Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893

Analisis Perbandingan Kinerja KNN Dan *Logistic Regression* Pada Klasifikasi Risiko Stroke

DOI: http://dx.doi.org/10.35889/jutisi.v15i4.3812

Asri Rahma Ayunanda[1*], Putry Wahyu Setyaningsih[2]
Sistem Informasi, Universitas Mercu Buana Yogyakarta, Yogyakarta, Indonesia
**e-mail Corresponding Author*: asrirahmaayunanda@gmail.com

Abstract

Stroke is a primary cause of mortality globally, making accurate risk classification vital for early intervention. This study aims to evaluate the effectiveness of K-Nearest Neighbor (KNN) compared to Logistic Regression in detecting stroke risk. Utilizing 4,981 medical records from Kaggle, the SMOTE technique was implemented to address data imbalance across 80:20 and 70:30 split scenarios. Results indicate that Logistic Regression outperforms KNN, providing more consistent performance in detecting high-risk classes. The most optimal performance was achieved by Logistic Regression at an 80:20 ratio, reaching 80.00% recall and 74.12% accuracy. This study demonstrates that Logistic Regression is a more effective and sensitive method for identifying clinical stroke risk factors in medical decision support systems.

Keywords: *Classification; KNN; Logistic Regression; SMOTE; Stroke*

Abstrak

Stroke merupakan penyebab kematian utama secara global, sehingga klasifikasi risiko yang akurat sangat penting untuk intervensi dini. Penelitian ini bertujuan mengevaluasi efektivitas algoritma *K-Nearest Neighbor* (KNN) dibandingkan Logistic Regression dalam mendeteksi risiko stroke. Dengan menggunakan dataset 4.981 rekam medis dari Kaggle, teknik SMOTE diterapkan untuk menangani ketidakseimbangan data pada skenario pembagian 80:20 dan 70:30. Hasil menunjukkan Logistic Regression mengungguli KNN dengan performa lebih konsisten dalam mendeteksi kelas risiko. Kinerja paling optimal dicapai Logistic Regression pada rasio 80:20 dengan nilai recall 80,00% dan akurasi 74,12%. Penelitian ini membuktikan Logistic Regression adalah metode yang lebih efektif dan sensitif dalam mengidentifikasi faktor risiko klinis stroke untuk mendukung sistem keputusan medis.

Kata kunci: *Klasifikasi; KNN; Logistic Regression; SMOTE; Stroke*

1. Pendahuluan

Stroke didefinisikan sebagai kegagalan fungsi sistem saraf akibat interupsi pasokan darah ke otak dengan durasi serangan yang umumnya menetap minimal 24 jam [1]. Fenomena ini merupakan penyebab kematian ketiga terbesar di dunia serta pemicu utama kecacatan fisik jangka panjang. Mengingat risikonya yang berkorelasi dengan faktor usia, medis, dan gaya hidup, identifikasi potensi bahaya secara dini sangat krusial untuk menurunkan kemungkinan komplikasi serta meningkatkan kualitas hidup pasien melalui langkah pencegahan yang lebih awal.

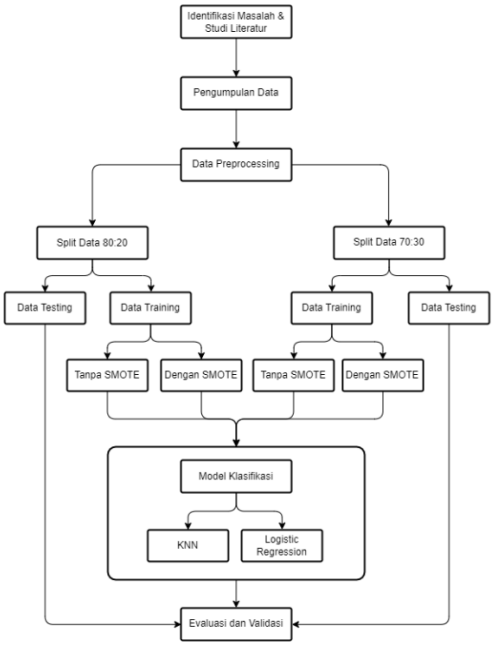
Hingga saat ini, langkah penanggulangan stroke, baik dari aspek edukasi maupun pengelolaan klinis, dinilai belum efektif [2]. Kompleksitas penanganan diperparah oleh tingginya frekuensi keterlambatan diagnosa, terutama di wilayah yang mengalami defisit infrastruktur pemindaian tingkat lanjut seperti CT-scan dan MRI [3]. Selain itu, identifikasi risiko konvensional seringkali memakan waktu lama dan rumit, terutama ketika harus mengolah informasi medis dalam skala masif. Hambatan teknis lainnya muncul dari fenomena ketidakseimbangan kelas (data imbalance) pada basis data klinis yang dapat mengganggu efektivitas pelatihan model prediksi.

Literatur sebelumnya telah mengeksplorasi penggunaan pembelajaran mesin dalam klasifikasi berbagai penyakit kronis. Nasien dkk. melakukan riset klasifikasi penyakit diabetes dengan menguji algoritma KNN, Naive Bayes, dan Regresi Logistik menggunakan ekstraksi fitur Principal Component Analysis (PCA) [4]. Dalam ranah penyakit kardiovaskular, Sitanggang dkk. mengimplementasikan metode KNN dan Regresi Logistik untuk memprediksi penyakit jantung dengan memanfaatkan data dari Kaggle [5]. Secara spesifik pada kasus stroke, Azhar dkk. mengeksekusi studi komparatif terhadap algoritma Regresi Logistik, Random Forest, SVM, dan KNN untuk mengidentifikasi potensi serangan melalui data mining [6]. Meskipun penelitian-penelitian tersebut telah berhasil mencapai tingkat akurasi tinggi, terdapat gap atau celah yang masih tersisa. Riset sebelumnya cenderung berfokus pada perbandingan banyak metode sekaligus untuk kebutuhan diagnosis umum dan sering kali menghadapi kendala akurasi saat dihadapkan pada data yang tidak seimbang (unbalanced data). Selain itu, masih terbatas kajian yang secara mendalam mengoptimasi perbandingan spesifik antara KNN dan Regresi Logistik dengan mengintegrasikan teknik penyeimbangan data khusus untuk analisis risiko awal.

Determinasi algoritma dalam membangun model proyeksi sangat krusial karena berdampak langsung pada akurasi keluaran [7]. Sebagai solusi, penelitian ini mengomparasikan metode K-Nearest Neighbor (KNN) yang bekerja berdasarkan kedekatan spasial dan kemiripan objek [8][9], dengan Regresi Logistik yang efisien dalam memproses variabel biner untuk klasifikasi dua kategori [10][11]. Kebaruan dalam penelitian ini terletak pada penggabungan kedua algoritma tersebut dengan teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk mengatasi ketidakseimbangan data melalui produksi sampel artifisial [12][13]. Pendekatan ini dirancang untuk menghasilkan evaluasi kinerja yang objektif guna menentukan algoritma yang paling optimal dalam memetakan kategori kerawanan stroke secara berbasis data.

2. Metodologi

Dalam penelitian ini terdapat sejumlah langkah yang meliputi hal-hal berikut:



Gambar 1. Metodologi

2.1. Pengumpulan Dataset

Langkah pengumpulan data dijalankan di fase awal sebagai dasar permulaan dalam proses analisis dan perolehan hasil penelitian [14]. Dataset yang berkaitan dengan stroke yang digunakan diambil secara terbuka melalui platform Kaggle, terdiri dari 11 atribut yang mencakup 10 variabel independen dan satu variabel dependen. Komponen variabel bebas di dalam penelitian ini mencakup determinan gender, rentang usia, riwayat tekanan darah tinggi, rekam medis patologi jantung, status konjugal, kategorisasi lapangan pekerjaan, klasifikasi wilayah

domisili, konsentrasi glukosa darah rata-rata, pengukuran indeks massa tubuh, serta klasifikasi perilaku konsumsi rokok. Sementara itu, variabel dependen adalah stroke, yang menunjukkan klasifikasi tingkat risiko individu. Data tersebut akan diproses dan dimanfaatkan sebagai sumber informasi utama yang mendukung pelaksanaan penelitian ini [5].

Dataset pertama yang diterapkan dalam penelitian ini terdiri dari 4.981 baris catatan medis. Usai menjalani proses preprocessing yang mencakup pembersihan data, jumlah data tetap sebanyak 4.981 baris, namun kini berada dalam format yang ideal untuk analisis matematis. Selanjutnya, pada tahap pembagian data (split data), data dipisahkan ke dalam dua skenario utama. Pada skenario 80:20, data terbagi menjadi 3.984 data untuk pelatihan dan 997 data untuk pengujian. Di sisi lain, dalam skenario 70:30, distribusi data mengalami perubahan menjadi 3.486 data pelatihan dan 1.495 data pengujian untuk mengevaluasi konsistensi kinerja model dalam rasio yang berbeda. Kumpulan data ini termasuk variabel klinis yang akan dianalisis menggunakan metode KNN dan Logistic Regression. Dataset dapat dilihat dalam Tabel 1.

Tabel 1. Dataset

gender	age	hypertension	...	bmi	smoking_status	stroke
Male	67.0	0	...	36.6	Formerly smoked	1
Male	80.0	0	...	32.5	Never smoked	1
Female	49.0	0	...	34.4	smokes	1
Female	79.0	1	...	24.0	Never smoked	1
Male	81.0	0	...	29.0	Formerly smoked	1

Tabel 2. Dataset Deskripsi

Nama Fitur	Deskripsi
gender	Jenis kelamin pasien.
age	Usia pasien (dalam tahun).
hypertension	Apakah pasien memiliki darah tinggi (0 = tidak; 1 = iya).
heart_disease	Apakah pasien memiliki riwayat penyakit jantung (0 = tidak; 1 = iya).
ever_married	Apakah pasien pernah atau sedang menikah.
work_type	Jenis pekerjaan atau status ketenagakerjaan.
residence_type	Tipe area tempat tinggal pasien.
avg_glucose_level	Kadar glukosa (gula) rata-rata dalam darah.
bmi	Indeks Massa Tubuh pasien.
smoking_status	Status merokok pasien saat ini atau di masa lalu.
stroke	Apakah pasien terkena stroke atau tidak (0 = tidak; 1 = iya).

2.2. Data Preprocessing

Proses mengubah data mentah menjadi format yang mudah dipahami [5]. Pada fase pra-pemrosesan, pembersihan data (data cleaning) dieksekusi secara spesifik untuk mengeliminasi anomali berupa nilai yang hilang (missing value) [15]. Di samping itu, intervensi terhadap disparitas proporsi kelas dijalankan melalui implementasi teknik SMOTE sebelum tahapan konstruksi model dimulai. Fase ini memegang peranan krusial mengingat data mentah pada umumnya belum berada dalam format yang terstandardisasi. Ketiadaan standardisasi tersebut menyebabkan proses penambangan data tidak dapat berjalan secara efektif pada basis data yang belum diolah, sehingga tahapan ini wajib diintegrasikan guna mempersiapkan variabel data sekaligus menyederhanakan alur komputasi berikutnya[6]. Oleh karena itu, pelaksanaan fase pra-pemrosesan yang dilakukan secara presisi dan terstruktur menjadi prasyarat mutlak dalam menjamin validitas serta kredibilitas hasil pengujian yang akan diperoleh pada tahapan analisis lanjutan [16].

Tahapan preprocessing diawali dengan memastikan tidak adanya data yang kosong (missing value) pada seluruh atribut. Selanjutnya, Label Encoding akan diterapkan untuk mengkonversi variabel kategori ('gender', 'ever_married', 'work_type', 'Residence_type', dan 'smoking_status') menjadi format angka. Setelah itu, Standarisasi dilakukan dengan menggunakan StandardScaler untuk menyesuaikan rentang nilai di antara fitur-fitur dan

mengurangi dampak dari pencilan, sehingga dataset siap untuk digunakan dalam proses klasifikasi dengan presisi yang tinggi.

Tabel 3. Label Encoding pada gender

Gender	Gender (encoding)
Male	1
Male	1
Female	0
Female	0
Male	1

Tabel 3, merupakan representasi contoh proses encoding. Prosedur yang sama juga diterapkan pada variabel kategorikal lainnya seperti ever_married, work_type, Residence_type, dan smoking_status.

Tabel 4. StandardScaler pada Gender setelah encoding

Gender (encoding)	Gender (scaler)
1	1.183909
1	1.183909
0	-0.844660
0	-0.844660
1	1.183909

Sebagai kelanjutan dari proses preprocessing, Tabel 4 menunjukkan contoh hasil akhir standarisasi pada variabel 'gender'. Dengan penerapan StandardScaler, rentang nilai yang semula berbeda-beda diseragamkan sehingga model machine learning dapat melakukan klasifikasi secara lebih optimal dan objektif.

2.3. Rasio Pembagian Data

Skenario Split Data 80:20 dilakukan dengan mengalokasikan 80% data untuk proses pelatihan (training) agar model mengenali pola data, dan 20% sisanya untuk pengujian (testing) guna mengukur tingkat akurasi prediksi. Sedangkan skenario Split Data 70:30 menggunakan proporsi 70% informasi digunakan untuk proses pelatihan dan 30% dialokasikan untuk pengujian guna mengevaluasi stabilitas serta konsistensi performa algoritma pada jumlah data yang diuji yang lebih banyak.

2.4. Kondisi Penyeimbangan Data

Kondisi Tanpa SMOTE menerapkan pengujian pada data asli yang masih memiliki ketidakseimbangan kelas (imbalance). Tahap ini berfungsi sebagai pembanding untuk melihat kecenderungan model dalam memprediksi kelas mayoritas saja. Sebaliknya, kondisi berbasis SMOTE yang mengintegrasikan teknik *Synthetic Minority Over-sampling Technique* diorientasikan untuk menyelaraskan kuantitas data antara kelompok minoritas dan mayoritas. Langkah ekualisasi ini bertujuan mengamplifikasi sensitivitas arsitektur model dalam mendeteksi dan mengidentifikasi populasi pasien yang memiliki risiko keterjangkitan penyakit stroke.

2.5. Modelling

Proses pemodelan pada tahapan ini menerapkan algoritma K-Nearest Neighbor (KNN) dan Logistic Regression untuk memprediksi risiko penyakit stroke. Penerapan kedua metode tersebut bertujuan untuk menghasilkan model klasifikasi yang valid.

2.5.1. K-Nearest Neighbor (KNN)

Algoritma KNN melakukan klasifikasi berdasarkan kedekatan jarak antar data. Perhitungan jarak dalam penelitian ini menggunakan Euclidean Distance dengan formula sebagai berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \textbf{(1)}$$

Keterangan:
- d(x,y): Jarak (Euclidean) antara data uji x dan data latih y
- n: Jumlah fitur atau dimensi data.
- x_i: Nilai fitur ke-i pada data uji.
- y_i: Nilai fitur ke-i pada data latih.

2.5.2. Logistic Regression

Regresi Logistik digunakan untuk memprediksi probabilitas variabel dependen biner. Model ini menggunakan fungsi sigmoid yang dirumuskan sebagai:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad \textbf{(2)}$$

Keterangan:
- P: Probabilitas variabel dependen (misalnya: probabilitas risiko stroke).
- e: Bilangan Euler atau konstanta eksponensial (≈2,718).
- β_0: Konstanta atau intercept (titik potong garis pada sumbu Y).
- β_1: Koefisien regresi (menunjukkan besarnya pengaruh variabel x).
- x: Variabel independen (fitur atau data input pasien).

2.6. Evaluasi dan Validasi

Langkah ini bertujuan menguji kapasitas model dalam mengelompokkan data baru menggunakan confusion matrix. Instrumen ini berfungsi mengukur efektivitas model dengan menyajikan data komparatif antara hasil prediksi sistem dan kondisi riil di lapangan [17]. Evaluasi dilakukan melalui metrik akurasi, presisi, recall, dan F1-score yang didasarkan pada identifikasi nilai True Positive (TP), True Negative (TN), False Positive (FP), serta False Negative (FN). Adapun formula evaluasi yang digunakan adalah sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(3)}$$

$$\text{Presisi} = \frac{TP}{TP + FP} \quad \textbf{(4)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \textbf{(5)}$$

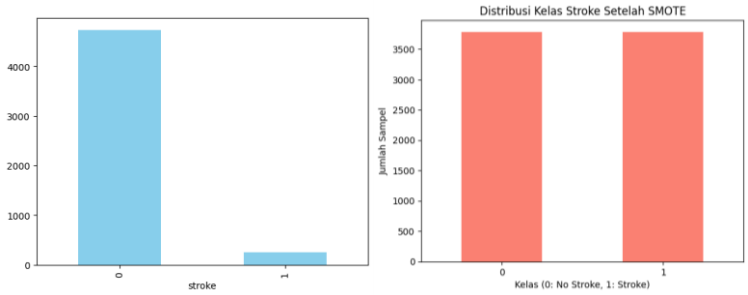
$$\text{F1-Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad \textbf{(6)}$$

Keterangan:
- TP (True Positive): Data positif yang diprediksi benar oleh model.
- TN (True Negative): Data negatif yang diprediksi benar oleh model.
- FP (False Positive): Data negatif yang salah diprediksi sebagai positif.
- FN (False Negative): Data positif yang salah diprediksi sebagai negatif.

3. Hasil dan Pembahasan

3.1. Distribusi Data Target Sebelum dan Sesudah SMOTE

Data awal menunjukkan ketidakseimbangan yang signifikan antara pasien yang mengalami stroke dan yang tidak, yang menyebabkan model tidak mampu mendeteksi risiko stroke dengan tepat. Dengan menggunakan metode SMOTE, jumlah sampel dari kelas minoritas diimbangi secara buatan agar model dapat lebih efektif dalam memahami pola dari pasien stroke. Tujuan dari penyeimbangan ini adalah untuk meningkatkan nilai pengingat atau sensitivitas model dalam mendeteksi pasien yang sebenarnya memiliki risiko stroke. Perbandingan data sebelum dan setelah penerapan SMOTE dapat dilihat pada gambar 2.



(a) (b)

Gambar 2. (a) Sebelum SMOTE; (b) Setelah SMOTE

3.2. Optimasi Parameter Nilai K (K-Nearest Neighbor)

Penetuan nilai tetangga terdekat atau K (neighbor) yang paling optimal pada algoritma KNN wajib dilakukan sebelum melangkah ke tahap pengujian komparasi antaralgoritma. Proses pengujian ini berlangsung dengan menguji serangkaian representasi angka ganjil pada rentang nilai 1 sampai 10. Pemilihan bilangan ganjil tersebut bertujuan untuk mengeliminasi potensi munculnya hasil voting yang berimbang atau seri saat penentuan kelas data. Adapun data perbandingan mengenai performa dari masing-masing nilai K yang diuji telah dirangkum secara sistematis pada Tabel 5.

Tabel 5. Perbandingan nilai K

Nilai K	Akurasi	Recall
K =1	90.57%	14.00%
K= 3	93.88%	2.00%
K = 5	94.48%	0.00%
K = 7	94.78%	2.00%
K = 9	94.98%	0.00%

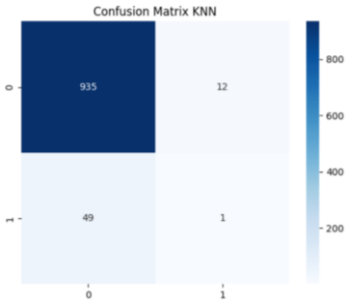
Berdasarkan Tabel 5, Nilai K=3 ditetapkan sebagai parameter terbaik karena menghasilkan performa klasifikasi yang lebih konsisten dibandingkan K=7, meskipun keduanya memiliki nilai recall yang sama sebesar 2,00%. Alasan utama tidak memilih K=7 adalah karena munculnya kembali angka recall setelah sempat menyentuh 0,00% pada K=5 mengindikasikan adanya hasil yang kurang stabil, sementara K=3 berada pada rentang awal yang lebih mampu mempertahankan kemampuan deteksi secara alami. Melalui pemilihan K=3, model berhasil mencapai akurasi yang tinggi tanpa kehilangan sensitivitas terhadap data positif sebelum performa tersebut benar-benar hilang pada nilai K yang lebih besar akibat dominasi kelompok data mayoritas.

3.3. Evaluasi dan Validasi

Bagian ini menyajikan hasil pengujian algoritma KNN dan Logistic Regression dengan skenario rasio data 80:20 (3.984 data pelatihan dan 997 data pengujian) dan 70:30 (3.486 data pelatihan dan 1.495 data pengujian), baik tanpa maupun dengan penerapan SMOTE. Analisis mendalam dilakukan menggunakan confusion matrix untuk membedah akurasi klasifikasi pada setiap model. Sebagai bagian dari tahap evaluasi ini, akan dipaparkan pula penjelasan mengenai perbandingan hasil akurasi serta recall dari seluruh skenario pengujian guna menentukan model dengan performa yang paling optimal.

3.3.1. Pengujian 80:20 KNN tanpa SMOTE

Hasil evaluasi menggunakan algoritma KNN tanpa SMOTE dengan rasio 80:20 dapat diperhatikan pada Gambar 3 yang menampilkan confusion matrix.

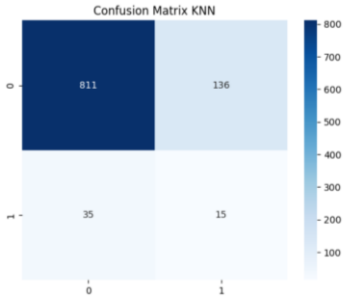


Gambar 3. Confusion Matrix KNN 80:20 tanpa SMOTE

Visualisasi data pada Gambar 3 memaparkan rincian capaian dari pengujian confusion matrix. Berdasarkan grafik tersebut, akurasi prediksi model menunjukkan angka True Negative (TN) sebesar 935 kasus dan True Positive (TP) sebanyak 1 kasus. Sementara itu, tingkat kekeliruan model mencakup False Positive (FP) yang teridentifikasi sebanyak 12 kasus serta False Negative (FN) yang menyentuh angka 49 kasus. Dominasi nilai FN yang lebih tinggi dibanding TP mengindikasikan bahwa model tidak berhasil menemukan mayoritas pasien terkena stroke, hanya bisa mengenali 1 dari 50 kasus stroke yang sebenarnya. Hal ini membuktikan bahwa tanpa penyeimbangan data, model cenderung hanya mengenali pola kelas mayoritas sehingga tidak efektif dalam mendiagnosa risiko stroke meskipun nilai akurasi keseluruhannya terlihat tinggi.

3.3.2. Pengujian 80:20 KNN dengan SMOTE

Hasil pengujian menggunakan algoritma KNN dengan SMOTE dengan proporsi 80:20 dapat diperhatikan pada Gambar 4 yang menunjukkan confusion matrix.

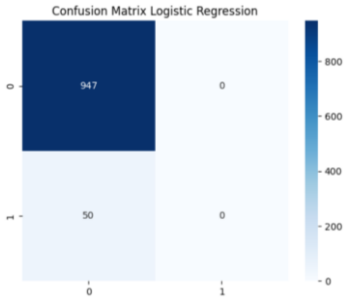


Gambar 4. Confusion Matrix KNN 80:20 dengan SMOTE

Representasi visual pada Gambar 4 memaparkan metrik confusion matrix yang menghasilkan perolehan nilai True Negative (TN) sebanyak 811 kasus, False Positive (FP) sebesar 136 kasus, False Negative (FN) sejumlah 35 kasus, serta kenaikan True Positive (TP) yang mencapai 15 kasus. Jika disandingkan dengan capaian data sebelum penerapan SMOTE, kuantitas pengidap stroke yang mampu diidentifikasi secara tepat (TP) mengalami lonjakan signifikan dari yang semula hanya 1 kasus menjadi 15 kasus, yang dibarengi dengan penurunan angka kegagalan prediksi (FN) dari 49 kasus menjadi 35 kasus. Meskipun intervensi ini berdampak pada penurunan tingkat akurasi secara umum, implementasi metode SMOTE teruji efektif dalam memosisikan model untuk mengenali data kelas minoritas dengan lebih sensitif, sehingga klasifikasi yang dihasilkan menjadi jauh lebih relevan untuk mendeteksi potensi risiko stroke.

3.3.3. Pengujian 80:20 Logistic Regression tanpa SMOTE

Berikut merupakan hasil pengujian yang dilakukan dengan algoritma Logistic Regression tanpa SMOTE dengan proporsi 80:20. Detail performa klasifikasi ditampilkan pada Gambar 5.

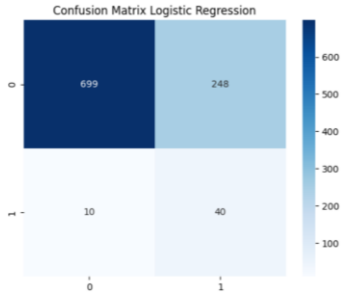


Gambar 5. Confusion Matrix Logistic Regression 80:20 tanpa SMOTE

Gambar 5 menunjukkan hasil dari pengujian algoritma Regresi Logistik dengan proporsi 80:20 tanpa SMOTE. Berdasarkan perhitungan pada confusion matrix, diperoleh hasil empiris yang menunjukkan angka True Negative (TN) mencapai 947 kasus dan False Positive (FP) sebesar 0 kasus. Sebaliknya, performa model dalam mengenali data positif menghasilkan nilai False Negative (FN) sebanyak 50 kasus tanpa ada satu pun data yang teridentifikasi sebagai True Positive (TP) atau bernilai 0 kasus. Data ini menunjukkan bahwa model mengalami kegagalan total dalam mendeteksi kelas positif karena mengklasifikasikan seluruh sampel sebagai data tidak stroke. Meskipun akurasinya tinggi, model ini tidak memiliki kemampuan deteksi risiko stroke sama sekali pada kondisi data asli yang tidak seimbang.

3.3.4. Pengujian 80:20 Logistic Regression dengan SMOTE

Berikut merupakan hasil pengujian yang dilakukan dengan algoritma Logistic Regression dengan SMOTE dengan proporsi 80:20. Detail performa klasifikasi ditampilkan pada Gambar 6.

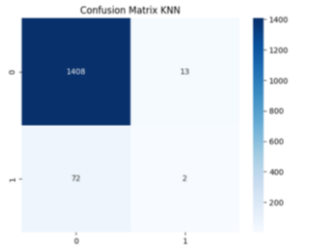


Gambar 6. Confusion Matrix Logistic Regression 80:20 dengan SMOTE

Diagram pada Gambar 6 menyajikan rincian data hasil pengujian confusion matrix, yang menunjukkan perolehan True Negative (TN) menyentuh angka 699 kasus serta True Positive (TP) yang berada pada tingkat 40 kasus. Di sisi lain, akumulasi kekeliruan prediksi dari model ini meliputi data False Positive (FP) yang terdata sebanyak 248 kasus bersama dengan False Negative (FN) yang berada pada jumlah 10 kasus. Jika dibandingkan dengan pengujian sebelum menggunakan SMOTE, jumlah pasien stroke yang terdeteksi secara tepat meningkat drastis dari 0 menjadi 40 kasus, sementara kegagalan prediksi (FN) berkurang dari 50 menjadi hanya 10 kasus. Meskipun terjadi penurunan akurasi, penerapan SMOTE berhasil mentransformasi model Logistic Regression menjadi sangat sensitif dan efektif dalam mengenali risiko stroke dibandingkan kondisi data asli.

3.3.5. Pengujian 70:30 KNN tanpa SMOTE

Hasil pengujian menggunakan algoritma KNN tanpa SMOTE dengan proporsi 70:30 dapat dilihat pada Gambar 7 yang menunjukkan confusion matrix.

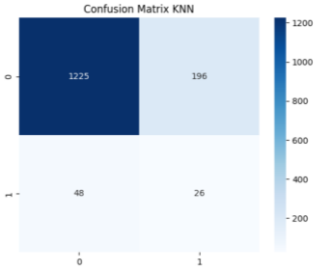


Gambar 7. Confusion Matrix KNN 70:30 tanpa SMOTE

Gambar 7 menunjukkan hasil dari confusion matrix, yang mencakup rincian dengan nilai True Negative (TN) sebesar 1408, False Positive (FP) sebanyak 13, False Negative (FN) sejumlah 72, dan True Positive (TP) hanya 2. Hasil ini menunjukkan bahwa meskipun pembagian data diubah menjadi 70:30, model tetap mengalami kegagalan signifikan dalam mendeteksi kelas minoritas karena hanya mampu mengidentifikasi 2 dari 74 total kasus stroke. Tingginya angka False Negative dibandingkan True Positive mempertegas bahwa tanpa penyeimbangan data, model cenderung hanya memprediksi kelas mayoritas dan tidak efektif untuk tujuan deteksi medis.

3.3.6. Pengujian 70:30 KNN dengan SMOTE

Hasil pengujian menggunakan algoritma KNN dengan SMOTE dengan proporsi 70:30 dapat diperhatikan pada Gambar 8 yang menunjukkan confusion matrix.

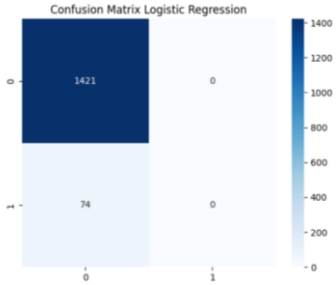


Gambar 8. Confusion Matrix KNN 70:30 dengan SMOTE

Data visual yang disajikan pada Gambar 8 memaparkan rincian capaian dari pengujian confusion matrix. Berdasarkan grafik tersebut, model berhasil mengidentifikasi kelas negatif dengan tepat yang ditunjukkan oleh angka True Negative (TN) sebesar 1225 kasus, diikuti dengan kenaikan nilai True Positive (TP) yang menyentuh angka 26 kasus. Sementara itu, tingkat kekeliruan klasifikasi pada pengujian ini mencakup False Positive (FP) sebanyak 196 kasus serta False Negative (FN) sejumlah 48 kasus. Dibandingkan dengan hasil sebelum menggunakan SMOTE pada rasio yang sama, jumlah pasien stroke yang berhasil teridentifikasi secara benar (TP) meningkat signifikan dari 2 menjadi 26 kasus. Meskipun terjadi penurunan akurasi sebagai dampak dari penyeimbangan data, penggunaan SMOTE terbukti memberikan kontribusi positif dalam meningkatkan sensitivitas model KNN untuk mengenali kelas minoritas pada pembagian data 70:30.

3.3.7. Pengujian 70:30 Logistic Regression tanpa SMOTE

Berikut merupakan hasil pengujian yang dilakukan dengan algoritma Logistic Regression tanpa SMOTE dengan proporsi 70:30. Confusion Matrix dilihat pada Gambar 9.

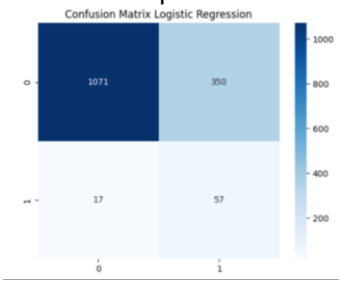


Gambar 9. Confusion Matrix Logistic Regression 70:30 tanpa SMOTE

Hasil pengujian melalui confusion matrix yang disajikan pada Gambar 9 memaparkan rincian performa model dengan capaian True Negative (TN) sebesar 1421 kasus, sedangkan nilai False Positive (FP) berada pada angka 0 kasus. Sebaliknya, kemampuan model dalam mendeteksi data positif menghasilkan nilai False Negative (FN) sebanyak 74 kasus tanpa adanya data yang berhasil dikategorikan sebagai True Positive (TP) atau bernilai 0 kasus. Hasil ini menunjukkan bahwa meskipun akurasi yang diperoleh sangat tinggi, model mengalami kegagalan total dalam mendeteksi kelas stroke karena memprediksi seluruh sampel ke dalam kelas mayoritas. Hal ini mempertegas bahwa tanpa penyeimbangan data, model Logistic Regression pada pembagian data 70:30 tidak memiliki kemampuan deteksi risiko penyakit yang efektif bagi kelas minoritas.

3.3.8. Pengujian 70:30 Logistic Regression dengan SMOTE

Berikut merupakan hasil pengujian yang dilakukan dengan algoritma Logistic Regression dengan SMOTE dengan proporsi 70:30. Detail performa klasifikasi dilihat pada Gambar 10.



Gambar 10. Confusion Matrix Logistic Regression 70:30 dengan SMOTE

Ilustrasi grafis pada Gambar 10 menyajikan perincian hasil pengujian melalui confusion matrix. Data tersebut menunjukkan kemampuan model dalam mengklasifikasikan data secara tepat dengan perolehan True Negative (TN) sebanyak 1071 kasus dan True Positive (TP) yang menyentuh angka 57 kasus. Adapun margin kesalahan prediksi dalam model ini meliputi indikasi False Positive (FP) yang mencapai 350 kasus serta False Negative (FN) sejumlah 17 kasus. Perbandingan dengan hasil sebelum menggunakan SMOTE memperlihatkan perubahan drastis di mana model kini mampu mendeteksi secara tepat 57 dari total 74 kasus stroke, sementara kegagalan prediksi (FN) berkurang dari 74 menjadi hanya 17 kasus. Meskipun terjadi penurunan akurasi, penerapan SMOTE berhasil mengoptimalkan kemampuan model dalam mengenali risiko stroke secara jauh lebih efektif dibandingkan kondisi data asli pada pembagian data 70:30.

3.3.9. Perbandingan Hasil

Implementasi pengujian yang komprehensif melalui instrumen confusion matrix pada seluruh skenario eksperimen berhasil mengukur dan memetakan parameter performa model secara terperinci. Hasil analisis tersebut menyajikan data kuantitatif yang valid mengenai tingkat akurasi serta daya ingat (recall) dari algoritma yang diuji. Untuk memberikan gambaran perbandingan yang lebih komprehensif dan sistematis, seluruh metrik tersebut dirangkum ke dalam Tabel 6. Hal ini bertujuan untuk mempermudah analisis efektivitas antara algoritma KNN dan Logistic Regression, sebagaimana rinciannya disajikan sebagai berikut:

Tabel 6. Perbandingan Hasil

Rasio Data	Algoritma	Skenario	Akurasi	Recall
80:20	KNN	Tanpa SMOTE	93.88%	2.00%
		Dengan SMOTE	82.85%	30.00%
	Logistic Regression	Tanpa SMOTE	94.98%	0.00%
		Dengan SMOTE	74.12%	80.00%
70:30	KNN	Tanpa SMOTE	94.31%	2.70%
		Dengan SMOTE	83.68%	35.14%
	Logistic Regression	Tanpa SMOTE	95.05%	0.00%
		Dengan SMOTE	75.45%	77.03%

Merujuk pada data empiris yang disajikan dalam Tabel 6, algoritma Logistic Regression menunjukkan tingkat efektivitas yang lebih unggul daripada metode KNN dalam hal

mengklasifikasikan potensi risiko penyakit stroke. Keunggulan ini dibuktikan pada rasio 80:20 dengan perolehan recall tertinggi sebesar 80,00%, yang mencerminkan sensitivitas deteksi jauh lebih baik daripada KNN. Secara keseluruhan, Logistic Regression pada skenario 80:20 merupakan model paling optimal karena memberikan hasil identifikasi risiko yang lebih stabil dan relevan untuk kebutuhan medis.

3.4. Pembahasan

Hasil pengujian yang telah dipaparkan sebelumnya memberikan sudut pandang menarik mengenai kinerja algoritma pada data medis yang tidak seimbang. Fokus utama riset ini sejak awal adalah menemukan metode paling efektif untuk mengenali risiko stroke secara dini, dan temuan empiris menunjukkan bahwa konfigurasi Logistic Regression yang dipadukan dengan teknik SMOTE pada rasio 80:20 adalah pilihan yang paling unggul. Dengan perolehan recall sebesar 80,00% dan akurasi 74,12%, model ini berhasil menjawab tujuan penelitian dengan sangat baik. Dalam praktik medis, kemampuan algoritma untuk meminimalkan kegagalan deteksi pada orang sakit atau false negative menjadi tolok ukur yang sangat krusial. Kegagalan mengenali pasien yang sebenarnya berisiko stroke bisa berakibat fatal bagi keselamatan jiwa, sedangkan kesalahan deteksi pada orang sehat atau false positive dipandang sebagai konsekuensi klinis yang lebih ringan karena masih dapat divalidasi ulang melalui pemeriksaan laboratorium lanjutan.

Secara teoretis, keunggulan Logistic Regression dibandingkan KNN dalam studi ini berakar pada cara kedua metode tersebut membangun batas keputusan. Logistic Regression bekerja menggunakan fungsi sigmoid yang memetakan probabilitas berdasarkan bobot variabel secara global dan mantap. Sifatnya yang parametrik membuat Logistic Regression tetap stabil meskipun data telah dimanipulasi melalui teknik oversampling SMOTE. Di sisi lain, kinerja KNN sangat bergantung pada perhitungan jarak spasial antar tetangga terdekat secara lokal. Ketika SMOTE menciptakan banyak sampel sintetis di wilayah yang padat, lingkungan di sekitar titik data menjadi sangat berisik dan memicu tumpang tindih antar kelas atau overlapping. Kondisi ini membuat KNN seringkali kesulitan menentukan batas klasifikasi yang tegas, sehingga konsistensi prediksinya pada data baru cenderung menurun. Fenomena ini memperkuat pandangan bahwa model berbasis statistik linear seringkali memberikan transparansi serta stabilitas yang lebih baik pada data tabular medis dibandingkan model yang hanya mengandalkan kedekatan jarak antar objek.

Jika dibandingkan dengan jejak riset sebelumnya, temuan ini memberikan kontribusi diskursus yang signifikan. Maskuri dkk. (2022) sebelumnya mencatat adanya kendala besar pada penggunaan KNN dalam menghadapi variabilitas data stroke yang tinggi [18], dan riset ini seolah mengonfirmasi hal tersebut dalam skala data yang jauh lebih besar. Meskipun Mualfah dkk. (2022) telah memvalidasi keampuhan SMOTE dalam menyeimbangkan proporsi kelas minoritas pada kasus stroke [19], penelitian ini melangkah lebih jauh dengan membuktikan bahwa pemilihan algoritma setelah penyeimbangan data adalah kunci utama stabilitas model. Hal ini selaras dengan argumen Rachmatullah dkk. (2025) mengenai urgensi meminimalkan angka false negative pada deteksi dini kondisi medis yang mematikan [20]. Selain itu, pemilihan rasio pembagian data dalam studi ini juga memperkuat analisis Zuama dkk. (2022) yang menyatakan bahwa distribusi data yang tepat sangat memengaruhi reliabilitas model pembelajaran mesin pada dataset cerebrovascular [21].

Ada perbedaan menarik jika disandingkan dengan riset Bakari dkk. (2025) yang menggunakan ANN dan SVM [22]. Meskipun riset mereka mencapai akurasi sangat tinggi, mereka menggunakan dataset berukuran kecil. Studi ini, dengan dataset yang mencakup hampir 5.000 record, membuktikan bahwa untuk data skala besar, Logistic Regression menawarkan reliabilitas yang lebih bisa dipertanggungjawabkan bagi sistem deteksi dini di lapangan. Temuan ini juga didukung oleh Kusuma dan Prasetyo (2024) yang menekankan bahwa integrasi SMOTE dengan algoritma klasifikasi mampu mengubah parameter klinis menjadi informasi penyelamat nyawa pasien [23]. Lebih lanjut, Fajriati (2026) menjelaskan bahwa metode berbasis statistik memberikan keuntungan dalam hal interpretabilitas klinis [24]. Hal ini dipertegas oleh Sulaksono dkk. (2026) yang menemukan bahwa meskipun algoritma kompleks seperti Random Forest memberikan akurasi tinggi, Logistic Regression tetap memiliki posisi strategis karena memudahkan tenaga medis memahami alasan di balik sebuah prediksi melalui koefisien fitur yang dihasilkan [25].

Walaupun sensitivitas yang dihasilkan sudah sangat memuaskan bagi kebutuhan medis, penelitian ini tidak lepas dari kendala teknis berupa penurunan nilai presisi. Munculnya angka positif palsu dalam jumlah yang cukup signifikan merupakan konsekuensi yang sulit dihindari saat teknik SMOTE berupaya memaksakan keseimbangan data pada area perbatasan kelas yang cenderung ambigu. Namun, keterbatasan ini justru menjadi pemantik bagi peluang penelitian lanjutan yang lebih mendalam. Arah pengembangan ke depan dapat difokuskan pada penerapan teknik Hybrid Sampling, yang menyinergikan SMOTE dengan metode pembersihan data untuk menapis noise sintetis yang berpotensi menyesatkan model. Melalui pendekatan tersebut, diharapkan lahir arsitektur prediktif yang tidak hanya tajam dalam menangkap risiko stroke, tetapi juga memiliki tingkat akurasi yang lebih stabil demi menjaga standar keselamatan pasien yang menjadi pilar utama dalam inovasi teknologi kesehatan.

4. Simpulan

Penelitian ini menemukan bahwa algoritma Regresi Logistik lebih efektif dan dapat diandalkan dibandingkan K-Nearest Neighbor (KNN) dalam mengidentifikasi risiko stroke. Hasil pengujian menunjukkan Logistic Regression lebih sensitif dan konsisten, dengan performa terbaik pada rasio data 80:20 yang menghasilkan recall 80,00% dan akurasi 74,12%. Secara matematis, Logistic Regression terbukti lebih efektif dalam menangani variabel klinis dibandingkan pendekatan berbasis jarak pada KNN. Oleh karena itu, Logistic Regression direkomendasikan sebagai metode utama dalam pengembangan sistem deteksi dini risiko stroke karena kemampuan deteksinya yang lebih stabil dan akurat.

Daftar Referensi

[1] L. Nugraheni, J. T. Atmojo, and A. S. Mubarok, “Efektivitas Pemberian Elevasi Kepala 30 Derajat Dalam Peningkatan Saturasi Oksigen Pada Pasien Stroke: Literature Review,” *Journal of Language and Health*, vol. 5, no. 2, pp. 435–444, 2024, [Online]. Available: http://jurnal.globalhealthsciencegroup.com/index.php/JLH

[2] A. H. Siregar, A. A. Tumanggor, and A. Rahmadani, “Penerapan K-Nearest Neighbors (KNN) dalam Memprediksi dan Menghitung Akurasi Data Penyakit Stroke,” *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 2, no. 4, pp. 146–154, Nov. 2023, doi: 10.55606/juprit.v2i4.3040.

[3] E. F. Nopitasari, S. P. A. Alkadri, and R. W. S. Insani, “Implementasi Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor,” *KREATIF: Jurnal Pengabdian Masyarakat Nusantara*, vol. 5, no. 3, pp. 344–365, Sep. 2025, doi: 10.55606/kreatif.v5i3.8215.

[4] D. Nasien *et al.*, “Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes,” *JURNAL TEKNIK INFORMATIKA*, vol. 4, no. 1, pp. 10–17, 2024.

[5] D. Sitanggang *et al.*, “Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor Dan Logistic Regression,” *Jurnal TEKINKOM*, vol. 5, no. 2, pp. 493–501, 2022, doi: 10.37600/tekinkom.v5i1.698.

[6] Y. Azhar, A. K. Firdausy, and P. J. Amelia, “Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke,” *SINTECH*, vol. 5, no. 2, pp. 191–197, 2022, [Online]. Available: https://doi.org/10.31598

[7] Zuriati Z and Qomariyah N, “Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN),” *Jurnal Sistem dan Teknologi Informasi*, vol. 1, no. 1, pp. 1–8, 2023, doi: 10.20222/rt.v1i1.2665.

[8] H. A. D. Fasnuari, H. Yuana, and M. T. Chulkamdi, “Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Klasifikasi Penyakit Diabetes Melitus Studi Kasus : Warga Desa Jatitengah,” *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 16, no. 2, pp. 133–142, Oct. 2022, doi: 10.35457/antivirus.v16i2.2445.

[9] L. N. Mukarromah, Z. Fatah, and I. Yunita, “Klasifikasi Penyakit Diabetes Menggunakan Metode K-Nearest Neighbors (KNN),” *JURNAL RISET TEKNIK KOMPUTER*, vol. 1, no. 4, pp. 41–46, 2024, doi: 10.69714/jgq41610.

[10] A. A. Putri Lo and V. J. E. Tjioe, “Penerapan Model CRISP-DM untuk Prediksi Penyakit Diabetes Menggunakan Metode K-Nearest Neighbor dan Logistic regression,” *Prosiding SENAM*, vol. 4, pp. 48–57, 2024.

[11] N. H. Setyawan and N. Wakhidah, “Analisis Perbandingan Metode Logistic Regression, Random Forest, Gradient Boosting Untuk Prediksi Diabetes,” *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 150–162, Jan. 2025, doi: 10.29100/jipi.v10i1.5743.
[12] H. Hatta Irsyad, M. I. Syafwan, and D. Ramadhani, “Analisis Perbandingan Kinerja Algoritma K-Nearest Neighbors DAN Support Vector Machine Untuk Klasifikasi Penyakit Diabetes,” *Journal of System & Technology*, vol. 1, no. 2, pp. 43–48, 2025, [Online]. Available: https://doi.org/XX.XXXXX/systec.X.X.X-XX
[13] Z. Khairi, R. Yanti, T. A. Fitri, and E. Fatdha, “Optimasi Algoritma Knn Menggunakan Smote Untuk Prediksi Stroke,” *Jurnal Algoritma*, vol. 22, no. 2, pp. 164–175, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2474.
[14] A. Aryanti, F. P. Daviana, N. D. Ulhaq, M. Riswanto, M. D. Alfajri, and N. T. Novita, “Perbandingan Metode K-Nearst Neighbor dan Logistic Regression dalan Klasifikasi Dini Anemia Berdasarkan Parameter Hematologi Dasar,” *Jurnal Riset Rekayasa Elektro*, vol. 7, no. 2, pp. 251–258, 2025.
[15] Devian, P. N. Sabrina, and A. Komarudin, “Prediksi Penyakit Diabetes Dengan Metode K-Nearest Neighbor (KNN) Dan Seleksi Fitur Information Gain,” *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 6, pp. 11320–11326, 2024.
[16] V. Artanti, M. Faisal, and F. Kurniawan, “Klasifikasi Cardiovascular Diseases Menggunakan Algoritma K-Nearest Neighbors (KNN),” *Techno.COM*, vol. 23, no. 2, pp. 469–481, 2024.
[17] A. C. Firdaus and I. Veritawati, “Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Stroke pada Rumah Sakit Pusat Otak Nasional Prof. Dr. dr. Mahar Mardjono Jakarta,” *TEKINFO*, vol. 26, no. 1, pp. 144–153, 2025, doi: 10.37817/Tekinfo.v26i1.
[18] M. N. Maskuri, Harliana, K. Sukerti, and H. R. M. Bhakti, “Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke,” *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 4, no. 1, pp. 130–140, 2022.
[19] D. Mualfah, W. Fadila, and R. Firdaus, “Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 2, pp. 107–113, Aug. 2022, doi: 10.37859/coscitech.v3i2.3912.
[20] M. I. C. Rachmatullah, S. Armiati, and Mubassiran, “Menerapkan SMOTE Pada Klasifikasi Data Penyakit Stroke,” *Jurnal Ilmiah Manajemen Informatika*, vol. 17, no. 1, pp. 9–12, 2025.
[21] R. A. Zuama, S. Rahmatullah, and Y. Yuliani, “Analisis Performa Algoritma Machine Learning pada Prediksi Penyakit Cerebrovascular Accidents,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, pp. 531–534, Jan. 2022, doi: 10.30865/mib.v6i1.3488.
[22] S. Nurfatimah Bakari, A. Lasarudin, and W. Hasyim, “Penerapan Data Mining Dalam Mengklasifikasi Resiko Penyakit Stroke Menggunakan Algoritma SVM Dan ANN,” *JURNAL ILMU KOMPUTER (JUIK)*, vol. 5, no. 2, pp. 114–125, 2025, [Online]. Available: https://journal.umgo.ac.id/index.php/juik/index
[23] F. Ananda Kusuma and B. Prasetyo, “Pemodelan Klasifikasi Anemia Aplastik Menggunakan Teknik Oversampling Dan K-Nearest Neighbors,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 1587–1592, Aug. 2024, doi: 10.23960/jitet.v12i3.4326.
[24] D. Fajriati, “Implementasi Metode Machine Learning Berbasis Statistik untuk Klasifikasi Data Medis,” *Jurnal Matematikan dan Sains Alam*, vol. 1, no. 1, pp. 15–21, 2026, [Online]. Available: https://ojs.feliciajurnal.com/index.php/jmsa/index
[25] J. Sulaksono, D. Putra Pamungkas, and D. W. Widodo, “Analisis Perbandingan Model Machine Learning Untuk Prediksi Penyakit Jantung,” *Prosiding Seminar Nasional Teknologi Dan Sains*, vol. 5, pp. 707–712, 2026.