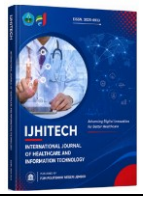
Content lists available at PublikasiPolije

International Journal of Healthcare and Information Technology

journal homepage : https://publikasi.polije.ac.id/index.php/ijhitech



Article

Application of the Naive Bayes Algorithm in Data Mining for Predicting Stroke Disease

Nur Anggun S[1], Nadhifa Dwi Rahmalia[2], Nurrahma Harris[3], Wita Windari[4], Andi Muhammad Zulkifli[5], Mochammad Choirur Roziqin[6] and Ziani Said[7]

[1] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; nanggun86@gmail.com
[2] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; nadhifadwirahmalia@gmail.com
[3] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; nurrahmaharris@gmail.com
[4] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; witawindari27@gmail.com
[5] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; andimuhzulkifli592@gmail.com
[6] Health Information Manajemen Study Program, Health Departement, Politeknik Negeri Jember, Indonesia; irul@polije.ac.id
[7] Department of Health Technologies Engineering, Research Group in Biomedical Engineering and Pharmaceutical Sciences, ENSAM, Mohammed V University, Agdal, Morocco; zsaid@um5.ac.ma

* Correspondence: nanggun86@gmail.com

Abstract: Stroke is a major global health problem that may lead to permanent disability or death. In Indonesia, analytical approaches such as data mining have been increasingly used to support early identification of stroke risk based on patient health records. This study applies the Naive Bayes algorithm as the main classification method. The research procedure includes data collection from the Kaggle repository, data selection based on predetermined clinical criteria, data cleaning to remove duplicates and missing values, and data transformation by converting categorical attributes into numerical form. The dataset was then split into training and testing subsets for model development. The final dataset consisted of 3,256 patient records containing variables such as gender, age group, hypertension, heart disease, average glucose level, body mass index (BMI), smoking status, and stroke occurrence. After completing these preprocessing stages, the Naive Bayes model achieved an accuracy of 89.49%. This result indicates that the model was able to classify stroke and non-stroke cases in the dataset with a satisfactory level of accuracy. The findings suggest that the model may serve as a baseline component for developing decision-support systems aimed at facilitating early identification of stroke risk in targeted population groups.

Keywords: Stroke, Data Mining, Naive Bayes

Citation: N. Anggun S., N. D. Rahmalia, N. Harris, W. Windari, A. M. Zulkifli, M. C. Roziqin and Z. Said, "Application of the Naive Bayes Algorithm in Data Mining for Predicting Stroke Disease", *ijhitech*, vol. 4, no. 1, pp. 26–34, Jul. 2026.

Received: 27-11-2025
Accepted: 25-06-2026
Published: 31-07-2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY SA) license (http://creativecommons.org/licenses/by-sa/4.0/).

1. Introduction

According to the World Health Organization (WHO), stroke or cerebrovascular disease is defined as a clinical syndrome that develops rapidly due to a focal or global disturbance of brain function, caused by the blockage or rupture of blood vessels in the brain. The resulting symptoms persist for 24 hours or longer [1]. Meanwhile, According to the American Stroke Association, stroke is defined as permanent cell death in the brain, spinal cord, or retina resulting from a vascular cause, as evidenced by pathology or imaging, with or without accompanying clinical symptoms [2]. In general, stroke can be classified into two main categories, ischemic

stroke and hemorrhagic stroke. Ischemic stroke occurs due to an interruption of blood supply to a part of the brain, resulting in a sudden loss of a blood vessel or structural abnormalities of the vasculature. Approximately 80% of all stroke cases are ischemic, while about 20% are hemorrhagic, although these proportions vary depending on the population.

The latest Global Burden of Disease (GBD) 2021 estimates show that among noncommunicable diseases, stroke is the second leading cause of death (approximately 7 million deaths) and the third leading cause of death and disability combined (measured in disability-adjusted life-years (DALYs), which is more than 160 million cases) worldwide [3]. Based on various epidemiological studies, stroke risk factors are divided into two categories: modifiable and non-modifiable risk factors. Modifiable risk factors include hypertension, diabetes, high cholesterol, cardiovascular disease, lifestyle, atrial fibrillation, smoking, and alcohol consumption. Non-modifiable risk factors include age and gender [2].

Stroke is a serious health threat worldwide that causes permanent damage or death. This problem requires immediate attention from medical personnel, because the sooner treatment is given, the greater the chances of recovery by preserving the affected area of the brain and preventing further deficits. Stroke detection systems can be used as an important tool to help healthcare professionals make quick and accurate treatment decisions [4]. Innovation in stroke prevention, control, and treatment is crucial. A number of studies have also been conducted in Indonesia on stroke control. These studies include predicting stroke based on patient history data using data mining, a method in which classification is used. One of the algorithms used in classification is Naive Bayes [5]**.**

The Naive Bayes theory is named after an 18th-century English mathematician, Thomas Bayes. Naive Bayes logic can show a patient's test results with pre-test probabilities (from the population) to predict or determine the likelihood of finding a disease. The structure of Naive Bayes can be represented by a directed acyclic graph, where each node represents a variable and the lines connecting the nodes show the relationship of dependency between variables. Within a node, a variable can have many different values, each with a certain probability. One of the main problems with the Naive Bayes model is that the root node of the network has a relationship with all predictor variables and does not depend on other variables [6].

In the study by Chakraborty et al, the Naive Bayes method was used to predict stroke events by utilizing variables such as age, gender, history of hypertension, history of heart disease, marital status, type of occupation, type of residence, average glucose level, body mass index (BMI), and smoking status. This method yielded an accuracy of 72.8% with a precision of 67.0% and a recall of 73.6% [7]. These results show that Naive Bayes is capable of processing data patterns quickly and efficiently, while still providing informative prediction results.

This study utilizes machine learning techniques with a focus on the Naive Bayes algorithm to predict the vulnerability of individuals to disease at random, with the aim of improving early detection and more effective intervention [8]**.** This algorithm is called “naive” because it simplifies calculations by assuming that each feature is independent of one another. The essence of the Naive Bayes algorithm lies in its ability to make predictions based on probability distributions, making it a powerful tool in the field of machine learning and classification [9]**.** Therefore, this study uses the Naive Bayes algorithm because of its simple characteristics, ease of implementation, lack of heavy computation requirements, and suitability for preliminary analysis of stroke prediction in large-scale datasets. Despite its simplicity

and limited assumptions, Naive Bayes often performs well in various classification tasks, especially when used on large datasets [10].

2. Materials and Methods

Naive Bayes has a preprocessing stage before data processing. This preprocessing stage aims to clean and transform raw text data to make it more suitable for input into machine learning algorithms. Thru preprocessing, the raw data will become more standardized, clean, and structured [11]. The stages that will be carried out in this research include:

1. Data Collection: In this study, the data collection process was obtained from the dataset platform, Kaggle.
2. Data Pre-processing

   At this stage, several processes will be carried out, namely:

   a. Data Selection

      This stage aims to select which data will be collected and used in this study. Variable selection was performed by choosing attributes directly related to stroke prediction, namely: gender, age, hypertension, heart disease, glucose level, smoking status, body mass index (BMI), and stroke label.

   b. Data cleaning

      This stage aims to clean the data by removing duplicate data, handling missing values, and unusable data.
3. Data Transformation

   This stage converts categorical variables such as hypertension, heart disease, gender, and smoking status into a format (numerical) that can be read by Naive Bayes.
4. Process Data Mining

   The data mining stage is performed using the Naive Bayes algorithm to build a classification model for determining stroke diagnosis based on predictor variables.
5. Modeling Stages in RapidMiner

   The model was built using the Split Validation method with a 70% : 30% ratio of training to testing data.
6. Evaluation: The evaluation stage is the step to measure how well the algorithm performs in predicting or classifying data. Evaluation is done to ensure that the algorithm is truly capable of predicting stroke based on predictor variables such as age, hypertension, heart disease, glucose levels, and smoking status.

3. Results

3.1 Data Collection

The data in this study consists of patient data with symptoms indicating stroke, provided by the dataset platform Kaggle, and downloaded thru the link: *https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset*

This dataset contains 5110 rows of data with 12 columns, including: id, gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose level, bmi, and smoking status. Then there is 1 column of dependent features, which is the predicted label, namely stroke

Table 1. Stroke Patient Data

No	Id	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Residence Tyoe	Avg glucose level	BMI	Smoking Status	Stroke
1.	9046	Male	67	1	1	Yes	Private	Urban	228.69	36.06	Formerly smoked	1
2.	51676	Female	61	0	0	Yes	Self umployed	Urban	202.21	n/a	Never smoked	1
3.	31112	Male	80	1	1	Yes	Private	Rural	105.92	32.5	Never smoked	1
...												
5110	44679	female	44	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

3.2 Data Pre-processing

3.2.1 Selection Data

The selection stage is performed to choose which variables will be used. In this study, the stroke prediction dataset obtained from Kaggle contains 12 variables, including gender, age, hypertension, heart diseases, ever married, work type, residence type, average glucose level, BMI, smoking status, and stroke. In this study, only 8 variables will be used:

1. Age according to age requirements based on Kemenkes [12] as follows;
   a. Adult = age 18-59 years
   b. Elderly = starting at age 60 years
2. Average Glucose level based on American Diabetes Association Professional Practice Committee [13] as follows ;
   a. Normal = <140
   b. Pradiabetes = 140-199
   c. Diabetes = 200
3. Body Mass Index (BMI) Adult = age 18-59 years Elderly = from age 60 years is the method recommended by the World Health Organization (WHO) for classifying the nutritional status of adults based on body weight relative to height. BMI Classification According to WHO for Adults (≥ 18 years) [14] as follows;
   a. Underweight < 18,5,
   b. Normal 18,5 – 24,9,
   c. Overweight 25,0 – 29,9
   d. Obesity >30,0

   Additionally, this study uses the variables of gender, hypertension, heart diseases, smoking status, and stroke.

3.2.2 Data Cleaning

The next step is data cleaning. This step aims to remove data with no values, missing values, and redundant data. For example, the "other" gender category is cleaned, ages under 18 are cleaned, BMI values that are N/A (Not Available) are cleaned, and smoking status unknown and other variables with missing or empty

values are also cleaned. After going thru the data cleaning stage, the data, which initially consisted of 5110 rows, was reduced to 3256 rows.

Table 2. Data Cleaning

No	Gender	Age	Hypertension	Heart Disease	Avg glucose level	BMI	Smoking Status	Stroke
1.	Male	67	1	1	228.69	36.06	Formerly smoked	1
2.	Female	61	0	1	202.21	n/a	Never smoked	1
3.	Male	80	1	0	105.92	32.5	Never smoked	1
...								
5110	female	44	0	0	85.28	26.2	Unknown	0

3.3 Data Transformation

This stage is performed by adjusting the data to the data mining process that will be carried out, where text-based data is converted into numerical data, such as for the gender and smoking variables. Here is the information after transformation.

1. Gender : 0 if the patient “Female” and 1 if the patient “Male”
2. Age : 0 if the adult patient is aged 18-59 and 1 if the patient is elderly, 60 years or older.
3. Hypertention : 0 if the patient does not have hypertension, 1 if the patient has hypertension.
4. Heart diseases : 0 if the patient does not have heart disease, 1 if the patient has heart disease
5. Avg glucose level : 0 if the patient's blood sugar is normal, 1 if prediabetes, 2 if the patient has diabetes.
6. BMI : 0 if Underweight, 1 if Normal, 2 if the patient is Overweight, and 3 if the patient is Obesity.
7. Smoking status : 0 if the never smoked, 1 if the patient formerly smoked, and 2 if the patient currently smokes.
8. Stroke : 0 if the patient does not have a stroke and 1 if the patient has a stroke.

Tabel 3. Data Transformation

No	Gender	Age	Hypertension	Heart Disease	Avg glucose level	BMI	Smoking Status	Stroke
1.	1	1	0	1	2	1	1	1
2.	1	1	0	1	0	0	0	1
3.	0	0	0	0	1	2	2	1
...								
5110	1	0	0	0	1	1	1	0

3.4 Modeling stages in RapidMiner

Here are the steps for the modeling process using RapidMiner:

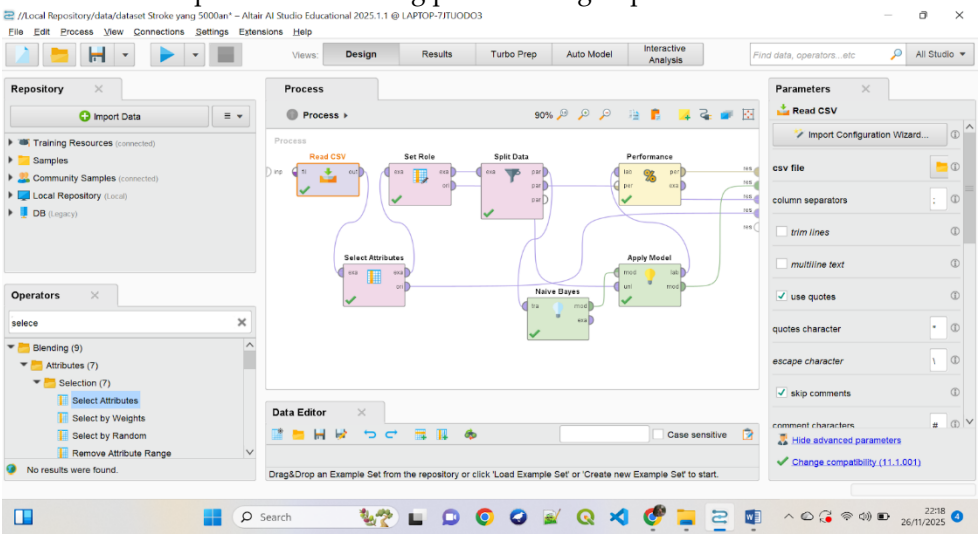


Figure 1. Modeling in RapidMiner

The image above shows the modeling process in RapidMiner using the Naive Bayes algorithm. The more detailed steps of the process in the image above are as follows:

. The first step is to open the RapidMiner software, then select the menu to create a new process.

a. Step two, input the data to be processed by clicking "import data," then browse for the location of the stroke dataset file. Ensure the target column (label) is set as a label (target attribute).

b. Step three: Select attributes to determine the attribute type or filter the attributes that will be used in the analysis process.

c. Step four: Specify the attribute names and target roles for the dataset that has been entered.

d. Step five: Insert the Split Data operator to divide the data into training and testing sets with a 7:3 ratio. Connect this operator with the Naïve Bayes operator and Apply Model.

e. Step six:, Insert the Performance operator and connect it to the Apply Model operator to find out the model's performance based on accuracy, precision, recall, and so on.

f. The final step: click play to run the classification process and view the model evaluation results.

3.5 Evaluation Stages

The evaluation stage is a step to measure the performance of the built model in order to determine how well the Naive Bayes algorithm predicts or classifies the stroke dataset. In the modeling process in RapidMiner, there is a performance operator used to evaluate model performance based on metrics such as accuracy, precision, and recall [16]. Here are the model performance results.

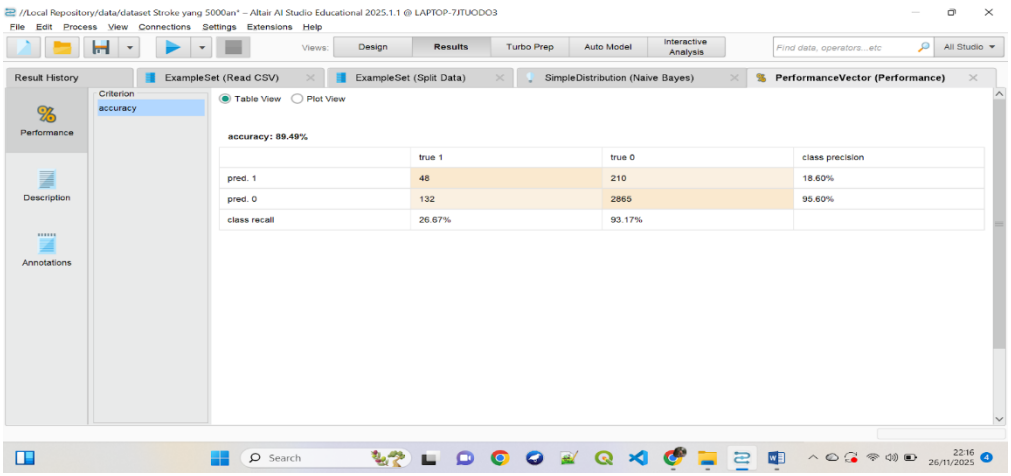


Figure 2. Evaluation Results of the Naïve Bayes Algorithm

	True 1	True 0	Class precision
pred. 1	48	210	18.60%
pred. 0	132	2865	95.60%
classrecall	26.67%	93.17%	

Based on the image above, it shows the model performance results, indicating good accuracy in predicting stroke with an accuracy of 89.49%, which means the model can correctly classify stroke data. As for the Naive Bayes operator used to determine the classes of the stroke dataset, the following are the results of the simple distribution model (Naive Bayes).

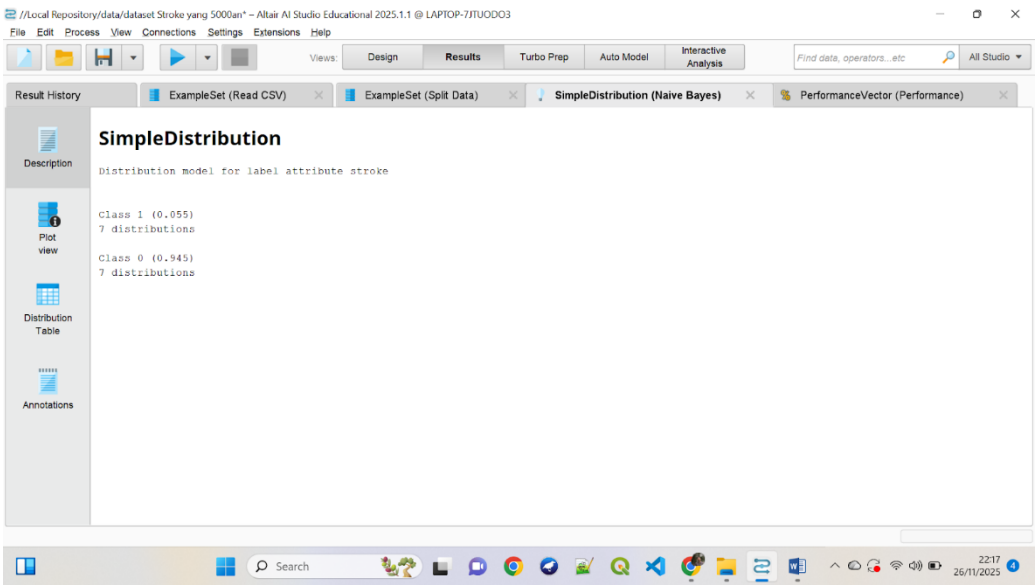


Figure 3. Distribution Model for Label Attribute Stroke

Based on the image above, it shows the results of simple distribution (Naive Bayes), which illustrates the division of classes. Class 1, representing stroke, shows a value of 0.055 or 5%, while class 0, representing no stroke, shows a value of 0.945 or 95%. This reflects the proportion of data in each class within the dataset used.

4. Discussion

The results of the study show that the Naive Bayes algorithm is capable of providing fairly good classification performance in predicting the risk of stroke with variables such as gender, age, hypertension, heart diseases, average glucose level, BMI, smoking status, and stroke. The accuracy obtained was 89.49% from 3256 pre-processed data. Thus, the results show that Naive Bayes is capable of utilizing patterns in physiological data (age, average glucose level, BMI), clinical data (hypertension, heart diseases, stroke, smoking status), and demographic data (gender), effectively classifying patients who are at higher risk of stroke and those who are not. This shows that these variables can help provide sufficient information for calculations in determining stroke classification. This model's performance aligns with previous research. Chakraborty et al. (2024) reported an accuracy of 72.8% using Naive Bayes, with a broader set of features and different preprocessing methods [7], while Sailasya and Gangavarapu (2021) achieved an accuracy of 82%, which was the highest among several machine learning algorithms evaluated[15]. The relatively higher accuracy in the present study may be attributed to more rigorous data cleaning and feature selection, further supporting Naive Bayes as a reliable baseline classifier for stroke prediction. In contrast, studies utilizing the C4.5 algorithm exhibit different performance patterns. While one study reported a high overall accuracy of 95.01%, the precision and recall values were low, indicating limitations in correctly identifying actual stroke cases. This pattern reflects a tendency for C4.5 to produce more accurate predictions for the non-stroke class, as evidenced by higher specificity and negative predictive values. Similar findings have been reported in other studies using C4.5 on datasets with imbalanced class distributions [16]. Despite the promising results, this study has several methodological limitations to note. The dataset was sourced from a single repository, Kaggle, making the model heavily dependent on the specific characteristics of that dataset. Additionally, external validation using independent datasets was not conducted, which limits the generalizability of the findings. This study also did not include direct comparisons with other machine learning algorithms, preventing a comprehensive assessment of the relative performance of Naive Bayes on this dataset.

5. Conclusions

Based on the implementation of the Naive Bayes machine learning algorithm in predicting a stroke disease dataset the result show good performance with an accuracy 89,49%. This indicates that the model is able to correctly classify stroke cases within the dataset. The modeling results can serve as a useful tool to help detect and predict stroke, which, as we know, is an acute condition that occurs suddenly and severely. This demonstrates that the use of the Naive Bayes machine learning algorithm can be a potential solution and support early prevention efforts for stroke by utilizing physiological data, clinical data, and demographic data to provide fast and efficient predictions.

References

[1] A. S. Arifianto, M. Sarosa, and O. Setyawati, "Klasifikasi Stroke Berdasarkan Kelainan Patologis dengan Learning Vector Quantization," *J. EECCIS Electr. Electron. Commun. Controls Inform. Syst.*, vol. 8, no. 2, pp. 117–122, Aug. 2014, doi: 10.21776/jeeccis.v8i2.248.

[2] E. S. Donkor, "Stroke in the 2 1 s t Century: A Snapshot of the Burden, Epidemiology, and Quality of Life," *Stroke Res. Treat.*, vol. 2018, pp. 1–10, Nov. 2018, doi: 10.1155/2018/3238165.
[3] V. L. Feigin *et al.*, "World Stroke Organization: Global Stroke Fact Sheet 2025," *Int. J. Stroke*, vol. 20, no. 2, pp. 132–144, Feb. 2025, doi: 10.1177/17474930241308142.
[4] O. Ozaltin, O. Coskun, O. Yeniay, and A. Subasi, "A Deep Learning Approach for Detecting Stroke from Brain CT Images Using OzNet," *Bioengineering*, vol. 9, no. 12, p. 783, Dec. 2022, doi: 10.3390/bioengineering9120783.
[5] L. Rahmawati, M. Hafid, and M. A. Sunandar, "Analisis Data Mining Untuk Memprediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," vol. 6, no. 2, 2023, doi: 10.17509/jatikom.v6i2.49051.
[6] M. Langarizadeh and F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review," *Acta Inform. Medica*, vol. 24, no. 5, p. 364, 2016, doi: 10.5455/aim.2016.24.364-369.
[7] P. Chakraborty *et al.*, "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," *BMC Bioinformatics*, vol. 25, no. 1, p. 329, Oct. 2024, doi: 10.1186/s12859-024-05866-8.
[8] A. Z. Arrayyan, H. Setiawan, and K. T. Putra, "Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations," *Semesta Tek.*, vol. 27, no. 1, pp. 28–36, Apr. 2024, doi: 10.18196/st.v27i1.21008.
[9] M. Hassan *et al.*, "Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness".
[10] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial Naive Bayes Classification Model for Sentiment Analysis".
[11] P. Zhang *et al.*, "Design of an Automatic Classification System for Educational Reform Documents Based on Naive Bayes Algorithm," *Mathematics*, vol. 12, no. 8, p. 1127, Apr. 2024, doi: 10.3390/math12081127.
[12] Kementerian Kesehatan Republik Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/33/2025 tentang Petunjuk Teknis Pemeriksaan Kesehatan Gratis Hari Ulang Tahun," HK.01.07/MENKES/33/2025, 2025. Accessed: Nov. 26, 2025. [Online]. Available: https://peraturan.bpk.go.id/Download/393217/Permendagri%20No.%2014%20Tahun%202025.pdf
[13] American Diabetes Association Professional Practice Committee *et al.*, "2. Diagnosis and Classification of Diabetes: *Standards of Care in Diabetes—2024*," *Diabetes Care*, vol. 47, no. Supplement_1, pp. S20–S42, Jan. 2024, doi: 10.2337/dc24-S002.
[14] World Health Organization, "WHO Discussion Paper: Obesity – Key Analysis," Discussion Paper, 2021. Accessed: Nov. 26, 2025. [Online]. Available: https://cdn.who.int/media/docs/default-source/obesity/who-discussion-paper-on-obesity---final190821.pdf
[15] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120662.
[16] A. G. Ramadhan, H. S. Hidayatullah, and R. Ruba'i, "Implementasi Algoritma C4.5 pada Sistem Prediksi Stroke Berdasarkan Data Kesehatan," 2025.