

<p><b>Nama:</b> Mendari pertiwi</p> <p><b>NIM:</b> 064002200037</p>	 <p><b>Praktikum Statistika</b></p>	<p><b>MODUL 7</b></p> <p><b>Nama Dosen:</b> Dedy Sugiarto</p>
<p><b>Hari/Tanggal:</b> rabu, 2 agustus 2023</p>		<p><b>Nama Asisten Labratorium</b>  <b>1. Elen Fadilla Estri</b>  <b>064002000008</b>  <b>2. Rukhy Zaifa Aduhalim</b>  <b>064002000041</b></p>

## Data Preprocessing Menggunakan Python

### 1. Teori Singkat

Data Preprocessing adalah sebuah tahapan awal dalam sebuah pengolahan data sebelum data diaplikasikan dengan algoritma machine learning. Data yang biasanya kita gunakan dalam kehidupan sehari-hari — hari entah itu dari database, data excel dan sumber lainnya, merupakan data unstruktur (datanya tidak sempurna). Misalkan dalam sebuah dataset (kumpulan data) terdapat data yang kosong, tipe data yang berbeda dengan yang lain, dan sebagainya. Masalah tersebut harus bisa kita selesaikan terlebih dahulu agar data yang kita kelola lebih mudah dan outputnya sesuai dengan yang kita harapkan.

Terdapat beberapa case yang akan kita pelajari satu per satu, antara lain seperti:

- Mengimport libraries
- Mengimport dataset
- Menangani data kosong di dataset
- Mengolah data string menjadi kategori
- Membagi dataset menjadi training dan test set
- Feature Scaling

### Informasi Dataset

Sumber Data: Kaggle

Deskripsi: Memberikan informasi dari penumpang Titanic yang selamat dan tidak.

Jumlah data: 1309

Jumlah atribut: 12 (termasuk class)

Terdiri dari:

- PassengerId urutan nomor data dari penumpang
- Survived: status selamat (0:meninggal, 1:selamat)
- Pclass: kelas kamar dari penumpang (1: highclass, 2:midclass, 3:lowclass)
- Name: nama penumpang
- Sex: jenis kelamin penumpang (male, female)
- Age: umur penumpang
- SibSp: jumlah saudara kandung dan pasangan dari penumpang yang ada di kapal
- Parch: jumlah orangtua dan anak dari penumpang
- Ticket: kode tiket penumpang
- Fare: ongkos tiket yang dibeli penumpang
- Cabin: Kode kabin
- Embarked: Kota keberangkatan penumpang (C:Cherbourg, Q:Queenstown, S:Southampton)

## 2. Alat dan Bahan

Hardware : Laptop/PC

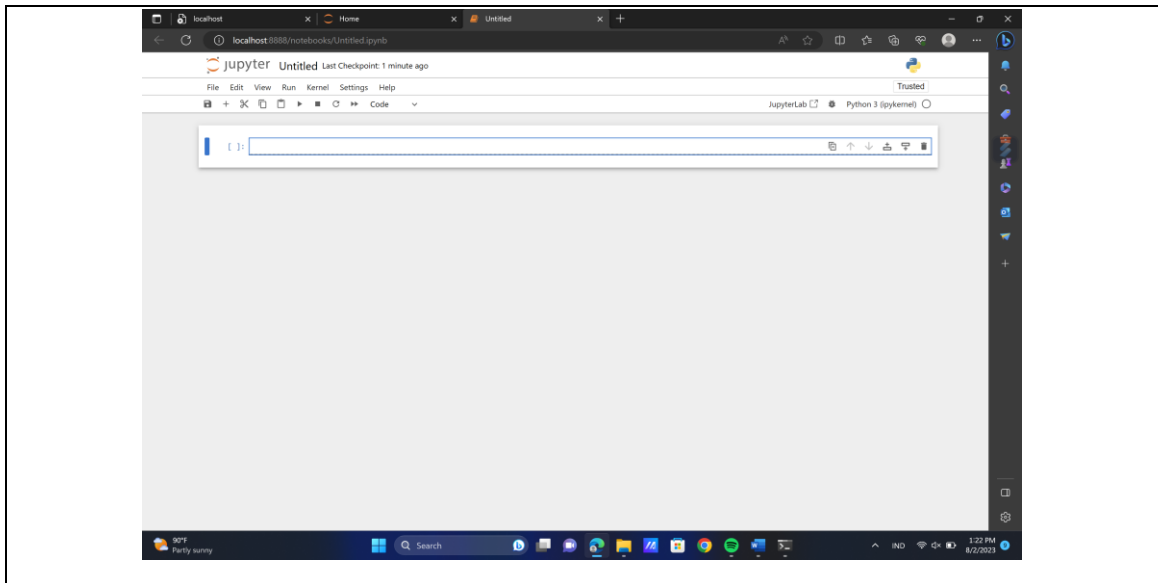
Software : R Studio

## 3. Elemen Kompetensi

a. Latihan pertama – Materi

1. Buka Jupyter Notebook atau gunakan Google Colab

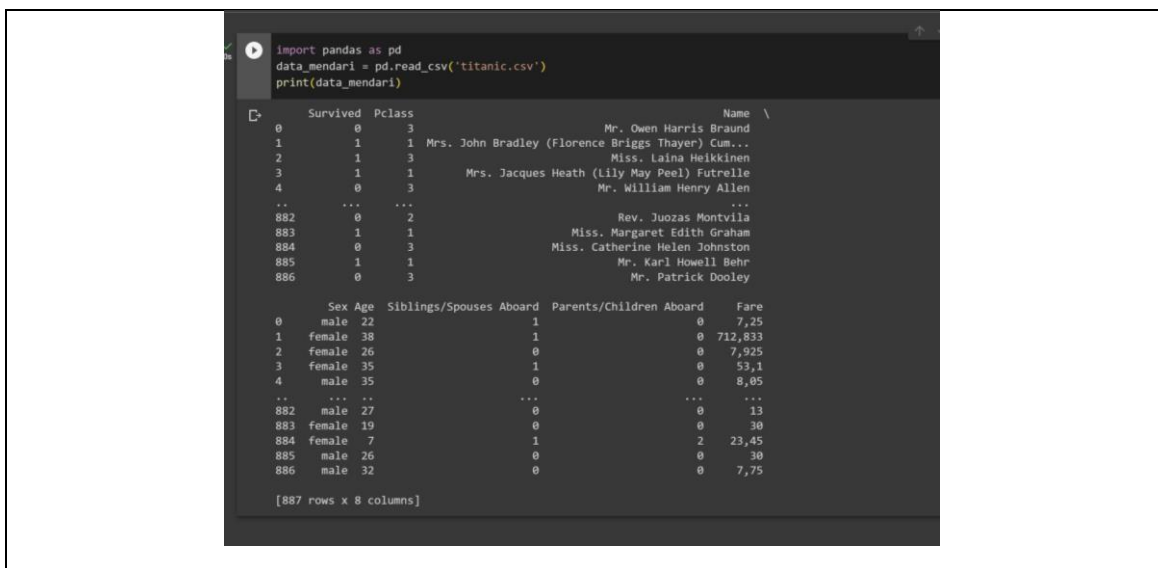




## 2. Script

```
import pandas as pd
#memanggil dan menampilkan dataset
data_nama = pd.read_csv('D:/dll/titanic.csv')
print(data_nama)
```

Output:



### 3. Script

```
#mengambil data pada kolom tertentu  
data1 = data_nama.loc[:,['Age','Pclass','Survived']]  
print(data1)
```

Output:

```
data1 = data_mendari.loc[:,['Age','Pclass','Survived']]  
print(data1)
```

	Age	Pclass	Survived
0	22	3	0
1	38	1	1
2	26	3	1
3	35	1	1
4	35	3	0
..	..	...	...
882	27	2	0
883	19	1	1
884	7	3	0
885	26	1	1
886	32	3	0

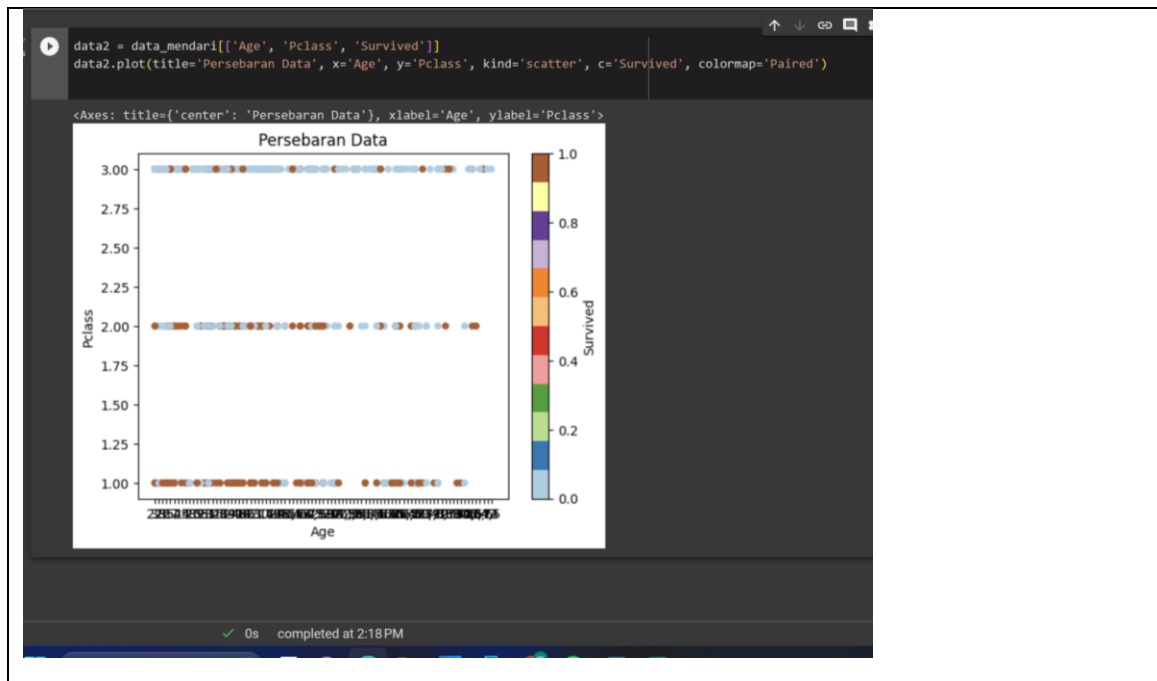
[887 rows x 3 columns]

### 4. Script

```
#memvisualisasikan data titanic  
data2 = data_nama[['Age', 'Pclass', 'Survived']]  
data2.plot(title='Persebaran Data', x='Age', y='Pclass', kind='scatter', c='Survived',  
colormap='Paired')
```

Output:





## 5. Script

```
#memanipulasi data jumlah penumpang berdasarkan group Pclass
data3 = data_nama[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
penumpang=data3.groupby('Pclass')['Name'].nunique()
print('Jumlah Penumpang:\n', penumpang)
```

Output:

```
data3 = data_nama[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
penumpang=data3.groupby('Pclass')['Name'].nunique()
print('jumlah penumpang:\n', penumpang)
```

jumlah penumpang:  
Pclass  
1 216  
2 184  
3 487  
Name: Name, dtype: int64

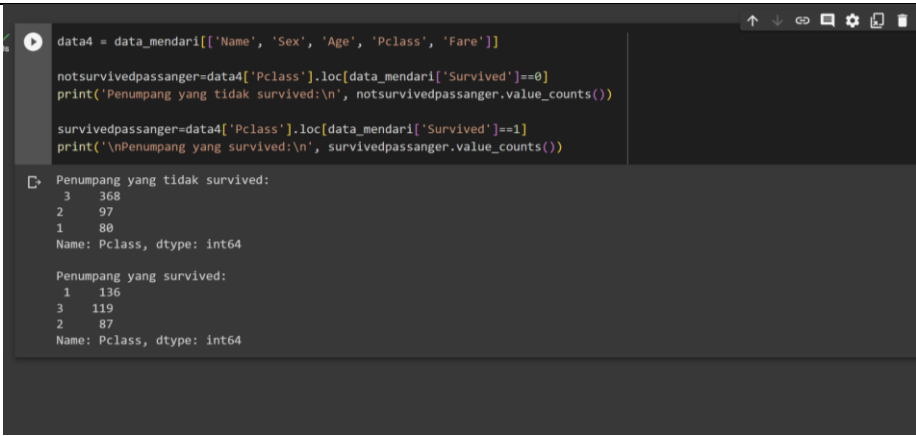
## 6. Script

```
#memfilter data penumpang yang selamat berdasarkan pclass
data4 = data_nama[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
```



```
notsurvivedpassanger=data4['Pclass'].loc[data_nama['Survived']==0]
print('Penumpang yang tidak survived:\n', notsurvivedpassanger.value_counts())
survivedpassanger=data4['Pclass'].loc[data_nama['Survived']==1]
print('\nPenumpang yang survived:\n', survivedpassanger.value_counts())
```

Output:



```
data4 = data_mendari[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
notsurvivedpassanger=data4['Pclass'].loc[data_mendari['Survived']==0]
print('Penumpang yang tidak survived:\n', notsurvivedpassanger.value_counts())
survivedpassanger=data4['Pclass'].loc[data_mendari['Survived']==1]
print('\nPenumpang yang survived:\n', survivedpassanger.value_counts())
```

Penumpang yang tidak survived:

3	368
2	97
1	80

Name: Pclass, dtype: int64

Penumpang yang survived:

1	136
3	119
2	87

Name: Pclass, dtype: int64

## b. Latihan Kedua – Tugas

1. Buatlah manipulasi data jumlah penumpang berdasarkan group sex

Script:

```
data4 = data_mendari[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
notsurvivedpassanger=data4['Sex'].loc[data_mendari['Survived']==0]
```

Output:



```
data3 = data_mendari[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]
penumpang=data3.groupby('Sex')['Name'].nunique()
print('Jumlah Penumpang:\n', penumpang)
```

Jumlah Penumpang:

Sex	
female	314
male	573

Name: Name, dtype: int64



Penjelasan: memanipulasi data jumlah penumpang berdasarkan group sex agar lebih teroganisir dan mudah dibaca. diatas terlihat jumlah penumpang:sex,female ada 314 dan male ada 573

2. Buatlah filter data penumpang yang selamat berdasarkan sex  
Script:

```
data4 = data_mendari[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]  
  
notsurvivedpassanger=data4['Sex'].loc[data_mendari['Survived']==0]  
print('Penumpang yang tidak survived:\n', notsurvivedpassanger.value_counts())  
survivedpassanger=data4['Sex'].loc[data_mendari['Survived']==1]  
print('\nPenumpang yang survived:\n', survivedpassanger.value_counts())
```

Output:



```
Name: Name, dtype: int64  
  
data4 = data_mendari[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]  
notsurvivedpassanger=data4['Sex'].loc[data_mendari['Survived']==0]  
print('Penumpang yang tidak survived:\n', notsurvivedpassanger.value_counts())  
survivedpassanger=data4['Sex'].loc[data_mendari['Survived']==1]  
print('\nPenumpang yang survived:\n', survivedpassanger.value_counts())  
  
Penumpang yang tidak survived:  
male    464  
female    81  
Name: Sex, dtype: int64  
  
Penumpang yang survived:  
female    233  
male     109  
Name: Sex, dtype: int64
```

Penjelasan: menyaring data dan mendapat informasi yang relevan yaitu sesuai dari data diatas yaitu ada penumpang yang tidak survived dan ada penumpang yang survived.

#### 4. File Praktikum

Github Repository:

#### 5. Soal Latihan

Soal:

1. Perintah apa yang digunakan untuk mengimport data CSV pada bahasa pemrograman python?
2. Apa kegunaan dari filter data?



Jawaban:

1. bisa menggunakan library pandas. Pandas yaitu salah satu library populer untuk manipulasi dan analisis data
2. untuk menyaring data berdasarkan kriteria tertentu sehingga hanya data yang memenuhi kriteria tersebut yang akan ditampilkan atau diproses.

## 6. Kesimpulan

- a. Dalam pengerjaan praktikum Statistika, belajar tentang berbagai teknik dan proses untuk membersihkan, mengubah, dan menyiapkan data agar siap untuk analisis lebih lanjut.
- b. Kita juga dapat mengetahui pentingnya tahap Data Preprocessing dalam analisis data dan pembuatan model machine learning. Langkah-langkah yang di pelajari dalam praktikum ini membantu memastikan bahwa data yang digunakan untuk analisis lebih lanjut adalah data yang berkualitas dan relevan.

## 7. Cek List (✓)

No	Elemen Kompetensi	Penyelesaian	
		Selesai	Tidak Selesai
1.	Latihan Pertama	✓	
2.	Latihan Kedua	✓	

## 8. Formulir Umpan Balik

No	Elemen Kompetensi	Waktu Pengerjaan	Kriteria
1.	Latihan Pertama	40 Menit	menarik
2.	Latihan Kedua	40 Menit	menarik

Keterangan:

1. Menarik
2. Baik
3. Cukup
4. Kurang