ASSIGNMENT 2 STUDI INDEPENDENT DSAI STARTUP CAMPUS

DATA PREPROCESSING

CREATED BY IRMA NURMALIA - KOMPI 18

BUSINESS UNDERSTANDING
HEART DISEASE DATASET

Menganalisis faktor-faktor apa saja yang paling berpengaruh sebagai basis untuk mendeteksi apakah seseorang terkena penyakit jantung atau tidak.

Hasil dari analisis ini juga bisa digunakan untuk meminimalkan jumlah cetakan brosur tentang penyakit jantung di beberapa rumah sakit.

DATASET

DATA UNDERSTANDING

	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
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
6	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
7	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
8	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
9	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 192 entries, 109 to 733
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       192 non-null    int64
 1   sex       192 non-null    int64
 2   cp        192 non-null    int64
 3   trestbps  192 non-null    int64
 4   chol      192 non-null    int64
 5   fbs       192 non-null    int64
 6   restecg   192 non-null    int64
 7   thalach   192 non-null    int64
 8   exang     192 non-null    int64
 9   oldpeak   192 non-null    float64
 10  slope     192 non-null    int64
 11  ca        192 non-null    int64
 12  thal      192 non-null    int64
 13  target    192 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 22.5 KB
```

Dataset ini terdiri dari **14** column dan **1025** baris. Semua kolom memiliki tipe data yang sama, yakni integer (int) kecuali pada kolom **oldpeak** yang memiliki tipe data float (angkanya berupa bilangan desimal).

DATA UNDERSTANDING

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	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

Berikut merupakan hasil *stastical summaries* dari ke-empat belas kolom dalam dataset.

DATA UNDERSTANDING
ATTRIBUTE INFORMATION

1. age = umur pasien
2. sex = jenis kelamin (0 = female, 1=male)
3. cp = chest pain type
    - 0 = typical angina
    - 1 = atypical angina
    - 2 = non-anginal pain
    - 3 = asymptomatic
4. trestbps = resting blood pressure
5. chol = serum cholesterol (mg/dl)
6. fbs = fasting blood sugar whose value is >120 mg/dl
7. restcg = resting electrocardiographic results
    - 0 = normal
    - 1 = having ST-T wave abnormally
    - 2 = showing probable
8. thalach = maxium heart rate achieved
9. exang = exercise induced angina (0 = no, 1 = yes)
10. oldpeak = ST depression induced by exercise relative to test

DATA UNDERSTANDING
ATTRIBUTE INFORMATION

11. slope = the slope of the peak exercise ST sergment
    - 0 = usloping
    - 1 = flat
    - 2 = downsloping
12. ca = number of major vessels colored by fluroscopy
    - 0 = tidak ada pembuluh darah utama yang terlihat
    - 1 = satu pembuluh terlihat
    - 2 = dua pembuluh terlihat
    - 3 = tiga pembuluh terlihat
    - 4 = empat pembuluh terlihat

13 thal = a blood disorder **called thalassemia**
    - **0 = tidak ada gejala thalassemia terdeteksi**
    - **1 = thalasemia ringan**
    - **2 = thalasemia sedang**
    - **3 = thalasemia berat**

14. target = the presence of heart disease in the patient (0 = no, 1 = disease)

DATA PREPARATION

1

MISSING VALUE

Terdapat dua cara, yakni **deletion** atau **imputation**

2

DUPLICATE VALUE

Bertujuan untuk menghilangkan data duplikat

3

OUTLIER

Mencari data yang nilainya secara signifikan berbeda

4

IMBALANCE DATA

Dilakukan jika ada fitur target yang jumlahnya tidak seimbang

DATA PREPARATION

1

```
data.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
```

Pada dataset tidak ada *missing value*

```
duplikat = data.duplicated()
duplikat
```

```
0       False
1       False
2       False
3       False
4       False
        ...
1020     True
1021     True
1022     True
1023     True
1024     True
Length: 1025, dtype: bool
```

2

```
duplikat = df.duplicated()
duplikat
```

```
0       False
1       False
2       False
3       False
4       False
        ...
723     False
733     False
739     False
843     False
878     False
Length: 302, dtype: bool
```

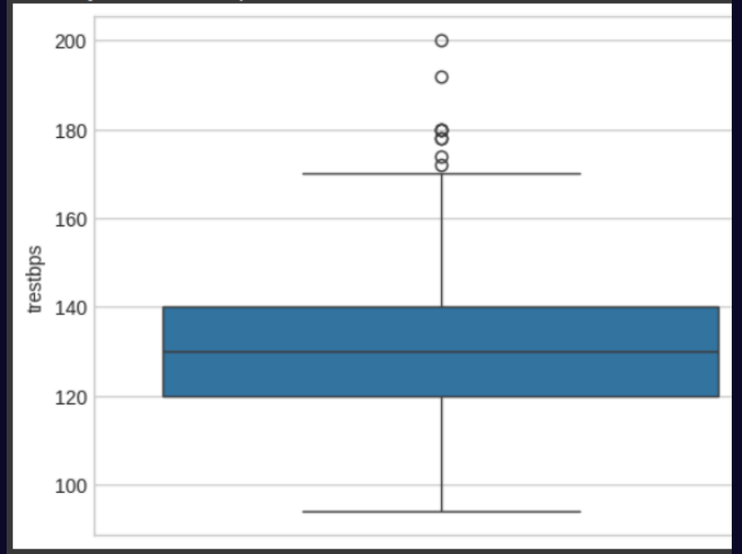
Jumlah data berkurang dari **1025** menjadi **302** setelah dilakukan *cleaning* dengan code
df = data.drop_duplicates()

DATA PREPARATION

3

```
sns.boxplot(df['trestbps'])
```

```
<Axes: ylabel='trestbps'>
```



Terdapat beberapa data outlier namun tidak perlu ditangani

```
# Imbalance Data
df['sex'].value_counts()
```

```
1    206
0     96
Name: sex, dtype: int64
```

4

```
# gabungkan data laki-laki (down-sample) dan perempuan
df_downsampled = pd.concat([df_male_dwsampled, df_female])
df_downsampled['sex'].value_counts()
```

```
1    96
0    96
Name: sex, dtype: int64
```

Metode yang dipilih adalah **Down-Sample Majority Class** agar terhindar dari data duplikat.

FEATURE ENGINEERING

MEAN

Nilai yang diperoleh dari penjumlahan semua data dibagi dengan banyaknya data

MEDIAN

Nilai tengah dari keseluruhan data yang diurutkan

MODUS

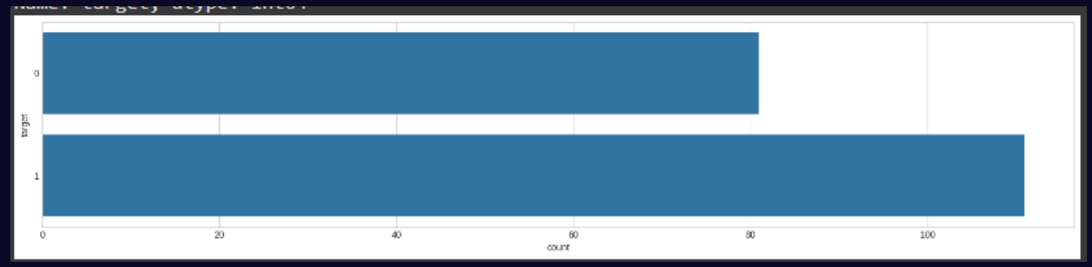
Nilai dengan frekunsei terbanyak dalam data

FEATURE ENGINEERING

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000
mean	54.963542	0.500000	0.932292	132.182292	252.390625	0.161458	0.531250	149.156250	0.312500	0.973437	1.437500	0.625000	2.312500	0.578125
std	8.771904	0.501307	0.997693	18.143786	54.238648	0.368915	0.540566	22.259925	0.464724	1.207469	0.602134	0.918113	0.602134	0.495150
min	29.000000	0.000000	0.000000	94.000000	141.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	49.000000	0.000000	0.000000	120.000000	212.000000	0.000000	0.000000	133.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	57.000000	0.500000	1.000000	130.000000	244.500000	0.000000	1.000000	154.500000	0.000000	0.600000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	282.250000	0.000000	1.000000	164.250000	1.000000	1.500000	2.000000	1.000000	3.000000	1.000000
max	76.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Berikut merupakan hasil *stastical summaries* dari ke-empat belas kolom dalam dataset setelah dilakukan proses **inbalance data**

EXPLORATORY DATA ANALYSIS



Feature : target

81
No

111
Disease

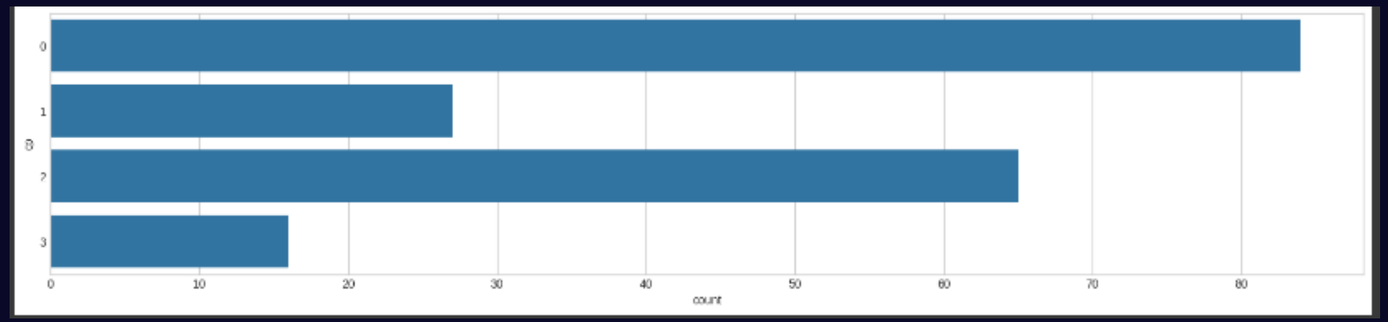
Feature : cp

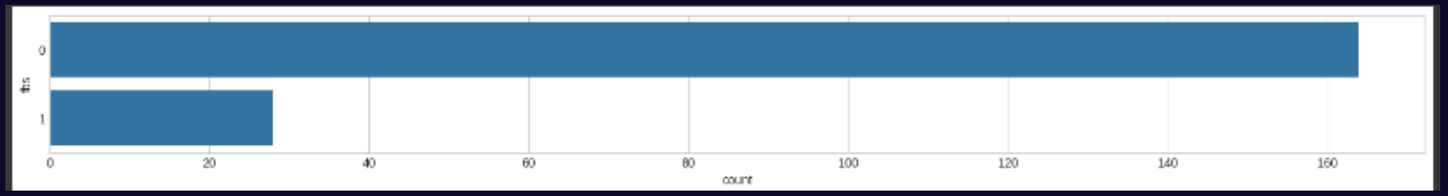
84
typical angina

27
atypical angina

65
non-anginal

16
asympto-matic





Feature : fbs > 120 mg/dl

164
False

28
Yes

EXPLORATORY DATA ANALYSIS

Feature : restecg

89
Normal

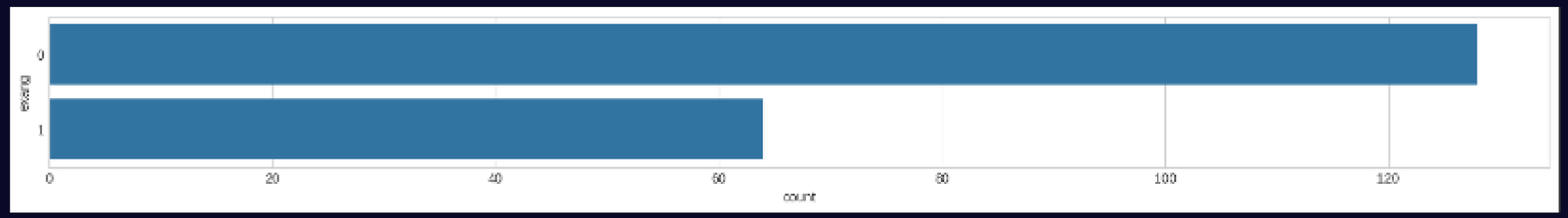
100
ST-T wave abnormal

3
Showing probable

Feature : exang

128
Tidak nyeri dada

64
Nyeri dada



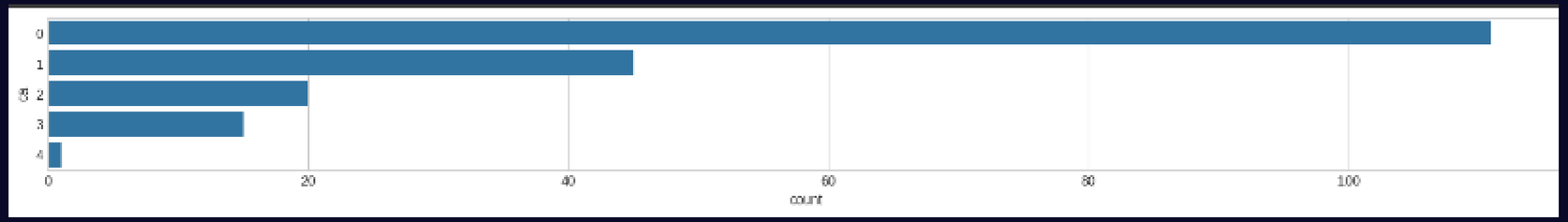
Feature : slope

11
Up-slopping

90
Flat

91
Down-slopping

EXPLORATORY DATA ANALYSIS



Feature : ca (jumlah pembuluh darah besar yang tampak

105 Tidak ada

43 Satu

31 Dua

11 Tiga

2 Empat

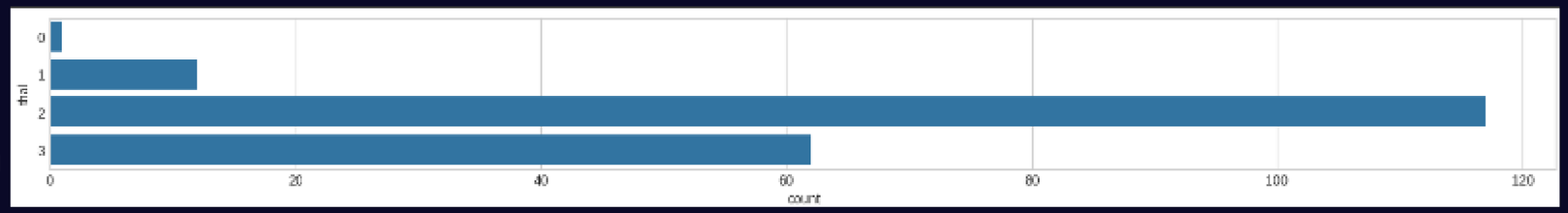
Feature : ca (jumlah pembuluh darah besar yang tampak

1 Tidak thalasemia

12 Thalasemia ringan

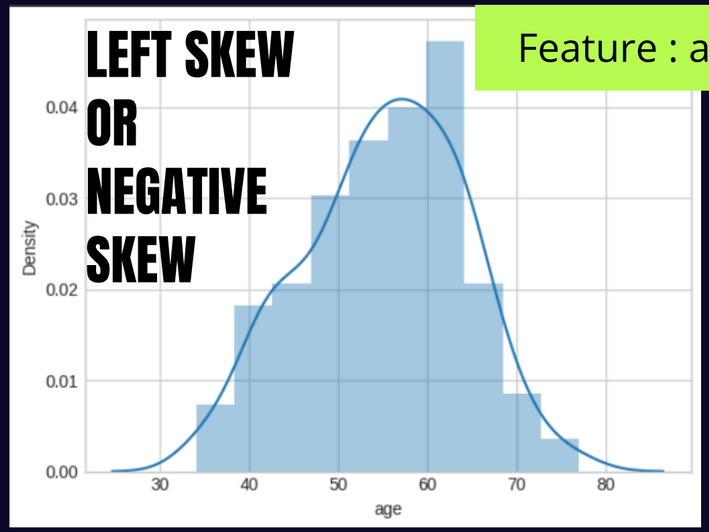
117 Thalasemia sedang

62 Thalasemia berar

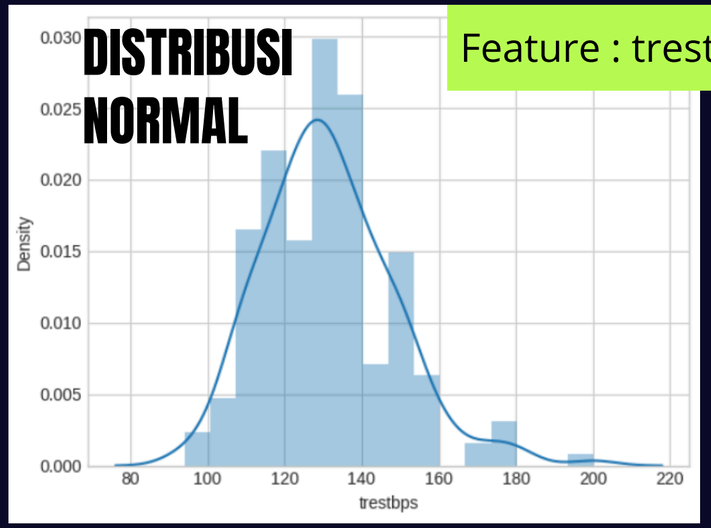


EXPLORATORY DATA ANALYSIS

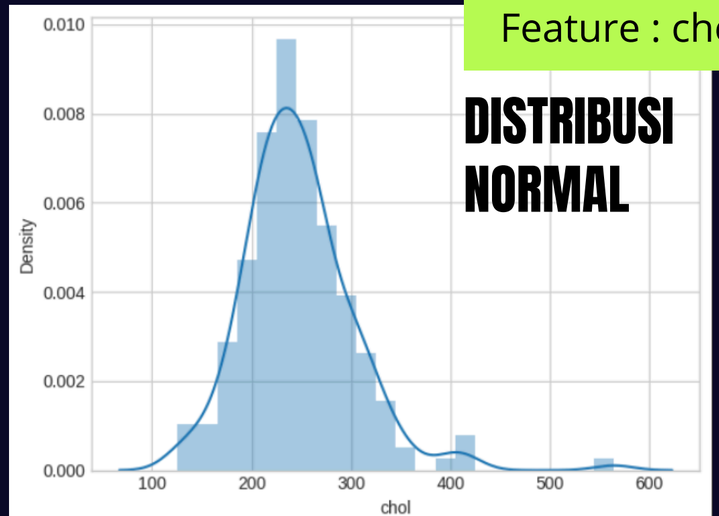
Feature : age



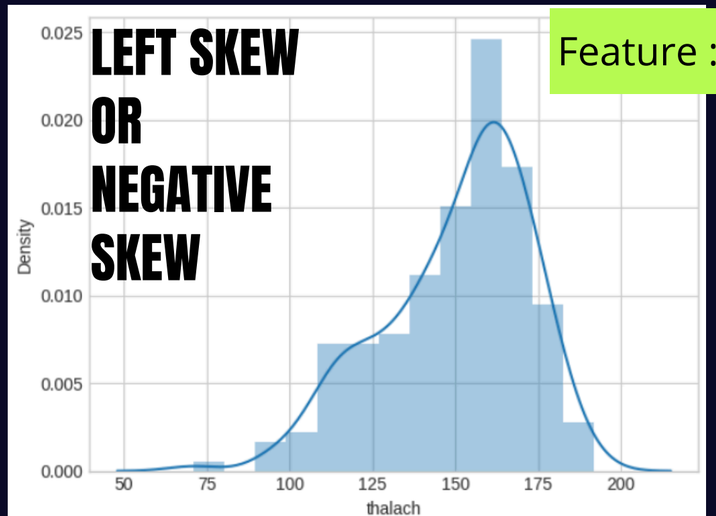
Feature : trestbps



Feature : chol



Feature : thalach



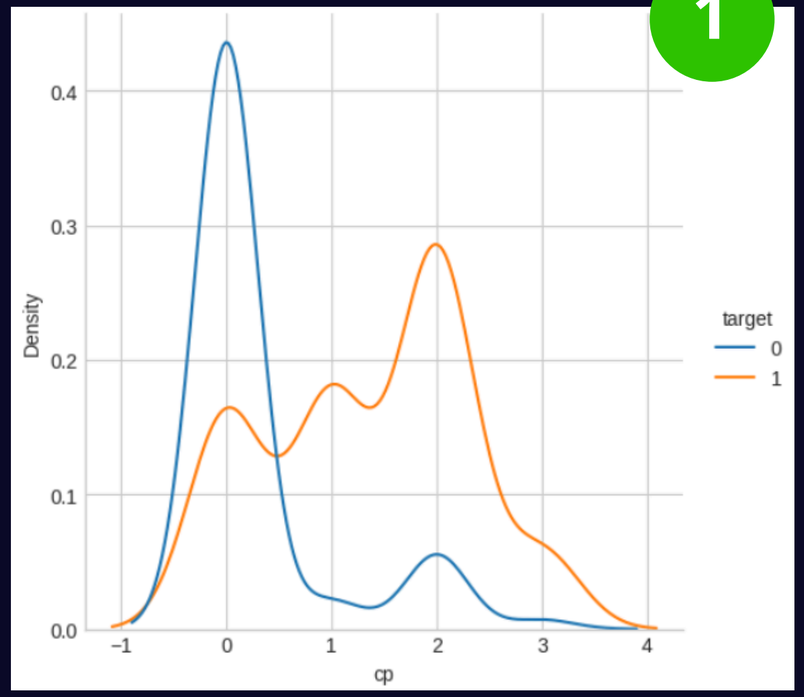
CORRELATION ANALYSIS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.06	-0.09	0.30	0.23	0.10	-0.06	-0.40	0.12	0.27	-0.18	0.32	0.08	-0.28
sex	-0.06	1.00	-0.09	-0.10	-0.25	0.10	-0.02	-0.04	0.18	0.10	0.02	0.21	0.21	-0.35
cp	-0.09	-0.09	1.00	0.01	-0.03	0.08	0.04	0.23	-0.38	-0.19	0.18	-0.21	-0.12	0.53
trestbps	0.30	-0.10	0.01	1.00	0.12	0.14	-0.07	-0.04	0.12	0.25	-0.15	0.13	0.02	-0.16
chol	0.23	-0.25	-0.03	0.12	1.00	0.09	-0.18	0.04	0.04	0.10	0.01	0.04	0.06	-0.01
fbs	0.10	0.10	0.08	0.14	0.09	1.00	-0.04	0.06	0.10	0.02	0.03	0.21	0.08	-0.08
restecg	-0.06	-0.02	0.04	-0.07	-0.18	-0.04	1.00	0.08	-0.06	-0.10	0.11	-0.07	0.08	0.07
thalach	-0.40	-0.04	0.23	-0.04	0.04	0.06	0.08	1.00	-0.32	-0.32	0.38	-0.18	0.04	0.40
exang	0.12	0.18	-0.38	0.12	0.04	0.10	-0.06	-0.32	1.00	0.20	-0.21	0.12	0.21	-0.49
oldpeak	0.27	0.10	-0.19	0.25	0.10	0.02	-0.10	-0.32	0.20	1.00	-0.56	0.31	0.25	-0.47
slope	-0.18	0.02	0.18	-0.15	0.01	0.03	0.11	0.38	-0.21	-0.56	1.00	-0.11	-0.09	0.36
ca	0.32	0.21	-0.21	0.13	0.04	0.21	-0.07	-0.18	0.12	0.31	-0.11	1.00	0.06	-0.49
thal	0.08	0.21	-0.12	0.02	0.06	0.08	0.08	0.04	0.21	0.25	-0.09	0.06	1.00	-0.27
target	-0.28	-0.35	0.53	-0.16	-0.01	-0.08	0.07	0.40	-0.49	-0.47	0.36	-0.49	-0.27	1.00

Berdasarkan hasil analisis diatas, terdapat beberapa faktor yang berkorelasi positifi dengan target, yakni **Chest pain type (cp), Maximum heart rate (thalach),** dan **slope**

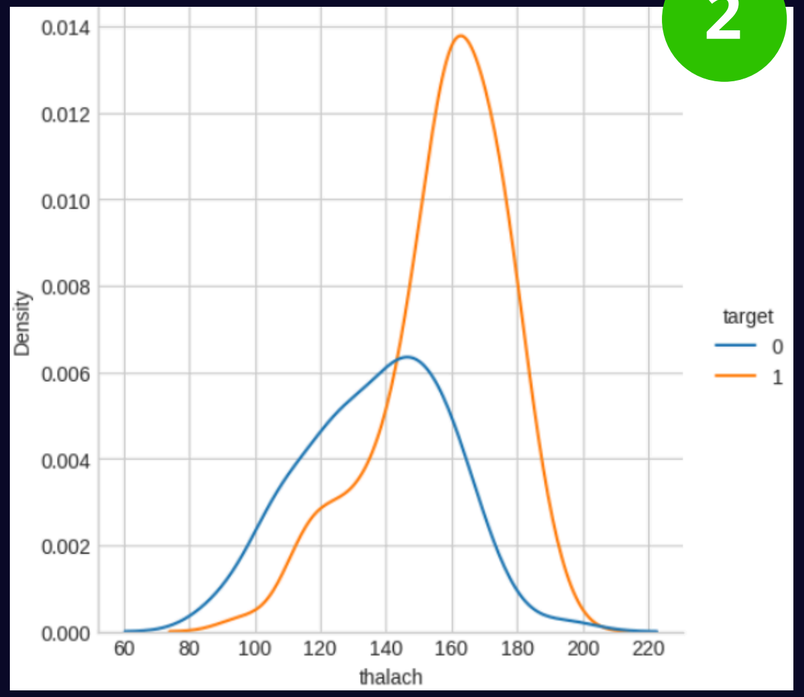
FEATURE ENGINEERING

1



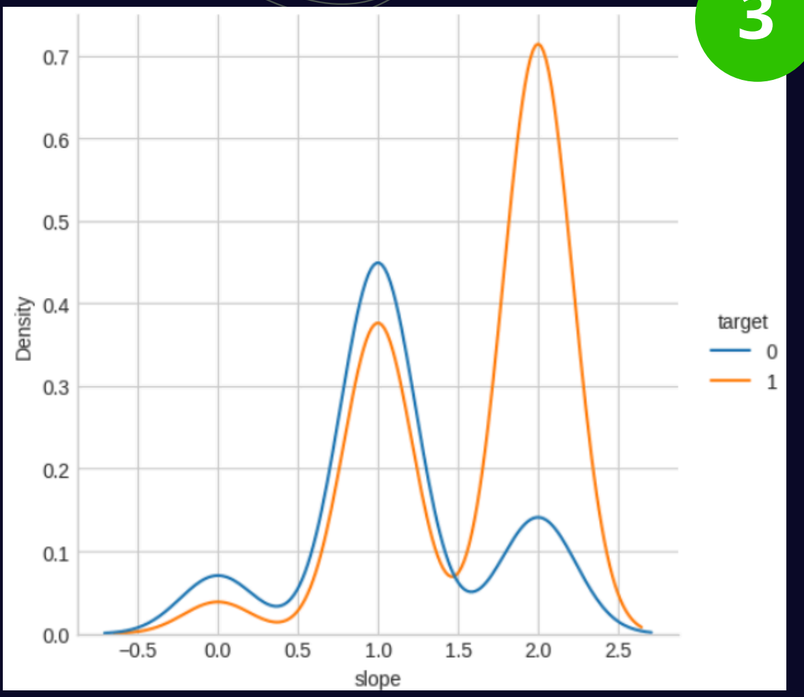
Orang yang terindikasi penyakit jantung (1) memiliki tipe nyeri dada **non-anginal pain (2)**

2



Orang yang terindikasi penyakit jantung (1) memiliki detak jantung maksimum **±160**

3

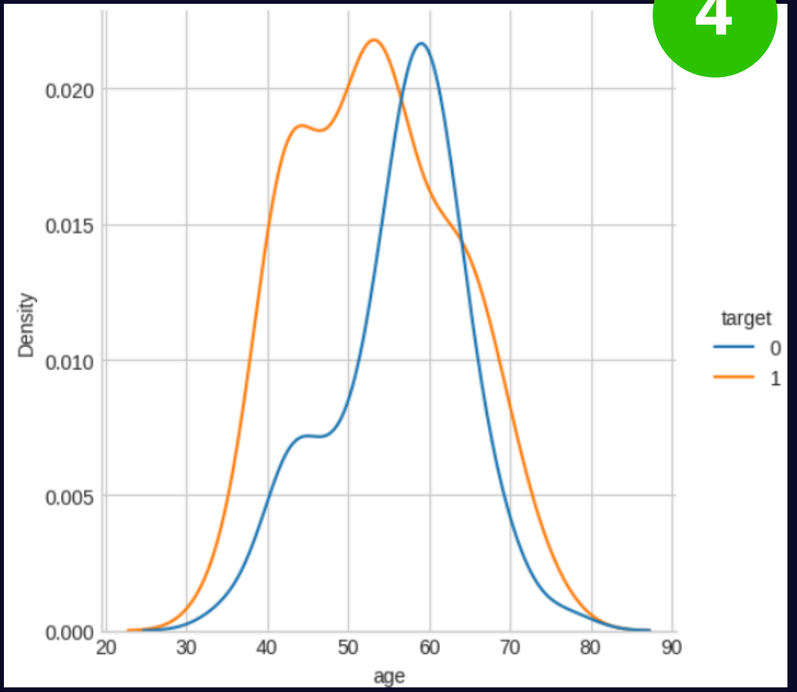


Orang yang terindikasi penyakit jantung (1) memiliki ST segment dengan kemiringan **downsloping** (2). Hal itu menandakan aliran darah ke jantung berkurang selama aktivitas fisik

TOP 3 ON CORRELATION MATRIX

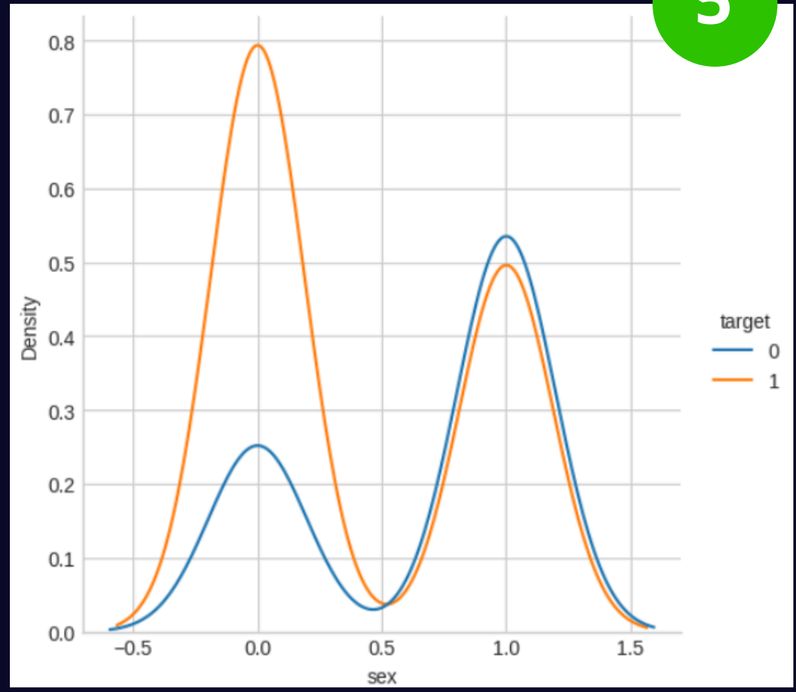
FEATURE ENGINEERING

4



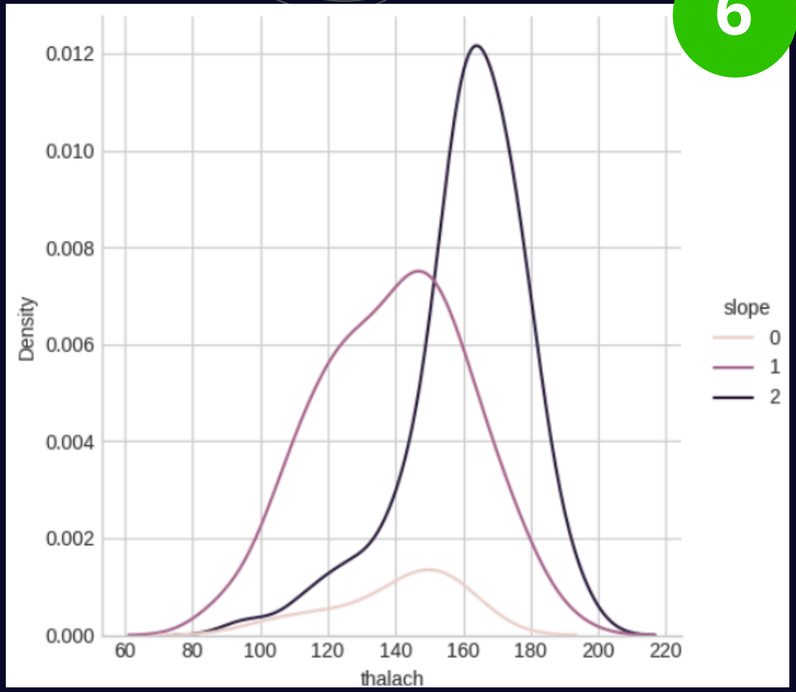
Orang yang terindikasi penyakit jantung (1) berada di rentang usia **50 - 55** tahun

5



Orang yang terindikasi jantung (1) dominan berjenis kelamin **perempuan (0)**

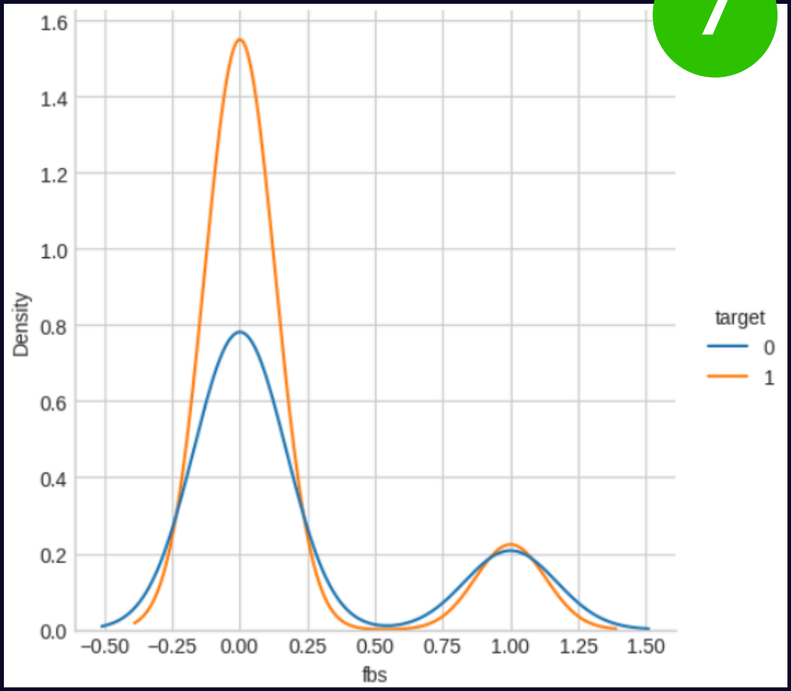
6



Orang dengan detak jantung maksimum **160** memiliki kemiringan ST segment **downslopping**

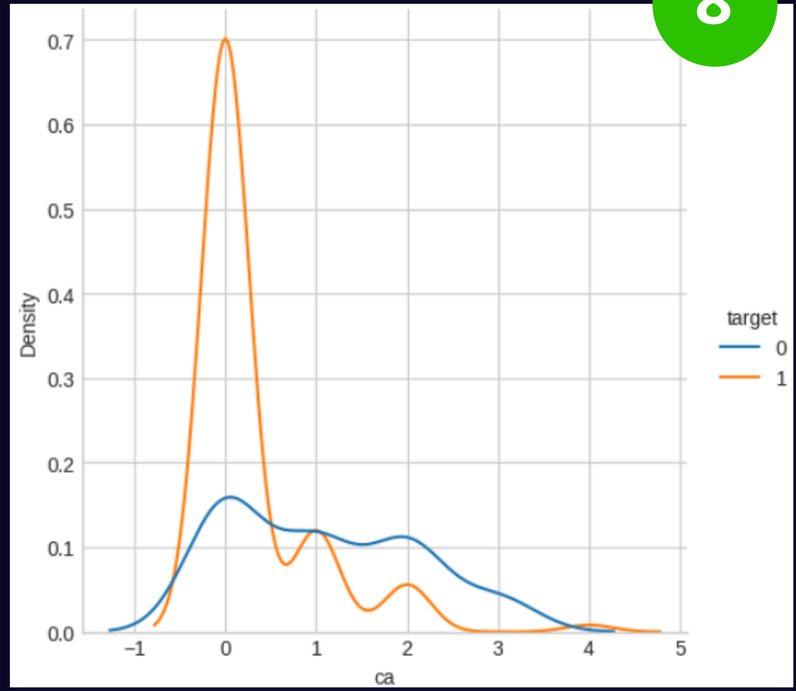
FEATURE ENGINEERING

7



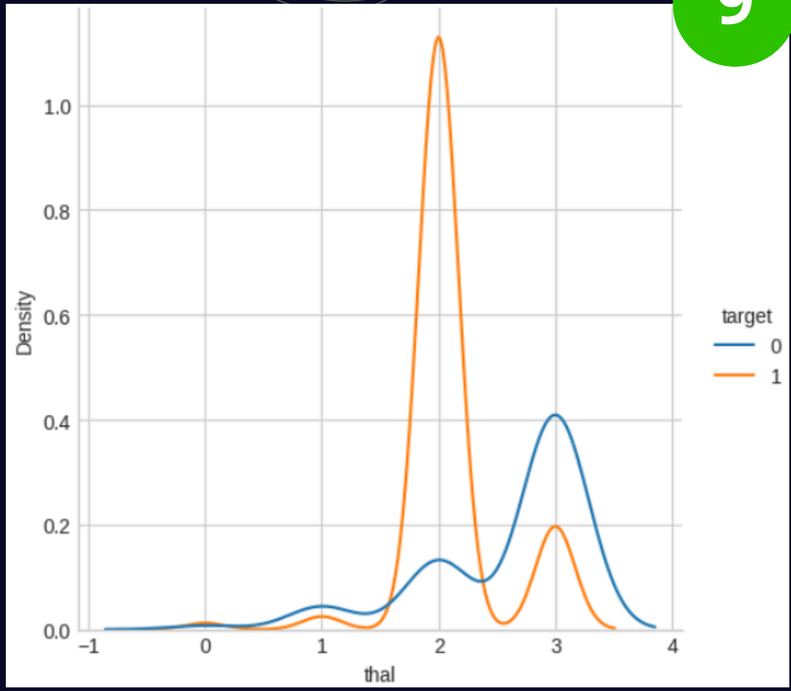
Orang yang terindikasi penyakit jantung (1) memiliki **fasting blood sugar <120 mg/dl** (0)

8



Orang yang terindikasi penyakit jantung (1) saat dilakukan pengecekan **tidak ada pembuluh darah utama yang terlihat** (0)

9

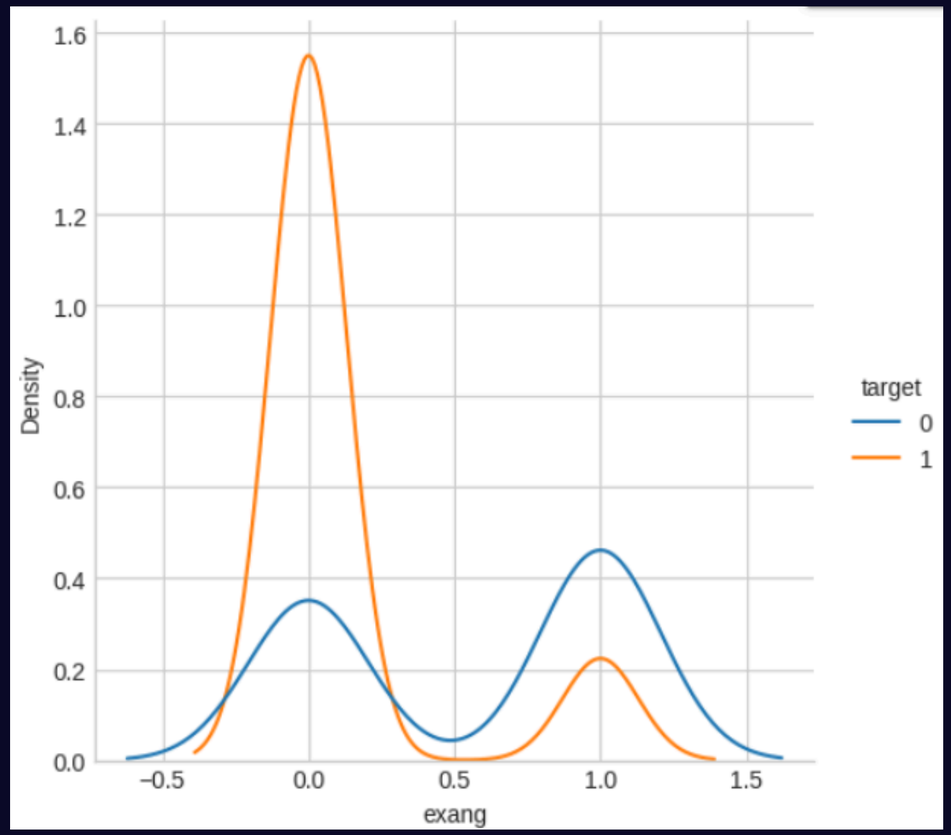


Orang yang terindikasi penyakit jantung (1) memiliki **thalasemia sedang** (2)

FEATURE ENGINEERING

Target vs Exang

Orang yang terindikasi penyakit jantung (1) **tidak mengalami nyeri dada saat *exercise* (0)**



DATA REDUCTION

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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

1025 rows × 14 columns

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	KNN_Predictions	DT_Predictions	RF_Predictions	XGB_Predictions
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1	1	1	1	1
10	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1	0	0	1	1
12	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1	1	1	1	1
16	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1	0	0	1	1
18	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
691	55	0	1	135	250	0	0	161	0	1.4	1	0	2	1	1	1	1	1
708	60	0	2	120	178	1	1	96	0	0.0	2	0	2	1	1	1	1	1
719	52	1	0	108	233	1	1	147	0	0.1	2	3	3	1	0	1	0	0
723	68	0	2	120	211	0	0	115	0	1.5	1	0	2	1	0	1	1	1
733	44	0	2	108	141	0	1	175	0	0.6	1	0	2	1	1	1	1	1

164 rows × 18 columns

SUMMARY
HEART DISEASE DATASET

- Terdapat 3 ciri utama jika seseorang terkena penyakit jantung, yakni tipe nyeri dadanya (cp) **non-anginal pain**, detak jantung maksimum (thalach) **160,** dan gelombang ST-nya (slope) **downslopping**
- Jumlah cetakan brosur tentang penyakit jantung berkurang hingga **86,39%.** Hal tersebut dapat mengurangi cost rumah sakit.

THANK'S

FOR READING

See you for another assignment 😉