



Perbandingan algoritma C4.5, Naïve Bayes, dan K-Nearest Neighbors untuk prediksi penyakit jantung

Mohamad Syafri Tuloli, Titin Seh Kinanti, Lanto Ningrayati Amali

Program Studi Sistem Informasi, Universitas Negeri Gorontalo, Indonesia

Riwayat Artikel:

Diterima 14 April 2025

Direvisi 25 April 2025

Disetujui 26 April 2025

Kata Kunci:

Prediksi penyakit jantung

C4.5

Naïve Bayes

K-Nearest Neighbors

Klasifikasi

ABSTRACT. Heart disease is one of the leading causes of death worldwide, making early detection crucial to reducing fatality risks. This study aims to compare the performance of three classification algorithms, namely C4.5, Naïve Bayes, and K-Nearest Neighbors (KNN), in predicting heart disease. The dataset used is the Heart Failure Prediction Dataset obtained from Kaggle and the UCI Repository, comprising a total of 1,211 patient records. After preprocessing, the data was split into training and testing sets with a 70:30 ratio. Model performance was evaluated using accuracy, precision, and recall metrics based on the confusion matrix. The results indicate that the C4.5 algorithm achieved the best performance with an accuracy of 81.07%, outperforming Naïve Bayes (79.10%) and KNN (75.68%). C4.5 also demonstrated a higher recall rate in detecting positive heart disease cases, suggesting its effectiveness in handling clinical datasets with characteristics similar to those used in this study. The main contribution of this research is providing a recommendation for utilizing the C4.5 algorithm in developing decision support systems for early heart disease detection. These findings are expected to help improve early diagnostic accuracy and support faster and more accurate medical decision-making.

ABSTRAK. Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga deteksi dini menjadi sangat penting untuk mengurangi risiko fatalitas. Penelitian ini bertujuan membandingkan kinerja tiga algoritma klasifikasi, yaitu C4.5, Naïve Bayes, dan K-Nearest Neighbors (KNN), dalam memprediksi penyakit jantung. Dataset yang digunakan adalah Heart Failure Prediction Dataset yang diperoleh dari Kaggle dan UCI Repository, dengan total 1211 data pasien. Setelah dilakukan preprocessing, data dibagi dengan rasio 70:30 untuk pelatihan dan pengujian. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, dan recall berbasis confusion matrix. Hasil penelitian menunjukkan bahwa algoritma C4.5 memiliki performa terbaik dengan akurasi 81,07%, mengungguli Naïve Bayes (79,10%) dan KNN (75,68%). C4.5 juga menunjukkan nilai recall yang lebih tinggi dalam mendeteksi kasus positif penyakit jantung, mengindikasikan efektivitasnya dalam menangani data dengan karakteristik klinis seperti dataset yang digunakan. Kontribusi utama penelitian ini adalah memberikan rekomendasi pemanfaatan algoritma C4.5 dalam pengembangan sistem pendukung keputusan untuk deteksi dini penyakit jantung. Temuan ini diharapkan dapat membantu meningkatkan akurasi diagnosis awal dan mendukung pengambilan keputusan medis secara lebih cepat dan tepat.

This is an open-access article under the [CC-BY-SA](#) license.



Penulis Korespondensi:

Titin Seh Kinanti,

Jurusan Teknik Informatika,

Fakultas Teknik Universitas Negeri Gorontalo

Jl. Prof. Dr. Ing. B. J. Habibie, Bone Bolango, Indonesia.

Email: sehkinantititin@gmail.com

PENDAHULUAN

Penyakit jantung adalah penyebab kematian nomor satu di dunia, merenggut nyawa hampir 2 juta orang Amerika setiap tahunnya (Nurkholifah et al., 2023). Berdasarkan laporan RISKESDAS tahun 2020, 20 dari 1000 penduduk Indonesia mengalami penyakit jantung, dan jumlahnya terus meningkat

setiap tahunnya, tidak terbatas pada kelompok usia tertentu (Indriyani dkk., 2024). Penyakit kardiovaskular merupakan salah satu isu kesehatan secara global, yang menurut data WHO menyebabkan sekitar 17,9 juta kematian setiap tahun atau sekitar 32% dari total angka kematian global (Bouqentar et al., 2024). Pernyataan Direktur Pencegahan dan Pengendalian Penyakit Tidak Menular Kementerian Kesehatan Republik Indonesia, Eva Susanti (2023) yang dikutip dari situs Dinas Kesehatan Provinsi Aceh, beliau menegaskan bahwa penyakit jantung menjadi penyebab kematian tertinggi di Indonesia.

Dengan semakin meningkatnya kasus penyakit jantung dari tahun ke tahun, maka pencegahan, deteksi awal, dan penanganan penyakit jantung yang tepat menjadi semakin penting dalam upaya meningkatkan kesehatan masyarakat. Salah satu solusi yang potensial adalah penerapan algoritma data mining dalam analisis data dan pemodelan prediksi. Data mining merupakan proses mengekstraksi pengetahuan dari kumpulan data berukuran besar dengan memanfaatkan algoritma serta teknik dari bidang statistik, machine learning, dan sistem manajemen database (Maulana dan Yahya, 2019). Pemodelan prediksi memungkinkan keputusan diambil dengan lebih cepat dan tepat berdasarkan data besar khususnya di bidang kesehatan.

Berdasarkan permasalahan tersebut, perlu dilakukan analisis perbandingan algoritma yaitu C4.5, Naïve Bayes, dan K- Nearest Neighbors (KNN) untuk menentukan algoritma mana yang paling efektif dalam konteks prediksi penyakit jantung sehingga diharapkan angka kematian akibat penyakit jantung dapat berkurang. Algoritma C4.5 dipilih dikarenakan cukup populer dan mudah untuk diinterpretasikan (Muthohhar & Prihanto, 2023). Naïve Bayes dipilih karena cukup sederhana dan efisien, terutama untuk dataset berukuran besar (Narulita dan Adi, 2024). Naïve Bayes juga merupakan algoritma/metode yang berlandaskan pada Teorema Bayes, dimana setiap atribut dianggap independen satu dengan yang lain, sehingga masing-masing atributnya dianggap tidak saling berhubungan dan bersifat bebas (Permana dkk., 2021). KNN dipilih karena mampu diterapkan secara efektif pada data berukuran besar, sambil tetap memberikan hasil yang akurat (Permana dkk., 2021). KNN juga adalah sebuah algoritma klasifikasi yang sederhana namun sangat efektif dalam menganalisis data dan dapat bekerja dengan cara mengelompokkan data berdasar pada kesamaan dan pola yang terdapat pada data pelatihan yang ada (Akbarollah dkk., 2023).

Penelitian ini bertujuan untuk membandingkan efektivitas algoritma C4.5, Naïve Bayes, dan KNN dalam memprediksi penyakit jantung. Dengan melakukan analisis perbandingan ini, diharapkan dapat dievaluasi dan diidentifikasi algoritma klasifikasi yang paling efektif dalam memprediksi penyakit jantung. Hasil dari penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem pendukung keputusan yang dapat digunakan untuk deteksi dini dan pencegahan penyakit jantung, serta memberikan pemahaman lebih lanjut tentang aplikasi analisis data dalam bidang kesehatan masyarakat.

METODE

Tahapan Penelitian

Penelitian ini dilakukan melalui langkah-langkah yang diawali dari studi literatur, pengumpulan dataset, preprocessing data, implementasi algoritma, dan evaluasi. Tahapan dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Alur penelitian

Studi Literatur

Pada tahap ini akan dilakukan pencarian dan pemahaman terhadap penelitian serupa yang telah ada sebelumnya untuk memperoleh teori pendukung dan informasi yang relevan dengan penelitian ini. Studi literatur yang dilakukan yaitu terkait berbagai metode klasifikasi pada machine learning dan mengenai kasus penyakit jantung.

Pengumpulan Dataset

Proses berikutnya yaitu pengumpulan dataset. Data yang digunakan yaitu Heart Failure Prediction Dataset yang diperoleh dari situs Kaggle dan UCI Repository. Dataset pada penelitian ini menggunakan format csv dengan total data yang digunakan dalam penelitian ini yaitu sebanyak 1211 data pasien penyakit jantung.

Preprocessing Data

Tahap berikutnya yaitu preprocessing data. Preprocessing adalah proses dimana data yang awalnya tidak terstruktur diubah menjadi format terstruktur agar dapat diproses lebih lanjut sesuai dengan kebutuhan (Muthohhar & Prihanto, 2023). Preprocessing data dilakukan untuk menghilangkan data yang bersifat noisy dan redundansi. Pada preprocessing data juga akan dilakukan transformasi data pada atribut HeartDisease yang awalnya dalam bentuk numerik (1 & 0) diubah ke dalam bentuk kata "yes" dan "no". Hal ini dilakukan untuk mempermudah pemahaman hasil penelitian oleh pihak-pihak yang terlibat atau yang akan memanfaatkan informasi tersebut. Software yang digunakan untuk preprocessing data yaitu RapidMiner Studio. Dalam situs web www.rapidminer.com, perangkat lunak Rapidminer dimanfaatkan untuk membangun alur kerja visual guna menganalisis *data science* dan pembelajaran mesin di dalam kelompok, mulai dari level analisis hingga ahli. Rapidminer mudah digunakan dan mampu mengakumulasi data dari berbagai sumber, termasuk database, penyimpanan *cloud*, dokumen, platform medsos, dan juga aplikasi bisnis, juga memungkinkan untuk eksplorasi dan visualisasi data secara statistik (Widaningsih, 2019).

Implementasi Algoritma

Dalam tahap implementasi algoritma, dataset akan dibagi menjadi data training dan data testing dengan rasio sebesar 70:30. Dikutip dari Nalug, pembagian data training dan data testing dengan rasio sebesar 70:30 dapat membantu mencegah overfitting dan underfitting serta memastikan bahwa model dapat digeneralisasi dengan baik ke data yang baru. Penelitian ini menggunakan 3 model yakni C4.5, Naïve Bayes, dan KNN. Tiga model tersebut akan dibandingkan dengan menggunakan aplikasi RapidMiner Studio. Parameter yang digunakan pada algoritma KNN yaitu menggunakan nilai k-3, k-5, k-7, dan k-9. Secara umum, nilai k yang tinggi akan mengurangi efek noisy pada klasifikasi. Nilai k yang baik dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan cross validation (Lestari, 2014).

Evaluasi

Tahap selanjutnya adalah tahap evaluasi yang dilakukan pada Rapidminer dengan melihat hasil implementasi ketiga algoritma. Evaluasi ini bertujuan untuk menemukan model yang paling tepat dalam memprediksi penyakit jantung dengan mengukur nilai *accuracy*, *precision*, dan *recall* dari ketiga algoritma yang dapat dihitung dengan menggunakan *Confusion Matrix* dan akan divisualisasikan menggunakan *bar chart*. *Confusion Matrix* akan menghitung akurasi dalam data mining yang berfungsi untuk menganalisis seberapa baik *classifier* dalam mengidentifikasi berbagai kelas (Derisma, 2020).

- *Accuracy*

Accuracy merupakan jumlah rasio prediksi dengan jumlah total prediksi. Pengukuran nilai *accuracy* dapat dilihat pada persamaan (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (1)$$

- **Precision**
Precision merupakan rasio jumlah prediksi benar dibanding jumlah total prediksi positif. Pengukuran nilai *precision* dapat dilihat pada persamaan (2).

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots (2)$$

- **Recall**
Recall merupakan rasio jumlah prediksi positif dibanding dengan total prediksi benar. Pengukuran nilai *Recall* dapat dilihat pada persamaan (3).

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots (3)$$

Keterangan:

- TP : (*true positive*), jumlah *record* data bernilai positif dan prediksi positif
- TN : (*true negative*), jumlah *record* data bernilai negatif dan prediksi negatif
- FP : (*false positive*), jumlah *record* data bernilai negatif dan prediksi positif
- FN : (*false negative*), jumlah *record* data bernilai positif dan prediksi negatif

HASIL DAN DISKUSI

Pengumpulan Dataset

Pada proses pengumpulan dataset, data yang digunakan yaitu Heart Failure Prediction Dataset yang diperoleh dari situs Kaggle dan UCI Repository. Dikutip dari InfoWorld, kaggle merupakan sebuah komunitas daring yang didirikan oleh Anthony Goldbloom (CEO) dan Ben Hamner (CTO) pada tahun 2010. Dataset pada penelitian ini menggunakan format csv. Data dari Kaggle berjumlah 918 data dengan masing-masing data pasien memiliki 12 atribut klinis. Data dari UCI Repository berjumlah 293 data dengan masing-masing data pasien memiliki 14 atribut klinis. Sehingga total data yang digunakan dalam penelitian ini yaitu 1211 data pasien. Dataset yang digunakan dalam penelitian ini dapat dilihat pada Gambar 2.

Row No.	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.050	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
7	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
8	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.050	Flat	1
10	48	F	ATA	120	284	0	Normal	120	N	0	Up	0
11	37	F	NAP	130	211	0	Normal	142	N	0	Up	0
12	58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
13	39	M	ATA	120	204	0	Normal	145	N	0	Up	0

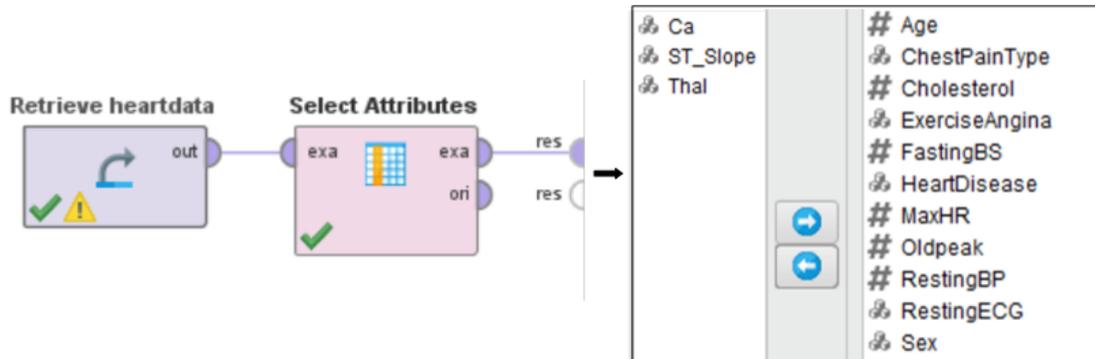
ExampleSet (918 examples, 0 special attributes, 12 regular attributes)

Gambar 2. Dataset prediksi penyakit jantung

Preprocessing Data

Setelah proses pengumpulan dataset, langkah selanjutnya yaitu preprocessing data yang dilakukan untuk mengolah data mentah. Tahap preprocessing melibatkan pengidentifikasian data, penanganan missing values, serta pengurangan fitur dengan menghapus atribut yang tidak relevan. Pada tahap preprocessing juga akan dilakukan transformasi data pada atribut HeartDisease yang awalnya dalam

bentuk numerik (1 & 0) diubah ke dalam bentuk kata “yes” dan “no”. Hal pertama yang dilakukan dalam preprocessing adalah penghapusan fitur atau atribut yang tidak relevan atau tidak dibutuhkan, yaitu atribut STSlope, Ca, dan Thal. Proses penghapusan fitur yang tidak relevan dapat dilihat pada Gambar 3.



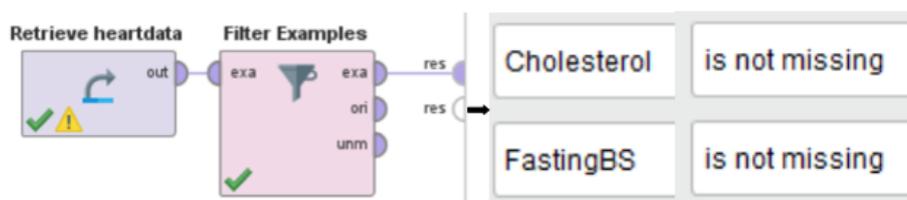
Gambar 3. Proses *select attributes*

Hal selanjutnya yang dilakukan dalam preprocessing adalah pengidentifikasian data (Gambar 4).

Name	Type	Missing	Statistics		(11 / 11 attributes): <input type="text" value="Search for Attributes"/>
Age	Integer	0	Min: 28	Max: 77	Average: 52.135
Sex	Polynomial	0	Least: F (273)	Most: M (938)	Values: M (938), F (273)
ChestPainType	Polynomial	0	Least: TA (57)	Most: ASY (619)	Values: ASY (619), ATA (278), ...[2 more]
RestingBP	Integer	0	Min: 0	Max: 200	Average: 132.442
Cholesterol	Integer	23	Min: 0	Max: 603	Average: 210.581
FastingBS	Integer	8	Min: 0	Max: 1	Average: 0.195

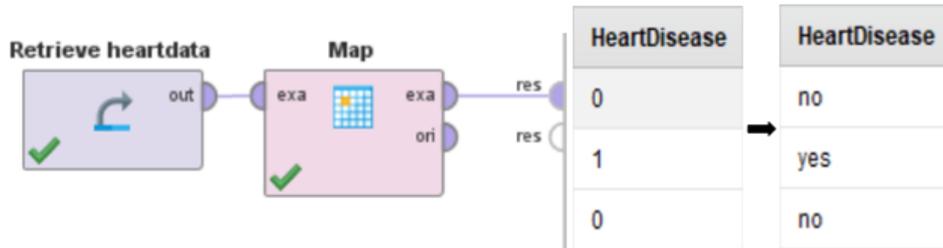
Gambar 4. Hasil identifikasi data

Berdasarkan hasil identifikasi pada Gambar 4, diketahui bahwa dataset yang digunakan dalam penelitian ini memiliki beberapa nilai yang missing, yaitu pada atribut Cholesterol dan FastingBS. Dua atribut tersebut akan dilakukan penghapusan missing values yang prosesnya dapat dilihat pada Gambar 5.



Gambar 5. Proses penanganan missing values

Setelah dilakukan penghapusan fitur yang tidak relevan atau tidak dibutuhkan dan penanganan missing value, maka 1180 data dari 1211 data yang ada dapat dikatakan bersih dan siap untuk diolah. Selanjutnya akan dilakukan transformasi data pada Atribut HeartDisease yang awalnya dalam bentuk numerik (1 & 0) diubah ke dalam bentuk kata “yes” dan “no”. Transformasi data akan menggunakan operator Map, yang prosesnya dapat dilihat pada Gambar 6.

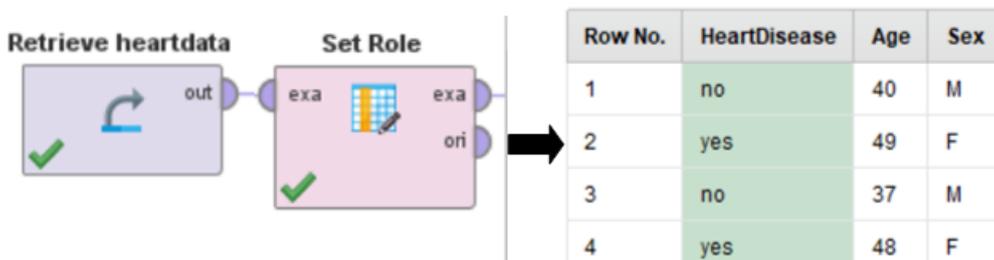


Gambar 6. Proses transformasi data

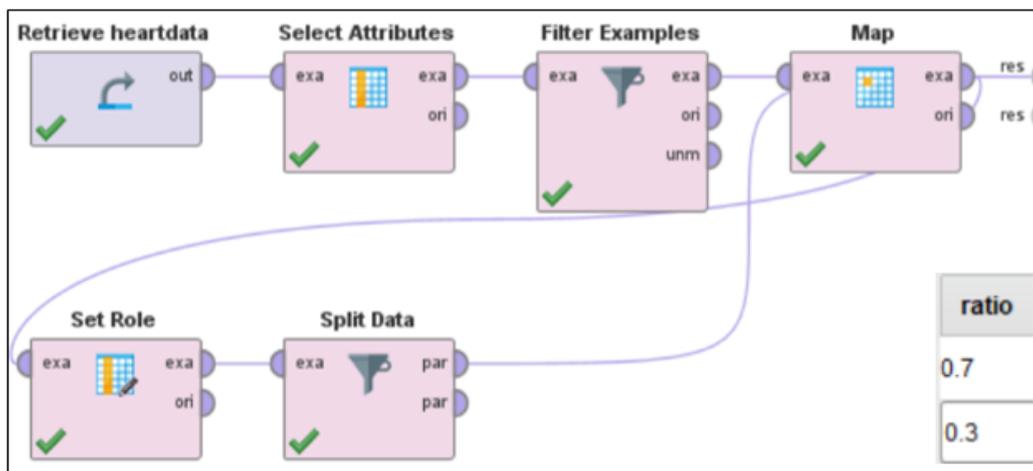
Berdasarkan Gambar 6, dapat dilihat bahwa nilai yang awalnya 1 & 0 berhasil diubah menjadi “yes” dan “no” dengan menggunakan operator *Map*.

Implementasi Algoritma

Setelah data dikatakan “bersih” dan siap untuk diolah, tahap berikutnya yaitu implementasi algoritma. Pada tahap ini, hal pertama yang dilakukan yaitu menentukan label kelas dengan menggunakan operator *Set Role* yang dapat dilihat pada Gambar 7.



Gambar 7. Penentuan label kelas



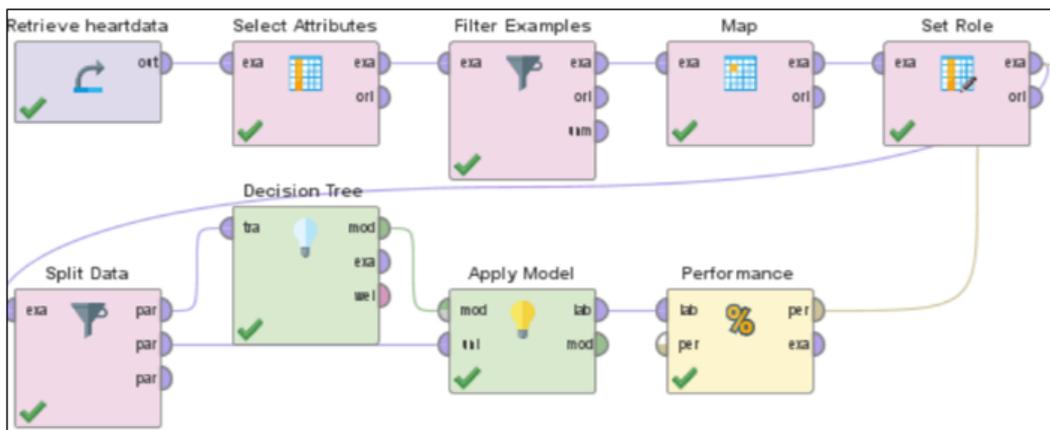
Gambar 8. Pembagian dataset

Berdasarkan Gambar 7 dapat dilihat bahwa kolom atribut HeartDisease sudah berwarna hijau menunjukkan bahwa kolom atribut ini merupakan label kelas atau target dalam proses implementasi model. Dengan menetapkan atribut tertentu sebagai label kelas, algoritma akan tahu bahwa ini merupakan variabel yang harus diprediksi.

Langkah selanjutnya yaitu pembagian dataset menjadi data training dan data testing dengan rasio sebesar 70:30, sehingga dari 1180 data, 826 data digunakan sebagai data training dan 354 data digunakan sebagai data testing. Pembagian dataset ini menggunakan operator Split Data yang dapat dilihat pada Gambar 8.

Implementasi Algoritma C4.5

Gambar 9 menunjukkan alur implementasi algoritma C4.5 pada dataset yang telah dipersiapkan.



Gambar 9. Modeling Algoritma C4.5

Maka didapati hasil dari C4.5 yaitu nilai akurasi sebesar 81,07%, yang dapat dilihat pada Gambar 10.

accuracy: 81.07%

	true no	true yes	class precision
pred. no	135	30	81.82%
pred. yes	37	152	80.42%
class recall	78.49%	83.52%	

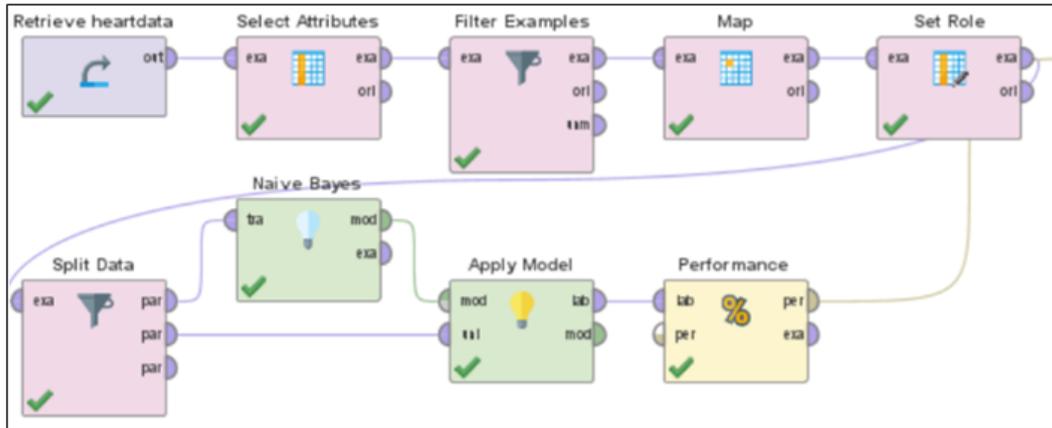
Gambar 10. Tabel hasil Algoritma C4.5

Berdasarkan confusion matrix di atas, model C4.5 mampu mengenali sebagian besar kasus penyakit jantung dengan benar. Hal ini tercermin dari recall yes yang tinggi yaitu 83,52%, yang menunjukkan kemampuan model dalam menangkap kasus positif secara efektif.

Algoritma ini unggul karena mampu menghasilkan model yang jelas dan mudah dipahami, serta menggunakan metode pemilihan atribut terbaik berdasarkan nilai informasi. Hal ini membuat C4.5 efektif dalam mengolah data yang berisi kombinasi angka dan kategori.

Implementasi Algoritma Naïve Bayes

Gambar 11 berikut memperlihatkan alur pemodelan menggunakan algoritma Naïve Bayes yang dimulai dari prapemrosesan hingga evaluasi performa model.



Gambar 11. Algoritma pemodelan Naive Bayes

Maka diperoleh hasil dari Naive Bayes yaitu nilai akurasi sebesar 79,10%, yang dapat dilihat pada Gambar 12.

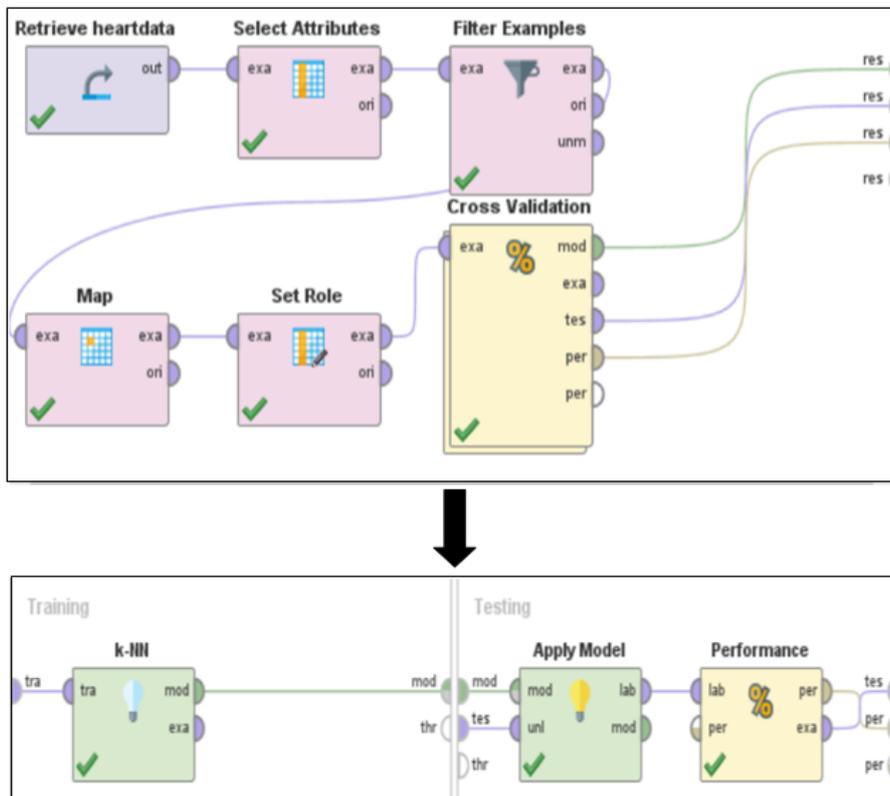
accuracy: 79.10%

	true no	true yes	class precision
pred. no	130	32	80.25%
pred. yes	42	150	78.12%
class recall	75.58%	82.42%	

Gambar 12. Hasil perhitungan Algoritma Naive Bayes

Implementasi Algoritma KNN

Gambar 13 menunjukkan alur implementasi algoritma KNN menggunakan validasi silang untuk evaluasi model.



Gambar 13. Modeling Algoritma KNN

Maka didapati hasil dari KNN (K-3) yaitu nilai akurasi sebesar 75,68%, yang dapat dilihat pada Gambar 14.

accuracy: 75.68% +/- 4.59% (micro average: 75.68%)

	true no	true yes	class precision
pred. no	456	170	72.84%
pred. yes	117	437	78.88%
class recall	79.58%	71.99%	

Gambar 14. Hasil Algoritma KNN (K-3)

Kemudian didapati hasil dari KNN (K-5) yaitu nilai akurasi sebesar 72,71%, yang dapat dilihat pada Gambar 15.

accuracy: 72.71% +/- 3.89% (micro average: 72.71%)

	true no	true yes	class precision
pred. no	413	162	71.83%
pred. yes	160	445	73.55%
class recall	72.08%	73.31%	

Gambar 15. Hasil Algoritma KNN (K-5)

Kemudian didapati hasil dari KNN (K-7) yaitu nilai akurasi sebesar 72,46%, yang dapat dilihat pada Gambar 16.

accuracy: 72.46% +/- 3.65% (micro average: 72.46%)

	true no	true yes	class precision
pred. no	410	162	71.68%
pred. yes	163	445	73.19%
class recall	71.55%	73.31%	

Gambar 16. Hasil Algoritma KNN (K-7)

Kemudian didapati hasil dari KNN (K-9) yaitu nilai akurasi sebesar 71,95%, yang dapat dilihat pada Gambar 17.

accuracy: 71.95% +/- 3.80% (micro average: 71.95%)

	true no	true yes	class precision
pred. no	420	178	70.23%
pred. yes	153	429	73.71%
class recall	73.30%	70.68%	

Gambar 17. Hasil Algoritma KNN (K-9)

Evaluasi

- *Confusion matrix*

Tabel 1. Hasil akurasi C4.5, Naïve Bayes, dan KNN

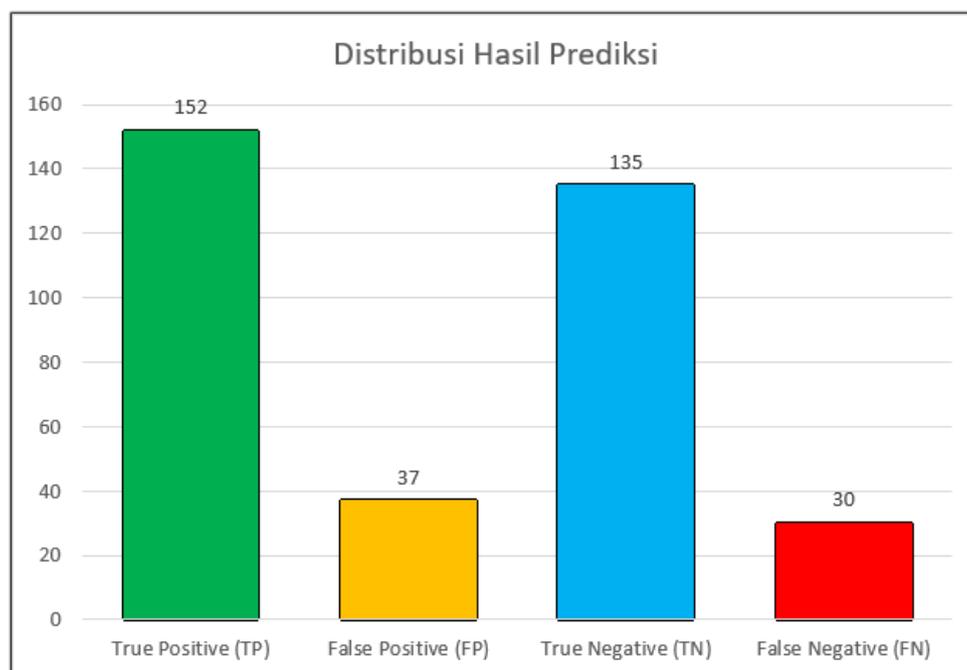
Algoritma	Accuracy	Precision		Recall	
		Yes	No	Yes	No
C4.5	81,07%	80,42%	81,82%	83,52%	78,49%
Naïve Bayes	79,10%	78,12%	80,25%	82,42%	75,58%
KNN (k-3)	75,68%	78,88%	72,84%	71,99%	79,58%

Algoritma	Accuracy	Precision		Recall	
		Yes	No	Yes	No
KNN ($k=5$)	72,71%	73,55%	71,83%	73,31%	72,08%
KNN ($k=7$)	72,46%	73,19%	71,68%	73,31%	71,55%
KNN ($k=9$)	71,95%	73,71%	70,23%	70,68%	70,30%

Setelah dilakukan pengujian menggunakan tabel confusion matrix, maka didapatkan hasil bahwa metode C4.5 memiliki nilai akurasi paling tinggi dari metode Naïve Bayes (79,10%) maupun KNN (75,68%) yaitu sebesar 81,07%.

- Hasil visualisasi

Pada tahap visualisasi, *bar chart* digunakan untuk memberikan gambaran yang jelas dan mudah dipahami mengenai hasil implementasi algoritma yang memiliki performa paling baik yaitu algoritma C4.5 yang dapat dilihat pada Gambar 18.



Gambar 18. *Bar Chart*

Hasil *bar chart* pada Gambar 18 menampilkan hasil implementasi algoritma C4.5 yang menunjukkan bahwa 152 pasien diprediksi positif (ada penyakit) maka hasilnya sesuai dengan prediksi yaitu positif (ada penyakit), 37 pasien diprediksi positif (ada penyakit) tetapi hasilnya adalah negatif (tidak ada penyakit), 135 pasien diprediksi negatif (tidak ada penyakit) maka hasilnya sesuai dengan prediksi yaitu negatif (tidak ada penyakit), dan 30 pasien diprediksi negatif (tidak ada penyakit) tetapi hasilnya adalah positif (ada penyakit).

KESIMPULAN

Penelitian ini membandingkan kinerja algoritma C4.5, Naïve Bayes, dan K-Nearest Neighbors (KNN) dalam memprediksi penyakit jantung menggunakan dataset klinis. Hasil evaluasi menunjukkan bahwa algoritma C4.5 memiliki performa terbaik, dengan akurasi mencapai 81,07%, mengungguli Naïve Bayes (79,10%) dan KNN (75,68%). Selain akurasi, C4.5 juga menunjukkan nilai recall yang tinggi dalam mengidentifikasi kasus positif, yang mencerminkan kemampuannya

dalam mendeteksi risiko penyakit jantung secara efektif. Kontribusi utama dari penelitian ini adalah memberikan rekomendasi pemanfaatan algoritma C4.5 dalam pengembangan sistem pendukung keputusan medis berbasis data mining, khususnya untuk deteksi dini penyakit jantung. Hasil ini diharapkan dapat memperkuat pengambilan keputusan medis yang lebih cepat dan tepat, sekaligus berkontribusi dalam menurunkan angka fatalitas akibat keterlambatan diagnosis. Namun, penelitian ini memiliki keterbatasan pada ukuran dataset dan ruang lingkup algoritma. Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar dan beragam, serta mengeksplorasi algoritma lain seperti Random Forest, SVM, dan metode deep learning. Pengembangan lebih lanjut juga dapat diarahkan pada aplikasi prediksi berbasis web atau mobile yang mudah diakses oleh tenaga medis maupun masyarakat.

REFERENSI

- Akbarollah, M. F., Wiyanto, W., Ardiatma, D., & Zy, A. T. (2023). Penerapan algoritma k-nearest neighbor dalam klasifikasi penyakit jantung. *Journal of Computer System and Informatics (JoSYC)*, 4(4), 850–860. <https://doi.org/10.47065/josyc.v4i4.4071>
- Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., & Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, 10, e38731. <https://doi.org/10.1016/j.heliyon.2024.e38731>
- Derisma. (2020). Perbandingan kinerja algoritma untuk prediksi penyakit jantung dengan teknik data mining. *Journal of Applied Informatics and Computing (JAIC)*, 4(1). <https://doi.org/10.30871/jaic.v4i1.2152>
- Dinas Kesehatan Provinsi Aceh. (2023). *Kemenkes: Penyakit kardiovaskular penyebab kematian tertinggi di Indonesia*. Dinas Kesehatan Provinsi Aceh. <https://dinkes.acehprov.go.id/detailpost/kemenkes-penyakit-kardiovaskular-penyebab-kematian-tertinggi-di-indonesia>
- Indriyani, T., Rozi, F. F., Hakimah, M., Rozi, N. F., & Muhima, R. R. (2024). Metode decision tree c4.5 untuk klasifikasi penyakit jantung. *Prosiding Seminar Nasional Sains dan Teknologi Terapan (SNTEK-PAN) XII* (ISSN 2685-6875). Institut Teknologi Adhi Tama Surabaya.
- InfoWorld. (2020). Kaggle: Where data scientists learn and compete. <https://www.infoworld.com/article/2258785/kaggle-where-data-scientists-learn-and-compete.html>
- Khoirudin. (2024). Klasifikasi penyakit jantung menggunakan perbandingan algoritma Decision Tree dan Naïve Bayes. *Kopertip: Jurnal Ilmiah Manajemen Informatika dan Komputer*, 8(1), 19–25. <https://doi.org/10.32485/kopertip.v8i1.351>
- Lestari, M. (2014). Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung. *Faktor Exacta*, 7(4), 366–371. <https://doi.org/10.30998/faktorexacta.v7i4.290>
- Maulana, D., & Yahya, R. (2019). Implementasi algoritma Naïve Bayes untuk klasifikasi penderita penyakit jantung di Indonesia menggunakan Rapid Miner. *Jurnal Teknik Informatika dan Sistem Informasi (JURTIF)*, 10(2), 67–75.
- Muthohhar, J. D., & Prihanto, A. (2023). Analisis perbandingan algoritma klasifikasi untuk penyakit jantung. *Journal of Informatics and Computer Science (JINACS)*, 4(3), 298-304. <https://doi.org/10.26740/jinacs.v4n03.p298-304>
- Narulita, S., & Adi, P. N. (2024). Feature selection information gain pada klasifikasi pasien penyakit jantung (heart disease). *Jurnal Rekam Medis & Manajemen Informasi Kesehatan*, 4(1), 13–19. <https://doi.org/10.53416/jurmik.v4i1.240>
- Nurkholifah, M., NofiarAm, A., & Oktorina, F. K. (2023). Analisa penyakit jantung menggunakan algoritma Naïve Bayes. *Journal of System and Computer Engineering (JSCE)*, 4(1), 26–36. <https://doi.org/10.47650/jsce.v4i1.671>
- Permana, A. P., Ainiyah, K., & Holle, K. F. H. (2021). Analisis perbandingan algoritma Decision Tree, kNN, dan Naive Bayes untuk prediksi kesuksesan start-up. *JISKA: Jurnal Informatika Sunan Kalijaga*, 6(3), 178–188. <https://doi.org/10.14421/jiska.2021.6.3.178-188>
- Scholar, A. (2023, June 11). Understanding the importance of the 70/30 rule in machine learning. *Insight Tribune*. <https://www.nalug.net/understanding-the-importance-of-the-70-30-rule-in-machine-learning/>
- Widaningsih, S. (2019). Perbandingan metode data mining untuk prediksi nilai dan waktu kelulusan mahasiswa prodi Teknik Informatika dengan algoritma C4.5, Naïve Bayes, KNN dan SVM. *Jurnal Tekno Insentif*, 13(1), 16–25. <https://doi.org/10.36787/jti.v13i1.78>