Indonesian Journal of Digital Business

Journal homepage: https://ejournal.upi.edu/index.php/IJDB/index

Implementasi Metode Klasifikasi C4.5 Penyebab Faktor Resiko Penyakit Stroke

Rizqi Alfiyan[1], Missi Hikmatyar[2], Shinta Siti Sundari [3]

[1, 2] Teknik Informatika, Universitas Perjuangan Tasikmalaya

[3] Universitas Pendidikan Indonesia

Correspondence: E-mail:

[1] alfiyanrizqi77@gmail.com, [2] missi@unper.ac.id, [3]shintasiti@unper.ac.id

A B S T R A K

Stroke merupakan masalah kesehatan utama bagi masyarakat modern. Penyakit Stroke merupakan jenis penyakit yang mematikan dimana masuk kedalam 10 jenis penyakit yang paling mematikan. Untuk membantu mempercepat hasil diagnosa penderita stroke, maka dibutuhkan suatu algoritma klasifikasi yang dapat mengklasifikasikan suatu data yang banyak dalam waktu yang singkat. Di antara banyaknya algoritma klasifikasi, algoritma Decision Tree C4.5 dinilai cocok untuk digunakan dalam penelitian ini berdasarkan tujuan penelitian ini yaitu mengklasifikasi algoritma Decision Tree C4.5 yang memiliki kecepatan dan ketepatan yang tinggi untuk proses klasifikasi walaupun data yang dipakai bervolume besar. Pengujian dilakukan dengan membagi data menjadi dua set, yaitu data training (90% atau 80%) dan data testing (10% atau 20%). Hasil pengujian menunjukkan bahwa Rapidminer mencapai tingkat akurasi sebesar 93.50% pada pembagian data training 90% dan data testing 10%, serta meningkat menjadi 94.30% saat pembagian data training menjadi 80%. Di sisi lain, penggunaan Python memberikan tingkat akurasi yang sedikit lebih rendah, yaitu 92% pada data training 90% dan testing 10%, dan 91% pada data training 80% dan testing 20%. Meskipun demikian, baik Rapidminer maupun Python memberikan hasil yang cukup baik dalam memprediksi kasus stroke.

Informasi Artikel

Submitted/Received 19 Mei 2024
First Revised 1 September 2024
Accepted 15 October 2024
First Available online 28 October 2024
Publication Date 30 October 2024

Kata Kunci:
Data Mining, RapidMiner, Python Stroke, Decision Tree C4.5

1. PENDAHULUAN

Stroke merupakan masalah kesehatan utama bagi masyarakat modern. Pada saat ini, stroke semakin menjadi masalah serius yang dihadapi hampir diseluruh dunia. Hal tersebut dikarenakan serangan stroke yang mendadak dapat mengakibatkan kematian, kecacatan fisik dan mental baik pada usia produktif maupun usia lanjut. Penyakit Stroke merupakan jenis penyakit yang mematikan dimana masuk kedalam 10 jenis penyakit yang paling mematikan. Menurut World Health Organization (2024) Penyakit stroke membawa risiko kematian yang tinggi. Penyintas dapat mengalami kehilangan penglihatan bicara, kelumpuhan, dan kebingungan. Disebut stroke karena cara penyakitnya menyerang orang. Risiko serangan stroke lebih lanjut meningkat secara signifikan pada orang yang pernah mengalami stroke sebelumnya.

Untuk membantu mempercepat hasil diagnosa penderita stroke, maka dibutuhkan suatu algoritma klasifikasi yang dapat mengklasifikasikan suatu data yang banyak dalam waktu yang singkat. Di antara banyaknya algoritma klasifikasi, algoritma Decision Tree C4.5 dinilai cocok untuk digunakan dalam penelitian ini berdasarkan tujuan penelitian ini yaitu mengklasifikasi algoritma Decision Tree C4.5 yang memiliki kecepatan dan ketepatan yang tinggi untuk proses klasifikasi walaupun data yang dipakai bervolume besar.

Beberapa penelitian dari Fazrin Meila Azzahra Sofyan dkk (Sofyan et al. 2023). Berkonsentrasi untuk mengembangkan model prediksi penyakit stroke dengan menggunakan metode Knowledge Discovery in Databases(KDD). Dataset yang digunakan berasal dari situs Kaggle dengan 12 atribut dan 4.000 data pasien, namun hanya delapan atribut yang digunakan karena relevansinya dengan prediksi stroke. Algoritma Decision Tree C4.5 digunakan dengan pembagian data training sebesar 80% dan data testing sebesar 20%. Hasil penelitian menunjukkan akurasi sebesar 95%, recall sebesar 96%, dan presisi sebesar 99%. Dengan hasil yang cukup baik ini, algoritma C4.5 dapat efektif dalam prediksi penyakit stroke dan menunjukkan potensi untuk pengembangan lebih lanjut dengan menggunakan algoritma lain.

Penelitian serupa dilakukan oleh (Pambudi, Sriyanto, and Firmansyah 2022). Berkonsentrasi menggunakan algoritma Decision Tree C4.5. Tujuan penelitian ini adalah untuk menganalisis data mengenai faktor penyebab stroke dan menilai akurasi serta performa algoritma dalam bentuk confusion matrix dan nilai Area Under Curve (AUC). Hasil penelitian menunjukkan bahwa algoritma C4.5 berhasil memprediksi penyakit stroke dengan akurasi sebesar 96.05%, menunjukkan efektivitas algoritma dalam menganalisis dan klasifikasi faktor penyebab penyakit stroke.

Penelitian yang sama dilakukan oleh (Iskandar, Ernawati, and Widiastiwi 2022). Menyarankan menggunakan metode Random Forest. Tujuan utama penelitian ini adalah untuk mengevaluasi performa Random Forest dengan variasi jumlah pohon yang berbeda dalam mendiagnosis penyakit stroke. Hasil penelitian menunjukkan bahwa meskipun variasi jumlah pohon tidak signifikan meningkatkan akurasi, model dengan 90 pohon memberikan hasil optimal dengan akurasi 95,2%, sensitivity 4,1%, specificity 99,8%, precision 66,7%, dan F-measure 7,6%. Selain itu, nilai ROC Curve sebesar 0,8048 menunjukkan bahwa model tersebut masuk ke dalam kategori Good Classification. Ini menegaskan bahwa Random Forest efektif sebagai metode dalam mendiagnosis penyakit stroke dengan penggunaan optimal jumlah pohon.

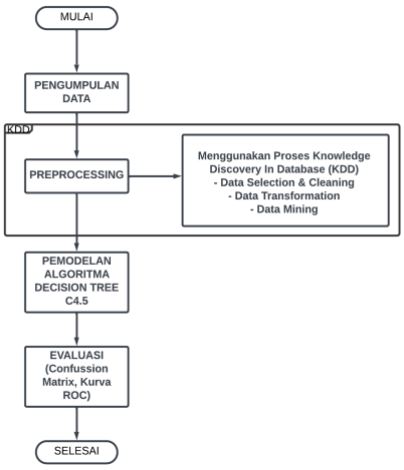
Penelitian serupa dilakukan oleh (Kohsasih and Situmorang 2022). Menyarankan untuk mengkaji perbandingan antara algoritma C4.5 dan Naïve Bayes dalam memprediksi Penyakit Cerebrovascular atau stroke. Dengan membagi dataset menjadi 60% data training dan 40% data testing, penelitian ini menunjukkan bahwa algoritma C4.5 memiliki performa yang lebih baik dibandingkan Naïve Bayes. Algoritma C4.5 mencapai tingkat akurasi sebesar 95%, dengan nilai presisi 90%, recall 95%, dan f1-score 93%. Sedangkan Naïve Bayes memiliki akurasi 91%, presisi 92%, recall 91%, dan f1-score 92%. Dalam hal log loss, Naïve Bayes memiliki nilai 0.205, sedangkan C4.5 memiliki nilai 0.190. Selain itu, spesifisitas Naïve Bayes adalah 0.213, sementara C4.5 memiliki nilai yang lebih rendah yaitu 0.047. Dengan demikian, algoritma C4.5 menunjukkan kinerja yang lebih unggul dalam memprediksi penyakit stroke dibandingkan dengan Naïve Bayes berdasarkan parameter-parameter evaluasi yang digunakan.

Penelitian yang dilakukan oleh (Sebastian and Juliane 2023). Berkonsentrasi pada perbandingan kinerja tiga algoritma klasifikasi data mining, yaitu Naïve Bayes, Decision Tree C4.5, dan K-Nearest Neighbor (KNN), dalam memprediksi penyakit stroke dengan menggunakan metode SMOTE upsampling. Hasil penelitian menunjukkan bahwa penerapan SMOTE upsampling memberikan pengaruh signifikan terhadap performa ketiga algoritma tersebut. Naïve Bayes menunjukkan selisih performa yang paling rendah dibandingkan dua algoritma lainnya, dengan selisih sebesar 53.48% untuk class precision, 63.13% untuk recall, 58.28% untuk F1 score, dan 0.02 untuk nilai AUC. Algoritma C4.5 dan KNN menunjukkan selisih performa yang lebih besar, di mana C4.5 memiliki selisih sebesar 80.93% untuk class precision, 79.34% untuk recall, 80.11% untuk F1 score, dan 0.37 untuk nilai AUC, sedangkan KNN memiliki selisih sebesar 83.53% untuk class precision, 59.14% untuk recall, 70.31% untuk F1 score, dan 0.28 untuk nilai AUC. Dengan demikian, Algoritma C4.5 menunjukkan inerja yang paling stabil dan konsisten dibandingkan dengan Naïve Bayes dan KNN dalam memprediksi penyakit stroke dengan teknik SMOTE upsampling.

Berdasarkan latar belakang yang telah dijelaskan di atas, maka pada penelitian ini akan menggunakan algoritma C4.5 untuk klasifikasi penyakit stroke.

2. METODE PENELITIAN

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.



Gambar 1. Siklus Hidup Proses Penelitian.

Langkah-langkah dalam metode ini adalah sebagai berikut. (bidin A 2017)

2.1. Pengumpulan Data

Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses. Dengan atribut dari penyakit stroke adalah Jenis Kelamin, Umur, Hipertensi, Riwayat Penyakit Jantung, Status Menikah, Tipe Pekerjaan, Tipe Tempat Tinggal, Rata-rata Kadar Glukosa, Indeks Masa Tubuh, Riwayat Merokok, dan Stroke.

Tabel 1. Atribut dan Nilai Kategori.

Atribut	Nilai
Jenis Kelamin	Laki-laki Perempuan
Umur	Umur Pasien
Hipertensi	0 = Tidak menderita penyakit hipertensi 1 = Menderita penyakit hipertensi
Penyakit Jantung	0 = Tidak mengidap penyakit jantung 1 = Mengidap penyakit
Status Menikah	Ya = Menikah Tidak = Belum Menikah
Tipe Pekerjaan	Swasta, Wiraswasta, Pemerintahan, Anak-anak dan Tidak Bekerja.
Tipe Tempat Tinggal	Perkotaan dan Pedesaan
Rata-rata Kadar Glukosa	55.25 – 271.74 Indikator Kadar Glukosa
Indeks Massa Tubuh (BMI)	12 – 78 Indikator Indeks Masa Tubuh
Riwayat Merokok	Merokok, Tidak pernah merokok, Tidak diketahui, dan Mantan Merokok
Stroke	0 = Tidak terkena penyakit stroke dan 1 = Terkena penyakit stroke

2.2. Preprocessing

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 5110 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (preparation data). (Zhang, Zhang, and Yang 2003)

a. Database
   Data studi kasus penyakit stroke bersumber dari repository https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Data yang digunakan pada penelitian ini ialah sebanyak 5110 record.
b. Data Cleaning
   Pada umumnya, data yang diperoleh, baik dari database maupun survey, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau hanya sekedar salah ketik. Data yang tidak relevan itu juga lebih baik dibuang karena keberadaanya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya.

Pembersihan data juga akan mempengaruhi performasi dari system data mining karena data yang akan ditangani akan berkurang jumlah dan kompleksitasnya.

c. Data Transformation
Transformasi data dilakukan dari bentuk excel dengan format .xlsx menjadi format .csv agar dapat di operasikan pada platform RapidMiner dan Pyton.

d. Data Mining
Bagian ini adalah proses mengeksplorasi dan menganalisa data dalam jumlah yang besar yang bertujuan untuk menemukan suatu pola atau informasi yang menarik dari data yang tersimpan dalam jumlah yang besar dengan menggunakan teknik atau metode tertentu (Hidayat 2015). Teknik, metode, atau algoritma yang tepat sangat bergantung pada tujuan dan proses knowledge discovery in databases (KDD) secara keseluruhan.

e. Pattern Evaluation
Dalam tahap ini, merupakan hasil dari teknik data mining berupa pola pola yang khas maupun model dievaluasi untuk menilai apakah ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai maka ada beberapa alternatif yang dapat diambil seperti menjadikanya umpan balik untuk memperbaiki data mining lain yang lebih sesuai, atau menerima hasilnya sebagai suatu hasil yang diluar dugaan yang mungkin bermanfaat.

2.3. Pemodelan

Tahap modeling untuk menyelesaikan klasifikasi penyakit stroke dengan menggunakan algoritma C4.5. Penelitian ini termasuk penelitian eksperimen, dimana penelitian ini dimulai dengan menentukan model yang digunakan, memasukan data training dan testing kedalam model dan mengujinya dengan tools rapidminer dan python.

Decision Tree berdasarkan algoritma C4.5 adalah teknik klasifikasi yang umum digunakan untuk mengekstrak hubungan yang relevan dalam data (Quinlan 1986). Algoritma C4.5 adalah program yang membuat pohon keputusan berdasarkan pada set data input berlabel. Kelebihannya adalah modelnya dapat dengan mudah ditafsirkan dan diimplementasikan dengan nilai kontinu dan nilai diskri. Algoritma C4.5 membagi data training dengan bantuan perolehan informasi. Atribut yang memiliki frekuensi tinggi dipertimbangkan untuk memisahkan data berdasarkan informasi yang tersedia dalam dataset.

Sebelum menghitung nilai gain terlebuh dahulu untuk mengetahui nilai entropy yaitu dengan persamaan untuk sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} - pi \cdot \log_2 pi$$

Keterangan:

S = Himpunan Kasus

A = Atribut

n = Jumlah partisi S

pi = Proporsi dari Si terhadap S

Persamaan yang digunakan untuk menghitung Information Gain:

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

Keterangan:

S = Himpunan kasus

A = Atribut

n = Jumlah partisi atribut

A |Si| = Jumlah kasus pada partisi ke-i

|S| = Jumlah kasus dalam S

Secara ringkas, tahapan algoritma Decision Tree dapat digambarkan sebagai berikut:

a. Menghitung nilai Information Gain dari setiap atribut
b. Memilih atribut yang memiliki nilai Information Gain paling besar
c. Membentuk simpul yang berisi atribut tersebut

Proses perhitungan Information Gain akan terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai Information Gain selanjutnya.

2.4. Evaluasi

Pada tahap ini dilakukan pengujian terhadap model-model yang dikomparasi untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan metode cross validation, confusion matrix, dan, kurva ROC. Validasi adalah proses mengevaluasi accuracy prediksi dari sebuah model, validasi mengacu untuk mendapatkan prediksi dengan menggunakan model yang ada kemudian membandingkan hasil yang diperoleh dengan hasil yang diketahui (Syariah and Ilmu n.d.). Mengevaluasi accuracy dari model klasifikasi sangat penting, accuracy dari sebuah model mengindikasikan kemampuan model tersebut untuk memprediksi class target. Untuk membuktikan performa algoritma yang digunakan, kita dapat menggunakan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP: Jumlah positif sebenarnya yang diklasifikasikan sebagai positif
FP: Jumlah negatif sebenarnya yang diklasifikasikan sebagai positif
TN: Jumlah negatif sebenarnya yang diklasifikasikan sebagai negative
FN: Jumlah positif sebenarnya yang diklasifikasikan sebagai negative

Untuk mengevaluasi model digunakan metode confusion matrix, dan kurva ROC

(Receiver Operating Characteristic). AUC dapat dibagi menjadi beberapa kelompok (Bramer 2016).

a. 0.90-1.00 = Excellent Classification
b. 0.80-0.90 = Good Classification
c. 0.70-0.80 = Fair Classification
d. 0.60-0.70 = Poor Classification
e. 0.50-0.60 = Failure

3. HASIL DAN PEMBAHASAN

Pengumpulan data yang didapat berasal dari kumpulan data penyakit stroke yang melibatkan data medis dan demografi dari pasien yang telah didiagnosis atau berisiko terkena penyakit stroke. Jumlah data yang digunakan sebanyak 5110 record, memiliki sebelas atribut dan satu atribut stroke sebagai label atau kelas (class) yang menyatakan 249 data pasien yang menderita penyakit stroke, 4861 data pasien yang tidak

menderita penyakit stroke. Dapat dilihat pada tabel 4.1

Tabel 2. Sample Data Training.

Jenis kelamin	Umur	HiperTensi	riwayat penyakit jantung	Status menikah	tipe pekerjaan	tipe tempat tinggal	Rata-rata kadar glukosa	Indeks masa tubuh	Riwayat merokok	Stroke
Laki-laki	67	tidak	ya	Meni Kah	Swasta	Perkotaan	228.69	36.6	mantan merokok	Ya
Perempuan	61	tidak	tidak	Menikah	Wiraswasta	Pedesaan	202.21		tidak pernah merokok	Ya
Laki-laki	80	tidak	ya	Meni Kah	Swasta	Pedesaan	105.92	32.5	tidak pernah merokok	Ya
Perempuan	49	tidak	tidak	Meni Kah	Swasta	Perkotaan	171.23	34.4	merokok	Ya
Perempuan	79	Ya	tidak	Meni Kah	Wiras Wasta	Pedesaan	174.12	24	tidak pernah merokok	Ya
Laki-laki	81	tidak	tidak	Meni Kah	Swasta	Perkotaan	186.21	29	mantan merokok	Ya
Laki-laki	74	ya	ya	Meni Kah	Swasta	Pedesaan	70.09	27.4	tidak pernah merokok	Ya
Perempuan	69	tidak	tidak	belum menikah	Swasta	Perkotaan	94.39	22.8	tidak pernah merokok	Ya
Perempuan	59	tidak	tidak	Meni Kah	Swasta	Pedesaan	76.15		tidak diketahui	Ya
Perempuan	78	tidak	tidak	Men Ikah	Swasta	Perkotaan	58.57	24.2	tidak diketahui	Ya
Laki-laki	3	Tidak	Tidak	Belum menikah	Anak-anak	Pedesaan	95.12	18	Tidak pernah Merokok	Tidak
Laki-laki	58	Ya	Tidak	Meni Kah	Swasta	Perkotaan	87.96	39.2	tidak diketahui	Tidak
Perempuan	8	Tidak	Tidak	Belum menikah	Swasta	Perkotaan	110.89	17.6	tidak diketahui	Tidak
Laki-laki	70	Tidak	Tidak	Meni Kah	Swasta	Pedesaan	69.04	35.9	Mantan Merokok	Tidak
Perempuan	14	Tidak	Tidak	Belum menikah	Tidak Bekerja	Pedesaan	161.28	19.1	Tidak Dike tahui	Tidak
Perempuan	47	Tidak	Tidak	Meni	Swa	Perko	210.95	50.1	Tidak Dike	Tidak

Puan				kah	sta	taan			tahui	
Perem				Meni	Swa	Perko			Mantan	
Puan	52	Tidak	Tidak	kah	sta	taan	77.59	17.7	Merokok	Tidak
									Tidak	
Perem				Meni	Wiras	Pede			Pernah	
Puan	75	Tidak	Ya	kah	wasta	saan	243.53	27	Merokok	Tidak
Perem				Meni	Swa	Pede				
Puan	32	Ya	Tidak	kah	sta	saan	77.67	32.3	Merokok	Tidak
Perem				Meni	Swa	Perko			Tidak	
Puan	74	Tidak	Tidak	kah	sta	taan	205.84	54.6	Diketahui	Tidak
Perem				Meni	Pemeri	Perko			Mantan	
Puan	79	Tidak	Tidak	kah	ntahan	taan	77.08	35	Merokok	Tidak
Pere				Meni	Swa	Pede			Tidak	
Puan	37	Tidak	ya	kah	sta	saan	57.08	22	Diketahui	Tidak
Laki				Meni	Swa	Perko			Mantan	
-laki	40	Tidak	Tidak	kah	sta	taan	73.5	26.1	Merokok	Tidak
									Tidak	
Pere				Meni	Swa	Perko			Pernah	
Puan	35	Tidak	Tidak	kah	sta	taan	95.04	42.4	Merokok	Tidak
Laki				Meni	Swa	Perko				
-laki	20	Tidak	Tidak	kah	sta	taan	84.62	19.7	Merokok	Tidak

Pada penelitian ini, eksperimen yang dilakukan bertujuan untuk mengetahui tingkat accuracy yang terbaik dalam algoritma C4.5 diantara data training 90% dan data testing 10% dan data training 80%

dan data tesing 20%. Setelah diolah dan menghasilkan model.

Tabel 3. Hasil Klasifikasi Algoritma Berdasarkan Data Training dan Data Testing.

Platform	**Algoritma**	***Data training***	***Data testing***	***Accuracy***
RapidMiner	C4.5	90	10	93.50%
		80	20	94.30%
Python		90	10	91%
		80	20	92%

Dari tabel 3 dapat kita lihat hasil algoritma C4.5 Berdasarkan pengujian dengan pembagian data training 80:20 dan 90:10, algoritma C4.5 menunjukkan kinerja yang baik dalam mengklasifikasikan data. Dalam percobaan dengan data training 80%, akurasi mencapai 94.30% menggunakan RapidMiner dan 92% dengan Python. Sedangkan pada data training 90%, algoritma ini mencatatkan akurasi 93.50% dengan RapidMiner dan 91% dengan Python. Kesimpulan ini menunjukkan konsistensi dalam performa algoritma C4.5 dalam berbagai konteks dan platform pengimplementasiannya, serta menyoroti pentingnya pembagian data yang proporsional untuk hasil klasifikasi yang optimal.

Model confusion matrix akan membentuk matrix yang terdiri dari true positif atau false positif dan true negatif atau false negatif. Berikut dibawah ini merupakan hasil confusion matrix dari

algoritma klasifikasi C4.5 untuk data training 90% data testing 10% pada platform RapidMiner didapatkan

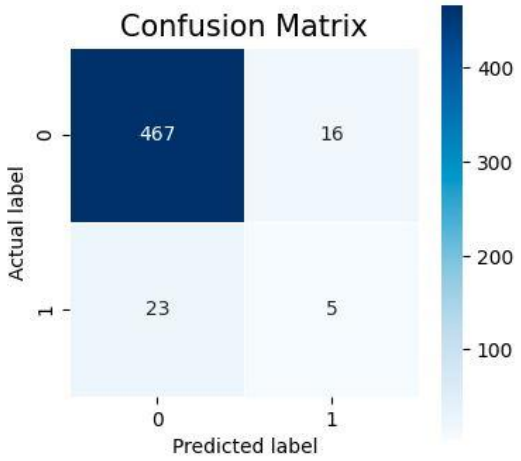
accuracy: 93.50%

	true Ya	true Tidak	class precision
pred. Ya	1	9	10.00%
pred. Tidak	24	474	95.18%
class recall	4.00%	98.14%	

Gambar 2. Confussion Matrix Algoritma C4.5 (Data Training 90%, Data Testing 10%) pada RapidMiner.

Penjelasan gambar 2 menunjukan bahwa, diketahui terdapat sebanyak 1 jumlah data yang diprediksi stroke dan pada kenyataannya memang menderita penyakit stroke, 474 data diprediksi tidak sroke dan pada kenyataannya memang tidak menderita penyakit stroke, 9 data yang diprediksi stroke tetapi kenyataannya tidak menderita penyakit stroke, dan 24 data diprediksi tidak menderita stroke tetapi kenyataannya menderita penyakit stroke. Berdasarkan gambar 2 menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma C4.5 untuk data training dan data testing 90% : 10% adalah sebesar 93.50%.

Model confusion matrix akan membentuk matrix yang terdiri dari true positif atau false positif dan true negatif atau false negatif. Berikut dibawah ini merupakan hasil confusion matrix dari algoritma klasifikasi C4.5 untuk data training 90% data testing 10% pada platform Python sebesar 92% didapatkan hasil pada gambar 3 sebagai berikut:



	precision	recall	f1-score	support
0	0.95	0.97	0.96	483
1	0.24	0.18	0.20	28
accuracy			0.92	511
macro avg	0.60	0.57	0.58	511
weighted avg	0.91	0.92	0.92	511

Gambar 3. Confussion Matrix Algoritma C4.5 (Data Training 90%, Data Testing 10%) pada Python.

Penjelasan pada gambar 3 dibawah ini menunjukan bahwa, tingkat accuracy dengan menggunakan algoritma C4.5 untuk perbandingan data training dan data testing 90% : 10% menggunakan python adalah sebesar 92%. Dari keselurahan 508 dataset yang diolah ini diambil dari data testing, sebanyak 5 jumlah data yang diprediksi stroke dan pada kenyataannya memang menderita penyakit stroke, 467 data diprediksi tidak stroke dan pada kenyataannya memang tidak menderita penyakit stroke, 16 data yang diprediksi stroke tetapi kenyataannya tidak menderita penyakit stroke, dan 23 data diprediksi tidak menderita stroke tetapi kenyataannya stroke.

Model confussion matrix yang kedua dengan perbandingan data training dan data testing 80% : 20% pada platform RapidMiner sehingga didapatkan hasil pada gambar 4 sebagai berikut:

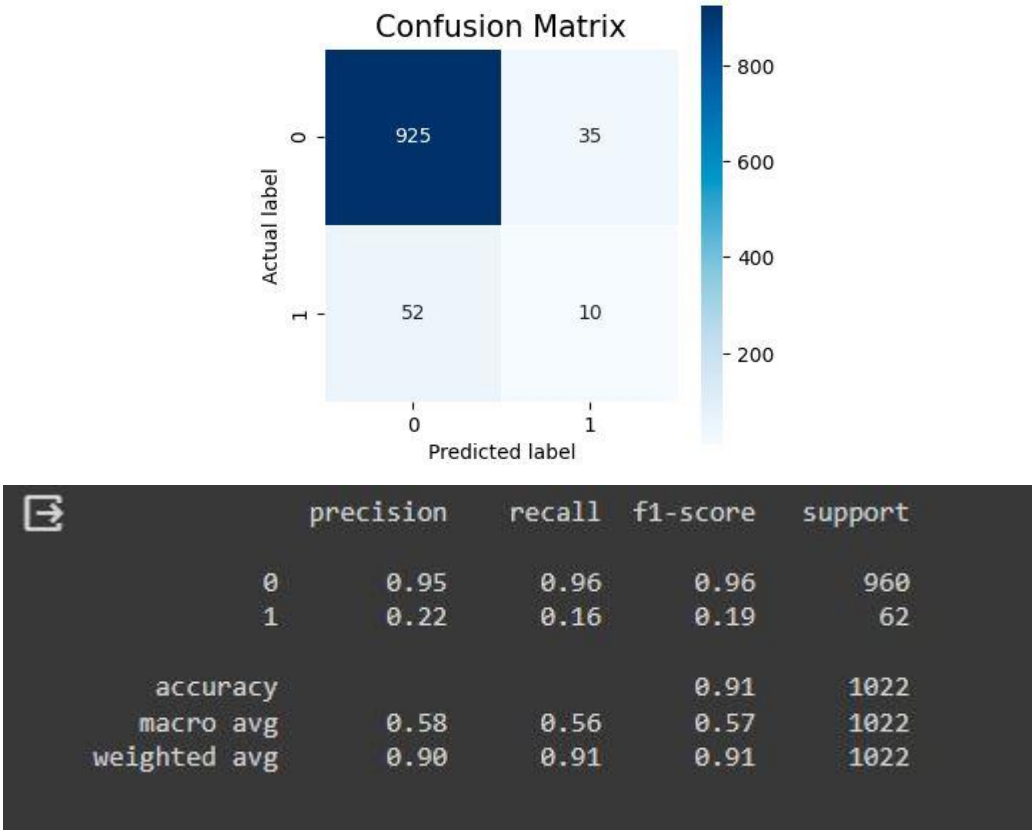
accuracy: 94.30%

	true Ya	true Tidak	class precision
pred. Ya	1	9	10.00%
pred. Tidak	49	958	95.13%
class recall	2.00%	99.07%	

Gambar 4. Confussion Matrix Algoritma C4.5 (Data Training 80%, Data Testing 20%) pada RapidMiner.

Penjelasan dari gambar 4 menunjukan bahwa, diketahui dari keselurahan 1017 dataset yang diolah ini diambil dari data testing, sebanyak 1 jumlah data yang diprediksi stroke dan pada kenyataannya memang menderita penyakit stroke, 958 data diprediksi tidak stroke dan pada kenyataannya memang tidak menderita penyakit stroke, 9 data yang diprediksi stroke tetapi kenyataannya tidak menderita penyakit stroke, dan 49 data diprediksi tidak menderita stroke tetapi kenyataannya stroke. Berdasarkan gambar 4.16 menunjukan bahwa, tingkat akurasi yang paling tinggi pada platform RapidMiner menggunakan algoritma C4.5 untuk data training dan data testing 80% : 20% adalah sebesar 94.30%.

Model confussion matrix yang kedua dengan perbandingan data training dan data testing 80% : 20% pada platform Python sebesar 91% sehingga didapatkan hasil pada gambar 5 sebagai berikut:



	precision	recall	f1-score	support
0	0.95	0.96	0.96	960
1	0.22	0.16	0.19	62
accuracy			0.91	1022
macro avg	0.58	0.56	0.57	1022
weighted avg	0.90	0.91	0.91	1022

Gambar 5. Confussion Matrix Algoritma C4.5 (Data Training 80%, Data Testing 20%) pada pyhton.

Untuk perbandingan data training dan data testing 80% : 20% menggunakan python adalah sebesar 91%. Dari keselurahan 1017 dataset yang diolah ini diambil dari data testing, sebanyak 10 jumlah data yang diprediksi stroke dan pada kenyataannya memang menderita penyakit stroke, 925 data diprediksi tidak stroke dan pada kenyataannya memang tidak menderita penyakit stroke, 35 data yang diprediksi stroke tetapi kenyataannya tidak menderita penyakit stroke, dan 52 data diprediksi tidak menderita stroke tetapi kenyataannya stroke.

4. KESIMPULAN

Berdasarkan hasil pengujian dan analisis bahwa pengujian ini bertujuan untuk mengetahui model algoritma C4.5 yang memiliki akurasi untuk klasifikasi prediksi penyebab resiko penyakit stroke, diantaranya:

Hasil pengujian menunjukkan tingkat akurasi algoritma C4.5 dengan data training 90% dan data testing 10% sebesar 93.50%, sedangkan dengan data training 80% dan data testing 20% sebesar 94.30%.

Pengujian menggunakan python juga dilakukan untuk membandingkan tingkat akurasi. Hasilnya menunjukkan tingkat akurasi sebesar 92% untuk perbandingan data training 90% dan data testing 10%, dan akurasi sebesar 91% untuk perbandingan data training 80% dan data testing 20%.

Faktor yang paling dominan dalam prediksi resiko penyakit stroke berdasarkan Algoritma C4.5 adalah usia, dengan nilai gain sebesar 3.295977.

5. DAFTAR PUSTAKA

bidin A. 2017. "Опыт Аудита Обеспечения Качества и Безопасности Медицинской Деятельности в Медицинской Организации По Разделу «Эпидемиологическая БезопасностьNo Title." *Вестник Росздравнадзора* 4(1): 9–15.

Bramer, Max. 2016. *Introduction to Data Mining*.

Hidayat, Muhammad Mahaputra. 2015. "Data Mining Data Mining." *Mining of Massive Datasets* 2(January 2013): 5–20. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part.

Iskandar, Nur Aliffiyanti, Iin Ernawati, and Yuni Widiastiwi. 2022. "Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest." *Seminak Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*: 432–41. https://conference.upnvj.ac.id/index.php/senamika/article/view/2190.

Kohsasih, Kelvin Leonardi, and Zakarias Situmorang. 2022. "Comparative Analysis of C4.5 and Naïve Bayes Algorithms in Predicting Cerebrovascular Disease." *Jurnal Informatika* 9(1): 13–17.

Pambudi, Estian R, Sriyanto, and Firmansyah. 2022. "Teknika 16 (02): 221-226." *Ijccs* 16, No.02(x): 221–26.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1(1): 81–106.

Sebastian, Ronald, and Christina Juliane. 2023. "Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method." *JUITA : Jurnal Informatika* 11(2): 311.

Sofyan, Fazrin Meila Azzahra, Affani Putri Riyandoro, Devi Fitriani Maulana, and Jajam Haerul Jaman. 2023. "Penerapan Data Mining Dengan Algoritma C5.0 Untuk Prediksi Penyakit Stroke." *J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD)* 6(2): 619.

Syariah, Kelembagaan Bank, and Graha Ilmu. *No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title*.

Zhang, Shichao, Chengqi Zhang, and Qiang Yang. 2003. 17 Applied Artificial Intelligence *Data Preparation for Data Mining*.