



Application of the Naive Bayes Algorithm in Stroke Risk Classification Using Patient Clinical Data

**Ahmad Izzuddin¹, Siti Rokhmah², Basuki Nur Apriliano³, Karnsa
Lambang Sesami⁴**

Institut Teknologi Bisnis AAS Indonesia, Sukoharjo, Indonesia
ahmadizzuddin2345@gmail.com¹, basukinurapriliano@gmail.com²,
sesami1822@gmail.com³

Abstract

Stroke is a serious medical condition and is one of the leading causes of death and long-term disability worldwide, including in Indonesia. The ability to predict stroke risk early can help in prevention efforts and timely medical intervention. This study applies the Naive Bayes classification algorithm to build a stroke risk prediction model. The dataset used in this study is 'healthcare-dataset-stroke-data' sourced from the Kaggle platform, including 5,110 patient data with 11 relevant clinical and demographic attributes, such as age, gender, hypertension, heart disease, average glucose level, body mass index (BMI), and smoking status. The Naive Bayes method was chosen because of its computational efficiency, its ability to handle high-dimensional data, and its solid performance in many medical diagnostic applications. The research process includes several stages: data preprocessing to handle missing values and discretize continuous attributes, implementation of the Naive Bayes algorithm by calculating prior probabilities and likelihood probabilities for each attribute against the target class (stroke and non-stroke), and classification on the test data. The results of the study indicate that the Naive Bayes model is capable of classifying stroke risk using the evaluation metrics discussed below. Analysis of the likelihood probability table also confirmed that factors such as age, hypertension, and heart disease significantly influence the prediction. This study demonstrates the potential of Naive Bayes as a practical and informative initial screening tool for healthcare practitioners.

Keywords : *Stroke; classification; machine learning; naive Bayes; disease prediction; data mining; medical diagnosis.*

Introduction

Stroke is a significant global health challenge. According to the World Health Organization (WHO), stroke is the second leading cause of death worldwide and a leading cause of disability [1]. The burden of this disease is felt not only by individuals and families but also by national health systems. In Indonesia, the prevalence of stroke is also alarming. Based on data from the 2018 Basic Health Research (Riskesdas), the prevalence of stroke in



Indonesia, based on doctor diagnosis in the population aged 15 years and above, was 10.9 per 1,000, or approximately 2.1 million people [2]. This figure highlights the urgency of effective methods for prevention and early detection.

Primary prevention, which aims to reduce the risk of stroke in healthy individuals, is the most effective strategy. Identifying individuals at high risk is the first step in this prevention. Many risk factors for stroke have been identified and can be divided into two categories: non-modifiable (such as age, gender, and family history) and modifiable (such as hypertension, diabetes, smoking, obesity, and heart disease) [3]. Analysis of these risk factors can provide a more accurate prediction of a person's likelihood of experiencing a stroke.

Advances in information technology and computing have enabled the collection of large-scale health data (Big Data). This data, when properly analyzed, can reveal hidden patterns valuable for disease prediction and diagnosis. This is where machine learning plays a crucial role.

Machine learning is a branch of artificial intelligence that enables computer systems to "learn" from data without being explicitly programmed [4]. In a medical context, machine learning algorithms can be trained on historical patient datasets to build predictive models. These models can then be used to estimate disease risk in new patients based on their clinical and demographic profiles. This approach offers the potential for faster, more objective, and more personalized diagnoses. Several previous studies have explored the use of various machine learning techniques for stroke prediction. Govindarajan et al. [5] used several models, including Logistic Regression and Support Vector Machine (SVM), to predict functional outcomes after stroke. They found that the machine learning model was capable of providing accurate predictions. Meanwhile, a study by Sailasya and Kumari [6] compared several classification algorithms such as Decision Tree, Naive Bayes, and K-Nearest Neighbors (KNN) for stroke prediction and reported that Decision Tree provided the highest accuracy on their dataset.

Another study by Kadam and Jadhav [7] specifically focused on the use of Naive Bayes for predicting heart disease, a major risk factor for stroke. They highlighted the speed and simplicity of Naive Bayes as key advantages. Although numerous studies have been conducted, there is a need for studies that transparently demonstrate the algorithm's operation on commonly used stroke datasets, so that it can be replicated and better understood.

Based on the background and a review of related research, this study has the following objectives:

1. Apply the Naive Bayes classification algorithm to build a stroke risk prediction model based on the healthcare-dataset-stroke-data dataset.
2. Describe in detail and transparently each step of the manual calculation of the Naive Bayes algorithm, from pre-processing to the final classification.

3. Analyze the classification results and evaluate the influence of each attribute on stroke risk prediction.

The main contribution of this research is the presentation of a detailed and easy-to-follow case study on the implementation of Naive Bayes for a real-life medical problem, which can serve as an educational reference and a foundation for further research.

Results and Discussion

A. Data pre-processing results

Tabel III menunjukkan distribusi data setelah proses diskretisasi atribut 'age'."

**TABLE III
DISTRIBUSI DATA SETELAH DISKRETISASI USIA**

Age Category	Range	number of patient
	S=0	S=1
Child	35	1
teeneger	16	1
Adult	132	37
elderly	66	210

child	0-12	[36]
teeneger	13-18	[17]
Adult	19-55	[169]
elderly	> 55	[276]

B. Probability Calculation Results

1. Prior Probability: Based on the dataset analysis, it was found that out of 5,110 patients, 249 had a stroke and 4,861 did not. The results of the prior probability calculation are presented in Table IV.

**TABLE IV
PRIOR PROBABILITY CALCULATION RESULTS**

Class	Number of Cases	Prior Probability
Stroke = 1	249	0.0487
Stroke = 0	4861	0.9513
Total	5110	1.0

It can be seen that this dataset is highly imbalanced, with the minority class (stroke) only accounting for approximately 4.9% of the total data. This imbalance can affect model performance.



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

2. Likelihood Probability: A likelihood probability table was calculated for each attribute. Tables V through X present some of the calculation results as examples.

P(gender stroke)	0	1
Female	0,58	0,57
Male	0,42	0,43
P(hypertension st	0	1
0	0,91	0,73
1	0,09	0,27
P(heart_disease s	0	1
0	0,96	0,81
1	0,04	0,19
P(ever_married st	0	1
Yes	0,63	0,88
No	0,37	0,12
P(work_type strok	0	1
Private	0,60	0,60
Self-employed	0,12	0,26
Children	0,17	0,01
Govt_job	0,10	0,13
Never_worked	0,01	0,00
P(Residence_type	0	1
Rural	0,44	0,46
Urban	0,56	0,54
P(smoking_status	0	1
never smoked	0,36	0,36
formerly smoked	0,16	0,28
smokes	0,16	0,17
Unknown	0,33	0,19

TABLE V

LIKELIHOOD PROBABILITY FOR THE ATTRIBUTE OF
HYPERTENSION

| Value | P(Value | Stroke=1) | P(Value | Stroke=0) |

P(hypertension stroke)	0	1
0	0,91	0,73
1	0,09	0,27

| 1 (Yes) | [0,91] | [0,73] |

| 0 (No) | [0,09] | [0,27] |

TABLE VI

LIKELIHOOD PROBABILITY FOR THE ATTRIBUTE age_category

| Value | P(Value | Stroke=1) | P(Value | Stroke=0) |

Age Category	Range	number od
--------------	-------	--------------



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

	<i>patient</i>	
	S=0	S=1
<i>Child</i>	35	1
<i>teeneger</i>	16	1
<i>Adult</i>	132	37
<i>elderly</i>	66	210

Children	[35]	[1]
Teenagers	[16]	[1]
Adults	[132]	[37]
Seniors	[66]	[210]

C. Classification Case Study To demonstrate the classification process, one hypothetical test data is taken as follows:

1. Test Patient Data:

- gender: male*
- age_category: Lansia*
- hypertension: 1*
- heart_disease: 0*
- ever_married: Yes*
- work_type: Private*
- Residence_type: Rural*
- avg_glucose_category: Pra-diabetes*
- bmi_category: Gemuk*
- smoking_status: smoked*

2. Calculation for Stroke class=1:

- $P(\text{textstroke}=1/\text{texdata}) \propto P(\text{textstroke}=1) \times P(\text{textgender}=F/\text{texts}=1) \times \dots$
- $= 0.0487 \times [P(\text{gender}=F/s=1)] \times [P(\text{age}=Lansia/s=1)] \times \dots$
- $= 2,4158E-08$

3. Calculation for Stroke class=0:

- $P(\text{textstroke}=0/\text{texdata}) \propto P(\text{textstroke}=0) \times P(\text{textgender}$



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

=F/texts=0)times...

b. =0.9513times[textnilaiP(gender=F/s=0)]times[textnilaiP(age=La
nsia/s=0)]times...

c. = 1,3276E-09

4. Decision:

Hasil
Jadi, berdasarkan hasil prediksi data baru yang telah dilakukan, terlihat bahwa nilai P(stroke=1 data) yaitu 2.41583E-08 lebih besar dibandingkan dengan nilai P(stroke=0 data) yaitu 1.32759E-09.
Maka dari itu, dapat disimpulkan bahwa pasien kemungkinan besar akan mengalami stroke (stroke=1)

D. Performance Evaluation Results

id	gender	age	hypertension	heart_dise	ever_marriec	work_type	Residence_type	avg_gluco	bmi	smoking	stroke
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly s	1
31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smi	1
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
1665	Female	79	1	0	Yes	Self-empl	Rural	174,12	24	never smi	1
56669	Male	81	0	0	Yes	Private	Urban	186,21	29	formerly s	1
53882	Male	74	1	1	Yes	Private	Rural	70,09	27,4	never smi	1
60491	Female	78	0	0	Yes	Private	Urban	58,57	24,2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80,43	29,7	never smi	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120,46	36,8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104,51	27,3	smokes	1
5317	Female	79	0	1	Yes	Private	Urban	214,09	28,2	never smi	1
58202	Female	50	1	0	Yes	Self-empl	Rural	167,41	30,9	never smi	1
56112	Male	64	0	1	Yes	Private	Urban	191,61	37,5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221,29	25,8	smokes	1
20165	Female	77	0	0	Yes	Private	Urban	250,8	32,9	never smi	0
6988	Female	52	0	0	Yes	Self-empl	Urban	113,21	38,3	never smi	0
13328	Female	45	0	0	Yes	Private	Rural	106,95	33,4	Unknown	0
5121	Male	30	0	0	Yes	Private	Urban	96,84	21,1	Unknown	0
18040	Female	49	0	0	Yes	Govt_job	Rural	89,61	27,7	never smi	0
44759	Male	57	0	0	Yes	Private	Urban	215,92	27,4	smokes	0
18412	Male	41	0	0	Yes	Private	Rural	82,32	27,9	Unknown	0
67431	Female	52	0	0	Yes	Private	Urban	73,73	34,4	formerly s	0


normadist age	age S=1	normadist glukc	glukosa S=1	normadist bmi	bmi S=1	likelihood 0	likelihood 1	PROB S=0	PROB S=1
0,008185292	0,031453973	0,000212526	0,001924127	0,029042833	0,03743761	3,04585E-13	1,36597E-11	0,021812	0,978188
0,003387538	0,01969866	0,008827626	0,005884146	0,046672732	0,064340264	6,65874E-12	3,55798E-11	0,157646	0,842354
0,016038724	0,010552909	0,003047552	0,005307411	0,038962393	0,053354539	4,37813E-10	6,86986E-10	0,389236	0,610764
0,00366846	0,021199338	0,002770465	0,005148608	0,045619522	0,038639313	8,36972E-12	7,61316E-11	0,099048	0,900952
0,003121981	0,018190408	0,00177872	0,004427902	0,054499524	0,06810007	4,86544E-11	8,81819E-10	0,05229	0,94771
0,005304754	0,027868955	0,00651454	0,003873888	0,054184357	0,061716742	9,09166E-13	3,23512E-12	0,219378	0,780622
0,003964869	0,022672492	0,005169523	0,003153225	0,046361536	0,040107851	4,42368E-10	1,33484E-09	0,248911	0,751089
0,003121981	0,018190408	0,007585658	0,004524043	0,053817666	0,06937292	2,30076E-11	1,03058E-10	0,182505	0,817495
0,010994396	0,027358072	0,008346656	0,006334262	0,028150082	0,035986773	2,29107E-10	5,3093E-10	0,2929	0,7071
0,014183691	0,017504671	0,008826975	0,005825053	0,054078271	0,061179517	1,55649E-09	1,43411E-09	0,520461	0,479539
0,00366846	0,021199338	0,000487327	0,002702647	0,054669478	0,065465446	8,21498E-13	3,15269E-11	0,025395	0,974605
0,015710953	0,011823494	0,003435151	0,005506341	0,051552241	0,069164491	5,02249E-11	8,1286E-11	0,381907	0,618093
0,009570786	0,030169588	0,001425964	0,004088597	0,025087147	0,031038021	2,11702E-12	2,36731E-11	0,082086	0,917914
0,004946952	0,026716422	0,000327907	0,002301721	0,05134076	0,051732517	1,39752E-12	5,33827E-11	0,025511	0,974489
0,004276806	0,024097255	4,95798E-05	0,001034219	0,045185869	0,062421857	2,1476E-12	3,48693E-10	0,006121	0,993879
0,014986655	0,014566901	0,00869491	0,006148085	0,021745992	0,025734909	6,35149E-10	5,16602E-10	0,551464	0,448536
0,017080278	0,006292036	0,008822671	0,00592574	0,043210632	0,059693213	2,39889E-09	8,19942E-10	0,745267	0,254733
0,016338458	0,000372076	0,008674246	0,005463785	0,033178981	0,019616195	1,57007E-09	1,33153E-11	0,991591	0,008409
0,016038724	0,010552909	0,008310568	0,00507172	0,054442448	0,063241102	1,67335E-08	7,80503E-09	0,681927	0,318073
0,012867911	0,022004896	0,000441704	0,002597907	0,054184357	0,061716742	5,07808E-11	5,81745E-10	0,080283	0,919717
0,017625828	0,003395252	0,007755666	0,004640107	0,054563899	0,064180296	1,9709E-09	2,67172E-10	0,880624	0,119376
0,014986655	0,014566901	0,006914325	0,00410444	0,038962393	0,053354539	9,04954E-10	7,15021E-10	0,558622	0,441378

After applying the model to 20 test data, the results were obtained which are summarized in the Confusion Matrix in Table XI.

TABLE XI

CONFUSION MATRIX TEST RESULTS |

| Predicted: Stroke | Predicted: No Stroke |

METODE PENGUJIAN CONFUSION MATRIK						
Kelas	Positif	Negatif		Kelas	Positif	Negatif
Positif	TP	FN		Positif	13	1
Negatif	FP	TN		Negatif	2	6
				Akurasi	Presisi	Recall
				86%	87%	93%
						F-Measure
						90%

| Actual: Stroke | TP = [13] | FN = [1] |

| Actual: No Stroke | FP = [2] | TN = [6] |

Based on this matrix, the performance metrics were calculated as follows:

- Accuracy = 86%
- Precision = 87%
- Recall = 93%
- F1-Score = 90%

E. Discussion: This section discusses the meaning of the results obtained, relating them back to the theory and context of the problem.

1. Interpretation of Performance Results: A model accuracy of 86% indicates that the model was able to predict correctly in most cases. However, accuracy can be a misleading metric in imbalanced datasets. In this case, the model could simply predict 'no stroke' for all data and still achieve high accuracy (around 95%). Therefore, the Precision and Recall metrics are more informative. A Recall value of 93% is crucial; this means the model successfully identified 93% of all patients who actually had a stroke. A Precision value of 87% indicates that of all patients predicted to have a stroke, 87% actually had a stroke. There is a trade-off between Precision and Recall that needs to be considered. Increasing Recall (reducing False Negatives) often comes at the expense of Precision (increasing False Positives), and vice versa.
2. Attribute Influence Analysis: By analyzing likelihood probability tables, we can identify the most influential attributes. An attribute is considered influential if the value of $P(\text{attribute}|\text{textstroke}=1)$ is significantly different from $P(\text{attribute}|\text{textstroke}=0)$.
 - a. Age: From Table VI, it can be seen that the probability of being in the 'Elderly' category is significantly higher for stroke=1 compared to stroke=0. This confirms that age is a major risk factor.



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

- b. Hypertension & Heart Disease: Similar to age, having a history of hypertension or heart disease drastically increases the likelihood for stroke=1.
- c. Glucose Level: Patients with glucose levels in the 'Diabetes' category also show a higher probability of having a stroke.

This analysis not only validates the model from a technical standpoint but also demonstrates that the model successfully captures existing medical knowledge from the data.

3. Error Analysis:

- a. False Negatives (FN): Cases where a stroke patient is predicted not to have a stroke. This is the most fatal error. A possible cause could be that the patient has an unusual profile, for example, being young and not having hypertension, but having other risk factors that are underrepresented in the data.
- b. False Positives (FP): Cases where a healthy patient is predicted to have a stroke. This error, while not fatal, can cause anxiety in patients and unnecessary costs for further testing. This can occur if a patient has many common risk factors (elderly, hypertension) but does not experience a stroke.

4. Research Limitations: This study has several limitations that need to be acknowledged:

- a. Naive Bayes Independence Assumption: As discussed, the assumption that all features are independent is not entirely accurate. Complex interactions between risk factors (e.g., obesity increases the risk of hypertension) are ignored by the model.
- b. Class Imbalance: The dataset is highly imbalanced. This can lead the model to favor the majority class ('no stroke'). Imbalance management techniques such as SMOTE (Synthetic Minority Oversampling Technique) were not applied in this study.
- c. Discretization: The process of converting continuous data into categories can lose information. The choice of category boundaries (e.g., the age boundaries for 'Adult' and 'Elderly') is subjective and may affect the results.
- d. Small Scale: The calculations presented are manual and evaluated on a small sample, not through more robust cross-validation on the entire data.

Conclusion and Recommendation

Conclusion

Based on the analysis and testing conducted, the following conclusions can be drawn:



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

1. The Naive Bayes classification model was successfully applied to predict stroke risk with promising performance. The model demonstrated an accuracy of 86% and an F1-score of 90% on the test data.
2. The high recall metric (93%) was a key highlight, indicating that the model was highly effective in identifying the majority of patients truly at risk of stroke. This is crucial in a medical context to minimize missed fatal cases (false negatives).
3. The likelihood probability analysis confirmed that attributes such as age (especially the elderly category), hypertension, heart disease, and high glucose levels were the most significant risk factors increasing the chance of stroke, in line with existing medical knowledge.
4. Despite demonstrating good performance, the model has inherent limitations, particularly the Naive Bayes assumption of independence between attributes, class imbalance in the dataset, and potential information loss due to the discretization process of continuous attributes.

Recommendation

For further research, several steps can be taken to address existing limitations and improve model quality:

1. Addressing Data Imbalance: It is recommended to apply imbalanced data handling techniques such as SMOTE (Synthetic Minority Oversampling Technique). This technique can help the model learn better from the minority class (stroke patients) and potentially reduce the number of false negatives.
2. Using Alternative Algorithms: Future research can compare the performance of Naive Bayes with other classification algorithms that do not assume feature independence, such as Decision Tree, Random Forest, or Support Vector Machine (SVM), which may be able to capture the complex relationships between risk factors.
3. More Robust Model Validation: For a more reliable and generalizable performance evaluation, it is recommended to use cross-validation (k-fold cross-validation) on the entire dataset, rather than just a limited sample of test data.
4. Exploring Discretization Methods: Further studies are needed to investigate the effect of different discretization methods on model performance. Using entropy-based discretization methods or other more objective methods can be considered as an alternative to subjective range division.

References

- [1] World Health Organization, "The top 10 causes of death," 9 Desember 2020. [Online]. Available: <https://www.who.int/news->



Proceedings of the International Multidisciplinary Seminar of ITB AAS Indonesia

Website: <https://prosiding.itbaas.ac.id/index.php/psd>

[room/fact-sheets/detail/the-top-10-causes-of-death](#). [Diakses: 16 Juli 2025].

- [2] Kementerian Kesehatan Republik Indonesia, "Laporan Nasional Riskesdas 2018," Badan Penelitian dan Pengembangan Kesehatan, Jakarta, 2019.
- [3] American Stroke Association, "Stroke Risk Factors," 2021. [Online]. Available: <https://www.stroke.org/en/about-stroke/stroke-risk-factors>. [Diakses: 16 Juli 2025].
- [4] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
- [5] P. Govindarajan et al., "Machine learning to predict functional outcomes in large-vessel-occlusion stroke," *Journal of the American Heart Association*, vol. 9, no. 4, e014113, 2020.
- [6] G. Sailasya and G. L. A. Kumari, "Analysis of Stroke Prediction using Classification Algorithms," *3rd International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 917-922, 2020.
- [7] V. J. Kadam and S. M. Jadhav, "Heart Disease Prediction using Naive Bayes," *International Journal of Computer Applications*, vol. 182, no. 47, pp. 31-35, 2018.
- [8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann Publishers, 2012.
- [9] I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001.
- [10] F. Fedesor, "Stroke Prediction Dataset," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>. [Diakses: 16 Juli 2025].