Pembelajaran Mesin (Praktikum) SI –B4

FAKULTTAS VOKASI
UNIVERSITAS AIRLANGGA

[152111283042] | [Nela Anjani] | [21 September 2023]

LOGBOOK

TUGAS :

DATA
MISSING VALUE
OUTLIER

1. Tugas dikerjakan secara mandiri

2. Carilah data bebas

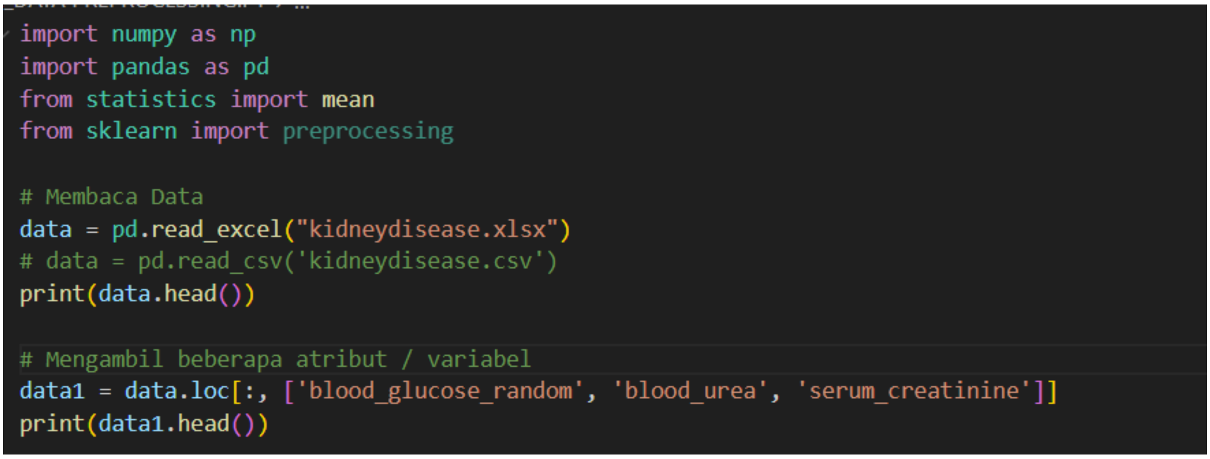
3. Lakukan PreProcessing dengan memeriksa missing value dan outlier.

4. Lakukan penanganan missing value dan outlier (dihapus dan direplace dengan Mean)

5. Tugas terdiri dari Laporan, file phyton, dan data aslinya dan dikumpulkan dengan nama “Tugas PreProcessing_NIM. Zip

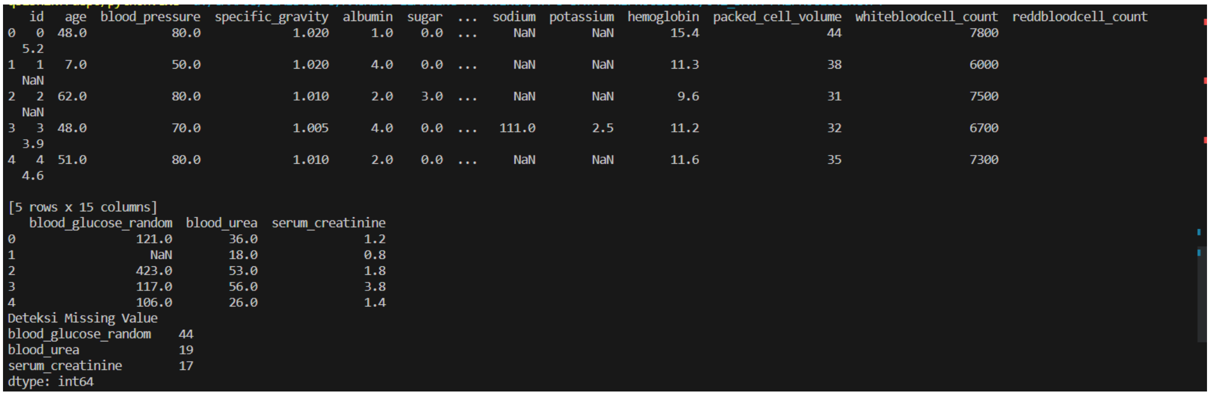
6. Tugas dikumpulkan paling lambat hari Kamis / 14 Sepetember 2023 pukul 20.00 Wib

- Program



```python
import numpy as np
import pandas as pd
from statistics import mean
from sklearn import preprocessing

# Membaca Data
data = pd.read_excel("kidneydisease.xlsx")
# data = pd.read_csv('kidneydisease.csv')
print(data.head())

# Mengambil beberapa atribut / variabel
data1 = data.loc[:, ['blood_glucose_random', 'blood_urea', 'serum_creatinine']]
print(data1.head())
```

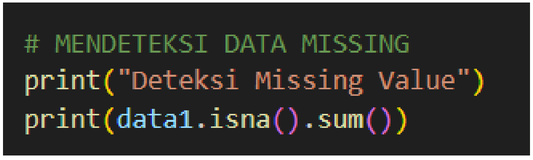
- Ouput yang dihasilkan



```
   id   age  blood_pressure  specific_gravity  albumin  sugar  ...  sodium  potassium  hemoglobin  packed_cell_volume  whitebloodcell_count  reddbloodcell_count
0   0  48.0            80.0             1.020      1.0    0.0  ...     NaN        NaN        15.4                  44                  7800
  5.2
1   1   7.0            50.0             1.020      4.0    0.0  ...     NaN        NaN        11.3                  38                  6000
  NaN
2   2  62.0            80.0             1.010      2.0    3.0  ...     NaN        NaN         9.6                  31                  7500
  NaN
3   3  48.0            70.0             1.005      4.0    0.0  ...   111.0        2.5        11.2                  32                  6700
  3.9
4   4  51.0            80.0             1.010      2.0    0.0  ...     NaN        NaN        11.6                  35                  7300
  4.6

[5 rows x 15 columns]
   blood_glucose_random  blood_urea  serum_creatinine
0                 121.0        36.0               1.2
1                   NaN        18.0               0.8
2                 423.0        53.0               1.8
3                 117.0        56.0               3.8
4                 106.0        26.0               1.4
Deteksi Missing Value
blood_glucose_random    44
blood_urea              19
serum_creatinine        17
dtype: int64
```

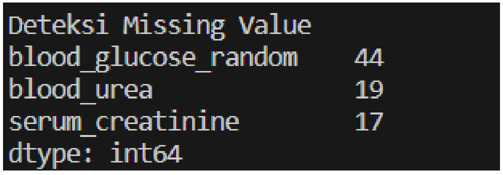
MISSING VALUE

- Deteksi missing value



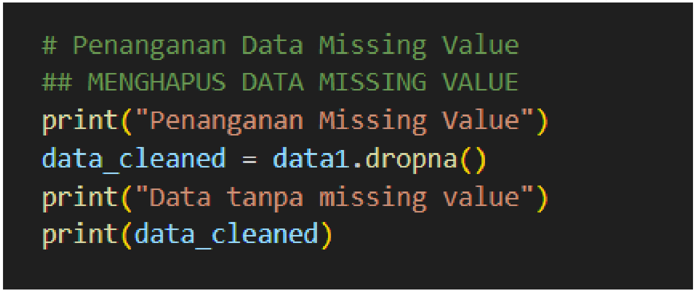
```python
# MENDETEKSI DATA MISSING
print("Deteksi Missing Value")
print(data1.isna().sum())
```

Output yang dihasilkan



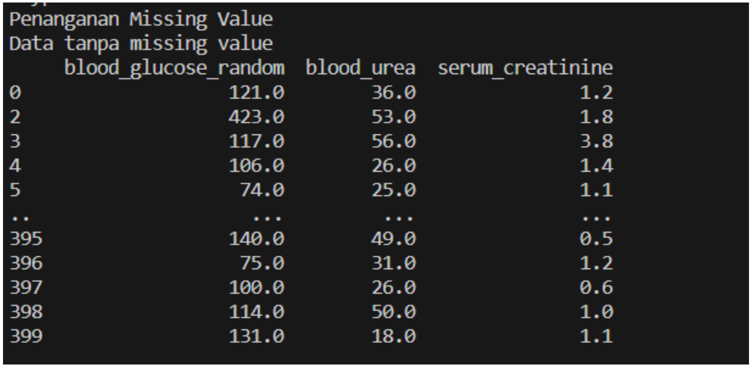
```
Deteksi Missing Value
blood_glucose_random    44
blood_urea              19
serum_creatinine        17
dtype: int64
```

- Penanganan missing value (menghapus missing value)



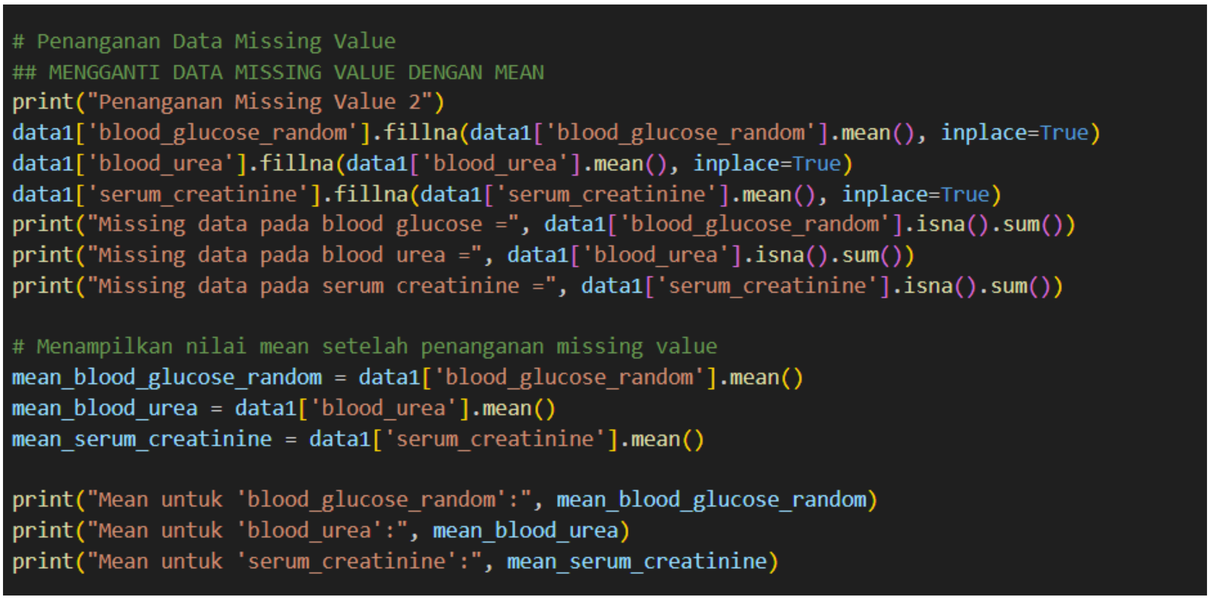
```python
# Penanganan Data Missing Value
## MENGHAPUS DATA MISSING VALUE
print("Penanganan Missing Value")
data_cleaned = data1.dropna()
print("Data tanpa missing value")
print(data_cleaned)
```

- Output



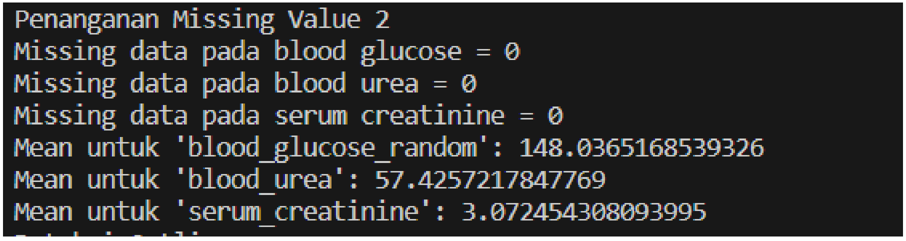
```
Penanganan Missing Value
Data tanpa missing value
     blood_glucose_random  blood_urea  serum_creatinine
0                   121.0        36.0               1.2
2                   423.0        53.0               1.8
3                   117.0        56.0               3.8
4                   106.0        26.0               1.4
5                    74.0        25.0               1.1
..                    ...         ...               ...
395                 140.0        49.0               0.5
396                  75.0        31.0               1.2
397                 100.0        26.0               0.6
398                 114.0        50.0               1.0
399                 131.0        18.0               1.1
```

- Penanganan Missing Value (mengganti missing value dengan nilai rata-rata (mean)



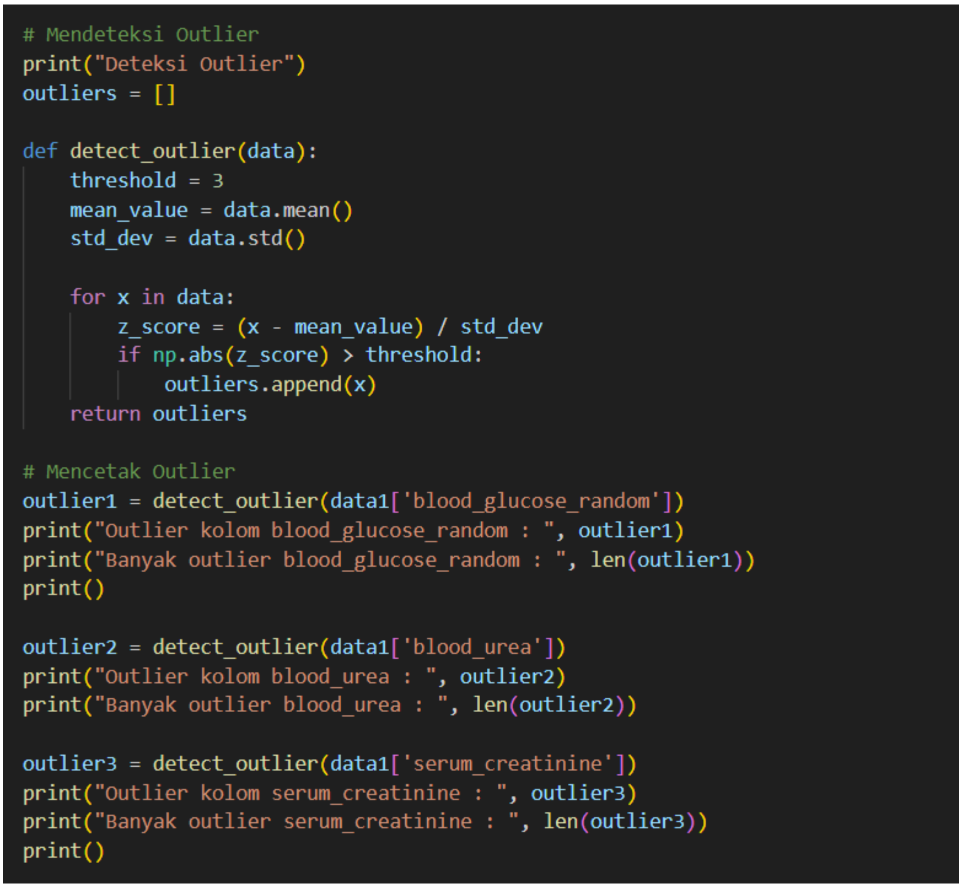
```
# Penanganan Data Missing Value
## MENGGANTI DATA MISSING VALUE DENGAN MEAN
print("Penanganan Missing Value 2")
data1['blood_glucose_random'].fillna(data1['blood_glucose_random'].mean(), inplace=True)
data1['blood_urea'].fillna(data1['blood_urea'].mean(), inplace=True)
data1['serum_creatinine'].fillna(data1['serum_creatinine'].mean(), inplace=True)
print("Missing data pada blood glucose =", data1['blood_glucose_random'].isna().sum())
print("Missing data pada blood urea =", data1['blood_urea'].isna().sum())
print("Missing data pada serum creatinine =", data1['serum_creatinine'].isna().sum())

# Menampilkan nilai mean setelah penanganan missing value
mean_blood_glucose_random = data1['blood_glucose_random'].mean()
mean_blood_urea = data1['blood_urea'].mean()
mean_serum_creatinine = data1['serum_creatinine'].mean()

print("Mean untuk 'blood_glucose_random':", mean_blood_glucose_random)
print("Mean untuk 'blood_urea':", mean_blood_urea)
print("Mean untuk 'serum_creatinine':", mean_serum_creatinine)
```

- Ouput



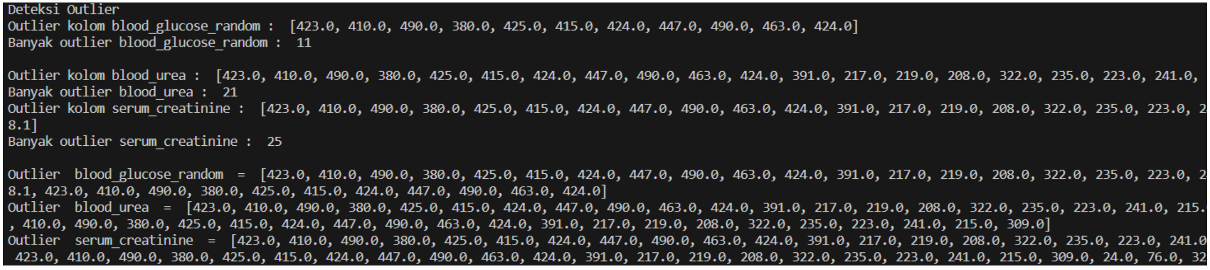
```
Penanganan Missing Value 2
Missing data pada blood glucose = 0
Missing data pada blood urea = 0
Missing data pada serum creatinine = 0
Mean untuk 'blood_glucose_random': 148.0365168539326
Mean untuk 'blood_urea': 57.4257217847769
Mean untuk 'serum_creatinine': 3.072454308093995
```

- Deteksi outlier



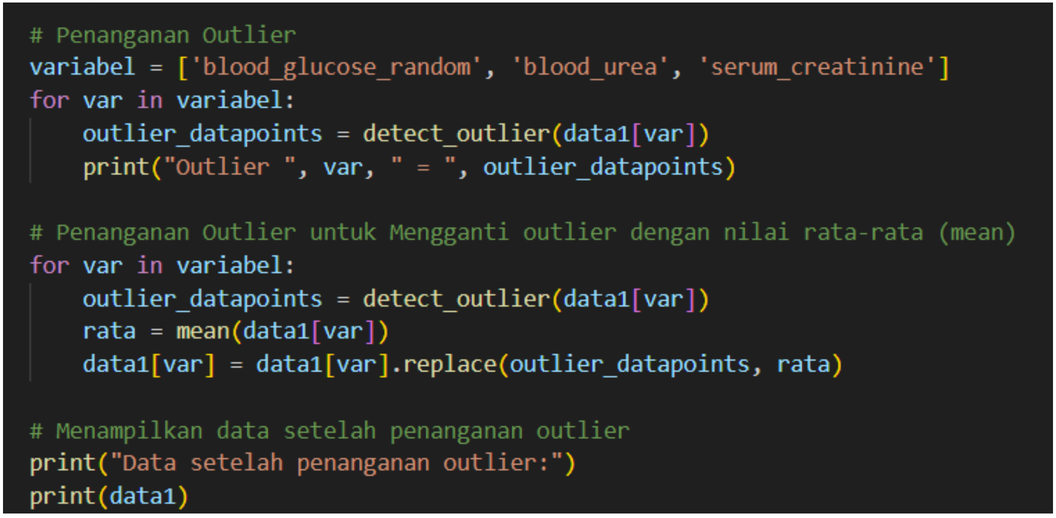
```
# Mendeteksi Outlier
print("Deteksi Outlier")
outliers = []

def detect_outlier(data):
    threshold = 3
    mean_value = data.mean()
    std_dev = data.std()

    for x in data:
        z_score = (x - mean_value) / std_dev
        if np.abs(z_score) > threshold:
            outliers.append(x)
    return outliers

# Mencetak Outlier
outlier1 = detect_outlier(data1['blood_glucose_random'])
print("Outlier kolom blood_glucose_random : ", outlier1)
print("Banyak outlier blood_glucose_random : ", len(outlier1))
print()

outlier2 = detect_outlier(data1['blood_urea'])
print("Outlier kolom blood_urea : ", outlier2)
print("Banyak outlier blood_urea : ", len(outlier2))

outlier3 = detect_outlier(data1['serum_creatinine'])
print("Outlier kolom serum_creatinine : ", outlier3)
print("Banyak outlier serum_creatinine : ", len(outlier3))
print()
```

- Output



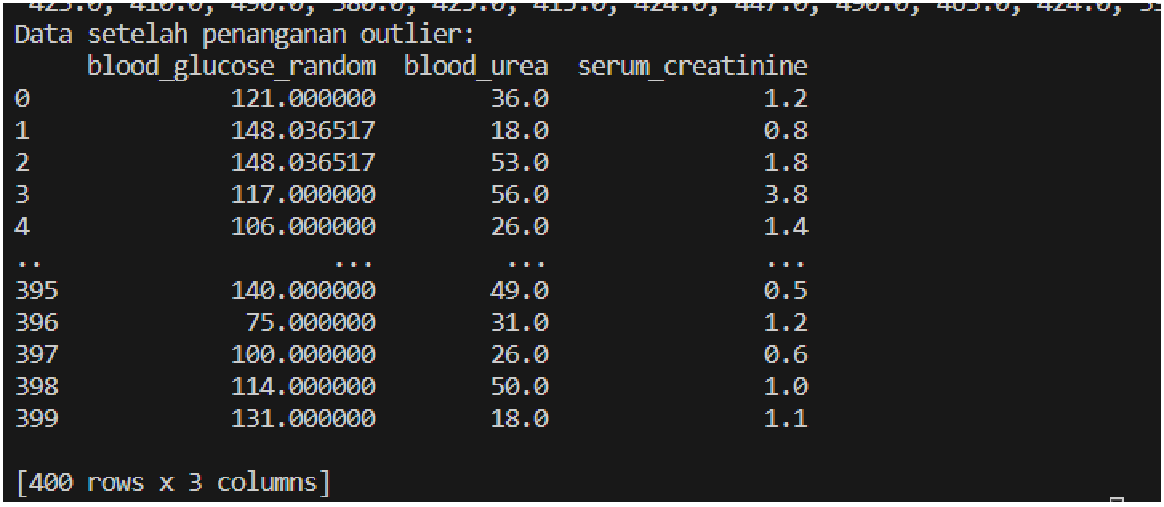
```
Deteksi Outlier
Outlier kolom blood_glucose_random :  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0]
Banyak outlier blood_glucose_random :  11

Outlier kolom blood_urea :  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 241.0,
Banyak outlier blood_urea :  21
Outlier kolom serum_creatinine :  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 2
8.1]
Banyak outlier serum_creatinine :  25

Outlier  blood_glucose_random  =  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 2
8.1, 423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0]
Outlier  blood_urea  =  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 241.0, 215.
, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 241.0, 215.0, 309.0]
Outlier  serum_creatinine  =  [423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 241.0
 423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 391.0, 217.0, 219.0, 208.0, 322.0, 235.0, 223.0, 241.0, 215.0, 309.0, 24.0, 76.0, 32
```

- Penanganan outlier



```
# Penanganan Outlier
variabel = ['blood_glucose_random', 'blood_urea', 'serum_creatinine']
for var in variabel:
    outlier_datapoints = detect_outlier(data1[var])
    print("Outlier ", var, " = ", outlier_datapoints)

# Penanganan Outlier untuk Mengganti outlier dengan nilai rata-rata (mean)
for var in variabel:
    outlier_datapoints = detect_outlier(data1[var])
    rata = mean(data1[var])
    data1[var] = data1[var].replace(outlier_datapoints, rata)

# Menampilkan data setelah penanganan outlier
print("Data setelah penanganan outlier:")
print(data1)
```

- Ouput



```
 423.0, 410.0, 490.0, 380.0, 425.0, 415.0, 424.0, 447.0, 490.0, 463.0, 424.0, 3
Data setelah penanganan outlier:
     blood_glucose_random  blood_urea  serum_creatinine
0               121.000000        36.0               1.2
1               148.036517        18.0               0.8
2               148.036517        53.0               1.8
3               117.000000        56.0               3.8
4               106.000000        26.0               1.4
..                     ...         ...               ...
395             140.000000        49.0               0.5
396              75.000000        31.0               1.2
397             100.000000        26.0               0.6
398             114.000000        50.0               1.0
399             131.000000        18.0               1.1

[400 rows x 3 columns]
```