UMT PRESS

Journal of Mathematical Sciences and Informatics
Journal Homepage: https://journal.umt.edu.my/index.php/jmsi
eISSN: 2948-3697
DOI: http://doi.org/10.46754/jmsi.2025.06.007

JMSI
Journal of Mathematical Sciences and Informatics

CLASSIFICATION FOR DIAGNOSING STROKE USING ORANGE DATA MINING

WARTIKA* AND AGUS NURSIKUWAGUS

Faculty of Engineering and Computer Science, Indonesian Computer University, Bandung, West Java, Indonesia.

**Corresponding author: wartika@email.unikom.ac.id*

ARTICLE INFO

Article History:
Received 4 June 2024
Revised 15 April 2025
Accepted 15 April 2025
Published 15 June 2025

Keywords:
Classification;
K-Nearest Neighbor;
naive bayes
neural network;
orange;
stroke.

ABSTRACT

Stroke is a disease with the highest mortality rate among people aged over 45 years in Indonesia. The problem with stroke in Indonesia requires very serious attention because the number of cases continues to increase, and the death rate is very high. The treatment needed is to maintain health and also detect strokes early. This research aims to build a classification model that can predict whether someone is at risk of having a stroke based on available clinical data. Factors that influence the diagnosis of stroke are needed. Based on the data obtained, several factors are used as sources of analysis such as BMI (Body Mass Index), hypertension, heart disease, glucose level, smoker or not, age, gender, type of work, and type of residence that can be used classified as input variables and output variables. Data mining can be used to classify whether a patient has had a stroke. This research aims to apply the orange data mining application using the K-Nearest Neighbor (K-NN), Naive Bayes, and Neural Network models. Next, the patient data will be analysed using the Orange Data Mining application with K-NN, Naive Bayes and Neural Network models. This research contribution can be used by health services to detect stroke early so that it is known earlier.

2020 Mathematics Subject Classification:

Introduction

Stroke is a condition when the blood supply to the brain is disrupted due to a blockage (ischemic stroke) or rupture of a blood vessel (hemorrhagic stroke). This condition causes certain areas of the brain not to receive a supply of oxygen and nutrients, resulting in the death of brain cells [1].

Strokes are increasing in Indonesia and are a burden for society and the country. WHO estimates that the number of stroke patients in several European countries will increase from 1.1 million per year in 2000 to 1.5 million per year in 2025 [1]. According to the 2007 Riskesdas report, stroke is the main cause of death in Indonesia compared to other diseases, namely 15.4% which increased to 12.1% in 2013 [1].

Early detection of stroke usually takes a long time. With advances in technology, stroke can be prevented by detecting the risk early, so that it can be treated quickly and increase the chances of recovering from a stroke. Another advantage of having fast detection is efficiency in treatment costs, and more people can diagnose the risk quickly [2].

Choosing the right method for detecting the level of risk of stroke is very necessary because it influences the results that will be displayed. However, the data taken is in the form of mixed data, namely numerical and categorical data. So, this research can utilise a combination of the K-Nearest Neighbor and Naive Bayes methods.

The K-Nearest Neighbor algorithm, abbreviated as KNN, can be applied in classifying objects based on learning data that has small difference values and the distance of the nearest neighbour to the object.

Based on the background above, this research aims to classify stroke accuracy by analysing several activities carried out by patients. Next, it will be analysed to ensure the level of accuracy using the K-Nearest Neighbour and Naive Bayes methods because the data obtained uses numerical and categorical attributes. In this application, there are characteristic symptoms of stroke, such as high blood pressure, diabetes, cholesterol, and smoking. So, it is hoped that when this method is applied, it can help determine the level of risk of stroke so that it can be treated quickly.

Literature Review

Definition of Stroke

Stroke can be defined as a disturbance in the function of the nervous system that occurs suddenly and is caused by disturbances in cerebral blood circulation. Stroke occurs due to blood vessel disorders in the brain, which can be in the form of blocked blood vessels or rupture of blood vessels in the brain.

Stroke symptoms often appear without warning and occur suddenly. There are several main symptoms including Headache, accompanied by vomiting and loss of consciousness, difficulty walking, including symptoms of dizziness and reduced coordination, visual disturbances (blurred eyes), numbness in several parts of the body, especially the face, difficulty speaking and understanding conversations [5].

Data Mining

The term data mining has several views, such as knowledge discovery or pattern recognition, both of which actually have accuracy in accordance with their role. The term knowledge discovery is appropriate because the main goal of data mining is to obtain knowledge that is still hidden in pieces of data. The term pattern recognition is also applicable to use because there is knowledge that needs to be learned.

Data mining or information mining is a software tool that is used to discover hidden patterns, trends or rules that exist on a large-dimensional basis and create rules that are used to predict future behaviour. Classification of information is very important in creating a set of rules called rules, which will be used as markers to predict the class of information that wants to be predicted [6].

The process stages in using data mining are one of a series of KDD (knowledge discovery in the database), which is related to the integration and visualisation techniques of patterns in a number of data sets.

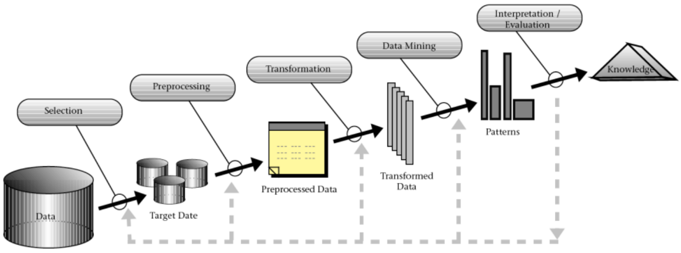


Figure 1: Data mining stages [7]

This series of processes has stages, including:

Data cleaning is to remove inconsistent data noise where fragmented data sources can be combined. Data Selection, namely where data that is relevant to the analysis task is returned to the database. Data transformation, namely when data is changed or collected into the correct form for mining with aggregate performance or operational actions, data mining is an essential process where intelligent methods are used to extract data patterns used, pattern evaluation, after completing the data mining process, the patterns resulting from the process need to be evaluated. The purpose of the assessment is to test the initial hypothesis. After testing, the data can be presented to users [7].

K-Nearest Neighbour (K-NN)

K-Nearest Neighbor (K-NN) is a supervised procedure, which requires training information to classify objects that neighbours will evaluate. The K-NN algorithm can be called “Lazy learner”, where K-NN applies “Lazy learn” or “Instant Based Learn” which means the algorithm does not need a training process and building a model. K-NN is often used for Classification and Regression cases, however, K-NN is more often used in the Classification process [8].

The simplest data mining technique, one of which is KNN. This is usually called memory-based classification; for example, the training data needs to be stored in memory at run-time.

The KNN algorithm also aims to classify based on attributes and training data. The KNN algorithm uses the distance between neighbours and neighbours as a prediction value. The KNN algorithm formula is defined as follows:

Definition 2.1. The KNN algorithm formula

$$dis\ (x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2} \tag{1}$$

Naive Bayes

Naive Bayes is a statistical classification that is used to predict the probability of membership of a class. To overcome the problem of uncertainty in data, you can use the Bayes probability equations (2) and (3) below, which show the Naive Bayes equation [9].

Definition 2.2. Naive Bayes equation [9]

$$P(hj|x) = \frac{p(x|h)p(h)}{p(x)} \tag{2}$$

Neural Network

With the help of artificial neural networks, we can give systems a kind of intelligence when the system is given time to "train" and then expect from the learning process [10].

Measurement

Confusion Matrix is a performance measurement for machine learning classification problems where the output can be in the form of 2 or more classes. Confusion Matrix can be explained that there are 4 terms representing the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TF), False Positive (FP), and False Negative (FN). Confusion Matrix is also often called an error matrix. Basically, the confusion matrix shares data comparing the results of the classification attempted by the system with the actual classification.

Table 1: Confusion matrix testing [7]

Classification	True	False
Actual True	True Positif(TP)	False Negatif (FN)
Actual False	False Positif (FP)	True Negatif (TN)

Accuracy is defined as the level of correlation between predicted values and actual values. Precision is the level of accuracy between the data expected by the user and the answers provided by the system. On the other hand, recall is the level of success of the system in recreating data.

Definition 2.3. Accuracy = $\frac{(TP + TN\ )}{(TP + TN + FP + FN)}\ X\ 100\%$ (3)

Definition 2.4. Precision = $\frac{(TP\ )}{(TP + FP)}\ X\ 100\%$ (4)

Definition 2.5. Recall = $\frac{(TP\ )}{(TP + FN)}\ X\ 100\%$ (5)

Research Methodology

This study aims to conduct a comparative analysis of the KNN, Naive Bayes, and Neural Network methods used to classify stroke in patients using medical record data. The application used for this simulation is Orange Data Mining, which is an open source application for processing data that has been proven to help researchers analyse data.

The steps in the first research are problem identification, formulation and literature review. This is done in order to develop research objectives and contributions. Second is the process of collecting data, namely compiling training data and test data as a source of data classification. The third step is designing the orange data mining widget for the classification process. The fifth classification process uses the KNN, Naive Bayes, and Neural Network models. Finally, the process of evaluating the performance of classification methods and analysing the results of comparison of these methods.

Data Set

In this research, researchers carried out analysis using quantitative analysis methods to produce stronger analysis if used according to the rules and can be used to estimate or predict. The data collected for this research was obtained from the Central Bureau of Statistics.

In diagnosing stroke, factors that influence the stroke are needed. Based on the data obtained, there are several factors such as those contained in the data which are used as a source of analysis such as BMI (Body Mass Index), hypertension, heart disease, glucose level, smoker or not, age, gender, type of work, and type of residence that can be classified as input variables and output variables. Input variables and output variables can be seen in Figure 2 and Table 2. Based on the Figure 2, there is 1 output variable to help predict the classification of stroke.

No.	Stroke	BMI	ID	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Incidence Type	Glucose Level	Smoking Status
1	1	36.6	9046	Male	67.00	0	1	Yes	Private	Urban	228.09	Never
2	1	N/A	55676	Female	61.00	0	0	Yes	Self employed	Rural	202.21	Never
3	1	32.5	31112	Male	80.00	0	1	Yes	Private	Rural	105.92	Smokes
4	1	34.4	60182	Female	49.00	0	0	Yes	Private	Urban	171.23	Never
5	1	24	1665	Female	79.00	1	0	Yes	Self employed	Rural	174.12	Never
6	1	29	54449	Male	81.00	0	0	Yes	Private	Urban	186.21	Never
7	1	27.4	53882	Male	74.00	1	1	Yes	Private	Rural	70.09	never
8	1	22.8	10434	Female	69.00	0	0	Yes	Private	Urban	94.39	Never
9	1	N/A	27459	Female	59.00	0	0	Yes	Private	Rural	76.15	Unknown
10	1	24.2	60491	Female	78.00	0	0	Yes	Private	Urban	58.57	Unknown
11	1	29.7	12109	Female	81.00	1	0	Yes	Private	Rural	80.43	Never
12	1	36.8	12095	Female	61.00	0	1	Yes	Private	Rural	120.46	Smokes
13	1	27.3	12175	Female	54.00	0	0	Yes	Private	Urban	104.51	Smokes
14	1	N/A	8213	Male	78.00	0	1	Yes	Private	Urban	219.84	Unknown
15	1	24.2	5317	Male	79.00	0	1	Yes	Private	Urban	214.09	Never

Figure 2: Input variable

Table 2: Output variable

No.	Variable Name	Description
1	Determining stroke	1 = Having a Stroke 0 = Not Having a Stroke

Problem Analysis

In the K-Nearest Neighbour analysis, there are two types of variables, namely: (1) The dependent or predicted variable, symbolised by Y, is a variable whose condition is influenced by the condition of other variables. (2) The independent variable or predictor, symbolised by X is an independent variable whose condition is not influenced by other variables.

Meanwhile, Naive Bayes analysis has rules such as that the results (C, target) can be estimated based on several test samples (X, attributes) that are being observed. There are several important things about Bayes' rule, namely: (1) Initial probability (C, target) or P(C) is the probability of the hypothesis before the evidence is observed. (2) A final probability (C, target) or P(C|X) is the probability of a hypothesis after the evidence is observed. Note that the classification process requires some guidance to determine which category is appropriate for the sample data being analysed.

Data Cleaning

Enhancing the relevance of our analysis involves implementing a filtering process to extract only the necessary information from the dataset. It is important to recognise that the data may contain incomplete entries, including instances of missing data. Apart from that, the attributes are not suitable for the data mining processing that will be used. It is also better to throw away unused data because its presence can reduce the quality or accuracy of data mining results later. Data cleaning will also affect the performance of the data mining system because the amount of data handled will decrease.

Data Selection

Data is selected to determine what variables will be taken so that there are no unnecessary similarities and repetitions in processing data mining techniques.

Data Transformation

Converting data into a format suitable for data mining processing. Some data mining methods require special data formats before they can be processed by the data mining method. Some methods such as analysis can only accept categorical data input. Therefore, data in the form of continuous numeric numbers needs to be divided into several intervals.

Mining Process

To process the main techniques when the method is applied to find valuable knowledge, the data collected according to the procedure must be applied to the mining process after the data has gone through the transformation stage.

In analysing the performance of several classification models in the Orange tool, a comparison of several data mining methods was conducted to select the best method with high accuracy in classifying stroke disease status datasets in patients.

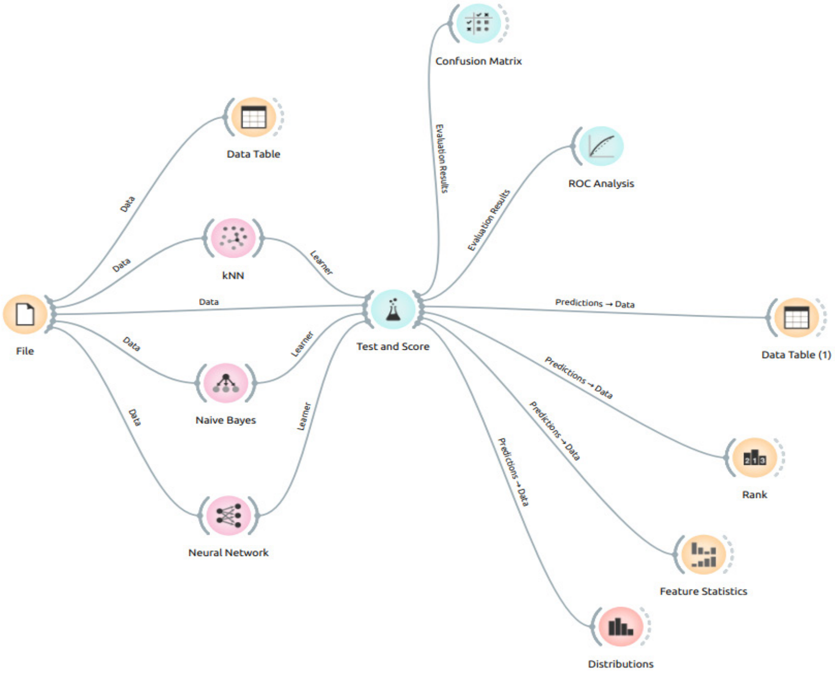


Figure 3: Classification model design [7]

In Figure 3, a widget is designed using a classification model in data mining software in the form of K-NN, Naive Bayes, and Neural Network, which is inputted by a dataset that has been previously processed. Then, the dataset is processed into classification mode.

Classification Model Testing Process

The next process is testing the model that has been created previously. A collection of data is needed to find out the classification results. In Figure 4, the widget design has been added to the classification trial process. In red is a set of trial data that is entered into the classification process to determine the results of stroke classification.

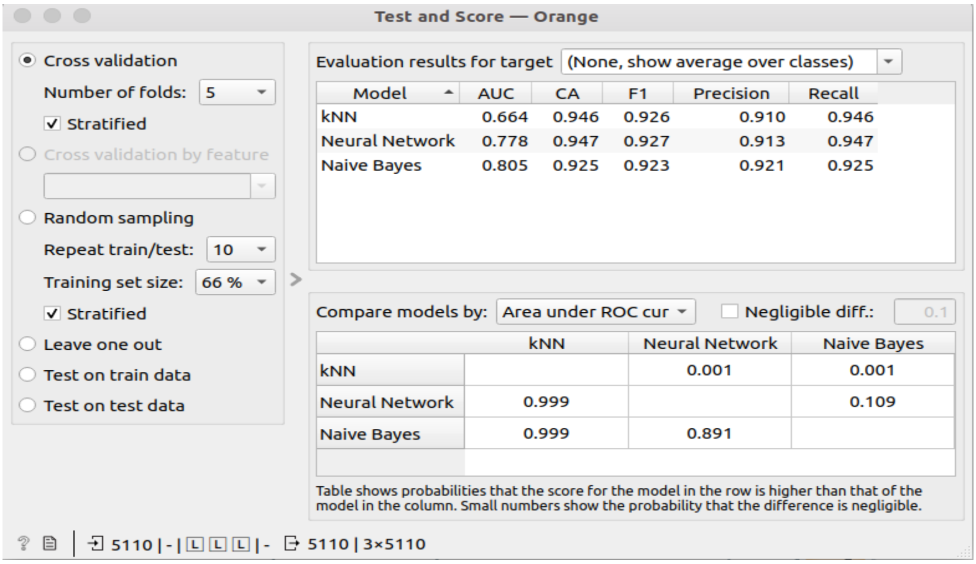
Evaluation

This stage is identifying dance patterns in the identified knowledge base. In this stage, the results of data mining techniques in the form of typical patterns and prediction models are evaluated to assess whether existing studies have met the desired targets.

Results and Discussion

Classification Model Simulation Results

Classification model simulation results using a test data set with 1 attribute as a target, 2 attributes as numeric, and 8 attributes as categorical. So that the test score results are obtained, as shown in Figure 4.



Test and Score — Orange

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.664	0.946	0.926	0.910	0.946
Neural Network	0.778	0.947	0.927	0.913	0.947
Naive Bayes	0.805	0.925	0.923	0.921	0.925

Compare models by: Area under ROC cur □ Negligible diff.: 0.1

	kNN	Neural Network	Naive Bayes
kNN		0.001	0.001
Neural Network	0.999		0.109
Naive Bayes	0.999	0.891	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Figure 4: Test score

Based on 5,000 patient data that have been tested, the calculation results for precision, recall and accuracy for each are obtained as shown in Figure 6. The classification results of the K-NN, Decision Tree, and Naive Bayes models show that the Naive Bayes accuracy value is the highest.

Based on Figure 4, which also shows 3 AUC models, it is known that the highest AUC value is the Naive Bayes method, with a value of 0.805. AUC is used to measure performance by estimating the output probability of randomly selected outcomes. The greater the AUC, the better the classification results used [9].

Evaluation Results with Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification problems where the output can be in the form of two or more classes. The evaluation results for each classification model can be seen in the image below for the K-NN, Naive Bayes, and Neural Network models.

The value of True Positive (TP) is 4851, True Negative (TN) is 1, False Positive (FP) is 10, and False Negative (FN) is 248. So, the Accuracy Precision and Recall values of the K-method NN are as follows:

Definition 2.5. Accuracy $= \frac{(4851 + 10\ )}{(4851 + 248 + 1 + 10)}\ X\ 100$ (6)

Definition 2.6. Precision $= \frac{(4851\ )}{(4851 + 248)}\ X\ 100\%$ (7)

Definition 2.7. Recall $= \frac{(4851\ )}{(4851 + 10)}\ X\ 100\%$ (8)

The value of True Positive is 4672, True Negative is 41, False Positive is 189, False Negative is 208. So, the Precision and Recall accuracy values of the Naive Bayes method are as follows:

Definition 2.8. Accuracy = $\frac{(4672 + 189\ )}{(4672 + 208 + 41 + 189)}\ X\ 100\%$ (9)

Definition 2.9. Precision = $\frac{(4672\ )}{(4672 + 208)}\ X\ 100\%$ (10)

Definition 2.10. Precision = $\frac{(4672\ )}{(4672 + 208)}\ X\ 100\%$ (11)

The value of True Positive is 4827, True Negative is 9, False Positive 34, False Negative 240.

Definition 2.11. Accuracy = $\frac{(4827 + 34\ )}{(4827 + 240 + 9 + 34)}\ X\ 100\%$ (12)

Definition 2.12. Precision = $\frac{(4827\ )}{(4827 + 240)}\ X\ 100\%$ (13)

Definition 2.13. Recall = $\frac{(4827\ )}{(4827 + 34)}\ X\ 100\%$ (14)

Based on the results of evaluation and validation using the Confusion Matrix, comparison values for Accuracy, Precision and Recall were obtained from the three methods of K-NN, Naive Bayes and Neural Network.

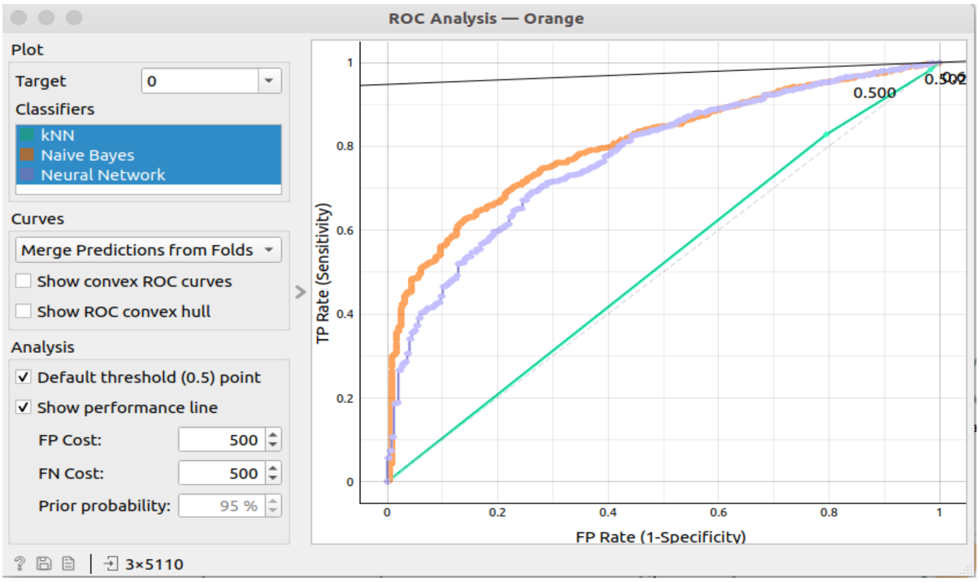
Table 3: Performance comparison

Method	Accuracy	Precision	Recall
K-NN	95%	95%	99%
Naive Bayes	95%	96%	96%
Neural Network	95%	96%	99%

Based on Table 3, it can be seen that the performance of the Neural Network model is better than K-NN and Naive Bayes. Classification accuracy cannot achieve perfect results because there must be error values. This can be influenced by the amount of test data and training data used in the simulation process being performed.

Evaluation Results with ROC Curve

Manual accuracy values can be done by looking at the ROC curve comparison visualised from the Confusion Matrix. Model viewing ROC curves is the most easily visible way to compare the accuracy values of each classification model graphically. The ROC graphic results can be seen in the image below.



ROC Analysis — Orange

Plot

Target: 0

Classifiers

- 🟩 kNN
- 🟧 Naive Bayes
- 🟪 Neural Network

Curves

Merge Predictions from Folds

□ Show convex ROC curves

□ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior probability: 95 %

3×5110

Figure 5: ROC analysis

Figure 5 shows that the results of the ROC analysis for each model are as follows: K-NN is 0.6, Naive Bayes is 0.500 and Neural Network is 0.502. Therefore, for this case study, the models that have the best accuracy values are Naive Bayes and Neural Network because the curves are close to the 0.1 point.

After conducting the research above, it was found that in determining stroke by using output variables and input variables, namely factors such as age, gender, glucose levels, hypertension, and heart disease, which can cause stroke. Using the K-Nearest Neighbor, Neural Network and Naive Bayes methods.

Factors that can cause stroke are, for example, age, gender, diabetes, genetic factors, obesity, physical condition, hypertension, heart disease, emotional factors, high blood pressure, drugs, and genetics [7]. So, in the data classification above, determining whether a patient has a stroke or not is based on the factors mentioned.

Conclusions

The classification analysis for diagnosing stroke using Orange Data Mining effectively identifies significant risk factors such as age, hypertension, diabetes, and smoking, allowing for an in-depth understanding of their relationship to stroke risk. By building accurate predictive models based on clinical data, Orange Data Mining enables early identification of individuals at risk of stroke. The model's performance, evaluated through metrics such as accuracy, precision, recall, and AUC, demonstrates its reliability in predicting stroke occurrences. Among the models tested—K-Nearest Neighbor, Naive Bayes, and Neural Network—Naive Bayes was found to be the most effective, achieving 95% accuracy and 96% precision. This superior performance suggests its utility in early stroke detection, which can inform timely interventions, including lifestyle changes and more intensive medical supervision for high-risk individuals. Furthermore, this study contributes to public health by providing valuable insights that can guide policy efforts in stroke prevention and management, offering benefits not only to individuals but also to healthcare systems at large.

Acknowledgements

Thank you to the Chancellor of the Indonesian Computer University, Prof. Dr. Eddy Soeryanto Soegoto., MT, Vice Chancellor for Academic Affairs, Prof. Dr. Ummi Narimawati., SE., Dr. Senny Lukyardi and team, fellow lecturers of the Information Systems Study Program who have supported this research.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

[1] Adelina, V., Ratnawati, D. E., & Fauzi, M. A. (2018). Klasifikasi tingkat risiko penyakit stroke menggunakan metode GA-Fuzzy Tsukamoto. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(9), 3015-3021. Diambil dari https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2513

[2] Romli, I. (2021). *Penerapan data mining menggunakan algoritma K-means untuk klasifikasi penyakit ISPA*. *Indonesian Journal of Business Intelligence*, *4*(1), Artikel 10. https://doi.org/10.21927/ijubi.v4i1.1727

[3] Byna, A., & Basit, M. (2020). *Penerapan metode ADABOOST untuk mengoptimasi prediksi penyakit stroke dengan algoritma naïve bayes*. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(3), 407-411. https://doi.org/10.32736/sisfokom.v9i3.1023

[4] Argina, A. M. (2020). Penerapan metode klasifikasi K-Nearest neigbor pada dataset penderita penyakit diabetes. *Indonesian Journal of Data and Science*, *1*(2), 29-33. https://doi.org/10.33096/ijodas.v1i2.11

[5] Riyadina, W., & Rahajeng, E. (2013). Determinan penyakit stroke. *Kesmas National Public Health Journal*, *7*(7), Artikel 324. https://doi.org/10.21109/kesmas.v7i7.31

[6] Susanto, S., & Suryadi, D. (2010). *Pengantar Data Mining*. Yogyakarta: Penerbit Andi.https://repository.unpar.ac.id/bitstream/handle/123456789/1551/Sani_129277-p.pdf?sequence=1&isAllowed=y

[7] Santoso, H., Hariyadi, I. P., Prayitno. (2016). *Data Mining Analisa Pola Pembelian Produk Dengan Menggunakan Metode Algoritma Apriori*. *Seminar Nasional Teknologi Informasi dan Multimedia, STMIK AMIKOM Yogyakarta, 6-7 Februari 2016* (pp. 19-24). https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/download/1267/1200

[8] Hozairi, H., Anwari, A., & Alim, S. (2021). *Implementasi orange data mining untuk klasifikasi kelulusan mahasiswa dengan model K-nearest neighbor*, *decision tree serta naïve bayes*. *Network Engineering Research Operation*, *6*(2), Artikel 133. https://doi.org/10.21107/nero.v6i2.237

[9] Manalu, E., Sianturi, A., & Manalu, M. R. (2017). *Penerapan Algoritma Naive Bayes untuk memprediksi jumlah produksi barang berdasarkan data persediaan*. *Jurnal Teknologi Dan Informasi*, *1*(2), 16-21.

[10] Retnowati, & Danang A. N. W. (1970). *Sistem pendukung keputusan penjurusan di SMA menggunakan metode Neural Network Backpropagation (Studi Kasus Sma Islam Kepanjen Malang)*. Bimasakti. https://ejournal.unikama.ac.id/index.php/JFTI/article/view/831/519