| **Nama:**  **(Adrian Halim)**  **NIM: (064002200043)** | C:\Users\RPL-SI 02\Pictures\288px-Trisakti_Logo.svg.png | **MODUL 7**  **Nama Dosen:**  **Dedy Sugiarto** |
| --- | --- | --- |
| **Hari/Tanggal:**  **Rabu, 2 AGUSTUS 2023** | **Praktikum Statistika** | **Nama Asisten Labratorium 1. Elen Fadilla Estri**  **064002000008**  **2. Rukhy Zaifa Aduhalim**  **064002000041** |

**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)

1. **Alat dan Bahan**

Hardware : Laptop/PC

Software : R Studio

1. **Elemen Kompetensi**
   1. 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:

|  |
| --- |

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:

|  |
| --- |

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:

|  |
| --- |

* 1. Latihan Kedua – Tugas

1. Buatlah manipulasi data jumlah penumpang berdasarkan group sex

Script:

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

Output:

|  |
| --- |

Penjelasan: Kode di atas memanipulasi data penumpang berdasarkan jenis kelamin (Sex) dari sebuah dataset yang disimpan dalam variabel "data\_adrian". Berikut adalah penjelasan singkat langkah-langkahnya:

1 .Dari dataset "data\_adrian", dibuat DataFrame baru "data3" yang hanya memuat kolom-kolom tertentu ('Name', 'Sex', 'Age', 'Pclass', 'Fare').

2 .Data tersebut dikelompokkan berdasarkan kolom "Sex" (jenis kelamin) menggunakan metode "groupby()".

3 .Menghitung jumlah penumpang unik (berbeda) dalam setiap kelompok menggunakan metode "nunique()" berdasarkan kolom "Name".

4 .Hasil jumlah penumpang per jenis kelamin dicetak menggunakan "print('Jumlah Penumpang:\n', penumpang)".

Dengan kode ini, kita dapat melihat jumlah penumpang unik per jenis kelamin dari dataset "data\_adrian".

2. Buatlah filter data penumpang yang selamat berdasarkan sex

Script:

| #memfilter data penumpang yang selamat berdasarkan sex  data4 = data\_adrian[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]  survivedpassanger=data4['Sex'].loc[data\_adrian['Survived']==1]  print('\nPenumpang yang survived:\n', survivedpassanger.value\_counts()) |
| --- |

Output:

|  |
| --- |

Penjelasan: Kode tersebut mengambil dataset penumpang yang disimpan dalam variabel "data\_adrian" dan melakukan filter berdasarkan jenis kelamin (sex) untuk memisahkan penumpang yang selamat. Hasilnya adalah kelompok data: untuk penumpang yang selamat. Setelah itu, kode mencetak jumlah penumpang selamat dan tidak selamat berdasarkan jenis kelamin dari dataset tersebut.

1. **File Praktikum**

Github Repository:

|  |
| --- |

1. **Soal Latihan**

Soal:

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

Jawaban:  
1. import pandas as pd

#memanggil dan menampilkan dataset

data\_nama = pd.read\_csv('titanic.csv')

2. Kalau dilihat dari Latihan 2 bagian b dengan memfilter data yang sealamat berdasarkan sex.. melakukan filter data untuk memisahkan penumpang yang selamat dan tidak selamat berdasarkan jenis kelamin (Sex) dari dataset "data\_adrian". Hasilnya adalah dua kelompok data: satu untuk penumpang yang selamat dan satu lagi untuk penumpang yang tidak selamat. Setelah itu, kode mencetak jumlah penumpang selamat dan tidak selamat berdasarkan jenis kelamin.

1. **Kesimpulan**
   1. Dalam pengerjaan praktikum Statistika, Kita mempelajari memproses data, memanipulasi, memfilter dan memvisualisasi data.
   2. Kita juga dapat mengetahui 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).
2. **Cek List (✔)**

| **No** | **Elemen Kompetensi** | **Penyelesaian** | |
| --- | --- | --- | --- |
| **Selesai** | **Tidak Selesai** |
| **1.** | Latihan Pertama | **✅** |  |
| **2.** | Latihan Kedua | **✅** |  |

1. **Formulir Umpan Balik**

| **No** | **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| --- | --- | --- | --- |
| **1.** | Latihan Pertama | 10 Menit | Menarik |
| **2.** | Latihan Kedua | 30 Menit | Menarik |

Keterangan:

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