|  |  |  |
| --- | --- | --- |
| **Nama:**  **Dimas Dwi Saputtra**  **NIM: 064002200041** | C:\Users\RPL-SI 02\Pictures\288px-Trisakti_Logo.svg.png | **MODUL 7**  **Nama Dosen:**  **Dedy Sugiarto** |
| **Hari/Tanggal:**  **Kamis, 10 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:

|  |
| --- |
| # Import library pandas  import pandas as pd  # Membaca dataset dari file "dimas.csv" yang telah diunggah  data\_nama = pd.read\_csv('dimas.csv')  # Memilih kolom yang akan digunakan  data5 = data\_nama[['Name', 'Sex', 'Age', 'Pclass', 'Fare']]  # Mengelompokkan data berdasarkan Sex dan menghitung jumlah penumpang berdasarkan jenis kelamin  penumpang\_by\_sex = data5.groupby('Sex')['Name'].nunique()  # Menampilkan jumlah penumpang berdasarkan jenis kelamin  print('Jumlah Penumpang Berdasarkan Jenis Kelamin:\n', penumpang\_by\_sex) |

Output:

|  |
| --- |
|  |

Penjelasan: untuk mengelompokkan data berdasarkan kolom "Sex" (jenis kelamin). Kemudian, dengan menggunakan metode nunique(), kita menghitung jumlah penumpang unik dalam setiap kelompok jenis kelamin.

Menampilkan Hasil: Hasil dari perhitungan jumlah penumpang berdasarkan jenis kelamin ditampilkan menggunakan perintah print.

Hasil dari kode ini adalah informasi tentang jumlah penumpang yang unik dalam setiap kelompok jenis kelamin. Ini memberikan gambaran tentang distribusi jenis kelamin di antara penumpang dalam dataset.

2. Buatlah filter data penumpang yang selamat berdasarkan sex

Script:

|  |
| --- |
| # Import library pandas  import pandas as pd  # Membaca dataset dari file "dimas.csv" yang telah diunggah  data\_nama = pd.read\_csv('dimas.csv')  # Memilih kolom yang akan digunakan  data6 = data\_nama[['Name', 'Sex', 'Age', 'Pclass', 'Fare', 'Survived']]  # Memfilter penumpang yang selamat berdasarkan jenis kelamin (Sex)  survived\_passengers\_by\_sex = data6.loc[data6['Survived'] == 1, 'Sex']  # Menampilkan jumlah penumpang selamat berdasarkan jenis kelamin  print('Penumpang yang Selamat Berdasarkan Jenis Kelamin:\n', survived\_passengers\_by\_sex.value\_counts()) |

Output:

|  |
| --- |
|  |

Penjelasan: Memfilter Penumpang yang Selamat: Dengan menggunakan kondisi boolean data6['Survived'] == 1, kita dapat memfilter baris-baris dari DataFrame yang hanya memiliki nilai "Survived" yang sama dengan 1 (yang berarti penumpang yang selamat).

1. **File Praktikum**

Github Repository:

|  |
| --- |
| https://github.com/Dimasassr/Praktikum\_Statistika.git |

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. Dalam perintah di atas, import pandas as pd mengimpor pustaka pandas dan memberikan alias 'pd' untuk memudahkan penggunaan. Kemudian, pd.read\_csv('nama\_file.csv') digunakan untuk membaca data dari file CSV dengan nama "nama\_file.csv" dan menyimpannya dalam bentuk DataFrame.  
2. Filter data digunakan untuk mengambil subset tertentu dari data berdasarkan kriteria atau kondisi tertentu.

1. **Kesimpulan**

**Dalam pembahasan ini, kita telah melihat berbagai aspek terkait analisis data menggunakan bahasa pemrograman Python dan pustaka pandas.**

1. **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 | … Menit | menarik |
| **2.** | Latihan Kedua | … Menit | menarik |

Keterangan:

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