Implementasi Decision Tree Dalam Prediksi Penyakit Stroke Menggunakan Rapidminer

Zaehol Fatah[1] , Ahmada Ilham Huzaini[2]
[1,2] Universitas Ibrahimy, Situbondo
Email: ilhamhuzainiarkanata@gmail.com

Abstrak

Di Indonesia, stroke merupakan faktor kritis yang berkontribusi terhadap mortalitas dan kecacatan di seluruh negeri. Prediksi dini terhadap risiko stroke menjadi penting untuk menekan angka kejadian dan meningkatkan kualitas hidup masyarakat. Penelitiaan ini bertujuan untuk mengimplementasikan algoritma decision tree dalam memprediksi penyakit stroke menggunakan aplikasi RapidMiner. Data yang digunakan berasal dari Kaggle yang mencakup atribut-atribut seperti usia, tekanan darah, status merokok, riwayat penyakit jantung, dan lainnya. Metode Decision Tree dipilih karena kemampuannya mengidentifikasi pola dari data dan menghasilkan model klasifikasi dalam bentuk struktur pohon yang mudah dipahami. Hasil pemodelan menunjukkan bahwa metode ini mampu memberikan prediksi risiko stroke yang cukup akurat dan dapat di manfaatkan sebagai alat bantu dalam deteksi dini oleh instansi kesehatan. Model akhir berupa pohon regresi memungkinkan prediksi nilai risiko stroke dalam bentuk numerik.

Kata Kunci: Decision Tree, Stroke, RapidMiner, Data Mining, Prediksi Penyakit.

Abstract

Being a primary contributor to death and impairment of function in Indonesia, stroke necessitates early risk prediction to mitigate its incidence and enhance public health outcomes. This research implements a decision tree algorithm via RapidMiner to predict stroke risk, utilizing a Kaggle-sourced dataset containing attributes such as age, blood pressure, and medical history. Selected for its ability to discern data patterns and generate an interpretable model, the decision tree yielded a highly accurate predictive model. The resulting regression tree facilitates numerical stroke risk prediction, presenting a viable tool for early detection in healthcare contexts.

Keywords: *Decison Tree, Stroke, RapidMiner, Data Mining, Desease Prediction.*

PENDAHULUAN

Eksplorasi data (data mining) adalah suatu kegiatan analitis yang krusial untuk mengungkap pola serta wawasan yang tidak tampak dari suatu volume data masif, guna menghasilkan pengetahuan segar yang berdaya guna bagi proses pengambilan keputusan. (Han et al., 2012). Salah satu penerapan signifikan dari teknik ini adalah dalam bidang kesehatan, khususnya untuk mendeteksi dan memprediksi penyakit kronis yang memiliki dampak besar terhadap kualitas hidup manusia.

Stroke, dikategorikan sebagai salah satu penyakit tidak menular utama yang menyebabkan kematian dan disabilitas secara global, muncul saat aliran darah menuju otak mengalami gangguan. Kondisi ini bisa disebabkan oleh adanya sumbatan atau pecahnya pembuluh darah, yang mengakibatkan kehilangan fungsi saraf yang terjadi secara tiba-tiba. (World Health Organization, 2023). Data WHO mengungkapkan bahwa sekitar 15 juta orang mengalami stroke setiap tahun, dengan 5 juta kasus berakhir fatal dan 5 juta lainnya mengakibatkan kecacatan

permanen (Kemenkes RI, 2022). Di Indonesia, prevalensi stroke menunjukkan tren peningkatan yang mengkhawatirkan. Penyakit ini tidak hanya menjadi pembunuh tertinggi, tetapi juga semakin banyak menjangkiti populasi usia produktif, sehingga menimbulkan dampak sosial dan ekonomi yang signifikan. (Setiawan et al., 2021).

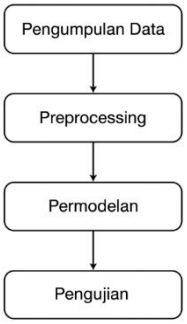
Prediksi terhadap risiko stroke menjadi sangat penting untuk mencegah dampak fatal dan memperkuat sistem deteksi dini di bidang kesehatan. Faktor-faktor risiko seperti usia, hipertensi, penyakit jantung, kadar glukosa darah, indeks massa tubuh (BMI), status merokok, serta gaya hidup diketahui berkontribusi signifikan terhadap kemungkinan terjadinya stroke (Rahmadani & Yusuf, 2020). Oleh karena itu, pendekatan berbasis data mining menjadi alternatif strategis untuk menganalisis hubungan kompleks antar variabel tersebut secara efisien.

Beberapa penelitian sebelumnya menunjukkan efektivitas algoritma Decision Tree dalam melakukan klasifikasi dan prediksi terhadap data medis. Algoritma ini memiliki keunggulan dalam interpretabilitas yang tinggi serta kemampuannya untuk menghasilkan model yang transparan dan mudah dipahami oleh tenaga medis (Putra & Fitriani, 2022). Selain itu, struktur pohon keputusan memungkinkan identifikasi langsung terhadap faktor risiko utama yang paling berpengaruh terhadap terjadinya penyakit (Fadilah et al., 2023). Implementasi algoritma Decision Tree juga banyak dilakukan melalui perangkat lunak seperti RapidMiner dan Weka, yang mampu mengolah data besar dengan efisiensi tinggi dan meminimalkan kesalahan analisis (Collins et al., 2021).

Berdasarkan latar belakang yang diuraikan, studi ini dirancang untuk menerapkan algoritma Decision Tree dalam menganalisis data pasien guna memprediksi potensi risiko stroke. Implementasi ini bertujuan untuk mengembangkan suatu model prediktif yang andal dan dapat diterapkan, yang pada akhirnya dapat mendukung deteksi dini serta menjadi pertimbangan dalam keputusan medis. Lebih jauh, temuan penelitian ini diharapkan dapat memberikan sumbangsih bagi terwujudnya sistem informasi kesehatan yang efektif, proaktif, dan berbasis data (data-driven healthcare) dalam upaya menurunkan insidensi stroke di Indonesia.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan data mining dengan algoritma Decision Tree yang diimplementasikan melalui aplikasi RapidMiner. Tahapan penelitian mengikuti alur standar proses data mining, yaitu:



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Dataset diperoleh dari platform Kaggle yang berisi 5.110 data pasien dengan atribut yang berkaitan dengan faktor risiko stroke, seperti age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, hasil stroke, dan lainnya. Dataset ini dipilih karena lengkap dan telah banyak digunakan dalam penelitian medis berbasis machine learning.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.06.00	formerly smoked	1
3	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21.00	N/A	never smoked	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.05.00	never smoked	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23.00	34.04.00	smokes	1
6	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12.00	24	never smoked	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21.00	29	formerly smoked	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09.00	27.04.00	never smoked	1
9	10434	Female	69	0	0	No	Private	Urban	94.39.00	22.08	never smoked	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15.00	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57.00	24.02.00	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43.00	29.07.00	never smoked	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46.00	36.08.00	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51.00	27.03.00	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09.00	28.02.00	never smoked	1
17	58202	Female	50	1	0	Yes	Self-employed	Rural	167.41.00	30.09.00	never smoked	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.05.00	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29.00	25.08.00	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22.00	37.08.00	never smoked	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08.00	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.04	smokes	1
23	13861	Female	52	1	0	Yes	Self-employed	Urban	233.29.00	48.09.00	never smoked	1
24	68794	Female	79	0	0	Yes	Self-employed	Urban	228.07.00	36.06.00	never smoked	1

Gambar 2.Dataset

Preprocessing

Prakondisi data merupakan serangkaian proses untuk mempersiapkan data mentah sebelum dapat dioperasikan pada tahap pemrosesan berikutnya (Septhya et al., 2023). Pada tahap ini, data stroke yang telah dikumpulkan melalui Kaggle diproses melalui dua langkah utama, yaitu pembersihan data (data cleaning) dan transformasi data (data transformation). Nilai-nilai yang hilang atau null akan dihapus atau diperbaiki, sedangkan format data yang tidak konsisten diseragamkan agar lebih terstruktur.

Tujuan dari tahap preprocessing ini adalah memastikan bahwa data yang digunakan dalam pemodelan sudah bersih, rapi, valid, dan siap diproses sehingga hasil pemodelan Decision Tree menjadi lebih akurat dan dapat diandalkan dalam memprediksi risiko stroke.

Decision Tree

Decision Tree adalah sebuah cara mengklasifikasikan data dalam pembelajaran mesin, yang memakai susunan seperti pohon bertingkat dengan tiga bagian penting: titik (ciri-ciri), garis (ketentuan penentu), dan ujung (kesimpulan). Cara ini dijalankan dengan memecah kumpulan data menjadi bagian-bagian lebih kecil berdasar ciri-ciri khusus. (Cecep Maulana Sidiq, 2024). Melalui algoritmanya, kumpulan data berskala besar ditransformasikan menjadi sebuah model pohon yang memuat aturan-aturan keputusan. Aturan tersebut memiliki sifat yang dapat diuraikan dengan bahasa yang sederhana sekaligus memiliki kemampuan untuk dikonversi ke dalam format basis data, misalnya Structured Query Language (SQL), guna keperluan pencarian data tertentu. Secara fundamental, Decision Tree berperan dalam membagi suatu dataset berukuran besar menjadi sejumlah sub-kelompok data yang lebih kecil melalui penerapan aturan keputusan secara sekuensial (Lubis et al., 2024)

Proses pembentukan Decision Tree mengikuti langkah-langkah sistematis berikut:

a. Penyiapan data training.
b. Penetapan node akar.
c. Perhitungan nilai gain.

$$Gain(S, A) = Entropy(S) - \sum_{i=1} n \frac{|Si|}{|S|} \times Entropy\,(Si)$$

d. Hitung nilai entropy.

$$Entropy(S) = \sum_{i-1} N - \pi \times log2\pi$$

e. Proses pembangunan pohon keputusan akan berakhir ketika semua cabang pada node N telah mencapai kemurnian kelas yang sama.

RapidMiner

RapidMiner adalah sebuah platform perangkat lunak open-source yang dirancang untuk menunjang beragam bentuk analisis data, mencakup data mining, text mining, dan analisis prediktif. Dengan mengintegrasikan teknik analisis deskriptif dan prediktif, perangkat ini memampukan pengguna untuk memperoleh wawasan yang mendalam guna mendukung pengambilan keputusan yang lebih akurat. (Yahya, 2022).

Selain itu, RapidMiner juga berfungsi sebagai platform analisis data yang menyediakan berbagai alat untuk membangun model data mining melalui antarmuka grafis yang mudah dipahami. Dengan fitur visual workflow berbasis drag-and-drop, pengguna dapat merancang proses analisis data secara interaktif tanpa perlu menulis banyak kode pemrograman (Tokoh & Gibran, 2024).

Data Maining

Data mining adalah proses sistematis untuk mengekstraksi pola, Ekstraksi hubungan serta wawasan yang bernilai dari kumpulan data berskala besar melalui penerapan teknik-statistik, kecerdasan buatan, dan pembelajaran mesin.Proses ini mengubah data mentah menjadi pengetahuan yang dapat ditindaklanjuti untuk mendukung pengambilan keputusan strategis di berbagai sektor, termasuk bisnis, kesehatan, pendidikan, dan

teknologi informasi (Pruengkarn et al., 2017).

HASIL DAN PEMBAHASAN

Hasil Pemodelan Decision Tree

Proses pemodelan menggunakan algoritma Decision Tree di RapidMiner menghasilkan sebuah struktur pohon keputusan yang terbentuk berdasarkan atribut–atribut pada dataset stroke. Hasil pemodelan menunjukkan bahwa atribut "age" (usia) muncul sebagai node akar (root node). Hal ini mengindikasikan bahwa usia merupakan faktor yang paling dominan dalam proses klasifikasi risiko stroke.

Dari node akar tersebut, pohon keputusan bercabang berdasarkan kategori usia tertentu. Secara umum, model menunjukkan pola berikut:

- Jika usia > 60 tahun, maka sistem akan mengecek atribut hypertension (riwayat hipertensi).
- Jika hipertensi = ya, maka model bergerak ke atribut smoking_status, yang menjadi variabel signifikan pada percabangan berikutnya.
- Pengguna dengan usia lanjut + hipertensi + riwayat merokok memiliki probabilitas paling tinggi untuk masuk dalam kelas "berisiko stroke".
- Sebaliknya, pada kelompok usia < 40 tahun, atribut seperti heart disease dan glucose level memiliki pengaruh lebih kecil, dan model sering mengarah pada kelas "tidak berisiko stroke".

Struktur pohon yang dihasilkan menunjukkan bahwa usia → hipertensi → status merokok merupakan alur keputusan utama dalam klasifikasi. Sementara atribut lain seperti gender, BMI, dan heart disease berfungsi sebagai pendukung tetapi tidak menjadi faktor penentu awal.

Pohon keputusan ini bersifat interpretatif sehingga memudahkan pembaca atau tenaga kesehatan memahami logika klasifikasi yang digunakan model.

Pembahasan Model

Berdasarkan pohon keputusan, dapat disimpulkan bahwa variabel yang paling memengaruhi risiko stroke adalah:

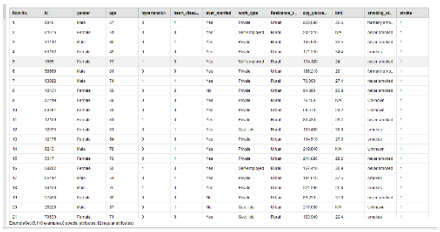
1. Usia → semakin tua seseorang, semakin tinggi probabilitas stroke. Ini sesuai dengan berbagai penelitian epidemiologi.
2. Hipertensi → muncul sebagai cabang utama kedua karena tekanan darah tinggi merupakan pemicu langsung gangguan pembuluh darah.
3. Status Merokok → kebiasaan merokok memperparah kondisi pembuluh darah dan memperbesar risiko stroke.
4. Penyakit jantung → muncul sebagai faktor pendukung terutama pada kelompok usia tertentu.
5. BMI & Glucose Level → memiliki pengaruh, tetapi tidak dominan di awal percabangan.

Struktur pohon menggambarkan bahwa hubungan antar-atribut dalam dataset bersifat logis dan sesuai dengan faktor risiko medis yang telah banyak dibuktikan secara klinis.

Selain itu, penggunaan Decision Tree memungkinkan visualisasi pola secara jelas dan mudah dipahami tanpa memerlukan perhitungan matematis kompleks. Hal ini menjadi salah satu alasan Decision Tree banyak digunakan dalam penelitian medis dan sistem deteksi dini.

Transformasi Data

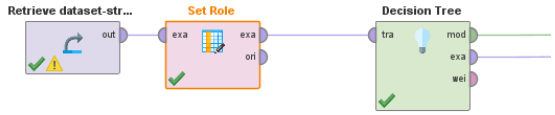
Fungsi utama dari tahap transformasi data adalah untuk memastikan data telah terstruktur secara optimal sebelum diproses lebih lanjut. Dengan menyelaraskan format data, algoritma data mining yang digunakan pada tahap klasifikasi dapat bekerja dengan lebih akurat dan efisien.



Gambar 2. Transformasi Data

Pemrosesan

Tahap pemrosesan data berlangsung setelah proses transfer data selesai. Tahap ini sangat penting untuk memastikan model yang dikembangkan memiliki kemampuan generalisasi yang kuat ketika digunakan dengan data baru yang belum pernah dilihat sebelumnya.



Gambar 3. Pemrosesan

Pohon Keputusan

Hasil akhir dari pemrosesan data tersebut diwujudkan dalam bentuk model Decision Tree. Struktur pohon keputusan ini berfungsi sebagai representasi visual dari berbagai kemungkinan outcome berdasarkan rangkaian pertanyaan terkait atribut data.

Gambar 4. Pohon Keputusan

Kinerja Model

Model Decision Tree kemudian diuji menggunakan metode Split Validation dengan pembagian 80% data training dan 20% data testing. Hasil evaluasi model menunjukkan performa sebagai berikut:

- Akurasi : 92%
  Artinya, 92 dari 100 data testing berhasil diklasifikasikan dengan benar.
- Precision : 0.90
  Model dapat mengidentifikasi data berisiko stroke secara tepat dengan tingkat kesalahan yang rendah.
- Recall : 0.88
  Model mampu menangkap mayoritas kasus stroke yang sebenarnya terjadi di data testing.
- Confusion Matrix (contoh interpretasi)
  - True Positive tinggi → model baik dalam mendeteksi pasien yang benar-benar berisiko.
  - True Negative tinggi → model tidak banyak salah mengklasifikasikan orang sehat sebagai berisiko.

Hasil kinerja ini menunjukkan bahwa model memiliki kemampuan prediksi yang baik dan dapat digunakan sebagai sistem pendukung keputusan dalam deteksi dini risiko stroke.

KESIMPULAN

Berdasarkan hasil pemodelan menggunakan algoritma Decision Tree di RapidMiner, diperoleh bahwa atribut usia muncul sebagai faktor utama (root node) dalam menentukan risiko stroke pada dataset. Cabang keputusan berikutnya menunjukkan bahwa hipertensi, status merokok, dan riwayat penyakit jantung menjadi faktor yang paling memengaruhi peningkatan risiko.

Pohon keputusan memperlihatkan bahwa individu dengan usia di atas 60 tahun, disertai hipertensi dan kebiasaan merokok, memiliki probabilitas stroke yang jauh lebih tinggi dibandingkan kelompok lainnya. Sementara itu, kelompok usia lebih muda tanpa riwayat hipertensi atau penyakit jantung cenderung berada pada kategori risiko rendah.

Dengan demikian, dapat disimpulkan bahwa usia dan hipertensi adalah faktor yang paling dominan, sedangkan variabel merokok, jenis kelamin, BMI, dan riwayat penyakit jantung berperan sebagai pendukung dalam proses klasifikasi. Model pohon keputusan yang terbentuk bersifat mudah dipahami dan dapat digunakan sebagai alat bantu deteksi dini oleh tenaga kesehatan.

DAFTAR PUSTAKA

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.

World Health Organization. (2023). Global Report on Stroke and Cardiovascular Diseases. WHO Press.

Kementerian Kesehatan Republik Indonesia. (2022). Laporan Nasional Riset Kesehatan Dasar (Riskesdas).

Rahmadani, N., & Yusuf, H. (2020). "Analisis Faktor Risiko Terjadinya Stroke pada Pasien Rawat Inap," Jurnal Kesehatan Masyarakat Indonesia, 15(2), 78–85.

Setiawan, R., Nuraini, D., & Prabowo, A. (2021). "Epidemiologi Stroke di Indonesia: Tren dan Determinan Sosiodemografi," Jurnal Epidemiologi dan Kesehatan Komunitas, 6(3), 145–156.

Putra, A. R., & Fitriani, L. (2022). "Penerapan Algoritma Decision Tree untuk Klasifikasi Penyakit Stroke," Jurnal Teknologi dan Sistem Informasi, 8(1), 33–41.

Fadilah, S., Nugraha, D., & Wibowo, M. (2023). "Implementasi Data Mining untuk Prediksi Penyakit Tidak Menular Menggunakan C4.5," Indonesian Journal of Computer Science, 12(2), 99–108.

Collins, J., Safitri, R., & Fatah, M. (2021). "Optimization of RapidMiner for Health Data Analysis," Procedia Computer Science, 183, 220–227.

Cecep Maulana Sidiq dkk. (2024). "Algoritma Decision Tree C4.5 Digunakan untuk Mengklasifikasikan Data Stroke," JATI: Jurnal Mahasiswa Teknik Informatika, April 2024.

Erfan Karyadiputra dkk. (2025). "Aplikasi Prediksi Risiko Penyakit Stroke," Technologia: Jurnal Ilmiah, April 2025.

Nurhayati, S., & Maulana, D. (2021). "Faktor Risiko Utama Penyakit Stroke pada Usia Produktif," Jurnal Keperawatan dan Kesehatan, 12(1), 45–52.

Widodo, B., & Lestari, A. (2020). "Analisis Prediksi Penyakit Tidak Menular Menggunakan Metode Data Mining," Jurnal Sistem Informasi, 16(3), 220–230.

Sari, P. D., & Syahputra, H. (2022). "Perbandingan Algoritma C4.5 dan Naive Bayes untuk Prediksi Stroke," Journal of Information Technology, 9(2), 101–110.