MODUL 3 DATASET PREPARATION MENGGUNAKAN PANDAS

TUJUAN PEMBELAJARAN

1. Mahasiswa mampu konsep dasar Pandas, seperti DataFrame dan Series, serta cara membuat, mengakses, dan memanipulasi data menggunakan struktur data.
2. Mahasiswa mampu melakukan analisis data sederhana, seperti pengurutan, penyaringan, dan pengelompokan data menggunakan Pandas.

TINJUAN PUSTAKA

Pandas adalah sebuah library di Python yang berlisensi BSD dan open source yang menyediakan struktur data dan analisis data yang mudah digunakan Pandas biasa digunakan untuk membuat tabel, mengubah dimensi data, mengecek data, dan lain sebagainya. Struktur data dasar pada Pandas dinamakan DataFrame, yang memudahkan kita untuk membaca sebuah file dengan banyak jenis format seperti file .txt, .csv, dan .tsv. Fitur ini akan menjadikannya table dan juga dapat mengolah suatu data dengan menggunakan operasi seperti join, distinct, group by, agregasi, dan teknik lainnya yang terdapat pada SQL.

Library Pandas memiliki dua tipe struktur data untuk versi terbaru yaitu Series dan Data Frame serta satu deprecated struktur data yaitu Panel (deprecated). Series diibaratkan sebagai array satu dimensi sama halnya dengan numpy array, hanya bedanya mempunyai index dan kita dapat mengontrol index dari setiap elemen tersebut. Sedangkan data frame merupakan array dua dimensi dengan baris dan kolom. Struktur data ini merupakan cara paling standar untuk menyimpan data dalam bentuk tabel/data tabular. Dapat disimpulkan, bahwa Pandas merupakan library analisis data yang diperlukan untuk membersihkan data mentah ke dalam sebuah bentuk yang bisa untuk diolah.

PERSIAPAN

ALAT DAN BAHAN:

- Kebutuhan Peralatan
  1. Komputer PC Laptop
  2. Visual Studio Code
  3. Python dan Jupyter Notebook
- Kebutuhan bahan Praktikum
  1. Dataset

PROSEDUR KERJA

Percobaan 1

Jalankan kode berikut

```
1  import pandas
2
3  a = [1, 7, 2, 3, 5]
4  myvar = pd.Series(a)
5  print("Indeks ke-0 : ", myvar[0])
6  print(myvar)
7  myvar.values
8
9  myvar = pd.Series(a, index = ["a", "b", "c", "d", "e"])
10 print("Indeks ke-a : ", myvar["a"])
11 print("Indeks ke-c : ", myvar["c"])
12 print(myvar)
13 myvar.index
14
15 myvar.loc["d"]
16 myvar.iloc[3]
17
18 dict_populasi = {'Jakarta':750,
19                  'Bogor':400,
20                  'Depok':350,
21                  'Tanggerang':270,
22                  'Bekasi':670}
23 dict_populasi
24
25 populasi = pd.Series(dict_populasi)
26 populasi.loc['Bogor']
27 populasi.iloc[1]
28 populasi
29
30 data = {
31   "calories": [420, 380, 390, 450, 360],
32   "duration": [50, 40, 45, 60, 35]
33 }
34
35 myvar = pd.DataFrame(data)
36 print(myvar)
37 print(myvar.loc[0])
38 print(myvar.loc[[1, 3]])
39
40 df = pd.DataFrame(data, index = ["day1", "day2", "day3", "day4",
41 "day5"])
42 print(df)
43 print(df.loc["day2"])
44
45 calories = {"day1": 420, "day2": 380, "day3": 390}
46 myvar = pd.Series(calories)
47 print(myvar)
48
49 users = {
50   'fist_name': ['John', 'Andrew', 'Maria', 'Helen'],
51   'last_name': ['Brown', 'Purple', 'White', 'Blue'],
52   'age': [25, 48, 76, 19]
53 }
54
55 df = pd.DataFrame(users)
56 df
57
58 fruits = ["apple", "pineaple", "orange", "grapes", "banana"]
59 series = pd.Series(fruits)
60 df = pd.DataFrame(series)
61 df
```

1. Jelaskan fungsi dari library pandas

2. Jelaskan fungsi perintah pada baris 4. Jelaskan pengertian series.
3. Jelaskan perbedaan index pada baris 5-6 dan 9-12
4. Buatlah series berikut 18, 17, 8, 10, 18, 8 menggunakan pengindeksan biasa. Cetaklah data ke 2, 4, dan 5
5. Buatlah series berikut 18, 17, 8, 10, 18, 8 menggunakan pengindeksan bil1, bil2, dst. Cetaklah data ke bil2, bil4, dan bil5
6. Jelaskan pengertian dataframe
7. Jelaskan perintah baris 18 - 38
8. Jelaskan perbedaan index pada baris 35-38 dengan baris 40 - 43
9. Buatlah data frame dari 2 series berikut nilai 1 : 18, 17, 8, 10, 18, 8 dan nilai 2 : 16, 16, 13, 12, 1, 4. Untuk setiap baris data frame gunakan pengindeksan berikut urutan 1, urutan 2, dst.
10. Jelaskan fungsi dari baris 45 – 47
11. Jelaskan perbedaan data frame yang dibuat di baris 45 – 47 dengan 49 – 56
12. Jelaskan perintah baris 58 - 61
13. Buatlah data frame dari dictionary berikut "suhu1": 35, "suhu2": 35, "suhu3": 35, "suhu4": 35. Kemudian cetaklah suhu2 dan suhu4
14. Buatlah dataframe dari data berikut. kemudian cetaklah setiap data dengan indeks ganjil

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Percobaan 2

Jalankan kode berikut

```
1  import pandas as pd
2  df = pd.read_csv('Titanic.csv')
3  print(df)
4  print(df.to_string())
5  print(df.head(10))
6  print(df.head())
7  print(df.tail())
8  print(df.info())
```

1. Jelaskan kode baris 3
2. Jelaskan perbedaan baris 4 – 5
3. Jelaskan perbedaan baris 6 – 7
4. Jelaskan fungsi baris 8
5. Jelaskan fungsi baris 9
6. Ulangi soal tersebut dengan iris.csv

Percobaan 3

Jalankan kode berikut

```
1  import pandas as pd
2
3  df = pd.read_csv('Titanic.csv')
4  df.isnull().sum()
5  df.duplicated().sum()
6
7  df = pd.read_csv('Titanic.csv')
8  df.Cabin.value_counts()
9  df.drop('Cabin', axis=1, inplace = True)
10
11 df = pd.read_csv('Titanic.csv')
12 df.Age.plot(kind='hist')
13 val = df.Age.median()
14 df['Age'] = df.Age.fillna(val)
15
16 df = pd.read_csv('Titanic.csv')
17 df.Embarked[df.Embarked.isnull()]
18 df.Embarked.value_counts()
19 val = df.Embarked.mode().values[0]
20 df['Embarked'] = df.Embarked.fillna(val)
21
22 df = pd.read_csv('Titanic.csv')
23 df['Age'] = df['Age'].astype(int)
24 df.info()
25
26 df = pd.read_csv('Titanic.csv')
27 df.loc[4]
28 df.loc[4, 'Age'] = 36
29
30 df = pd.read_csv('Titanic.csv')
31 for i in df.index:
32   if df.loc[i, "Sex"] == "female":
33     df.loc[i, "Pclass"] = 3
```

1. Jelaskan tujuan dari proses data cleansing
2. Bad data terdiri atas 4 jenis : empty cell, data in wrong format, wrong data, duplicates. Jelaskan masing masing
3. Jelaskan fungsi baris 3 - 5
4. Jelaskan fungsi baris 7 – 9

5. Jelaskan fungsi baris 11-14
6. Jelaskan fungsi baris 16-20
7. Jelaskan fungsi baris 22-24
8. Jelaskan fungsi baris 26-28
9. Jelaskan fungsi baris 30-33
10. Jelaskan cara mengakses elemen tertentu data frame seperti pada baris 28

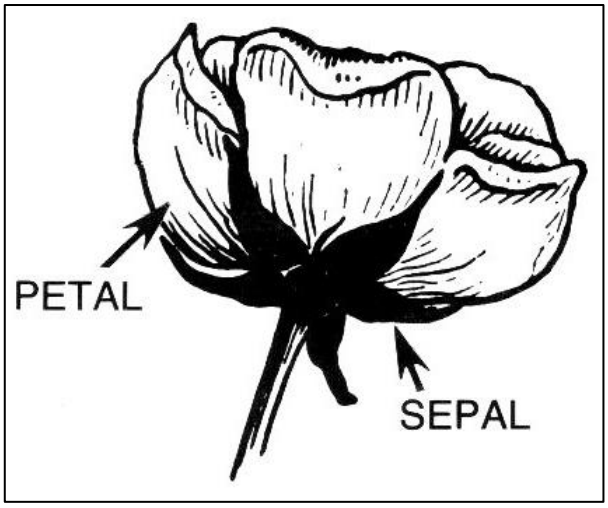
HASIL DAN ANALISIS DATA

Seluruh hasil dari prosedur kerja ditulis dan dianalisis dalam bagian ini. Analisis dilakukan pada data yang dihasilkan dari prosedur kerja baik itu data kualitatif ataupun kuantitatif. Mahasiswa diharapkan mampu melakukan komparasi hasil dengan rumus atau teori pendukung. Mahasiswa dilatih untuk mengolah data mentah menjadi data yang dapat dilaporkan. Berikut adalah contoh hasil dan analisis data.

KESIMPULAN

Mahasiswa diharapkan mampu menarik kesimpulan dari hasil percobaan yang telah dilakukan dengan mengacu pada teori pendukung yang ada. Hal ini untuk melatih mahasiswa untuk membuat gambaran umum tentang hasil percobaan, analisis, pembahasan, dan/atau pengujian hipotesis yang ada. Apabila diperlukan mahasiswa juga dapat memberikan rekomendasi berdasarkan temuan dari hasil percobaan yang dilakukan.

TUGAS MANDIRI



a. **Masukkan iris.csv ke dataframe. Normalisasikan setiap atribut iris ke dataframe menggunakan normalisasi Min-Max sehingga setiap nilai atribut memiliki range antara 0 sampai 1. Dengan Vbaru adalah nilai data disuatu sel setelah normalisasi, Vlama adalah nilai data asli sebelum normalisasi, Vmax adalah nilai max dari kolom data tersebut, dan Vmin adalah nilai min dari kolom data tersebut.**

b. **Carilah centroid dari data yang telah dinormalisasi. Centroid dihitung dari rata-rata seluruh data suatu kolom.**

c. **Masukkan iristesting.csv ke dataframe. Normalisasikan iristesting.csv menggunakan normalisasi Min-Max.**

d. **Prediksilah kelas data di iristesting.csv menggunakan kriteria kedekatan distance antara setiap data di iristesting.csv ke centroid masing-masing class di iris. Gunakan euclidean distance.**

e. **Hitunglah berapa persen klasifikasi yang benar (akurasi) saat memprediksi class**

DAFTAR PUSTAKA

Hans, R. (2021). Pengenalan dan Tutorial Dasar Numpy Array dengan Python. [online]. Available at: https://dqlab.id/pengenalan-dan-tutorial-dasar-numpy-anay-dengan-pyt hon [Accessed 12 Nov. 2023].

Numpy (2009). NumPy. [online]. Available at: https://numpy.org/.

Oliphant, T.E., 2006. Guide to numpy (Vol. 1, p. 85). USA: Trelgol Publishing.

W3Schools (2023). Introduction to NumPy. [online]. Available at: https://www.w3schools.com/pvthon/numpv/numpv intro. asp.