NAMA : Azzuhri Wisnu Akhsan

NIM : A11.2020.13187

Kelas : BKDS02

Projek Bimbingan Karir Data Science

1) Pengumpulan Data

Data yang digunakan merupakan dataset penyakit jantung yang diambil melalui link : https://archive.ics.uci.edu/dataset/45/heart+disease Dataset yang dipakai adalah dataset dengan nam file "Hungarian.data", diharapkan untuk membaca dokumentasi pada "heart-disease.name"

2) Menelaah data

Masukan library yang diperlukan

```
import pandas as pd
import re
import numpy as np
import itertools
```

Load Dataset

```
dir = 'hungarian.data'
```

```
with open (dir, encoding='Latin1') as file :
  lines = [line.strip() for line in file]

lines[0:10]
```

```
['1254 0 40 1 1 0 0',
 '-9 2 140 0 289 -9 -9 -9',
 '0 -9 -9 0 12 16 84 0',
 '0 0 0 0 150 18 -9 7',
 '172 86 200 110 140 86 0 0',
 '0 -9 26 20 -9 -9 -9 -9',
 '-9 -9 -9 -9 -9 -9 -9 12',
 '20 84 0 -9 -9 -9 -9 -9',
 '-9 -9 -9 -9 -9 1 1 1',
 '1 1 -9. -9. name']
```

Rubah bentuk data menjadi dataframe agar lebih mudah dipahami

```
data = itertools.takewhile(
    lambda x: len(x) == 76,
    (' '.join(lines[i:(i + 10)]).split() for i in range(0, len(lines), 10))
)

df = pd.DataFrame.from_records(data)

df.head()
```

	0	1	2	3	4	5	6	7	8	9	...	66	67	68	69	70	71	72	73	74	75
0	1254	0	40	1	1	0	0	-9	2	140	...	-9	-9	1	1	1	1	1	-9.	-9.	name
1	1255	0	49	0	1	0	0	-9	3	160	...	-9	-9	1	1	1	1	1	-9.	-9.	name
2	1256	0	37	1	1	0	0	-9	2	130	...	-9	-9	1	1	1	1	1	-9.	-9.	name
3	1257	0	48	0	1	1	1	-9	4	138	...	2	-9	1	1	1	1	1	-9.	-9.	name
4	1258	0	54	1	1	0	1	-9	3	150	...	1	-9	1	1	1	1	1	-9.	-9.	name

5 rows × 76 columns

menampilkan informasi dari file dataset yang sudah dimasukkan dalam dataframe

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 76 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   0       294 non-null    object
 1   1       294 non-null    object
 2   2       294 non-null    object
 3   3       294 non-null    object
 4   4       294 non-null    object
 5   5       294 non-null    object
 6   6       294 non-null    object
 7   7       294 non-null    object
 8   8       294 non-null    object
 9   9       294 non-null    object
 10  10      294 non-null    object
 11  11      294 non-null    object
 12  12      294 non-null    object
 13  13      294 non-null    object
 14  14      294 non-null    object
 15  15      294 non-null    object
 16  16      294 non-null    object
 17  17      294 non-null    object
 18  18      294 non-null    object
 19  19      294 non-null    object
 20  20      294 non-null    object
 21  21      294 non-null    object
 22  22      294 non-null    object
 23  23      294 non-null    object
 24  24      294 non-null    object
 25  25      294 non-null    object
 26  26      294 non-null    object
 27  27      294 non-null    object
 28  28      294 non-null    object
 29  29      294 non-null    object
 30  30      294 non-null    object
 31  31      294 non-null    object
 32  32      294 non-null    object
 33  33      294 non-null    object
```

```
 34  34      294 non-null    object
 35  35      294 non-null    object
 36  36      294 non-null    object
 37  37      294 non-null    object
 38  38      294 non-null    object
 39  39      294 non-null    object
 40  40      294 non-null    object
 41  41      294 non-null    object
 42  42      294 non-null    object
 43  43      294 non-null    object
 44  44      294 non-null    object
 45  45      294 non-null    object
 46  46      294 non-null    object
 47  47      294 non-null    object
 48  48      294 non-null    object
 49  49      294 non-null    object
 50  50      294 non-null    object
 51  51      294 non-null    object
 52  52      294 non-null    object
```

```
df = df.iloc[:,:-1]
df = df.drop(df.columns[0], axis=1)
```

mengubah tipe file dataset menjadi tipe data float sesuai dengan nilai null yaitu -0.9

```
df = df.astype(float)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 74 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   1       294 non-null    float64
 1   2       294 non-null    float64
 2   3       294 non-null    float64
 3   4       294 non-null    float64
 4   5       294 non-null    float64
 5   6       294 non-null    float64
 6   7       294 non-null    float64
 7   8       294 non-null    float64
 8   9       294 non-null    float64
 9   10      294 non-null    float64
 10  11      294 non-null    float64
 11  12      294 non-null    float64
 12  13      294 non-null    float64
 13  14      294 non-null    float64
 14  15      294 non-null    float64
 15  16      294 non-null    float64
 16  17      294 non-null    float64
 17  18      294 non-null    float64
 18  19      294 non-null    float64
 19  20      294 non-null    float64
 20  21      294 non-null    float64
 21  22      294 non-null    float64
 22  23      294 non-null    float64
 23  24      294 non-null    float64
 24  25      294 non-null    float64
 25  26      294 non-null    float64
 26  27      294 non-null    float64
 27  28      294 non-null    float64
 28  29      294 non-null    float64
 29  30      294 non-null    float64
 30  31      294 non-null    float64
```

```
 31  32      294 non-null    float64
 32  33      294 non-null    float64
 33  34      294 non-null    float64
 34  35      294 non-null    float64
 35  36      294 non-null    float64
 36  37      294 non-null    float64
 37  38      294 non-null    float64
 38  39      294 non-null    float64
 39  40      294 non-null    float64
 40  41      294 non-null    float64
 41  42      294 non-null    float64
 42  43      294 non-null    float64
 43  44      294 non-null    float64
 44  45      294 non-null    float64
 45  46      294 non-null    float64
 46  47      294 non-null    float64
 47  48      294 non-null    float64
 48  49      294 non-null    float64
 49  50      294 non-null    float64
 50  51      294 non-null    float64
 51  52      294 non-null    float64
```

3) Validasi Data

Bertujuan untuk mengetahu kondisi dataset untuk mengetahui langkah apa yang harus dilakukan

Dalam kasus dataset ini mengubah nilai -9.0 menjadi nilai nilai null valuse sesuai dengan deskripsi dataset

```
df.replace(-9.0, np.nan, inplace=True)
```

megnghitung jumlah nilai null value

```
df.isnull().sum()
```

```
1       0
2       0
3       0
4       0
5       0
      ...
70      0
71      0
72      0
73    266
74    294
Length: 74, dtype: int64
```

```
df.head()
```

	1	2	3	4	5	6	7	8	9	10	...	65	66	67	68	69	70

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 74 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   1       294 non-null    float64
 1   2       294 non-null    float64
 2   3       294 non-null    float64
 3   4       294 non-null    float64
 4   5       294 non-null    float64
 5   6       294 non-null    float64
 6   7       0 non-null      float64
 7   8       294 non-null    float64
 8   9       293 non-null    float64
 9   10      293 non-null    float64
 10  11      271 non-null    float64
 11  12      12 non-null     float64
 12  13      1 non-null      float64
 13  14      0 non-null      float64
 14  15      286 non-null    float64
 15  16      21 non-null     float64
 16  17      1 non-null      float64
 17  18      293 non-null    float64
 18  19      294 non-null    float64
 19  20      294 non-null    float64
 20  21      294 non-null    float64
 21  22      293 non-null    float64
 22  23      292 non-null    float64
 23  24      293 non-null    float64
 24  25      293 non-null    float64
 25  26      293 non-null    float64
 26  27      285 non-null    float64
 27  28      292 non-null    float64
 28  29      104 non-null    float64
 29  30      292 non-null    float64
 30  31      293 non-null    float64
 31  32      293 non-null    float64
 32  33      293 non-null    float64
 33  34      293 non-null    float64
 34  35      293 non-null    float64
 35  36      293 non-null    float64
 36  37      293 non-null    float64
 37  38      292 non-null    float64
 38  39      294 non-null    float64
 39  40      104 non-null    float64
 40  41      293 non-null    float64
 41  42      294 non-null    float64
 42  43      4 non-null      float64
 43  44      0 non-null      float64
 44  45      0 non-null      float64
 45  46      0 non-null      float64
 46  47      3 non-null      float64
 47  48      0 non-null      float64
 48  49      2 non-null      float64
 49  50      28 non-null     float64
 50  51      27 non-null     float64
 51  52      17 non-null     float64
 52  53      0 non-null      float64
```

4) Menentukan Object Data

Memilih 14 fitur yang akan digunakan sesuai dengan deskripsi dataset

```
df_selected = df.iloc[:, [1, 2, 7, 8, 10, 14, 17, 30, 36, 38, 39, 42, 49, 56]]
```

```
df_selected.head()
```

	2	3	8	9	11	15	18	31	37	39	40	43	50	57
0	40.0	1.0	2.0	140.0	289.0	0.0	0.0	172.0	0.0	0.0	NaN	NaN	NaN	0.0
1	49.0	0.0	3.0	160.0	180.0	0.0	0.0	156.0	0.0	1.0	2.0	NaN	NaN	1.0
2	37.0	1.0	2.0	130.0	283.0	0.0	1.0	98.0	0.0	0.0	NaN	NaN	NaN	0.0
3	48.0	0.0	4.0	138.0	214.0	0.0	0.0	108.0	1.0	1.5	2.0	NaN	NaN	3.0
4	54.0	1.0	3.0	150.0	NaN	0.0	0.0	122.0	0.0	0.0	NaN	NaN	NaN	0.0

```
df_selected.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 14 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   2       294 non-null    float64
 1   3       294 non-null    float64
 2   8       294 non-null    float64
 3   9       293 non-null    float64
 4   11      271 non-null    float64
 5   15      286 non-null    float64
 6   18      293 non-null    float64
 7   31      293 non-null    float64
 8   37      293 non-null    float64
 9   39      294 non-null    float64
 10  40      104 non-null    float64
 11  43      4 non-null      float64
 12  50      28 non-null     float64
 13  57      294 non-null    float64
dtypes: float64(14)
memory usage: 32.3 KB
```

mengganti nama 14 kolom sesuai dengan deskripsi dataset

```
column_mapping = { 2: 'age',
        3: 'sex',
        8: 'cp',
        9: 'trestbps',
        11: 'chol',
        15: 'fbs',
        18: 'restecg',
        31: 'thalach',
        37: 'exang',
        39: 'oldpeak',
        40: 'slope',
        43: 'ca',
        50: 'thal',
        57: 'target'
    }
```

```
df_selected.rename(columns=column_mapping, inplace=True)
```

```
<ipython-input-17-e9a4003b4301>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/inde
  df_selected.rename(columns=column_mapping, inplace=True)
```

```
df_selected.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       294 non-null    float64
 1   sex       294 non-null    float64
 2   cp        294 non-null    float64
 3   trestbps  293 non-null    float64
 4   chol      271 non-null    float64
 5   fbs       286 non-null    float64
 6   restecg   293 non-null    float64
 7   thalach   293 non-null    float64
 8   exang     293 non-null    float64
 9   oldpeak   294 non-null    float64
 10  slope     104 non-null    float64
 11  ca        4 non-null      float64
 12  thal      28 non-null     float64
 13  target    294 non-null    float64
dtypes: float64(14)
memory usage: 32.3 KB
```

menghitung jumlah fitur pada dataset

```
df_selected.value_counts()
```

```
age   sex  cp   trestbps  chol   fbs  restecg  thalach  exang  oldpeak  slope  ca   thal  target
47.0  1.0  4.0  150.0     226.0  0.0  0.0      98.0     1.0    1.5      2.0    0.0  7.0   1.0
1
dtype: int64
```

```
df_selected
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	c
0	40.0	1.0	2.0	140.0	289.0	0.0	0.0	172.0	0.0	0.0	NaN	Nal

5) Membersihkan data

3	48.0	0.0	4.0	138.0	214.0	0.0	0.0	108.0	1.0	1.5	2.0	Nal

menghitung jumlah null values pada dataset

```
df_selected.isnull().sum()
```

```
age           0
sex           0
cp            0
trestbps      1
chol         23
fbs           8
restecg       1
thalach       1
exang         1
oldpeak       0
slope       190
ca          290
thal        266
target        0
dtype: int64
```

Berdasarkan output kode program diatas ada beberapa fitur yang hampir 90% datanya memiliki nilai null (cont kolom "slope", "ca", "thal") sehingga perlu dilakukan penghapusan fitur menggunakan fungsi drop

```
columns_to_drop = ['ca', 'slope','thal']
df_selected = df_selected.drop(columns_to_drop, axis=1)
```

```
df_selected.isnull().sum()
```

```
age          0
sex          0
cp           0
trestbps     1
chol        23
fbs          8
restecg      1
thalach      1
exang        1
oldpeak      0
target       0
dtype: int64
```

Keterangan diatas mennjukan bahwa masih ada nilai null, maka dari itu akan diisi dengan nilai mean atau rata-rata setiap kolom

```
meanTBPS = df_selected['trestbps'].dropna()
meanChol = df_selected['chol'].dropna()
meanfbs = df_selected['fbs'].dropna()
meanRestCG = df_selected['restecg'].dropna()
meanthalach = df_selected['thalach'].dropna()
meanexang = df_selected['exang'].dropna()
```

```
meanTBPS = meanTBPS.astype(float)
meanChol = meanChol.astype(float)
meanfbs = meanfbs.astype(float)
meanthalach = meanthalach.astype(float)
meanexang = meanexang.astype(float)
meanRestCG = meanRestCG.astype(float)
```

```
meanTBPS = round(meanTBPS.mean())
meanChol = round(meanChol.mean())
meanfbs = round(meanfbs.mean())
meanthalach = round(meanthalach.mean())
meanexang = round(meanexang.mean())
meanRestCG = round(meanRestCG.mean())
```

mengubah nilai null menjadi nilai mean yang sudah ditentukan sebelumnya

```
fill_values = {'trestbps': meanTBPS, 'chol': meanChol, 'fbs': meanfbs, 'thalach':meanthalach,'exang':m
dfClean = df_selected.fillna(value=fill_values)
```

```
dfClean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       294 non-null    float64
 1   sex       294 non-null    float64
 2   cp        294 non-null    float64
 3   trestbps  294 non-null    float64
 4   chol      294 non-null    float64
 5   fbs       294 non-null    float64
 6   restecg   294 non-null    float64
 7   thalach   294 non-null    float64
 8   exang     294 non-null    float64
 9   oldpeak   294 non-null    float64
 10  target    294 non-null    float64
dtypes: float64(11)
memory usage: 25.4 KB
```

```
dfClean.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
target      0
dtype: int64
```

melalukan pengecekan terhadap duplikaksi data

```
duplicate_rows = dfClean.duplicated()
dfClean[duplicate_rows]
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	target
163	49.0	0.0	2.0	110.0	251.0	0.0	0.0	160.0	0.0	0.0	0.0

```
print("All Duplicate Rows:")
dfClean[dfClean.duplicated(keep=False)]
```

```
All Duplicate Rows:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	target
90	49.0	0.0	2.0	110.0	251.0	0.0	0.0	160.0	0.0	0.0	0.0
163	49.0	0.0	2.0	110.0	251.0	0.0	0.0	160.0	0.0	0.0	0.0

Menghapus data yang memiliki duplikat

```
dfClean = dfClean.drop_duplicates()
print("All Duplicate Rows:")
dfClean[dfClean.duplicated(keep=False)]
```

```
All Duplicate Rows:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	target

```
dfClean.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	target
0	40.0	1.0	2.0	140.0	289.0	0.0	0.0	172.0	0.0	0.0	0.0
1	49.0	0.0	3.0	160.0	180.0	0.0	0.0	156.0	0.0	1.0	1.0
2	37.0	1.0	2.0	130.0	283.0	0.0	1.0	98.0	0.0	0.0	0.0
3	48.0	0.0	4.0	138.0	214.0	0.0	0.0	108.0	1.0	1.5	3.0
4	54.0	1.0	3.0	150.0	251.0	0.0	0.0	122.0	0.0	0.0	0.0

```
dfClean['target'].value_counts()
```

```
0.0    187
1.0     37
3.0     28
2.0     26
4.0     15
Name: target, dtype: int64
```

```
import seaborn as sns
import matplotlib.pyplot as plt
```

mencari korelasi antar fitur

```
dfClean.corr()
```

	age	sex	cp	trestbps	chol	fbs	restecg	tha
age	1.000000	0.014516	0.146616	0.246571	0.087101	0.181130	0.050672	-0.46
sex	0.014516	1.000000	0.245769	0.082064	0.027695	0.044372	-0.108656	-0.10
cp	0.146616	0.245769	1.000000	0.081293	0.134697	0.031930	-0.016372	-0.36
trestbps	0.246571	0.082064	0.081293	1.000000	0.080818	0.096222	0.011256	-0.18
chol	0.087101	0.027695	0.134697	0.080818	1.000000	0.107686	0.048081	-0.12
fbs	0.181130	0.044372	0.031930	0.096222	0.107686	1.000000	0.047988	-0.06
restecg	0.050672	-0.108656	-0.016372	0.011256	0.048081	0.047988	1.000000	0.00
thalach	-0.460514	-0.106959	-0.367819	-0.181824	-0.122038	-0.069722	0.006084	1.00
exang	0.239223	0.154925	0.494674	0.211507	0.161055	0.115503	0.041290	-0.40
oldpeak	0.178172	0.115959	0.351735	0.204000	0.106743	0.063179	0.042193	-0.30
target	0.210429	0.220732	0.427536	0.214898	0.256027	0.154319	0.042643	-0.36

```
cor_mat=dfClean.corr()
fig,ax=plt.subplots(figsize=(15,10))
sns.heatmap(cor_mat,annot=True,linewidths=0.5,fmt=".3f")
```

```
<Axes: >
```