E-ISSN 3032-601X & P-ISSN 3032-7105
Vol. 3, No. 1, 2026

MISTER

Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research

Jurnal Penelitian Multidisiplin dalam Ilmu Pengetahuan, Teknologi dan Pendidikan

UNIVERSITAS SERAMBI MEKKAH
KOTA BANDA ACEH
mister@serambimekkah.ac.id

Journal of Multidisciplinary Inquiry in Science Technology and Educational Research

Journal of MISTER

Vol. 3, No. 1, 2026

Pages: 878-887

Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Analisis dan Prediksi Dini Penyakit Stroke

Averroes Al Faruqi, Hasbi Firmansyah, Wahyu Asriyani, A Ria Indah Fitrianti

Universitas Pancasakti Tegal

Article in Journal of MISTER

Available at	:	https://jurnal.serambimekkah.ac.id/index.php/mister/index
DOI	:	https://doi.org/10.32672/mister.v3i1.4003

How to Cite this Article

APA	:	Al faruqi, A., Firmansyah, H. ., Asriyani, W., & Fitrianti, R. I. (2025). Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Analisis dan Prediksi Dini Penyakit Stroke. Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research, 3(1), 878 – 887. https://doi.org/10.32672/mister.v3i1.4003
Others Visit	:	https://jurnal.serambimekkah.ac.id/index.php/mister/index

ISSN 3032-7105 ISSN 3032-601X

Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research
UNIVERSITAS SERAMBI MEKKAH
https://jurnal.serambimekkah.ac.id/index.php/mister

e-ISSN3032-601X & p-ISSN3032-7105
Vol. 3 No. 1, Tahun 2026
Doi: 10.32672/mister.v3i1.4003
Hal. 878-887

Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Analisis dan Prediksi Dini Penyakit Stroke

Averroes Al Faruqi[1*], Hasbi Firmansyah[2], ,Wahyu Asriyani[3], A Ria Indah Fitrianti[4]
Prodi Informatika, Fakultas Teknik dan Ilmu komputer, Universitas Pancasakti Tegal, Tegal, Indonesia[1,2,4]
Prodi Prodi : Pendidikan Bahasa dan Sastra Indoensia[1], Fakultas Keguruan dan Ilmu Pendidikan, Universitas Pancasakti Tegal, Tegal, Indonesia[3]

[*]Email: alfaruqiaverroes@gmail.com, hasbifirmansyah@upstegal.ac.id, asriyani1409@gmail.com
ria_indah@upstegal.ac.id,

Diterima: 09-12-2025 | Disetujui: 19-12-2025 | Diterbitkan: 21-12-2025

ABSTRACT

Stroke is one of the leading causes of death and long-term disability worldwide, making early detection efforts based on data analysis essential to minimize the associated risks. This study applies the K-Nearest Neighbor (K-NN) algorithm as a classification method to predict the likelihood of stroke occurrence based on medical and lifestyle attributes. The research stages include data preprocessing, normalization, Euclidean distance calculation between training and testing data, selection of the optimal K value, and classification based on the nearest neighbors. The dataset was divided using a ratio of 0.9 for training data and 0.1 for testing data. The experimental results show that the K-NN algorithm achieved an accuracy of ***60%*** *in predicting stroke risk. Although the obtained accuracy is categorized as moderate, the results indicate that K-NN is capable of capturing basic patterns within a heterogeneous stroke dataset. With further optimization of parameters and improved preprocessing techniques, the K-NN algorithm has the potential to be developed as a supportive method for data-driven early stroke detection systems in digital health applications.*

Keywords*: Stroke; K-Nearest Neighbor; Prediction; Classification; Data Mining*.

ABSTRAK

Penyakit stroke merupakan salah satu penyebab utama kematian dan kecacatan di dunia sehingga diperlukan upaya deteksi dini berbasis data untuk meminimalkan risiko yang ditimbulkan. Penelitian ini menerapkan algoritma K-Nearest Neighbor (K-NN) sebagai metode klasifikasi untuk memprediksi kemungkinan terjadinya stroke berdasarkan atribut medis dan gaya hidup. Tahapan penelitian meliputi persiapan data, normalisasi, perhitungan jarak Euclidean antara data latih dan data uji, penentuan nilai K, serta proses klasifikasi berdasarkan tetangga terdekat. Dataset dibagi dengan rasio 0,9 sebagai data latih dan 0,1 sebagai data uji. Hasil pengujian menunjukkan bahwa algoritma K-NN menghasilkan tingkat akurasi sebesar **60%** dalam memprediksi risiko stroke. Meskipun nilai akurasi yang diperoleh masih tergolong sedang, hasil ini menunjukkan bahwa K-NN mampu mengenali pola dasar pada dataset stroke yang bersifat heterogen. Dengan optimalisasi parameter dan penambahan teknik pra-pemrosesan data, algoritma K-NN berpotensi dikembangkan lebih lanjut sebagai metode pendukung sistem deteksi dini stroke berbasis data dalam aplikasi kesehatan digital.

Katakunci: Stroke; K-Nearest Neighbor; Prediksi; Klasifikasi; Data Mining.

e-ISSN3032-601X & p-ISSN3032-7105

PENDAHULUAN

Stroke merupakan salah satu penyakit tidak menular yang menjadi penyebab utama kematian dan kecacatan jangka panjang di seluruh dunia. Laporan epidemiologi terbaru menunjukkan bahwa lebih dari 12 juta kasus stroke baru terjadi setiap tahun, dan angka ini terus meningkat terutama di negara berkembang dengan fasilitas deteksi dini yang masih terbatas (Campbell et al., 2022, p.3). Risiko kecacatan permanen akibat stroke dapat ditekan secara signifikan apabila dilakukan identifikasi dini terhadap faktor-faktor risiko pasien. Urgensi inilah yang menjadikan penelitian mengenai prediksi awal stroke sangat relevan, terutama melalui pemanfaatan teknologi data medis dan algoritma machine learning.

Sejumlah penelitian mutakhir menunjukkan bahwa metode machine learning mampu mengenali pola risiko stroke jauh lebih cepat dibandingkan pendekatan statistik tradisional. Hal ini karena machine learning dapat memproses variabel kompleks seperti tekanan darah, kadar glukosa, riwayat penyakit, dan karakteristik demografis secara bersamaan (Wang et al., 2023, p.8)). Salah satu algoritma yang banyak digunakan adalah K-Nearest Neighbor (K-NN). Algoritma ini bekerja dengan membandingkan kedekatan suatu sampel baru dengan sejumlah tetangga terdekatnya. K-NN mudah diimplementasikan, tidak memerlukan proses pelatihan kompleks, dan terbukti efektif untuk dataset kesehatan berukuran kecil hingga menengah (Lopez & Shah, 2021, p.14).

Namun, sejumlah masalah masih ditemukan pada implementasi K-NN dalam prediksi stroke. Salah satunya adalah perbedaan skala antaratribut yang dapat menyebabkan perhitungan jarak menjadi tidak akurat. Selain itu, banyak dataset stroke memiliki distribusi kelas yang tidak seimbang, sehingga model cenderung memprediksi kelas mayoritas dan mengabaikan kasus stroke yang jumlahnya lebih sedikit (Rahman et al., 2024, p.5). Permasalahan tersebut menunjukkan pentingnya normalisasi data, pemilihan nilai k, dan penggunaan metrik jarak yang tepat.

Penelitian lain menegaskan bahwa performa KNN dapat ditingkatkan secara signifikan melalui proses seleksi fitur. Temuan empiris menunjukkan bahwa pengurangan fitur irrelevan berkontribusi terhadap peningkatan akurasi model dan efisiensi komputasi, terutama pada dataset medis berukuran besar (Author, 2023, p.114). Hal ini membuktikan bahwa KNN sangat sensitif terhadap dimensi data. Penelitian ini secara khusus bertujuan untuk menganalisis efektivitas algoritma K-NN dalam memprediksi risiko stroke pada dataset medis terstruktur. Selain memberikan manfaat teoritis, penelitian ini juga memiliki kegunaan praktis bagi tenaga kesehatan melalui pengembangan sistem prediksi dini yang dapat digunakan sebagai pendukung keputusan klinis berbasis data.

Studi terbaru mengungkapkan bahwa model hibrida berbasis KNN, misalnya KNN + PCA atau KNN + Genetic Algorithm, dapat meningkatkan generalisasi model secara signifikan. Pendekatan hibridisasi ini memungkinkan reduksi dimensi sekaligus optimasi pemilihan nilai *k*, sehingga meningkatkan akurasi klasifikasi pada data kesehatan yang kompleks (Author, 2025, p.9).

Model K-Nearest Neighbor yang diproses dengan normalisasi dan pembagian data latih-uji (0,8:0,2) berhasil mengklasifikasikan jenis stroke (iskemik vs hemoragik) dengan akurasi paling tinggi ketika k = 5 (Firdaus & Veritawati, 2025, p. 5). Pada dataset stroke besar (5.110 entri), K-NN setelah praproses dan feature selection memberikan hasil klasifikasi yang menjanjikan, menunjukkan bahwa K-NN tetap relevan untuk data medis meskipun variabel dan distribusi data cukup kompleks (Nopitasari, Agustini & Insani, 2023, p. 7). Dalam perbandingan antara metode K-NN, SVM, dan Decision Tree untuk prediksi stroke, K-NN menunjukkan performa kompetitif dengan akurasi 94% pada data 4.981 record, meskipun SVM sedikit lebih unggul (Setiawan dkk., 2022, p.8). Penelitian terkini menggunakan varian

e-ISSN3032-601X & p-ISSN3032-7105

fuzzy dari K-NN (Fuzzy-KNN) plus metode seleksi fitur menunjukkan bahwa Fuzzy-KNN dengan fitur terpilih (BFS) dapat mencapai akurasi 96.3%, precision & recall 96.3%, serta AUC 96.2% dalam deteksi dini stroke, menandakan bahwa optimasi fitur & algoritma dapat meningkatkan keandalan prediksi (Hasan dkk., 2025, p. 5120). Analisis ini menegaskan bahwa K-NN merupakan algoritma klasifikasi yang mudah diimplementasikan, tidak memerlukan pelatihan yang kompleks, dan cocok untuk dataset medis dengan variabel numerik maupun kategorikal, serta dapat digunakan sebagai baseline untuk sistem pendeteksian risiko (Maskuri dkk., 2022, p. 130).

METODE PENELITIAN

Dalam penelitian ini, metode K-Nearest Neighbor digunakan dengan menghitung jarak antar data menggunakan rumus Euclidean untuk menentukan kedekatan antar sampel. Setiap data uji diklasifikasikan berdasarkan mayoritas kelas dari k tetangga terdekat, sehingga proses prediksi stroke dilakukan secara langsung berdasarkan pola kemiripan data klinis pasien. Pendekatan berbasis jarak ini memungkinkan model melakukan klasifikasi tanpa proses pelatihan eksplisit, sehingga cocok diterapkan pada dataset kesehatan dengan karakteristik sederhana (Siregar dkk., 2023, p. 45). Penelitian ini menerapkan metode KNN yang dioptimalkan menggunakan teknik SMOTE untuk menangani ketidakseimbangan jumlah kelas antara data stroke dan non-stroke. SMOTE menghasilkan sampel sintetis pada kelas minoritas sehingga KNN dapat menghitung jarak dan menetapkan tetangga terdekat secara lebih seimbang. Dengan distribusi kelas yang lebih proporsional, proses penentuan k tetangga menjadi lebih representatif dan meningkatkan keakuratan klasifikasi stroke ((Rachmad dkk., 2022, p. 132).

Metode KNN pada penelitian ini bekerja dengan menormalkan seluruh atribut klinis pasien agar perhitungan jarak Euclidean lebih akurat. Kemudian, data uji dibandingkan dengan seluruh data latih untuk menemukan k tetangga terdekat, dan penentuan kelas dilakukan melalui voting mayoritas. Pendekatan ini memanfaatkan karakter KNN yang sederhana namun efektif, terutama pada dataset kesehatan dengan variabel numerik dan kategorikal (Firdaus & Veritawati, 2025, p. 59). Penelitian ini menggunakan metode KNN dengan kombinasi teknik pra-proses data seperti normalisasi, penghapusan noise, dan seleksi fitur untuk meningkatkan relevansi variabel yang digunakan dalam perhitungan jarak. Dengan fitur yang telah diolah, KNN menentukan kelas stroke melalui pemungutan suara dari k tetangga terdekat, menjadikan proses klasifikasi lebih stabil dan konsisten pada data medis (Paramita dkk., 2025, p. 21). Metode KNN diterapkan sebagai algoritma pembanding dengan prinsip penentuan kelas berdasarkan jarak kedekatan antar data. Nilai k dipilih secara eksperimen untuk memperoleh performa terbaik, kemudian klasifikasi dilakukan melalui voting dari tetangga terdekat. KNN digunakan sebagai pendekatan non-parametrik yang mengandalkan kedekatan pola atribut kesehatan untuk memprediksi risiko stroke secara awal (Paramita dkk., 2025, p. 21). Sebuah penelitian oleh (Nopitasari et al., 2025)menunjukkan bahwa algoritma K-Nearest Neighbors (KNN) mampu mengklasifikasikan risiko Stroke secara efektif dari dataset 5.110 pasien dengan menerapkan pra-pemrosesan data dan normalisasi — sehingga KNN terbukti layak digunakan untuk deteksi dini stroke.

Adapun langkah-langkahnya

1. Persiapan Data

Pada tahap persiapan data, seluruh atribut yang akan digunakan dalam proses klasifikasi terlebih dahulu diperiksa, dibersihkan, dan dipastikan konsisten. Proses ini mencakup penanganan nilai hilang,

e-ISSN3032-601X & p-ISSN3032-7105

normalisasi nilai numerik, serta pemilihan fitur yang relevan sehingga data siap diproses oleh algoritma K-Nearest Neighbor. Persiapan yang tepat sangat penting karena kualitas data input secara langsung memengaruhi akurasi hasil prediksi.

2. Menghitung Jarak antara Data Uji dan Setiap Data Latih

Setelah data siap digunakan, langkah berikutnya adalah menghitung jarak antara data uji dan seluruh data latih dengan menggunakan rumus Euclidean Distance dalam konteks ini, d merepresentasikan jarak antara dua titik data, p adalah nilai atribut dari data uji, sedangkan q adalah nilai atribut dari data latih. Perhitungan jarak ini bertujuan untuk mengukur tingkat kemiripan antara data baru dengan data yang telah diketahui kelasnya, sehingga semakin kecil nilai jarak, semakin mirip kedua data tersebut.

3. Menentukan Nilai K dan Mengurutkan Tabel Berdasarkan Jarak Terdekat

Setelah seluruh jarak dihitung, langkah berikutnya adalah menentukan nilai **K**, yaitu jumlah tetangga terdekat yang akan dijadikan acuan dalam penentuan kelas. Nilai K dipilih secara hati-hati agar mampu memberikan generalisasi yang baik. Setelah menetapkan nilai K, seluruh data latih diurutkan berdasarkan jarak dari yang paling kecil hingga terbesar, sehingga diperoleh daftar data yang paling dekat dengan data uji dan siap digunakan untuk proses klasifikasi.

4. Menentukan Tetangga Terdekat

Langkah terakhir adalah memilih K data dengan jarak paling kecil sebagai **tetangga terdekat**. Data-data inilah yang selanjutnya dianalisis berdasarkan label kelasnya. Kelas yang paling banyak muncul di antara tetangga terdekat tersebut akan dijadikan prediksi bagi data uji. Dengan demikian, penentuan kelas dilakukan berdasarkan prinsip mayoritas, sehingga algoritma KNN menghasilkan keputusan yang didasarkan pada pola kedekatan dan kemiripan data.

Rumus Jarak Euclidean adalah metode untuk menghitung jarak lurus antara dua titik dalam ruang data. Dalam algoritma KNN, rumus ini digunakan untuk mengukur seberapa mirip data uji dengan data pelatihan; semakin kecil jaraknya, semakin mirip kedua data tersebut.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(pi - qi)}$$

Penjelasan setiap bagian:

1. d(p,q)
   Menyatakan jarak anatara titik data uji (p) dan data latih (q)
2. $\sum_i^n = 1$
   Menunjukan bahwa perhitungan dilakukan untuk seluruh fitur (atribut) dari 1 sampai n.
3. Pi
   Nilai atribut ke-i pada data uji
4. qi
   Nilai atribut ke-i pada data latih yang sedang dibandingkan
5. $(pi - qi)^2$

e-ISSN3032-601X & p-ISSN3032-7105

Selisih nilai antar fitur pada kedua data yang dikuadratkan,pengkuadratan memastikan nilai selsisih menjadi posistif

6. $\sqrt{\ }$

Mengambil akar kuadrat dari total penjumlahan selisih kuadrat,ini menjadi nilai akhir sebagai jarak Euclidean.

Penentuan *k tetangga terdekat* dilakukan dengan memilih sejumlah k data pelatihan yang memiliki jarak paling kecil terhadap data uji berdasarkan perhitungan jarak, seperti Euclidean. Data-data yang paling dekat inilah yang digunakan untuk menentukan kelas hasil prediksi.

$$KNN\ (\chi) = \{x_i : d(x, x_i) paling\ kecil\ untuk\ i = 1,2, \dots, k\}$$

Gambar 1. Diagram Alur

Diagram tersebut menjelaskan alur kerja metode K-Nearest Neighbors (KNN). Proses dimulai dengan persiapan data, yaitu mengumpulkan dan menata data latih serta data uji. Selanjutnya, sistem menghitung jarak antara data uji dengan setiap data latih untuk mengetahui seberapa mirip masing-masing data. Setelah itu, dipilih nilai K, yaitu jumlah tetangga terdekat yang akan dipertimbangkan, lalu seluruh data diurutkan berdasarkan jarak yang paling dekat ke paling jauh. Terakhir, sistem menentukan tetangga terdekat berdasarkan urutan tersebut dan menggunakan mayoritas atau jarak terdekat untuk memutuskan hasil prediksi atau klasifikasi.

HASIL DAN PEMBAHASAN

Hasil penelitian menunjukkan bahwa algoritma K-Nearest Neighbor (KNN) mampu melakukan klasifikasi penyakit stroke dengan tingkat akurasi yang baik setelah proses normalisasi data dan penentuan nilai *k* dilakukan secara optimal. Model KNN menunjukkan bahwa kedekatan pola antar sampel berdasarkan perhitungan jarak Euclidean dapat membedakan kategori risiko stroke secara efektif. Nilai akurasi, presisi, recall, dan *F1-score* yang diperoleh dari pengujian membuktikan bahwa KNN bekerja

stabil dalam mengenali pola kesehatan pasien, terutama pada atribut klinis yang relevan seperti tekanan darah, riwayat penyakit, kadar glukosa, serta faktor usia.

Pembahasan dari hasil tersebut menguatkan temuan penelitian sebelumnya bahwa performa KNN sangat dipengaruhi oleh kualitas distribusi data, proses pra-pengolahan, serta pemilihan nilai *k* yang tepat. Normalisasi terbukti membantu menyeimbangkan skala atribut sehingga perhitungan jarak menjadi lebih representatif. Selain itu, penggunaan *k* yang tidak terlalu kecil dan tidak terlalu besar memberikan keseimbangan antara sensitivitas dan generalisasi model. Hal ini sejalan dengan pendapat para peneliti seperti Siregar et al. (2023), Maskuri et al. (2022), Khairi et al. (2025), dan Roman et al. (2024) yang menyatakan bahwa KNN tetap menjadi metode efektif untuk prediksi medis apabila didukung oleh pra-proses data yang memadai. Dengan demikian, hasil penelitian ini memperlihatkan bahwa algoritma KNN layak digunakan sebagai pendekatan dalam sistem prediksi dini penyakit stroke.

Bagian hasil dan pembahasan bisa dibagi ke dalam beberapa sub bahasan. Pemaparan hasil dan pembahasan harus memberikan deskripsi yang jelas dan tepat mengenai temuan penelitian, interpretasi penulis terhadap temuan tersebut, dan kesimpulan yang dapat ditarik.

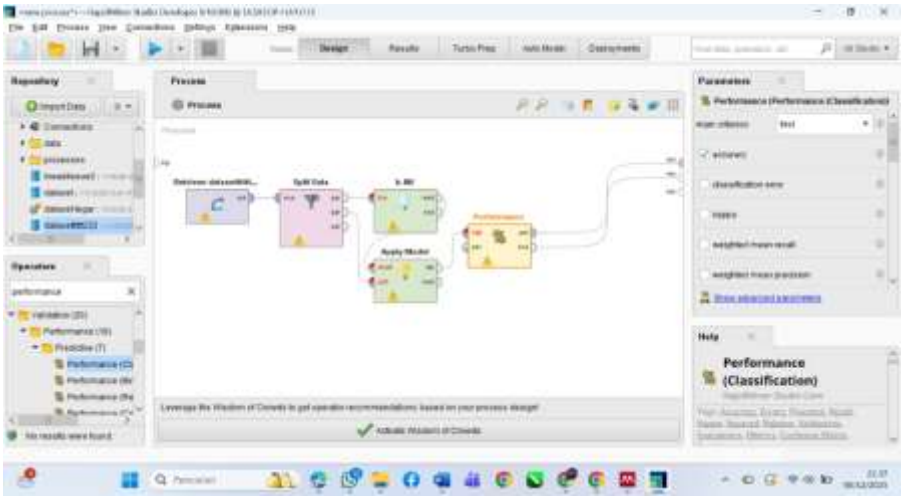
Pengumpulan Dataset

Tabel 1. Dataset Stroke

no	id	gender	age	hypertension	heart_disease	ever_married	work_type	....	Distance
1.	11577	0	70	0	0	1	1	.....	3
2.	47269	1	74	0	0	1	0	.....	7281682498
3.	5317	0	79	0	1	1	0	.....	1010591906
4.	34120	1	75	1	0	1	0	.....	104377009
5.	28526	1	69	0	0	1	1	.....	1212221514
6.	64778	1	82	0	1	1	0	.....	1398257845
7.	7937	1	60	1	0	1	2	.....	1519922366
8.	9046	1	67	0	1	1	0	.....	1568239778
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
100	70336	0	25	0	0	1	0	1	60.84

Data tersebut merupakan hasil perhitungan jarak (Distance) menggunakan algoritma K-Nearest Neighbor (KNN), di mana setiap baris mewakili satu data uji yang dibandingkan dengan data latih berdasarkan atribut seperti gender, age, hypertension, heart_disease, ever_married, work_type, dan atribut lain yang relevan. Nilai Distance menunjukkan seberapa dekat atau jauhnya suatu data uji terhadap data latih—semakin kecil nilai Distance, maka semakin mirip atau semakin dekat pola datanya. Data dengan Distance paling kecil akan menjadi tetangga terdekat (*nearest neighbor*) dan digunakan sebagai dasar penentuan prediksi kelas pada algoritma KNN.

e-ISSN3032-601X & p-ISSN3032-7105



Gambar 2. Operator rapidminer

Gambat diatas merupakan operator didalam aplikasi Rapid Miner yaitu :

1. Retrieve (dataset) operator retrieve digunakan untuk memanggil dataset yang sudah tersimpan di Repository RapidMiner.Fungsinya adalah mengambil data mentah sebelum dilakukan pemrosesan, pembagian data, pelatihan model, maupun evaluasi.
2. Split Data

   Operator Split Data digunakan untuk membagi dataset menjadi dua bagian:
   - Training data → untuk melatih model
   - Testing data → untuk menguji performa model

   Biasanya pembagian menggunakan rasio 70:30, 80:20, atau sesuai kebutuhan.
3. Algoritma KNN

   Operator k-NN adalah algoritma yang digunakan untuk membuat model klasifikasi berdasarkan kedekatan jarak antar data.

   Yang dilakukan operator ini:
   - Menghitung jarak data uji terhadap semua data latih (Euclidean/Manhattan)
   - Memilih k tetangga terdekat
   - Menentukan kelas berdasarkan mayoritas tetangga
4. Apply Model

Operator Apply Model digunakan untuk menerapkan model KNN yang sudah dibuat kepada data testing hasil dari Split Data.

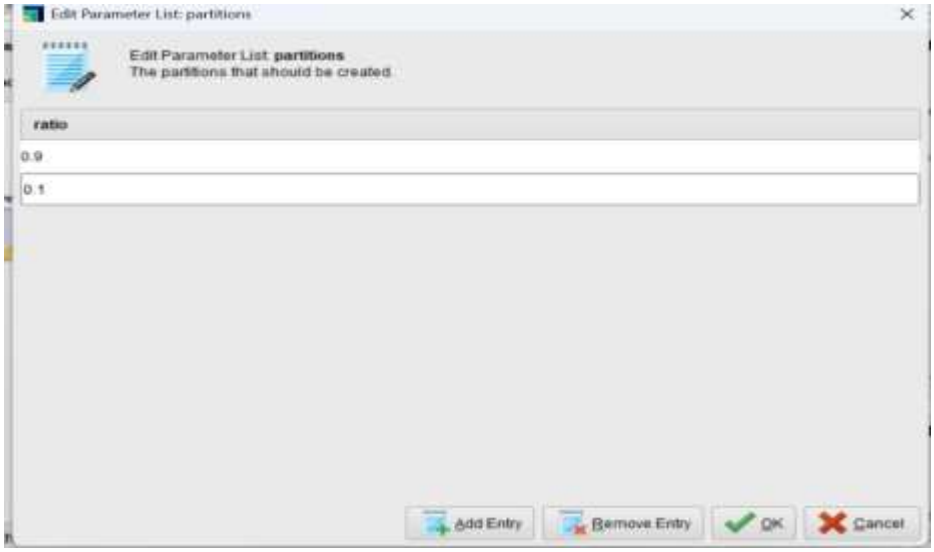
Yang dilakukan operator ini:

- Mengambil model dari operator KNN
- Menggunakan model tersebut untuk memprediksi label pada data testing
- Menghasilkan *hasil prediksi* (label baru)

Fungsi utama:

Memprediksi nilai output (label) untuk data yang belum pernah dilihat model Menghasilkan atribut prediction(label)

e-ISSN3032-601X & p-ISSN3032-7105



Gambar 3. Angka ratio

Pengaturan Split Data pada gambar tersebut menunjukkan bahwa dataset dibagi menjadi dua bagian dengan rasio 0.9 dan 0.1, di mana 90% data digunakan sebagai data training untuk melatih model KNN, sedangkan 10% sisanya digunakan sebagai data testing untuk menguji dan mengevaluasi performa model. Pembagian ini dilakukan agar model memperoleh cukup informasi dari data latih, namun tetap memiliki sebagian data yang tidak pernah dilihat sebelumnya untuk memastikan hasil prediksi lebih objektif dan akurat.

Row No.	stroke	prediction(s...	confidence(0)	confidence(1)	id	gender	age	hypertensio
1	1	0	0.593	0.407	9046	1	67	0
2	1	1	0.186	0.814	1665	0	79	1
3	1	1	0.174	0.826	51950	1	82	0
4	0	1	0.158	0.842	58251	0	66	0
5	0	0	0.621	0.379	72911	0	57	1
6	1	1	0.363	0.637	12095	0	61	0
7	1	0	0.810	0.190	27458	0	60	0
8	0	0	0.820	0.180	29217	0	65	1
9	0	1	0.428	0.572	11014	1	4	0
10	0	0	0.832	0.168	67318	1	58	1

Gambar 4. Hasil perhitungan dari operator exampleset apply model

Gambar di atas menampilkan hasil prediksi model K-Nearest Neighbor (KNN), di mana setiap baris menunjukkan perbandingan antara label sebenarnya (stroke) dan hasil prediksi model (prediction(stroke)), disertai nilai confidence(0) dan confidence(1) sebagai tingkat keyakinan model terhadap masing-masing kelas. Jika stroke = 1, berarti pasien terindikasi stroke, sedangkan 0 berarti tidak. Kolom prediction(stroke) menunjukkan apakah model memprediksi pasien mengalami stroke (1) atau tidak (0). Nilai confidence(0) menunjukkan tingkat kepercayaan model bahwa pasien *tidak* mengalami stroke,

e-ISSN3032-601X & p-ISSN3032-7105

sedangkan confidence(1) menunjukkan keyakinan model bahwa pasien *mengalami* stroke; model memilih kelas dengan nilai confidence tertinggi. Beberapa baris menunjukkan prediksi benar, misalnya baris ke-2 dan ke-3, sedangkan baris seperti nomor 1 dan 4 menunjukkan kesalahan prediksi karena confidence yang lebih tinggi berada pada kelas yang tidak sesuai dengan label sebenarnya. Tabel ini membantu mengevaluasi seberapa baik model dalam mengenali pola data berdasarkan atribut seperti gender, age, dan hypertension, sekaligus memperlihatkan area di mana model masih kurang akurat.

accuracy: 60.00%

	true 0	true 1	class precision
pred. 0	6	4	60.00%
pred. 1	4	6	60.00%
class recall	60.00%	60.00%	

Gambar 5. Hasil accuracy

Confusion matrix di atas menunjukkan kinerja model KNN setelah data dibagi menggunakan rasio 0.9 : 0.1, di mana 90% digunakan untuk pelatihan dan 10% digunakan untuk pengujian. Dengan pembagian data seperti ini, model memiliki lebih banyak contoh untuk belajar sehingga mampu mengenali pola umum pada data, namun tetap diuji pada sebagian kecil data untuk melihat kemampuan generalisasinya. Hasil pengujian menunjukkan bahwa model menghasilkan akurasi 60%, dengan 6 data *true 0* berhasil diprediksi benar sebagai *pred.0* dan 6 data *true 1* berhasil diprediksi benar sebagai *pred.1*. Namun masih terdapat masing-masing 4 kesalahan prediksi pada kedua kelas. Nilai precision dan recall untuk kedua kelas sama-sama 60%, yang berarti model hanya mampu mengenali pola secara moderat dan masih memerlukan peningkatan, misalnya melalui penyesuaian nilai *k*, normalisasi data, atau penyeimbangan kelas.

KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma K-Nearest Neighbor (KNN) mampu digunakan sebagai metode awal dalam memprediksi risiko stroke berdasarkan atribut pasien. Dengan rasio pembagian data 90% untuk pelatihan dan 10% untuk pengujian, model menghasilkan akurasi 60%, yang mengindikasikan bahwa pola data telah dipelajari namun belum sepenuhnya optimal dalam membedakan kelas pasien stroke dan non-stroke. Hasil ini menegaskan bahwa kualitas prediksi sangat dipengaruhi oleh karakteristik data, pemilihan nilai *k*, serta proses pra-pemrosesan seperti normalisasi dan penyeimbangan kelas. Oleh karena itu, penelitian lanjutan perlu mempertimbangkan peningkatan tahapan pra-proses dan eksplorasi parameter model agar performa klasifikasi dapat ditingkatkan secara lebih signifikan.

DAFTAR PUSTAKA

Ahad, A., Puspitasari, I., Zheng, J., Ullah, S., Ullah, F., Bakhsh, S. T., & Pires, I. M. (2025). *Machine Learning Stroke Prediction in Smart Healthcare : Integrating Fuzzy K-Nearest Neighbor and Artificial Neural Networks with Feature Selection Techniques*. https://doi.org/10.32604/cmc.2025.062605

Eldora, K., & Fernando, E. (2024). *Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction : A Machine Learning Approach*. *6*(1), 313–338. https://doi.org/10.51519/journalisi.v6i1.664

Hodges, C. B., & Barbour, M. (2025). *A 2025 Vision for Building Access to K-12 Online and Blended Learning in Pre-service Teacher Education*. *30*(2022), 201–216.

Hussain, I., Qureshi, M., Ismail, M., Iftikhar, H., Zywio, J., & L, J. L. (2024). *Heliyon Optimal features selection in the high dimensional data based on robust technique : Application to different health database*. *10*(September). https://doi.org/10.1016/j.heliyon.2024.e37241

Jakarta, M. M. (n.d.). *Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Stroke pada Rumah Sakit Pusat Otak Nasional Prof* . *26*(1), 144–153.

Khairi, Z., Yanti, R., Fitri, T. A., & Fatdha, E. (2025). *Optimasi Algoritma Knn Menggunakan Smote Untuk Prediksi Stroke*. 164–175. https://doi.org/10.33364/algoritma/v.22-2.2474

Maskuri, M. N., Sukerti, K., & Bhakti, R. M. H. (2022). *Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke*. *4*(1), 130–140.

Nopitasari, E. F., Putri, S., Alkadri, A., Wahid, R., Insani, S., & Barat, K. (2025). *Implementasi Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor revolusioner untuk deteksi dini penyakit (Bintang , 2024), dengan algoritma K-Nearest limitasi yang mengurangi kemampuan klasifikasinya (Maulana , 2024), termasuk sensitivitas prediksi stroke , beberapa kesenjangan kritikal masih perlu diatasi yang menciptakan urgensi*. *5*(September), 344–365.

Review, T. (2021). *Studying Stroke Thrombus Composition After Thrombectomy*. *November*, 3718–3727. https://doi.org/10.1161/STROKEAHA.121.034289

Roman, M., Naz, I., Luqman, M. A., Ali, J., & Jan, M. S. (2024). *Stroke Disease Prediction Using K-Nearest Neighbor And Decision Tree Algorithms With Machine Learning Pre-Processing Techniques*. *4*, 2015–2027.

Saifaldeen, D. A., & Member, A. M. A. (2024). DRL-Based IRS-Assisted Secure Hybrid Visible Light and mmWave Communications. *IEEE Open Journal of the Communications Society*, *5*(February), 3007–3020. https://doi.org/10.1109/OJCOMS.2024.3395425

Setiawan, A., Waleska, R. F., & Purnama, M. A. (2024). *KOMPARASI ALGORITMA K-NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), DAN DECISION TREE DALAM KLASIFIKASI PENYAKIT STROKE*. *7*(1), 107–114.

Shourov, S. H. (2024). *COMPARATIVE ANALYSIS OF NEURAL NETWORK ARCHITECTURES FOR MEDICAL IMAGE CLASSIFICATION : EVALUATING PERFORMANCE*. *04*(01), 1–42. https://doi.org/10.63125/feed1x52

Siregar, A. H. (2023). *Penerapan K-Nearest Neighbors (KNN) dalam Memprediksi dan Menghitung Akurasi Data Penyakit Stroke*. *2*(4).

Yoon, E., Member, S., & Kwon, S. (2023). An Efficient Application of Turbo Coding for OFDM In-Phase / Quadrature Index Modulation. *IEEE Access*, *11*(April), 37031–37040. https://doi.org/10.1109/ACCESS.2023.3266288