J-CEKI : Jurnal Cendekia Ilmiah
Vol.4, No.2, Februari 2025

Implementasi Algoritma C4.5 pada Sistem Prediksi Stroke Berdasarkan Data Kesehatan

Alpito Gilang Ramadhan[1], Hiroki Setiawan Hidayatullah[2], Rizal Ruba'i[3]
Prodi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
E-mail: alpitogilang07@gmail.com[1], kurocky1234@gmail.com[2], rizal354r@gmail.com[3]

Article History:
Received: 21 Desember 2024
Revised: 17 Januari 2025
Accepted: 22 Januari 2025

Keywords: *Stroke, Prediction, C4.5 Algorithm, Decision Tree, Early Detection.*

Abstract: *Stroke is one of the leading causes of death and disability worldwide. Early detection of the risk of stroke is crucial to reducing its impact and improving the effectiveness of treatment. This study aims to predict the probability of a person being at risk of stroke using the C4.5 Algorithm, known as a decision tree-based method. The dataset used includes variables such as age, blood pressure, blood sugar levels, medical history, smoking habits, and physical activity. After data preprocessing, including handling missing data and normalization, a predictive model was built using the C4.5 Algorithm. The results show that the model achieved an accuracy of 95.01%, a precision of 33.33%, a recall of 2.04%, and an F1-score of 3.81%. Additionally, the Negative Predictive Value (NPV) of 95.10% and specificity of 99.79% demonstrate the model's ability to detect risks effectively in negative cases. The model also generates decision rules that are easily interpretable by medical professionals, making it a potential tool to support medical decision-making in the early identification of stroke risks.*

PENDAHULUAN

Kesehatan adalah aspek penting yang perlu mendapat perhatian. Jika seseorang mengabaikan kesehatannya, ia akan lebih rentan terkena berbagai jenis penyakit. Cara terbaik untuk menghindari hal ini adalah dengan melakukan pencegahan sejak dini. Penyakit memiliki beragam jenis dengan dampak yang bervariasi, termasuk yang dapat mengancam jiwa. Salah satu penyakit yang berpotensi menyebabkan kematian adalah stroke (Rivliansyah et al., 2022), atau penyakit yang menyerang pembuluh darah di dalam otak (Abidin et al., 2024).

Stroke merupakan kondisi medis yang terjadi ketika aliran darah menuju otak terhambat atau terputus, sehingga menyebabkan kerusakan pada sel-sel otak (Maulana Sidiq et al., 2024). Stroke merupakan salah satu masalah kesehatan utama dalam masyarakat modern dan semakin menjadi perhatian serius di berbagai belahan dunia. Hal ini disebabkan oleh sifat stroke yang terjadi secara mendadak dan dapat menyebabkan kematian, kecacatan fisik, serta gangguan mental, baik pada usia produktif maupun lanjut. Stroke termasuk dalam daftar 10 penyakit paling mematikan.

Menurut laporan World Health Organization (2024), stroke membawa risiko kematian yang tinggi dan dapat menyebabkan berbagai komplikasi, seperti kehilangan kemampuan berbicara

atau melihat, kelumpuhan, serta kebingungan (Alfiyan et al., 2024). Penyakit ini dinamakan stroke karena cara serangannya yang tiba-tiba. Selain itu, risiko serangan stroke kedua meningkat secara signifikan pada penyintas yang pernah mengalaminya sebelumnya.

Untuk mempercepat proses diagnosis pada penderita stroke, diperlukan algoritma Prediksi yang mampu mengolah data dalam jumlah besar secara cepat dan akurat. Di antara berbagai algoritma Prediksi yang tersedia, algoritma Decision Tree C4.5 dianggap paling sesuai untuk penelitian ini. Hal ini didasarkan pada keunggulan algoritma tersebut dalam kecepatan dan ketepatan proses Prediksi, bahkan ketika digunakan pada dataset dengan volume yang besar.

Menurut Firman Akbar et al., dalam jurnal "Implementation of Decision Tree Algorithm C4.5 and Support Vector Regression for Stroke Disease Prediction", penelitian ini bertujuan untuk membandingkan kinerja algoritma Decision Tree C4.5 dan Support Vector Regression (SVR) dalam memprediksi penyakit stroke menggunakan dataset pasien stroke dari Kaggle yang berjumlah 5110 data dengan pembagian data latih dan uji sebesar 70:30.

Hasil penelitian menunjukkan bahwa algoritma Decision Tree C4.5 memiliki tingkat kesalahan yang lebih rendah dibandingkan dengan algoritma SVR. Nilai root mean square error (RMSE) untuk Decision Tree C4.5 adalah 0,235, sedangkan SVR menghasilkan RMSE sebesar 0,399. Dengan demikian, Decision Tree C4.5 lebih akurat dan lebih efektif digunakan dalam prediksi penyakit stroke dibandingkan dengan SVR, terutama karena hasilnya mendekati nol yang menunjukkan tingkat prediksi yang baik.

1. Keakuratan Analisis: Analisis dalam penelitian ini cukup akurat karena: Metodologi: Penelitian menggunakan dataset yang besar (5110 data) yang valid dan relevan untuk kasus stroke.
2. Evaluasi Kinerja: Membandingkan dua algoritma dengan parameter evaluasi yang jelas seperti RMSE.
3. Penggunaan Algoritma: Algoritma Decision Tree C4.5 yang digunakan dikenal andal untuk analisis berbasis data kategoris dan regresi.

Namun, keterbatasan penelitian dapat muncul jika algoritma ini diterapkan pada dataset yang lebih kompleks atau tidak terstruktur, yang memerlukan eksplorasi lebih lanjut untuk validitas hasil.

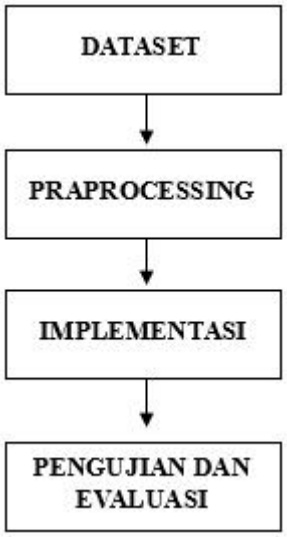
Menurut Randi Estian Pambudi et al., dalam jurnal "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C4.5", penelitian ini bertujuan untuk mengklasifikasikan penyakit stroke menggunakan algoritma Decision Tree C4.5. Dataset yang digunakan berasal dari repositori Kaggle dengan 5110 data pasien, di mana atribut seperti usia, jenis kelamin, kadar glukosa, hipertensi, dan status merokok dianalisis untuk menentukan faktor-faktor yang paling signifikan dalam memprediksi stroke.

Hasil penelitian menunjukkan bahwa algoritma Decision Tree C4.5 menghasilkan tingkat akurasi sebesar 96,05%. Algoritma ini juga menghasilkan model berupa pohon keputusan dengan 23 aturan klasifikasi yang teridentifikasi dari 11 atribut awal, yang kemudian diseleksi menjadi 8 atribut utama berdasarkan pengaruhnya terhadap prediksi stroke. Selain itu, hasil analisis menggunakan kurva ROC menunjukkan nilai AUC sebesar 0,755, yang mengindikasikan performa yang baik dalam klasifikasi.

METODE PENELITIAN

Desain penelitian adalah langkah-langkah atau proses yang dirancang untuk merencanakan sebuah penelitian. Tujuan dari desain ini adalah untuk membantu peneliti dalam melaksanakan

penelitian secara lebih terstruktur. Tahapan-tahapan dalam penelitian ini dapat dilihat pada Gambar



Gambar 1. Metode Penelitian

Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining untuk memprediksi kemungkinan stroke berdasarkan dataset kesehatan. Penelitian kuantitatif digunakan untuk menganalisa masalah berdasarkan pengujian sebuah teori yang terdiri dari berbagai variabel dan data berupa sesuatu yang dapat dihitung (Abadi et al., 2024). Data ini dianalisis menggunakan algoritma C4.5 untuk membangun model prediksi yang terstruktur dalam bentuk pohon keputusan (Saputra et al., 2020).

Dataset

Dataset adalah sekumpulan data yang dapat digunakan sebagai bahan percobaan riset (Yuliska & Syaliman, 2020). Dataset yang digunakan bersumber dari Kaggle dan mencakup variabel-variabel penting seperti ***gender***, ***age***, ***hypertension***, ***heart_disease***, ***ever_married***, ***work_type***, ***Residence_type***, ***avg_glucose_level***, ***bmi***, ***smoking_status***, dan ***stroke*** sebagai label target. Dataset ini dipilih karena menyediakan data yang relevan untuk analisis risiko stroke.

Preprocessing Data

Tahapan preprocessing dilakukan untuk memastikan data yang digunakan berkualitas dan siap untuk diolah. Langkah-langkahnya meliputi pembersihan data untuk mengatasi data yang hilang atau tidak relevan, transformasi variabel kategorikal menggunakan metode encoding, serta pembagian dataset menjadi 80% data latih dan 20% data uji untuk memastikan model dapat diuji dengan baik.

Algoritma C4.5

Salah satu pendekatan yang bisa digunakan untuk menangani masalah dengan kategori atribut dalam klasifikasi data adalah algoritma C4.5. Algoritma ini digunakan untuk membangun pohon keputusan yang berisi aturan-aturan untuk proses klasifikasi. Dalam penerapannya, algoritma C4.5 membentuk pohon keputusan dengan langkah-langkah utama: memilih atribut sebagai akar, membuat cabang untuk setiap nilai atribut, membagi kasus pada setiap cabang, dan

mengulangi proses tersebut hingga semua kasus pada setiap cabang memiliki kelas yang sama (Gusrialni Fitri et al., 2023). *Decision Tree* merupakan algoritma pengklasifikasian dengan struktur sederhana serta mudah diinterpretasikan (Abidin et al., 2024). Struktur *Decision Tree* terdiri dari node, cabang, dan daun, Dimana node mewakili atribut, banag mewakili aturan keputusan, dan daun mewakili hasil Keputusan (Maulana Sidiq et al., 2024).

Implementasi Algoritma C4.5

Pada tahap ini, algoritma C4.5 diterapkan untuk membangun model prediksi karena dinilai lebih efektif (Alfiyan et al., 2024). Proses diawali dengan perhitungan gain informasi untuk menentukan atribut yang paling berpengaruh dalam memprediksi stroke. Atribut ini menjadi dasar pembentukan pohon keputusan, yang memetakan hubungan antar variabel secara hierarkis.

Pengujian dan Evaluasi

Model yang dihasilkan diuji menggunakan data uji untuk mengukur performanya. Evaluasi dilakukan menggunakan metrik seperti akurasi, *precision, recall, F1-score, Negative Predictive Value* (NPV), dan *Specificity*. Metrik-metrik ini membantu menilai seberapa efektif algoritma C4.5 dalam memprediksi risiko stroke lebih komprehensif (Nabila et al., 2023).

Tools dan Software

Data mining ialah proses identifikasi hubungan, pola, maupun perkembangan data besar dengan memanfaatkan teknologi pengenalan pola serta metode statistik ataupun matematika (Pradhana et al., 2024). Penelitian ini dilakukan menggunakan RapidMiner, sebuah perangkat lunak data science yang mempermudah proses pengolahan data, implementasi algoritma, dan evaluasi hasil. RapidMiner dipilih karena kemampuannya yang *user-friendly* serta mendukung implementasi algoritma C4.5 secara efisien. RapidMiner merupakan alat mining data yang berkembang dengan menggunakan bahasa pemrograman Java (Nadira & Farida Utami, 2024).

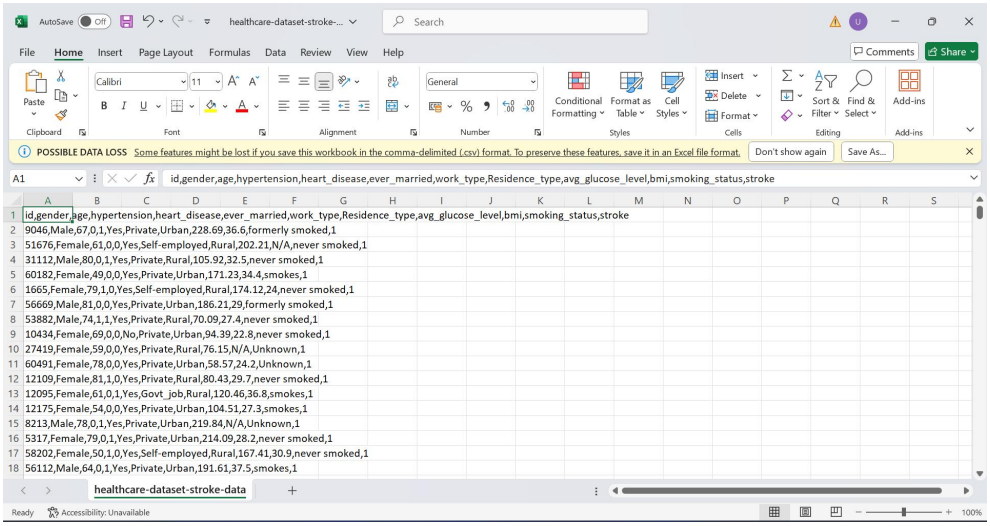
HASIL DAN PEMBAHASAN

Dataset (Pengumpulan Data)

Data yang digunakan dalam penelitian ini diperoleh dari Kaggle, yang menyediakan Informasi terkait faktor risiko stroke.

Dataset : *https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset*

Dataset ini terdiri dari beberapa variabel, di antaranya ***id, gender***, ***age***, ***hypertension***, ***heart_disease***, ***ever_married***, ***work_type***, ***Residence_type***, ***avg_glucose_level***, ***bmi***, ***smoking_status***, dan ***stroke*** sebagai variabel target. Dataset ini dipilih karena relevansinya dalam mendukung analisis prediksi stroke menggunakan algoritma C4.5.



Gambar 2. Tampilan Dataset Analisa Prediksi Stroke

Preprocessing Data

Pada tahap *preprocessing*, beberapa langkah penting dilakukan untuk mempersiapkan data yang digunakan dalam pemodelan. Pertama, kami menghapus variable **id** karena kami nilai kurang relevan. Kemudian, satu record data dengan nilai gender "*Other*" dihapus karena dianggap tidak relevan dengan tujuan penelitian. Selanjutnya, variabel ***Stroke*** diubah menjadi label untuk memudahkan proses klasifikasi. Beberapa variabel lainnya juga diubah menjadi tipe ***Binominal***, seperti ***Stroke***, ***Gender***, ***Hypertension***, ***Heart_disease***, ***Ever_married***, dan ***Residence_type***. Selain itu, data dibagi menjadi dua bagian: 80% digunakan untuk pelatihan model (*training set*) dan 20% sisanya digunakan untuk pengujian model (*testing set*). Tahap ini memastikan bahwa data yang digunakan untuk pelatihan dan pengujian dalam kondisi yang sesuai.

Row No.	stroke	gender	age	hypertension	heart_disea...	ever_married	work_type	Residenc
1	1	Male	67	0	1	Yes	Private	Urban
2	1	Female	61	0	0	Yes	Self-employed	Rural
3	1	Male	80	0	1	Yes	Private	Rural
4	1	Female	49	0	0	Yes	Private	Urban
5	1	Female	79	1	0	Yes	Self-employed	Rural
6	1	Male	81	0	0	Yes	Private	Urban
7	1	Male	74	1	1	Yes	Private	Rural
8	1	Female	69	0	0	No	Private	Urban
9	1	Female	59	0	0	Yes	Private	Rural
10	1	Female	78	0	0	Yes	Private	Urban
11	1	Female	81	1	0	Yes	Private	Rural

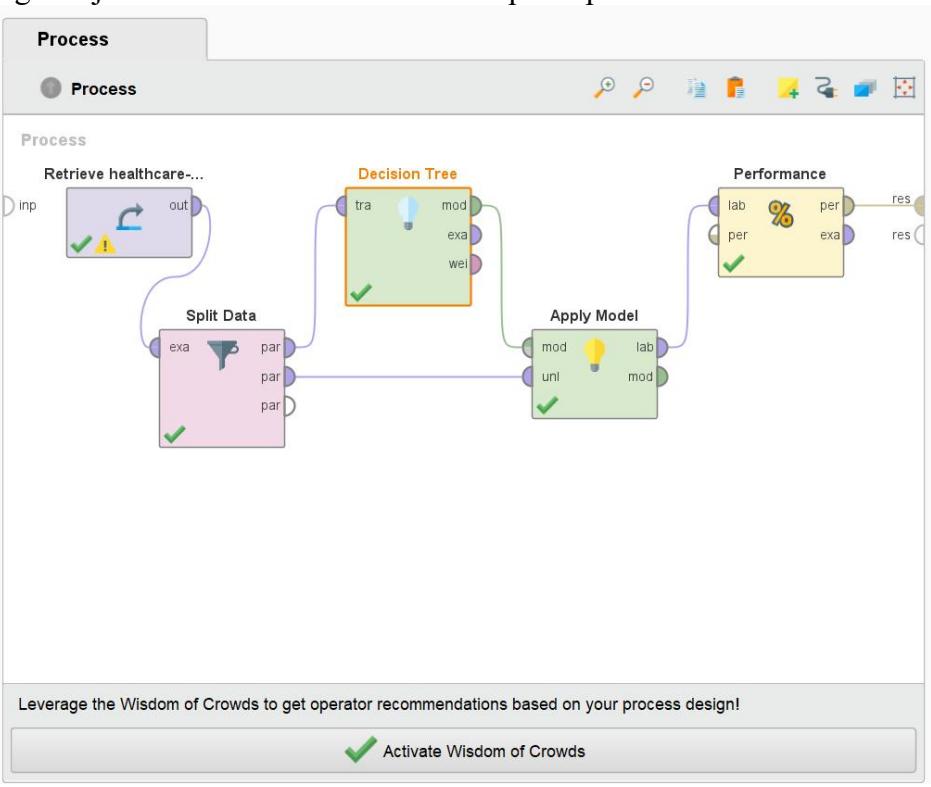
ExampleSet (5,109 examples,1 special attribute,10 regular attributes)

Gambar 3. Preprocessing Data dengan Dataset yang sudah dipersiapkan

Implementasi

Algoritma C4.5 diterapkan pada data yang telah diproses untuk membangun model prediksi stroke. Proses ini menggunakan software RapidMiner untuk membangun pohon keputusan yang akan digunakan untuk memprediksi apakah seorang individu akan mengalami stroke atau tidak,

berdasarkan data yang tersedia. Pada tahap ini, diterapkan teknik ***pruning*** dan ***pre-pruning*** untuk menghindari overfitting. ***Apply Pruning* 0.1** diterapkan untuk mengurangi ukuran pohon keputusan, sementara ***Apply Pre-pruning* 0.01** digunakan untuk membatasi pertumbuhan pohon keputusan agar lebih sederhana dan lebih mudah untuk diinterpretasikan. C4.5 bekerja dengan cara membuat pohon keputusan berdasarkan pembagian data yang optimal menggunakan atribut yang ada, dengan tujuan untuk memaksimalkan ketepatan prediksi.



Gambar 4. Implementasi Algoritma C4.5 terhadap data yang sudah diproses

Pengujian dan Evaluasi

Setelah model selesai dibangun, tahap pengujian dilakukan menggunakan data uji yang telah dipisahkan sebelumnya. Hasil pengujian model diperoleh melalui ***PerformanceVector*** dengan nilai evaluasi sebagai berikut:

- Akurasi : 95.01%
- Confusion Matrix :

Tabel 1. Nilai Evaluasi PerformanceVector

Prediksi \ Aktual	**1 (Stroke)**	**0 (Tidak Stroke)**
1 (Prediksi Stroke)	1	2
0 (Prediksi Tidak Stroke)	49	970

Dengan menggunakan Confusion Matrix, berikut adalah perhitungan untuk metrik evaluasi lainnya:

- **Precision** : Precision dihitung dengan rumus :

$$Precision = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} \ + \ \text{False Positive (FP)}}$$

Precision menunjukkan hasil 33.33%.

- **Recall** : Recall dihitung dengan rumus :

$$Recall = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} \ + \ \text{False Negative (FN)}}$$

Recall menunjukkan hasil 2.04%.

- **F1-score** : F1-score dihitung sebagai rata-rata harmonis dari Precision dan Recall :

$$F1 - Score = 2\ x\ \frac{\text{Precision x Recall}}{\text{Precision} + \ \text{Recall}}$$

F1-score menunjukkan hasil 3.81%.

- **Negative Predictive Value** (NPV) : NPV dihitung dengan rumus :

$$NPV = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} \ + \ \text{False Negative (FN)}}$$

NPV menunjukkan hasil 95.10%

- **Specificity** : Specificity dihitung dengan rumus :

$$Specificity = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} \ + \ \text{False Positive (FP)}}$$

Specificity menunjukkan hasil 99.79%.

Berdasarkan hasil evaluasi ini, meskipun model menunjukkan akurasi yang tinggi (95.01%), terdapat beberapa kelemahan dalam mendeteksi kasus stroke. Hal ini terlihat dari nilai ***recall*** dan ***precision*** yang rendah. Model cenderung lebih baik dalam mengidentifikasi individu yang tidak terkena stroke, seperti yang tercermin dari ***Specificity*** yang tinggi dan ***NPV*** yang juga baik.

Namun, kemampuan model dalam mendeteksi individu yang benar-benar mengalami stroke (yang tercermin dalam ***precision*** dan ***recall***) masih dapat ditingkatkan. Oleh karena itu, penelitian ini menyarankan agar dilakukan penyempurnaan pada model, terutama dengan meningkatkan ***recall***, sehingga model dapat lebih efektif dalam mengidentifikasi kasus stroke yang sebenarnya.

KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma C4.5 untuk memprediksi stroke berdasarkan data kesehatan. Model yang dihasilkan menunjukkan akurasi tinggi, yaitu 95.01%, yang mengindikasikan kemampuan model dalam memprediksi dengan baik. Namun, meskipun akurasi model tinggi, terdapat tantangan dalam mendeteksi individu yang benar-benar terkena stroke, yang tercermin dari nilai ***precision dan recall*** yang rendah.

Sebaliknya, model lebih baik dalam memprediksi individu yang tidak terkena stroke, terbukti dari nilai ***specificity*** dan ***NPV*** yang tinggi. Meskipun demikian, penelitian ini menyarankan adanya penyempurnaan, terutama untuk meningkatkan nilai recall. Langkah ini diperlukan agar model menjadi lebih efektif dalam mendeteksi kasus stroke secara akurat dan membantu upaya pencegahan lebih lanjut. Algoritma C4.5 terbukti efisien, namun perbaikan pada aspek tertentu masih dibutuhkan agar memberikan hasil yang lebih optimal.

DAFTAR REFERENSI

Abadi, A. F., Alamsyah, N., Retnanto, F. G., Daniati, E., & Ristyawan, A. (2024). Penerapan Data Mining dalam Mengklasifikasi Penyakit Stroke Menggunakan Algoritma. *Agustus*, *7*, 2549–7952.

Abidin, Z., Tamrin, T., Harsono, V., Aziza, D. N., & Kansania, I. (2024). *OPTIMALISASI DIAGNOSIS STROKE DENGAN ALGORITMA C4.5 DAN STRATEGI IMPUTASI k-NN UNTUK MENGATASI MISSING VALUE OPTIMIZATION OF STROKE DIAGNOSIS USING C4.5 ALGORITHM AND k-NN IMPUTATION STRATEGY TO OVERCOME MISSING VALUE*. *15*(2), 2548–4168.

Alfiyan, R., Hikmatyar, M., & Sundari, S. S. (2024). *Indonesian Journal of Digital Business Implementasi Metode Klasifikasi C4 . 5 Penyebab Faktor*. *4*(October), 37–48.

Gusrialni Fitri, N., Adilya, S., & Azizi, F. (2023). Comparison of the Naive Bayes Classification System and C4.5 for the Diagnosis of Stroke Perbandingan Sistem Klasifikasi Naive Bayes dan C4.5 Untuk Diagnosa Penyakit stoke. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 49–55.

Maulana Sidiq, C., Faqih, A., & Dwilestari, G. (2024). Algoritma Decision Tree C4.5 Digunakan Untuk Mengklasifikasikan Data Stroke. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1869–1874. https://doi.org/10.36040/jati.v8i2.8388

Nabila, F., Afrianty, I., Sanjaya, S., & Syafria, F. (2023). Implementasi Algoritma C4.5 dalam Melakukan Klasifikasi Penyakit Stroke Otak. *Jurnal Informatika Universitas Pamulang*, *8*(2), 229–235. https://doi.org/10.32493/informatika.v8i2.31361

Nadira, A., & Farida Utami, S. (2024). Implementasi Data Mining Dalam Mengidentifikasi Faktor Pasien Yang Berpotensi Mengalami Obesitas Menggunakan Algoritma C4.5. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(4), 7872–7876. https://doi.org/10.36040/jati.v8i4.10498

Pradhana, A. H., Irfa, M., Ali, A., Ristyawan, A., & Daniati, E. (2024). *Penerapan Regresi Linear Menggunakan RapidMiner Untuk Memprediksi Penjualan dan Persediaan*. *8*, 291–297.

Rivliansyah, Z., Persyaratan, U., Dan, P., & Tugas, P. (2022). *UNTUK MEMPREDIKSI PENYAKIT STROKE OTAK PROGRAM STUDI TEKNIK INFORMATIKA*.

Saputra, R. A., Agustina, C., Puspitasari, D., Ramanda, R., Warjiyono, Pribadi, D., Lisnawanty, & Indriani, K. (2020). Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization. *Journal of Physics: Conference Series*, *1641*(1), 61–67. https://doi.org/10.1088/1742-6596/1641/1/012025

Yuliska, Y., & Syaliman, K. U. (2020). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688