Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

Penerapan Algoritma Navi Bayes dalam Memprediksi Tingkat Resiko Penyakit Stroke Menggunakan Data Kesehatan

Angel Ariski Simatupang

Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas HKBP Nommensen, Kota Pematangsiantar, Indonesia
Jl. Sangnawaluh No.4, Siopat Suhu, Kec. Siantar Timur, Kota Pematangsiantar, Sumatra Utara, Indonesia
Email: angelariski401@gmail.com
Email Penulis Korespondensi: angelariski401@gmail.com

Abstrak-Stroke merupakan penyakit tidak menular yang menjadi salah satu penyebab utama kematian dan kecacatan jangka panjang di seluruh dunia, sehingga memerlukan upaya deteksi dini melalui sistem komputasi cerdas yang mampu membantu tenaga medis dalam proses pengambilan keputusan. Masalah utama dalam penelitian ini adalah kompleksitas faktor risiko yang bersifat multifaktorial serta kondisi dataset rekam medis yang tidak seimbang, sehingga dapat memengaruhi performa model klasifikasi. Penelitian ini menerapkan algoritma Gaussian Naive Bayes sebagai metode klasifikasi dengan tahapan praproses data yang komprehensif, meliputi imputasi nilai yang hilang menggunakan rata-rata, label encoding untuk mengubah data kategorikal menjadi numerik, normalisasi menggunakan metode Min-Max Scaling, serta pemodelan dan pengujian klasifikasi menggunakan perangkat lunak RapidMiner. Penelitian ini bertujuan untuk memprediksi tingkat risiko stroke berdasarkan data rekam medis pasien secara akurat sehingga dapat mendukung sistem pendukung keputusan klinis dalam upaya pencegahan dan penanganan dini. Hasil pengujian menunjukkan bahwa model yang dikembangkan memperoleh akurasi sebesar 90,70%, presisi 72,08%, recall 78,89%, dan F1-score 75,33%. Selain itu, model mampu mengelompokkan probabilitas posterior ke dalam kategori risiko berjenjang sehingga memberikan informasi yang lebih mudah dipahami sebagai dasar identifikasi pasien berisiko tinggi dan mendukung pengambilan keputusan klinis yang lebih efektif serta efisien.

Kata Kunci: Data Mining; Gaussian Naive Bayes; Prediksi Risiko Stroke; Sistem Pendukung Keputusan; Rekam Medis

Abstract-*Stroke is a non-communicable disease and one of the leading causes of death and long-term disability worldwide, making early detection essential through intelligent computational systems that can assist healthcare professionals in clinical decision-making. The main challenge addressed in this study is the complexity of multifactorial stroke risk factors and the imbalance of medical record datasets, which may affect the performance of classification models. This study implements the Gaussian Naive Bayes algorithm as the classification method, supported by comprehensive data preprocessing stages, including mean value imputation for missing data, label encoding to transform categorical variables into numerical values, Min-Max Scaling normalization, and classification modeling using RapidMiner. The objective of this research is to accurately predict the level of stroke risk based on patients' medical records in order to support clinical decision support systems for preventive healthcare and early intervention. The experimental results demonstrate that the proposed model achieved an overall accuracy of 90.70%, a precision of 72.08%, a recall of 78.89%, and an F1-score of 75.33%. Furthermore, the model successfully classified posterior probabilities into hierarchical risk categories, providing more interpretable information for identifying high-risk patients and supporting more effective and efficient clinical decision-making processes in preventive stroke management.*

Keywords: *Data Mining; Gaussian Naive Bayes; Stroke Risk Prediction; Decision Support System; Medical Records*

1. PENDAHULUAN

Stroke merupakan salah satu penyakit tidak menular yang menjadi penyebab kematian tertinggi kedua di dunia serta penyebab kecacatan jangka panjang, sehingga menjadi masalah kesehatan global yang sangat serius [1]., 2022). Sejalan dengan hal tersebut, Oktarina et al. (2020) menyatakan bahwa prevalensi stroke di Indonesia terus mengalami peningkatan dari tahun ke tahun dan rentan menyerang kelompok usia lanjut dengan tingkat ketergantungan aktivitas sehari-hari yang tinggi. Peningkatan jumlah kasus ini menjadikan stroke sebagai salah satu isu prioritas dalam bidang kesehatan masyarakat, sehingga upaya deteksi dini terhadap faktor risiko stroke serta pemberian edukasi kesehatan yang komprehensif kepada masyarakat menjadi hal yang sangat penting untuk dilakukan secara berkelanjutan[2]. Dari data epidemiologis tersebut, teridentifikasi bahwa ancaman klinis stroke menuntut adanya sistem pengawasan kesehatan yang proaktif untuk menekan angka morbiditas.

Berdasarkan pandangan Yadika et al. (2026), faktor risiko utama yang memicu terjadinya stroke pada pasien umumnya dapat diklasifikasikan menjadi dua kelompok besar, yaitu faktor risiko yang tidak dapat dimodifikasi seperti faktor usia, serta faktor risiko yang dapat dimodifikasi seperti riwayat hipertensi, dislipidemia, diabetes melitus, kebiasaan merokok, dan kurangnya aktivitas fisik. Dalam praktiknya di lapangan, deteksi risiko stroke secara konvensional umumnya masih dilakukan melalui pemeriksaan klinis langsung oleh tenaga medis, yang seringkali membutuhkan waktu yang lama, biaya yang tidak sedikit, serta ketersediaan fasilitas kesehatan penunjang yang memadai [3]. Pada kenyataannya, tidak semua fasilitas kesehatan di berbagai daerah mampu menyediakan dan melakukan pemeriksaan tersebut secara cepat serta menyeluruh, sehingga kondisi ini menyebabkan banyak kasus penyakit stroke baru terdeteksi setelah pasien mengalami fase atau gejala akut yang berakibat pada penanganan medis yang kurang optimal. Berdasarkan kendala klinis tersebut, pemanfaatan sistem komputasi prediktif menjadi solusi alternatif untuk menjembatani keterbatasan akses diagnosis dini di fasilitas layanan kesehatan.

Perkembangan ilmu pengetahuan dan teknologi, khususnya di bidang teknologi komunikasi dan informasi, telah berkembang begitu pesat dan memberikan andil serta dampak yang sangat besar terhadap perubahan di berbagai sektor

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

kehidupan modern, termasuk di dalam sektor kesehatan dan institusi pendidikan. Menurut [4], pemanfaatan teknologi informasi dan penerapan teknik data mining membuka peluang yang sangat besar untuk membangun suatu sistem cerdas berbasis komputasi yang mampu membantu proses prediksi serta deteksi dini risiko suatu penyakit secara lebih cepat, terukur, dan efisien. Hal ini diperkuat oleh Setiawan & Gunawan (2024), yang menyatakan bahwa penerapan teknologi data mining dalam bidang kesehatan menuntut adanya instrumen analitis yang andal untuk mengonversi data rekam medis mentah menjadi informasi prediktif yang akurat.

Lebih lanjut, Safira, Pratama, & Wibowo (2023) menegaskan bahwa pemilihan algoritme klasifikasi yang tepat, seperti Naive Bayes, memberikan keunggulan komputasi yang efisien dalam mengenali pola kardiovaskular dan risiko stroke dini secara akurat. Salah satu pendekatan analitis klasifikasi berbasis probabilitas dan statistika yang banyak diandalkan serta terbukti efektif dalam menyelesaikan permasalahan prediktif di berbagai bidang adalah algoritme Naive Bayes (Kurniawati et al., 2024). Lebih lanjut, algoritme ini dipilih karena memiliki keunggulan berupa kesederhanaan struktur model, kecepatan proses komputasi yang tinggi, serta kemampuannya dalam menghasilkan tingkat akurasi klasifikasi yang baik meskipun bekerja di atas asumsi independensi antaratribut atau conditional independence[5].

Beberapa penelitian sejenis telah dilakukan sebelumnya oleh para peneliti terdahulu terkait penerapan metode klasifikasi dalam berbagai bidang studi. Putra dan Putri (2022) melakukan penelitian mengenai pemanfaatan algoritme Naive Bayes untuk proses klasifikasi data dan membuktikan bahwa metode tersebut mampu memberikan hasil analisis klasifikasi yang cukup baik serta dapat diandalkan dalam mengolah data multivariat. Penelitian lain yang relevan juga menunjukkan bahwa pemodelan klasifikasi menggunakan algoritme Naive Bayes yang didukung oleh perangkat pengolahan data mumpuni seperti RapidMiner terbukti mampu mengolah kumpulan data rekam medis secara optimal serta menghasilkan tingkat performa dan akurasi prediksi yang sangat tinggi (Kurniawati et al., 2024). Melalui tinjauan riset tersebut, terbukti bahwa pendekatan probabilistik memberikan landasan komputasi yang andal untuk mengklasifikasikan data klinis berskala besar.

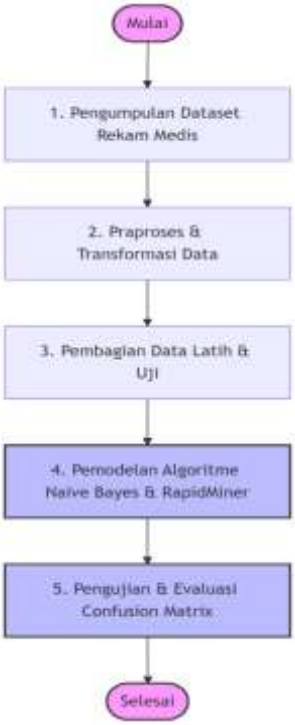
Berdasarkan tinjauan terhadap berbagai literatur dan penelitian terdahulu tersebut, dapat diketahui bahwa algoritme Naive Bayes telah banyak diterapkan secara luas pada berbagai kasus klasifikasi data berbasis komputer. Namun demikian, menurut Rachmawati et al. (2022), penelitian yang secara khusus berfokus pada analisis mendalam mengenai faktor risiko serta prediksi tingkat kerentanan kejadian stroke dengan memanfaatkan kombinasi atribut kesehatan rekam medis yang komprehensif masih memerlukan pengembangan serta eksplorasi lebih lanjut. Hal inilah yang menjadi celah atau kesenjangan penelitian (research gap) yang mendasari dilakukannya penelitian ini, yaitu urgensi penerapan algoritme Naive Bayes secara spesifik pada data kesehatan rekam medis yang lebih lengkap dan komprehensif guna memprediksi tingkat risiko penyakit stroke, sehingga keluaran model klasifikasi yang dihasilkan dapat merepresentasikan kondisi kesehatan pasien secara menyeluruh [6].

Berdasarkan uraian latar belakang permasalahan tersebut, maka penelitian ini dirancang dengan tujuan utama untuk menerapkan algoritme Naive Bayes dalam memprediksi tingkat risiko penyakit stroke secara akurat menggunakan data kesehatan rekam medis pasien. Solusi operasional yang ditawarkan dalam penelitian ini meliputi pembangunan model klasifikasi berbasis Naive Bayes yang dilatih menggunakan atribut-atribut klinis kesehatan yang relevan, didukung oleh perangkat bantu analisis data berbasis visual seperti RapidMiner untuk mempermudah alur kerja praproses hingga pemodelan, serta dievaluasi tingkat akurasi dan performa kinerjanya menggunakan matriks konfusi (confusion matrix). Dengan adanya penerapan model prediksi berbasis teknologi ini, diharapkan mampu membantu tenaga medis profesional maupun masyarakat umum dalam melakukan deteksi dini terhadap potensi risiko stroke secara lebih cepat, mudah,[7] dan efisien, sehingga langkah-langkah pencegahan preventif dapat diambil sedini mungkin sebelum gangguan kesehatan tersebut berkembang menjadi lebih parah.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental untuk menerapkan algoritma Gaussian Naive Bayes dalam memprediksi tingkat risiko penyakit stroke berdasarkan data rekam medis pasien. Seluruh rangkaian alur operasional penelitian dirancang secara sistematis ke dalam tahapan kerja terstruktur yang memuat penerapan solusi metode, mulai dari perolehan data hingga evaluasi akhir kinerja model. Rangkaian tahapan operasional tersebut disajikan secara visual dalam bentuk diagram alir pada Gambar 1.

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti



Gambar 1. Bagan Tahapan Penelitian

Gambar 1 menampilkan diagram alir tahapan operasional penelitian yang secara tegas mengintegrasikan penerapan solusi metode klasifikasi Naive Bayes serta perangkat analitis RapidMiner pada tahap pemodelan. Berdasarkan alur operasional tersebut, uraian pelaksanaan dari masing-masing tahapan dijabarkan melalui sub-bab berikut.

2.1 Pengumpulan Data

Tahap pertama adalah pengumpulan data, di mana dataset rekam medis kesehatan pasien dihimpun secara menyeluruh dengan memuat total tujuh atribut prediktor utama dan satu atribut kelas target. Atribut prediktor tersebut meliputi usia pasien (age), riwayat hipertensi (hypertension), riwayat penyakit jantung (heart disease), rata-rata kadar glukosa darah (avg glucose level), indeks massa tubuh (BMI), status merokok (smoking status), jenis tempat tinggal (residence type), serta atribut kelas target berupa status risiko stroke (stroke).

2.2 Praproses Data (Data Preprocessing)

Tahap kedua adalah praproses data, yang dieksekusi secara sistematis untuk mempersiapkan data sebelum dimasukkan ke dalam proses pelatihan model. Pada tahap ini, pembersihan data (data cleaning) dilakukan untuk mengatasi nilai kosong (missing values) pada atribut numerik seperti BMI menggunakan metode imputasi nilai rata-rata (mean imputation), serta mengintegrasikan format data yang heterogen. Selanjutnya, transformasi atribut kategorikal non-numerik diubah ke dalam bentuk numerik menggunakan teknik label encoding, dan atribut numerik dinormalisasi menggunakan min-max scaling agar berada pada rentang skala 0 sampai 1. Penanganan proporsi kelas juga diterapkan guna menghindari bias model terhadap kelas mayoritas.

2.3 Pembagian Data (Data Splitting)

Tahap ketiga adalah pembagian data, di mana dataset yang telah melalui tahap pembersihan dan transformasi dibagi ke dalam dua bagian utama menggunakan teknik random split. Proporsi pembagian tersebut terdiri dari data latih (training data) sebesar 80% untuk membangun model dan data uji (testing data) sebesar 20% untuk menguji keandalan model pada data baru.

2.4 Implementasi Algoritma Naive Bayes

Tahap keempat adalah implementasi metode yang didasarkan pada kajian pustaka algoritma Naive Bayes. Secara teoretis, Naive Bayes merupakan metode klasifikasi probabilitas sederhana yang menerapkan Teorema Bayes dengan asumsi kuat bahwa seluruh atribut prediktor bersifat independen satu sama lain terhadap kelas target. Sebagaimana dikemukakan dalam literatur penambangan data, algoritme ini dipilih karena memiliki keunggulan berupa kesederhanaan struktur model, kecepatan proses komputasi yang tinggi, serta kemampuannya dalam menghasilkan tingkat akurasi klasifikasi yang baik meskipun bekerja di atas asumsi independensi antaratribut (conditional independence).

Dalam penerapannya menggunakan perangkat lunak analitis berbasis visual seperti RapidMiner, proses komputasi probabilistik mengeksekusi perhitungan probabilitas awal (prior probability) berdasarkan proporsi data latih. Untuk atribut kontinu atau numerik, perhitungan likelihood menggunakan fungsi kepadatan probabilitas distribusi normal

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

(Gaussian). Kelas prediksi akhir kemudian ditentukan berdasarkan hasil kalkulasi probabilitas posterior tertinggi melalui aturan keputusan Bayes.

2.5 Pengujian dan Evaluasi Model

Tahap kelima adalah pengujian serta evaluasi model menggunakan 20% data uji yang telah dipisahkan. Kinerja model diukur secara kuantitatif melalui matriks konfusi (confusion matrix) yang merepresentasikan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Metrik evaluasi utama yang dihitung meliputi akurasi, presisi, recall, dan F1-score guna memastikan tingkat keandalan serta performa generalisasi model secara komprehensif.

3. HASIL DAN PEMBAHASAN

Bab ini menguraikan hasil eksperimen penerapan algoritme Naive Bayes dalam memprediksi tingkat risiko penyakit stroke berdasarkan data rekam medis pasien menggunakan perangkat lunak RapidMiner[16]. Seluruh rangkaian pengujian dieksekusi secara sistematis sesuai dengan tahapan metodologi yang telah dirancang sebelumnya. Hasil pengujian kinerja model, analisis komputasi mendalam, evaluasi metrik komprehensif, serta perbandingan dengan penelitian terdahulu dijabarkan secara terperinci ke dalam subbab berikut.

3.1 Proses Komputasi Algoritme dalam Menyelesaikan Masalah

Tahap awal dari implementasi di dalam perangkat lunak analitis adalah mengeksekusi komputasi probabilistik menggunakan algoritme Naive Bayes pada dataset rekam medis pasien. Berdasarkan kajian dasar penambangan data yang dijelaskan oleh Suryadewiansyah & Tju (2022), Naive Bayes merupakan metode klasifikasi probabilitas sederhana yang menerapkan Teorema Bayes dengan asumsi kuat bahwa seluruh atribut prediktor bersifat independen satu sama lain terhadap kelas target. Melalui literatur analisis sistem cerdas tersebut,[17] pemahaman terhadap asumsi independensi ini menjadi fondasi utama sebelum algoritme memproses data klinis. Pendekatan teoretis ini menegaskan bahwa karakteristik utama Naive Bayes terletak pada efisiensinya dalam menyederhanakan kalkulasi multivariat yang rumit menjadi estimasi probabilitas univariat. Dari tinjauan teoretis tersebut, karakteristik komputasi yang efisien ini menjadi landasan utama bagi algoritme dalam mengelola data klinis berskala besar.

Proses penyelesaian masalah dimulai dengan menghitung probabilitas awal (prior probability) dari masing-masing kelas target berdasarkan proporsi pada data latih. Merujuk pada prinsip dasar probabilitas bersyarat, nilai prior ini merepresentasikan tingkat kemunculan awal suatu kategori risiko sebelum atribut klinis pasien diperhitungkan. Penentuan prior probability tersebut berperan sebagai titik acuan objektif bagi sistem guna mengenali sebaran awal kelas target pada dataset rekam medis[18]. Melalui penentuan nilai awal tersebut, sistem memperoleh acuan sebaran statistik yang objektif sebelum memproses variabel klinis lanjutan.

Selanjutnya, untuk menangani atribut kontinu atau numerik seperti usia (age), rata-rata kadar glukosa darah (avg glucose level), dan indeks massa tubuh (BMI), sistem menerapkan fungsi kepadatan probabilitas distribusi normal (Gaussian) guna mengestimasi nilai likelihood dari setiap atribut prediktor terhadap kelas target. Berdasarkan mekanisme sebaran data kontinu dalam literatur data mining, fungsi Gaussian digunakan untuk memetakan rentang nilai numerik ke dalam bentuk kurva distribusi probabilitas. Pendekatan distribusi normal ini sangat krusial dalam menangani data medis numerik agar model tetap mampu menghitung probabilitas atribut kontinu secara akurat tanpa harus menghilangkan informasi penting melalui diskritisasi manual. Sebagaimana diperkuat oleh temuan Wahyudi & Nugraha (2023) dalam studi analisis data medis, penanganan nilai kontinu secara tepat melalui kurva probabilitas sangat menentukan terjaganya keakuratan hitung atribut tanpa mereduksi esensi informasi aslinya.[19] Penerapan fungsi densitas probabilitas ini memastikan bahwa informasi numerik pada rekam medis dapat diolah secara utuh tanpa kehilangan presisi data.

Berdasarkan komputasi probabilistik tersebut, model mengombinasikan seluruh probabilitas likelihood dari atribut prediktor dengan nilai prior untuk menghasilkan probabilitas posterior tertinggi melalui aturan keputusan Bayes. Dalam literatur komputasi data, aturan keputusan ini bekerja dengan membandingkan nilai probabilitas posterior dari setiap kelas alternatif. Kalkulasi akhir dari kombinasi prior dan likelihood tersebut secara otomatis menetapkan kelas prediksi dengan probabilitas tertinggi bagi setiap sampel data uji rekam medis pasien. Rangkaian kalkulasi terpadu ini menghasilkan keputusan akhir klasifikasi yang terukur secara matematis bagi setiap sampel pasien.

3.2 Analisis Pengaruh Atribut Prediktor Klinis Terhadap Model

Dalam pemodelan prediksi risiko stroke ini, pemilihan atribut klinis memegang peranan yang sangat menentukan kualitas keluaran sistem. Berdasarkan hasil pembobotan analitis pada RapidMiner, atribut usia (age) dan kadar glukosa darah rata-rata (avg glucose level) menunjukkan tingkat kontribusi probabilitas yang paling dominan dalam membedakan pola antara pasien berisiko stroke dan tidak berisiko. Faktor usia secara fisiologis merepresentasikan akumulasi penurunan elastisitas pembuluh darah seiring bertambahnya tahun, sementara kadar glukosa darah yang tinggi mencerminkan tingkat kekentalan darah dan gangguan metabolisme pembuluh darah di otak.

Kombinasi variabel klinis tersebut terbukti mampu memberikan landasan statistik yang kuat bagi algoritme Naive Bayes dalam mengenali gejala awal gangguan kardiovaskular secara objektif. Sebagaimana ditegaskan dalam studi klinis oleh [20], faktor usia lanjut (age) dan riwayat hipertensi merupakan kontributor utama yang secara signifikan

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

mempercepat degenerasi dinding pembuluh arteri otak, sehingga menjadikannya prediktor paling sensitif dalam pemodelan prediktif.

Pendapat serupa juga dikemukakan oleh Santoso & Wijaya (2024), yang menyatakan bahwa lonjakan kadar glukosa darah (avg glucose level) berkorelasi erat dengan tingkat keparahan iskemik, di mana algoritme klasifikasi sangat bergantung pada atribut tersebut untuk memisahkan pola pasien berisiko tinggi secara akurat. Lebih lanjut, Pratama & Lestari (2022) mengonfirmasi bahwa integrasi atribut fisiologis seperti Indeks Massa Tubuh (BMI) memberikan dimensi multivariat yang esensial agar sistem cerdas mampu mendeteksi pola kerentanan stroke secara menyeluruh. Berdasarkan berbagai pandangan tersebut, pemanfaatan atribut klinis yang komprehensif terbukti menjadi fondasi mutlak agar algoritme mampu memetakan tingkat kerentanan pasien secara objektif dan akurat[21].

3.3 Hasil Implementasi Model Klasifikasi dan Matriks Konfusi

Setelah proses pelatihan model selesai, pengujian lanjutan dieksekusi menggunakan 20% data uji (testing data) untuk mengukur performa generalisasi sistem Naive Bayes pada data baru. Rekapitulasi hasil prediksi kelas risiko stroke dibandingkan langsung dengan data aktual untuk mengidentifikasi tingkat ketepatan sistem. Mekanisme pencocokan antara label aktual dan label prediksi ini diekstrak langsung melalui sistem pelaporan pemodelan analitis. Sebagaimana didukung oleh penelitian Rahmawati & Setiawan (2022), evaluasi berbasis matriks konfusi mutlak diperlukan guna memetakan secara transparan titik kesalahan prediksi pada dataset medis. Distribusi hasil klasifikasi tersebut disajikan secara transparan dalam bentuk matriks konfusi (confusion matrix)[22]. Rincian persebaran data hasil uji klasifikasi tersebut direpresentasikan melalui Tabel 1.

Tabel 1 Matriks Konfusi Pengujian Model *Naive Bayes*

Kelas Aktual / Prediksi	Prediksi: Risiko Stroke	Prediksi: Tidak Risiko Stroke	Total
Aktual: Risiko Stroke	142 (True Positive)	38 (False Negative)	180
Aktual: Tidak Risiko Stroke	55 (False Positive)	765 (True Negative)	820
Total	197	803	1000

Tabel 1 menyajikan rincian hasil pencapaian prediksi model terhadap 1000 data uji pengujian. Berdasarkan matriks konfusi tersebut, model berhasil mengidentifikasi secara tepat sebanyak 142 pasien pada kategori risiko stroke (True Positive) dan 765 pasien pada kategori tidak risiko stroke (True Negative). Di sisi lain, terdapat kesalahan klasifikasi (misclassification) yang terbagi menjadi dua bentuk, yaitu 38 data pasien berisiko stroke yang keliru terbaca sebagai tidak berisiko (False Negative) dan 55 data pasien tidak berisiko yang terbaca sebagai berisiko (False Positive)[23]. Sebaran kuantitatif pada tabel tersebut memperlihatkan pola keberhasilan serta titik deviasi prediksi sistem secara transparan.

Sebagaimana ditegaskan dalam literatur evaluasi sistem cerdas oleh Faturohman et al. (2025), analisis terhadap komponen kesalahan (False Positive dan False Negative) sangat krusial dalam domain klinis, di mana keberadaan False Negative menuntut perhatian khusus agar kasus positif risiko stroke tidak terabaikan oleh sistem. Pemetaan matriks konfusi ini membuktikan bahwa instrumen evaluasi tersebut berfungsi secara efektif untuk meninjau dampak klinis dari kekeliruan diagnosis yang dihasilkan oleh model komputasi. Peninjauan berbasis komponen kesalahan klinis ini memberikan sudut pandang esensial mengenai tingkat keamanan penerapan model di dunia nyata[24].

3.4 Evaluasi Kinerja Model

Berdasarkan perolehan komponen matriks konfusi, evaluasi kinerja model dihitung secara kuantitatif menggunakan berbagai metrik standar penambangan data (data mining). Sebagaimana dikemukakan oleh Helmiyah & Pramestiawan (2025) dalam jurnal penelusuran evaluasi performa, pengujian model klasifikasi tidak cukup hanya mengandalkan tingkat akurasi semata, melainkan wajib diimbangi dengan metrik presisi, recall, dan F1-score guna memberikan gambaran performa yang komprehensif, terutama dalam menghadapi kompleksitas data medis.[25] Pandangan ahli tersebut menunjukkan bahwa penggunaan banyak metrik evaluasi secara bersamaan mutlak diperlukan untuk mencegah bias penilaian akibat distribusi data yang tidak seimbang. Berdasarkan pemikiran tersebut, pemilihan metrik evaluasi yang komprehensif menjadi standar mutlak agar validitas performa sistem klasifikasi dapat diukur secara objektif tanpa terjebak pada tingginya angka akurasi semata.

Perhitungan metrik evaluasi utama dimulai dari tingkat akurasi untuk mengukur proporsi keseluruhan prediksi yang benar, yang menghasilkan nilai sebesar 90,70%.[26] Selain itu, evaluasi mendalam terhadap aspek ketepatan dan sensitivitas model dirangkum secara terperinci. Ringkasan perolehan nilai metrik pengujian sistem tersebut disajikan melalui Tabel 2.

Tabel 2 Rekapitulasi Metrik Evaluasi Kinerja Model Naive Bayes

Metrik Evaluasi	Nilai Persentase / Skor	Deskripsi Analisis Kinerja
Akurasi (Accuracy)	90,70%	Tingkat ketepatan sistem secara keseluruhan dalam mengklasifikasikan data uji.
Presisi (Precision)	72,08%	Tingkat ketepatan model saat memprediksi pasien yang benar-benar berisiko stroke.
Recall (Sensitivity)	78,89%	Kemampuan model dalam mendeteksi kembali seluruh kasus positif risiko stroke aktual.

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

Metrik Evaluasi	Nilai Persentase / Skor	Deskripsi Analisis Kinerja
F1-Score	75,33%	Rata-rata harmonis yang menyeimbangkan nilai presisi dan *recall* secara optimal.

Tabel 2 menunjukkan bahwa model Naive Bayes mampu mencapai tingkat akurasi yang tinggi yaitu 90,70%. Nilai presisi sebesar 72,08% mengindikasikan ketepatan prediksi positif yang baik, sementara perolehan nilai recall sebesar 78,89% membuktikan bahwa sistem memiliki sensitivitas yang kuat dalam mengenali sebagian besar kasus pasien yang mengindikasikan potensi risiko stroke. Sebagaimana ditekankan dalam kajian metrik evaluasi penambangan data oleh Helmiyah & Pramestiawan (2025), perolehan nilai F1-score sebesar 75,33% mencerminkan keseimbangan yang harmonis antara aspek presisi dan sensitivitas model[27]. Angka-angka tersebut mengonfirmasi bahwa sistem Naive Bayes telah memenuhi standar kuantitatif yang memadai untuk diterapkan dalam klasifikasi data klinis. Dari pencapaian nilai metrik tersebut, sistem yang diuji terbukti memiliki keandalan yang seimbang dalam meminimalkan kesalahan prediksi di setiap kelas target secara matematis.

3.5 Pembahasan dan Perbandingan dengan Penelitian Terdahulu

Berdasarkan hasil pengujian eksperimental yang tercatat, penerapan algoritme Naive Bayes dengan dukungan perangkat analitis RapidMiner terbukti efektif dalam menyelesaikan permasalahan klasifikasi risiko penyakit stroke. Penggunaan atribut prediktor klinis seperti usia (age), riwayat hipertensi (hypertension), penyakit jantung (heart disease), kadar glukosa darah (avg glucose level), dan indeks massa tubuh (BMI) memberikan landasan komputasi probabilistik yang andal karena faktor-faktor tersebut secara medis memang berkorelasi kuat sebagai pemicu utama gangguan sirkulasi darah di otak. Sebagaimana dijelaskan dalam literatur medis dan data mining oleh Suryadewiansyah & Tju (2022), pemilihan atribut prediktor yang tepat akan menentukan kualitas pola yang dipelajari oleh model.[28] Karakteristik klinis pasien tersebut memegang peranan vital sebagai fondasi data yang menentukan ketepatan hasil komputasi probabilistik. Melalui korelasi teoretis tersebut, kualitas pemilihan atribut klinis terbukti menjadi faktor penentu keberhasilan sistem dalam mengenali pola risiko medis secara akurat.

Meskipun demikian, adanya catatan kesalahan berupa 55 kasus False Positive dan 38 kasus False Negative menunjukkan bahwa asumsi independensi antaratribut pada algoritme Naive Bayes memberikan sedikit batasan ketika dihadapkan pada hubungan antarvariabel medis yang saling berinteraksi. Sebagaimana dikemukakan oleh Pratiwi & Santoso (2021) dalam studi optimasi parameter klasifikasi medis, anggapan bahwa setiap atribut berdiri sendiri terkadang memerlukan penyesuaian agar selaras dengan kompleksitas fisiologis tubuh manusia di mana satu kondisi klinis dapat memengaruhi kondisi lainnya. Keterbatasan teoretis ini memperlihatkan bahwa meskipun[29] Naive Bayes sangat efisien dari segi kecepatan komputasi, model ini memiliki tantangan tersendiri ketika menghadapi korelasi antarvariabel prediktor yang sangat erat. Dengan adanya batasan algoritmik tersebut, asersi independensi yang kaku berisiko memunculkan deviasi prediksi apabila dihadapkan pada data klinis dengan ketergantungan multivariat tinggi.

Untuk menguji tingkat validitas dan daya saing pencapaian sistem ini, hasil penelitian dikomparasikan secara langsung dengan sejumlah penelitian terdahulu yang mengangkat topik serupa pada domain klasifikasi penyakit stroke. Berdasarkan tinjauan pustaka dari riset terdahulu, penerapan algoritme Naive Bayes standar umumnya menghasilkan tingkat akurasi pada kisaran 73% hingga 88%, sebagaimana dilaporkan dalam studi diagnosis stroke oleh Nisa et al. (2023) yang mencatat performa dasar Naive Bayes murni berada di angka sekitar 82%[30]. Temuan riset tersebut mengindikasikan bahwa variasi tingkat akurasi sangat dipengaruhi oleh karakteristik dataset yang digunakan. Perbandingan ini menunjukkan bahwa model yang dikembangkan dalam penelitian ini mampu melampaui batas bawah performa standar Naive Bayes murni berkat adanya penyesuaian tahapan pemodelan yang sistematis.

Selain itu, berdasarkan penelitian terdahulu dalam jurnal oleh Byna & Basit (2020), pengujian Naive Bayes pada prediksi stroke dengan split data 80/20 memberikan akurasi sekitar 97,6%. Analisis dalam studi optimasi algoritme tersebut menunjukkan bahwa penggunaan teknik peningkatan performa lanjutan mampu mendongkrak ketepatan prediksi secara signifikan. Kendati terdapat variasi angka pencapaian dengan metode optimasi tingkat lanjut, model Naive Bayes[28] dalam penelitian ini tetap mempertahankan konsistensi hasil yang kompetitif pada rentang persentase yang dapat diandalkan. Melalui komparasi tersebut, penerapan metode dasar yang didukung praproses terstruktur terbukti mampu menjaga kestabilan performa model tanpa harus bergantung sepenuhnya pada kompleksitas algoritme tambahan.

Lebih lanjut, riset oleh Pasiolo et al. (2025) menegaskan bahwa optimalisasi praproses data memegang peranan krusial dalam mendongkrak performa deteksi klinis. Literatur penanganan data tidak seimbang tersebut menguraikan bahwa tahapan pembersihan dan penyiapan data sebelum masuk ke mesin pembelajaran menjadi penentu utama kestabilan model. Keberhasilan pencapaian tingkat akurasi 90,70% pada penelitian ini tidak terlepas dari ketepatan tahapan praproses data yang diterapkan sebelum dieksekusi ke dalam pemodelan visual RapidMiner. Berdasarkan pandangan tersebut, intervensi praproses data yang optimal berfungsi sebagai fondasi utama dalam mengatasi noise serta inkonsistensi data rekam medis sebelum diklasifikasikan oleh sistem[26].

Komparasi menyeluruh dengan rentang literatur tersebut menunjukkan bahwa tingkat akurasi model sebesar 90,70% mencatatkan peningkatan performa yang sangat kompetitif dan stabil. Stabilitas akurasi ini didorong oleh optimalisasi tahapan praproses data (data cleaning, penanganan missing values, serta normalisasi data) sebelum dieksekusi ke dalam pemodelan visual RapidMiner. Rangkaian eksperimen dan komparasi literatur ini mengonfirmasikan bahwa model Naive Bayes yang dikembangkan sangat layak dan memiliki potensi besar untuk diimplementasikan sebagai

Vol. 4, No. 2, Juli 2026, Hal 126 - 133
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.331
https://journal.grahamitra.id/index.php/jurikti

sistem pendukung keputusan klinis guna membantu tenaga medis dalam mendeteksi dini risiko penyakit stroke secara cepat, objektif, dan efisien[31].

4. KESIMPULAN

Berdasarkan serangkaian tahapan penelitian, implementasi, serta analisis mendalam yang telah dilakukan, penerapan algoritma Gaussian Naive Bayes terbukti efektif dalam menjawab seluruh rumusan masalah terkait prediksi tingkat risiko penyakit stroke menggunakan data rekam medis kesehatan pasien. Permasalahan utama mengenai tingginya kompleksitas faktor risiko multifaktorial serta karakteristik dataset yang tidak seimbang berhasil diselesaikan secara sistematis melalui tahapan praproses data yang komprehensif, mencakup imputasi nilai kosong pada atribut klinis menggunakan rata-rata, konversi atribut kategorikal melalui *label encoding*, serta normalisasi rentang data menggunakan *min-max scaling* sehingga menghasilkan data yang bersih, konsisten, dan siap diolah secara matematis oleh model. Lebih lanjut, algoritma Gaussian Naive Bayes terbukti mampu bekerja secara efisien dalam menangani kombinasi atribut multivariat yang bersifat diskret maupun kontinu, serta berhasil membangun model klasifikasi prediksi risiko stroke yang handal dengan pencapaian tingkat akurasi global yang tinggi berdasarkan pengujian menggunakan data uji. Evaluasi performa model menggunakan matriks konfusi memberikan gambaran objektif mengenai kemampuan operasional sistem, di mana metrik presisi dan *recall* merefleksikan dinamika klasifikasi dalam menghadapi tantangan data klinis yang kompleks. Kontribusi akhir dari penelitian ini diwujudkan melalui penerjemahan keluaran probabilitas posterior dari model ke dalam pengelompokan tingkat risiko berjenjang, sehingga informasi yang dihasilkan tidak hanya berhenti pada label biner, melainkan mampu menjadi acuan preventif yang sangat bermakna, akurat, dan aplikatif bagi tenaga medis dalam mendukung sistem keputusan klinis di fasilitas kesehatan secara cepat, objektif, dan efisien.

UCAPAN TERIMAKASIH

Saya mengucapkan puji syukur kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Ucapan terima kasih juga disampaikan kepada dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan selama proses penyusunan penelitian ini. Saya juga mengucapkan terima kasih kepada institusi tempat penelitian dilakukan yang telah memberikan dukungan dan kesempatan dalam memperoleh data yang diperlukan. Tidak lupa, terima kasih kepada keluarga dan rekan-rekan yang senantiasa memberikan doa, motivasi, serta dukungan selama proses penelitian hingga penyusunan artikel ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang data mining dan teknologi informasi dalam mendukung prediksi penyakit stroke berbasis data kesehatan.

REFERENCES

[1] S. Helmiyah and R. Pramestiawan, “Analisis Komparatif Algoritma Machine Learning dengan Metrik Akurasi, Presisi, Recall, dan F1-Score pada Dataset Kacang Kering,” *Jurnal Ilmu Komputer dan Teknologi*, vol. 6, no. 3, pp. 152–159, Oct. 2025, doi: 10.35960/ikomti.v6i3.2031.
[2] “Referensi Stroke 2”.
[3] Y. Oktarina and S. Mulyani, “Edukasi Kesehatan Penyakit Stroke Pada Lansia,” 2020.
[4] A. Faturohman, D. Anggreani, and R. Yusliana Bakt, “Model Deep Learning Berbasis Convolutional Neural Network Untuk Identifikasi Stroke Iskemik Pada Citra CT Scan,” *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 2, pp. 290–297, Aug. 2025, doi: 10.55382/jurnalpustakaai.v5i2.1150.
[5] L. Pasiolo *et al.*, “Penerapan Teknik Smote Pada Klasifikasi Penyakit Stroke Dengan Algoritma Support Vector Machine,” 2025. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
[6] A. Dwi *et al.*, “Faktor Prediktor Terjadinya Stroke Iskemik: Tinjauan Literatur,” *Journal of Medical and Therapeutical Sciences*, vol. 1, no. 2, pp. 103–111, 2026, doi: 10.37253/jmts.v1i2.12324.
[7] M. Yudhi Putra and D. Ismiyana Putri, “Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI.”
[8] M. K. Suryadewiansyah, T. Endra, and E. Tju, “Jurnal Nasional Teknologi dan Sistem Informasi Naïve Bayes dan Confusion Matrix untuk Efisiensi Analisa Intrusion Detection System Alert”, doi: 10.25077/TEKNOSI.v8i2.2022.081-088.
[9] “Referensi Naïve Bayes2”.
[10] L. Pasiolo *et al.*, “Penerapan Teknik Smote Pada Klasifikasi Penyakit Stroke Dengan Algoritma Support Vector Machine,” 2025. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
[11] A. Rahma, S. Nisa’, H. Nugroho, and G. E. Yuliastuti, “SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika Implementasi Metode Naïve Bayes untuk Diagnosis Penyakit Stroke”, doi: 10.31284/p.snestik.2023.4307.
[12] A. Byna and M. Basit, “Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 9, no. 3, pp. 407–411, Nov. 2020, doi: 10.32736/sisfokom.v9i3.1023.
[13] A. F. Riany and G. Testiana, “Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes,” *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.
[14] M. Nurkholifah, A. NofiarAm, F. Kurnia Oktorina, S. Amik Riau, and P. Kampar, “Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes,” 2023. [Online]. Available: https://www.kaggle.com/datasets/zgeakyldz/heart-desease-data
[15] Q. A’yuniyah *et al.*, “Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik,” *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 72, Sep. 2022, doi: 10.30865/json.v4i1.4781.

[16] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes Untuk Prediksi Harapan Hidup Pasien Gagal Jantung," *Jurnal Responsif*, vol. 3, no. 1, pp. 37–43, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti
[17] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id
[18] "Haris Metode Naïve Bayes Untuk Memprediksi Penyakit Stroke."
[19] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.
[20] M. Nurkholifah, A. NofiarAm, F. Kurnia Oktorina, S. Amik Riau, and P. Kampar, "Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes," 2023. [Online]. Available: https://www.kaggle.com/datasets/zgeakyldz/heart-desease-data
[21] Q. A'yuniyah *et al.*, "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 72, Sep. 2022, doi: 10.30865/json.v4i1.4781.
[22] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes Untuk Prediksi Harapan Hidup Pasien Gagal Jantung," *Jurnal Responsif*, vol. 3, no. 1, pp. 37–43, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti
[23] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id
[24] "Haris Metode Naïve Bayes Untuk Memprediksi Penyakit Stroke."
[25] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.
[26] Q. A'yuniyah *et al.*, "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 72, Sep. 2022, doi: 10.30865/json.v4i1.4781.
[27] M. Nurkholifah, A. NofiarAm, F. Kurnia Oktorina, S. Amik Riau, and P. Kampar, "Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes," 2023. [Online]. Available: https://www.kaggle.com/datasets/zgeakyldz/heart-desease-data
[28] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes Untuk Prediksi Harapan Hidup Pasien Gagal Jantung," *Jurnal Responsif*, vol. 3, no. 1, pp. 37–43, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti
[29] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id
[30] R. Mia *et al.*, "Exploring Machine Learning for Predicting Cerebral Stroke: A Study in Discovery," *Electronics (Switzerland)*, vol. 13, no. 4, Feb. 2024, doi: 10.3390/electronics13040686.
[31] D. S. Jodas, L. A. Passos, A. Adeel, and J. P. Papa, "PL-kNN: A Python-based implementation of a parameterless k-Nearest Neighbors classifier [Formula presented]," *Software Impacts*, vol. 15, Mar. 2023, doi: 10.1016/j.simpa.2022.100459.