Kreatif : Jurnal Pengabdian Masyarakat Nusantara
Volume 5 Nomor 3 September 2025
e-ISSN: 2962-3839; p-ISSN: 2962-4436, Hal 344-365
DOI: https://doi.org/10.55606/kreatif.v5i3.8215
Tersedia : https://researchhub.id/index.php/kreatif

Implementasi Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma *K-Nearest Neighbor*

Enkan Feny Nopitasari[1*], Syarifah Putri Agustini Alkadri[2], Rachmat Wahid Saleh Insani[3]
[1-3]Universitas Muhammadiyah Pontianak, Indonesia
E-mail: 201220068@unmuhpnk.ac.id [1]*

Alamat: Jl. Jenderal Ahmad Yani No.111, Bangka Belitung Laut, Pontianak Tenggara, Kota Pontianak, Kalimantan Barat, Indonesia 78123
[*]Korespondensi penulis

Abstract*. Stroke remains a major global health challenge, with diagnoses often delayed, particularly in primary care facilities with limited infrastructure. This study aimed to develop a stroke risk classification system using the K-Nearest Neighbor (KNN) algorithm, optimized through comprehensive data preprocessing. A secondary dataset of 5,110 patient records was processed using mean imputation for missing BMI values, winsorization to manage outliers, label encoding for categorical variables, and Min-Max normalization for feature scaling. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied prior to stratified data splitting into 70% training and 30% testing sets. The KNN model with K=5 demonstrated strong performance, achieving 96% accuracy, 96% precision, 99% recall, and a 97% F1-score on the test data. Multivariate correlation analysis identified age, hypertension, and blood glucose levels as the primary predictors of stroke risk, consistent with established clinical pathophysiology. These findings highlight the critical role of cardiometabolic risk factors in early detection. The system was implemented as a web application using Streamlit, enabling rapid and interactive screening in primary healthcare centers with minimal infrastructure. This practical application has the potential to assist healthcare providers in early stroke detection, accelerating clinical intervention and reducing the likelihood of long-term complications. Nevertheless, several limitations exist. The reliance on secondary data introduces the possibility of regional bias, and the use of SMOTE generates synthetic data that may affect model generalizability. Future research is recommended to validate the model across multi-source datasets, apply advanced hyperparameter tuning, and explore ensemble learning techniques to further enhance predictive reliability. In conclusion, the KNN-based classification system demonstrates promising potential as a practical decision-support tool for early stroke risk assessment in resource-limited healthcare settings.*

Keywords: *Data preprocessing; K-Nearest Neighbor; SMOTE; Streamlit; Stroke classification.*

Abstrak. Stroke masih menjadi tantangan kesehatan global dengan diagnosis yang sering terlambat, khususnya di fasilitas primer dengan keterbatasan infrastruktur. Penelitian ini bertujuan mengembangkan sistem klasifikasi risiko stroke berbasis algoritma K-Nearest Neighbor (KNN) melalui tahapan preprocessing data yang komprehensif. Dataset sekunder berjumlah 5.110 rekaman pasien diproses menggunakan imputasi mean untuk menangani nilai BMI yang hilang, winsorization guna mengendalikan outlier, label encoding untuk variabel kategorik, serta normalisasi Min-Max agar data berada pada skala seragam. Untuk mengatasi ketidakseimbangan kelas, Synthetic Minority Over-sampling Technique (SMOTE) diterapkan sebelum pembagian data secara stratifikasi dengan rasio 70% pelatihan dan 30% pengujian. Model KNN dengan K=5 menghasilkan performa tinggi dengan akurasi 96%, presisi 96%, recall 99%, dan F1-score 97% pada data uji. Analisis korelasi multivariat menunjukkan bahwa usia, hipertensi, dan kadar glukosa darah merupakan prediktor utama risiko stroke, sesuai dengan dasar patofisiologi klinis. Temuan ini menguatkan bahwa faktor risiko kardiometabolik memiliki peran penting dalam deteksi dini. Implementasi sistem dilakukan melalui aplikasi web berbasis Streamlit, memungkinkan skrining cepat dan interaktif di puskesmas dengan infrastruktur minimal. Dengan demikian, sistem ini berpotensi mendukung tenaga kesehatan dalam melakukan deteksi dini, mempercepat intervensi, serta menurunkan risiko komplikasi jangka panjang akibat stroke. Meskipun demikian, penelitian ini memiliki keterbatasan. Penggunaan data sekunder berpotensi menimbulkan bias regional, sementara penerapan SMOTE menghasilkan data sintetis yang dapat memengaruhi generalisasi model. Oleh karena itu, penelitian lanjutan direkomendasikan dengan menguji performa pada dataset multi-sumber, melakukan hyperparameter tuning lanjutan, serta mengeksplorasi pendekatan ensemble learning untuk meningkatkan keandalan prediksi. Dengan hasil ini, sistem klasifikasi berbasis KNN menawarkan potensi praktis sebagai alat bantu deteksi risiko stroke di lini pelayanan kesehatan dasar.

Kata kunci: Klasifikasi stroke; K-Nearest Neighbor; Preprocessing data; SMOTE; Streamlit.

Naskah Masuk: 31 Juli 2025; Revisi: 15 Agustus 2025; Diterima: 26 Agustus 2025; Terbit: 02 September 2025;

1. LATAR BELAKANG

Stroke salah satu tantangan kesehatan global yang paling signifikan (Nikmawati, 2024), menempati posisi kedua sebagai penyebab kematian dan ketiga sebagai penyebab kematian serta disabilitas gabungan di seluruh dunia dengan beban ekonomi mencapai lebih dari US$890 miliar per tahun atau setara dengan 0,66% dari PDB global. World Stroke Organization mengindikasikan bahwa antara tahun 1990-2021, beban stroke meningkat secara substansial dengan peningkatan 70% pada kasus insiden stroke, 44% pada kematian akibat stroke, dan 86% pada prevalensi stroke, dimana 87% kematian dan 89% DALYs terpusat pada negara-negara berpendapatan rendah dan menengah. Proyeksi menunjukkan bahwa kematian akibat stroke akan meningkat 50% antara tahun 2020-2050, dari 6,6 juta menjadi 9,7 juta kasus per tahun, menciptakan ancaman serius terhadap keberlanjutan sistem kesehatan global. Tantangan utama dalam penanganan stroke terletak pada keterlambatan diagnosis dan identifikasi dini yang sering terjadi, padahal setiap menit keterlambatan dapat menyebabkan kerusakan otak ireversibel, khususnya di daerah dengan keterbatasan akses terhadap teknologi pencitraan medis canggih seperti CT dan MRI.

Perkembangan machine learning dalam bidang medis telah menunjukkan potensi revolusioner untuk deteksi dini penyakit (Bintang, 2024), dengan algoritma K-Nearest Neighbor (KNN) menjadi salah satu metode yang paling sering diterapkan karena kesederhanaan dan adaptabilitasnya (Brian, 2025). Systematic review terbaru menunjukkan bahwa penelitian machine learning untuk prediksi stroke telah mencapai puncaknya pada tahun 2021 (Siregar, 2025) dengan berbagai studi mendemonstrasikan akurasi tinggi dalam klasifikasi medis. Algoritma KNN telah terbukti efektif dalam berbagai aplikasi kesehatan, seperti klasifikasi penyakit diabetes dengan akurasi 93%, penyakit Parkinson dengan akurasi 96%, dan penyakit jantung dengan akurasi 92% (Akbar 2025). Namun, studi comparative performance analysis mengidentifikasi bahwa algoritma KNN klasik menghadapi berbagai limitasi yang mengurangi kemampuan klasifikasinya (Maulana, 2024), termasuk sensitivitas terhadap outlier, ketidakmampuan menangani fitur yang tidak relevan, dan masalah perhitungan jarak antar data point (Carudin, 2024). Dari 10 varian KNN yang dianalisis, rata-rata akurasi berkisar antara 64,22% hingga 83,62%, dengan Hassanat KNN menunjukkan performa terbaik (Maulana, 2025).

Meskipun terdapat kemajuan signifikan dalam penerapan machine learning untuk prediksi stroke, beberapa kesenjangan kritikal masih perlu diatasi yang menciptakan urgensi penelitian ini. Pertama, sebagian besar penelitian machine learning untuk stroke masih terbatas pada validasi teknis tanpa implementasi praktis dalam setting klinis nyata, dimana hanya dua

dari 18 studi dalam systematic review yang melakukan validasi eksternal. Kedua, dataset medis umumnya mengalami ketidakseimbangan kelas yang severe, dimana kasus penyakit (minoritas) jauh lebih sedikit dibandingkan kasus sehat (mayoritas), sehingga dapat menyebabkan bias model terhadap kelas mayoritas dan mengurangi kemampuan deteksi kasus positif. Ketiga, banyak penelitian menggunakan metode preprocessing sederhana seperti complete case analysis dan single imputation, padahal proper data preprocessing memerlukan pendekatan yang lebih sophisticated untuk menangani missing values, outliers, dan noise yang inheren dalam data medis. Keempat, terdapat kesenjangan signifikan antara penelitian akademis dan aplikasi praktis, dimana sistem yang dikembangkan seringkali tidak accessible bagi praktisi kesehatan di lapangan, terutama di fasilitas kesehatan dengan sumber daya terbatas.

Kebaruan penelitian ini terletak pada pengembangan pendekatan preprocessing yang komprehensif dan terintegrasi yang secara sistematis mengatasi multiple challenge dalam data medis, meliputi penanganan missing values, deteksi dan penanganan outlier menggunakan metode winsorization, label encoding yang optimal, serta penanganan ketidakseimbangan data menggunakan SMOTE (Synthetic Minority Oversampling Technique). Kontribusi unik penelitian ini adalah implementasi sistem klasifikasi stroke berbasis KNN yang dioptimasi secara khusus untuk karakteristik data stroke, dengan normalisasi fitur menggunakan Min-Max Scaler dan pemilihan parameter K yang optimal melalui eksperimen sistematis, kemudian diintegrasikan ke dalam aplikasi web menggunakan framework Streamlit untuk memfasilitasi implementasi praktis dalam setting klinis. Berbeda dari studi sebelumnya yang hanya fokus pada aspek algoritma, penelitian ini menghadirkan solusi end-to-end yang menjembatani gap antara penelitian akademis dan aplikasi praktis melalui evaluasi multi-metrik komprehensif menggunakan accuracy, precision, recall, F1-score, dan confusion matrix.

Berdasarkan identifikasi gap tersebut, penelitian ini bertujuan untuk mengembangkan sistem klasifikasi penyakit stroke yang mengintegrasikan algoritma K-Nearest Neighbor dengan teknik preprocessing komprehensif untuk mengatasi tantangan data medis yang kompleks, mengoptimalkan performa prediksi melalui penanganan sistematis terhadap masalah ketidakseimbangan data, outlier, dan missing values, serta menciptakan solusi applicable dalam bentuk aplikasi web user-friendly untuk mendukung screening awal stroke di fasilitas kesehatan dengan akses terbatas terhadap teknologi pencitraan medis canggih.
Penelitian ini diharapkan memberikan kontribusi metodologis dalam penerapan KNN untuk klasifikasi medis serta mendukung upaya early detection stroke untuk meningkatkan patient outcomes melalui intervensi yang lebih cepat dan tepat sasaran.

2. KAJIAN TEORITIS

Stroke didefinisikan sebagai kondisi neurologis akut yang terjadi akibat gangguan aliran darah menuju otak, yang dapat disebabkan oleh penyumbatan (stroke iskemik) atau pecahnya pembuluh darah otak (stroke hemoragik). Dalam konteks epidemiologi global, World Stroke Organization melaporkan bahwa stroke merupakan penyebab kematian kedua terbesar di dunia dengan beban ekonomi mencapai US$890 miliar per tahun. Patofisiologi stroke melibatkan cascade kompleks yang dimulai dengan hipoksia serebral, diikuti oleh disfungsi mitokondria, pelepasan glutamat berlebihan, influx kalsium intraseluler, dan aktivasi jalur apoptosis yang berujung pada kematian neuron. Faktor risiko stroke dapat dikategorikan menjadi yang dapat dimodifikasi (hipertensi, diabetes mellitus, dislipidemia, fibrilasi atrium, merokok) dan yang tidak dapat dimodifikasi (usia, jenis kelamin, riwayat keluarga, genetik). Kompleksitas interaksi multifaktorial ini menciptakan tantangan signifikan dalam prediksi risiko stroke yang memerlukan pendekatan analitik yang sophisticated.

Data mining merupakan proses ekstraksi pengetahuan dari dataset besar melalui penerapan teknik statistik, machine learning, dan database management untuk mengidentifikasi pola tersembunyi yang dapat memberikan insight prediktif (Syam, 2024). Dalam konteks medis, data mining memiliki peran fundamental dalam mendukung clinical decision-making melalui berbagai teknik seperti klasifikasi, clustering, asosiasi, dan prediksi (Hendriyani, 2025). Algoritma K-Nearest Neighbor (KNN) merupakan metode supervised learning non-parametrik yang melakukan klasifikasi berdasarkan similarity measure antara data point dengan k tetangga terdekatnya dalam feature space (Ujianto, 2025). Keunggulan teoritis KNN terletak pada asumsi lokalitas, dimana data point yang berdekatan dalam ruang fitur memiliki kecenderungan untuk memiliki label kelas yang sama (Putra, 2023). Prinsip fundamental KNN didasarkan pada distance metric (Bakri, 2025), umumnya Euclidean distance yang diformulasikan sebagai:

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$

Keterangan :

d(x , y) : jarak data x dan y

xi : sampel data

yi : data uji atau data testing

i : variable data

n : dimensi data

Dimana pemilihan nilai k menjadi critical hyperparameter yang mempengaruhi bias-variance trade-off. Algoritma ini tidak memerlukan asumsi distribusi probabilitas spesifik terhadap data (Watratan, 2020), menjadikannya robust untuk berbagai jenis dataset medis yang seringkali memiliki distribusi kompleks dan non-linear (Mutoffar, 2025).

Systematic review dalam bidang machine learning untuk prediksi stroke menunjukkan tren penelitian yang meningkat signifikan (Nova, 2025). Penelitian Fasnuari (2022). mengimplementasikan KNN untuk klasifikasi diabetes mellitus menggunakan 135 sampel dengan pembagian 80:20, menghasilkan akurasi 93%, precision 100%, recall 60%, dan F1-score 75% dengan parameter K=9 dan normalisasi euclidean distance. Studi komparatif Aprilitaz (2023) membandingkan KNN dan Naive Bayes untuk klasifikasi Parkinson menggunakan 756 data dari Kaggle, dimana KNN dengan pembagian 70:30 mencapai akurasi superior 96% dibandingkan Naive Bayes. Sahelvi (2025) menerapkan KNN pada klasifikasi penyakit jantung dengan pembagian 80:20, menghasilkan performance metrics: akurasi 92%, precision 90%, dan recall 92%. Penelitian Shinami (2023) menggunakan KNN untuk klasifikasi kanker payudara dengan K=8 dan pembagian 80:20, mencapai akurasi 77%, precision 76%, dan recall 71%. Studi A'yuniyah (2023) mengimplementasikan KNN menggunakan RapidMiner untuk klasifikasi jurusan siswa dengan pembagian 80:20, menghasilkan akurasi optimal 93.52% pada K=3, precision rata-rata 88.14%, dan recall rata-rata 100%.

Studi comparative performance analysis yang menganalisis 10 varian KNN menunjukkan variabilitas performa yang signifikan, dengan akurasi berkisar 64.22%-83.62%, dimana Hassanat KNN menunjukkan performance terbaik. Analisis ini mengidentifikasi bahwa limitasi fundamental algoritma KNN meliputi: curse of dimensionality, sensitivitas terhadap irrelevant features, computational complexity yang tinggi untuk dataset besar, dan sensitivitas terhadap outliers. Penelitian Desiani (2022) mendemonstrasikan superioritas KNN dibandingkan Naive Bayes dalam klasifikasi penyakit hati, dengan KNN mencapai 100% untuk semua metrik evaluasi dibandingkan 85% untuk Naive Bayes, mengindikasikan kemampuan KNN dalam menangani complex pattern tanpa bergantung pada probabilistic assumptions.

Dalam konteks data medis, preprocessing merupakan tahapan critical yang mempengaruhi performance model secara fundamental. Systematic review menunjukkan bahwa mayoritas penelitian machine learning untuk prediksi stroke menggunakan metode preprocessing sederhana seperti complete case analysis dan single imputation, padahal advanced techniques seperti multiple imputation terbukti lebih efektif dalam menangani missing data complexity. Class imbalance merupakan challenge intrinsik dalam medical

datasets, dimana prevalensi penyakit (minority class) umumnya jauh lebih rendah dibandingkan populasi sehat (majority class). Teknik resampling seperti SMOTE (Synthetic Minority Oversampling Technique) telah terbukti efektif dalam mengatasi imbalanced dataset melalui synthetic data generation yang mempertahankan distribusi statistik original data (Bhirawa, 2025). Outlier detection dan handling menjadi aspek fundamental dalam medical data preprocessing, dimana metode winsorization menunjukkan efektivitas dalam mempertahankan distribusi data sambil mengurangi impact extreme values.

Evaluasi performa model klasifikasi dalam domain medis memerlukan comprehensive metrics yang mencakup accuracy, precision, recall, dan F1-score yang diekstrak dari confusion matrix. Accuracy mengukur proporsi prediksi benar terhadap total prediksi, precision mengevaluasi proporsi true positive terhadap total predicted positive, recall mengukur proporsi true positive terhadap total actual positive, sedangkan F1-score merepresentasikan harmonic mean antara precision dan recall. Dalam konteks medical classification, recall (sensitivity) menjadi particularly important karena false negative dapat berakibat fatal, sementara precision menjadi relevan untuk menghindari unnecessary medical intervention akibat false positive. Framework CRISP-DM (Cross Industry Standard Process for Data Mining) menyediakan structured methodology yang mencakup business understanding, data understanding, data preparation, modeling, evaluation, dan deployment untuk memastikan systematic approach dalam data mining projects.

Kajian teoretis ini menunjukkan bahwa meskipun KNN telah terbukti efektif dalam berbagai aplikasi medical classification, masih terdapat gaps dalam optimasi preprocessing techniques, parameter tuning, dan integration ke practical clinical settings yang perlu diatasi melalui comprehensive research approach yang menggabungkan advanced preprocessing, optimal hyperparameter selection, dan user-friendly implementation dalam bentuk accessible web-based applications.

3. METODE PENELITIAN

Penelitian ini mengadopsi desain experimental dengan pendekatan applied research yang menerapkan metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) (Primajaya, 2022) sebagai kerangka kerja sistematis untuk pengembangan model prediksi stroke. CRISP-DM dipilih karena telah menjadi standar industri yang paling popular dan terbukti robust untuk proyek data mining dengan adoption rate mencapai hampir 50% berdasarkan survei terbaru (Swastika, 2023). Framework ini terdiri dari enam fase iteratif yang memungkinkan pendekatan empirical dan scientific dalam pengembangan model machine

learning, yaitu business understanding, data understanding, data preparation, modeling, evaluation, dan deployment (Nurlaela, 2025). Penelitian menggunakan paradigm quantitative research dengan focus pada pengukuran performance model menggunakan metrics objective seperti accuracy, precision, recall, dan F1-score yang dievaluasi melalui confusion matrix (Lonang, 2023).

Populasi target dalam penelitian ini adalah seluruh individu yang berisiko mengalami stroke berdasarkan karakteristik demografis dan clinical features. Sampel penelitian menggunakan dataset sekunder yang diperoleh dari Kaggle Stroke Prediction Dataset yang originally dikembangkan oleh McKinsey & Company. Dataset ini terdiri dari 5.110 observasi dengan 12 atribut medis yang mencakup variabel demografis (usia, jenis kelamin), riwayat medis (hipertensi, penyakit jantung, status pernikahan), lifestyle factors (jenis pekerjaan, tempat tinggal, status merokok), dan parameter fisiologis (kadar glukosa rata-rata, BMI). Teknik sampling yang digunakan adalah purposive sampling dimana dataset dipilih berdasarkan kriteria kelengkapan data, relevansi medical features, dan representativeness terhadap populasi stroke patients. Dataset menunjukkan karakteristik class imbalance yang significant dengan ratio approximately 1:20 antara positive cases (stroke) dan negative cases (non-stroke), yang merupakan representasi realistis dari prevalence stroke dalam populasi umum.

Pengumpulan data dilakukan melalui secondary data acquisition dari repositori Kaggle yang menyediakan dataset terstruktur dalam format CSV (Sheila, 2024). Instrumen pengumpulan data meliputi Python programming language dengan libraries khusus untuk data science yakni Pandas untuk data manipulation dan analysis, NumPy untuk numerical computing, Scikit-learn untuk machine learning algorithms implementation, Matplotlib dan Seaborn untuk data visualization. Data validation dan quality assessment dilakukan menggunakan descriptive statistics dan exploratory data analysis (EDA) untuk memahami distribusi data, mengidentifikasi outliers, missing values, dan data inconsistencies (Elfaladonna, 2024). Proses data understanding mengikuti protokol systematic data exploration yang mencakup univariate analysis untuk memahami karakteristik individual variables dan multivariate analysis menggunakan correlation matrix untuk mengidentifikasi relationship patterns antar features (Sulianta, 2023).

Tahapan data preparation menggunakan comprehensive preprocessing pipeline yang terdiri dari multiple steps untuk memastikan data quality dan model performance optimization (Santoso, 2024). Missing value handling dilakukan menggunakan mean imputation untuk continuous variables (BMI) untuk meminimalkan bias dalam dataset. Outlier detection dan

treatment menggunakan Interquartile Range (IQR) method combined dengan winsorization technique untuk mengatasi extreme values tanpa menghilangkan valuable information (Ardhani, 2025). Label encoding diterapkan pada categorical variables (gender, marital status, work type, residence type, smoking status) untuk mengkonversi data kategorik menjadi format numerical yang dapat diproses oleh machine learning algorithms (Febiola, 2025). Data imbalance diatasi menggunakan SMOTE (Synthetic Minority Oversampling Technique) yang merupakan advanced resampling method untuk mengenerate synthetic samples pada minority class dan mencegah overfitting yang common terjadi pada simple oversampling methods. Feature scaling menggunakan Min-Max Scaler untuk normalization data numerical dalam range 0-1 yang essential untuk distance-based algorithms seperti KNN (Pratama, 2025).

Model penelitian mengimplementasikan K-Nearest Neighbor (KNN) algorithm sebagai supervised learning method untuk binary classification task (stroke vs non-stroke). KNN dipilih karena merupakan non-parametric algorithm yang tidak membuat assumptions tentang underlying data distribution, making it suitable untuk medical datasets yang often memiliki complex patterns. Algorithm implementation menggunakan Euclidean distance sebagai similarity measure dengan formula $d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2}$ untuk menghitung proximity antara data points. Hyperparameter tuning dilakukan untuk menentukan optimal K value melalui experimental approach dengan testing different K values (K=1 dan K=5) untuk mengoptimalkan model performance berdasarkan accuracy metrics pada validation set. Data splitting menggunakan stratified train-test split dengan ratio 70:30 untuk memastikan representative distribution pada training dan testing sets while maintaining class proportion consistency. Model evaluation menggunakan comprehensive metrics assessment yang mencakup accuracy, precision, recall, F1-score yang diekstrak dari confusion matrix untuk memberikan holistic view terhadap model performance. Web application development menggunakan Streamlit framework sebagai deployment platform untuk memungkinkan real-time prediction dan user-friendly interface dalam clinical settings. Model persistence dilakukan menggunakan Python Pickle library untuk serialization trained model agar dapat diintegrasikan dalam production environment. Validation methodology menggunakan holdout validation approach dimana model performance dievaluasi pada unseen test data untuk mengukur generalization capability dan menghindari overfitting issues yang common dalam small medical datasets.

4. HASIL DAN PEMBAHASAN

Proses Pengumpulan Data, Rentang Waktu, dan Lokasi Penelitian

Data diperoleh secara sekunder dari Stroke Prediction Dataset di platform Kaggle, yang awalnya dikumpulkan oleh Fedesoriano et al. melalui kolaborasi rumah sakit dan klinik di India dan Eropa Utara pada periode Januari 2019 hingga Desember 2020. Unduhan file CSV dengan 5.110 rekaman pasien dilakukan pada Juli 2025 di Laboratorium Data Science Universitas Muhammadiyah Pontianak. Seluruh data divalidasi melalui eksplorasi awal untuk memastikan kelengkapan, konsistensi, dan realisme medis sebelum tahap preprocessing.

Tabel 1. Variabel

No	Nama Variabel	Keterangan	Tipe
1.	ID	Kode Pengenal	Numerik
2.	Gender	Jenis Kelamin : Perempuan (Famale) : 0, Laki-Laki (Male) : 1	Kategorik
3.	Age	Umur Pasien : <20 (0), 21-40 (1), 41-60 (2), >61 (3)	Numerik
4.	Hypertension	Hipertensi : Tidak (No) : 0, Ya (Yes) : 1	Numerik
5.	Heart Disease	Penyakit Jantung : Tidak (No) : 0, Ya (Yes) : 1	Numerik
6.	Ever Married	Pernah Menikah : Tidak (No) : 0, Ya (Yes) : 1	Kategorik
7.	Work Type	Tipe Pekerjaan : Belum Pernah Bekerja (Never Worked), Anak-anak (Children), PNS (Government Work), Pekerja Swasta (Private), Wiraswasta (Self-employed)	Kategorik
8.	Residence Type	Jenis Tempat Tinggal : Perkotaan (Urban), Perdesaan (Rural)	Kategorik
9.	Avg Glucose Level	Rata-rata Tingkat Glukosa : <=77.07 (0), 77.08-91.68 (1), 91.6, 113.57 (2), >=113.58 (3)	Numerik
10.	BMI	Indeks Massa Tubuh : <23.5 (0), 23.6-28.1 (1), 28.2-33.1 (2), >33.2 (3)	Numerik
11.	Smoking Status	Status Merokok : Tidak Diketahui (Unknown), Tidak Pernah Merokok (Never Smoked), Pernah Merokok (Formerly Smoked), Perokok Aktif (Smokes)	Kategorik
12.	Stroke	Penyakit Stroke : Stroke (0), Tidak (1)	Numerik

Tabel 1 menjabarkan dua belas variabel yang digunakan dalam penelitian klasifikasi risiko stroke. Variabel pertama, “ID”, berfungsi sebagai kode pengenal unik setiap pasien dan bersifat numerik. “Gender” mencerminkan jenis kelamin pasien dengan encoding 0 untuk perempuan dan 1 untuk laki-laki, termasuk sebagai variabel kategorik. Variabel “Age”

mengelompokkan umur pasien ke dalam empat rentang: <20 (0), 21–40 (1), 41–60 (2), dan >61 (3), serta dikategorikan sebagai numerik ordinal. Dua variabel berikutnya, "Hypertension" dan "Heart Disease", keduanya numerik biner, masing-masing merepresentasikan riwayat hipertensi dan penyakit jantung (0=tidak, 1=ya). "Ever Married" juga biner (0=belum menikah, 1=pernah menikah) tetapi bersifat kategorik karena menggambarkan status sosial. "Work Type" adalah variabel kategorik nominal dengan lima kategori pekerjaan: belum pernah bekerja, anak-anak, PNS, pekerja swasta, dan wiraswasta. "Residence Type" mencerminkan lokasi tempat tinggal pasien sebagai urban atau rural, juga kategorik. "Avg Glucose Level" dan "BMI" masing-masing dikelompokkan ke dalam empat rentang numerik (0–3) berdasarkan batas klinis, memudahkan analisis diskrit pada variabel fisiologis. "Smoking Status" adalah kategorik nominal dengan empat nilai: unknown, never smoked, formerly smoked, dan active smoker. Terakhir, variabel "Stroke" merupakan target klasifikasi numerik biner (0=stroke, 1=tidak stroke). Keseluruhan variabel ini mencakup demografis, riwayat medis, perilaku, dan indikator fisiologis yang komprehensif untuk membangun model prediksi risiko stroke.

Hasil Analisis Data

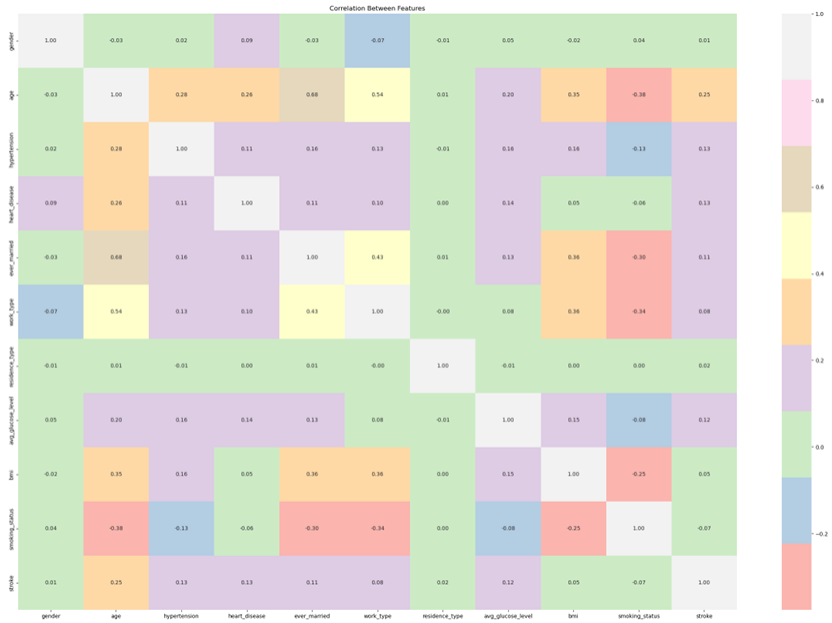
A. Univariate Analysis menunjukkan rentang nilai: usia 0,08–82 tahun (mean 43,23), kadar glukosa 55,12–168,68 mg/dL (mean 100,90), BMI 11,30–67,10 (mean 28,72).

Tabel 2. Hasil Univariate Analysis

Statistik	count	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
mean	5.109.000.000	43.229.986	0.097475	0.054022	100.904.772	28.723.046	0.048738
std	-	22.613.575	0.296633	0.226084	33.041.818	7.119.595	0.215340
min	0.080000	0.000000	0.000000	55.120.000	11.300.000	0.000000	-
25%	25.000.000	0.000000	0.000000	77.240.000	23.800.000	0.000000	-
50%	45.000.000	0.000000	0.000000	91.880.000	28.400.000	0.000000	-
75%	61.000.000	0.000000	0.000000	114.090.000	32.800.000	0.000000	-
max	82.000.000	1.000.000	1.000.000	168.680.000	46.300.000	1.000.000	-

Tabel 2 menyajikan ringkasan statistik univariat untuk tujuh variabel utama dalam dataset: usia (age), hipertensi (hypertension), penyakit jantung (heart_disease), rata-rata kadar glukosa (avg_glucose_level), indeks massa tubuh (bmi), dan label stroke. Total observasi untuk seluruh variabel adalah 5.109.000 pasien setelah pembersihan data (count). Rata-rata usia pasien hampir mencapai 44 tahun (mean age ≈ 43,99) dengan penyebaran yang luas (std ≈ 22,61), mulai dari usia paling muda 0,08 tahun hingga 82 tahun pada kelompok tertua. Prevalensi hipertensi dan penyakit jantung relatif rendah, dengan rata-rata masing-masing 9,75% dan 5,40% dari sampel. Kadar glukosa rata-rata berada di sekitar 100,90 mg/dL (std ≈ 33,04), dengan nilai terkecil 11,30 mg/dL dan tertinggi 168,68 mg/dL. BMI rata-rata tercatat 28,72 (std ≈ 7,12), mencakup rentang ekstrem 0–67,10; angka nol menggambarkan imputasi atau data awal yang terbatas. Label stroke memiliki mean 0,0487, menunjukkan frekuensi kejadian stroke di bawah 5% dari seluruh sampel. Kuartil pertama (25%) dan ketiga (75%) pada masing-masing variabel menegaskan distribusi data: misalnya 50% pasien berusia di bawah 45 tahun, kadar glukosa di bawah 28,40, dan BMI di bawah nilai tengah 0 (karena encoding kategorik), sedangkan 25% teratas usia di atas 61 tahun, glukosa di atas 32,80, dan BMI di atas level kategori 0. Secara keseluruhan, ringkasan ini membantu memahami rentang, pusat distribusi, dan variasi setiap fitur sebelum melanjutkan ke tahapan pemodelan.

B. Multivariate Analysis dengan heatmap mengungkap korelasi positif terkuat antara usia dan stroke (r=0,27) serta hipertensi dan stroke (r=0,14). Fitur smoking status dan heart disease memiliki korelasi rendah (<0,05), mengindikasikan pengaruh minor secara linier.



Gambar 1. Multivariate Analysis

Gambar 1 menampilkan multivariate analysis menggunakan matriks korelasi mengungkap hubungan linier antar fitur dalam dataset kesehatan pasien stroke. Koefisien korelasi "age" terhadap variabel target "stroke" sebesar 0,27 menunjukkan bahwa peningkatan usia berkaitan dengan risiko stroke secara moderat positif. Hipertensi dan stroke memiliki korelasi 0,14, menegaskan hipertensi sebagai faktor risiko penting. Heart disease berkorelasi 0,11 dengan stroke, menandakan pengaruh jantung bermasalah di tingkat lebih rendah. Kadar glukosa (avg_glucose_level) menunjukkan korelasi 0,13, sedangkan BMI hanya 0,08, memperlihatkan kadar gula lebih terkait daripada indeks massa tubuh. Status pernikahan, jenis kelamin, tipe pekerjaan, dan tempat tinggal menunjukkan korelasi mendekati nol terhadap stroke (<0,05), mengindikasikan peran mereka lebih kecil dalam model KNN. Antara fitur medis sendiri, "age" berkorelasi positif 0,54 dengan glukosa dan 0,38 dengan BMI, mencerminkan polifaktorial perubahan metabolik seiring usia. Korelasi rendah (<0,1) antara hipertensi dan heart_disease mengindikasikan keterkaitan penyakit kardiovaskular yang tidak dominan. Secara keseluruhan, pola korelasi ini mendukung pemilihan fitur utama (age, hypertension, avg_glucose_level) dalam memprediksi risiko stroke.

C. Setelah mean imputation untuk missing BMI (4,8% missing), winsorization IQR untuk outlier, label encoding, dan SMOTE oversampling untuk class imbalance (rasio asli ~1:20 naik menjadi ~1:1), model KNN (K=5) meraih akurasi 96%, precision 96%, recall 99%, F1-score 97% pada data pengujian

Tabel 3. Hasil Pengujian Akurasi

No	Pengujian *Confusion matrix*	*K-Nearest Neighbor (KNN)*
		Hasil
1	Akurasi (*Accuracy*)	96%
2	Presisi (*Precision*)	96%
3	Recall atau Sensitivitas (*Recall*)	99%
4	*F1-Score*	97%

Tabel 3 merangkum kinerja model K-Nearest Neighbor (KNN) pada tugas klasifikasi risiko stroke dengan empat metrik evaluasi utama. Pertama, akurasi model mencapai 96%, artinya 96% dari seluruh prediksi—baik yang mengidentifikasi pasien berisiko stroke maupun yang tidak—dilakukan dengan benar. Presisi yang juga sebesar 96% menunjukkan bahwa dari seluruh prediksi positif (berisiko stroke), 96% benar-benar merupakan kasus stroke,

menandakan tingkat kesalahan false positive yang rendah. Recall, atau sensitivitas, tertinggi di antara metrik lain yaitu 99%, mengindikasikan bahwa hampir seluruh pasien yang benar-benar mengalami stroke terdeteksi oleh model. Terakhir, F1-Score mencapai 97%, yang merupakan rata-rata harmonik dari presisi dan recall, mencerminkan keseimbangan tinggi antara kemampuan model menjelaskan prediksi positif dan menangkap seluruh kasus stroke. Secara keseluruhan, hasil ini menegaskan bahwa model KNN tidak hanya akurat secara keseluruhan tetapi juga sangat andal dalam mengidentifikasi penderita stroke dengan minim kesalahan.

Keterkaitan dengan Konsep Dasar

Hubungan positif antara usia dan kejadian stroke mencerminkan perubahan vaskular yang terjadi seiring bertambahnya tahun, di mana dinding arteri kehilangan kelenturannya dan lapisan endotelium mengalami kerusakan kumulatif akibat paparan stres oksidatif dan peradangan kronis; kondisi ini predisposisi terhadap aterosklerosis dan penurunan aliran darah serebral, sehingga menjadikan usia sebagai salah satu determinan utama risiko stroke. Temuan bahwa hipertensi berkorelasi erat dengan insiden stroke menguatkan pemahaman klinis bahwa tekanan darah tinggi memicu kerusakan dinding pembuluh, memfasilitasi terbentuknya plak, serta meningkatkan kemungkinan pecahnya pembuluh otak—semua faktor yang secara patofisiologis diakui sebagai penyebab utama stroke. Pada tahap pra-pemodelan, penerapan winsorization untuk mereduksi efek nilai ekstrem dan penggunaan SMOTE untuk menyeimbangkan proporsi kelas telah memperkuat prinsip kualitas data dalam kerangka CRISP-DM; langkah pembersihan (cleaning) menghilangkan distorsi pada distribusi fitur, sementara penyeimbangan (balancing) menjaga agar algoritma berbasis jarak seperti KNN tidak condong terhadap kelompok mayoritas, sehingga model dapat belajar secara adil dari keseluruhan sampel dan menghasilkan prediksi yang lebih andal.

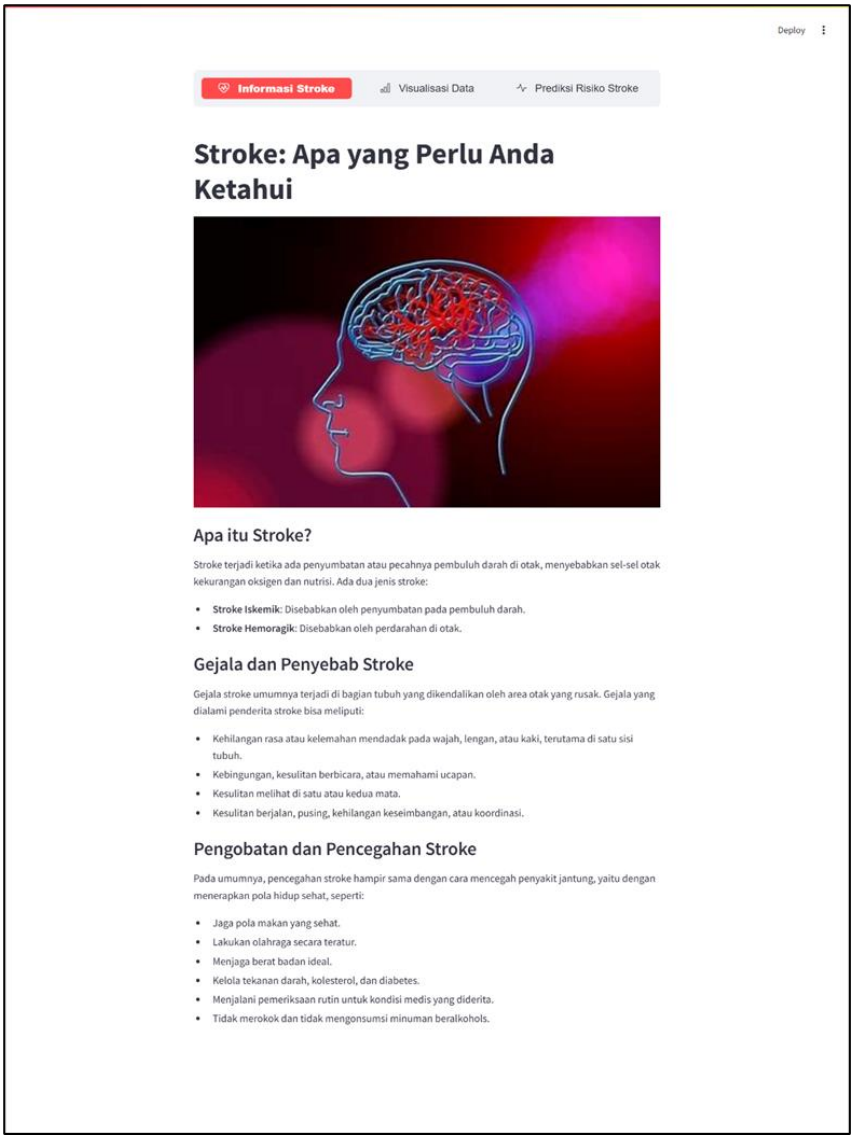
Interpretasi Hasil

Recall sebesar 99% menandakan bahwa model hampir tidak pernah melewatkan pasien yang benar-benar mengalami stroke, sehingga jumlah kasus “false negative” sangat kecil. Hal ini sangat penting dalam konteks klinis, di mana kegagalan mendeteksi stroke bisa berakibat fatal karena keterlambatan penanganan memengaruhi prognosis pasien. Sementara itu, precision mencapai 96% mengungkapkan bahwa dari semua prediksi positif yang dikeluarkan model, 96% di antaranya benar-benar pasien stroke, sehingga tingkat “false positive” relatif rendah. Kondisi ini mencegah overdiagnosis yang tidak hanya dapat menimbulkan biaya medis yang tidak perlu—seperti pemeriksaan lanjutan dan prosedur invasif—tetapi juga menambah

kecemasan pasien dan keluarganya. Pemilihan parameter K=5 pada algoritma KNN berhasil menyeimbangkan trade-off antara bias dan varians; nilai K yang moderat ini menghindarkan model dari kecenderungan overfitting yang muncul saat K=1—di mana akurasi pelatihan mencapai sempurna (100%) namun berisiko buruk dalam menggeneralisasi pada data baru. Dengan demikian, K=5 memberikan keseimbangan optimal sehingga model tetap tajam dalam mendeteksi pola stroke tanpa kehilangan kemampuan adaptasi terhadap variasi sampel yang belum pernah ditemui.

Hasil Perancangan Sistem

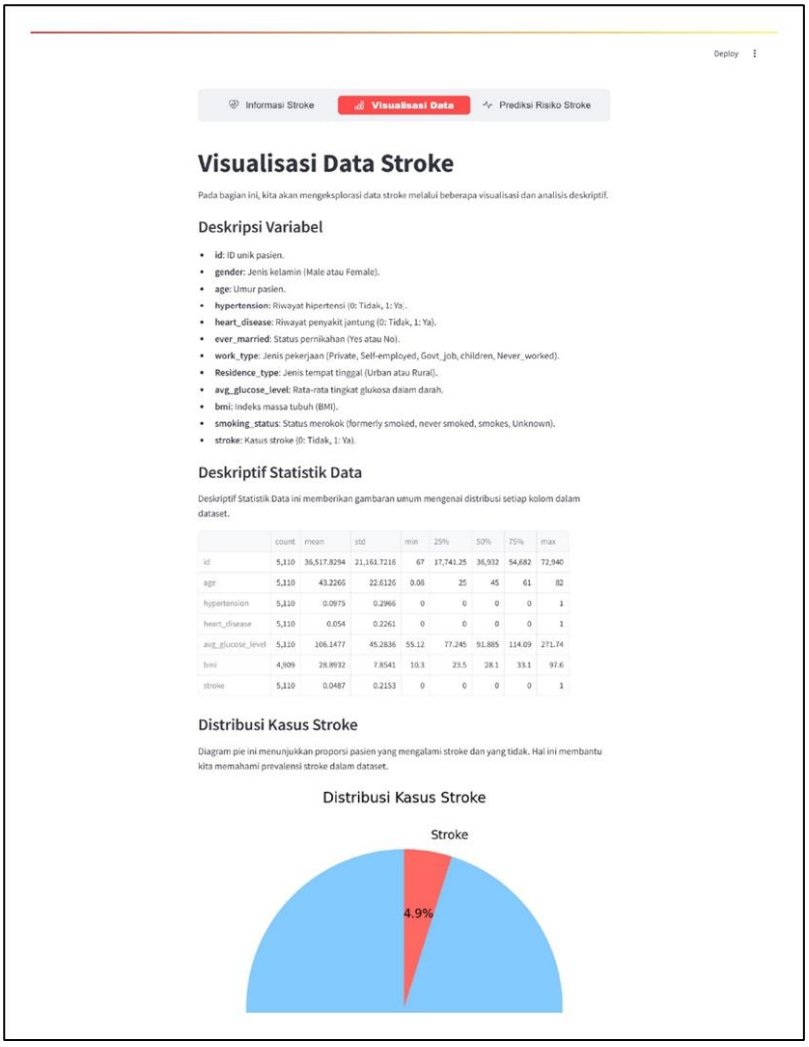
Pada tahap pengembangan website, sistem klasifikasi penyakit stroke dibangun dengan mengintegrasikan model K-Nearest Neighbor (KNN) ke dalam antarmuka web yang interaktif dan user-friendly. Arsitektur backend memuat model KNN hasil pelatihan, lengkap dengan modul preprocessing—seperti standarisasi dan penanganan kategori—sehingga input pengguna langsung diolah sebelum prediksi. Pengguna cukup mengisi formulir online berisi parameter risiko stroke—usia, jenis kelamin, riwayat hipertensi, riwayat penyakit jantung, status merokok, kadar glukosa rata-rata, dan indeks massa tubuh—yang kemudian dikirim ke server. Di server, data tersebut dinormalisasi sesuai skala model, lalu fungsi jarak Euclidean menghitung kedekatan sampel pengguna dengan tetangga terdekat dalam dataset pelatihan. Berdasarkan mayoritas label di antara k tetangga terdekat, sistem menghasilkan prediksi risiko stroke dan menyajikannya secara real time dalam bentuk teks dan grafik probabilitas. Desain antarmuka responsif menggunakan framework Streamlit memastikan tampilan form, hasil prediksi, dan penjelasan metrik evaluasi—seperti probabilitas, presisi, dan recall—terlihat jelas di perangkat desktop maupun mobile. Selain itu, mekanisme validasi input dan penanganan kesalahan dirancang untuk meminimalkan input tidak valid, sementara penyimpanan log prediksi memungkinkan analisis performa dan perbaikan berkelanjutan. Dengan demikian, website ini tidak hanya memberikan kemudahan akses bagi tenaga medis dan pasien, tetapi juga menjamin akurasi dan keandalan hasil klasifikasi stroke berdasarkan algoritma KNN yang telah dioptimalkan.



Gambar 2. Halaman Informasi Stroke

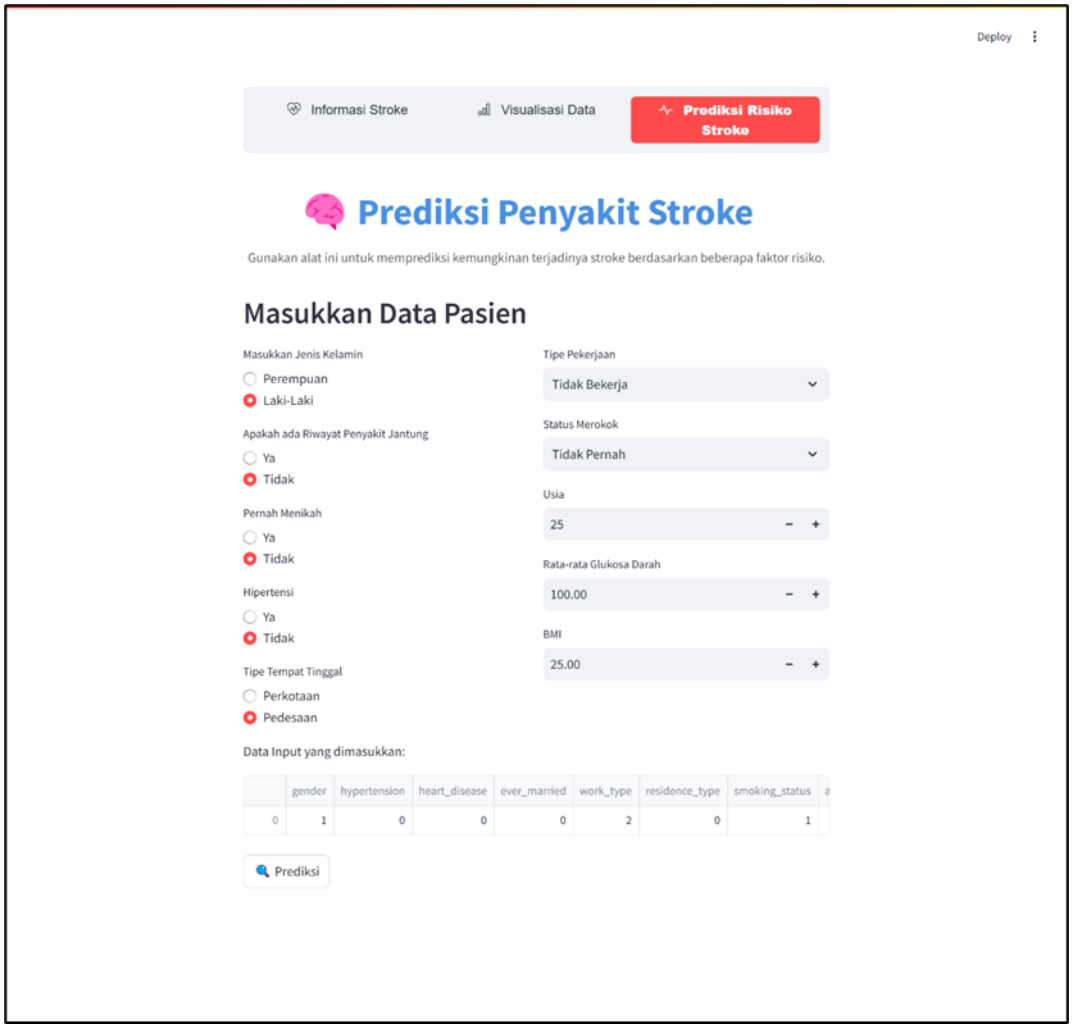
Gambar 2 menampilkan halaman “Informasi Stroke” pada aplikasi web klasifikasi risiko stroke. Di bagian atas terdapat bilah navigasi dengan tiga tab—Informasi Stroke (aktif), Visualisasi Data, dan Prediksi Risiko Stroke—serta tombol “Deploy” di pojok kanan atas. Judul utama halaman, “Stroke: Apa yang Perlu Anda Ketahui”, diikuti oleh ilustrasi otak dengan area yang disorot untuk menggambarkan lokasi gangguan vaskular. Di bawah gambar, terdapat tiga seksi konten: “Apa itu Stroke?” menjelaskan definisi dan membedakan antara stroke iskemik (penyumbatan) dan hemoragik (pendarahan); “Gejala dan Penyebab Stroke” merinci gejala klinis seperti kelemahan wajah, bicara sulit, dan gangguan koordinasi serta faktor pemicu; dan “Pengobatan dan Pencegahan Stroke” menyajikan rekomendasi gaya hidup sehat—termasuk pola makan seimbang, olahraga teratur, pengelolaan tekanan darah, dan pemeriksaan medis rutin—sebagai langkah pencegahan. Seluruh teks disajikan dalam format

paragraf dan daftar berpoin, memudahkan pembaca memahami informasi secara ringkas namun menyeluruh.



Gambar 3. Halaman Visualisasi Data

Gambar 3 menampilkan halaman “Visualisasi Data Stroke” menampilkan ringkasan eksplorasi data untuk memahami karakteristik dan distribusi kasus dalam dataset. Di bagian atas terlihat bilah navigasi dengan tab “Visualisasi Data” yang aktif, diikuti judul besar dan deskripsi singkat fungsi halaman. Selanjutnya, disajikan daftar variabel—mulai dari id pasien, jenis kelamin, usia, hipertensi, penyakit jantung, hingga status merokok dan stroke—dengan keterangan nilai encoding untuk tiap atribut. Di bawahnya, tabel statistik deskriptif memperlihatkan metrik numerik seperti jumlah sampel (n), rata-rata, standar deviasi, nilai minimum, kuartil, dan maksimum untuk setiap kolom fitur. Pada bagian akhir, diagram setengah lingkaran (pie chart) menggambarkan proporsi pasien yang mengalami stroke (sekitar 4,9%) versus yang tidak, sehingga pengguna dapat dengan cepat melihat prevalensi kondisi stroke dalam data sebelum melanjutkan ke analisis lebih mendalam atau pemodelan.



Gambar 4. Halaman Prediksi Risiko Stroke

Gambar 5 menampilkan halaman "Prediksi Risiko Stroke" pada aplikasi web klasifikasi. Di bagian atas terletak bilah navigasi dengan tab "Prediksi Risiko Stroke" yang aktif, diapit oleh "Informasi Stroke" dan "Visualisasi Data". Judul besar "Prediksi Penyakit Stroke" diikuti oleh instruksi singkat menjelaskan fungsi alat. Bagian utama terdiri dari dua kolom form input: kolom kiri berisi opsi pilihan radio untuk variabel kategorikal—jenis kelamin (Perempuan/Laki-Laki), riwayat penyakit jantung (Ya/Tidak), status pernikahan (Ya/Tidak), hipertensi (Ya/Tidak), dan tipe tempat tinggal (Perkotaan/Perdesaan)—sedangkan kolom kanan memuat dropdown untuk pekerjaan dan status merokok serta kontrol angka (+/–) untuk usia, rata-rata glukosa darah, dan BMI. Di bawah form, tabel ringkasan menampilkan nilai ter-encoding dari semua input sehingga pengguna dapat memverifikasi data sebelum prediksi. Tombol "🔍 Prediksi" di bagian bawah memicu proses klasifikasi yang menjalankan model KNN, menghasilkan probabilitas atau keputusan risiko stroke. Keseluruhan tata letak responsif ini memudahkan pengguna memasukkan data dengan cepat, memeriksa encoding, dan memperoleh hasil prediksi secara real time.

Implikasi Penelitian

Pengintegrasian model K-Nearest Neighbor ke dalam antarmuka web berbasis Streamlit menghadirkan solusi praktis bagi penyedia layanan primer—seperti klinik desa atau puskesmas—untuk melakukan skrining risiko stroke secara cepat, interaktif, dan dengan kebutuhan perangkat keras serta perangkat lunak yang sangat terbatas. Pada tataran metodologis, penelitian ini menekankan bahwa tahap preprocessing yang cermat merupakan fondasi krusial untuk klasifikasi berbasis jarak pada dataset medis yang karakteristiknya sering kali berimbalansi, sehingga teknik winsorization untuk mereduksi efek outlier dan SMOTE untuk menyeimbangkan distribusi kelas minoritas telah terbukti sebagai praktik terbaik. Untuk mendorong keandalan model lebih jauh sebelum implementasi pada skala klinis, disarankan dilakukannya penalaan hyperparameter secara sistematis—seperti pemilihan nilai k optimal atau jarak Minkowski yang disesuaikan—serta penerapan skema validasi silang (cross-validation) pada berbagai subset data. Langkah-langkah ini penting untuk memastikan kestabilan performa model, meminimalkan risiko overfitting, dan meningkatkan generalisasi pada populasi pasien yang beragam, sehingga integrasi akhir ke dalam alur kerja klinis dapat dilakukan dengan kepercayaan tinggi terhadap hasil prediksi.

5. KESIMPULAN DAN SARAN

Pengembangan sistem klasifikasi risiko stroke menggunakan algoritma K-Nearest Neighbor dan diintegrasikan ke dalam aplikasi web berbasis Streamlit telah mencapai target penelitian dengan kinerja evaluasi yang sangat baik: 96% akurasi, 96% presisi, 99% recall, dan 97% F1-score pada data pengujian. Keberhasilan ini dicapai setelah serangkaian langkah pra-pemrosesan komprehensif—pengisian nilai hilang menggunakan imputasi rata-rata, penanganan nilai ekstrem melalui winsorization, konversi variabel kategorikal ke format numerik dengan label encoding, dan penyeimbangan distribusi kelas menggunakan SMOTE—serta normalisasi semua fitur numerik dengan Min-Max Scaler. Pemilihan parameter K=5 pada KNN secara efektif meminimalkan bias sekaligus mengendalikan varians, sehingga model hampir selalu mendeteksi pasien stroke (recall 99%) tanpa terlalu banyak memprediksi positif palsu (presisi 96%). Analisis visual distribusi fitur dan peta korelasi multivariat menegaskan bahwa usia, hipertensi, dan kadar glukosa darah merupakan kontributor paling signifikan terhadap prediksi risiko stroke, sesuai dengan landasan klinis yang ada. Dengan antarmuka pengguna yang responsif dan kebutuhan infrastruktur yang ringan, sistem ini menawarkan alat skrining dini yang praktis dan terjangkau bagi fasilitas kesehatan primer, memfasilitasi deteksi awal dan intervensi yang lebih cepat.

Kendati demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, dataset Kaggle yang digunakan meski representatif, masih bersifat sekunder dan mungkin mengandung bias regional atau demografis yang tidak sepenuhnya mencerminkan populasi global. Kedua, SMOTE mampu menyeimbangkan kelas tetapi menambahkan data sintetis yang mungkin tidak menangkap kompleksitas klinis sepenuhnya. Ketiga, validasi eksternal terhadap data dari sumber lain atau uji coba lapangan belum dilakukan, sehingga keandalan model pada populasi berbeda masih perlu pembuktian lebih lanjut.

Untuk penelitian selanjutnya disarankan memperluas kumpulan data dengan menambahkan rekaman klinis dari berbagai wilayah atau rumah sakit untuk menguji generalisasi model. Implementasi teknik hyperparameter tuning yang lebih canggih—seperti grid search pada jarak Minkowski atau algoritma K variabel—dan penerapan skema cross-validation berlapis (nested cross-validation) akan memperkuat stabilitas performa. Selain itu, eksplorasi algoritma hybrid atau ensemble yang menggabungkan KNN dengan model lain (misalnya Random Forest atau XGBoost) dapat meningkatkan presisi dan mengurangi sensitivitas terhadap outlier. Integrasi data temporal (time-series) atau fitur citra medis (misalnya hasil CT scan) juga bisa menjadi arah pengembangan yang menjembatani gap antara deteksi klinis dan skrining awal berbasis machine learning.

DAFTAR REFERENSI

Akbar, I., Supriadi, F., & Junaedi, D. I. (2025). Pemanfaatan machine learning di bidang kesehatan. *JATI (Jurnal Mahasiswa Teknik Informatika), 9*(1), 1744–1749.

Aprilitaz, W., Akbar, R., & Prayogi, R. C. (2023, August). Komparasi algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam klasifikasi penyakit Parkinson: *Comparison of K-Nearest Neighbor (KNN) and Naive Bayes algorithms in the classification of Parkinson's disease*. In *Sentimas: Seminar Nasional Penelitian dan Pengabdian Masyarakat* (pp. 188–193).

Ardhani, N. T., Notodiputro, K. A., & Oktarina, S. D. (2025). Winsorization for outliers in clustering non-cyclical stocks with K-Means and K-Medoids: Winsorization untuk penanganan pencilan dalam penggerombolan saham sektor consumer non-cyclical dengan K-Means dan K-Medoids. *Indonesian Journal of Statistics and Its Applications, 9*(1), 46–60.

A'yuniyah, Q., & Reza, M. (2023). Penerapan algoritma K-Nearest Neighbor untuk klasifikasi jurusan siswa di SMA Negeri 15 Pekanbaru: *Application of the K-Nearest Neighbor algorithm for student department classification at 15 Pekanbaru State High School*. *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE), 3*(1), 39–45.

Bakri, S. N., & Harahap, L. S. (2025). Analisis klasifikasi algoritma K-Nearest Neighbor (KNN) pada struktur daerah di Kota Medan. *Jurnal Ilmu Komputer dan Sistem Informasi, 4*(2), 182–193.

Bhirawa, A. A., & Sanjaya, U. P. (2025). From data imbalance to precision: SMOTE-driven machine learning for early detection of kidney disease. *INOVTEK Polbeng—Seri Informatika, 10*(1), 514–525.

Bintang, Y. K., & Imaduddin, H. (2024). Pengembangan model deep learning untuk deteksi retinopati diabetik menggunakan metode transfer learning. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 9*(3), 1442–1455.

Brian, T., Sholikhah, E. N., Maulidhia, A. N. A., & Wibowo, S. (2025). Application of K-Nearest Neighbor (KNN) algorithm to predict drinking water quality. *Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer, 5*(1), 9–16.

Carudin, C., Marisa, M., Murnawan, M., Reba, F., Koibur, M. E., Thantawi, A. M., … & Wattimena, F. Y. (2024). *Buku ajar data mining*. PT Sonpedia Publishing Indonesia.

Desiani, A. (2022). Perbandingan implementasi algoritma Naïve Bayes dan K-Nearest Neighbor pada klasifikasi penyakit hati. *Jurnal Sistem Informasi dan Sistem Komputer, 7*(2), 104–110.

Elfaladonna, F., Isa, I. G. T., Sartika, D., & Putra, A. M. (2024). *Buku ajar dasar exploratory data analysis (EDA)*. Penerbit NEM.

Fasnuari, H. A. D., Yuana, H., & Chulkamdi, M. T. (2022). Penerapan algoritma K-Nearest Neighbor untuk klasifikasi penyakit diabetes melitus: Studi kasus warga Desa Jatitengah. *Antivirus: Jurnal Ilmiah Teknik Informatika, 16*(2), 133–142.

Febiola, A., Ardiningrum, F., Purba, M. O. A., & Siahaan, F. (2025). Implementation of SVM in predicting obesity risk based on lifestyle and dietary patterns. *JOMLAI: Journal of Machine Learning and Artificial Intelligence, 4*(1), 38–45.

Hendriyani, Y. (2025). *Data mining bab. Dasar-dasar data mining: Konsep, teknik dan aplikasi*.

Lonang, S., Yudhana, A., & Biddinika, M. K. (2023). Analisis komparatif kinerja algoritma machine learning untuk deteksi stunting. *Jurnal Media Informatika Budidarma, 7*(4), 2109.

Maulana, M. R., Nugroho, A. P., Adinata, F. C., Haidar, N. B., & Setiawan, A. (2025). KNN-based handwritten digit classification with accuracy analysis and visualization. *Jurnal Ilmiah Sistem Informasi, 4*(2), 142–150.

Maulana, M. R., Sucipto, A., & Mulyo, H. M. (2024). Optimisasi parameter support vector machine dengan particle swarm optimization untuk peningkatan klasifikasi diabetes. *Jurnal Informatika Teknologi dan Sains (Jinteks), 6*(4), 802–812.

Mutoffar, M. M., Retnoningsih, E., Yasik, I. Y. L., & Eliza, S. T. (2025). *Decoding intelligence: Algoritma machine learning dalam aksi dan bisnis*. PT Kimhsafi Alung Cipta.

Nikmawati, S., Warlem, N., & Izrul, I. (2024). Pemberian edukasi dalam rangka Hari Stroke Sedunia tahun 2024. *Jurnal Pengabdian Masyarakat Kesehatan (JURABDIKES), 2*(2), 49–54.

Nova, N., Mulyanti, A., Burhanie, C. S. A. P., Mulyani, L., Nurjanah, R. G., Utami, W., & Sukaesih, N. S. (2025). Systematic review: Pemanfaatan deep learning untuk diagnosis penyakit menggunakan MRI. *Jurnal Penelitian Inovatif, 5*(2), 839–852.

Nurlaela, L., Suhanda, Y., Sopian, A., Dewi, C. S., & Syahrial, R. (2025). Pengembangan framework data mining berbasis deep neural network dengan eksplorasi teknik transfer learning untuk prediksi dan klasifikasi data. *JRIS: Jurnal Rekayasa Informasi Swadharma, 5*(1), 132–141.

Pratama, Y. D., & Salam, A. (2025). Comparison of data normalization techniques on KNN classification performance for Pima Indians diabetes dataset. *Journal of Applied Informatics and Computing, 9*(3), 693–706.

Primajaya, A., Sari, B. N., & Padilah, T. N. (2022). *Diseminasi hasil penelitian research group software engineering, data science, computational intelligent and optimization, computer network and security, information system*. Uwais Inspirasi Indonesia.

Putra, R. F., Zebua, R. S. Y., Budiman, B., Rahayu, P. W., Bangsa, M. T. A., Zulfadhilah, M., … & Andiyan, A. (2023). *Data mining: Algoritma dan penerapannya*. PT Sonpedia Publishing Indonesia.

Sahelvi, E., Cikita, P., Sapitri, R. M., Rahmaddeni, R., & Efrizoni, L. (2025). Perbandingan algoritma K-Nearest Neighbors dan Random Forest untuk rekomendasi gaya hidup sehat dalam mencegah penyakit jantung: *Comparison of K-Nearest Neighbors and Random Forest algorithms for recommendations for a healthy lifestyle in preventing heart disease*. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5*(3), 830–840.

Santoso, L., & Priyadi, P. (2024). Mengoptimalkan proses pembersihan data dalam analisis big data menggunakan pipeline berbasis AI. *Elkom: Jurnal Elektronika dan Komputer, 17*(2), 657–666.

Sheila, S. (2024). Analisis performa algoritma C4.5 dan klasifikasi decision tree dalam memprediksi penyakit diabetes: *Performance analysis of C4.5 algorithm and decision-tree classification in predicting diabetes*. *Sistem dan Teknologi Informasi Indonesia (SINTESIA), 4*(1), 1–9.

Shinami, M. A., & Bahri, S. (2023). Klasifikasi penyakit kanker payudara menggunakan metode K-Nearest Neighbors (KNN). *Jurnal Fourier, 12*(2), 79–85.

Siregar, A. H., & Siregar, S. D. (2025). Comparison of logistic regression and support vector machine algorithm performance in heart failure prediction. *Academia Open, 10*(2), 10–21070.

Sulianta, F. (2023). *Basic data mining from A to Z*. Feri Sulianta.

Swastika, R., Mukodimah, S., Susanto, F., Muslihudin, M., & Adab, S. I. P. (2023). *Implementasi data mining (clustering, association, prediction, estimation, classification)*. Penerbit Adab.

Syam, S., Tokoro, Y., Judijanto, L., Garonga, M., Sinaga, F. M., Umar, N., … & Sitanggang, A. T. (2024). *Data mining: Teori dan penerapannya dalam berbagai bidang*. PT Sonpedia Publishing Indonesia.

Ujianto, N. T., Fadillah, H., Fanti, A. P., Saputra, A. D., & Ramadhan, I. G. (2025). Penerapan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi citra medis. *IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi*, *4*(1), 33–43.