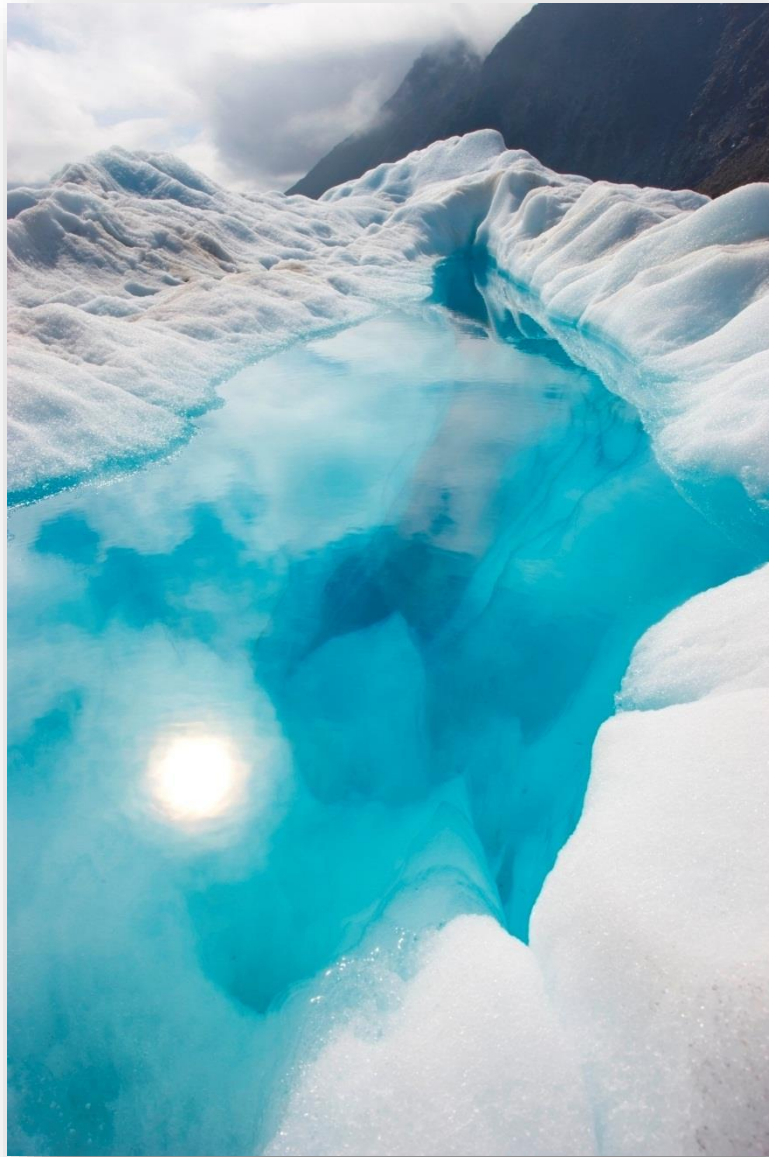


2023



BUKU KERJA/JOB SHEET

ASSOCIATE DATA SCIENTIST

Nama Peserta	:	Nur Magfirah Ramadani
Nomor Urut	:	17

DAFTAR ISI

DAFTAR ISI	1
BUKTI 1-ADS	3
1. Kebutuhan Data.....	3
2. Pengambilan Data	4
3. Pengintegrasian Data	5
BUKTI 2-ADS	5
1. Analisis Tipe dan Relasi Data	6
2. Analisis Karakteristik Data	6
3. Laporan Telaah Data.....	7
BUKTI 3-ADS	8
1. Pengecekan Kelengkapan Data	8
2. Rekomendasi Kelengkapan Data	9
BUKTI 4-ADS	10
1. Kriteria dan Teknik Pemilihan Data	10
2. Attributes (Columns) dan Records (Row) Data	11
BUKTI 5-ADS	12
1. Pembersihan Data Kotor	12
2. Laporan dan Rekomendasi Hasil Pembersihan Data Kotor	16
BUKTI 6-ADS	18
1. Analisis Teknik Transformasi Data.....	18
2. Transformasi Data	19
3. Dokumentasi Konstruksi Data	20
BUKTI 7-ADS	21
1. Pelabelan Data	21
2. Laporan Hasil Pelabelan Data.....	22
BUKTI 8-ADS	24
1. Parameter Model	24
2. Tools Pemodelan	24
BUKTI 9-ADS	26
1. Penggunaan Model dengan Data Riil	26

2.	Penilaian Hasil Pemodelan	26
----	---------------------------------	----

Link google colab :

https://colab.research.google.com/drive/13cRGP9-JuHweY1XuZhexl-bkf8aN_j6H#scrollTo=6SCclYta7241

BUKTI 1-ADS

Kode Unit	:	J.62DMI00.004.1
Judul Unit	:	Mengumpulkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengumpulkan data untuk data science.

Langkah Kerja:

- 1) Menentukan kebutuhan data
- 2) Mengambil data
- 3) Mengintegrasikan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengubah teks
 - Aplikasi basis data
 - Tools pengambilan data

1. KEBUTUHAN DATA

Instruksi Kerja:

- Identifikasi kebutuhan data sesuai tujuan teknis data science
- Periksa ketersediaan data berdasarkan kebutuhan data sesuai aturan yang berlaku
- Tentukan volume data berdasarkan kebutuhan data sesuai tujuan teknis data science

Jawab :

1. Identifikasi kebutuhan data

Identifikasi variabel : terdapat 12 variabel dengan 11 variabel prediktor (x) dan 1 variabel target (y) yaitu heartdisease. Adapun variabel tersebut adalah :

1. Age: Umur pasien dalam tahun.
2. Sex: Jenis kelamin pasien, dengan nilai M untuk Male dan F untuk Female.

3. Chest Pain Type : Jenis nyeri dada yang dirasakan oleh pasien, dengan nilai:
 - TA: Typical Angina
 - ATA: Atypical Angina
 - NAP: Non-Anginal Pain
 - ASY: Asymptomatic
4. Resting Blood Pressure (Resting BP): Tekanan darah istirahat pasien dalam mm Hg.
5. Cholesterol: Kolesterol serum pasien dalam mm/dl.
6. Fasting Blood Sugar (Fasting BS): Status gula darah pasien setelah puasa, dengan nilai 1 jika Fasting BS > 120 mg/dl, dan 0 sebaliknya.
7. Resting Electrocardiogram Results (Resting ECG): Hasil elektrokardiogram (ECG) pada istirahat, dengan nilai:
 - Normal: Normal
 - ST: Terdapat abnormalitas gelombang ST-T (inversi gelombang T dan/atau elevasi atau depresi ST > 0,05 mV)
 - LVH: Menunjukkan hipertrofi ventrikel kiri yang mungkin atau pasti menurut kriteria Estes'.
8. Maximum Heart Rate Achieved (MaxHR): Denyut jantung maksimum yang dicapai oleh pasien.
9. Exercise-Induced Angina (Exercise Angina): Apakah pasien mengalami angina selama latihan, dengan nilai numerik antara 60 dan 202.
10. Old Peak (Oldpeak): Nilai numerik yang mengukur depresi pada segmen ST.
11. ST Slope: Kemiringan segmen ST selama latihan puncak, dengan nilai:
 - Up: Meningkat
 - Flat: Datar
 - Down: Menurun
12. Heart Disease (Heartdisease): Kelas output yang menunjukkan apakah pasien memiliki penyakit jantung (1) atau tidak (0).

Sumber data : github (<https://github.com/arubhasy/dataset/blob/main/heart.csv>)

2. ketersediaan data
Data berasal dari github yang bisa diakses.
3. volume data
Jumlah baris : 918 dan jumlah kolom : 12

2. PENGAMBILAN DATA

Instruksi Kerja:

- Identifikasi metode dan tools pengambilan data sesuai tujuan teknis data science

- Tentukan tools pengambilan data sesuai tujuan teknis data science
- Siapkan tools pengambilan data sesuai tujuan teknis data science
- Jalankan proses pengambilan data sesuai dengan tools yang telah disiapkan

Jawab :

Proses pengambilan data dari github

```
# Loading Data
import warnings; warnings.simplefilter('ignore')
import pandas as pd # Loading Module yang dibutuhkan

file_ = 'data/heart.csv'
try: # Running Locally, yakinkan "file_" berada di folder "data"
    heart = pd.read_csv(file_)
except: # Running in Google Colab
    !mkdir data
    !wget -P data/ https://raw.githubusercontent.com/arubhasy/dataset/main/heart.csv
    heart = pd.read_csv(file_)
```

3. PENGINTEGRASIAN DATA

Instruksi Kerja:

- Periksa integritas data sesuai tujuan teknis data science
- Integrasikan data sesuai tujuan teknis data science

Jawab :

integritas data adalah mengacu pada keberlanjutan data yaitu data harus dapat diandalkan dan tersedia ketika dibutuhkan, akurasi/ketepatan yaitu data harus akurat dan mencerminkan keadaan sebenarnya, dan konsistensi data yaitu data harus konsisten di semua tempat di mana itu digunakan.

BUKTI 2-ADS

Kode Unit	:	J.62DMI00.005.1
Judul Unit	:	Menelaah Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menelaah data untuk data science.

Langkah Kerja:

- 1) Menganalisis tipe dan relasi data
- 2) Menganalisis karakteristik data
- 3) Membuat laporan telaah data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Tools pengolahan data
 - Tools pembuat grafik

1. ANALISIS TIPE DAN RELASI DATA

Instruksi Kerja:

- Identifikasi tipe data yang terkumpul sesuai tujuan teknis
- Uraikan nilai atribut data yang terkumpul sesuai dengan batasan konteks bisnisnya
- Identifikasi relasi antar data yang terkumpul sesuai dengan tujuan teknis

Jawab :

```
[4] heart.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
#   Column             Non-Null Count  Dtype
---  -
0   Age                 911 non-null   float64
1   Sex                 908 non-null   object
2   ChestPainType       918 non-null   object
3   RestingBP           918 non-null   int64
4   Cholesterol          918 non-null   int64
5   FastingBS           918 non-null   int64
6   RestingECG          918 non-null   object
7   MaxHR               918 non-null   int64
8   ExerciseAngina      918 non-null   object
9   Oldpeak             918 non-null   float64
10  ST_Slope            918 non-null   object
11  HeartDisease        918 non-null   int64
dtypes: float64(2), int64(5), object(5)
memory usage: 86.2+ KB
```

```
[6] for col in varObjects.columns:
      heart[col] = heart[col].astype('category')
      heart.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
#   Column             Non-Null Count  Dtype
---  -
0   Age                 911 non-null   float64
1   Sex                 908 non-null   category
2   ChestPainType       918 non-null   category
3   RestingBP           918 non-null   int64
4   Cholesterol          918 non-null   int64
5   FastingBS           918 non-null   int64
6   RestingECG          918 non-null   category
7   MaxHR               918 non-null   int64
8   ExerciseAngina      918 non-null   category
9   Oldpeak             918 non-null   float64
10  ST_Slope            918 non-null   category
11  HeartDisease        918 non-null   int64
dtypes: category(5), float64(2), int64(5)
memory usage: 55.5 KB
```

Interpretasi :

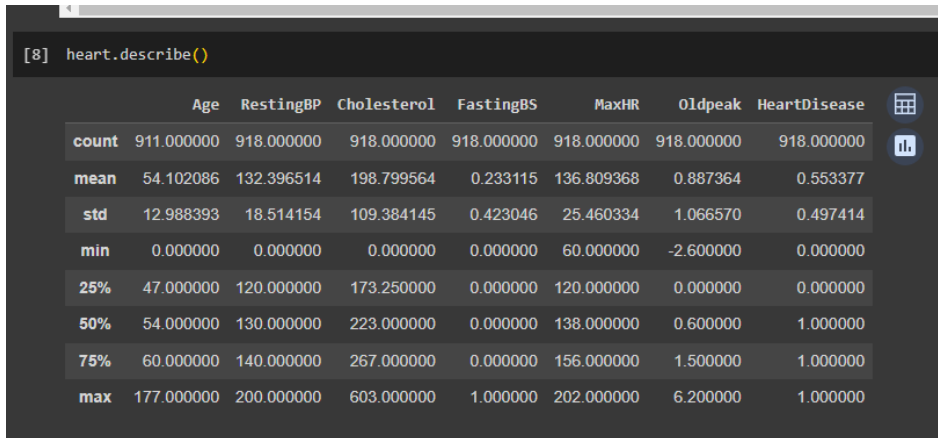
untuk melihat tipe data setiap variabel, dapat dilihat bahwa terdapat 2 variabel tipe float yaitu age dan oldpeak, kemudian terdapat 5 variabel tipe integer yaitu resting bp, cholestrol, fasting bs, MaxHR, dan heartdisease, kemudian terdapat 5 variabel tipe object yang sudah diubah ke kategori yaitu sex, chest pain type, resting ecg, exercise angina, dan ST_Slope.

2. ANALISIS KARAKTERISTIK DATA

Instruksi Kerja:

- Sajikan karakteristik data yang terkumpul dengan deskripsi statistik dasar
- Sajikan karakteristik data yang terkumpul dengan visualisasi grafik
- Analisis karakteristik data dari hasil penyajian data untuk telaah data

Jawab :



```
[8] heart.describe()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	911.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	54.102086	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	12.988393	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	0.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	177.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Interpretasi :

Dari tabel statistika deskriptif tsb dpt dilihat bahwa variabel age terdapat missing value karena total barisnya ada 911 sedangkan seharusnya total data adalah 918.

3. LAPORAN TELAHAH DATA

Instruksi Kerja:

- Dokumentasikan hasil analisis dalam bentuk laporan sesuai dengan tujuan teknis
- Susun hipotesis berdasar hasil analisis sesuai tujuan teknis data science

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) menganalisis tipe dan relasi data; dan (2) menganalisis karakteristik data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat laporan telaah data; dapat diabaikan.

BUKTI 3-ADS

Kode Unit	:	J.62DMI00.006.1
Judul Unit	:	Memvalidasi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memvalidasi data untuk data science.

Langkah Kerja:

- 1) Melakukan pengecekan kelengkapan data
- 2) Membuat rekomendasi kelengkapan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengubah teks

1. PENGECEKAN KELENGKAPAN DATA

Instruksi Kerja:

- Sajikan penilaian kualitas data dari hasil telaah sesuai tujuan teknis data science
- Sajikan penilaian tingkat kecukupan data dari hasil telaah sesuai tujuan teknis data science

Jawab :

```
#validasi data
numVar_heart = heart.select_dtypes(include=['float64', 'int64'])
numVar_heart.head()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
0	NaN	140	289	0	172	0.0	0
1	49.0	160	180	0	156	1.0	1
2	37.0	130	283	0	98	0.0	0
3	48.0	138	214	0	108	1.5	1
4	54.0	150	195	0	122	0.0	0

```
[108] #validasi data
varCategory_heart = heart.select_dtypes(include=['category'])
varCategory_heart.head()
```

	Sex	ChestPainType	RestingECG	ExerciseAngina	ST_Slope
0	M	ATA	Normal	N	Up
1	F	NAP	Normal	N	Flat
2	M	ATA	ST	N	Up
3	F	ASY	Normal	Y	Flat
4	M	NAP	Normal	N	Up

Interpretasi :

untuk memeriksa apakah variabel-variabel numerik dan kategorik yang dipilih sesuai dengan data dan apakah data tersebut memiliki format dan nilai yang konsisten dengan analisis atau pemodelan yang akan dilakukan selanjutnya.

2. REKOMENDASI KELENGKAPAN DATA

Instruksi Kerja:

- Susun rekomendasi hasil penilaian kualitas sesuai tujuan teknis data science
- Susun rekomendasi hasil penilaian kecukupan data sesuai tujuan teknis data science

Jawab :

```
[75] # mengecek apakah ada duplikat data sangatlah mudah menggunakan Pandas
# Bayangkan jika menggunakan Excel.
print(heart.shape)
print("jumlah data yang duplikat", heart.duplicated().sum())
heart[heart.duplicated() == True].head()
# Perhatikan kalau sebelumnya kita tidak "Drop" var observasi,
# maka kita tidak akan mendapatkan duplikasi dengan cara ini.

(746, 12)
jumlah data yang duplikat 0
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
--	-----	-----	---------------	-----------	-------------	-----------	------------	-------	----------------	---------	----------	--------------

Interpretasi :

Kelengkapan data diperiksa dengan tujuan untuk memastikan bahwa data yang digunakan dalam analisis atau pemodelan memiliki kualitas yang baik dan memiliki jumlah data yang mencukupi untuk mencapai tujuan data science yang diinginkan. data yang berkualitas salah satunya adalah data yang tidak terdapat duplikat, pada data heart ini diketahui bahwa tidak ada duplikasi.

BUKTI 4-ADS

Kode Unit	:	J.62DMI00.007.1
Judul Unit	:	Menentukan Objek Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilah dan memilih data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

- 1) Memutuskan kriteria dan teknik pemilihan data
- 2) Menentukan attributes (columns) dan records (row) data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi spreadsheet
 - Aplikasi notepad plus
 - Aplikasi SQL (Structured Query Language)

1. KRITERIA DAN TEKNIK PEMILIHAN DATA

Instruksi Kerja:

- Identifikasi kriteria pemilihan data sesuai dengan tujuan teknis dan aturan yang berlaku
- Tetapkan teknik pemilihan data sesuai dengan kriteria pemilihan data

Jawab :

```
[137] # menentukan objek data
heart.sample(10)
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
749	54.0	M	NAP	150	232	0	LVH	165	N	1.6	Up	0
37	41.0	F	ATA	110	250	0	ST	142	N	0.0	Up	0
263	59.0	M	ASY	130	126	0	Normal	125	N	0.0	Flat	1
615	70.0	M	ASY	130	322	0	LVH	109	N	2.4	Flat	1
748	64.0	M	ASY	120	246	0	LVH	96	Y	2.2	Down	1
182	52.0	M	ASY	140	404	0	Normal	124	Y	2.0	Flat	1
693	42.0	F	NAP	120	209	0	Normal	173	N	0.0	Flat	0
771	55.0	M	ASY	140	217	0	Normal	111	Y	5.6	Down	1
887	43.0	M	ASY	132	247	1	LVH	143	Y	0.1	Flat	1
550	55.0	M	ASY	172	260	0	Normal	73	N	2.0	Flat	1

Interpretasi :

Teknik pemilihan data bisa dilakukan dengan pengambilan sample acak pada data atau dengan pemilihan berdasarkan kriteria yaitu dengan slicing.

2. ATTRIBUTES (COLUMNS) DAN RECORDS (ROW) DATA

Instruksi Kerja:

- Identifikasi attributes (columns) data sesuai dengan kriteria pemilihan data
- Identifikasi records (row) data sesuai dengan kriteria pemilihan data

Jawab :

```
# Menampilkan informasi atribut (columns) dan jumlah records (rows)
print("Attributes (Columns):", heart.columns)
print("Number of Records (Rows):", len(heart))

# Menampilkan contoh records (rows)
print("Sample Records:")
print(heart.head())
```

Attributes (Columns): Index(['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS', 'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope', 'HeartDisease'], dtype='object')

Number of Records (Rows): 746

Sample Records:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	NaN	M	ATA	140	289	0	Normal					
1	49.0	F	NAP	160	180	0	Normal					
2	37.0	M	ATA	130	283	0	ST					
3	48.0	F	ASY	138	214	0	Normal					
4	54.0	M	NAP	150	195	0	Normal					

	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	172	N	0.0	Up	0
1	156	N	1.0	Flat	1
2	98	N	0.0	Up	0
3	108	Y	1.5	Flat	1
4	122	N	0.0	Up	0

BUKTI 5-ADS

Kode Unit	:	J.62DMI00.008.1
Judul Unit	:	Membersihkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membersihkan data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

- 1) Melakukan pembersihan data yang kotor
- 2) Membuat laporan dan rekomendasi hasil membersihkan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi spreadsheet
 - Aplikasi text editor
 - Aplikasi SQL (Structured Query Language)

1. PEMBERSIHAN DATA KOTOR

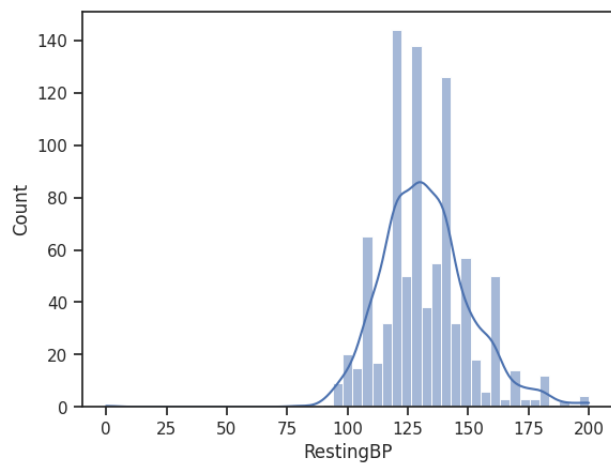
Instruksi Kerja:

- Tentukan strategi pembersihan data berdasarkan hasil telaah data
- Koreksi data yang kotor berdasarkan strategi pembersihan data

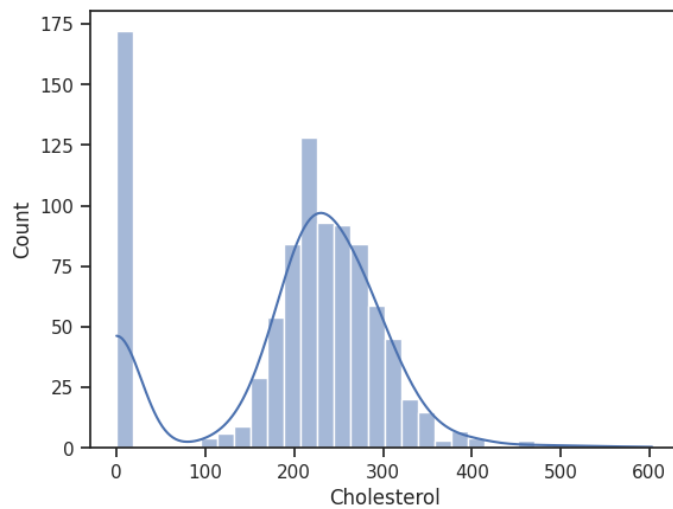
Jawab :

1. Noise

```
[12] # Visual Python: Visualization > Seaborn
sns.histplot(data=heart, x='RestingBP', kde=True)
plt.show()
heart[['RestingBP']].describe()
```

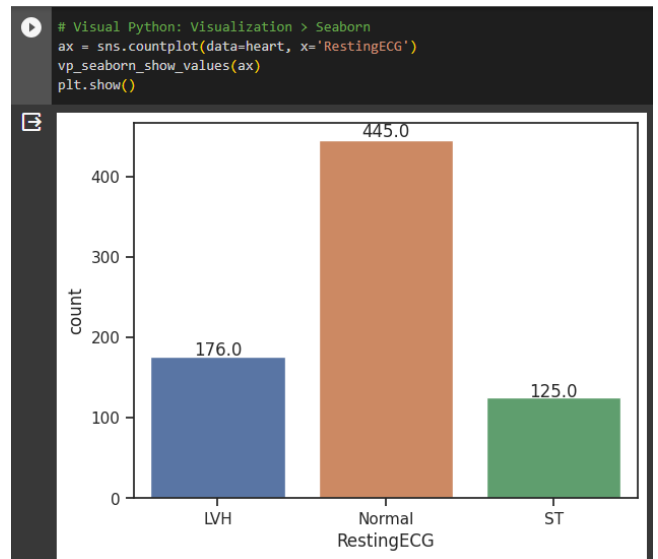
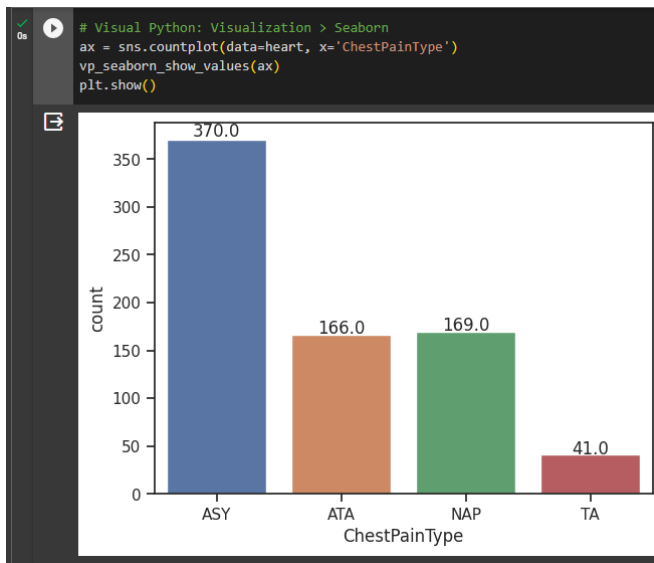
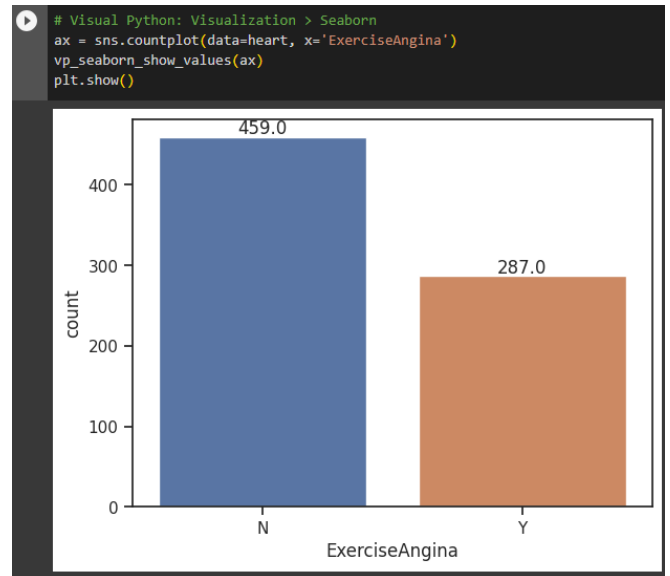
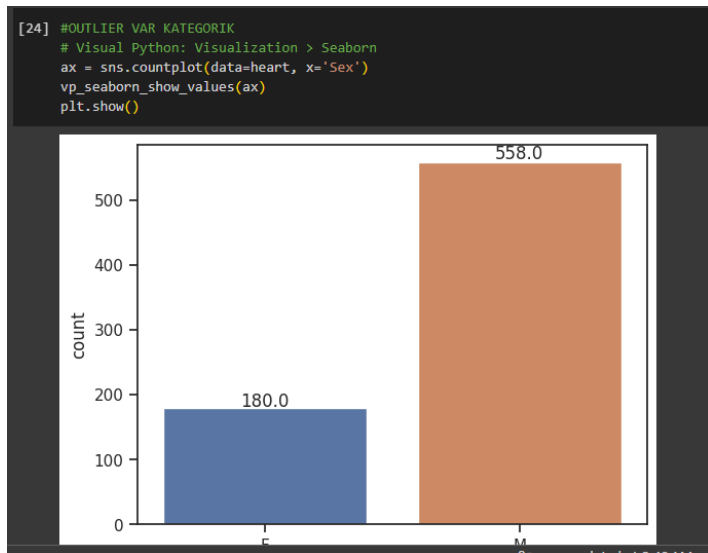


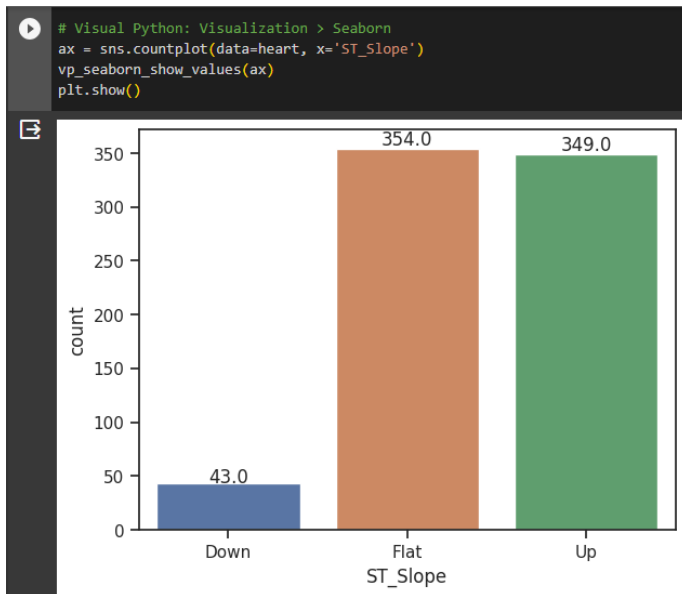
```
# Visual Python: Visualization > Seaborn
sns.histplot(data=heart, x='Cholesterol', kde=True)
plt.show()
heart[['Cholesterol']].describe()
```



2. Outlier

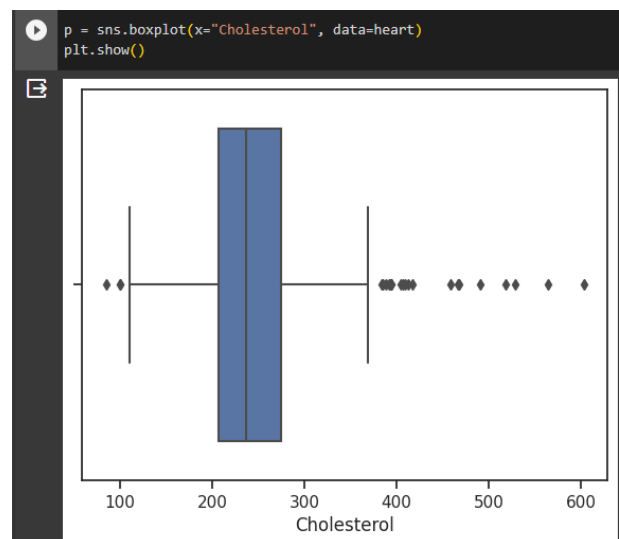
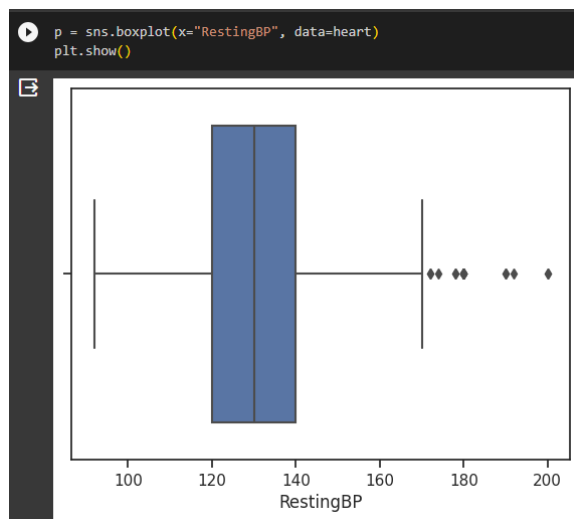
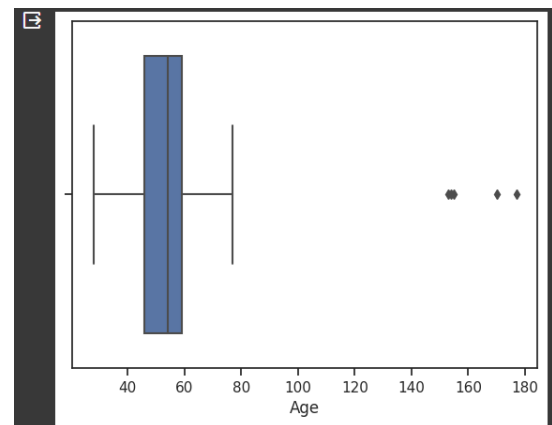
- Var kategorik

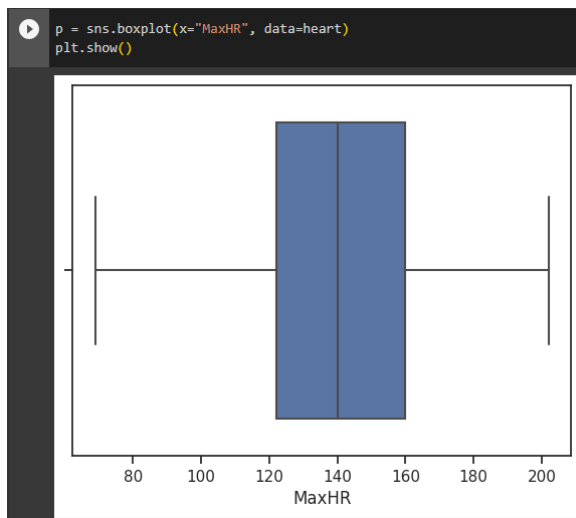
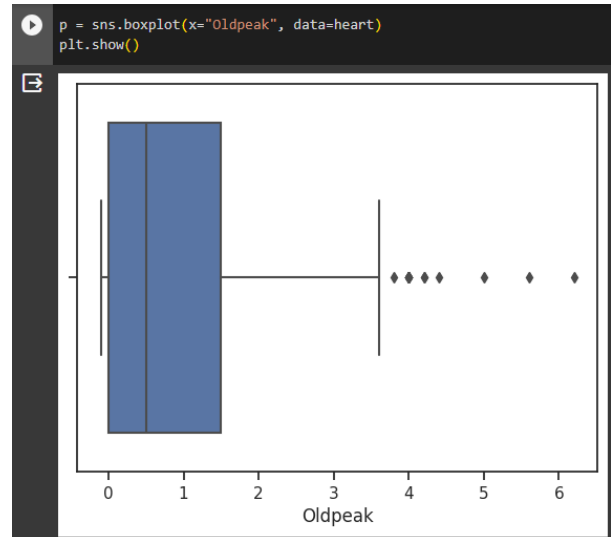
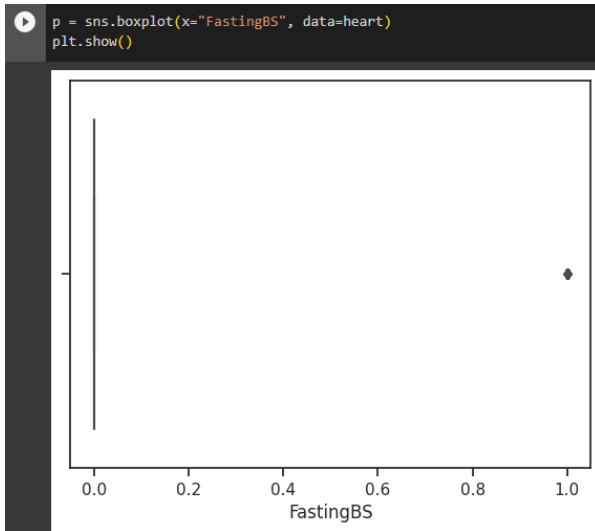




-var numerik

```
#OUTLIER VAR NUMERIK
# Visual Python: Visualization > Seaborn
p = sns.boxplot(x="Age", data=heart)
plt.show()
```



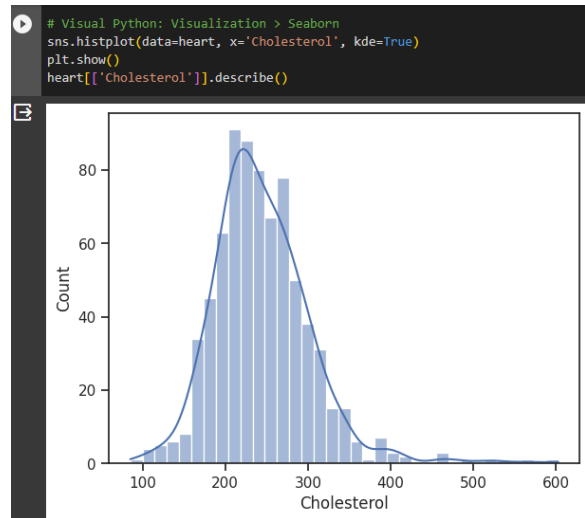
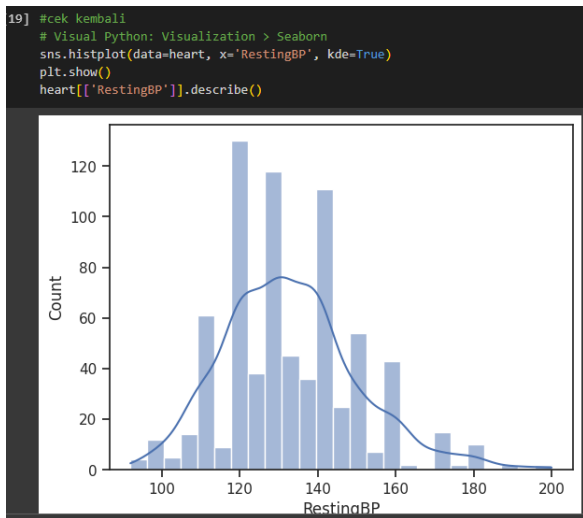


2. LAPORAN DAN REKOMENDASI HASIL PEMBERSIHAN DATA KOTOR

Instruksi Kerja:

- Deskripsikan masalah dan teknis koreksi data sesuai dengan kondisi data dan strategi pembersihan data
- Lakukan evaluasi berdasarkan analisis koreksi yang telah dilakukan
- Dokumentasikan evaluasi proses dan hasil pembersihan data kotor

Jawab :



Interpretasi Penanganan :

1. Noise

-Var kategorik

Pada data kategorik tidak ada noise hanya ditemukan missing value pada sex yaitu nan

-Var numerik

Pada var numerik resting bp = tekanan darah terdapat nilai 0. Dilakukan penghapusan baris pada nilai 0 di setiap kolom resting bp, karena tidak mungkin tekanan darah manusia 0.

Sedangkan Pada var numerik cholestrol = lemak kolestrol terdapat nilai 0. Dilakukan penghapusan baris pada nilai 0 di setiap kolom kolestrol, karena tidak mungkin kolestrol manusia 0.

2. Outlier

Dilihat dari output bahwa outlier tidak terlalu banyak jadi diputuskan untuk dibiarkan saja.

BUKTI 6-ADS

Kode Unit	:	J.62DMI00.009.1
Judul Unit	:	Mengkonstruksi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengkonstruksi data untuk proyek data science.

Langkah Kerja:

- 1) Menganalisis teknik transformasi data
- 2) Melakukan transformasi data
- 3) Membuat dokumentasi konstruksi data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolahan kata
 - Tools pengolahan kata

1. ANALISIS TEKNIK TRANSFORMASI DATA

Instruksi Kerja:

- Lakukan analisis data untuk menentukan representasi fitur data awal
- Lakukan analisis representasi fitur data awal untuk menentukan teknik rekayasa fitur yang diperlukan untuk pembangunan model data science

Jawab :

```
#mengatasi missing value var kategorik di sex
heart1 = heart.dropna() # jika ada MV minimal satu di salah satu kolom, maka baris di hapus
#df.dropna(how='all') # jika ada MV di semua kolom, maka baris di hapus
#df.dropna(thresh=2) # jika ada MV minimal di salah 2 kolom, maka baris di hapus
#df.dropna(subset=['Pekerjaan'])[:7] # jika ada MV minimal satu di salah kolom Dist_Hospital
print(heart.shape, heart1.shape)
heart1.head()
```

(746, 12) (732, 12)

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	49.0	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37.0	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48.0	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54.0	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
5	39.0	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

Interpretasi :

Terdapat missing value di var kategorik 'sex' dan var numerik 'age'. Penanganannya adalah cukup menghapus baris yang terdapat missing value var 'sex' dan 'age'.

Dicek kembali dan dapat dilihat bahwa missing value sudah tidak ada

```
print(heart1.isnull().sum())
```

Age	0
Sex	0
ChestPainType	0
RestingBP	0
Cholesterol	0
FastingBS	0
RestingECG	0
MaxHR	0
ExerciseAngina	0
Oldpeak	0
ST_Slope	0
HeartDisease	0
dtype: int64	

2. TRANSFORMASI DATA

Instruksi Kerja:

- Lakukan transformasi untuk mendapatkan fitur data awal
- Lakukan rekayasa fitur data untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science

Jawab :

```
print(heart1.shape)
for col in catVar.columns:
    if col != 'HeartDisease':
        transformasi = pd.get_dummies(heart1[col], prefix='')
        heart1 = pd.concat([heart1, transformasi], axis = 1)
        # Hapus Variabel Kategorik Awal, Sudah tidak diperlukan
    try:
        heart1.drop([col], axis=1, inplace=True)
    except Exception as err_:
        print(err_)
print(heart1.shape)
heart1.head()
```

(732, 12)
(732, 21)

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	_F	_M	_ASY	...	_NAP	_TA	_LVH	_Normal	_ST	_N	_Y	_Down	_Flat	_Up
1	49.0	160	180	0	156	1.0	1	1	0	0	...	1	0	0	1	0	1	0	0	1	0
2	37.0	130	283	0	98	0.0	0	0	1	0	...	0	0	0	0	1	1	0	0	0	1
3	48.0	138	214	0	108	1.5	1	1	0	1	...	0	0	0	1	0	0	1	0	1	0
4	54.0	150	195	0	122	0.0	0	0	1	0	...	1	0	0	1	0	1	0	0	0	1
5	39.0	120	339	0	170	0.0	0	0	1	0	...	1	0	0	1	0	1	0	0	0	1

5 rows x 21 columns

Interpretasi :

Transformasi variabel kategorikal menjadi variabel dummy menggunakan one-hot encoding. Ini dapat memiliki pengaruh terhadap model machine learning, terutama jika model tersebut sensitif terhadap representasi kategorikal. Transformasi ini umumnya dilakukan pada variabel kategorikal karena sebagian besar algoritma machine learning bekerja lebih baik dengan data numerik. One-hot encoding memungkinkan model untuk menangkap hubungan antar kategori tanpa memberikan bobot yang keliru pada variabel kategorikal. Setelah melakukan transformasi data, data jadi mempunyai 732 baris dan 21 kolom.

3. DOKUMENTASI KONSTRUKSI DATA

Instruksi Kerja:

- Jabarkan teknis transformasi data dalam bentuk tertulis
- Tuangkan hasil transformasi data dan rekomendasi hasil transformasi dalam bentuk tertulis

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) menganalisis teknik transformasi data; dan (2) melakukan transformasi data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat dokumentasi konstruksi data; dapat diabaikan.

BUKTI 7-ADS

Kode Unit	:	J.62DMI00.010.1
Judul Unit	:	Menentukan Label Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menentukan label data untuk pembangunan model data science.

Langkah Kerja:

- 1) Melakukan pelabelan data
- 2) Membuat laporan hasil pelabelan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi pelabelan data

1. PELABELAN DATA

Instruksi Kerja:

- Uraikan kesesuaian antara analisis hasil pelabelan data sejenis yang sudah ada dengan Standard Operating Procedure (SOP) pelabelan
- Lakukan pelabelan data sesuai dengan SOP pelabelan

Jawab :

```
[43] predictor = heart1.loc[:, ~heart1.columns.isin(['HeartDisease'])]
      target = heart1['HeartDisease']

      xTrain, xTest, yTrain, yTest = train_test_split(predictor, target, test_size=0.3, random_state=33)
      print(xTrain.shape, yTrain.shape)
      print(xTest.shape, yTest.shape)

(512, 20) (512,)
(220, 20) (220,)
```

Interpretasi :

Sebelum melakukan pemodelan, data akan dilakukan training testing data dengan membagi dataset menjadi dua subset: set pelatihan (training set) dan set pengujian (testing set). Tujuan utama dari pembagian ini adalah untuk mengukur kinerja model pada data.

1. (512, 20) (512,): Ini adalah bentuk (shape) dari set pelatihan.

- Jumlah baris (instances) dalam set pelatihan adalah 512.
- Jumlah fitur (features) atau kolom dalam set pelatihan adalah 20.

2. (220, 20) (220,): Ini adalah bentuk dari set pengujian.

- Jumlah baris dalam set pengujian adalah 220.
- Jumlah fitur atau kolom dalam set pengujian adalah 20.

Jadi, output ini memberi informasi tentang ukuran dataset set pelatihan dan set pengujian setelah dilakukan pembagian menggunakan `train_test_split`. Jumlah baris mencerminkan berapa banyak sampel data yang ada dalam masing-masing set, sedangkan jumlah fitur mencerminkan berapa banyak variabel independen yang digunakan untuk pelatihan dan pengujian model.

2. LAPORAN HASIL PELABELAN DATA

Instruksi Kerja:

- Uraikan statistik hasil pelabelan pada laporan
- Uraikan evaluasi proses pelabelan pada laporan

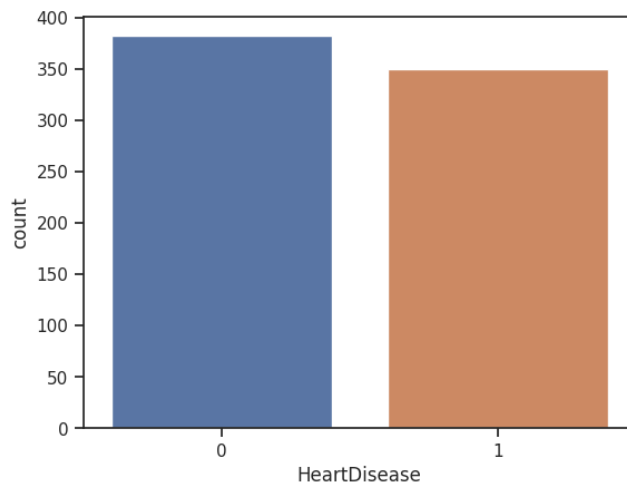
Jawab :

```
# Visual Python: Visualization > Seaborn
from collections import Counter

sns.countplot(data=heart1, x='HeartDisease')
plt.show()

D = Counter(heart1['HeartDisease'])
print(D)
print("0 = ", D[0]*100/(len(heart1['HeartDisease'])), '% 1 = ', D[1]*100/(len(heart1['HeartDisease'])), '%')
```

Dengan output :



```
Counter({0: 382, 1: 350})
0 = 52.185792349726775 % 1 = 47.814207650273225 %
```

Interpretasi :

1. Counter menunjukkan jumlah masing-masing kelas dalam bentuk dictionary, dengan kelas 0 memiliki 382 sampel dan kelas 1 memiliki 350 sampel.
2. Output persentase memberikan informasi tentang seberapa besar proporsi masing-masing kelas terhadap total jumlah sampel. Dalam kasus ini, kelas 0 (tidak ada penyakit jantung) menyumbang sekitar 52.19%, sedangkan kelas 1 (ada penyakit jantung) menyumbang sekitar 47.81%.
3. Distribusi kelas : terdapat sedikit lebih banyak sampel dengan tabel 0 (tidak ada penyakit jantung) dibandingkan dengan label 1 (ada penyakit jantung), tetapi perbedaannya tidak signifikan.

BUKTI 8-ADS

Kode Unit	:	J.62DMI00.013.1
Judul Unit	:	Membangun Model

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membangun model.

Langkah Kerja:

- 1) Menyiapkan parameter model
- 2) Menggunakan tools pemodelan

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer dan peralatannya
 - Perangkat lunak data science di antaranya: rapid miner, weka, atau development untuk bahasa pemrograman tertentu seperti Python atau R.
- Perlengkapan
 - Dokumen best practices kriteria dan evaluasi penilaian

1. PARAMETER MODEL

Instruksi Kerja:

- Identifikasi parameter-parameter yang sesuai dengan model
- Tetapkan nilai toleransi parameter evaluasi pengujian sesuai dengan tujuan teknis

2. TOOLS PEMODELAN

Instruksi Kerja:

- Identifikasi tools untuk membuat model sesuai dengan tujuan teknis data science
- Bangun algoritma untuk teknik pemodelan yang ditentukan menggunakan tools yang dipilih
- Eksekusi algoritma pemodelan sesuai dengan skenario pengujian dan tools untuk membuat model yang telah ditetapkan
- Optimasi parameter model algoritma untuk menghasilkan nilai parameter evaluasi yang sesuai dengan skenario pengujian

Jawab :

```
Model Regresi Logistik

[46] reglog = LogisticRegression().fit(xTrain, yTrain)
prediksi_regLog = reglog.predict(xTest)
print(confusion_matrix(yTest, prediksi_regLog))
print(classification_report(yTest, prediksi_regLog))

[[100  23]
 [ 12  85]]
      precision    recall  f1-score   support

      0       0.89      0.81      0.85        123
      1       0.79      0.88      0.83         97

 accuracy          0.84          220
 macro avg       0.84      0.84      0.84          220
 weighted avg    0.85      0.84      0.84          220
```

Interpretasi :

Output yang didapatkan dari model regresi logistik mencakup confusion matrix dan classification report.

1. Confusion matriks yang artinya True Positive (TP) = 85, False Positive (FP) = 23, True Negative (TN) = 100, False Negative (FN) = 12
2. Accuracy : Model memiliki tingkat akurasi sekitar 84%, yang berarti sekitar 84% dari prediksi model benar.
3. Precision : Dari yang diprediksi sebagai tidak ada penyakit jantung kelas(0), sekitar 89% adalah benar-benar tidak ada penyakit jantung kelas (0). Dari yang diprediksi sebagai ada penyakit jantung kelas (1), sekitar 79% adalah benar-benar ada penyakit jantung kelas (1).
4. Recall : Dari semua kasus yang sebenarnya tidak ada penyakit jantung, model mengidentifikasi sekitar 81% dengan benar. Dari semua kasus yang sebenarnya ada penyakit jantung, model mengidentifikasi sekitar 88% dengan benar.
5. F1-Score : Merupakan rata-rata harmonik dari precision dan recall. Semakin tinggi, semakin baik keseimbangan antara precision dan recall.
6. Support : Jumlah sampel dalam setiap kelas.
7. Model regresi logistik memiliki tingkat akurasi sekitar 84%, yang merupakan persentase prediksi yang benar dari seluruh dataset.
8. Model cenderung lebih baik dalam mengidentifikasi kelas 0 daripada kelas 1 berdasarkan nilai precision dan recall.
9. Precision, recall, dan F1-score yang baik dapat menunjukkan bahwa model ini memiliki kinerja yang baik dalam memprediksi kelas 0 dan 1.

BUKTI 9-ADS

Kode Unit	:	J.62DMI00.014.1
Judul Unit	:	Mengevaluasi Hasil Pemodelan

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengevaluasi hasil pemodelan.

Langkah Kerja:

- 1) Menggunakan model dengan data riil
- 2) Menilai hasil pemodelan

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Tools untuk mengeksekusi model
 - Tools untuk pengumpulan data riil

1. PENGGUNAAN MODEL DENGAN DATA RIIL

Instruksi Kerja:

- Kumpulkan data baru untuk evaluasi pemodelan sesuai kebutuhan yang mengacu kepada parameter evaluasi
- Uji model dengan menggunakan data riil yang telah dikumpulkan

2. PENILAIAN HASIL PEMODELAN

Instruksi Kerja:

- Nilai keluaran pengujian model berdasarkan metrik kesuksesan
- Dokumentasikan hasil penilaian sesuai standar yang berlaku

Jawab :

```
#crossvalidation
mulai = time.time()
scores_reglog = cross_val_score(reglog, predictor, target, cv=10) # perhatikan sekarang kita menggunakan seluruh data
waktu = time.time() - mulai
print("Accuracy Regresi Logistik: %0.2f (+/- %0.2f), Waktu = %0.3f detik" % (scores_reglog.mean(), scores_reglog.std() * 2, waktu))

Accuracy Regresi Logistik: 0.86 (+/- 0.12), Waktu = 1.203 detik
```

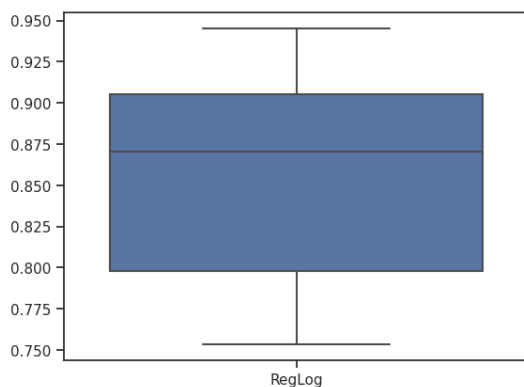
Interpretasi :

1. CROSS VALIDASI adalah membagi train dan tes data berkali-kali. Hal ini lebih akurat dibandingkan hanya 1x training dan testing data.
2. Dengan predictor adalah var x dan target adalah variabel y (heartdisease) cross validasi paling umum dilakukan sebanyak 10 x dan minimal 5x
3. Tujuannya adalah utk mengukur resiko dari menggunakan Machine Learning
4. Akurasi rata-rata dari model regresi logistik selama 10 kali validasi silang adalah sekitar 86%. Ini menunjukkan seberapa baik model ini dapat melakukan prediksi secara keseluruhan pada dataset.
5. Rentang (interval) +/- 0.12, dinyatakan utk mengukur sebaran (variasi) nilai akurasi antar iterasi validasi silang. Rentang ini memberikan gambaran sejauh mana model konsisten dalam kinerjanya. Semakin kecil rentangnya, semakin konsisten model tersebut.
6. Tingkat akurasi yang relatif tinggi (86%) menunjukkan bahwa model regresi logistik konsisten dalam melakukan prediksi pada dataset, dan hasilnya dapat dianggap sebagai indikasi kinerja yang baik.
7. Waktu yang diperlukan (0.348 detik) memberikan gambaran tentang efisiensi komputasi dari proses cross-validation ini.

Secara keseluruhan, hasil cross-validation menunjukkan bahwa model regresi logistik memiliki kinerja yang baik pada dataset ini, dengan akurasi yang tinggi dan konsistensi yang baik antar iterasi validasi silang.

```
[48] # Visualisasi untuk mengevaluasi & membandingkan model dengan lebih baik lagi
heart1_ = pd.DataFrame({'RegLog': scores_regLog})
p = sns.boxplot(data = heart1_)
heart1_.min()

RegLog    0.753425
dtype: float64
```



Interpretasi :

1. Nilai minimum sekitar 75% adalah nilai terendah yang dicapai oleh model regresi logistik selama proses cross-validation. Ini menunjukkan bahwa bahkan pada situasi terburuk (nilai terendah), model masih memiliki tingkat akurasi sekitar 75%.
2. Boxplot memberikan gambaran visual tentang distribusi nilai-nilai hasil cross-validation. Ini mencakup kuartil, median, dan adanya pencilan (outliers).
3. Meskipun nilai minimum memberikan gambaran terburuk dari model, nilai tersebut masih cukup tinggi, menunjukkan bahwa model cenderung memberikan prediksi yang baik dalam mendeteksi penyakit jantung pada dataset ini.

Berdasarkan hasil pemodelan, dapat memberikan kontribusi yang berharga dalam tujuan data scientist yaitu mendeteksi pasien dengan penyakit jantung. Berikut beberapa cara interpretasi hasil model dapat diterapkan ke dalam konteks dunia nyata:

1. Evaluasi Kinerja Model

Tingkat akurasi sekitar 86% pada regresi logistik menunjukkan bahwa model dapat secara efektif memprediksi apakah seorang pasien mengidap penyakit jantung atau tidak berdasarkan fitur-fitur yang digunakan.

2. Rekomendasi untuk Dokter

Model ini dapat membantu dokter dalam mendeteksi dini kasus penyakit jantung. Identifikasi awal pasien dengan risiko tinggi memungkinkan pemberian perhatian lebih intensif atau pengawasan lebih ketat.

3. Faktor Risiko Individu

Model regresi logistik mungkin menggunakan fitur-fitur seperti tekanan darah, kadar kolesterol, usia, dan variabel lain yang dapat memberikan wawasan tentang faktor risiko individual.

4. Pertimbangan Praktis

Pasien yang mendapatkan prediksi positif dapat direkomendasikan untuk konsultasi lebih lanjut dengan dokter untuk pemeriksaan lebih intensif atau penanganan lanjutan.

5. Gaya Hidup Sehat

Pasien dengan risiko tinggi yang teridentifikasi dapat diberikan saran dan rekomendasi untuk mengadopsi gaya hidup sehat guna mengurangi risiko penyakit kardiovaskular, seperti perubahan dalam pola makan, olahraga teratur, dan mengurangi faktor risiko lainnya.