Jurnal
Kolaborasi Riset Sarjana

Artikel

Penerapan Metode *Naïve Bayes* dengan SMOTE pada Sistem Pendukung Keputusan untuk Prediksi Risiko Stroke

Adam Fathurrohman Arya Bakhti*[1], Berliana Rahmadhani[2], Khoirun Nisa[3].
[1,2,3] Program Studi Informatika, Universitas Harapan Bangsa, Purwokerto, Indonesia
* Korespondensi: adamfathurrohman50@gmail.com

Abstrak: Stroke merupakan salah satu penyebab kematian dan kecacatan terbesar di dunia, sehingga prediksi dini menjadi kritis untuk mencegah komplikasi serius. Penelitian ini mengembangkan sistem pendukung keputusan untuk memprediksi risiko stroke menggunakan algoritma *Naïve Bayes* yang dikombinasikan dengan *Synthetic Minority Oversampling Technique* (SMOTE) guna mengatasi ketidakseimbangan data pada *Stroke Prediction Dataset* (5110 sampel, 4,87% kasus stroke). Metode penelitian mencakup *preprocessing data*, penghapusan fitur non-informatif, *encoding* variabel kategorikal, *oversampling* menggunakan SMOTE, serta evaluasi performa model menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa SMOTE meningkatkan sensitivitas model secara signifikan, dengan nilai *recall* 93% dan F1-score 81%, meskipun *precision* mengalami penurunan akibat bertambahnya prediksi positif palsu. Temuan ini menegaskan pentingnya pemilihan metrik evaluasi yang tepat pada data tidak seimbang. Studi ini memberikan kontribusi dalam pengembangan *pipeline* prediksi medis berbasis *Naïve Bayes* dan menawarkan dasar bagi pengembangan model yang lebih akurat melalui optimasi parameter dan algoritma alternatif.

Kata kunci: stroke; *Naïve Bayes*; SMOTE; sistem pendukung keputusan; klasifikasi.

***Received**: 30 Mei 2024*
***Revised**: 30 Juni 2024*
***Accepted**: 20 Juli 2024*
***Published**: 30 Juli 2024*

Pendahuluan

Stroke merupakan salah satu penyebab utama kematian dan disabilitas di dunia, sehingga deteksi dan prediksi dini menjadi sangat penting dalam upaya mencegah komplikasi serius serta mendukung penanganan medis yang lebih cepat dan efektif (Bathla & Kumar, 2022). Tantangan utama dalam membangun sistem prediksi risiko stroke adalah sifat dataset yang cenderung tidak seimbang, di mana hanya sekitar 5% data merupakan kasus stroke

(Advithi & Umadevi, 2024). Ketidakseimbangan ini mengakibatkan model klasifikasi sering kali lebih bias terhadap kelas mayoritas, sehingga kurang sensitif dalam mendeteksi kasus stroke yang sebenarnya sangat kritis.

Salah satu pendekatan yang dapat digunakan untuk mengatasi ketidakseimbangan kelas adalah *Synthetic Minority Oversampling Technique* (SMOTE), sebuah teknik *oversampling* yang terbukti mampu meningkatkan performa algoritma prediksi medis dengan memperbaiki representasi kelas minoritas (Das & Chowdhury, 2024). Dalam konteks klasifikasi, algoritma *Naïve Bayes* (NB) menjadi pilihan karena kemampuannya dalam memodelkan probabilitas dengan cepat dan efisien pada data dengan fitur numerik maupun kategorikal. Berbagai penelitian menunjukkan bahwa meskipun NB tidak selalu melampaui model yang lebih kompleks seperti *Random Forest*, performanya dapat meningkat signifikan setelah diterapkan teknik *balancing* seperti SMOTE, dengan akurasi mencapai sekitar 82% serta peningkatan pada nilai *recall* dan *F1-score* (Khansa & Gunawan, 2024). Meskipun demikian, efektivitas SMOTE dapat bervariasi bergantung pada karakteristik dataset dan dalam beberapa kasus justru menurunkan akurasi (Damari et al., 2025).

Penggunaan SMOTE terbukti dapat meningkatkan sensitivitas dan kemampuan deteksi model dalam mengklasifikasi kelas minoritas. Peningkatan ini terlihat dari metrik evaluasi seperti *F1-score* dan *recall* yang menjadi lebih seimbang, meskipun akurasi total model bisa sedikit menurun karena perubahan distribusi data. Sebuah studi mencatat bahwa setelah penerapan SMOTE, nilai F1-score model NB meningkat dari 70% menjadi 87%, meski akurasi menurun sedikit (Prameswara & Gunawan, 2024). Oleh karena itu, evaluasi model tidak cukup hanya menggunakan metrik akurasi, melainkan perlu mempertimbangkan metrik lain yang lebih mencerminkan kemampuan model dalam mendeteksi kelas minoritas, seperti *precision, recall*, dan *F1-score*.

Dalam sistem pendukung keputusan medis, *trade-off* antara peningkatan *recall* dan potensi penurunan *precision* tetap dapat diterima, mengingat prioritas utama adalah meminimalkan kasus stroke yang tidak terdeteksi (*false negative*), sebagaimana ditekankan pada studi Mutmainah (2021) terkait penanganan data *imbalance*. Hal ini menegaskan pentingnya pendekatan yang tidak hanya mempertimbangkan akurasi keseluruhan, tetapi juga menekankan metrik sensitivitas. Namun, di sisi lain, sintesis data sintetis dari SMOTE juga berpotensi memengaruhi stabilitas model secara keseluruhan sehingga perlu dilakukan evaluasi secara cermat terhadap kualitas data hasil *oversampling* (Damari et al., 2025; Rivaldo et al., 2024). Validasi yang ketat dengan berbagai metrik diperlukan agar model yang dihasilkan tidak hanya optimal di atas kertas, tetapi juga andal dalam praktik klinis.

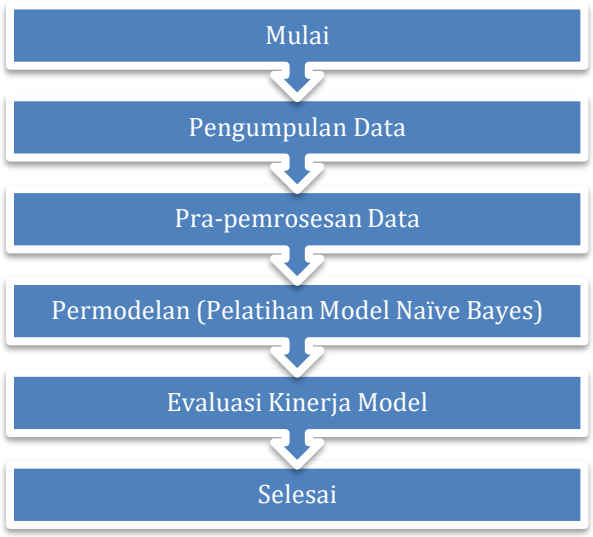
Berdasarkan latar belakang tersebut, penelitian ini berfokus pada pengembangan model prediksi risiko stroke menggunakan *Naïve Bayes* yang dipadukan dengan SMOTE untuk mengatasi ketidakseimbangan data. Penelitian ini diharapkan dapat memberikan pemahaman yang lebih baik mengenai efektivitas kombinasi kedua metode tersebut, serta kontribusi bagi peningkatan sistem pendukung keputusan dalam skrining awal risiko stroke.

Metode Penelitian

Metode penelitian dirancang secara sistematis untuk memperoleh model prediksi risiko stroke yang valid, reliabel, dan *robust*. Tahapan penelitian meliputi desain penelitian, karakteristik dan sumber data, proses pra-pemrosesan, penanganan ketidakseimbangan data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), pembangunan model *Naïve Bayes*, serta penyusunan skenario evaluasi model.

Desain Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental, yang memanfaatkan dataset publik untuk membangun dan menguji performa model klasifikasi. Seluruh proses analisis dilakukan secara terstruktur, dimulai dari pengolahan data mentah hingga pengujian performa model. Penelitian berfokus pada pengembangan *pipeline* klasifikasi menggunakan *Naïve Bayes* dengan penanganan ketidakseimbangan data melalui SMOTE, sebagaimana dijelaskan pada bagian sebelumnya mengenai relevansi teknik tersebut dalam meningkatkan performa deteksi kasus minoritas. Kerangka kerja ini menjadi panduan agar setiap langkah penelitian dapat dieksekusi secara logis dan berurutan, Gambar 1 mengilustrasikan alur kerja penelitian secara visual.



Gambar 1. Desain Penelitian

Sumber Data dan Deskripsi Variabel

Penelitian ini memanfaatkan dataset publik "*Stroke Prediction Dataset*" yang bersumber dari platform repositori data *Kaggle* (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data). Dataset ini merupakan fondasi utama untuk membangun model prediksi, terdiri dari 5.110 data observasi pasien dengan 12 atribut yang relevan. Penggunaan dataset publik ini bertujuan untuk menjamin transparansi dan reprodusibilitas hasil penelitian. Atribut yang ada mencakup atribut-atribut klinis seperti usia, riwayat hipertensi, riwayat penyakit jantung, kadar glukosa darah, indeks massa tubuh, serta variabel demografis lainnya. Data bersifat tidak seimbang, dengan proporsi kelas stroke hanya sekitar 5%, yang sesuai dengan tantangan distribusi kelas yang telah dijelaskan pada kajian awal. Hal ini menjadi dasar pentingnya penggunaan teknik *balancing* seperti SMOTE untuk menghindari bias klasifikasi terhadap kelas mayoritas.

Struktur detail dari setiap atribut disajikan pada Tabel 1, yang mengklasifikasikan setiap variabel berdasarkan tipe data dan deskripsi fungsionalnya. Atribut-atribut ini mencakup pengidentifikasi unik (*id*) yang akan diabaikan selama pemodelan, fitur-fitur prediktif, dan variabel target biner (*stroke*). Fitur-fitur ini merupakan

kombinasi dari data numerik (seperti usia dan *BMI*) dan data kategorikal (seperti jenis kelamin dan status merokok), yang memerlukan perlakuan pra-pemrosesan yang berbeda.

Tabel 1. Deskripsi Atribut

Atribut	Deskripsi
id	Identitas unik pasien (dieliminasi saat pemodelan).
gender	Jenis kelamin (*'Male', 'Female', 'Other'*).
age	Usia pasien (Numerik).
hypertension	Riwayat hipertensi (0: tidak, 1: ya).
heart_disease	Riwayat penyakit jantung (0: tidak, 1: ya).
ever_married	Status pernikahan (*'Yes', 'No'*).
work_type	Jenis pekerjaan (*'Private', 'Self-employed', 'Govt_job', 'children', 'Never_worked'*).
Residence_type	Tipe tempat tinggal (*'Urban', 'Rural'*).
avg_glucose_level	Rata-rata kadar glukosa dalam darah (Numerik).
bmi	Indeks Massa Tubuh (Numerik).
smoking_status	Status merokok (*'formerly smoked', 'never smoked', 'smokes', 'Unknown'*).
stroke	Kejadian stroke (1: ya, 0: tidak).

Tahapan Pra-pemrosesan Data

Pra-pemrosesan dilakukan sebelum proses pemodelan, mencakup pembersihan data, transformasi variabel kategorikal menjadi numerik melalui pengkodean (*encoding*), serta penyelarasan format data agar sesuai dengan kebutuhan algoritma *Naïve Bayes*. Tahap ini juga memastikan bahwa setiap fitur yang digunakan tetap informatif dan tidak menimbulkan gangguan pada proses pembelajaran model.

Penanganan Ketidakseimbangan Data dengan SMOTE

Untuk mengatasi tantangan ini, penelitian ini mengadopsi *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE adalah algoritma *oversampling* yang diakui secara luas dan terbukti efektif untuk mengatasi masalah ketidakseimbangan kelas (Wongvorachan et al., 2023). Berbeda dengan metode naif seperti *random over-sampling* yang hanya menduplikasi data minoritas dan berisiko tinggi menyebabkan *overfitting*, SMOTE bekerja dengan menciptakan data sintetis baru. Metode ini tidak menyalin data, melainkan menghasilkan sampel baru yang plausibel secara statistik di dalam ruang fitur (*feature space*), sehingga membantu memperluas region keputusan untuk kelas minoritas (Sakho et al., 2025).

Kelebihan utama SMOTE adalah kemampuannya menghasilkan data latih yang lebih seimbang tanpa kehilangan informasi (seperti pada *under-sampling*) dan dengan risiko *overfitting* yang lebih rendah dibandingkan *random over-sampling* (Carvalho et al., 2025). Efektivitas SMOTE dalam meningkatkan metrik evaluasi krusial seperti *Recall*, *F1-Score*, dan *AUC* pada data medis telah divalidasi secara ekstensif dalam berbagai studi terkini. Sebagai contoh, penelitian dalam domain prediksi penyakit kardiovaskular menunjukkan bahwa SMOTE secara konsisten mampu meningkatkan sensitivitas model dalam mendeteksi kasus-kasus langka namun kritis (Tompra et al., 2024).

Permodelan (Pelatihan Model *Naïve Bayes*)

Pemilihan metode klasifikasi merupakan inti dari pengembangan sistem pendukung keputusan ini. Penelitian ini menggunakan *Naïve Bayes Classifier*, sebuah keluarga algoritma klasifikasi probabilistik yang didasarkan pada Teorema *Bayes*. Metode ini dipilih bukan hanya karena efisiensi komputasinya yang tinggi, tetapi juga karena performanya yang telah terbukti solid di berbagai domain, termasuk dalam analisis data klinis yang kompleks. Sifatnya yang probabilistik memungkinkan model untuk tidak hanya memberikan label prediksi (misalnya, *'stroke'* atau *'tidak stroke'*), tetapi juga menyajikan probabilitas dari prediksi tersebut, yang merupakan informasi krusial untuk pengambilan keputusan klinis (Masood et al., 2024).

Naïve Bayes merupakan metode klasifikasi yang tidak bergantung pada aturan tertentu, melainkan memanfaatkan teori probabilitas dalam matematika untuk menentukan kemungkinan tertinggi dari suatu klasifikasi. Proses ini dilakukan dengan menganalisis frekuensi masing-masing kelas dalam data pelatihan. Sebagai teknik klasifikasi statistik, *Naïve Bayes* digunakan untuk memperkirakan probabilitas suatu data termasuk dalam kelas tertentu. Pendekatan ini didasarkan pada Teorema *Bayes* dan memiliki performa klasifikasi yang sebanding dengan metode seperti *decision tree* dan *neural network* (Azeraf et al., 2021).

Aturan *Bayes* (*Bayes Rule*) digunakan untuk memperkirakan probabilitas dari suatu kelas berdasarkan informasi sebelumnya. Algoritma *Naïve Bayes* menawarkan pendekatan sistematis untuk menggabungkan probabilitas awal (*prior*) dengan probabilitas bersyarat (*likelihood*), sehingga membentuk rumus yang dapat digunakan untuk menghitung kemungkinan setiap kelas secara matematis. Secara umum, bentuk dasar dari Teorema *Bayes* dinyatakan pada Persamaan 1.

$$P(H|X) = \frac{P(x|H)P(H)}{P(x)} \quad (1)$$

Rumus Teorema *Naïve bayes*:
X : Data dengan *class* yang belum diketahui
H : Hipotesis data X merupakan suatu *class* spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi x (posteriori *probability*)
P(H) : Probabilitas hipotesis H (*prior probability*)
P(X|H) : Probabilitas X berdasarkan kondisi tersebut P(X) = Probabilitas dari X

Pembagian Data dan Skema Evaluasi

Dataset dibagi menjadi 80% data pelatihan dan 20% data pengujian menggunakan parameter `random_state` agar proses dapat di replikasi dengan hasil yang konsisten. Pembagian ini menghasilkan 7777 sampel untuk pelatihan dan 1945 sampel untuk pengujian, dengan distribusi kelas yang seimbang pada tahap pelatihan. Evaluasi model dilakukan menggunakan empat metrik utama, yaitu akurasi, *precision, recall*, dan *F1-score*, yang memberikan gambaran menyeluruh mengenai performa model khususnya pada konteks data tidak seimbang. Mengingat sifat dataset yang sangat tidak seimbang, evaluasi tidak akan bergantung pada metrik akurasi yang dapat memberikan gambaran keliru. Sebaliknya, analisis akan berpusat pada *Confusion Matrix* sebagai landasan evaluasi. Matriks ini merangkum performa model dengan mengategorikan prediksi menjadi empat kuadran: *True Positive* (TP) untuk kasus stroke yang terdeteksi benar, *True Negative* (TN) untuk kasus non-stroke yang

terdeteksi benar, *False Positive* (FP) untuk alarm palsu, dan yang paling krusial, *False Negative* (FN) untuk kasus stroke yang gagal terdeteksi oleh model.

Dari *Confusion Matrix*, diturunkan dua metrik yang sangat penting untuk evaluasi klinis. Pertama adalah *recall* (juga dikenal sebagai *Sensitivity*), yang mengukur kemampuan model untuk mengidentifikasi semua kasus stroke yang sebenarnya. Metrik ini menjadi prioritas karena kegagalan mendeteksi penyakit (*False Negative*) merupakan risiko paling fatal dalam konteks medis. Selanjutnya *precision*, yang mengukur seberapa akurat prediksi positif yang dibuat oleh model. Formula untuk kedua metrik ini masing-masing dijelaskan pada Persamaan 2 dan 3.

$$Recall = \frac{TP}{TP+TN} \ ; \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Untuk menyeimbangkan pertukaran (*trade-off*) antara *precision* dan *recall*, *F1-score* digunakan. Sebagai rata-rata harmonik dari kedua metrik tersebut, *F1-score* memberikan skor tunggal yang efektif untuk evaluasi pada data tidak seimbang. Formula untuk *F1-score* dijelaskan pada Persamaan 4.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Lalu, akurasi dibutuhkan untuk mengukur seberapa banyak prediksi model yang benar dibandingkan dengan total prediksi. Formula untuk akurasi dijelaskan dalam Persamaan 5.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

Hasil dan Pembahasan

Pemahaman Data

Tahap pemahaman data dimulai dengan eksplorasi fitur-fitur seperti `gender`, `age`, `hypertension`, dan `avg_glucose_level` yang digunakan untuk prediksi risiko stroke. Fitur `id` dan `bmi` dihapus karena dianggap tidak relevan atau memiliki banyak nilai hilang. Analisis deskriptif menunjukkan variasi signifikan pada `age` dan `avg_glucose_level`, serta kemungkinan adanya *outlier* (pencilan). Distribusi target sangat tidak seimbang, sehingga digunakan teknik SMOTE untuk mengatasi ketimpangan ini. Pemahaman ini menjadi dasar penting dalam pemilihan fitur dan strategi *preprocessing* sebelum penerapan model *Naïve Bayes*. Proses eksperimen menghasilkan model klasifikasi berbasis *Naive Bayes* dengan dua skenario, yaitu pelatihan menggunakan data asli yang tidak seimbang dan pelatihan menggunakan data *oversampling* melalui SMOTE. Pembagian data sebesar 80% untuk pelatihan dan 20% untuk pengujian menghasilkan 7777 data latih dan 1945 data uji, dengan distribusi kelas stroke yang sangat timpang pada set uji. Tabel 2 menjabarkan karakteristik fitur pada *dataset* yang digunakan dalam penelitian ini.

Tabel 2. Karakteristik Fitur

Fitur	Karakteristik Data (Awal)
id	5110 nilai unik.
gender	*Female'*: 2994, *'Male'*: 2115, *'Other'*: 1.

age	Rentang: 0.08 hingga 82 tahun.
hypertension	0' (Tidak): 4612, '1' (Ya): 498.
heart_disease	0' (Tidak): 4834, '1' (Ya): 276.
ever_married	*Yes'*: 3353, *'No'*: 1757.
work_type	*Private'*: 2925, *'Self-employed'*: 819, *'children'*: 687, *'Govt_job'*: 657, *'Never_worked'*: 22.
Residence_type	*Urban'*: 2596, *'Rural'*: 2514.
avg_glucose_level	Rentang: 55.12 hingga 271.74.
bmi	Terdapat *missing values*: 4909 *non-null* dari 5110 total entri. Rentang: 10.3 hingga 97.6.

Identifikasi karakteristik dataset ini, termasuk prevalensi *missing values* pada kolom *bmi* dan ketidakseimbangan kelas yang signifikan pada variabel target stroke, menjadi landasan penting untuk keputusan dalam tahapan *pre-processing* data selanjutnya. Hal ini krusial untuk memastikan bahwa data dipersiapkan secara optimal agar model prediktif yang dibangun dapat memberikan hasil yang akurat dan tidak bias.

Pra-pemrosesan Data

1. Penanganan *Missing Values*

Langkah awal *pre-processing* difokuskan pada pembersihan dataset dari fitur-fitur yang tidak relevan atau memiliki isu kualitas data. Kolom *'id'* yang berfungsi sebagai pengidentifikasi unik pasien, tidak memberikan kontribusi informatif terhadap prediksi stroke, sehingga diputuskan untuk dihapus. Selain itu, kolom *'bmi'* (*Body Mass Index*) teridentifikasi memiliki sejumlah *missing values* yang dapat mempengaruhi integritas analisis, sebagaimana dapat dilihat pada Gambar 2. Meskipun terdapat berbagai metode imputasi, kolom *'bmi'* juga dikeluarkan dari dataset untuk menyederhanakan model dan menghindari potensi bias yang mungkin timbul dari imputasi data yang tidak sempurna.

Checking Missing/Null Values:

	0
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
smoking_status	0
stroke	0

Gambar 2. Cuplikan *Missing Value*

Setelah penghapusan kedua kolom yang dapa dilihat pada Gambar 2, verifikasi menyeluruh dilakukan untuk memastikan tidak ada lagi *missing values* pada fitur-fitur yang tersisa. Pemeriksaan ini mengonfirmasi bahwa dataset kini bebas dari nilai yang hilang, menjamin kelengkapan data untuk tahap-tahap selanjutnya.

2. *Encoding* Variabel Kategorikal

Dataset awal mengandung beberapa fitur yang bersifat kategorikal, seperti *'gender'*, *'ever_married'*, *'work_type'*, *'Residence_type'*, dan *'smoking_status'*. Algoritma *machine learning* umumnya memerlukan input dalam bentuk numerik. Oleh karena itu, *encoding* diterapkan untuk mengubah representasi tekstual ini menjadi format numerik yang dapat dipahami oleh model. Metode *label encoding* digunakan untuk mengonversi setiap kategori menjadi nilai integer unik, sesuai dengan pemetaan yang telah didefinisikan pada Gambar 3.

```
gender :  ['Male' 'Female' 'Other']

ever_married :  ['Yes' 'No']

work_type :  ['Private' 'Self-employed' 'Govt_job' 'children' 'Never_worked']

Residence_type :  ['Urban' 'Rural']

age :  [6.70e+01 6.10e+01 8.00e+01 4.90e+01 7.90e+01 8.10e+01 7.40e+01 6.90e+01
 5.90e+01 7.80e+01 5.40e+01 5.00e+01 6.40e+01 7.50e+01 6.00e+01 5.70e+01
 7.10e+01 5.20e+01 8.20e+01 6.50e+01 5.80e+01 4.20e+01 4.80e+01 7.20e+01
 6.30e+01 7.60e+01 3.90e+01 7.70e+01 7.30e+01 5.60e+01 4.50e+01 7.00e+01
 6.60e+01 5.10e+01 4.30e+01 6.80e+01 4.70e+01 5.30e+01 3.80e+01 5.50e+01
 1.32e+00 4.60e+01 3.20e+01 1.40e+01 3.00e+00 8.00e+00 3.70e+01 4.00e+01
 3.50e+01 2.00e+01 4.40e+01 2.50e+01 2.70e+01 2.30e+01 1.70e+01 1.30e+01
 4.00e+00 1.60e+01 2.20e+01 3.00e+01 2.90e+01 1.10e+01 2.10e+01 1.80e+01
 3.30e+01 2.40e+01 3.40e+01 3.60e+01 6.40e-01 4.10e+01 8.80e-01 5.00e+00
 2.60e+01 3.10e+01 7.00e+00 1.20e+01 6.20e+01 2.00e+00 9.00e+00 1.50e+01
 2.80e+01 1.00e+01 1.80e+00 3.20e-01 1.08e+00 1.90e+01 6.00e+00 1.16e+00
 1.00e+00 1.40e+00 1.72e+00 2.40e-01 1.64e+00 1.56e+00 7.20e-01 1.88e+00
 1.24e+00 8.00e-01 4.00e-01 8.00e-02 1.48e+00 5.60e-01 4.80e-01 1.60e-01]

hypertension :  [0 1]

heart_disease :  [1 0]

avg_glucose_level :  [228.69 202.21 105.92 ...  82.99 166.29  85.28]
```

Gambar 3. Cuplikan *Encoding* Variabel

3. Penanganan *Imbalanced Class* menggunakan SMOTE

Salah satu tantangan utama adalah *imbalanced class* pada variabel target *'stroke'*, dengan hanya 4.87% kasus stroke yang tercatat. Ketidakseimbangan ini dapat membiasakan model untuk memprediksi kelas mayoritas. Untuk mengatasinya, teknik *oversampling* SMOTE (*Synthetic Minority Over-sampling Technique*) diimplementasikan. SMOTE menghasilkan sampel sintetis dari kelas minoritas (pasien stroke), sehingga setelah penerapannya, distribusi kelas menjadi seimbang dengan 4861 sampel untuk setiap kelas. Keseimbangan ini penting untuk meningkatkan sensitivitas dan akurasi model dalam mendeteksi stroke. Perbandingan detail dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan *Class* Sebelum dan Sesudah SMOTE

Variabel Target '`stroke`'	Jumlah Sampel Sebelum SMOTE	Jumlah Sampel Setelah SMOTE
0 (Tidak Stroke)	4861	4861
1 (Stroke)	249	4861
Total Sampel	5110	9722

4. Pembagian Data untuk Pelatihan dan Pengujian

Langkah terakhir dalam tahap *pre-processing* adalah membagi dataset yang sudah bersih dan seimbang ke dalam set pelatihan (*training set*) dan set pengujian (*testing set*). Pembagian ini dilakukan untuk mengevaluasi kemampuan generalisasi model. Data pelatihan digunakan untuk melatih algoritma *machine learning* agar mempelajari pola dari fitur-fitur *input* dan hubungannya dengan variabel target. Sementara itu, data pengujian berfungsi sebagai data yang belum pernah dilihat oleh model selama pelatihan, digunakan untuk mengukur kinerja model secara objektif.

Dataset dibagi dengan proporsi 80% untuk set pelatihan dan 20% untuk set pengujian. Penggunaan `random_state`=105 memastikan bahwa pembagian data bersifat konsisten dan dapat di replikasi, menjamin bahwa setiap kali kode dijalankan, hasil pembagian data akan sama. Setelah pembagian, set pelatihan (`X_train`, `Y_train`) memiliki 7777 sampel, dan set pengujian (`X_test`, `Y_test`) memiliki 1945 sampel. Kedua set ini mempertahankan distribusi kelas yang seimbang, memastikan model dilatih dan dievaluasi berdasarkan representasi yang adil dari kedua kelas.

Evaluasi Model *Naïve Bayes*

Hasil pengujian performa model ditunjukkan melalui empat metrik utama: akurasi, *precision, recall*, dan *F1-score*. Pada data tidak seimbang, model *Naive Bayes* cenderung bias terhadap kelas mayoritas sehingga nilai *recall* untuk kelas stroke relatif rendah. Setelah penerapan SMOTE pada data latih, performa model mengalami peningkatan signifikan khususnya pada metrik *recall* dan *F1-score*. Peningkatan *recall* menunjukkan bahwa model menjadi lebih sensitif dalam mendeteksi pasien berisiko stroke, yang merupakan aspek kritis dalam konteks prediksi medis. Hasil evaluasi lengkap disajikan dalam Tabel 4 berikut.

Tabel 4. Hasil Evaluasi Model *Naïve Bayes*

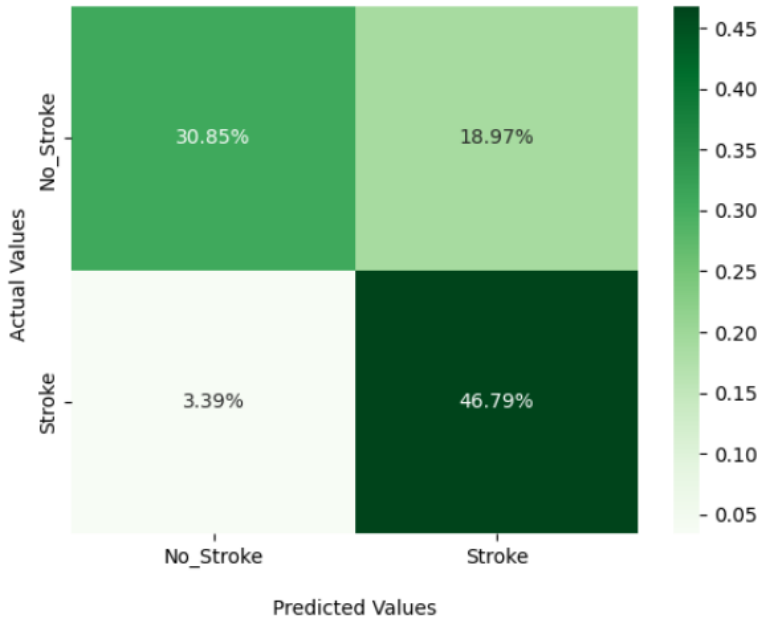
Metrik	Nilai
Accuracy	0.78
Precision	0.71
Recall	0.93
F1 Score	0.81

Secara keseluruhan, *pipeline SMOTE–Naive Bayes* memberikan hasil yang lebih stabil dalam mendeteksi kelas minoritas. Tabel performa yang telah disajikan memperlihatkan bahwa meskipun *precision* menurun akibat meningkatnya prediksi positif palsu, hasil ini masih dapat diterima untuk aplikasi klinis di mana prioritas utamanya adalah meminimalkan kesalahan *false negative*. Dengan demikian, hasil penelitian ini menunjukkan efektivitas SMOTE dalam meningkatkan performa *Naive Bayes* pada data stroke yang tidak seimbang.

Berdasarkan hasil tersebut, model menunjukkan kemampuan yang baik dalam mendeteksi kasus stroke, tercermin dari nilai *recall* yang tinggi yaitu 93%. Hal ini sejalan dengan temuan dalam penelitian oleh Damari et al. (2025) yang menyatakan bahwa penerapan SMOTE memang dapat meningkatkan sensitivitas model terhadap kelas minoritas. Namun, penelitian tersebut juga menegaskan bahwa penambahan data sintetis melalui SMOTE dapat mengubah distribusi data, yang dalam beberapa kasus justru berdampak pada penurunan akurasi model.

Dalam penelitian ini, meskipun akurasi model setelah penerapan SMOTE sebesar 78% tergolong cukup baik, terdapat penurunan dibandingkan akurasi tanpa *balancing* data. Temuan ini konsisten dengan hasil penelitian Damari et al. (2025), yang menunjukkan bahwa penerapan SMOTE perlu dilakukan dengan hati-hati karena potensi perubahan distribusi data yang dapat memengaruhi efektivitas model, terutama ketika dikombinasikan dengan teknik optimasi seperti PSO. Selain itu, nilai *precision* sebesar 71% menunjukkan adanya sejumlah *false positive*, yaitu kasus non-stroke yang terdeteksi sebagai stroke. Hal ini merupakan konsekuensi umum dalam penerapan teknik *oversampling*, di mana model cenderung meningkatkan deteksi kasus minoritas namun berpotensi meningkatkan prediksi positif palsu.

Meskipun demikian, dalam konteks sistem pendukung keputusan medis, *trade-off* ini masih dapat diterima, mengingat prioritas utama adalah meminimalkan kasus stroke yang tidak terdeteksi (*false negative*), sebagaimana juga ditekankan dalam penelitian Mutmainah, 2021 terkait penanganan data *imbalance* pada klasifikasi penyakit stroke. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa penerapan *Naïve Bayes* dengan metode SMOTE dapat meningkatkan kemampuan deteksi risiko stroke, terutama pada kelas minoritas. Namun, diperlukan perhatian khusus terhadap dampak distribusi data sintetis terhadap performa keseluruhan model, sebagaimana juga diuraikan dalam studi Damari et al. (2025). Penggunaan teknik optimasi tambahan seperti PSO berpotensi menjadi alternatif solusi untuk meningkatkan akurasi dan stabilitas model prediksi. Gambar 4 menunjukkan grafik *Confusion Matrix.*



Gambar 4. Visualisasi *Confusion Matrix*

Visualisasi *Confusion Matrix* pada Gambar 4 menunjukkan bahwa dari seluruh data pasien dengan status sebenarnya stroke, sebanyak 46,79% berhasil diklasifikasikan dengan benar oleh model sebagai stroke, sedangkan 3,39% diklasifikasikan salah sebagai tidak stroke (*false negative*). Di sisi lain, dari pasien yang sebenarnya tidak mengalami stroke, sebesar 30,85% berhasil dikenali dengan tepat sebagai tidak stroke, sementara 18,97% diklasifikasikan salah sebagai stroke (*false positive*).

Temuan ini sangat penting dalam konteks sistem pendukung keputusan medis, di mana keberhasilan model dalam mendeteksi pasien stroke (*true positive*) sebesar 46,79% menunjukkan performa yang baik. Terlebih lagi, proporsi kesalahan model dalam mengklasifikasikan pasien stroke sebagai tidak stroke (*false negative*) tergolong rendah, yaitu hanya 3,39%, sehingga potensi keterlambatan diagnosis dan penanganan stroke dapat diminimalkan.

Kesimpulan

Penelitian ini menunjukkan bahwa integrasi metode *Naïve Bayes* dengan teknik *oversampling* SMOTE mampu meningkatkan performa model dalam mendeteksi risiko stroke pada dataset yang tidak seimbang, terutama melalui peningkatan nilai *recall* dan *F1-score* yang menggambarkan kemampuan model dalam mengenali kasus stroke secara lebih sensitif dan akurat. Meskipun terjadi penurunan nilai *precision* akibat meningkatnya prediksi positif palsu, *trade-off* ini masih dapat diterima pada konteks klinis, karena meminimalkan *false negative* dinilai jauh lebih penting dibandingkan risiko *false positive* dalam deteksi penyakit yang berpotensi fatal seperti stroke. Secara keseluruhan, penelitian ini berkontribusi pada pengembangan *pipeline* prediksi medis berbasis data tidak seimbang dan menegaskan bahwa *Naïve Bayes* tetap merupakan algoritma yang efisien dan kompetitif apabila didukung oleh teknik *balancing* yang tepat. Ke depan, peningkatan kualitas model dapat dilakukan melalui optimasi parameter SMOTE, penggunaan algoritma *ensemble learning*, serta penerapan metode *explainable AI* (XAI) seperti SHAP atau LIME untuk memberikan interpretasi klinis yang lebih jelas, sehingga sistem pendukung keputusan yang dikembangkan tidak hanya akurat tetapi juga transparan, terpercaya, dan mudah diadopsi oleh praktisi kesehatan.

Referensi

Azeraf, E., Monfrini, E., & Pieczynski, W. (2021). Using the Naïve Bayes as a discriminative model. *2021 13th International Conference on Machine Learning and Computing*, 106–110. https://doi.org/10.1145/3457682.3457697

Bathla, P., & Kumar, R. (2022). Artificial Intelligence based Model for Brain Stroke Prediction. *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 1–6. https://doi.org/10.1109/IATMSI56455.2022.10119373

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1), 71. https://doi.org/10.1186/s40537-025-01119-4

Advithi, D., & Umadevi, V. (2024). Integrating SMOTE and LIME Techniques for Enhanced Stroke Prediction using Machine Learning Approaches. *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 1–6. https://doi.org/10.1109/ICETCS61022.2024.10544182

Damari, A., Taghfirul Azhima Yoga Siswa, & Wawan Joko Pranoto. (2025). Implementation of the PSO-SMOTE Method on the Naïve Bayes Algorithm to Address Class Imbalance in Landslide Disaster Data. *INOVTEK Polbeng - Seri Informatika*, *10*(1), 332–343. https://doi.org/10.35314/7wcvrb72

Das, D. K., & Chowdhury, S. (2024). Brain Stroke Prediction by Machine Learning Algorithm Along with Boosting Classifier. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 1576–1581. https://doi.org/10.1109/ICSES63445.2024.10763085

Khansa, G. A. F., & Gunawan, P. H. (2024). Predicting Stunting in Toddlers Using KNN and Naïve Bayes Methods. *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 17–21. https://doi.org/10.1109/ICoDSA62899.2024.10651676

Masood, A., Naseem, U., Rashid, J., Kim, J., & Razzak, I. (2024). Review on enhancing clinical decision support system using machine learning. *CAAI Transactions on Intelligence Technology*. https://doi.org/10.1049/cit2.12286

Prameswara, A., & Gunawan, P. H. (2024). Predicting Toddler Stunting at Bandarharjo Health Center: Applications of K-Nearest Neighbors and Naïve Bayes Algorithms. *2024 8th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, 30–34. https://doi.org/10.1109/ELTICOM64085.2024.10864390

Rivaldo, V. J., Siswa, T. A. Y., & Pranoto, W. J. (2024). Perbaikan Akurasi Naïve Bayes dengan Chi-Square dan SMOTE Dalam Mengatasi High Dimensional dan Imbalanced Data Banjir. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(3), 1656. https://doi.org/10.30865/mib.v8i3.7886

Sakho, A., Malherbe, E., & Scornet, E. (2025). *Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants*. http://arxiv.org/abs/2402.03819

Tompra, K.-V., Papageorgiou, G., & Tjortjis, C. (2024). Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification. *Algorithms*, *17*(5), 178. https://doi.org/10.3390/a17050178

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054