eISSN : 2621-4970
pISSN : 2621-4962

Volume 9 | No 2
Juli | **2026**

Prediksi Risiko Penyakit Stroke Menggunakan Logistic Regression Dengan Dashboard Interaktif

Aldillah[1], Khairunnas[2], Irma Eryanti Putri[3]
Fakultas Teknik dan Ilmu Komputer, Program Studi Ilmu Komputer
Universitas Muhammadiyah Bima
Email : abdillah6juni@gmail.com

Abstrak

Stroke merupakan salah satu penyebab utama kematian dan kecacatan jangka panjang di dunia. Peningkatan kasus stroke mendorong pemanfaatan teknologi machine learning untuk membantu prediksi risiko penyakit secara cepat dan akurat. Penelitian sebelumnya telah menggunakan algoritma seperti Support Vector Machine dan Random Forest, namun sebagian besar masih berfokus pada performa model dan belum mengintegrasikan hasil prediksi dalam media visual yang interaktif. Penelitian ini bertujuan membangun model prediksi risiko stroke menggunakan algoritma Logistic Regression serta mengimplementasikannya dalam dashboard interaktif. Dataset yang digunakan berasal dari Kaggle dengan jumlah data sebanyak 5110. Tahapan penelitian meliputi preprocessing data, pembagian data latih 80% dan data uji 20%, serta pelatihan model menggunakan Logistic Regression. Hasil penelitian menunjukkan model memperoleh akurasi sebesar 92% dan ROC-AUC sebesar 84%. Dashboard interaktif yang dikembangkan mampu menampilkan probabilitas risiko stroke dan variabel yang berpengaruh sehingga membantu pengguna memahami hasil prediksi secara lebih efektif.
Kata kunci: Data Mining; Logistic Regression; Stroke

Abstract

Stroke is one of the leading causes of death and long-term disability worldwide. The rise in stroke cases has driven the use of machine learning technology to help predict disease risk quickly and accurately. Previous research has employed algorithms such as Support Vector Machines and Random Forests; however, most studies have focused primarily on model performance and have not integrated prediction results into interactive visual media. This study aims to build a stroke risk prediction model using the Logistic Regression algorithm and implement it in an interactive dashboard. The dataset used comes from Kaggle, containing 5,110 data points. The research stages include data preprocessing, splitting the data into an 80% training set and a 20% test set, and training the model using Logistic Regression. The results show that the model achieved an accuracy of 92% and an ROC-AUC of 84%. The developed interactive dashboard displays stroke risk probabilities and influencing variables, helping users understand the prediction results more effectively..
Keywords: ***Data Mining, Logistic Regression, Stroke***

PENDAHULUAN

Stroke merupakan salah satu penyebab kematian dan kecacatan jangka panjang tertinggi di dunia. Peningkatan jumlah penderita stroke setiap tahun menunjukkan bahwa penyakit ini masih menjadi masalah kesehatan global yang serius, termasuk di Indonesia (Annas et al., 2022). Selain berdampak pada kesehatan individu, stroke juga menimbulkan beban sosial dan ekonomi akibat tingginya biaya pengobatan, penurunan produktivitas, serta kebutuhan perawatan jangka panjang. Faktor-faktor seperti hipertensi, kadar

glukosa darah tinggi, penyakit jantung, obesitas, usia, dan kebiasaan merokok diketahui memiliki hubungan terhadap peningkatan risiko stroke (Guhdar et al., 2023).

Perkembangan teknologi informasi dan machine learning memberikan peluang dalam mendukung deteksi dini risiko stroke melalui pemanfaatan data medis dan demografis pasien. Berbagai penelitian sebelumnya telah menerapkan algoritma seperti *Random Forest, K-Nearest Neighbors, Neural Network,* dan *XGBoost* untuk memprediksi risiko stroke dengan tingkat akurasi yang cukup tinggi. Namun, sebagian besar penelitian masih berfokus pada peningkatan performa model tanpa memperhatikan aspek interpretabilitas hasil prediksi. Model yang kompleks cenderung sulit dipahami oleh tenaga medis maupun pengguna non-teknis sehingga kurang optimal untuk mendukung pengambilan keputusan klinis (Sitompul et al., 2025) dan (Sheng et al., 2024).

Selain itu, penelitian terdahulu umumnya hanya menampilkan hasil prediksi dalam bentuk nilai klasifikasi tanpa penyajian visual yang informatif. Padahal, informasi seperti persentase risiko dan faktor yang paling berpengaruh penting untuk membantu pengguna memahami hasil prediksi secara lebih mudah. Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma *Logistic Regression* untuk memprediksi risiko penyakit stroke menggunakan Stroke *Prediction* Dataset dari *Kaggle*. Algoritma Logistic Regression dipilih karena sesuai untuk permasalahan klasifikasi biner pada prediksi risiko stroke serta mampu menghasilkan probabilitas risiko yang mudah diinterpretasikan dalam konteks medis. Selain memiliki performa yang stabil pada dataset medis dengan jumlah fitur yang tidak terlalu besar, Logistic Regression juga memungkinkan interpretasi pengaruh masing-masing variabel terhadap hasil prediksi. Dibandingkan model yang lebih kompleks seperti Neural Network atau XGBoost, Logistic Regression lebih transparan sehingga lebih mudah digunakan sebagai pendukung pengambilan keputusan klinis dan evaluasi faktor risiko pasien.

Kebaruan penelitian ini terletak pada integrasi algoritma Logistic Regression yang bersifat interpretable dengan dashboard interaktif untuk mendukung visualisasi probabilitas risiko stroke dan pengambilan keputusan kesehatan berbasis data secara lebih transparan. Berbeda dengan penelitian terdahulu yang umumnya hanya berfokus pada peningkatan akurasi model klasifikasi, penelitian ini tidak hanya menghasilkan prediksi risiko stroke, tetapi juga menyajikan nilai probabilitas risiko serta visualisasi faktor-faktor yang berpengaruh terhadap hasil prediksi dalam sebuah dashboard interaktif. Pendekatan ini memungkinkan tenaga medis maupun pengguna non-teknis untuk memahami hasil prediksi secara lebih mudah, informatif, dan mendukung proses deteksi dini risiko stroke.

Rumusan Masalah

Berdasarkan fokus penelitian mengenai prediksi risiko stroke menggunakan algoritma *Logistic Regression,* maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan algoritma Logistic Regression untuk memprediksi risiko stroke pada dataset yang memiliki ketidakseimbangan kelas (imbalanced dataset)?

2. Bagaimana kemampuan Logistic Regression dalam menghasilkan prediksi risiko stroke yang akurat serta mudah diinterpretasikan dibandingkan model machine learning yang lebih kompleks?
3. Bagaimana mengembangkan dashboard interaktif yang dapat menyajikan hasil prediksi risiko stroke, persentase tingkat risiko, dan faktor-faktor yang paling berpengaruh sehingga lebih informatif dan mudah dipahami oleh tenaga medis maupun pengguna non-teknis?

Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan algoritma *Logistic Regression* untuk memprediksi risiko penyakit stroke menggunakan Dataset Stroke *Prediction* dari *Kaggle*.
2. Mengetahui tingkat akurasi algoritma *Logistic Regression* dalam memprediksi risiko penyakit stroke berdasarkan hasil pengujian model.
3. Menyajikan hasil prediksi risiko stroke dalam bentuk *dashboard interaktif* yang menampilkan persentase risiko stroke serta visualisasi variabel-variabel yang paling memengaruhi risiko stroke.

Manfaat Penelitian

Manfaat penelitian sebagai berikut:

1. Penelitian ini memberikan manfaat teoretis dalam pengembangan penerapan algoritma *Logistic Regression* untuk memprediksi risiko penyakit stroke menggunakan data klinis.
2. Penelitian ini juga memberikan pemahaman mengenai tingkat akurasi model *Logistic Regression* dalam memprediksi risiko stroke berdasarkan hasil pengujian model.
3. penelitian ini menambah dashboard interaktif untuk menyajikan hasil prediksi probabilitas risiko stroke dan prediksi yang berkelanjutan serta varibal yang berpengaruh.

Tinjauan Pustaka

Perkembangan *machine learning healthcare systems* memberikan peluang dalam mendukung deteksi dini penyakit melalui analisis data kesehatan pasien. Dalam bidang *healthcare predictive analytics,* data seperti usia, hipertensi, kadar glukosa, penyakit jantung, dan kebiasaan merokok banyak digunakan untuk memprediksi risiko stroke. Pemanfaatan teknologi ini membantu proses identifikasi pasien berisiko tinggi secara lebih cepat dan sistematis sehingga dapat mendukung pengambilan keputusan medis.

Penelitian oleh (Annas et al., 2022) menggunakan metode *Kernel Logistic Regression* untuk klasifikasi jenis stroke dan memperoleh akurasi sebesar 75,97%. Hasil penelitian menunjukkan bahwa pendekatan *regresi logistik* mampu mengidentifikasi variabel yang berpengaruh terhadap risiko stroke.

Penelitian lain oleh (Wang, 2023) menerapkan *Logistic Regression* dengan tahapan *preprocessing, oversampling*, dan penskalaan fitur sehingga menghasilkan akurasi lebih dari 95% dalam prediksi stroke.

Penelitian internasional juga menunjukkan bahwa model prediksi stroke berbasis *machine learning* memiliki performa yang baik dalam analisis risiko kesehatan. Penelitian oleh (Jo et al., 2023) menjelaskan bahwa model *machine*

learning yang bersifat interpretatif dapat membantu tenaga medis memahami hasil prediksi stroke secara lebih jelas. Selain itu, (Lolak et al., 2023) menyebutkan bahwa metode statistik seperti *Logistic Regression* tetap relevan karena mampu menghasilkan prediksi probabilistik yang lebih mudah dijelaskan dibandingkan model kompleks berbasis ensemble learning. Penelitian oleh (Zhang et al., 2023) juga menunjukkan bahwa penerapan kecerdasan buatan dalam prediksi stroke mampu meningkatkan efektivitas analisis data klinis. Sementara itu, (Jo et al., 2023) menjelaskan bahwa integrasi data kesehatan dengan sistem prediksi otomatis dapat mendukung proses pengambilan keputusan medis secara lebih efektif.

Berdasarkan penelitian terdahulu, sebagian besar penelitian masih berfokus pada performa algoritma dan akurasi klasifikasi. Oleh karena itu, penelitian ini mengintegrasikan *healthcare* data, klasifikasi menggunakan *Logistic Regression,* prediksi probabilistik, dan dashboard interaktif *decision support* untuk menampilkan persentase risiko stroke serta variabel yang paling berpengaruh sehingga hasil prediksi lebih informatif dan mudah dipahami oleh tenaga medis maupun pengguna non-teknis.

LANDASAN TEORI

Menurut (Zhang et al., 2023) healthcare data merupakan data medis dan demografis pasien yang digunaka untuk mendukung anDalam prediksi stroke, data seperti usia, hipertensi, penyakit jantung, kadar glukosa darah, indeks massa tubuh (BMI), dan kebiasaan merokok menjadi faktor penting karena memiliki hubungan terhadap peningkatan risiko stroke. Pemanfaatan healthcare data memungkinkan proses identifikasi risiko dilakukan lebih cepat dan sistematis sehingga dapat membantu deteksi dini penyakit strokealisis kesehatan berbasis machine learning.

Menurut (Wang, 2023) Dataset kesehatan umumnya memiliki karakteristik kompleks, seperti data tidak seimbang (imbalanced dataset), missing value, dan perbedaan skala antar variabel. Pada dataset stroke, jumlah data pasien non-stroke jauh lebih besar dibandingkan pasien stroke sehingga berpotensi menyebabkan model lebih dominan memprediksi kelas mayoritas dan mengurangi kemampuan deteksi pasien stroke

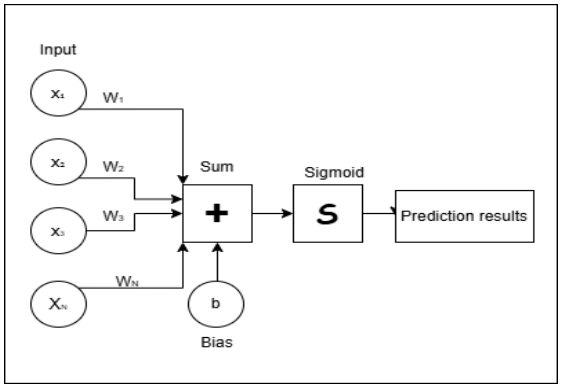
Menurut (Wang, 2023) Data preprocessing merupakan tahapan penting dalam machine learning untuk meningkatkan kualitas data sebelum proses pelatihan model dilakukan. Tahapan preprocessing meliputi penanganan missing value, transformasi data kategorikal, normalisasi data numerik, dan penyeimbangan distribusi kelas dataset.

Menurut (Lolak et al., 2023) Logistic Regression merupakan algoritma klasifikasi yang digunakan untuk memprediksi probabilitas suatu kejadian berdasarkan hubungan antara variabel independen dan variabel target. Algoritma ini banyak digunakan dalam bidang kesehatan karena memiliki interpretabilitas yang baik serta mampu menghasilkan prediksi probabilistik yang mudah dipahami oleh tenaga medis.

Menurut (Jo et al., 2023) Dalam bidang kesehatan, probabilistic prediction penting untuk mendukung proses deteksi dini dan pengambilan keputusan medis. Semakin tinggi probabilitas yang dihasilkan model, maka semakin tinggi pula tingkat risiko pasien terhadap stroke

sehingga dapat menjadi dasar pertimbangan tindakan preventif maupun pemeriksaan lanjutan.

Menurut (Rolansa et al., 2021) Dashboard interaktif merupakan media visualisasi data yang dirancang untuk mempermudah pengguna dalam memahami informasi secara cepat dan efisien melalui tampilan grafik, tabel, maupun indikator tertentu secara dinamis. Dalam penelitian kesehatan, Dashboard interaktif pada penelitian ini dirancang sebagai alat bantu prediksi lanjutan bagi tenaga medis dalam mengidentifikasi risiko stroke pasien menggunakan metode *Logistic Regression*. Dashboard ini mampu menampilkan persentase tingkat risiko stroke berdasarkan data pasien yang telah diinput, sekaligus menunjukkan variabel yang paling berpengaruh terhadap hasil prediksi. Pendekatan tersebut mendukung *decision support system* dalam bidang kesehatan karena hasil prediksi tidak hanya menampilkan klasifikasi, tetapi juga informasi probabilitas risiko yang lebih mudah dipahami oleh pengguna non-teknis. Berikut Gambar 1 alur algoritma *logistic regression*:



Gambar 1. Alur Algoritma *Logistic Regression* (Antor et al., 2021)

a. Input (X_1, X_2, X_3, …, X_n)
Bagian ini merupakan data masukan atau variabel yang digunakan dalam prediksi. Pada penelitian prediksi stroke, input dapat berupa usia, hipertensi, kadar glukosa, riwayat penyakit jantung, dan variabel kesehatan lainnya (Antor et al., 2021).

b. Weight (W_1, W_2, W_3, …, W_n)
Setiap variabel input memiliki bobot (*weight*) yang menunjukkan tingkat pengaruh variabel tersebut terhadap hasil prediksi. Semakin besar nilai bobot, maka semakin besar kontribusinya terhadap kemungkinan risiko stroke (Antor et al., 2021).

c. Bias (b)
Bias digunakan sebagai nilai tambahan untuk membantu model menyesuaikan hasil prediksi agar lebih akurat. *Bias* berfungsi mengoreksi hasil perhitungan sebelum masuk ke fungsi aktivasi (Antor et al., 2021).

d. Sum (Penjumlahan)
Pada tahap ini seluruh input dikalikan dengan bobot masing-masing, kemudian dijumlahkan bersama nilai *bias*. Hasil perhitungan ini menjadi nilai awal sebelum diproses lebih lanjut (Antor et al., 2021).

e. Sigmoid Function (S)
Fungsi *sigmoid* mengubah hasil penjumlahan menjadi nilai probabilitas antara 0 hingga 1. Nilai ini menunjukkan tingkat kemungkinan suatu data termasuk ke dalam kelas tertentu, misalnya risiko stroke tinggi atau rendah (Antor et al., 2021).

f. Prediction Results
Tahap terakhir adalah menghasilkan prediksi berdasarkan nilai probabilitas dari fungsi *sigmoid*. Jika probabilitas mendekati 1 maka risiko stroke dinilai tinggi, sedangkan jika mendekati 0 maka risiko stroke dinilai rendah (Antor et al., 2021).

Berdasarkan teori yang telah dijelaskan, prediksi risiko stroke memerlukan pemanfaatan healthcare data yang terdiri atas berbagai faktor kesehatan dan demografis pasien. Karakteristik data kesehatan yang cenderung kompleks, memiliki missing value, perbedaan skala data, serta ketidakseimbangan kelas menyebabkan proses preprocessing menjadi tahap penting sebelum pemodelan dilakukan. Setelah data dipersiapkan dengan baik, algoritma Logistic Regression digunakan untuk menghasilkan prediksi probabilitas risiko stroke yang mudah diinterpretasikan dan sesuai untuk mendukung pengambilan keputusan dalam bidang kesehatan. Hasil prediksi probabilistik tersebut kemudian diintegrasikan ke dalam dashboard interaktif sehingga informasi tingkat risiko stroke dan faktor-faktor yang memengaruhinya dapat disajikan secara lebih informatif, terstruktur, dan mudah dipahami oleh tenaga medis maupun pengguna non-teknis. Dengan demikian, keterkaitan antara healthcare data, preprocessing, Logistic Regression, probabilistic prediction, dan dashboard interaktif membentuk suatu sistem pendukung keputusan yang dapat membantu proses deteksi dini risiko stroke secara lebih efektif.

METODE PENELITIAN

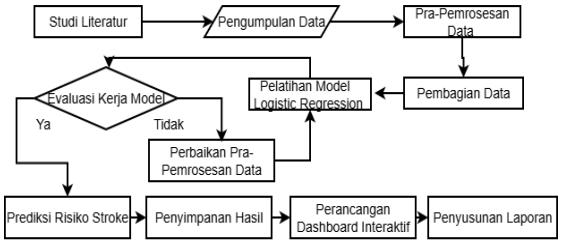
Penelitian ini menggunakan pendekatan kuantitatif melalui eksperimen untuk membangun model prediksi risiko stroke menggunakan algoritma *Logistic Regression.* Data yang digunakan berasal dari Stroke *Prediction* Dataset pada platform *Kaggle* yang mencakup variabel demografis, kondisi kesehatan, dan gaya hidup, seperti usia, jenis kelamin, hipertensi, penyakit jantung, kadar glukosa, BMI, dan status merokok. Dataset memiliki ketidakseimbangan kelas (*imbalanced dataset*), dimana jumlah data non-stroke lebih banyak dibandingkan data stroke sehingga berpotensi menimbulkan kesalahan klasifikasi seperti *false negative* dan *false positive*. Dalam konteks medis, *false negative* menjadi perhatian penting karena dapat menyebabkan pasien stroke tidak terdeteksi lebih awal, sedangkan *false positive* dapat menyebabkan tindakan pemeriksaan lanjutan yang tidak diperlukan (Mohanty & Acharya, 2023), (Akinwumi et al., 2025), (Antor et al., 2021). Oleh karena itu, penelitian ini mempertimbangkan penggunaan *class weighting* pada *Logistic Regression* untuk membantu model mengenali kelas minoritas serta mengevaluasi performa model menggunakan *accuracy*, *precision, recall*, *F1-score,* dan *ROC-AUC*. Pada tahap preprocessing dilakukan penanganan missing value pada variabel BMI menggunakan imputasi nilai rata-rata, normalisasi fitur numerik menggunakan *StandardScaler,* transformasi data kategorikal menggunakan *One-Hot Encoding,* serta identifikasi *outlier* menggunakan metode *Interquartile Range* (IQR) untuk meningkatkan kualitas data sebelum proses pelatihan model. Alat dan bahan yang digunakan pada penelitian ini.

1) Python 3.11, sebagai bahasa pemrograman utama yang menyediakan fleksibilitas dalam pengolahan data dan penerapan algoritma *machine learning*.
2) *Jupyter Notebook*, sebagai lingkungan pengolahan data, visualisasi, serta dokumentasi hasil secara terstruktur dalam satu platform.

3) *Visual Studio Code*, sebagai pembangunan *dashboard interaktif*.

Alur Penelitian

Tahapan penelitian meliputi pra-pemrosesan data, pembagian data menjadi data latih dan data uji, pelatihan model, serta evaluasi kinerja model. Hasil dari model kemudian diimplementasikan dalam bentuk dashboard interaktif untuk menyajikan informasi prediksi risiko stroke agar mudah dipahami. Alur penelitian yang digunakan ditunjukkan pada Gambar 2.



Gambar 2. Alur Penelitian

Berdasarkan Gambar 2. Alur penelitian dapat dijelaskan di bawah ini:

a. Studi Literatur

Studi literatur dilakukan sebagai tahap awal untuk memperoleh landasan teoritis yang relevan dengan penelitian. Tinjauan terhadap penelitian terdahulu digunakan untuk memahami perkembangan metode, mengidentifikasi celah penelitian, serta menentukan pendekatan yang sesuai dalam membangun model prediksi. Selain itu, studi ini juga menjadi dasar dalam pemilihan teknik analisis dan evaluasi agar penelitian yang dilakukan memiliki arah yang jelas dan terstruktur (Ebidor & Ikhide, 2024).

b. Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari platform Kaggle berupa Dataset prediksi stroke yang tersedia secara terbuka bisa di akses melalui tautan https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Dataset tersebut dipilih karena memiliki struktur data yang jelas serta telah banyak dimanfaatkan dalam penelitian terkait prediksi risiko stroke, sehingga dinilai relevan dan dapat mendukung kebutuhan analisis.

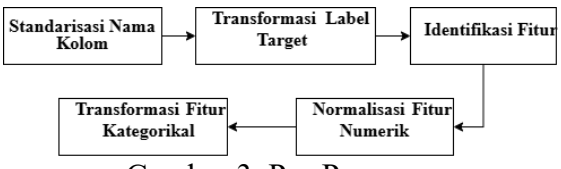
Data diunduh dalam format CSV dan diolah menggunakan lingkungan pemrograman Python melalui *Jupyter Notebook* untuk keperluan eksplorasi awal dan persiapan data. Pada tahap ini dilakukan penyesuaian nama variabel dan kategori ke dalam Bahasa Indonesia guna meningkatkan keterbacaan dan konsistensi penulisan, tanpa mengubah informasi yang terkandung dalam data.

Dataset memuat karakteristik individu yang mencakup aspek demografis, kondisi kesehatan, dan perilaku, yang digunakan sebagai variabel prediktor maupun target

c. Pra-Pemrosesan Data

Pra-pemrosesan data adalah tahap yang paling penting bertujuan untuk menyiapkan data supaya layak digunakan dalam proses pemodelan *Logistic Regression.* Tahapan ini dilakukan untuk memastikan data memiliki format yang konsisten, bersih, dan dapat diproses oleh algoritma machine learning (Tuppad & Patil, 2023).

Tahapan pra-pemrosesan data pada penelitian ini meliputi beberapa langkah dapat dilihat di gambar 3 dibawah sebagai berikut:



Gambar 3. Pra-Prosesan

1. Standarisasi Nama Kolom

Nama kolom pada dataset distandarisasi agar konsisten dan mudah diproses oleh sistem. Proses ini dilakukan dengan menghilangkan spasi, tanda baca, serta karakter khusus yang tidak diperlukan. Standarisasi nama kolom bertujuan untuk menghindari kesalahan pemanggilan variabel serta mempermudah proses analisis dan pemodelan data (Varma et al., 2023).

2. Transformasi Label Target
   Label target yang semula berbentuk teks, seperti yes dan no, diubah menjadi nilai numerik, yaitu 1 untuk kondisi stroke dan 0 untuk kondisi tidak stroke. Transformasi ini diperlukan karena algoritma *Logistic Regression* hanya dapat memproses data dalam bentuk numerik (Varma et al., 2023).
3. Identifikasi Fitur
   Pada tahap ini, seluruh variabel input diidentifikasi dan dikelompokkan berdasarkan jenis datanya, yaitu: Fitur numerik, variabel yang memiliki nilai berupa angka, seperti usia, kadar glukosa, dan indeks massa tubuh (BMI). Fitur kategorikal, yaitu variabel yang memiliki nilai berupa kategori atau teks, seperti jenis kelamin, status hipertensi, dan riwayat penyakit jantung. Pengelompokan ini dilakukan untuk menentukan metode transformasi yang sesuai bagi masing-masing jenis fitur (Pramakrisna et al., 2022).
4. Normalisasi Fitur Numerik
   Fitur numerik dinormalisasi menggunakan metode StandardScaler agar seluruh variabel memiliki skala yang seragam. Normalisasi bertujuan untuk mencegah dominasi fitur dengan skala besar terhadap fitur lain, serta membantu proses pelatihan model agar lebih stabil dan cepat mencapai (Putri, 2025).
5. Transformasi Fitur Kategorikal
   Fitur kategorikal diubah menjadi bentuk numerik menggunakan metode *One-Hot Encoding*. Metode ini merepresentasikan setiap kategori sebagai variabel biner, sehingga data kategorikal dapat diproses secara efektif oleh algoritma *Logistic Regression* tanpa menimbulkan asumsi urutan antar kategori (Guhdar et al., 2023).

d. Pembagian Data

Dataset yang telah dikonversi ke dalam bentuk vektor numerik selanjutnya dibagi menjadi dua kelompok data, yaitu data latih (*training set*) dan data uji (*testing set*). Data latih digunakan pada proses pembentukan model, sedangkan data uji dimanfaatkan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah digunakan sebelumnya. Proses pembagian data tersebut ditunjukkan pada Tabel 1 (Go et al., 2026).

Tabel 1. Pembagian Dataset

. Keterangan	Total data	Persentase	Data Stroke (80:20)
Total Dataset	5.110	100%	5.110
Data Latih (training set)	5.110	80%	4.088
Data Uji (testing uji)	5.110	20%	1.022

Pada penelitian ini, dataset dibagi menggunakan rasio 80% data latih dan 20% data uji. Pembagian tersebut dilakukan agar model memiliki jumlah data yang cukup dalam proses pelatihan sekaligus menyediakan data terpisah untuk tahap evaluasi. Data latih digunakan oleh algoritma *Logistic Regression* untuk mempelajari pola hubungan antarvariabel yang memengaruhi risiko stroke,

sedangkan data uji dimanfaatkan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data baru yang belum pernah diproses sebelumnya. Dengan komposisi pembagian tersebut, proses evaluasi model dapat dilakukan secara lebih optimal dan objektif (Go et al., 2026).

e. Pelatihan Model Logistic Regression

Pelatihan algoritma *logistic regression* ini berupa jenis regresi yang analisisnya menjelaskan antara hubungan variabel independen dan dependen. Hubungan ini menjelaskan antara satu atau lebih dari variabel bebas dan variabel yang terikat dari jenis kategori tertentu, dengan nilai yang berkisar antara 0 dan 1 atau benar dan salah, besar atau kecil (Pramakrisna et al., 2022).

$$ln(p/(1-p)) = B_0 + B_1X \quad (1)$$

$$p = \frac{e^{(B_0+B_1X)}}{1+e^{(B_0+B_1X)}} \quad (2)$$

B0 : konstanta.

B1 : Koefisien dari masing-masing nilai variabel.

p : Peluang (*Y=1*).

f. Evaluasi

Dalam penelitian ini, dilakukan evaluasi untuk mengukur kinerja algoritma *Logistic regression* dalam memprediksi risiko stroke. Untuk menampilkan prediksi yang salah dan benar *True Negative* (TN), *True Positive* (TP), *False Negative* (FN) dan *False Positive* (FP) yang menggunakan *Confusion Matrix* (Zul et al., 2025).

Tabel 2.*Confusion Metrix*

Actual		*Predicted*
	Positif	Negatif
Negatif	FP	TN
positif	TP	FN

Keterangan :

True Positive (TP): Kasus positif yang diprediksi benar sebagai positif.

True Negative (TN): Kasus negatif yang diprediksi benar sebagai negatif.

False Positive (FP): Kasus negatif yang salah diprediksi sebagai positif.

False Negative (FN): Kasus positif yang salah diprediksi sebagai negatif.

Prediksi positif akurasi model diukur dengan rasio prediksi positif asli (TP) terhadap prediksi positif total (FP+TP). Precision digunakan untuk resiko penyakit stroke dalam memastikan prediksi positif benar-benar akurat, mengurangi prediksi yang salah antara positif atau negatif, dan peningkatan kinerja model (Zul et al., 2025).

$$Precision = \frac{Tp}{Tp+Fp} \quad (3)$$

Recall mengukur sejauh mana model mampu mengidentifikasi seluruh kasus positif dengan menghitung perbandingan true positif dalam keseluruhan jumlah kasus positif. Metrik ini sangat penting untuk mengurangi false negatif agar semua kasus positif dapat terdeteksi dengan baik (Zul et al., 2025)

$$\text{Recall} = \frac{Tp}{Tp+Fn} \quad (4)$$

Evaluasi metrik menggunakan *F1-Score* rata-rata untuk menghitung *precision* dan *recall* untuk keseimbanan dalam memberikan hasil akurasi antara prediksi dan kemampuan dalam menilai positif mendeteksi semua kasus positif. *F1-Score* ideal untuk dataset dengan ketidakseimbangan kelas, dengan nilai berkisar antara 0 hingga 1; semakin tinggi nilainya, semakin baik keseimbangan model dalam menangani prediksi positif (Zul et al., 2025).

$$F1 - Score = 2\frac{Precision \cdot Recall}{Precision+Recall} \quad (5)$$

Evaluasi metrik adalah *accuracy* untuk menunjukan persentase prediksi yang tepat bagi seluruh hasil dibuat oleh model. Indekasi metrik ini umum memberikan tentang model seberapa sering

menghasilkan hasil akurat, untuk lebih baik kelas negatif atau positif (Zul et al., 2025).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Pada tahap ini, model *Logistic Regression* yang telah dilatih digunakan untuk menghasilkan prediksi risiko penyakit stroke pada data individu. Prediksi dilakukan dalam bentuk probabilitas, yang merepresentasikan tingkat kemungkinan seseorang mengalami stroke. Nilai probabilitas tersebut kemudian dikonversi ke dalam bentuk persentase risiko (1–100%) agar lebih mudah diinterpretasikan (Penyakit & Menggunakan, 2023).

g. Dashboard Interaktif

Pada tahap akhir penelitian, model yang telah dibangun diimplementasikan ke dalam sebuah dashboard interaktif sebagai media visualisasi dan sistem pendukung keputusan. Dashboard ini dirancang untuk mempermudah pengguna dalam melakukan prediksi risiko stroke berdasarkan data individu secara langsung.

Dashboard menyediakan beberapa kelompok input yang mencakup data demografis, data klinis, riwayat penyakit, dan gaya hidup. Variabel yang digunakan meliputi usia, jenis kelamin, status pernikahan, jenis pekerjaan, tempat tinggal, kadar glukosa, indeks massa tubuh (BMI), riwayat hipertensi, penyakit jantung, serta kebiasaan merokok. Data yang dimasukkan pengguna kemudian diproses menggunakan model *Logistic Regression* yang telah dilatih sebelumnya.

Hasil prediksi ditampilkan dalam bentuk nilai probabilitas yang menunjukkan tingkat risiko stroke, disertai dengan kategori risiko untuk memudahkan interpretasi. Selain itu, dashboard juga menampilkan informasi tambahan berupa ringkasan hasil dan faktor-faktor yang berkontribusi terhadap prediksi. Penyajian ini bertujuan agar hasil analisis dapat dipahami dengan lebih mudah, terutama oleh pengguna non-teknis.

IMPLEMENTASI DAN PEMBAHASAN

Data penelitian dibagi menjadi 80% data pelatihan sebanyak 4.088 data dan 20% data pengujian sebanyak 1.022 data. Dataset terdiri dari 4.861 data non-stroke dan 249 data stroke dengan 12 variabel prediktor yang meliputi fitur numerik dan kategorikal, serta 1 variabel target yaitu stroke seperti ditunjukkan pada Tabel 3. Ketidakseimbangan jumlah data antar kelas menyebabkan model lebih dominan mempelajari pola pada kelas non-stroke dibandingkan kelas stroke.

Tabel 3. Variabel Dataset

No	Atribut	Deskripsi	Keterangan
1	*Id*	Identitas unik setiap individu	Tidak digunakan dalam pemodelan
2	*Gender*	Jenis kelamin	0 = Perempuan, 1 = Laki-laki
3	*Age*	Usia individu (tahun)	Numerik
4	*Hypertension*	Riwayat hipertensi	0 = Tidak, 1 = Ya
5	*Heart_disease*	Riwayat penyakit jantung	0 = Tidak, 1 = Ya
6	*Ever_married*	Status pernikahan	0 = Belum, 1 = Sudah
7	*Work_type*	Jenis pekerjaan	Kategori (Private, Self-employed, dll.)
8	*Residence_type*	Tipe tempat tinggal	0 = Pedesaan, 1 = Perkotaan
9	*Avg_glucose_level*	Rata-rata kadar	Numerik (mg/dL)

		glukosa darah	
10	BMI	Indeks massa tubuh	Numerik
11	*Smoking_status*	Status merokok	Kategori (Tidak, Pernah, Aktif)
12	*Stroke*	Stroke dan Tidak	Data target stroke

Tabel 4. menunjukkan hasil evaluasi model Logistic Regression dengan nilai accuracy sebesar 92,27% dan ROC-AUC sebesar 84,44%. Nilai accuracy yang tinggi menunjukkan bahwa model mampu melakukan klasifikasi dengan baik secara umum, sedangkan nilai ROC-AUC menunjukkan bahwa model cukup baik dalam membedakan kelas stroke dan non-stroke. Tingginya accuracy dipengaruhi oleh dominasi data non-stroke sehingga model lebih mudah mengenali pola kelas mayoritas. Selain itu, Logistic Regression bekerja optimal pada pola data yang linear dan terstruktur, namun memiliki keterbatasan dalam mengenali pola yang lebih kompleks pada data stroke yang tidak seimbang.

Tabel 4. Hasil Akurasi

Accuracy	92,27
ROC-AUC	84,44

Berdasarkan hasil confusion matrix, model memperoleh performa yang sangat baik pada kelas non-stroke dengan precision sebesar 97%, recall sebesar 95%, dan F1-score sebesar 96%. Namun pada kelas stroke, precision hanya mencapai 30%, recall sebesar 42%, dan F1-score sebesar 35%. Hasil tersebut menunjukkan bahwa kemampuan model dalam mendeteksi pasien yang benar-benar mengalami stroke masih perlu ditingkatkan.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       972
           1       0.30      0.42      0.35        50

    accuracy                           0.92      1022
   macro avg       0.63      0.68      0.65      1022
weighted avg       0.94      0.92      0.93      1022
```

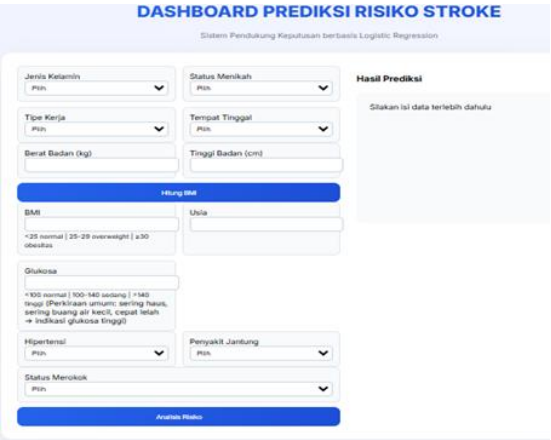
Gambar 4. Hasil Pengujian Model

Masih terdapat kesalahan klasifikasi berupa false positive dan false negative. Dalam konteks medis, false negative merupakan kondisi yang lebih kritis karena pasien yang sebenarnya berisiko stroke dapat tidak terdeteksi sehingga berpotensi menyebabkan keterlambatan penanganan. Oleh karena itu, evaluasi model tidak cukup hanya menggunakan accuracy, tetapi juga perlu mempertimbangkan recall, F1-score, dan ROC-AUC agar kemampuan deteksi risiko stroke dapat dinilai secara lebih komprnsif.

Dibandingkan dengan metode ensemble yang lebih kompleks seperti Random Forest, XGBoost, atau Gradient Boosting, Logistic Regression memiliki keunggulan pada aspek interpretabilitas karena hubungan antara variabel prediktor dan hasil prediksi dapat dijelaskan secara lebih mudah. Karakteristik ini penting dalam bidang kesehatan karena tenaga medis membutuhkan model yang transparan dan mudah dipahami untuk mendukung pengambilan keputusan. Meskipun demikian, performa prediksi Logistic Regression dapat sedikit lebih rendah dibandingkan metode ensemble yang mampu menangkap hubungan non-linear dan pola data yang lebih kompleks.

Model Logistic Regression kemudian diimplementasikan ke dalam dashboard interaktif untuk membantu pengguna melakukan prediksi risiko stroke berdasarkan data input seperti usia, hipertensi, penyakit jantung, kadar glukosa, BMI, dan status merokok. Dashboard

mampu menjalankan proses input data, prediksi otomatis, dan visualisasi probabilitas risiko stroke dalam bentuk persentase dan kategori risiko. Implementasi dashboard menunjukkan bahwa sistem dapat digunakan sebagai media pendukung keputusan awal dan edukasi risiko stroke secara dini. Namun, hasil prediksi tetap bergantung pada performa model sehingga keterbatasan seperti false negative masih dapat terjadi. Oleh karena itu, dashboard ini digunakan sebagai alat bantu prediksi awal dan bukan sebagai pengganti diagnosis medis secara langsung. Dashboard interaktif dapat diakses melalui https://warm-caramel-9c35f9.netlify.app/?utm_source=chatgpt.com



Gambar 5. Hasil Dasboard Interaktif

KESIMPULAN

Berdasarkan hasil penelitian, algoritma *Logistic Regression* mampu digunakan untuk prediksi risiko stroke dengan memperoleh nilai *accuracy* sebesar 92% dan *ROC-AUC* sebesar 84%, yang menunjukkan kemampuan model dalam membedakan kelas stroke dan non-stroke secara cukup efektif. Penelitian ini berkontribusi pada penerapan *interpretable machine learning* dan dashboard interaktif sebagai sistem pendukung keputusan untuk membantu visualisasi dan pemahaman risiko stroke secara lebih informatif. Namun, penelitian ini masih memiliki keterbatasan pada ketidakseimbangan dataset yang memengaruhi kemampuan model dalam mendeteksi kasus stroke, sehingga masih ditemukan *false positive* dan *false negative* pada hasil klasifikasi. Selain itu, penelitian ini belum menerapkan metode penanganan *imbalance* dataset seperti *SMOTE*, *undersampling*, maupun *class weighting*.

SARAN

Penelitian selanjutnya disarankan menerapkan teknik penanganan *imbalance* dataset seperti *SMOTE*, *undersampling*, atau *class weighting* untuk meningkatkan kemampuan model dalam mendeteksi kasus stroke pada kelas minoritas. Selain itu, penelitian berikutnya dapat membandingkan *Logistic Regression* dengan algoritma *machine learning* lainnya seperti *Random Forest, XGBoost*, maupun *deep learning* untuk mengevaluasi performa dan generalisasi model secara lebih komprehensif. Penggunaan *cross-validation, feature engineering*, serta evaluasi *usability* dashboard juga direkomendasikan untuk meningkatkan *robustness* model dan efektivitas sistem pendukung keputusan kesehatan berbasis *machine learning*.

DAFTAR PUSTAKA

Akinwumi, P. O., Ojo, S., Nathaniel, T. I., Wanliss, J., Karunwi, O., & Sulaiman, M. (2025). *Evaluating machine learning models for stroke prediction based on clinical variables*. Frontiers in Neurology, 16(September). *https://doi.org/10.3389/fneur.2025.1668420*

Annas, S., Poerwanto, B., Aswi, Abdy, M., & Fa'rifah, R. Y. (2022). *Classification Model for Type of*

Stroke Using Kernel Logistic Regression. Communications in Mathematical Biology and Neuroscience, *2022*, 1–17. https://doi.org/10.28919/cmbn/7752

Antor, M. B., Jamil, A. H. M. S., Mamtaz, M., Khan, M. M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). *A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer ' s Disease*. *2021*. *https://doi.org/10.1155/2021/9917919*

Ebidor, L., & Ikhide, I. G. (2024). *East African Journal of Education Studies Literature Review in Scientific Research : An Overview*. 7(2), 211–218. *https://doi.org/10.37284/eajes.7.2.1909.MLA*

Go, R. Y., Sarah, H., Gaspersz, D., & Kusumawardhana, R. H. (2026). *Analisis Sentimen Publik Program Makan Bergizi Gratis (MBG) di Youtube : Perbandingan Kinerja Algoritma Support Vector Machine (SVM) dan Random Forest*. *7*(1), 24–31.

Guhdar, M., Ismail Melhum, A., & Luqman Ibrahim, A. (2023). O*ptimizing Accuracy of Stroke Prediction Using Logistic Regression*. Journal of Technology and Informatics (JoTI), 4(2), 41–47. *https://doi.org/10.37802/joti.v4i2.278*

Jo, H., Kim, C., Gwon, D., Lee, J., Lee, J., Park, K. M., & Park, S. (2023). *Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke : an automated machine learning approach*. Scientific Reports, 1–8. *https://doi.org/10.1038/s41598-023-44201-8*

Lolak, S., Attia, J., Mckay, G. J., & Thakkinstian, A. (2023). *Comparing Explainable Machine Learning Approaches With Traditional Statistical Methods for Evaluating Stroke Risk Models : Retrospective Cohort Study Corresponding Author : 7*, 1–11. *https://doi.org/10.2196/47736*

Mohanty, S., & Acharya, A. A. (2023). ScienceDirect ScienceDirect MFBFST : *Building a stable ensemble learning model using MFBFST : Building a stable ensemble learning model using multivariate filter-based feature selection technique for detection of multivariate filter-based feature selection technique for detection of suspicious URL suspicious URL*. Procedia Computer Science, 218, 1668–1681. *https://doi.org/10.1016/j.procs.2023.01.145*

Pramakrisna, F. D., Adhinata, F. D., Annisa, N., & Tanjung, F. (2022). *Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression Web-based Classifying SMS Application Using Logistic Regression Algorithm*. 11(2), 90–97. *https://doi.org/10.34148/teknika.v11i2.466*

Putri, R. M. (2025). *Analisis Kategori Populasi Negara Menggunakan Random Forest dan Logistic Regression*. *5*(1), 26–31.

Rolansa, F., Yunita, & Suheri. (2021). *Pengembangan Interaktif Dashboard Kemahasiswaan Di Program Program Studi Teknik Informatika Politeknik Negeri Pontianak , JL . Ahmad Yani Pontianak*, Indonesia Program Studi Teknik Informatika Politeknik Negeri Pontianak , JL . Ahmad Yani Pontianak , Indo. Jurnal Pendidikan Informatika Dan Sains, 10(2), 110–118. *https://doi.org/10.31571/saintek.v10i2.2190*

Sheng, Z., Kuang, J., Yang, L., Wang, G., Gu, C., Qi, Y., Wang, R., Han, Y., Li, J., & Wang, X. (2024). *Predictive models for delay in medical decision-making among older patients with acute ischemic stroke: a comparative*

study using logistic regression analysis and lightGBM algorithm. *BMC Public Health*, *24*(1), 1–11. *https://doi.org/10.1186/s12889-024-18855-6*

Sitompul, L. R., Nababan, A. A., Manihuruk, M. L., & Ponsen, W. A. (2025). *Perbandingan XGBoost , Random Forest , dan Algoritma Regresi Logistik pada Stroke Klasifikasi Penyakit*. 9(April), 957–968.

Tuppad, A., & Patil, S. D. (2023). *Data Pre-processing Issues in Medical Data Classification*. Journal of Advanced Zoology, 44, 1079–1084.

Varma, D., Nehansh, A., & Guide, P. S. (2023). *Data Preprocessing Toolkit : An Approach to Automate Data Preprocessing*. 1–5. *https://doi.org/10.55041/IJSREM18270*

Wang, L. (2023). *Logistic Regression for Stroke Prediction: An Evaluation of its Accuracy and Validity*. Highlights in Science, Engineering and Technology, *39*, 1086–1092. *https://doi.org/10.54097/hset.v39i.6712*

Zhang, N., Zhang, X., Li, Q., Zhu, C., Zhou, M., Statistics, H., & Hospital, F. (2023). *The predictive performance of artificial intelligence on the outcome of stroke : a systematic review and meta-analysis*. September, 1–7. *https://doi.org/10.3389/fnins.2023.1256592*

Zul, Z., Al, H., & Subhiyakto, E. R. (2025). *Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression*. 300–311. *https://doi.org/10.33364/algoritma/v.22-1.2164*