Comparison Of The Performance Of K-Nearest Neighbors And Naive Bayes Algorithms For Stroke Disease Prediction.docx

by Miah Dahl

Submission date: 19-Jun-2025 12:21AM (UTC-0700)
Submission ID: 2606761192
File name: Comparison_Of_The_Performance_Of_K-Nearest_Neighbors_And_Naive_Bayes_Algorithms_For_Stroke_Disease_Prediction.docx (3.39M)
Word count: 4548
Character count: 27264

Jurnal CoreIT, Vol.X, No.X, Month Year
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Comparison Of The Performance Of K-Nearest Neighbors And Naive Bayes Algorithms For Stroke Disease Prediction

Baskoro[1*], Roby Novianto[2], Bambang Triraharjo[3]
[1]Dept. of Computer Science, Universitas Muhammadiyah Pringsewu, Indonesia
[2]Dept. of Computer Science, Universitas Muhammadiyah Pringsewu, Indonesia
[3]Dept. of Computer Science, Universitas Muhammadiyah Pringsewu, Indonesia
baskoro@umpri.ac.id , robynovianto@umpri.ac.id , bambangtriraharjo@umpri.ac.id

Abstract

Purpose: Stroke is a critical global health issue requiring early and accurate prediction to mitigate severe outcomes. This study aims to compare the performance of the K-Nearest Neighbors (KNN) and Naive Bayes algorithms in predicting stroke disease, addressing the challenge of imbalanced datasets and improving prediction accuracy for better clinical decision-making.
Methods/Study design/approach: The research followed the CRISP-DM model, utilizing a dataset of 5,110 patient records with 12 attributes from Kaggle. Data preprocessing included handling missing values and normalization. The KNN and Naive Bayes algorithms were implemented using RapidMiner, with performance evaluated through cross-validation, confusion matrices, and ROC-AUC curves.
Result/Findings: The KNN algorithm achieved an accuracy of 94.50%, but exhibited low precision (7.89%) and recall (1.20%) for stroke-positive cases due to dataset imbalance. Naive Bayes yielded an accuracy of 88.83% with an AUC of 0.767, demonstrating better probability modeling but similar challenges in minority class detection. Both algorithms highlighted the impact of data imbalance on predictive performance.
Novelty/Originality/Value: This study provides a comparative analysis of KNN and Naive Bayes for stroke prediction, emphasizing the need for data balancing and optimization techniques. The findings underscore the potential of these algorithms in healthcare applications while suggesting future improvements through ensemble methods or alternative algorithms like Random Forest.

Keywords: CRISP-DM, Data Mining, K-Nearest Neighbors, Naive Bayes, Stroke Prediction

Received July 2021 / **Revised** October 2021 / **Accepted** May 2022

INTRODUCTION

Stroke is a brain disease caused by high blood pressure. Stroke can occur at any time when bleeding occurs due to high intracranial pressure and embolus discharge from non-cerebral blood vessels and if there is an increase it can lead to hypertension[1]. This disease is classified as cerebrovascular disease (CVD) because it occurs suddenly and requires very quick treatment. All countries are still struggling to cope with stroke problems with limited resources and stroke neurologists. To overcome this problem, there are several studies that have shown that classification systems can overcome health problems such as stroke by combining methods or using a single classification method.

The World Health Organization defines stroke as a clinical sign that occurs due to local or global impaired brain function, with symptoms lasting 24 hours or more, which can lead to death, in the absence of any cause other than vascular[2]. Stroke occurs due to a cut off of blood supply to the brain, can occur due to a blockage in the form of blood clots or spurts in blood vessels.

Based on the results of Riskesdas in 2013, the prevalence of stroke in Indonesia increases with age. The highest stroke cases diagnosed by health workers are 75 years old and above (43.1%) and the lowest in the age group of 15-24 years is 0.2%[3]. The prevalence of stroke by 2 sexes is more males (7.1%)2 compared to females (6.8%). Based on place of residence, the prevalence of stroke in urban areas is higher (8.2%) compared to rural areas (5.7%). Based on data on the top 10 most common diseases in Indonesia in 2013, the prevalence of stroke cases in Indonesia based on the diagnosis of health workers is 7.0 per mill and 12.1 per mill for those diagnosed with stroke symptoms[4]. The highest prevalence of stroke cases is found in North Sulawesi Province (10.8%) and the lowest in Papua Province (2.3%), while Central Java Province is 7.7%. The prevalence of stroke between men and women is almost the same[5].

It can be seen from the data above that for the management of large data in Indonesia or the world, certain processing methods or methods are needed so that the data can be presented and viewed in general[6]. The

DOI: 10.24014/coreit.vxxx.xxxx

data processing usually uses data mining. Data mining itself is a process of extracting useful information and patterns from very large amounts of data. Data mining includes data collection, data extraction, data analysis, and data statistics[7]. Data mining is also known as *knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting*, and others

To be able to get information from existing data, it is necessary to carry out a data mining process such as classification[8]. Classification is a process to determine a model that explains or distinguishes a concept or class of data, with the aim of being able to estimate the class of an object whose class is unknown, in the classification is also given a number of records of which 3 are called training sets, which consist of several attributes, attributes can be continuous or categorical, one of the attributes indicates the class for records[9]. According to the K-Nears Neighbors Algorithm a simple algorithm that stores all available cases and classifies new cases based on decision functions (e.g., distance measures)[10]. The *Naive Bayes* algorithm is one of the classification methods that can predict future opportunities based on past experience[11]. *K-Nears Neighbors* dan Algoritma *Naive Bayes* to classify the most important factors for the disease. The test using the Fuzzy Tsukamoto-GA method with the title Classification of Stroke Risk Level Using the GA-Fuzzy Tsukamoto Method from the results of the study was carried out to optimize the limits of membership functions with an accuracy rate of 86.66% obtained[12]. With the best parameters that have optimal results, namely the number of popsizes of 500. Based on the results of research conducted by previous researchers, it still needs to be developed again so that the classification of stroke prediction gets a higher level of accuracy.

METHODS

A. Research Stage

The method used in this study follows the stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM) model[13]. The stages of the research can be seen in Figure 1 below:

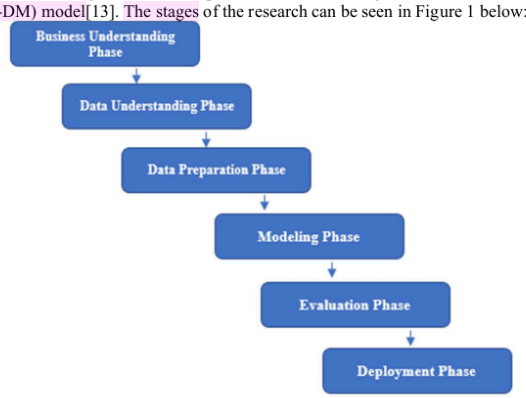

Figure 1 Research Flow

1. Business Understanding Phase
   This stage begins by defining the purpose of the study, which is to predict the risk of stroke with high accuracy. The focus is on building predictive models that can aid in early diagnosis and support clinical decision-making.
2. Data Understanding Phase
   In this stage, the dataset is explored to understand the patterns, distributions, and characteristics of each feature. The dataset includes parameters such as age, blood pressure, cholesterol levels, exercise habits, and disease history. In addition, preliminary analysis is carried out to detect missing values or anomalies.
3. Data Preparation Phase
   Data is processed to ensure quality and consistency, including handling missing values, encoding categorical features, and normalizing data. This step is important to maximize the performance of the machine learning algorithm used.
4. Modeling Phase

At this stage, the KNN and Naive Bayes algorithms are applied. KNN uses a distance-based approach for classification, while Naive Bayes utilizes probability theory to predict stroke status based on available parameters.

5. Evaluation Phase
   The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. This evaluation aims to determine which algorithm provides the best results in predicting stroke risk.
6. Deployment Phase
   Model developed can be applied in data-driven applications to support medical personnel in early diagnosis of stroke disease. In addition, the prediction results can be used to provide prevention recommendations to high-risk individuals.

B. Data Collection

The data collection method is an important thing in research and is a strategy or method used by researchers to collect the data needed in their research. Data used from Kaggle with url https://www.kaggle.com/code/docxian/stroke-prediction

C. Ekperimen

1. Algorithms KNN

Nearest Neighbor or k-Nearest Neighbor (kNN) is one of the classification algorithms in data mining that utilizes nearby data to make predictions on new data that is not yet known (test data)[14]. This algorithm works by finding a number of closest neighbors from the test data and determining the test data class based on the majority of the classes from the nearest neighbor (training data) found[15]. Nearest Neighbor can be used to handle various types of data, both numerical and categorical. In categorical data, the calculation of the distance of difference or similarity cannot be calculated using mathematical operations as can be done on numerical data. Given the training dataset D and spacing size

a. (x_i, y_i), $i = 1, 2, ..., N$
b. x_i is the training data in R^n
c. y_i is the appropriate class of the data x_i, and y_i $\{c_j, j\ 1, 2, ...\ M\}$
d. $dist\ (x - x_i) = ||\ x - x_i\ ||$

The new observation data x is classified into one of the yj classes using the following algorithm:

1. Enter new data x
2. Calculate the distance x to all xi training samples in the dataset: $dist\ (x - x_i)$
3. Sort $dist\ (x - x_i)\ (i = 1, 2, ..., N)$ In ascending order and order all xi according to: $x_{r1}, x_{r2}, ..., x_{rk}, ..., x_{rN}$
4. For the classification of the nearest neighbors (NN) classify x to y_{rl}
5. For *the K-NN* classification, classify x to the yrp majority class among the top k-rank data: $\{ x_{r1}, x_{r2}, ..., x_{rk}, \}$

Although the Euclidean ($L2$) and city block distance ($L1$) is a typical choice for distance measurements, other distances can be used depending on the application. The nearest neighbor (NN or 1-NN) produces too many classes, while *K-NN* provides more reliable classification results[16]. This is because the k value has a smoothing effect that makes the classifier more resistant to outliers. However, the performance of *the K-NN* classifier depends on the choice of k which is usually determined empirically.

2. Algorithms NB

Bayesian classification is a statistical classification that can be used to predict the probability of membership of a class discovered by the British scientist Thomas Bayes[17]. Naive Bayes is a fairly simple and easy-to-implement classification algorithm so it is very effective when tested with the correct data set, especially if Naive Bayes is combined with function selection, so that Naive Bayes can reduce redudant in the data, in addition Naive Bayes shows good results when combined with the clustering method[18]. Naive Bayes has proven to have high accuracy compared to support vector machines.

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

So X is the evidence, H is the hypothesis, P(H|X) i.e. probability is hypothesis H is true proof X or on P(H|X) is the posterior propiability H with the condition X, P(X|H) is the probability that the proof X is true or hypothesis H or the probability of Posterior X is the same as the condition H, P(H) is the probability prior to hypothesis H, and P(X) is the probability of the prior proof X.

$$P(C|F1 \ldots\ldots Fn) = \frac{P(C)P(F1.Fn|C)}{P(F1 \ldots Fn)}$$

So Variable C describes the class, while variable F1... Fn describes the character of the clue in classifying. Where this formula explains the chance that the sample enters the special character in class C (Posterior), namely the chance of leaving class C (before the entry of the sample, many are made priors), multiplied by the probability of the appearance of the character of the sample class C (also called likelihood), divided by the probability of the appearance of the sample character globally (also called evidence)[18]. The formula above can be made simply as follows

$$Posterior = \frac{Prior\ x\ likelihod}{evidence}$$

Continuous data classification is used Gauss Density formula:

$$P(Xi = Xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu i)^2}{2\sigma 2ij}}$$

Where: P: Opportunity
Xi : Attribute to i
xi : Value of attribute to i
Y : Searched class
yi : Subclass Y sought after
μ : mean, describing the average of all attributes
σ :Deviasi standar, menjelaskan varian across attributes.

D. Evaluation

Confusion matrix

This method only uses matrix tables as in Table 1, if the dataset consists of only two classes, one class is considered positive and the other is negative[19]. Evaluation with confusion matrix results in accuracy, precison, and recall values.

Table 1 Confusion Matrix

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

True Positive is the number of positive records classified as positive, false positive is the number of negative records classified as positive[20], false negative is the number of positive records classified as negative, true negative is the number of negative records classified as negative,

$$ACC = \frac{TP+TN}{TP+TN+FP+F}$$

$$P = \frac{TP}{TP+FP}$$

$$Sn = \frac{TP}{TP+FN}$$

$$Sp = \frac{TN}{TN+FP}$$

$$F - score = 2x\frac{P\ x\ Sn}{P+Sn}$$

Kurva ROC

The ROC curve is a graphical plot that illustrates the diagnostic capabilities of a binary classification system because its discrimination threshold varies. This method was originally developed for military radar receiving operators starting in 1941, which gave rise to its name. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. True positive levels are also known as sensitivity, memory, or detection probability. The false positive rate is also known as the probability of false alarms and can be calculated as (1 - specificity). ROC can also be thought of as a plot of strength as a function of the Type I Error of the decision rule (when performance is calculated only from a sample of the population, it can be considered as an estimator of this number). The performance of AUC accuracy can be classified into several groups, namely[21]:

1. 0.90 – 1.00 = *Exellent Classification*
2. 0.80 – 0.90 = *Good Classification*
3. 0.70 – 0.80 = *Fair Classification*
4. 0.60 – 0.70 = *Poor Classification*
5. 0.50 – 0.60 = *Failure Classification*

RESULT AND DISCUSSION

A. Data Preparation

Data on this Stroke disease research can be obtained directly from the official websitewww.kaggle.com(https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) which is accessed on July 17, 2024, The number of data records is 5,110 data which is the number of patients who experience symptoms and the number of dataset attributes is 12 attributes which are symptoms experienced by Stroke patients. The following table is the original data that will be processed starting from the preprocessing stage, the distribution of test data and training data, the implementation of the KNN method and optimization using the Bagging Technique. These attributes are described in Figure 4.1 below:

id	stroke	gender	age	hypertension	heart_disea...	ever_married	work_type	Residence_t...	avg_glucos...	bmi	smoking_st...
29540	0	Male	67	0	0	Yes	Private	Rural	97.040	26.9	smokes
53525	0	Female	72	0	0	Yes	Private	Urban	83.890	33.1	formerly smo...
65411	0	Female	51	0	0	Yes	Private	Urban	152.560	21.8	Unknown
26214	0	Female	63	0	0	Yes	Self-employed	Rural	75.930	34.7	formerly smo...
22190	0	Female	64	1	0	Yes	Self-employed	Urban	76.890	30.2	Unknown
56714	0	Female	1	0	0	No	children	Rural	62.130	16.8	Unknown
4211	0	Male	26	0	0	No	Govt_job	Rural	100.850	21	smokes
6369	0	Male	59	1	0	Yes	Private	Rural	95.050	30.9	never smoked
56799	0	Male	76	0	0	Yes	Govt_job	Urban	82.350	38.9	never smoked
32235	0	Female	45	1	0	Yes	Govt_job	Rural	95.020	N/A	smokes
28048	0	Male	13	0	0	No	children	Urban	82.380	24.3	Unknown
68598	0	Male	1	0	0	No	children	Rural	79.150	17.4	Unknown
41512	0	Male	57	0	0	Yes	Govt_job	Rural	76.620	28.2	never smoked
64520	0	Male	68	0	0	Yes	Self-employed	Urban	91.680	40.8	Unknown
579	0	Male	9	0	0	No	children	Urban	71.880	17.5	Unknown
7293	0	Male	40	0	0	Yes	Private	Rural	83.940	N/A	smokes
68398	0	Male	82	1	0	Yes	Self-employed	Rural	71.970	28.3	never smoked
36901	0	Female	45	0	0	Yes	Private	Urban	97.950	24.5	Unknown
45010	0	Female	57	0	0	Yes	Private	Rural	77.930	21.7	never smoked
22127	0	Female	18	0	0	No	Private	Urban	82.850	46.9	Unknown
14180	0	Female	13	0	0	No	children	Rural	103.080	18.6	Unknown

Figure 2 Dataset Snippets

The table above shows the data in this study, the data has a label on the Stroke attribute, the label on this data is 1 if the patient has a stroke or 0 if not. The tool used in this research is the RapidMiner application (Altair AI Studio).

B. Modeling

The selection and application of appropriate modeling techniques is carried out at this stage. The modeling in this study uses predictive data mining techniques.

1. Research Using the K-Nearest Neighbor Algorithm

The application of data on Rapidminer for Stroke Disease Prediction using the K-Nearest Neighbor algorithm is shown in figure 3 below:

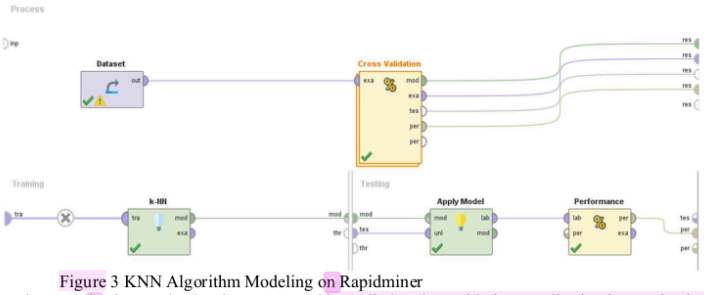

Figure 3 KNN Algorithm Modeling on Rapidminer

In Figure 3, the dataset that has been prepared is applied to the rapidminer application by conducting experiments using cross validation which can directly divide the data into training data and testing data because the data used is supervised and the algorithm used is KNN. The results of the experiment can be seen in figure 4.3 below.

accuracy: 94.50% +/- 0.28% (micro average: 94.50%)

	true 1	true 0	class precision
pred. 1	3	35	7.89%
pred. 0	246	4826	95.15%
class recall	1.20%	99.28%	

precision: 95.15% +/- 0.13% (micro average: 95.15%) (positive class: 0)

	true 1	true 0	class precision
pred. 1	3	35	7.89%
pred. 0	246	4826	95.15%
class recall	1.20%	99.28%	

recall: 99.28% +/- 0.35% (micro average: 99.28%) (positive class: 0)

	true 1	true 0	class precision
pred. 1	3	35	7.89%
pred. 0	246	4826	95.15%
class recall	1.20%	99.28%	

Figure 4 Confusion Matrix of KNN Algorithm

Figure 4 is the confusion matrix that shows the results of the experiment, in the confusion matrix we can see the results of accuracy, precision class, and recall class. The resulting accuracy is 94.50 %. The results of the accuracy in the previous study with this research experiment using the KNN algorithm for stroke disease prediction are shown in table 4.1 below:

2. Research Using Naive Bayesian Algorithm

A method that can improve the level of accuracy is the use of the naïve Bayes algorithm. The follow-up experiment carried out in this study is the use of the naïve bayes algorithm. The application of stroke prediction rapidminers using the Naïve Bayes Algorithm can be seen in figure 5 below:

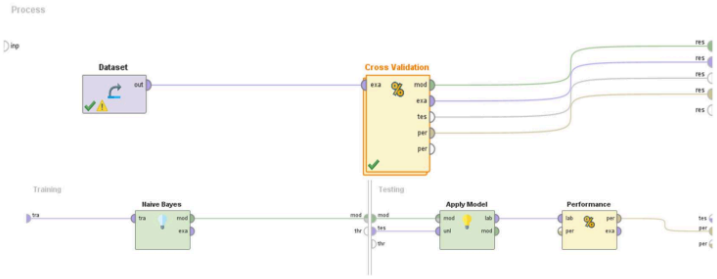

Figure 5 Application of Rapidminer naïve bayes algorithm

The experiment in figure 5 is a validation technique to divide training data and testing data using cross validation validation techniques. From the experiment, we get the results that we can see in figure 6 below

accuracy: 88.83% +/- 1.41% (micro average: 88.83%)

	true 1	true 0	class precision
pred. 1	75	397	15.89%
pred. 0	174	4464	96.25%
class recall	30.12%	91.83%	

precision: 96.25% +/- 0.44% (micro average: 96.25%) (positive class: 0)

	true 1	true 0	class precision
pred. 1	75	397	15.89%
pred. 0	174	4464	96.25%
class recall	30.12%	91.83%	

recall: 91.83% +/- 1.36% (micro average: 91.83%) (positive class: 0)

	true 1	true 0	class precision
pred. 1	75	397	15.89%
pred. 0	174	4464	96.25%
class recall	30.12%	91.83%	

Figure 6 Naïve Bayes' Algorithm's Confusion Matrix

The figure shows the results of the evaluation of the Naive Bayes model for stroke disease prediction. From these results, the model produces an accuracy level of 88.83% with small variability (±1.41%). This accuracy indicates that the model is able to classify data quite well.

However, there are significant differences between precision and recall for each class. For class 1 (patients with stroke), the accuracy is 15.89%, while the recall is 30.12%. This suggests that although the model was able to capture some patients who actually had a stroke, many of the class 1 predictions were wrong. In contrast, for class 0 (patients without stroke), precision and recall were 96.25% and 91.83%, respectively, showing much better performance in detecting patients without stroke.

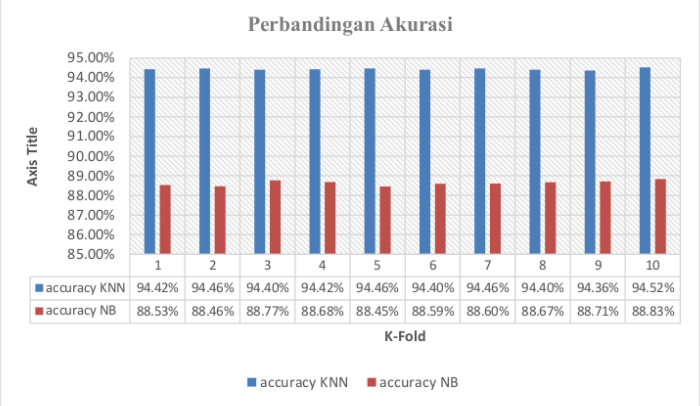

Figure 7 Accuracy Comparison

Based on the table and diagram above, it shows that the change in the number of K-Folds in Cross Validation in classifying will produce different accuracy so that we can produce the best accuracy. We can see that on the K-Fold 10 the accuracy and precision results are the highest, in contrast to the recall results that are not higher than other K-Fold uses.

In addition to the Confusion matrix to determine the performance of this experiment, we can rely on the resulting ROC-AUC curve. A comparison of the results of the ROC-AUC curve in studies that did not use Bagging and those that used Bagging can be seen in figures 8 and 9 below:

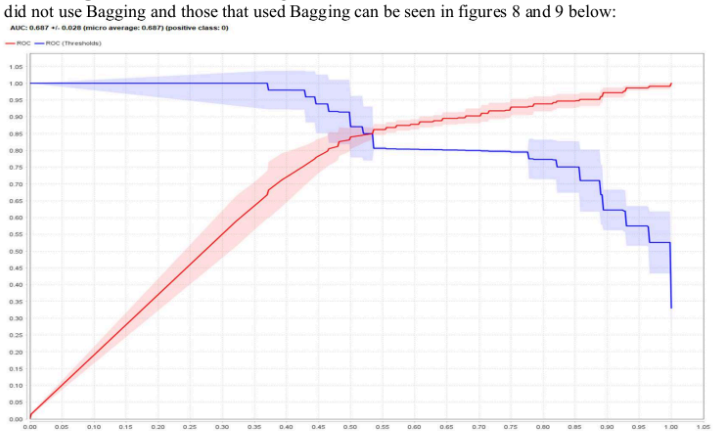

Figure 8 ROC-AUC Curve Experimental Results Using KNN

The figure shows the Receiver Operating Characteristic (ROC) curve for *the K-Nearest Neighbors (KNN)* model in predicting stroke disease. The resulting Area Under the Curve (AUC) value is **0.669 ± 0.048**, with a micro-average value of **0.669**. The ROC and AUC curves indicate the model's ability to distinguish between patients at risk of stroke (positive class) and without stroke risk (negative class).

The ROC curve plots *the True Positive Rate* (TPR) against the *False Positive Rate* (FPR) at various threshold values. Visually, the curve shows that the model has a moderate ability to distinguish between the two classes. An AUC of 0.669 indicates that the model has a 66.9% probability of correctly distinguishing between positive and negative classes, which are above the random line (AUC = 0.5), but still far from ideal performance (AUC close to 1.0).

This performance suggests that the KNN model has difficulty effectively detecting patients at risk of stroke (positive class), mainly due to class imbalances in the dataset. This imbalance is seen in the ROC curve closer to the diagonal line, which indicates a less than optimal performance in classifying minority classes.

These results indicate the need for optimization measures, such as the use of data balancing techniques (*oversampling* or *undersampling*), adjustment of KNN model parameters, or the use of alternative algorithms that are more sensitive to class imbalances. With these steps, the model's ability to detect patients at risk of stroke can be improved, making the prediction results more reliable and applicable in the real world.

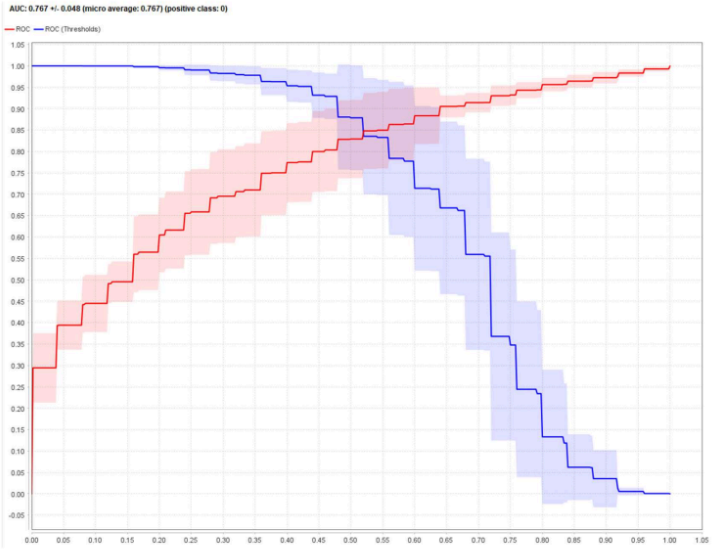

Figure 9 ROC-AUC Curve Experiment Using Naïve Bayes

The figure shows the Receiver Operating Characteristic (ROC) graph used to evaluate the performance of the Naive Bayes model in predicting stroke disease. The resulting Under the Curve (AUC) area is **0.767 ± 0.048**, with a micro mean value of **0.767**. This AUC value shows the model's ability to distinguish between positive classes (patients with stroke) and negative classes (patients without stroke).

ROC is a graph that plots *the True Positive Rate* (TPR) against *the False Positive Rate* (FPR) for various thresholds. The ROC curve near the top left corner shows good model performance. In this figure, the ROC curve shows that the model is performing quite well, but there is still room for improvement, especially in detecting positive classes.

An AUC of 0.767 indicates that the model has a 76.7% probability of correctly distinguishing between patients with stroke and without stroke. Although these results are quite good, the model's performance still needs to be improved, especially given the importance of early detection of stroke to reduce the risk of serious complications.

Errors in the prediction of false negatives can have a significant impact in the clinical context, as patients who actually have a stroke risk may go undetected. Therefore, further optimizations are needed, such as adjusting model parameters or using more complex algorithms, to improve the sensitivity and accuracy of predictions. In addition, balancing datasets or using ensemble methods can help improve model performance in predicting minority classes.

3. DISCUSSION

This study aims to predict stroke disease using *the K-Nearest Neighbors (KNN)* and *Naive Bayes* algorithms with datasets taken from Kaggle. This dataset consists of 5,110 patient data with 12 attributes that cover various symptoms experienced by patients. Modeling was carried out using the RapidMiner application with *cross-validation* validation to divide the data into training data and test data.

In the first experiment, the KNN algorithm produced an accuracy of **94.50%,** showing good performance in predicting stroke disease overall. However, the results of the confusion matrix show that this algorithm has a weakness in detecting patients at risk of stroke (class 1), with an accuracy of only **7.89%** and a recall **of 1.20%.** This occurs due to data imbalance, where the number of patients without stroke (class 0) dominates the dataset. In addition, the results of the ROC-AUC curve for KNN showed AUC values of **0.669 ± 0.048**, indicating moderate performance in differentiating patients with and without stroke. Class imbalance is the main obstacle in the performance of the KNN model, which can be overcome through data balancing techniques such as *oversampling* or *undersampling*.

In the second experiment, the Naive Bayes algorithm was applied to improve prediction performance. As a result, the accuracy of the model is **88.83% ± 1.41%,** lower than KNN. However, the model showed better performance in terms of prediction probability with AUC values of **0.767 ± 0.048**. Nonetheless, class imbalances still affected the performance of the Naive Bayes model, especially in detecting patients with stroke (class 1), where the precision was **15.89%** and the recall **was 30.12%.** This shows that although Naive Bayes has advantages in prediction efficiency and probability, the model also requires further optimization to improve sensitivity to minority classes.

Both algorithms show the advantages and disadvantages of each. KNN excels in overall accuracy, but is less sensitive to minority classes. In contrast, Naive Bayes has a better performance in modeling probabilities but is still affected by the unbalanced distribution of data. Model performance can be improved by optimizing model parameters, balancing datasets, or using ensemble techniques such as bagging or boosting.

CONCLUSION

This study shows that *the K-Nearest Neighbors (KNN)* and *Naive Bayes* algorithms can be used to predict stroke disease with a good degree of accuracy. The KNN algorithm produced the highest accuracy of **94.50**%, but was less effective in detecting patients with stroke (class 1) due to dataset imbalance. On the other hand, Naive Bayes gave quite competitive results with an AUC **of 0.767 ± 0.048**, showing a better ability in modeling class probabilities.

To overcome the weaknesses faced by both algorithms, optimization steps such as data balancing, the use of ensemble techniques, or exploration of other algorithms such as *Random Forest* or *Gradient Boosting* can be implemented. This study also emphasizes the importance of proper parameter selection

and thorough validation in building a more reliable and applicable stroke disease prediction model in the real world.

With the right optimization and integration, these predictive models have great potential to support early diagnosis, assist healthcare professionals, and provide preventive recommendations to high-risk individuals.

REFERENCES

[1] D. E. Cahyani and D. E. Cahyani, "PENERAPAN MACHINE LEARNING UNTUK PREDIKSI PENYAKIT STROKE," *Jurnal Kajian Matematika dan Aplikasinya VOLUME*, vol. 3, no. 1, 2022, doi: 10.17977/um055v3i1p15-22.

[2] I. Widharma *et al.*, "PERANCANGAN SISTEM INFORMASI PENYINTAS STROKE BERBASIS WEB DENGAN METODE SDLC," *DA Indah Cahya Dewi*, vol. 6, no. 2, p. 41.

[3] P. Bidang Komputer Sains dan Pendidikan Informatika, D. Akademi Perekam dan Informasi Kesehatan Iris Padang Jl Gajah Mada No, and S. Barat, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4.5 Yuli Mardi".

[4] Zuriati Z and Diterima, "Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN) INFORMASI ARTIKEL ABSTRAK Classification of Stroke Using the K-Nearest Neighbor (KNN) Algorithm," vol. 1, no. 1, pp. 1–8, 2023, doi: 10.20222/rt.v1i1.2665.

[5] S. and Communication Networks, "Retracted: Analysis and Application of Data Mining Technology for College English Education Integration," *Security and Communication Networks*, vol. 2024, pp. 1–1, Jan. 2024, doi: 10.1155/2024/9836129.

[6] G. Sanhaji and A. I. Hizbullah, "PEMANFAATAN ARTIFICIAL INTELLIGENCE DALAM BIDANG KESEHATAN," *EDUSAINTEK: Jurnal Pendidikan, Sains dan Teknologi*, vol. 11, no. 1, pp. 234–242, Aug. 2023, doi: 10.47668/edusaintek.v11i1.999.

[7] P. H. Trenggono and A. Bachtiar, "PERAN ARTIFICIAL INTELLIGENCE DALAM PELAYANAN KESEHATAN : A SYSTEMATIC REVIEW," 2023, [Online]. Available: http://journal.universitaspahlawan.ac.id/index.php/ners

[8] J. T. Atmojo *et al.*, "ARTIFICIAL INTELLIGENCE DALAM PRAKTIK KESEHATAN," 2024. [Online]. Available: http://journal.stikeskendal.ac.id/index.php/PSKM

[9] S. Hassani and U. Dackermann, "A Systematic Review of Advanced Sensor Technologies for Non-Destructive Testing and Structural Health Monitoring," Feb. 01, 2023, *MDPI*. doi: 10.3390/s23042204.

[10] X. Liu, J. Yan, S. Shan, and R. Wu, "A Blockchain-Assisted Electronic Medical Records by Using Proxy Reencryption and Multisignature," *Security and Communication Networks*, vol. 2022, 2022, doi: 10.1155/2022/6737942.

[11] X. Yan and X. Ren, "5G Edge Computing Enabled Directional Data Collection for Medical Community Electronic Health Records," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/5598077.

[12] C. C. A. Silva, G. S. Aquino, S. R. M. Melo, and D. J. B. Egdio, "A fog computing-based architecture for medical records management," *Wirel Commun Mob Comput*, vol. 2019, 2019, doi: 10.1155/2019/1968960.

[13] B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/art20203995.

[14] D. Berezkin, I. Kozlov, and P. Martynyuk, "Predictive analytics of scientific and technological trends for decision making in university management," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 270–277. doi: 10.1016/j.procs.2024.03.001.

[15] B. S. Wiguna, D. Purwitasari, and D. O. Siahaan, "Deep Learning Approach for Health Question and Answer Text Segmentation based on Physician-Patient Communication Aspect," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 213–221. doi: 10.1016/j.procs.2024.02.168.

[16] V. P. Prasetyo, M. F. A. Ulin Nuha, M. H. Hakiki, R. A. Vinarti, and A. Djunaidy, "Comparison of Data Mining Techniques on Stroke Clinical Dataset," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 502–511. doi: 10.1016/j.procs.2024.03.033.

[17] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds., Oxford: Academic Press, 2019, pp. 403–412. doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

[18] T. T. Sang Nguyen, "Model-based book recommender systems using Naïve Bayes enhanced with optimal feature selection," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2019, pp. 217–222. doi: 10.1145/3316615.3316727.

[19] Parteek Bhatia, "Data Mining and Data Warehousing," 2019.

[20] C. Karima and W. Anggraeni, "Performance Analysis of the Ada-Boost Algorithm For Classification of Hypertension Risk With Clinical Imbalanced Dataset," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 645–653. doi: 10.1016/j.procs.2024.03.050.

[21] F. Ridzuan and W. M. N. W. Zainon, "A Review on Data Quality Dimensions for Big Data," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 341–348. doi: 10.1016/j.procs.2024.03.008.

Comparison Of The Performance Of K-Nearest Neighbors And Naive Bayes Algorithms For Stroke Disease Prediction.docx

ORIGINALITY REPORT

20%	17%	15%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dokumen.pub Internet Source	3%
2	Dani Wilian, Sriyanto Sriyanto. "Comparison of the Performance of the C.45 Algorithm with Naive Bayes in Analyzing Book Borrowing at the Library Pringsewu Muhammadiyah University", Jurnal Sisfokom (Sistem Informasi dan Komputer), 2025 Publication	2%
3	ejournal.uin-suska.ac.id Internet Source	2%
4	id.123dok.com Internet Source	1%
5	jurnal.atmaluhur.ac.id Internet Source	1%
6	repository.uki.ac.id Internet Source	1%
7	www.frontiersin.org Internet Source	1%
8	fisiocrem.ro Internet Source	1%
9	Submitted to Polytechnic of Turin Student Paper	1%
10	jeeemi.org	

	Internet Source	1 %
11	peerj.com Internet Source	1 %
12	Harits Ar Rosyid, Utomo Pujianto, Bima Garis Invarian. "Performance Comparison of Naïve Bayes and Neural Network in Predicting Student Violation", 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021 Publication	1 %
13	conference.upnvj.ac.id Internet Source	1 %
14	ejurnal.teknokrat.ac.id Internet Source	<1 %
15	Jihadul Akbar, Ema Utami, Ainul Yaqin. "Multi-Label Classification of Film Genres Based on Synopsis Using Support Vector Machine, Logistic Regression and Naïve Bayes Algorithms", 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022 Publication	<1 %
16	core.ac.uk Internet Source	<1 %
17	scholar.its.ac.id Internet Source	<1 %
18	Nina Meliana, Sunardi, Abdul Fadlil. "Identification of Cyber Bullying by using Clustering Methods on Social Media Twitter", Journal of Physics: Conference Series, 2019 Publication	<1 %

19	ojs.trigunadharma.ac.id Internet Source	<1 %
20	Submitted to SRM University Student Paper	<1 %
21	jurnal.ugm.ac.id Internet Source	<1 %
22	ejurnal.ars.ac.id Internet Source	<1 %
23	Submitted to Universitas Brawijaya Student Paper	<1 %
24	Submitted to UIN Sunan Ampel Surabaya Student Paper	<1 %
25	Muhammad Risha, Mohamed Elsaadany, Paul Liu. "Uncertainty-Driven Modeling of Microporosity and Permeability in Clastic Reservoirs Using Random Forest", Qeios Ltd, 2025 Publication	<1 %
26	Submitted to The Hong Kong Polytechnic University Student Paper	<1 %
27	Submitted to Universitas Negeri Manado Student Paper	<1 %
28	www.vision.edu.mk Internet Source	<1 %
29	ejurnal.seminar-id.com Internet Source	<1 %
30	ioinformatic.org Internet Source	<1 %
31	mdpi-res.com Internet Source	

		<1%
32	"Selected Contributions on Statistics and Data Science in Latin America", Springer Science and Business Media LLC, 2019 Publication	<1%
33	jurnal.itscience.org Internet Source	<1%
34	Sutriawan, Wasis Haryo Sasoko, Zumhur Alamin, Ritzkal. "Benchmarking Text Embedding Models for Multi-Dataset Semantic Textual Similarity: A Machine Learning-Based Evaluation Framework", Acadlore Transactions on AI and Machine Learning, 2025 Publication	<1%
35	hig.diva-portal.org Internet Source	<1%
36	jurnal.suryanusantara.ac.id Internet Source	<1%
37	Indra Irawan, M Riski Qisthiano, Muhammad Syahril, Pamuji M. Jakak. "Optimasi Prediksi Kelulusan Tepat Waktu: Studi Perbandingan Algoritma Random Forest dan Algoritma K-NN Berbasis PSO", Jurnal Pengembangan Sistem Informasi dan Informatika, 2023 Publication	<1%
38	Saruni Dwiasnati, Yudo Devianto. "Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C.45 and Naive Bayes Algorithms", Journal of Physics: Conference Series, 2019 Publication	<1%

39	ejurnal.umri.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		