


Nama: Francisco NIM: 064002300044		MODUL 7 Nama Dosen: Drs. Joko Riyono, M.Si
Hari/Tanggal: Rabu, 08-05-2024		Nama Asisten Labratorium 1. Adzriel Yusak Noah (064.20.11) 2. Mohammad Akhmal Firdaus (064.20.34)

Deleted[A]:

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 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

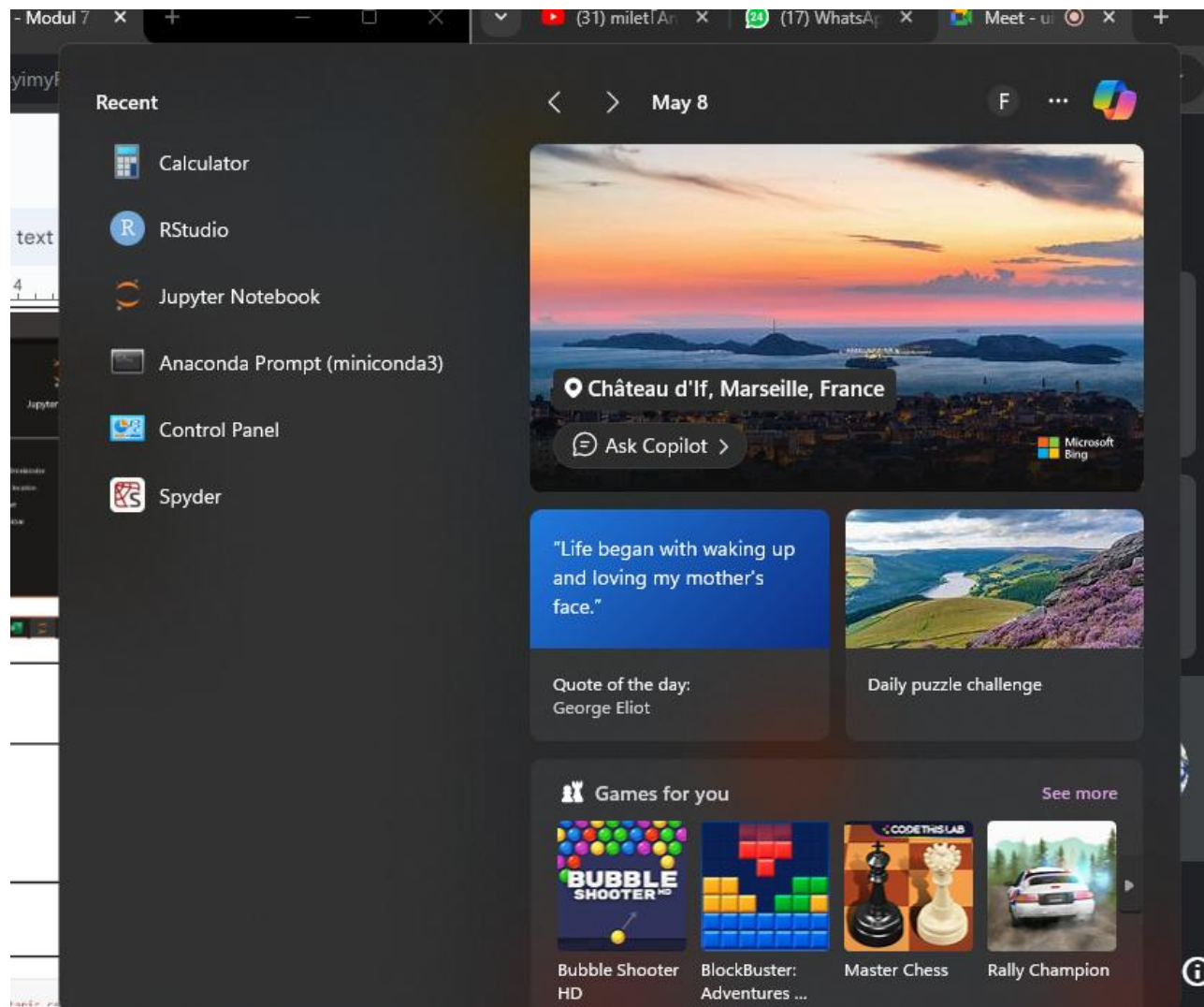
Hardware : Laptop/PC

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:

```
In [10]: import pandas as pd
data_francisco = pd.read_csv("C:/Users/Francisco/Downloads/titanic - titanic.csv")
print(data_francisco)
```

	Survived	Pclass	Name \
0	0	3	Mr. Owen Harris Braund
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...
2	1	3	Miss. Laina Heikkinen
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle
4	0	3	Mr. William Henry Allen
..
882	0	2	Rev. Juozas Montvila
883	1	1	Miss. Margaret Edith Graham
884	0	3	Miss. Catherine Helen Johnston
885	1	1	Mr. Karl Howell Behr
886	0	3	Mr. Patrick Dooley

	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare,,
0	male	22	1	0	7,25,
1	female	38	1	0	71,2833,
2	female	26	0	0	7,925,
3	female	35	1	0	53,1,
4	male	35	0	0	8,5,
..
882	male	27	0	0	13,,
---	-	-	-	-	--

3. Script

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

Output:

```
data1 = data_francisco.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]

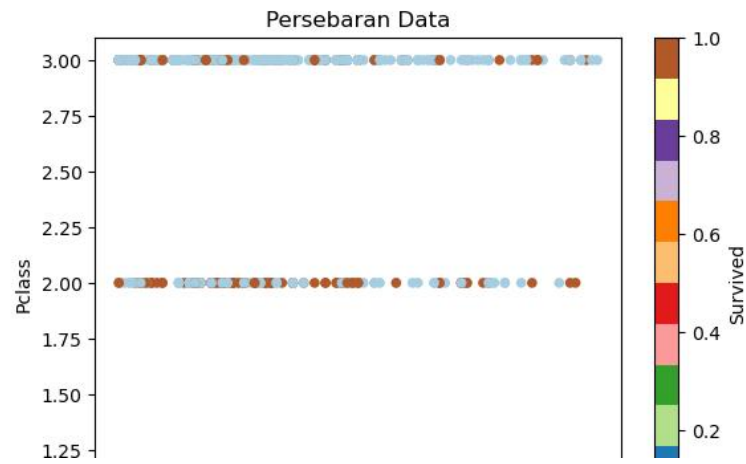
4. Script

```
#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:

```
in [12]: data2=data_francisco[['Age', 'Pclass', 'Survived']]
data2.plot(title='Persebaran Data', x='Age', y='Pclass', kind='scatter', c='Survived', colormap='Paired')

Out[12]: <Axes: title={'center': 'Persebaran Data'}, xlabel='Age', ylabel='Pclass'>
```



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_francisco[['Name', 'Sex', 'Age', 'Pclass', 'Fa
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_francisco[['Name', 'Sex', 'Age', 'Pclass', 'Fare',
notsurvivedpassanger=data4['Pclass'].loc[data_francisco['Su
print('Penumpang yang tidak survived:\n', notsurvivedpassan
survivedpassanger=data4['Pclass'].loc[data_francisco['Survi
print('\nPenumpang yang survived:\n', survivedpassanger.val
```

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:

```
data5 = data_francisco[['Name', 'Sex', 'Age', 'Pclass', 'Fare',,]]
penumpang = data3.groupby('Sex')['Name'].nunique()
print('Jumlah Penumpang\n', penumpang)
```

Output:

```
data5 = data_francisco[['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: secara keseluruhan, program ini memilih beberapa kolom dari DataFrame, mengelompokkan baris berdasarkan jenis kelamin, menghitung jumlah penumpang yang unik untuk setiap jenis kelamin, dan kemudian mencetak hasilnya.

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


```
data6 =  
data_francisco[['Name','Sex','Age','Pclass','Fare,,']]  
  
survivedpassanger =  
data6['Sex'].loc[data_francisco['Survived']==1]  
print('PenumpangYang Selamat:\n',  
survivedpassanger.value_counts())
```

Output:

```
data6 = data_francisco[['Name','Sex','Age','Pclass','Fare,,  
  
survivedpassanger = data6['Sex'].loc[data_francisco['Surviv  
print('PenumpangYang Selamat:\n', survivedpassanger.value_c
```

```
PenumpangYang Selamat:  
female    233  
male      109  
Name: Sex, dtype: int64
```

Penjelasan: secara keseluruhan, program ini memilih beberapa kolom dari DataFrame, memilih penumpang yang selamat, menghitung jumlah penumpang yang selamat untuk setiap jenis kelamin, dan kemudian mencetak hasilnya.

4. File Praktikum

Github Repository:

https://github.com/franafuk/Praktikum_stat

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. menggunakan import pandas as pd untuk mengimport data csvnya dan untuk melihat filecsvnya tunjukan lokasi dari filenya yang kita simpan
2. Kegunaannya yaitu untuk menganalisis dan memvisualisasi data untuk lebih spesifik dan sebagai solusi pemecahan masalah statistika.

6. Kesimpulan

- a. Dalam pengerjaan praktikum Statistika, tergolong sangat mudah
- b. Kita juga dapat mengetahui bahwasannya dalam materi dari manipulasi data di praktikum ini tergolong penting dalam statistika karna kita dapat memecahkan hal-hal penting untuk mengetahui apa saja yang kita ingin cari yang berkaitan dengan pertanyaan dalam konteks statistika dan matematika, dengan ini kita memahami lebih tentang materi di modul ini.

7. Cek List (✓)

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

Formulir Umpan Balik

No	Elemen Kompetensi	Waktu Pengerjaan	Kriteria
1.	Latihan Pertama	10 Menit	Menarik
2.	Latihan Kedua	10 Menit	Menarik

Keterangan:

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