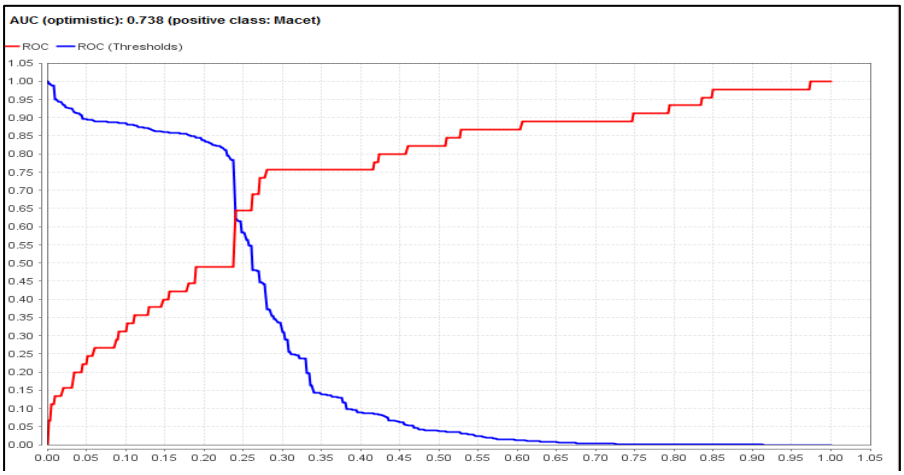

Figure 7. The AUC value on the ROC graph of the *Naive Bayes* Algorithm

In Figure 7 it can be seen that the AUC value is 0.783 with a fair classification accuracy, because it is in the range 0.70 – 0.80.

3.2.1 *Neural Network*

At this stage, it displays a process to determine the value of accuracy by using rapidminer tools. The stages of implementation are as follows:

1. Select the data to be used in both excel and csv formats. This study used data with excel format (.xls). then the drag and drop operator reads excel.
2. Import the data that is the object of research, then specify the label that will be the classification target.
3. Drag and drop operator to normal. Normalization is carried out in data mining, this operator works for the process of scaling the attribute values of the dataset so that it can be in a certain range.
4. Drag and drop operator split validation to run the data mining process.

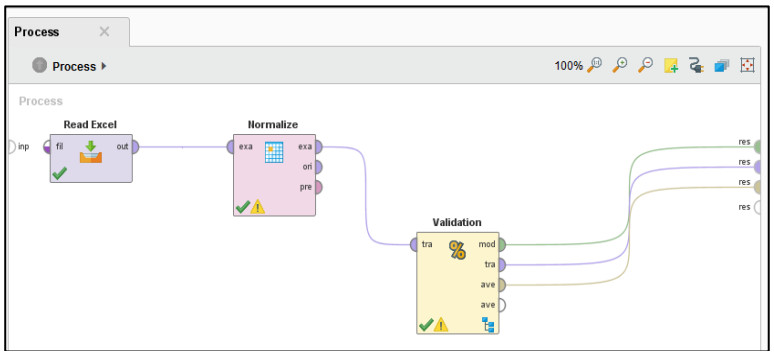

Figure 8. *Neural Network* Model Process

In Figure 8, validation after reading the data. The read excel operator is connected to the split validation operator. In the split validation operator there are training and testing sub-processes.

Drag and drop *Neural Network* operators in the training block, this block is used as a data training process with the Neural Network algorithm, while in the testing block, drag and drop to apply model operators and performance operators, this block is used to test data and produce accuracy values. , precision, recall, AUC, graph or classification pattern.

Based on the implementation of the dataset that has been carried out, data mining is processed by arranging operators and performing the results on the Neural Network algorithm to produce predictions for the dataset that has been inputted.

The results of the classification on the rapidminer application can be seen as follows:

1. Confusion Matrix

e-ISSN 2527-1682
p-ISSN 2527-9165

The Confusion Matrix model will form a matrix consisting of true positives, true negatives, false positives and false negatives.

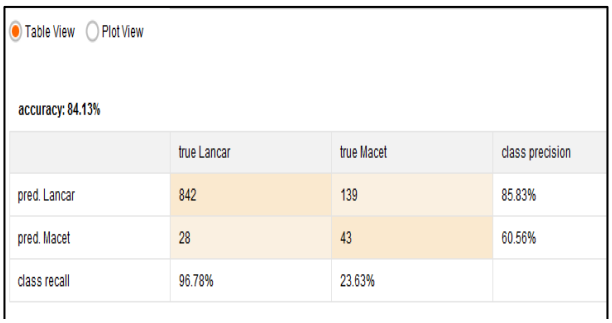
Table View Plot View

accuracy: 84.13%

	true Lancar	true Macet	class precision
pred. Lancar	842	139	85.83%
pred. Macet	28	43	60.56%
class recall	96.78%	23.63%	

Figure 9. Results of Confusion Matrix of Neural *Network* Algorithm

it can be seen that the accuracy value generated from the *Neural Network* algorithm process is an accuracy value of 84.13%, class recall is 85.83%, while for class precision the smooth prediction results are 96.78%. From 1,314 data, 842 true positives were predicted, 139 false negatives, 28 false positives and 43 true negatives.

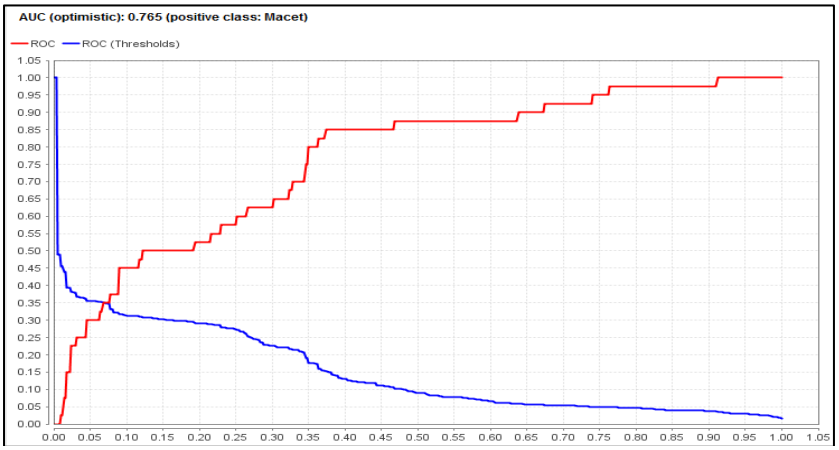

Figure 10. The AUC value on the ROC graph of the *Neural Network* Algorithm

In Figure 10 it can be seen that the AUC value is 0.765 with a fair classification accuracy, because it is in the range 0.70 – 0.80.

3.2.3 Comparison of N*aive Bayes* Algorithm and *Neural Network*

After testing using two algorithms, namely *Naive Bayes* and *Neural Network*, it can be seen that the Confusion Matrix comparison table is as follows:

Tabel 1. Comparison *Confusion Matrix* of *Naive Bayes* Algorithm and *Neural Network*

No	Algorithm	Accuracy	Recall	Precision
1.	*Naive Bayes*	72.62 %	91.48 %	73.85%
2.	*Neural Network*	84.13 %	85.83 %	96.78%

Table 1 shows the results of the Confusion Matrix from the *Naive Bayes* Algorithm and *Neural Network*, it can be seen from the comparison of the Confusion Matrix generated from the two algorithms tested.

Based on table 1, the *Neural Network* algorithm has a better Confusion Matrix value when compared to the *Naive Bayes* algorithm.

4. CONCLUSION

The data mining process that has been carried out in this study using KDD (Knowledge Discovery In Database) *Naive Bayes* and *Neural Network* methods and rapidminer tools in the Classification of the fluency multipurpose of bank mandiri credit payments based on debtor can be said as follows:

1. *Naive Bayes* can classify using probability and statistical methods, namely predicting opportunities in the past. Based on the classification of credit datasets in 2017, 2018, 2019 and 2020 with a total of 11 attributes, the accuracy value is 72.62%, class recall is 91.48%, while for class precision the current prediction result is 73.85%. Of the 1,314 data, 161 were predicted smoothly and 30 data were predicted.
2. The Neural *Network* algorithm in measuring credit payment measurements, the prediction process is more accurate in minimizing errors. The accuracy value is 84.13%, class recall is 85.83%, while for class precision, the current prediction results are 96.78%. From 1,314 data, 842 data were predicted smoothly and 43 data were predicted.
3. The comparison results show that the *Neural Network* algorithm has better results with an accuracy value of 84.13%, while the *Naive Bayes* algorithm only produces an accuracy value of 72.62%.

There is some improvement which can be implemented for further research. It is necessary to add several yahoos such as Random Forest, Linear Regression, Support Vector Machine and so on so that you can find out the comparison of the accuracy value of each yahoo to the data being tested. Adding more data and attributes so as to increase the value of better accuracy.

ACKNOWLEDGEMENTS

This research was partially supported by Dr. Edi Surya Negara, M. Kom as Advisor in writing this paper. and the second team of MMU Mikro Plaju 1 as partners and research sites. Last but not least, all my family who have been the motivation in completing this paper. The author realizes that in the preparation of this paper there are still many shortcomings. Therefore, the author accepts all constructive input and suggestions from various parties. Hopefully all the good deeds that have been given to the author will be rewarded according to Allah SWT.

5. REFERENCES

Achmad Rifai. (2016). *KAJIAN ALGORITMA C4 . 5 , NAIVE BAYES , NEURAL NETWORK DAN SVM DALAM PENENTUAN*. *2*, 176–182.

Amrin. (2015). Analisa Kelayakan Pemberian Kredit Mobil Dengan Menggunakan Neural Network Backpropagation. *Jurnal Techno Nusa Mandiri*, *XII*, 50.

Fayyad, . Usama. (1996). Advances in Knowledge Discovery and Data Mining. *MIT Press*.

Han, J,. & Kamber M, . (2006). Data Mining Concepts, Models and Techniques. *New Age International Limited*, *India*.

Hasan, M. (2017). *MENGGUNAKAN ALGORITMA NAIVE BAYES BERBASIS*. *9*, 317–324.

Informaatikalogi. (2017). *Algoritma Nive Bayes*. https://informatikalogi.com/algoritma-naive-bayes//

Ismail. (2010). *Akuntansi Bank*. Kencana.

Kasmir. (2015). *Manajemen Perbankan* (13th ed.).

Lailiyah A. (2014). *Urgensi Analis 5c Pada Pemberian Kredit Perbankan Untuk Meminimalisir Resiko*. *29(2)*, 217–232.

Liantoni, F., & Nugroho, . (2015). Klasifikasi Daun Herbal Menggunakan Metode Naive Bayes Classifier Dan Knearest Neigbor. *Jurnal Simantec*, *4*.

Mandala, E. P. W., & Putri, D. E. (2018). Prediksi Jumlah Pemberian Kredit Kepada Nasabah Di Bank Perkreditan Rakyat Dengan Algoritma C 4.5. *Komtekinfo*, *5*(1), 70–80.

Marie, F., & Supianto, A. (2018). *Clustering Credit Card Holder Berdasarkan Pembayaran Tagihan Menggunakan Improved K-Means dengan Particle Swarm Optimization TAGIHAN MENGGUNAKAN IMPROVED K-MEANS DENGAN PARTICLE SWARM*. *November*. https://doi.org/10.25126/jtiik.20185858

Murdianingsih, Y. (2015). *KLASIFIKASI NASABAH BAIK DAN BERMASALAH*. *2015*(November), 349–356.

Pratama, A. Z., Kurniawati, L., Larbona, S., & Haryanti, T. (2019). *Algoritma C4 . 5 Untuk Klasifikasi Nasabah Dalam*. *3*(2), 121–130.

Rani, L. N. (2016). *Klasifikasi Nasabah Menggunakan Algoritma C4 . 5 Sebagai Dasar Pemberian Kredit*.

Santoso, H., Hariyadi, I.P., & P. (2016). Data Mining oritma Apriori Pola Pembelan Produk Dengan Menggunakan Metode Algoritma Apriori. *Seminar Internasional Teknologi Informasi Dan Multimedia*, *2*.

Sutabri, T., Suryatno, A., Setiadi, D., & Negara, E. S. (2018). Improving naïve bayes in sentiment analysis for hotel industry in Indonesia. *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*, *March*, 1–6. https://doi.org/10.1109/IAC.2018.8780444

Sintia, S., Khautsar, A., Puspitasari, D., & Mustika, P. (2018). *Algoritma Naïve Bayes Untuk Memprediksi Kredit Macet Pada Koperasi Simpan Pinjam*. *4*(2).

D. Wanto, Anjar, *Data Mining : Algoritma dan Implementasi - Books*. Yayasan Kita Menulis, 2020.

Zola, F., Nurcahyo, G. W., & Jaringan, T. K. (2018). *Jaringan syaraf tiruan menggunakan algoritma backpropagation untuk memprediksi prestasi siswa*. *1*(1), 58–72.

M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "*A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users*' reviews," Future Gener. Comput. Syst., vol. 101, pp. 341–371, Dec. 2019, doi: 10.1016/j.future.2019.06.022

C. Manning, P. Raghavan, and H. Schuetze, "*Introduction to Information Retrieval*," p. 581, Apr. 2009

D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed. Canada: Wiley-Interscience, 2014.

Wahbeh, dkk (2014) "A *Comparison Study Between Data Mining Tools Over Some Classification Methods*"Jordan : Departement Of Computer Information System Yarmouk University.

Jiawei Han And Micheline Kamber. (2006). " *Data Mining :Concepts and Techiques* ". San Fransisco : Morgan Kaufmann Publishers

M. Allahyari, E. D. Trippe, and J. B. Gutierrez, "A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques," 2017

N. P. and M. P. J. Iran, "*Clustering Techniques and the Similarity used in Clustering: A survey*," Int. J. Comput. Appl., vol. 134, no. 7, 2016.

J. O. Yang, Y., Pedersen, "*Comparative Study on Feature Selection in Text Categorization*," Proc. Fourteenth Int. Conf. Mach. Learn., 1997.