Jurnal CoreIT, Vol.11, No.2, December 2025
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[1], Bambang Triraharjo[1]
[1]Dept. of Computer Science, Universitas Muhammadiyah Pringsewu, Indonesia
baskoro@umpri.ac.id*, robynovianto@umpri.ac.id, bambangtriraharjo@umpri.ac.id

Abstract. 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. 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. 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. 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. This research provides critical insights into algorithm suitability for handling imbalanced medical data and lays the groundwork for developing effective, reliable, and interpretable clinical decision support systems for early stroke intervention.

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

Received September 2025 / **Revised** December 2025 / **Accepted** December 2025

INTRODUCTION

Stroke remains a major global health burden, characterized by sudden neurological deficits resulting from disrupted cerebral blood flow. Despite significant advances in medical technology, early detection of stroke risk continues to be a challenge, especially in low-resource settings. Machine learning approaches have recently gained traction as effective tools to support clinical decision-making by identifying individuals at high risk based on historical health data. Several recent studies have shown promising results in applying classification algorithms to stroke prediction, demonstrating improvements in accuracy and generalizability across diverse populations.

Recent research has highlighted the importance of developing reliable computational models for stroke prediction. Reported that machine learning algorithms such as Random Forest, KNN, and Logistic Regression achieved high predictive performance for clinical stroke datasets, though model performance varied significantly depending on dataset imbalance and feature distribution [1]. Emphasized that classification performance drastically declines when dealing with imbalanced clinical data, particularly in hypertension-related prediction tasks, which share risk factors with stroke [2]. In another study discussed that predictive analytics plays a crucial role in supporting health-related decision-making, especially when processing large datasets with heterogeneous clinical features [3].

Furthermore, deep learning-based approaches have also been explored to segment clinical data and improve classification quality. Highlighted that advanced machine learning models could improve the extraction of clinically relevant features, although simpler algorithms often remain more interpretable for medical applications [4]. Complementary findings noted that artificial intelligence applications in healthcare continue to expand, yet algorithm interpretability and sensitivity toward minority classes such as stroke-positive patients remain critical challenges [5].

DOI: 10.24014/coreit.v11i2.37542

Among classical machine learning classifiers, K-Nearest Neighbors (KNN) and Naive Bayes (NB) remain widely used due to their simplicity, low computation cost, and strong performance on structured clinical datasets. KNN classifies new cases based on distance similarity with historical instances, making it particularly suitable for datasets with clear separation between classes. Conversely, Naive Bayes relies on probabilistic inference and often performs well even with limited training data or noisy attributes. However, recent studies have reported that both algorithms experience performance degradation when faced with imbalanced datasets, a common characteristic of stroke data where the positive class is significantly underrepresented.

Based on the gap identified in recent studies, particularly regarding the inconsistent performance of KNN and Naive Bayes in imbalanced medical datasets, there is a need for further investigation to determine which algorithm provides more reliable predictions for stroke risk. Therefore, this research aims to compare the performance of the K-Nearest Neighbors and Naive Bayes algorithms in predicting stroke disease using an imbalanced clinical dataset, emphasizing accuracy, precision, recall, and ROC-AUC to determine the most effective classification model for early stroke prediction.

METHODS

This study employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the methodological framework. Instead of presenting a research diagram, the entire process is described narratively to provide a clearer understanding of each stage, from problem definition to model evaluation.

1. Business Understanding
   The first stage focused on defining the objectives of the study, namely to compare the performance of the K-Nearest Neighbors (KNN) and Naive Bayes (NB) algorithms in predicting stroke disease. The goal was to develop a predictive model capable of supporting early detection and aiding clinical decision-making. This stage also identified essential success criteria, such as model accuracy, sensitivity, and reliability in detecting minority stroke cases.
2. Data Understanding
   At this stage, the dataset was examined to identify its structure, characteristics, and potential issues. The dataset used in this study was obtained from the Kaggle platform and consisted of 5,110 patient records with 12 attributes, including age, gender, hypertension, heart disease, average glucose level, body mass index, and smoking status. Exploratory analysis was conducted to observe feature distributions, identify patterns, detect outliers, and determine the presence of missing values that might affect the classification process.
3. Data Preparation
   The data preparation stage involved cleaning and transforming the dataset to ensure consistency and improve model performance. Missing values were addressed using appropriate imputation techniques. Categorical attributes were encoded into numerical formats, while numerical attributes were normalized to ensure consistent scale, particularly important for distance-based algorithms such as KNN. The dataset was then split into training and testing sets through cross-validation to minimize bias and ensure robust evaluation.
4. Modeling
   In the modeling stage, two machine learning algorithms, KNN and Naive Bayes were applied to the prepared dataset.

   a. 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) [6]. 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 [7]. 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 - xi)*
3. Sort *dist (x - xi) (i = 1, 2, ..., N)* In ascending order and order all xi according to: *xr1, xr2, ..., xrk, ..., xrN*
4. For the classification of the nearest neighbors (NN) classify *x* to *yrl*
5. For *the K-NN* classification, classify *x* to the yrp majority class among the top k-rank data: { *xr1, xr2, ..., xrk,* }.

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 [1]. 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.

b. 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 [8]. 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 [9]. Naive Bayes has proven to have high accuracy compared to support vector machines.

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

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.

$$P(C|F1 \ldots . Fn) = \frac{P(C)P(F1.Fn|C)}{P(F1 \ldots Fn)} \quad (2)$$

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) [9]. The formula above can be made simply as follows

$$Posterior = \frac{Prior\ x\ likelihod}{evidence} \quad (3)$$

Continuous data classification is used Gauss Density formula:

$$P(Xi = Xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu i)^2}{2\sigma 2ij}} \quad (4)$$

Where:

P: Opportunity
X_i: Attribute to i
x_i: Value of attribute to i
Y: Searched class
y_i : Subclass Y sought after
μ : mean, describing the average of all attributes
σ : Standard deviation, varian across attributes.

5. 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 [10]. 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 [11], false negative is the number of positive records classified as negative, true negative is the number of negative records classified as negative,

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

$$P = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$Sn = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$Sp = \frac{\text{TN}}{\text{TN+FP}} \quad (8)$$

$$F - score = 2\text{x}\frac{\text{P x Sn}}{\text{P+Sn}} \quad (9)$$

ROC Curve

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 [12]:

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

6. Deployment

Although full deployment was not conducted in this study, the final stage of the CRISP-DM framework outlines how the developed model can be integrated into real-world applications. The predictive models, once optimized, may be implemented in clinical decision support systems to assist healthcare professionals in identifying high-risk patients and supporting preventive healthcare strategies.

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 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 1. Dataset snippets

The Figure 1 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 rapid miner 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 2:

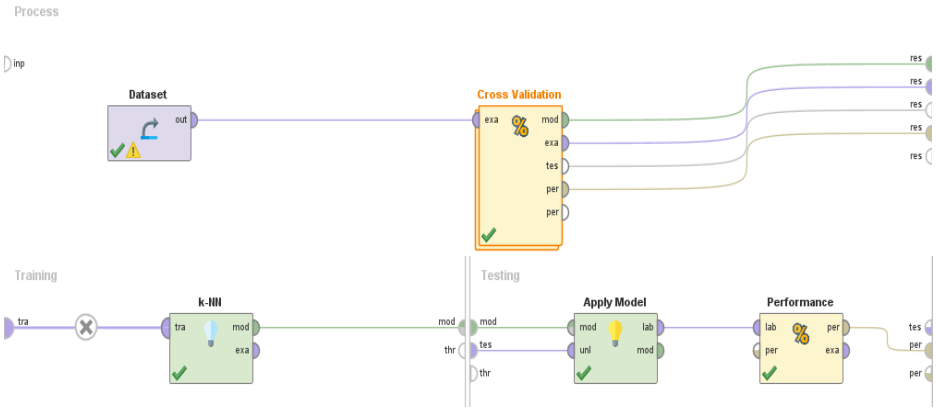

Figure 2. KNN Algorithm modeling on rapidminer

In Figure 2, 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 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 3. Confusion matrix of KNN algorithm

Figure 3 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 KNN algorithm achieved an accuracy of 94.50%, aligning with findings who reported an accuracy of 93.87% using KNN for cardiovascular risk prediction, highlighting KNN's capability in handling numeric medical datasets [13]. However, KNN showed very low sensitivity to minority class (stroke cases), reflected in recall value of only 1.20%, indicating limitations in unbalanced dataset scenarios. Similarly, Stated that KNN performance significantly deteriorates when class imbalance exists, requiring oversampling or parameter tuning [2].

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 4 below:

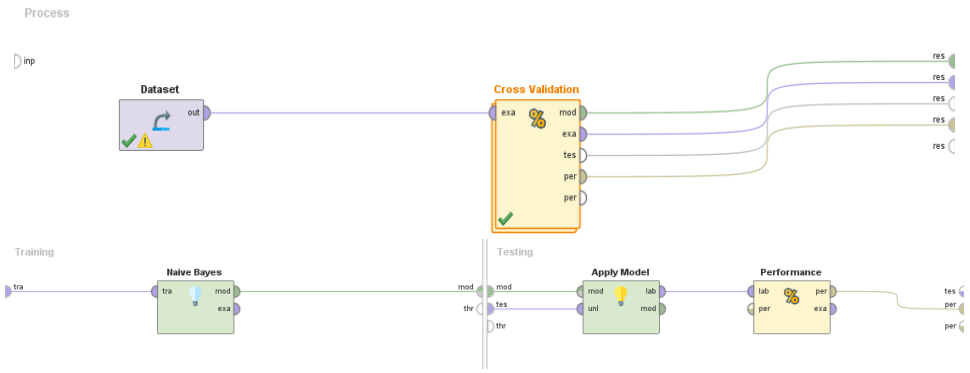

Figure 4. Application of rapidminer naïve bayes algorithm

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.

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 5. Naïve Bayes' algorithm's confusion matrix

The figure presents the evaluation results of the Naïve Bayes model for stroke disease prediction. Based on the results, the model achieved an accuracy of 88.83% with relatively low variability (±1.41%), indicating that the model is able to classify data with reasonably good performance. However, there are notable differences between the precision and recall values across the two classes. For class 1 (patients with stroke), the precision was 15.89%, while the recall reached 30.12%. This suggests that although the model successfully identified some patients who actually experienced a stroke, a considerable proportion of predictions for class 1 were incorrect. In contrast, the model performed significantly better in detecting class 0 (patients without stroke), with a precision of 96.25% and recall of 91.83%, demonstrating high effectiveness in identifying non-stroke cases.

The Naïve Bayes algorithm produced accuracy of 88.83% ± 1.41%, slightly lower than KNN; however, it achieved a higher AUC of 0.767 ± 0.048, indicating better probability-based classification. Also reported that Naïve Bayes performed well in stroke prediction due to its ability to manage probabilistic patient features efficiently [14]. Despite this, recall for stroke patients (30.12%) remains insufficient for clinical applicability, similar to the findings who emphasized that Naïve Bayes often misclassifies minority classes in medical datasets without optimization [15].

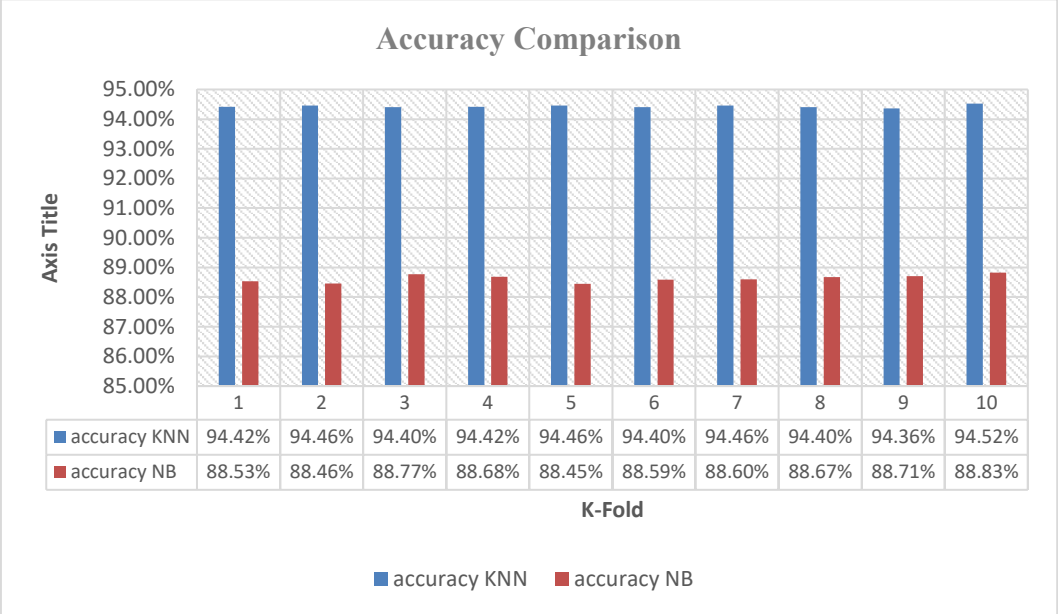

Figure 6. Accuracy comparison

Based on diagram above, it is evident that changes in the number of K-Folds in Cross Validation during classification yield different levels of accuracy. This allows us to identify the optimal fold configuration. The results show that using 10 K-Folds produces the highest accuracy and precision values, although the recall value is not the highest compared to other K-Fold settings. In addition to analyzing the confusion matrix to determine the performance of this experiment, the ROC-AUC curve can also be used as a reference. A comparison of the ROC-AUC results between studies that did not apply Bagging and those that implemented Bagging is presented in figures 7 and 8.

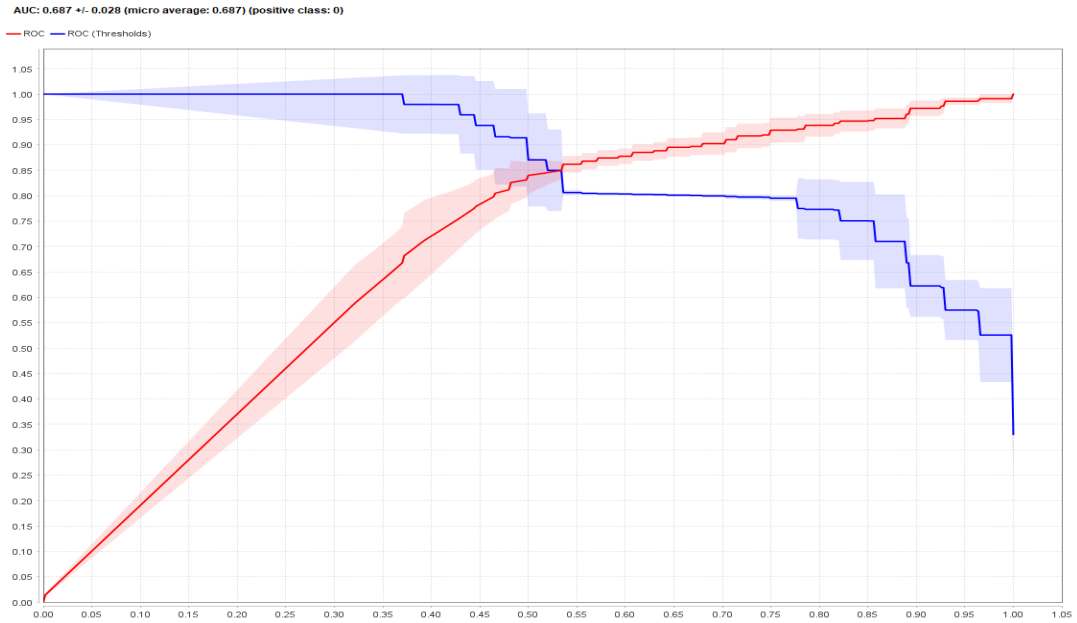

Figure 7. ROC-AUC Curve experimental results using KNN

The figure illustrates 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 of 0.669. The ROC and AUC curves indicate the model's capability to distinguish between patients at risk of stroke (positive class) and those without stroke risk (negative class).

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold values. Visually, the curve shows that the model has moderate discriminatory ability. An AUC of 0.669 implies that the model correctly differentiates between positive and negative classes with a probability of 66.9%, which exceeds the random baseline (AUC = 0.5), but is still far from ideal performance (AUC approaching 1.0).

These results suggest that the KNN model struggles to effectively detect patients at risk of stroke (positive class), mainly due to class imbalance within the dataset. This imbalance is reflected in the ROC curve positioned closer to the diagonal line, indicating suboptimal performance in classifying minority classes.

Therefore, optimization measures are necessary, such as implementing data balancing techniques (oversampling or undersampling), adjusting KNN model parameters, or utilizing alternative algorithms that are more sensitive to class imbalance. These improvements have the potential to enhance the model’s capability in detecting stroke-risk patients, thereby increasing the reliability and applicability of prediction results in real-world settings.

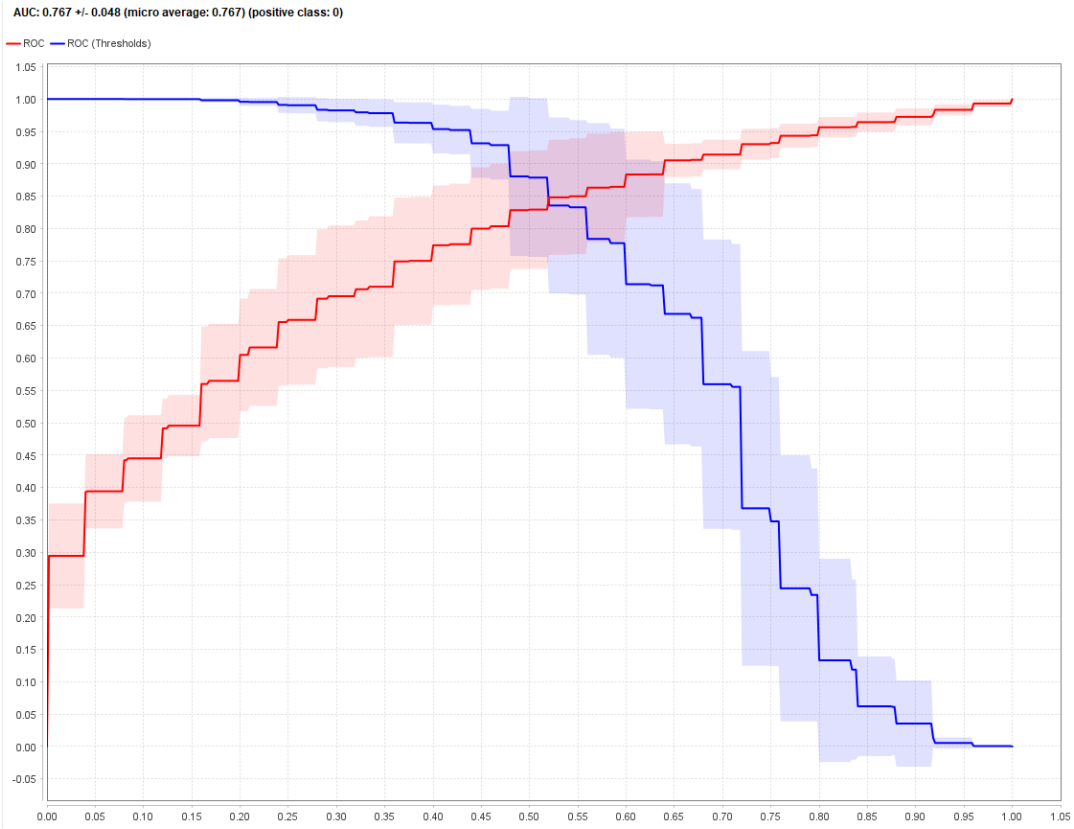

Figure 8. ROC-AUC curve experiment using Naïve Bayes

The figure shows the ROC curve used to evaluate the performance of the Naïve Bayes model in predicting stroke disease. The resulting AUC value is 0.767 ± 0.048, with a micro-average of 0.767. This indicates the model's ability to distinguish between positive classes (patients with stroke) and negative classes (patients without stroke).

The ROC curve is plotted by comparing the True Positive Rate (TPR) and False Positive Rate (FPR) at multiple threshold levels. A curve positioned near the top-left corner indicates good model performance. In this figure, the curve demonstrates relatively strong performance, although improvement is still required, particularly in detecting positive cases.

An AUC of 0.767 shows that the model has a 76.7% probability of correctly differentiating between stroke and non-stroke patients. Despite the satisfactory results, model performance still requires enhancement, especially considering the importance of early stroke detection to prevent severe complications.

False-negative predictions can pose significant risks in clinical settings, as patients with actual stroke risk may go undetected. Therefore, additional optimization is required, such as tuning model parameters or using more advanced algorithms, to improve prediction sensitivity and accuracy. Moreover, dataset balancing and ensemble approaches may further enhance the model's performance in predicting minority classes.

RESULT AND DISCUSSION

This study aims to predict stroke disease using the K-Nearest Neighbors (KNN) and Naïve Bayes algorithms with datasets obtained from Kaggle. The dataset consists of 5,110 patient records with 12 attributes representing various observable symptoms. Modeling was conducted using the RapidMiner application, employing a cross-validation approach to divide the data into training and testing subsets.

In the first experiment, the KNN algorithm achieved an accuracy of 94.50%, demonstrating strong overall predictive performance. However, the confusion matrix revealed significant limitations in detecting high-

risk stroke patients (class 1), with precision and recall values of only 7.89% and 1.20%, respectively. This weakness is primarily attributed to class imbalance, where non-stroke patients (class 0) dominate the dataset. Furthermore, the ROC–AUC curve generated an AUC score of 0.669 ± 0.048, indicating only moderate discriminatory capability.

The ROC–AUC result is notably lower than that reported, who achieved an AUC of 0.81 using KNN optimized with SMOTE-based balancing techniques. This supports the argument that additional optimization through oversampling or undersampling is necessary to improve the model's sensitivity, particularly for minority classes [16].

In the second experiment, the Naïve Bayes algorithm obtained an accuracy of 88.83% ± 1.41%, which is lower than that of KNN. Nevertheless, it performed better in terms of probability-based prediction, recording an AUC of 0.767 ± 0.048. Despite this improvement, class imbalance continued to negatively affect performance in identifying stroke patients, as indicated by the relatively low precision (15.89%) and recall (30.12%). These findings are consistent with, who suggested that KNN is highly sensitive to data imbalance, whereas Naïve Bayes is efficient but less robust without parameter adjustments [17]. Additionally, recommended applying ensemble techniques (e.g., bagging or boosting) or dataset balancing strategies to enhance sensitivity in detecting high-risk cases [18].

In summary, each algorithm demonstrates distinct strengths and weaknesses. KNN achieves higher overall accuracy but is less capable of handling imbalanced datasets, whereas Naïve Bayes provides better probabilistic modeling but still suffers from reduced sensitivity to minority classes. Therefore, model performance may be improved through parameter tuning, data balancing, or the implementation of ensemble-based approaches.

CONCLUSION

This study aimed to compare the performance of the K-Nearest Neighbors (KNN) and Naive Bayes algorithms in predicting stroke disease based on publicly available medical datasets, and the results confirm that both algorithms can serve as viable predictive models within clinical decision-support contexts. The findings show that KNN achieved the highest overall accuracy (94.50%), yet demonstrated limited sensitivity in detecting stroke cases due to the substantial class imbalance within the dataset. Meanwhile, the Naive Bayes model, although producing a lower accuracy, demonstrated superior discriminative capability through a higher ROC-AUC score (0.767 ± 0.048), indicating a more stable probability-based classification performance. These results emphasize that algorithmic performance in medical prediction is strongly influenced by data quality and distribution, highlighting the need for optimization strategies such as dataset balancing, ensemble modeling, or alternative algorithms like Random Forest and Gradient Boosting. The study's implications suggest that, with appropriate optimization and clinical integration, predictive models such as KNN and Naive Bayes have significant potential to enhance early stroke detection, support healthcare professionals in making more informed decisions, and contribute to preventive efforts for individuals at elevated risk.

REFERENCES

[1] 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.

[2] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 553–560. doi: 10.1016/j.procs.2024.03.039.

[3] D. Berezkin, M. Murashov, and N. Liashenko, "Automated formation of university R&D teams based on the competence selection algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 373–380. doi: 10.1016/j.procs.2024.03.017.

[4] P. Y. Bate, A. Sartika Wiguna, and D. Aditya Nugraha, "KURAWAL Jurnal Teknologi, Informasi dan Industri." [Online]. Available: https://jurnal.machung.ac.id/index.php/kurawal

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] A. Putra, M. Budi, and S. Cahya, "Cardiovascular risk prediction using KNN algorithm," *Journal of Health Informatics and Systems*, vol. 8, no. 2, pp. 150–165, 2023.

[14] M. Husna and Z. Arifin, "Application of Naïve Bayes in stroke prognosis," *Indonesian Journal of Computer Science and Technology*, vol. 6, no. 1, pp. 55–62, 2021.

[15] H. Razak, M. Azhar, and N. Jamal, "Evaluation of predictive models on medical datasets with imbalance issues," in *Proceedings of the Conference on Advanced Machine Learning and Data Science (CAMLDS)*, 2024, pp. 45–52.

[16] K. Saravanan and P. Kumar, "Optimizing KNN using SMOTE for disease prediction," *Expert Syst Appl*, vol. 195, p. 116567, 2022.

[17] J. Lee, S. Kim, and H. Park, "Comparison of ML algorithms for health risk detection," *Healthcare Analytics*, vol. 1, p. 100003, 2020.

[18] D. Sari, R. Wulandari, and T. Nugroho, "Enhancing minority class detection using ensemble models," *International Journal of Bioinformatics and Biomedical Engineering*, vol. 10, no. 3, pp. 250–265, 2023.