Naïve Bayes and K-Nearest Neighbor Algorithms Performance Comparison in Diabetes Mellitus Early Diagnosis

https://doi.org/10.3991/ijoe.v18i15.34143

Haviluddin(✉), Novianti Puspitasari, Aji Ery Burhandeny,
Andi Dhiya Awalia Nurulita, Dinnuhoni Trahutomo
Universitas Mulawarman, East Kalimantan, Indonesia
haviluddin@unmul.ac.id

Abstract—Diabetes Mellitus (DM) is a chronic disease that occurs when the body cannot effectively use the insulin it produces. The use of artificial intelligence (AI) can provide a means to diagnose. This study aims to obtain the best classification of the Naïve Bayes (NB) and K-Nearest Neighbors (KNN) methods so that accurate results are obtained in diagnosing DM disease using a dataset originating from The Abdul Moeis Hospital, Samarinda, East Kalimantan, Indonesia. The results showed that the KNN performed better in accuracy, precision, and specificity with an Area Under the Curve (AUC) value 10% higher than NB. Overall, KNN obtained a better recall compared to the NB in order to DM diagnosis.

Keywords—classification, Naïve Bayes, KNN, diabetes mellitus, confusion matrix

1 Introduction

Diabetes mellitus (DM) is a condition where the human body's pancreas cannot provide sufficient insulin which leads to an increase in sugar level, excessive thirst, appetite, and urine. Up to recently, diabetes has no permanent treatment, and yet in Indonesia Type-2 diabetes contributes to 24% of serious microvascular diseases namely retinopathy, nephropathy, and neuropathy among 5. Moreover, there is a trajectory of tuberculosis pandemic driven by diabetes within the upcoming decade [1].

Many studies have been conducted to analyze and predict an early diagnosis of serious illnesses as the crucial step to minimizing future complications and reducing treatment costs. The NB and KNN implemented on the Indian Liver Patient UCI dataset resulting dominancy of NB over KNN with the accuracy of 84% and 80.57% consecutively [2]. Similar prediction techniques are used for cryotherapy in wart treatment. In contrast to the previous findings, the KNN performance overwhelms NB with the accuracy of 90% and 86.67% consecutively [3]. In comparison among machine learning techniques, NB also performed well compared to Decision Tree, a support vector machine (SVM) using Pima Indian Diabetes Dataset (PIDD) to early detect diabetes in pregnancy [4]. Similar

performance results on comparing NB, KNN, SVM, and Random Forest (RF) using local and PIDD datasets. The resulting accuracy of KNN:90%, SVM, NB, and RF:98% on the local dataset. While the accuracy of KNN:81%, SVM:82%, NB:84%, and RF:82% on the PIDD dataset [5]. Researchers have applied the NB algorithm to classify the results of interviews into three categories, namely students' potential, talents, and interests. The results showed that with the validation of 50 respondents, the level of accuracy was obtained at 86.93%. This indicates that the NB and Reinforcement Phrase methods can be used to classify the results of interviews and make a positive contribution to students practicing interviews in English [6]. Then, researchers analyzed the sources of teaching English using the TF-IDF and KNN methods. Data classification based on 5,000 divided into training and testing (3,600:1,400) has changed the K value from 5 to 40. The results show that the enhanced kNN provides a feasible way to allocate English teaching resources. Research findings provide a reference for the storage and allocation of teaching resources [7]. Researchers have also utilized the enhanced KNN method with domain characteristics to analyze the teaching resources of the primary and secondary school classroom networks. The dataset is 3,000 (2,100:900), the feature dimension is 500, the K value is 15, and the WEKA software has been used. The results show that the KNN method can effectively ensure a uniform sample distribution and reduce the classification time [8].

This paper will examine the performance of NB and KNN methods measured by the Receiver Operating Characteristics Curve (ROC) and Area Under Curve (AUC) using local datasets from the Regional Public Hospital Abdul Moeis in Samarinda, East Kalimantan Province, Indonesia. The rest of this paper is organized as follows. Section 2 discusses NB and KNN techniques in classifying diabetes. Section 3 evaluates the results and accuracy and followed by some concluding remarks in Section 4.

2 Method

2.1 Naïve Bayes (NB) algorithm

The NB is a classification using probability and statistical methods based on Bayes' Theorem assuming strong independence. This method is proven accurate and provides high speed when applied to databases with extensive data. This method was put forward by British scientist Thomas Bayes, which predicts future opportunities based on previous experience. The NB method flowchart can be seen in Figure 1. This method works well compared to other classification models [9]–[12]. The NB algorithm uses Equation 1.

$$P(H|X) = \frac{P(X|C)P(C)}{P(X)} \tag{1}$$

Where, X is data with unknown classes; H is the hypothesis that data X is a specific class; P(H|X) is the hypothesis probability of H based on condition X (Posterior Probability); P(H) is the hypothesis probability of H (Prior Probability); P(X|H) is the probability of X based on the conditions in hypothesis H; P(X) is probability X. If the data used is continuous or numerical then the calculation uses Gauss Density in Equation 2.

$$P(Xi = xi|C = cj)$$
$$P = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(xi-\mu)^2}{2\sigma^2}} \quad (2)$$

Where, P is an opportunity; Xi is the *i-th* attribute; xi is the *i-th* attribute value; C is the class sought after; cj is the sought-after of C sub-class; μ is the mean, expressing the average of all attributes; σ standard deviation, expressing variants of all attributes; π is 3.141592654; and e is 2.718281828.

The likelihood value is obtained by multiplying the probability of attribute xi by the probability value of the category as in Equation 3.

$$P(C1) \times P(C2) \times P(C3) \ldots P(Cn) \times \text{Probability Value of Category} \quad (3)$$

The NB algorithm flow used in this study can be seen in Figure 1. Then, the stages of the NB algorithm are as follows:

Stage 1: Calculates the number and probability of each attribute of all classes

- Calculate the total record in each class.
- If the data used is numeric (positive/negative), look for each attribute's mean and standard deviation values based on the numerical data class. The mean value and standard deviation can be seen in Equations 4 and 5.

$$\mu = \frac{\sum_{i=1}^{n} xi}{n} \quad (4)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (xi - \mu)^2}{n-1}} \quad (5)$$

- If the data is not numeric (positive/negative), then find the probability value by calculating the corresponding amount of data from the same category divided by the amount of data in that category.

Stage 2: Calculates the posterior probability of each class category

- Calculate the posterior probability in each class category concerning the mean value and standard deviation obtained using Gauss Density in Equation (2) if the data used numerical (positive/negative).
- If the data is not numerical (positive/negative), then calculate the posterior probability of each class by dividing the amount of data of each class divided by the sum of the entire class.

Stage 3: Calculating Likelihood by class category

- Calculate the likelihood value using Equation (3), where the likelihood value will be used to find the probability value of the result.
- If the result is met, the class with the highest probability is the result of the prediction.

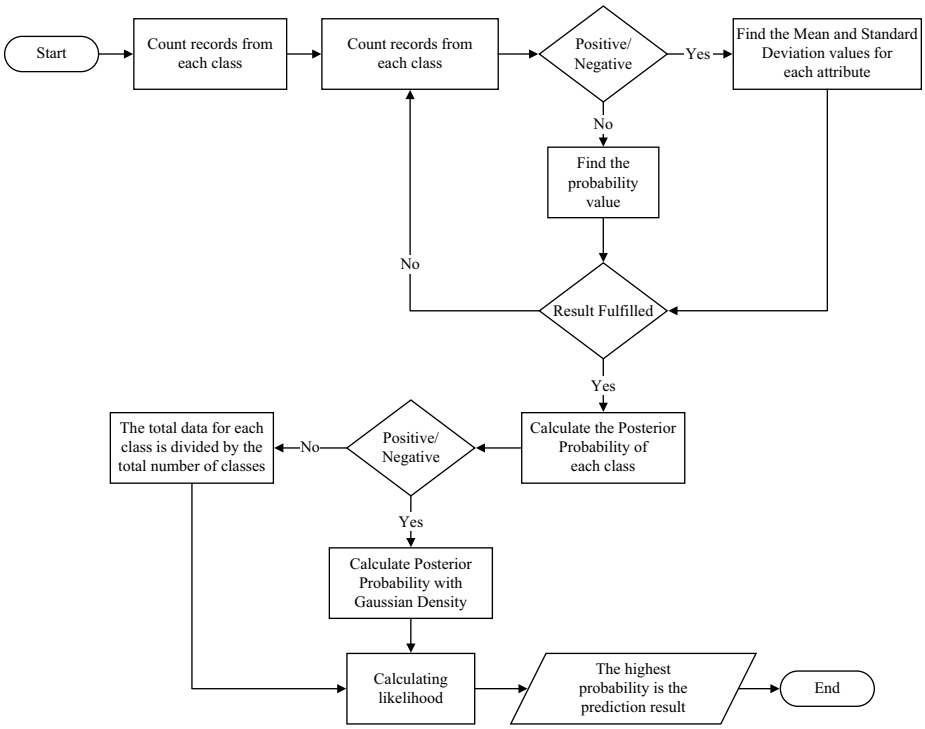

Fig. 1. NB method flowchart

2.2 K-Nearest Neighbor (KNN) algorithm

The KNN algorithm is a method for classifying new objects based on (K) their nearest neighbor [13]–[15]. KNN includes a supervised learning algorithm where training datasets are stored so that new unclassified records are obtained by comparing them with records that are most similar to training sets [16]. The most widely appeared class will be the class of classification results. The choice of value K is determined by the researcher. The selection of the value of K will affect the accuracy of the predictions made [17]–[19]. The KNN method flowchart can be seen in Figure 2.

KNN Algorithm steps are as follows:

Step 1: Determining Parameter K (number of nearby neighbors).

Step 2: Calculate the distance between training and testing data.

– Calculate distance using Manhattan Distance (MD), Equation 6.

$$d(x_1, x_2) = \sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right| \tag{6}$$

Step 3: Sort the results of the distance calculation.

– After calculating the following distance, find the smallest space by sorting the distance calculation results in ascending.
– Then collect the smallest distance that has been sorted by parameter K (nearest neighbor).

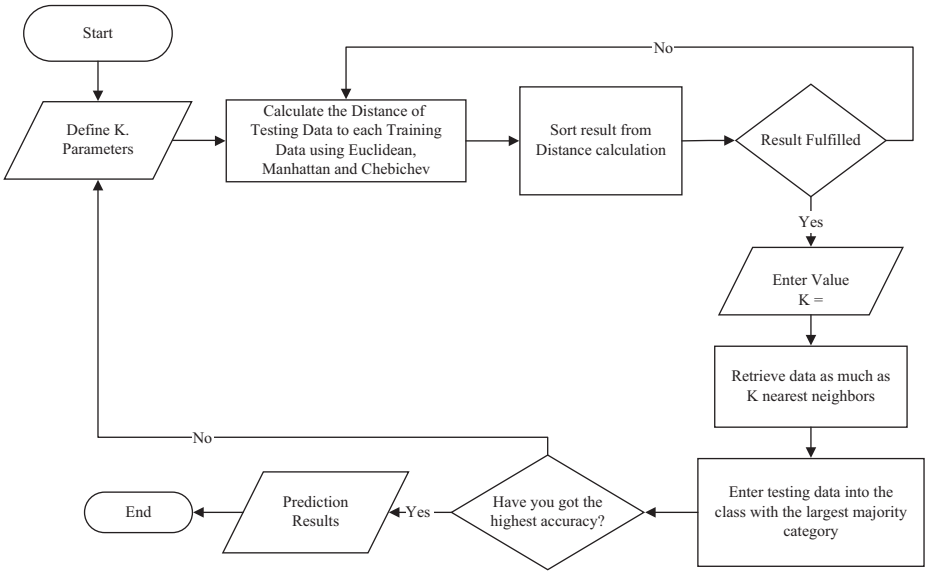

Fig. 2. KNN method flowchart

2.3 Accuracy

This study uses the Confusion Matrix (CM) method to compare the classification results. Since the CM method contains information on comparing classification labels with actual labels [20], [21]. This method can measure algorithm performance with the final accuracy result in percent units (%). The CM method can be viewed in Table 1.

Table 1. Confusion matrix (CM)

Classifications		**Predicated Class**	
		Class: (+)	**Class: (–)**
Observed Class	**(+)**	True Positive (TP)	True Negative (FN)
	(–)	False Positive (FP)	False Negative (TN)

The CM method is an evaluation of a classification of data mining represented in a table. This method is used to measure the performance of classification algorithms with an accuracy rate of a percent (%). Classification performance can be evaluated using Equation 7 to obtain accuracy, error rate, precision, recall, and specificity values [22]–[24].

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\
Error\ Rate &= \frac{FP+FN}{TP+FP+FN+TN} \\
Precision &= \frac{TP}{TP+FP} \\
Recall &= \frac{TP}{TP+FN}\times 100\% \\
Specificity &= \frac{TN}{(FP+TN)}\times 100\%
\end{aligned}
\tag{7}
$$

TP is a true positive value, namely how much data is actual positive class and model also predicts positive, *TN* is a true negative value, i.e., how much data is true negative class negative, and model also predicts negative, *FP* is a false positive value, i.e., how much data is actually negative class, but model predicts positive, *FN* is a false negative value, i.e., how much data the actual class is positive, but model predicts negative.

This study is a performance measurement tool in determining the threshold of a model, which has also been used. Where the ROC curve is based on the value obtained from the CM calculation between *FP* and *TP* rates. Meanwhile, comparing the performance values of the blue and green curves in the form of numbers by comparing the AUC has also been used [23], [25].

2.4 Dataset

In this study, the dataset came from The Regional Public Hospital Abdul Moeis Samarinda, East Kalimantan Province, Indonesia as many as 60 medical record data have been used. Then, the analysis parameters consist of 7 categories, namely age (K1), blood pressure (K2), blood sugar at the time (K3), urea (K4), and creatinine (K5), and leukocytes (K6) have also been applied.

3 Results and discussion

3.1 Naïve Bayes (NB) and K-Nearest Neighbor (KNN) results

In this experiment, we only present the best results from each test with various NB and KNN methods parameters. We have defined several parameters, including datasets ratio, 5-fold cross-validation, MD calculation, three samplings' parameters including linear, shuffled, and stratified, two categories classification including poor (0.600–0.700) and fair (0.700–0.800), and ROC and AUC then RapidMiner software has been explored.

For NB and KNN, modeling has used training, and testing data has been processed using the 5-fold schemes with datasets ratio of 80:20 (48:12) and 90:10 (54:6) implemented where the level of accuracy and other performance metrics are obtained through the evaluation stage of the NB algorithm model that produces CM, Table 2.

Table 2. NB and KNN algorithm classification results

Patient Code	NB					KNN				
	AC	CN	CP	PC	CR	AC	CN	CP	PC	CR
P1	–	0.351	0.649	+	I	–	0.201	0.799	+	I
P2	–	0.205	0.795	+	I	–	0.400	0.600	+	I
P3	–	0.355	0.645	+	I	+	0.000	1.000	+	A
P4	+	0.042	0.958	+	A	+	0.401	0.599	+	A
P5	–	0.096	0.904	+	I	+	0.196	0.804	+	A
P6	+	0.585	0.415	–	I	+	0.408	0.592	+	A
P7	–	0.186	0.814	+	I	–	0.604	0.396	–	A
P8	+	0.065	0.935	+	A	+	0.405	0.595	+	A
P9	+	0.646	0.354	–	I	–	0.583	0.417	–	A
P10	+	0.158	0.842	+	A	+	0.612	0.388	–	I
P11	–	0.241	0.759	+	I	–	0.578	0.422	–	A
P12	+	0.065	0.935	+	A	–	0.801	0.199	–	A
P13	–	0.672	0.328	–	A	–	0.586	0.414	–	A
P14	–	0.749	0.251	–	A	+	0.199	0.801	+	A
P15	–	0.548	0.452	–	A	–	0.577	0.423	–	A
P16	–	0.001	0.999	+	I	–	0.593	0.407	–	A
P17	–	0.864	0.136	–	A	–	0.597	0.403	–	A
P18	+	0.605	0.395	–	I	–	0.804	0.196	–	A
P19	+	0.993	0.007	–	I	+	0.599	0.401	–	I
P20	+	0.447	0.553	+	A	+	0.183	0.817	+	A
P21	+	0.003	0.997	+	A	+	0.193	0.807	+	A
P22	+	0.076	0.924	+	A	+	0.407	0.593	+	A
P23	–	1.000	0.000	–	A	–	1.000	0.000	–	A
P24	+	0.303	0.697	+	A	+	0.390	0.610	+	A
P25	+	0.005	0.995	+	A	–	0.605	0.395	–	A
P26	+	0.095	0.905	+	A	–	0.578	0.422	–	A
P27	–	0.079	0.921	+	I	–	0.599	0.401	–	A
P28	+	0.021	0.979	+	A	+	0.605	0.395	–	I
P29	+	0.000	1.000	+	A	+	0.392	0.608	+	A
P30	–	0.052	0.948	+	I	+	0.611	0.389	–	I
P31	+	0.268	0.732	+	A	–	0.395	0.605	+	I
P32	–	0.869	0.131	–	A	+	0.190	0.810	+	A
P33	–	0.051	0.949	+	I	–	1.000	0.000	–	A
P34	–	0.956	0.044	–	A	–	0.000	1.000	+	I
P35	–	0.000	1.000	+	I	+	0.419	0.581	+	A
P36	+	0.075	0.925	+	A	–	0.606	0.394	–	A
P37	+	0.331	0.669	+	A	–	0.800	0.200	–	A

(Continued)

Table 2. NB and KNN algorithm classification results *(Continued)*

Patient Code	NB					KNN				
	AC	**CN**	**CP**	**PC**	**CR**	**AC**	**CN**	**CP**	**PC**	**CR**
P38	+	0.592	0.408	–	I	+	0.598	0.402	–	I
P39	–	0.583	0.417	–	A	–	0.805	0.195	–	A
P40	–	1.000	0.000	–	A	+	0.380	0.620	+	A
P41	–	0.552	0.448	–	A	–	0.800	0.200	–	A
P42	+	0.366	0.634	+	A	+	0.402	0.598	+	A
P43	–	0.031	0.969	+	I	–	0.568	0.432	–	A
P44	+	0.250	0.750	+	A	–	1.000	0.000	–	A
P45	+	0.176	0.824	+	A	+	1.000	0.000	–	I
P46	+	0.137	0.863	+	A	+	0.399	0.601	+	A
P47	–	0.020	0.980	+	I	+	0.572	0.428	–	I
P48	–	0.538	0.462	–	A	+	0.598	0.402	–	I
P49	+	0.499	0.501	+	A	–	0.609	0.391	–	A
P50	–	0.276	0.724	+	I	–	0.585	0.415	–	A
P51	–	0.424	0.576	+	I	+	0.610	0.390	–	I
P52	+	0.275	0.725	+	A	+	0.607	0.393	–	I
P53	+	0.497	0.503	+	A	–	0.592	0.408	–	A
P54	+	0.161	0.839	+	A	+	0.372	0.628	+	A
P55	–	0.206	0.794	+	I	+	0.189	0.811	+	A
P56	–	0.392	0.608	+	I	–	0.202	0.798	+	I
P57	+	0.113	0.887	+	A	+	0.422	0.578	+	A
P58	–	0.045	0.955	+	I	+	0.194	0.806	+	A
P59	+	0.052	0.948	+	A	–	0.197	0.803	+	I
P60	–	0.259	0.741	+	I	–	0.605	0.395	–	A
Number of Accurate					**36**	**Number of Accurate**				**49**
Number of Inaccurate					**34**	**Number of Inaccurate**				**11**

Notes: Information: actual class (AC), confidence negative (CN), confidence positive (CP), prediction class (PC), classification result (CR), inaccurate (I), accurate (A).

The NB results showed that there were 25 patients who correctly declared DM positive in accordance with the prediction results, and were 19 DM negative but categorized as positive, then 5 DM positive but categorized as DM negative have been stated. The results showed that 22 patients correctly declared DM positive in accordance with the prediction results and were 21 DM negative and there were 9 DM negative but also categorized as positive, and as many as eight patients who were DM positive but were categorized as DM negative were confirmed. The following NB and KNN algorithms classification results can be seen in Table 3.

Table 3. CM of NB and KNN algorithm results

Prediction Class DM Datasets	Actual Class				
	NB			KNN	
		+	–	+	–
	+	25	19	22	9
	–	5	11	8	21

3.2 Accuracy performance of NB and KNN

In this experiment, the best accuracy of NB dan KNN methods with various experimental schemes is presented, so the results show in Tables 4–6.

Table 4. AUC evaluation results of NB

K-Fold	1	3	5	7	9
3	0.500	0.590	0.572	0.757	0.730
5	0.500	0.600	0.572	0.699	0.717
7	0.500	0.533	0.669	0.615	0.743
9	0.500	0.647	0.742	**0.778***	0.614

Table 5. AUC evaluation results of KNN

K-Fold	1	3	5	7	9
3	0.500	0.702	**0.767***	0.757	0.730
5	0.500	0.600	0.700	0.699	0.717
7	0.500	0.533	0.669	0.754	0.743
9	0.500	0.647	0.742	0.778	0.731

Table 6. NB and KNN AUC evaluation results

Algorithms	K-Fold	Parameter	AUC	Category Classification
NB	5	Stratified sampling	0.611	poor
KNN			**0.767**	fair

Table 7. NB and KNN algorithm accuracy results

Methods	K-Fold / Parameter	Accuracy	Precision	Recall	Specificity	Error Rate
NB	5 / Stratified sampling	**60.00%**	57%	83%	37%	40%
KNN		**73.54%**	77%	67%	80%	27%

Table 7 shows the accuracy result of NB algorithm testing using 5-folds and three sampling techniques that produce different accuracy and performance metrics. The highest accuracy data validation was produced at 60% with a standard deviation ± of 12.17%, precision at 57%, recall at 83%, specificity at 30%, and the error rate was

still relatively high at 40% using the stratified parameter. This means that the model was still not considered good in classifying. While the KNN results using the values K were 1, 3, 5, 7, and 9 have been implemented. The accuracy results of KNN algorithm testing using 5-folds and three sampling techniques that produce different accuracy and performance metrics. Then, the highest accuracy produced was 73.54% with a standard deviation ± of 16.98%, precision 77%, recall 67%, specificity 80%, and a reasonably low error rate of 27%. The accuracy level of KNN was influenced by the number of K values. In other words, the more K values, the lower accuracy. This was because the attributes used have a lot of similarities, so the more neighbors taken, the more data from other classes were taken into consideration for decisions, Tables 4 and 5.

It can be concluded that KNN gets superior accuracy results of 73.54% using K = 5 and precision of 77%, recall of 67%, specificity of 80%, and a fairly small error rate of 27% compared to NB, which gets an accuracy of 60%, precision of 56.82%, recall of 83%, specificity of 37%, and a fairly high error rate of 40% with validation of different data and parameters. The same but at the same warranty, NB still produces low accuracy. The AUC value of KNN was categorized as appropriate classification with a value of 0.767, while NB was categorized as poor classification with an AUC value of 0.611. The recall value from NB was still superior to KNN, but the accuracy results were inferior in predicting DM-positive patients. Then, the KNN evaluation value of 0.767 (fair) was achieved, Table 6.

The following visualizations of the ROC graph of the NB and KNN algorithms can be seen in Figures 3 and 4. From the ROC curve (red line), the y-axis is FP rate (1 – specificity), and the x-axis is TP rate (sensitivity). Based on the analysis of the tests that have been carried out, the level of accuracy, precision, recall, specificity, error rate, parameter, and AUC value from the classification of the two algorithms. It can be concluded that the model that has been made provides a poor performance with an AUC value of 0.611 or under the curve but quite far from 1, Figure 3. Furthermore, the model that has been made provides a fairly good performance with an AUC value of 0.767 or under the curve but close to 1, Figure 4.

Based on 60 patients, 54 DM positive and 6 DM negative according to the complication types. The classification results showed 19 DM positives with type hyperglycemia. Then, 15 DM positives with type hypertension. Then, 19 DM positives with hyperglycemia, nephropathy, and hypertension. It can be concluded that 83% of the KNN classification results have been correct. Meanwhile, the NB classification results were from 19 DM positive with hyperglycemia, nephropathy, and hypertension. It can be concluded that 81% of the NB classification results have also been precise.

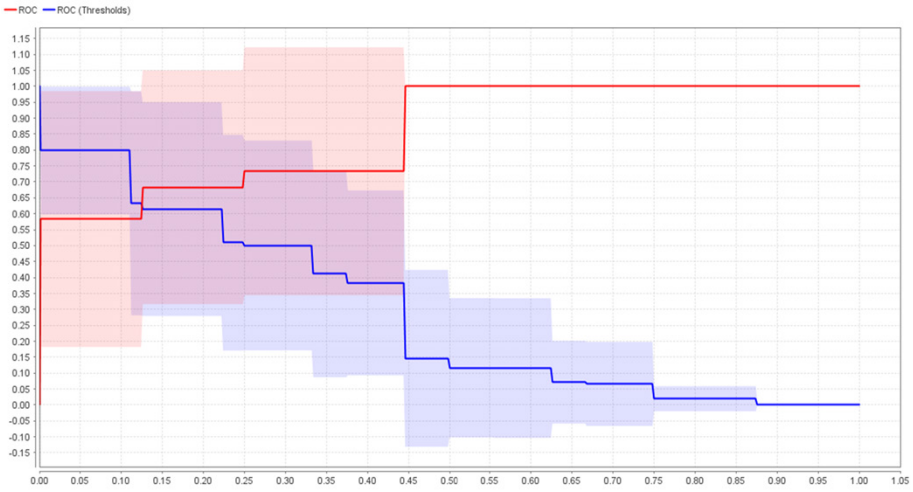

Fig. 3. ROC graph of NB algorithm

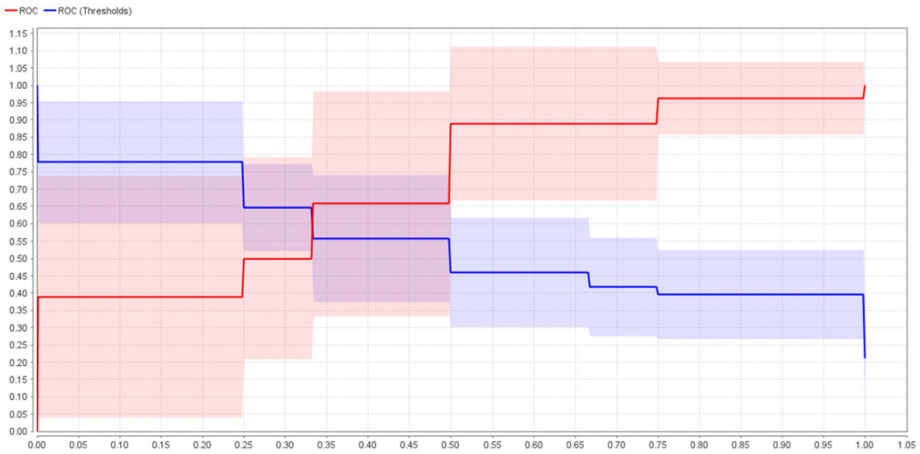

Fig. 4. ROC graph of KNN algorithm

4 Conclusion

This paper presented the accuracy of NB and KNN methods that can provide DM classification. The implementation of the NB and KNN methods has used parameters consisting of 5-fold, three distance measurement methods (i.e., ED, MD, and MinD), two classification categories (i.e., fair, and poor), CM method (i.e., accuracy, error rate, precision, recall, and specificity) as a classification measure, three sampling parameters (i.e., linear, shuffled, and stratified), and ROC and AUC methods by using RapidMiner software. Based on the experiment, the NB method has an accuracy level of 60%,

precision at 57%, recall at 83%, specificity at 30%, and the error rate was still relatively high at 40%. Meanwhile, the KNN method has an accuracy level of 73.54%, a precision of 77%, recall of 67%, specificity of 80%, and a reasonably low error rate of 27%. Then, the AUC evaluation value is achieved for NB of 0.611 (poor) and KNN of 0.767 (fair). Therefore, early detection of diabetes is the most important stage because it is useful to know the status of diabetes so that it can be treated quickly. Next is treatment, and the last is prevention by reducing risk triggers.

This study shows that the DM dataset captured from The Regional Public Hospital Abdul Moeis Samarinda, East Kalimantan Province, Indonesia that produce the accuracy performance of the KNN is superior to the NB methods. Based on method analysis, accuracy values can be improved in various ways, such as training and testing data settings and tuning K parameter values on KNN algorithms rather than trial methods. Our next focus is a deeper analysis of different machine learning techniques.

5 References

[1] S. F. Awad, J. A. Critchley, and L. J. Abu-Raddad, "Impact of diabetes mellitus on tuberculosis epidemiology in Indonesia: A mathematical modeling analysis," *Tuberculosis*, vol. 134, p. 102164, May 2022, https://doi.org/10.1016/j.tube.2022.102164

[2] L. A. Auxilia, "Accuracy prediction using machine learning techniques for Indian patient liver disease," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 45–50, https://doi.org/10.1109/ICOEI.2018.8553682

[3] H. Hartatik, M. B. Tamam, and A. Setyanto, "Prediction for diagnosing liver disease in patients using KNN and Naïve Bayes algorithms," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020, pp. 1–5, https://doi.org/10.1109/ICORIS50180.2020.9320797

[4] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018, https://doi.org/10.1016/j.procs.2018.05.122

[5] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: A comparative study," *Appl. Sci.*, vol. 11, no. 5, 2021, https://doi.org/10.3390/app11052218

[6] M. Sarosa, M. Junus, M. U. Hoesny, Z. Sari, and M. Fatnuriyah, "Classification technique of interviewer-bot result using Naïve Bayes and phrase reinforcement algorithms," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 2, pp. 33–47, 2018, https://doi.org/10.3991/ijet.v13i02.7173

[7] Y. Lou, "Storage and allocation of English teaching resources based on k-nearest neighbor algorithm," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 17, pp. 102–113, 2019, https://doi.org/10.3991/ijet.v14i17.11188

[8] Y. An, M. Xu, and C. Shen, "Classification method of teaching resources based on improved KNN algorithm," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 4, pp. 73–88, 2019, https://doi.org/10.3991/ijet.v14i04.10131

[9] D. L. Olson and D. Delen, *Advanced data mining techniques [electronic resource]*. 2008.

[10] G. Sharma and U. Hengaju, "Performance analysis of data mining classification techniques to predict diabetes," *Int. J. Adv. Netw. Appl.*, vol. 12, no. 1, pp. 4509–4518, 2020, https://doi.org/10.1016/j.procs.2016.04.016

[11] T. Winarti, H. Indriyawati, V. Vydia, and F. W. Christanto, "Performance comparison between naive bayes and k-nearest neighbor algorithm for the classification of Indonesian language articles," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2. pp. 452–457, 2021, https://doi.org/10.11591/ijai.v10.i2.pp452-457

[12] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," *Encycl. Mach. Learn.*, vol. 15, pp. 713–714, 2010, https://doi.org/10.1007/978-0-387-30164-8_576
[13] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of diabetes mellitus using k nearest neighbor algorithm," *Int. J. Comput. Sci. Trends Technol.*, vol. 2, no. 4, pp. 36–43, 2014.
[14] M. Alehegn, R. R. Joshi, and P. Mulay, "Diabetes analysis and prediction using random forest, KNN, Naïve Bayes and J48: An ensemble approach," *Int. J. Sci. Technol. Res*, vol. 8, no. 9, pp. 1346–1354, 2019.
[15] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, "Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: A comparison of four data mining approaches," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–13, 2020, https://doi.org/10.1186/s12859-020-03719-8
[16] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metab. Disord.*, vol. 19, no. 1, pp. 391–403, 2020, doi: https://doi.org/10.1007/s40200-020-00520-5
[17] R. Saxena, "Role of K-nearest neighbour in detection of diabetes mellitus," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 373–376, 2021.
[18] S. Suyanto, S. Meliana, T. Wahyuningrum, and S. Khomsah, "A new nearest neighbor-based framework for diabetes detection," *Expert Syst. Appl.*, vol. 199, p. 116857, 2022, https://doi.org/10.1016/j.eswa.2022.116857
[19] E. H. Hardi *et al.*, "Synbiotic application to enhance growth, immune system, and disease resistance toward bacterial infection in catfish (Clarias gariepinus)," *Aquaculture*, vol. 549, 2022, https://doi.org/10.1016/j.aquaculture.2021.737794
[20] M. A. Mohamed, A. B. Nassif, and M. AlShabi, "Classification of diabetes mellitus disease using machine learning," *Smart Biomed. Physiol. Sens. Technol.*, vol. XIV, no. 12123, pp. 104–112, 2022, https://doi.org/10.1117/12.2632632
[21] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016. https://doi.org/10.1016/B978-0-12-804291-5.00010-6
[22] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3, pp. 429–450, 2017, https://doi.org/10.1007/s10472-017-9564-8
[23] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, 2020, https://doi.org/10.1016/j.ins.2019.06.064
[24] I. H. Sarker, F. Faruque, H. Alqahtani, and A. Kalim, "K-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 26, 2018, https://doi.org/10.4108/eai.13-7-2018.162737
[25] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties," *Commun. Stat. Methods*, vol. 49, no. 9, pp. 2080–2093, 2020, https://doi.org/10.1080/03610926.2019.1568485

6 Authors

Haviluddin is a senior lecturer at Department of Informatics, Faculty of Engineering, Universitas Mulawarman, Indonesia, since 2002. He completed his Ph.D. in Computer Science from Universiti Malaysia Sabah, Malaysia, in 2016. He is the coordinator of publication and intellectual property rights of the Research Institute and Community Service of Mulawarman University. His research interest is in the artificial intelligent area. Email id: haviluddin@unmul.ac.id

Novianti Puspitasari is a lecturer at Department of Informatics, Faculty of Engineering, Universitas Mulawarman, Indonesia, since 2015. Her research interest is in the artificial intelligent area. Email id: novia.ftik.unmul@gmail.com

Aji Ery Burhandeny is a senior lecturer at Department of Electrical Engineering, Faculty of Engineering, Universitas Mulawarman, Indonesia, since 2005. He completed his Ph.D. in Computer Science from Ehime University, Japan in 2018. His research interest is in the artificial intelligent area. Email id: a.burhandenny@ft.unmul.ac.id

Andi Dhiya Awalia Nurulita is a student at Department of Informatics, Faculty of Engineering, Universitas Mulawarman, Indonesia. Her research interest is in the artificial intelligent area. Email id: dhiya.andi12@gmail.com

Dinnuhoni Trahutomo is a student at Department of Informatics System, Faculty of Engineering, Universitas Mulawarman, Indonesia. His research interest is in the artificial intelligent area. Email id: dinnuhoni.trahutomo@gmail.com

Article submitted 2022-07-21. Resubmitted 2022-09-04. Final acceptance 2022-09-12. Final version published as submitted by the authors.