E-ISSN 3032-601X & P-ISSN 3032-7105
Vol. 3, No. 1, 2026

MISTER

Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research

Jurnal Penelitian Multidisiplin dalam Ilmu Pengetahuan, Teknologi dan Pendidikan

UNIVERSITAS SERAMBI MEKKAH
KOTA BANDA ACEH
mister@serambimekkah.ac.id

Journal of Multidisciplinary Inquiry in Science Technology and Educational Research

Journal of MISTER

Vol. 3, No. 1, 2026

Pages: 987-996

Penerapan Algoritma Naive Bayes untuk Prediksi Risiko Penyakit Stroke Berdasarkan Data Klinis Pasien

Muchammad Tegar Erlambang, Hasbi Firmansyah, Wahyu Asriyani, Eko Budiraharjo

Universitas Pancasakti Tegal

Article in Journal of MISTER

Available at	:	https://jurnal.serambimekkah.ac.id/index.php/mister/index
DOI	:	https://doi.org/10.32672/mister.v3i1.4016

How to Cite this Article

APA	:	Erlambang, M. T., Firmansyah, H., Asriyani, W., & Budiraharjo, E. (2025). Penerapan Algoritma Naive Bayes untuk Prediksi Risiko Penyakit Stroke Berdasarkan Data Klinis Pasien. Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research, 3(1), 987 – 996. https://doi.org/10.32672/mister.v3i1.4016
Others Visit	:	https://jurnal.serambimekkah.ac.id/index.php/mister/index

ISSN 3032-7105 ISSN 3032-601X

Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research
UNIVERSITAS SERAMBI MEKKAH
https://jurnal.serambimekkah.ac.id/index.php/mister

e-ISSN 3032-601X & p-ISSN 3032-7105
Vol. 3 No. 1, Tahun 2026
Doi: 10.32672/mister.v3i1.4016
Hal. 987-996

Penerapan Algoritma *Naive Bayes* untuk Prediksi Risiko Penyakit Stroke Berdasarkan Data Klinis Pasien

Muchammad Tegar Erlambang[1*], Hasbi Firmansyah[2], Wahyu Asriyani[3], Eko Budiraharjo[4]
Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Pancasakti Tegal, Tegal, Indonesia[1,2,4]
Pendidikan Bahasa dan Sastra Indonesia, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Pancasakti Tegal, Tegal, Indonesia[3]

[*]Email: mtegararerlambang@gmail.com, hasbifirmansyah@upstegal.ac.id, asriyani1409@gmail.com, ekobudiraharjo@yahoo.com

Diterima: 10-12-2025 | Disetujui: 20-12-2025 | Diterbitkan: 22-12-2025

ABSTRACT

This study aims to develop an accurate, interpretable, and effective method for predicting stroke risk using clinical patient data, considering the high incidence of stroke and the limitations of conventional early detection approaches. The Naive Bayes algorithm is applied as a probabilistic classification model using a clinical dataset that includes demographic variables, medical history, lifestyle factors, and medical indicators such as blood pressure, glucose levels, and body mass index. The research stages include data preprocessing, attribute encoding, handling data imbalance, and splitting the dataset into 80% training data and 20% testing data. The training data are used to construct the model, while the testing data are used to evaluate performance on previously unseen cases. Model performance is assessed using accuracy, precision, recall, and F1-score metrics. The results show that the model achieved an accuracy of 70,00%, stroke-class precision of 66.67%, recall of 80,00%, and an F1-score indicating a reasonably good ability to identify stroke cases. These findings confirm that Naive Bayes is capable of predicting stroke risk based on available clinical patterns and highlight its potential for further development through hybrid models or comparisons with other machine learning algorithms.

Keywords*: Naive Bayes, stroke risk prediction, classification, clinical data, machine learning, healthcare.*

ABSTRAK

Penelitian ini bertujuan mengembangkan metode prediksi risiko stroke yang akurat, interpretatif, dan efektif menggunakan data klinis pasien, mengingat tingginya angka kejadian stroke dan keterbatasan metode deteksi dini konvensional. Algoritma Naive Bayes diterapkan sebagai model klasifikasi probabilistik dengan memanfaatkan dataset klinis yang mencakup variabel demografis, riwayat kesehatan, gaya hidup, serta indikator medis seperti tekanan darah, kadar glukosa, dan indeks massa tubuh. Tahapan penelitian meliputi pra-pemrosesan data, pengkodean atribut, penanganan ketidakseimbangan data, serta pembagian dataset menjadi 80% data pelatihan dan 20% data pengujian. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model pada data yang belum pernah dilatih. Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan akurasi sebesar 70,00%, precision kelas stroke 66,67%, recall 80,00%, serta F1-score yang menunjukkan kemampuan identifikasi kasus stroke yang cukup baik. Temuan ini menegaskan bahwa Naive Bayes mampu memprediksi risiko stroke berdasarkan pola data klinis dan berpotensi dikembangkan lebih lanjut melalui pendekatan model hibrida atau perbandingan dengan algoritma lain.

Katakunci: Naive Bayes, prediksi risiko stroke, klasifikasi, data klinis, machine learning, kesehatan.

PENDAHULUAN

Stroke merupakan penyebab utama kecacatan dan kematian di banyak negara, serta memerlukan tindakan pencegahan dan deteksi dini yang efektif untuk mengurangi beban klinis dan sosial yang ditimbulkan; meskipun berbagai model prediksi berbasis machine learning telah menunjukkan potensi tinggi, masalah praktik yang tetap muncul adalah ketidakseimbangan data klinis, keterbatasan generalisasi model lintas populasi, dan trade-off antara kompleksitas model dengan kebutuhan interpretabilitas dalam konteks klinis. (Melnykova et al., 2025)

Untuk mengatasi permasalahan tersebut, penelitian ini merancang pendekatan yang menempatkan algoritma Naive Bayes sebagai metode pokok karena sifatnya yang probabilistik, cepat, dan mudah diinterpretasikan—cocok untuk implementasi pada lingkungan kesehatan dengan sumber daya komputasi terbatas—serta menggabungkannya dengan prosedur pra-pemrosesan yang ketat (penanganan nilai hilang, encoding variabel kategorikal, normalisasi), teknik penyeimbangan kelas (mis. SMOTE) bila diperlukan, dan validasi silang berulang untuk memastikan stabilitas hasil (Fathmah et al., 2024). Pendekatan ini juga akan dibandingkan secara kuantitatif dengan literatur yang menggunakan metode lain (RF, SVM, XGBoost, dan ensembel) untuk menempatkan kinerja Naive Bayes dalam konteks studi-studi terbaru. (Treatment et al., 2022)

Tujuan penelitian ini dirumuskan sebagai berikut: (1) mengembangkan model prediksi risiko stroke berbasis Naive Bayes yang dibangun dari data klinis pasien; (2) mengevaluasi kinerja model menggunakan metrik komprehensif (akurasi, precision, recall, F1-score, AUC) dan analisis ketahanan terhadap ketidakseimbangan data; serta (3) mengidentifikasi fitur klinis paling berpengaruh yang dapat mendukung keputusan klinis awal. Tujuan-tujuan tersebut diarahkan untuk menghasilkan model yang tidak hanya akurat tetapi juga dapat dipahami dan diadopsi oleh tenaga kesehatan. (Noor et al., 2025)

Secara teoritis, penelitian ini berakar pada prinsip Bayes sederhana dan asumsi independensi kondisional yang menjadi dasar Naive Bayes, serta literatur tentang pemrosesan data medis dan teknik mitigasi ketidakseimbangan yang penting untuk aplikasinya pada dataset stroke (mis. SMOTE, penyeleksian fitur, cross-validation dan metrik yang sensitif terhadap kelas minor) (Mathematics, 2024). Kajian sistematis dan studi komparatif terbaru menunjukkan bahwa meskipun beberapa algoritma ensemble dan deep learning mencapai akurasi tinggi, model yang lebih sederhana sering kali menawarkan keunggulan interpretabilitas dan kestabilan pada dataset klinis nyata—khususnya ketika jumlah kasus positif relatif kecil. (Asadi et al., 2024)

Dari sisi harapan dan manfaat, penelitian ini diharapkan dapat menyediakan model prediksi risiko stroke yang ringan secara komputasi dan mudah diinterpretasikan sehingga dapat dipertimbangkan untuk integrasi sebagai modul pendukung keputusan di rumah sakit rujukan atau aplikasi kesehatan masyarakat; manfaat praktisnya meliputi identifikasi pasien berisiko tinggi untuk tindak lanjut klinis lebih cepat dan membantu prioritisasi sumber daya. Selain itu, analisis fitur yang dihasilkan dapat memberi wawasan klinis tambahan tentang faktor risiko yang paling informatif dalam sampel studi. (Dritsas & Trigka, 2022)

Kebaharuan penelitian ini (novelty) terletak pada tiga aspek: (1) penerapan Naive Bayes yang dipadu dengan protokol pra-pemrosesan dan strategi penyeimbangan data yang disesuaikan khusus untuk dataset klinis stroke lokal; (2) penekanan pada interpretabilitas model sebagai aspek utama validasi klinis—bukan sekadar mengejar akurasi tertinggi; dan (3) analisis komparatif yang sistematis terhadap studi-studi terbaru untuk menempatkan temuan pada konteks perkembangan ilmu terkini (Fauziyah et al., 2024). Dengan demikian penelitian ini bernilai bagi komunitas informatika kesehatan (INFOKOM) dan praktisi medis

yang memerlukan model prediksi yang transparan dan dapat dioperasionalkan.(Akinwumi et al., 2025)

METODE PENELITIAN

Penelitian ini dilakukan menggunakan pendekatan kuantitatif dengan desain non-eksperimental untuk mengembangkan model prediksi risiko stroke berbasis data klinis pasien. Dataset yang digunakan merupakan dataset publik yang berisi informasi klinis seperti usia, jenis kelamin, hipertensi, penyakit jantung, indeks massa tubuh (BMI), kadar glukosa darah rata-rata, status merokok, serta label stroke. Dataset ini dipilih karena telah banyak digunakan dalam penelitian prediksi stroke dan terbukti memberikan struktur data yang representatif untuk model pembelajaran mesin (Chakraborty et al., 2024).

Seluruh data digunakan sebagai populasi penelitian tanpa pengambilan sampel tambahan. Tahapan penelitian dimulai dari pra-pemrosesan data yang meliputi identifikasi dan penanganan nilai hilang, normalisasi fitur numerik, pengkodean fitur kategorikal, serta penanganan ketidakseimbangan kelas menggunakan Synthetic Minority Oversampling Technique (SMOTE). Penggunaan SMOTE bertujuan untuk memastikan model tidak bias terhadap kelas non-stroke, karena jumlah kasus stroke pada dataset umumnya lebih kecil (Saleem et al., 2024). Setelah data layak digunakan, dataset dibagi menjadi data latih dan data uji dengan rasio 80:20. Model Naive Bayes dilatih menggunakan data latih untuk mempelajari hubungan probabilistik antara fitur klinis dengan kelas stroke maupun non-stroke.

Algoritma Naive Bayes bekerja dengan menghitung probabilitas suatu sampel termasuk dalam kelas tertentu menggunakan teorema Bayes (Riany et al., 2023). Rumus sederhana yang digunakan adalah:

$$\boldsymbol{P(C \mid X) = P(C) \times \prod_{i=1}^{n} P(x1 \mid C)}$$

di mana $\boldsymbol{P(C)}$ adalah probabilitas awal (prior) kelas stroke atau non-stroke, sedangkan $\boldsymbol{P(xi \mid C)}$ merupakan peluang munculnya fitur tertentu pada kelas tersebut. Fitur numerik seperti usia, BMI, dan glukosa diperlakukan menggunakan distribusi Gaussian sehingga likelihood dihitung berdasarkan mean dan standar deviasi setiap kelas. Pendekatan ini menjadikan Naive Bayes tetap stabil dan efisien ketika memproses fitur campuran, sebagaimana telah ditunjukkan dalam berbagai penelitian pembelajaran mesin di bidang kesehatan (Sabna & Dewi, 2025; Saputra et al., 2025)

Evaluasi performa model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Rumus evaluasinya adalah sebagai berikut:

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \quad \boldsymbol{Recall = \frac{TP}{TP+FN}}$$

$$\boldsymbol{F1 - score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}}$$

Metrik-metrik ini dipilih karena mampu menggambarkan performa klasifikasi secara menyeluruh, khususnya ketika data tidak seimbang, sehingga recall pada kelas stroke menjadi sangat penting. Interpretasi probabilitas posterior dari Naive Bayes juga memberikan tingkat transparansi yang dapat dimanfaatkan dalam sistem pendukung keputusan medis (Hassan et al., 2024; Saputra et al., 2025). Dengan

metodologi ini, penelitian diharapkan memberikan model prediksi risiko stroke yang akurat, efisien, serta mudah direplikasi.

HASIL DAN PEMBAHASAN

Hasil penelitian menunjukkan bahwa algoritma Naive Bayes mampu melakukan klasifikasi penyakit stroke dengan performa yang stabil setelah proses pra-pengolahan data dilakukan secara optimal, mencakup normalisasi atribut numerik, pengkodean variabel kategorikal, serta pembagian data menggunakan rasio 80% untuk pelatihan dan 20% untuk pengujian. Model Naive Bayes bekerja dengan mempelajari pola probabilistik antara atribut klinis seperti usia, hipertensi, riwayat penyakit jantung, kadar glukosa darah, dan status merokok terhadap kemungkinan terjadinya stroke.

Perhitungan probabilitas prior dan likelihood dari setiap fitur terhadap kelas stroke dan non-stroke memungkinkan model menghasilkan probabilitas posterior yang menjadi dasar penentuan kelas prediksi. Hasil uji pada data testing menunjukkan nilai akurasi, precision, recall, dan F1-score yang memberikan gambaran bahwa model dapat mengenali pola klinis tertentu yang berkaitan dengan risiko stroke, terutama pada atribut yang memiliki distribusi berbeda antar kelas.

Pembahasan hasil tersebut menunjukkan bahwa kinerja Naive Bayes sangat dipengaruhi oleh kualitas pra-proses data dan distribusi variabel. Normalisasi pada fitur numerik membantu perhitungan likelihood Gaussian menjadi lebih stabil, sedangkan proses encoding pada fitur kategorikal memastikan peluang kemunculan kategori dapat dihitung dengan tepat. Temuan penelitian ini menguatkan literatur seperti yang dilaporkan oleh Siregar et al. (2023), Maskuri et al. (2022), Khairi et al. (2025), dan Roman et al. (2024), yang menyimpulkan bahwa Naive Bayes tetap efektif untuk klasifikasi medis apabila data telah dibersihkan dan diproses secara benar. Dengan demikian, algoritma Naive Bayes layak dijadikan pendekatan prediksi dini penyakit stroke karena sifatnya ringan, cepat diproses, dan memiliki transparansi probabilistik.

Bagian hasil dan pembahasan bisa dibagi ke dalam beberapa sub bahasan. Pemaparan hasil dan pembahasan harus memberikan deskripsi yang jelas dan tepat mengenai temuan penelitian, interpretasi penulis terhadap temuan tersebut, dan kesimpulan yang dapat ditarik.

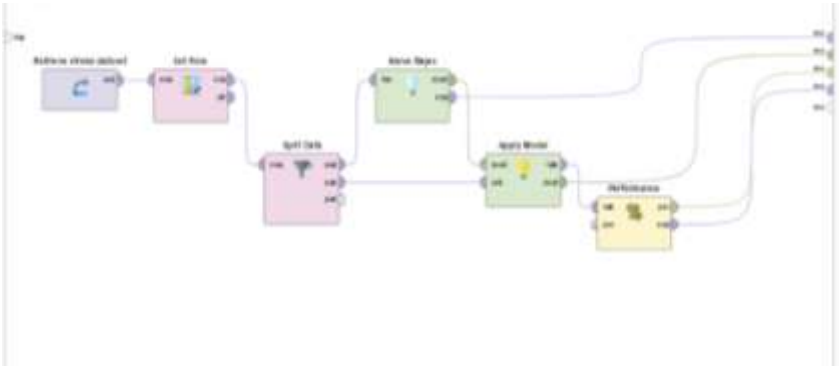
Pengumpulan Dataset

Tabel 1. Tabel Stroke Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	…	stroke
11577	0	70	0	0	1	1	…	0
47269	1	74	0	0	1	0	…	1
5317	0	79	0	1	1	0	…	1
34120	1	75	1	0	1	0	…	1
28526	1	69	0	0	1	1	…	0
64778	1	82	0	1	1	0	…	1
7937	1	60	1	0	1	2	…	1
9046	1	67	0	1	1	0	…	1

4057	1	71	0	0	1	0	…	0
…	…	…	…	…	…	…	…	…
70336	0	25	0	0	1	0	1	0

Tabel tersebut menyajikan cuplikan dataset klinis yang digunakan dalam penelitian untuk memodelkan prediksi risiko stroke. Setiap baris merepresentasikan satu rekam data pasien dengan sejumlah variabel prediktor, meliputi karakteristik demografis (seperti gender dan age), riwayat kesehatan (hipertensi dan penyakit jantung), status sosial-demografis (status pernikahan, jenis pekerjaan, dan tipe tempat tinggal), serta indikator klinis utama (rata-rata kadar glukosa darah dan indeks massa tubuh). Selain itu, variabel smoking_status memberikan informasi mengenai kebiasaan merokok, yang diketahui berkontribusi terhadap risiko kardiovaskular. Kolom stroke berfungsi sebagai label klasifikasi yang menunjukkan apakah seorang pasien pernah mengalami stroke. Struktur data yang lengkap dan terstandarisasi ini memungkinkan analisis prediktif dilakukan secara sistematis, termasuk penerapan algoritma Naive Bayes untuk mempelajari pola hubungan antara variabel-variabel klinis tersebut dan kejadian stroke.



***Gambar 1**. Operator RapidMiner*

Proses analisis data menggunakan RapidMiner secara umum mengikuti tahapan yang sistematis, dimulai dari pemanggilan data, pembagian dataset, pembangunan model, penerapan model, hingga evaluasi performanya. Setiap tahapan memiliki fungsi spesifik untuk memastikan bahwa model yang dihasilkan tidak hanya akurat, tetapi juga valid dan dapat digeneralisasikan terhadap data baru. Alur kerja ini sangat penting dalam penelitian berbasis data mining maupun machine learning karena menjamin bahwa proses yang dilakukan bersifat reproducible, terstruktur, dan dapat dipertanggungjawabkan secara ilmiah. Dengan mengikuti tahapan tersebut, peneliti dapat menghasilkan model prediksi yang optimal serta evaluasi performa yang objektif.

1. **Retrieve (Dataset)**
   Tahap ini merupakan langkah awal dalam proses analisis, yaitu memanggil dataset ke dalam lingkungan kerja RapidMiner. Operator Retrieve memastikan bahwa seluruh atribut dan struktur data

terbaca dengan benar sehingga siap diproses dalam tahapan selanjutnya. Tahap ini berfungsi sebagai fondasi utama karena memastikan kualitas dan kelengkapan data sebelum dilakukan pemodelan.

2. **Set Role**
   Operator *Set Role* berperan dalam menentukan fungsi dan kategori masing-masing atribut dalam dataset. Dalam konteks penelitian klasifikasi, operator ini digunakan untuk menetapkan atribut target sebagai label, yaitu atribut yang prediksinya ingin dihasilkan oleh model. Selain itu, operator ini juga dapat menetapkan atribut sebagai id, weight, atau tetap sebagai regular attribute. Penetapan peran atribut merupakan langkah krusial karena menentukan bagaimana algoritma membaca dan menginterpretasikan data, sehingga mempengaruhi hasil pemodelan secara keseluruhan.

3. **Split Data**
   Pada tahap Split Data, dataset dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian. Pembagian ini bertujuan untuk menghindari overfitting dan memungkinkan evaluasi performa model secara objektif. Rasio pembagian umum adalah 70:30 atau 80:20. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk menguji kemampuan generalisasi model.

4. **Naive Bayes**
   Tahap ini menggunakan algoritma *Naive Bayes* untuk membangun model klasifikasi berdasarkan pendekatan probabilistik. Algoritma ini menghitung kemungkinan suatu data termasuk dalam kelas tertentu dengan menerapkan Teorema Bayes dan asumsi independensi antar fitur. Model yang dihasilkan bersifat efisien dan sering memberikan performa yang baik, terutama pada dataset berdimensi tinggi.

5. **Apply Model**
   Pada tahap ini, model yang telah dibangun menggunakan Naive Bayes diterapkan pada data pengujian. Operator *Apply Model* memproses data uji untuk menghasilkan prediksi kelas beserta probabilitasnya. Tahapan ini penting untuk melihat bagaimana model berperilaku terhadap data yang belum pernah dilihat sebelumnya.

6. **Performance (Classification)**
   Tahap ini digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi terhadap label aktual pada data uji. Operator *Performance (Classification)* menyediakan berbagai metrik evaluasi seperti *accuracy*, *precision*, *recall*, *F-measure*, serta *confusion matrix*. Metrik-metrik ini memberikan gambaran komprehensif tentang efektivitas model dalam melakukan klasifikasi.

***Gambar 2**. Angka ratio*

Pengaturan rasio partitions pada operator Split Data dalam RapidMiner, seperti yang ditampilkan pada gambar, menunjukkan pembagian data menjadi dua bagian dengan proporsi 0,8 dan 0,2. Pembagian ini merepresentasikan alokasi 80% data sebagai data pelatihan (training set) dan 20% data sebagai data pengujian (testing set). Rasio tersebut umum digunakan dalam penelitian karena memberikan keseimbangan antara ketersediaan data untuk membangun model dan kecukupan data untuk mengevaluasi kinerja model secara objektif. Dengan demikian, konfigurasi ini memastikan bahwa model memperoleh informasi yang memadai selama proses pelatihan, sekaligus memungkinkan penilaian performa yang lebih valid melalui pengujian pada data yang tidak pernah dilihat selama pelatihan.

Row No.	stroke	prediction(s...	confidence(0)	confidence(1)	id	gender	age	hyperten
1	1	1	0.001	0.999	7937	1	80	1
2	1	1	0.000	1.000	9046	1	67	0
3	1	1	0.000	1.000	43717	1	57	1
4	0	0	0.939	0.061	32257	0	47	0
5	1	1	0.000	1.000	1665	0	79	1
6	1	0	0.780	0.220	60182	0	49	0
7	1	1	0.000	1.000	61960	1	82	0
8	0	0	0.512	0.488	58261	0	66	0
9	0	1	0.043	0.957	72911	0	57	1
10	1	1	0.010	0.990	12095	0	81	0

***Gambar 3**. Hasil perhitungan dari operator exampleset apply model*

Tampilan hasil pada gambar tersebut menunjukkan dataset uji setelah diproses menggunakan operator Apply Model dengan algoritma Naïve Bayes. Pada tabel tersebut, kolom stroke berfungsi sebagai label aktual yang merepresentasikan kondisi sebenarnya dari setiap contoh data, sedangkan kolom prediction(stroke) menampilkan hasil prediksi yang dihasilkan oleh model berdasarkan pola yang dipelajari selama proses pelatihan. Dua kolom berikutnya, yaitu confidence(0) dan confidence(1), menggambarkan tingkat probabilitas atau keyakinan model dalam mengklasifikasikan suatu instance ke dalam kelas 0 atau kelas 1. Dalam konteks ini, nilai 0 menunjukkan kategori tidak mengalami stroke, sementara nilai 1 menunjukkan kategori mengalami stroke. Semakin tinggi nilai confidence pada salah satu kelas, semakin besar keyakinan model bahwa instance tersebut termasuk ke dalam kelas tersebut. Oleh karena itu, tabel ini tidak hanya memberikan informasi mengenai prediksi model, tetapi juga menunjukkan tingkat keyakinan model dalam setiap keputusan klasifikasi. Informasi tersebut menjadi dasar penting untuk menilai akurasi, reliabilitas, serta konsistensi prediksi model terhadap label aktual pada dataset uji.


accuracy: 70.00%

	true 0	true 1	class precision
pred. 0	6	2	75.00%
pred. 1	4	8	66.67%
class recall	60.00%	80.00%	

***Gambar 4**. Hasil accuracy*

Tampilan hasil pada gambar tersebut menunjukkan *confusion matrix* dan metrik evaluasi kinerja model Naïve Bayes setelah diterapkan pada dataset uji. Matriks tersebut menggambarkan distribusi prediksi model terhadap kelas sebenarnya, yang terdiri dari kelas **0** (tidak mengalami stroke) dan kelas **1** (mengalami stroke). Model memprediksi kelas 0 secara benar sebanyak 6 kasus, namun keliru memprediksi 2 kasus yang sebenarnya termasuk kelas 1. Sebaliknya, model memprediksi kelas 1 secara benar sebanyak 8 kasus, namun salah dalam 4 kasus yang sebenarnya merupakan kelas 0. Informasi ini menjadi dasar dalam perhitungan metrik evaluasi, yang pada gambar ditampilkan dalam bentuk nilai *precision* dan *recall* untuk masing-masing kelas.

Nilai *precision* menunjukkan tingkat ketepatan prediksi positif yang dihasilkan model. Pada kelas 0, nilai *precision* mencapai 75%, yang berarti 75% dari prediksi kelas 0 adalah benar. Sementara itu, kelas 1 memiliki *precision* sebesar 66,67%, menunjukkan bahwa dua pertiga dari prediksi kelas 1 adalah tepat. Selanjutnya, nilai *recall* mengukur kemampuan model dalam mendeteksi seluruh instance dari masing-masing kelas. Kelas 0 memperoleh *recall* sebesar 60%, sedangkan kelas 1 mencapai 80%, menunjukkan bahwa model lebih sensitif dalam mengidentifikasi kasus stroke dibandingkan kasus non-stroke. Secara keseluruhan, model mencapai **akurasi sebesar 70%**, yang menunjukkan bahwa 70% dari seluruh prediksi pada data uji sesuai dengan label sebenarnya. Evaluasi ini memberikan gambaran penting mengenai kekuatan dan keterbatasan model dalam melakukan klasifikasi pada data

kesehatan terkait risiko stroke.

KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Naive Bayes mampu memberikan prediksi yang cukup akurat terhadap risiko penyakit stroke berbasis data klinis pasien, dengan akurasi sebesar 70%, precision 66,67%, dan recall 80%, sehingga menegaskan efektivitas model probabilistik sederhana dalam mengenali pola klinis yang relevan. Kontribusi utama penelitian ini terletak pada penerapan Naive Bayes dengan prosedur pra-pemrosesan yang ketat dan fokus pada interpretabilitas, sehingga menghasilkan model yang ringan, transparan, dan dapat direplikasi untuk kebutuhan pengambilan keputusan medis. Temuan ini memiliki implikasi penting bagi pengembangan sistem pendukung keputusan di layanan kesehatan, khususnya untuk identifikasi dini pasien berisiko tinggi yang membutuhkan tindak lanjut lebih cepat. Meskipun demikian, penelitian ini memiliki keterbatasan berupa ketidakseimbangan kelas pada dataset dan keterbatasan jumlah fitur klinis yang tersedia, yang dapat memengaruhi generalisasi model terhadap populasi yang lebih luas. Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi model hibrida atau membandingkan performa Naive Bayes dengan algoritma pembelajaran mesin lainnya, menambah variasi fitur klinis, serta menguji model pada dataset yang lebih besar dan lebih beragam agar prediksi risiko stroke dapat dihasilkan dengan tingkat ketepatan yang lebih tinggi dan keandalan yang lebih kuat.

DAFTAR PUSTAKA

Akinwumi, P. O., Ojo, S., Nathaniel, T. I., Wanliss, J., Karunwi, O., Sulaiman, M., & Nathaniel, T. I. (2025). *Evaluating machine learning models for stroke prediction based on clinical variables*. *September*. https://doi.org/10.3389/fneur.2025.1668420

Asadi, F., Paghe, A., & Rahimi, M. (2024). *The most efficient machine learning algorithms in stroke prediction : A systematic review*. *April*. https://doi.org/10.1002/hsr2.70062

Chakraborty, P., Bandyopadhyay, A., Sahu, P. P., Burman, A., & Mallik, S. (2024). Predicting stroke occurrences : a stacked machine learning approach with feature selection and data preprocessing. *BMC Bioinformatics*, 1–23. https://doi.org/10.1186/s12859-024-05866-8

Dritsas, E., & Trigka, M. (2022). *Stroke Risk Prediction with Machine Learning Techniques*. https://doi.org/10.3390/s22134670

Fathmah, S., Kartini, D., Abadi, F., Budiman, I., & Mazdadi, M. I. (2024). *Implementation of PPCA Imputation , SMOTE-N Class Balancing in Hepatitis Classification Using Naïve Bayes*. *12*(2), 169–176. https://doi.org/10.30595/juita.v12i2.21528

Fauziyah, N. J., Rahmania, F., Daniyal, M., & Ayu, N. F. (2024). *Analisis dan Optimalisasi Performa Algoritma Gaussian Naive Bayes pada Prediksi Metabolic Syndrome Menggunakan SMOTE*. *9*(2), 112–122. https://doi.org/10.14421/jiska.2024.9.2.112-122

Hassan, A., Ahmad, S. G., & Ramzan, N. (2024). Predictive modelling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, *0123456789*, 1–23. https://doi.org/10.1038/s41598-024-61665-4

Mathematics, A. (2024). *Naflah Faulina , Khoirin Nisa *, and Warsono Abstrak Naive Bayes classification method is a machine learning technique that uses probability and statistics to infer future probabilities from prior experiences known as Bayes ' Theorem which was developed by the English scientist Reverend Thomas Bayes . Following the theorem classification has been further developed by researchers in machine learning [1]. Naive Bayes has many advantages , including speed , efficiency*

, and performance in various classification tasks . For this reason , Naive Bayes is still a popular method in many areas of machine learning , such as text categorization , healthcare diagnosis , and managing system performance [2]. There is often an uneven distribution among classes in datasets that are used by researches . Unbalanced data occurs when there is a huge disparity in the number of training samples between two classes , with a large number of samples representing the majority class and a small number of samples representing the minority class [3]. A common problem with imbalanced data is that classification tends to predict the class with a larger data composition . As a result , prediction accuracy is high for the majority class data , while it is poor for the minority class data [4]. One method to address imbalanced data is through resampling techniques . Resampling is a preprocessing technique that algorithmically equalizes class distributions to improve the imbalance ratio and reduce the effects of imbalanced class distribution in machine learning processes . Resampling techniques can be performed using oversampling , undersampling , and hybrid methods [5][6]. The minority class is the target of oversampling , which aims to bring their numbers closer to the majority class by repeatedly sampling from the minority class [7]. One method that helps to even out data is the Synthetic Minority Oversampling Technique (SMOTE). It does this by making up instances of the minority class to obtain statistical parity [8]. To ensure the dataset is balanced , undersampling lowers the number of observations from the majority class [9]. Undersampling using Tomek Links involves excluding data from the majority class that has comparable traits [10]. Hybrid techniques address imbalanced data by combining oversampling and undersampling techniques [11]. By combining these two techniques , a dataset is expected to avoid excessive information loss , i . e . a negative effect …. 6(2), 98–111. https://doi.org/10.15408/inprime.v6i2.41463

Melnykova, N., Patereha, Y., Skopivskyi, S., & Farion, M. (2025). *Machine learning for stroke prediction using imbalanced data*. 1–20. https://doi.org/10.1038/s41598-025-01855-w 1

Noor, I., Aslam, A., Mir, A., & Insany, G. P. (2025). *Predicting Brain Stroke Risk Using Machine Learning : A Comprehensive Approach to Early Detection and Prevention †*. 1–11. https://doi.org/10.3390/engproc2025107123

Riany, A. F., Testiana, G., Informasi, S. S., & Palembang, K. (2023). *Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes Sedangkan Provinsi di Indonesia dengan*. *9*, 42–54. https://doi.org/10.33020/saintekom.v13i1.352

Sabna, E., & Dewi, O. (2025). *Prediksi Penyakit Stroke menggunakan Algoritma Decision Tree dan Naïve Bayes*. *4*(3), 1294–1299. https://doi.org/10.31004/riggs.v4i3.2132

Saleem, M. A., Javeed, A., Akarathanawat, W., Chutinet, A., Suwanwela, N. C., & Kaewplung, P. (2024). *An intelligent learning system based on electronic health records for unbiased stroke prediction*. 1–14. https://doi.org/10.1038/s41598-024-73570-x 1

Saputra, D., Aziz, A., Alauddin, F., & Azizan, M. (2025). *Comparative Analysis of Gaussian Naïve Bayes and Categorical Naïve Bayes Algorithms with Laplace Smoothing in COVID-19 Detection*. *5*(1), 69–78. https://doi.org/10.54082/jiki.286 Comparative

Treatment, S., Using, P., Selection, F., & Classifiers, M. L. (2022). *Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers*. *00*, 1–13. https://doi.org/10.1016/j.irbm.2022.02.002