



Comparison of K-Nearest Neighbor Algorithm Performance and Naïve Bayes in Predicting Stroke Disease

Abdul Roni¹, Maria Fransiska Fitriani^{2*}, Nazwa Aurellia Ainanur³, Sumanto⁴, Ade Surya Budiman⁵

^{1,2,3,4,5} Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia
1522086@bsi.ac.id¹, 15220815@bsi.ac.id^{2*}, 15220785@bsi.ac.id³, sumanto@bsi.ac.id⁴, ade.aum@bsi.ac.id⁵

Abstract

Stroke is one of the most dangerous diseases that can cause death and long-term disability. Early identification of *stroke* risk can help the prevention process. This study compares two classification algorithms, namely *K-Nearest Neighbor (K-NN)* and *Naïve Bayes*, in *prediction stroke* risk based on patient data. The dataset used is 1470 data that has several attributes such as age, hypertension, heart disease, glucose levels, and others. The evaluation results showed that *Naïve Bayes* algorithm performed better with 73.1% accuracy and 79.9% AUC, compared to *K-NN* which had 68.4% accuracy and 75.1% AUC. Based on these results, *Naïve Bayes* algorithm is considered more effective to be used in *stroke* risk prediction system.

Keywords: *Stroke, Prediction, K-Nearest Neighbor, Naïve Bayes, Data Mining*

1. Introduction

Cerebrovascular disease or better known as (*Stroke*) is the No.2 leading cause of death and the No.3 leading cause of disability worldwide [1]. Impaired nerve function in *stroke* is caused by impaired blood flow in the brain which can cause neurological disorders that are manifested in the form of paralysis of extremity muscles, weakness of swallowing muscles, weakness in verbal communication, visual disturbances, impaired consciousness and can even cause death [2].

Stroke is also a major cause of functional impairment, with 20% of survivors requiring institutional care after three months and 15-30% becoming permanently disabled [3]. Various risk factors have been identified to contribute to the incidence of *stroke*, including age, gender, hypertension, heart disease, obesity, smoking, and lack of physical activity. The good news is that *stroke* can be prevented through the implementation of healthy living behaviors such as exercising regularly, avoiding high-cholesterol foods, and not smoking [1].

Rapid and appropriate treatment helps *stroke* patients recover from the disease, while late treatment can result in long-term disability, prolonged brain damage, and even death [3]. Therefore, early identification of individuals at high risk of *stroke* is crucial to enable timely prevention efforts. In this context, machine learning offers great potential for developing accurate prediction models. Its ability to process complex data and identify hidden patterns can provide new insights into the likelihood of *stroke* occurrence.

Thus, this study aims to compare the *K-NN* and *Naïve Bayes* algorithms in predicting the likelihood of a person having a *stroke* based on historical patient data. Through the evaluation of accuracy, precision, recall, and f1-score of each algorithm, it is expected to know which method is more effective and efficient in the process of *stroke* disease classification. The results of this study are expected to be the basis for the development of decision support systems in the health sector, especially in efforts to prevent and treat early *stroke* disease.

2. Literature

2.1. Previous Research

In the process of preparing this research, the author refers to several previous studies that have links to the topics discussed. These studies are very helpful for the author to add insight, deepen understanding, and strengthen the theoretical basis used in this study. In addition, references from previous research also serve as a comparison and foundation in analyzing and developing the research being conducted. Therefore, the author summarizes several journals that are closely related to the topic of this research, which are then used as the main reference in preparing the study.

Research in 2024 by Baiq Andriska Candra Permana, Muhammad Sadali, Ramli Ahmad entitled “Application of Decision Tree Models Using Python for *Prediction* of Dominant Factors Causing *Stroke* Disease” where in this study obtained the results that the decision tree model has good accuracy in predicting the dominant factor causing *stroke*, namely with an accuracy rate of 91% and an AUC value of 0.5, in this study also obtained the most influential factor as the main cause of *stroke* is age which is triggered by several other factors such as sugar levels, weight and hypertension [4].

Research in 2023 by Nabilla Yolanda Paramitha, Aang Nuryaman, Ahmad Faisol, Eri Setiawan, and Dina Eka Nurvazly entitled “Classification of *Stroke* Disease Using the *Naïve Bayes* Method” which obtained the results of model evaluation calculations using confusion matrix on *stroke* disease data with a proportion of data sharing 80:20 resulting in higher accuracy than other proportions by. The proportion of 80:20 indicates that the best model is obtained. In addition, *determining* the proportion of division of training data and testing data can affect the results of testing the percentage of accuracy, because the training data pattern will be used as a reference in determining the class in the testing data [5].

Research in 2022 by Nur Aliffiyanti Iskandar, Iin Ernawati, Yuni Widiatiwi entitled “Classification of *Stroke* Disease Diagnosis Using the Random Forest Method” from this study produced optimal values, where the resulting accuracy value was 95.2%, sensitivity of 4.1%, specificity of 99.8%, precision of 66.7%, and F-measure of 7.6%. As well as the ROC Curve of 0.8048 which indicates that the model is included in the Good Classification [6].

2.2. Stroke

Stroke is a syndrome caused by cerebral circulatory disorders accompanied by clinical manifestations in the form of neurological deficits and not as a result of tumors, trauma or central nervous system infections (Dewanto, 2009) [7].

There are 4 risk factors that can cause *stroke*, namely:

- 1) Hypertension
A person is said to have hypertension when their blood pressure is more than 140/90 mmHg, or more than 135/85 mmHg in individuals with heart failure, renal insufficiency, or diabetes mellitus.
- 2) Diabetes
condition when a person's blood sugar level is at a very high level, which is equal to or more than 200 mg/dL.
- 3) Smoking
Smoking promotes increased blood viscosity, hardening of the blood vessel walls, and plaque buildup in the blood vessel walls. Smoking increases the risk of *stroke* up to 2 times.
- 4) Obesity
Research by Oki, et al (2006) concluded that a person with a body mass index ≥ 30 has a risk of *stroke* 2.46 times compared to those with a body mass index < 30 [8].

2.3. Classification

Classification is a grouping of data where the data used has a label or target class. Thus, algorithms for solving classification problems are categorized into supervised learning. Classification is a technique used to find models in order to explain or distinguish concepts or classes of data, with the aim of being able to estimate the class of an object whose label is unknown [9].

2.4. Prediction

Prediction is the process of estimating something that is most likely to be proven by comparing information owned in the past with information that is owned now with the aim that the error difference between something that happens and the estimated results can be minimized [10].

3. Research Methods

3.1. Research Stages



Fig. 1: Illustrates a series of stages that will be carried out in this research

The following is a discussion of the knowledge discovery stage in the database:

- 1 Data Collection: The selected data is the data of a person affected by *stroke* disease.
- 2 Preprocessing: Check the data by seeing if there is missing value data.
- 3 Transformation: Transform the selected data to make it suitable for the data mining process.
- 4 Proposed method: Perform data clustering using the *K-Nearest Neighbor* and *Naïve Bayes* algorithms in the *Orange Data Mining* application to aid analysis.
- 5 Evaluation: Evaluate the clustering results to assess the relevance and accuracy of the clustering results.

3.2. Data Source

In this study, 1470 relevant data were used. This data included attributes such as ID, gender, age, history of hypertension, heart disease, marital status, type of employment, type of residence, average glucose level, body mass index (BMI), and smoking status, all of which were used as explanatory and clustering information for the *K-Nearest Neighbor* and *Naïve Bayes* algorithms.

3.3. Classification Method

In this research, we will perform classification methods using the *K-Nearest Neighbor* and *Naïve Bayes* algorithms in *Orange Data Mining*.

1. Orange Data Mining
Orange *data mining* is an open source application that is able to assist researchers in analyzing data [11]
2. K-Nearest Neighbor
K-Nearest Neighbor is an efficient classification method in recognizing patterns, categorizing text, processing objects and can do large amounts of data training (Ulfatuldkk., 2022) [12].
3. Naïve Bayes
Naïve Bayes is a classification using probabilistic and statistical methods proposed by British scientist Thomas Bayes. *Naïve Bayes* classification assumes that the presence or absence of certain features of one class has nothing to do with the features of another class [10].
4. Confusion Matrix
Confusion Matrix is a tool to assess the performance of classification models in machine learning. This matrix is presented in the form of a table that compares model predictions with actual data [13].
5. Scatter Plot
Scatter plot is a graph that shows the relationship between two numerical variables. Each point represents one observation in the data, and its position reflects the value of the two variables. It helps us identify patterns of correlation or relationship between the variables[14].

4. Result and Discussion

4.1. Data Selection

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence	avg_glucose_level	bmi	smoking_status
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked
51676	Female	61	0	0	Yes	Self-empl	Rural	202.21	33	never smoked
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes
1665	Female	79	1	0	Yes	Self-empl	Rural	174.12	24	never smoked
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked
27419	Female	59	0	0	Yes	Private	Rural	76.15	38	smokes
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	never smoked

Fig. 2: Dataset View

In the early stages of the research, data selection involved the use of an excel read operator. The function of this operator is to read the information related to the data grouping of a person affected by *stroke* disease and stored in an MS-Excel file.

4.2. Preprocessing Data

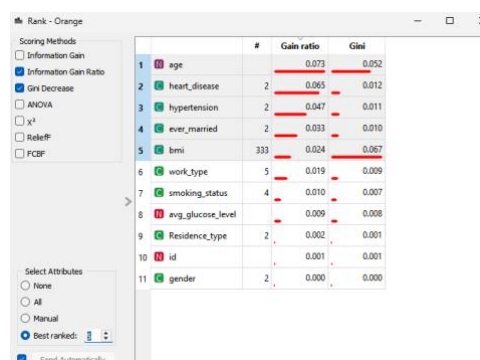


Fig. 3: Contents of the rank widget on orange data mining

Then after the dataset is entered into orange *data mining*, we will continue checking the data, whether the data is missing value or not.

Because of the data there is no missing value during the preprocessing stage, then we can immediately proceed to the next step, namely by selecting attributes. This selection will be based on rank considerations that exist for data processing.

4.3. Transformation Data

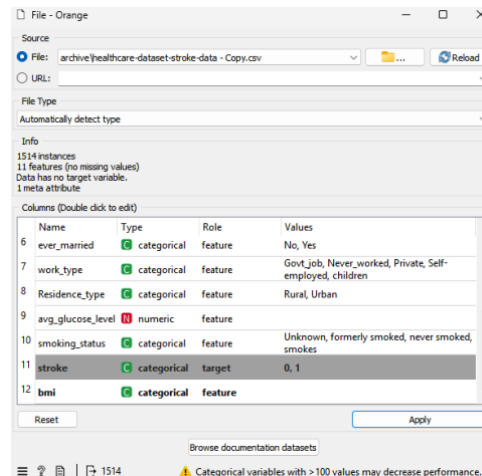


Fig. 4: Display when the role has been changed

Data that has gone through preprocessing will then be transformed by changing the role of the variables, in order to adjust the data format to the requirements of the algorithm to be used.

4.4. Proposed Method

		Predicted		Σ
		0	1	
Actual	0	617	219	836
	1	176	458	634
Σ		793	677	1470

Fig. 1 : Confusion matrix Naïve Bayes

Presents the confusion matrix of the *Naïve Bayes* model prediction results, which is used to evaluate the classification performance against two classes, namely classes 0 (negative) and 1 (positive). This matrix shows four important values that underlie the various metrics for evaluating model performance:

1. True Negative (TN): 617
2. False Positive (FP): 219
3. False Negative (FN): 176
4. True Positive (TP): 458

Based on the matrix, out of a total of 836 actual data with label 0, 617 were correctly predicted (TN), and 219 were wrongly classified as class 1 (FP). Meanwhile, of the 634 actual data with label 1, 458 were correctly classified (TP), and 176 were incorrectly classified as class 0 (FN).

From these results it can be concluded that the *Naïve Bayes* model has a fairly good classification ability, with an accuracy rate of 73.1%. Although there are a number of false positives (219) and false negatives (176), the high true positive and true negative values indicate that the model is quite capable of distinguishing between the two classes well.

However, the relatively large false positive value (219 cases) indicates that the model has a tendency to over-predict the positive class, i.e. classifying negative data as positive. This should be a concern if the application context is detection that requires high sensitivity to false positives, for example in the field of disease diagnosis.

Confusion matrix shows that *Naïve Bayes* algorithm is not only statistically superior in global metrics (AUC, F1, MCC), but also provides a fairly balanced and reliable prediction distribution. However, there is still room for improvement, especially in reducing the number of false positives, through setting thresholds, selecting more appropriate features, or combining models (*ensemble*).

		Predicted		Σ
		0	1	
Actual	0	700	136	836
	1	328	306	634
Σ		1028	442	1470

Fig. 2: Confusion matrix KNN

Displays the classification prediction results of the *K Nearest Neighbor (KNN)* algorithm in the form of a confusion matrix for two classes, namely class 0 (negative) and class 1 (positive). This matrix shows the number of correct and incorrect predictions based on the comparison between the actual label and the label predicted by the model.

The values in the confusion matrix are as follows:

- True Negative (TN) = 700
- False Positive (FP) = 136
- False Negative (FN) = 328
- True Positive (TP) = 306

It can be concluded that although *KNN* has a fairly good prediction ability in the negative class (TN = 700), the model's ability to recognize the positive class is still limited, as seen from the high number of False Negative (328) and the low Recall value (0.483). This means that the model often fails to recognize data that should belong to the positive class, which is risky when the model is used for important classification tasks such as fraud diagnosis or detection.

The Precision value (0.692) indicates that if the model predicts a data as a positive class, the prediction is likely to be correct. However, the low Recall value indicates that many positive data were not successfully recognized, which will directly impact the F1 Score value (0.566), which measures the balance between Precision and Recall.

With its total accuracy of 68.4%, the *KNN* model is still generally acceptable. However, compared to the *Naïve Bayes* model (whose accuracy reached 73.1%), the overall performance of *KNN* was lower, especially in classifying the positive class.

The *KNN* confusion matrix indicates that although the model is able to classify negative classes well (high true negative), it is less optimal in detecting positive classes, as seen from the high false negative. Therefore, *KNN* is less recommended if the application context requires a high level of sensitivity to minority or positive classes, for example in the case of early detection or risk classification.

Compared to the *Naïve Bayes* model, *KNN* tends to produce classifications that are more conservative towards positive classes, which can reduce the risk of false positives, but at the cost of increasing false negatives. This model is more suitable when the main priority is to minimize the error in predicting negative classes.

4.5. Evaluation

Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.799	0.731	0.732	0.734	0.731	0.457
kNN	0.751	0.684	0.672	0.686	0.684	0.346

Fig. 7: Results in the test and score widget

The results evaluate the performance of two classification algorithms, *Naive Bayes* and *k-Nearest Neighbor (kNN)*, based on tests using the Test and Score widget. The evaluation was conducted by measuring six key performance metrics: Area Under Curve (AUC), Classification Accuracy (CA), F1-Score (F1), Precision (Prec), Recall, and Matthews Correlation Coefficient (MCC). These six metrics provide a comprehensive overview of the accuracy and stability of the model in classifying the data.

Based on the evaluation results, the *Naive Bayes* algorithm performed better than *kNN* on all evaluation metrics. The AUC value for *Naive Bayes* is 0.799, while *kNN* only reaches 0.751. This shows that *Naive Bayes* has a higher ability to distinguish between positive and negative classes. Furthermore, in the Classification Accuracy metric, *Naive Bayes* obtained a value of 0.731, higher than *kNN* which was only 0.684.

In the F1-Score aspect which measures the balance between precision and recall, *Naive Bayes* obtained a value of 0.732, while *kNN* only reached 0.672. This shows that *Naive Bayes* is more stable in handling data that may be unbalanced or contain complex class distributions. The precision and recall values also show the superiority of *Naive Bayes* (0.734 and 0.731) over *kNN* (0.686 and 0.684), respectively.

Last but not least, the MCC (Matthews Correlation Coefficient) metric, which is often used as a performance indicator on unbalanced datasets, shows that *Naive Bayes* excels with a value of 0.457, while *kNN* only obtains 0.346. MCC takes into account all aspects of the confusion matrix (true/false positives and negatives), thus providing a fairer assessment than pure accuracy, especially when there is class imbalance.

Overall, these results indicate that *Naive Bayes* is more suitable for the dataset used in this study, both in terms of prediction accuracy and generalization ability. The superiority is most likely due to the suitability of the conditional independence assumption in *Naive Bayes* with the characteristics of the features in the dataset, as well as its computational efficiency compared to *kNN* which is based on distance calculation.

5. Conclusion

Based on the results of research that has been conducted on the comparison of the *K-Nearest Neighbor (KNN)* and *Naive Bayes* algorithms in *stroke* disease classification, it can be concluded that the *Naive Bayes* algorithm shows more optimal performance than *KNN*. This is shown from the performance evaluation results which show that *Naive Bayes* obtained an AUC value of 79.9% and an accuracy of 73.1%, higher than *KNN* which has an AUC of 75.1% and an accuracy of 68.4%. In addition, *Naive Bayes* also has better recall and F1-score values, which means it is more effective in detecting patients with *stroke* risk. These results prove that *Naive Bayes* is more effective in the classification process of *stroke* disease in the dataset used, especially in terms of detection ability of positive data (*stroke*). This finding is expected to be the basis for the development of a decision support system for early detection of *stroke*, which in turn can help medical authorities in the process of prevention and treatment more quickly.

References

- [1] S. Mutmainah, "Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 1, no. 1, pp. 10–16, 2021, doi: 10.20885/snati.v1i1.2.
- [2] T. TUNIK, "Faktor-Faktor Penyebab Dan Pencegahan Terjadinya Stroke Berulang," *Heal. J. Inov. Ris. Ilmu Kesehatan.*, vol. 1, no. 2, pp. 101–108, 2022, doi: 10.51878/healthy.v1i2.1114.
- [3] A. Rohman and M. Rochcham, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Mahasiswa," *Neo Tek.*, vol. 5, no. 1, pp. 73–79, 2019, doi: 10.37760/neoteknika.v5i1.1379.
- [4] B. A. C. Permana, M. Sadali, and R. Ahmad, "Penerapan Model Decision Tree Menggunakan Python Untuk Prediksi Faktor Dominan Penyebab Penyakit Stroke," *Infotek J. Inform. dan Teknol.*, vol. 7, no. 1, pp. 23–31, 2024, doi: 10.29408/jit.v7i1.23232.
- [5] N. Y. Paramitha, A. Nuryaman, A. Faisol, E. Setiawan, and E. Nurvazly, "Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes," vol. 04, no. 01, pp. 11–16, 2023.
- [6] N. A. Iskandar, I. Ernawati, and Y. Widiastiwi, "Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest," pp. 432–441, 2022.
- [7] S. P. Keluarga, *Buku Pencegahan Stroke dan Penatalaksanaan Pre Hospital*.
- [8] *Awas Stroke*.
- [9] P. Yayasan and K. Menulis, *Pengenalan Data Mining*.
- [10] L. Rahmawati, M. Hafid, and M. A. Sunandar, "Analisis Data Mining Untuk Memprediksi Penyakit Stroke Dengan Algoritma Naive Bayes," vol. 6, no. 2, pp. 55–60, 2023.
- [11] N. Ichsan *et al.*, "IMPLEMENTASI ORANGE DATA MINING UNTUK," vol. 4, no. 2, pp. 118–125, 2022.
- [12] M. M. Jakarta, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Stroke pada Rumah Sakit Pusat Otak Nasional Prof .," vol. 26, no. 1, pp. 144–153.
- [13] M. A. Salwa, "Optimasi Model Algoritma Klasifikasi Data Mining Menggunakan Metode Feature Selection Untuk Prediksi Penyakit Stroke," vol. 26, no. 1, pp. 11–20, 2025.
- [14] F. Y. Febrieta Ditta, *Statistika Dasar untuk Pemula*. 2023.