Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

Comparative Analysis of Naïve Bayes and K-Nearest Neighbor (KNN) Algorithms in Stroke Classification

Ida Bagus Ary Indra Iswara[1)*], Ida Bagus Gede Anandita[2)], Maria Dahul[3)]
[1)*2)3)]Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia
[1)*] indraiswara@instiki.ac.id, [2)] ida.bagus.anandita@gmail.com, [2)] mariadhl@gmail.com

ABSTRACT

Stroke, also known as cerebrovascular, is a type of Non-Communicable Disease (NCD). The symptoms of this disease arise due to a blockage (ischemic) or rupture (hemorrhagic) of a blood vessel that disrupts blood flow to the brain. This condition causes a lack of oxygen and nutrients to brain cells, resulting in damage and potentially death. This research aims to compare the use of Naive Bayes and K-Nearest Neighbor (K-NN) algorithms in classifying stroke diseases. The research process involves data collection, data validation, data preprocessing, data reading, data transformation, data splitting, model implementation, classification evaluation, application of Naive Bayes and K-Nearest Neighbor (K-NN) algorithms, and comparative analysis of results. The variables used in this study include: gender, age, hypertension, heart disease, ever married, work type, residence type, avg glucose level, bmi, smoking status, stroke. Sugar, BMI, Smoking Status, Stroke. Based on the experiments conducted, it was found that the Naive Bayes algorithm achieved an average accuracy rate of 91.67%, while the K-Nearest Neighbor (K-NN) algorithm achieved an average accuracy rate of 95.59%. Therefore, it can be concluded that the K-Nearest Neighbor (K-NN) algorithm has a higher average accuracy rate than the Naive Bayes algorithm, with a percentage difference in accuracy of 3.92%.

Keywords: Stroke Classification; Accuracy; Naive Bayes Algorithm; K-Nearest Neighbor (KNN)

1. INTRODUCTION

Stroke is a loss of brain function that results from the cessation of supply to part of the brain. Stroke can occur due to ischemia or hemorrhage. Stroke is a disease that occurs suddenly, progressively, quickly in the form of focal or global neurological deficit that lasts 24 hours or more or immediately causes death and is solely caused by non-traumatic cerebral blood disorders (Faridah, 2019). Stroke or cerebrovascular is a non-communicable disease (NCD) characterized by blockage (ischemic) or rupture of blood vessels (hemorrhagic) due to impaired blood flow to the brain. Blockage of the blood vessels to the brain causes the death of brain cells due to lack of oxygen and nutrient intake. The biggest risk experienced by stroke patients due to blood vessel damage is death. (Rahmadani and Muzafar, 2022)

The World Health Organization (WHO) defines stroke as a sudden impairment of brain function. According to the World Health Organization (WHO), stroke deaths account for 70% of total deaths in the world. More than 36 million people lose their lives, and 9 million of them occur before the age of 60. Stroke deaths are most common in low- and middle-income countries, including Indonesia. Stroke is the third leading cause of death in Indonesia after heart disease and cancer (Rahmadani and Muzafar, 2022).

The high number of deaths due to this disease is due to people's ignorance of the disease and the symptoms of stroke, even though if they have seen the symptoms and are treated as early or as soon as possible, there is still a possibility that stroke sufferers can be treated and gradually recover. Stroke or commonly called 'struk' is a condition where blood flow to the brain is blocked due to blockage of blood vessels. This stroke can occur due to blockage of blood vessels and for this blockage there are two types, the first is a blockage of blood vessels or called Ischemic Stroke, then the second is a rupture of blood vessels or called Hemorrhagic Stroke and almost 70% of Hemorrhagic Stroke cases occur in people with Hypertension (high blood pressure) (Haris, 2022).

Stroke deaths are difficult to estimate because the clinical symptoms are unpredictable and develop very quickly. Therefore, the involvement of technology called machine learning is needed to classify stroke diseases, in this case a data processing method is used, namely data mining. Data mining is a branch of science that combines

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

fields of computer science that are utilized to find patterns and interesting information from a set of data, or referred to as the process of deciphering knowledge in a database using certain methods, such as artificial intelligence, machine learning and statistics. One method that can be applied in data mining is classification. (Saputra et al., 2021). The classification methods applied in this study are the Naïve Bayes and K-Nearest Neighbor (K-NN) algorithms. The selection of these two algorithms is related to the dataset that the author uses in this study, namely unsupervised learning or data that already has a class label or result.

Naïve Bayes classification is an algorithm in data mining techniques that uses Bayes theory to classify. Naïve Bayes has a high level of speed and accuracy when applied to a data container with large enough data. While K-Nearest Neighbor (K-NN) is a data clustering method to determine categories based on most categories in K-Nearest Neighbor, K-Nearest Neighbor (K-NN) is done by looking for groups of K objects in training data that are closest to objects in new data or test data.

Many studies have been conducted to predict stroke disease, but it is not yet known which algorithm has the most accurate accuracy. One of them is research conducted by (Ulfatul et al., 2022) with the title Comparison of K-Nearest Neighbor and Gaussian Naive Bayes Methods for Stroke Disease Classification. The classification method used in this study is to compare the K-Nearest Neighbor and Gaussian Naive Bayes algorithms. Attributes used Age, Gender, Heredity or family history, Hypertension, Hypercholesterolemia and fat, History of Diabetes Mellitus, History of Heart Disease, Transient Ischemic Attack (TIA), Smoking, Obesity, Pregnancy, Drug abuse, and Alcohol consumption. From the comparison of accuracy, precision and recall, it can be seen that there is an increase in accuracy of 6.15%, precision of 6.81% and recall of 2.37%, thus proving that the performance of the Gaussian Naive Bayes algorithm is better.

Based on the analysis above, the research objectives are expected to help make it easier to find out which algorithm has the highest level of accuracy. In addition, the authors also calculate the percentage success rate of the Naïve Bayes K-Nearest Neighbor (K-NN) method comparison using the RapidMiner application to see how high the classification accuracy of the Naïve Bayes and K-Nearest Neighbor (K-NN) methods is on this dataset. The dataset that researchers use in this study comes from open source Kaggle data totaling 5110 data. The attributes used are Gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.

2. LITERATURE REVIEW

In the realm of machine learning, the comparison between Naïve Bayes and K-Nearest Neighbor (KNN) algorithms has been a subject of interest in various studies. (Shyla & Bhatnagar, 2023) conducted an analysis of different classification algorithms, including Naïve Bayes and KNN, using the HDTbNB algorithm. Their study compared the performance of these algorithms over the KDD 99 dataset. Similarly, (Veziroğlu, 2024) compared Naïve Bayes with KNN among other classifiers for news classification, highlighting the importance of performance evaluation in algorithm selection. (Nababan et al., 2018) found that Naïve Bayes outperformed KNN in terms of classification accuracy, indicating the superiority of Naïve Bayes over KNN. Additionally, (Salsabila, 2023) reported higher accuracy rates for Naïve Bayes compared to KNN in classifying the severity levels of traffic accident victims. These findings suggest that Naïve Bayes may generally exhibit better accuracy in classification tasks compared to KNN. On the other hand, (Handayani & Ikrimach, 2020) observed a higher accuracy for KNN in diagnosing breast cancer compared to Naïve Bayes. This discrepancy in results underscores the importance of considering specific application domains and datasets when selecting between Naïve Bayes and KNN. Furthermore, (Oktafriani, 2023) highlighted the superior accuracy of KNN over other algorithms in determining credit eligibility, indicating the effectiveness of KNN in certain classification tasks. In conclusion, while Naïve Bayes generally demonstrates good accuracy in classification tasks, the choice between Naïve Bayes and KNN should be made based on the specific requirements of the task at hand and the characteristics of the dataset being used.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

3. METHOD

Naive Bayes

Naïve Bayes Classifier is a static classifier, this method can predict a probability. The Naïve Bayes method is a classifier method based on Bayes' theorem. One of the advantages of using the Naïve Bayes method is that it only requires a small amount of training data to predict a dataset. Naïve Bayes is a statistical calculation to predict future opportunities based on previous experience or problems encountered so it is known as Bayes' Theorem. The equation for Naïve Bayes classification is as follows:

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

Description:
H= Data hypothesis
X = Data with unknown class
P(H) = Likelihood of hypothesis H
P(X) = Likelihood of X
P(H|X)= Likelihood of hypothesis H, based on condition X
P(X|H)= Probability of X, based on the hypothesis condition H

2.4.1

K-Nearest Neighbor

K-Nearest Neighbor is one method that can be applied in classifying data, by looking for data that has the closest distance to a research object, according to the number of nearest neighbors initialized with K. The closest distance search is usually calculated using the Euclidean distance. Euclidean distance has the following equation:

$$d(x,y)= \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (2)$$

Description:
$d(x, y)$ = Euclidean distance
xi = i-th Training Data
yi = i-th Testing Data

To calculate the distance between two points in the K-Nearest Neighbor (K-NN) algorithm, the Euclidean Distance method is used which can be used in 1-dimensional space, 2-dimensional space, or multi-dimensional space. 1-dimensional space means the distance calculation only uses one independent variable, 2-dimensional space means there are two independent variables, and multi-dimensional space means there are more than two variables. In general, the Euclidean distance formula in 1-dimensional space is as follows.

$$dis\ (x_i, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \quad (3)$$

The formula above can be used if there is only one independent variable. If there is more than one, we can add them up as below.

$$dis = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + \cdots} \quad (4)$$

Research Stages

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

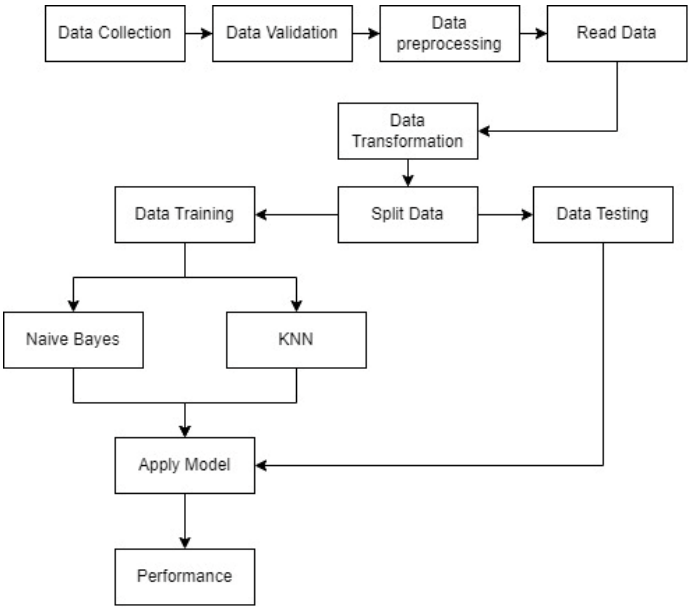

Fig.1 Research Stages

The flowchart illustrates the steps involved in the comparative analysis of Naïve Bayes and K-Nearest Neighbor (KNN) algorithms for stroke classification. The process begins with data collection, followed by data validation to ensure the quality and accuracy of the collected data. Next, the data undergoes data preprocessing, where it is cleaned and prepared for further analysis. The preprocessed data is then read into the system. Following this, data transformation is performed, and the data is split into training and testing datasets. Data training is carried out separately using the Naïve Bayes and KNN algorithms. Once trained, the models are applied to the testing data to evaluate their performance. Finally, the performance of both algorithms is compared to determine which is more effective for stroke classification. This structured approach ensures a comprehensive comparison between the two algorithms, highlighting their strengths and weaknesses in the context of stroke classification.

4. RESULT

Data Collecting

The result of this research is to find out the algorithm that has the highest stroke disease prediction accuracy. The data analyzed is data derived from open source kaggle which amounts to 4909 data. The data is analyzed using the Naive Bayes and K-Nearest Neighbor (K-NN) algorithm methods using RapidMiner assistance tools. The results of the analysis will then be compared to get the selected algorithm according to the criteria for selecting the best algorithm, namely the algorithm that has the highest accuracy.

Experiment Results with Naïve Bayes

The result of testing the model is predicting stroke disease with Naive Bayes to determine the accuracy value. In determining the value of the accuracy level in Naive Bayes by conducting several experiments using the help of the split data operator in Rapidminer 10.0. From the experimental data will be tested using the split data operator, where from 4909 data in the dataset will be formed into training data and testing data in the RapidMiner tool with the following model design:

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

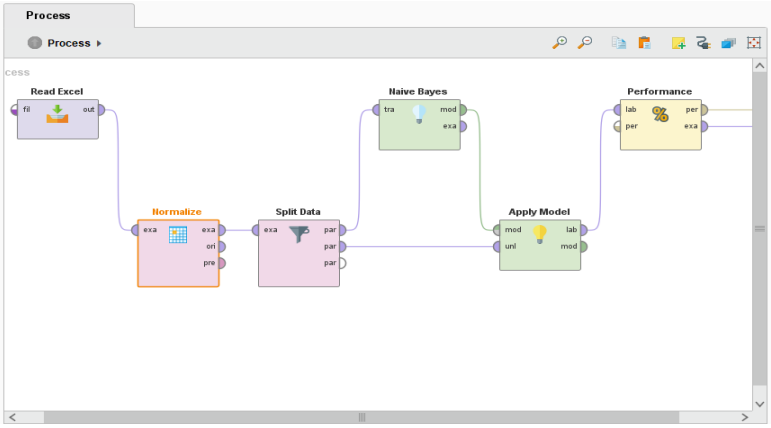

Fig.2 Naive Bayes Validation Testing Model

a. Experiment 1 For 90%:10% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the Naive Bayes model for 90% training data and 10% testing data or 491 data from a total of 4909 data. Based on result, it can be seen that true positive (TP) is all positive category data that is successfully classified or predicted positively, namely 5 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 436 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 34 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 16 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 89.82%.

b. Experiment 2 For 80%:20% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the Naive Bayes model for 80% training data and 20% testing data or 982 data from a total of 4909 data. Based on result, it can be seen that true positive (TP) is all positive category data that is successfully classified or predicted positively, namely 8 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 891 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 44 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 39 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 91.55%.

c. Experiment 3 For 70%:30% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the Naive Bayes model for 70% training data and 30% testing data or 1473 data from a total of 4909 data. Based on result, it can be seen that true positive (TP) is all positive category data that is successfully classified or predicted positively, namely 10 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 1342 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 65 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 56 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 91.79%.

d. Experiment 4 For 60%:40% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the Naive Bayes model for 60% training data and 40% testing data or 1964 data from a total of 4909 data. Based on result, it can be seen that true positive (TP) is all positive category data that is successfully classified or predicted positively, namely 15 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 1808 data. As for false positive (FP), which means

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

that all data that is categorized as negative but classified or predicted as positive, is 72 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 69 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 92.82%.

e. Experiment 5 For 50%:50% ratio

The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the Naive Bayes model for 50% training data and 50% testing data or 2454 data from a total of 4909 data. Based on result, it can be seen that true positive (TP) is all positive category data that is successfully classified or predicted positively, namely 24 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 2243 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 102 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 85 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 92.38%.

From all the experimental results of *training* data and *testing* data using *Naïve Bayes, the* following table is produced:

Table 1
Naive Bayes Experiment Results

Split Data Experiment	accuracy
Experiment 1 (90% Training and 10% Testing)	89.82%
Experiment 2 (80% Training and 20% Testing)	91.55%
Experiment 3 (70% Training and 30% Testing)	91.79%
Experiment 4 (60% Training and 40% Testing)	92.82%
Experiment 5 (50% Training and 50% Testing)	92.38%
Average	**91.67%**

Experiment Results with K-NN

The result of testing the model is predicting stroke disease with K-Nearest Neighbor (K-NN) to determine the accuracy value. In determining the value of the accuracy level in K-Nearest Neighbor (K-NN) by conducting several experiments using the help of the split data operator in Rapidminer 10.0. From the experimental data will be tested using the split data operator, where from 4909 data in the dataset will be formed into training data and testing data on RapidMiner tools. For the determination of the K value, we will use the K value with the highest accuracy value, namely K = 12.

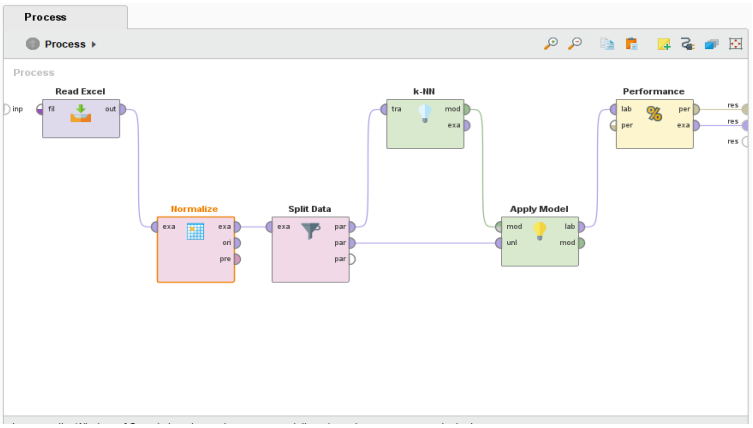

Fig.3 K-NN Validation Testing Model

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

a. Experiment 1. For 90%:10% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the K-Nearest Neighbor (KNN) model for 90% training data and 10% testing data or 491 data from a total of 4909 data. Based on result, known from 4909 stroke patient data, 90% is used as training data and 10% as testing data or 491 data. Seen true positive (TP) is all positive category data that is successfully classified or predicted positive, namely 0 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 470 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 0 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 21 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 95.72%.

b. Experiment 2 For 80%:20% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the K-Nearest Neighbor (K-NN) model for 80% training data and 20% testing data from a total of 4909 data. Based on result, known from 4909 stroke patient data, 80% is used as training data and 20% as testing data or 982 data. seen true positive (TP) is all positive category data that is successfully classified or predicted positive, namely 1 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 935 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 0 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 46 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 95.32%.

c. Experiment 3 For 70%:30% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the K-Nearest Neighbor (K-NN) model for 70% training data and 30% testing data from a total of 4909 data. Based on result, it shows that, known from 4909 stroke patient data, 70% is used as training data and 30% as testing data or 1473 data. Seen true positive (TP) is all positive category data that is successfully classified or predicted positive, namely 1 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 1407 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 0 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 65 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 95.59%.

d. Experiment 4 For 60%:40% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the K-Nearest Neighbor (K-NN) model for 60% training data and 40% testing data from a total of 4909 data. Based on result, known from 4909 stroke patient data, 60% is used as training data and 40% as testing data or 1964 data. Seen true positive (TP) is all positive category data that is successfully classified or predicted positive, namely 1 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 1880 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 0 data. While false negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 83 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 95.77%.

e. For 50%:50% ratio
The accuracy value of the split data process is calculated using RapidMiner. Confusion Matrix test results using the K-Nearest Neighbor (K-NN) model for 50% training data and 50% testing data from a total of 4909 data. Based on result, it shows that, known from 4909 stroke patient data, 50% is used as training data and 50% as testing data or 2454 data. Seen true positive (TP) is all positive category data that is successfully classified or predicted positive, namely 0 data. True negative (TN) is all negative category data that is successfully classified or predicted negatively, namely 2345 data. As for false positive (FP), which means that all data that is categorized as negative but classified or predicted as positive, is 0 data. While false

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

negative (FN) which means all data that is categorized as positive but classified or predicted as negative, here the data classified as false negative is 109 data. Accuracy is the amount of all data that is successfully classified correctly, both positive and negative data divided by the total amount of data, the result is 95.56%.

From all experimental results from training data and testing data using K-Nearest Neighbor (K-NN), the following table is produced:

Table 2
K-NN Experiment Results

Split Data Experiment	accuracy
Experiment 1 (90% Training and 10% Testing)	94,72%
Experiment 2 (80% Training and 20% Testing)	95,32%
Experiment 3 (70% Training and 30% Testing)	95,59%
Experiment 4 (60% Training and 40% Testing)	95,77%
Experiment 5 (50% Training and 50% Testing)	95,56%
Average	**95,59%**

Comparative Analysis of Results

The population rate of stroke is increasing. This shows that there are problems in diagnosing stroke disease. For this reason, in analyzing accurate stroke disease, the most appropriate algorithm method is needed. In this study the authors used two algorithms, namely Naïve Bayes and K-Nearest Neighbor (K-NN). Naive Bayes is used because it is known from previous research that Naive Bayes and K-Nearest Neighbor (K-NN) have the ability to analyze the classification of data. Based on the results of experiments that have been carried out to solve the problem of predicting stroke disease prediction results, it can be concluded that the results of experiments using the Naive Bayes method have an average accuracy rate of **91.67%**, while using the K-Nearest Neighbor (K-NN) algorithm in diagnosing stroke disease produces an average accuracy rate of **95.59%**.

From all the experimental results of training data and testing data using naïve Bayes and K-Nearest Neighbor (K-NN), the following table is produced:

Table 3
Comparative Analysis of Results

No.	Experiment	Accuracy	
		Naïve Bayes	K-NN
1	Experiment 1	89.82%	95,72%
2	Experiment 2	91.55%	95,32%
3	Experiment 3	91.79%	95,59%
4	Experiment 4	92.82%	95,77%
5	Experiment 5	92.38%	95,56%
6	**Average**	**91.67%**	**95,59%**

Visualization of Analysis Results

Visualization of Results The analysis of data testing of stroke disease classification results will be visualized in the form of a bar chart. Based on the results of experiments that have been carried out using the Naïve Bayes and K-Nearest Neighbor (K-NN) classification algorithms, the results of stroke disease classification accuracy with the division of training data and testing data 90%: 10%, 80%: 20%, 70%: 30%, 60%: 40%, 50%: 50% will be depicted in the following graph.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

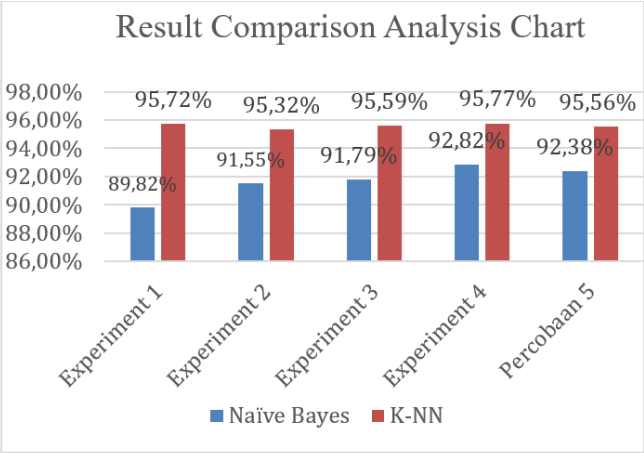

Fig.4 Comparative Analysis Chart of Results

From the results of Figure 4.13 above with a dataset taken from the open source kaggle, it gives the result that the K-Nearest Neighbor (K-NN) algorithm has a greater average accuracy value than the Naive Bayes algorithm with a difference in accuracy percentage of **3.92%**. There are several possibilities that affect this, such as the number of datasets used or the number of attributes used will affect the accuracy rate.

5. CONCLUSION

Based on the research findings, it can be concluded that out of a total of 491 available comment data, 314 (63.95%) were labeled as positive, while 177 comments (36.05%) were labeled as negative. The dataset was split with a ratio of 75:25%, where 271 data were used for training the model and 120 data for testing the model. The model evaluation showed an accuracy of 72.5%, indicating that the Naive Bayes algorithm effectively predicted and classified the data accurately. The study demonstrated the effectiveness of the Naive Bayes algorithm in designing sentiment analysis models for the presence of Coldplay in Indonesia using Twitter. As a further suggestion, it is recommended to consider using other classification methods such as Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Decision Tree, or other classification methods to compare accuracy with the current algorithm.

6. REFERENCES

Fatmawati, K., dan Windarto, A. P. 2018. "Data Mining : Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi", *3*(2), 173–178.

Haris 2022. "Metode Naïve Bayes Untuk Memprediksi Penyakit Stroke".

Maulid, R. 2021. "Kursus Belajar Data: Mengenal Apa Itu Missing Value". diambil 6 Maret 2023, dari https://www.dqlab.id/kursus-belajar-data-mengenal-apa-itu-missing-value.

Mutiarasari, D. 2019. "Ischemic Stroke: Symptoms, Risk Factors, And Prevention", *6*(1).

Nugroho, K. S. 2020. "Menerapkan Model Klasifikasi Machine Learning pada RapidMiner". diambil 16 Maret 2023, dari https://ksnugroho.medium.com/menerapkan-model-machine-learning-pada-rapidminer-142259846e13.

Pambudi, R. E. S. F. 2022. "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45", *16*(x), 221–226.

Putri, R. W., Ristyawan, A., dkk. 2018. "Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease".

Rahmadani, D., dan Muzafar, A. A. 2022. "Comparative Analysis of C4 . 5 and CART Algorithms for Classification of Stroke", 198–206.

Rerung, R. R. 2018. "Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4395

Submitted : 25 July 2024
Accepted : 30 July 2024
Published : 31 July 2024

Produk", *3*(1), 89–98. https://doi.org/10.31544/jtera.v3.i1.2018.89-98.

Rezkia, S. M. 2020. "Tingkatkan Kompetensi dengan Mengulik Sumber Dataset Untuk Membangun Model Pada Data Science". diambil 11 Februari 2023, dari https://www.dqlab.id/data-scientist-mengenal-dataset-datascience.

Saputra, D., Irmayani, W., dkk. 2021. "A Comparative Analysis of C4 . 5 Classification Algorithm , Naïve Bayes and Support Vector Machine Based on Particle Swarm Optimization (PSO) for Heart Disease Prediction", *2*(2), 84–95. https://doi.org/10.25008/ijadis.v2i2.1221.

Sari, M., dan Ikhwani, Y. 2018. "Komparasi Algoritma K-Nearest Neighbor Dan Naive Baiyes Untuk Mendeteksi Dini Resiko Kanker Serviks Pada", *2*(2).

Ulfatul, D., Rachmad, M., dkk. 2022. "Jurnal Smart Teknologi Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke", *3*(4), 405–412.

Utomo, D. P. 2020. "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung", *4*(April), 437–444. https://doi.org/10.30865/mib.v4i2.2080.

Handayani, I., & Ikrimach, I. (2020). Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer. *Jurnal Infotel*, *12*(4), 151–159. https://doi.org/10.20895/infotel.v12i4.547

Nababan, A. A., Sitompul, O. S., & Tulus. (2018). Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. *Journal of Physics Conference Series*, *1007*, 12007. https://doi.org/10.1088/1742-6596/1007/1/012007

Oktafriani, Y. (2023). Analysis of Data Mining Applications for Determining Credit Eligibility Using Classification Algorithms C4.5, Naïve Bayes, K-Nn, and Random Forest. *Asian Journal of Social and Humanities*, *1*(12), 1139–1158. https://doi.org/10.59888/ajosh.v1i12.119

Salsabila, N. A. (2023). Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas Di Kota Samarinda Menggunakan Algoritma K-Nearest Neighbor Dan Naive Bayes. *Eksponensial*, *14*(2), 99. https://doi.org/10.30872/eksponensial.v14i2.1085

Shyla, & Bhatnagar, V. (2023). Perspicacious Apprehension of HDTbNB Algorithm Opposed to Security Contravention. *Intelligent Automation & Soft Computing*, *35*(2), 2431–2447. https://doi.org/10.32604/iasc.2023.029126

Veziroğlu, M. (2024). *Performance Comparison Between Naive Bayes and Machine Learning Algorithms for News Classification*. https://doi.org/10.5772/intechopen.1002778

* Corresponding author