

FINAL PROJECT

DATA MINING & VISUALISASI

**KLASIFIKASI RISIKO SERANGAN JANTUNG
MENGGUNAKAN METODE NAIVE BAYES
DAN DECISION TREE**



ANGGOTA KELOMPOK



ANDREW PUTRA HARTANTO
5003211016



RAHMANNUAJI SATUHU
5003211125

DEMO R SHINY

SCAN ME



<https://its.id/m/DashboardHeartAnalysis>

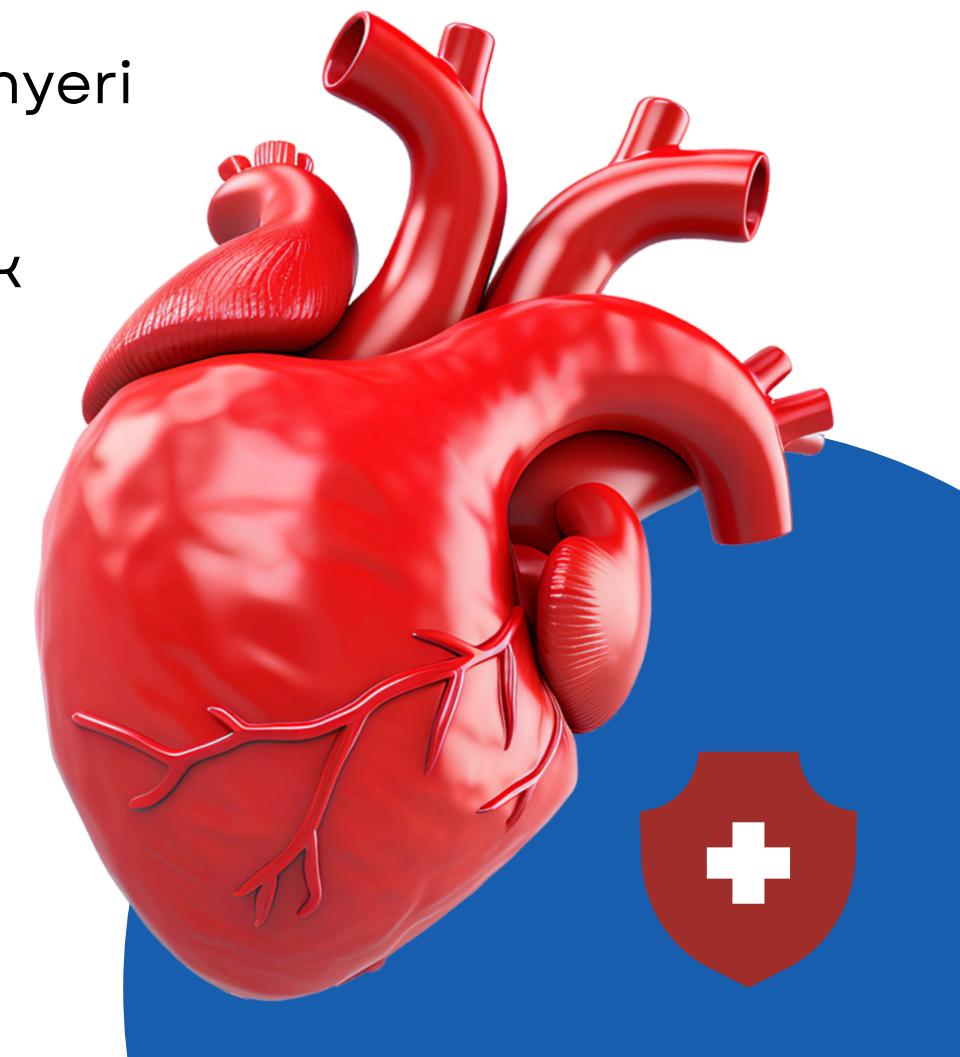


PROBLEM DATASET

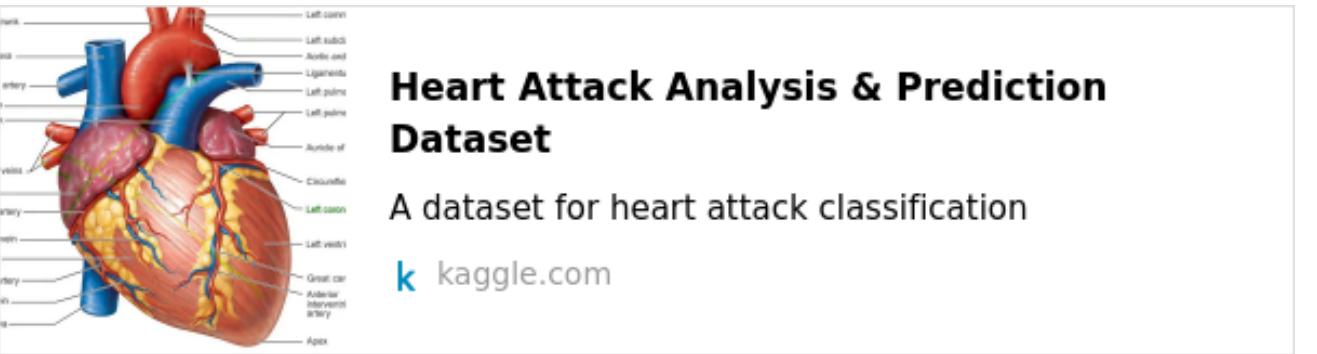
Penyakit kardiovaskular, terutama serangan jantung, merupakan salah satu penyebab utama kematian di seluruh dunia, yang menimbulkan tantangan kesehatan masyarakat yang signifikan. Penelitian ini mengeksplorasi hubungan yang kompleks antara berbagai indikator kesehatan dan potensi kontribusinya terhadap penyakit jantung dengan tujuan mengembangkan model prediksi yang akurat untuk deteksi dini risiko serangan jantung.

Dataset yang digunakan mencakup berbagai faktor seperti usia, jenis kelamin, jenis nyeri dada, tekanan darah istirahat, kolesterol serum, gula darah puasa, dan hasil elektrokardiografi istirahat. Masing-masing fitur ini memberikan perspektif yang unik tentang profil kesehatan individu, menawarkan wawasan penting tentang berbagai atribut perkembangan penyakit jantung.

Dengan menyertakan variabel yang mengindikasikan diagnosis medis penyakit kardiovaskular, data ini menjadi sangat berguna untuk prediksi yang ditargetkan untuk deteksi dan pencegahan dini.

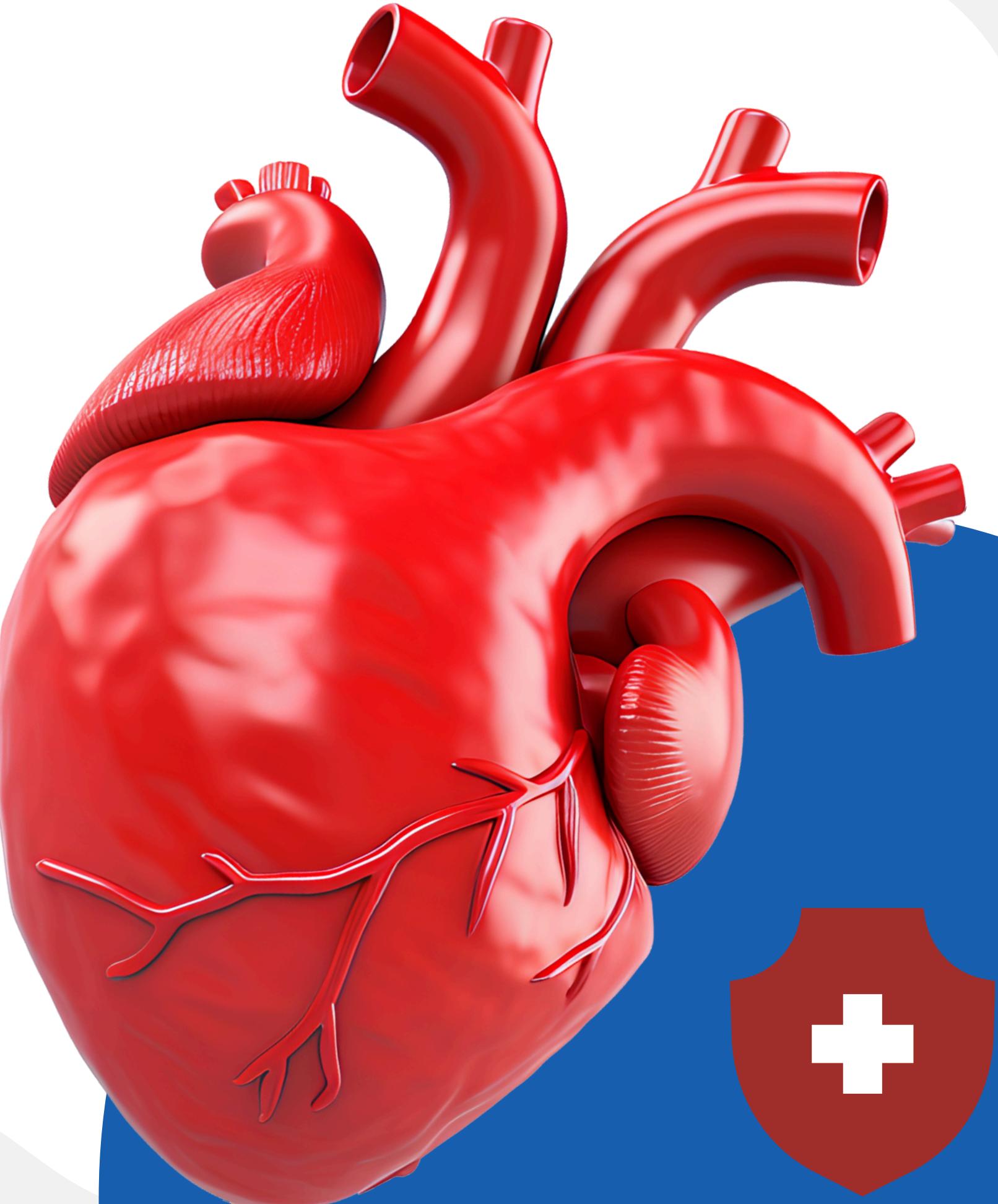


SUMBER DATA



Data yang digunakan pada penelitian ini merupakan data sekunder yang bersumber dari situs Kaggle. Data yang digunakan adalah data analisis dan prediksi serangan jantung.

Source :
<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

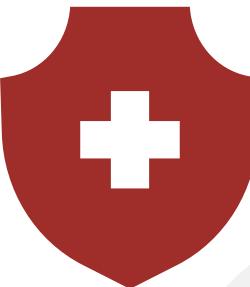


VARIABEL PENELITIAN

Variabel penelitian yang digunakan pada penelitian ini adalah sebagai berikut :

Variabel	Definisi	Jenis Data	Keterangan
age	Usia pasien	Numerik	-
sex	Jenis kelamin pasien	Kategorik	[1] laki-laki [0] perempuan
cp	Tipe nyeri dada	Kategorik	[0] angina tipikal [1] angina atipikal [2] nyeri non-angina [3] tanpa gejala
trtbps	Tekanan darah istirahat (dalam mm Hg)	Numerik	-
chol	Kolesterol dalam mg/dl	Numerik	-
fbs	Gula darah puasa >120 mg/dl	Kategorik	[1] ya [0] tidak
restecg	Hasil elektrokardiografi istirahat	Kategorik	[0] normal [1] mengalami kelainan gelombang ST-T [2] menunjukkan kemungkinan atau pasti hipertrofi ventrikel kiri menurut kriteria Estes

Variabel	Definisi	Jenis Data	Keterangan
thalachh	Denyut jantung maksimal tercapai	Numerik	-
oldpeak	Depresi ST yang disebabkan oleh olahraga relatif terhadap istirahat	Numerik	-
slp	Kemiringan puncak latihan segmen ST	Kategorik	[0] menanjak [1] datar [2] menurun
caa	Jumlah pembuluh darah	Kategorik	0-4
thall	Kelainan darah yang disebut thalassemia	Kategorik	[0] null [1] cacat tetap [2] aliran darah normal [3] cacat reversibel
exng	Angina akibat olahraga	Kategorik	[1] ya [0] tidak
output	Penyakit jantung	Kategorik	[0] less chance of heart attack [1] more chance of heart attack



STRUKTUR DATA

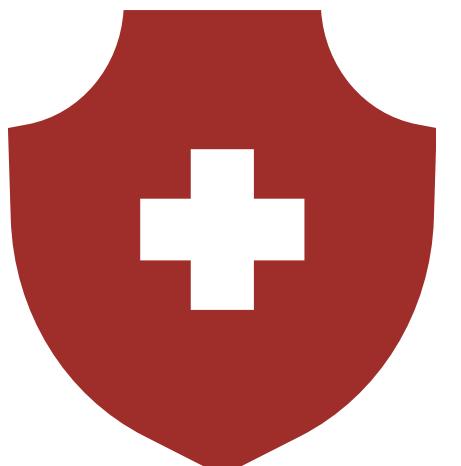
Struktur data yang digunakan pada penelitian ini adalah sebagai berikut:

Observasi	<i>Variabel</i>				
	<i>Y</i>	<i>X₁</i>	<i>X₂</i>	...	<i>X_m</i>
1	<i>Y₁</i>	<i>X_{1.1}</i>	<i>X_{1.2}</i>	...	<i>X_{1.m}</i>
2	<i>Y₂</i>	<i>X_{2.1}</i>	<i>X_{2.2}</i>	...	<i>X_{2.m}</i>
3	<i>Y₃</i>	<i>X_{3.1}</i>	<i>X_{3.2}</i>	...	<i>X_{3.m}</i>
...
<i>n</i>	<i>Y_n</i>	<i>X_{n.1}</i>	<i>X_{n.2}</i>	...	<i>X_{n.m}</i>

Keterangan :

1,2,...,13 : m

1,2,...,303 : n



LANGKAH ANALISIS

Langkah analisis yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Mengidentifikasi masalah dan tujuan penelitian
2. Menginput data yang akan digunakan
3. Melakukan identifikasi tipe data
4. Melakukan pre-processing berupa data cleaning
5. Menganalisis statistika deskriptif
6. Melakukan eksplorasi dan visualisasi data
7. Melakukan analisis feature selection dengan metode Chi-Square dan LDA
8. Membagi data dengan metode repeated holdout dan k-fold
9. Melakukan analisis klasifikasi dengan metode Decision Tree dan Naïve Bayes
10. Melakukan model evaluation and selection
11. Membuat kesimpulan



PRE-PROCESSING (1)

Sebelum dilakukan cleaning dilakukan penginputan data terlebih dahulu.

Data

```
[ ] # Import Data  
df = pd.read_csv('heart.csv')
```

Shape Data

```
▶ # Df Shape  
df.shape
```

→ (303, 14)

Preview Data

```
[ ] df.head()
```

→

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

PRE-PROCESSING (2)

Melakukan cek duplicates untuk mengetahui data yang terdapat duplikasi, karena terdapat data duplicate maka data dilakukan drop duplicate.

Duplicates

```
[ ] # Check Duplicates  
print(f'Count of Duplicates: {df.duplicated().sum()}')
```

→ Count of Duplicates: 1

```
[ ] # Drop Duplicates  
df.drop_duplicates(inplace = True)
```

PRE-PROCESSING (3)

Karena terdapat ketidaksesuaian tipe data dengan description data, maka dilakukan change data type.

Change Datatype

```
[ ] df['age']=df['age'].astype(float)
df['trtbps']=df['trtbps'].astype(float)
df['chol']=df['chol'].astype(float)
df['thalachh']=df['thalachh'].astype(float)
df['oldpeak']=df['oldpeak'].astype(float)
df['sex']=df['sex'].astype(int)
df['exng']=df['exng'].astype(int)
df['caa']=df['caa'].astype(int)
df['cp']=df['cp'].astype(int)
df['fbs']=df['fbs'].astype(int)
df['restecg']=df['restecg'].astype(int)
df['slp']=df['slp'].astype(int)
df['thall']=df['thall'].astype(int)
df['output']=df['output'].astype(int)
```

PRE-PROCESSING (4)

Melakukan pengecekan apakah terdapat missing value.

Missing value

```
[ ] # Check Missing Value  
df.isnull().sum()
```

```
→ age      0  
    sex     0  
    cp      0  
    trtbps  0  
    chol    0  
    fbs     0  
    restecg 0  
    thalachh 0  
    exng   0  
    oldpeak 0  
    slp     0  
    caa     0  
    thall   0  
    output  0  
    dtype: int64
```

Pada data tersebut tidak terdapat missing value, maka lanjut pada tahap berikutnya.

SUMMARY DATA

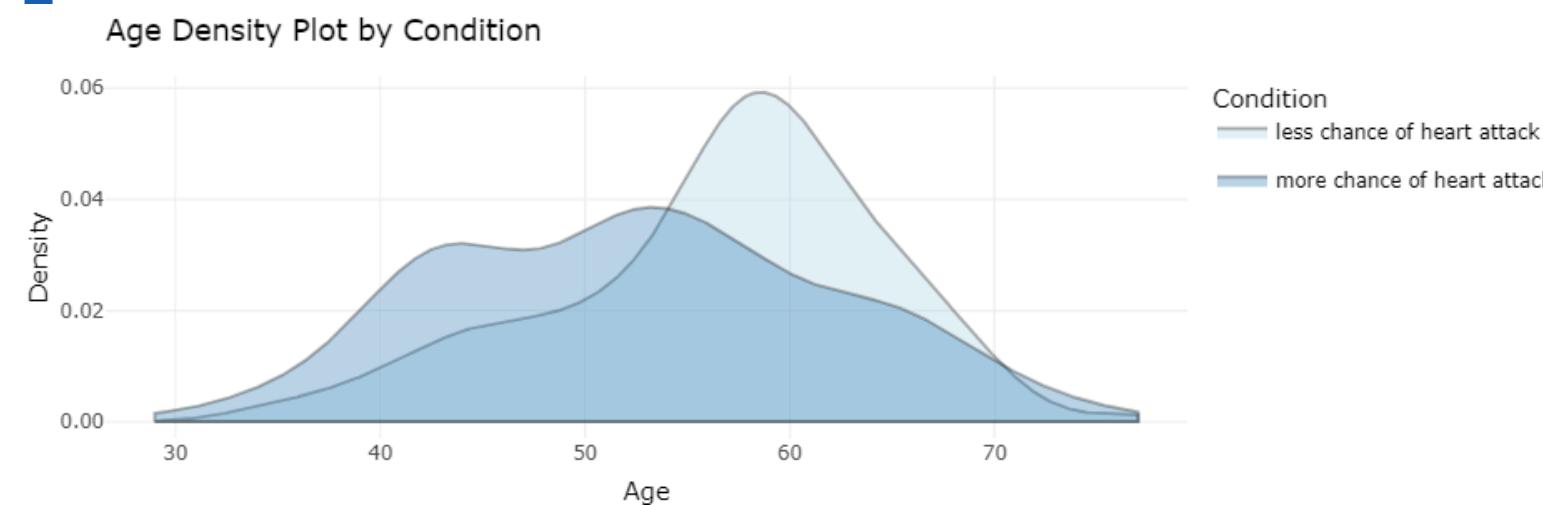
Berikut adalah hasil analisis statistika deskriptif:

Dataset ini terdiri dari 302 observasi dengan lima variabel numerik: Usia, Tekanan Darah Saat Istirahat, Kolesterol, Detak Jantung Maksimal, dan Previous Peak. Rata-rata usia adalah 54.4 tahun dengan standar deviasi 9 tahun. Rata-rata tekanan darah saat istirahat adalah 131.6 mmHg (SD = 17.6), sementara rata-rata kadar kolesterol adalah 246.5 mg/dL (SD = 51.8). Rata-rata detak jantung maksimal adalah 149.6 bpm (SD = 22.9), dan nilai previous peak rata-rata adalah 1 (SD = 1.2). Distribusi variabel menunjukkan variasi dengan beberapa skewness, namun tidak ada data yang hilang pada setiap variabel.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Age [numeric]	Mean (sd) : 54.4 (9) min ≤ med ≤ max: $29 \leq 55.5 \leq 77$ IQR (CV) : 13 (0.2)	41 distinct values		302 (100.0%)	0 (0.0%)
2	Resting Blood Pressure [numeric]	Mean (sd) : 131.6 (17.6) min ≤ med ≤ max: $94 \leq 130 \leq 200$ IQR (CV) : 20 (0.1)	49 distinct values		302 (100.0%)	0 (0.0%)
3	Cholesterol [numeric]	Mean (sd) : 246.5 (51.8) min ≤ med ≤ max: $126 \leq 240.5 \leq 564$ IQR (CV) : 63.8 (0.2)	152 distinct values		302 (100.0%)	0 (0.0%)
4	Max. Heart Rate [numeric]	Mean (sd) : 149.6 (22.9) min ≤ med ≤ max: $71 \leq 152.5 \leq 202$ IQR (CV) : 32.8 (0.2)	91 distinct values		302 (100.0%)	0 (0.0%)
5	Previous Peak [numeric]	Mean (sd) : 1 (1.2) min ≤ med ≤ max: $0 \leq 0.8 \leq 6.2$ IQR (CV) : 1.6 (1.1)	40 distinct values		302 (100.0%)	0 (0.0%)

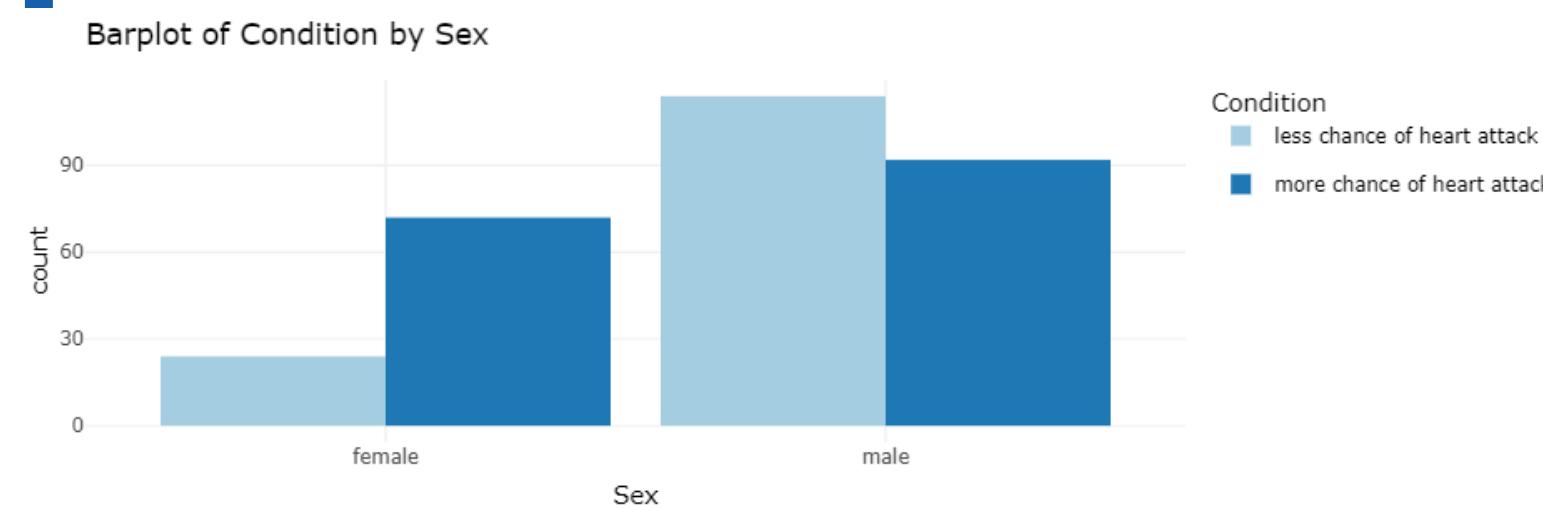
VISUALISASI DATA (1)

AGE DENSITY PLOT BY CONDITION



Plot densitas menunjukkan puncak pada sekitar umur 60 tahun. Ini menunjukkan bahwa frekuensi individu dengan risiko lebih rendah terkena serangan jantung cenderung berumur sekitar 60 tahun. Selain itu, risiko lebih tinggi tersebar di rentang umur yang lebih luas, mulai dari sekitar 40 tahun hingga 70 tahun.

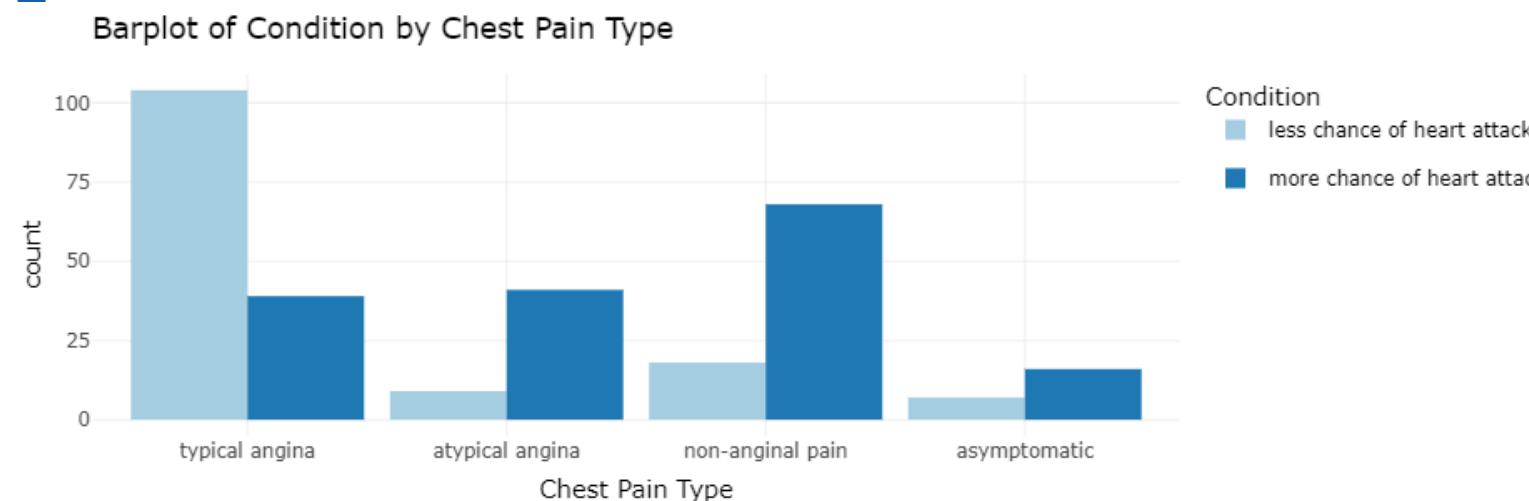
BARPLOT OF CONDITION BY SEX



Lebih banyak wanita yang berada pada kategori "more chance of heart attack" dibandingkan dengan kategori "less chance of heart attack". Kemudian, ada ketidakseimbangan yang signifikan dalam jumlah pria dan wanita yang memiliki risiko lebih tinggi terkena serangan jantung, dengan pria menunjukkan jumlah yang lebih tinggi.

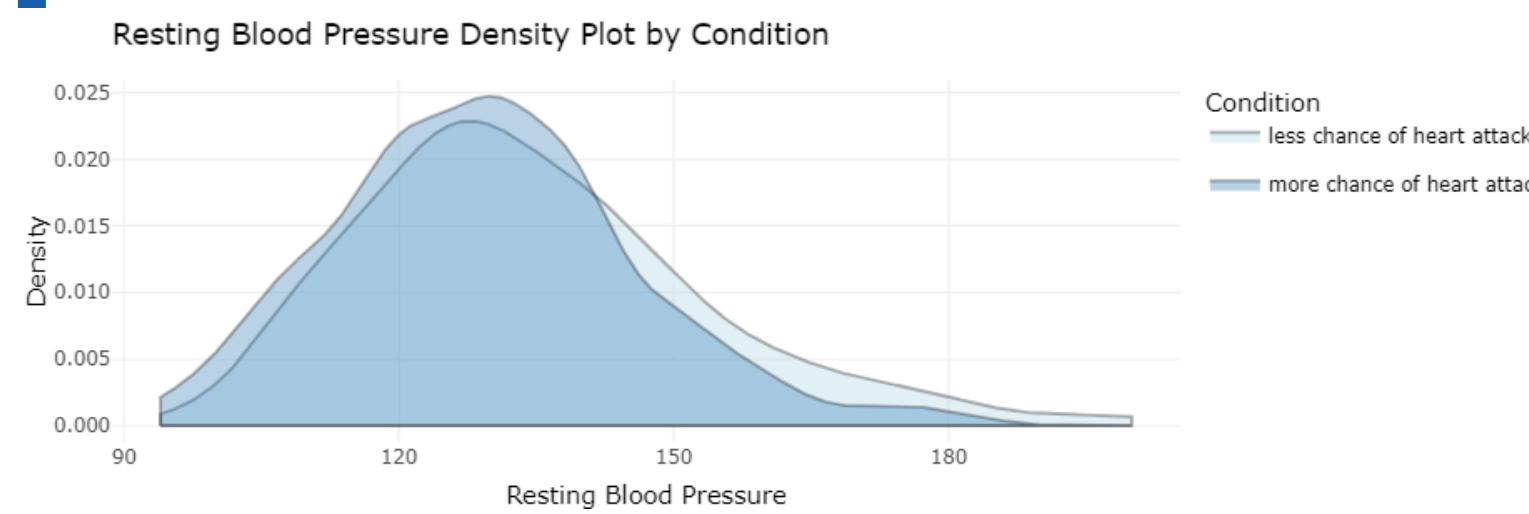
VISUALISASI DATA (2)

BARPLOT OF CONDITION BY CHEST PAIN



Frekuensi tipe nyeri dada "Typical angina" lebih banyak ditemukan pada individu dengan risiko lebih rendah, sementara tipe nyeri dada "atypical angina", "non-angina pain" dan "asymptomatic" lebih banyak ditemukan pada individu dengan risiko lebih tinggi terkena serangan jantung.

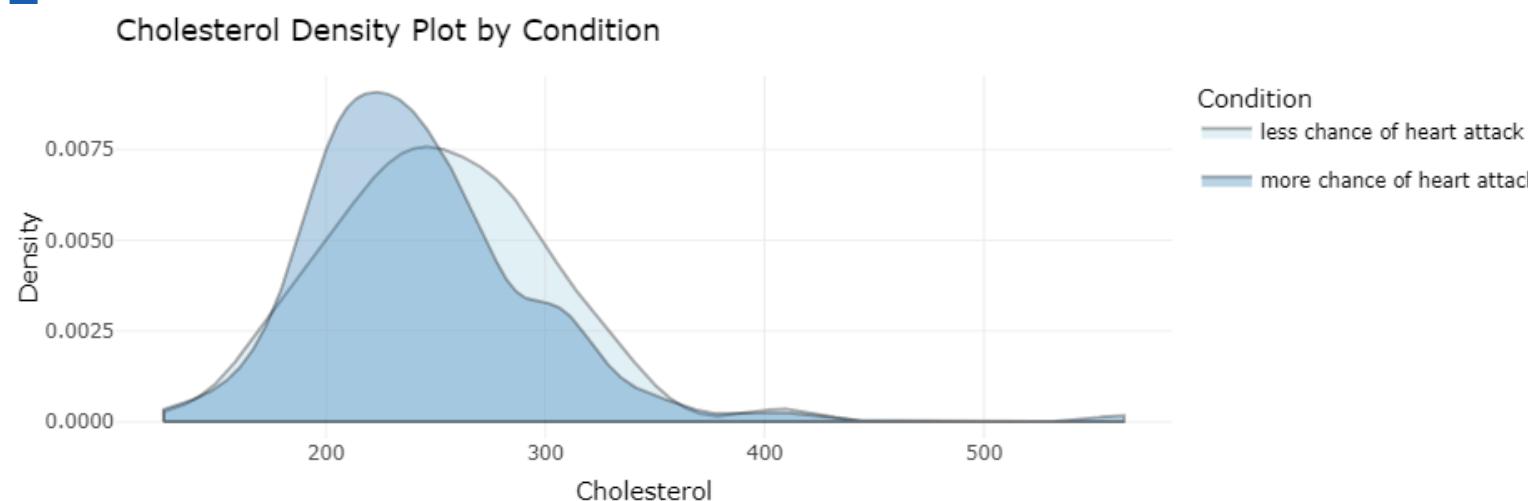
RESTING BLOOD PRESSURE DENSITY PLOT BY CONDITION



Resiko rendah memiliki puncak sekitar 125 untuk tekanan darah istirahat. Sementara itu, resiko tinggi memiliki puncak sedikit lebih tinggi, sekitar 130. Visualisasi ini menunjukkan korelasi antara tingkat tekanan darah istirahat dan risiko mengalami serangan jantung, dengan tekanan darah istirahat yang lebih tinggi mengindikasikan risiko yang lebih besar.

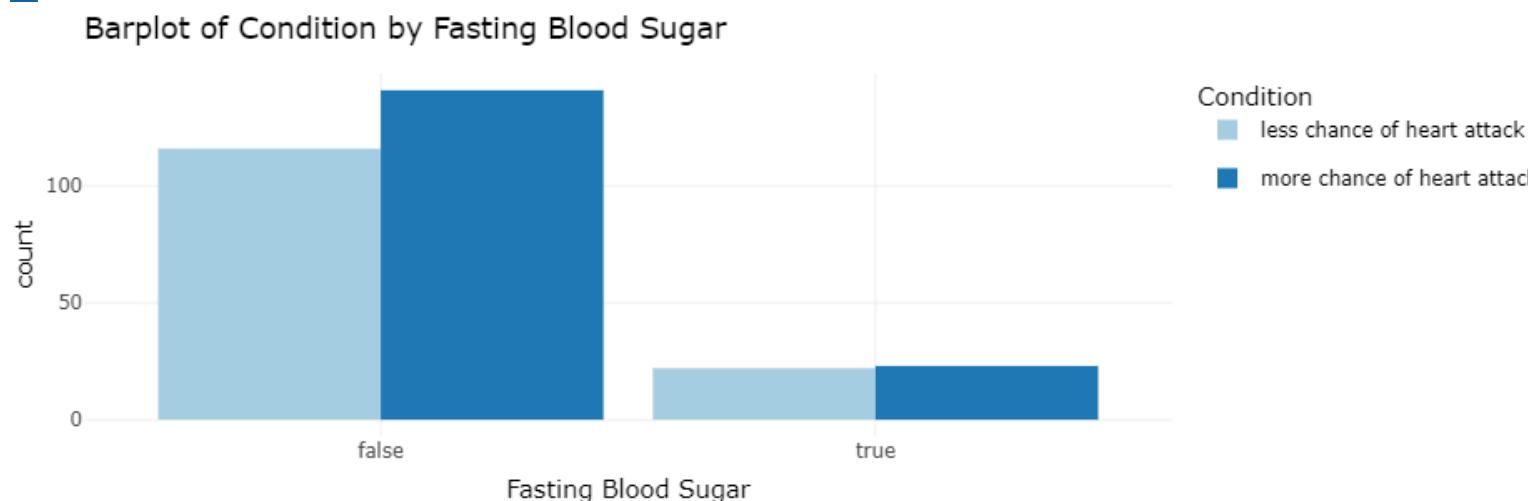
VISUALISASI DATA (3)

CHOLESTROL DENSITY PLOT BY CONDITION



Resiko tinggi memiliki puncak sekitar 220 untuk tingkat kolesterol. Sementara itu, resiko rendah memiliki puncak sedikit setelah 250 dan memiliki penyebaran yang lebih lebar. Visualisasi ini korelasi antara tingkat kolesterol dan resiko serangan jantung, dimana pasien dengan kolesterol di tingkat 220 lebih beresiko terkena serangan jantung.

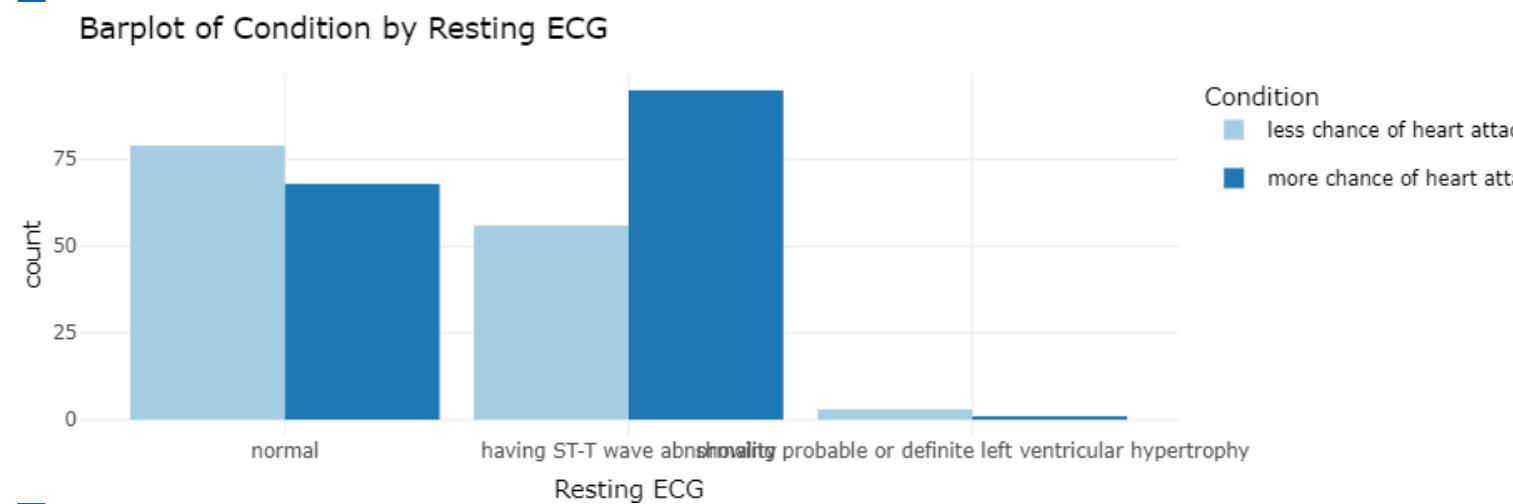
BARPLOT OF CONDITION BY FASTING BLOOD SUGAR



Dalam kategori "false" menunjukkan lebih banyak kasus dengan peluang serangan jantung lebih tinggi ketika pasien tidak melakukan puasa kadar gula darah atau pasien memiliki kondisi tidak melakukan puasa gula darah cenderung lebih besar terkena serangan jantung.

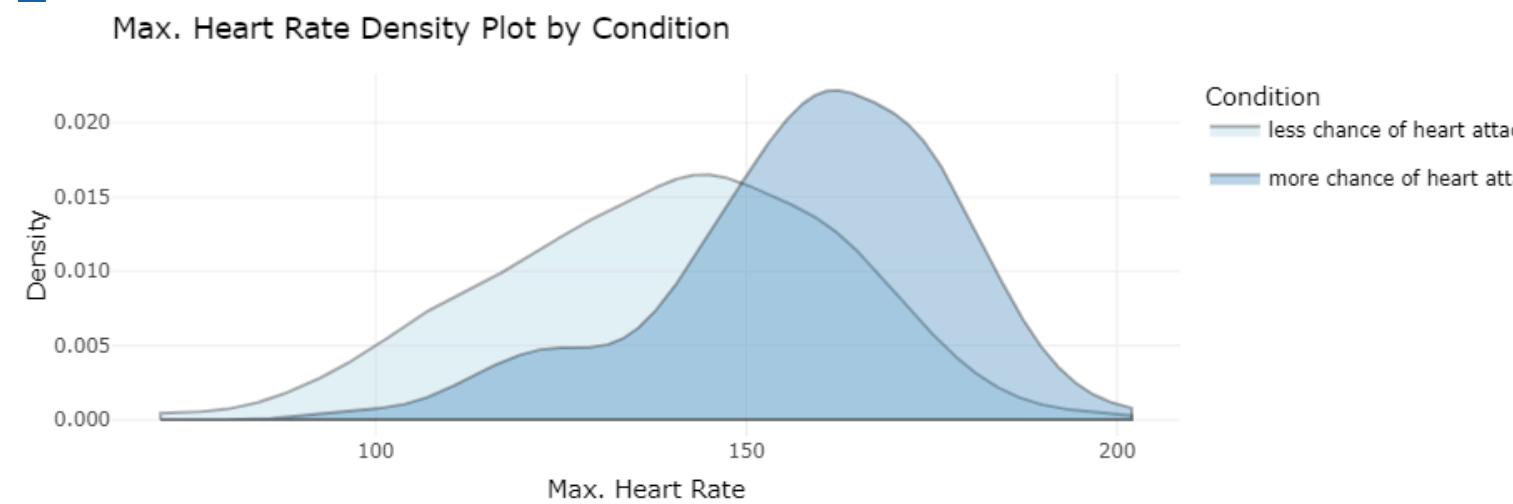
VISUALISASI DATA (4)

BARPLOT OF CONDITION BY RESTING ECG



Kategori ‘normal’ memiliki jumlah yang lebih tinggi untuk kedua kondisi dibandingkan dengan kategori lainnya. Kemudian, pasien memiliki kondisi yang sama antara normal dan memiliki kelainan pada gelombang, dari kedua kondisi pasien yang seperti ini kategori yang memiliki kelainan cenderung terkena serangan jantung.

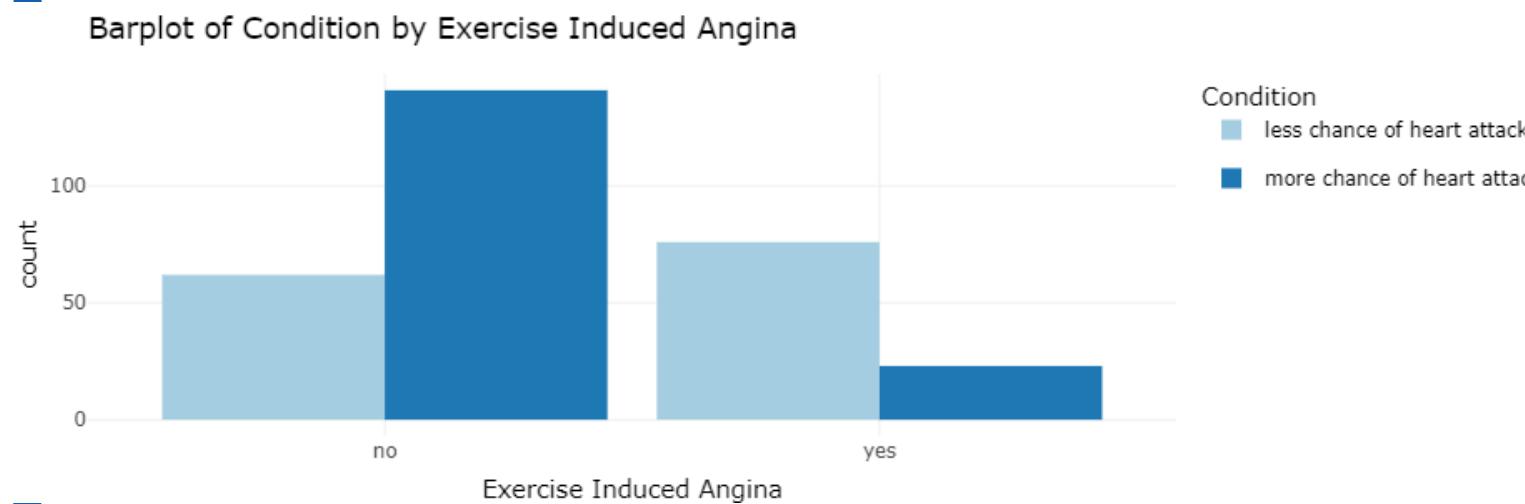
MAX. HEART RATE DENSITY PLOT BY CONDITION



Pada max. heart rate density plot, menunjukkan pasien yang memiliki denyut jantung sekitar 160-165 memiliki kondisi resiko tinggi terkena serangan jantung.

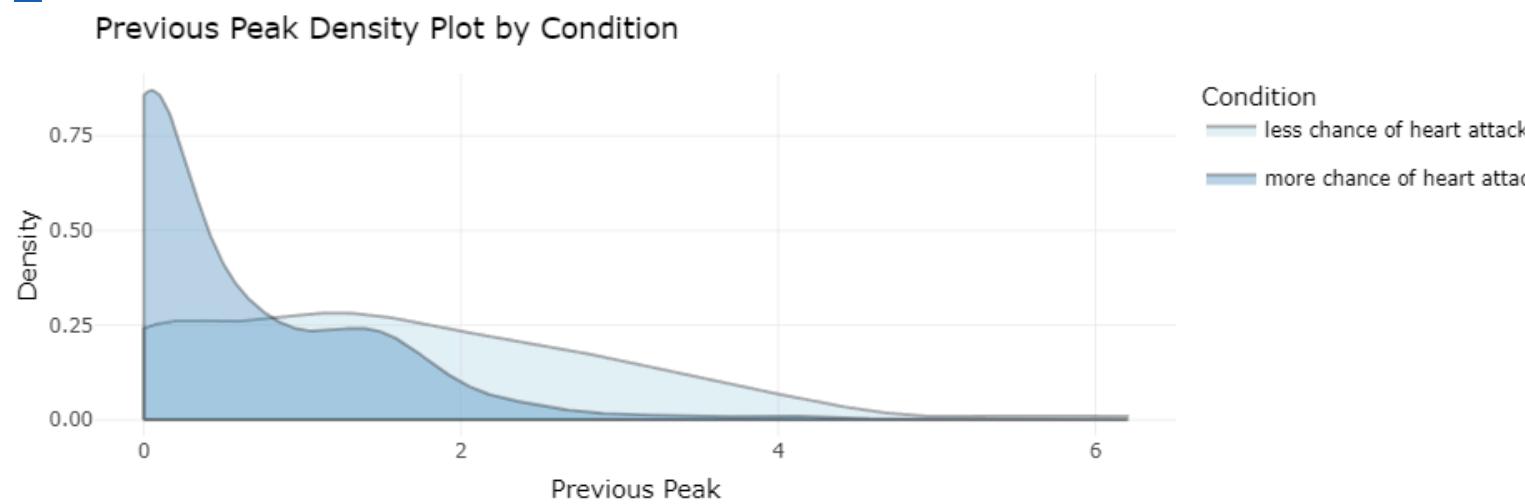
VISUALISASI DATA (5)

BARPLOT OF CONDITION BY EXERCISE INCLUDED ANGINA



Pada exercise angina, menunjukkan frekuensi pasien yang tidak mengalami exercise angina lebih banyak berpotensi mengalami serangan jantung, sedangkan pasien yang mengalami exercise angina lebih kecil beresiko mengalami serangan jantung.

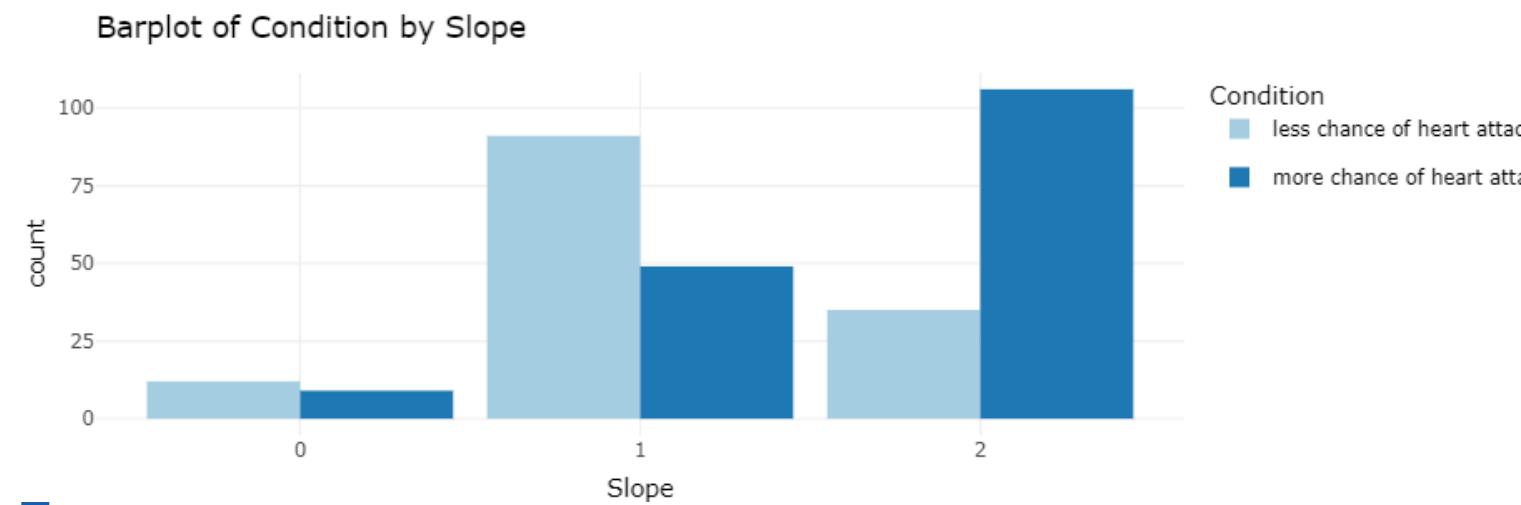
PREVIOUS PEAK DENSITY PLOT BY CONDITION



Pada previous peak, menunjukkan pasien yang memiliki previous peak rendah memiliki resiko lebih besar terkena serangan jantung, sedangkan pada previous peak menengah hingga tinggi lebih beresiko rendah mengalami serangan jantung.

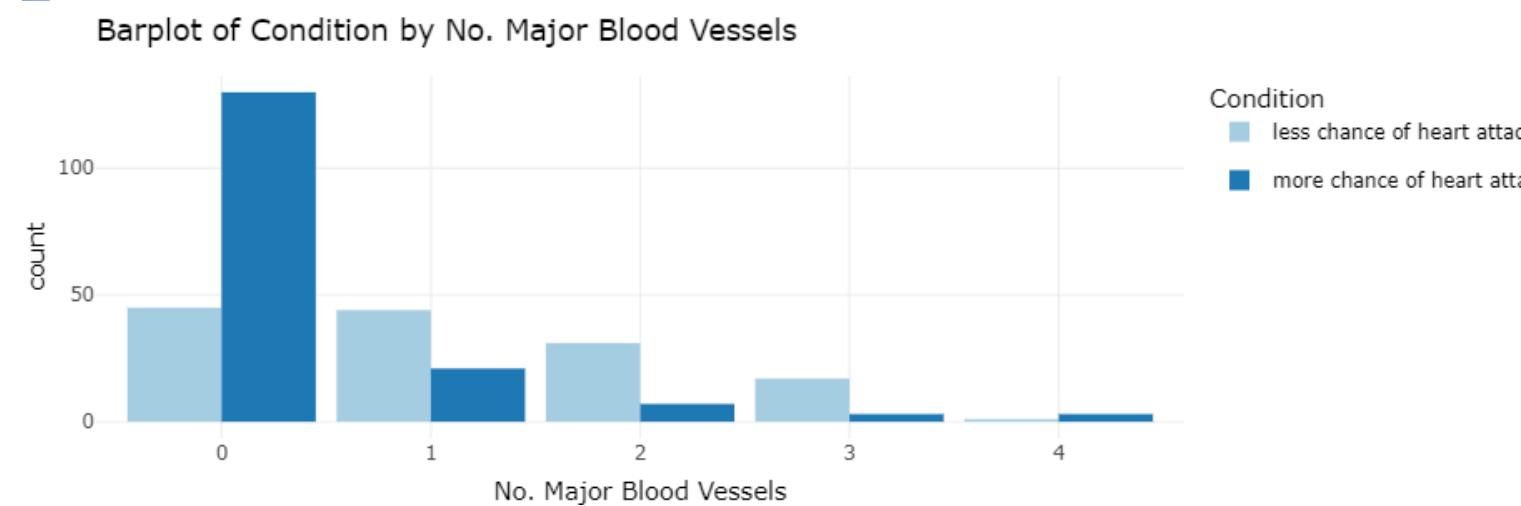
VISUALISASI DATA (6)

BARPLOT OF CONDITION BY SLOPE



Pada slope, menunjukkan pasien yang memiliki slope 2 (menurun) lebih beresiko terkena serangan jantung daripada slope 0 (menanjak) dan 1 (datar).

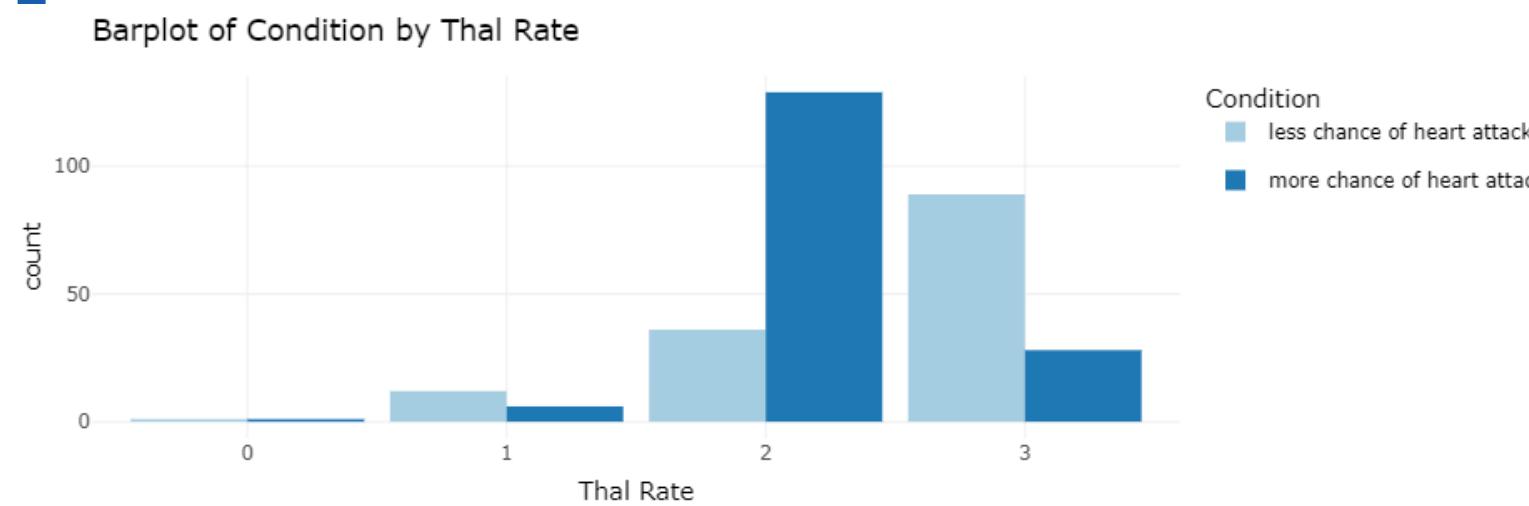
BARPLOT OF CONDITION BY NO. MAJOR BLOOD VESSELS



Pada Number Major Blood Vessels, menunjukkan bahwa pasien yang memiliki jumlah Major Blood Vessels rendah lebih beresiko tinggi terkena serangan jantung, dibandingkan pasien yang memiliki jumlah Major Blood Vessels tinggi.

VISUALISASI DATA (7)

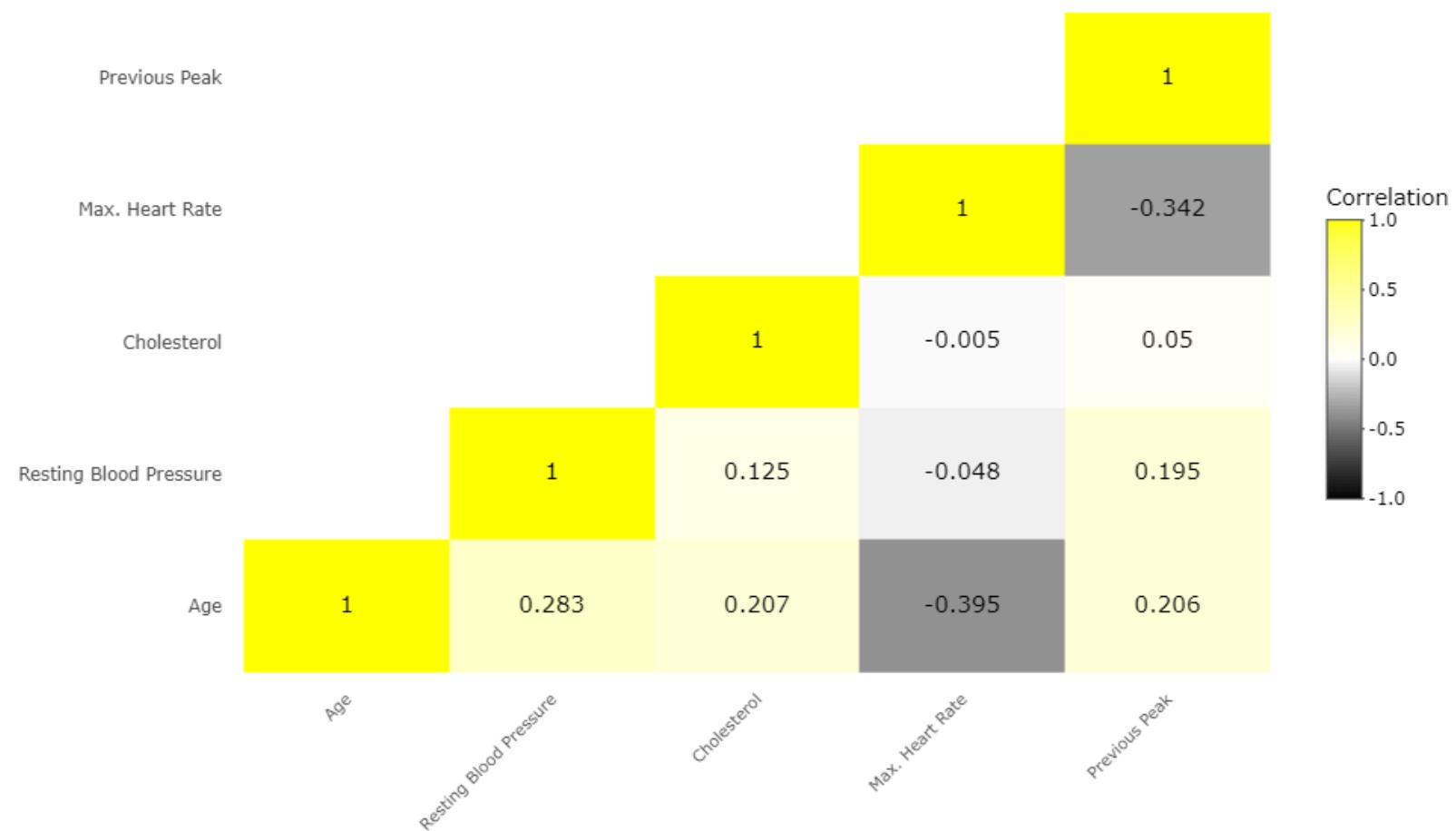
BARPLOT OF CONDITION BY THAL RATE



Pada Thall Rate, menunjukkan frekuensi pasien yang memiliki aliran darah normal cukup banyak yang memiliki resiko tinggi terkena serangan jantung. Sedangkan pasien yang mengalami cacat reversibel lebih beresiko rendah terkena serangan jantung.

VISUALISASI DATA (8)

CORRELATION PLOT



Berdasarkan correlation plot dapat diketahui bahwa korelasi positif tertinggi terjadi di antara variabel “Resting Blood Pressure” dan “Age” yaitu sebesar 0,283. Sedangkan korelasi negatif tertinggi terjadi di antara variabel “Max. Heart Rate” dan “Age” yaitu sebesar -0,395

FEATURE SELECTION



Feature selection digunakan untuk mengetahui variabel manakah yang paling berpengaruh terhadap variabel respon yaitu variabel output. Variabel-variabel yang paling berpengaruh akan digunakan dalam pemodelan untuk mendapatkan nilai klasifikasi yang tinggi. Pada langkah ini kami bagi data menjadi data kategorik dan data numerik.



VARIABEL KATEGORIK

Variabel-variabel prediktor kategorik difilter dengan metode Chi-Square, sehingga terpilih lima variabel prediktor kategorik yang paling berpengaruh terhadap variabel heart attack

	Selected_columns	Score_chi2
2	caa	71.020719
3	cp	62.116086
1	exng	38.518849
6	s1p	9.677715
0	sex	7.721690



VARIABEL NUMERIK

Melakukan penyederhanaan fitur pada variabel numerik dengan metode Linear Discriminant Analysis (LDA) sehingga didapatkan tiga variabel prediktor numerik yang paling berpengaruh terhadap variabel heart attack

	Selected_columns	Score_ANOVA
4	chol	67.721931
3	trtbps	64.237793
0	age	15.474511

SPLIT DATA TRAINING & TESTING



CROSS VALIDATION K-FOLD

Pada metode ini data pengamatan akan dibagi menjadi 5 bagian yang sama besar (menyesuaikan jika jumlah data pengamatan bukan kelipatan 5), dimana 4 bagian menjadi data training dan 1 sisanya menjadi data testing, sehingga terdapat 5 pasangan data training dan data testing.



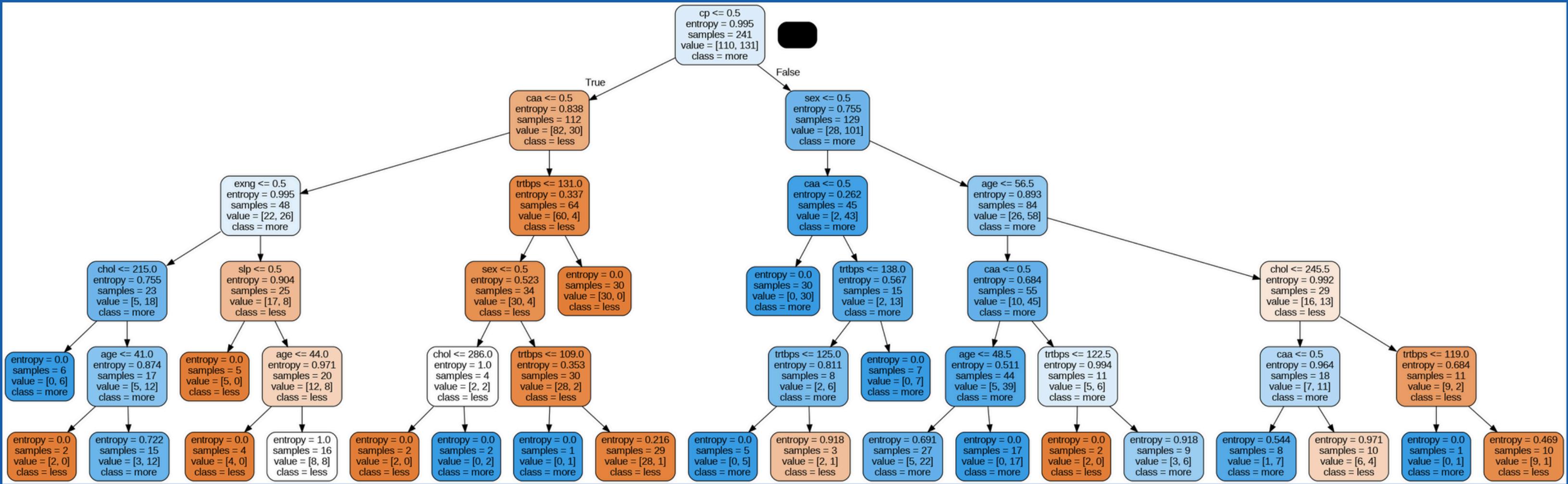
REPEATED HOLDOUT

Pada metode ini data pengamatan akan dibagi menjadi 2 yaitu data training sebesar 70% dan data testing 30%, dan diulang sebanyak 5 kali, sehingga terdapat 5 pasangan data training dan data testing.

DECISION TREE (K-FOLD)



FOLD 1



Berdasarkan output decision tree fold 1 diperoleh variabel "cp" sebagai root node.

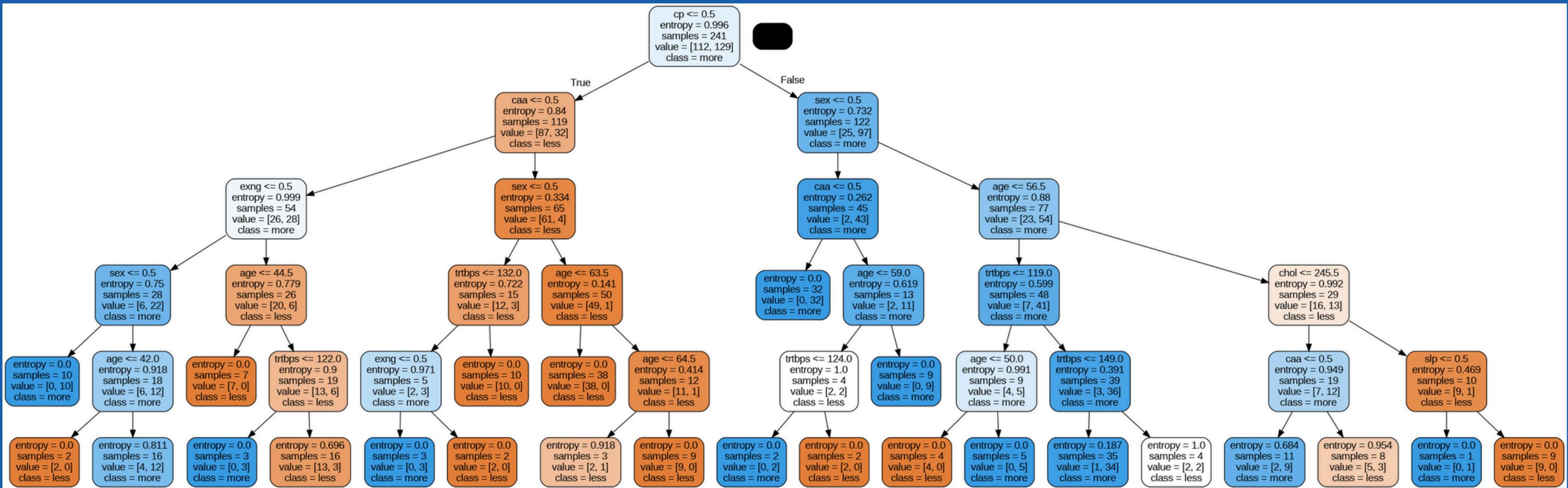
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "trtbps" dan "exng" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "sex" menjadi decision node, diikuti variabel "caa" dan "age" sebagai sub decision node.

DECISION TREE (K-FOLD)



FOLD 2



Berdasarkan output decision tree fold 2 diperoleh variabel "cp" sebagai root node.

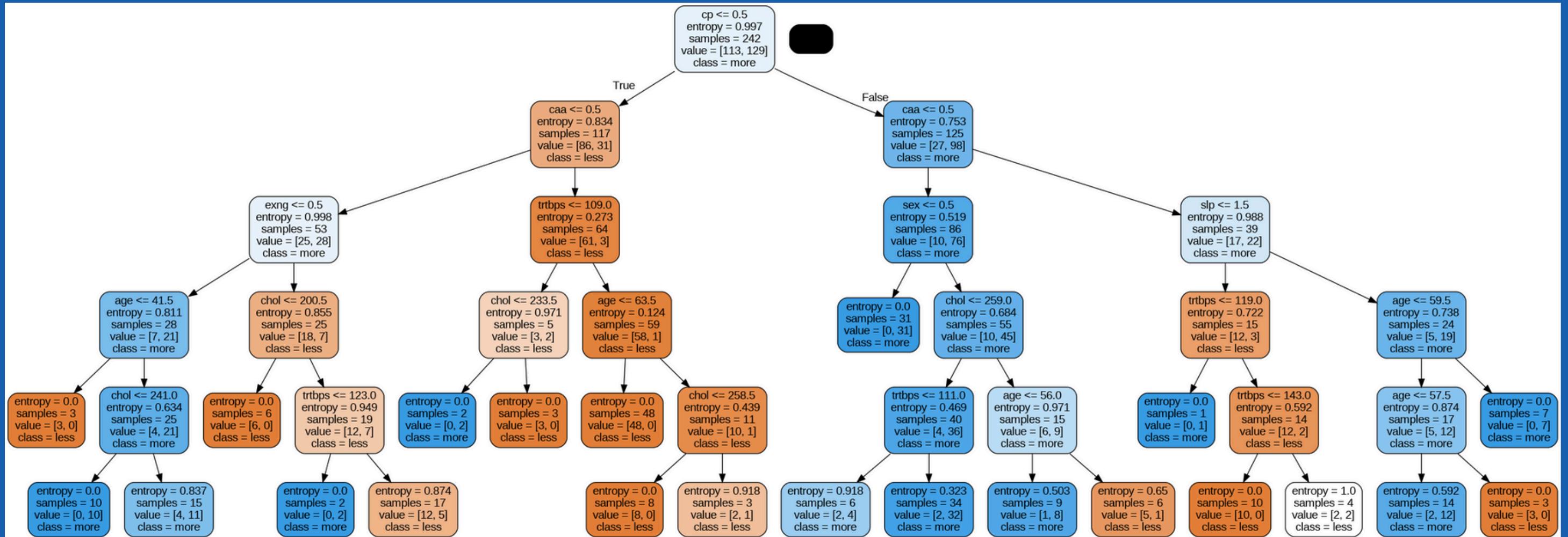
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "exng" dan "sex" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "sex" menjadi decision node, diikuti variabel "caa" dan "age" sebagai sub decision node.

DECISION TREE (K-FOLD)



FOLD 3



Berdasarkan output decision tree fold 3 diperoleh variabel "cp" sebagai root node.

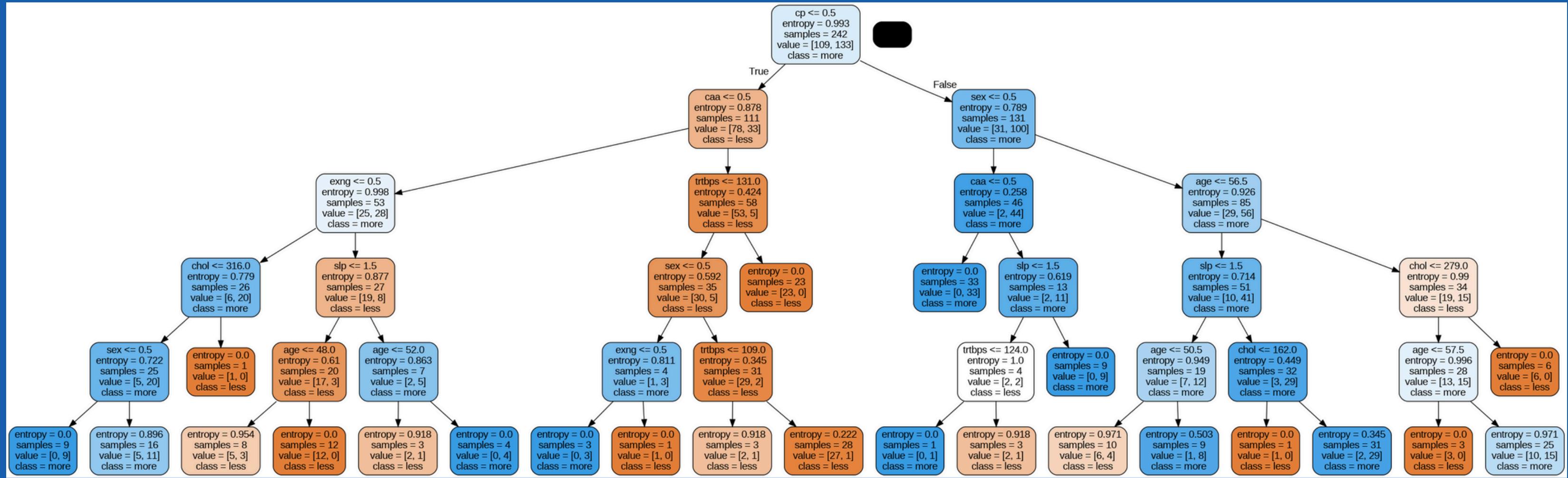
Jika variabel "cp" bernilai true maka variabel "caa-less" menjadi decision node, diikuti variabel "trtbps" dan "exng" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "caa-more" menjadi decision node, diikuti variabel "sex" dan "slp" sebagai sub decision node.

DECISION TREE (K-FOLD)



FOLD 4



Berdasarkan output decision tree fold 4 diperoleh variabel "cp" sebagai root node.

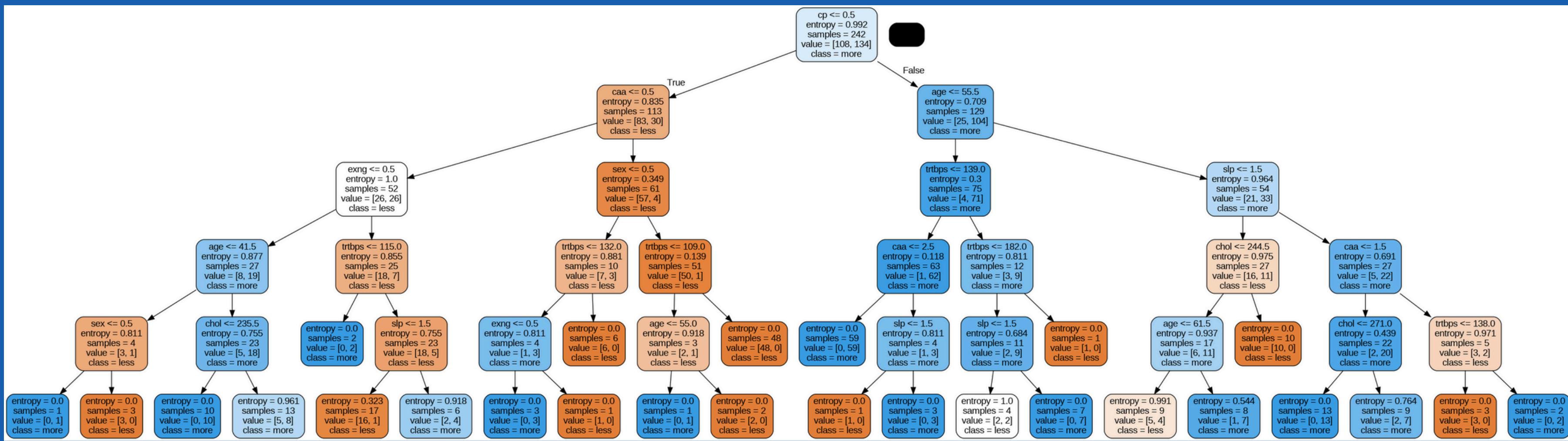
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "trtbps" dan "exng" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "sex" menjadi decision node, diikuti variabel "caa" dan "age" sebagai sub decision node.

DECISION TREE (K-FOLD)



FOLD 5



Berdasarkan output decision tree fold 5 diperoleh variabel "cp" sebagai root node.

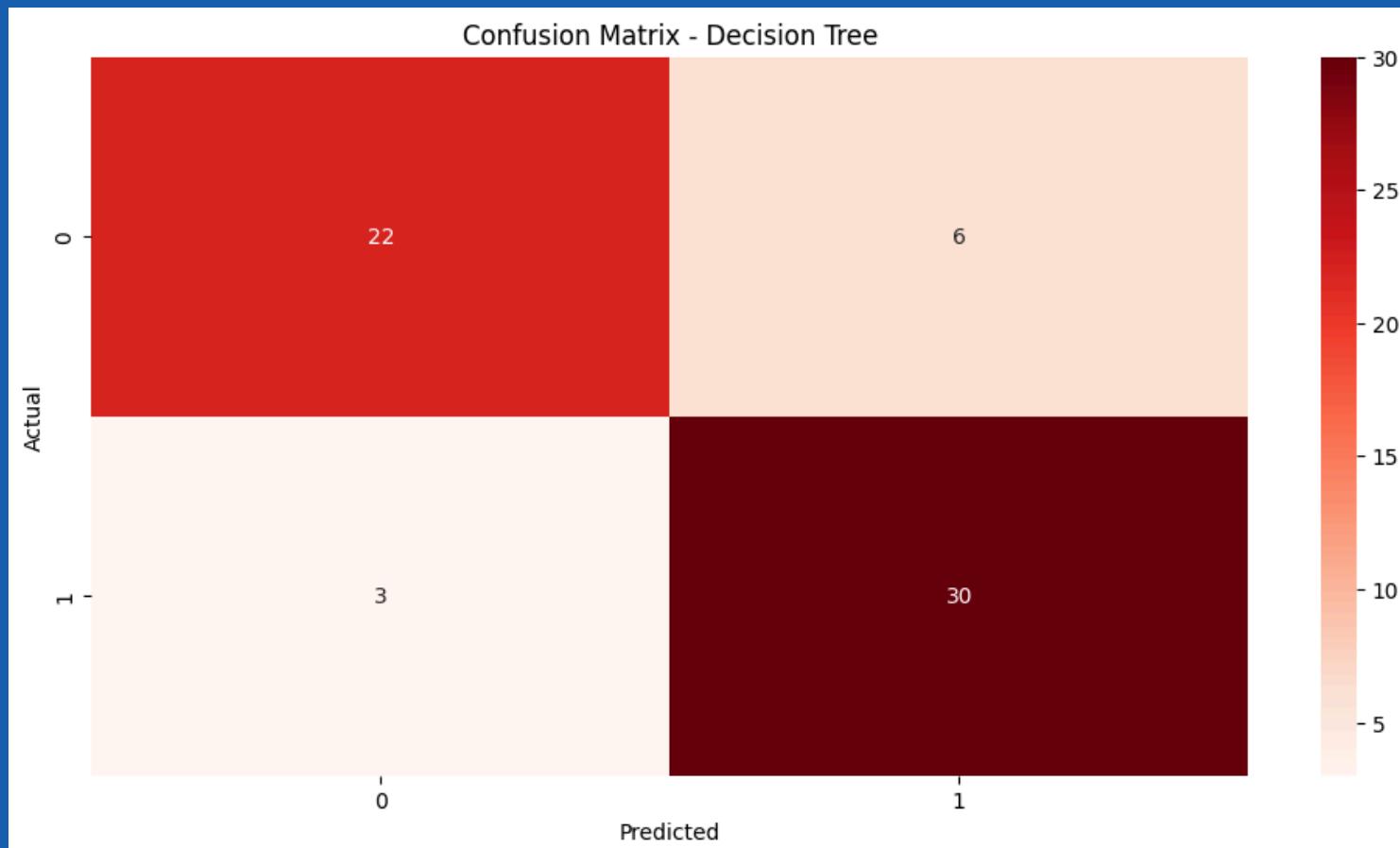
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "sex" dan "exng" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "age" menjadi decision node, diikuti variabel "trtbps" dan "slp" sebagai sub decision node.

DECISION TREE (K-FOLD)



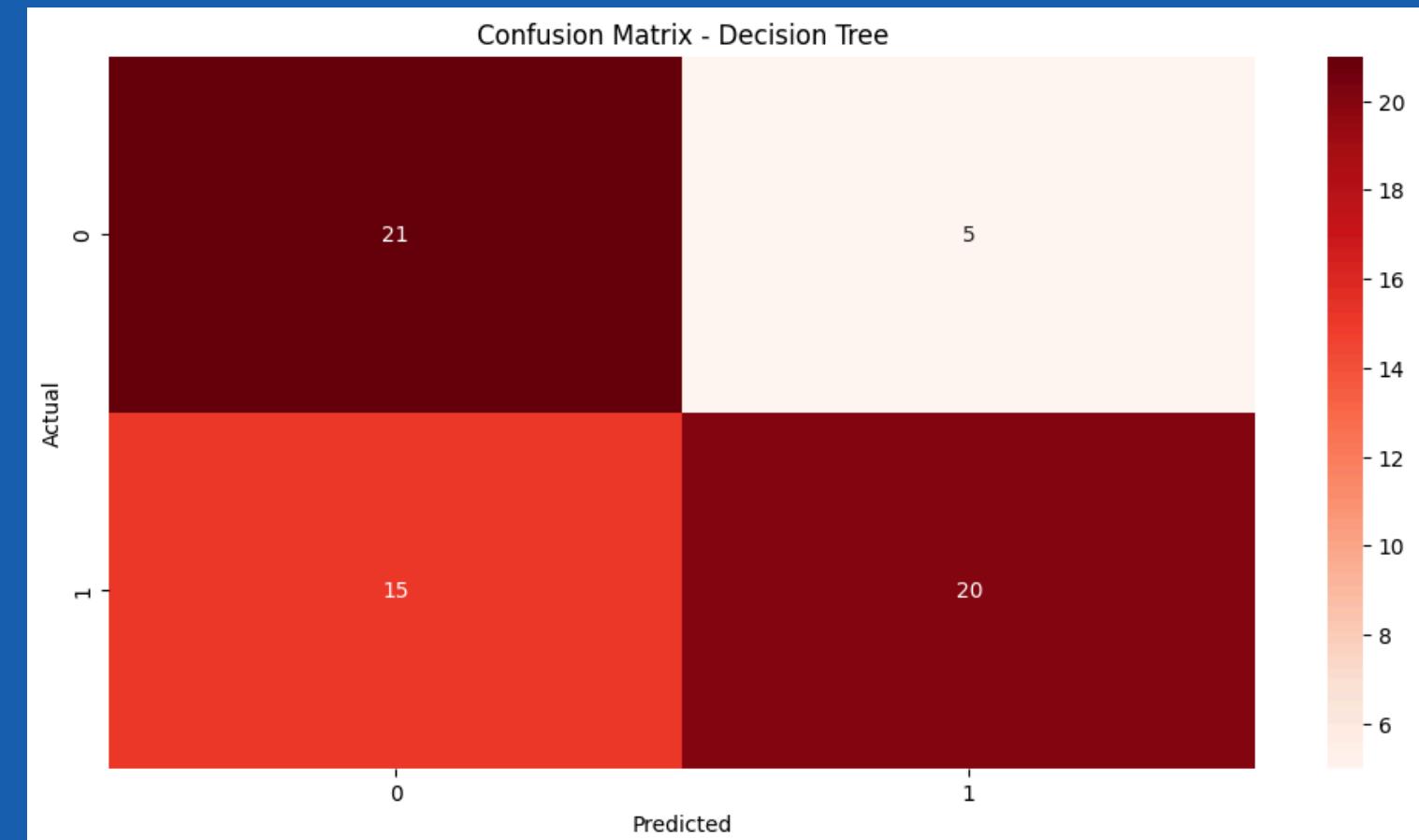
FOLD 1



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 30. Untuk kasus False Positif adalah sebanyak 6, kasus True Negatif sebanyak 22 dan yang terjadi pada kasus False Negatif sebanyak 3.



FOLD 2

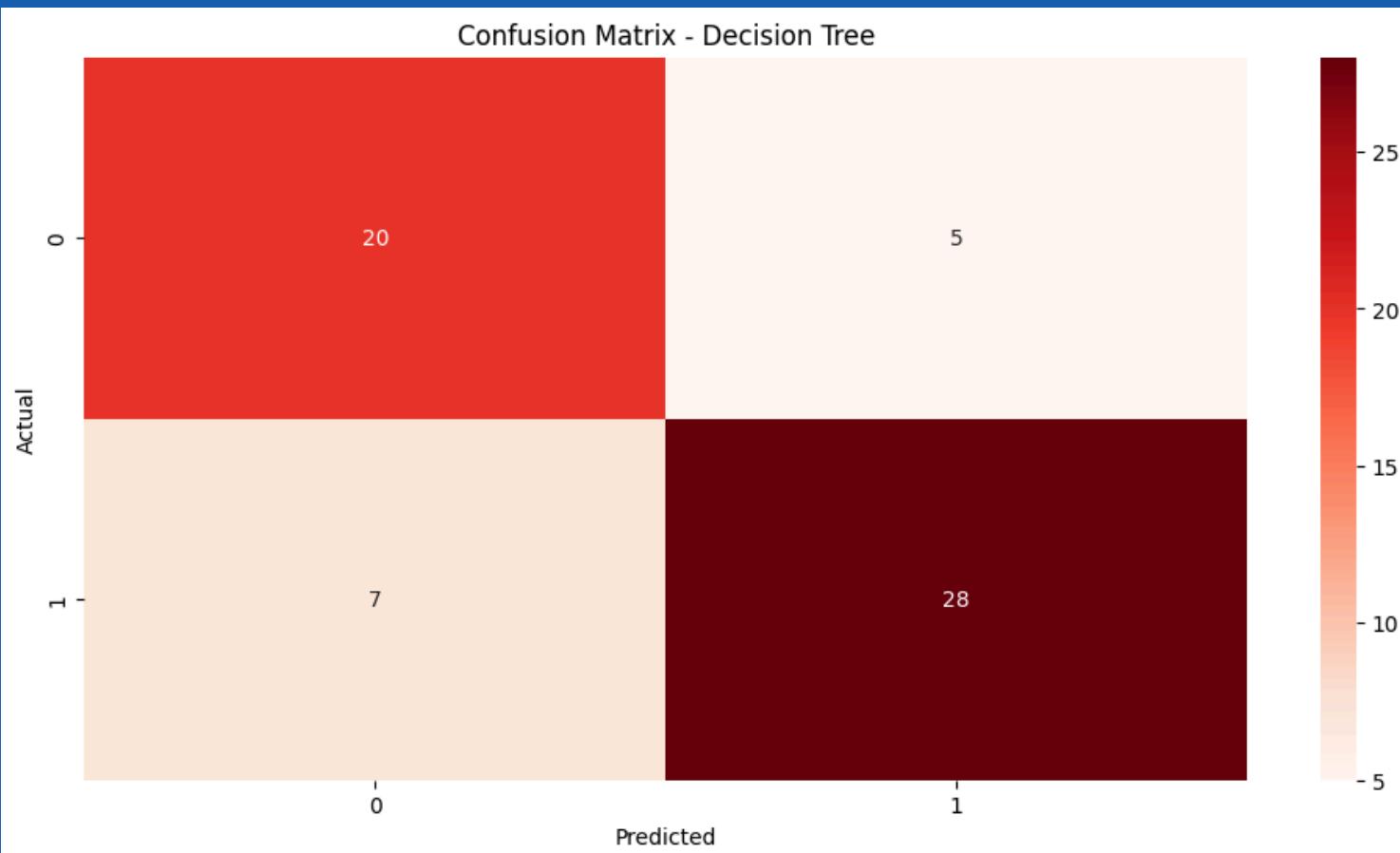


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 20. Untuk kasus False Positif adalah sebanyak 5, kasus True Negatif sebanyak 21 dan yang terjadi pada kasus False Negatif sebanyak 15.

DECISION TREE (K-FOLD)



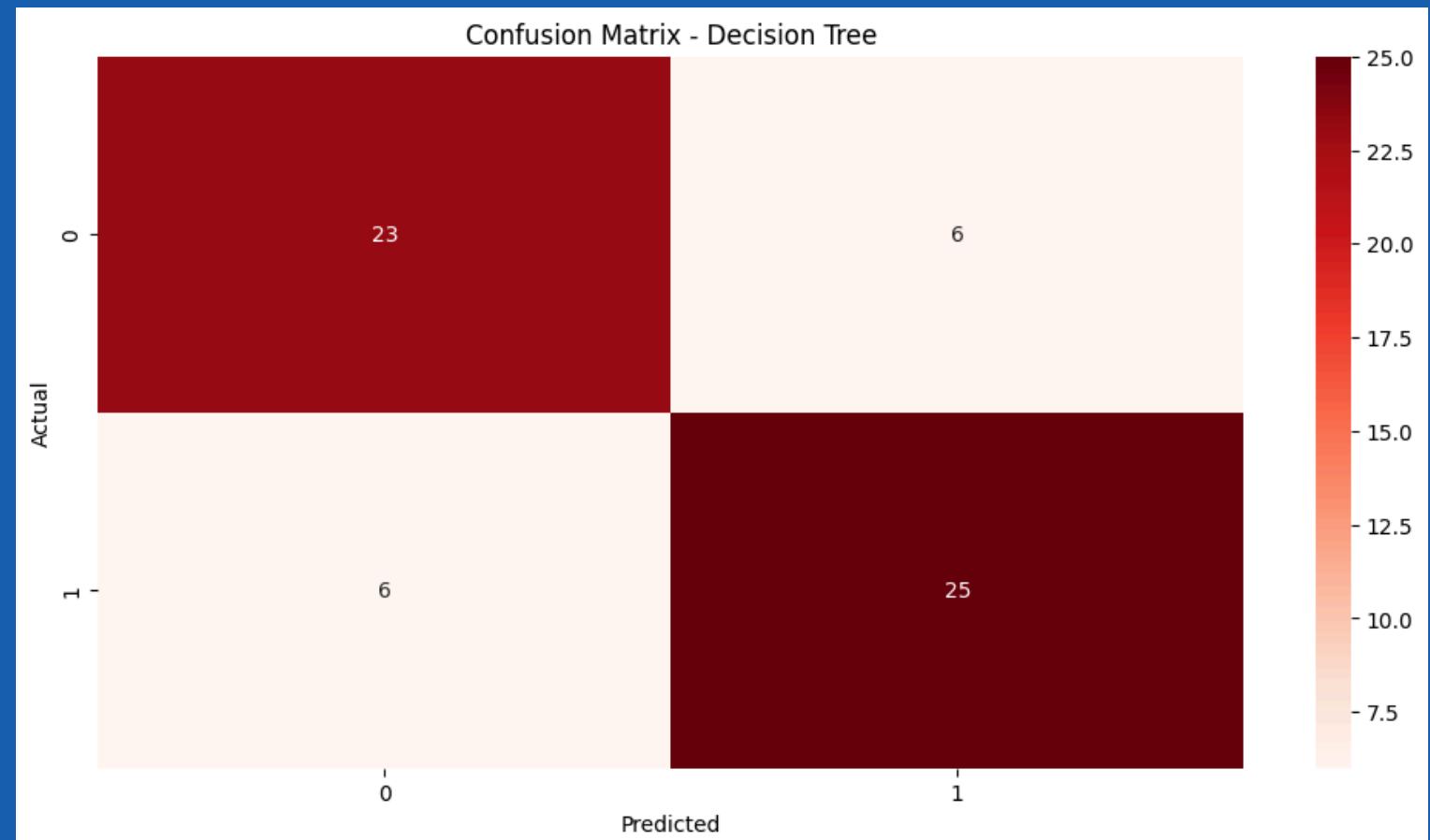
FOLD 3



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 28. Untuk kasus False Positif adalah sebanyak 5, kasus True Negatif sebanyak 20 dan yang terjadi pada kasus False Negatif sebanyak 7.



FOLD 4

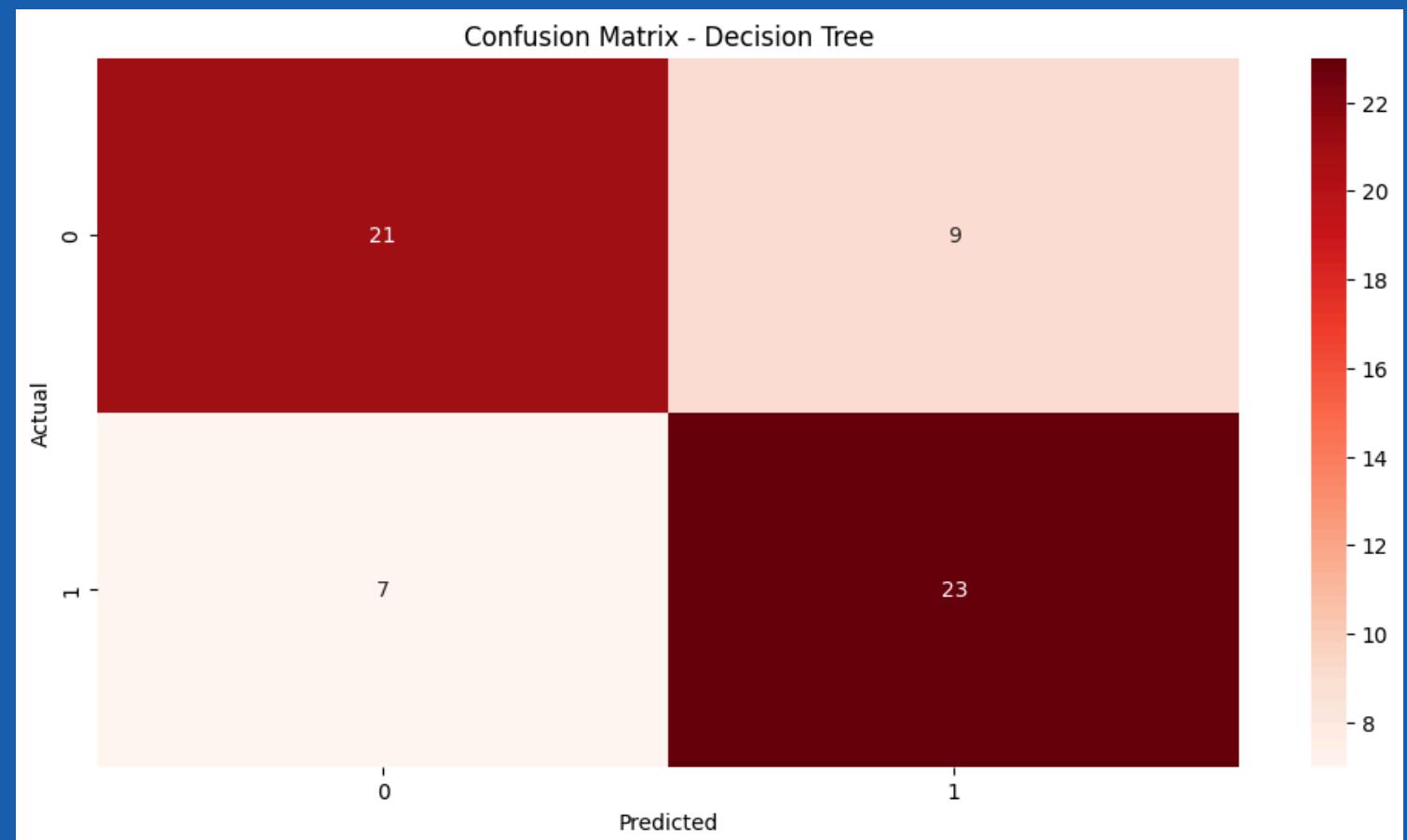


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 25. Untuk kasus False Positif adalah sebanyak 6, kasus True Negatif sebanyak 23 dan yang terjadi pada kasus False Negatif sebanyak 6.

DECISION TREE (K-FOLD)



| FOLD 5

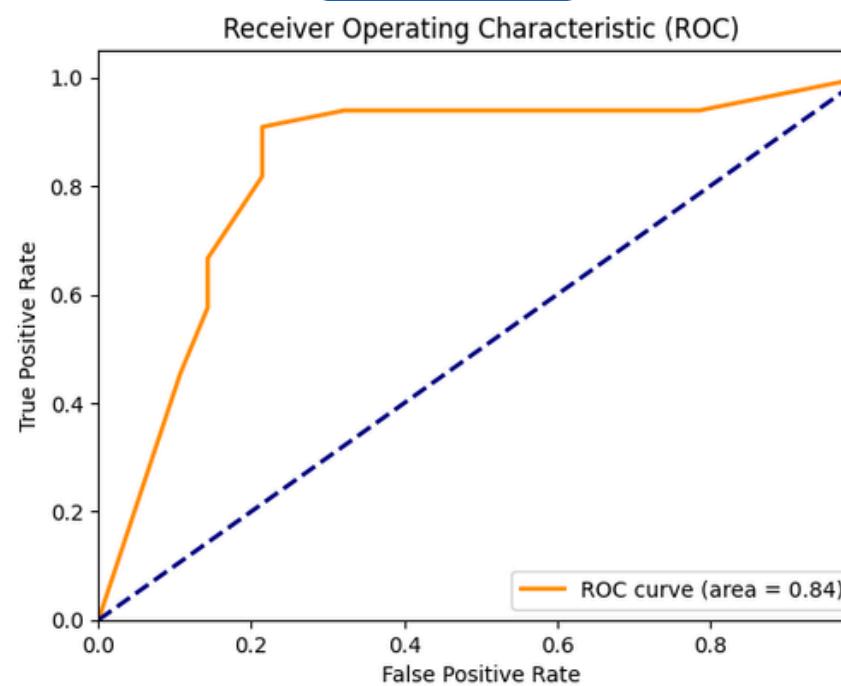


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 23. Untuk kasus False Positif adalah sebanyak 9, kasus True Negatif sebanyak 21 dan yang terjadi pada kasus False Negatif sebanyak 7.

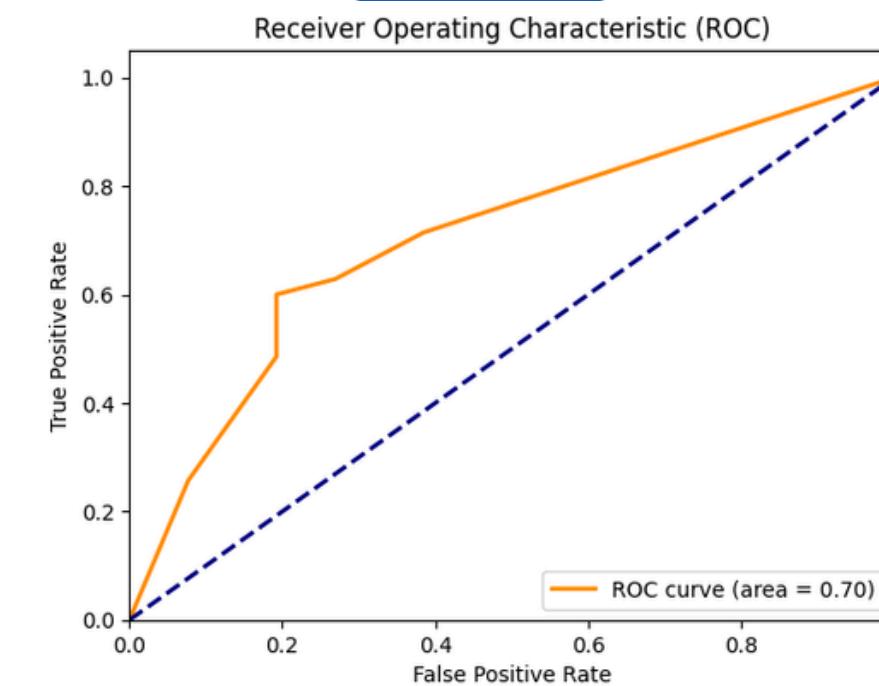
HASIL ROC-AUC

Selanjutnya akan dilakukan evaluasi untuk mengetahui model terbaik yang digunakan. Berikut ROC untuk masing-masing fold.

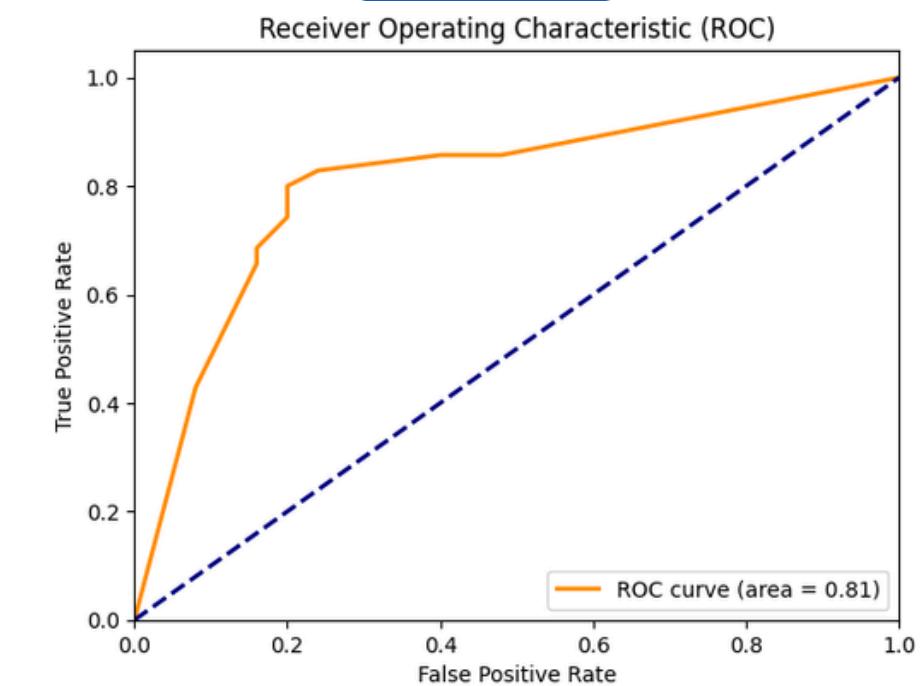
FOLD 1



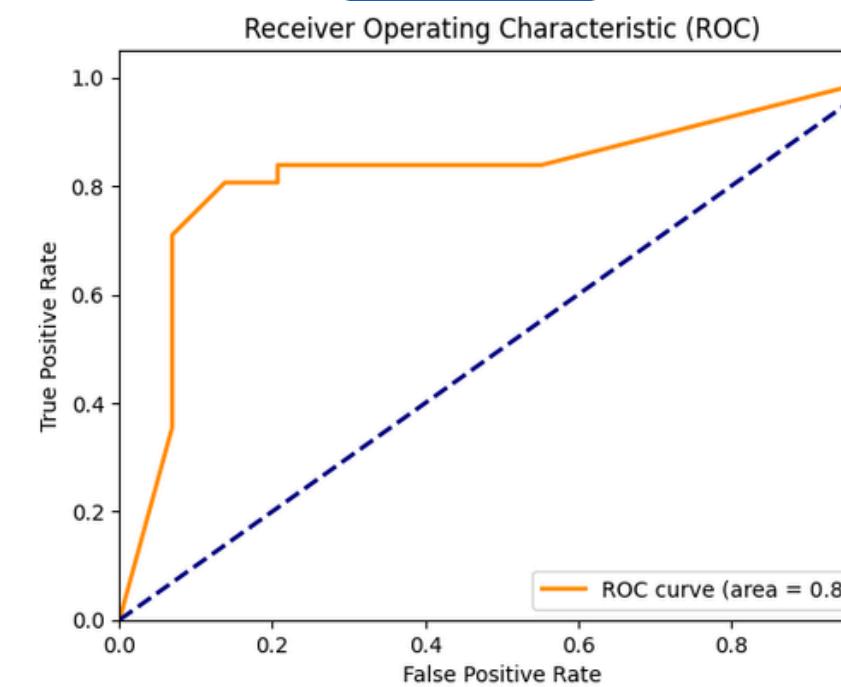
FOLD 2



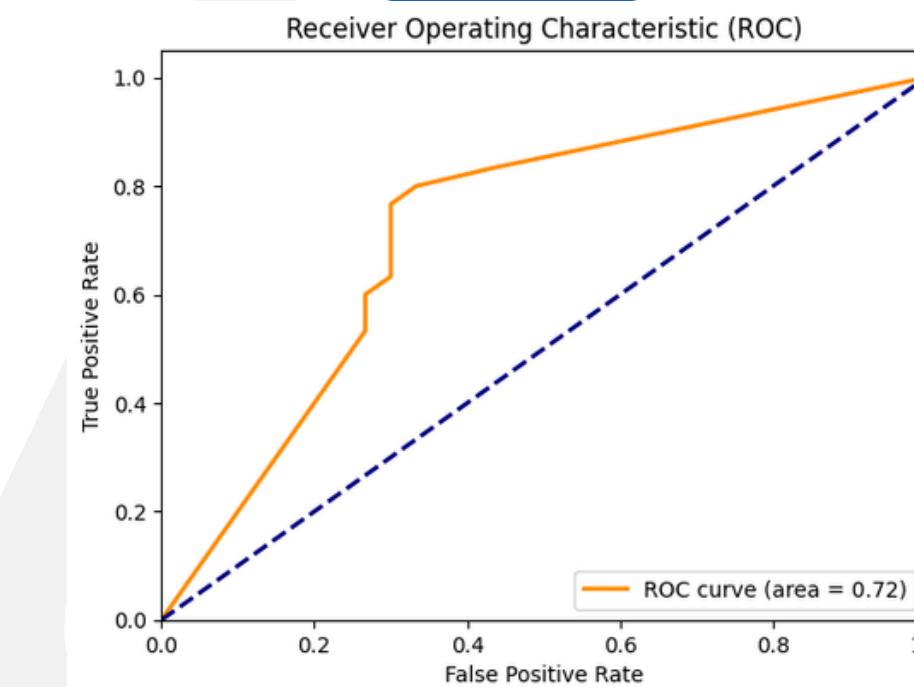
FOLD 3



FOLD 4



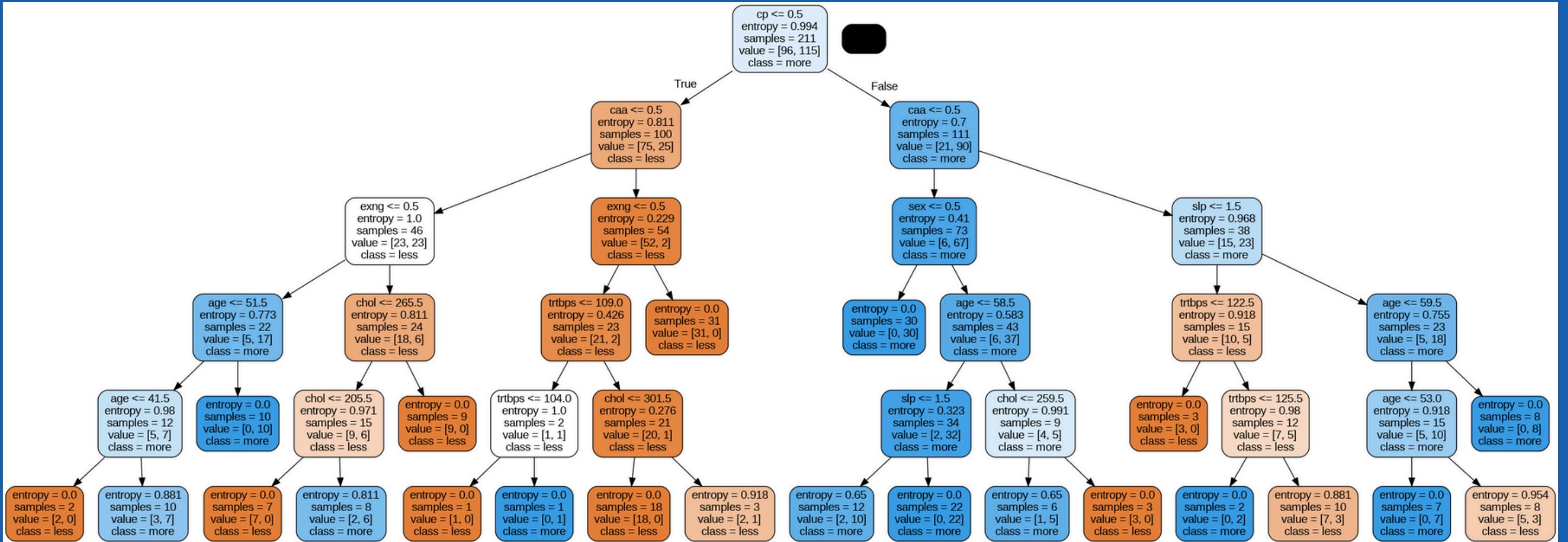
FOLD 5



DECISION TREE (HOLDOUT)



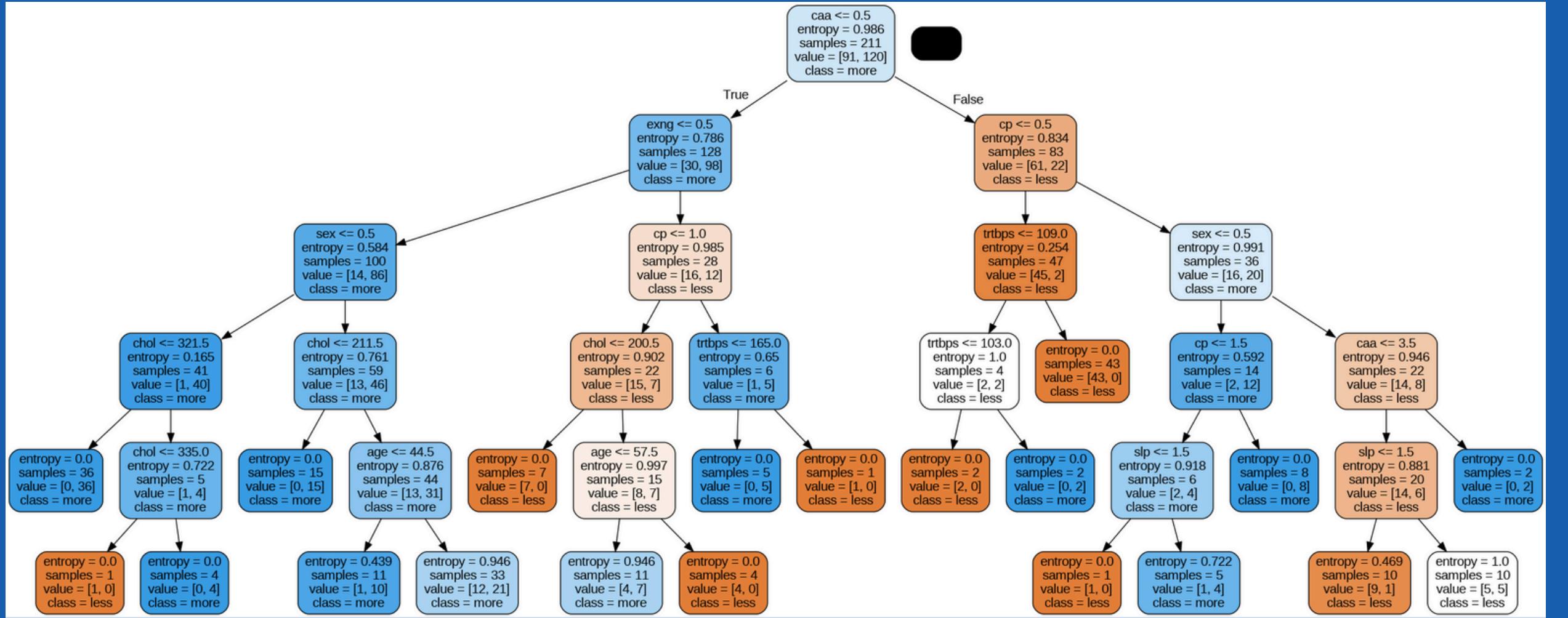
HOLDOUT 1



DECISION TREE (HOLDOUT)



HOLDOUT 2



Berdasarkan output decision tree holdout 2 diperoleh variabel "caa" sebagai root node.

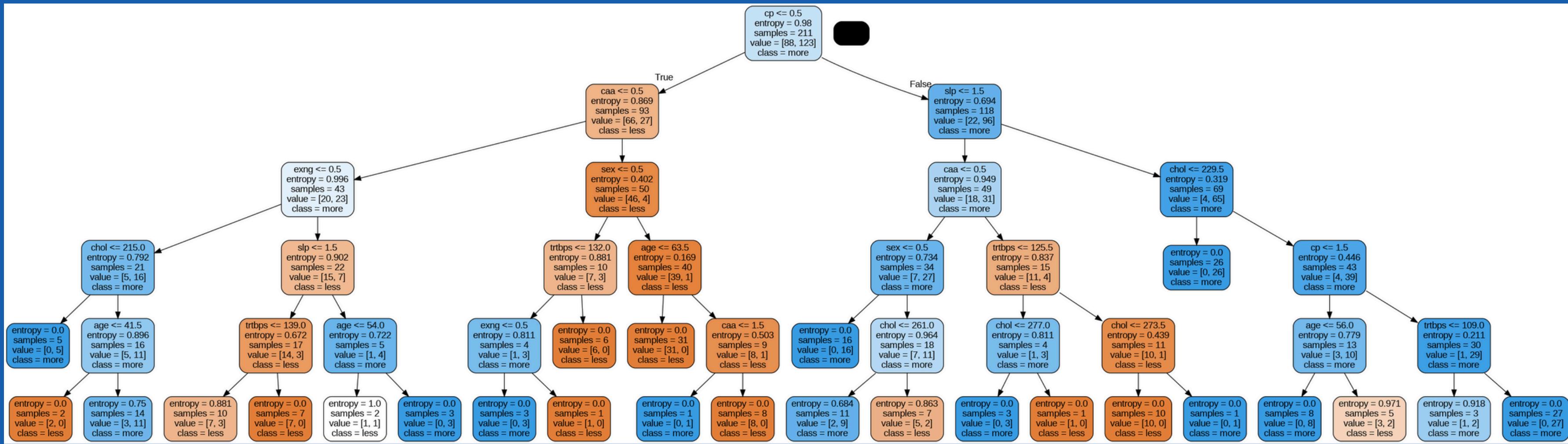
Jika variabel "caa" bernilai true maka variabel "exng" menjadi decision node, diikuti variabel "sex" dan "cp" sebagai sub decision node.

Jika variabel "caa" bernilai false maka variabel "cp" menjadi decision node, diikuti variabel "trtbps" dan "sex" sebagai sub decision node.

DECISION TREE (HOLDOUT)



HOLDOUT 3



Berdasarkan output decision tree holdout 3 diperoleh variabel "cp" sebagai root node.

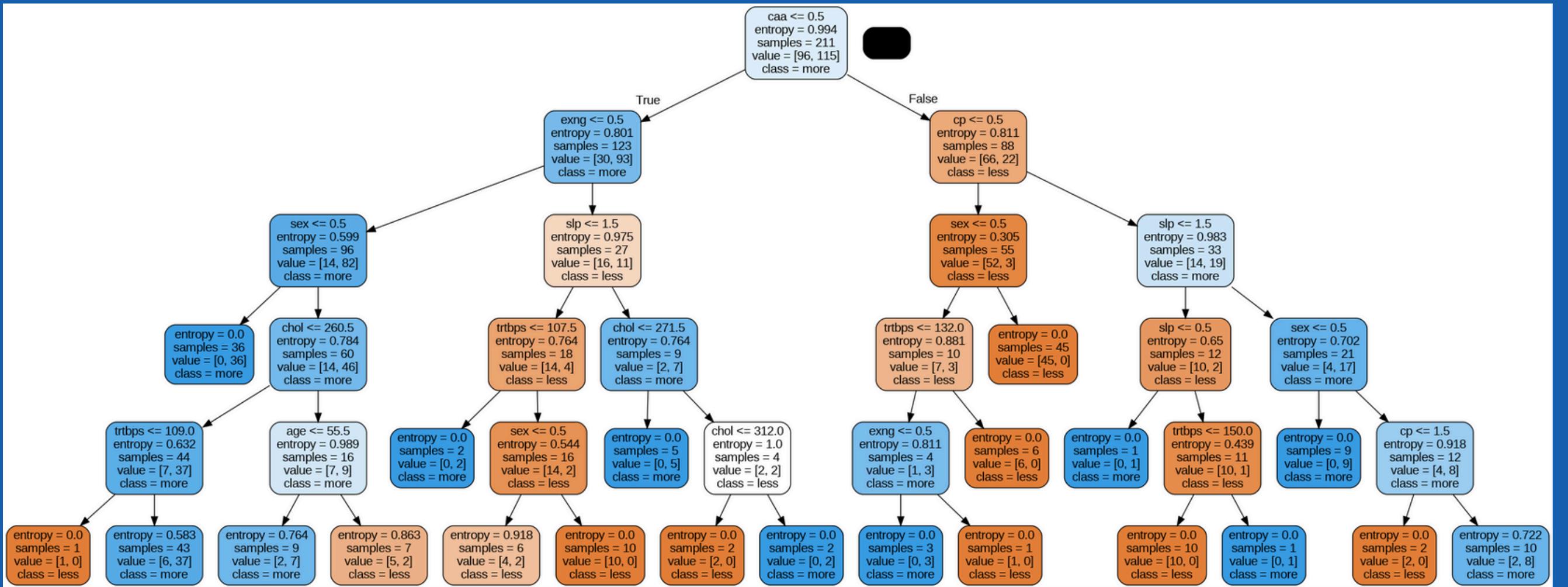
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "exng" dan "sex" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "slp" menjadi decision node, diikuti variabel "caa" dan "chol" sebagai sub decision node.

DECISION TREE HOLDOUT



HOLDOUT 4



Berdasarkan output decision tree holdout 4 diperoleh variabel "caa" sebagai root node.

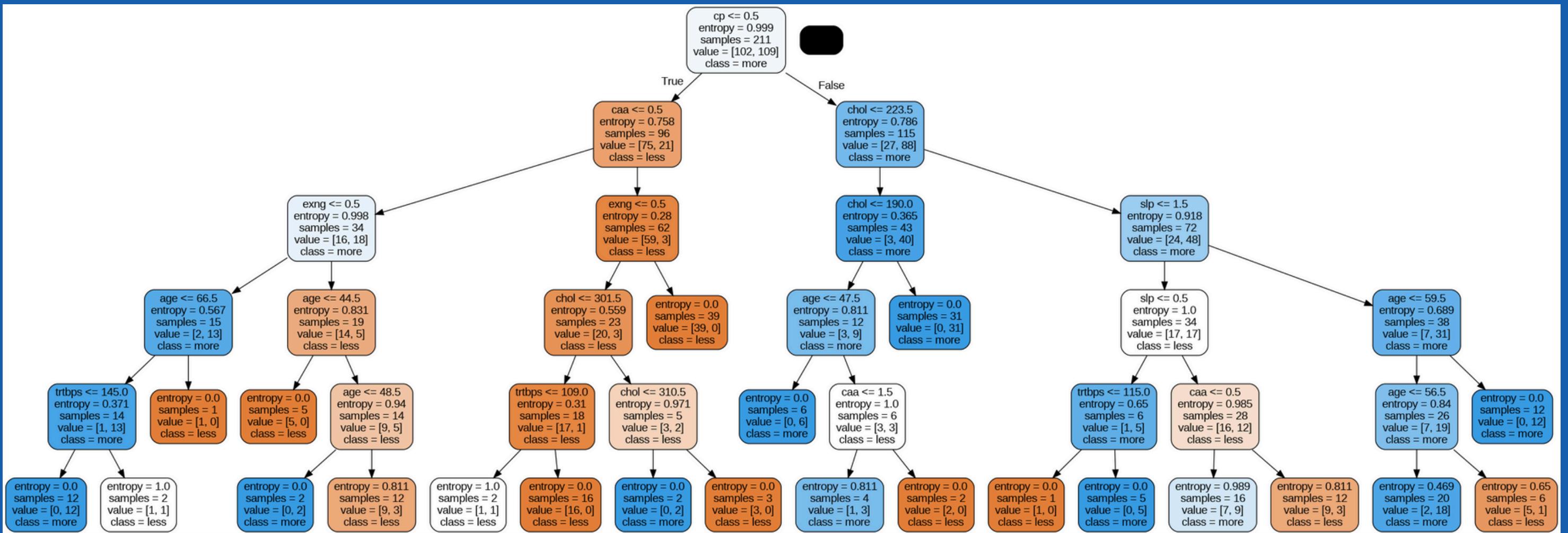
Jika variabel "caa" bernilai true maka variabel "exng" menjadi decision node, diikuti variabel "sex" dan "slp" sebagai sub decision node.

Jika variabel "caa" bernilai false maka variabel "cp" menjadi decision node, diikuti variabel "sex" dan "slp" sebagai sub decision node.

DECISION TREE (HOLDOUT)



HOLDOUT 5



Berdasarkan output decision tree holdout 5 diperoleh variabel "cp" sebagai root node.

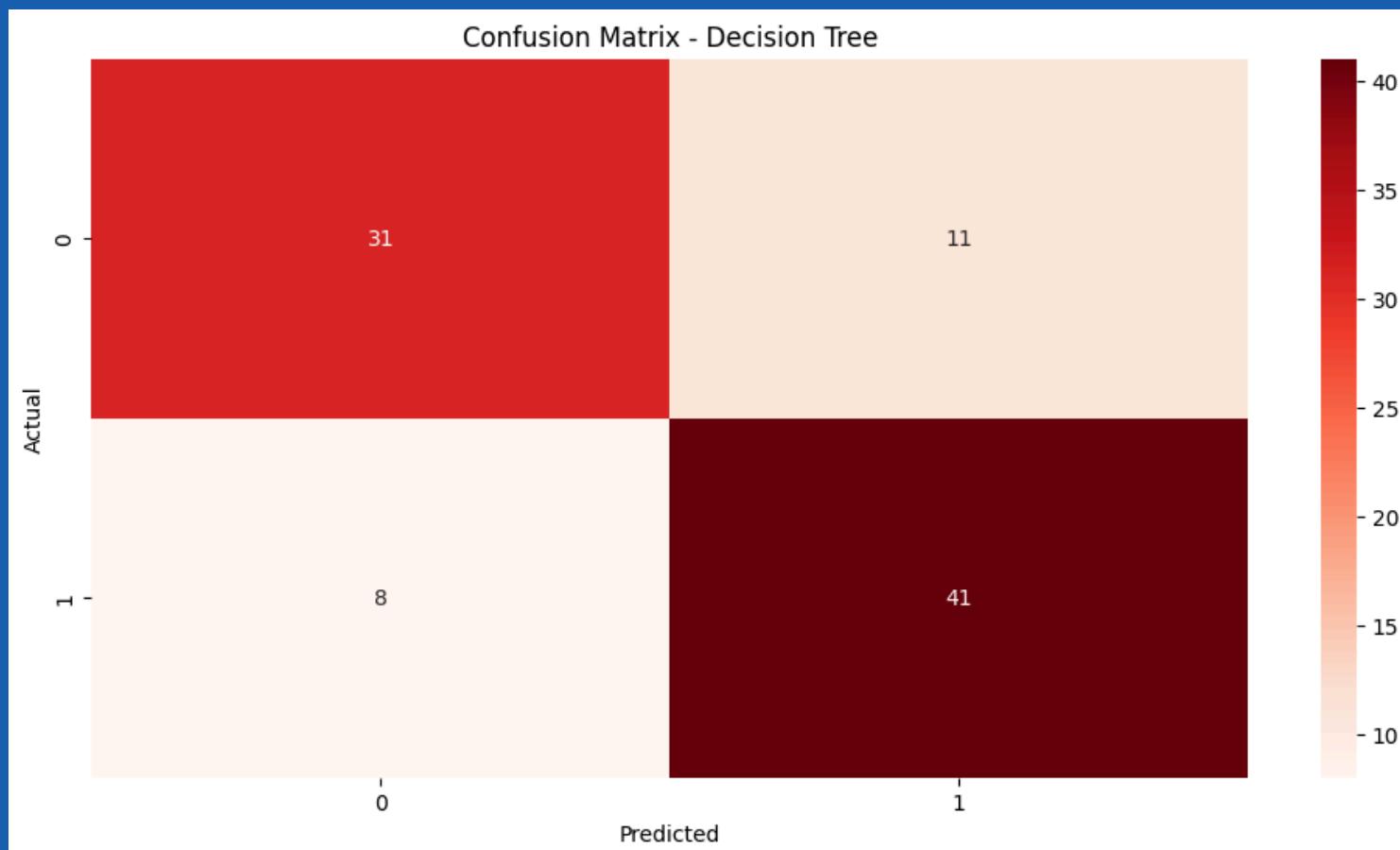
Jika variabel "cp" bernilai true maka variabel "caa" menjadi decision node, diikuti variabel "exng-more" dan "exng-less" sebagai sub decision node.

Jika variabel "cp" bernilai false maka variabel "chol" menjadi decision node, diikuti variabel "slp" sebagai sub decision node.

DECISION TREE (HOLDOUT)



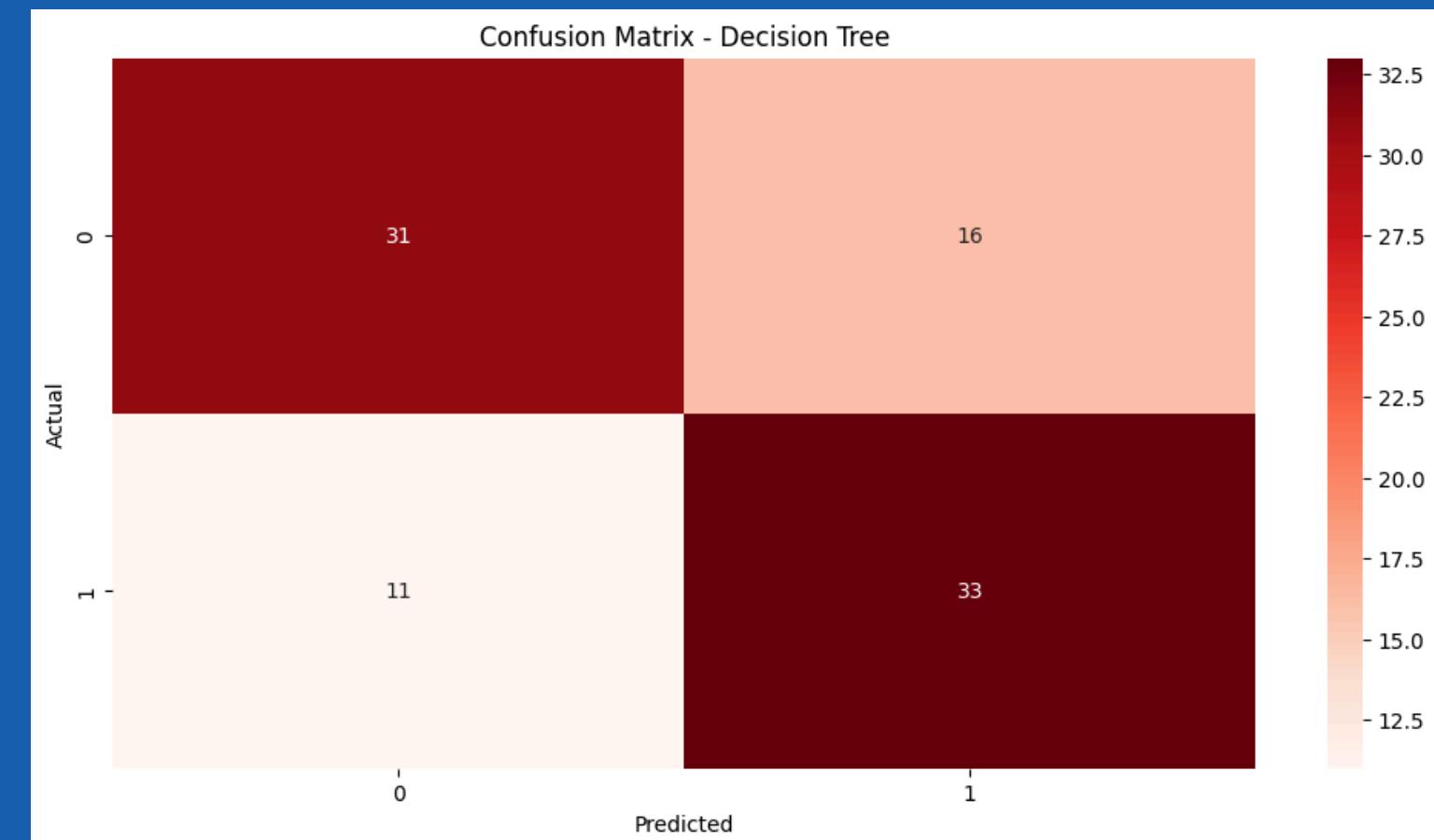
HOLDOUT 1



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 41. Untuk kasus False Positif adalah sebanyak 11, kasus True Negatif sebanyak 31 dan yang terjadi pada kasus False Negatif sebanyak 8.



HOLDOUT 2

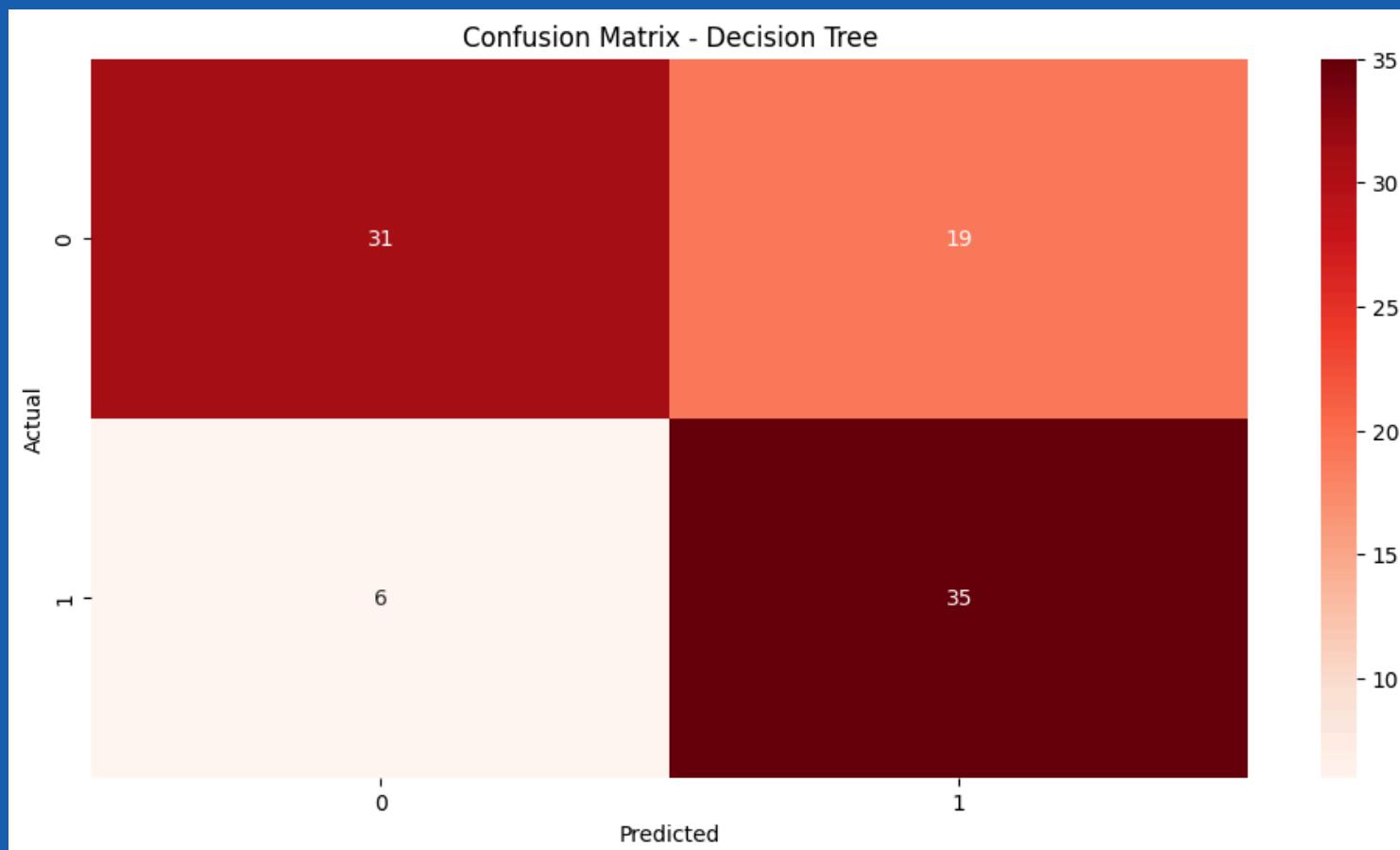


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 33. Untuk kasus False Positif adalah sebanyak 16, kasus True Negatif sebanyak 31 dan yang terjadi pada kasus False Negatif sebanyak 11.

DECISION TREE (HOLDOUT)



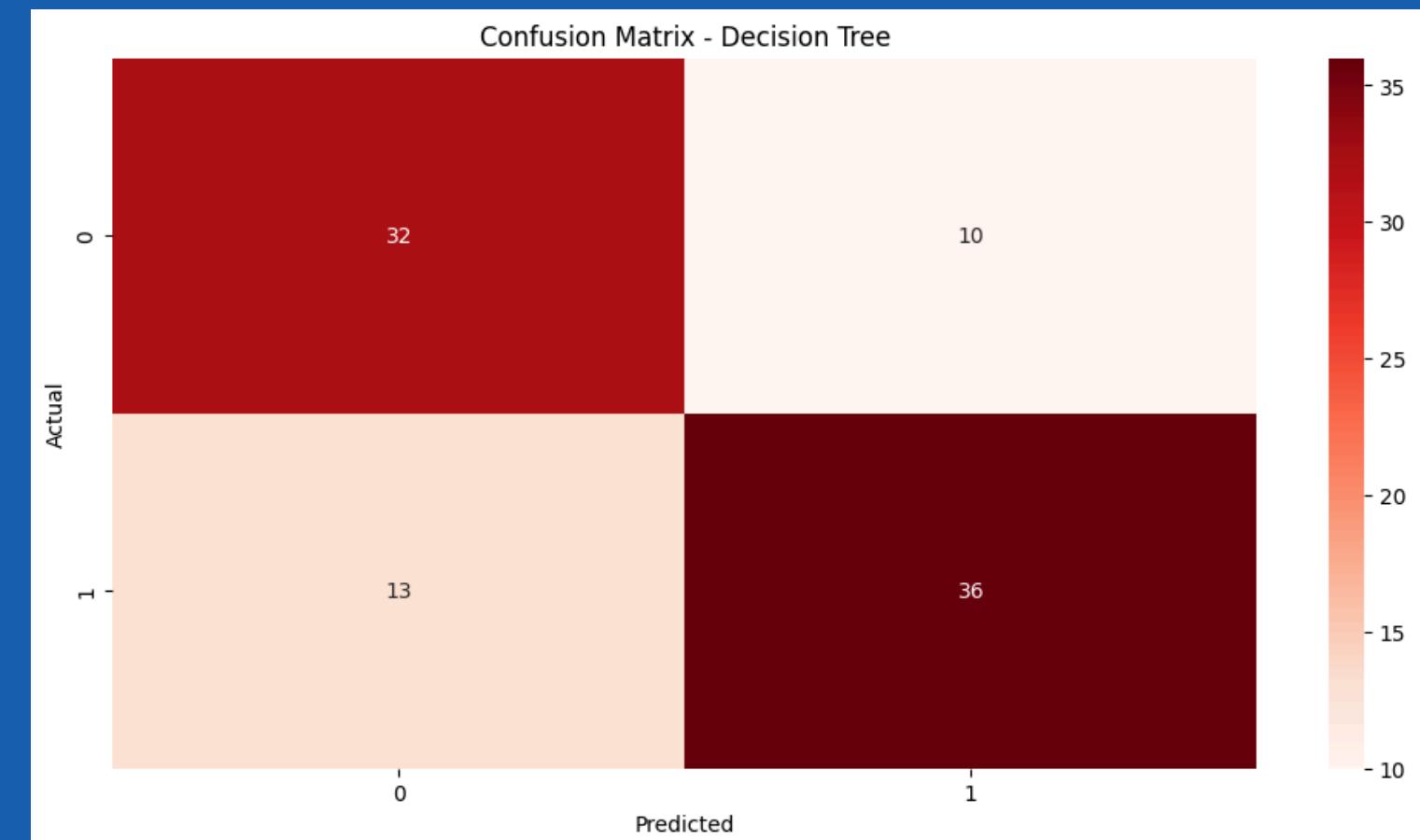
HOLDOUT 3



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 35. Untuk kasus False Positif adalah sebanyak 19, kasus True Negatif sebanyak 31 dan yang terjadi pada kasus False Negatif sebanyak 6.



HOLDOUT 4

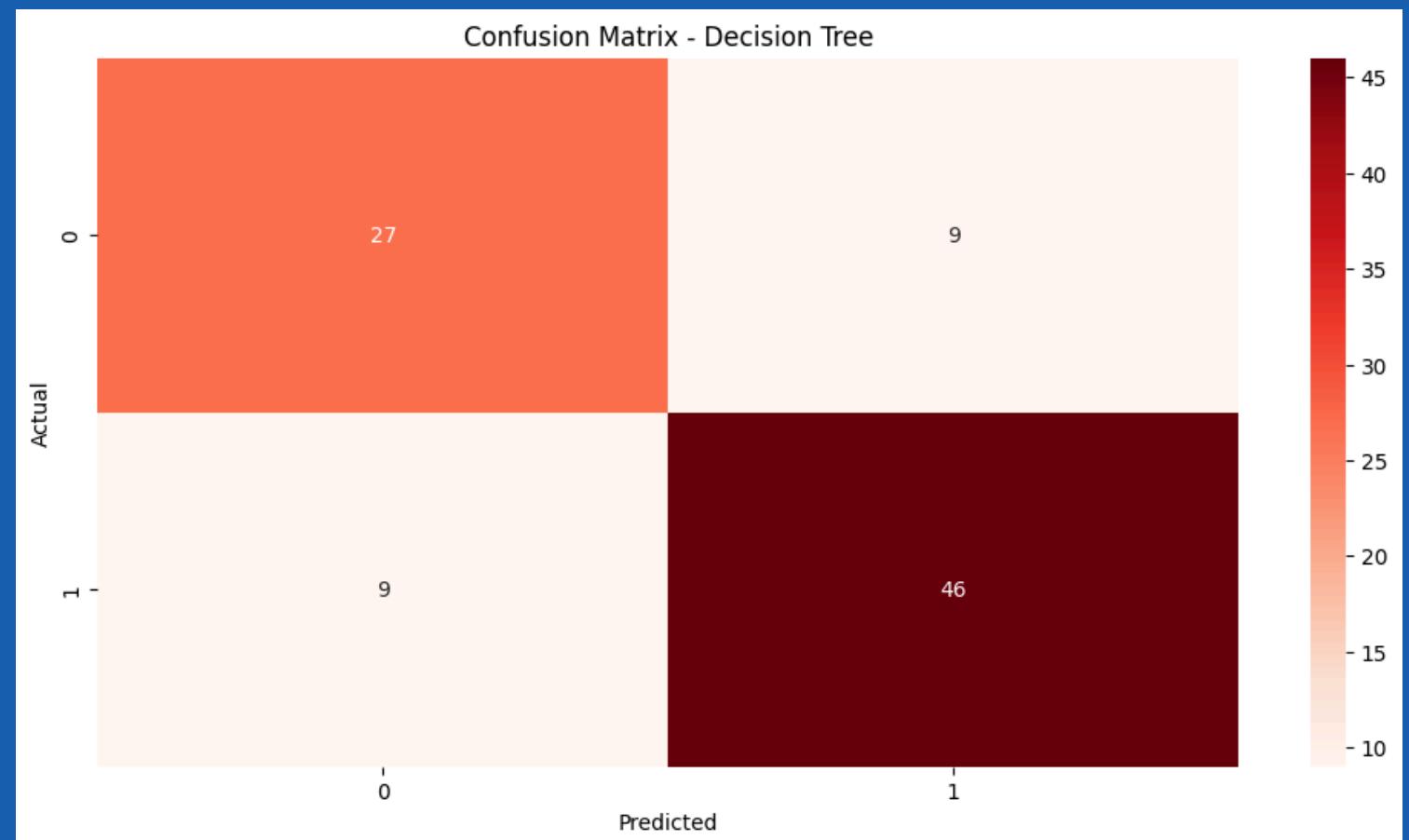


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 36. Untuk kasus False Positif adalah sebanyak 10, kasus True Negatif sebanyak 32 dan yang terjadi pada kasus False Negatif sebanyak 13.

DECISION TREE (HOLDOUT)



HOLDOUT 5

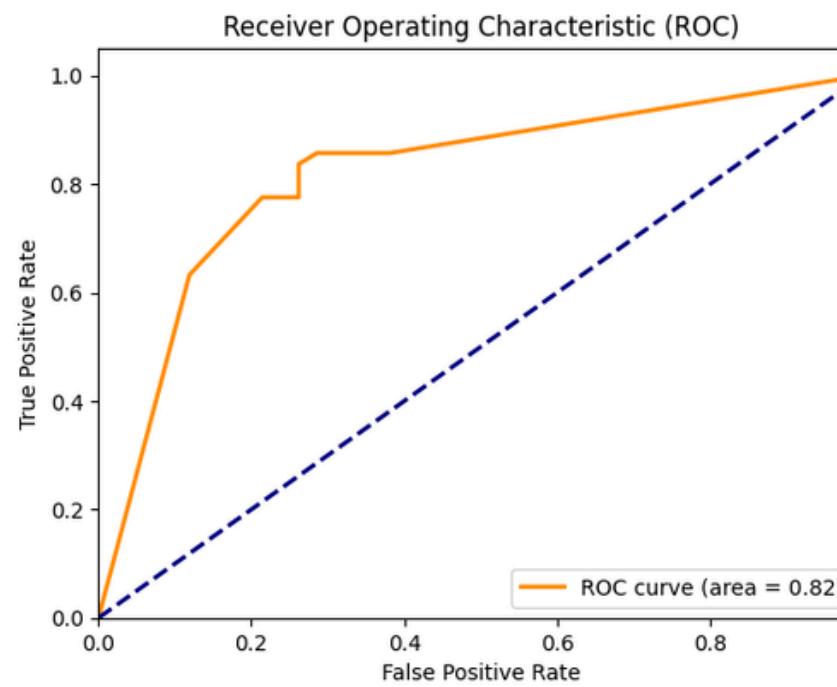


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 46. Untuk kasus False Positif adalah sebanyak 9, kasus True Negatif sebanyak 27 dan yang terjadi pada kasus False Negatif sebanyak 9.

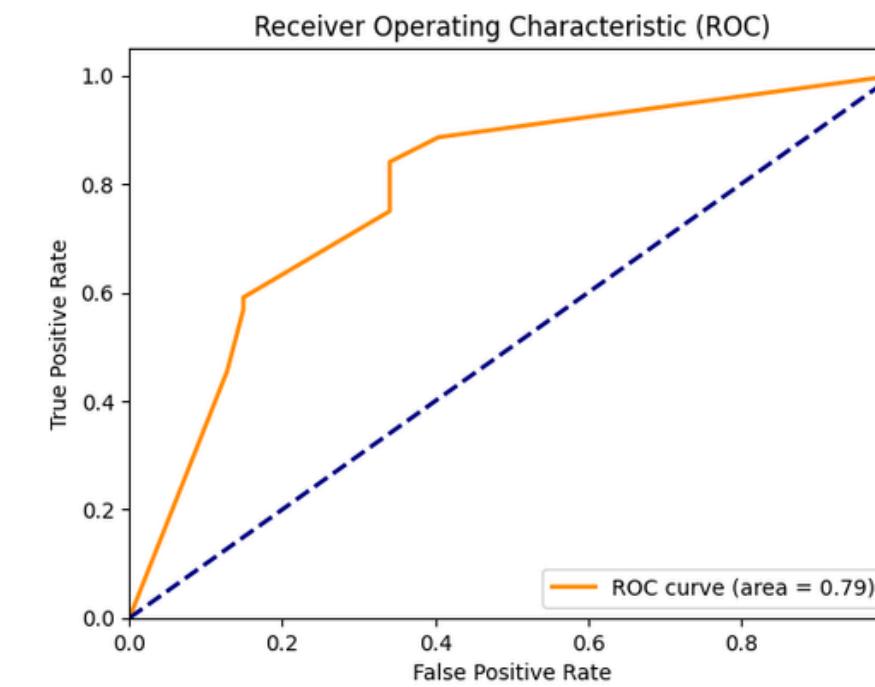
HASIL ROC-AUC

Selanjutnya akan dilakukan evaluasi untuk mengetahui model terbaik yang digunakan. Berikut ROC untuk masing-masing holdout.

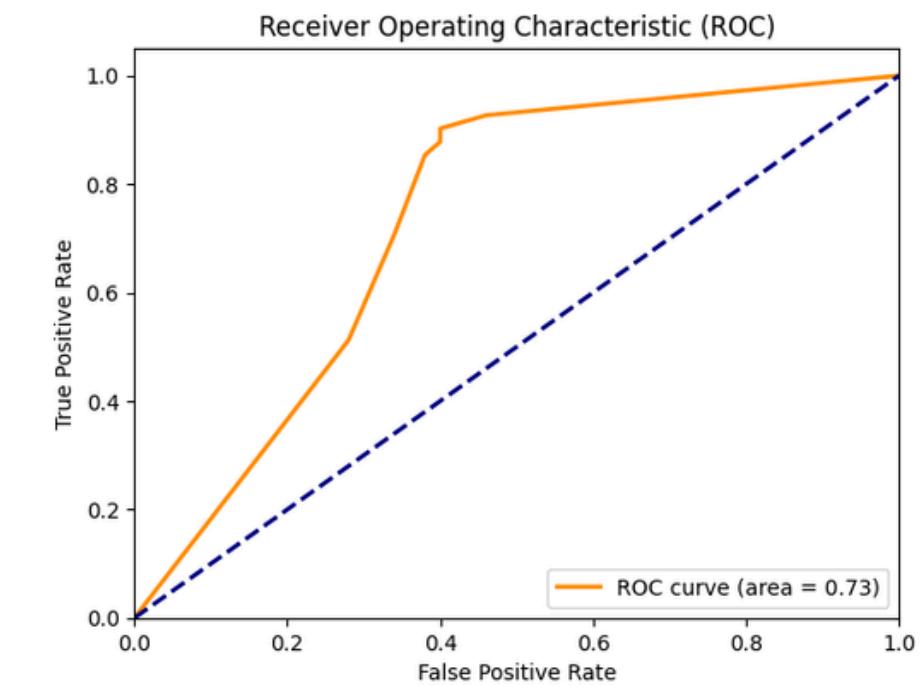
HOLDOUT 1



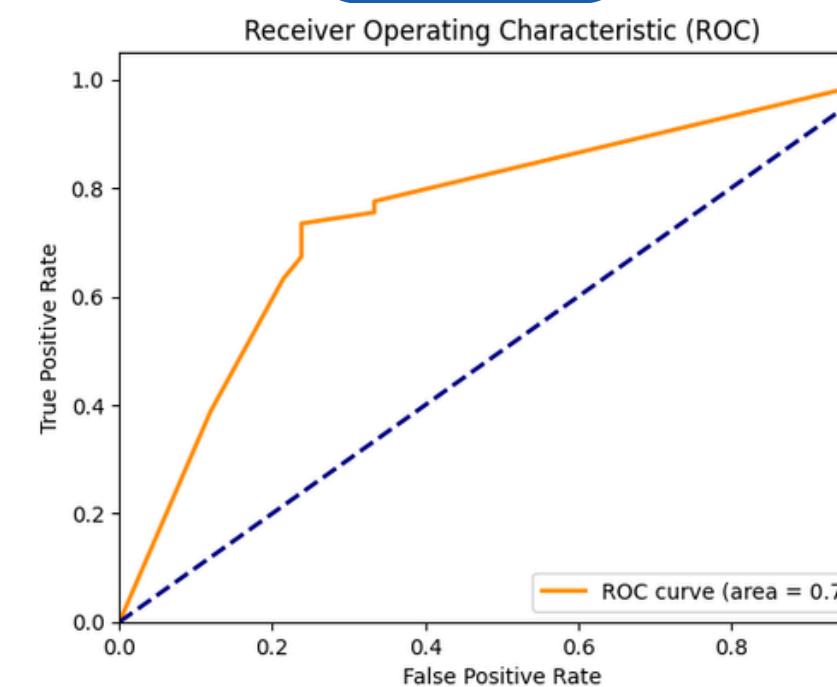
HOLDOUT 2



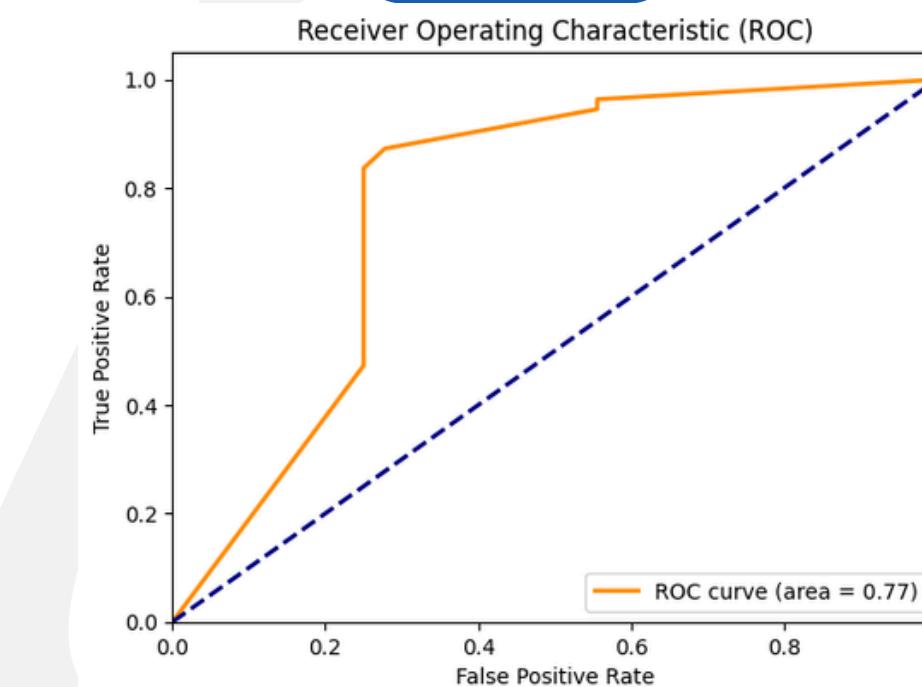
HOLDOUT 3



HOLDOUT 4



HOLDOUT 5



HASIL ANALISIS DECISION TREE



CROSS VALIDATION K-FOLD

Fold	Sensitifitas	Spesifitas	Akurasi	F-Measure	AUC
1	0,85	0,79	0,85	0,85	0,84
2	0,69	0,81	0,67	0,67	0,70
3	0,80	0,80	0,80	0,80	0,81
4	0,80	0,79	0,80	0,80	0,82
5	0,73	0,70	0,73	0,73	0,72

Diperoleh bahwa AUC tertinggi adalah fold 1 yaitu 0.84, dimana mempunyai F- Measure tertinggi yaitu 0.85. Maka, dapat disimpulkan fold 1 merupakan model terbaik dalam pengklasifikasianya (good classification).



REPEATED HOLDOUT

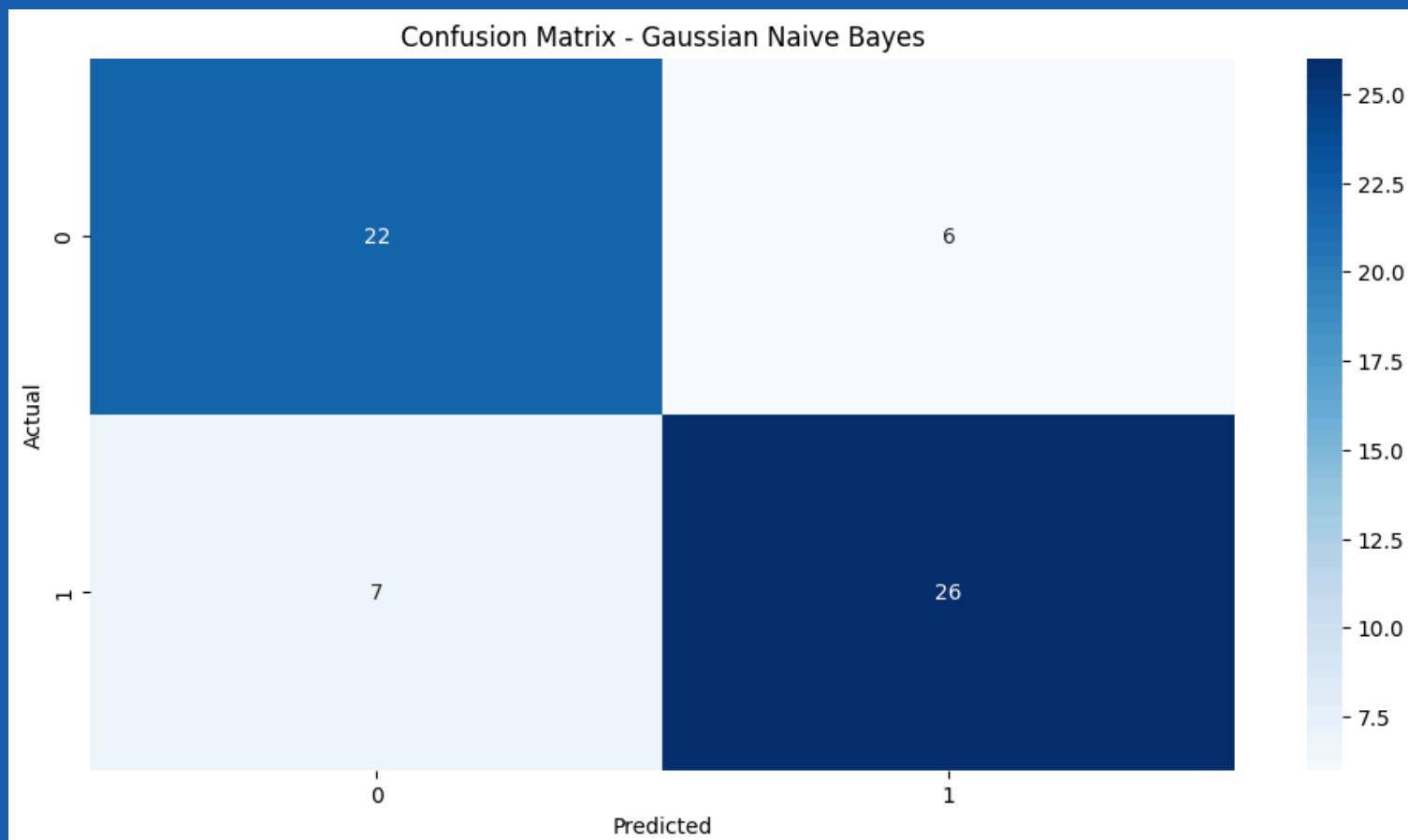
Holdout	Sensitifitas	Spesifitas	Akurasi	F-Measure	AUC
1	0,79	0,74	0,79	0,79	0,82
2	0,70	0,66	0,70	0,70	0,79
3	0,74	0,62	0,73	0,72	0,73
4	0,75	0,76	0,75	0,75	0,75
5	0,79	0,75	0,80	0,79	0,77

Diperoleh bahwa AUC tertinggi adalah holdout 1 yaitu 0.82, dimana mempunyai F-Measure tertinggi yaitu 0.79. Maka, dapat disimpulkan holdout 1 merupakan model terbaik dalam pengklasifikasianya (good classification).

NAIVE BAYES (K-FOLD)



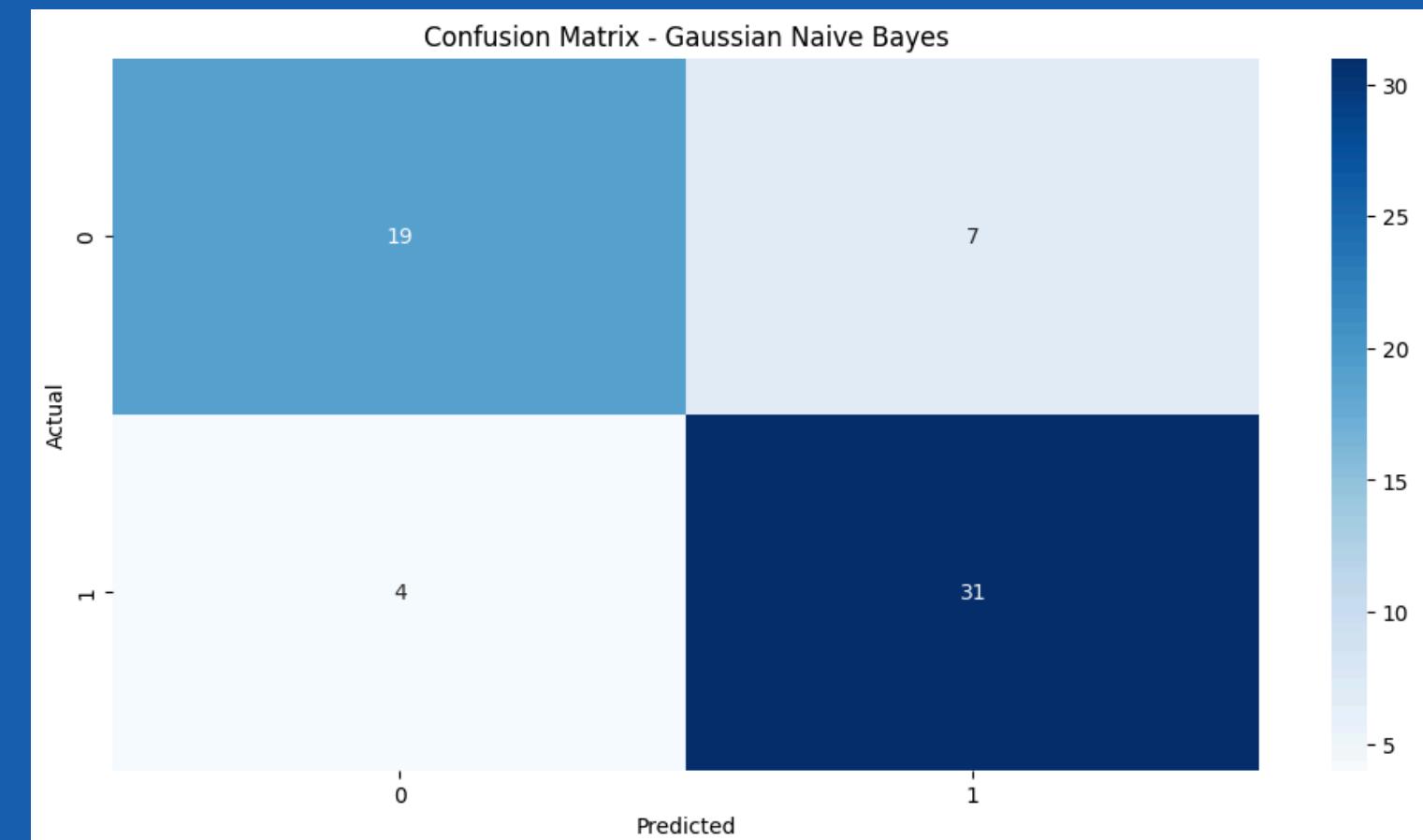
FOLD 1



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 26. Untuk kasus False Positif adalah sebanyak 6, kasus True Negatif sebanyak 22 dan yang terjadi pada kasus False Negatif sebanyak 7.



FOLD 2

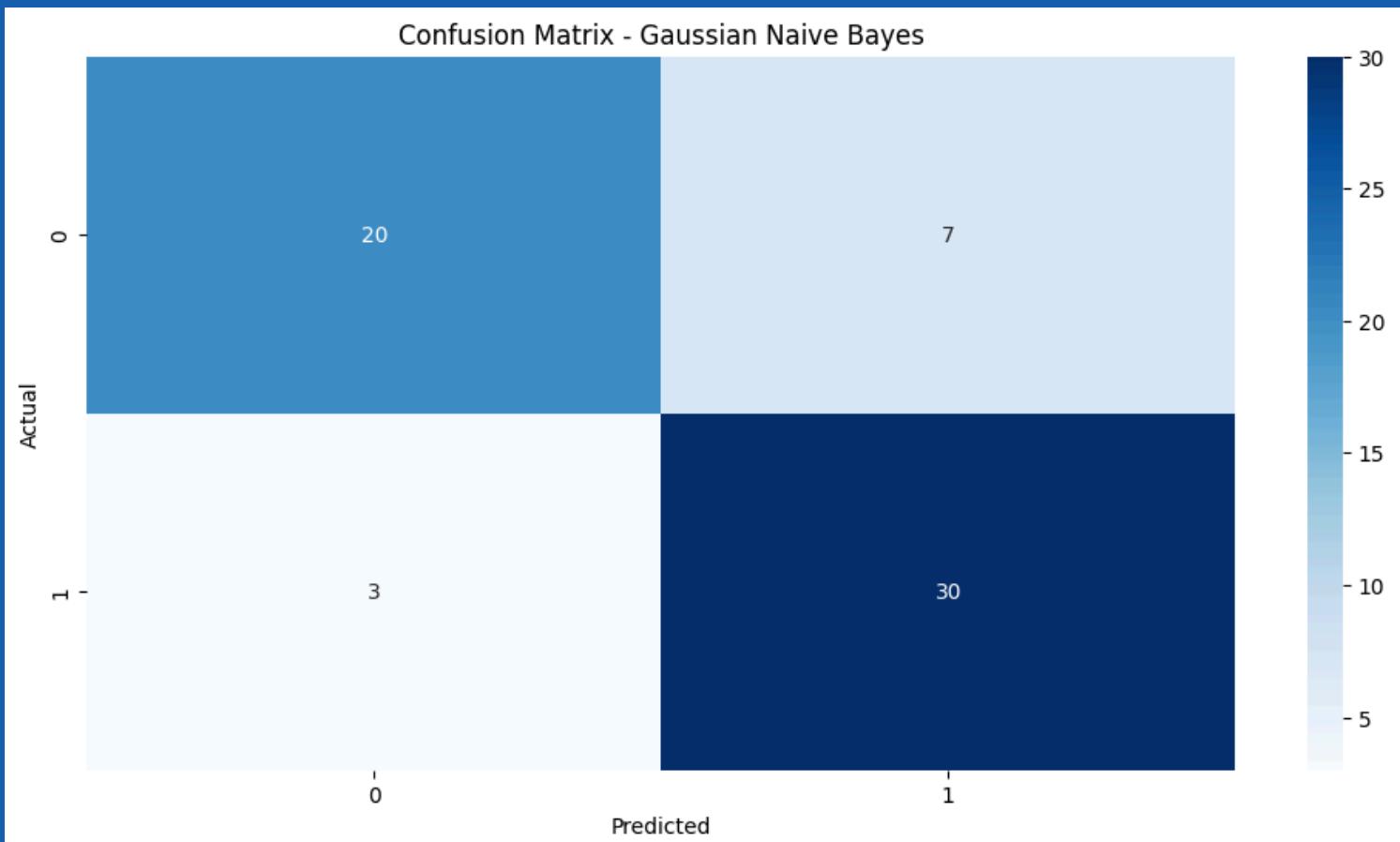


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 31. Untuk kasus False Positif adalah sebanyak 7, kasus True Negatif sebanyak 19 dan yang terjadi pada kasus False Negatif sebanyak 4.

NAIVE BAYES (K-FOLD)



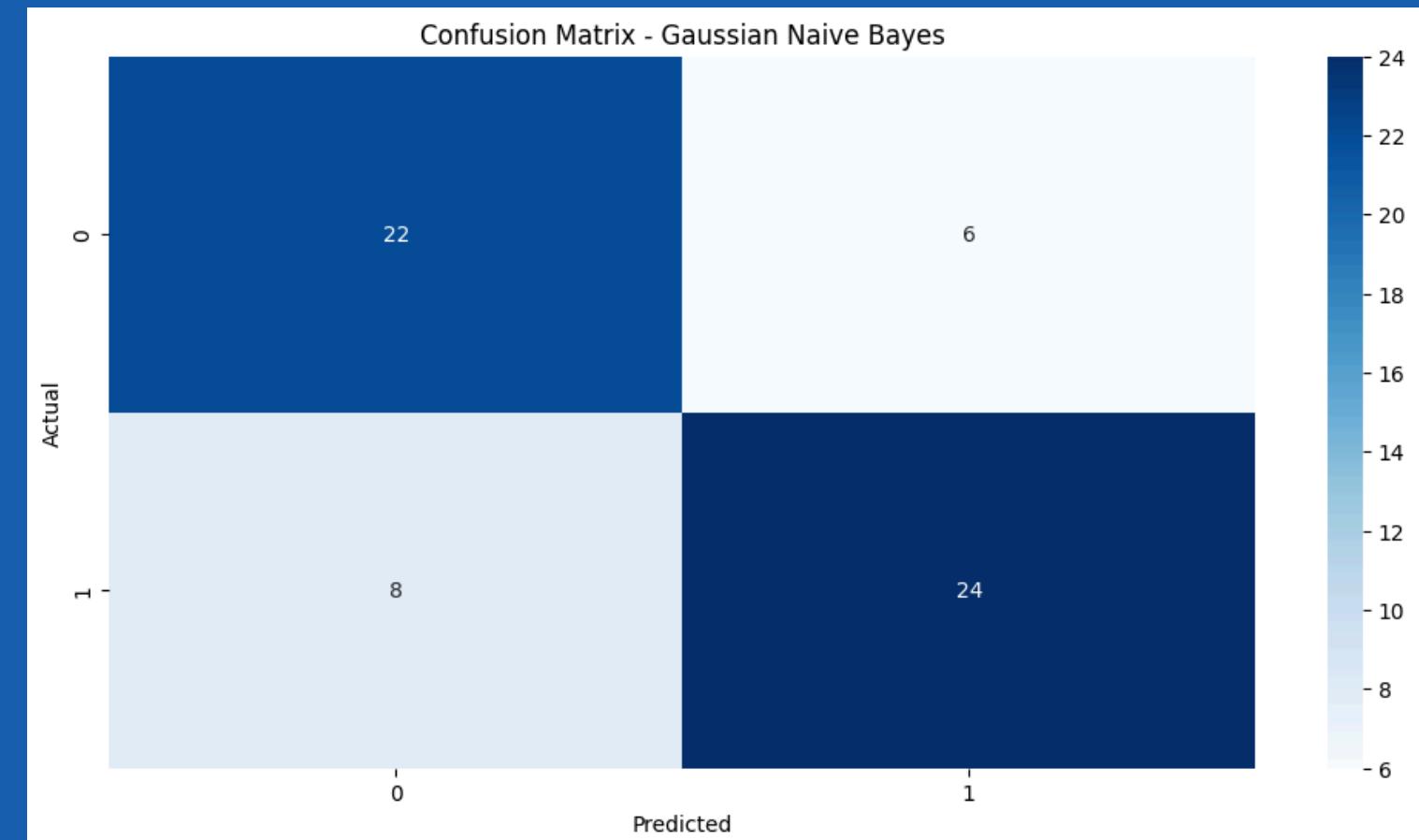
FOLD 3



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 30. Untuk kasus False Positif adalah sebanyak 7, kasus True Negatif sebanyak 20 dan yang terjadi pada kasus False Negatif sebanyak 3.



FOLD 4

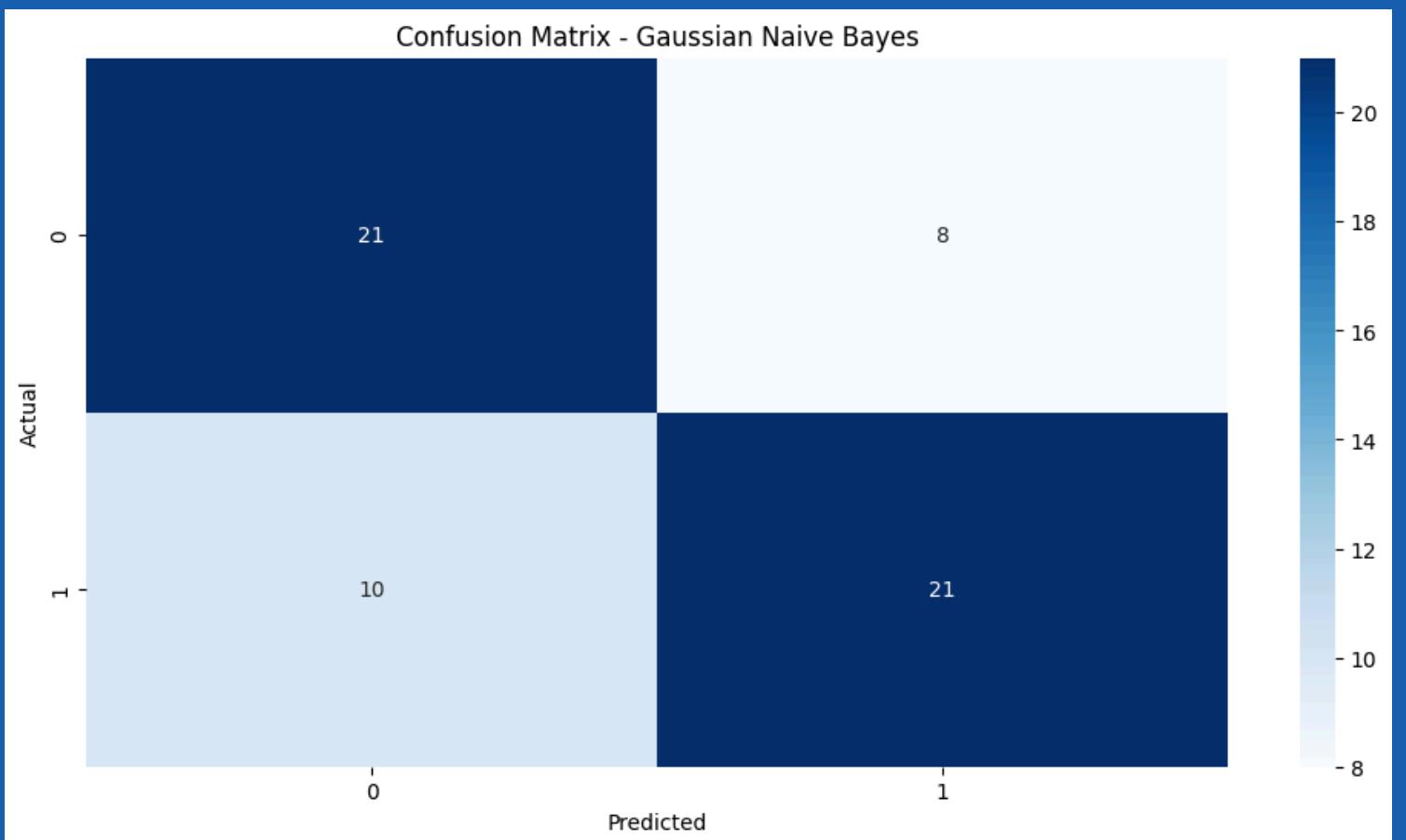


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 24. Untuk kasus False Positif adalah sebanyak 6, kasus True Negatif sebanyak 22 dan yang terjadi pada kasus False Negatif sebanyak 8.

NAIVE BAYES (K-FOLD)



| FOLD 5



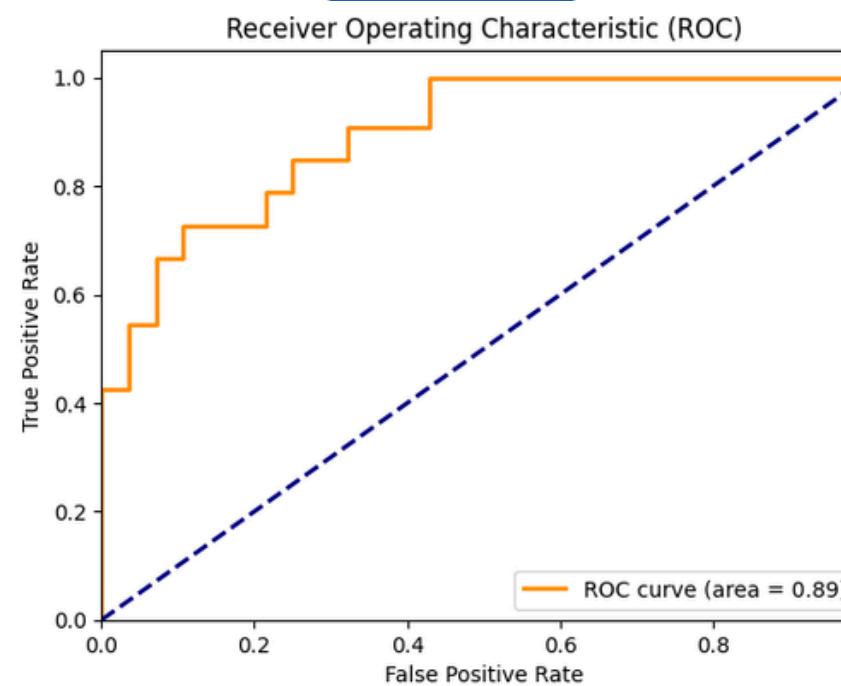
Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 21. Untuk kasus False Positif adalah sebanyak 8, kasus True Negatif sebanyak 21 dan yang terjadi pada kasus False Negatif sebanyak 10.



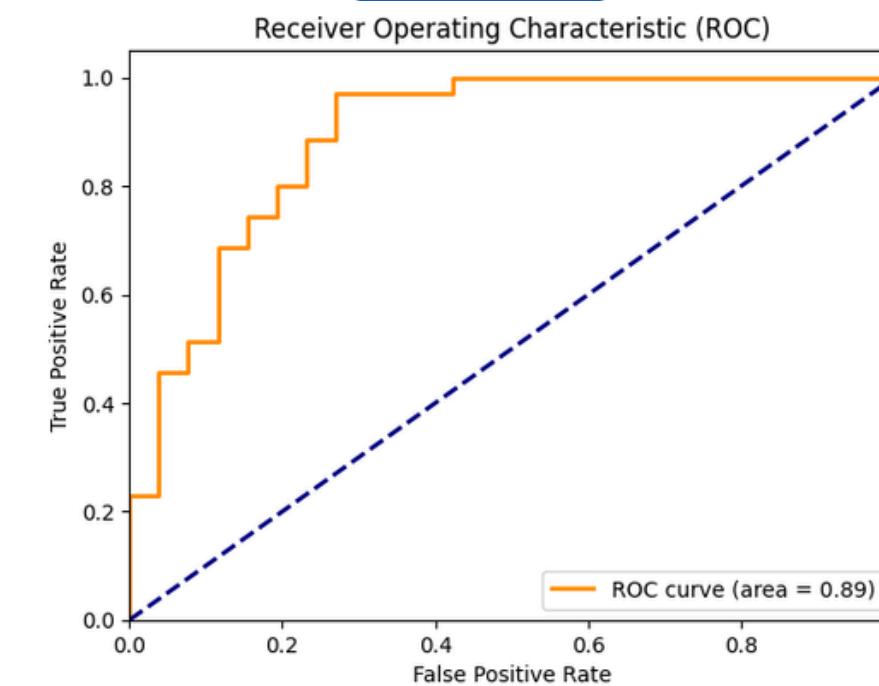
HASIL ROC-AUC

Selanjutnya akan dilakukan evaluasi untuk mengetahui model terbaik yang digunakan. Berikut ROC untuk masing-masing fold.

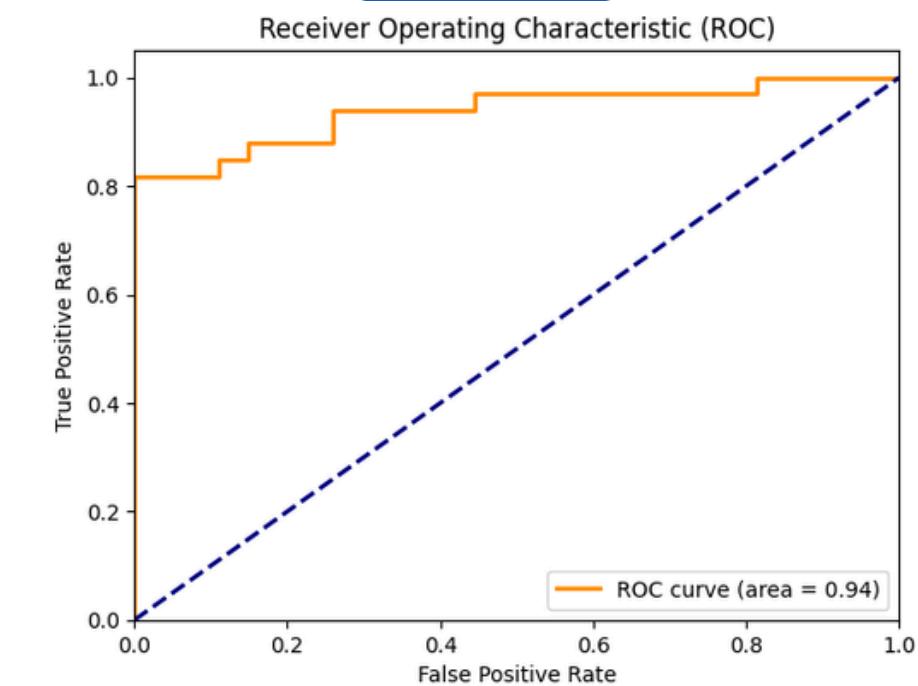
FOLD 1



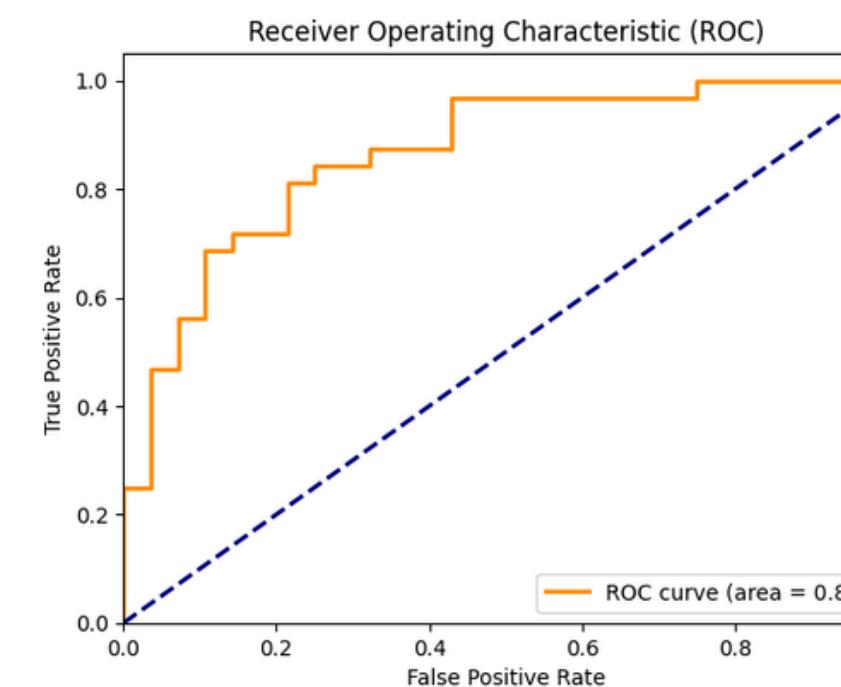
FOLD 2



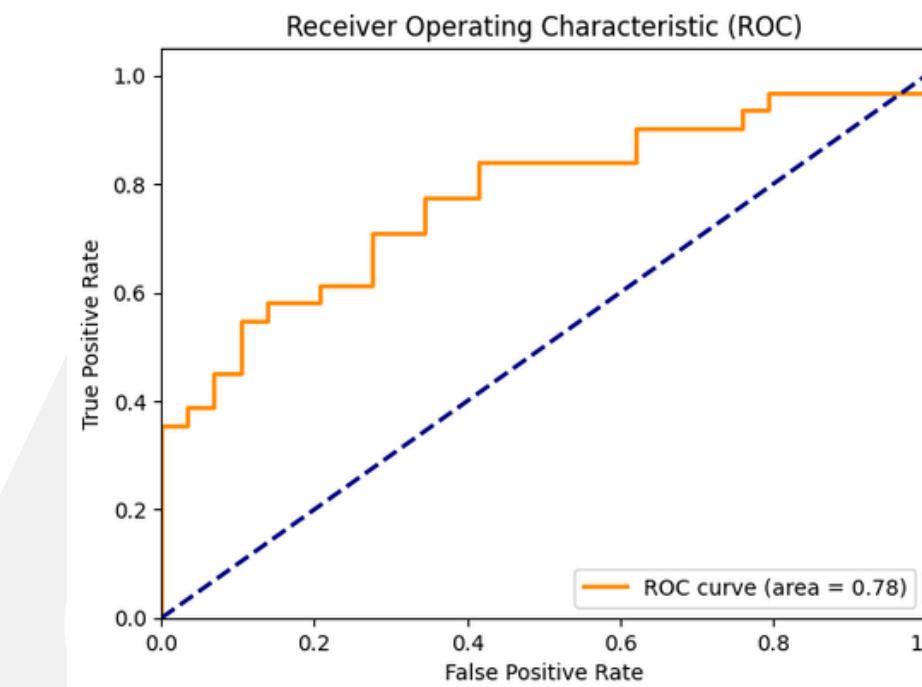
FOLD 3



FOLD 4



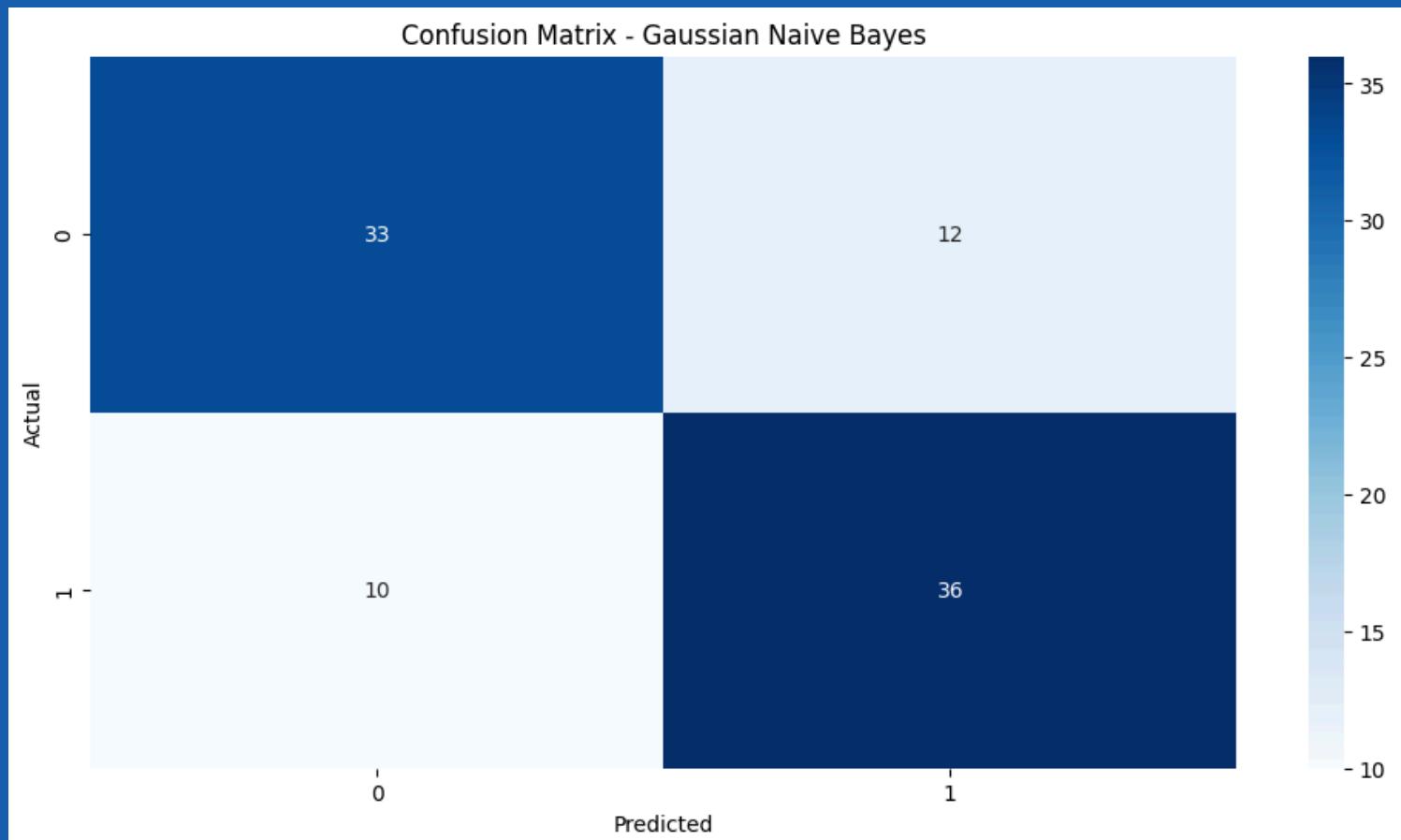
FOLD 5



NAIVE BAYES (HOLDOUT)



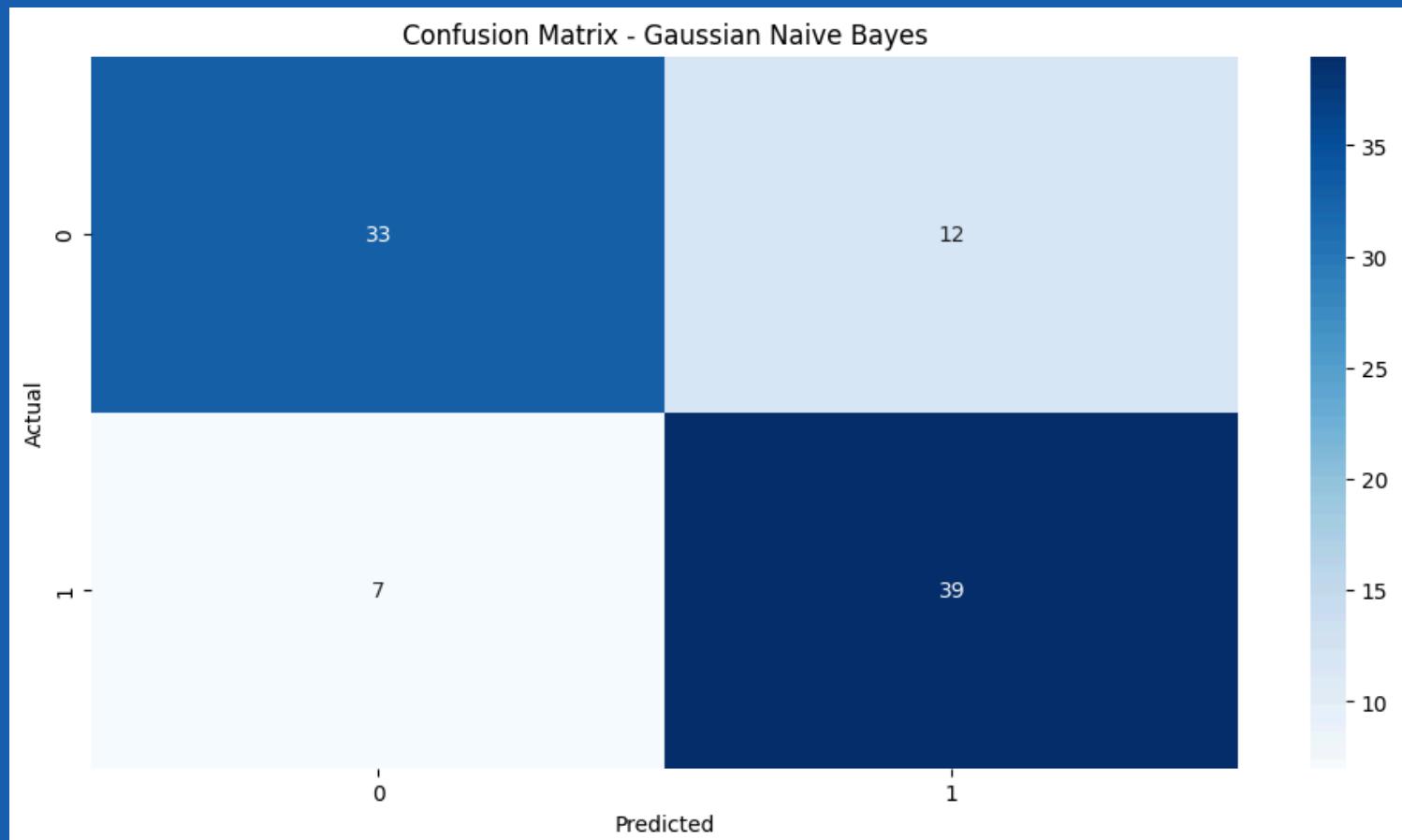
HOLDOUT 1



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 36. Untuk kasus False Positif adalah sebanyak 12, kasus True Negatif sebanyak 33 dan yang terjadi pada kasus False Negatif sebanyak 10.



HOLDOUT 2

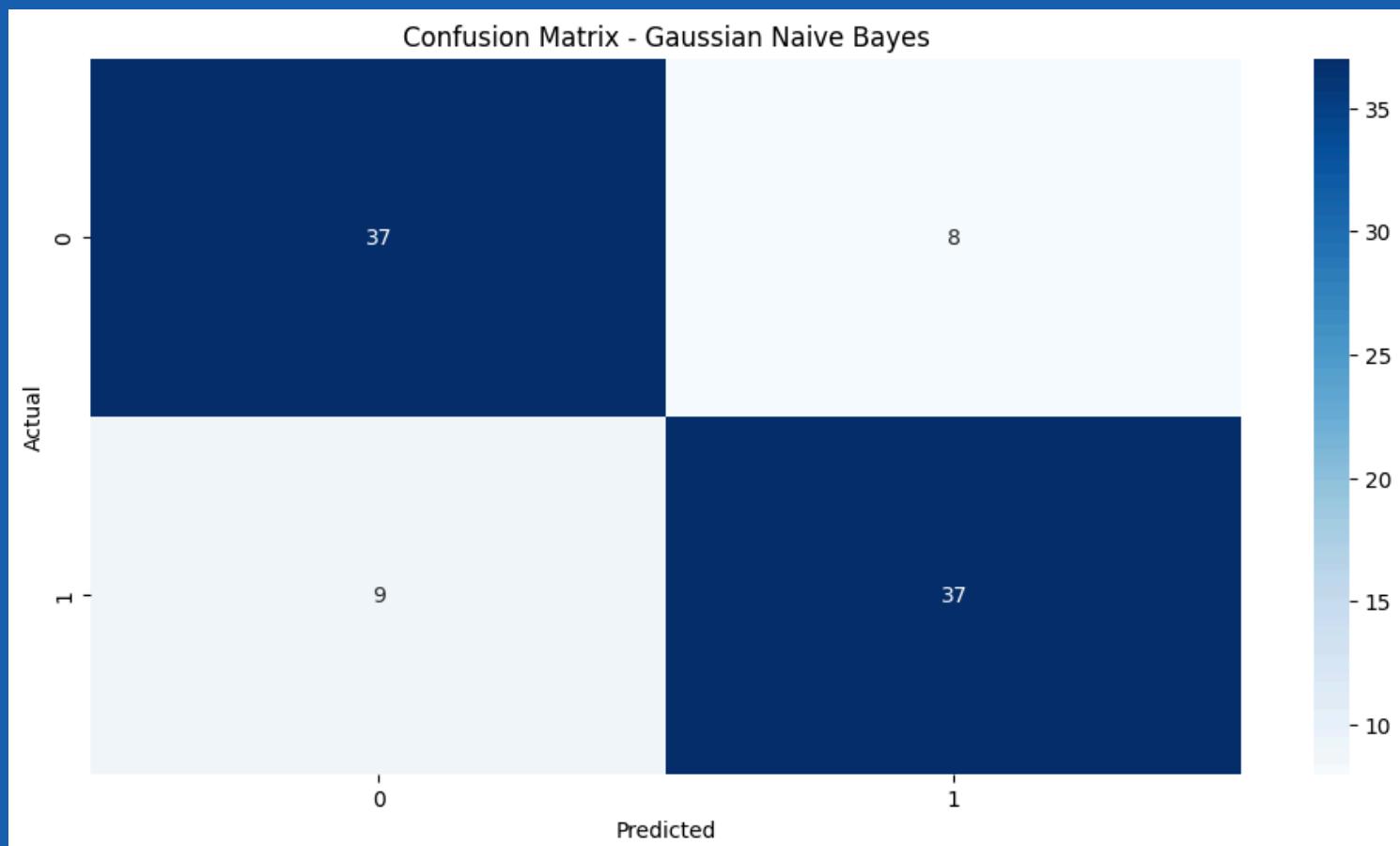


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 39. Untuk kasus False Positif adalah sebanyak 12, kasus True Negatif sebanyak 33 dan yang terjadi pada kasus False Negatif sebanyak 7.

NAIVE BAYES (HOLDOUT)



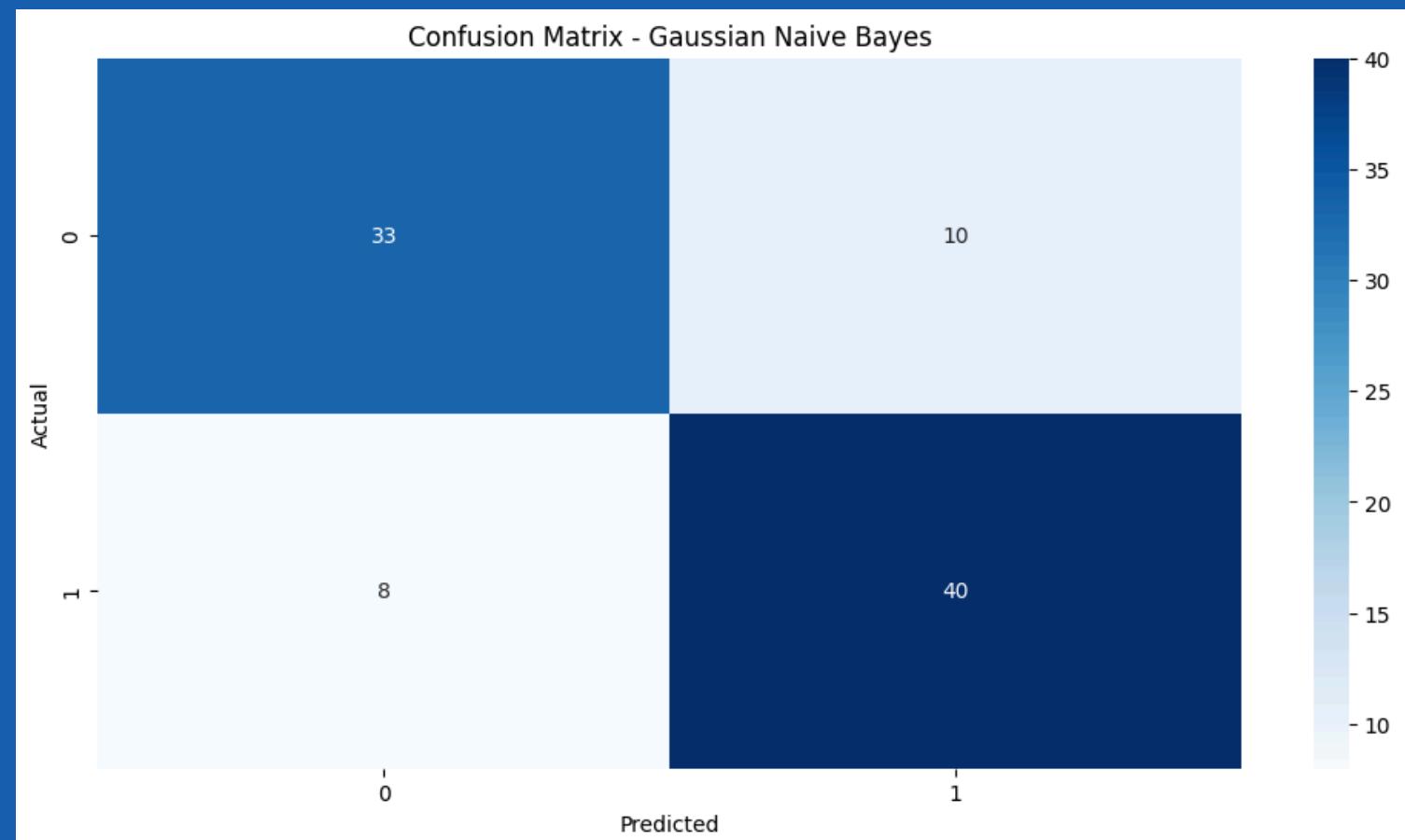
HOLDOUT 3



Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 37. Untuk kasus False Positif adalah sebanyak 8, kasus True Negatif sebanyak 37 dan yang terjadi pada kasus False Negatif sebanyak 9.



HOLDOUT 4

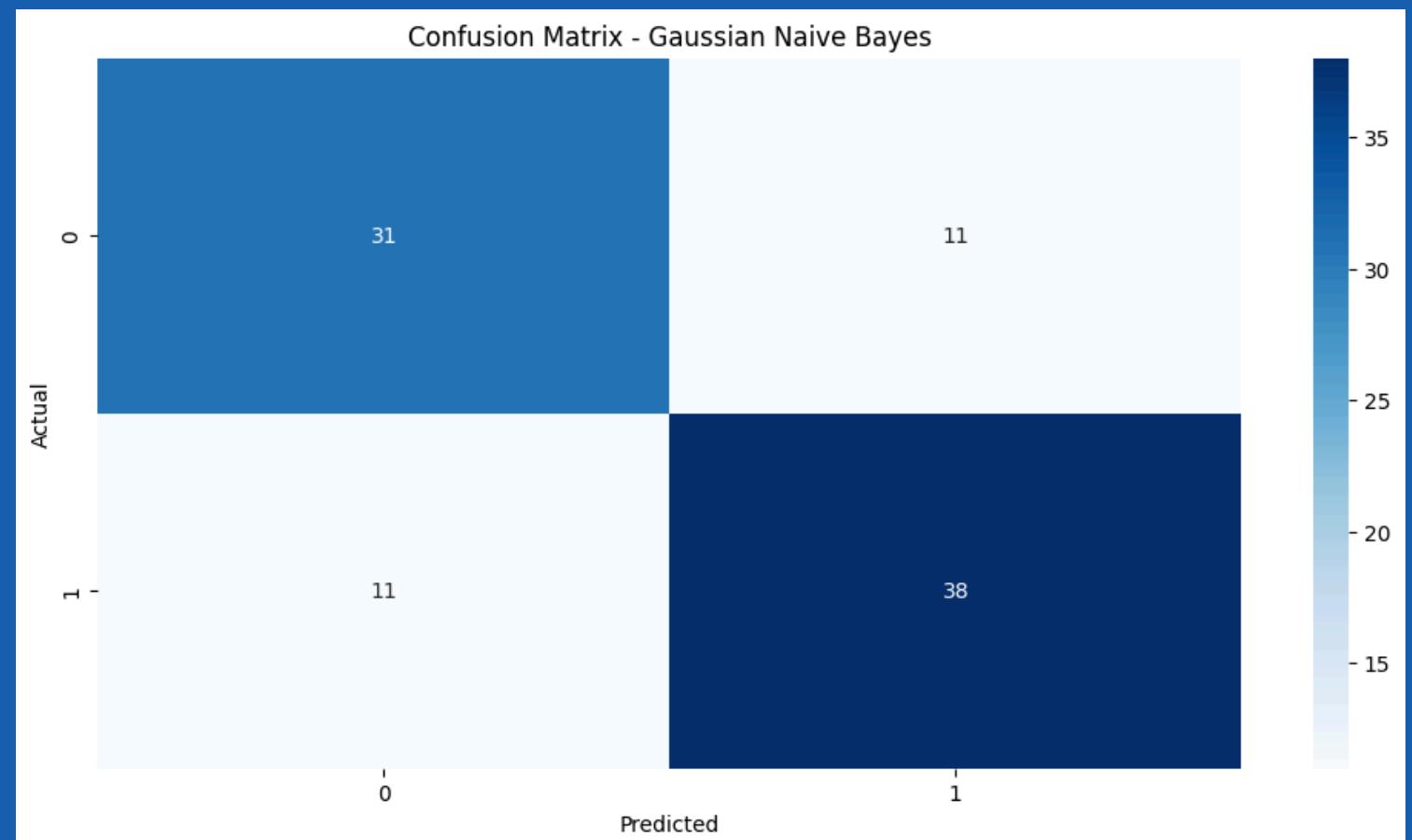


Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 40. Untuk kasus False Positif adalah sebanyak 10, kasus True Negatif sebanyak 33 dan yang terjadi pada kasus False Negatif sebanyak 8.

NAIVE BAYES (HOLDOUT)



HOLDOUT 5



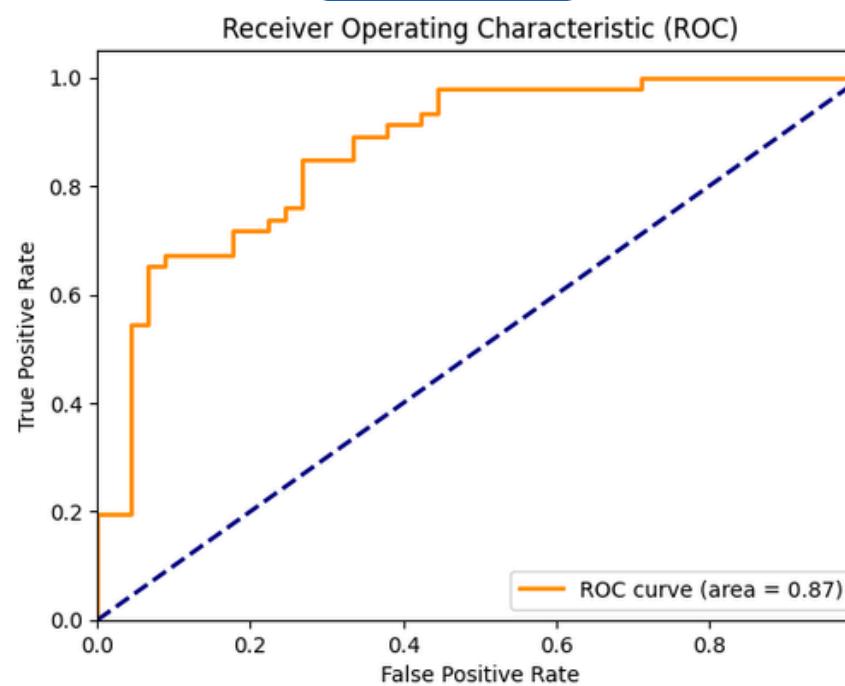
Apabila terjadi kasus True positif, maka nilai yang didapatkan adalah sekitar 38. Untuk kasus False Positif adalah sebanyak 11, kasus True Negatif sebanyak 31 dan yang terjadi pada kasus False Negatif sebanyak 11.



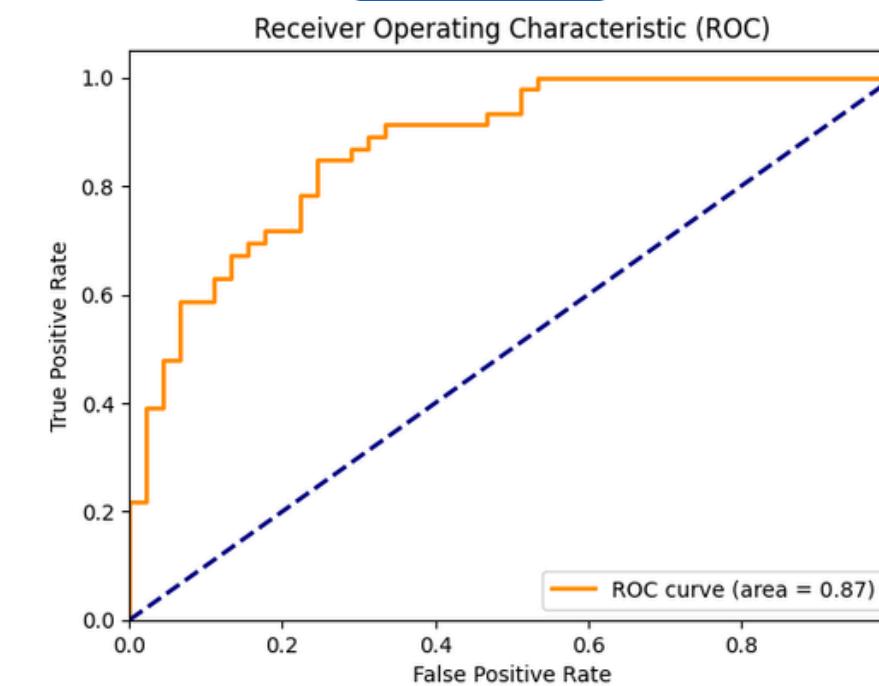
HASIL ROC-AUC

Selanjutnya akan dilakukan evaluasi untuk mengetahui model terbaik yang digunakan. Berikut ROC untuk masing-masing holdout.

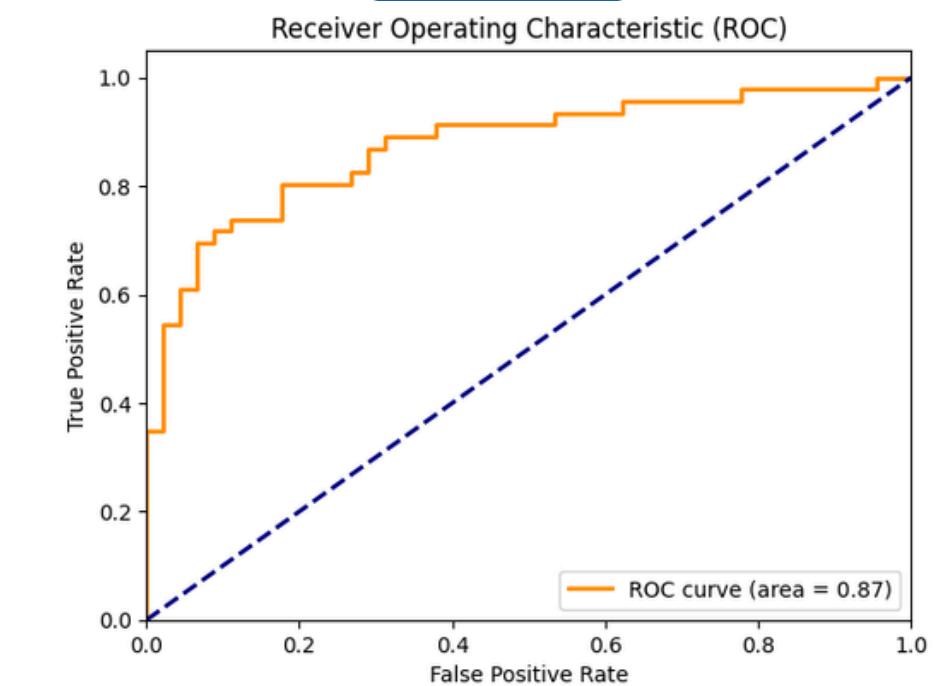
HOLDOUT 1



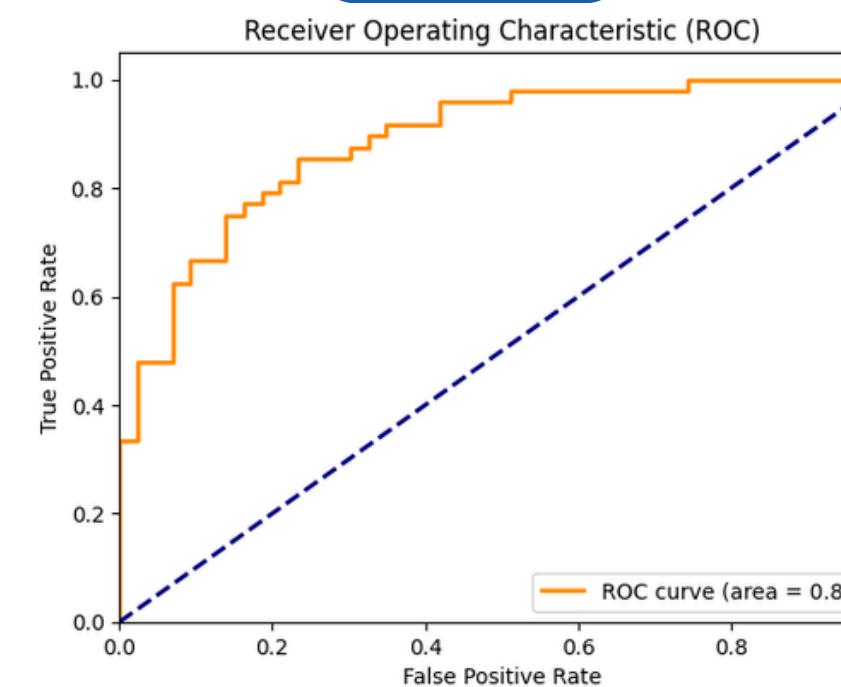
HOLDOUT 2



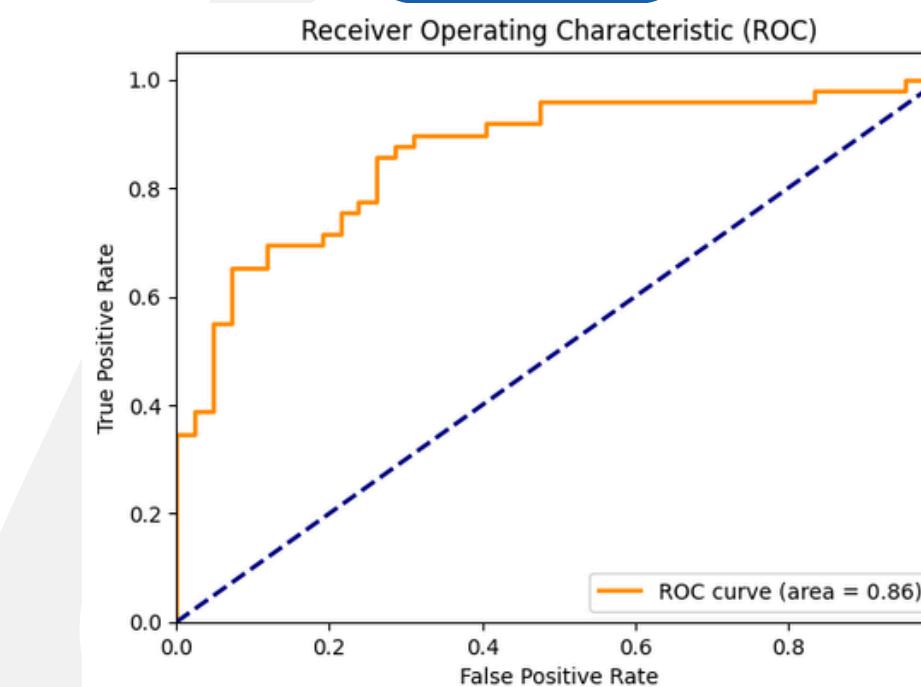
HOLDOUT 3



HOLDOUT 4



HOLDOUT 5



HASIL ANALISIS NAIVE BAYES



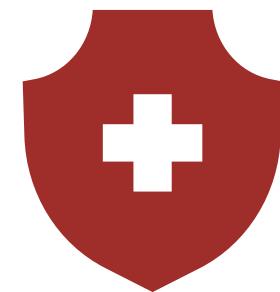
CROSS VALIDATION K-FOLD

Fold	Sensitifitas	Spesifitas	Akurasi	F-Measure	AUC
1	0,79	0,79	0,79	0,79	0,89
2	0,81	0,73	0,82	0,81	0,89
3	0,82	0,74	0,83	0,83	0,94
4	0,77	0,79	0,77	0,77	0,87
5	0,70	0,72	0,70	0,70	0,78



REPEATED HOLDOUT

Holdout	Sensitifitas	Spesifitas	Akurasi	F-Measure	AUC
1	0,76	0,73	0,76	0,76	0,87
2	0,79	0,73	0,79	0,79	0,87
3	0,81	0,82	0,81	0,81	0,87
4	0,80	0,77	0,80	0,80	0,89
5	0,76	0,74	0,76	0,76	0,86



Diperoleh bahwa AUC tertinggi adalah fold 3 yaitu 0.94, dimana mempunyai F-Measure tertinggi yaitu 0.83. Maka, dapat disimpulkan fold 3 merupakan model terbaik dalam pengklasifikasianya (good classification).

Diperoleh bahwa AUC tertinggi adalah holdout 4 yaitu 0.89, dimana mempunyai F-Measure kedua tertinggi yaitu 0.80. Maka, dapat disimpulkan holdout 4 merupakan model terbaik dalam pengklasifikasianya (good classification).



KESIMPULAN

AUC		<i>Decision Tree</i>	<i>Naive Bayes</i>
<i>K-Fold</i>	1	0,84	0,89
	2	0,70	0,89
	3	0,81	0,94
	4	0,82	0,87
	5	0,72	0,78
<i>Holdout</i>	1	0,82	0,87
	2	0,79	0,87
	3	0,73	0,87
	4	0,75	0,89
	5	0,77	0,86

1. Berdasarkan feature selection metode filter chisquare dan LDA diperoleh variabel – variabel prediktor yang berpengaruh dan dilakukan analisis, yaitu variabel CAA, CP, EXNG, SLP, SEX, CHOL, TRTBPS, AGE
2. Data splitting dengan metode K-Fold Cross Validation memberikan hasil AUC yang lebih tinggi dari pada metode Repeated Holdout
3. Metode klasifikasi terbaik untuk memodelkan klasifikasi data “serangan jantung” adalah Naive Bayes dengan nilai AUC sebesar 0,94 atau 94%



TERIMA KASIH

Presentation by
ANDREW PUTRA HARTANTO (5003211016)
RAHMANNUAJI SATUHU (5003211125)