Jurnal Sistem Komputer dan Kecerdasan Buatan
Volume VII Nomor 1 September 2023

Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier

Hidayat[1], Andi Sunyoto[2], Hanif Al Fatta[3]

[123] Magister Teknik Informatika, Universitas Amikom Yogyakarta
[1]hidayat.densaiko@students.amikom.ac.id
[2]andi@amikom.ac.id
[3]hanif.a@amikom.ac.id

Diterima : 01 September 2023
Disetujui : 28 September 2023

***Abstract*— Penelitian ini bertujuan untuk meningkatkan peforma model saat melakukan klasifikasi peyakit jantung. Algoritma random fores digunakan untuk melakukan klasifikasi peyakit jantung berdasarkan fitur-fitur yang ada pada dataset. Klasifikasi ini dilakukan dengan menggunakan Heart Disease Dataset dari kaggle yang mempunyai 2 class diantaranya 0 (tidak terindikasi penyakit), dan 1 (terindikasi penyakit). Selanjutnya dataset tersebut dilakukan teknik pre-prosessing data, normalisasi data, split data, klasifikasi dan yang terakhir evaluasi metode. Penelitian ini mengungkapkan bahwa metode random forest berhasil menghasilkan tingkat akurasi yang lebih tinggi dalam proses klasifikasi penyakit jantung, dibandingkan dengan hasil penelitian sebelumnya yaitu mencapai akurasi sebesar 94%. Hal ini menunjukkan bahwa penggunaan Random forest dibantu dengan teknik pre-prosessing, dan normalisasi data dapat menjadi alternatif yang baik dalam melakukan klasifikasi. Penelitian ini memberikan manfaat saat klasifikasi penyakit jantung secara cepat dan akurat.**

Keywords* —Penyakit Jantung, *Random Forest

I. PENDAHULUAN

Penyakit jantung adalah suatu kondisi di mana bagian jantung, seperti pembuluh darah jantung, lapisan jantung, katup jantung, dan otot jantung menjadi tidak berfungsi, penyakit ini dapat memiliki banyak penyebab, seperti penyumbatan di arteri jantung, peradangan, infeksi, atau cacat lahir[1]. Penyakit kardiovaskular meningkat setiap tahun, sedikitnya 15 dari setiap 1.000 orang atau sekitar 2.784.064 orang di Indonesia menderita penyakit jantung[2].

Faktor risiko penyakit jantung antara lain pola hidup yang tidak sehat seperti mengonsumsi makanan tinggi karbohidrat atau berlemak, obesitas, jarang berolahraga, merokok, dan riwayat keluarga yang memainkan peran penting dalam risiko penyakit jantung[3]. Penyakit jantung juga merupakan penyakit yang memiliki beban biaya terbesar. Berdasarkan data BPJS kesehatan, penyakit jantung akan menjadi pengeluaran kesehatan terbesar di tahun 2021 sebesar Rp 7,7 triliun[4].

Terdapat beberapa riset terkait penyakit jantung ini dimana riset tersebut mengungkapkan bahwa penyakit jantung ialah salah satu penyakit yang perlu mendapatkan perhatian serius karena Serangan jantung yang parah atau terlambat ditangani bisa menyebabkan beberapa komplikasi berbahaya. Komplikasi tersebut antara lain gangguan irama jantung atau aritmia, gagal jantung, syok kardiogenik, dan henti jantung[5].

Beberapa cara yang dapat dilakukan untuk dapat membantu para petugas medis dalam menemukan apakah seseorang terindikasi penyakit jantung atau tidak agar ketika pasien yang mengalami penyakit tersebut ini dapat ketahui dengan cepat, salah satunya ialah dengan penggunaan Machine learning, dengan penggunaan ini terbukti mampu menyelesaikan topik klasifikasi, dan optimasi dalam pembuatan sebuah system penyedia layanan kesehatan[6].

Terdapat beberapa penelitian sebelumnya dengan study kasus yang sama yaitu klasifikasi penyakit jantung, dari penelitian yang ada

menggunakan beberapa metode machine learning untuk dapat memprediksi seseorang terindikasi penyakit jantung yaitu random forest classifier, ann, svm, naïve bayes, support vector machine, decision tree dll sebagainya, hasil penelitian sebelumnya memiliki hasil akurasi yang paling baik adalah 90% dengan menggunakan teknik seperti preprosessing data, penentuan hyperparameter, kombinasi metode balancing data dll sebagainya [7] [8] [9] [10] [11].

Rata-rata penelitian sebelumnya belum menerapkan teknik pre-prosessing data seperti mengubah nilai kosong pada atribut dataset, hal ini perlu dilakukan untuk mencegah hasil klasifikasi yang bias[12]. berikutnya terkait dengan belum adanya penerapan teknik normalisasi data, kenapa teknik ini perlu dilakukan karena untuk mencegah perbedaan skala yang terjadi pada dataset, jika hal ini terjadi maka model machine learning rata-rata menghasilkan hasil yang kurang optimal[13].

Tujuan dari penelitian ini adalah meningkatkatkan hasil akurasi pada identifikasi seseorang terindikasi penyakit jantung atau tidak dengan mengusulkan metode machine learning yaitu random forest, teknik pre-prosessing data dengan mengatasi data kosong pada dataset, dan teknik normalisasi data untuk mendapatkan hasil akurasi yang lebih baik dari akurasi yang dihasilkan sebelumnya, untuk menilai kinerja model maka penelitian ini menggunakan Confusion matrix untuk melakukan klasifikasi terhadap metode yang di pakai pada proses identifikasi penyakit jantung.

II. TEORI

A. Penyakit Jantung

Penyakit jantung merupakan kondisi ketika jantung manusia mengalami gangguan. Beberapa jenis penyakit jantung meliputi: [14]:

1. Penyakit jantung koroner, merupakan suatu penyakit jantung yang terjadi akibat penyempitan pembuluh darah di jantung.
2. Penyakit jantung bawaan, merupakan suatu masalah jantung yang ditemukan sejak bayi, yang paling umum ditemukan adalah kebocoran katup jantung.
3. Infeksi jantung (endokarditis), merupakan suatu infeksi pada lapisan dalam jantung.

Penelitian ini membahas topik penyakit jantung yang lebih mengarah ke jenis penyakit jantung coroner. Penyakit jantung coroner adalah penyakit yang disebabkan oleh penyumbatan pembuluh darah utama yang mengalirkan pasokan oksigen, darah, dan nutrisi untuk jantung[15]. Umumnya, kondisi ini merupakan dampak dari adanya plak kolesterol dan peradangan pada pembuluh darah arteri di jantung. Terdapat beberapa penyebab terjadinya penyakit jantung koroner di antaranya sebagai berikut [16]:

1. Hipertensi (tekanan darah tinggi)
2. Diabetes
3. Berat badan berlebih
4. Peradangan pada pembuluh darah
5. Kebiasaan merokok
6. Dan Kadar kolesterol dan trigliserida tinggi.

B. *Machine Learning*

Pembelajaran mesin (Machine Learning) adalah metode komputer yang tidak perlu didefinisikan oleh manusia dan dapat belajar dari data, metode ini menjadi lebih pintar saat lebih banyak data diproses (belajar berdasarkan pengalaman)[17]. Metode ini banyak digunakan untuk menyelesaikan kasus klasifikasi dan clustering dan biasanya digunakan untuk mengolah dataset yang besar atau dataset yang besar[18]. Ada tiga cabang utama pembelajaran mesin yaitu [19]:

1. Suvervised machine learning.
   Sistem ini awalnya diberi data yang telah diberi label, dan kemudian mengelompokkan setiap titik data ke dalam satu atau beberapa kelompok berdasarkan label tersebut. Sistem kemudian menggunakan data yang telah dikelompokkan ini sebagai data pelatihan terstruktur untuk belajar bagaimana data tersebut diproses. Dari data pelatihan ini, sistem dapat

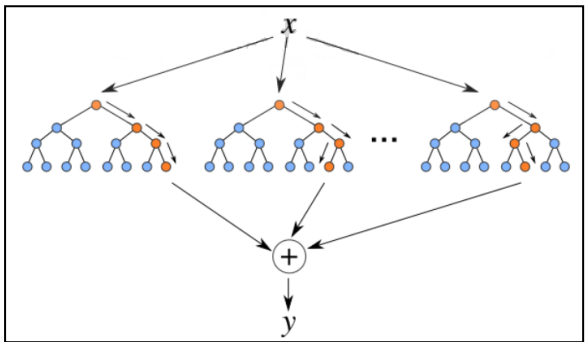
memprediksi atau mengklasifikasikan data uji atau data test.

2. Unsupervised learning.
   Supervised machine learning adalah kebalikan dari Unsupervised learning, di mana Unsupervised learning merupakan jenis pembelajaran tanpa pengawasan. Ini berarti bahwa analisis dilakukan pada data yang tidak memiliki label atau informasi klasifikasi sebelumnya.
3. Reinforcement learning.
   Pembelajaran yang menguatkan pemahaman berdasarkan pengalaman dapat ditemukan baik dalam pembelajaran supervised maupun unsupervised.

C. Random Forest

Breiman memperkenalkan algoritma Random Forest pada tahun 2001, algoritma ini memiliki kemampuan untuk menangani dua jenis masalah, yakni klasifikasi dan regresi[20]. Random Forest merupakan hasil pengembangan dari metode Classification dan Regression Tree (CART), yang menggunakan metode bag or bootstrap aggregation dan random feature selection. Bagging merupakan salah satu teknik yang dapat digunakan untuk memperbaiki hasil dari suatu algoritma klasifikasi. Metode bagging ini didasarkan pada metode ensemble [21]. Metode algoritma hutan acak dapat dibagi menjadi dua tahap, tahap pertama melibatkan pembuatan "k" pohon untuk membentuk hutan acak, sedangkan tahap kedua menggunakan hutan acak yang telah dibentuk untuk membuat prediksi[22]. Proses penerapan metode Random Forest melibatkan beberapa langkah, seperti yang dijelaskan dalam [23]:

1. Pertama-tama, data sampel dibuat dengan mengambil contoh data secara acak dari dataset, dan ini dilakukan dengan pengembalian data yang sudah diambil.
2. Setelah itu, sampel data digunakan untuk membangun pohon ke-i, di mana i adalah iterasi yang berjalan dari 1 hingga k.
3. Langkah 1 dan 2 diulangi sebanyak k kali sesuai dengan jumlah pohon yang ingin dibangun dalam hutan acak.



Gambar 1. Random Forest
Sumber: Morioh.com

Dalam konstruksi pohon keputusan menggunakan metode CART, komputasi melibatkan verifikasi informasi yang menjelaskan seberapa penting atribut dalam mengklasifikasikan setiap simpul pohon. Secara khusus, jika kita menganggap N sebagai simpul yang memisahkan kelas data D berdasarkan atribut-atributnya, maka komputasi ini membantu mengukur seberapa relevan atau informatif setiap atribut tersebut dalam proses pemisahan kelas data. Proses pembagian simpul dilakukan dengan memilih atribut yang memiliki tingkat informasi validasi tertinggi. Rumus yang digunakan untuk menghitung tingkat informasi validasi adalah sebagai berikut :

$$Gain\,(A) = Info\,(D) - Info(D) \quad (1)$$

Untuk mendapatkan nilai info(D), kita dapat menghitungnya menggunakan rumus 2 dan 3, yang akan menghasilkan nilai info A(D). :

$$Info(D) = -\sum_{i=1}^{n} p_i log2(p_i) \quad (2)$$

Keterangan :
n = jumlah kelas target
pi = proporsi kelas i terhadap partisi D

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (3)$$

Keterangan :
v = jumlah partisi.
Dj = total partisi ke j.
D = total baris pada semua partisi.

Untuk atribut yang berisi nilai kontinu atau numerik, perlu menentukan titik pembagian terbaik untuk mengelompokkan nilai. Proses menemukan resolusi terbaik dimulai dengan mengurutkan data. Median atau rata-rata dari setiap pasangan nilai yang berdekatan digunakan sebagai titik pembagian yang mungkin. Sebagai contoh, jika atribut A adalah atribut bernilai

kontinu, maka semua nilai A diurutkan dan nilai tengahnya menjadi salah satu titik pembagian yang mungkin. Hal ini dapat menghasilkan dua atau lebih partisi, di mana dalam contoh ini v = 2 (dengan j=1 dan 2) adalah kemungkinan jumlah partisi[24].

III. METODOLOGI

Penelitian ini bertujuan untuk meningkatkan akurasi dalam mendeteksi indikasi penyakit jantung dengan melakukan proses klasifikasi menggunakan metode Random Forest Classifier, sehingga diharapkan hasilnya dapat mengungguli tingkat akurasi penelitian sebelumnya. berikut ini deskripsi langkah-langkah yang dilakukan oleh penelitian ini, sebagai berikut :

A. Pengambilan dataset.

Proses awal ini penulis mencari dataset yang sesuai dengan topik penelitian ini, maka dengan itu penulis mengambil dataset publik yang berasal dari kaggle yang nantinya akan digunakan untuk proses pengolahan klasifikasi.

B. *Pre-prosessing data*

Proses ini dilakukan untuk memeriksa data pada dataset yang dipilih dan memperbaiki kesalahan yang ditemukan pada dataset tersebut, sehingga dapat dilanjutkan pada langkah berikutnya. Teknik pada proses ini yang digunakan dalam penelitian yaitu mengatasi missing value dan normalisasi data. Penting untuk mengatasi missing value pada proses klasifikasi karena akan terdapat informasi yang hilang, bias dalam analisis, dan dapat menurunkan hasil akurasi saat proses klasifikasi[25]. Selain itu data yang memiliki skala yang berbeda-beda dapat menyebabkan masalah dalam analisis dan klasifikasi, normalisasi membantu menghilangkan perbedaan skala ini dan mengatur semua variabel pada rentang yang serupa, untuk memastikan bahwa tidak ada variabel yang mendominasi proses klasifikasi hanya karena memiliki skala yang lebih besar, sehingga mencegah bias dalam model[26].

C. *Normalisasi data*

Proses mengubah atribut numerik dalam dataset sehingga memiliki skala yang seragam atau sebanding. proses ini menggunakan metode min-max normalisasi, Metode ini dapat menggunakan rumus sebagai berikut :

$$N = \frac{MinRange + (X - MinValue)(MaxRange - MinRange)}{MaxValue - MinValue} \quad (4)$$

Dimana :

N = Normalisasi Min_Max

MinRange = Nilai Konversi Kecil Yang ditentukan.

MaxRange = Nilai Konversi Terbesar yang ditentukan.

MaxValue = Nilai Terbesar pada atribut yang dibandingkan.

MinValue = Nilai Terkecil pada atribut yang dibandingkan.

D. *Split data*

Tahap ini merupakan tahapan pembagian dataset menjadi data training dan data testing, pembagian pada penelitian ini dibagikan data menjadi 80/20. Secara umum model machine learning mendapatkan hasil akurasi yang baik jika memiliki jumlah data testing yang sedikit. Maka dalam penelitian ini meningkatkan data test dan menguji apakah model mendapatkan hasil yang baik atau tidak.

E. Klasifikasi dengan *Random forest*

Selanjutnya yaitu melakukan klasifikasi dengan metode random fores, tahap ini melakukan identifikasi pasien yang terindikasi penyakit atau tidak.

F. *Evaluasi Method*

Proses ini dilakukan untuk menilai kinerja dari model yang telah dirancang sebelumnya untuk melakukan klasifikasi penyakit jantung, jika hasil akurasi yang dihasilkan tidak melebihi hasil dari penelitian sebelumnya maka kembali ke proses klasifikasi random forest dan mengatur parameter hingga menghasilkan akurasi yang lebih baik. Jika hasil akurasi sudah melebihi hasil dari penelitian sebelumnya maka dilanjutkan pada proses berikutnya yaitu pengambilan kesimpulan. Confusion matrix digunakan untuk menilai peforma dari model yang digunakan, matriks evaluasi ini, merupakan tabel yang digunakan untuk menghitung dan mengevaluasi kinerja model klasifikasi, tabel ini didasarkan pada jumlah kasus yang

diklasifikasikan benar dan salah menurut model dan informasi ini dapat dilihat pada Tabel 1.

TABEL 1 Confusion Matrix
Sumber: [27]

Classification	*Predicted Positive*	*Predicted Negative*
Actual Positive	*True Positive (TP)*	*False Negative (FN)*
Actual Negative	***False Positive (FP)***	***True Negative (TN)***

Dalam pengukuran kinerja menggunakan confusion matrix, terdapat empat komponen yang digunakan untuk mengidentifikasi hasil prediksi, yaitu: [28]:

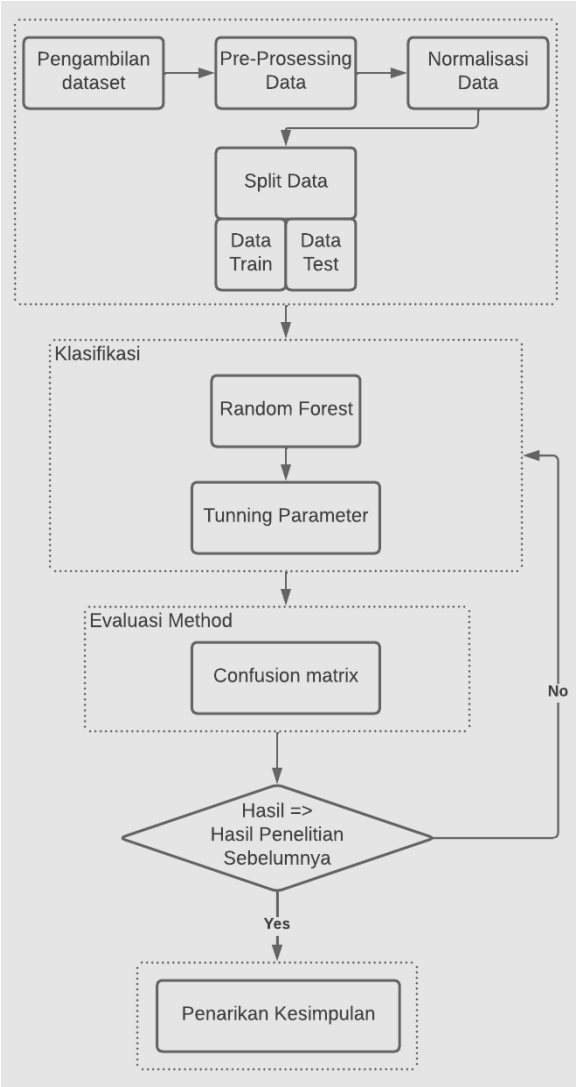
1. TP (True Positive) ini adalah jumlah data yang memiliki nilai aktual positif dan telah diprediksi dengan benar sebagai positif.
2. TN (True Negative) adalah merupakan jumlah data yang memiliki nilai aktual negatif dan telah diprediksi dengan benar sebagai negatif..
3. FP (False Positive) ini adalah jumlah data yang memiliki nilai aktual negatif, tetapi diprediksi sebagai positif dengan tidak benar.
4. FN (False Negative) merupakan jumlah data yang memiliki nilai aktual positif, tetapi diprediksi sebagai negatif dengan tidak benar.

Accuracy merupakan nilai evaluasi yang sering dipakai pada klasifikasi biner, akurasi dapat ditemukan berdasarkan nilai yang ada pada confusion matrix, Accuracy (ACC) adalah efektivitas dari hasil yang didapatkan dalam proses klasifikasi, accuracy dilakukan menggunakan persamaan (5) berikut[29].

$$Accuracy\ (\%) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (5)$$

G. *Penarikan Kesimpulan*

Tahap ini merupakan tahap pemberian kesimpulan berdasarkan hasil dari pengujian yang telah dilakukan. Hasil penelitian berupa fakta yang diperoleh metode yang di terapkan untuk pediksi penyakit jantung. Hasil pengujian dan evaluasi dijadikan kesimpulan akhir mengenai penerapan teknik pre-prosessing data, algoritma random forest untuk proses klasifikasi, dan parameter yang diujikan. Keseluruhan proses penelitian ini dapat dilihat pada Gambar 2. Metodologi Penelitian.



Gambar 2. Metodologi Penelitian

IV. HASIL DAN DISKUS

A. Pengambilan dataset

Penelitian ini menggunakan dataset penyakit jantung yang diambil dari kaggle https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download, data ini dipakai untuk melatih dan menguji model saat melakukan klasifikasi khususnya pada klasfikasi penyakit jantung. berikut ini merupakan tampilan dataset yang dapat dilihat pada gambar 3.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0

Gambar 3. Dataset.

Pada gambar 3 dataset ini mempunyai 2 class yaitu 0 (tidak ada penyakit) sebanyak 499 data dan 1 (penyakit) sebanyak 526 data.

B. Pre-prosessing data

Terdapat suatu teknik pre-prosessing yang dilakukan sebelum dilakukannya proses berikutnya, teknik tersebut yaitu mengatasi missing value. Dataset ini terdapat nilai kosong pada fitur dataset yaitu pada atribut thalach sebanyak 36 data yang kosong, bisa dilihat pada gambar 4 berikut ini.

```
1  data.isna().sum()

age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach    36
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
```

Gambar 4. Missing value

atribut missing value tersebut diubah dengan rata-rata atribut thalach mengalami penyakit dan thalach tidak mengalami penyakit menjadi pilihan, dibandingkan dengan menghapus baris data yang terdapat nilai kosong, karena tidak ada data yang terbuang. hasil dari proses ini terdapat pada gambar 5.

```
1  mean_thalach_has_jantung = data[data['target']==1]['thalach'].mean()
2  mean_thalach_has_jantung

158.61023622047244

1  mean_thalach_has_No_jantung = data[data['target']==0]['thalach'].mean()
2  mean_thalach_has_No_jantung

138.985446985447

data.loc[data['target']==1,'thalach'] =
data.loc[data['target']==1, 'thalach'].fillna(mean_thalach_has_jantung)
data.loc[data['target']==0,'thalach'] =
data.loc[data['target']==0, 'thalach'].fillna(mean_thalach_has_No_jantung)
```

Gambar 5. hasil Replace missing value

pada gambar 5. merupakan proses mengubah nilai missing pada fitur dataset dengan rata-rata rata-rata atribut thalach mengalami penyakit dan thalach tidak mengalami penyakit. Setelah proses replace missing value dilakukan maka kembali untuk pengecekan nilai missing value pada dataset apakah sudah berhasil diganti atau tidak yang dapat dilihat pada gambar 6.

```
1  data.isna().sum()

age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Gambar 6. Missing value

C. Normalisasi Data

Teknik normalisasi data yang digunakan adalah Min Max normalization yang merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar atribut, atribut-atribut ini ketika dikonversi dan menghasilkan perhitungan yang similaritas, maka nantinya dapat berada di rentang 0 hingga 1. Berikut ini hasil standarisasi menggunakan teknik normalisasi min_max pada dataset penyakit jantung yang dapat dilihat pada gambar 7.

```
1  dataNormalisasi = min_max_scaler.fit_transform(X) #transformasi MinMax untuk fitur

1  dataNormalisasi

array([[0.47916667, 1.        , 0.        , ..., 1.        , 0.5       ,
        1.        ],
       [0.5       , 1.        , 0.        , ..., 0.        , 0.        ,
        1.        ],
       [0.85416667, 1.        , 0.        , ..., 0.        , 0.        ,
        1.        ],
       ...,
       [0.375     , 1.        , 0.        , ..., 0.5       , 0.25      ,
        0.66666667],
       [0.4375    , 0.        , 0.        , ..., 1.        , 0.        ,
        0.66666667],
       [0.52083333, 1.        , 0.        , ..., 0.5       , 0.25      ,
        1.        ]])
```

Gambar 7. Normalisasi data.

D. Split Data

Membagikan dataset menjadi data training dan data testing, pembagian ini dilakukan menjadi data training dan data testing. pembagian tersebut mempunyai tujuan melihat model dalam memprediksi ketika mempunyai data test 20% dengan total 205 dan peningkatan pada data test menjadi 30% dengan total 308. Tabel 2 menggambarkan pembagian data yang di lakukan.

Tabel 2. Train/Test Split

Keterangan	**DataTrain**	**Data Test**	**Jumlah**
Proporsi	80%	20%	100%
Jumlah	820	205	1025
Keterangan	**DataTrain**	**Data Test**	**Jumlah**
Proporsi	70%	30%	100%
Jumlah	717	308	1025

Tabel 2. di jelaskan bahwa terdapat 2 pembagian data training dan data testing yang dilakukan pertama yaitu membagikan/split data menjadi 80/20, 80% untuk data training yang berjumlah 820 data dan 20% untuk data testing yang berjumlah 205 data. Pembagian kedua memiliki tujuan untuk menguji model saat melakukan klasifikasi dengan data test yang dinaikan menjadi 30% yang berjumlah 308 data, apakah model masih dapat melakukan klasifikasi dengan tepat atau tidak ketika data uji ditingkatkan.

E. Klasifikasi dengan Random Forest.

Metode yang digunakan pada penelitian ini adalah metode random forest untuk melakukan klasifikasi penyakit jantung. Kekuatan sistem saat melakukan klasifikasi bergantung pada proses pelatihan untuk menghasilkan akurasi. Algoritma Random Forest, memiliki beberapa parameter yang umumnya diujikan atau disesuaikan untuk meningkatkan kinerja dan mengoptimalkan model.

Beberapa parameter tersebut diantaranya [30]:

1. *n_estimators*.
   Jumlah pohon keputusan dalam ensemble (kumpulan) Random Forest. Semakin banyak pohon yang digunakan, semakin baik kemungkinan model akan tahan terhadap overfitting, tetapi juga semakin mahal waktu komputasinya.
2. *max_depth*.
   max_depth (kedalaman maksimum) pada setiap pohon keputusan mengatur kompleksitas model. Tanpa pengaturan ini, pohon akan terus tumbuh hingga semua daunnya murni atau sampai jumlah sampel minimum di setiap daun tercapai.
3. *min_samples_leaf*.
   Jumlah minimum sampel yang diperlukan dalam setiap leaf (simpul terakhir) dari pohon. Nilai ini membantu mencegah pembentukan leaf dengan sedikit sampel yang mungkin menyebabkan overfitting.
4. *bootstrap*.
   Mengontrol apakah sampel akan diambil dengan penggantian saat membangun setiap pohon. Jika diatur sebagai True, setiap pohon akan dibangun dari sampel bootstrap, dan jika diatur sebagai False, setiap pohon akan dibangun dari seluruh dataset.
5. dan *random_state*.
   Merupakan bilangan bulat untuk mengontrol inisialisasi angka acak yang digunakan oleh model. Pengaturan nilai ini memastikan hasil yang konsisten setiap kali model dilatih, sehingga memudahkan perbandingan antara model yang berbeda.

Beberapa pengujian dengan pengaturan parameter random forest yang telah dilakukan pada klasifikasi penyakit jantung dapat dilihat pada Tabel 3 dan Tabel 4.

Tabel 3. Pengujian Random forest Pertama.

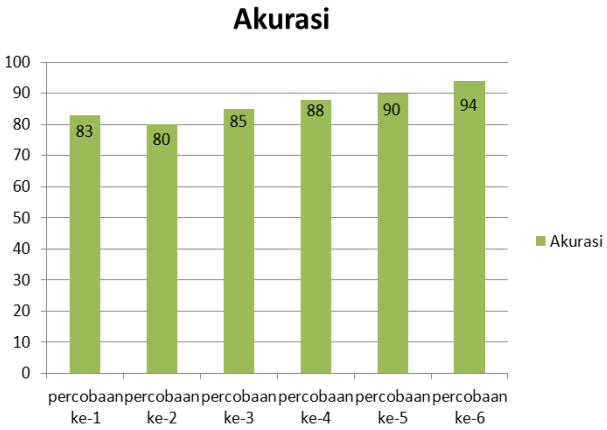
Split data 80/20						
Pengujian ke	*1.*	*2.*	*3.*	*4.*	*5.*	*6.*
n_ estimators	50	70	80	90	100	120
max_depth	5	7	9	12	14	16
min_ samples_ leaf	2	1	5	3	4	6
bootstrap	true	true	true	true	true	true
random_ state	32	32	42	12	32	42
Acc	83%	80%	85%	88%	90%	94%

Tabel 4. Pengujian Random forest Kedua.

Split data 70/30						
Pengujian ke	*7*	*8.*	*9.*	*10.*	*11.*	*12.*
n_ estimators	50	70	80	90	100	120
max_depth	5	7	9	12	14	16
min_ samples_ leaf	2	1	5	3	4	6
bootstrap	true	true	true	true	true	true
random_ state	32	32	42	12	32	42
Acc	81%	80%	82%	86%	89%	92%

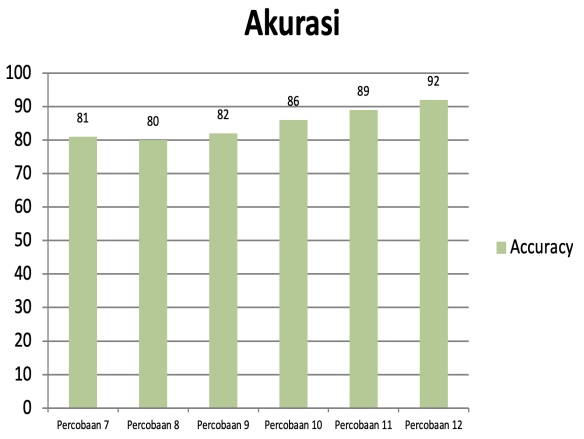
Pada Tabel 3 dan Tabel 4 diatas dapat dilihat bahwa parameter yang diujikan pada proses klasifikasi adalah *max_depth, n_estimator, random_state, boostrap,* dan *min_samples_leaf.* Split data yang digunakan pada proses klasifikasi adalah 80/20 dan 70/30. Pengujian pertama memiliki 6 kali percobaan, dan untuk pengujian kedua juga memiliki 6 kali percobaan. Percobaan di atur nilai parameter yang rendah lalu pada percobaan berikutnya diatur dengan nilai parameter yang meningkat. parameter yang terjadi peningkatan value hanya pada *max_depth,* dan *n_estimator*. Parameter yang terjadi penurunan value hanya pada parameter *random_state.* Terakhir parameter yang nilai nya tetap hanya pada parameter *boostrap* yang di atur sebagai "*true*".

Pengujian dengan pembagian data 80/20 ini mendapatkan hasil terbaik yaitu pada pengujian ke-6 dengan tingkat akurasi yang dimiliki adalah 94%, yang mempunyai 120 pohon yang digunakan, kemudian memiliki tingkat kedalaman maksimum sebanyak 16, setiap leaf (simpul terakhir) dari setiap pohon keputusan dalam ensemble *Random Forest* harus memiliki setidaknya 6 sampel (data) dalamnya, lalu sampel yang digunakan untuk membangun setiap pohon keputusan dalam ensemble *Random Forest* akan diambil dengan penggantian dari dataset pelatihan, dan 42 digunakan sebagai nilai inisialisasi untuk generator angka acak yang digunakan oleh model. hasil akurasi dari keseluruhan percobaan-percobaan menggunakan split data 80/20. Dapat dilihat pada gambar 8.



Gambar 8. Hasil akurasi split data 80/20.

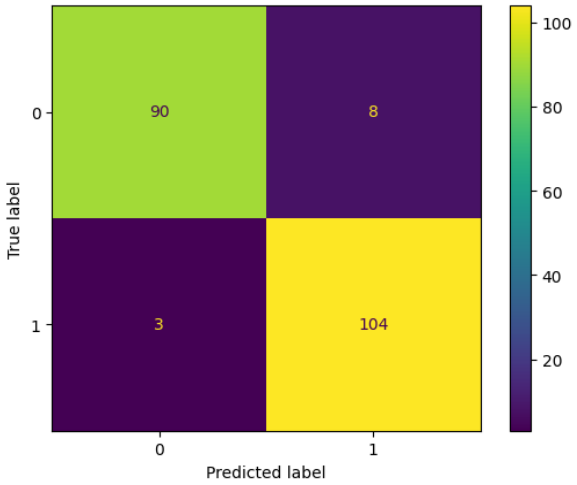
Untuk pengujian klasifikasi pada pembagian data 70/30 mendapatkan hasil terbaik yaitu pada pengujian ke-12 dengan tingkat akurasi 92%. Hasil terbaik ini memiliki nilai parameter yang sama seperti hasil terbaik pada pengujian pertama yang diujikan pada pembagian data 80/20 di Tabel 3. Hasil keseluruhan dari proses klasifikasi ini dapat dilihat pada gambar 9 yang menggunakan split data 70/30.



Gambar 9. Hasil akurasi split data 70/30.

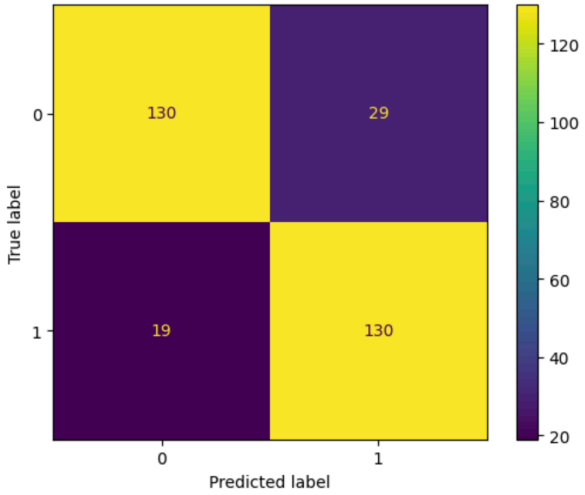
F. *Evaluasi Methood*

Penelitian ini melakukan 12 pengujian dengan menggunakan metode klasifikasi Random Forest, percobaan tersebut menghasilkan akurasi terbaik adalah percobaan pertama yaitu pada pembagian data 80/20 yang keenam. Confusion matrix dari percobaan pertama yang keenam (Percobaan terbaik) ditunjukkan pada Gambar 9.



Gambar 9. Hasil confusion matrix split data 80/20.

Gambar 9. merupakan hasil evaluasi confusion matrix, dari hasil tersebut memperoleh sebanyak delapan data yang salah diprediksi oleh model yang seharusnya class 0 (tidak terindikasi penyakit) tetapi di prediksi sebagai class 1 (terindikasi penyakit), kemudian terdapat 3 data yang seharusnya class 1 (terindikasi penyakit) tetapi diprediksi sebagai class 0 (tidak terindikasi penyakit). Eksperimen pertama mendapatkan hasil terbaik dengan nilai akurasi sebesar 94% yang dapat dilihat pada Gambar 9. Hasil pengujian kedua menggunakan split data 70/30 mendapatkan hasil terbaik yaitu pada eksperimen ke-12 dengan akurasi 92%, pengujian ini menggunakan parameter yang sama dengan pengujian yang dilakukan saat menggunakan split data 80/20, hasil dari pengujian ini tidak lebih baik dari pengujian pertama menggunakan split data 80/20 yang dikarenakan jumlah data training yang berkurang dan data testing yang bertambah hal membuat model kekurangan pembelajaran untuk menentukan data uji. Berikut ini hasil evaluasi confusion matrix terbaik dari pengujian menggunakan split data 70/30 yang dapat dilihat pada gambar 10.



Gambar 10. Hasil confusion matrix split data 80/20.

Gambar 10. merupakan hasil evaluasi dari pengujian terbaik yaitu pada pengujian ke dua belas menggunakan confusion matrix. Sebanyak 29 data yang salah diprediksi oleh model seharusnya class 0 (tidak terindikasi penyakit) tetapi di prediksi sebagai class 1 (terindikasi penyakit), kemudian terdapat 19 data yang seharusnya class 1 (terindikasi penyakit) tetapi diprediksi sebagai class 0 (tidak terindikasi penyakit).

Akurasi ini merupakan rasio prediksi yang tepat dalam mengidentifikasi seseorang terindikasi penyakit dan tidak terindikasi penyakit secara keseluruhan dalam dataset. Jumlah data yang diprediksi salah pada pengujian menggunakan split data 80/20 adalah 11 data, dan total data yang benar di klasifikasikan adalah 194 data dari total seluruh data uji yang berjumlah 205 data. Kesalahan prediksi pada split data 70/30 adalah 48 data, dan total data yang benar di klasifikasikan adalah 260 data dari total seluruh data uji yang berjumlah 308 data. Rata-rata terjadi penurunan akurasi 2% saat melakukan uji data split 70/30, penurunan tersebut terdapat pada pengujian ke 7, 9, 10 dan 12.

V. SIMPULAN

Berdasarkan analisis yang dilakukan dengan menggunakan dataset penyakit jantung yang memiliki 2 class, dapat disimpulkan bahwa :

- Penelitian ini terdapat dua belas pengujian, hasil yang terbaik yaitu pada pengujian pertama yang ke-6 menggunakan split data 80/20 dengan akurasi 94% dibandingkan dengan pengujian menggunakan pembagian

data 70/30 yang menghasilkan akurasi terbaik yaitu 92%. akurasi terbaik ini terjadi peningkatan 4% dan 2% dari hasil akurasi yang dihasilkan penelitian sebelumnya yaitu 90% dan akurasi yang dihasilkan menggunakan split data 70/30.

- Terdapat beberapa teknik penelitian yang membantu menemukan kondisi model yang tepat sehingga menghasilkan akurasi yang baik saat proses klasifikasi dilakukan yaitu : pre-prosessing data untuk memperbaiki kesalahan dari dataset, normalisasi data untuk mengatasi selisih nilai yang ada pada fitur dataset, berikutnya klasifikasi serta pengaturan parameter model.

DAFTAR PUSTAKA

[1] Pittara, "Pengertian Penyakit Jantung," 2023. https://www.alodokter.com/penyakit-jantung.

[2] P. K. RI, "Hari Jantung Sedunia (World Heart Day): Your Heart is Our Heart Too," 2019.

[3] R. Fadli, "Penyakit Jantung," 2022. https://www.halodoc.com/kesehatan/penyakit-jantung.

[4] Rokom, "Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer," 2022.

[5] puskesmas sungai durian, "bahaya serangan jantung," 2023. Https://puskesmas.kuburayakab.go.id/sungai-durian/read/173/bahaya-serangan-jantung.

[6] L. A. Dewi, "Klasifikasi Machine Learning Untuk Mendeteksi Penyakit Jantung Dengan Algoritma K-Nn, Decision Tree dan Random Forest," *Repository.Uinjkt.Ac.Id*, 2023.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] A. K. S. & S. C. M. Kshyanaprava Panda Panigrahi, Himansu Das, "Maize Leaf Disease Detection and Classification Using Machine Learning Algorithms," *Prog. Comput. Anal. Netw.*, vol. volume 111, 2020.

[9] M. A. Bianto, K. Kusrini, and S. Sudarmawan, "Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 75, 2020, doi: 10.24076/citec.2019v6i1.231.

[10] A. Nurmasani and Y. Pristyanto, "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class," *Pseudocode*, vol. 8, no. 1, pp. 21–26, 2021, doi: 10.33369/pseudocode.8.1.21-26.

[11] P. Singh and I. S. Virk, "Heart Disease Prediction Using Machine Learning Techniques," *2023 Int. Conf. Artif. Intell. Smart Commun. AISC 2023*, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.

[12] R. A. Maula *et al.*, "Handling Missing Value dengan Pendekatan Regresi pada Dataset Akuakultur Berukuran Kecil," *J. Rekayasa Elektr.*, vol. 18, no. 3, pp. 175–184, 2022, doi: 10.17529/jre.v18i3.25903.

[13] E. Irawan and R. S. Wahono, "Penggunaan Random Under Sampling untuk Penanganan Ketidakseimbangan Kelas pada Prediksi Cacat Software Berbasis Neural Network," *J. Softw. Eng.*, vol. 1, no. 2, pp. 92–100, 2015.

[14] Lely Puspita Candra Dewi, "Jenis, Gejala, dan Penyebab Penyakit Jantung," 2021. https://rs-soewandhi.surabaya.go.id/jenis-gejala-dan-penyebab-penyakit-jantung/.

[15] F. R. Makarim, "Penyakit Jantung Koroner," 2023. https://www.halodoc.com/kesehatan/penyakit-jantung-koroner.

[16] M. S. Hospitals, "Kenali Faktor Risiko Penyakit Jantung Koroner Sejak Dini," 2023. https://www.siloamhospitals.com/informasi-siloam/artikel/faktor-risiko-penyakit-jantung-koroner.

[17] J. J. Pangaribuan, H. Tanjaya, and Kenichi, "Mendeteksi Penyakit Jantung Menggunakan Mechine Learning Dengan Algoritma Logistic Regression," *Mach. Learn.*, vol. 45, no. 13, pp. 40–48, 2021.

[18] Dqlab, "4 Algoritma Data Science untuk Klasifikasi dan Clustering," 2022. https://dqlab.id/4-algoritma-data-science-untuk-klasifikasi-dan-clustering.

[19] I. D. Id, *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python*. 2021.

[20] L. Fadilah, *Klasifikasi Random Forest pada Data Imbalanced Program Studi Matematika Universitas Islam Negeri Syarif Hidayatullah 2018 / 1439 H Klasifikasi Random Forest*. 2018.

[21] N. K. Dewi, S. Y. Mulyadi, and U. D. Syafitri, "Penerapan Metode Random Forest Dalam Driver Analysis," *Forum Stat. Dan Komputasi*, vol. 16, no. 1, pp. 35–43, 2012, [Online]. Available: http://journal.ipb.ac.id/index.php/statistika/article/view/5443.

[22] M. L. Suliztia, "Penerapan Analisis Random Forest Pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask," *Fak. Mat. Dan Ilmu Pengetah. Alam*, pp. 1–107, 2020.

[23] P. Choirunisa, "Implementasi Artificial Inteligence Untuk Memprediksi Harga Penjualan Rumah Menggunakan Metode Random Forest dan Flask (Tugas Akhir)," 2020.

[24] R. A. Haristu and P. H. P. Rosa, "Penerapan Metode Random Forest untuk Prediksi Win Ratio Pemain Player Unknown Battleground," *MEANS (Media Inf. Anal. dan Sist.*, vol. 4, no. 2, pp. 120–128, 2019, doi: 10.54367/means.v4i2.545.

[25] D. B. Little, R. J. A., & Rubin, *Statistical analysis with missing data (2nd ed.). Wiley.* 2002.

[26] J. Hastie, T., Tibshirani, R., & Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer US, 2009.

[27] Kuncahyo Setyo Nugroho, "Confusion Matrix untuk Evaluasi Model pada Supervised Learning," 2019. https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f.

[28] Maria Susan Anggreany, "Confusion Matrix," 2020. .

[29] L. Afifah, "Apa itu Confusion Matrix di Machine Learning," 2023. https://ilmudatapy.com/apa-itu-confusion-matrix/.

[30] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.