

ANALISIS FAKTOR RISIKO PEMICU SERANGAN JANTUNG DI INDONESIA, MENGGUNAKAN METODE KLASIFIKASI (DECISION TREE, NAIVE BAYES, DAN RANDOM FOREST)

Cheisyah Andini Bahri, Zahra Tri Zafira, Pratiwi Ayuningtiyas,
M. Hadi Al-Farissy, M. Alfari, Ken Ditha Tania, Allsela Meiriza

Sistem Informasi, Universitas Sriwijaya

Jalan Palembang-Prabumulih, KM 32 Indralaya, Kabupaten Ogan Ilir, Sumatera Selatan

cheisyahandiniabahri@gmail.com

ABSTRAK

Serangan jantung merupakan penyakit yang menyebabkan tingginya angka kematian di Indonesia, di mana penyakit tersebut dipengaruhi oleh beberapa faktor dan risiko, yaitu konsumsi alkohol, kebiasaan merokok, tingkat depresi, dan juga hipertensi. Dilakukannya penelitian ini dengan tujuan untuk melakukan kegiatan analisis pengaruh dari faktor-faktor tersebut dengan penyakit serangan jantung, penelitian ini menggunakan metode klasifikasi (*Decision Tree*, *Naïve Bayes*, dan *Random Forest*, berdasarkan data yang terdapat pada aplikasi Kaggle dengan judul *Heart Attack Indonesia*. Analisis pada penelitian ini dilakukan dengan bantuan *tools* (*RapidMiner*) untuk membandingkan performa dari ketiga metode klasifikasi dengan perbandingan rasio 70/30, 80/20, dan 90/10. Sehingga mendapatkan hasil analisis tertinggi yaitu metode *Decision Tree* dan *Naïve Bayes* memiliki akurasi yang sama, yaitu 74.97%, sedangkan metode *Random Forest* memiliki akurasi yang lebih rendah, yaitu 67.14%. Berdasarkan evaluasi menggunakan kurva ROCs, *Decision Tree* terbukti lebih efektif dalam mengklasifikasikan faktor risiko dibandingkan metode lainnya.

Kata kunci: serangan jantung, *Decision Tree*, *Naïve Bayes*, *Random Forest*, *RapidMiner*, Kurva ROCs

1. PENDAHULUAN

Jantung merupakan sebuah organ yang memiliki otot secara teratur bertugas untuk memompa darah melalui pembuluh darah pada manusia, di mana jantung ini berperan penting dalam peredaran darah, apabila terjadi gangguan maka akan mengakibatkan risiko kematian atau kelainan [1].

Data Organisasi Kesehatan Dunia (WHO) pada tahun 2009 menyatakan sekitar 17.1 juta orang meninggal akibat penyakit jantung. Sedangkan data yang diterbitkan oleh Kementerian Kesehatan RI pada Riset Kesehatan Dasar tahun 2007 sebesar 9.3% kematian akibat penyakit jantung di Indonesia. Penyakit jantung terkadang tidak memperlihatkan gejala tertentu dan menyebabkan kematian yang tiba-tiba, dalam beberapa kasus lain terjadi saat menjalani aktivitas fisik yang berat [2].

Faktor yang menyebabkan penyakit jantung dapat diprediksi dengan berbagai macam metode, salah satunya metode klasifikasi dalam data mining. Klasifikasi adalah model analisis data yang digunakan dalam prediksi label kelas sampel [3].

Untuk mengidentifikasi faktor yang paling mempengaruhi terjadinya penyakit jantung, metode klasifikasi merupakan pendekatan yang tepat.

Penelitian ini memiliki tujuan untuk menguji keakuratan metode klasifikasi dengan algoritma *Decision Tree*, *Naïve Bayes*, dan *Random Forest* dalam memprediksi faktor yang paling mempengaruhi penyakit jantung dengan beberapa faktor yang telah ditentukan yaitu kebiasaan merokok, tingkat depresi, dan hipertensi. Ketiga metode tersebut dipilih karena memiliki karakteristik yang berbeda dalam klasifikasi.

Decision Tree digunakan karena bisa menghasilkan model yang mudah diinterpretasikan, *Naïve Bayes* dipilih karena efisien dalam menghitung probabilitas, dan *Random Forest* digunakan untuk meningkatkan akurasi dengan pendekatan *ensemble learning* yang mengurangi risiko *overfitting*. Melalui perbandingan akan terlihat faktor yang paling berpengaruh dari tiga faktor yang telah ditentukan sebelumnya. Penelitian ini menggunakan dataset yang didapatkan dari Kaggle, yang terdiri dari 23 atribut, namun penelitian ini hanya akan menggunakan 5 atribut utama berupa *HeartAttack* (label kelas), *SmokingHabits* (Atribut), *StressLevel* (Atribut), *AlcoholConsumption* (Atribut), dan *Hypertension* (Atribut). Hasil penelitian ini diharapkan dapat memberikan wawasan bagi masyarakat untuk menghindari faktor yang menyebabkan terjadinya serangan jantung. Sebelumnya telah dilakukan penelitian serupa yang membandingkan tiga model yaitu *Decision Tree*, *Naïve Bayes*, dan *Random Forest* untuk memprediksi banjir di Desa Dayeuhkolot, Penelitian yang dilakukan oleh [4]. mendapatkan hasil bahwa metode terbaik untuk diterapkan adalah model *Random Forest* dengan perolehan nilai rata-rata untuk akurasi sebesar 99.05%, *precision* sebesar 97.91%, *recall* sebesar 99.18%, dan *f-1 score* sebesar 98%.

2. TINJAUAN PUSTAKA

Heart Attack merupakan salah satu bentuk penyakit organ jantung dengan angka kematian yang cukup besar di dunia, dengan berbagai faktor risiko yang dapat berpengaruh terhadap peningkatan jumlah kasus. Risiko serangan jantung dapat meningkat

karena beberapa faktor yaitu konsumsi alkohol, kebiasaan merokok, tingkat depresi, dan hipertensi. Identifikasi dan klasifikasi dini terhadap faktor-faktor risiko ini menjadi sangat penting untuk mencegah dampak lebih lanjut. Berbagai metode *machine learning* telah diterapkan dalam pengelompokan penyakit jantung untuk meningkatkan akurasi prediksi. *Decision Tree*, *Naïve Bayes*, dan *Random Forest* merupakan beberapa algoritma yang sering digunakan dalam klasifikasi data medis, khususnya dalam memprediksi penyakit jantung.

Sebuah penelitian yang dilakukan oleh [5], menunjukkan bahwa kombinasi *Knowledge Management System (KMS)* dan pendekatan *Knowledge Data Discovery (KDD)* dapat meningkatkan efisiensi manajemen dan analisis data kesehatan. Dalam konteks klasifikasi penyakit, pemanfaatan algoritma *Decision Tree (C4.5)* terbukti mampu mengidentifikasi pola penting dalam dataset dengan akurasi mencapai 91.47%. Penelitian yang dilakukan oleh [6] menggunakan metrik akurasi untuk membandingkan metode *Decision Tree* dan *Naïve Bayes*. Hasil penelitian menunjukkan bahwa metode *Decision Tree (C4.5)* dalam klasifikasinya menghasilkan akurasi 85.78% dan nilai AUC 0.784 yang termasuk dalam *Fair Classification*. *Naïve Bayes* menghasilkan akurasi 77.32% dan nilai AUC 0.784.

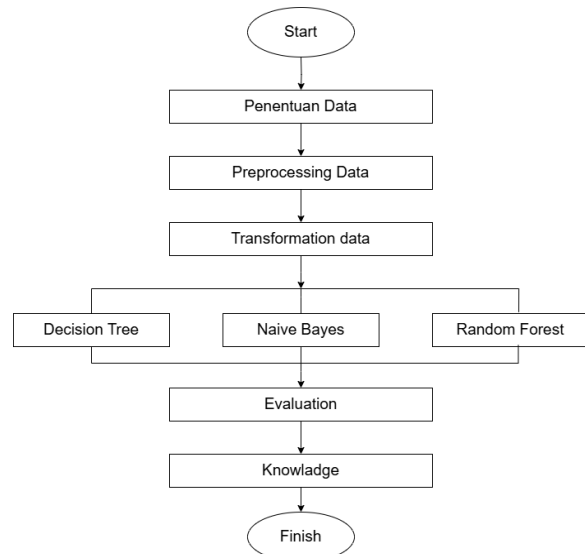
Selain itu, studi oleh [7] membandingkan performa *Random Forest*, *Naïve Bayes*, dan *Random Forest* terhadap data penyakit jantung untuk membuat klasifikasi. Penelitian ini mendapatkan hasil bahwa *Naïve Bayes* memiliki nilai rata-rata akurasi tertinggi, yaitu 84%, dibandingkan dengan algoritma lainnya. Hal ini mengindikasikan bahwa dalam beberapa kasus, *Naïve Bayes* dapat lebih unggul dalam klasifikasi penyakit jantung, terutama jika dataset memiliki variabel kategori yang dominan.

Sementara itu, penelitian terbaru oleh [8] membandingkan tiga metode klasifikasi utama dalam menganalisis dataset penyakit jantung. Hasil penelitian mereka menunjukkan bahwa *KNN* memiliki akurasi sebesar 94.92%, *Decision Tree* memiliki akurasi sebesar 84.77%, diikuti oleh *Naïve Bayes* dengan 81.64%. Berdasarkan hasil ini, *KNN* lebih efektif dalam mengidentifikasi pola dari dataset dengan jumlah fitur yang cukup besar, sementara *Random Forest* lebih cocok untuk dataset yang memiliki distribusi merata. Sebaliknya, *Naïve Bayes* lebih efisien dalam menangani dataset dengan variabel kategori, tetapi kurang optimal untuk data dengan variasi tinggi.

Menurut penelitian yang dilakukan oleh [9], di mana hasil penelitian tersebut menunjukkan bahwa bentuk model *Random Forest* memiliki performa terbaik dalam melakukan klasifikasi data, dengan akurasi sebesar 82%, model tersebut unggul dibandingkan model lainnya, temuan ini dapat mendukung dan memperkuat alasan pemilihan algoritma *Random Forest* untuk penelitian selanjutnya.

Dari berbagai penelitian yang telah dikaji, dapat disimpulkan bahwa tidak ada satu metode yang selalu unggul dalam semua kondisi. *Decision Tree* lebih efektif untuk dataset dengan banyak fitur, *Random Forest* lebih baik dalam dataset dengan distribusi yang merata, dan *Naïve Bayes* dapat bekerja dengan baik pada dataset kategori. Oleh karena itu, dalam penelitian lebih lanjut, pemilihan metode klasifikasi yang tepat harus disesuaikan dengan karakteristik dataset yang digunakan.

3. METODE PENELITIAN



Gambar 1. Alur penelitian

Metode penelitian ini diawali dengan membandingkan algoritma *Decision Tree*, *Naïve Bayes*, dan *Random Forest* untuk menganalisis faktor risiko seperti konsumsi alkohol, kebiasaan merokok, tingkat depresi dan hipertensi terhadap kemungkinan terjadinya serangan jantung. Dataset yang digunakan berasal dari situs online seperti Kaggle.com. Atribut dataset yang digunakan yaitu variabel independen (x): konsumsi alkohol, kebiasaan merokok, tingkat depresi dan hipertensi, sedangkan variabel dependen (y): indikasi penyakit jantung (1 = Yes 0 = No). Setelah data ditentukan kemudian data tersebut dilanjutkan ke tahap *preprocessing* data menggunakan RapidMiner, seperti pembersihan data (*Data cleaning*), transformasi data (*Data transformation*), penanganan data yang hilang (*Missing value handling*), normalisasi data jika diperlukan, data dibagi menjadi beberapa pembagian perbandingan rasio data latih dan data uji yaitu 70/30, 80/20, dan 90/10. Selanjutnya setiap algoritma akan digunakan untuk membangun model prediksi dan dievaluasi menggunakan akurasi.

3.1. Decision Tree

Decision Tree merupakan algoritma berbasis pohon keputusan yang digunakan dalam berbagai penelitian untuk mengklasifikasikan data berdasarkan aturan keputusan [10].

Algoritma ini memilih atribut terbaik sebagai akar pohon berdasarkan *Information Gain* atau *Gini Index* (Quinlan, 1996).

Rumus *Information Gain*:

$$IG(S, A) = Entropy(S) - \sum_{|S_v|} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

Dengan:

S = Dataset awal

A = Atribut yang diuji

S_v = Subset data berdasarkan atribut A

Rumus *Entropy* :

$$Entropy(S) = -\sum P_i \log_2 P_i \quad (2)$$

Jika digunakan *Gini Index*, maka perhitungannya :

$$Gini(S) = 1 - \sum P_i^2 \quad (3)$$

Metode *Decision Tree* telah banyak digunakan dalam penelitian klasifikasi, termasuk dalam prediksi kualitas udara dan analisis sentimen [11].

3.2. Naïve Bayes

Naïve Bayes merupakan algoritma probabilitas yang bekerja berdasarkan *Teorema Bayes*. Algoritma ini mengasumsikan bahwa setiap fitur bersifat independen, sehingga dapat digunakan untuk data kategori seperti riwayat merokok dan konsumsi alkohol [12].

Rumus *Teorema Bayes*:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (4)$$

Dengan:

y = variabel kelas (serangan jantung: ya/tidak)

X = faktor risiko serangan jantung

Jika fitur X berupa data numerik, maka probabilitas bersyarat $P(X|y)$ dapat dihitung menggunakan distribusi *Gaussian*:

$$P(X|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

Dengan:

μ = rata-rata (*mean*) dari data dalam kelas tertentu

σ^2 = varians dari data dalam kelas tertentu

Naïve Bayes sering digunakan dalam klasifikasi karena kecepatan komputasi dan kemampuannya menangani data kategori.

3.3. Random Forest

Random Forest merupakan sebuah teknik yang digunakan untuk menggabungkan beberapa bentuk *Decision tree* untuk bekerja sama dalam menentukan hasil dari prediksi suatu kelas. Dengan penentuan hasil akhir berdasarkan mayoritas suara dari seluruh pohon yang ada dalam model prediktif [13]

Algoritma ini juga dapat mengurangi *varians* yang terjadi pada satu pohon keputusan, sehingga dalam penanganan data besar lebih stabil [14].

Rumus *accuracy* dalam *Random Forest*:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Dengan:

TP (*True positive*) = Jumlah prediksi positif yang benar

TN (*True negative*) = Jumlah prediksi negatif yang benar

FP (*False positive*) = Jumlah prediksi positif yang salah

FN (*False negative*) = Jumlah prediksi negatif yang salah

4. HASIL DAN PEMBAHASAN

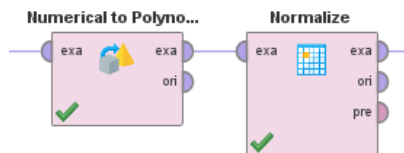
4.1. Penentuan Data

Tahapan pertama yang akan dilakukan adalah menentukan data yang dibutuhkan untuk proses klasifikasi. Penelitian ini menggunakan dataset *HeartAttack* yang didapatkan dari situs *web open source* berupa kaggle.

ID	State	Age	Gender	Cholesterol	SmokingHabit	PhysicalActivity	BMI	Hypertension	Diabetes	AlcoholConsumption	DetType	OccupationType	StressLevel	EducationLevel	MaritalStatus	FamilyHistoryHeartDisease	IncomeLevel	HealthcareAccess	SleepHours	UrbanOrRural	AirQualityIndex	PollutionLevel	EmploymentStatus	HeartAttack
1	California	45	Male	235	Smoker	High	28.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.5	Urban	150	High	Full-time	Yes
2	California	52	Female	220	Non-smoker	High	25.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	140	High	Full-time	Yes
3	California	48	Male	240	Smoker	Low	30.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	160	High	Full-time	Yes
4	California	55	Female	210	Non-smoker	High	26.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.5	Urban	130	High	Full-time	Yes
5	California	42	Male	250	Smoker	High	29.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
6	California	50	Female	230	Non-smoker	High	27.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	140	High	Full-time	Yes
7	California	47	Male	245	Smoker	Low	31.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.5	Urban	155	High	Full-time	Yes
8	California	53	Female	215	Non-smoker	High	25.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.2	Urban	135	High	Full-time	Yes
9	California	49	Male	235	Smoker	High	28.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.2	Urban	150	High	Full-time	Yes
10	California	51	Female	225	Non-smoker	High	26.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.1	Urban	145	High	Full-time	Yes
11	California	46	Male	240	Smoker	Low	30.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
12	California	54	Female	210	Non-smoker	High	25.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	140	High	Full-time	Yes
13	California	43	Male	255	Smoker	High	29.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
14	California	56	Female	205	Non-smoker	High	24.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.5	Urban	130	High	Full-time	Yes
15	California	44	Male	245	Smoker	Low	30.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.5	Urban	150	High	Full-time	Yes
16	California	57	Female	200	Non-smoker	High	24.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	135	High	Full-time	Yes
17	California	41	Male	260	Smoker	High	31.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
18	California	58	Female	195	Non-smoker	High	23.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.5	Urban	125	High	Full-time	Yes
19	California	40	Male	265	Smoker	High	32.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
20	California	59	Female	190	Non-smoker	High	23.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	130	High	Full-time	Yes
21	California	39	Male	270	Smoker	High	32.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
22	California	60	Female	185	Non-smoker	High	22.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	125	High	Full-time	Yes
23	California	38	Male	275	Smoker	High	33.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
24	California	61	Female	180	Non-smoker	High	22.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	120	High	Full-time	Yes
25	California	37	Male	280	Smoker	High	33.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
26	California	62	Female	175	Non-smoker	High	21.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	115	High	Full-time	Yes
27	California	36	Male	285	Smoker	High	34.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
28	California	63	Female	170	Non-smoker	High	21.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	110	High	Full-time	Yes
29	California	35	Male	290	Smoker	High	34.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
30	California	64	Female	165	Non-smoker	High	20.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	105	High	Full-time	Yes
31	California	34	Male	295	Smoker	High	35.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
32	California	65	Female	160	Non-smoker	High	20.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	100	High	Full-time	Yes
33	California	33	Male	300	Smoker	High	35.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
34	California	66	Female	155	Non-smoker	High	19.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	95	High	Full-time	Yes
35	California	32	Male	305	Smoker	High	36.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
36	California	67	Female	150	Non-smoker	High	19.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	90	High	Full-time	Yes
37	California	31	Male	310	Smoker	High	36.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
38	California	68	Female	145	Non-smoker	High	18.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	85	High	Full-time	Yes
39	California	30	Male	315	Smoker	High	37.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
40	California	69	Female	140	Non-smoker	High	18.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	80	High	Full-time	Yes
41	California	29	Male	320	Smoker	High	37.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
42	California	70	Female	135	Non-smoker	High	17.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	75	High	Full-time	Yes
43	California	28	Male	325	Smoker	High	38.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
44	California	71	Female	130	Non-smoker	High	17.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	70	High	Full-time	Yes
45	California	27	Male	330	Smoker	High	38.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
46	California	72	Female	125	Non-smoker	High	16.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	65	High	Full-time	Yes
47	California	26	Male	335	Smoker	High	39.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
48	California	73	Female	120	Non-smoker	High	16.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	60	High	Full-time	Yes
49	California	25	Male	340	Smoker	High	39.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
50	California	74	Female	115	Non-smoker	High	15.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	55	High	Full-time	Yes
51	California	24	Male	345	Smoker	High	40.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
52	California	75	Female	110	Non-smoker	High	15.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	50	High	Full-time	Yes
53	California	23	Male	350	Smoker	High	40.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
54	California	76	Female	105	Non-smoker	High	14.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	45	High	Full-time	Yes
55	California	22	Male	355	Smoker	High	41.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
56	California	77	Female	100	Non-smoker	High	14.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	40	High	Full-time	Yes
57	California	21	Male	360	Smoker	High	41.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
58	California	78	Female	95	Non-smoker	High	13.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	35	High	Full-time	Yes
59	California	20	Male	365	Smoker	High	42.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
60	California	79	Female	90	Non-smoker	High	13.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	30	High	Full-time	Yes
61	California	19	Male	370	Smoker	High	42.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
62	California	80	Female	85	Non-smoker	High	12.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	25	High	Full-time	Yes
63	California	18	Male	375	Smoker	High	43.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
64	California	81	Female	80	Non-smoker	High	12.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	20	High	Full-time	Yes
65	California	17	Male	380	Smoker	High	43.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
66	California	82	Female	75	Non-smoker	High	11.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	15	High	Full-time	Yes
67	California	16	Male	385	Smoker	High	44.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
68	California	83	Female	70	Non-smoker	High	11.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	10	High	Full-time	Yes
69	California	15	Male	390	Smoker	High	44.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
70	California	84	Female	65	Non-smoker	High	10.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	5	High	Full-time	Yes
71	California	14	Male	395	Smoker	High	45.0	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
72	California	85	Female	60	Non-smoker	High	10.0	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	0	High	Full-time	Yes
73	California	13	Male	400	Smoker	High	45.5	No	No	Heavy	Professional	High	High	High	Married	No	High	Good	7.0	Urban	150	High	Full-time	Yes
74	California	86	Female	55	Non-smoker	High	9.5	No	No	Light	Professional	High	High	High	Married	No	High	Good	8.0	Urban	0	High	Full	

Gambar tersebut menggambarkan proses pemilihan label data dependen dari atribut tertentu dengan menetapkan label menggunakan operator *set role*, kemudian menggunakan operator *multiply* untuk menduplikasi dataset agar dapat menggunakan model yang berbeda pada satu proses, kemudian menggunakan operator *split data* yang digunakan untuk membagi dataset menjadi beberapa bagian (*training* dan *testing*).

4.3. Transformation Data

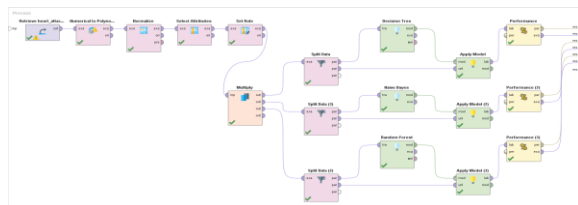


Gambar 5. Numerical to polynomial dan normalization

Gambar tersebut menunjukkan tahap *transformation* dua operator pada Rapid Miner yaitu *Numerical to Polynomial* digunakan untuk mengubah atribut bertipe *numeric* menjadi tipe *polynomial*, operator selanjutnya yaitu *Normalization* digunakan untuk menormalisasi data yang digunakan.

4.4. Implementasi Model

Setelah data dibersihkan dan ditransformasi, tahap selanjutnya yaitu implementasi model. Terdapat tiga model klasifikasi yang akan dibandingkan yaitu *Decision Tree*, *Naïve Bayes*, dan *Random Forest*.



Gambar 6. Multiply model

Pada tahap implementasi model akan dibuat dengan tiga operator klasifikasi yaitu *Decision tree*,

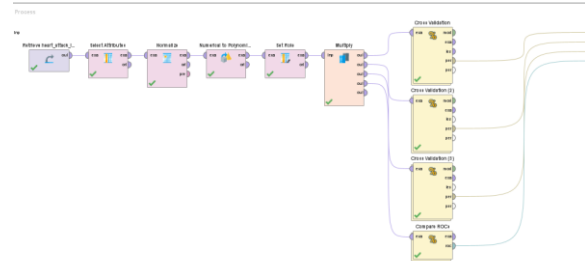
4.5. Evaluation

Tabel 1. Hasil Pengujian Menggunakan Rasio Data (Latih/Uji) 70/30

Model	Akurasi (%)	Recall No-Yes (%)	Recall Yes-Yes (%)	Precision No (%)	Precision Yes (%)
Decision Tree	74.97%	100.00%	0.00%	74.97%	0.00%
Naïve Bayes	74.97%	100.00%	0.00%	74.97%	0.00%
Random Forest	67.08%	84.04%	16.28%	75.04%	25.40%

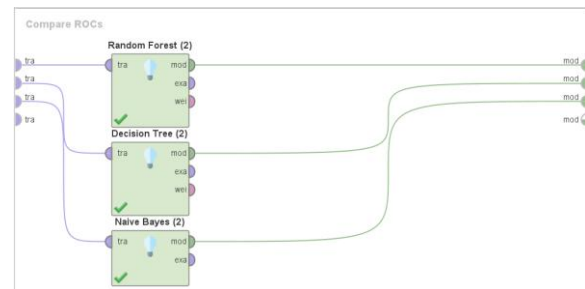
Hasil pengujian pertama menggunakan rasio 70/30 dengan model *Decision Tree* dan *Naïve Bayes* dapat terlihat bahwa akurasi dari model ini sebesar 74.97% dengan *recall* untuk kelas *no-yes* sebesar 100.00% dan untuk kelas *yes-yes* adalah 0.00%. Serta memiliki *precision* sebesar 74.97% untuk kelas *no* dan kelas *yes* sebesar 0.00%. Sedangkan pengujian

Naïve Bayes, dan *Random Forest*, kemudian digabungkan dengan menggunakan operator *multiply*. Sebelum model diterapkan akan dilakukan *split data* terlebih dahulu untuk membagi antara data pelatihan dan data pengujian beberapa rasio perbandingan yaitu 70/30, 80/20, dan 90/10. Operator *performance* digunakan setelah model diterapkan, *performance* ini digunakan untuk melihat nilai akurasi dari masing-masing model.



Gambar 7. Cross validation model

Selanjutnya dilakukan pula evaluasi model dengan menggunakan operator *cross validation* untuk mengetahui seberapa besar akurasi dari model yang telah dirancang, model ini juga ditambahkan operator *Compare ROCs* untuk menampilkan kurva dari performa ketiga model tersebut.



Gambar 8. Compare ROCs

Pada *Compare ROCs* menggunakan model *Random Forest*, *Decision Tree*, dan *Naïve Bayes* akan dimasukkan untuk dilakukan pengujian dengan melihat perbandingan akurasi dari ketiga model tersebut.

menggunakan model *Random Forest* dapat terlihat bahwa akurasi dari model ini sebesar 67.08% dengan *recall* untuk kelas *no-yes* sebesar 84.04% dan untuk kelas *yes-yes* adalah 16.28%. Serta memiliki *precision* sebesar 75.04% untuk kelas *no* dan kelas *yes* sebesar 25.40%.

Tabel 2. Hasil Pengujian Menggunakan Rasio Data (Latih/Uji) 80/20

Model	Akurasi (%)	Recall No-Yes (%)	Recall Yes-Yes (%)	Precision No (%)	Precision Yes (%)
Decision Tree	74.97%	100.00%	0.00%	74.97%	0.00%
Naïve Bayes	74.97%	100.00%	0.00%	74.97%	0.00%
Random Forest	67.14%	84.18%	16.11%	75.03%	25.37%

Hasil pengujian pertama menggunakan rasio 80/20 dengan model *Decision Tree* dan *Naïve Bayes* dapat terlihat bahwa akurasi dari model ini sebesar 74.97% dengan *recall* untuk kelas *no-yes* sebesar 100.00% dan untuk kelas *yes-yes* adalah 0.00%. Serta memiliki *precision* sebesar 74.97% untuk kelas *no* dan kelas *yes* sebesar 0.00%. Sedangkan pengujian

menggunakan model *Random Forest* dapat terlihat bahwa akurasi dari model ini sebesar 67.14% dengan *recall* untuk kelas *no-yes* sebesar 84.18% dan untuk kelas *yes-yes* adalah 16.11%. Serta memiliki *precision* sebesar 75.03% untuk kelas *no* dan kelas *yes* sebesar 25.37%.

Tabel 3. Hasil Pengujian Menggunakan Rasio Data (Latih/Uji) 90/10

Model	Akurasi (%)	Recall No-Yes (%)	Recall Yes-Yes (%)	Precision No (%)	Precision Yes (%)
Decision Tree	74.97%	100.00%	0.00%	74.97%	0.00%
Naïve Bayes	74.97%	100.00%	0.00%	74.97%	0.00%
Random Forest	66.75%	83.65%	16.12%	74.92%	24.77%

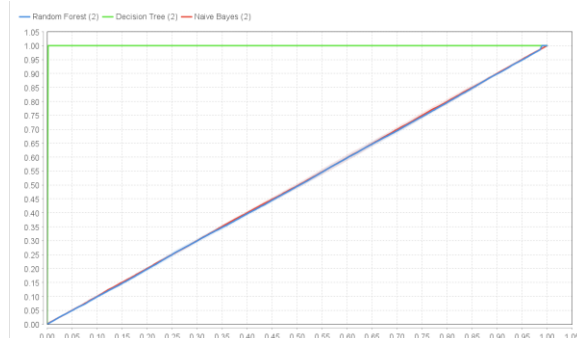
Hasil pengujian pertama menggunakan rasio 90/10 dengan model *Decision Tree* dan *Naïve Bayes* dapat terlihat bahwa akurasi dari model ini sebesar 74.97% dengan *recall* untuk kelas *no-yes* sebesar 100.00% dan untuk kelas *yes-yes* adalah 0.00%. Serta memiliki *precision* sebesar 74.97% untuk kelas *no* dan kelas *yes* sebesar 0.00%. Sedangkan pengujian

menggunakan model *Random Forest* dapat terlihat bahwa akurasi dari model ini sebesar 66.75% dengan *recall* untuk kelas *no-yes* sebesar 83.65% dan untuk kelas *yes-yes* adalah 16.12%. Serta memiliki *precision* sebesar 74.92% untuk kelas *no* dan kelas *yes* sebesar 24.77%.

Tabel 4. Hasil Pengujian Cross Validation

Model	Akurasi (%)	Recall No-Yes (%)	Recall Yes-Yes (%)	Precision No (%)	Precision Yes (%)
Decision Tree	74.97%	100.00%	0.00%	74.97%	0.00%
Naïve Bayes	74.97%	100.00%	0.00%	74.97%	0.00%
Random Forest	74.25%	98.57%	1.42%	74.97%	24.89%

Hasil pengujian model *Decision Tree* didapatkan nilai akurasi sebesar 74.97% dengan *recall* untuk kelas *no-yes* sebesar 100% dan 0% untuk kelas *yes-yes*. *precision* yang didapatkan sebesar 74.97% untuk kelas *no* dan 0% untuk kelas *yes*. Sedangkan Pada model *Naïve Bayes* didapatkan nilai akurasi sebesar 74.97% dengan *recall* untuk kelas *no-yes* sebesar 100% dan kelas *yes-yes* sebesar 0%. Pada *precision* didapatkan hasil sebesar 74.97% untuk kelas *no* dan 0% untuk kelas *yes*. Pada model *Random Forest* dapat terlihat bahwa akurasi dari model ini sebesar 74.25% dengan *recall* untuk kelas *no-yes* sebesar 98.57% dan untuk kelas *yes-yes* adalah 1.42%. Serta memiliki *precision* sebesar 74.97% untuk kelas *no* dan kelas *yes* sebesar 24.98%.



Gambar 15. Kurva ROCs

Kurva ROCs (*Receiver Operating Characteristic*) merupakan salah satu cara dalam mengevaluasi akurasi dari klasifikasi secara visual. Evaluasi ROCs akan dilakukan pada tiga model klasifikasi, yaitu *Random Forest*, *Decision Tree*, dan *Naïve Bayes*. Berdasarkan hasil perhitungan ROCs dapat terlihat bahwa algoritma yang mendekati angka 1.00 dengan nilai FPR yang rendah adalah model *Decision Tree*. Dibandingkan pada model *Random Forest* dan *Naïve Bayes*. Berdasarkan hasil tersebut dapat diketahui bahwa model *Decision Tree* merupakan model yang paling efektif dalam klasifikasi dataset serangan jantung.

5. KESIMPULAN DAN SARAN

Dengan dilakukannya pengujian dengan menggunakan beberapa rasio perbandingan yaitu 70/30, 80/20, dan 90/10, mendapatkan hasil tertinggi dari ketiga metode yaitu *Decision Tree* dan *Naïve Bayes* memiliki akurasi sebesar 74.97% dalam memprediksi risiko serangan jantung di Indonesia, sedangkan *Random Forest* menunjukkan hasil akurasi yang lebih rendah 67.14%. *Cross Validation* menunjukkan bahwa *Decision Tree* paling efektif dalam mendeteksi faktor risiko dengan nilai ROCs mendekati 1,00 dan nilai FPR yang rendah. *Random Forest* lebih seimbang dalam *recall* dan *precision*,

tetapi *Decision Tree* lebih akurat dalam klasifikasi. Analisis berbasis data diharapkan dapat memberikan kontribusi yang signifikan terhadap pencegahan serangan jantung. Penelitian selanjutnya akan menerapkan *deep learning* dan *ensemble learning* dengan dataset yang lebih luas dan algoritma yang lebih canggih untuk meningkatkan akurasi prediksi. Penelitian ini membuka peluang untuk pengembangan sistem rekomendasi kesehatan berbasis data.

DAFTAR PUSTAKA

- [1] D. Haganta Depari, Y. Widiastiwi, and M. Mega Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *JURNAL INFORMATIK Edisi ke-18*, vol. 18, no. 3, Dec. 2022, Accessed: Mar. 27, 2025. [Online]. Available: <https://ejournal.upnvj.ac.id/informatik/article/view/4694/1852>
- [2] Ardiansyah Sembiring M, "ANALISIS FAKTOR PREDIKSI DIAGNOSA TINGKAT SERANGAN JANTUNG MENGGUNAKAN METODE REGRESSION," *JURNAL TEKNIKI (Jurnal Teknologi Komputer dan Sistem Informasi)*, vol. 4, no. 1, pp. 16–22, Feb. 2024, [Online]. Available: <http://jurnal.goretanpena.com/index.php/teknisi>
- [3] A. Yogiarto, A. Homaidi, and Z. Fatah, "Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 3, pp. 1720–1728, Jul. 2024, doi: 10.33379/gtech.v8i3.4495.
- [4] M. B. A. Darmawan, F. Dewanta, and S. Astuti, "Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naive Bayes untuk Prediksi Banjir di Desa Dayeuhkolot Comparative Analysis of Decision Tree, Random Forest, and Naive Bayes Algorithm for Flood Prediction at Dayeuhkolot Village," *TELKA*, vol. 9, no. 1, pp. 52–61, 2023, Accessed: Apr. 03, 2025. [Online]. Available: <https://telka.ee.uinsgd.ac.id/index.php/TELKA/article/view/telka.v9n1.52-61>
- [5] N. Eka Saputra, K. Ditha Tania, and R. Izwan Heroza, "PENERAPAN KNOWLEDGE MANAGEMENT SYSTEM (KMS) MENGGUNAKAN TEKNIK KNOWLEDGE DATA DISCOVERY (KDD) PADA PT PLN (PERSERO) WS2JB RAYON KAYU AGUNG," *Jurnal Sistem Informasi (JSI)*, vol. 8, no. 2, pp. 2355–4614, 2016, [Online]. Available: <http://ejournal.unsri.ac.id/index.php/jsi/index>
- [6] I. Hamzah, Z. Sitorus, and Khairul, "Analisa Classification Decision Tree C45 dan Naive Bayes Pada Indikasi Penyakit Diabetes Menggunakan Rapid Miner," *Jurnal Nasional Teknologi Komputer*, vol. 4, no. 1, 2024, Accessed: Apr. 05, 2025. [Online]. Available: <https://publikasi.hawari.id/index.php/jnastek/article/view/126>
- [7] A. Samosir, M. Hasibuan, W. E. Justino, and T. Hariyono, "Komparasi Algoritma Random Forest, Naive Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung," in *Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat*, Institut Informatika dan Bisnis Darmajaya, Aug. 2021. Accessed: Mar. 27, 2025. [Online]. Available: <https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2955>
- [8] E. Astama Putra and R. Supriyanto, "Analisis Algoritma Decesion Tree, KNN dan Naive Bayes Pada Dataset Penyakit Jantung," *Jurnal Ilmiah KOMPUTASI*, vol. 22, no. 2, Jun. 2023, Accessed: Mar. 27, 2025. [Online]. Available: <https://ejournal.jak-stik.ac.id/index.php/komputasi/article/view/3353>
- [9] A. I. Fadilla, K. T. Ditha, A. Wedhasmara, and A. Meiriza, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Knowledge Extraction Using Aspect-Based Sentiment Analysis and Classification Method," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 10, no. 3, pp. 378–388, Dec. 2024, Accessed: Apr. 10, 2025. [Online]. Available: <https://jurnal.untan.ac.id/index.php/jepin/article/view/83389>
- [10] F. S. Pattiiha and Hendry, "Perbandingan Metode K-NN, Naive Bayes, Decision Tree untuk Analisis Sentimen Tweet Twitter Terkait Opini Terhadap PT PAL Indonesia," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 2, pp. 506–514, Apr. 2022, doi: 10.30865/jurikom.v9i2.4016.
- [11] N. L. W. S. R. Ginantra, C. P. Yanti, G. D. Prasetya, I. B. G. Sarasvananda, and I. K. A. G. Wiguna, "Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan K-NN," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 205–215, Dec. 2022, doi: 10.23887/janapati.v11i3.49450.
- [12] Kartianom, Arpandi, G. K. Kassymova, and O. Ndayizeye, "Prediction model of teacher candidate student graduation status: Decision Tree C4.5, Naive Bayes, and k-NN," *EKSPOSE: Jurnal Penelitian Hukum dan Pendidikan*, vol. 21, no. 2, pp. 1407–1418, 2022, Accessed: Mar. 27, 2025. [Online]. Available: <https://jurnal.iain-bone.ac.id/index.php/ekspose/article/view/3407>
- [13] T. Gori and A. Hestiningtyas, "View of Optimasi Pemilihan Fitur untuk Prediksi Penyakit Jantung Menggunakan Algoritma Genetika dan Random Forest," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, Oct. 2024, doi: doi.org/10.33022/ijcs.v13i5.4214.
- [14] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007