e- ISSN: 2828–299X

Klasifikasi Penyakit Jantung Menggunakan *Decision Tree* dan *Random Forest*

Sabrina Adnin Kamila[1], RR Sri Sulistijowati[2], Irwan Susanto[3]
[1,2,3]Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
E-mail: *[1]*sabrinakamila158@student.uns.ac.id*, [2]rr_ssh@staff.uns.ac.id,
[3]irwansusanto@staff.uns.ac.id

Abstrak – *Salah satu penyakit yang paling banyak menyebabkan kematian adalah penyakit jantung (heart disease). Penyakit jantung juga merupakan penyakit yang paling besar dibiayai oleh BPJS Kesehatan. Sebagai upaya preventif dalam penanganan penyakit jantung, perlu dilakukan prediksi penyakit jantung pada pasien. Proses klasifikasi untuk memprediksi penyakit jantung dilakukan dengan menggunakan decision tree dan random forest. Objek penelitian ini menggunakan Heart Disease Cleveland UCI Dataset dengan 297 record data. Kemudian melakukan k-fold cross validation dengan nilai k = 9 yang menghasilkan data training sebanyak 264 sampel dan data testing sebanyak 33 sampel. Hasil dari kedua klasifikasi akan dibandingkan dengan melihat performa akurasi, precision, recall, dan F1 score.*

Kata Kunci *— K-fold Cross Validation, Klasifikasi, Penyakit jantung*

1. PENDAHULUAN

Penyakit jantung merupakan suatu kondisi dimana terdapat gangguan pada organ jantung. Penyakit jantung dapat ditandai dengan beberapa gejala yang perlu diwaspadai. Gejala-gejala penyakit jantung diantaranya adalah adanya rasa mual dan muntah, berkeringat dingin dan perasaan mudah lelah, sakit kepala, nyeri dada sebelah kiri, sesak napas, lemas, jantung berdebar, dan dada terasa seperti diremas-remas.

Sebelum pandemi Covid-19, penyakit jantung menjadi penyebab utama kematian di dunia dan di Indonesia. Menurut data yang diterbitkan oleh WHO pada tahun 2021, terdapat 17.8 juta kematian akibat penyakit jantung. Jumlah kasus penyakit jantung pada tahun 2021 mencapai 12.934.931 kasus. Penyakit jantung merupakan penyakit yang paling banyak dibiayai oleh BPJS Kesehatan dengan jumlah pembiayaan hampir mencapai 7.7 triliun rupiah [1]. Melihat tingginya angka dari kasus, kematian, serta biaya untuk penyakit jantung maka perlu adanya deteksi dini atau prediksi mengenai penyakit jantung dengan klasifikasi *machine learning*.

Penelitian terkait deteksi penyakit jantung menggunakan *machine learning* sudah pernah dilakukan. Penelitian oleh Annisa (2019) tentang klasifikasi data mining untuk prediksi penyakit jantung menggunakan *decision tree, naïve bayes, k-nearest neighbors, random forest,* dan *decision stump*. *Random forest* memiliki nilai akurasi tertinggi dari hasil penelitian yang diperoleh dengan *10-fold cross validation* dan *t-test* dengan akurasi sebesar 80.38% [2].

Azhima dkk (2022) dalam penelitiannya tentang *hybrid machine learning* untuk memprediksi penyakit jantung, disimpulkan bahwa *hybrid model* dapat meningkatkan akurasi. Klasifikasi dengan *random foresr* memberikan nilai akurasi sebesar 83.16% dan klasifikasi dengan *logistic regression* memberikan skor akurasi 77.88%. Dengan *hybrid model,* skor akurasi klasifikasi sebesar 84.48% yang berarti skor akurasi *random forest* meningkat sebesar 1.32% [2].

Alham dkk (2021) melakukan penelitian terkait sistem diagnosis penyakit jantung koroner di RSUD dr. Soedarso Pontianak dengan *decision tree.* Dari hasil penelitian didapatkan nilai akurasi sebesar 94.4%. Penelitian ini membandingkan hasil klasifikasi menggunakan *decision tree* dan *random forest* dengan *k-fold cross validation* untuk menemukan hasil prediksi terbaik untuk penyakit jantung [4].

2. METODE PENELITIAN

2.1 Sumber Data

Dataset yang digunakan berjudul *Heart Disease Cleveland* UCI dan dipublikasikan di *Kaggle* [5]. *Dataset* terdiri dari 14 atribut yang ditunjukkan pada tabel 1. Kesimpulan terdapatnya penyakit jantung pada kolom *condition* dengan nilai 0 atau 1.

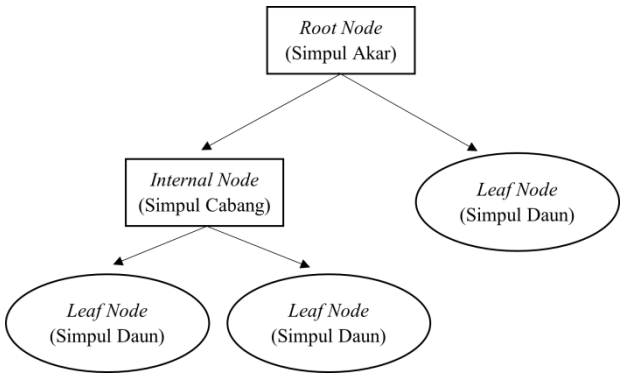
e- ISSN: 2828–299X

Tabel 1. Daftar Atribut *Dataset*

Atribut	Keterangan
Age	Umur
Sex	Jenis Kelamin
CP (Chest Pain)	Rasa sakit dada
Trestbps	Tekanan darah saat istirahat (mmHg)
Chol	Kolesterol (mg/dl)
Fbs	Gula darah puasa (>120mg/dl) (1=ya; 0=tidak)
Restecg	Hasil elektrokardiografi saat istirahat
Thalach	Detak jantung maksimal
Exang	Latihan yang diinduksi angina (1=ya;0=tidak)
Oldpeak	Depresi yang diinduksi oleh latihan relatif
Slope	Kemiringan puncak ST Segmen
Ca	Jumlah pembuluh darah yang berwarna setelah diwarnai flourosopy
Thal	Tipe kerusakan pembuluh darah (2=cacat sementara;1=cacat tetap;0=normal)
Condition	Indikasi penyakit jantung (1=ya;0=tidak)

2.2 *Decision Tree*

Decision tree adalah implementasi yang dirancang untuk menemukan dan memperoleh keputusan dengan mempertimbangkan berbagai faktor yang terkait dengan masalah. *Decision tree* membangun pohon keputusan dengan mengambil atribut sebagai akar, membuat cabang untuk setiap nilai kasus di cabang, dan mengulangi proses untuk setiap cabang hingga kasus di cabang memiliki kelas yang sama. Struktur *decision tree* digambarkan pada gambar 1.



Gambar 1. Struktur *Decision Tree*

Penentuan besarnya keefektifan suatu atribut dalam klasifikasi disebut *information gain* yang berdasarkan nilai *gain* tertinggi, menggunakan persamaan 3 [6].

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad \text{.....(3)}$$

dengan S adalah set kasus, A adalah atribut, N adalah sejumlah partisi atribut A, $|S_v|$ adalah jumlah kasus pada partisi ke-i, dan $|S|$ adalah jumlah kasus di S. Nilai *Entropy* dihitung sebelum nilai *gain* diperoleh. Rumus *Entropy* dituliskan pada persamaan 4.

$$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \quad \text{.....(4)}$$

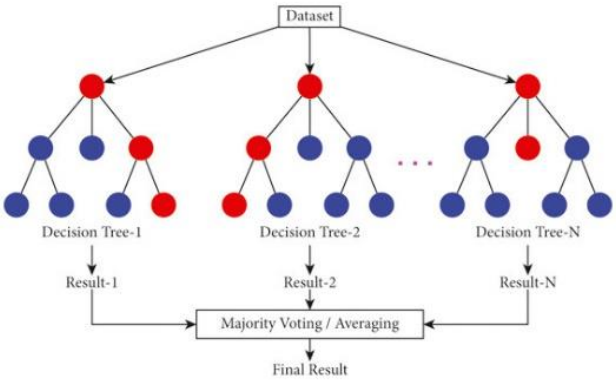
2.3 *Random Forest*

Random forest adalah pengklasifikasi yang terdiri dari kumpulan pohon klasifikasi. Misalkan $\{h(x, \Theta_k), k = 1, \dots\}$ dimana $\{\Theta_k\}$ adalah vektor random independen yang didistribusikan secara identik (*independenct identically distributed*) dan setiap pohon memilih kelas yang paling populer dari data (*majority vote*). Misalkan *ensemble* pengklasifikasi $h_1(x), h_2(x), \dots, h_K(x)$, dan dengan data *training* dipilih secara random dari distribusi random Y dan X, fungsi margin ($mg(X,Y)$) dari *random forest* didefinisikan pada persamaan 5 [7].

$$mg(X,Y) = \frac{\sum_{K=1}^{K} I(h_K(X)=Y)}{K} - \max_{j \neq Y} \left[\frac{\sum_{K=1}^{K} I(h_k(X)=j)}{K} \right] \quad \text{.....(5)}$$

dengan I adalah fungsi indikasi dan K adalah banyaknya pohon. Struktur dari *random forest* digambarkan pada gambar 2.

e- ISSN: 2828–299X



Gambar 2. Struktur *Random Forest*

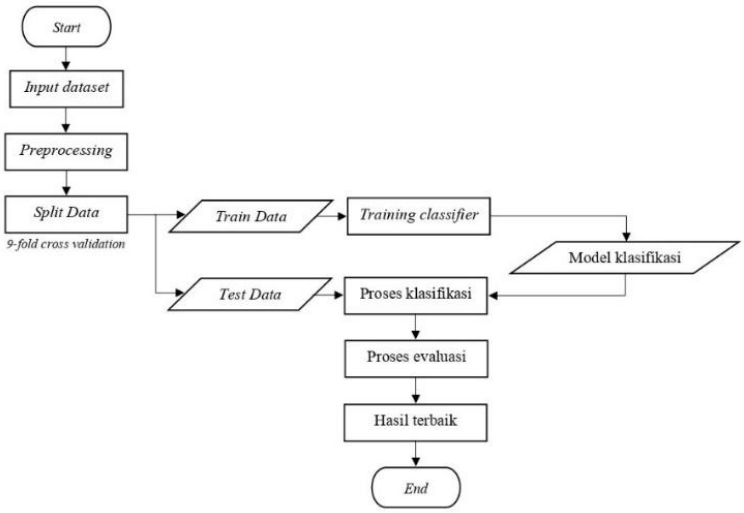
2.4 *K-Fold Cross Validation*

K-fold cross validation adalah metode validasi pengujian sistematis yang membagi data menjadi *k* bagian dan mengestimasi kesalahan satu siklus. Langkah-langkah dari *k-fold cross validation* yaitu:

1. Membagi jumlah data menjadi *k* bagian.
2. *Fold* ke-1 adalah ketika bagian ke-1 menjadi data *testing* dan sisanya menjadi data *training*, kemudian menghitung nilai evaluasi berdasarkan bagian data tersebut.
3. *Fold* ke-2 adalah ketika bagian ke-2 menjadi data *testing* dan sisanya menjadi data *training*, kemudian menghitung nilai evaluasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai *fold* ke-*k*, nilai evaluasi final adalah hasil perhitungan rata-rata dari *k* buah *fold.*

2.5 Langkah Penelitian

Penelitian ini terdiri dari beberapa tahap diantaranya membagi *dataset* menjadi *k* bagian sesuai dengan *k-fold cross validation*, melakukan *preprocessing* data, kemudian melakukan klasifikasi dengan *decision tree* dan *random forest*, menghitung nilai evaluasi untuk mendapatkan hasil klasifikasi terbaik. Alur dari penelitian ini dapat dilihat pada gambar 3.



Gambar 3. Alur Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Data

Dataset pada penelitian ini dilakukan pemeriksaan adanya *missing value* dengan menggunakan *software Rstudio* didapatkan hasil pada gambar 4.

e- ISSN: 2828–299X

```
> colSums(is.na(heart))
     age      sex       cp trestbps     chol      fbs  restecg  thalach    exang
       0        0        0        0        0        0        0        0        0
 oldpeak    slope       ca     thal condition
       0        0        0        0        0
```

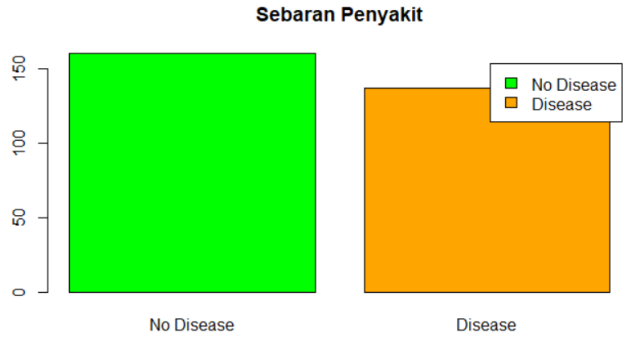
Gambar 4. Hasil Pemeriksaan *Missing Value*

Berdasarkan hasil yang diperoleh pada gambar 4, *dataset* tidak memiliki *missing value* sehingga dapat dilanjutkan ke tahap proses klasifikasi. Statistik deskriptif dari *dataset* yang digunakan ditunjukkan pada tabel 2.

Tabel 2. Statistik Deskriptif

	Min	***1st Qu***	***Median***	***Mean***	***3rd Qu***	***Max***
Age	29	48	56	54.54	61	77
Sex	0	0	1	0.6768	1	1
CP	0	2	2	2.158	3	3
Trestbps	94	120	130	131.7	140	200
Chol	126	211	243	247.4	276	564
Fbs	0	0	0	0.1448	0	1
Restecg	0	0	1	0.9966	2	2
Thalach	71	133	153	149.6	166	202
Exang	0	0	0	0.3266	1	1
Oldpeak	0	0	0.8	1.056	1.6	6.2
Slope	0	0	1	0.6027	1	2
Ca	0	0	0	0.6768	1	3
Thal	0	0	0	0.835	2	2
Condition	0	0	0	0.4613	1	1

Analisis data visualisasi dilakukan untuk mengetahui perbandingan rasio dari kelas. Gambar 5 menggambarkan *balanced data* dengan persentase pasien tidak mempunyai penyakit jantung sebesar 53.87% dan persentase pasien mempunyai penyakit jantung sebesar 46.13%.



Gambar 5. *Balanced Class Distribution*

Penelitian ini menggunakan *9-fold cross validation* dengan membagi *dataset* menjadi *9-fold* dengan ukuran yang sama, dimana *8-fold* akan digunakan sebagai data *training* dan *1-fold* akan diguanakan sebagai data *testing*. Pembagian *dataset* menjadi data *training* dan data *testing* digambarkan pada tabel 3.

Tabel 3. Pembagian *Dataset*

k-fold	*Dataset*								
1	Test	Train	Train	Train	Train	Train	Train	Train	Train
2	Train	Test	Train	Train	Train	Train	Train	Train	Train
3	Train	Train	Test	Train	Train	Train	Train	Train	Train
4	Train	Train	Train	Test	Train	Train	Train	Train	Train
5	Train	Train	Train	Train	Test	Train	Train	Train	Train
6	Train	Train	Train	Train	Train	Test	Train	Train	Train
7	Train	Train	Train	Train	Train	Train	Test	Train	Train
8	Train	Train	Train	Train	Train	Train	Train	Test	Train
9	Train	Train	Train	Train	Train	Train	Train	Train	Test

3.2 Hasil Klasifikasi

Tahap pertama yang dilakukan untuk klasifikasi adalah melakukan pemodelan dengan data *training*. Setelah model klasifikasi pada data *training* terbentuk, model tersebut diuji menggunakan data *testing*. Performa

e- ISSN: 2828–299X

klasifikasi dengan *decision tree* dapat dilihat pada tabel 4 dan performa klasifikasi dengan *random forest* dapat dilihat pada tabel 5.

Tabel 4. Performa Klasifikasi *Decision Tree*

	Accuracy	*Precision*	*Recall*	*F1 Score*
Validation 1	72.73%	82.35%	70.00%	75.67%
Validation 2	81.82%	88.24%	78.95%	83.34%
Validation 3	75.76%	83.33%	75.00%	78.95%
Validation 4	84.85%	100.00%	73.68%	84.85%
Validation 5	87.88%	94.74%	85.71%	90.00%
Validation 6	81.82%	87.50%	77.78%	82.35%
Validation 7	66.67%	73.91%	77.27%	75.55%
Validation 8	72.73%	100.00%	65.38%	77.59%
Validation 9	72.73%	94.74%	69.23%	80.00%

Tabel 5. Performa Klasifikasi *Random Forest*

	Accuracy	*Precision*	*Recall*	*F1 Score*
Validation 1	72.73%	82.35%	70.00%	75.67%
Validation 2	81.82%	88.24%	78.95%	83.34%
Validation 3	75.76%	83.33%	75.00%	78.95%
Validation 4	84.85%	100.00%	73.68%	84.85%
Validation 5	87.88%	94.74%	85.71%	90.00%
Validation 6	81.82%	87.50%	77.78%	82.35%
Validation 7	66.67%	73.91%	77.27%	75.55%
Validation 8	72.73%	100.00%	65.38%	77.59%
Validation 9	72.73%	94.74%	69.23%	80.00%

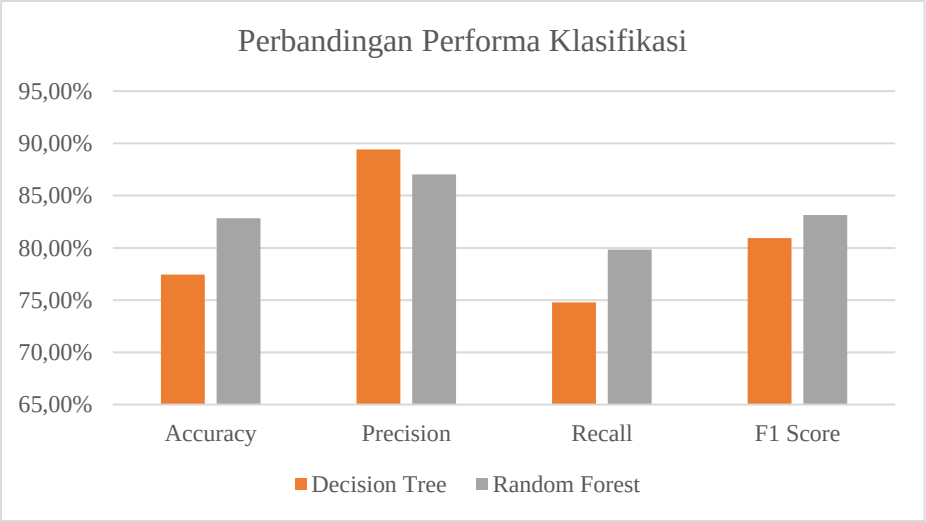
Klasifikasi *random forest* menujukkan hasil *mtry* sebesar 2. *Mtry* menjelaskan jumlah variabel yang akan digunakan sebagai *split* pada setiap pohon yang terbentuk. *Ntree* pada klasifikasi ini adalah 500 yang merupakan nilai *default* untuk *ntree. Ntree* menujukkan jumlah pohon yang dibentuk dalam klasifikasi *random forest.* Semakin banyak jumlah pohon yang dibentuk maka akan semakin baik.

Berdasarkan perhitungan menggunakan *9-fold cross validation*, maka didapatkan performa klasifikasi penyakit jantung menggunakan *decision tree* dan *random forest* seperti yang dituliskan pada tabel 6.

Tabel 6. Performa Klasifikasi

	Accuracy	*Precision*	*Recall*	*F1 Score*
Decision Tree	77.44%	89.42%	74.78%	80.92%
Random Forest	81.82%	87.04%	79.82%	83.13%

Berdasarkan pada tabel 6, perbandingan performa klasifikasi penyakit jantung menggunakan *decision tree* dan *random forest* juga digambarkan pada grafik pada gambar 6.



Gambar 6. Perbandingan Nilai Evaluasi Klasifikasi

e- ISSN: 2828–299X

Berdasarkan gambar 6 diperoleh nilai akurasi *random forest* lebih tinggi dari nilai akurasi *decision tree. Random forest* menghasilkan nilai akurasi sebesar 81.82% dan *decision tree* menghasilkan nilai akurasi sebesar 77.44%. Nilai *precision* pada *decision tree* adalah sebesar 89.42% dan nilai *precision* dari *random forest* adalah sebesar 87.04%. Nilai *recall* yang dihasilkan dari *random forest* adalah sebesar 79.82% dan nilai *recall decision tree* adalah sebesar 74.78%. Kemudian *F1 Score* dari *random forest* adalah sebesar 83.13% dan *F1 score decision tree* adalah sebesar 80.92%.

4. SIMPULAN

Berdasarkan pada hasil analisis, maka diperoleh klasifikasi penyakit jantung dengan *random forest* adalah lebih baik jika dibandingkan dengan klasifikasi penyakit jantung dengan *decision tree* karena ketepatan klasifikasi dengan *random forest* memberikan hasil lebih baik. Klasifikasi dengan *random forest* menghasilkan nilai akurasi 81.82%, nilai *precision* sebesar 87.04%, nilai *recall* sebesar 79.82%, dan *F1 score* sebesar 83.13%.

5. SARAN

Dengan berbagai keterbatasan dalam penelitian ini maka beberapa saran yang dapat diberikan untuk penelitian selanjutnya adalah sebagai berikut:

1. Pada penelitian selanjutnya dapat menambah atau memperbarui metode atau algoritma yang digunakan sehingga hasil yang diperoleh dapat lebih akurat.
2. Menggunakan *dataset* yang berjumlah lebih banyak untuk meningkatkan akurasi.

DAFTAR PUSTAKA

[1] Kementerian Kesehatan Republik Indonesia. 2022. https://sehatnegeriku.kemkes.go.id/ diakses pada 29 November 2022.

[2] Annisa, R. 2019. Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Penderita Penyakit Jantung. *Jurnal Teknik Informatika Kaputama (JTIK)*. No.1. Vol.3. 22-28.

[3] Azhima, dkk. 2022. *Hybrid Machine Learning Model* untuk Memprediksi Penyakit Jantung dengan Metode *Logistic Regression* dan *Random Forest. Jurnal Teknologi Terpadu.* No.1. Vol.8. 40-46.

[4] Alham S. R. J. I., Efy Y., dan Rizqia C. 2021. Sistem Diagnosis Penyakit Jantung Koroner Dengan Algoritma C4.5 Berbasis Website (Studi Kasus: RSUD Dr. Soedarso Pontianak). *PETIR : Jurnal Pengkajian dan Penerapan Teknik Informatika.* No.2. Vol.14. 214-222.

[5] Kaggle. Heart Disease Cleveland UCI. https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci diakses pada 3 November 2022.

[6] Mitchell, T. M. 1997. Machine Learning. McGraw-Hill.

[7] Breiman, L. 2002. Random Forest. *Machine Learning* (45, 5-32).