SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK III - Surabaya, 11 Maret 2023
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2023.4307

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Implementasi Metode Naïve Bayes untuk Diagnosis Penyakit Stroke

Alphinda Rahma Safira Nisa', Hendro Nugroho, Gusti Eka Yuliastuti

Jurusan Teknik Informatika, Institut Teknologi Adhi Tama Surabaya
e-mail: alphindarahmasn@gmail.com

ABSTRACT

Stroke is a disease that is the second-leading cause of death and the third-leading cause of disability in the world. A stroke occurs due to a blockage of blood vessels leading to the brain. The number of stroke diseases is increasing because some people think that the mild symptoms that occur are normal. This happens because the symptoms of a stroke are not diagnosed early. Therefore, the purpose of this study was to detect strokes early using one of the data mining methods. The stroke dataset was derived from Surabaya Hajj Hospital with a total of 100 observations in 10 variables and 2 classification classes i.e., stroke and non-stroke. This research employed the Naive Bayes method and 10-fold cross-validation, producing an accuracy value of 82%. The accuracy resulting from the oversampling and undersampling techniques was 71% and 64%, respectively. The conclusion of the three scenarios stated that standard Naïve Bayes had better performance compared to Naïve Bayes with unbalanced data handling.

Keywords: *Naïve Bayes; stroke; data mining; classification.*

ABSTRAK

Stroke merupakan salah satu penyakit yang menjadi penyebab kematian kedua dan disabilitas ketiga di dunia. Penyakit stroke terjadi karena adanya penyumbatan pembuluh darah yang menuju ke otak. Meningkatnya angka penyakit stroke disebabkan karena sebagian orang menganggap bahwa gejala ringan yang terjadi merupakan hal yang wajar. Oleh karena itu, penelitian ini dilakukan untuk mendiagnosis penyakit stroke secara dini menggunakan salah satu metode pada data *mining*. Dataset stroke yang digunakan yaitu berasal dari Rumah Sakit haji Surabaya dengan jumlah 100 data yang memiliki 10 variabel dan 2 kelas klasifikasi, yaitu stroke dan tidak stroke. Metode yang digunakan pada penelitian ini yaitu *Naïve Bayes*. Hasil penelitian ini menggunakan 10-*Folds Cross Validation* dengan nilai akurasi sebesar 82%. Hasil

akurasi yang dihasilkan dari teknik *oversampling* dan *undersampling* masing-masing sebesar 71% dan 64%. Kesimpulan dari ketiga skenario menyatakan bahwa *Naïve Bayes* standar lebih bagus kinerjanya dibandingkan dengan *Naïve Bayes* dengan penanganan data tidak seimbang.

Kata kunci: *Naïve Bayes*, penyakit stroke, data mining, klasifikasi.

PENDAHULUAN

Stroke merupakan penyakit yang terjadi karena adanya penyumbatan pada pembuluh darah yang menuju ke otak. Berdasarkan data dari WHO, penyakit stroke menjadi penyebab kematian kedua dan disabilitas ketiga di dunia karena setiap tahun kasus stroke semakin meningkat hingga 13,7 juta kasus dan sekitar 5,5 juta angka kematian akibat penyakit stroke [1]. Sebagian orang yang menderita stroke menganggap bahwa gejala ringan yang terjadi sebelumnya merupakan hal yang wajar. Hal tersebut terjadi karena gejala stroke yang dialami tidak dilakukan diagnosis secara dini. Menurut WHO, setiap tahun jumlah orang menderita stroke mencapai 15 juta orang dan sekitar 5 juta orang mengalami kelumpuhan [2].

Stroke menjadi salah satu penyakit yang berfokus pada otak dan sistem saraf paling utama di dunia serta Indonesia merupakan negara di Asia dengan jumlah penderita paling tinggi [3]. Saat ini penanganan seseorang yang menderita stroke juga masih manual dengan datang secara langsung untuk melakukan pemeriksaan yang dilakukan oleh dokter spesialis syaraf. Diagnosis tersebut dilakukan dengan cara mengajukan pertanyaan-pertanyaan tentang keluhan apa saja yang dialami dan nantinya akan menarik sebuah kesimpulan mengenai hasil diagnosis penyakit stroke yang dialami oleh pasien [4].

Faktor-faktor yang menyebabkan stroke yaitu adanya stroke di masa lalu, penyakit jantung, usia, hipertensi, kolesterol, darah tinggi, diabetes, obesitas, gaya hidup, konsumsi alkohol, dan gangguan pembekuan darah [5]. Tingkat kematian dan jumlah orang yang terkena penyakit stroke diperkirakan akan terus bertambah seiring dengan bertambahnya jumlah penduduk dunia. Namun, angka kematian yang lebih tinggi ini dapat angka kematian ini dapat dicegah dengan pengobatan dini dan prediksi dini [6].

Berdasarkan penjelasan yang dipaparkan sebelumnya, maka akan lebih mudah jika ada suatu sistem yang membantu masyarakat awam dalam mendiagnosis stroke secara dini, agar nantinya dapat mengetahui dan melakukan penanganan terhadap penyakit stroke yang dialaminya. Dalam hal ini, diagnosis penyakit stroke dapat dibedakan menjadi 2, yaitu stroke dan tidak stroke. Penentuan metode yang tepat dalam diagnosis penyakit stroke sangat berpengaruh terhadap hasil akhir yang akan ditampilkan.

Metode yang digunakan dalam melakukan diagnosis penyakit stroke yaitu metode *Naïve Bayes* karena metode *Naïve Bayes* mudah dalam proses implementasinya dan metode *Naïve Bayes* hanya membutuhkan jumlah yang sedikit untuk data *training*, sehingga cocok digunakan dalam menentukan parameter yang dibutuhkan dalam proses klasifikasi, serta metode *Naïve Bayes* memiliki tingkat akurasi yang cukup tinggi berdasarkan penelitian-penelitian sebelumnya [7]. Naïve Bayes diintegrasikan dengan lingkungan pemicu dan menunjukkan bahwa metode tersebut adalah solusi prediksi yang baik ketika berhadapan dengan parameter kesehatan [8].

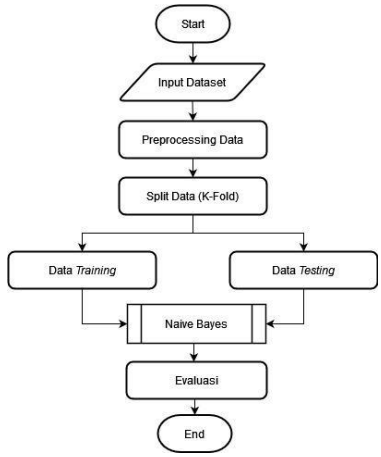
Penelitian yang membandingkan antara metode *Naïve Bayes* dan KNN untuk prediksi penyakit jantung menunjukkan bahwa algoritma *Naïve Bayes* lebih akurat dan lebih baik dalam melakukan klasifikasi untuk penyakit jantung, KNN memiliki tingkat akurasi 80% sedangkan *Naïve Bayes* sebesar 90% [9]. Penelitian selanjutnya yaitu, membahas tentang perancangan sistem untuk memprediksi penyakit jantung menggunakan algoritma data mining yang melibatkan *Naïve Bayes* dengan hasil akurasi yang diperoleh sebesar 85,25% [10]. Penelitian yang membahas tentang diagnosis pertumbuhan janin dalam kandungan yang terhambat pada kehamilan ini menghasilkan tingkat akurasi sebesar 84% dengan *precision* 52,5%, *recall* 86,7%, dan *specificity* sebesar 83,8% [11]. Penelitian lain yang membahas tentang diagnosis penyakit jantung menggunakan metode *Naïve Bayes* juga menghasilkan nilai akurasi sebesar 89.77% [12].

Penelitian selanjutnya yaitu tentang identifikasi penyakit stroke berdasarkan gejala yang dialami menggunakan metode Naïve Bayes menghasilkan akurasi sebesar 80% [13].

Pada penelitian ini, maka penulis melakukan penelitian terkait diagnosis penyakit stroke menggunakan metode *Naïve Bayes*. Dari penelitian sebelumnya yang menggunakan metode *Naïve Bayes* menghasilkan akurasi yang cukup baik, sehingga diharapkan metode *Naïve Bayes* yang digunakan dalam penelitian ini dapat diterapkan dengan baik dalam membantu diagnosis dini penyakit stroke.

METODE

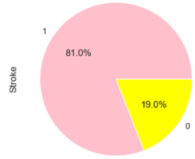
Klasifikasi berhubungan erat dengan pengelompokan, pengelompokan merupakan penentuan kelompok-kelompok data yang serupa sedangkan klasifikasi mempelajari struktur kumpulan data yang sudah dipartisi menjadi kelompok-kelompok yang disebut sebagai kategori atau kelas [14]. Proses klasifikasi metode *Naïve Bayes* digambarkan menggunakan *flowchart* sistem, dimana *flowchart* ini menjelaskan tentang gambaran proses klasifikasi menggunakan metode *Naïve Bayes* untuk menghasilkan keluaran berupa hasil evaluasi model untuk mengukur keberhasilan sistem dalam melakukan diagnosis penyakit stroke. Klasifikasi tersebut digambarkan menggunakan *flowchart* pada Gambar 1.



Gambar 1. Alur Kerja Sistem menggunakan Metode *Naïve Bayes*

Pengumpulan Data

Data yang diperoleh untuk penelitian berasal dari sebuah rumah sakit yaitu Rumah Sakit Umum Haji Surabaya. Dataset terdiri dari 10 variabel seperti jenis kelamin, usia, hipertensi, penyakit jantung, diabetes, GDP (Gula Darah Puasa), LDL (kolesterol jahat), 2JPP (gula darah 2 jam setelah makan), GDS (Gula Darah Sewaktu), dan 1 target indikator penyakit stroke. Perbandingan jumlah kelas data stroke dan tidak stroke dapat dilihat pada Gambar 2 dengan persentase stroke sebesar 81.0% dan tidak stroke 19.0%.



Gambar 2. Persentase Pasien Stroke

***Preprocessing* Data**

Pada tahap ini dilakukan *preprocessing* data untuk mengolah data mentah menjadi data yang efisien dan berkualitas sebelum dilakukan ke proses selanjutnya. Dalam dataset yang digunakan terdapat beberapa data yang tidak konsisten seperti data yang kosong pada suatu variabel. Tahap *preprocessing* juga dilakukan dengan mengubah format tipe data pada suatu data dengan menyesuaikan tipe data yang lainnya.

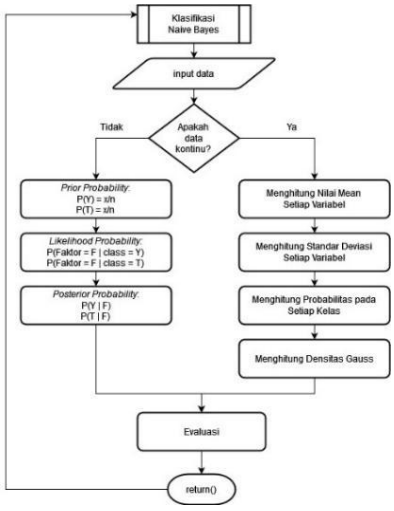
Klasifikasi *Naïve Bayes*

Naïve Bayes merupakan metode klasifikasi yang menggunakan probabilitas sederhana dengan menghitung sekumpulan probabilitas yang menjumlahkan kombinasi nilai dan frekuensi dari dataset yang diberikan, metode *Naïve Bayes* juga memiliki keuntungan mudah untuk dibangun karena tidak membutuhkan estimasi parameter yang rumit dan mudah diterapkan pada kumpulan data yang besar, serta hasil klasifikasi mudah diinterpretasikan oleh orang awam [15]. Persamaan dari metode *Naïve Bayes* dituliskan menggunakan persamaan 1:

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)} \quad (1)$$

Diketahui:
P(A) : Probabilitas hipotesis A (*prior probability*)
P(B) : Probabilitas hipótesis B (*prior probability*)
P(A|B) : Probabilitas hipotesis A jika diberikan kondisi B (*likelihood probability*)
P(B|A) : Probabilitas hipotesis B jika diberikan kondisi A (*posterior probability*)

Proses klasifikasi algoritma untuk mendiagnosis penyakit stroke menggunakan metode *Naïve Bayes* digambarkan dalam *flowchart* pada Gambar 3.



Gambar 3. *Flowchart Naïve Bayes*

K-Fold Cross Validation

Proses validasi klasifikasi metode *Naïve Bayes* menggunakan 10-*Folds Cross Validation*. Hasil yang diperoleh dari pengukuran berupa nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Nilai k yang digunakan yaitu 10, hasil yang diperoleh pada masing-masing iterasi dilakukan perhitungan rata-rata. Skenario dari 10-*Folds Cross Validation* ditunjukkan pada Tabel 1.

Tabel 1. Skenario 10-*Folds Cross Validation*

Fold	***Training* Data**	***Testing* Data**	***Fold***	***Training Data***	***Testing* Data**
1	K2, K3, K4, K5, K6, K7, K8, K9, K10	K1	6	K1, K2, K3, K4, K5, K7, K8, K9, K10	K6
2	K1, K3, K4, K5, K6, K7, K8, K9, K10	K2	7	K1, K2, K3, K4, K5, K6, K8, K9, K10	K7
3	K1, K2, K4, K5, K6, K7, K8, K9, K10	K3	8	K1, K2, K3, K4, K5, K6, K7, K9, K10	K8
4	K1, K2, K3, K5, K6, K7, K8, K9, K10	K4	9	K1, K2, K3, K4, K5, K6, K7, K8, K10	K9
5	K1, K2, K3, K4, K6, K7, K8, K9, K10	K5	10	K1, K2, K3, K4, K5, K6, K7, K8, K9	K10

Confusion Matrix

Metode ini dilakukan untuk mengetahui seberapa baik kinerja model klasifikasi yang digunakan. Pengukuran kinerja klasifikasi menggunakan *confusion matrix* terdiri dari representasi proses klasifikasi, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Tabel *confusion matrix* ditunjukkan pada Tabel 2.

Tabel 2. *Confusion Matrix*

Prediksi	**Aktual**	
	P	**N**
T	TP	TN
F	FP	FN

Pada tabel 2, ditunjukkan bahwa baris pada tabel confusion matrix merupakan nilai pada kelas prediksi sedangkan kolom merupakan nilai dari kelas aktual, dimana:

- TP (True Positive): Banyaknya data positif yang terklasifikasi benar.
- TN (True Negative): Banyaknya data negatif yang terklasifikasi benar.
- FP (False Positif): Banyaknya data positif yang terklasifikasi salah.
- FN (False Negative): Banyaknya data negatif yang terklasifikasi salah

Berdasarkan tabel *confusion matrix* diatas, untuk melakukan evaluasi kinerja klasifikasi dapat dilakukan menggunakan perhitungan menggunakan persamaan sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Skenario Pengujian

Dilakukan 3 skenario pengujian, yang pertama yaitu melakukan klasifikasi *Naïve Bayes* menggunakan 10-*Folds Cross Validation*, yang kedua dilakukan dengan menggunakan dataset yang telah dilakukan proses *oversampling* dengan kelas data sebanyak 81 data stroke dan 81 data tidak stroke, skenario terakhir dilakukan menggunakan dataset yang telah dilakukan proses *undersampling* dengan kelas data masing-masing berjumlah 19 data untuk kelas stroke dan tidak

stroke. Hasil yang digunakan juga dilakukan menggunakan 10-*Folds Cross Validation*. Dari ketiga skenario yang dilakukan akan dibandingkan dari setiap hasil yang diperoleh menggunakan *Confusion Matrix*.

HASIL DAN PEMBAHASAN

Pengolahan Data

Tahap preprocessing dilakukan dengan dua tahap yaitu tahap transformasi dan pembersihan data. *Transformation* data dilakukan untuk mengubah format tipe data pada suatu data dengan menyesuaikan tipe data yang lainnya. Tahap untuk mengubah tipe data pada variabel jenis kelamin dilakukan dengan menggunakan *LabelEncoder()* dari sklearn.preprocessing. *Cleaning* Data dilakukan dengan mengisi data kosong pada variabel LDL menggunakan perhitungan *mean*. Data yang telah dilakukan *preprocessing* data ditunjukkan pada Tabel 3.

Tabel 3. Data yang Sudah Diolah

	Jenis Kelamin	Usia	Hipertensi	Penyakit Jantung	Diabetes	GDP	LDL	2JPP	GDS	Stroke
0	0	43	1	1	1	120	136.0	130	129	1
1	0	67	1	1	1	119	172.0	128	88	1
2	0	54	1	0	0	85	138.2	125	85	1
3	1	66	1	1	0	90	138.2	132	90	1
4	0	48	1	1	1	118	178.0	127	166	1
….	….	….	….	….	….	….	….	….	….	….
95	0	63	1	0	1	116	113.0	162	141	1
96	0	60	0	0	1	278	138.2	225	228	0
97	1	63	0	1	0	112	164.0	212	188	0
98	0	73	1	1	1	192	131.0	314	185	1
99	1	65	1	1	1	221	204.0	334	257	1

Uji Validitas dan Evaluasi

Tahap validasi menggunakan 10-*folds cross validation* dilakukan dengan cara menghitung nilai rata-rata dari keseluruhan nilai pada setiap iterasi. Tabel 4 menunjukkan hasil pengujian akurasi menggunakan 10-*folds cross validation*.

Tabel 4. Hasil Pengujian Akurasi *Naïve Bayes* dengan 10-*Folds Cross Validation*

Data *Testing*	**Akurasi *Naïve Bayes***	**Data *Testing***	**Akurasi *Naïve Bayes***
1	1.0	6	0.8
2	1.0	7	0.5
3	1.0	8	0.5
4	1.0	9	0.8
5	1.0	10	0.6
Rata-Rata			**0.82**

Selanjutnya dilakukan Tahap analisis hasil untuk evaluasi dilakukan dengan membandingkan 3 skenario pengujian Nilai yang dibandingkan berupa nilai *accuracy*, *precision*, *recall*, dan *f1 Score* dari masing-masing skenario. Hasil yang diperoleh dari ketiga skenario klasifikasi dataset stroke menggunakan 10-*Folds Cross Validation* ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian *Naïve Bayes*

UJI VALIDITAS	*NB*	*NB + OVERSAMPLING*	*NB + UNDERSAMPLING*
Accuracy	82%	71%	64%
Precision	82%	57%	65%
Recall	97%	52%	60%
F1-Score	88%	52%	61%

Berdasarkan 3 skenario yang telah dilakukan, dapat disimpulkan bahwa proses pengujian untuk diagnosis penyakit stroke menggunakan metode *Naïve Bayes* yang diperoleh dari proses validasi 10-*Folds Cross Validation* menunjukkan bahwa hasil yang diperoleh dari data awal tanpa dilakukan teknik *oversampling* dan *undersampling* lebih baik daripada menggunakan penanganan untuk data seimbang.

KESIMPULAN

Dalam Tugas Akhir ini telah berhasil melakukan implementasi metode *Naïve Bayes* untuk diagnosis penyakit stroke. Diagnosis penyakit stroke metode *Naïve Bayes* menghasilkan akurasi yang baik sebesar 82% untuk pengujian menggunakan 10-*Folds Cross Validation*. Dari hasil implementasi metode yang digunakan, diketahui bahwa terdapat 3 skenario yaitu menggunakan data awal, *undersampling*, dan *oversampling*. Berdasarkan hasil perbandingan analisa yang telah dilakukan, metode *Naïve Bayes* untuk dataset penyakit stroke yang digunakan menghasilkan nilai akurasi, presisi, *recall*, dan *f1-score* masing-masing sebesar 82%, 82%, 97%, dan 88%. Dataset yang digunakan dalam penelitian ini menggunakan data yang tidak seimbang. Sehingga, penelitian ini dilakukan pengujian data menggunakan *oversampling* dan *undersampling*. *Naïve Bayes* dengan penanganan data seimbang tersebut masing-masing menghasilkan akurasi, presisi, recall, dan f1-scrore, untuk *undersampling* sebesar 64%, 65%, 60%, 61% sedangkan untuk *oversampling* sebesar 71%, 57%, 52%, 52%. Dari hasil tersebut *Naïve Bayes* standar lebih bagus kinerjanya dibandingkan dengan *Naïve Bayes* dengan penanganan data tidak seimbang.

DAFTAR PUSTAKA

[1] Kemenkes RI, “Stroke Dont Be The One.” p. 10, 2018.

[2] N. M. K. Heny Siswanti, S.Kep., *KENALI TANDA GEJALA STROKE*, vol. ث ققثق, no. ثق ثقثقثق. 2021.

[3] F. Rahmawati, Y. V. Via, and E. Y. Puspaningrum, “IMPLEMENTASI METODE NAIVE BAYES DAN CERTAINTY FACTOR DALAM MENDIAGNOSA PENYAKIT KULIT,” vol. 1, no. 1, pp. 631–641, 2020.

[4] J. Kanggeraldo, R. P. Sari, and M. I. Zul, “Sistem Pakar Untuk Mendiagnosis Penyakit Stroke Hemoragik dan Iskemik Menggunakan Metode Dempster Shafer,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 498–505, 2018, doi: 10.29207/resti.v2i2.268.

[5] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[6] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, “A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach,” *Healthc. Anal.*, vol. 2, no. August, p. 100116, 2022, doi: 10.1016/j.health.2022.100116.

[7] F. Fathonah and A. Herliana, “Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid - 19 Menggunakan Metode Naïve Bayes,” *J. Sains dan Inform.*, vol. 7, no. 2, pp. 155–164, 2021, doi: 10.34128/jsi.v7i2.331.

[8] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, “Development of Big Data

Predictive Analytics Model for Disease Prediction using Machine learning Technique," *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1398-y.

[9] D. A. Langga and Dkk, "Perbandingan Algoritma Naive Bayes Dengan Algoritma K-Nearest Neighbor Untuk Prediksi Penyakit Jantung," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.

[10] D. M. Al Hafiz, K. Amaly, J. Jonathan, and M. T. Rachmatullah, "Sistem Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," vol. 2, no. 2, pp. 151–157.

[11] T. Badriyah, "Application of Naive Bayes Method for IUGR (Intra Uterine Growth Restriction) Diagnosis on The Pregnancy," no. June, pp. 12–13, 2020.

[12] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 2019-April, no. Icoei, pp. 292–297, 2019, doi: 10.1109/icoei.2019.8862604.

[13] F. Karim, G. W. Nurcahyo, and S. Sumijan, "Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes," *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 221–226, 2021, doi: 10.37034/jsisfotek.v3i4.69.

[14] A. A. Mahran, R. K. Hapsari, and H. Nugroho, "Penerapan Naive Bayes Gaussian Pada Klasifikasi Jenis Jamur Berdasarkan Ciri Statistik Orde Pertama," *Netw. Eng. Res. Oper.*, vol. 5, no. 2, p. 91, 2020, doi: 10.21107/nero.v5i2.165.

[15] G. E. Yuliastuti, A. N. Alfiyatin, A. M. Rizki, A. Hamdianah, H. Taufiq, and W. F. Mahmudy, "Performance analysis of data mining methods for sexually transmitted disease classification," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 3933–3939, 2018, doi: 10.11591/ijece.v8i5.pp3933-3939.