Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 3 (2025) pp: 1294-1299
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Prediksi Penyakit Stroke menggunakan Algoritma Decision Tree dan Naïve Bayes

Eka Sabna[1], Octavia Dewi[2]
[12]Universitas Hang Tuah Pekanbaru
es3jelita@gmail.com[1], dewitavia@yahoo.com[2]

Abstrak

*Stroke salah satu penyakit yang dapat menyebabkan kematian dan kecacatan permanen apabila tidak dideteksi sejak dini. Stroke merupakan penyebab kematian kedua terbanyak dan kecacatan ketiga terbanyak di dunia. Teknologi data mining dapat dimanfaatkan untuk membantu proses prediksi penyakit stroke dengan lebih cepat dan akurat. Penelitian ini bertujuan menerapkan dua Algoritma klasifikasi yaitu **Decision Tree** dan **Naïve Bayes** untuk memprediksi risiko stroke berdasarkan data kesehatan pasien. Dataset yang digunakan sebanyak 4.149 data yang terdiri dari atribut usia, tekanan darah, kadar glukosa, indeks massa tubuh, dan kebiasaan merokok. Penelitian ini dataset dibagi menjadi 2 bagian yaitu data training 80% dan data testing 20% menggunakan split validation. Pengujian menggunakan Confusion Matrics, derdasarkan hasil pengujian yang telah dilakukan pada penelitian ini, algoritma Decision Tree mempunyai nilai akurasi dan Recall tertinggi dibandingkan dengan Algoritma Naïve Bayes yaitu Akurasi 97,23% dan Recall 99,63%. Dengan demikian, Decision Tree lebih direkomendasikan sebagai metode klasifikasi untuk prediksi penyakit stroke pada data ini. Temuan ini diharapkan dapat menjadi dasar pengembangan sistem pendukung keputusan medis dalam deteksi dini penyakit stroke. Penelitian ini dapat dilakukan pengembangan lebih lanjut dengan menambah fitur tambahan seperti negara, status gaya hidup untuk meningkatkan ketepatan prediksi.*

Kata kunci: *Prediksi, Stroke, Naive Bayes, Decision Tree, Evaluasi*

1. Pendahuluan

Menurut Global Organisasi Stroke Dunia (WSO) stroke merupakan penyebab kematian terbanyak kedua di dunia [1] [2]. Kematian global akibat stroke dapat mencapai sekitar 9,7 juta per tahun pada tahun 2050, yang merupakan peningkatan hampir 50% dari tingkat kematian tahun 2020, menurut analisis baru dari Komisi Neurologi Lancet-Organisasi Stroke Dunia [3]. Menurut data World Health Organization (WHO), setiap tahun 15 juta orang di seluruh dunia menderita stroke. Dari jumlah tersebut 5 juta orang meninggal dunia dan 5 juta lainnya mengalami cacat permanen yang membebani keluarga dan masyarakat [4].

Stroke terjadi ketika pasokan darah ke otak terganggu atau berkurang akibat penyumbatan atau pecahnya pembuluh darah, sehingga jaringan otak tidak mendapatkan oksigen dan nutrisi yang cukup. Deteksi dini terhadap risiko stroke sangat penting agar tindakan pencegahan dapat dilakukan sedini mungkin, khususnya bagi individu yang memiliki faktor risiko seperti usia lanjut, hipertensi, diabetes, obesitas, kebiasaan merokok, dan riwayat penyakit jantung [1][5].

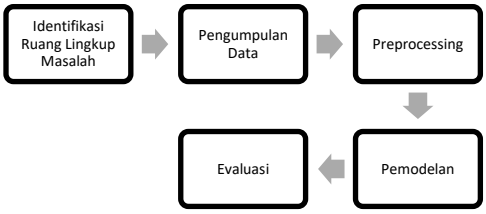
Data mining berhubungan erat dengan analisa data dan penggunaan perangkat lunak untuk mencari pola dan kesamaan dalam sekumpulan data. Salah satu metode data mining adalah klasifikasi. Pada persoalan klasifikasi data akan diprediksi dengan beberapa atribut dan satu atribut target [6] [7]. Dalam beberapa tahun terakhir, perkembangan teknologi informasi khususnya di bidang **data mining** dan **machine learning** telah membuka peluang besar untuk pengembangan sistem prediksi penyakit. Algoritma pembelajaran mesin mampu mengenali pola dari data riwayat kesehatan pasien dan melakukan klasifikasi terhadap risiko penyakit secara otomatis dan efisien. Data mining menjadi solusi potensial untuk mengembangkan sistem prediksi penyakit yang akurat [8]. Studi menggunakan Algoritma Klasifikasi yang diimplementasikan untuk penyakit telah banyak dilakukan, diantara nya adalah klasifikasi penyakit jantung [9], ISPA[10] dan Diabetes [11] . Data mining dapat digunakan untuk menemukan pola tersembunyi dalam data kesehatan pasien yang dapat digunakan untuk memperkirakan kemungkinan terjadinya stroke. Dua algoritma yang sering digunakan dalam

klasifikasi medis adalah **Naive Bayes** dan **Decision Tree.** Algoritma klasifikasi seperti Decision Tree dan Naïve Bayes merupakan metode popular yang digunakan untuk prediksi dalam berbagai kasus medis .

Penelitian ini menggunakan data public dari platform Kaggle sebanyak 4.149 data dengan 11 atribut yaitu Jenis Kelamin, Umur, Hipertensi, Penyakit Jantung, Pernah Menikah, Tipe Pekerjaan, Tempat Tinggal, Kadar Glukosa rata-rata dalam darah, Indeks Massa Tubuh, Status Merokok, Stroke.. Pertanyaan yang ingin dijawab dalam penelitian ini adalah bagaimana menerapkan algoritma klasifikasi Pohon Keputusan dan Naïve Bayes untuk memprediksi resiko penyakit stroke dan kinerja dari kedua Algoritma tersebut.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan klasifikasi data mining untuk memprediksi kemungkinan terjadinya resiko stroke pada seseorang berdasarkan atribut Jenis Kelamin, Umur, Hipertensi, Penyakit Jantung, Pernah Menikah, Tipe Pekerjaan, Tempat Tinggal, Kadar Glukosa rata-rata dalam darah, Indeks Massa Tubuh, Status Merokok. Terdapat beberapa tahapan yang dilakukan dalam proses penelitian ini yaitu :



Gambar 1. Metode Penelitian

2.1 Identifikasi Ruang Lingkup Masalah

Pada tahap ini, dilakukan penentuan batasan terhadap masalah yang akan diteliti agar penelitian lebih terarah, fokus dan sistematis. **Ruang lingkup masalah** menjelaskan **batasan** dan **cakupan** dari penelitian yang dilakukan agar pembahasan tidak melebar dan bisa diselesaikan secara realistis sesuai tujuan yang ingin dicapai. Analisis difokuskan untuk membangun model prediksi risiko stroke.

2.2 Pengumpulan Data

Pengumpulan data adalah proses mengumpulkan informasi yang diperlukan untuk menjawab rumusan masalah, menguji hipotesis, dan mencapai tujuan penelitian. Data yang dikumpulkan harus relevan, valid, dan dapat diandalkan. Pada penelitian ini pengumpulan data dilakukan untuk mendapatkan sumber data. Dataset yang digunakan merupakan data sekunder yang berasal dari web penyedia datasets yaitu Kaggle yang berjudul *"Stroke Prediction Dataset"*. Dataset ini terdiri dari informasi dari pasien dengan atribut seperti usia, jenis kelamin, tekanan darah, kadar glukosa, BMI, riwayat hipertensi, penyakit jantung, pekerjaan, tempat tinggal dan status merokok.

Tabel 1. Keterangan Dataset

Variabel (Atribut)	Deskripsi Variabel
id	Pengenalan unik
Gender	Female=0, Male=1
Age	Usia Pasien
Hypertensi	Jika pasien tidak hypertensi=0, hypertensi=1
Heart Disease	Jika pasien tidak penyakit jantung=0, penyakit jantung=1
Ever Married	Tidak dan Ya
Work Type	Tipe Pekerjaan
Residence Type	Rural & Urban
Average Glucose Level	Kadar Glukosa rata-rata dalam darah
BMI	Indeks Massa Tubuh
Smoking Status	Merokok atau Tidak
Stroke	Stroke = 1, Tidak Stroke = 0

2.3 Preprocessing

Tahap ini adalah langkah pembersihan data sebelum ke tahap berikutnya. Tahap preprocessing (pra-pemrosesan) merupakan langkah penting dalam proses data mining yang bertujuan untuk membersihkan,

DOI: https://doi.org/ 10.31004/riggs.v4i3.2132

menyiapkan, dan mengubah data mentah menjadi format yang layak digunakan oleh algoritma pembelajaran mesin. Pada penelitian ini, tahapan preprocessing yang dilakukan adalah pemeriksaaan dan penanganan missing value dan duplikasi data.

2.4 Pemodelan

Tahap pemodelan adalah proses menerapkan model klasifikasi berdasarkan data yang telah dipreproses. Tahap **pemodelan** merupakan proses membangun dan melatih model algoritma machine learning berdasarkan data yang telah dipersiapkan, dengan tujuan untuk memprediksi atau mengklasifikasikan hasil tertentu, dalam konteks penelitian ini untuk memprediksi apakah seorang pasien berisiko mengalami **stroke** atau tidak. Penelitian ini menggunakan dua algoritma klasifikasi, yaitu **Naïve Bayes** dan **Decision Tree**, dengan atribut yang disesuaikan untuk prediksi penyakit stroke.

Algoritma Naïve Bayes :

Bayesian classification adalah pengklasifikasian statistik yang dapat di gunakan untuk memprediksi probabilitas keanggotaan suatu class. Bentuk umum Teorema Bayes adalah :

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$$

P(C|X) adalah Probabilitas hipotesis C (kelas, seperti "stroke") diberikan data X (fitur-fitur pasien), P(X|C) adalah Probabilitas data X muncul jika diketahui kelas C, P(C) adalah Probabilitas awal (prior probability) kelas C dan P(X) adalah Probabilitas awal data X (konstanta untuk semua kelas)

Algoritma Decision Tree (C4.5) :

C4.5 adalah algoritma pohon keputusan yang dikembangkan oleh **Ross Quinlan** sebagai pengembangan dari algoritma **ID3**. C4.5 digunakan untuk tugas **klasifikasi** dan menghasilkan pohon keputusan dari dataset pelatihan. C4.5 adalah algoritma pengembangan dari ID3 yang menggunakan Gain Ratio untuk pemilihan atribut terbaik.

Langkah-langkah C4.5 adalah Hitung Entropy Awal, hitung Information Gain setiap Atribut, hitung Split Info dan Gain Ratio, Hitung GainRatio(A) = Gain(A) /SplitInfo(A), Pilih atribut dengan Gain Ratio tertinggi dan bagi data berdasarkan atribut terpilih dan ulangi proses rekursif.

2.5 Evaluasi

Tahap **evaluasi klasifikasi** adalah proses untuk mengukur seberapa baik model klasifikasi (seperti Decision Tree atau Naive Bayes) dalam memprediksi kelas target berdasarkan data uji. Evaluasi ini penting untuk mengetahui akurasi dan keandalan model sebelum digunakan pada data nyata. Evaluasi klasifikasi merupakan tahap penting dalam pengujian model prediksi untuk menilai seberapa baik model dalam memprediksi kelas target. Evaluasi dilakukan menggunakan berbagai metrik berdasarkan hasil dari confusion matrix.

Confusion matrix adalah matriks yang menunjukkan perbandingan antara hasil prediksi model dan nilai actual kelas target. Matriks ini terdiri dari empat komponen yaitu True Positive (TP) adalah data positif yang terprediksi benar, False Positive (FP) adalah data negatif tapi terprediksi sebagai data positif, False Negative (FN) adalah data positif yang terprediksi sebagai data negatif dan True Negative (TN) adalah data negatif yang terprediksi dengan benar.

Tabel 2. Confusion Matrix

	Prediksi Positif	Prediksi Negatif
Aktual Positif (Stroke)	True Positive (TP)	False Negative (FN)
Aktual Negatif (Tidak Stroke)	False Positive (FP)	True Negative (TP)

Dari confusion matrix, berbagai metric dapat dihitung :

- Akurasi (Accuracy) adalah **proporsi prediksi yang benar** dibandingkan dengan seluruh jumlah data. Accuracy = (TP +TN) / (TP+TN+FP+FN)
- Presisi menunjukkan **seberapa banyak dari hasil prediksi positif (stroke)** yang benar-benar **stroke**. Precision = TP / (TP+FP)

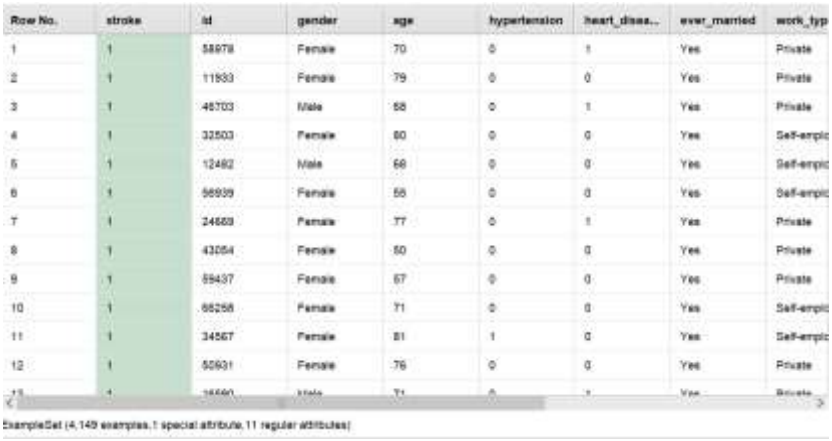
DOI: https://doi.org/ 10.31004/riggs.v4i3.2132

- Recall menunjukkan **seberapa banyak dari semua pasien yang benar-benar stroke,** yang berhasil **terdeteksi oleh model.**
  Recall = TP / (TP+FN)

3. Hasil dan Diskusi

3.1 Hasil

Data yang digunakan sebanyak 4.149 data dengan 11 atribut, berikut tampilan data resiko stoke [12].

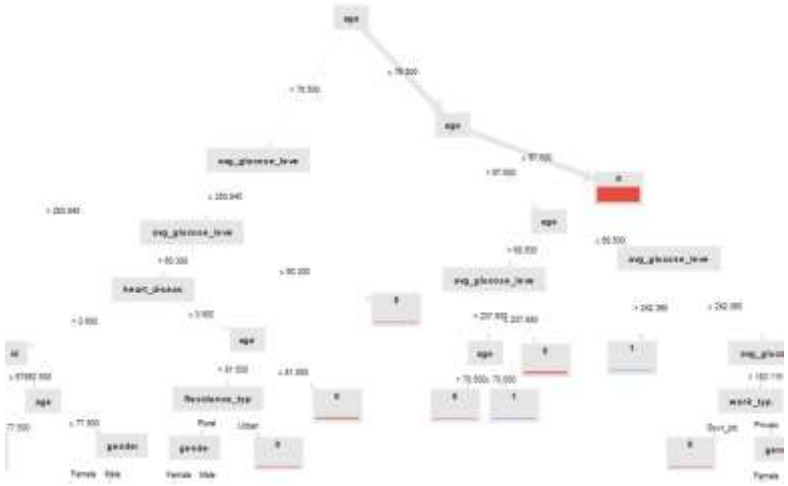


Row No.	stroke	id	gender	age	hypertension	heart_disea...	ever_married	work_typ
1	1	58978	Female	70	0	1	Yes	Private
2	1	11933	Female	79	0	0	Yes	Private
3	1	46703	Male	68	0	1	Yes	Private
4	1	32503	Female	80	0	0	Yes	Self-empl
5	1	12482	Male	68	0	0	Yes	Self-empl
6	1	56939	Female	55	0	0	Yes	Self-empl
7	1	24669	Female	77	0	1	Yes	Private
8	1	43054	Female	60	0	0	Yes	Private
9	1	59437	Female	67	0	0	Yes	Private
10	1	66256	Female	71	0	0	Yes	Self-empl
11	1	34567	Female	81	1	0	Yes	Self-empl
12	1	52631	Female	76	0	0	Yes	Private

ExampleSet (4.149 examples,1 special attribute,11 regular attributes)

Gambar 2. Data resiko stroke

Pemodelan dilakukan dengan **Split data** yaitu proses **membagi dataset** menjadi beberapa bagian untuk tujuan pelatihan (**training**) dan pengujian (**testing**) model machine learning. Proporsi untuk Data latih (training) dan data uji (testing) dengan rasio 80:20. Pemodelan dengan Algoritma Decision Tree menghasilkan pohon keputusan sebagai berikut :

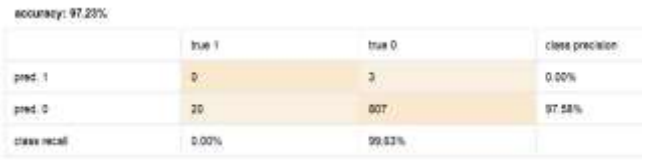


Gambar 3. Model Algoritma Decision Tree

Pohon keputusannya pada Gambar 3 memperlihatkan bahwa Atribut Age (umur) menjadi Node Awal artinya Atribut Age (umur) merupakan atribut yang paling menentukan dalam memprediksi penyakit Stroke.

Setelah dilakukan pelatihan model Algoritma Decision Tree dievaluasi dengan tiga metrik utama yaitu Presisi, Recall dan Akurasi, diperoleh hasil evaluasi sebagai berikut:

DOI: https://doi.org/ 10.31004/riggs.v4i3.2132


accuracy: 97.23%

	true 1	true 0	class precision
pred. 1	0	3	0.00%
pred. 0	20	807	97.58%
class recall	0.00%	99.63%	

Gambar 3. Hasil Evaluasi Algoritma Decision Tree

Hasil evaluasi di menghasilkan dapatkan nilai Akurasi, Recall dan Presisi adalah :
Akurasi = (0+807)/830 = 97,23 %
Recall= 827 / 830 = 99,63%
Presisi = 807/827 =97,58%

Pemodelan dengan Algoritma Naïve Bayes menghasilkan **Simple Distribution. Simple Distribution adalah menghitung probabilitas berdasarkan frekuensi kemunculan** di data.

Tabel 3. Hasil Simple Distribution

Class	Hasil
Stroke (1)	0,024
Tidak Stroke	0,976

Setelah dilakukan pelatihan model Algoritma Naïve Bayes diperoleh hasil evaluasi sebagai berikut:


accuracy: 92.89%

	true 1	true 0	class precision
pred. 1	7	46	13.21%
pred. 0	13	764	98.33%
class recall	35.00%	94.32%	

Gambar 4. Hasil Evaluasi Algoritma Naïve Bayes

Hasil evaluasi di dapatkan nilai Akurasi, Recall dan Presisi yaitu :

Akurasi = 771/830 = 92,89%
Recall = 764 / 810 = 94,32 %
Presisi = 764 / 777 = 98,33%

3.2 Diskusi dan Implikasi Penelitian

Hasil penelitian ini menunjukkan bahwa algoritma Decision Tree memberikan nilai akurasi dan recall yang lebih tinggi untuk prediksi resiko penyakit Stroke dibandingkan dengan Algoritma Naïve Bayes. Hal ini menunjukkan bahwa Decesion Tree lebih baik dalam menangani atribut-atribut yang memiliki hubungan kompleks terhadap target (Stroke). Meskipun demikian Algoritma Naïve Bayes menunjukkan kinerja yang cukup baik dalam proses klasifikasi.

Tabel 4. Hasil Evaluasi model dengan Algoritma Decision Tree dan Naïve Bayes

	Algoritma Decision Tree	Algoritma Naïve Bayes
Akurasi	97,23%	92,89%
Presisi	97,58%	98,33%
Recall	99,63%	94,32%

Namun, terdapat beberapa hal yang dapat ditingkatkan dalam penelitian .
1. Menambah Fitur baru seperti negara sebagai historis data untuk dapat memprediksi berdasarkan negara asal dan fitur Status Gaya Hidup.

DOI: https://doi.org/ 10.31004/riggs.v4i3.2132

2. Algoritma Naive Bayes dapat bekerja lebih baik dengan data kategorikal misalnya kategori untuk Usia (muda,dewasa, tua) dan kadar Glukosa (Rendah, Normal, Tinggi)

4. Kesimpulan

Prediksi penyakit stroke menggunakan algoritma klasifikasi Decision Tree menghasilkan tingkat akurasi sebesar 97,23 % dan nilai Recall 99,63% hasil ini lebih tinggi di bandingkan dengan Algoritma Naïve Bayes. Pemodelan dengan Algoritma Decesion Tree dapat digunakan sebagai Prediksi penyakit resiko stroke dengan Baik dan diharapkan dapat menjadi acuan bagi penderita stroke, dokter dan masyarakat . Model ini sebagai pengetahuan agar Masyarakat menjaga pola hidup dan menghindari penyakit stroke melalui variabel yang mempengaruhi terjadinya penyakit tersebut. Hasil penelitian ini diharapkan dapat menjadi referensi bagi masyarakat. Selain itu, disarankan untuk melakukan penelitian tentang prediksi penyakit stroke menggunakan algoritma lain agar mendapatkan kinerja model yang lebih tinggi seperti Gaussian NB dan Categorical NB.

Referensi

[1] V. L. Feigin *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022, doi: 10.1177/17474930211065917,.

[2] V. L. Feigin *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet Neurol.*, vol. 23, no. 10, pp. 973–1003, Oct. 2024, doi: 10.1016/S1474-4422(24)00369-7.

[3] T. Reed, "Stroke deaths could jump 50% by 2050, study warns," 2023. https://www.axios.com/2023/10/10/stroke-deaths-could-jump-50-by-2050-study-warns (accessed Aug. 05, 2025).

[4] "WHO EMRO | Stroke, Cerebrovascular accident | Health topics," 2025. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html (accessed Aug. 05, 2025).

[5] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659/ASSET/7C114CC3-D4F4-4270-B285-D12BA9378850/ASSETS/CIR.0000000000000659.FP.PNG.

[6] Suyanto, "Machine Learning : Tingkat Dasar dan Lanjut," 2018, Accessed: Aug. 28, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/146400/slug/machine-learning-tingkat-dasar-dan-lanjut.html.

[7] G. S. Thejas, Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "Machine Learning with Applications An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," *Mach. Learn. with Appl.*, vol. 8, no. January, p. 100267, 2022, doi: 10.1016/j.mlwa.2022.100267.

[8] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[9] A. B. Wibisono and A. Fahrurozi, "PERBANDINGAN ALGORITMA KLASIFIKASI DALAM PENGKLASIFIKASIAN DATA PENYAKIT JANTUNG KORONER," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, Feb. 2020, doi: 10.35760/TR.2019.V24I3.2393.

[10] D. Meylitasari, B. Tarigan, P. Dian, M. T. Rini, and V. Puspita, "Perancangan Data Mining untuk Klasifikasi Prediksi Penyakit ISPA dengan Algoritma C4.5," *Comput. Sci. ICT*, vol. ISBN, no. 1, pp. 979–587, 2017.

[11] E. Sabna, "PENERAPAN ALGORITMA KLASIFIKASI DATA MINING POHON KEPUTUSAN UNTUK PREDIKSI PENYAKIT DIABETES," 2024. https://com.ojs.co.id/index.php/jikr/article/view/105/117 (accessed Aug. 05, 2025).

[12] Fedesoriano, "Stroke Prediction Dataset," 2020. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed Aug. 06, 2025).

DOI: https://doi.org/ 10.31004/riggs.v4i3.2132