DigitalSkola

Predicting Heart Disease Using Machine Learning

Digital Talent | SIB A
Jumat, 3 Desember 2021
Mentor : Rifyal Tumber

DATASET 11 - HEALTHCARE

DigitalSkola

Our Members

DIGITAL TALENT

Reza Sefti Damayanti

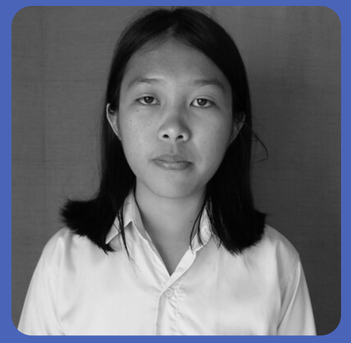
Pend. Teknologi Informasi - Universitas Bhinneka PGRI

Tiara Asa Wellana

Sistem Informasi - Universitas Indo Global Mandiri

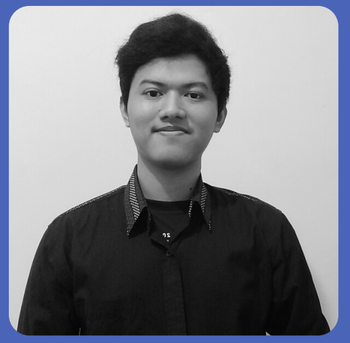
Irsandi Nur Habibie Mukmin

Teknik Industri - Universitas Brawijaya



Maria Stefani Br Simbolon

Akuntansi - Universitas HKBP Nommemsen



Taufik Aji Putra

Statistika - Universitas Diponegoro

DigitalSkola

Project Steps

01 Data Understanding

02 Identify Activities

03 Analyze the data

04 Data Pre-Processing

05 Modelling & Evaluation

Objective

Penyakit jantung merupakan **salah satu penyakit yang paling banyak diderita di dunia**. Terkadang banyak penyakit jantung yang tidak dapat terdiagnosa lebih awal, sehingga pasien mengalami penyakit jantung ketika sudah parah dengan gejala yang muncul.

Dengan **melakukan prediksi** terhadap penyakit jantung, hal ini **dapat meminimalisir** terjadinya hal serupa. Oleh karena itu, akan dilakukan pendekatan model menggunakan **Machine Learning** untuk **membantu dalam melakukan prediksi** apakah pasien terindikasi terkena penyakit jantung atau tidak.

STAGE 1

Data Understanding

Understanding the Dataset

DigitalSkola

DigitalSkola

Dataset 11 - Healthcare

1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1

300	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
301	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
302	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
303	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
304	57	0	1	130	236	0	0	174	0	0	1	1	2	0

Deskripsi Dataset

Age

Menunjukkan umur dari masing-masing individu

Sex

Menunjukkan gender dari masing-masing individu
1 = male, 0 = female

Cp

Jenis Nyeri Dada yang dialami

1 = typical angina
2 = atypical angina
3 = non — anginal pain
4 = asymptotic

Trestbps

Menunjukkan nilai tekanan darah dari masing-masing individu dalam satuan mmHG

Restecg

Hasil ECG

0 = Normal
1 = Kelainan gelombang ST-T
2= Hipertrofi Ventrikel Kiri

Thalach

Detak jantung maksimal

Chal

Nilai Cholestrol dalam satuan mg/dl

Fbs

Membandingkan nilai gula darah puasa dengan 120 mg/dl

Jika > 120mg/dl maka : 1 (true), lain : 0 (false)

DigitalSkola

HEALTHCARE DATASET

Deskripsi Dataset

Exang

Angina yang diinduksi oleh olahraga

1 = Ya
0 = Tidak

Oldpeak

Depresi ST yang disebabkan oleh latihan relatif terhadap istirahat

Slope

Latihan puncak segmen ST

1 = menanjak
2 = mendatar
3 = menurun

Thal

Thalasemia

1,3 = normal
6 = cacat tetap
7 = cacat reversibel

Target

Diagnosis penyakit jantung

0 = tidak
1 = Ya

ca

Jumlah major vessel yang diwarnai oleh floursopy

STAGE 2

IDENTIFY ACTIVITIES

Describe what the activities should be done to get what to do on Stage 1

DigitalSkola

Import Library and Dataset to Notebook

Import Library Pandas, Lalu import dataset format CSV dengan source berasal dari raw data yang telah diunggah melalui Github. Link Google Collab: https://bit.ly/GoogleCollab_DMP

```
[10] import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import numpy as np
     from scipy import stats
```

```
[11] data = pd.read_csv('https://raw.githubusercontent.com/habib238/My-DigitalSkola/main/Dataset_11%20-%20Healthcare%20.csv')
     data.head()
```

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

Aktivitas Selanjutnya yang perlu dilakukan

- Exploratory & Vizualization Data Analysis
- Data Preprocessing
- Modelling & Evaluation

STAGE 3

ANALYZE THE DATA

Exploratory & Data Visualization

DigitalSkola

Data Exploration


```
data.dtypes
```

```
age          int64
sex          int64
cp           int64
trestbps     int64
chol         int64
fbs          int64
restecg      int64
thalach      int64
exang        int64
oldpeak    float64
slope        int64
ca           int64
thal         int64
target       int64
dtype: object
```

```
data.nunique()
```

```
age          41
sex           2
cp            4
trestbps     49
chol        152
fbs           2
restecg       3
thalach      91
exang         2
oldpeak      40
slope         3
ca            5
thal          4
target        2
dtype: int64
```

```
data.shape
```

```
(303, 14)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

- Dapat dilihat bahwa data memiliki **14 features** dan **303 data input.**
- Tipe data berupa integer dan float
- Feature data memiliki nilai unik yang jumlahnya tidak sama dengan data input sehingga semua feature dapat digunakan.

Data Exploration

DigitalSkola

```
data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.37	0.68	0.97	131.62	246.26	0.15	0.53	149.65	0.33	1.04	1.40	0.73	2.31	0.54
std	9.08	0.47	1.03	17.54	51.83	0.36	0.53	22.91	0.47	1.16	0.62	1.02	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	47.50	0.00	0.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	2.00	0.00
50%	55.00	1.00	1.00	130.00	240.00	0.00	1.00	153.00	0.00	0.80	1.00	0.00	2.00	1.00
75%	61.00	1.00	2.00	140.00	274.50	0.00	1.00	166.00	1.00	1.60	2.00	1.00	3.00	1.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

Data Exploration


```
[16] #Membedakan data yang kategorik dan kontinu
     categorical_val = []
     continous_val = []
     for column in data.columns:
         if len(data[column].unique()) <= 10:
             categorical_val.append(column)
         else:
             continous_val.append(column)

[17] categorical_val

     ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal', 'target']

[ ]  continous_val

     ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
```

Data Kategorik

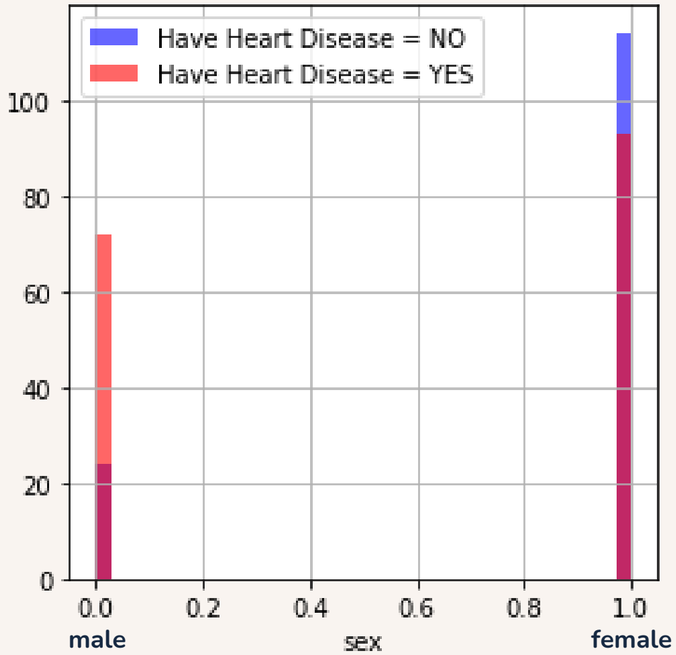
- Sex
- Fbs
- Restecg
- Exang
- Slope
- Ca
- Thal
- Target

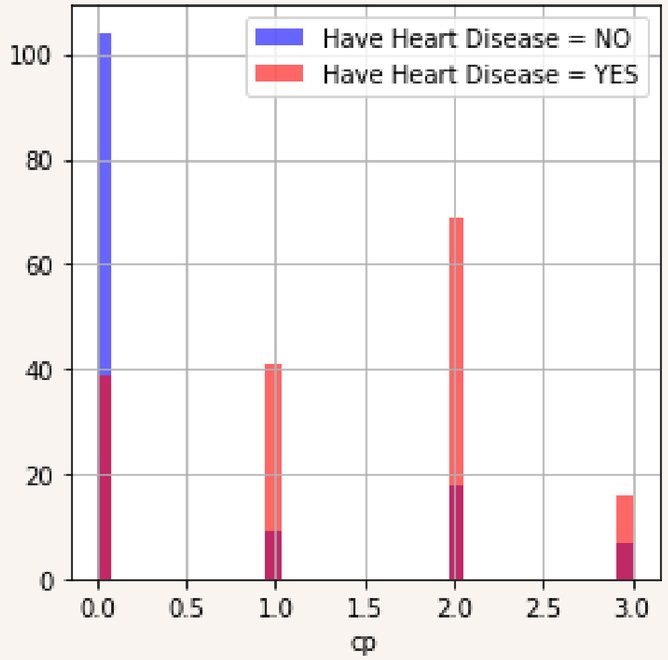
Data Numerik (Kontinu)

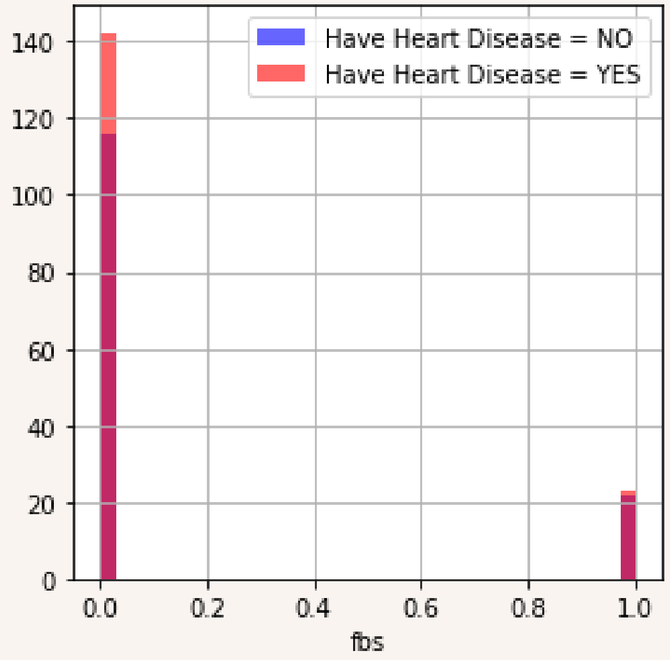
- Age
- Trestbps
- Chol
- Thalach
- Oldpeak

Data Visualization

- Bagaimana hubungan variabel-variabel kategorik ['**sex**', '**cp**', '**fbs**', '**restecg**', '**exang**', '**slope**', '**ca**', '**thal**'] dengan variabel '**target**'?

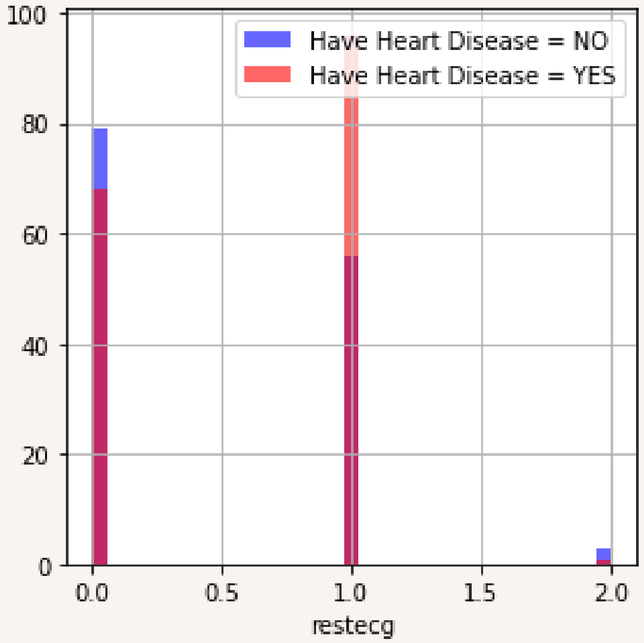


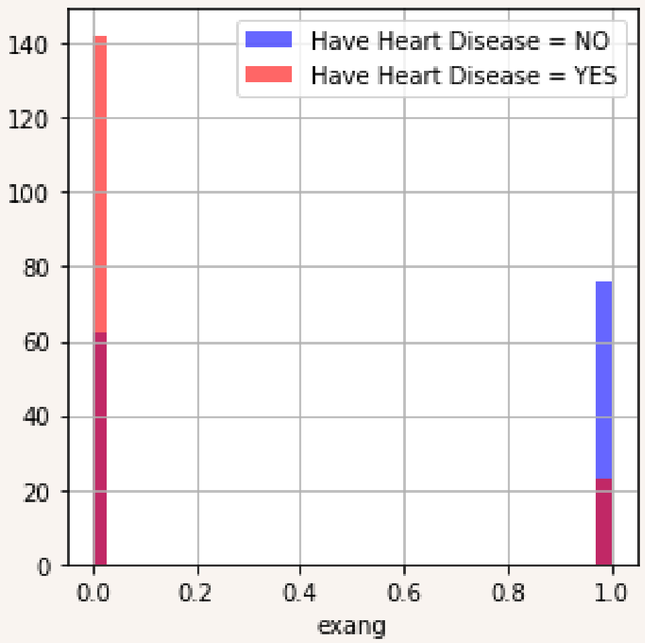


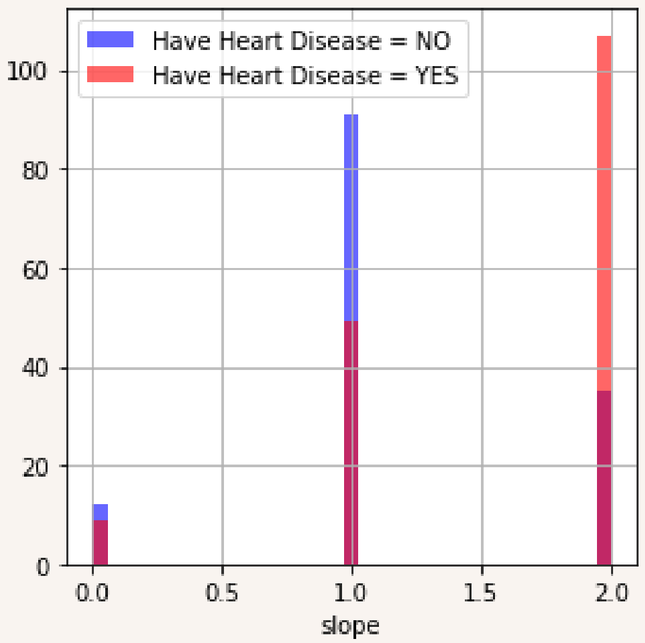


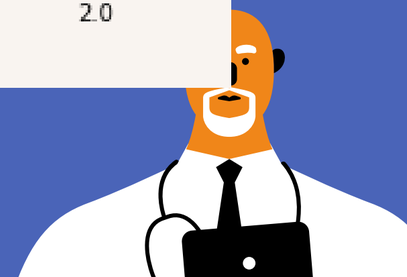
Data Visualization

- Bagaimana hubungan variabel-variabel kategorik **['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal']** dengan variabel **'target'**?



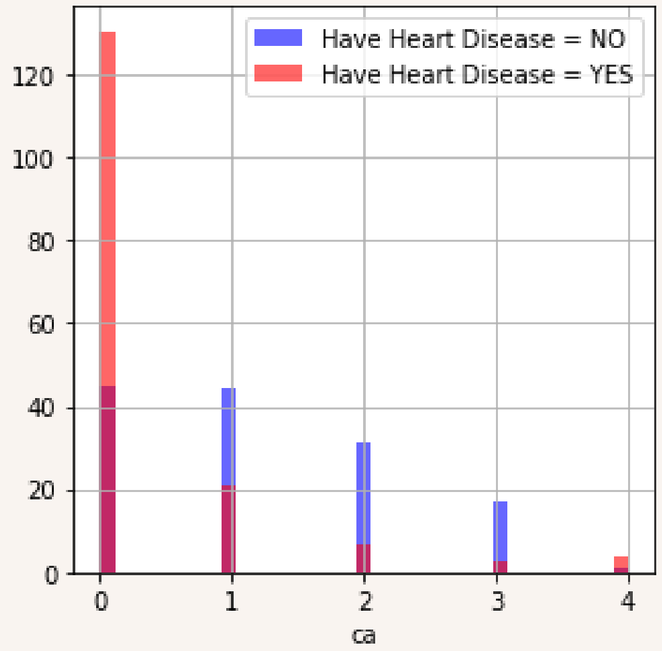


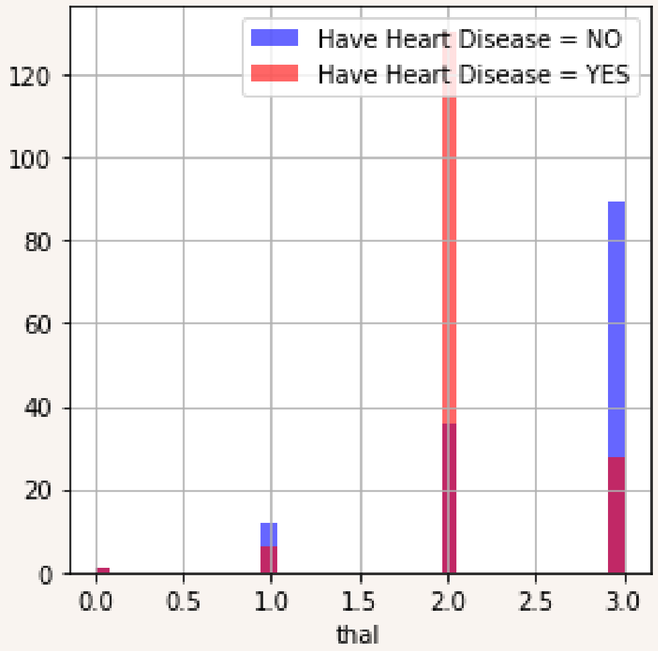




Data Visualization

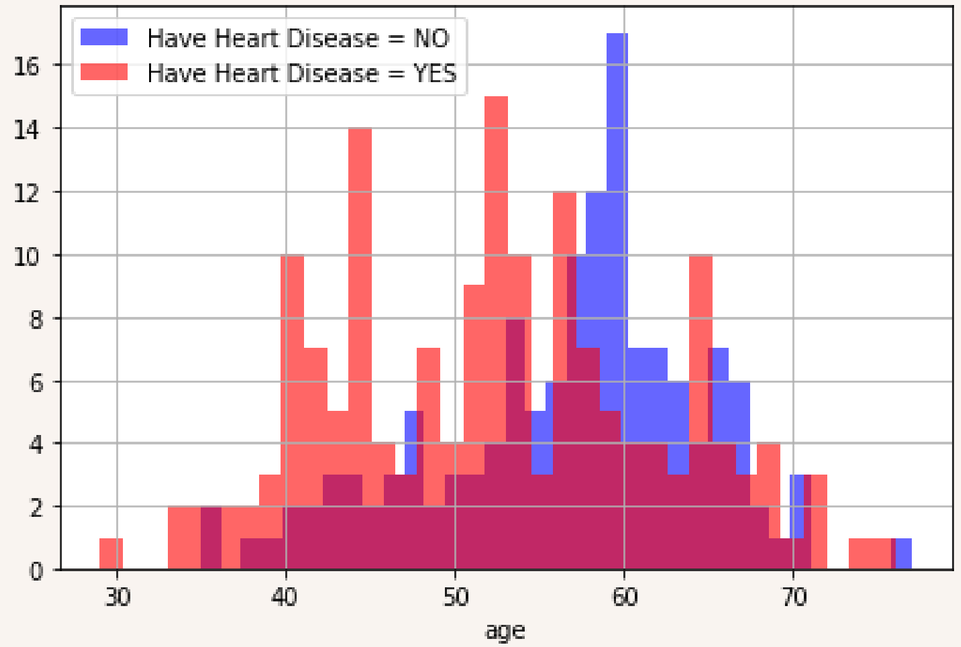
- Bagaimana hubungan variabel-variabel kategorik **['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal']** dengan variabel **'target'**?

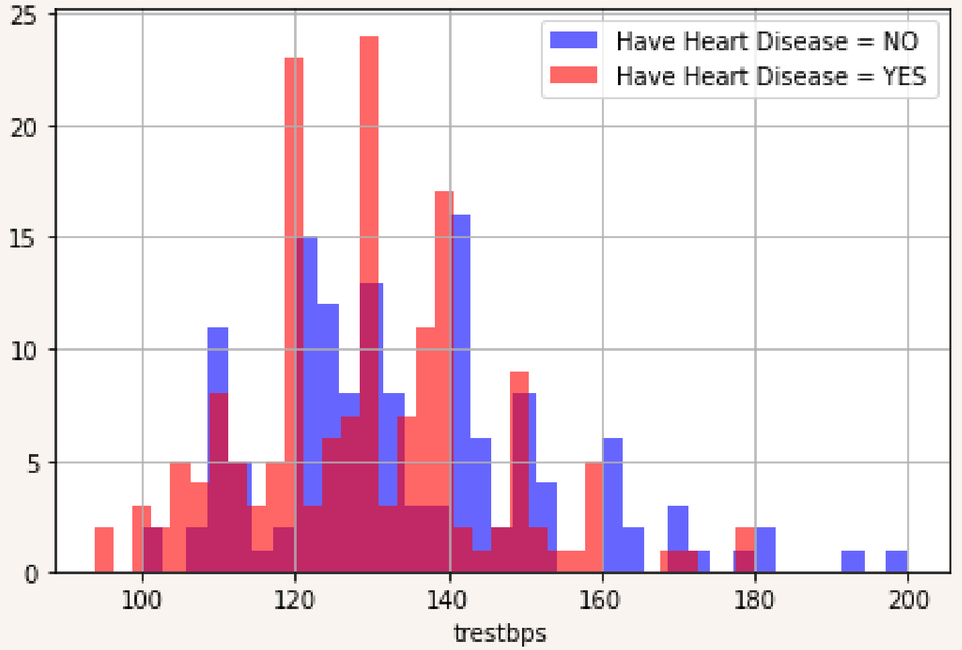


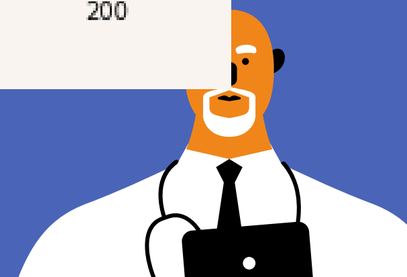


Data Visualization

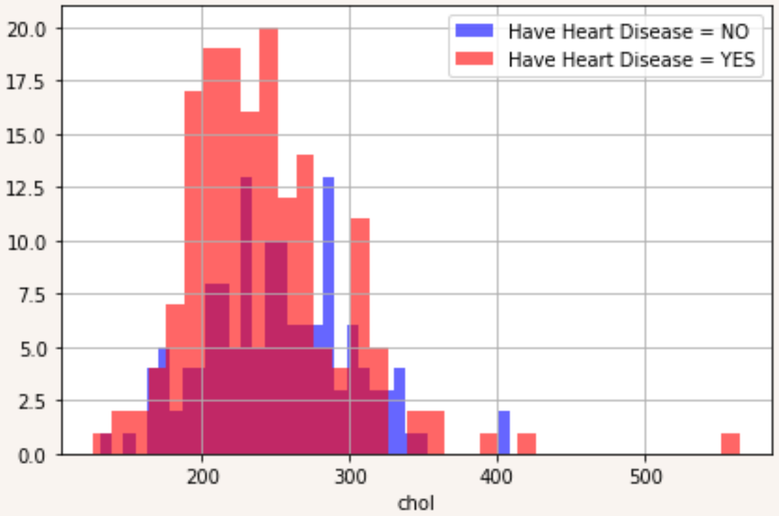
- Bagaimana hubungan variabel-variabel kontinu **['age', 'trestbps', 'chol', 'thalach', 'oldpeak']** dengan variabel **'target'**?

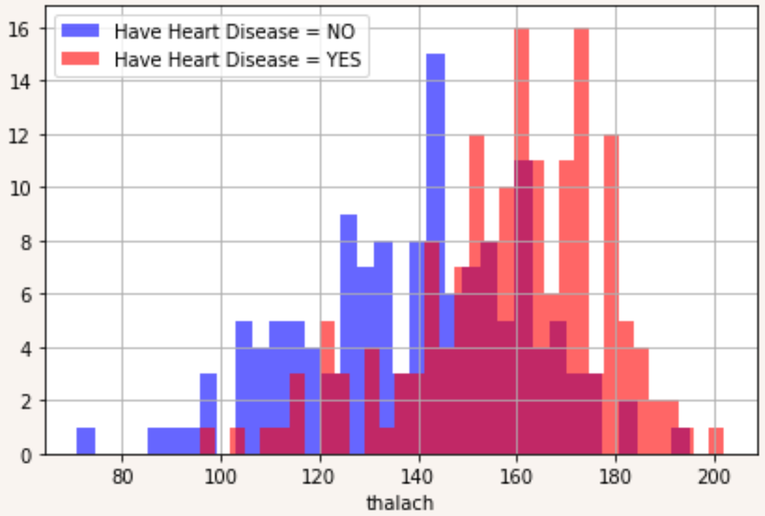


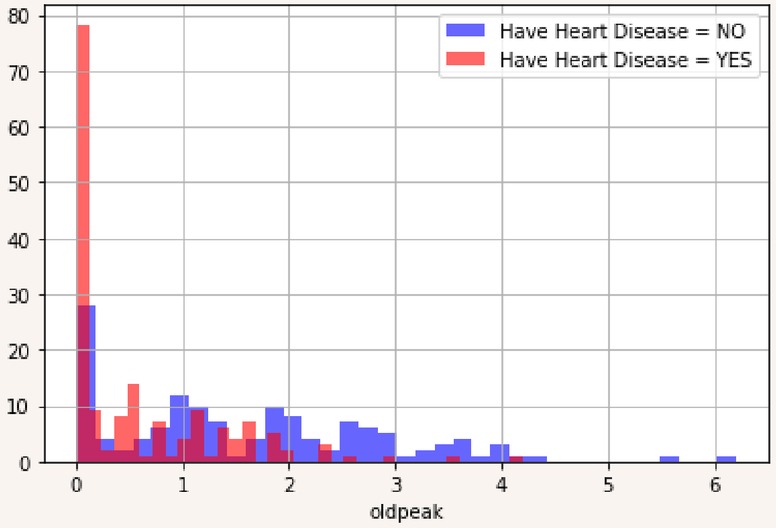




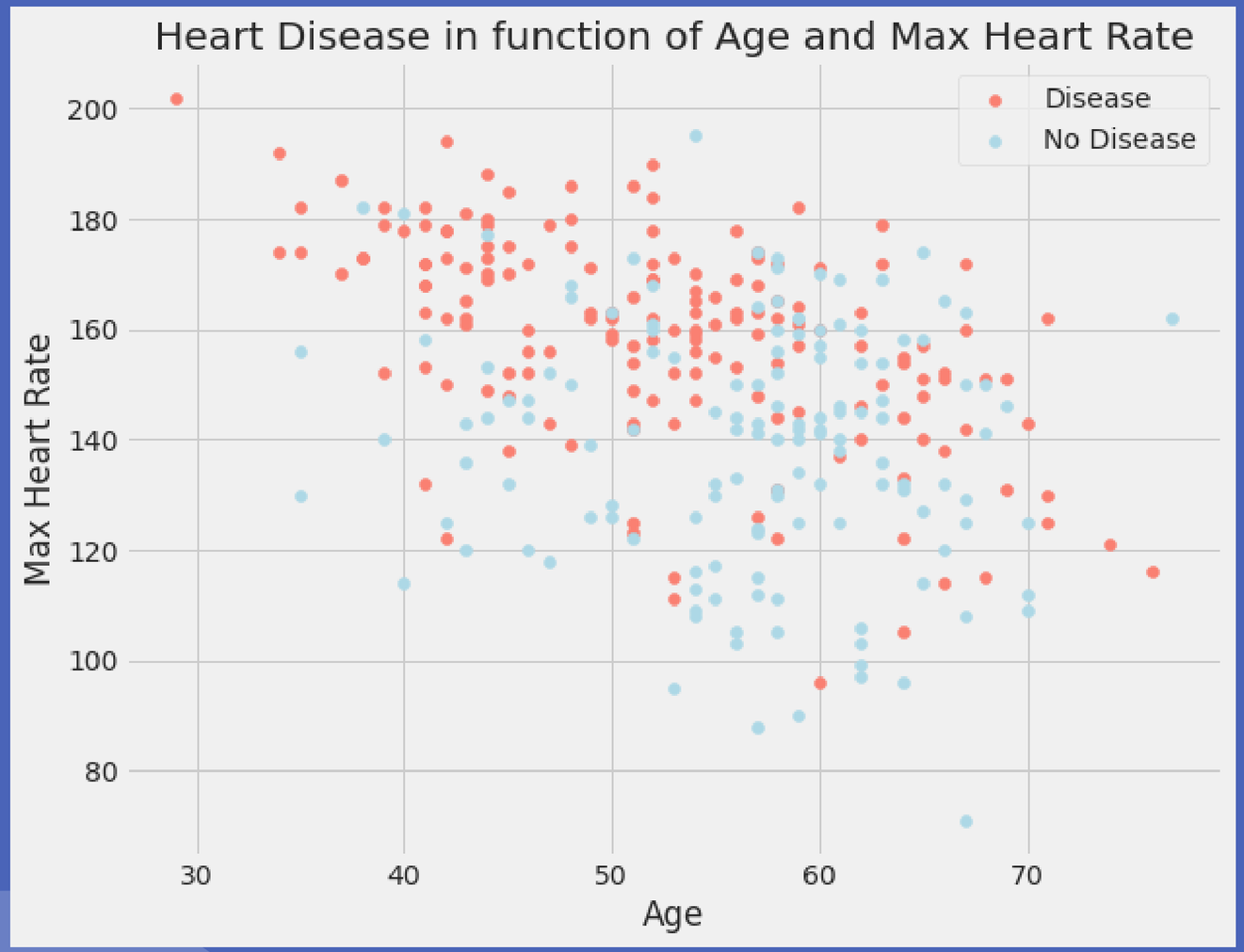
Data Visualization







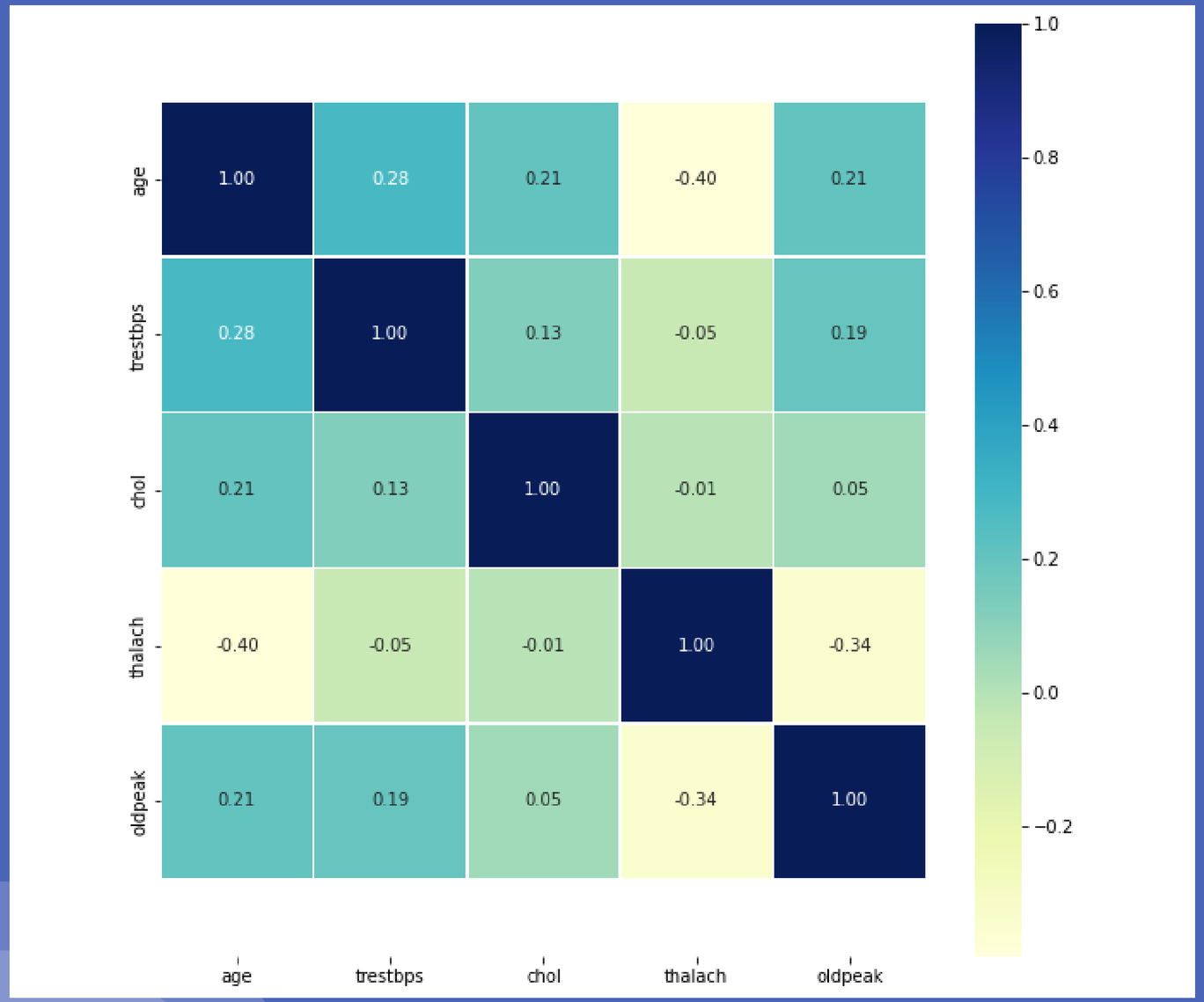
Data Visualization



Dapat dilihat bahwa penyakit jantung (Heart Disease) banyak diderita oleh orang yang berusia rentang 30-70+ tahun.

Secara visual dapat dilihat juga bahwa rentang usia 30-60 tahun lebih rentan terkena penyakit jantung disertai dengan max heart rate (thalach) yang tinggi.

Data Visualization



Correlation (Heatmap)

(Continous Feature)

Semakin besar nilai korelasi/warna semakin gelap

=

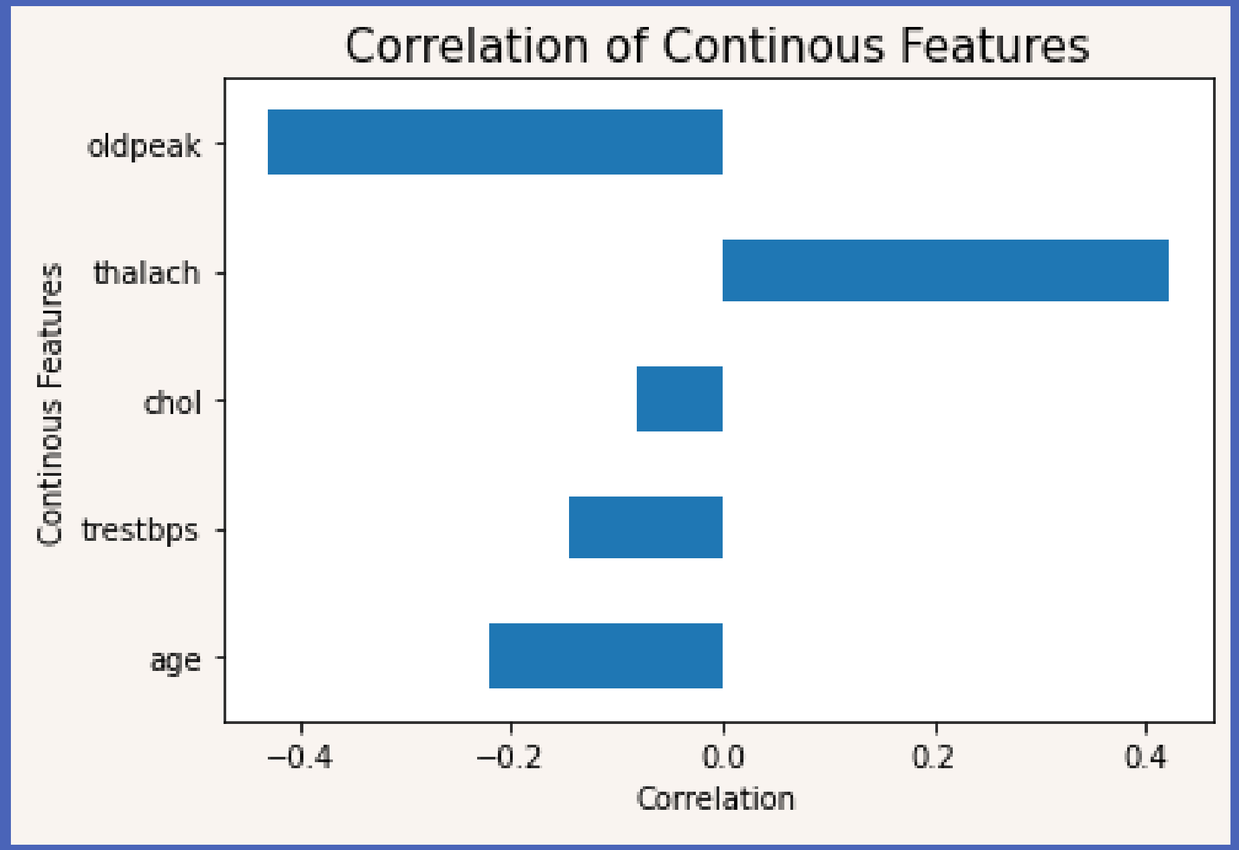
Semakin menuju **Korelasi Positif (+)**

Semakin kecil nilai korelasi / warna semakin terang

=

Semakin Menuju **Korelasi Negatif (-)**

Data Visualization



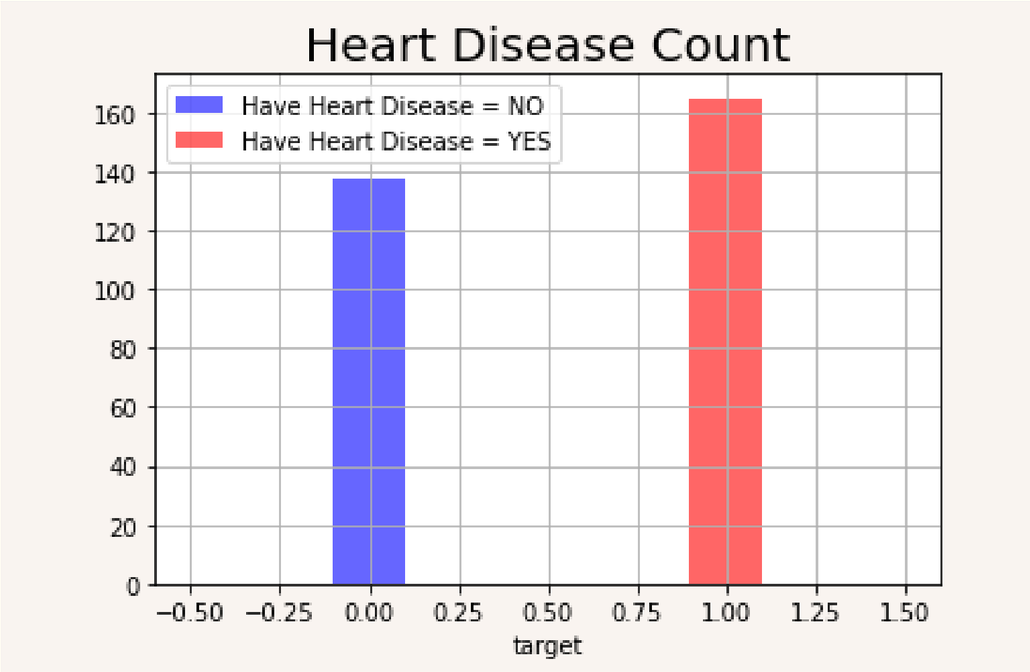
Correlation (Bar)

(Continous Feature)

Bagaimana korelasi antar variabel-variabel numerik dengan variabel "target"?

- **trestbps** (tekanan darah) and **chol** (kolestrol) adalah feature dengan korelasi terkecil.
- Semua variabel lainnya memiliki korelasi yang cukup.

Data Visualization




```
data.target.value_counts()
```

```
1    165
0    138
Name: target, dtype: int64
```

Data Target

Dapat disimpulkan bahwa:

person **with** heart disease = 165
person **without** heart disease = 138

STAGE 4

DATA PREPROCESSING

Data akan di proses sedemikian rupa sehingga menghasilkan data yang bersih dan siap untuk dilanjutkan ke tahap selanjutnya.

DigitalSkola

Missing Value & Duplicated Data

DigitalSkola

```
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
```

```
data.duplicated().sum()
```

```
1
```

```
data[data.duplicated()]
```

	age	sex	cp	trestbps	chol	fbs
164	38	1	2	138	175	0

```
data = data.drop_duplicates(subset=None, keep='first', inplace=False)
```

```
data.duplicated().sum()
```

```
0
```

- Dataset tidak memiliki missing value.
- Dataset memiliki data duplikat, tepatnya di index 164. Sehingga dilakukan penghapusan data duplikat.

Penambahan Variabel Dummy


```
#Convert Catgorical Variables into Dummy

categorical_val.remove('target')
dataset = pd.get_dummies(data, columns = categorical_val)
```

Akan ada penambahan variabel dummy pada feature yang bersifat kategorikal.

```
[15] print(data.columns)
     print(dataset.columns)

     Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
            'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
           dtype='object')
     Index(['age', 'trestbps', 'chol', 'thalach', 'oldpeak', 'target', 'sex_0',
            'sex_1', 'cp_0', 'cp_1', 'cp_2', 'cp_3', 'fbs_0', 'fbs_1', 'restecg_0',
            'restecg_1', 'restecg_2', 'exang_0', 'exang_1', 'slope_0', 'slope_1',
            'slope_2', 'ca_0', 'ca_1', 'ca_2', 'ca_3', 'ca_4', 'thal_0', 'thal_1',
            'thal_2', 'thal_3'],
           dtype='object')
```

Feature Scaling

Feature scaling menggunakan standar scaler diterapkan pada feature selain kategorikal kecuali feature target, dengan cara menghapus rata-rata dan melakukan scaling menjadi range tertentu.

	age	trestbps	chol	thalach	oldpeak
0	63	145	233	150	2.3
1	37	130	250	187	3.5
2	41	130	204	172	1.4
3	56	120	236	178	0.8
4	57	120	354	163	0.6

Sebelum Feature Scaling

	age	trestbps	chol	thalach	oldpeak
0	0.952197	0.763956	-0.256334	0.015443	1.087338
1	-1.915313	-0.092738	0.072199	1.633471	2.122573
2	-1.474158	-0.092738	-0.816773	0.977514	0.310912
3	0.180175	-0.663867	-0.198357	1.239897	-0.206705
4	0.290464	-0.663867	2.082050	0.583939	-0.379244

Sesudah Feature Scaling

Train-test split


```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
y_train.value_counts()
```

```
1    115
0     96
Name: target, dtype: int64
```

Mengatasi Imbalance (SMOTE)

```
from imblearn.over_sampling import SMOTE

sm = SMOTE()
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)
y_train_sm.value_counts()
```

```
1    115
0    115
Name: target, dtype: int64
```

STAGE 5

DATA MODELLING

Modelling dilakukan untuk membuat dan menentukan model yang dapat membantu dalam melakukan prediksi penyakit jantung.

DigitalSkola

Logistic Regression

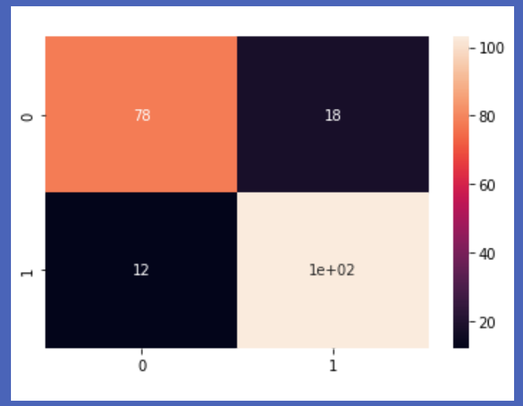
Logistic Regression adalah sebuah algoritma klasifikasi untuk mencari hubungan antara fitur (input) diskrit/kontinu dengan probabilitas hasil output diskrit tertentu.

Logistic Regression

Before SMOTE

Metrics Evaluation (Train)

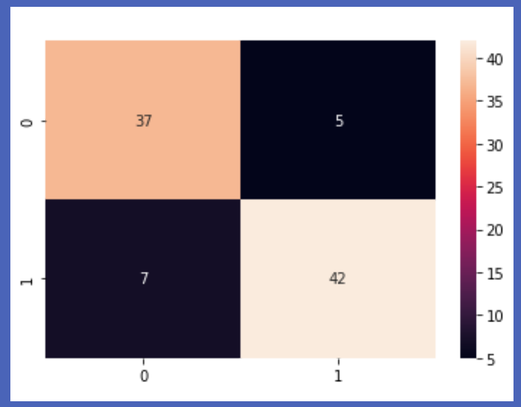
Accuracy	85.78%
Precision	85.12%
Recall	89.56%
F-1 Score	87.28%




```
Confusion Matrix:
 [[ 78  18]
 [ 12 103]]
```

Metrics Evaluation (Test)

Accuracy	86.81%
Precision	89.36%
Recall	85.71%
F-1 Score	87.50%



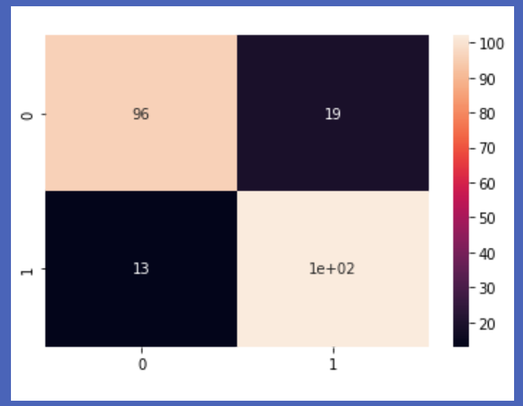

```
Confusion Matrix:
 [[37  5]
 [ 7 42]]
```

Logistic Regression

After SMOTE

Metrics Evaluation (Train)

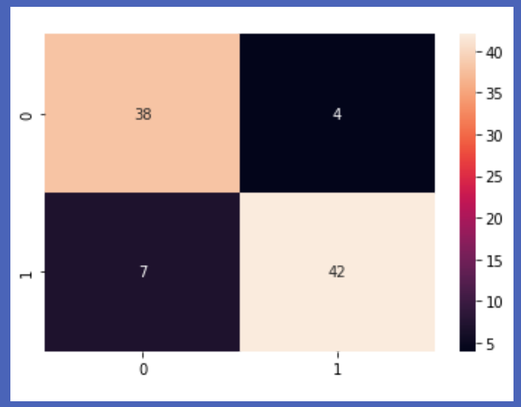
Accuracy	86.09%
Precision	84.29%
Recall	88.69%
F-1 Score	86.44%




```
Confusion Matrix:
 [[ 96  19]
 [ 13 102]]
```

Metrics Evaluation (Test)

Accuracy	87.91%
Precision	91.30%
Recall	85.71%
F-1 Score	88.42%




```
Confusion Matrix:
 [[38  4]
 [ 7 42]]
```

```
auc_lr
0.9275996112730807
```

Support Vector Machine

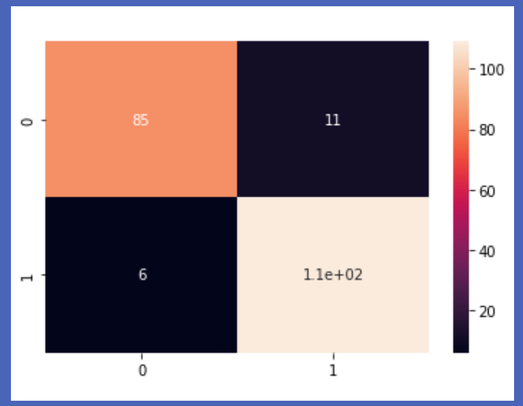
Support Vector Machines (SVM) membuat model yang menetapkan titik data baru ke salah satu kategori yang diberikan. Dengan demikian, ini dapat dipandang sebagai pengklasifikasi linear biner non-probabilistik.

Support Vector Machine

Before SMOTE

Metrics Evaluation (Train)

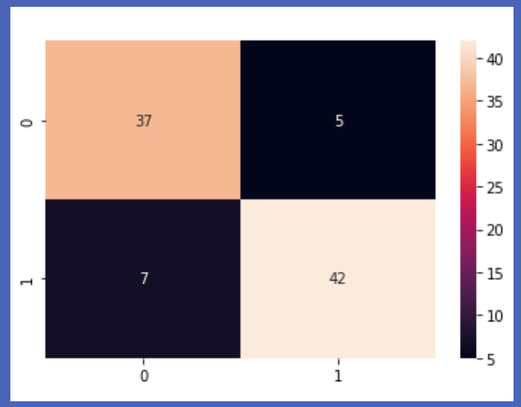
Accuracy	91.94%
Precision	90.83%
Recall	94.78%
F-1 Score	92.76%




```
Confusion Matrix:
 [[ 85  11]
 [  6 109]]
```

Metrics Evaluation (Test)

Accuracy	86.81%
Precision	89.36%
Recall	85.71%
F-1 Score	87.50%



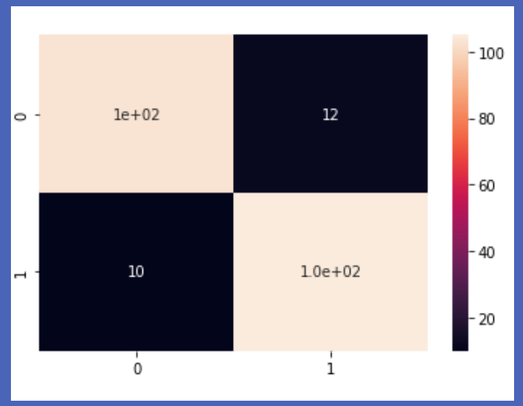

```
Confusion Matrix:
 [[37  5]
 [ 7 42]]
```

Support Vector Machine

After SMOTE

Metrics Evaluation (Train)

Accuracy	90.43%
Precision	89.74%
Recall	91.30%
F-1 Score	90.51%



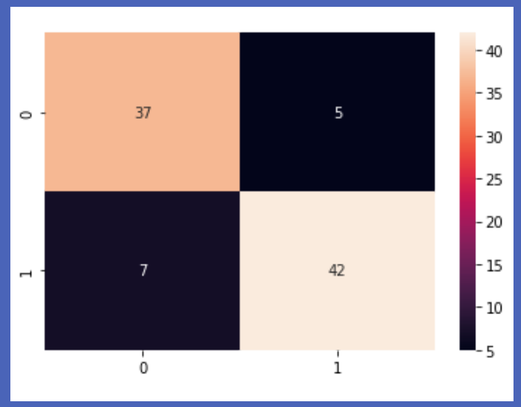

```
Confusion Matrix:
 [[103  12]
 [ 10 105]]
```

```
auc_svm
```

```
0.9217687074829931
```

Metrics Evaluation (Test)

Accuracy	86.81%
Precision	89.36%
Recall	85.71%
F-1 Score	87.50%




```
Confusion Matrix:
 [[37  5]
 [ 7 42]]
```

Decision Tree Classifier

Decision Tree Classifier membagi data menjadi himpunan bagian berdasarkan variabel inputnya. Algoritma ini merupakan jenis diagram alir yang membantu dalam proses pengambilan keputusan.

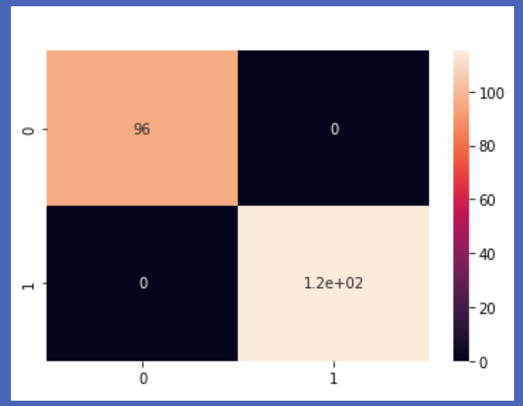
Decision Tree ini menjadi alat pendukung keputusan yang menggunakan grafik atau model seperti pohon. Grafik ini terdiri dari jumlah minimum ya/tidak pertanyaan dari sebuah pertanyaan, untuk menilai masing-masing probabilitasnya.

Decision Tree Classifier

Before SMOTE

Metrics Evaluation (Train)

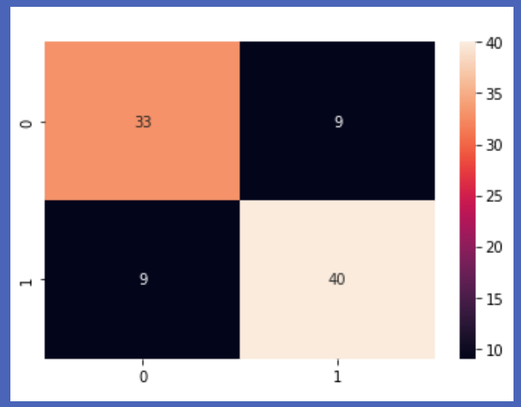
Accuracy	100.00%
Precision	100.00%
Recall	100.00%
F-1 Score	100.00%




```
Confusion Matrix:
 [[ 96   0]
 [  0 115]]
```

Metrics Evaluation (Test)

Accuracy	80.22%
Precision	81.63%
Recall	81.63%
F-1 Score	81.63%



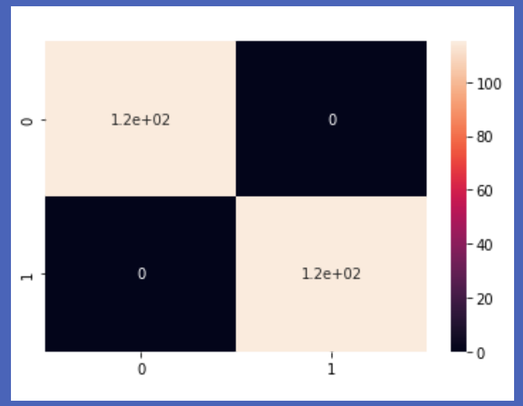

```
Confusion Matrix:
 [[33  9]
 [15 34]]
```

Decision Tree Classifier

After SMOTE

Metrics Evaluation (Train)

Accuracy	100.00%
Precision	100.00%
Recall	100.00%
F-1 Score	100.00%



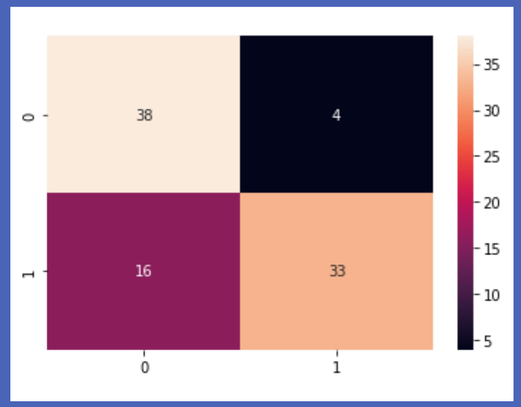

```
Confusion Matrix:
 [[115   0]
 [  0 115]]
```

```
auc_dt

0.7397959183673468
```

Metrics Evaluation (Test)

Accuracy	78.02%
Precision	89.18%
Recall	67.34%
F-1 Score	76.74%



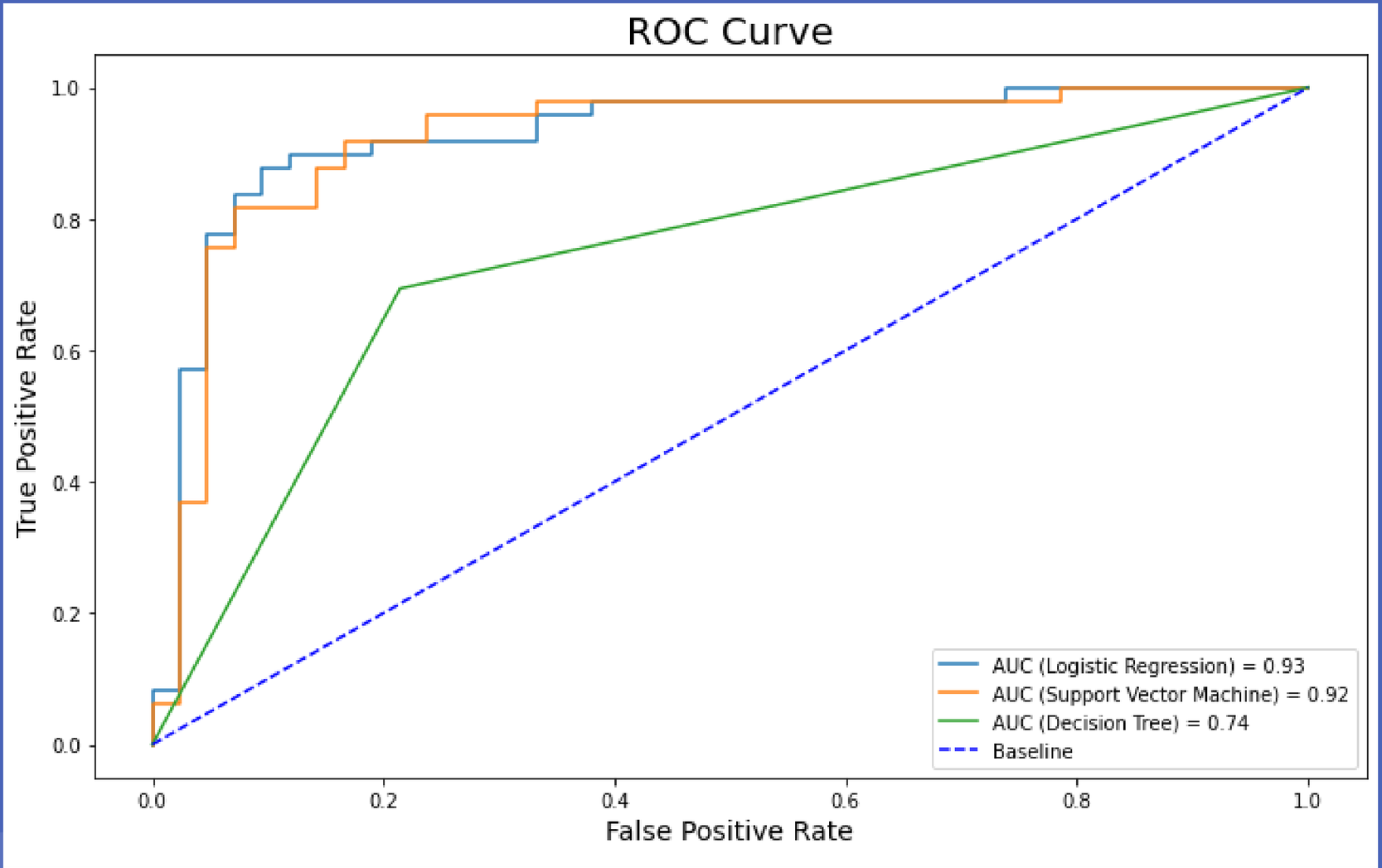

```
Confusion Matrix:
 [[38  4]
 [16 33]]
```

Summary #1

Model	Metrics	Before SMOTE		After SMOTE	
		Train	Test	Train	Test
Logistics Regression	Accuracy	85.78%	86.81%	**86.09%**	**87.91%**
	Precision	85.12%	89.36%	84.29%	91.30%
	Recall	89.56%	86.71%	88.69%	85.71%
	F-1 Score	87.28%	87.5%	86.44%	88.42%
SVM	Accuracy	91.94%	86.81%	**90.43%**	**86.81%**
	Precision	90.83%	89.26%	89.74%	89.36%
	Recall	94.78%	85.71%	91.30%	85.71%
	F-1 Score	92.76%	87.50%	90.51%	87.50%
Decision Tree	Accuracy	100.00%	80.22%	**100.00%**	**78.02%**
	Precision	100.00%	81.63%	100.00%	89.18%
	Recall	100.00%	81.63%	100.00%	67.34%
	F-1 Score	100.00%	8.63%	100.00%	76.74%

Model Terbaik

Summary #2



AUC (Logistic Regression) = 0.93
AUC (Support Vector Machine) = 0.92
AUC (Decision Tree) = 0.74

Hyperparameter Tuning

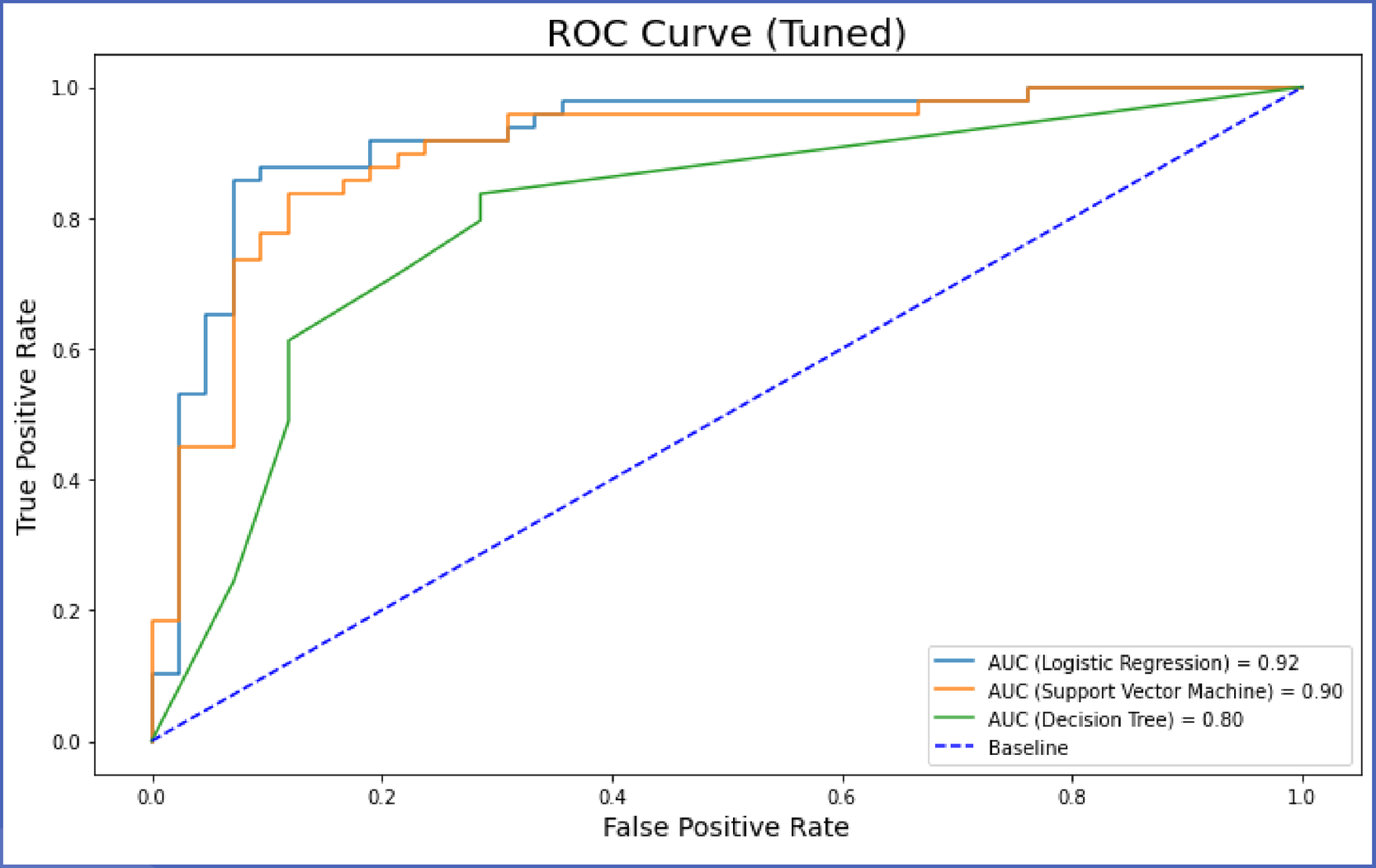
Model	Training Accuracy %	Testing Accuracy %
Tuned Logistic Regression (SMOTE)	88.26	89.01
Tuned Support Vectore Machine (SMOTE)	97.83	83.52
Tuned Decision Tree Classifier (SMOTE)	93.04	76.92

auc_lr
0.9237123420796891

auc_svm
0.9037900874635569

auc_dt
0.7993197278911565

#Summary 3



AUC (Logistic Regression) = 0.92
AUC (Support Vector Machine) = 0.90
AUC (Decision Tree) = 0.80

Rekomendasi

Model terbaik adalah: Tuned Logistic Regression (SMOTE)

Model	Training Accuracy %	Testing Accuracy %
Tuned Logistic Regression (SMOTE)	88.26	89.01

Dengan menggunakan model terbaik, diharapkan bisa membantu dokter atau tim ahli kesehatan dalam memprediksi/mengklasifikasikan pasien berdasarkan feature/kriteria yang memungkinkan apakah pasien terindikasi mengalami penyakit jantung atau tidak, dengan lebih cepat dan dengan kemungkinan klasifikasi tepat sasaran yang besar.

Thank You!