



# **SISTEM PREDIKSI PENYAKIT JANTUNG BMENGGUNAKAN METODE NAIVE BAYES**

**Nama : Fitrinur Indriyana Salsabila**

**NPM : 14210009**

**Prodi : Sistem Informasi**

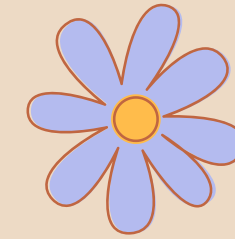
**Dimsyar M Al Hafiz<sup>1</sup> , Khoirul Amaly<sup>1</sup> , Javen Jonathan<sup>1</sup> , M. Teranggono Rachmatullah<sup>1</sup> , Rosidi<sup>1</sup>  
Teknik Elektro, Fakultas Teknik Universitas Sriwijaya Palembang, Indonesia**

**Penulis korespondensi: [khoirulamaly@gmail.com](mailto:khoirulamaly@gmail.com)**

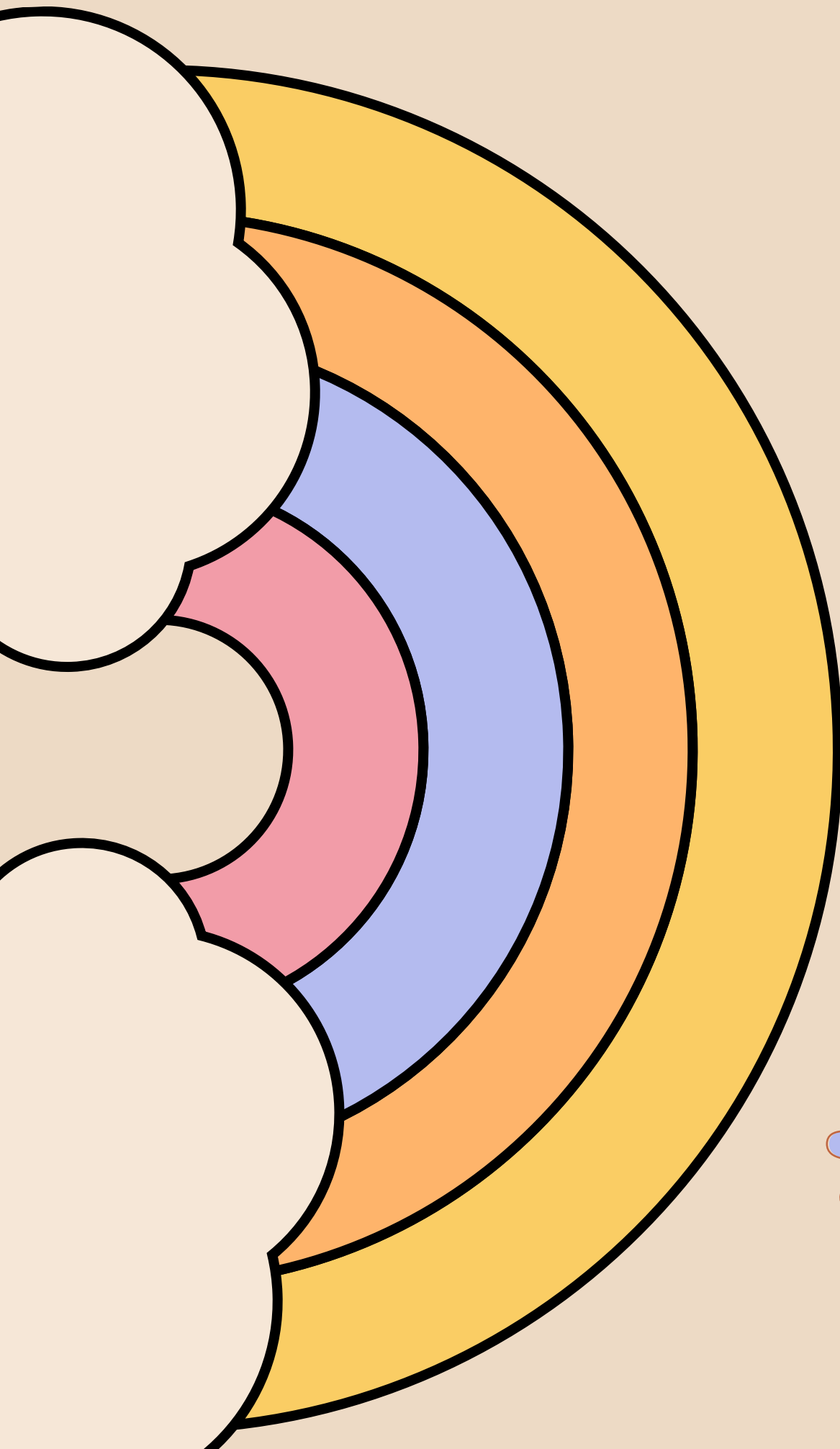
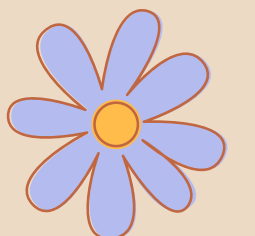
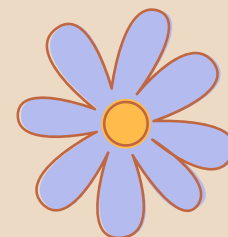
**eISSN 2716-4063**



# PENDAHULUAN



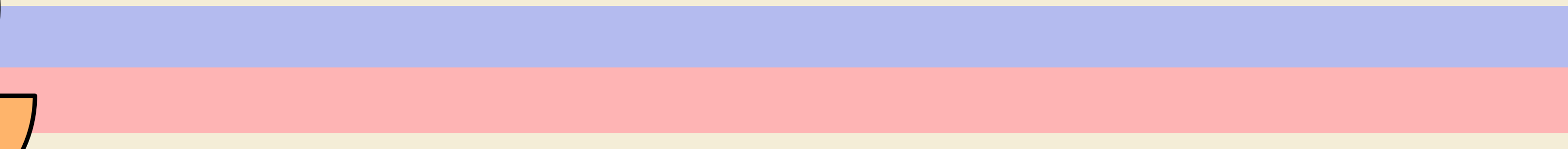
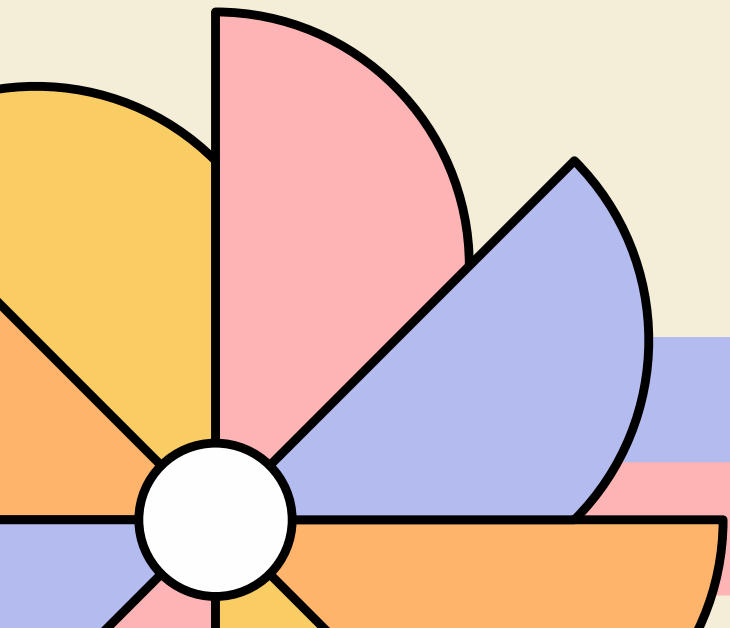
Pada tahun 2016, Indonesia mengalami total kematian sebesar 1.863.000 jiwa, dengan 73% disebabkan oleh penyakit tidak menular. Penyakit jantung koroner merupakan penyebab kematian nomor 1 sebanyak 35%. Oleh sebab itu diagnosis penyakit jantung yang efisien, akurat, dan dini sangat penting untuk mengambil langkah pencegahan yang tepat guna mencegah kematian





# TUJUAN PENELITIAN

Menyajikan data tentang tingkat kematian akibat penyakit tidak menular di Indonesia, fokus pada penyakit jantung, serta penekanan pada pentingnya diagnosis dini dan peran data mining dalam memprediksi penyakit jantung untuk pengembangan solusi pencegahan dan pengobatan.



# METODELOGI PENELITIAN

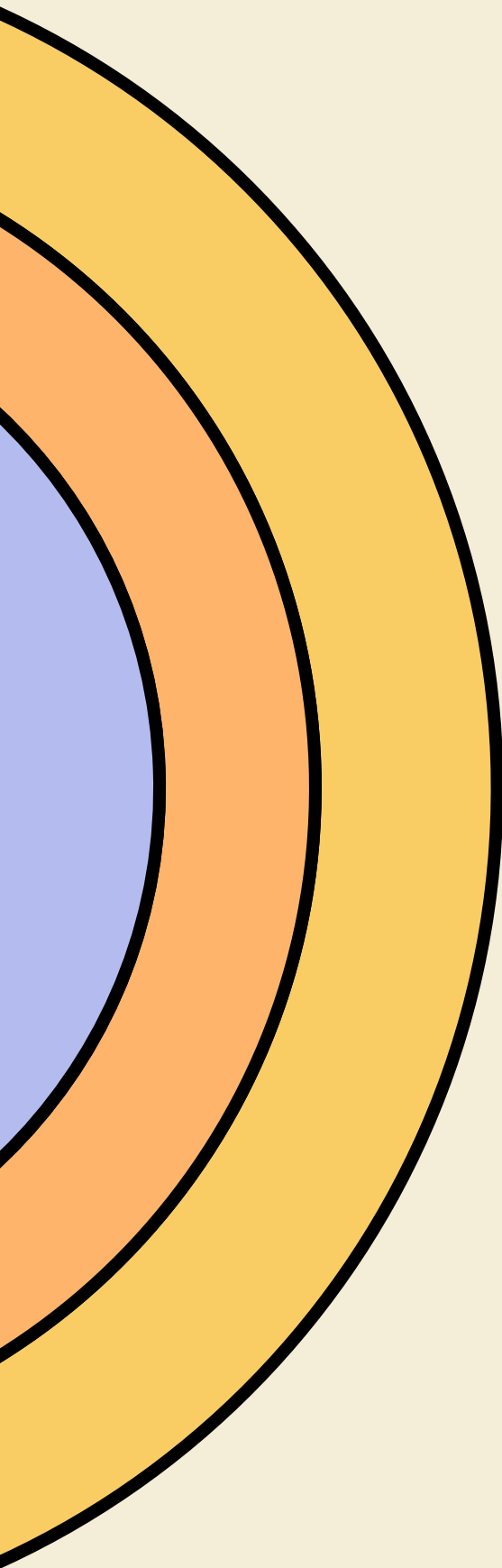
## Metode Eksperimen

Dimana peneliti melakukan percobaan menggunakan dataset yang disediakan oleh UCI Machine learning. Dataset ini digunakan sebagai data training dengan menggunakan pemrograman bahasa Python dan menggunakan library Sklearn untuk pengklasifikasian target berupa diagnosis ada atau tidaknya penyakit jantung menggunakan metode Naive Bayes. Peneliti menguji keakuratan dari sistem prediksi yang dibangun dengan menghitung persentase benar atau salah dari hasil prediksi sistem

## Pengumpulan Dataset

Dataset yang digunakan adalah Heart Disease dari publik yang sudah tersebar di internet. Menggunakan dataset untuk melakukan proses testing dan juga training data yang terdiri dari dataset nyata 304 contoh data dengan 13 fitur, deskripsi, dan values yang ditampilkan pada tabel

# METODELOGI PENELITIAN



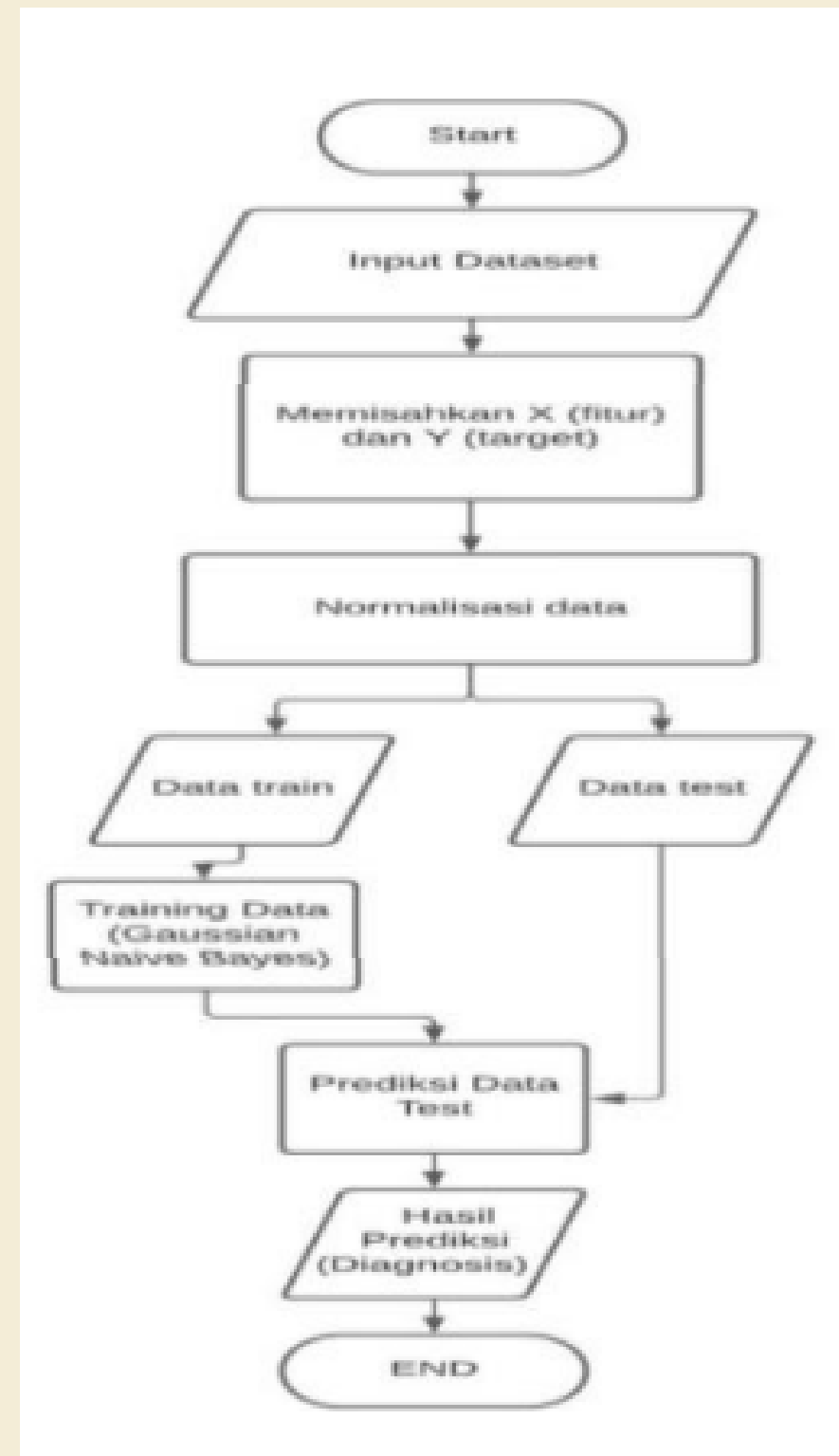
TABEL I. FITUR DATASET

Sr. no	Attribute	Description	Values
1	Age	Age in years	Continuous
2	Sex	Male or female	1 = male
3	Cp	Chest pain type	2 = typical type agina 3 = non-agina pain 4=asymptomatic
4	threstbps	Resting blood pressure	Continuous value in mm hg
5	chol	Serum cholesterol	Continuous value in mm/dl
6	Restecg	Resting electrographic results	0 = normal 1 = having ST_T wave abnormal 2 = left ventricular hypertrophy
7	FBS	Fasting blood sugar	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl
8	thalach	Maximum heart rate achieved	Continuous value
9	exang	Exercise-induced agina	0= no 1 = yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
9	exang	Exercise-induced agina	0= no 1 = yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	slope	The slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 = downsloping
12	Ca	Number of major vessels colored by floursopy	0-3 value
13	thal	Defect type	3 = normal 6 = fixed 7 = reversible defect

# METODELOGI PENELITIAN

## Rancangan Algoritma Sistem

Pada penelitian ini saya merancang skema sistem prediksi. Penelitian ini memiliki tujuan untuk menghasilkan produk berupa sistem prediksi penyakit jantung, sehingga dapat membantu untuk mendiagnosa pasien. flowchart dimulai dengan input dataset yang selanjutnya akan dilakukan pemisahan fitur (X), dan target (Y). Kemudian dilakukan normalisasi data pada fitur dan target tersebut akan mendapatkan hasil data training dan juga data test. Langkah terakhir dilakukan pengujian data test tersebut dan didapatkan hasil prediksi.



# HASIL DAN PEMBAHASAN

## Normalize Data

```
[ [0.70833333 1.          1.          ... 0.          0.          0.33333333 ]  
 [0.16666667 1.          0.66666667 ... 0.          0.          0.66666667 ]  
 [0.25         0.          0.33333333 ... 1.          0.          0.66666667 ]  
 ...  
 [0.8125        1.          0.          ... 0.5         0.5         1.          ]  
 [0.58333333 1.          0.          ... 0.5         0.25        1.          ]  
 [0.58333333 0.          0.33333333 ... 0.5         0.25        0.66666667 ] ]
```

Tahap ini adalah tahapan yang penting dalam pra proses Machine learning. Normalize data berfungsi untuk membuat beberapa variable data memiliki rentang nilai yang sama, sehingga tidak ada data yang terlalu besar maupun terlalu kecil.

# HASIL DAN PEMBAHASAN

## Training Data

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
74	0.291667	0.0	0.666667	0.264151	0.198630	0.0	0.5	0.717557	0.0	0.032258	0.5	0.00	0.666667
153	0.770833	0.0	0.666667	0.490566	0.347032	0.0	0.0	0.618321	0.0	0.000000	0.5	0.25	0.666667
64	0.604167	1.0	0.666667	0.433962	0.194064	1.0	0.0	0.717557	0.0	0.000000	1.0	0.00	0.666667
296	0.708333	0.0	0.000000	0.283019	0.162100	0.0	0.5	0.496183	1.0	0.000000	0.5	0.00	0.666667
287	0.583333	1.0	0.333333	0.568038	0.242009	0.0	0.0	0.709924	0.0	0.000000	1.0	0.25	0.666667
...	...	...	...	...	...	...	...	...	...	...	...	...	...
251	0.291667	1.0	0.000000	0.359491	0.278256	1.0	0.0	0.549618	1.0	0.018129	0.5	1.00	1.000000
192	0.520833	1.0	0.000000	0.245283	0.141553	0.0	0.5	0.320611	0.0	0.225806	0.5	0.25	1.000000
117	0.562500	1.0	1.000000	0.245283	0.152968	0.0	0.0	0.694656	0.0	0.306452	0.5	0.00	1.000000
47	0.375000	1.0	0.666667	0.415094	0.299087	0.0	0.0	0.648855	0.0	0.000000	1.0	0.00	0.666667
172	0.604167	1.0	0.333333	0.245283	0.380731	0.0	0.0	0.679389	0.0	0.290323	0.5	0.00	0.666667

Training data adalah data yang akan digunakan untuk melatih program yang dibuat nantinya dapat membuat prediksi sehingga mampu mencari korelasi data sendiri atau belajar pola dari data yang diberikan.



# HASIL DAN PEMBAHASAN

## Pemisahan Data

TABEL II. PERBANDINGAN DATA LATIH DAN DATA TES

	Data Latih	Data Tes	Total
Persentase	80	20	100
Jumlah	243	61	304

Data yang digunakan selanjutnya akan dipisah menjadi dua kelompok yaitu data latih dan data tes

# HASIL DAN PEMBAHASAN

## Data Tes

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
226	0.854167	1.0	0.000000	0.481132	0.109589	0.0	0.5	0.412214	1.0	0.419355	0.0	0.00	1.000000
152	0.729167	1.0	1.000000	0.716981	0.230594	0.0	0.0	0.641221	0.0	0.096774	0.5	0.00	1.000000
228	0.825000	1.0	1.000000	0.716981	0.369083	0.0	0.0	0.671758	0.0	0.032258	0.5	0.00	1.000000
201	0.645833	1.0	0.000000	0.292453	0.301370	0.0	0.0	0.534351	1.0	0.451613	0.5	0.25	1.000000
52	0.687500	1.0	0.666667	0.339623	0.239726	0.0	0.5	0.572519	0.0	0.290323	0.5	0.75	1.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
146	0.312500	0.0	0.666667	0.228415	0.264840	0.0	0.5	0.695420	0.0	0.048387	0.5	0.25	0.666667
302	0.583333	0.0	0.333333	0.339623	0.251142	0.0	0.0	0.786260	0.0	0.000000	0.5	0.25	0.666667
36	0.825000	1.0	0.666667	0.528302	0.196347	1.0	0.5	0.658489	0.0	0.258065	1.0	0.00	0.666667
108	0.437500	0.0	0.333333	0.245283	0.269406	0.0	0.5	0.694858	0.0	0.177419	1.0	0.00	0.666667
89	0.604167	0.0	0.000000	0.056604	0.278539	0.0	0.0	0.389313	0.0	0.161290	0.5	0.00	0.666667

Data tes merupakan data yang digunakan untuk mengetes program yang telah dibuat apakah mampu mencari korelasi data sehingga dapat dilihat keakuratannya. Seperti pada Gambar 5, sisa 20% dari keseluruhan data digunakan sebagai data tes yaitu berjumlah 61 data. Data tes tidak boleh merupakan data yang pernah dilihat oleh model sebelumnya

# HASIL DAN PEMBAHASAN

## Data Prediksi

```
y_pred = nb.predict(x_test)
y_pred

array([1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1,
       1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1,
       0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0])
```

Data ini adalah data uji yang digunakan untuk melakukan klasifikasi berdasarkan dataset tes. Pada Gambar 7, total dari data prediksi sama dengan data tes yaitu 61 data.

## Confusion Matrix

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay

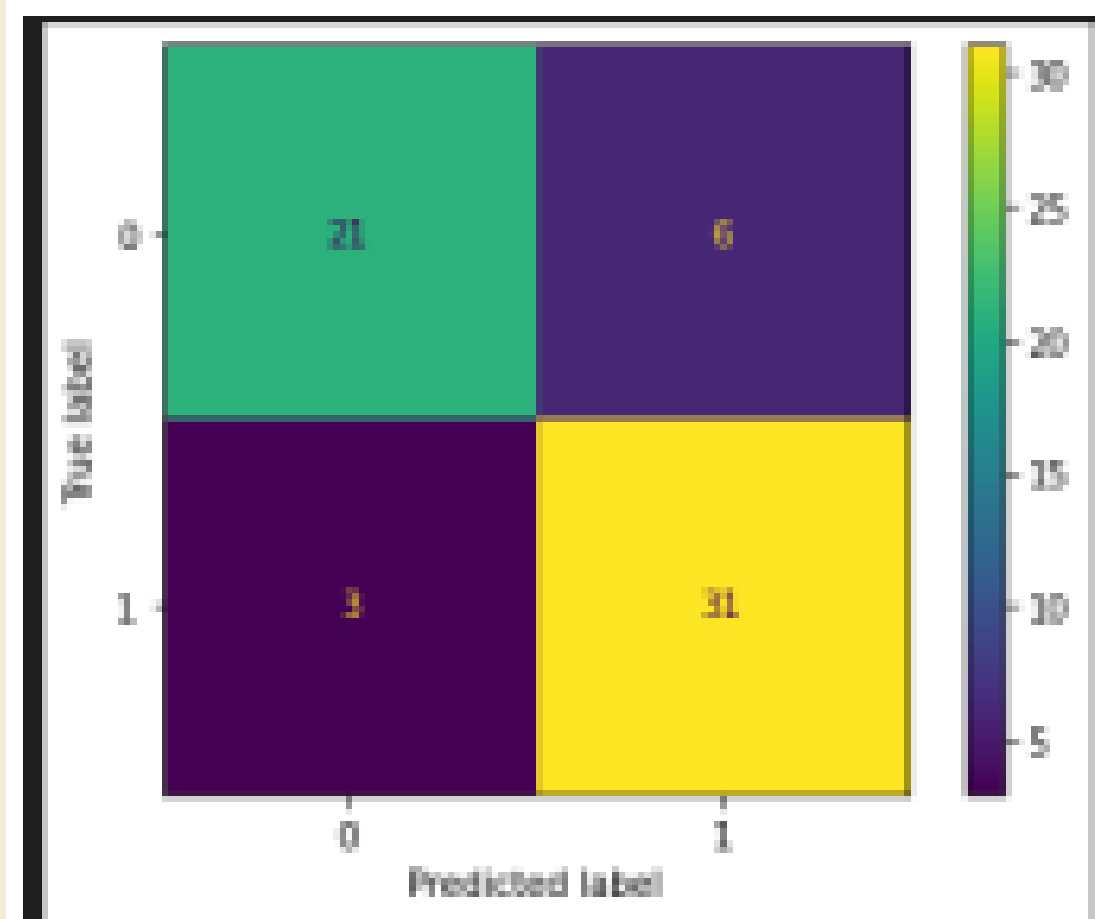
y_pred = nb.predict(x_test)

cm = confusion_matrix(y_test, y_pred)
cm_display = ConfusionMatrixDisplay(cm).plot()
```

Pada Gambar diatas digunakan fungsi `confusion_matrix` dan `ConfusionMatrixDisplay` yang terdapat di dalam modul `sklearn`. "y\_pred" merupakan variabel pengganti fungsi `nb.predict()` yang digunakan untuk menguji model menggunakan data yang akan diuji. Selanjutnya dilakukan pengecekan hasil data uji dengan "hasil yang sebenarnya" menggunakan fungsi `confusion_matrix`. Variabel "cm\_display" digunakan untuk menggantikan fungsi `ConfusionMatrixDisplay().plot()` yang berfungsi untuk menampilkan hasil dari `confusion_matrix`.

# HASIL DAN PEMBAHASAN

Confusion Matrix



Gambar disamping menunjukkan bahwa 52 data yang terprediksi dinyatakan benar sesuai dengan “hasil yang sebenarnya”. Di antaranya sebanyak 21 data yang terprediksi benar tidak mengidap penyakit jantung dan data yang terprediksi benar mengidap sakit jantung berjumlah 31 data. Kemudian terdapat 6 data terprediksi mengidap sakit jantung dan 3 data terprediksi tidak mengidap penyakit jantung adalah salah. Sehingga total data yang terprediksi salah berjumlah 9 data

# HASIL DAN PEMBAHASAN

## Tingkat Akurasi

$$\begin{aligned} \text{Akurasi} &= \frac{\Sigma \text{Prediksi Benar}}{\Sigma \text{Prediksi}} \times 100\% \\ &= \frac{52}{61} \times 100\% \\ &= 85.2459\% \\ &= 85.25\% \end{aligned}$$

Dari 61 data yang diprediksi, data yang benar berjumlah 52 sedangkan data yang salah berjumlah 9 buah data. Sehingga nilai akurasi dari model ini menggunakan persamaan 3

```
from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred)*100
accuracies['Naive Bayes'] = acc
print("Akurasi: {:.2f}%".format(acc))
```

✓ 0.3s

Akurasi: 85.25%

dilakukan penginputan perhitungan untuk menentukan nilai dari akurasi (`accuracy_score`) yang terdapat didalam modul `sklearn`. Setelah itu dibuat variabel “acc” sebagai pengganti dari fungsi `accuracy_score` yang digunakan untuk menghitung akurasi dari data yang telah diuji. Pada perhitungan akurasi ini, hasilnya berupa nilai 1 angka dibelakang koma sehingga dikalikan dengan 100 untuk menampilkan bilangan puluhan

# KESIMPULAN

dari penelitian ini menunjukkan bahwa meskipun model Naive Bayes berhasil memperoleh tingkat keakuratan yang relatif tinggi, perlu adanya evaluasi lebih lanjut dengan metode klasifikasi lain untuk memastikan keefektifan dan kehandalan prediksi penyakit jantung. Hal ini dapat membantu dalam menemukan metode yang paling optimal untuk diagnosis penyakit jantung.



**THANK YOU**