Teknika 17 (2): 419 – 426

JURNAL TEKNIKA
ISSN: 0854-3143
e-ISSN: 2622-3481
Journal homepage: http://jurnal.polsri.ac.id/index.php/teknika
Journal Email: teknika@polsri.ac.id

Perbandingan Kinerja Algoritma C4.5. Naive Bayes Dan Random Forest Dalam Prediksi Penyakit Jantung

Khodijah[1], Sriyanto*[2]

[1,2]IIB Darmajaya Lampung; JL.ZA Pagar Alam Gedung Meneng No 93 .35142, telp 0721-787214
Email Penulis Korespondensi: khodijahdj05@gmail.com*

Abstrak

Data mining adalah proses menemukan pola dan pengetahuan dari data yang berjumlah besar. Salah satu bagian penting pada data mining adalah pengklasifikasian data. Klasifikasi digunakan untuk menggolongkan data berdasarkan sifat data yang sudah dikenali masing-masing kelasnya. Ada berbagai macam teknik yang digunakan untuk mengklasifikasikan data, diantaranya yaitu C4.5,Naive bayes.dan Random Forest.. Berdasarkan beberapa peneliti, metode C4.5 dan Naive Bayes memiliki kinerja yang baik sehingga dibuat sistem dengan tujuan untuk membandingkan kinerja C4.5 dan Naive Bayes. Data yang akan digunakan pada sistem yang akan dibuat adalah penyakit jantung. Berdasarkan hasil penelitian diketahui bahwa kinerja dari algoritma Radom Forest lebih baik dibandingkan dengan kinerja dari algoritma C4.5 dan Naïve Bayes . Hal ini dapat dilihat dari nilai akurasi 98,60%, recal 99,44%l, precision 98,60%yang dihasilkan oleh Radom Forest lebih besar dibandingkan oleh C4.5 dan Naïve Bayes.

Kata kunci*—Radom Forest, Naive Bayes, C4.5, akurasi*

Abstract

Data mining is the process of finding patterns and knowledge from large amounts of data. One important part of data mining is data classification. Classification is used to classify data based on the characteristics of the data that are recognized by each class. There are various techniques used to classify data, including C4.5, Naive Bayes and Random Forest. Based on several researchers, the C4.5 and Naive Bayes methods have good performance so a system was created with the aim of comparing the performance of C4.5 and Naive Bayes. The data that will be used in the system that will be created is heart disease. Based on the research results, it is known that the performance of the Radom Forest algorithm is better than the performance of the C4.5 and Naïve Bayes algorithms. This can be seen from the accuracy value of 98.60%, recall of 99.44%l, precision of 98.60% produced by Radom Forest which is greater than that of C4.5 and Naïve Bayes.

Keywords—*Radom Forest, Naive Bayes, C4.5, accuracy*

1. PENDAHULUAN

Salah satu penyakit yang sering terjadi dan tidak menular (PTM) adalah penyakit jantung, Penyakit jantung tersebut merupakan penyakit merupakan kondisi yang terjadi ketika pembuluh darah utama yang menyuplai darah ke jantung mengalami kerusakan dan tidak dapat bekerja dengan semestinya hal tersebut dikarenakan banyak faktor. Tumpukan kolesterol pada pembuluh darah serta proses peradangan diduga salah satu menjadi faktor penyebab penyakit ini. Ketika terjadi penumpukan kolesterol (plak), pembuluh darah koroner akan menyempit sehingga aliran darah dan suplai oksigen menuju jantung pun akan terhambat sehingga mempengaruhi kestabilan proses bekerja jantung. Kurangnya aliran darah ini akan menyebabkan rasa nyeri pada dada (*angina*) dan sesak napas, hingga suatu saat terjadi hambatan total pada aliran darah menuju jantung atau yang disebut juga dengan serangan jantung. Menurut banyak penelitian penyakit jantung koroner termasuk salah satu penyebab kematian tertinggi di dunia.

Federasi Jantung Dunia memperkirakan angka kematian akibat penyakit jantung koroner di Asia Tenggara mencapai 1,8 juta kasus pada tahun 2014. Di Indonesia sendiri pada tahun 2013 tercatat ada setidaknya 883.447 orang yang terdiagnosis penyakit jantung koroner di Indonesia dengan mayoritas penderita berusia 55-64 tahun. Angka kematian akibat penyakit jantung pun menjadi cukup tinggi, yakni sekitar 45 persen dari seluruh angka kematian di Indonesia [1].

Pada acara peringatan hari jantung sedunia (*world heart day*) yang dicetuskan pertama kali oleh World Heart Federation pada tahun 2000 menginformasikan kepada orang-orang di seluruh dunia bahwa penyakit jantung dan stroke merupakan penyebab utama kematian di dunia yang saat ini di klaim mencapai 17,3 juta kematian setiap tahunnya. Angka kematian akibat penyakit jantung tersebut diperkirakan akan terus meningkat hingga mencapai 23,3 juta pada tahun 2030. Berdasarkan Hasil Riset Kesehatan Dasar (Riskesdas) Kemenkes RI Tahun 2013, prevalensi penyakit jantung koroner di Indonesia mencapai 0,5% dan gagal jantung sebesar 0,13% dari total penduduk berusia 18 tahun keatas sehingga harus penyakit tersebut harus selalu diwaspadai (dinkes.inhukab.go.id, 2015). Banyak faktor yang dapat meningkatkan risiko terkena penyakit jantung. Faktor risiko tersebut terdiri dari faktor risiko yang tidak dapat dimodifikasi seperti riwayat keluarga, umur serta jenis kelamin dan faktor risiko yang dapat dimodifikasi seperti hipertensi, kebiasaan merokok, penyakit diabetes, dislipidemia, obesitas, kurang aktifitas fisik, pola makan serta stres. Penyakit pembuluh darah, atau disebut dengan penyakit kardiovaskular atau disebut juga penyakit jantung. Penyakit ini berhubungan dengan proses aterosklerosis, yaitu suatu kondisi pada organ tubuh yang terjadi ketika zat yang disebut plak menumpuk di dinding arteri. Penumpukan ini menyebabkan penyempitan pada arteri sehingga aliran darah terhenti atau tersumbat dan tidak dapat beredar dengan semestinya, hal ini dapat mengakibatkan serangan jantung atau stroke [2].

Salah satu bagian penting dari pengobatan atau tindakan medis adalah pengambilan keputusan dan proses klasifikasi atau prediksi pada suatu hal yang menjadi fokus seperti pendeteksian penyakit, namun klasifikasi medis atau prediksi biasanya merupakan proses yang sangat kompleks dan sulit dilakukan jika tak dapat mengetahui metode yang tepat dan terbaik dalam memberikan solusinya[3] Tantangan yang dihadapi oleh organisasi kesehatan adalah mendiagnosa pasien dengan benar, diagnosa atau prediksi yang buruk dapat menyebabkan konsekuensi yang mendatangkan malapetaka yang kemudian tidak dapat diterima. Untuk menjawab tantangan tersebut beberapa penelitian telah dilakukan dalam bidang [4]kesehatan untuk mendapatkan prediksi penyakit dengan lebih akurat, namun belum diketahui metode apa yang paling akurat dalam memprediksi penyakit pasien. Berbagai macam teknik analisa yang secara konvensional dan manual yang selama ini digunakan tidak lagi begitu efektif digunakan untuk hal mendiagnosa suatu penyakit. Seiring dengan perkembangan teknologi dan system. berbasis pengetahuan terutama medis tuntutan akan adanya penggunaan sistem pengetahuan berbasis komputer sebagai teknik analisa dalam mendiagnosa penyakit menjadi semakin penting dan harus selalu dikembangkan. Oleh karenanya, saat inilah waktu yang tepat untuk mengembangkan sistem pengetahuan berbasis komputer yang modern, efektif dan efisien dalam mendiagnosa masalah penyakit[5].

Oleh karena itu, penelitian ini dilakukan untuk membantu menyelesaikan permasalahan tersebut dengan data mining klasifikasi untuk penyakit jantung, diperlukan suatu metode atau teknik yang dapat mengolah data-data yang sudah ada. Salah satu metodenya menggunakan teknik data mining klasifikasi. Penggunaan data mining Algoritma C4.45, Naive Bayes dan Random Forest dalam Klasifikasi Penderita Penyakit Jantung sebagai pilihan untuk diagnosa penyakit jantung dapat menjadi alternatif pilihan yang tepat, tetapi sampai saat ini belum diketahui algoritma klasifikasi yang paling akurat dalam prediksi penyakit jantung.

Pada penelitian sebelumnya membandingkan algoritma klasifikasi data mining Naive Bayes Berbasis PSO untuk deteksi penyakit jantung. Pengukuran dengan Naives Bayes menghasilkan akurasi 82.14%, sementara dengan Naive Bayes Berbasis Particle Swarm Optimization akurasi meningkat menjadi 92.86%. [6]

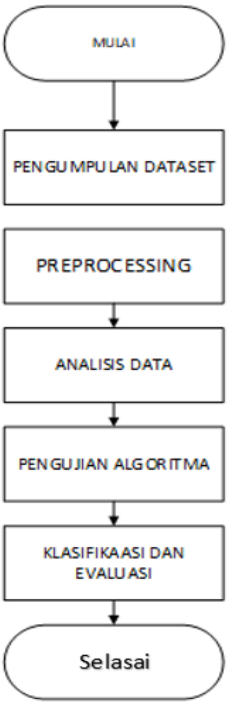
Pada penelitian algoritma menggabungkan k-NN dengan algoritma genetika untuk klasifikasi yang lebih efektif. Algoritma genetika melakukan proses yang kompleks dan berikan solusi optimal. Hasil percobaan menunjukkan bahwa algoritma kami Meningkatkan akurasi diagnosis penyakit jantung[7].

Berdasarkan penelitian tersebut penulis melakukan penelitian dengan melakukan perbandingan data mining dengan metode Algoritma C4.5,Naive Bayes dan Random Forest dalam Klasifikasi Penderita Penyakit Penyakit Jantung untuk mengetahui algoritma yang memiliki memiliki akurasi yang lebih tinggi dalam hal ini yaitu prediksi penyakit jantung.

2. METODE PENELITIAN

Adapun perancangan pada penelitian telah disusun sehingga alur dari proses pnelitian ini sesuai dengan yang telah ditentukan.

Metode penelitian kuantitatif digunakan untuk meneliti pada populasi atau sample tertentu, pengumpulan data menggunakan instrumen penelitian, analisis data bersifat kuantitatif/statistik dengan tujuan untuk menguji hipotesis yang telah ditentukan. Metode penelitian kuantitatif disebut juga dengan metode discovery karena dengan metode ini dapat ditemukan dan dikembangkan berbagai hal baru sehingga memunculkan ilmu pengetahuan baru [2]



Gambar .1 Tahapan Penelitian[8]

2.1 Pengumpulan Data

Banyak penelitian yang telah dilakukan untuk memprediksi penyakit gagal jantung, namun belum diketahui algoritma mana yang paling akurat. Oleh karena itu dalam penelitian ini akan dilakukan komparasi algoritma Naive Bayes dan C4.5 untuk mengetahui algoritma mana yang lebih akurat dalam memprediksi penyakit gagal jantung. Sampel Penelitian Sampel dari Penelitian ini adalah data profil Prediksi penyakit gagal jantung, data tersebut yang bersifat publik yang didapatkan dari Kaggle.com (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction) Data tersebut berisi 918 Observasi dengan 12 Atribut.

2.2 Pengolahan Awal Data

Pada tahap ini data yang masih berupa nilai numerik dan juga continue ditransformasikan kedalam bentuk kategorikal dan dibuat skala atau interval sehingga menghasilkan range yang lebih kecil sebagai bahan pembelajaran algoritma Naive Bayes dan C4.5 akan lebih mudah di klasifikasikan dengan menggunakan Rapid Miner sehingga memudahkan dalam memproses klasifikasi pada dataset yang telah tersedia sebelumnya. Ada 12 atribut yang digunakan dalam mendiagnosa penyakit jantung, yaitu: *Age*, *Sex*, *Chest paint type*, *resting blood pressure*, *Serum cholesterol* dalam mg/dl, *fasting blood sugar*>120 mg/dl, *resting electrocardiographic result*, *Maximum heart rate achieved*, *Exercise induced angina*, *oldpeak*, *The slope of the peak exercise ST segment*, *Number of major vessels*, dan *Thal*.

2.3 Preprocesing

Preprocessing merupakan suatu tahap untuk mempersiapkan data yang telah diperoleh dari tahap pengumpulan data sebelumnya sebelum data tersebut digunakan untuk tahap selanjutnya. Persiapan data yang dilakukan berupa membersihkan data dengan menghilangkan noise, menghapus data duplikat, memeriksa data untuk ketidakkonsistenan, dan memperbaiki kesalahan dalam data [9].

2.4 Analisis Data

Tahap analisis adalah prosedur sistematis untuk menentukan data yang sesuai dan cara terbaik untuk memanfaatkannya. Tahap ini melibatkan kombinasi metode atau teknik dalam mengolah data. Proses analisis data dimulai dengan membandingkan hasil eksperimen. Data yang telah dikumpulkan kemudian dibersihkan dan dinormalisasi sebelum dilakukan transformasi data pengklusteran. Hasil transformasi ini akan dibagi menjadi beberapa kelompok atau grup data.

2.5 Pengujian Algoritma Decision Tree C4.5

Pengujian data yang digunakan dalam algoritma C4.5 adalah dataset prediksi gagal jantung yang didapatkan dari Kaggle [10].

$Gaint\ (S.A\ ) = Entropy(S)\frac{Si}{S}$* Entropy (Si) …… (.1)

Keterangan:
S: Himpunan kasus
A: Data Atribut
n: Jumlah partisi di dalam atribut
|Si|: Jumlah kasus pada partisi ke-i
|S|: Jumlah kasus

$Entropy\ (S) = \sum_{pi}^{n}$ -pi=$\log_2$ pi …… (2.)

Keterangan:
S: Himpunan kasus
n: Jumlah partisi dalam atribut
pi: Proposi dari Si terhadap S

2.6 Pengujian Algoritma Naïve Bayes

Pengujian data yang digunakan dalam algoritma Naïve Bayes adalah dataset prediksi gagal jantung yang didapatkan dari Kaggle[9]

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad \text{.....(3)}$$

Keterangan:

X	= data dengan kelas yang belum diketahui
H	= hipotesis data X, merupakan suatu kelas yang spesifik
P(H|X)	= probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H)	= probabilitas hipotesis H (posteriori probability)
P(X|H)	= probabilitas X berdasar kondisi H
P(X)	= probabilitas dari X

atau

$$Posterior\ Probability = \frac{Prior\ Probability\ X\ likelihood}{evidance} \quad \text{.....(4)}$$

2.7 Pengujian Algoritma Radom Forest

Pengujian data yang digunakan dalam algoritma Naïve Bayes adalah dataset prediksi gagal jantung yang didapatkan dari Kaggle

$l(y) = argmaxc(\sum Ihn(y)=c\ N\ n=1\ )$ ...(5)

keterangan

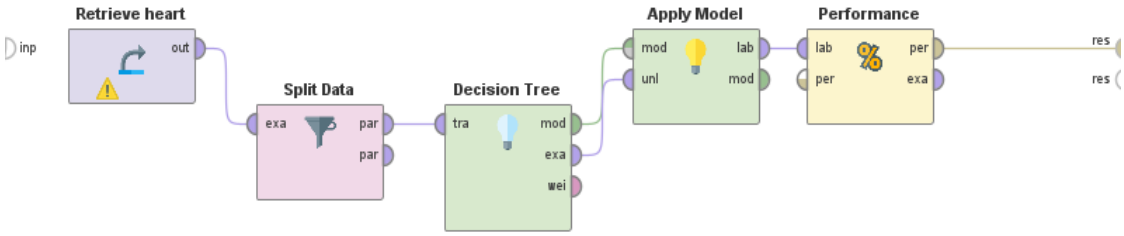
Dimana I adalah fungsi indikator dan *hn* adalah tree ke-n dari RF (Liparas, 2014). Random Forest memiliki mekanisme internal yang menyediakan estimasi dari generalization error-nya sendiri yang disebut out-of-bag (OOB) error estimate. Dalam pembentukan tree hanya 2/3 dari data asli yang digunakan dalam pengambilan sampel bootstrap. Sedangkan 1/3 sisanya diklasifikasikan oleh tree yang terbentuk dan digunakan untuk menguji performanya [10].

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini penulis akan membahas mengenai hasil dari penelitian, data tersebut akan dihitung menggunakan feature selection dengan menggunakan algoritma Decision Tree C4.5 ,Algoritma Nive bayes dan Algoritma Radom Forest yang kemudian akan diuji menggunakan Split Validation menggunakan algoritma Decision Tree C.45 ,Algoritma Nive bayes dan Algoritma Radom Forest Dataset yang digunakan merupakan dataset publik yang berasal dari https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction Yang sudah melalui proses preprocessing data dan analisis data.

Penerapan data pada Rapidminer untuk klasifikasi dengan menggunakan split validation dengan nilai akurasi, precision, confusion matrix atau nilai recall dan nilai AUC dengan pembagian Data training dan testing sebesar 70:30. dapat dilihat pada gambar berikut:

3.1 Algoritma C4.5



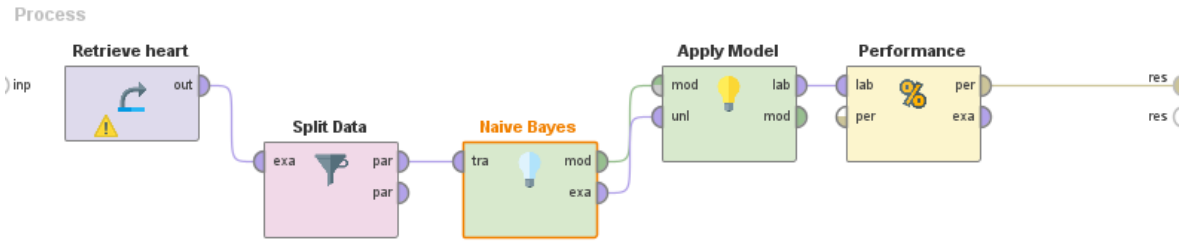
Gambar 2 Proses menggunakan Algoritma C4.5 Split Validation

Split validation Algoritma C4.5 dengan nilai akurasi, precision, confusion matrix atau nilai recall dan nilai AUC dengan pembagian Data training dan testing sebesar 70:30. dapat dilihat pada Table berikut:

Table .1 Nilai Akurasi, Precision, Confusion matrix atau nilai recall dan nilai AUC

Algoritma	**Akurasi**	**AUC**	**Recall**	**Precission**
Algoritma C4.5	**88,34%**	**0,903**	**89,80%**	**89,80%**

3.2 Algoritma Naïve Bayes



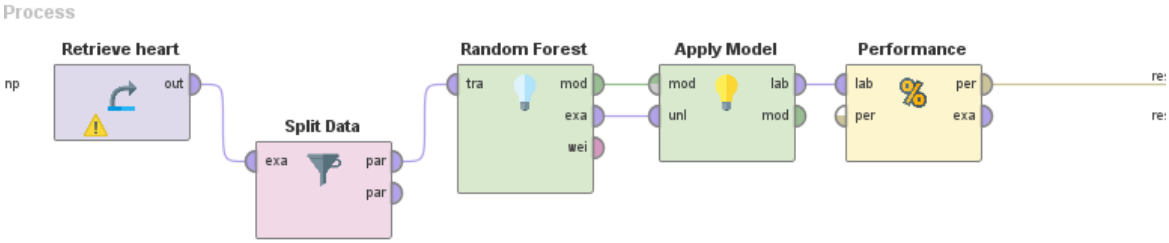
Gambar 3 Proses menggunakan Algoritma Naïve Bayes Split Validation

Split validation Algoritma Naïve Bayes dengan nilai akurasi, precision, confusion matrix atau nilai recall dan nilai AUC dengan pembagian Data training dan testing sebesar 70:30. dapat dilihat pada Tabel 2.

Tabel 2 Nilai Akurasi, Precision, Confusion matrix atau nilai recall dan nilai AUC

Algoritma	**Akurasi**	**AUC**	**Recall**	**Precission**
Algoritma Naïve Bayes	**87,25%**	**0,934**	**88,20%**	**88,70%**

3.3 Algoritma Random Forest



Gambar .4 Proses menggunakan Algoritma Random Forest Split Validation

Split validation Algoritma Random Forest dengan nilai akurasi, precision, confusion matrix atau nilai recall dan nilai AUC dengan pembagian Data training dan testing sebesar 70:30. dapat dilihat pada Tabel 3.

Table 3.Nilai Akurasi, Precision, Confusion matrix atau nilai recall dan nilai AUC

Algoritma	**Akurasi**	**AUC**	**Recall**	**Precission**
Algoritma Random Forest	**98,60%**	**1.000**	**99,44%**	**98,60%**

4. KESIMPULAN

Dari penjelasan diatas ini dapat diambil kesimpulan penelitian dengan menggunakan Algoritma C.45 dan Algoritma Naïve Bayes dan Random Forest .Pengujian dalam penelitian ini menggunakan Accuracy untuk melihat akurasi perbandingan yang didapatkan dari hasil pengujian. Dari nilai akurasi yang terdapat diatas, dapat diketahui algoritma yang lebih baik digunakan pada perbandingan penyakit jantung Kesimpulan akhir dari peerbandingan algoritma C.45 Mendapatkan nilai Kurasi 88,34%,Algoritma Naïve Bayes Mendapatkan nilai akurasi 87.25% dan Random Forest Mendapatkan Nilai akurasi 98,60 , tingkat akurasi 98,60%.yang terbaik pada algoritma Random Forest.

5. SARAN

Dataset *public* masih terdapat error sehingga diharapkan nantinya akan diperoleh analisis yang lebih tepat. Selain itu untuk meningkatkan nilai akurasi pada prediksi penyakit jantung menggunakan algoritma C4.5 Algoritma Naïve Bayes dan Random Forest dalam pengimprove akurasi dapat mengubah parameter sampling linier disesuaikan dengan dataset. Efisien dan menghasilkan akurasi yang optimum. Namun apabila data missing value tersebut tidak dapat dihindarkan maka untuk penelitian selanjutnya dapat dilakukan pengembangan metode yang lain untuk data missing value dan penentuan parameter tanpa trial and error yang diharapkan nantinya akan memberikan akurasi yang lebih tinggi

.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tim Redaksi Jurnal Teknika Politeknik Negeri Sriwijaya yang telah memberi memberi kesempatan, sehingga artikel ilmiah ini dapat diterbitkan.

DAFTAR PUSTAKA

[1] J. Brier and lia dwi jayanti, “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,” vol. 21, no. 1, pp. 1–9, 2020, [Online]. Available: http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203

[2] hidayah nur umi and dkk, “Hidayah Nur Umi Dkk,” *Nalisis Metod. K Nearest Neighbor Terhadap Klasifikasi Data Pasien Penderita Gagal Jantung*, vol. 28, no., pp. 145–158, 2021, [Online]. Available: http://www.riss.kr/link?id=A99932365

[3] T. Suharsono, Y. Krisna, and L. Sukmarini, “Dampak Home Based Exercise Training Terhadap Kapasitas Fungsional Pasien Gagal Jantung Di Rsud Ngudi Waluyo Wlingi,” *Ilmu Keperawatan*, vol. 1, no. 1, pp. 1–18, 2013.

[4] S. Hasbrima, S. E. Rahayuningsih, and D. Hilmanto, “Korelasi antara Neutrophil-Lymphocyte Ratio dan NT-proBNP pada Pasien Gagal Jantung Anak Akibat Penyakit Jantung Rematik,” *Sari Pediatr.*, vol. 23, no. 3, p. 191, 2021, doi: 10.14238/sp23.3.2021.191-6.

[5] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, “Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi,” *J. Teknoinfo*, vol. 14, no. 2, p. 115, 2020, doi: 10.33365/jti.v14i2.679.

[6] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, “Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke,” *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 1, p. 155, 2020, doi: 10.24114/cess.v5i1.15225.

[7] D. Prajarini, S. Tinggi, S. Rupa, D. Desain, and V. Indonesia, “Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit,” *Informatics J.*, vol. 1, no. 3, p. 137, 2016.

[8] A. Syukron, “Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” vol. 10, no. 1, pp. 47–50, 2023.

[9] Julia Triani, Yovi Pratama, and E. Yanti, “Komparasi Dalam Prediksi Gagal Jantung Dengan Menggunakan Metode C4.5 dan Naïve Bayes,” *J. Inform. Dan Rekayasa Komputer(JAKAKOM)*, vol. 3, no. 1, pp. 394–402, 2023, doi: 10.33998/jakakom.2023.3.1.759.

[10] E. Nurlia and U. Enri, “Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5,” *J. Tek. Inform. Musirawas) Elin Nurlia*, vol. 6, no. 1, p. 42, 2021.