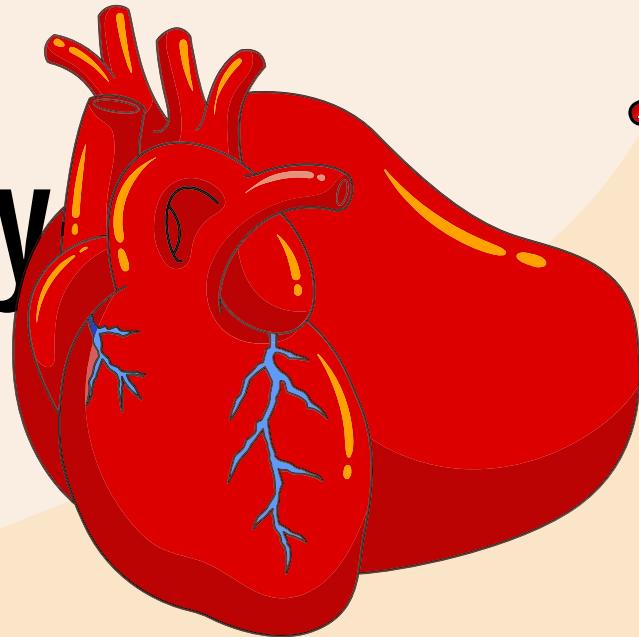


Health Case Study

Heart Disease



By Rofik

Link Colab:

<https://colab.research.google.com/drive/1NTXb4iaTwQoCtS215TReQmHLk6nRUJgp?usp=sharing>

Heart Disease Prediction

Background

Penyakit jantung adalah penyebab utama kematian di banyak negara. Identifikasi faktor-faktor risiko dan pemahaman mendalam tentang data pasien yang terkena penyakit jantung penting untuk pembuatan model prediksi penyakit jantung yang memiliki performa optimal.

Tujuan

- Tujuan dari analisis ini adalah untuk memahami lebih baik faktor-faktor yang berkontribusi terhadap penyakit jantung.
- Mengidentifikasi pola atau tren dalam data yang dapat membantu pembuatan model yang mampu dengan baik membedakan orang-orang dalam keadaan sehat dan penderita penyakit jantung melalui data-data pengukurannya.

Deskripsi Dataset

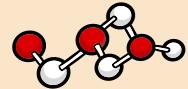
Dataset berasal dari Kaggle dari URL: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>. Terdiri dari 14 atribut. Dengan targetnya mengacu pada keberadaan penyakit jantung pasien. Target dengan kondisi 0 adalah tidak ada penyakit dan 1 adalah punya penyakit. Dataset terdiri dari 1025 record.

- Age (age in years)
- Sex (1 = male; 0 = female)
- CP (chest pain type)
- TRESTBPS (resting blood pressure (in mm Hg on admission to the hospital))
- CHOL (serum cholestoral in mg/dl)
- FPS (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- RESTECG (resting electrocardiographic results)
- THALACH (maximum heart rate achieved)
- EXANG (exercise induced angina (1 = yes; 0 = no))
- OLDPEAK (ST depression induced by exercise relative to rest)
- SLOPE (the slope of the peak exercise ST segment)
- CA (number of major vessels (0-3) colored by flourosopy)
- THAL (3 = normal; 6 = fixed defect; 7 = reversable defect)
- TARGET (1 or 0)

Alat bantu kisaran angka (range) dalam fitur.

Intelligent Heart Disease Prediction System

Age	<input type="text"/> [Range: 25-110]
Gender	<input type="text" value="Male"/>
Chest Pain Type	<input type="text" value="Typical Angina"/>
Blood Pressure (In mmHg)	<input type="text"/> [Range: 60-200]
Cholesterol (In mg/dl)	<input type="text"/> [Range: 120-600]
Fasting Blood Sugar	<input type="text" value=">120 mg/dl"/>
Resting ECG	<input type="text" value="Normal"/>
Maximum Heart Rate	<input type="text"/> [Range: 60-200]
Exercise Induced Angina	<input type="text" value="Please select"/>
Old Peak	<input type="text"/> [Range: 0-6]
Slope	<input type="text" value="Upsloping"/>
Number of Vessels Colored	<input type="text" value="0"/>
Thal	<input type="text" value="Normal"/>



Gambaran Data

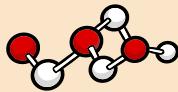


1 df.head()

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

[6] 1 df.tail()

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0



```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   age         1025 non-null    int64  
 1   sex          1025 non-null    int64  
 2   cp           1025 non-null    int64  
 3   trestbps     1025 non-null    int64  
 4   chol          1025 non-null    int64  
 5   fbs           1025 non-null    int64  
 6   restecg       1025 non-null    int64  
 7   thalach        1025 non-null    int64  
 8   exang          1025 non-null    int64  
 9   oldpeak        1025 non-null    float64 
 10  slope          1025 non-null    int64  
 11  ca              1025 non-null    int64  
 12  thal            1025 non-null    int64  
 13  target          1025 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Data terdiri dari 1025 record. Dengan 14 target diantaranya adalah

1. age (usia)
2. sex (jenis kelamin)
3. cp (jenis nyeri dada)
4. trestbps (tekanan darah)
5. chol (kolesterol serum dalam mg)
6. fbs (gula darah)
7. restecg (hasil elektrokardiografi)
8. thalach (detak jantung maksimum)
9. exang (angina yang disebabkan oleh olahraga)
10. oldpeak (depresi ST yang diinduksi oleh olahraga)
11. slope (kemiringan segmen ST)
12. ca (jumlah pembuluh darah utama)
13. thal (acat yang dapat diperbaiki)
14. Target

Masing-masing bertipe data int, kecuali oldpeak dengan tipe data float





Deskripsi Statistik

1 df.describe()

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000



Data Preparation

01 Identifikasi dan penyelesaian Missing Value

```
1 df.isnull().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

Didapatkan bahwa tidak ada missing value dalam dataset.
Tidak perlu diatasi

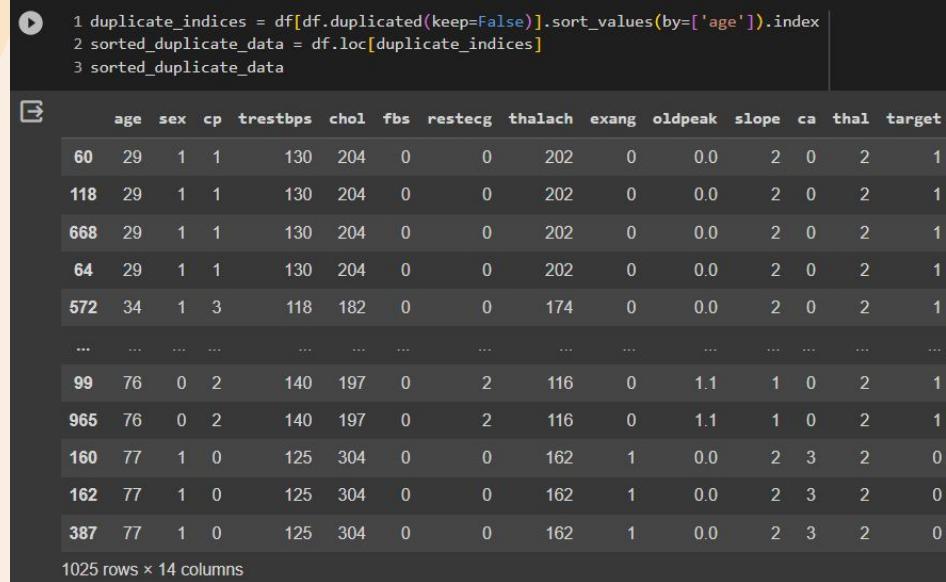


Data Preparation

02 Identifikasi dan penyelesaian Duplicate Value

```
1 df.duplicated().sum()  
2  
723
```

Terdapat 723 baris data yang duplikat



```
1 duplicate_indices = df[df.duplicated(keep=False)].sort_values(by=['age']).index  
2 sorted_duplicate_data = df.loc[duplicate_indices]  
3 sorted_duplicate_data
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
60	29	1	1	130	204	0	0	202	0	0.0	2	0	2	1
118	29	1	1	130	204	0	0	202	0	0.0	2	0	2	1
668	29	1	1	130	204	0	0	202	0	0.0	2	0	2	1
64	29	1	1	130	204	0	0	202	0	0.0	2	0	2	1
572	34	1	3	118	182	0	0	174	0	0.0	2	0	2	1
...
99	76	0	2	140	197	0	2	116	0	1.1	1	0	2	1
965	76	0	2	140	197	0	2	116	0	1.1	1	0	2	1
160	77	1	0	125	304	0	0	162	1	0.0	2	3	2	0
162	77	1	0	125	304	0	0	162	1	0.0	2	3	2	0
387	77	1	0	125	304	0	0	162	1	0.0	2	3	2	0

1025 rows × 14 columns

Ternyata data duplikat dalam 1 baris itu sama perkolomnya (identik).

Data duplikat tersebut tidak memberikan informasi tambahan yang bermanfaat dan hanya mengakibatkan redundansi



```
[12] 1 df = df.drop_duplicates(keep='first')
```

1 df

2

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...
723	68	0	2	120	211	0	0	115	0	1.5	1	0	2	1
733	44	0	2	108	141	0	1	175	0	0.6	1	0	2	1
739	52	1	0	128	255	0	1	161	1	0.0	2	1	3	0
843	59	1	3	160	273	0	0	125	0	0.0	2	0	2	0
878	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

302 rows × 14 columns

Data duplikat dihapus, dengan menyisakan 1 data awal (yang sebelumnya sama atau duplikat).

Sisa data sekarang menjadi 302 record/baris.



Data Preparation

02 Identifikasi dan penyelesaian Outlier –Pada fitur age

Perbandingan summary statistic dari fitur age, z-scorenya dan IQRnya

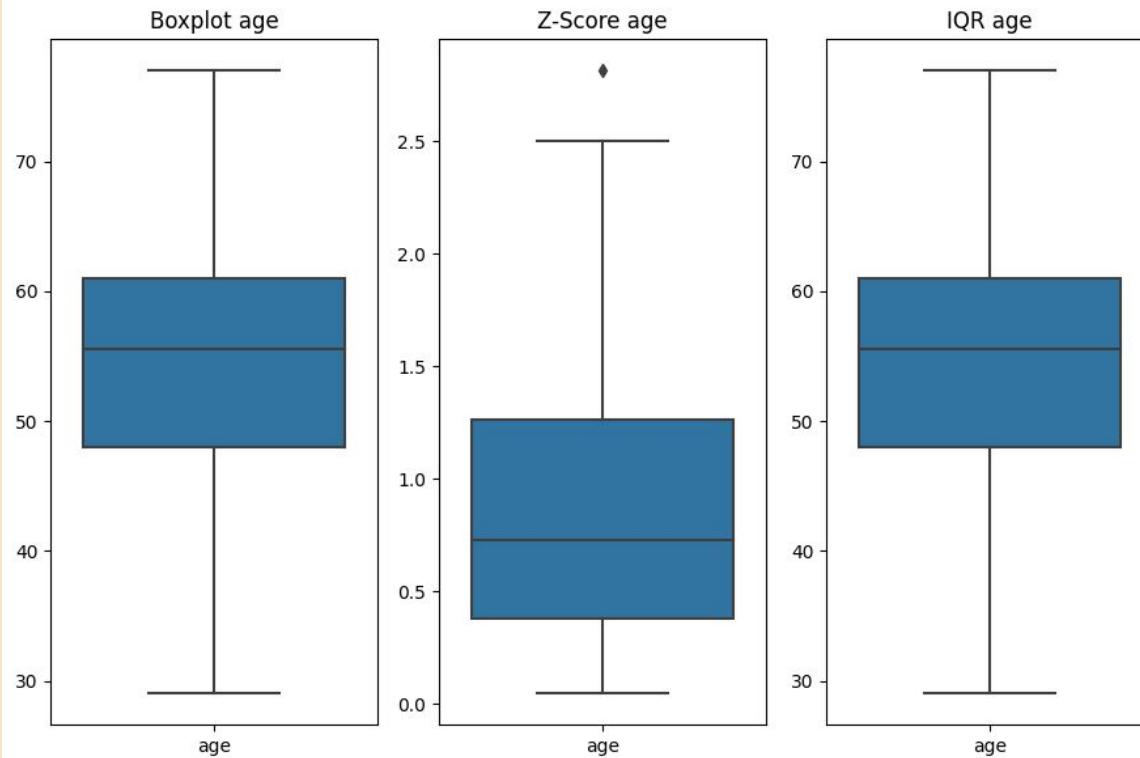
	Age	Age Z-Score	IQR
count	302.00000	302.000000	302.0
mean	54.42053	0.821959	13.0
std	9.04797	0.570491	0.0
min	29.00000	0.046555	13.0
25%	48.00000	0.378671	13.0
50%	55.50000	0.728383	13.0
75%	61.00000	1.264315	13.0
max	77.00000	2.814192	13.0

- Data berjumlah 302 record
- Rata-rata usianya adalah 54
- Terdapat variasi yang cukup besar di fitur age, yang mana standar deviasinya menunjukkan hasil sebesar 9.05
- Usia paling rendah adalah 29
- Kuartil pertama adalah 48
- Kuartil kedua atau mediannya adalah 55.5
- Kuartil ketiganya adalah 61
- Dan Usia tertinggi adalah 77



1. **Z-Score:** Data Age dengan mean 0.821959 dan standar deviasi 0.570491. Z-Score maksimum adalah 2.814192. Secara umum, jika nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3, itu dianggap sebagai outlier. Dalam kasus Heart Disease ini, Z-Score tertinggi (2.814192) jauh lebih kecil dari 3, sehingga dengan kriteria ini, data Age tidak memiliki outlier berdasarkan Z-Score.
2. **IQR:** Data Age dengan nilai IQR (Interquartile Range) sebesar 13.0. IQR adalah perbedaan antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Jika kita tidak memiliki nilai Q1 dan Q3, IQR tidak dapat digunakan untuk mengidentifikasi outlier. Namun, jika IQR sebenarnya adalah 13.0 dan seluruh rentang data sebesar 13.0, maka ini juga menunjukkan bahwa tidak ada outlier dalam data.

Perbandingan menggunakan visualisasi bloxpot



```
1 df['age'].max()  
77
```

Nilai maximum 77, dalam konteks ini adalah orang yang usianya 77 tahun. Maka apabila mendeteksi penyakit jantung, hal ini wajar. Jika yang diperiksa adalah orang dengan usia tersebut. Jadi saya menganggap itu bukan outlier.

Apalagi pada gambar awal rentang usia adalah 25-110

Data Preparation

02 Identifikasi dan penyelesaian Outlier –Pada fitur trestbps(tekanan darah)

Perbandingan summary statistic dari fitur trestbps, z-scorenya dan IQRnya

	trestbps	trestbps Z-Score	IQR
count	302.000000	302.000000	302.0
mean	131.602649	0.772395	20.0
std	17.563394	0.636197	0.0
min	94.000000	0.022661	20.0
25%	120.000000	0.364848	20.0
50%	130.000000	0.661712	20.0
75%	140.000000	1.117961	20.0
max	200.000000	3.900776	20.0

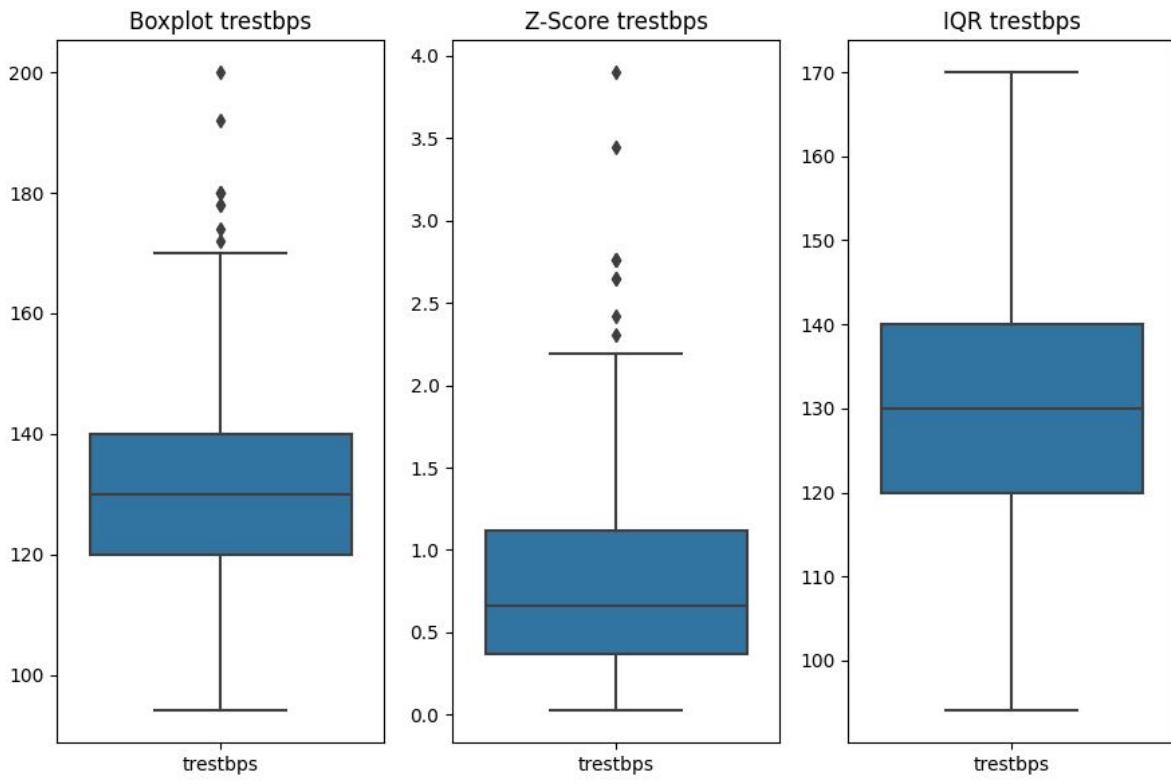
- Data berjumlah 302 record
- Rata-rata tekanan darah adalah 131.60
- Terdapat variasi yang signifikan pada tekanan darah, yang mana standar deviasinya menunjukkan hasil sebesar 17.56
- Tekanan darah paling rendah adalah 94
- Kuartil pertama adalah 120
- Kuartil kedua atau mediannya adalah 130
- Kuartil ketiganya adalah 140
- Dan tekanan darah tertinggi adalah 200

- Z-Score:** Data tekanan darah dengan mean 0.77 dan standar deviasi 0.64. Z-Score maksimum adalah 3.90. Umumnya, jika nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3, itu dianggap sebagai outlier. Dalam kasus ini, Z-Score tertinggi (3.90) jauh lebih besar dari 3, sehingga berdasarkan kriteria ini, data **trestbps** memiliki outlier berdasarkan Z-Score.
- IQR:** Data tekanan darah dengan nilai IQR (Interquartile Range) sebesar 20.0. IQR adalah perbedaan antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Jika IQR adalah 20.0 dan seluruh rentang data sebesar 20.0, maka ini juga menunjukkan bahwa tidak ada outlier dalam data **trestbps**. Tapi dengan stdnya 0, z-score dipertimbangkan sebagai pengukuran ada tidaknya outlier

Blood Pressure Stages			
Blood Pressure Category	Systolic mm Hg (upper #)		Diastolic mm Hg (lower #)
Normal	less than 120	and	less than 80
Elevated	120-129	and	less than 80
High Blood Pressure (Hypertension) Stage 1	130-139	or	80-89
High Blood Pressure (Hypertension) Stage 2	140 or higher	or	90 or higher
Hypertensive Crisis (Seek Emergency Care)	higher than 180	and/or	higher than 120

Jika dilihat pada gambar tersebut, nilai yang tinggi pada fitur **trestbps** tersebut mungkin saja terjadi. Dan range pada gambar di awal untuk **trestbps** adalah 60-200. Berarti nilai-nilai pada fitur tersebut tidak outlier.

Perbandingan menggunakan visualisasi bloxpot



```
1 df['trestbps'].max()  
200
```

Nilai maximum 200, meskipun ada titik poin di atas bloxpot, tapi nilai IQR menunjukkan konstan. Dan karena ini konteksnya tekanan darah, maka tekanan darah yang mencapai 200 itu benar ada. Apalagi untuk deteksi penyakit jantung. Itu mungkin saja terjadi pada pasien. Jadi dianggap itu bukan outlier.

Data Preparation

02 Identifikasi dan penyelesaian Outlier –Pada fitur chol (kolesterol serum dalam mg)

Perbandingan summary statistic dari fitur chol, z-scorenya dan IQRnya

	chol	chol Z-Score	IQR
count	302.000000	302.000000	302.00
mean	246.500000	0.759309	63.75
std	51.753489	0.651810	0.00
min	126.000000	0.009677	63.75
25%	211.000000	0.299994	63.75
50%	240.500000	0.667728	63.75
75%	274.750000	1.049978	63.75
max	564.000000	6.145034	63.75

- Data berjumlah 302 record
- Rata-rata chol adalah 246.50
- Terdapat variasi yang cukup besar pada fitur chol, yang mana standar deviasinya menunjukkan hasil sebesar 51.75 (relatif tinggi)
- Chol paling rendah adalah 126.0
- Kuartil pertama adalah 211.0
- Kuartil kedua atau mediannya adalah 240.5
- Kuartil ketiganya adalah 274.75
- Dan chol tertinggi adalah 564.0

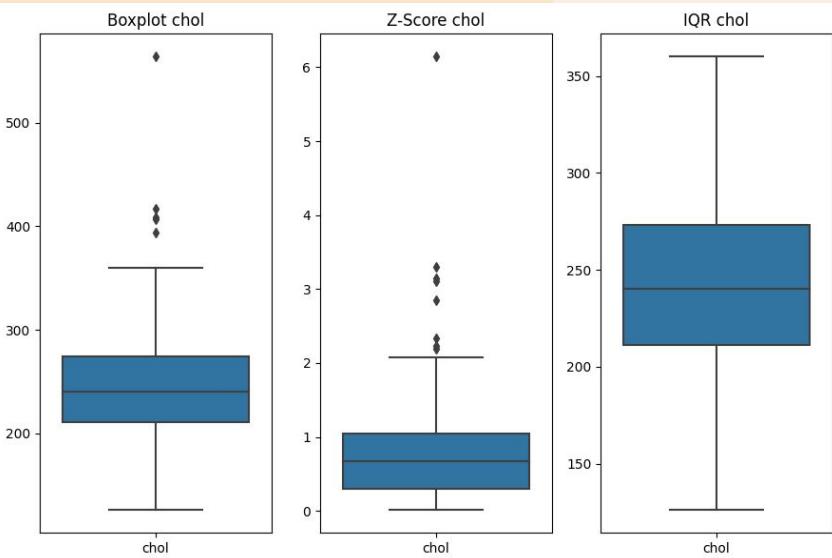


1.Z-Score: Data kolesterol dengan mean 0.76 dan standar deviasi 0.65. Z-Score maksimum adalah 6.15. Secara umum, jika nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3, itu dianggap sebagai outlier. Dalam kasus ini, Z-Score tertinggi (6.15) jauh lebih besar dari 3, sehingga dengan kriteria ini, data kolesterol (**chol**) memiliki outlier berdasarkan Z-Score.

2.IQR: Data kolesterol dengan nilai IQR (Interquartile Range) sebesar 63.75. IQR adalah perbedaan antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Jika kita tidak memiliki nilai Q1 dan Q3, IQR tidak dapat digunakan untuk mengidentifikasi outlier. Namun, jika IQR sebenarnya adalah 63.75 dan seluruh rentang data sebesar 63.75, maka ini juga menunjukkan bahwa tidak ada outlier dalam data **chol**. Karena stdnya juga 0, maka z-score dipertimbangkan kembali untuk mendeteksi outlier.



Perbandingan menggunakan visualisasi bloxpot



```
1 # Mendapatkan data terindikasi outlier pada fitur 'chol'
2 Q1_chol = df['chol'].quantile(0.25)
3 Q3_chol = df['chol'].quantile(0.75)
4 IQR_chol = Q3_chol - Q1_chol
5 lower_bound_chol = Q1_chol - 1.5 * IQR_chol
6 upper_bound_chol = Q3_chol + 1.5 * IQR_chol
7
8 outlier_data_chol = df[(df['chol'] < lower_bound_chol) | (df['chol'] > upper_bound_chol)]
9
10 # Menampilkan data terindikasi outlier
11 outlier_data_chol
12
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
123	65	0	2	140	417	1	0	157	0	0.8	2	1	2
158	67	0	2	115	564	0	0	160	0	1.6	1	0	3
179	56	0	0	134	409	0	0	150	1	1.9	1	2	3
255	62	0	0	140	394	0	0	157	0	1.2	1	0	2
450	63	0	0	150	407	0	0	154	0	4.0	1	3	3

Ditemukan 5 baris yang terindikasi outlier.

LDL ("Kolesterol jahat")	
Kurang dari 100	Optimal
100-129	Mendekati optimal
130-159	Batas normal tertinggi
160-189	Tinggi
Lebih dari 190	Sangat tinggi
HDL ("Kolesterol Baik")	
Kurang dari 40	Rendah
Lebih dari 60	Tinggi
Total kolesterol (TC)	
Kurang dari 200	Yang diperlukan
200-239	Batas normal tertinggi
Lebih dari 240	Tinggi
Trigliserida (TGA)	
Kurang dari 150	Normal
150-199	Batas normal tertinggi
200-499	Tinggi
Sama atau lebih dari 500	Sangat tinggi

Berdasarkan gambar tersebut maka ada mungkin seseorang memiliki kolesterol dengan nilai-nilai mencapai tersebut. Dan berdasarkan IQR juga, maka dianggap tidak ada outlier. Banyaknya data yang terindikasi outlier juga tidak lebih dari 10. Jadi dianggap itu bukan outlier.

Gambar diawal juga menunjukkan bahwa range kolesterol adalah 120-600. Jadi nilai-nilai tinggi yang ada pada data bukan outlier.

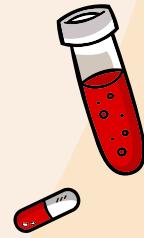
Data Preparation

02 Identifikasi dan penyelesaian Outlier –Pada fitur thalach (detak jantung maksimum)

Perbandingan summary statistic dari fitur thalach, z-scorenya dan IQRnya

	thalach	thalach Z-Score	IQR
count	302.000000	302.000000	302.00
mean	149.569536	0.808115	32.75
std	22.903527	0.590003	0.00
min	71.000000	0.018826	32.75
25%	133.250000	0.331045	32.75
50%	152.500000	0.718568	32.75
75%	166.000000	1.116734	32.75
max	202.000000	3.436149	32.75

- Data berjumlah 302 record
- Rata-rata chol adalah 149.57
- Estándar deviasi 22.90
merupakan nilai yang moderat.
-> variasinya cukup signifikan.
- Thalach paling rendah adalah 71
- Kuartil pertama adalah 133.25
- Kuartil kedua atau mediannya adalah 152.50
- Kuartil ketiganya adalah 166.00
- Dan thalach tertinggi adalah 202

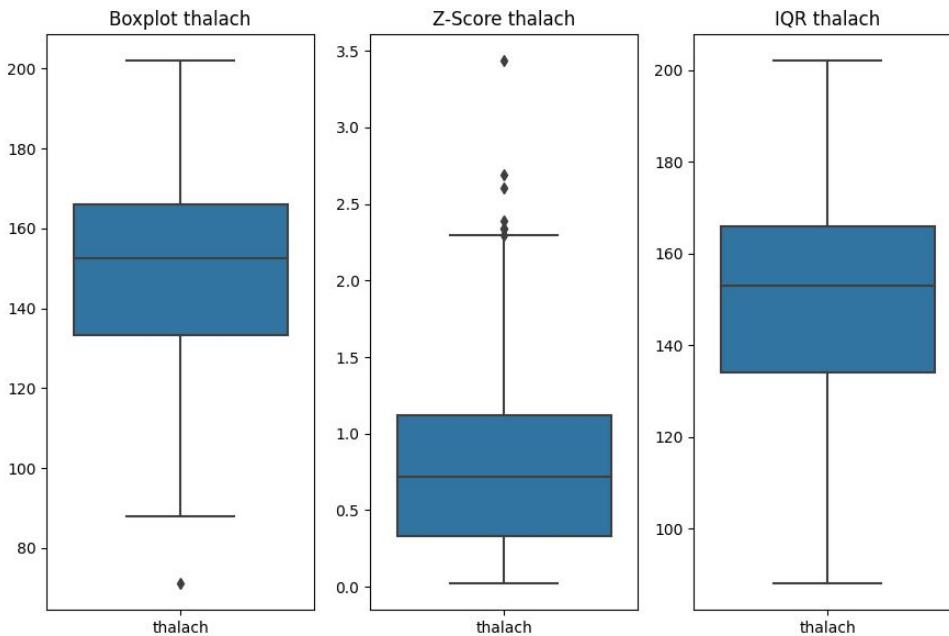


1. Z-Score: Data detak jantung maksimal dengan mean 0.81 dan standar deviasi 0.59. Z-Score maksimum adalah 3.44. Umumnya, jika nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3, itu dianggap sebagai outlier. Dalam kasus Anda, Z-Score tertinggi (3.44) jauh lebih besar dari 3, sehingga dengan kriteria ini, data detak jantung maksimal (**thalach**) memiliki outlier berdasarkan Z-Score.

2. IQR: Data detak jantung maksimal dengan nilai IQR (Interquartile Range) sebesar 32.75. IQR adalah perbedaan antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Jika kita tidak memiliki nilai Q1 dan Q3, IQR tidak dapat digunakan untuk mengidentifikasi outlier. Namun, jika IQR sebenarnya adalah 63.75 dan seluruh rentang data sebesar 32.75, maka ini juga menunjukkan bahwa tidak ada outlier dalam data **thalach**.



Perbandingan menggunakan visualisasi bloxpot



Jika dilihat ternyata dalam data, ada orang yang memiliki detak jantung maksimum sebesar 71 dengan usia 67. Namun ini juga bisa terjadi.

Intelligent Heart Disease Prediction System

Age	[Range: 25-110]
Gender	Male
Chest Pain Type	Typical Angina
Blood Pressure (In mmHg)	[Range: 60-200]
Cholesterol (In mg/dl)	[Range: 120-600]
Fasting Blood Sugar	>120 mg/dl
Resting ECG	Normal
Maximum Heart Rate	[Range: 60-200]
Exercise Induced Angina	Please select
Old Peak	[Range: 0-6]
Slope	Upsloping
Number of Vessels Colored	0
Thal	Normal

Berdasarkan gambar tersebut, thalach rangenya adalah 60-200. Jadi titik poin tidak dianggap sebagai outlier

Data Preparation

02 Identifikasi dan penyelesaian Outlier –Pada fitur oldpeak

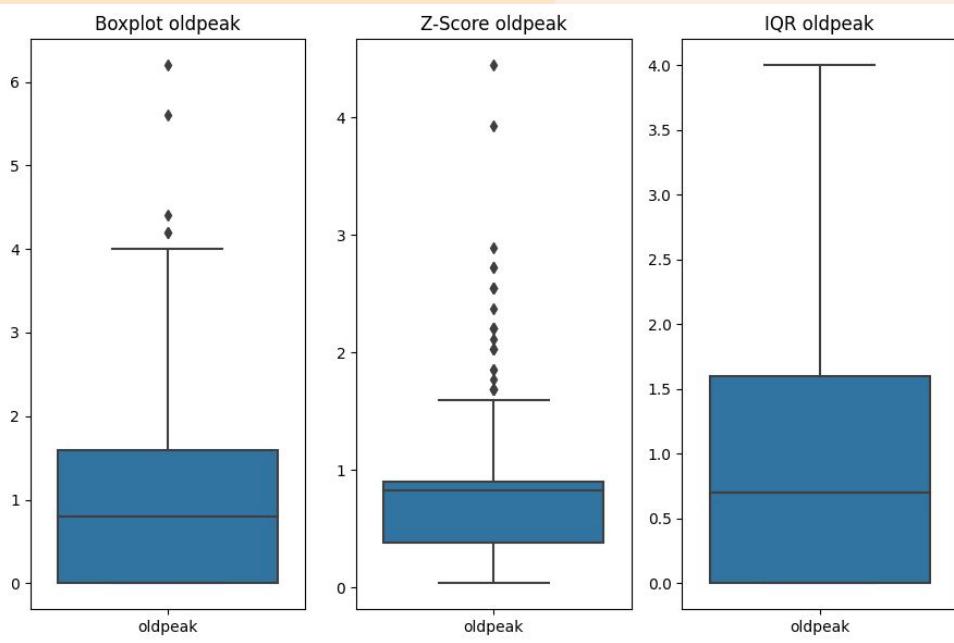
Perbandingan summary statistic dari fitur oldpeak, z-scorenya dan IQRnya

	oldpeak	oldpeak Z-Score	IQR
count	302.00	302.00	302.00
mean	1.04	0.80	1.60
std	1.16	0.60	0.00
min	0.00	0.04	1.60
25%	0.00	0.38	1.60
50%	0.80	0.83	1.60
75%	1.60	0.90	1.60
max	6.20	4.45	1.60

- Data berjumlah 302 record
- Rata-rata oldpeak adalah 1.04
- Estándar deviasi 1.16
merupakan nilai yang moderat.
-> variasinya cukup signifikan.
- Oldpeak paling rendah adalah 0.
- Kuartil pertama adalah 0
- Kuartil kedua atau mediannya adalah 0.80
- Kuartil ketiganya adalah 1.60
- Dan olpeka tertinggi adalah 6.20



Perbandingan menggunakan visualisasi bloxpot



Terlihat indikasi outlier pada boxplot oldpeak dan z-score oldpeak.

```
1 # Mendapatkan data outlier pada fitur 'oldpeak'
2 Q1_oldpeak = df['oldpeak'].quantile(0.25)
3 Q3_oldpeak = df['oldpeak'].quantile(0.75)
4 IQR_oldpeak = Q3_oldpeak - Q1_oldpeak
5 lower_bound_oldpeak = Q1_oldpeak - 1.5 * IQR_oldpeak
6 upper_bound_oldpeak = Q3_oldpeak + 1.5 * IQR_oldpeak
7
8 outlier_data_oldpeak = df[(df['oldpeak'] < lower_bound_oldpeak) | (df['oldpeak'] > upper_bound_oldpeak)]
9
10 # Menampilkan data outlier
11 outlier_data_oldpeak
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	
6	58	1	0	114	318	0	2	140.00	0	4.40	0	3	1	0
13	51	1	0	140	298	0	1	122.00	1	4.20	1	3	3	0
54	55	1	0	140	217	0	1	111.00	1	5.60	0	0	3	0
69	62	0	0	160	164	0	0	145.00	0	6.20	0	3	3	0
528	59	1	3	178	270	0	0	145.00	0	4.20	0	0	3	1

Identifikasi data outlier

Intelligent Heart Disease Prediction System

Age	<input type="text"/> [Range: 25-110]
Gender	<input type="text" value="Male"/>
Chest Pain Type	<input type="text" value="Typical Angina"/>
Blood Pressure (In mmHg)	<input type="text"/> [Range: 60-200]
Cholesterol (In mg/dl)	<input type="text"/> [Range: 120-600]
Fasting Blood Sugar	<input type="text" value=">120 mg/dl"/>
Resting ECG	<input type="text" value="Normal"/>
Maximum Heart Rate	<input type="text"/> [Range: 60-200]
Exercise Induced Angina	<input type="text" value="Please select"/>
Old Peak	<input type="text"/> [Range: 0-6]
Slope	<input type="text" value="Upsloping"/>
Number of Vessels Colored	<input type="text" value="0"/>
Thal	<input type="text" value="Normal"/>

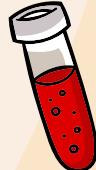
Berdasarkan gambar tersebut rentangng oldpeak adalah 0-6. Pada data ada data dengan oldpeak 6.20. Maka dilakukan penginputan nilai median untuk data dengan oldpeak tersebut

```
[38] 1 # Ganti nilai 'oldpeak' yang sama dengan 6.20 dengan nilai median
      2 median_oldpeak = df['oldpeak'].median()
      3 df.loc[df['oldpeak'] == 6.20, 'oldpeak'] = median_oldpeak
      4
```

```
[39] 1 df['oldpeak'].max()
```

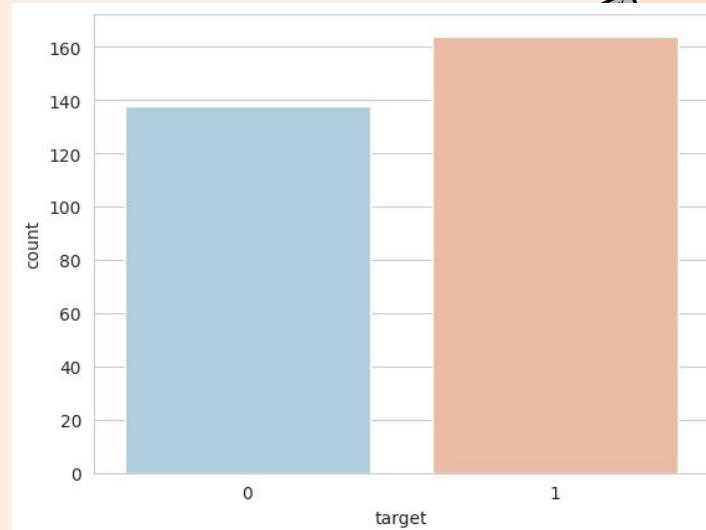
```
5.6
```

Periksa Data Tidak Seimbang



```
[43] 1 # Melihat jumlah penderita penyakit jantung dan tidak  
2 sick_counts = df['target'].value_counts()  
3 print('Jumlah Penderita Heart Disaese:', sick_counts[1])  
4 print('Jumlah Healthy people:', sick_counts[0])
```

```
Jumlah Penderita Heart Disaese: 164  
Jumlah Healthy people: 138
```



Setelah melalui tahap-tahap sebelumnya, ternyata data yang dihasilkan tidak seimbang. Data dengan kelas (1) yaitu orang dengan penderita heart disaese memiliki data yang lebih banyak dibanding orang yang dalam kondisi sehat (kelas =0)



Periksa Data Tidak Seimbang



```
[44] 1 from imblearn.over_sampling import SMOTE  
2  
3 # Memisahkan fitur dan target  
4 X = df.drop('target', axis=1)  
5 y = df['target']  
6  
7 # Menerapkan SMOTE  
8 smote = SMOTE(sampling_strategy='auto', random_state=42)  
9 X_resampled, y_resampled = smote.fit_resample(X, y)  
10  
11 # Menggabungkan fitur dan target kembali  
12 df_resampled = pd.concat([pd.DataFrame(X_resampled, columns=X.columns), pd.DataFrame({'target': y_resampled})], axis=1)
```

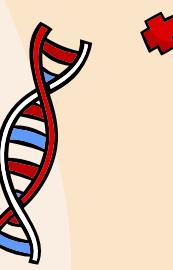
```
[45] 1 # Melihat distribusi kelas setelah SMOTE  
2 class_counts = df_resampled['target'].value_counts()  
3 print('Jumlah Penderita Heart Disease:', class_counts[1])  
4 print('Jumlah Orang Sehat:', class_counts[0])
```

```
Jumlah Penderita Heart Disease: 164  
Jumlah Orang Sehat: 164
```



Diterapkan SMOTE, untuk membuat data sintesis dari data minoritas (data orang dalam keadaan sehat). Dan dihasilkan data yang seimbang, yaitu sama-sama sejumlah 164.



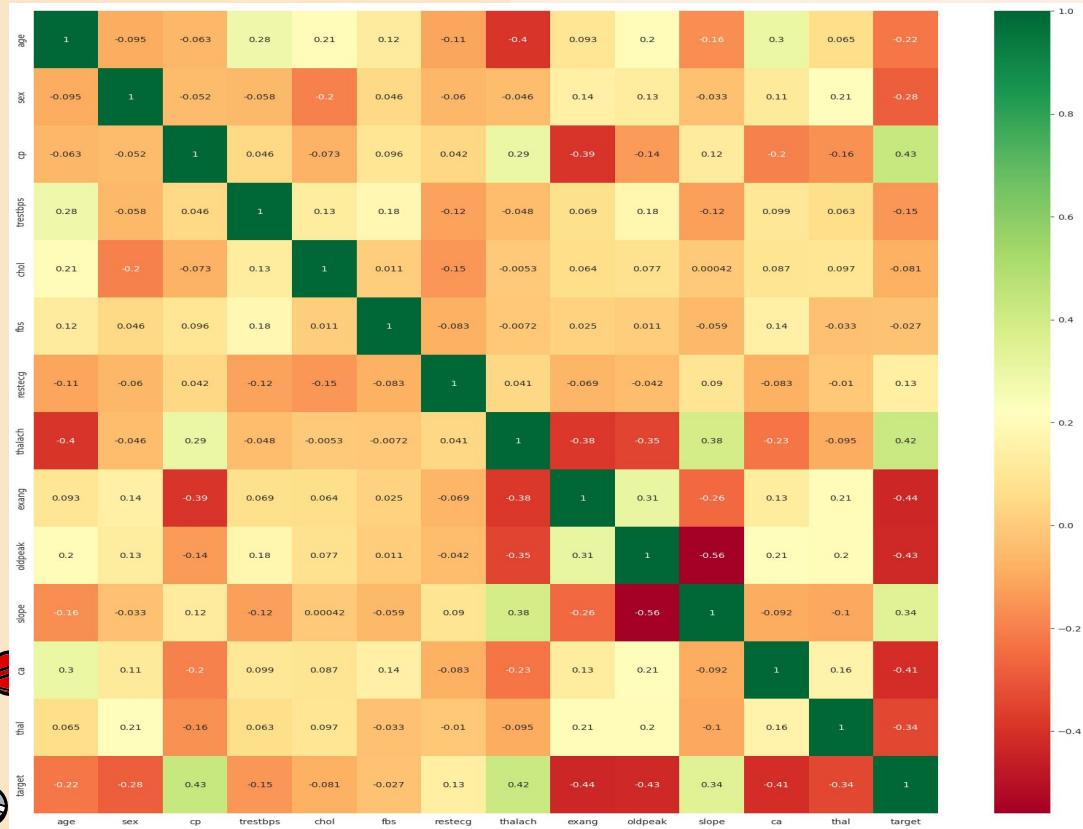


1 df

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212.0	0	1	168.0	0	1.0	2	2	3	0
1	53	1	0	140	203.0	1	0	155.0	1	3.1	0	0	3	0
2	70	1	0	145	174.0	0	1	125.0	1	2.6	0	0	3	0
3	61	1	0	148	203.0	0	1	161.0	0	0.0	2	1	3	0
4	62	0	0	138	294.0	1	1	106.0	0	1.9	1	3	2	0
...
723	68	0	2	120	211.0	0	0	115.0	0	1.5	1	0	2	1
733	44	0	2	108	141.0	0	1	175.0	0	0.6	1	0	2	1
739	52	1	0	128	255.0	0	1	161.0	1	0.0	2	1	3	0
843	59	1	3	160	273.0	0	0	125.0	0	0.0	2	0	2	0
878	54	1	0	120	188.0	0	1	113.0	0	1.4	1	1	3	0

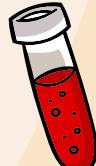
Data sudah berupa numerik semua, jadi tidak perlu dilakukan pengkodean lagi.

Korelasi antar Fitur

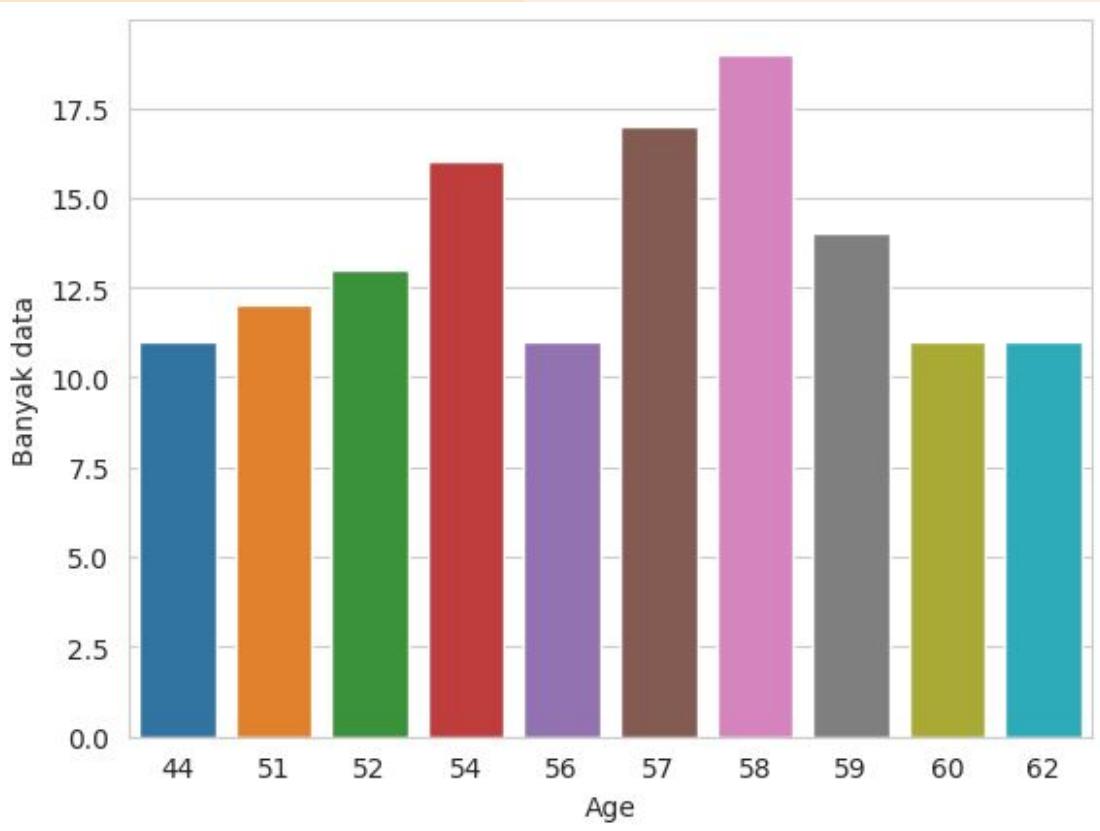


Fitur yang paling berpengaruh terhadap target adalah

- Pengaruh positif (urutan dari yang paling kuat) = cp, thalach, dan slope. Sedangkan
- Penagruh negatif (berkebalikan, dari urutan terkuat) = exang, oldpeak, ca

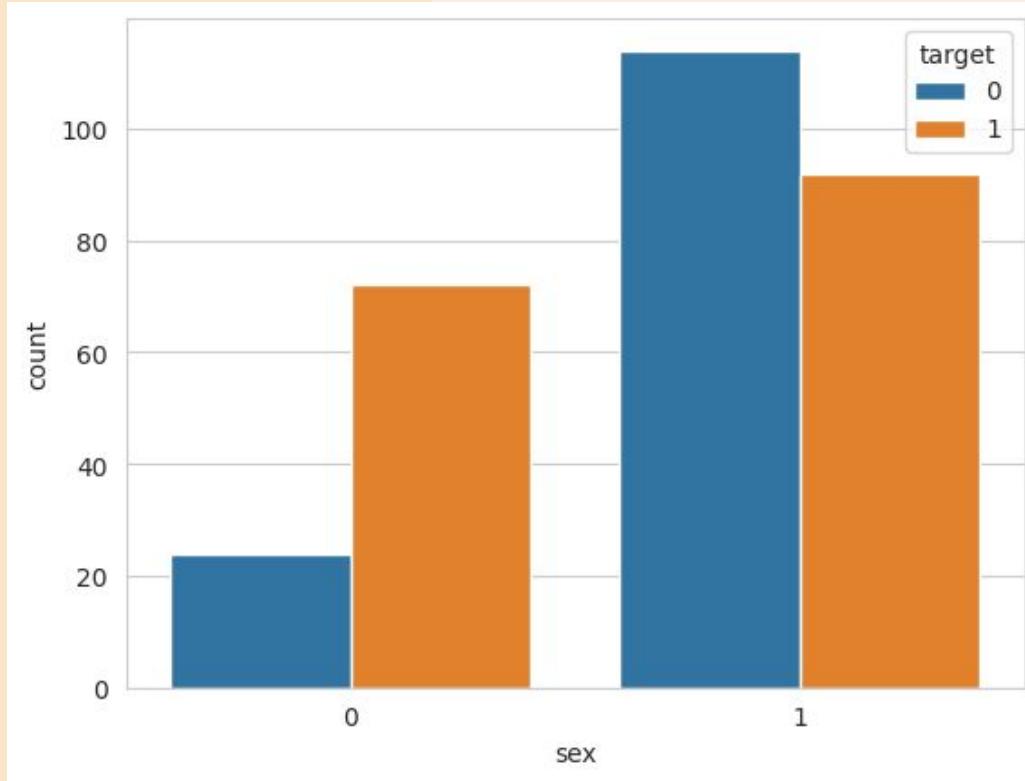


Gambaran usia yang ada di dataset



Usia Minimum: 29
Usia Maximum: 77
Rata-rata Usia: 54.42

Perbandingan jumlah data berdasarkan gender dengan targetnya.



1	206
0	96

0 = female
1 = male

1	68.21%
0	31.79%

Melihat Ketidakseimbangan data

```
Jumlah Orang Keadaan Sehat: 138  
Jumlah Penderita Heart Disease: 164
```

Dilakukan penyeimbangan data menggunakan SMOTE

```
1 from imblearn.over_sampling import SMOTE  
2  
3 # Memisahkan fitur dan target  
4 X = df.drop('target', axis=1)  
5 y = df['target']  
6  
7 # Menerapkan SMOTE  
8 smote = SMOTE(sampling_strategy='auto', random_state=42)  
9 X_resampled, y_resampled = smote.fit_resample(X, y)  
10  
11 # Menggabungkan fitur dan target kembali  
12 df_resampled = pd.concat([pd.DataFrame(X_resampled, columns=X.columns), pd.DataFrame(y_resampled, columns=['target'])], axis=1)
```

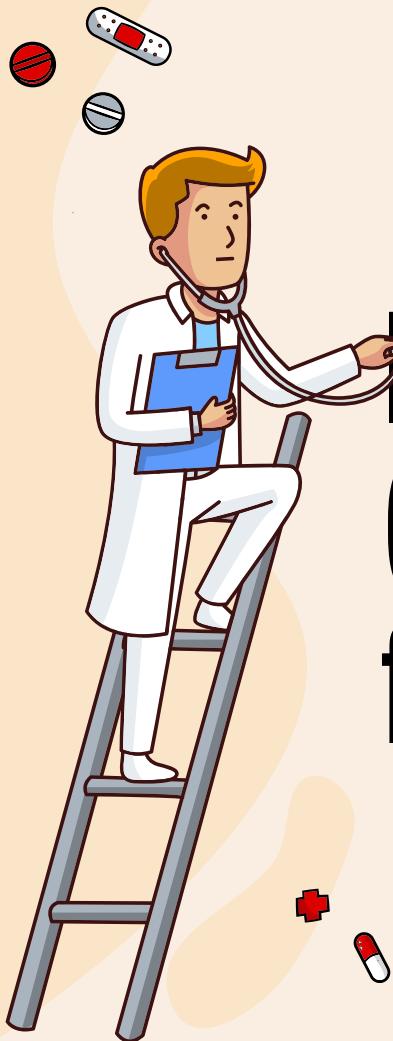
Hasil penyeimbangan data

```
Jumlah Penderita Heart Disease: 164  
Jumlah Orang Sehat: 164
```

Deskripsi statistic setelah dilakukan penyeimbangan data menggunakan SMOTE

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490	149.569536	0.327815	1.025166	1.397351	0.718543	2.314570	0.543046
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027	22.903527	0.470196	1.122717	0.616274	1.006748	0.613026	0.498970
min	29.00000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.00000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.250000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.50000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.500000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.00000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.00000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	5.600000	2.000000	4.000000	3.000000	1.000000

Feature Engineering (membuat minimal 10 fitur baru)



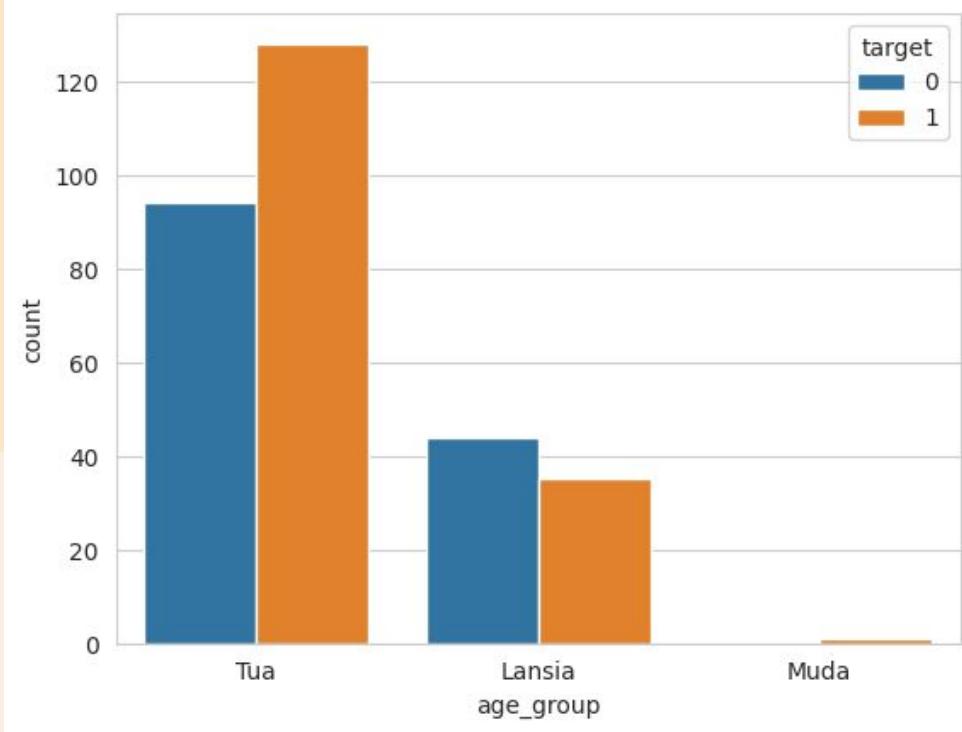
1. age_group

```
1 #Membuat Kelompok Usia  
2  
3 df['age_group'] = np.where(df['age'] < 30, 'Muda', np.where(df['age'] <= 60, 'Tua', 'Lansia'))  
4
```

```
1 correlation = df['age_group'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
0.1160499123656859
```

Hasil uji korelasi dengan target. Berkorelasi positif dengan tingkat kekuatan 0.116



Paling banyak penderita heart disease berasal dari kalangan orang yang sudah tua.

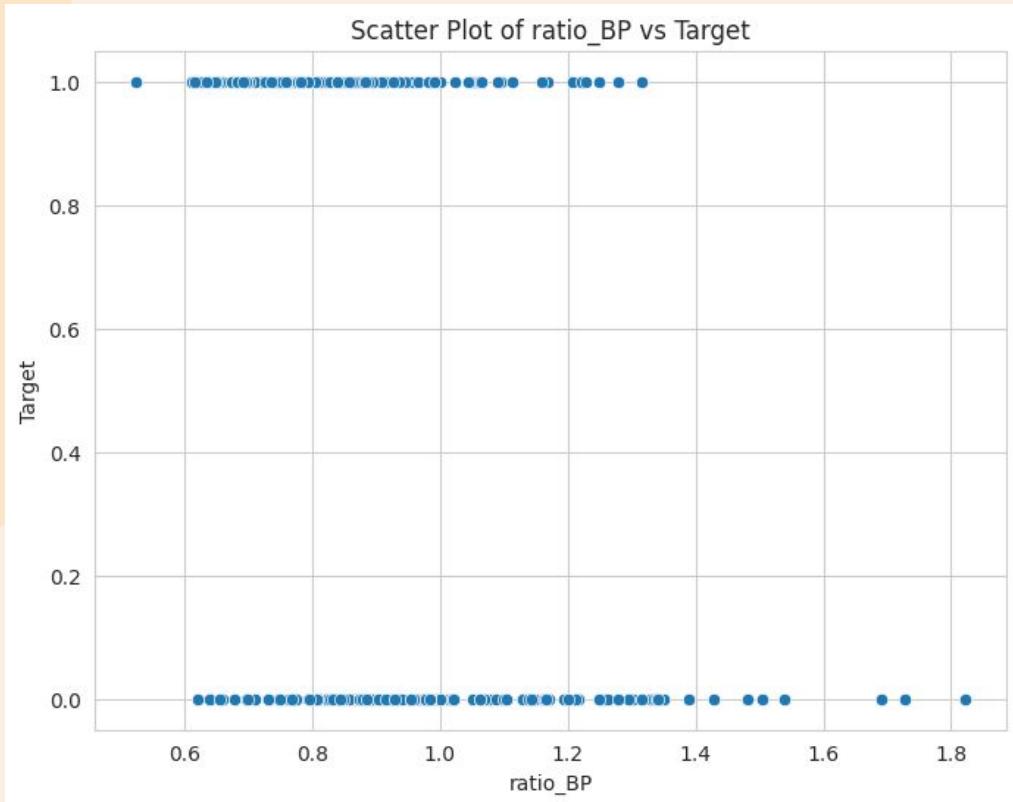
2. ratio_BP (Rasio Tekanan Darah (trestbps) terhadap Detak Jantung Maksimal (thalach))

```
1 df['ratio_BP'] = df['trestbps'] / df['thalach']
```

```
1 correlation = df['ratio_BP'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
-0.40410691343130134
```

- Hasil uji korelasi dengan target. Dan ternyata berkorelasi negatif dengan target dengan tingkatan sebesar -0,2021



Koefisien Korelasi (Pearson):
-0.39666687706012066

Cenderung sama pada ratio_BP rendah untuk kelas 0 dan 1. Namun semakin tinggi ratio_Bpnya maka semakin berpeluang menjadi orang yang sehat.

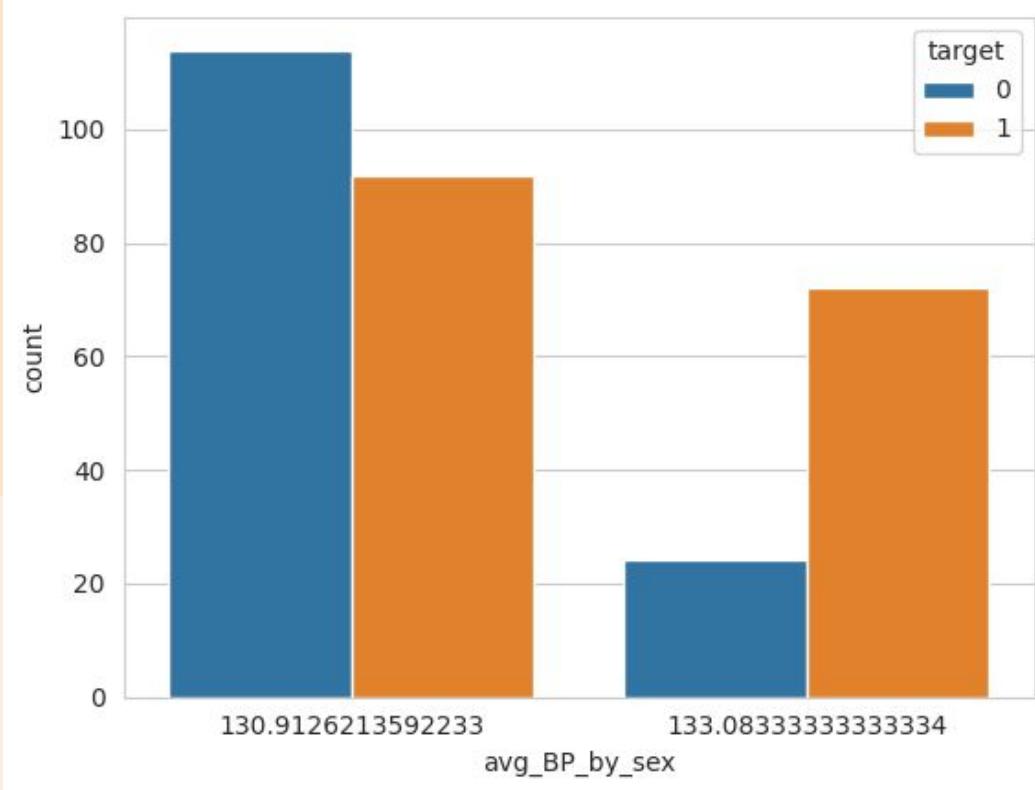
3. Rata-rata Tekanan Darah berdasarkan sex

```
1 avg_bp_by_sex = df.groupby('sex')['trestbps'].mean().to_dict()
2 df['avg_BP_by_sex'] = df['sex'].map(avg_bp_by_sex)
```

```
1 correlation = df['avg_BP_by_sex'].astype('category').cat.codes.corr(df['target'])
2 correlation
```

```
0.28360935779586227
```

Hasil uji korelasi dengan target. Berkorelasi positif dengan tingkat kekuatan 0.2836



130.9 merupakan rata-rata laki-laki. Dan 133.08 merupakan Perempuan. Dan memang terlihat presentase jumlah dan penderita heart disease lebih banyak dibanding Perempuan.

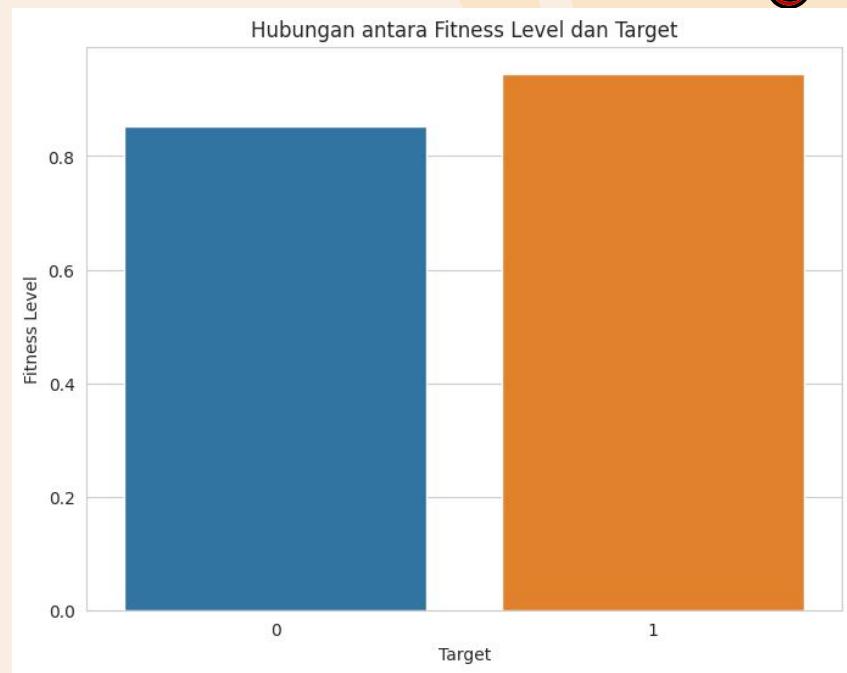
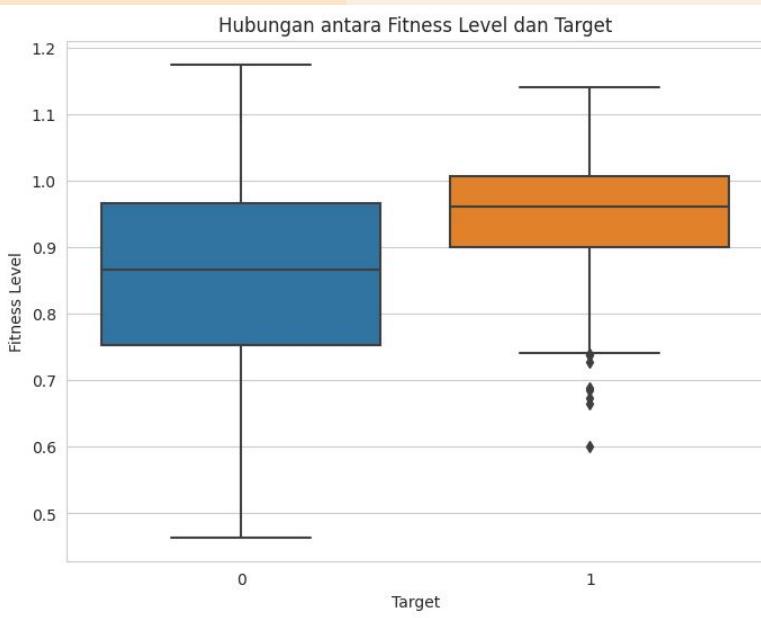
4. Fitness Level

```
1 df['fitness_level'] = df['thalach'] / (220 - df['age'])
```

```
1 correlation = df['fitness_level'].astype('category').cat.codes.corr(df['target'])
2 correlation
```

```
0.35783530399060565
```

Hasil uji korelasi dengan target. Berkorelasi positif dengan tingkat kekuatan 0.3578



Gambaran hubungan Fitness Level dengan target.

5. Rata-rata Kolesterol (chol) berdasarkan Jenis Kelamin (sex)

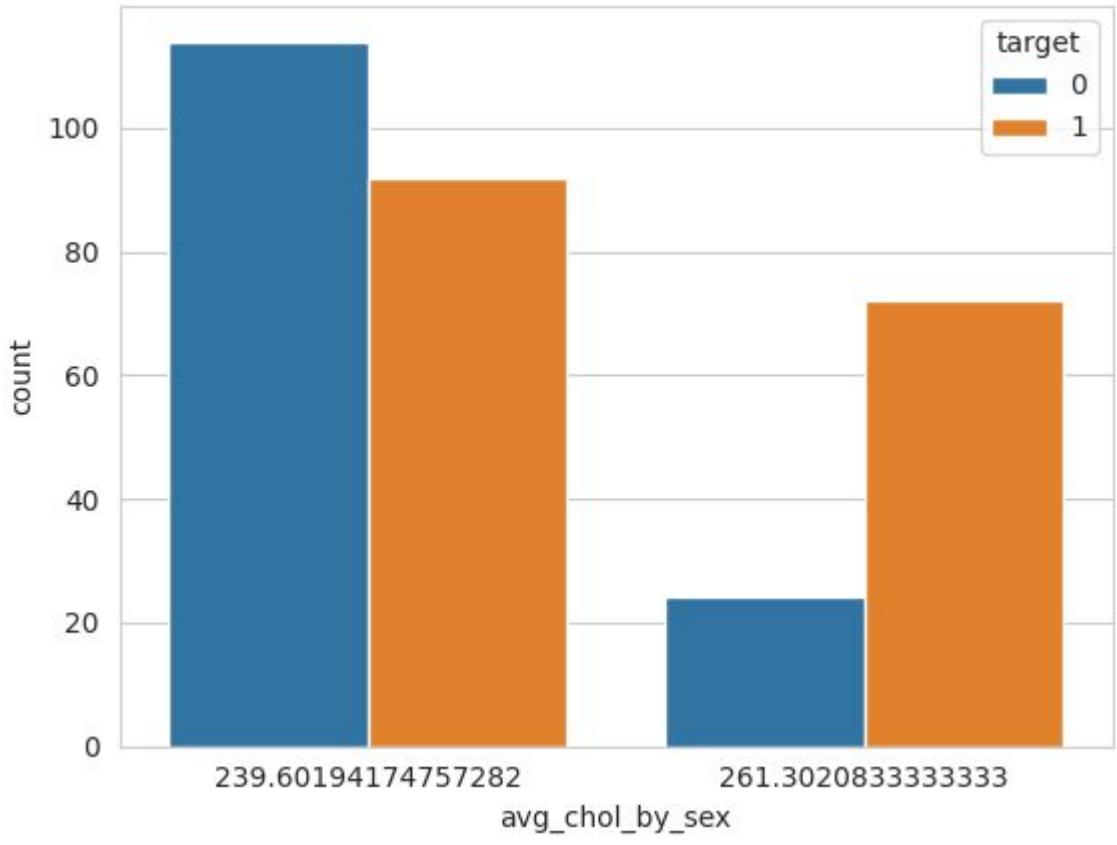
```
1 avg_chol_by_sex = df.groupby('sex')['chol'].mean().to_dict()
2 df['avg_chol_by_sex'] = df['sex'].map(avg_chol_by_sex)
```

```
1 correlation = df['avg_chol_by_sex'].astype('category').cat.codes.corr(df['target'])
2 correlation
```

```
0.28360935779586227
```

Hasil uji korelasi dengan target. Berkorelasi positif dengan tingkat kekuatan 0.2836





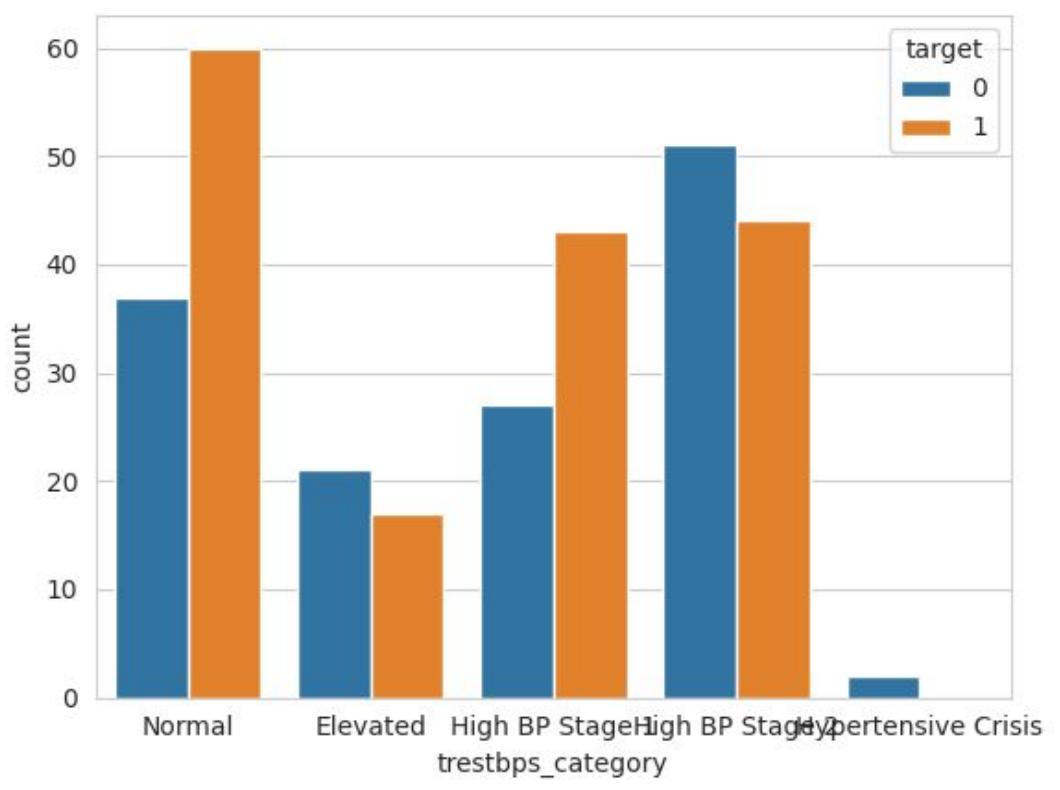
Rata-rata chol laki-laki cenderung lebih rendah dibanding rata-rata chol Perempuan.

6. Rasio Kolesterol terhadap tekanan darah

```
[101] 1 df['trestbps_category'] = pd.cut(  
2     df['trestbps'],  
3     bins=[0, 120, 129, 139, 180, float('inf')], # Definisikan batas kategori sesuai kebutuhan  
4     labels=['Normal', 'Elevated', 'High BP Stage 1', 'High BP Stage 2', 'Hypertensive Crisis']  
5 )
```

```
[103] 1 correlation = df['trestbps_category'].astype('category').cat.codes.corr(df['target'])  
2 correlation  
  
-0.11064540565467902
```

Hasil uji korelasi dengan target. Berkorelasi negatif dengan tingkat kekuatan -0,1106



Penderita Heart Disease cenderung menimpa orang dengan kategori trestbpsnya adalah normal, High BP Stage 1 dan High BP stage 2.

7. Skor Risiko Kardiovaskular

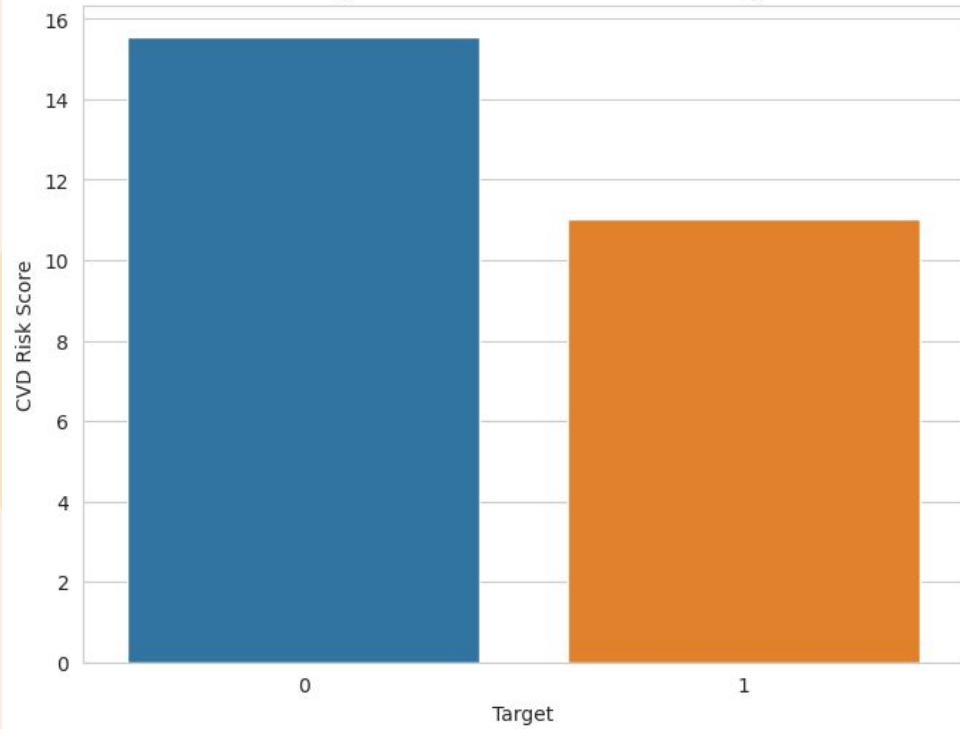
```
1 df['cvd_risk_score'] = (df['age'] * 0.2) + (df['chol'] * 0.1) - (df['thalach'] * 0.15)
```

```
1 correlation = df['cvd_risk_score'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
-0.33290664261815184
```

Hasil uji korelasi dengan target menunjukkan hubungan yang berkebalikan, dengan tingkatan -0.3329

Hubungan antara CVD Risk Score dan Target



Bisa dibilang bahwa semakin tinggi resikonya, maka semakin tinggi orang dikategorikan dalam keadaan sehat.

8. Rasio Tekanan Darah (trestbps) terhadap kolesterol (chol)

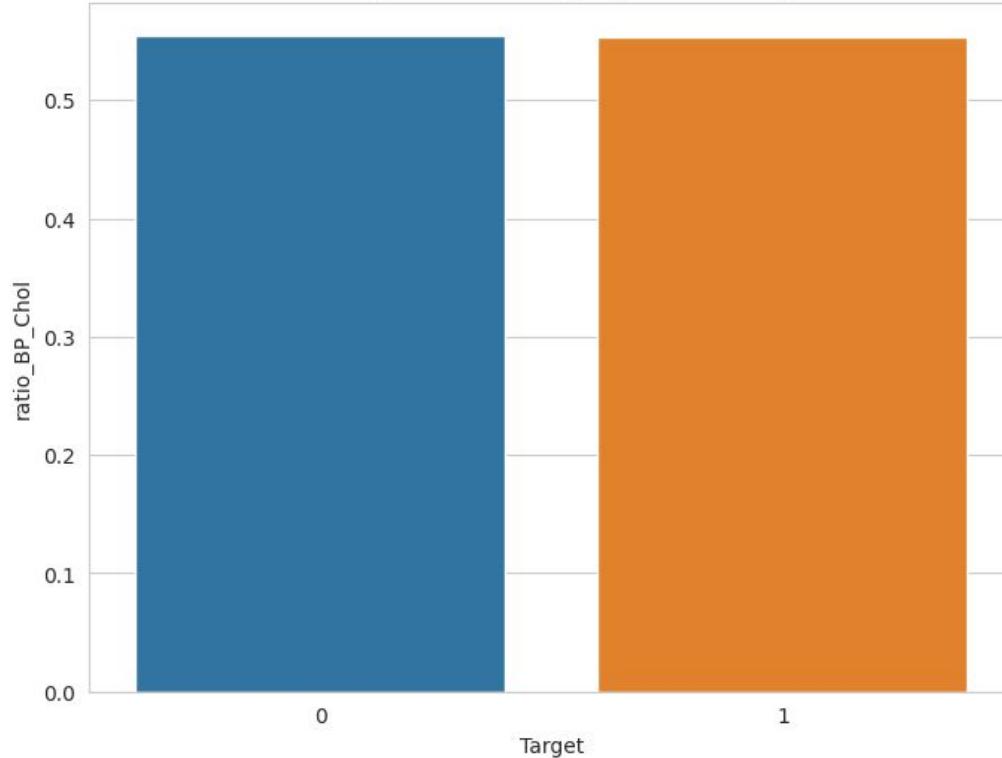
```
1 df['ratio_BP_Chol'] = df['trestbps'] / df['chol']
```

```
1 correlation = df['ratio_BP_Chol'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
0.01322900264032367
```

Hasil uji korelasi dengan target menunjukkan hubungan yang positif, namun hanya sebesar 0.01322.

Hubungan antara ratio_BP_Chol dan Target



Bisa dibilang tidak ada perbedaan antara ratio_BP_Chol dengan target.

9. Indeks Glikemik

```
1 df['glycemic_index'] = df['chol'] + (df['fbs'] * 100)  
2
```

```
1 correlation = df['glycemic_index'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
-0.1007912312396445
```

Hasil uji korelasi dengan target menunjukkan hubungan yang negative/berkebalikan, dengan tingkat korelasi sebesar -0.1008



Kelas untuk orang dalam keadaan sehat untuk glycemic_index presentasenya bisa dibilang cukup tinggi.

10. Tingkat Kemiringan (Slope) dari serangan jantung

```
1 df['slope_risk'] = df['slope'] * (df['ca'] + 1)
```

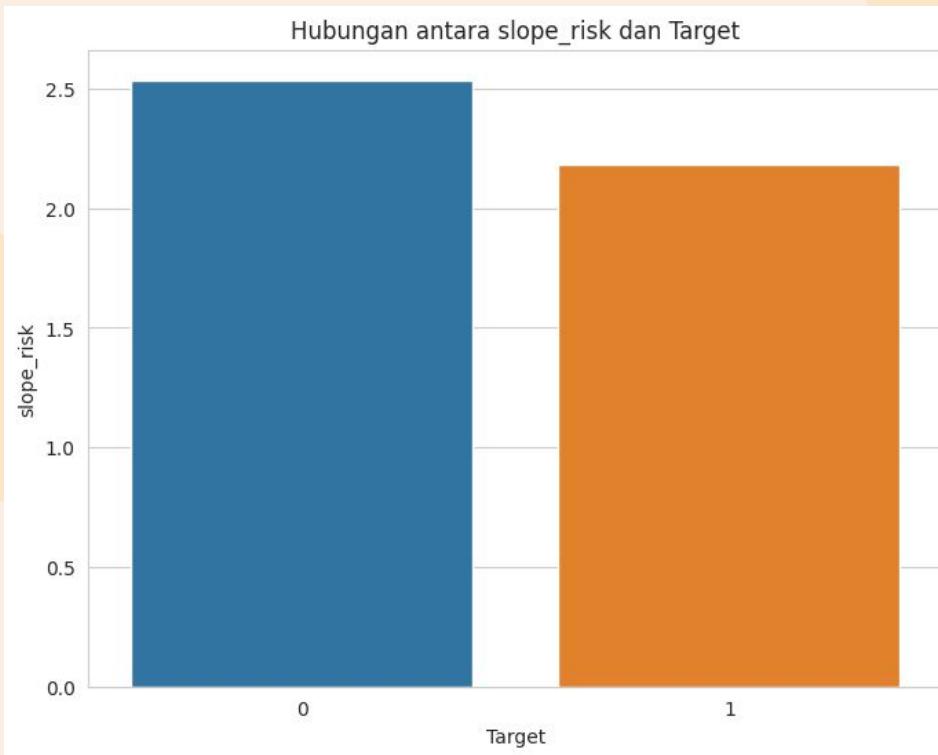
```
1 correlation = df['slope_risk'].astype('category').cat.codes.corr(df['target'])  
2 correlation
```

```
-0.11536898533453091
```



- Hasil uji korelasi dengan target menunjukkan hubungan yang negative/berkebalikan, dengan tingkat korelasi sebesar -0.1153

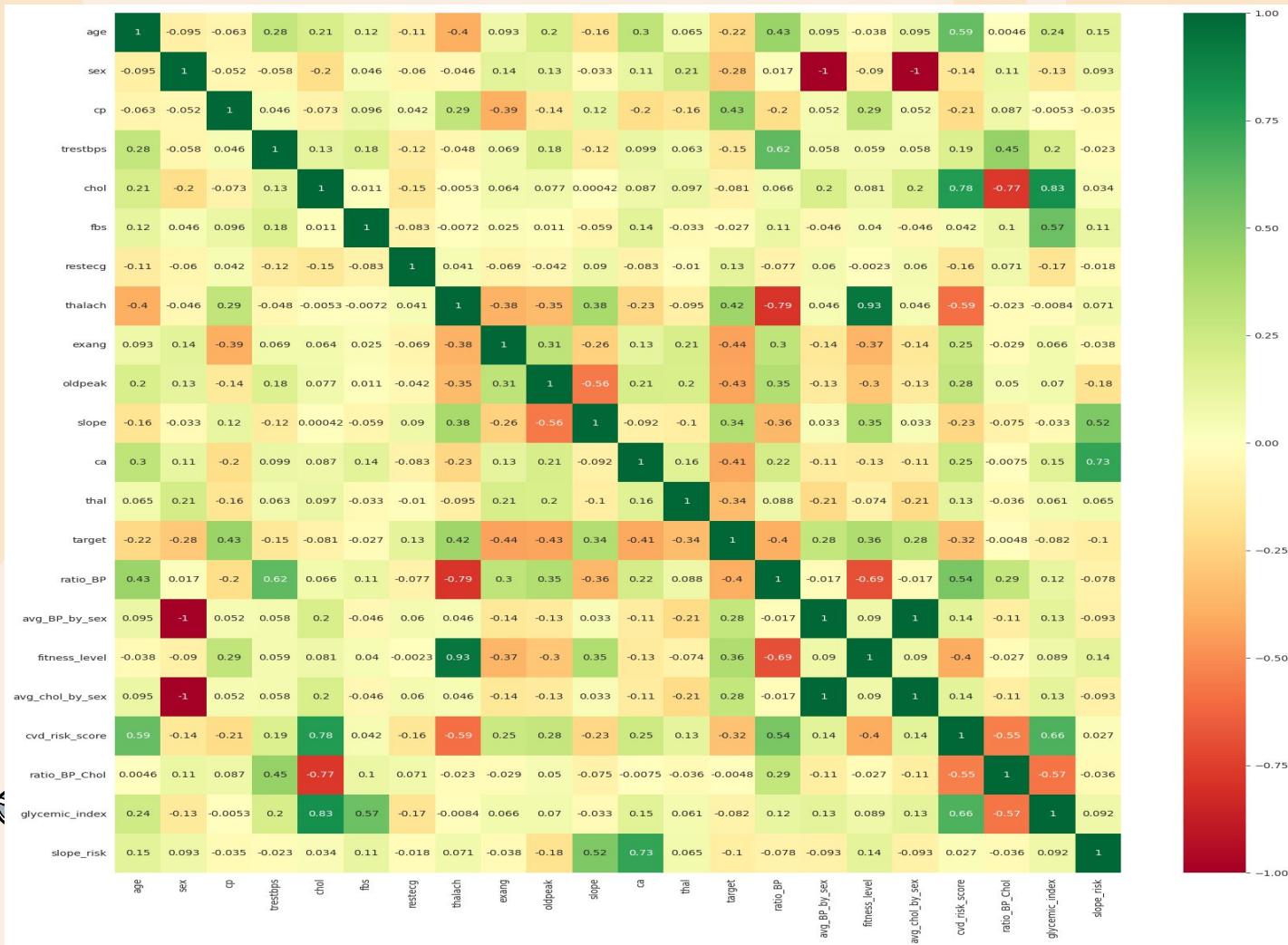




Kelas untuk orang dalam keadaan sehat jika dilihat dari slope_risk nya juga cenderung lebih tinggi jika dibanding penderita heart disease.

Hasil Featur Engineering

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 302 entries, 0 to 878
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   age              302 non-null    int64  
 1   sex              302 non-null    int64  
 2   cp               302 non-null    int64  
 3   trestbps         302 non-null    int64  
 4   chol              302 non-null    int64  
 5   fbs               302 non-null    int64  
 6   restecg          302 non-null    int64  
 7   thalach           302 non-null    int64  
 8   exang             302 non-null    int64  
 9   oldpeak           302 non-null    float64 
 10  slope             302 non-null    int64  
 11  ca                302 non-null    int64  
 12  thal              302 non-null    int64  
 13  target            302 non-null    int64  
 14  age_group         302 non-null    object  
 15  ratio_BP          302 non-null    float64 
 16  avg_BP_by_sex    302 non-null    float64 
 17  fitness_level    302 non-null    float64 
 18  avg_chol_by_sex  302 non-null    float64 
 19  trestbps_category 302 non-null    category 
 20  cvd_risk_score   302 non-null    float64 
 21  ratio_BP_Chol    302 non-null    float64 
 22  glycemic_index   302 non-null    int64  
 23  slope_risk        302 non-null    int64  
dtypes: category(1), float64(7), int64(15), object(1)
```



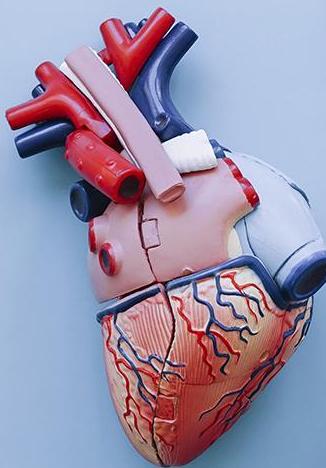
Berdasarkan heatmap yang ditampilkan, beberapa fitur yang paling berpengaruh terhadap target adalah:

Berkorelasi positif: cp, thalach, fitness_level dan slope

Berkorelasi negatif :exang, oldpeak, ca, ratio_BP

Dan mungkin dari fitur-fitur yang paling berpengaruh tersebut, bisa dijadikan rekomendasi untuk fitur-fitur tersebut saja yang dimasukkan dalam pemodelan.

EDA





Uraian fitur dan tipe data statistiknya (nominal, ordinal, numerik) pada Dataset Heart Disease

```
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype  
 --- 
 0   age               302 non-null    int64  
 1   sex               302 non-null    int64  
 2   cp                302 non-null    int64  
 3   trestbps          302 non-null    int64  
 4   chol              302 non-null    int64  
 5   fbs               302 non-null    int64  
 6   restecg           302 non-null    int64  
 7   thalach            302 non-null    int64  
 8   exang             302 non-null    int64  
 9   oldpeak            302 non-null    float64 
 10  slope              302 non-null    int64  
 11  ca                302 non-null    int64  
 12  thal               302 non-null    int64  
 13  target             302 non-null    int64  
 14  age_group          302 non-null    object  
 15  ratio_BP           302 non-null    float64 
 16  avg_BP_by_sex      302 non-null    float64 
 17  fitness_level       302 non-null    float64 
 18  avg_chol_by_sex     302 non-null    float64 
 19  trestbps_category  302 non-null    category 
 20  cvd_risk_score      302 non-null    float64 
 21  ratio_BP_Chol       302 non-null    float64 
 22  glycemic_index       302 non-null    int64  
 23  slope_risk           302 non-null    int64  
dtypes: category(1), float64(7), int64(15), object(1)
memory usage: 65.2+ KB
```

1. Age (Tipe data int, numerik)
2. Sex (Tipe data int, nominal, dengan kategori 1(laki-laki) dan 0 (perempuan))
3. Cp (Tipe data int, nominal)
4. Trestbps (Tipe data int, numerik)
5. Chol (Tipe data float, numerik)
6. Fbs (Tipe data int, nominal)
7. Restecg (Tipe data int, nominal)
8. Thalach (Tipe data int, numerik)
9. Exang (Tipe data int, nominal)
10. Oldpeak (Tipe data float, numerik)
11. Slope (Tipe data int, nominal)
12. Ca (Tipe data int, nominal)
13. Thal (Tipe data int, nominal)
14. Target (Tipe data int, nominal)

Data columns (total 24 columns):

#	Column	Non-Null Count	Dtype
0	age	302 non-null	int64
1	sex	302 non-null	int64
2	cp	302 non-null	int64
3	trestbps	302 non-null	int64
4	chol	302 non-null	int64
5	fbs	302 non-null	int64
6	restecg	302 non-null	int64
7	thalach	302 non-null	int64
8	exang	302 non-null	int64
9	oldpeak	302 non-null	float64
10	slope	302 non-null	int64
11	ca	302 non-null	int64
12	thal	302 non-null	int64
13	target	302 non-null	int64
14	age_group	302 non-null	object
15	ratio_BP	302 non-null	float64
16	avg_BP_by_sex	302 non-null	float64
17	fitness_level	302 non-null	float64
18	avg_chol_by_sex	302 non-null	float64
19	trestbps_category	302 non-null	category
20	cvd_risk_score	302 non-null	float64
21	ratio_BP_Chol	302 non-null	float64
22	glycemic_index	302 non-null	int64
23	slope_risk	302 non-null	int64

dtypes: category(1), float64(7), int64(15), object(1)
memory usage: 65.2+ KB

15. age_group(Tipe data object, ordinal)
16. ratio_BP(Tipe data float, numerik)
17. avg_BP_by_sex(tipe data float, numerik)
18. fitness_level(Tipe data float, numerik)
19. avg_chol_by_sex(Tipe data float, numerik)
20. trestbps_category(Tipe data category, ordinal)
21. cvd_risk_score (Tipe data float, numerik)
22. ratio_BP_Chol (Tipe data float, numerik)
23. glycemic_index (Tipe data int, numerik)
24. slope_risk (Tipe data int, numerik)

Summary Statistical Setelah data diberi treatment

	age	sex	cp	trestbps	chol	fb	restecg	thalach	exang	oldpeak	...	thal	target	ratio_BP	avg_BP_by_sex	fitness_level	avg_chol_by_sex	cvd_risk_score	ratio_BP_Chol	glycemic_index	slope_risk
count	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	...	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490	149.569536	0.327815	1.025166	...	2.314570	0.543046	0.904987	131.602649	0.903048	246.500000	13.098675	0.554407	261.400662	2.344371
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027	22.903527	0.470196	1.122717	...	0.613026	0.498970	0.208604	1.012476	0.128144	10.121505	7.130641	0.127403	63.189125	1.698334
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	...	0.000000	0.000000	0.525140	130.912621	0.464052	239.601942	-4.100000	0.203901	131.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.250000	0.000000	0.000000	...	2.000000	0.000000	0.758242	130.912621	0.832557	239.601942	8.225000	0.462034	216.250000	1.000000
50%	55.500000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.500000	0.000000	0.800000	...	2.000000	1.000000	0.865432	130.912621	0.928358	239.601942	12.550000	0.537889	251.000000	2.000000
75%	61.000000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000	166.000000	1.000000	1.600000	...	3.000000	1.000000	0.992955	133.083333	1.000000	261.302083	17.487500	0.634177	301.500000	3.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	5.600000	...	3.000000	1.000000	1.822222	133.083333	1.174699	261.302083	45.800000	1.190476	564.000000	10.000000

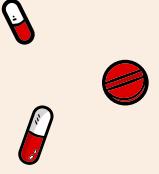
8 rows × 22 columns

Data columns (total 24 columns):

#	Column	Non-Null Count	Dtype
0	age	302 non-null	int64
1	sex	302 non-null	int64
2	cp	302 non-null	int64
3	trestbps	302 non-null	int64
4	chol	302 non-null	int64
5	fbs	302 non-null	int64
6	restecg	302 non-null	int64
7	thalach	302 non-null	int64
8	exang	302 non-null	int64
9	oldpeak	302 non-null	float64
10	slope	302 non-null	int64
11	ca	302 non-null	int64
12	thal	302 non-null	int64
13	target	302 non-null	int64
14	age_group	302 non-null	object
15	ratio_BP	302 non-null	float64
16	avg_BP_by_sex	302 non-null	float64
17	fitness_level	302 non-null	float64
18	avg_chol_by_sex	302 non-null	float64
19	trestbps_category	302 non-null	category
20	cvd_risk_score	302 non-null	float64
21	ratio_BP_Chol	302 non-null	float64
22	glycemic_index	302 non-null	int64
23	slope_risk	302 non-null	int64

dtypes: category(1), float64(7), int64(15), object(1)
memory usage: 65.2+ KB

15. age_group(Tipe data object, ordinal)
16. ratio_BP(Tipe data float, numerik)
17. avg_BP_by_sex(tipe data float, numerik)
18. fitness_level(Tipe data float, numerik)
19. avg_chol_by_sex(Tipe data float, numerik)
20. trestbps_category(Tipe data category, ordinal)
21. cvd_risk_score (Tipe data float, numerik)
22. ratio_BP_Chol (Tipe data float, numerik)
23. glycemic_index (Tipe data int, numerik)
24. slope_risk (Tipe data int, numerik)



Penjelasan situasi dan perbedaan penggunaan mean, median, dan modus



Mean adalah nilai yang diperoleh dengan menjumlahkan semua data lalu membaginya dengan jumlah data.

$$\bar{X} = \frac{\sum X}{n}$$

Mean (Rata-rata)

Mean sangat sensitif terhadap nilai-nilai ekstrem atau outlier dalam data. Jika ada outlier yang signifikan, mean dapat menjadi representasi yang tidak baik untuk data tersebut.

Mean dapat digunakan untuk mengukur "nilai tengah" dari data numerik, seperti usia, tekanan darah, kolesterol, dll.

Mean berguna untuk mendapatkan perkiraan pusat distribusi data. Misalnya, mean usia pada data Anda adalah sekitar 54.42 tahun.

Median

Median adalah nilai tengah dalam data ketika data telah diurutkan dari nilai terkecil hingga terbesar.

$$me = X_{\frac{(n+1)}{2}}$$

Median biasanya digunakan ketika distribusi data terdistorsi oleh nilai-nilai ekstrem.

Median digunakan untuk mengukur "nilai tengah" yang lebih tahan terhadap outlier.

Median berguna ketika kita ingin mengidentifikasi titik tengah distribusi data yang tidak terpengaruh oleh outlier. Misalnya, median usia pada data adalah sekitar 55.5 tahun.



Modus adalah nilai yang paling sering muncul dalam data.

Modus digunakan untuk mengidentifikasi nilai yang paling umum dalam data kategorikal atau nominal.

Modus

Dalam data , karena sebagian besar fitur adalah numerik, modus mungkin lebih relevan untuk fitur yang bersifat kategori, seperti "cp" (jenis nyeri dada) atau "thal" (jenis cacat thalassemia).

Jika terdapat dua atau lebih nilai yang memiliki frekuensi tertinggi yang sama, data tersebut dikatakan bimodal atau multimodal.



Statistical Five Summaries

	age	sex	cp	trestbps	chol	fbp	restecg	thalach	exang	oldpeak	...	thal	target	ratio_BP	avg_BP_by_sex	fitness_level	avg_chol_by_sex	cvd_risk_score	ratio_BP_Chol	glycemic_index	slope_risk
count	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	...	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490	149.569536	0.327815	1.025166	...	2.314570	0.543046	0.904987	131.602649	0.903048	246.500000	13.098675	0.554407	261.400662	2.344371
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027	22.903527	0.470196	1.122717	...	0.613026	0.498970	0.208604	1.012476	0.128144	10.121505	7.130641	0.127403	63.189125	1.698334
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	...	0.000000	0.000000	0.525140	130.912621	0.464052	239.601942	-4.100000	0.203901	131.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.250000	0.000000	0.000000	...	2.000000	0.000000	0.758242	130.912621	0.832557	239.601942	8.225000	0.462034	216.250000	1.000000
50%	55.500000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.500000	0.000000	0.800000	...	2.000000	1.000000	0.865432	130.912621	0.928358	239.601942	12.550000	0.537889	251.000000	2.000000
75%	61.000000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000	166.000000	1.000000	1.600000	...	3.000000	1.000000	0.992955	133.083333	1.000000	261.302083	17.487500	0.634177	301.500000	3.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	5.600000	...	3.000000	1.000000	1.822222	133.083333	1.174699	261.302083	45.800000	1.190476	564.000000	10.000000

8 rows × 22 columns

Statistical Five Summaries

Nilai Min

- **Definisi:** Nilai terkecil dalam set data, menunjukkan titik terendah data tersebut.
- **Contoh:** Jika kita memiliki data pengukuran tinggi badan (dalam cm) dari lima orang: 150, 155, 160, 145, dan 170, maka nilai minimumnya adalah 145 cm.

Q1 (First Quartile):

- **Definisi:** Nilai yang membagi 25% data terendah dari yang tertinggi. Dalam kata lain, 25% data berada di bawah nilai Q1.
- **Contoh:** Jika kita mengurutkan tinggi badan dari yang terkecil ke yang terbesar dan menemukan bahwa Q1 adalah 155 cm, itu berarti 25% dari orang-orang dalam sampel memiliki tinggi di bawah 155 cm.

Median (Second Quartile):

- **Definisi:** Nilai tengah dalam set data saat diurutkan. Juga dikenal sebagai nilai Q2.
- **Contoh:** Dalam data tinggi badan yang sama, jika mediannya adalah 160 cm, ini berarti setengah dari orang-orang dalam sampel memiliki tinggi di bawah 160 cm, dan setengahnya di atas 160 cm.

Statistical Five Summaries

Q3 (Third Quartile):

- **Definisi:** Nilai yang membagi 75% data terendah dari yang tertinggi. 75% data berada di bawah nilai Q3.
- **Contoh:** Jika Q3 dalam data tinggi badan adalah 165 cm, ini berarti 75% orang dalam sampel memiliki tinggi di bawah 165 cm.

Nilai Maximum

- **Definisi:** Nilai terbesar dalam set data, menunjukkan titik tertinggi data tersebut.
- **Contoh:** Dalam data tinggi badan, jika nilai maksimumnya adalah 170 cm, itu berarti ada orang dalam sampel yang memiliki tinggi 170 cm, yang merupakan nilai tertinggi dalam sampel tersebut.

Statistical Five Summaries

```
1 df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490	149.569536	0.327815	1.025166	1.397351	0.718543	2.314570	0.543046
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027	22.903527	0.470196	1.122717	0.616274	1.006748	0.613026	0.498970
min	29.00000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.00000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.250000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.50000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.500000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.00000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.00000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	5.600000	2.000000	4.000000	3.000000	1.000000

Statistical Five Summaries

Fitur age

- Rata-rata usia pasien adalah sekitar 54.42 tahun, yang merupakan usia tengah dalam dataset ini.
- Standar deviasi yang sekitar 9.05 menunjukkan variasi yang signifikan dalam usia pasien, dengan rentang usia dari 29 hingga 77 tahun.
- Nilai minimum usia pasien adalah 29 tahun, sementara nilai maksimumnya adalah 77 tahun. Ini mencerminkan kisaran usia dalam dataset.
- Kuartil pertama (Q1) adalah sekitar 48 tahun, menandakan bahwa 25% pasien memiliki usia di bawah nilai ini.
- Median (kuartil kedua / Q2) adalah sekitar 55.5 tahun, yang merupakan nilai tengah dari dataset. Setengah dari pasien memiliki usia di bawah 55.5 tahun, dan setengahnya di atasnya.
- Kuartil ketiga (Q3) adalah sekitar 61 tahun, menunjukkan bahwa 75% pasien memiliki usia di bawah nilai ini.

Statistical Five Summaries

Fitur trestbps

- Rata-rata tekanan darah dalam keadaan istirahat adalah sekitar 131.60 mm Hg, yang merupakan indikasi tekanan darah "normal" dalam dataset ini.
- Standar deviasi yang relatif rendah (sekitar 17.56 mm Hg) menunjukkan variasi data yang terbatas, menandakan bahwa sebagian besar pasien memiliki tekanan darah yang relatif seragam di sekitar rata-rata.
- Nilai minimum tekanan darah dalam keadaan istirahat adalah 94 mm Hg, sementara nilai maksimumnya adalah 200 mm Hg. Ini mencerminkan rentang tekanan darah dalam dataset.
- Kuartil pertama (Q1) adalah sekitar 120 mm Hg, menandakan bahwa 25% pasien memiliki tekanan darah dalam keadaan istirahat di bawah nilai ini.
- Median (kuartil kedua / Q2) adalah sekitar 130 mm Hg, yang merupakan nilai tengah dari dataset. Setengah dari pasien memiliki tekanan darah dalam keadaan istirahat di bawah 130 mm Hg, dan setengahnya di atasnya.
- Kuartil ketiga (Q3) adalah sekitar 140 mm Hg, menunjukkan bahwa 75% pasien memiliki tekanan darah dalam keadaan istirahat di bawah nilai ini.
- Dengan kata lain, sebagian besar pasien memiliki tekanan darah dalam kisaran antara 120 hingga 140 mm Hg, dengan rata-rata sekitar 131.60 mm Hg. Standar deviasi yang rendah menandakan konsistensi tekanan darah di sekitar rata-rata ini.

Statistical Five Summaries

Fitur chol

- Rata-rata tingkat kolesterol adalah sekitar 246.50 mg/dL, yang mencerminkan rata-rata tingkat kolesterol pasien dalam dataset ini.
- Standar deviasi yang sekitar 51.75 mg/dL menunjukkan variasi yang cukup signifikan dalam tingkat kolesterol pasien. Variabilitas ini menandakan bahwa beberapa pasien memiliki tingkat kolesterol yang jauh lebih tinggi atau lebih rendah daripada rata-rata.
- Nilai minimum untuk tingkat kolesterol adalah 126 mg/dL, sementara nilai maksimumnya adalah 564 mg/dL. Ini mencerminkan kisaran tingkat kolesterol dalam dataset.
- Kuartil pertama (Q1) adalah sekitar 211 mg/dL, menandakan bahwa 25% pasien memiliki tingkat kolesterol di bawah nilai ini.
- Median (kuartil kedua / Q2) adalah sekitar 240.5 mg/dL, yang merupakan nilai tengah dari dataset. Setengah dari pasien memiliki tingkat kolesterol di bawah 240.5 mg/dL, dan setengahnya di atasnya.
- Kuartil ketiga (Q3) adalah sekitar 274.75 mg/dL, menunjukkan bahwa 75% pasien memiliki tingkat kolesterol di bawah nilai ini.

Statistical Five Summaries

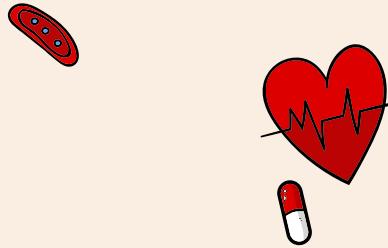
Fitur thalach

- Rata-rata tingkat detak jantung maksimum adalah sekitar 149.57 denyut per menit (bpm), yang mencerminkan rata-rata detak jantung pasien dalam dataset ini.
- Standar deviasi yang sekitar 22.90 bpm menunjukkan variasi yang cukup signifikan dalam tingkat detak jantung maksimum pasien. Variabilitas ini menandakan bahwa beberapa pasien memiliki tingkat detak jantung maksimum yang jauh lebih tinggi atau lebih rendah daripada rata-rata.
- Nilai minimum untuk tingkat detak jantung maksimum adalah 71 bpm, sementara nilai maksimumnya adalah 202 bpm. Ini mencerminkan kisaran tingkat detak jantung maksimum dalam dataset.
- Kuartil pertama (Q1) adalah sekitar 133.25 bpm, menandakan bahwa 25% pasien memiliki tingkat detak jantung maksimum di bawah nilai ini.
- Median (kuartil kedua / Q2) adalah sekitar 152.50 bpm, yang merupakan nilai tengah dari dataset. Setengah dari pasien memiliki tingkat detak jantung maksimum di bawah 152.50 bpm, dan setengahnya di atasnya.
- Kuartil ketiga (Q3) adalah sekitar 166.00 bpm, menunjukkan bahwa 75% pasien memiliki tingkat detak jantung maksimum di bawah nilai ini.

Statistical Five Summaries

Fitur oldpeak

- Rata-rata tingkat depresi ST (oldpeak) adalah sekitar 1.03. Ini menggambarkan rata-rata tingkat depresi ST pada pasien dalam dataset ini.
- Standar deviasi yang sekitar 1.12 menunjukkan variasi yang signifikan dalam tingkat depresi ST pasien. Variabilitas ini menandakan bahwa beberapa pasien memiliki tingkat depresi ST yang jauh lebih rendah atau lebih tinggi daripada rata-rata.
- Nilai minimum untuk tingkat depresi ST adalah 0.0, sementara nilai maksimumnya adalah 5.6. Ini mencerminkan kisaran tingkat depresi ST dalam dataset, yang berkisar dari tidak ada depresi ST hingga depresi ST yang signifikan.
- Kuartil pertama (Q1) adalah sekitar 0.0, menunjukkan bahwa 25% pasien memiliki tingkat depresi ST di bawah atau sama dengan nilai ini.
- Median (kuartil kedua / Q2) adalah sekitar 0.8, yang merupakan nilai tengah dari dataset. Setengah dari pasien memiliki tingkat depresi ST di bawah 0.8, dan setengahnya di atasnya.
- Kuartil ketiga (Q3) adalah sekitar 1.6, menandakan bahwa 75% pasien memiliki tingkat depresi ST di bawah atau sama dengan nilai ini.

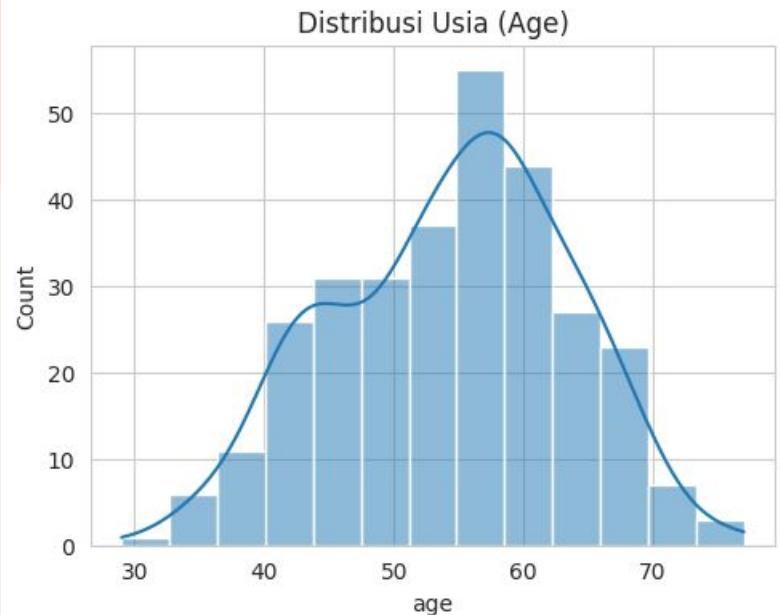


Distribusi dalam Statistika

Dalam statistika, distribusi merujuk pada cara nilai-nilai dalam suatu set data tersebar atau didistribusikan.

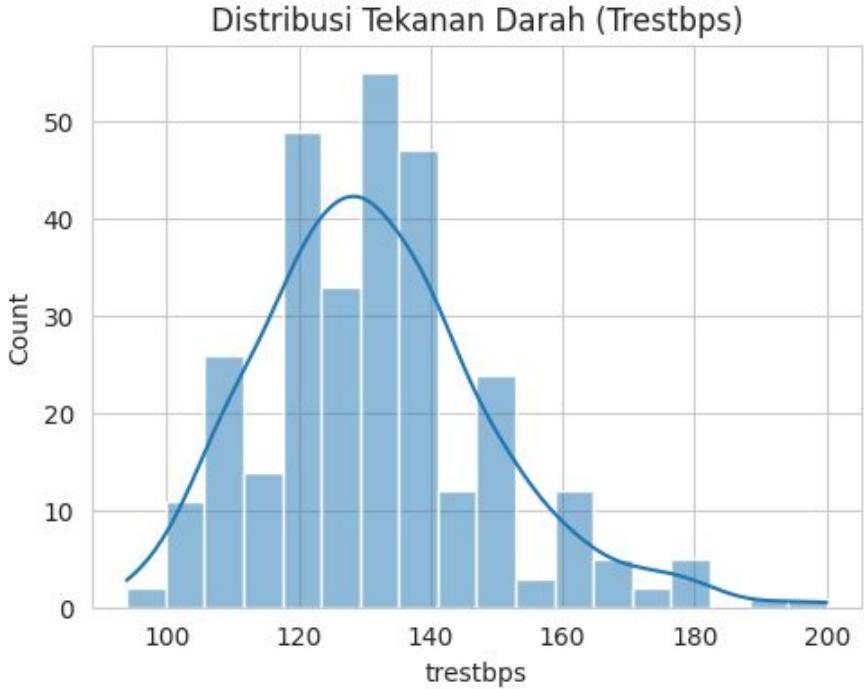


Distribusi fitur age



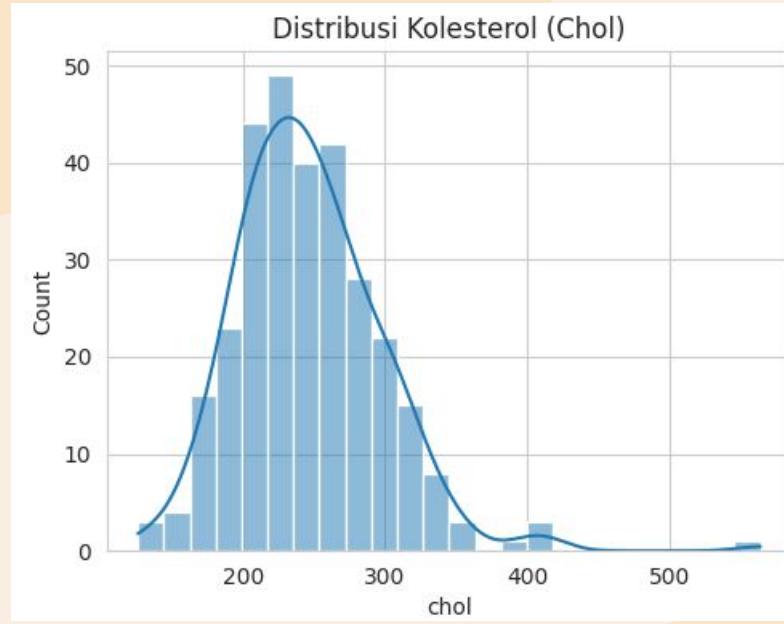
Dataset berasal dari pengukuran orang dalam rentang usia 29 hingga 77 tahun. Terlihat distribusi fitur age hampir mendekati distribusi normal, berarti data menyebar dari rentang usia tersebut

Distribusi fitur trestbps



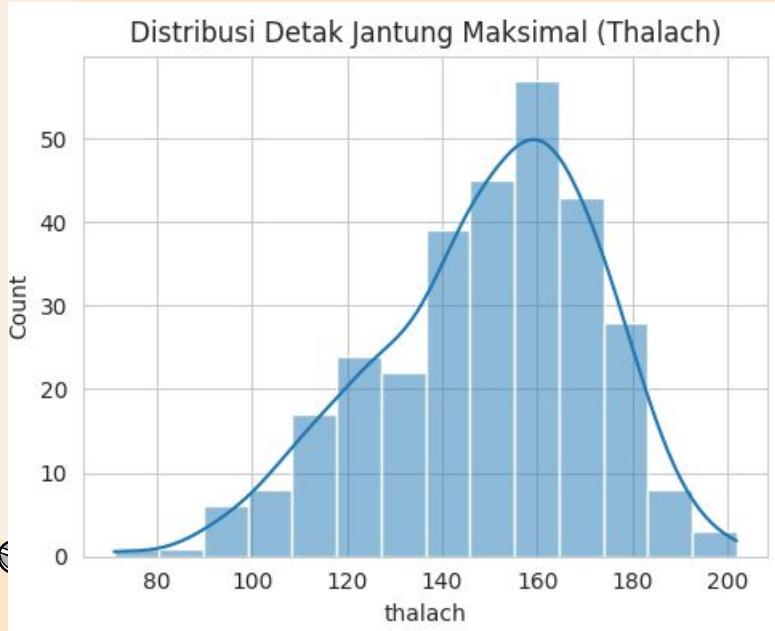
Terlihat distibusi agak right skew (banyak data poin yang mengumpul di nilai trestbps yang minim). Namun trestbps yang tinggi juga bisa terjadi pada seseorang dan mungkin saja mengindikasikan adanya penyakit. Apabila datanya lebih banyak, dan dicoba mempertahankan nilai trestbps yang tinggi itu membuat model berperforma kurang, maka mungkin selanjutnya bisa untuk menghilangkan nilai trestbps yang tinggi tersebut.

Distribusi fitur chol



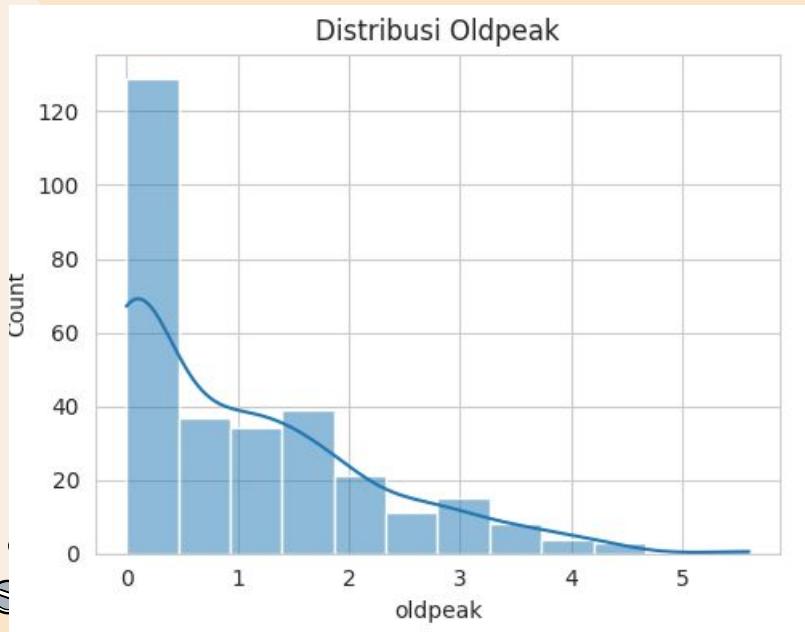
Hal yang sama terjadi pada fitur chol. Bisa juga chol dihapus untuk membuat performa model lebih optimal.

Distribusi fitur thalach

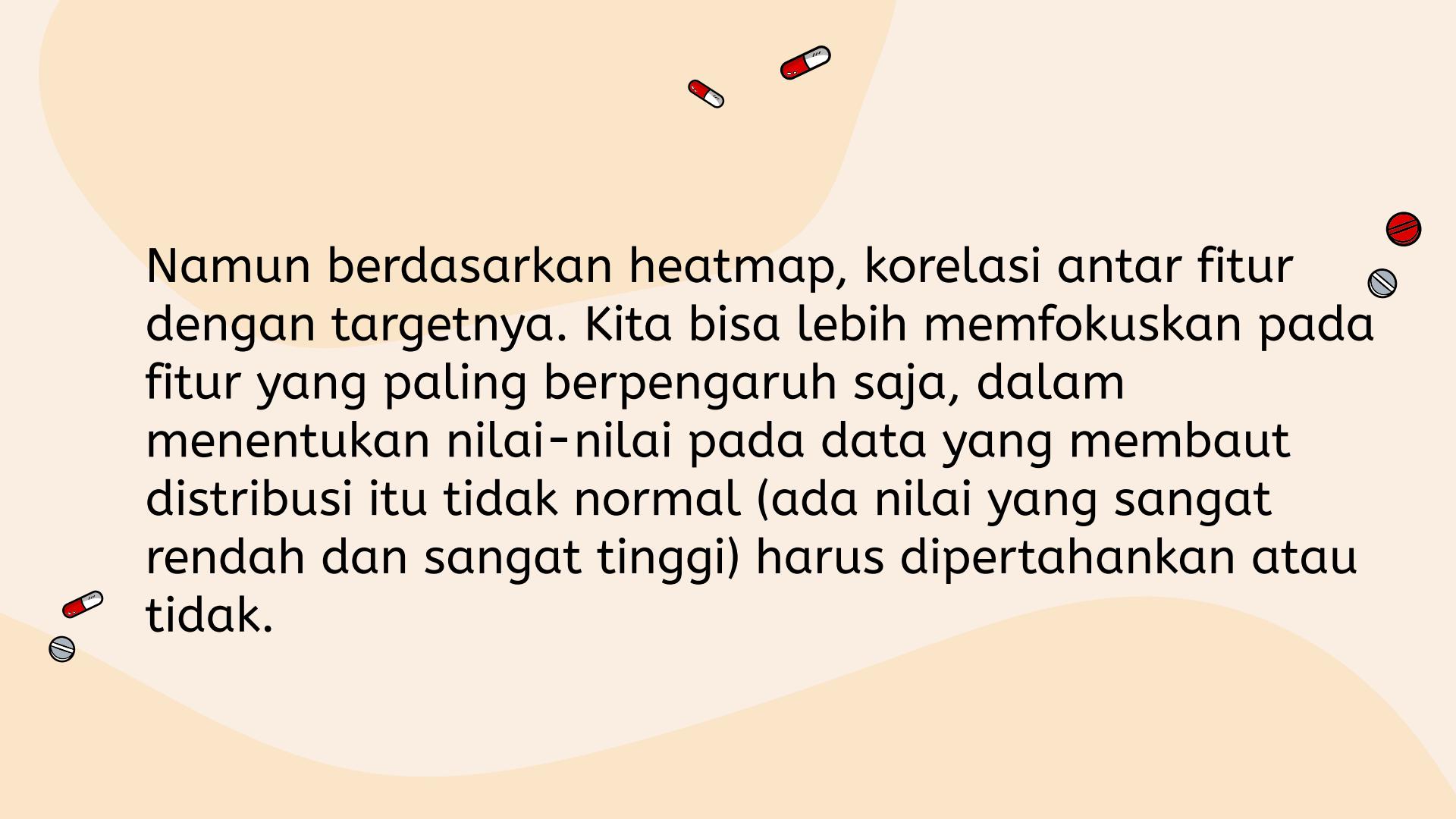


Terlihat bahwa distribusi sedikit mengarah ke left skew. Banyak data poin yang banyak berkumpul di titik maksimum data. Apabila performa model kurang jika mempertahankan thalach yang rendah, maka bisa dicoba untuk menghilangkan nilai rendah pada thalach tersebut untuk mencapai performa yang optimal. Dan bisa di pastikan kembali selanjutnya apakah pengukuran thalach seseorang bisa pada nilai-nilai rendah pada distribusi tersebut.

Distribusi fitur oldpeak



Terlihat bahwa distribusi menunjukkan right skew. Banyak data point yang ada dalam nilai oldpeak yang minimum. Apabila data lebih banyak dan performa menerapkan data ini menunjukkan hasil yang kurang. Bisa dicoba untuk menghapus saja nilai tinggi oldpeak ini, karena jumlahnya juga tidak terlalu banyak.



Namun berdasarkan heatmap, korelasi antar fitur dengan targetnya. Kita bisa lebih memfokuskan pada fitur yang paling berpengaruh saja, dalam menentukan nilai-nilai pada data yang membuat distribusi itu tidak normal (ada nilai yang sangat rendah dan sangat tinggi) harus dipertahankan atau tidak.

1. Ternyata banyak fitur yang tidak berdistribusi normal. Namun berdasarkan heatmap, korelasi antar fitur dengan targetnya. Kita bisa lebih memfokuskan pada fitur yang paling berpengaruh saja, dalam menentukan nilai-nilai pada data yang membuat distribusi itu tidak normal (ada nilai yang sangat rendah dan sangat tinggi) harus dipertahankan atau tidak.
2. Keputusan mempertahankan data yang membuat distribusi pada fitur-fitur tersebut juga bisa dilihat pada banyaknya data. Jumlah data yang minim karena nilai yang bisa dibilang sangat rendah dan sangat tinggi dihilangkan. Mampu membuat model juga berperforma rendah, karena model tidak terlalu banyak belajar dari data.

Referensi



- <https://www.sentrakalibrasiindustri.com/mean-median-modus-pengertian-rumus-dan-contoh-soalnya/#:~:text=Mean%20digunakan%20untuk%20menggambarkan%20rata,nilai%20yang%20paling%20sering%20muncul.>
- [https://www.researchgate.net/publication/368661532 Mean Median dan Modus](https://www.researchgate.net/publication/368661532_Mean_Median_dan_Modus)
- <https://socs.binus.ac.id/2018/12/08/distribusi-peluang-binomial/>
- chrome-extension://efaidnbmnnibpcajpcglclefindmkaj/https://repository.unikom.ac.id/61031/1/Pertemuan%208.pdf
- <https://jagostat.com/statistika-matematika-1/distribusi-chi-square>



Thank you

CRÉDITOS: Esta plantilla para presentaciones es una creación de Slidesgo, e incluye íconos de Flaticon, infografías e imágenes de Freepik

Por favor, conserva esta diapositiva para atribuirnos

