JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

ANALISIS PERBANDINGAN MODEL NAÏVE BAYES DAN C4.5 UNTUK PREDIKSI STROKE BERDASARKAN RIWAYAT DATA MEDIS DENGAN PENDEKATAN MATRIKS KORELASI

Samuel*[1)], Idmi[2)], Gandung Triyono[3)]
1. Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia
2. Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia
3. Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia

Article Info
Kata Kunci : C4.5, *Data Mining*, Matriks Korelasi, Naïve Bayes, Prediksi Stroke.

Keywords : *Correlation Matrix*, C4.5, *Data Mining*, Naïve Bayes, *Stroke Prediction*.

Article history:
Received 7 July 2025
Revised 27 October 2025
Accepted 11 November 2025
Available online 1 December 2025

DOI :
https://doi.org/10.29100/jipi.v10i4.8653

* Corresponding author.
Samuel
E-mail address:
2311601138@student.budiluhur.ac.id

ABSTRAK
Stroke merupakan satu diantara penyakit mematikan yang bisa terjadi secara mendadak dan bisa menyebabkan kematian/kecacatan. Prediksi dini risiko stroke sangat krusial guna mendukung tindakan antisipasi dan penanganan yang tepat. Penelitian ini membandingkan akurasi dua algoritma klasifikasi, yaitu Naïve Bayes dan C4.5, dalam memprediksi risiko stroke berdasarkan data medis pasien. Metode pemilihan atribut menggunakan matriks korelasi diterapkan untuk memilih fitur yang paling relevan guna meningkatkan akurasi model. Data yang digunakan merupakan *dataset* stroke dari situs *Kaggle*. Proses penelitian mengikuti tahapan *Knowledge Discovery in Database* (KDD). Hasil penelitian memperlihatkan penerapan matriks korelasi sebagai teknik seleksi atribut meningkatkan akurasi kedua algoritma. Algoritma C4.5 memberikan akurasi tertinggi mencapai 95%. Atribut yang berpengaruh signifikan dalam prediksi stroke antara lain tipe tempat tinggal, jenis kelamin, penyakit jantung, hipertensi, rata-rata kadar glukosa, dan status merokok. Dengan demikian, kombinasi seleksi fitur berbasis matriks korelasi dan algoritma C4.5 efektif untuk membangun model prediksi risiko stroke yang akurat dan dapat menjadi alat bantu diagnosis medis.

ABSTRACT
Stroke is one of the most deadly diseases that can occur suddenly and lead to death or disability. Early prediction of stroke risk is crucial to support effective prevention and treatment efforts. This study compares the accuracy of two classification algorithms namely, Naïve Bayes and C4.5 in predicting stroke risk based on patient medical data. A correlation matrix based attribute selection method was applied to identify the most relevant features, improving model accuracy. The data used was a stroke dataset obtained from Kaggle. The research process followed the Knowledge Discovery in Databases (KDD) framework. The results show that applying a correlation matrix as a feature selection technique improves the accuracy of both algorithms. The C4.5 algorithm achieved the highest accuracy of 95%. Significant attributes influencing stroke prediction include residence type, gender, heart disease, hypertension, average glucose level, and smoking status. Thus, the combination of correlation matrix based feature selection and the C4.5 algorithm is effective in building an accurate stroke risk prediction model and can serve as a valuable tool for medical diagnosis.

I. PENDAHULUAN

Memahami gejala suatu penyakit merupakan langkah awal penting dalam upaya mencegah penyakit yang dapat membahayakan kesehatan dan berpotensi mengancam nyawa. Stroke menjadi salah satu masalah serius yang dihadapi hampir seluruh belahan dunia [1]. Hal ini disebabkan oleh serangan stroke yang dapat terjadi secara tiba-tiba dan menyebabkan kematian, serta berisiko menimbulkan kecacatan fisik dan mental, baik pada orang yang berada di usia produktif maupun lanjut usia [2]. Stroke terjadi ketika aliran darah ke otak terganggu akibat adanya penyumbatan atau pecahnya pembuluh darah, sehingga menyebabkan sebagian otak tidak memperoleh pasokan oksigen yang dibutuhkan. Kondisi ini mengakibatkan kematian sel atau jaringan otak

dan menjadikan stroke sebagai penyakit yang berpotensi fatal bagi manusia [3]. Berdasarkan data WHO (*World Health Organization*), stroke menempati posisi ke-2 sebagai penyebab kematian tertinggi setelah penyakit jantung. Cahyani [4] menyebutkan bahwa setiap tahun sekitar lima belas juta orang di seluruh dunia mengalami stroke, dengan satu orang meninggal setiap 4-5 menit akibat serangan tersebut. Di Indonesia, data yang berasal dari Kementerian Kesehatan menerangkan bahwa lebih dari 500.000 orang terkena stroke di tiap tahunnya, dengan 12.500 di antaranya meninggal dan selebihnya mengalami kecacatan ringan.

Hampir setiap hari atau paling tidak setiap tiga hari sekali, terjadi kematian akibat stroke pada masyarakat Indonesia, baik pada usia lanjut maupun usia muda. Kematian yang disebabkan oleh stroke sulit diprediksi karena gejalanya muncul secara tiba-tiba dan berkembang dengan cepat [5]. Meskipun rumah sakit menyimpan data pasien stroke dalam jumlah yang besar, data tersebut tidak akan berguna jika tidak diolah menjadi informasi yang bermakna. Oleh karena itu, penting untuk melakukan analisis lebih mendalam guna menggali informasi baru dari data rekam medis pasien stroke untuk melakukan prediksi penyakit ini [6]. Rahman [7] menyatakan bahwa salah satu teknik yang bisa diperoleh manfaatnya dalam analisis data kesehatan ialah *data mining*. Adapun *data mining* adalah ilmu yang mengkombinasikan berbagai disiplin dalam ilmu komputer untuk menemukan pola-pola dan informasi esensial dari sejumlah besar data. Proses ini dikenal sebagai ekstraksi pengetahuan dari *database* besar dengan mempergunakan bermacam metode, seperti kecerdasan buatan, *machine learning*, dan statistik. Satu diantara metode yang dapat diterapkan pada *data mining* ialah metode klasifikasi, yaitu teknik guna mengenali kelas atau model data yang kemudian dipergunakan untuk memprediksi suatu kelas/label [8]. Penelitian ini mempergunakan algoritma Naïve Bayes dan C4.5 untuk melaksanakan klasifikasi. Naïve Bayes memanfaatkan probabilitas guna memperkirakan peluang berdasarkan data historis [9]. Sementara itu, Algoritma C4.5 membangun pohon keputusan untuk mengidentifikasi hubungan antar variabel-variabel [10].

Penelitian terdahulu yang menerapkan algoritma klasifikasi *data mining* yaitu Naïve Bayes dan C4.5 untuk memprediksi stroke, antara lain adalah penelitian yang dilakukan oleh Riany dan Testiana [11] menggunakan algoritma Naïve Bayes untuk mengklasifikasikan risiko stroke berdasarkan data medis pasien, hasilnya Naïve Bayes dapat digunakan untuk mengelompokkan pasien guna menentukan apakah mereka terdiagnosis stroke atau tidak, dengan akurasi mencapai 92,48% yang dikategorikan sebagai klasifkasi yang baik. Penelitian lainnya yang menggunakan algoritma Naïve Bayes, antara lain oleh Airi et al [12] dengan akurasi 71,9%, Akmal et al [13] dengan akurasi 90,10%, serta Abadi et al [14] dengan akurasi 87,22%. Adapun Penelitian yang menggunakan algoritma C4.5 untuk melakukan klasifikasi data stroke, seperti yang dilakukan oleh Maulana Sidiq et al [15] hasil penelitian ini, algoritma C4.5 menunjukkan kemampuan yang signifikan dalam melakukan klasifikasi awal terhadap penyakit stroke dengan tingkat akurasi yang dihasilkan sebesar 93,64%. Penelitian serupa juga dilakukan oleh Hendriyansyah et al [16] dengan akurasi 85,81%. Alfiyan et al [17] dengan rata-rata akurasi 93%. Permana et al [18] dengan akurasi 91%.

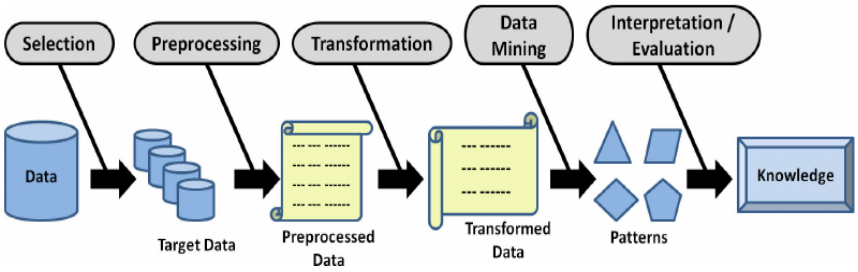
Meskipun berbagai penelitian sebelumnya menunjukkan hasil akurasi yang tinggi. Namun, sebagian besar penelitian tersebut menggunakan seluruh atribut dalam *dataset* tanpa melakukan seleksi fitur berbasis matriks korelasi. Akibatnya, model yang dihasilkan berpotensi mengalami beberapa kelemahan, seperti pada asumsi algoritma Naïve Bayes yang mengharuskan atribut-atribut *input* bersifat *independent* terhadap kelas target. Jika atribut yang digunakan saling berkorelasi tinggi, maka akan melanggar asumsi tersebut [8]. Disisi lain pada algoritma C4.5, atribut yang berkorelasi tinggi satu sama lain dapat menyebabkan pohon keputusan menjadi lebih kompleks dan sulit digeneralisasi ke data baru [19]. Matriks korelasi digunakan untuk melihat hubungan antar atribut dalam *dataset*. Dengan cara ini, atribut yang paling berpengaruh terhadap hasil prediksi dapat dipertahankan, sedangkan atribut yang memiliki hubungan terlalu kuat antar sesamanya dihapus karena membawa informasi yang sama, sehingga membuat model lebih fokus pada fitur penting [20][21]. Dengan begitu, Naïve Bayes dapat menghitung probabilitas dengan lebih tepat dan C4.5 dapat membentuk pohon keputusan yang lebih sederhana, sehingga hasil prediksi menjadi lebih akurat dan stabil.

Oleh karena itu, penelitian ini dilakukan untuk mengatasi keterbatasan tersebut melalui penerapan algoritma Naïve Bayes dan C4.5 dengan teknik seleksi atribut menggunakan operator *correlation matrix* di *RapidMiner* guna mengidentifikasi faktor-faktor risiko yang lebih relevan terhadap penyakit stroke, sehingga model yang dihasilkan menjadi lebih fokus dan akurat. Pendekatan ini diharapkan mampu mendukung penerapan analisis berbasis data dalam bidang kesehatan. Hasil penelitian ini juga diharapkan dapat menjadi referensi pengembangan sistem prediksi penyakit berbasis *data mining*, khususnya untuk mendukung pengambilan keputusan medis berbasis data. Dengan demikian, penelitian ini tidak hanya memperkaya kajian akademik di bidang *data mining*, tetapi juga memberikan manfaat bagi dunia medis.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

II. METODOLOGI PENELITIAN

Metodologi penelitian ini berfungsi sebagai panduan sistematis dalam mencapai tujuan penelitian. Data yang digunakan berupa data sekunder, yakni *dataset stroke* yang termuat di situs *Kaggle* (https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset). *Dataset* stroke yang digunakan terdiri dari 4.981 *record* data, dengan kelas non-stroke sebanyak 4.733 data dan kelas stroke sebanyak 248 data. Penelitian ini menggunakan kerangka kerja *Knowledge Discovery in Database* (KDD) yang mengandung tahapan-tahapan yaitu *selection* (pemilihan atribut), *preprocessing* (pembersihan dan pengolahan data), *transformation* (transformasi data ke format yang sesuai), *data mining* (penerapan algoritma klasifikasi Naïve Bayes dan C4.5), dan *evaluation* (penilaian akurasi model). Adapun pengertian KDD sendiri adalah proses sistematis untuk mengekstraksi pengetahuan yang berguna dari kumpulan data besar [8]. Penjelasan rinci tahapan-tahapan mengenai KDD dapat dilihat Gambar 1.



Gambar 1. *Knowledge Discovery in Database* (KDD)

A. *Tahap selection*

Di tahapan awal, dilakukan pemilihan data yang sesuai untuk dianalisis. [9]. Pemilihan atribut yang tepat sangat penting, terutama dalam menentukan variabel target (yang akan diprediksi) dan label (yang digunakan untuk klasifikasi atau pelabelan), karena atribut-atribut ini akan digunakan dalam membangun model.

B. *Tahap preprocessing*

Data yang sudah dipilih kemudian dilakukan *cleaning* dari berbagai masalah kualitas misalnya data ganda, inkonsistensi, nilai yang hilang, serta mencakup perbaikan kesalahan penulisan. Tujuannya agar data yang digunakan benar-benar representatif dan tidak mengandung bias yang dapat mengganggu hasil analisis [9]. Pada tahap ini juga digunakan matriks korelasi untuk menganalisis hubungan antar atribut dalam *dataset*. Matriks ini menunjukkan kekuatan dan arah hubungan antar variabel melalui nilai korelasi, sehingga atribut yang kurang relevan dapat diabaikan untuk meningkatkan efisiensi dan akurasi model prediksi [22]. Penelitian ini menggunakan ambang batas korelasi $\geq$ 0,7 untuk seleksi atribut, karena nilai di atas batas tersebut menunjukkan hubungan yang kuat antar variabel [23]. Dengan ambang ini, atribut yang paling relevan terhadap variabel target (stroke) dapat dipertahankan sehingga model menjadi lebih fokus dan representatif. Adapun rumus korelasi matriks [22] dapat dilihat pada Persamaan (1).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \ \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

Keterangan :
x_i dan y_i = nilai individu untuk variabel x dan y.
$\bar{x}$ dan $\bar{y}$ = rata-rata dari masing-masing variabel.

C. *Tahap transformation*

Tahap ini dilakukan penyesuaian format data supaya selaras dengan keperluan teknik analisis yang dijalankan. Contohnya ialah mengubah tipe data kategorikal menjadi numerik, serta mengelompokkan data menjadi dua kelompok yaitu *training* dan *testing* agar dapat dilaksanakan pelatihan dan pengujian model [24].

D. *Tahap data mining*

Pada tahap *data mining*, penelitian ini mempergunakan dua algoritma klasifikasi, yaitu Naïve Bayes dan C4.5. Algoritma Naïve Bayes merupakan algoritma klasifikasi dengan memanfaatkan prinsip-prinsip teorema Bayes, yang mengasumsikan bahwa setiap variabel bersifat independen satu sama lain dalam menentukan nilai *output* tertentu. Dengan arti lainnya, Naïve Bayes menganggap bahwa setiap variabel *input* memberikan kontribusi

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

secara terpisah dan bebas dalam menentukan probabilitas kelas target, sehingga perhitungan probabilitas menjadi lebih sederhana dan efisien [5]. Secara matematis, teorema Bayes [25] terlihat pada Persamaan (2).

$$P(c|d) = \frac{P(d|c)\ x\ P(c)}{P(d)} \quad (2)$$

Dimana :
$P(c|d)$ = *posterior*/peluang kelas c diberikan kondisi d.
$P(d|c)$ = peluang kelas d diberikan kondisi c (*likelihood*).
$P(c)$ = *prior*/peluang awal munculnya kelas c.
$P(d)$ = *evidence*/peluang munculnya kelas d.

Sementara itu, algoritma C4.5 ialah algoritma pohon keputusan yang membangun model klasifikasi dengan memilih atribut terbaik berdasarkan nilai *Information Gain Ratio* tertinggi. Algoritma ini sangat populer karena kemampuannya dalam menangani berbagai tipe data, sekaligus menghasilkan pohon keputusan yang mudah dipahami. Atribut yang dipilih untuk dijadikan *node*/akar berdasarkan pada nilai *gain* yang paling besar dari seluruh atribut [15]. Persamaan (3) merupakan cara melakukan perhitungan nilai *gain* [25].

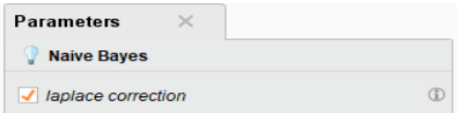
$$Gain\ (S, A) = Entrophy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entrophy(Si) \quad (3)$$

Adapun untuk perhitungan mencari nilai *entrophy* [25], seperti pada Persamaan (4).

$$Entrophy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (4)$$

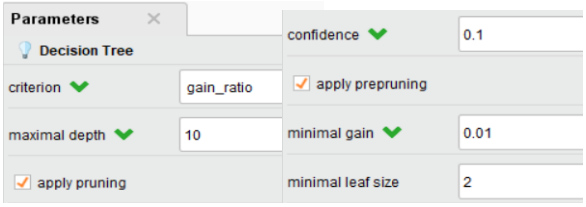
Keterangan :
$Gain\ (S, A)$ = informasi yang didapatkan dari atribut (A) terhadap *output* data (S).
$Entrophy\ (S)$ = ukuran ketidakpastian dalam suatu himpunan data (S).
$|Si|$ = jumlah kasus pada partisi ke-i.
$|S|$ = jumlah kasus pada S.
pi = proporsi dari Si terhadap S.
n = jumlah partisi atribut A.

Berikut adalah rincian parameter dan konfigurasi yang digunakan untuk algoritma Naïve Bayes dan C4.5 pada perangkat lunak *RapidMiner*.



Gambar 2. Parameter dan Konfigurasi Naïve Bayes Pada *RapidMiner*

Gambar 2 menampilkan konfigurasi parameter operator Naïve Bayes dengan opsi *laplace correction* aktif, sesuai dengan pengaturan awal *RapidMiner*. Opsi ini berfungsi untuk mencegah munculnya nilai probabilitas nol ketika suatu atribut tidak memiliki nilai tertentu pada kelas target.



Gambar 3. Parameter dan Konfigurasi C4.5 Pada *RapidMiner*

Gambar 3 menunjukkan konfigurasi parameter operator *Decision Tree* (C4.5) dengan *criterion* yaitu *gain_ratio* sebagai dasar pemilihan atribut terbaik. Adapun pengaturan lainnya menggunakan nilai *default* dari *RapidMiner*.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

E. Tahap evaluation

Tahap terakhir berupa pengujian dan penilaian kinerja model menggunakan *confusion matrix*. Adapun *confusion matrix* ialah sebuah teknik yang umum dalam melakukan penilaian terhadap performa model klasifikasi pada *data mining* dengan menghitung akurasi serta metrik lainnya. *Confusion matrix* memuat informasi tentang perbandingan antara kelas sebenarnya (aktual) dan kelas prediksi yang dihasilkan oleh pengolahan klasifikasi, sehingga dapat memberikan gambaran lengkap mengenai performa model [26].

III. HASIL DAN PEMBAHASAN

Pada bagian hasil dan pembahasan, diterangkan langkah-langkah detail teknik *data mining* pada algoritma Naïve Bayes dan C4.5 untuk prediksi risiko penyakit stroke. Penelitian ini juga memakai pendekatan menggunakan teknik matriks korelasi guna memilih atribut yang paling relevan sehingga dapat meningkatkan akurasi model klasifikasi. Semua proses tersebut dilakukan dan diuji menggunakan perangkat lunak *RapidMiner* dengan *dataset* stroke yang diperoleh dari situs *Kaggle* yang dapat dilihat pada Tabel I.

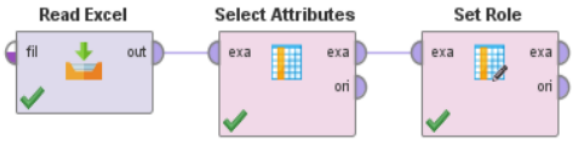
TABEL I
CONTOH DATASET STROKE

Gender	Age	Hypertension	Heart disease	Ever married	Work type	Residence type	Avg Glucose level	BMI	Smoking status	Stroke
Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	1
Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	1
Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
Female	79	1	0	Yes	Self-em-ployed	Rural	174,12	24	never smoked	1
Male	81	0	0	Yes	Private	Urban	186,21	29	formerly smoked	1
Male	74	1	1	Yes	Private	Rural	70,09	27,4	never smoked	1
Female	69	0	0	No	Private	Urban	94,39	22,8	never smoked	1
Female	78	0	0	Yes	Private	Urban	58,57	24,2	Unknown	1
Female	81	1	0	Yes	Private	Rural	80,43	29,7	never smoked	1
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..

Dataset stroke yang digunakan dalam penelitian ini melewati beberapa tahapan, yaitu *Selection*, *Preprocessing*, hingga *Evaluation*, sebagai bagian dari kerangka kerja *Knowledge Discovery in Database* (KDD). Tahapan ini sangat penting untuk memastikan bahwa data siap dianalisis lebih lanjut menggunakan metode klasifikasi dengan algoritma Naïve Bayes dan C4.5.

A. Tahap selection

Tahap ini melibatkan pemilihan data yang mencakup 10 atribut prediktor serta 1 atribut target (label). Adapun atribut prediktor tersebut adalah *Gender*, *Age*, *Hypertension*, *Heart disease*, *Ever married*, *Work type*, *Residence type*, *Avg glucose level*, BMI, dan *Smoking status*. Adapun atribut target (label) yaitu Stroke.



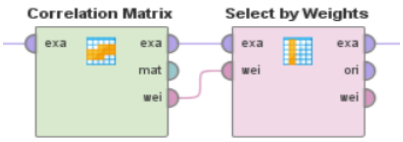
Gambar 4. Proses Mengimpor *Dataset*, Memilih Atribut, dan Menetapkan Label pada *RapidMiner*

Proses peng*input*an *dataset*, pemilihan atribut serta penentuan atribut target dilakukan melalui sejumlah operator pada perangkat lunak *RapidMiner* seperti Gambar 4. *Read excel* untuk mengimpor *dataset* dari *file Excel* ke dalam *RapidMiner*. *Select attributes* digunakan untuk memilih atribut-atribut yang akan dipakai dalam analisis. Kemudian untuk memilih atribut target (label) menggunakan operator *set role*.

B. Tahap preprocessing

Setelah tahap seleksi, pengecekan kesalahan dilakukan untuk memastikan tidak ditemukan *error* atau inkonsistensi data. Atribut yang lolos pengecekan tersebut, kemudian dianalisis untuk mengetahui hubungan

korelasi antar atribut. Atribut yang menunjukkan korelasi signifikan akan digunakan dalam proses *data mining*, sementara atribut yang kurang relevan akan dikesampingkan. Dalam penelitian ini, operator *Correlation matrix* dan *Select by weight* pada *RapidMiner* digunakan untuk membantu menentukan atribut-atribut yang memiliki hubungan kuat dengan atribut target. Atribut yang lolos kriteria ini kemudian dipertahankan untuk diproses pada tahap analisis berikutnya.



Gambar 5. *Preprocessing* Menggunakan *Correlation Matrix* dan *Select by Weights*

Pada Gambar 5, operator *Correlation matrix* berfungsi untuk menghitung hubungan antar atribut dalam *dataset*, sedangkan *Select by weights* untuk menyeleksi atribut berdasarkan bobot atau relevansi yang tinggi. Atribut yang berkorelasi signifikan terhadap atribut target dengan nilai korelasi ≥ 0,7 dipilih untuk analisis lebih lanjut, sementara atribut dengan nilai di bawah batas ambang tersebut diabaikan. Dengan demikian, proses *preprocessing* menggunakan operator *Correlation matrix* dan *Select by weights* dapat membantu meningkatkan kinerja model klasifikasi dengan memfokuskan pada atribut yang benar-benar berkontribusi/berkorelasi.

TABEL II
BOBOT ATRIBUT

Atribut	Weight
Residence type	1
Gender	0,980
Heart disease	0,881
Avg glucose level	0,848
Hypertension	0,834
Smoking status	0,719

Berdasarkan hasil penggunaan operator *Correlation matrix* dan *Select by weights* pada Tabel II, atribut yang memiliki korelasi signifikan terhadap atribut target antara lain *Residence type*, *Gender*, *Heart disease*, *Avg glucose level*, *Hypertension*, dan *Smoking status*, yang kemudian dipertahankan dalam proses *data mining*. Sebaliknya, atribut dengan bobot rendah seperti BMI, *Work type*, *Ever married*, dan *Age* akan diabaikan pada tahap berikutnya.

Secara medis, jenis tempat tinggal (*residence type*) dapat berpengaruh secara tidak langsung terhadap risiko stroke. Individu yang tinggal di wilayah perkotaan cenderung memiliki pola hidup kurang aktivitas fisik dan tingkat stres yang lebih tinggi dibandingkan mereka yang tinggal di pedesaan, sehingga berpotensi meningkatkan risiko stroke [27]. Jenis kelamin (*gender*) juga berperan sebagai faktor risiko stroke, laki-laki memiliki kemungkinan lebih tinggi mengalami stroke dibandingkan perempuan, terutama karena kebiasaan merokok, pola makan, dan tingkat hipertensi yang lebih tinggi [28]. Faktor penyakit jantung (*heart disease*) juga penyebab stroke karena gangguan pada jantung dapat menghambat aliran darah ke otak dan menyebabkan penyumbatan pembuluh darah [29]. Rata-rata kadar glukosa (*average glucose level*) yang tinggi juga berhubungan erat dengan gangguan metabolisme, di mana kadar gula berlebih dapat merusak dinding pembuluh darah dan meningkatkan risiko penyumbatan otak [27]. Selain itu, tekanan darah tinggi (*hypertension*) dapat merusak pembuluh darah dan memicu pecahnya pembuluh darah di otak. Terakhir, kebiasaan merokok dapat mempercepat penumpukan plak pada dinding pembuluh darah serta meningkatkan risiko penggumpalan darah, yang menjadi salah satu penyebab utama stroke iskemik [29].

C. *Tahap transformation*

Tahap selanjutnya adalah *transformation* yang bertujuan untuk menyiapkan data agar sesuai dengan kebutuhan algoritma klasifikasi Naïve Bayes dan C4.5.

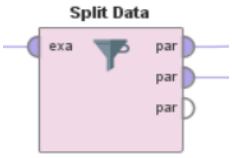
TABEL III
INISIALISASI ATRIBUT

Atribut	Keterangan	Inisialisasi
Gender	Female	0
	Male	1
Residence type	Rural	0

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

Atribut	Keterangan	Inisialisasi
Smoking status	Urban	1
	Never smoked	0
	Formerly smoked	1
	Smokes	2
	Unknown	3

Pada Tabel III, beberapa atribut kategorikal seperti *Gender*, *Residence type*, dan *Smoking status* diubah menjadi representasi numerik melalui proses inisialisasi karena algoritma *machine learning* umumnya memerlukan *input* berupa numerik. Dengan mengubah nilai kategorikal menjadi angka, data menjadi kompatibel dengan algoritma tersebut [30]. Selanjutnya, data dibagi menjadi dua bagian yaitu data *training* dan data *testing* dengan proporsi 70% dan 30% menggunakan operator *split data* pada perangkat *RapidMiner*.

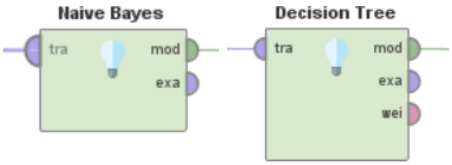


Gambar 6. Pembagian Data *Training* dan *Testing*

Melalui operator *Split data* pada Gambar 6, pembagian ini memungkinkan model dilatih pada data *training* dan diuji pada data *testing* yang tidak dikenal oleh model, sehingga dapat mengukur kinerja model secara objektif. Dengan tahap *transformation* ini, data yang digunakan sudah siap untuk tahap pemodelan dan evaluasi berikutnya.

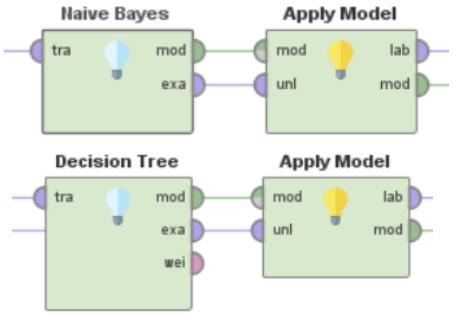
D. Tahap data mining

Tahap ini dilakukan pemodelan pada *dataset* stroke untuk proses klasifikasi menggunakan dua algoritma, yaitu Naïve Bayes dan C4.5.



Gambar 7. Proses *Data Mining*

Pada Gambar 7, proses *data mining* dilakukan dengan operator Naïve Bayes dan operator *Decision Tree* (C4.5) pada perangkat lunak *RapidMiner*. Adapun atribut yang digunakan untuk tahap *data mining* ini, berdasarkan atribut yang berkorelasi signifikan (≥ 0,7) dengan atribut target sehingga dapat menghasilkan model klasifikasi yang lebih akurat dan menghindari pengaruh atribut kurang relevan yang dapat menurunkan performa model.



Gambar 8. Proses Penerapan Model

Dalam Gambar 8, sesudah proses pemodelan selesai. Langkah selanjutnya ialah menggunakan operator *Apply model*. Operator tersebut berfungsi untuk menerapkan model yang sudah dilatih dengan data *training* ke data *testing* yang belum memiliki label. Tujuannya adalah menghasilkan prediksi pada data uji tersebut sekaligus mengevaluasi performa model dalam mengklasifikasikan data.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3749-3759

	true 0	true 1
pred. 0	1399	67
pred. 1	21	7

Gambar 9. Hasil Prediksi Stroke Pada Algoritma Naïve Bayes

Gambar 9 merupakan hasil prediksi stroke pada algoritma Naïve Bayes. Dalam matriks tersebut, model mengklasifikasikan dengan tepat sebanyak 1399 kasus yang sebenarnya tidak mengalami stroke. Namun, ada 21 kasus yang sebenarnya bebas stroke tetapi diklasifikasikan sebagai stroke oleh model. Selain itu, model juga mengklasifikasikan 67 pasien yang memang mengalami stroke tetapi memprediksi mereka sebagai tidak stroke. Sedangkan untuk pasien yang benar-benar mengalami stroke dan berhasil dikenali dengan benar oleh model sebanyak 7 kasus.

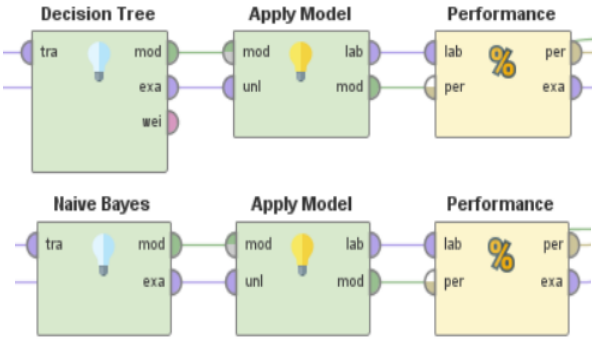
	true 0	true 1
pred. 0	1416	70
pred. 1	4	4

Gambar 10. Hasil Prediksi Stroke Pada Algoritma C4.5

Gambar 10 merupakan hasil prediksi stroke pada algoritma C4.5, dimana model berhasil mengklasifikasikan 1416 kasus non-stroke dengan benar, tetapi menghasilkan 4 pasien non-stroke yang diprediksi sebagai stroke dan 70 pasien stroke yang tidak terdeteksi stroke. Selain itu, 4 kasus stroke yang terprediksi dengan benar.

E. Tahap evaluation

Pada tahap evaluasi, dilakukan evaluasi kinerja model algoritma Naïve Bayes dan C4.5 dengan operator *Performace* seperti pada Gambar 11.



Gambar 11. Evaluasi Kinerja Model

Berikut ini disajikan hasil *confusion matrix* yang diperoleh dari penggunaan operator *Performance* pada aplikasi *RapidMiner* dengan proporsi pembagian data 70% untuk *training* dan 30% untuk *testing*. Selain itu, juga disajikan perbandingan akurasi model ketika menggunakan matriks korelasi sebagai metode seleksi fitur dan tanpa menggunakan matriks korelasi. Perbandingan ini akan memperlihatkan sejauh mana pengaruh matriks korelasi dalam meningkatkan akurasi dan kinerja model, sebagaimana dirangkum pada Tabel IV.

TABEL IV
HASIL AKURASI

Algorima	Menggunakan matriks korelasi Akurasi	Tanpa matriks korelasi Akurasi
Naïve Bayes	94,11%	89,63%
C4.5	95,05%	93,57%

Tabel IV memperlihatkan perbandingan tingkat akurasi model klasifikasi yang dibangun dengan dua pendekatan berbeda dalam seleksi fitur, yaitu menggunakan dan tanpa menggunakan matriks korelasi. Hasil menunjukkan bahwa penerapan matriks korelasi sebagai metode seleksi fitur berhasil meningkatkan akurasi pada kedua algoritma yang diuji, yakni Naïve Bayes dan C4.5. Pada Naïve Bayes, akurasi naik dari 89,63%

menjadi 94,11%, sementara pada algoritma C4.5, akurasi meningkat dari 93,57% menjadi 95,05%. Peningkatan tersebut mengindikasikan bahwa seleksi fitur menggunakan matriks korelasi mampu mengeliminasi atribut yang kurang relevan, sehingga membuat model lebih efisien dan akurat dalam melakukan prediksi. Dengan demikian, dari hasil tersebut algoritma C4.5 memperoleh hasil terbaik dengan akurasi tertinggi sebesar 95,05% dibandingkan Naïve Bayes.

TABEL V
HASIL AKURASI LANJUTAN

Data *training* & data *testing*	Algorima	Akurasi menggunakan matriks korelasi	Akurasi tanpa matriks korelasi	Peningkatan akurasi
70 : 30	Naïve Bayes	94,11%	89,63%	4,48%
	C4.5	95,05%	93,57%	1,48%
80 : 20	Naïve Bayes	94,18%	89,37%	4,81%
	C4.5	95,19%	93,98%	1,21%
90 : 10	Naïve Bayes	94,38%	90,36%	4,02%
	C4.5	95,38%	94,18%	1,20%

Tabel V menyajikan hasil akurasi lanjutan dari uji coba dengan tambahan dua variasi pembagian data *training* dan *testing*, yaitu 80% : 20% dan 90% : 10%. Tujuan dari pengujian ini adalah untuk melihat konsistensi pengaruh penggunaan matriks korelasi dalam meningkatkan akurasi model, serta untuk mengevaluasi apakah algoritma C4.5 tetap memberikan hasil terbaik pada berbagai proporsi data tersebut. Berdasarkan hasil Tabel V, dapat dilihat bahwa pada ketiga proporsi pembagian data, penggunaan matriks korelasi selalu menghasilkan akurasi yang lebih tinggi dibandingkan tanpa matriks korelasi untuk kedua algoritma yang diujikan yaitu Naïve Bayes maupun C4.5. Selain itu, algoritma C4.5 secara konsisten menunjukkan akurasi tertinggi, baik dengan penggunaan matriks korelasi maupun tanpa penggunaan matriks korelasi. Hal ini disebabkan, algoritma C4.5 merupakan metode berbasis pohon keputusan yang mampu menangani kombinasi atribut numerik dan kategorikal secara langsung. Pemilihan atribut menggunakan *information gain ratio* menjadikan algoritma ini lebih efektif dalam mengidentifikasi variabel yang paling berpengaruh terhadap kelas target (stroke). Selain itu, mekanisme *pruning* C4.5 mampu mengurangi *overfitting* dan menangkap interaksi antar variabel tanpa bergantung pada asumsi independensi, sehingga model yang dihasilkan lebih stabil dan adaptif terhadap data medis [15]. Di sisi lain, algoritma Naïve Bayes menggunakan pendekatan probabilistik berdasarkan teorema Bayes dengan asumsi bahwa setiap atribut bersifat independen [11]. Pendekatan ini membuatnya sederhana dan efisien, tetapi pada *dataset* medis seperti stroke, banyak atribut yang saling berkaitan, seperti *hypertension* dan *heart disease*. Hal ini dapat menyebabkan perhitungan probabilitas menjadi kurang akurat, sehingga akurasinya cenderung lebih rendah dibandingkan C4.5.

Adapun dari hasil pengujian, akurasi tertinggi diperoleh oleh algoritma C4.5 pada pembagian data 90% *training* dan 10% *testing* yang memanfaatkan matriks korelasi, yaitu mencapai 95,38%. Pada Tabel V terlihat juga peningkatan akurasi pada algoritma Naïve Bayes mencapai 5%, sedangkan algoritma C4.5 peningkatan akurasinya mencapai 1,5%. Berdasarkan penelitian terdahulu yang telah dibahas pada Pendahuluan serta menggunakan *dataset* serupa yang diperoleh dari situs *Kaggle* untuk memprediksi stroke tanpa menggunakan matriks korelasi. Hasil akurasi yang diperoleh pada algoritma Naïve Bayes antara lain 92,48% [11], 71,9% [12], 90,10% [13], dan 87,22% [14]. Sementara itu, pada algoritma C4.5, akurasi yang dihasilkan berkisar 93,64% [15], 85,81% [16], 93% [17], dan 91% [18]. Hasil-hasil tersebut menunjukkan bahwa kedua algoritma dapat menghasilkan akurasi yang baik, tetapi jika dibandingkan dengan hasil penelitian ini yang menggunakan matriks korelasi, akurasi pada algoritma Naïve Bayes dapat mencapai 94% dan pada algoritma C4.5 bisa mencapai 95%.

Jadi secara keseluruhan, hasil ini menunjukkan bahwa matriks korelasi efektif dalam meningkatkan performa model, dan algoritma C4.5 menghasilkan akurasi tertinggi dibandingkan Naïve Bayes dalam berbagai variasi proporsi data yang diuji, seperti yang terlihat pada Tabel V. Oleh karena itu, Peningkatan akurasi ini mengindikasikan bahwa teknik matriks korelasi memberikan kontribusi signifikan dalam meningkatkan kemampuan prediksi, terutama pada data yang bersifat kompleks seperti *dataset* yang digunakan dalam penelitian ini. Meskipun hasil penelitian ini menunjukkan bahwa penggunaan matriks korelasi mampu meningkatkan akurasi model klasifikasi, terdapat beberapa keterbatasan yang perlu diperhatikan. Penelitian ini menggunakan *dataset* publik dari situs *Kaggle* yang bersifat sekunder, sehingga kualitas dan kelengkapan datanya sangat bergantung pada sumber aslinya. Selain itu, distribusi data yang tidak seimbang antara kelas stroke dan non-stroke dapat berpotensi menimbulkan bias prediksi terhadap kelas mayoritas, terutama jika tidak diterapkan teknik penyeimbangan data.

IV. KESIMPULAN

Penelitian ini membuktikan bahwa penerapan matriks korelasi sebagai metode seleksi fitur dapat secara signifikan meningkatkan akurasi model klasifikasi untuk memprediksi stroke pada algoritma Naïve Bayes dan C4.5. Di antara keduanya, algoritma C4.5 menunjukkan performa terbaik dengan akurasi tertinggi rata-rata sebesar 95% ketika menggunakan matriks korelasi. Pengujian dengan variasi proporsi data memperlihatkan bahwa metode seleksi fitur dengan matriks korelasi tetap efektif dalam meningkatkan akurasi hasil klasifikasi. Adapun beberapa atribut yang memiliki hubungan korelasi kuat dan berpengaruh besar terhadap prediksi stroke adalah *Residence type*, *Gender*, *Heart disease*, *Hypertension*, *Avg glucose level*, dan *Smoking status*. Pemilihan atribut berdasarkan korelasi tertinggi tersebut terbukti dapat memaksimalkan akurasi model. Oleh karena itu, kombinasi penggunaan matriks korelasi untuk seleksi fitur dan algoritma C4.5 menjadi strategi yang efektif dalam membangun model prediksi risiko stroke yang handal berbasis data medis pasien. Penelitian selanjutnya dapat mengoptimasi model pohon keputusan C4.5 untuk meningkatkan performa yang lebih signifikan lagi. Selain itu, dapat dikembangkan aplikasi berbasis *web* atau *mobile* yang mengintegrasikan model prediksi risiko stroke untuk tenaga medis, serta menerapkan teknik SMOTE untuk menyeimbangkan distribusi kelas.

DAFTAR PUSTAKA

[1] A. G. Ramadhan, S. H. Hidayatullah, and R. Ruba'i, "Implementasi Algoritma C4.5 pada Sistem Prediksi Stroke Berdasarkan Data Kesehatan," *J-CEKI J. Cendekia Ilm.*, vol. 4, no. 2, pp. 1617–1624, 2025.

[2] G. Sailasya and G. L. Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," 2021. [Online]. Available: www.ijacsa.thesai.org

[3] F. Akbar, H. Wira Saputra, A. Karel Maulaya, and M. Fikri Hidayat, "Implementation of Decision Tree Algorithm C4.5 and Support Vector Regression for Stroke Disease Prediction," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, pp. 61–67, 2022.

[4] D. E. Cahyani, "PENERAPAN MACHINE LEARNING UNTUK PREDIKSI PENYAKIT STROKE," *J. Kaji. Mat. dan Apl. Vol.*, vol. 3, no. 1, 2022, doi: 10.17977/um055v3i1p15-22.

[5] L. Rahmawati, M. H. T, and M. A. Sunandar, "Analisis Data Mining Untuk Memprediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," *JATIKOM J. Apl. dan Teor. Ilmu Komput.*, vol. 6, no. 2, pp. 55–60, 2023.

[6] D. Rahma Ente, S. Astuti Thamrin, H. Kuswanto, and S. Arifin, "KLASIFIKASI FAKTOR-FAKTOR PENYEBAB PENYAKIT DIABETES MELITUS DI RUMAH SAKIT UNHAS MENGGUNAKAN ALGORITMA C4.5 *," 2020.

[7] T. N. Rahman, "ANALISA ALGORITMA DECISION TREE DAN NAÏVE BAYES PADA PASIEN PENYAKIT LIVER," *J. FASILKOM*, vol. Volume 10, pp. 144–151, 2020.

[8] I. Tahyudin, I. M. Putra, and A. Y. Syafa'at, *DATA MINING DAN DATA WAREHOUSE MENGGUNAKAN APLIKASI KNIME*. Purwokerto: Zahira Media, 2021. [Online]. Available: https://play.google.com/store/books/details?id=7AtBEAAAQBAJ

[9] S. W. Audria, I. Farikhah, R. M. Saputra, and N. Purwati, "Prediksi Penyakit Paru-Paru Menggunakan Algoritma Naïve Bayes," *J. Data Sci. Methods Appl.*, vol. 01, no. 01, pp. 33–41, 2025, doi: 10.30873/jodmapps.v1i1.pp33-41.

[10] Z. D. R. Sari, Jasmir, and Y. Arvita, "Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5," vol. 4, no. April, pp. 827–834, 2024.

[11] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *J. SAINTEKOM*, vol. 13, no. 1, pp. 42–54, 2023, doi: 10.33020/saintekom.v13i1.352.

[12] F. A. H. Airi, T. Suprapti, and A. Bahtiar, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *E-Link J. Tek. Elektro dan Inform.*, vol. 18, no. 1, p. 73, 2023, doi: 10.30587/e-link.v18i1.5271.

[13] K. Akmal, A. Faqih, and F. Dikananda, "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 470–477, 2023, doi: 10.36040/jati.v7i1.6367.

[14] A. F. Abadi, N. Alamsyah, F. G. Retnanto, E. Daniati, and A. Ristyawan, "Penerapan Data Mining dalam Mengklasifikasi Penyakit Stroke Menggunakan Algoritma Naive Bayes," *INOTEK*, vol. 7, pp. 2549–7952, 2024.

[15] C. Maulana Sidiq, A. Faqih, and G. Dwilestari, "Algoritma Decision Tree C4.5 Digunakan Untuk Mengklasifikasikan Data Stroke," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1869–1874, 2024, doi: 10.36040/jati.v8i2.8388.

[16] H. Hendriyansyah, A. Irma Purnamasari, and T. Suprapti, "Penerapan Algoritma Decision Tree Dalam Klasifikasi Penyakit Stroke Otak," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3038–3043, 2024, doi: 10.36040/jati.v8i3.9602.

[17] R. Alfiyan, M. Hikmatyar, and S. S. Sundari, "Implementasi Metode Klasifikasi C4.5 Penyebab Faktor Resiko Penyakit Stroke," *Indones. J. Digit. Bus. J.*, vol. 4, no. October, pp. 37–48, 2024.

[18] B. A. C. Permana, M. Sadali, and R. Ahmad, "Penerapan Model Decision Tree Menggunakan Python Untuk Prediksi Faktor Dominan Penyebab Penyakit Stroke," *Infotek J. Inform. dan Teknol.*, vol. 7, no. 1, pp. 23–31, 2024, doi: 10.29408/jit.v7i1.23232.

[19] M. Muhsi, S. Suprapto, and R. Rofiuddin, "Node Selection Method for Split Attribute in C4.5 Algorithm Using the Coefficient of Determination Values for Multivariate Data Set," *J. Penelit. Pendidik. IPA*, vol. 9, no. 7, pp. 5574–5583, 2023, doi: 10.29303/jppipa.v9i7.4031.

[20] A. Bengnga and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," *Jambura J. Electr. Electron. Eng.*, vol. 4, no. 2, pp. 169–174, 2022, doi: 10.37905/jjeee.v4i2.14403.

[21] F. Kurniawati and D. Brahma Arianto, "Analisis Implementasi Seleksi Fitur Pada Klasifikasi Diabetes dengan Metode Corellation Matrix dan Algoritma Logistic Regression," *Inform. J. Ilmu Komput.*, vol. 19, no. 3, pp. 157–164, 2023, doi: 10.52958/iftk.v19i3.6019.

[22] E. N. R. Khakim, A. Hermawan, and D. Avianto, "Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine," *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 1, p. 158, 2023, doi: 10.26798/jiko.v7i1.771.

[23] J. Jiarpakdee, C. Tantithamthavorn, and C. Treude, "The impact of automated feature selection techniques on the interpretation of defect models," *Empir. Softw. Eng.*, vol. 25, no. 5, pp. 3590–3638, 2020, doi: 10.1007/s10664-020-09848-1.

[24] M. Faisal Fahrul and W. Hadikurniawati, "Klasifikasi Diabetes Pada Wanita Menggunakan Metode Naive Bayes Classifier," *J. Ilm. Inform.*, vol. 10, no. 01, pp. 70–73, 2022, doi: 10.33884/jif.v10i01.4705.

[25] B. A. Candra Permana and I. K. Dewi Patwari, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes," *Infotek J. Inform. dan Teknol.*, vol. 4, no. 1, pp. 63–69, 2021, doi: 10.29408/jit.v4i1.2994.

[26] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *J. Inform.*, vol. 9, no. 1, pp. 13–17, 2022, doi: 10.31294/inf.v9i1.11931.

[27] I. Ikhtiar, M. W. Rosyich, M. A. Ardhanu, D. S. Bastiana, D. Kurniawan, and S. Setyowatie, "Stroke Risk Factor Profile in an Urban Population: A Community-Based Descriptive Study in Mojo Sub-District, Surabaya, Indonesia," *Aksona*, vol. 3, no. 1, pp. 1–6, 2023, doi:

10.20473/aksona.v3i1.40764.
[28] TUNIK, R. NININGASIH, and E. YULIDANINGSIH, "FAKTOR-FAKTOR PENYEBAB DAN PENCEGAHAN TERJADINYA STROKE BERULANG," *Heal. J. Inov. Ris. Ilmu Kesehat.*, vol. 1, no. 2, pp. 101–108, 2022.
[29] T. G. Rahayu, "The Analysis of Stroke Risk Factors and Stroke Types," *Faletehan Heal. J.*, vol. 10, no. 01, pp. 48–53, 2023.
[30] C. Mirna Wati -, A. Charis Fauzan -, and H. -, "PERFORMANCE COMPARISON OF MUSHROOM TYPE CLASSIFICATION BASED ON MULTI-SCENARIO DATASET USING DECISION TREE C4.5 AND C5.0", doi: 10.34288/jri.v4i3.XXX.