

**LAPORAN DATA MINING**

**DATA PREPROCESSING**



DISUSUN OLEH:

Gilbert Ciady (09011182328005)

DOSEN PENGAMPU :  
Dr. Firdaus, M.Kom.  
Anggun Islami, M.Kom.

**PROGRAM STUDI SISTEM KOMPUTER**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS SRIWIJAYA**

**2024**

## **KATA PENGANTAR**

Puji syukur penulis ucapkan kepada Tuhan Yang Maha Esa karena berkat rahmat-Nya sehingga penulis bisa menyelesaikan penyusunan laporan ini dengan tepat waktu. Laporan ini disusun dan dibuat untuk memenuhi tugas mata kuliah.

Tidak lupa ucapkan terima kasih kepada selaku dosen pengampu mata kuliah yaitu Bapak Dr. Firdaus, M.Kom. dan Ibu Anggun Islami, M.Kom. yang telah membimbing penulis.

Penulis menyadari bahwa masih terdapat kekurangan, baik dari penyusunan maupun tata Bahasa. Oleh karena itu, penulis meminta maaf sebesar-besarnya atas kekurangan tersebut. Semoga laporan ini bermanfaat bagi kita semua, sekian dan Terima kasih.

Indralaya, 11 September 2024  
Penulis,

Gilbert Ciady

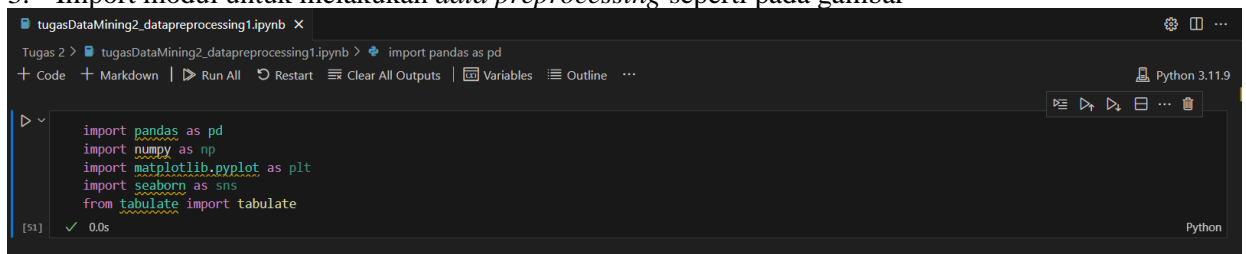
## I. Pendahuluan

Data mining merupakan proses untuk mengekstrak informasi penting dari data yang banyak dan kompleks. Tujuan dari data mining sendiri adalah untuk menemukan pola tersembunyi untuk membantu dalam pengambilan keputusan dalam berbagai bidang seperti bisnis, kesehatan, kehidupan sosial, dll.

Salah satu tahap yang penting dalam data mining adalah *data preprocessing*. *Data preprocessing* diperlukan sebagai Langkah awal yang penting, dikarenakan data yang diperoleh biasanya tidak rapi, memiliki *missing values* (nilai yang hilang), mengandung *outliers* (nilai tidak wajar/terpaut jauh), dll. Tujuan utama dari *data preprocessing* adalah mengubah data mentah menjadi lebih bersih dan konsisten agar mudah untuk diolah dan dianalisis.

## II. Langkah - Langkah Pengerjaan

1. Buka aplikasi untuk memulai proses *data preprocessing*, bisa menggunakan Jupyter dan Anaconda, pada kesempatan ini, penulis menggunakan Visual Studio Code dengan ekstensi Jupyter dan Python sebagai Bahasa pemrogramannya
2. Buat file dengan format “ ‘nama’.ipynb ”
3. Import modul untuk melakukan *data preprocessing* seperti pada gambar

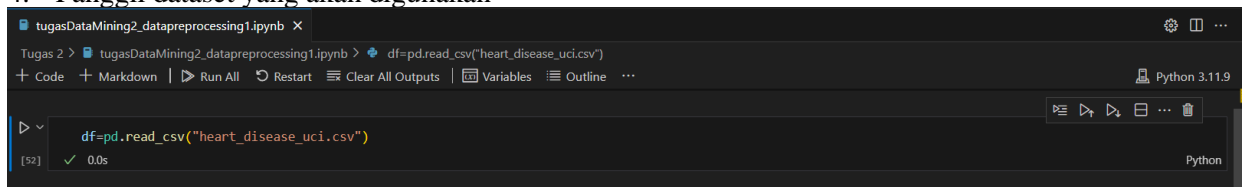


```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > import pandas as pd
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ...
Python 3.11.9

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from tabulate import tabulate

[51] ✓ 0.0s Python
```

4. Panggil dataset yang akan digunakan

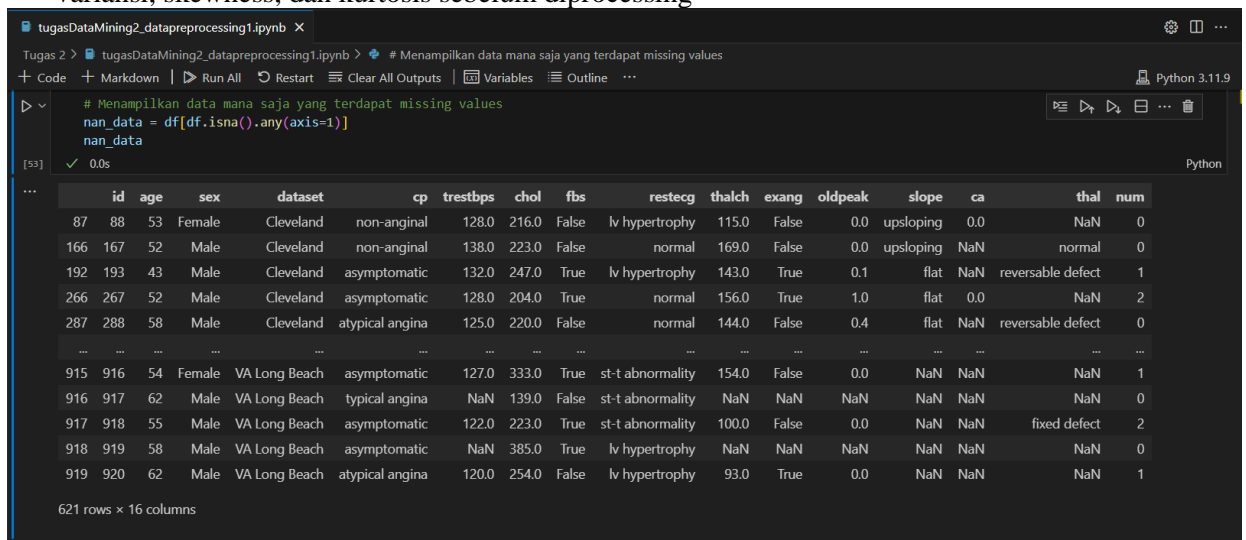


```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > df=pd.read_csv("heart_disease_uci.csv")
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ...
Python 3.11.9

df=pd.read_csv("heart_disease_uci.csv")

[52] ✓ 0.0s Python
```

5. Tampilkan terkait dengan *missing values* dan hitungan mean, mode, median, standar deviasi, variansi, skewness, dan kurtosis sebelum diprocessing



```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > # Menampilkan data mana saja yang terdapat missing values
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ...
Python 3.11.9

# Menampilkan data mana saja yang terdapat missing values
nan_data = df[df.isna().any(axis=1)]
nan_data

[53] ✓ 0.0s Python
```

id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num	
87	88	53	Female	Cleveland	non-anginal	128.0	216.0	False	lv hypertrophy	115.0	False	0.0	upsloping	0.0	NaN	0
166	167	52	Male	Cleveland	non-anginal	138.0	223.0	False	normal	169.0	False	0.0	upsloping	NaN	normal	0
192	193	43	Male	Cleveland	asymptomatic	132.0	247.0	True	lv hypertrophy	143.0	True	0.1	flat	NaN	reversable defect	1
266	267	52	Male	Cleveland	asymptomatic	128.0	204.0	True	normal	156.0	True	1.0	flat	0.0	NaN	2
287	288	58	Male	Cleveland	atypical angina	125.0	220.0	False	normal	144.0	False	0.4	flat	NaN	reversable defect	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
915	916	54	Female	VA Long Beach	asymptomatic	127.0	333.0	True	st-t abnormality	154.0	False	0.0	NaN	NaN	NaN	1
916	917	62	Male	VA Long Beach	typical angina	NaN	139.0	False	st-t abnormality	NaN	NaN	NaN	NaN	NaN	NaN	0
917	918	55	Male	VA Long Beach	asymptomatic	122.0	223.0	True	st-t abnormality	100.0	False	0.0	NaN	NaN	fixed defect	2
918	919	58	Male	VA Long Beach	asymptomatic	NaN	385.0	True	lv hypertrophy	NaN	NaN	NaN	NaN	NaN	NaN	0
919	920	62	Male	VA Long Beach	atypical angina	120.0	254.0	False	lv hypertrophy	93.0	True	0.0	NaN	NaN	NaN	1

621 rows x 16 columns

tugasDataMining2\_datapreprocessing1.ipynb
Python 3.11.9
# Menampilkan informasi terkait non missing value dan tipe data sebelum di processing
df.info()
Python
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 920 entries, 0 to 919  
Data columns (total 16 columns):  
# Column Non-Null Count Dtype  
---  
0 id 920 non-null int64  
1 age 920 non-null int64  
2 sex 920 non-null object  
3 dataset 920 non-null object  
4 cp 920 non-null object  
5 trestbps 861 non-null float64  
6 chol 890 non-null float64  
7 fbs 830 non-null object  
8 restecg 918 non-null object  
9 thalch 865 non-null float64  
10 exang 865 non-null object  
11 oldpeak 858 non-null float64  
12 slope 611 non-null object  
13 ca 309 non-null float64  
14 thal 434 non-null object  
15 num 920 non-null int64  
dtypes: float64(5), int64(3), object(8)  
memory usage: 115.1+ KB

tugasDataMining2\_datapreprocessing1.ipynb
Python 3.11.9
#Menampilkan Hasil Menghitung Mean, Median, Mode, Standar Deviasi, Variansi, Skewness, dan Kurtosis sebelum di processing
desc\_stats = df.describe().reset\_index()
desc\_stats.index = range(1, len(desc\_stats) + 1)
print(tabulate(desc\_stats, headers='keys', tablefmt='outline'))
Python

	index	id	age	trestbps	chol	thalch	oldpeak	ca	num
1	count	920	920	861	890	865	858	309	920
2	mean	460.5	53.5109	132.132	199.13	137.546	0.878788	0.676375	0.995652
3	std	265.725	9.42469	19.0661	110.781	25.9263	1.09123	0.935653	1.14269
4	min	1	28	0	0	60	-2.6	0	0
5	25%	230.75	47	120	175	120	0	0	0
6	50%	460.5	54	130	223	140	0.5	0	1
7	75%	690.25	60	140	268	157	1.5	1	2
8	max	920	77	200	603	202	6.2	3	4

## 6. Ganti *missing values* sesuai dengan tipe objek dari data

tugasDataMining2\_datapreprocessing1.ipynb
Python 3.11.9
# Mengganti Missng Value Numerik
df['trestbps'] = df['trestbps'].fillna(df['trestbps'].mean())
df['chol'] = df['chol'].fillna(df['chol'].mean())
df['thalch'] = df['thalch'].fillna(df['thalch'].mean())
for col in ['oldpeak', 'ca']:
df[col] = df[col].fillna(df[col].mode()[0])
Python
# Mengganti Missng Value Boolean
for col in ['fbs', 'exang']:
df[col] = df[col].fillna(df[col].mode()[0])
Python
# Mengganti Missng Value Ordinal / kategorial
for col in ['restecg', 'slope', 'thal']:
df[col] = df[col].fillna(df[col].mode()[0])
Python

## 7. Tampilkan terkait dengan *missing values*

```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > # Mengganti Missing Value Numerik
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ... Python 3.11.9

# Menampilkan informasi terkait non missing value dan tipe data setelah di processing
df = df.infer_objects()
df.info()

[59] ✓ 0.0s Python

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 16 columns):
#   Column      Non-Null Count  Dtype
---  -
0    id          920 non-null    int64
1    age         920 non-null    int64
2    sex         920 non-null    object
3    dataset     920 non-null    object
4    cp          920 non-null    object
5    trestbps    920 non-null    float64
6    chol        920 non-null    float64
7    fbs         920 non-null    bool
8    restecg     920 non-null    object
9    thalch      920 non-null    float64
10   exang       920 non-null    bool
11   oldpeak     920 non-null    float64
12   slope       920 non-null    object
13   ca          920 non-null    float64
14   thal        920 non-null    object
15   num         920 non-null    int64
dtypes: bool(2), float64(5), int64(3), object(6)
memory usage: 102.6+ KB
```

## 8. Ganti nilai yang terdapat outliers (nilai tidak wajar/terpaut jauh)

```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > # Mengganti Outliers 'chol'
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ... Python 3.11.9

# Mengganti Outliers 'chol'
Q1 = df['chol'].quantile(0.25)
Q3 = df['chol'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Mengganti outliers dengan batas bawah dan atas
df['chol'] = df['chol'].apply(lambda x: upper_bound if x > upper_bound else (lower_bound if x < lower_bound else x))

[75] ✓ 0.0s Python

# Mengganti Outliers 'oldpeak'
df['oldpeak'] = df['oldpeak'].abs()

[76] ✓ 0.0s Python

# Mengganti Outliers 'trestbps'
mean_trestbps = df['trestbps'].mean()
df['trestbps'] = df['trestbps'].replace(0, mean_trestbps)

[77] ✓ 0.0s Python
```

## 9. Tampilkan terkait hitungan mean, mode, median, standar deviasi, variansi, skewness, dan kurtosis setelah diprocessing

```
tugasDataMining2_datapreprocessing1.ipynb X
Tugas 2 > tugasDataMining2_datapreprocessing1.ipynb > # Mengganti Outliers 'chol'
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ... Python 3.11.9

#Menampilkan Hasil Menghitung Mean, Median, Mode, Standar Deviasi, Variansi, Skewness, dan Kurtosis setelah di processing
desc_stats = df.describe().reset_index()
desc_stats.index = range(1, len(desc_stats) + 1)
print(tabulate(desc_stats, headers='keys', tablefmt='outline'))

[63] ✓ 0.0s Python

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| | index | id | age | trestbps | chol | thalch | oldpeak | ca | num |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | count | 920 | 920 | 920 | 920 | 920 | 920 | 920 | 920 |
| 2 | mean | 460.5 | 53.5109 | 132.276 | 206.317 | 137.546 | 0.835978 | 0.227174 | 0.995652 |
| 3 | std | 265.725 | 9.42469 | 17.9209 | 91.493 | 25.1385 | 1.05453 | 0.628936 | 1.14269 |
| 4 | min | 1 | 28 | 80 | 43.875 | 60 | 0 | 0 | 0 |
| 5 | 25% | 230.75 | 47 | 120 | 177.75 | 120 | 0 | 0 | 0 |
| 6 | 50% | 460.5 | 54 | 130 | 221 | 138 | 0.2 | 0 | 1 |
| 7 | 75% | 690.25 | 60 | 140 | 267 | 156 | 1.5 | 0 | 2 |
| 8 | max | 920 | 77 | 200 | 400.875 | 202 | 6.2 | 3 | 4 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

## 10. Tampilkan dataset yang telah diprocessing

```
tugasDataMining2.dataprocessing1.ipynb X
Tugas 2 > tugasDataMining2.dataprocessing1.ipynb > # Menampilkan dataset lengkap yang telah di processing
+ Code + Markdown | Run All | Restart | Clear All Outputs | Variables | Outline ... Python 3.11.9

# Menampilkan dataset lengkap yang telah di processing
pd.set_option("display.precision", 2)
print(df.to_string())

[44] ✓ 0.0s Python

...
   id  age  sex  dataset  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  num
0    1   63  Male  Cleveland  typical angina  145.00  233.00  True  lv hypertrophy  150.00  False  2.3  downsloping  0.0  fixed defect  0
1    2   67  Male  Cleveland  asymptomatic  160.00  286.00  False  lv hypertrophy  100.00  True  1.5  flat  3.0  normal  2
2    3   67  Male  Cleveland  asymptomatic  120.00  220.00  False  lv hypertrophy  120.00  True  2.6  flat  2.0  reversable defect  1
3    4   37  Male  Cleveland  non-anginal  130.00  250.00  False  normal  187.00  False  3.5  downsloping  0.0  normal  0
4    5   41  Female  Cleveland  atypical angina  130.00  204.00  False  lv hypertrophy  172.00  False  1.4  upsloping  0.0  normal  0
5    6   56  Male  Cleveland  atypical angina  120.00  236.00  False  normal  178.00  False  0.8  upsloping  0.0  normal  0
6    7   62  Female  Cleveland  asymptomatic  140.00  268.00  False  lv hypertrophy  160.00  False  3.6  downsloping  2.0  normal  3
7    8   57  Female  Cleveland  asymptomatic  120.00  354.00  False  normal  163.00  True  0.6  upsloping  0.0  normal  0
8    9   63  Male  Cleveland  asymptomatic  130.00  254.00  False  lv hypertrophy  147.00  False  1.4  flat  1.0  reversable defect  2
9   10   53  Male  Cleveland  asymptomatic  140.00  203.00  True  lv hypertrophy  155.00  True  3.1  downsloping  0.0  reversable defect  1
10   11   57  Male  Cleveland  asymptomatic  140.00  192.00  False  normal  148.00  False  0.4  flat  0.0  fixed defect  0
11   12   56  Female  Cleveland  atypical angina  140.00  204.00  False  lv hypertrophy  153.00  False  1.3  flat  0.0  normal  0
12   13   56  Male  Cleveland  non-anginal  130.00  256.00  True  lv hypertrophy  142.00  True  0.6  flat  1.0  fixed defect  2
13   14   44  Male  Cleveland  atypical angina  120.00  263.00  False  normal  173.00  False  0.0  upsloping  0.0  reversable defect  0
14   15   52  Male  Cleveland  non-anginal  172.00  199.00  True  normal  162.00  False  0.5  upsloping  0.0  reversable defect  0
15   16   57  Male  Cleveland  non-anginal  150.00  168.00  False  normal  174.00  False  1.6  upsloping  0.0  normal  0
16   17   48  Male  Cleveland  atypical angina  110.00  229.00  False  normal  168.00  False  1.0  downsloping  0.0  reversable defect  1
17   18   54  Male  Cleveland  asymptomatic  140.00  239.00  False  normal  160.00  False  1.2  upsloping  0.0  normal  0
18   19   48  Female  Cleveland  non-anginal  130.00  275.00  False  normal  139.00  False  0.2  upsloping  0.0  normal  0
19   20   49  Male  Cleveland  atypical angina  130.00  266.00  False  normal  171.00  False  0.6  upsloping  0.0  normal  0
20   21   40  Male  Cleveland  typical angina  110.00  211.00  False  lv hypertrophy  144.00  True  1.8  flat  0.0  normal  0
21   22   58  Female  Cleveland  typical angina  150.00  283.00  True  lv hypertrophy  162.00  False  1.0  upsloping  0.0  normal  0
22   23   58  Male  Cleveland  atypical angina  120.00  284.00  False  lv hypertrophy  160.00  False  1.8  flat  0.0  normal  1
23   24   58  Male  Cleveland  non-anginal  132.00  224.00  False  lv hypertrophy  173.00  False  3.2  upsloping  2.0  reversable defect  3
...
916  917   62  Male  VA Long Beach  typical angina  132.13  139.00  False  st-t abnormality  137.55  False  0.0  flat  0.0  normal  0
917  918   55  Male  VA Long Beach  asymptomatic  122.00  223.00  True  st-t abnormality  100.00  False  0.0  flat  0.0  fixed defect  2
918  919   58  Male  VA Long Beach  asymptomatic  132.13  385.00  True  lv hypertrophy  137.55  False  0.0  flat  0.0  normal  0
919  920   62  Male  VA Long Beach  atypical angina  120.00  254.00  False  lv hypertrophy  93.00  True  0.0  flat  0.0  normal  1

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

## III. Kesimpulan

Dataset diatas merupakan dataset mengenai *heart disease* atau penyakit jantung. Dataset diatas memiliki *missing value* yang tergolong cukup parah karena beberapa variabel pada data tersebut seperti ‘ca’ dan ‘thal’ *missing value*-nya mencapai lebih dari 50% dan variabel ‘slope’ yang memiliki *missing value* sekitar 33%. Dataset ini juga terdapat *outliers* seperti pada variabel ‘chol’, ‘oldpeak’, dan ‘trestbps’. Variabel ‘chol’ dan ‘trestbps’ memiliki nilai minimum yang tidak mungkin terjadi dan tidak mungkin dimasukan di dataset jika itu benar, yaitu “0”, variabel ‘oldpeak’ juga memiliki nilai minimum tidak wajar di “-2.6” yang sangat tidak mungkin terjadi.

Penanganan untuk *missing value* pada dataset ini adalah dengan menggantinya dengan nilai “mean” untuk variabel ‘trestbps’, ‘chol’, dan ‘thalch’ dan dengan nilai “mode” untuk variabel ‘oldpeak’, ‘ca’, ‘fbs’, ‘exang’, ‘restecg’, ‘slope’, dan ‘thal’. Penanganan untuk *outliers* pada variabel ‘chol’ menggunakan nilai *Interquartile Range*, variabel ‘oldpeak’ dengan mengganti nilai minus menjadi nilai positif, dan variabel ‘trestbps’ dengan nilai “mean”.