Classification of Bank Deposit Using Naïve Bayes Classifier (NBC) and *K*–Nearest Neighbor (*K*-NN)

Muhammad Hafidh Effendy, Dian Anggraeni[*], Yuliani Setia Dewi, Alfian Futuhul Hadi

Departement of Mathematics, Faculty of Mathematics and Natural Science, University of Jember, Indonesia.
[*]*Corresponding author. Email:* dian_a.fmipa@unej.ac.id

ABSTRACT

Banks are financial institutions whose activities are to collect funds from the public in the form of deposits (saving deposit, demand deposit, and time deposit) and distribute them to the public in the form of credit or other forms. Deposits are an alternative for customers because the interest offered on deposits is higher than regular savings. Naïve Bayes Classification (NBC) is a statistical classification method based on Bayes' theorem that can be used to predict the probability of membership of a class. *K*-Nearest Neighbor (*K*-NN) is a method for classifying objects based on the learning data that is closest to the object. This study will use bank customer data consisting of 4521 records and 17 variables. The data is divided into 3 types of training-testing processes, namely 70%:30%, 75%:25%, and 80%:20% and the $\acute{K}$ -fold cross validation method is used with a value of $\acute{K}$ =10. The results of this study indicate that the *K*-NN method is better than the NBC method. Where the best classification of the *K*-NN method is in the training-testing process of 70%:30% which has an accuracy rate of 89.23%. While the best classification of the NBC method is in the training-testing process of 80%:20% which has an accuracy rate of 84.51%. the *K*-Nearest Neighbor and the Naïve Bayes Classifier method show the same results on the importance variables. Where out of 16 variables in classifying banking customers, of which 5 which have the most influence are the duration of time the bank contacted its customers, the results of the previous deposit offer, the last month contacted the customer, the type of communication used by the customer, and the number of contacts the bank had made prior to the promotion of opening a deposit..

Keywords: *Classification, Naive Bayes classifier, K-nearest neighbor, Importance variables.*

1. INTRODUCTION

One of the economic developments in the world can be seen through the emergence of financial institutions, especially in the banking sector. Banks are financial institutions whose main activities are collecting funds from the public (funding) and channeling these funds back to the community (lending) and providing other bank services [1]. The strategy of the bank as a channel for bank funds must first collect funds so that from the difference in interest the bank makes a profit. Generally, banks themselves benefit from customers that can be used as a source of funds in the form of checking accounts, savings and time deposits. Furthermore, the form of source of funds that became one of the bank's mainstay is deposits. Time deposits are deposits of other parties in banks whose withdrawals are only made based on an agreement. Time deposits can be an alternative for customers because the interest offered on deposits is higher than ordinary savings [2].

In managing customer data, of course, the amount of data used is very large, so a bank customer data classification system is needed that can classify between customers who have the opportunity to open a deposit or who do not have the opportunity to open a deposit. Several methods that can be used to classify bank customer data in statistics include the Naïve Bayes Classifier (NBC) and K-Nearest Neighbor (*K*-NN). The algorithm was chosen as the classification algorithm compared because both have a relatively high level of accuracy. This is evidenced by several previous studies, including determining the status of a volcano using k-fold cross validation, with results showing that the accuracy rate of the NBC method is better than the *K*-NN method with the NBC accuracy rate reaching 79.71%, and the *K*-NN accuracy reaching 63.68% [3]. Subsequent research was conducted on the prediction of creditworthiness, with results showing that the accuracy rate of the NBC method is better than the *K*-NN method with the *K*-NN accuracy value reaching 81.82%, and the NBC accuracy reaching 81.83% [4]. The last research on the Intrusion Detection System (IDS), with results showing that the accuracy of the *K*-NN method is better than the NBC method with the accuracy value of the *K*-NN method reaching 99.70%, while the NBC accuracy reaches 88.55% [5].

Based on this description, the authors are interested in conducting research to determine the highest level of accuracy and importance variables from the classification of banking data using the *K*-NN and the NBC method. It is hoped that this research can help the bank in identifying customers who will potentially open a time deposit so that it can be used to assist the performance and operations of the bank.

2. METHOD

2.1 Banking

Banking is everything related to banks, institutions, and business activities. Banking is the core of every country's financial system. Banks are places for companies, government and private agencies, as well as individuals to store their finances. The purpose of banking is to implement national development in order to increase equity, economic growth, and national stability towards increasing the welfare of the people at large [6].

Generally, banks themselves need additional funds from various parties, including from customers which can be used as a source of funds in the form of checking accounts, savings and time deposits. The form of source of funds that is one of the bank's mainstays is deposits. Deposit itself is a place for customers to make transactions in the form of securities. Time deposits can be an alternative for customer because time deposits have a period of time, but a consideration for customers to choose deposits is interest because the interest offered on deposits is higher than ordinary savings [1].

2.2 Classification

Classification is a job of assessing data objects to include them in a certain class from a number of available classes [7]. The system in the classification is expected to be able to classify all datasets correctly, but it cannot be denied that errors will occur in the classification process so it is necessary to measure the performance of the classification system. One method that can be used to measure the accuracy of the classification algorithm is the Confusion Matrix method. Confusion Matrix is a table of classification results. This method uses a matrix table such as Table 1. Where TP is the number of positive records classified as positive. FN is the number of positive records classified as negative. FP is the number of negative records classified as positive. TN is the number of negative records classified as negative.

Table 1. *Confusion Matrix*

f_{ij}		Prediction Class(*j*)	
		Positive	Negative
Actual Class (*i*)	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Based on Table 1, classification performance can be obtained, including Accuracy, Precision, and Sensitivity. Accuracy is the ratio of correct predictions (positive and negative) to the overall data. Precision is the ratio of positive correct predictions to the overall positive predicted results. Sensitivity is the ratio of true positive predictions compared to the overall data that are true positive. The following formulas for calculating Accuracy, Precision, Sensitivity, and Specificity are shown in equation (1), equation (2), equation (3) [8]

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FN}} * 100\% \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} * 100\% \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} * 100\% \quad (3)$$

2.3 Naive Bayes Classifier (NBC)

Naïve Bayes was first proposed by Revered Thomas Bayes. The use of Naïve Bayes has been introduced since 1702-1761 which is basically used to predict future opportunities based on previous experience. NBC is a statistical classification method based on the Bayes theorem that can be used to predict the probability of membership of a class [9]. NBC is one of the algorithms in data mining techniques in classification that is reliable in handling large datasets and can handle irrelevant data [10]. The symbol for *X* is the input vector containing the data and *Y* is the class label. The equation of Bayes' theorem (equation (4)) is as follows [8]:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (4)$$

Where: *X* is data with unknown class, *Y* is *X* with a specific class, $P(Y|X)$ is probability of hypothesis *Y* based on condition *X* (posterior probability), $P(Y)$ is prior probability, $P(X|Y)$ is the probability of *X* based on the condition *Y* (likelihood), $P(X)$ is the probability of *X*.

2.4 K–Nearst Neighbor (K-NN)

The K-Nearest Neighbor algorithm is an algorithm that performs classification based on the proximity of the location (distance) of one data to other data. Near or far a location (distance) is usually calculated based on the Euclidean distance with the following formula (equation (5)) [8].

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (5)$$

where: x_{1i} is the -i variable of object 1, x_{2i} is the -i variable of object 2, $dist(x_1, x_2)$ is the distance, *n* is number of objects for each variable, and *i* is the index from 1 to *n*.

3. MODEL DEVELOPMENT

The data used by the author in this study is public data obtained from the web https://archive.ics.uci.edu/ml/datasets/Bank+Marketing or site UCI Machine Learning Repository, entitled "Banking Marketing". The type of data used is primary data with a total of 4521 records and 17 variables. The data obtained consist of *age*, *job*, *material*, *education*, *default*, *balance*, *housing*, *loan*, *contact*, *day*, *month*, *duration*, *campaign*, *pday*, *previous and poutcome* which are independent varaible and *y* is dependent variable.

The data analysis technique used in this research is the *K*-NN method and the NBC method using R studio software. The steps taken in the research are as follows:

1. Processing data according to the *K*-NN and NBC methods.
   The data obtained is first entered through the microsoft excel program which is adjusted into a matrix form. The data according to the method is inputted into the R Studio program via the `readx1` package.
2. Dividing the data into two parts, training and testing. Splitting the data by dividing the training-testing process by 70%:30%, 75%:25%, and 80%:20% with the same proportions for each class. Data sharing using caret package with createDataPartition function. Then create a data frame from training data and testing data.
3. Carrying out the classification process using NBC on the training data test with the following stages:
   a. Conduct training tests on the R Studio program using the e1071 package
   b. Determine the *Ḱ*-fold cross validation method with the value *Ḱ*=10
   c. Get the results of the training data test
4. Carrying out the classification process using the *K*-NN method with the following stages:
   a. Conduct training tests on the R Studio program using the caret package
   b. Determine the *Ḱ*-fold cross validation method with the value *Ḱ*=10
   c. Getting the best *K* value from the *K* value of the nearest neighbor on *K*-NN
   d. Get the results of the training data test
5. Testing the results of the NBC and *K*-NN model using testing data set
6. Getting the results of the classification and the importance variables for the NBC and the *K*-NN model
7. Comparing the results of the classification of the NBC and the *K*-NN model

4. RESULT AND DISCUSSION

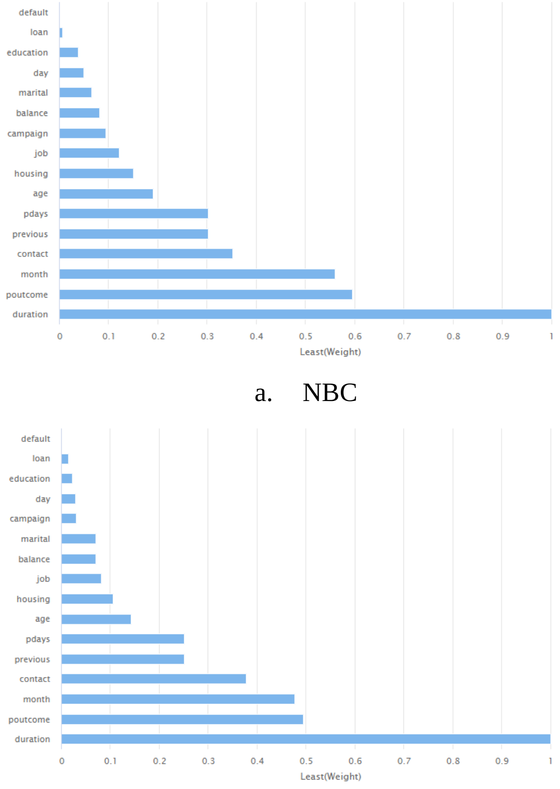
In this study, we use two methods the Naive Bayes Classifier (NBC) and the *K*-Nearest Neighbor (*K*-NN). Based on the prediction results in Table 2, it shows that the best classification prediction results for the NBC method are in the training-testing process of 80%:20%, with an accuracy value of 84.51% where this accuracy value is obtained by dividing the ratio of correct predictions (positive and negative). with all data or with $\frac{702+62}{702+95+45+62} * 100\%$.. The precision value is 93.98% where the precision value is obtained by dividing the ratio of true positive predictions by the overall positive predicted results or by $\frac{702}{702+45} * 100\%$. Sensitivity value is 93.98% where the sensitivity value is obtained by dividing the ratio of true positive predictions compared to the overall data that is true positive or by $\frac{702}{702+95} *$ 100%.

Table 2. NBC and *K*-NN Classification Performance

Results	70 %:30%		75 %:25%		80 %:20%	
	NBC (%)	*K*-NN (%)	NBC (%)	*K*-NN (%)	NBC (%)	*K*-NN (%)
Accuracy	83,78	**89,23**	83,36	89,20	**84,51**	87,94
Precision	94,24	89,95	94,54	90,02	93,98	88,91
Sensitivity	87,04	98,92	86,25	98,80	88,08	98,62

The best prediction results of the *K*-NN method are found in the training-testing process of 70%:30%, with an Accuracy value of 89.23% where this accuracy value is obtained by dividing the ratio of correct predictions (positive and negative) with the whole data or by $\frac{1191+19}{1191+13+133+19} * 100\%$. The Precision value is 89.95% where the precision value is obtained by dividing the ratio of true positive predictions by the overall positive predicted results or by $\frac{1191}{1191+133} * 100\%$. Sensitivity value is 98.92% where the sensitivity value is obtained by dividing the ratio of true positive predictions compared to the overall data that is true positive or by $\frac{1191}{1191+19} * 100\%$.

Figure 1 below shows that from bank customer data, there are several influential and less influential variables, this is indicated by the weight value of each variable. The higher the weight value, the higher the influence of these variables on the classification of banking customers. In addition, both methods show the same results on the importance variable, where from 16 variables there are 5 of the same importance variable, including the duration of time the bank contacted its customers (*duration*), the results of the previous deposit offer (*poutcome*), the last month contacted the customer (*month*), the type of communication used by the customer (*contact*), and the number of contacts the bank had made prior to the promotion of opening a deposit (*previous*).



a. NBC

b. K-NN

Figure 1 Importance variable Bank of customer data using NBC and *K*-NN methods.

5. CONCLUSION

Based on the results of the discussion, it can be concluded that the *K*-Nearest Neighbor method is better than the Naïve Bayes Classifier method in classifying bank customer deposit data. It is shown that the *K*-Nearest Neighbor method has the best accuracy rate of 89.23% in the training-testing process of 70%:30%. Meanwhile, the Naïve Bayes Classifier method has the best accuracy rate of 84.51% in the 80%:20% training-testing process. In addition, the second method also shows the same results on the influential variable (variable of interest), where the variables that have more influence are the duration of time the bank contacted its customers, the results of the previous deposit offer, the last month contacted the customer, the type of communication used by the customer, and the number of contacts the bank had made prior to the promotion of opening a deposit.

ACKNOWLEDGMENTS

This paper and the research behind it would not have been possible without the exceptional support funding from LP2M University of Jember.

REFERENCES

[1] N. Qomariah, Bank dan Lembaga Keuangan Lainnya, Penerbit Cahaya Ilmu, 2015.

[2] Kasmir, Bank dan Lembaga Keuangan Lainnya. Penerbit PT Raja Grafindo Persada, 2014

[3] S. Wahyuningsih, D.R. Utari. Perbandingan Metode K-Nearest Neigbor, Naive Bayes, dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit, Jurnal Konferensi Nasional Sistem Informasi, 2018.

[4] F.Tempola, M. Muhammad, A. Khairan. Perbandingan Klasifikasi Antara KNN dan *Naive Bayes* Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) 5, 2018, pp: 577-587.

[5] A. D. Afifaturahman, F. Maulana, Perbandingan Algoritma K-Nearest Neigbour (KNN) dan Naive Bayes pada Instrusion Detection System (IDS). Jurnal Innovation in Research of Informatics Vol. 3 No. 1, 2021, pp: 17-25.

[6] Undang-Undang Republik Indonesia Nomor 10 Tahun 1998. Perbankan. 10 November 1998. Lembaga Republik Indonesia Tahun 1998. Nomor 3790. Jakarta.

[7] E. Prasetyo, Data Mining Konsep dan Aplikasi Menggunakan MATLAB, ANDI Yogyakarta, 2014.

[8] J. Han, M. Kamber, Data Mining Concepts and Techniques Second Edition, Morgan Kaufman California, 2006.

[9] D. J. Hand, K. Yu. Idiot's Bayes: Not So Stupid after All? International Statistical Review, 69 (3), 2001, pp: 385- 398.

[10] M. Ridwan, H. Suyono, M. Sarosa, Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Jurnal EECCIS, 2013, pp: 59-64.