Application of Data Mining for Classification of Customer Eligibility at XYZ Bank in Credit Agreements Using Naives Bayes and Random Forest Methods

M. Arief Rahman[1], Ade Sukma Wati[2], Aurantia Marina[3], Nurul Ilma Hasana Kunio[4], and Nur Jumrituniisah[5]

[1-5] State Polytechnic of Sriwijaya, Palembang, Indonesia
m.arief.rahman@polsri.ac.id

Abstract. This research addresses the challenges faced by Bank XYZ in assessing loan eligibility through manual analysis of customer data. Utilizing Data Mining techniques, specifically Naive Bayes and Random Forest algorithms, the study aims to enhance the accuracy of classifying customer patterns. The implemented models were evaluated using the CRISP-DM methodology, revealing Naive Bayes with an accuracy of 78.10% and Random Forest with 57.14%. The comparison suggests that Naive Bayes outperforms Random Forest in accuracy. The findings emphasize the potential of Naive Bayes for future implementations in customer classification at XYZ Bank, providing a promising solution to streamline the loan evaluation process and minimize the risk associated with late payments or bad loans.

Keywords: Data Mining, Naive Bayes, Random Forest, Customer Satisfaction

1 Introduction

1.1 Background

In today's technological development, the amount of data is both a problem and an opportunity for an agency [1]. Data becomes a problem if it cannot be stored, managed, or processed properly [2]. Data that always appears every time will continue to accumulate and if it is not properly documented, the data will become useless for the company. Meanwhile, data becomes an opportunity if it can be stored, managed and processed to be more meaningful to the agency [3]. With data, a trend or structure can be found which can later be used to obtain information in the future [4].

Bank XYZ is one of the companies that feel the development of technology. Palembang Bank is a business entity that collects funds from the public in the form of deposits and distributes them to the public in the form of credit or other forms in order to improve the lives of many people. One of the main tasks of a financial body including XYZ Bank is to develop some set of models and techniques to enable them or the financial body to determine the feasibility of a loan.

N. L. Husni et al. (eds.), *Proceedings of the 7th FIRST 2023 International Conference on Global Innovations (FIRST-ESCSI 2023)*, Advances in Engineering Research 232,
https://doi.org/10.2991/978-94-6463-386-3_57

The problem that often occurs at this time is that the behavior patterns or characteristics of customers are not good. In its implementation [5], the credit contract analysis at XYZ bank makes decisions on prospective customers by analyzing them. To analyze the eligibility of prospective customers receiving loans, an analyst must first request data from BI via the web and analyze it manually. So that a loan analyst still has to determine the customer's eligibility through a manual process. To determine the eligibility of prospective loan recipient customers will take a long time or cannot directly determine which customers are eligible to receive loans. Analysis of bank loan data is needed with the aim of minimizing the risk of customers who are late in paying loans or customers who cause bad loans.

Based on the resulting loan parameters, an assessment can be made of the status of the credit agreement on XYZ Bank data, namely prospective customers who are eligible or not to receive loans. To determine future loan eligibility, accurate forecasting is needed, one of which uses technology in the field of Data Mining. Many studies discuss the determination of loan recipient eligibility with Data Mining algorithms. As research conducted by [6] states that, the risk for financial institutions to provide the requested loan depends on how well they distinguish good loan applicants from bad loan applicants. The usual effort to reduce the number of bad loans is to carefully analyze loans or by improving the quality of employees to handle a prudent attitude in providing escort when paying credit contracts.

2 RESEARCH METHODS

In this research, the authors conducted research using the CRISP-DM (Cross-Standard Industry for Data Mining) research method. Crisp-Dm stages consist of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment [7].

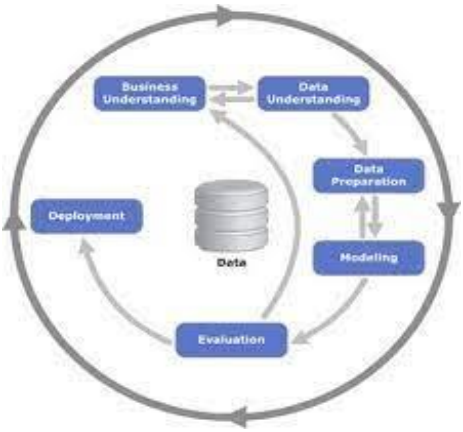


Figure.1 CRISPM-DM

3 RESULTS AND DISCUSSION

3.1 Organization Context

This research was conducted to determine the classification of customer patterns. The use of parameters affects the accuracy and model results that will be generated by the naives bayes algorithm. The appearance of the data mining model structure used consists of several objects or operators including read XLSX file, selected attribute discretization and Cross Validation which appear in the main process to run the data mining process. To determine the accuracy of the algorithm, this research uses Cross Validation. The design of the naives bayes model contained in RapidMiner is shown in Figure 2.

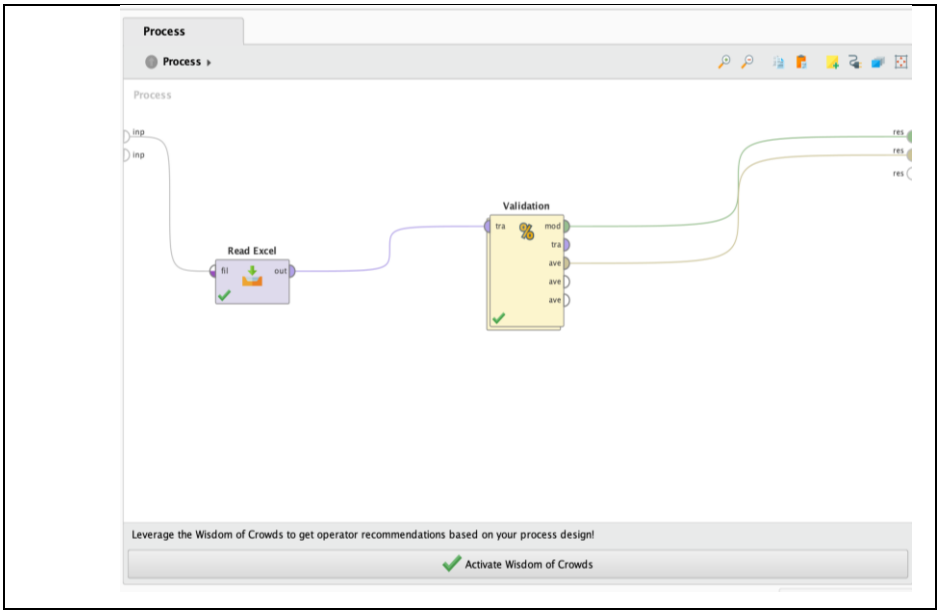


Figure 2. Cross Validation

In the Discretize object, the selected attribute is a subnet that has a continuous value only such as a smooth or stuck label. In the Cross Validation object or operator, there is a sub-process consisting of the naives bayes operator as an algorithm used in the Classification data mining process to be carried out, the Apply Model operator and the Performance operator as an operator to produce data mining processing in the form of naives bayes. The sub-process model can be seen in Figure 3

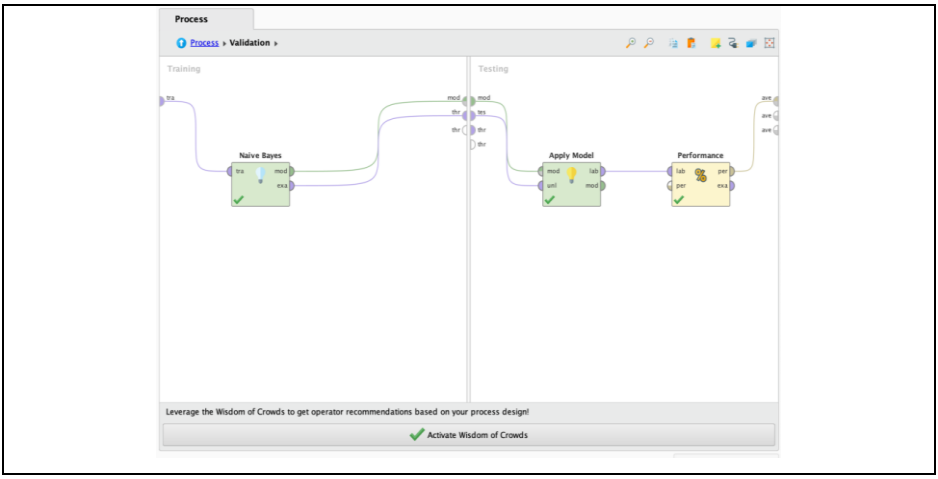


Figure 3. Rapidminer Model Sub Proses Clasification Naives Bayes

One of the objectives of this research is to determine the accuracy value of the naives bayes algorithm used to classify labels. In the training column there is a classification algorithm applied, namely naives bayes, Performance to measure the performance of the naives bayes model. This research was conducted to determine the classification of customer patterns. The use of parameters affects the accuracy and model results that will be generated by the random forest algorithm. The display of the data mining model structure used consists of several objects or operators including read XLSX file, selected attribute discretization and Cross Validation which appear in the main process to run the data mining process. To determine the accuracy of the algorithm, this research uses Cross Validation. The design of the random forest model contained in RapidMiner is shown in Figure 4.

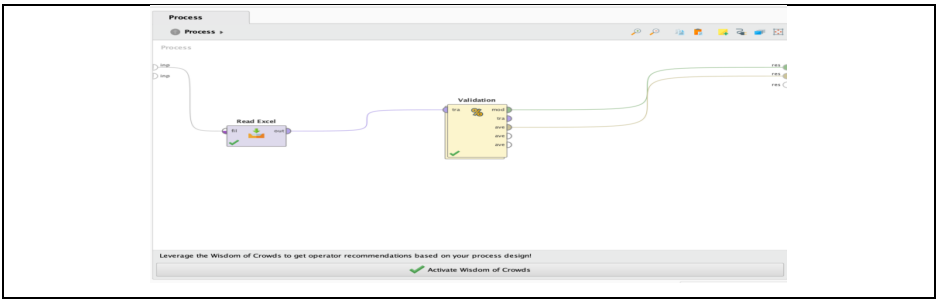


Figure 4. Cross Validation

In the Discretize object, the selected attribute is a subnet that has a continuous value only such as a smooth or stuck label. In the Cross Validation object or operator, there is a sub-process consisting of the naives bayes operator as an algorithm used in the Classification data mining process to be carried out, the Apply Model operator and the Performance operator as an operator to produce data mining processing in the form of random forest. The sub-process model can be seen in Figure 5.

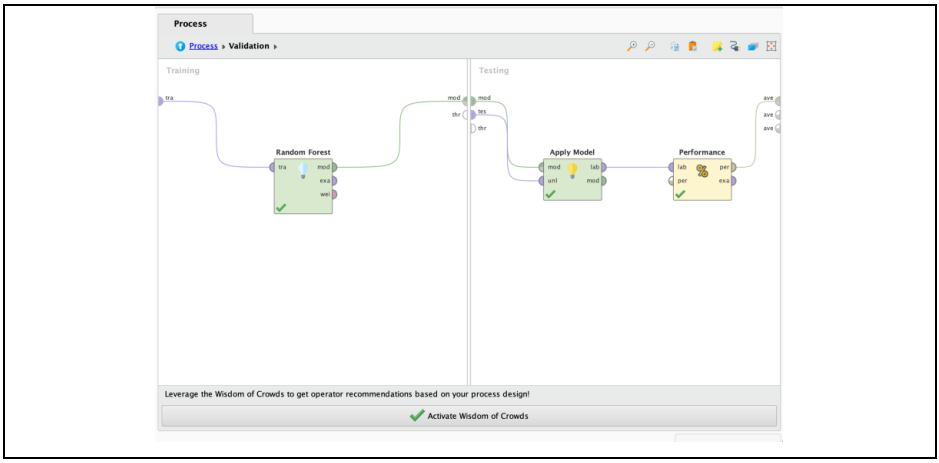
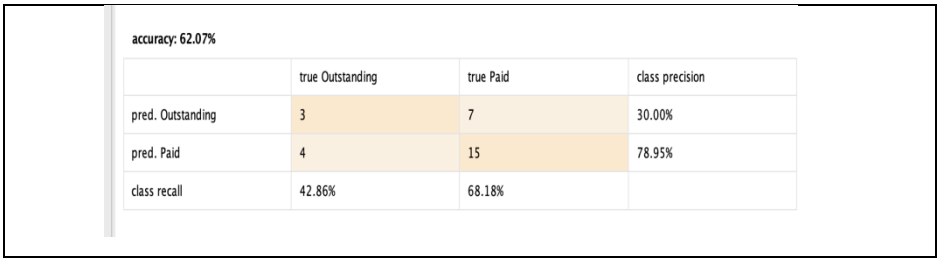


Figure 5. Rapidminer Model Sub Process Clasification Random Forest

One of the objectives of this research is to determine the accuracy value of the random forest algorithm used to classify labels. In the training column there is a classification algorithm applied, namely random forest, Performance to measure the performance of the random forest model. The implementation of the data processing process uses data discretization which is used to reduce the number of numeric attribute values by dividing the attribute range into intervals. Interval labels can then be used to replace actual data values. Modeling results that have been processed by the RapidMiner Tool in addition to producing in the form of modeling patterns can also determine the level of accuracy, recall, and precision. The results of the accuracy, recall, and precision values of the naives bayes algorithm can be seen as follows:



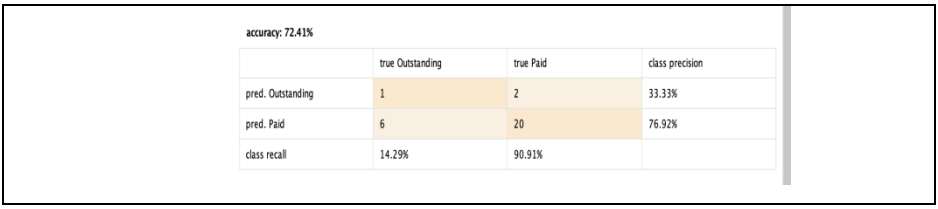
accuracy: 62.07%

	true Outstanding	true Paid	class precision
pred. Outstanding	3	7	30.00%
pred. Paid	4	15	78.95%
class recall	42.86%	68.18%	

Figure 6. Rapidminer Performance Naives Bayes

From the picture above, it can be seen that the existing accuracy value is 62.07%, class recall for outsanding is 42.86%, class recall paid is 68.18%. The implementation of the data processing process uses data discretization which is used to reduce the number of numeric attribute values by dividing the attribute range into intervals. Interval labels can then be used to replace actual data values. Modeling results that have been processed by the RapidMiner Tool in addition to producing in the form of modeling patterns can also determine the level of accuracy, recall, and precision. The

results of the accuracy, recall, and precision values of the random forest algorithm can be seen as follows:


accuracy: 72.41%

	true Outstanding	true Paid	class precision
pred. Outstanding	1	2	33.33%
pred. Paid	6	20	76.92%
class recall	14.29%	90.91%	

Figure 7. Rapidminer Performance Random Forest

From the picture above, it can be seen that the existing accuracy value is 72.41%, the class recall for outsanding is 14.29%, the class recall paid is 90.95%. After processing the analysis and obtaining the results of testing training data from the naives bayes algorithm and the random forest algorithm, the two algorithms have comparative results so that it can be concluded which algorithm is accurate. The results of the two algorithms can be seen as follows:

Table 1. Comparison of naives bayes algorithm and random forest algorithm

No.	Algoritma Naives Bayes	Algoritma Random Forest
1.	Has an accuracy value of 62.07%	Has an accuracy value 72.41%%
2.	Own class recall Paid 68.18%	Own value recall paid 90.95%
3.	Own class recall outsanding 42.8%	Own value class recall outsanding 14.29%

Source: Researcher, 2023

Based on the table above, it can be seen that the highest accuracy value based on rapidminer is the naives bayes algorithm of 78.10% while random forest is 57.14%. So that researchers suggest that future development implementations can use the naives bayes algorithm. This is because the accuracy value of naives bayes is quite high.

4 CONCLUSION

After analysing, designing, and testing, it can be concluded that the evaluation of the previous discussion and from the tests that have been carried out, the application

of data mining customer classification at XYZ Bank produces accuracy for naives bayes 78.10% and random forest 57.14% and the research objectives have been answered with this accuracy value. The attributes used in naives bayes and random forest are the same attributes label, instalment, income, gender and occupation.

References

[1] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, 2009, doi: 10.1109/TKDE.2008.239.

[2] J. He and H. Chen, "An LSM-Tree Index for Spatial Data," *Algorithms*, vol. 15, no. 4, 2022, doi: 10.3390/a15040113.

[3] A. Sukma Wati *et al.*, "Perbandingan Implementasi Algoritma CT-PRO dan Algoritma C45 Untuk Menentukan Pola Nasabah," *Seminar Nasional Informatika*, pp. 16–17, 2022.

[4] C. Anam and H. B. Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," *Energy : Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 8, no. 1, pp. 13–19, May 2018, Accessed: Dec. 13, 2023. [Online]. Available: https://ejournal.upm.ac.id/index.php/energy/article/view/111

[5] S. Anand and K. Mishra, "Identifying potential millennial customers for financial institutions using SVM," *Journal of Financial Services Marketing*, vol. 27, no. 4, 2022, doi: 10.1057/s41264-021-00128-7.

[6] A. Heiat, "Modeling Consumer Credit Scoring Through Bayes Network," 2011.

[7] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, 2021, doi: 10.1109/TKDE.2019.2962680.