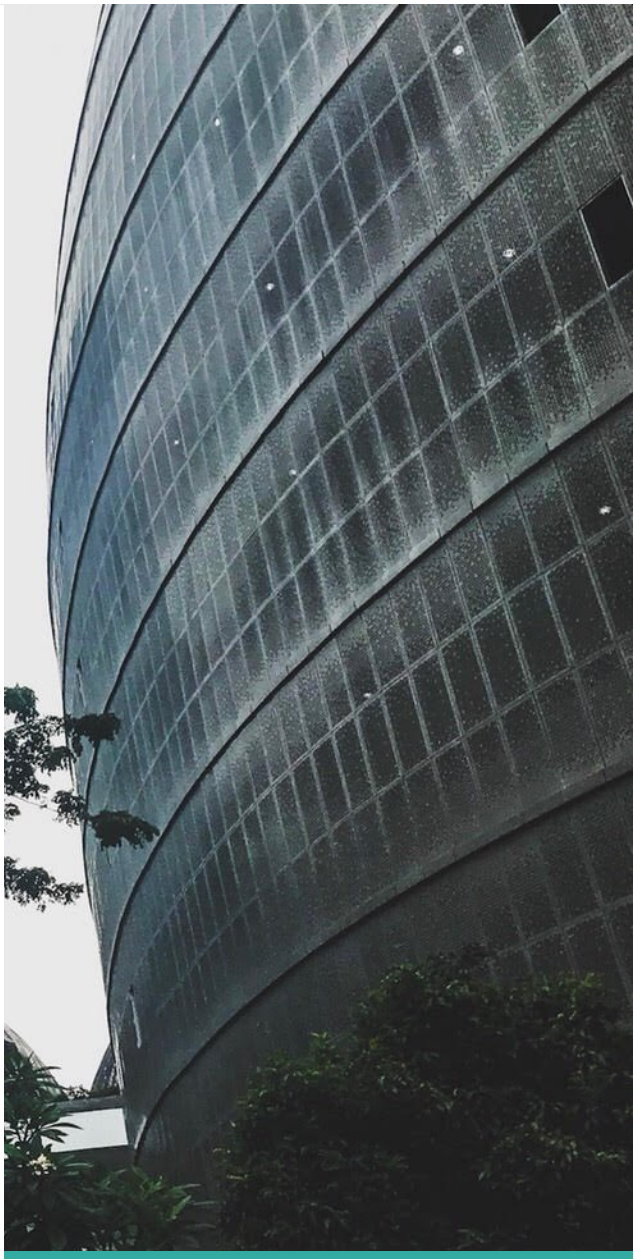
MODUL PRAKTIKUM

IS411 – DATA MODELING
PROGRAM SARJANA S1 SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA



PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA

Gedung B Lantai 5, Kampus UMN
Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten-15811 Indonesia
Telp: +62-21.5422.0808 (ext. 1803), email: ict.lab@umn.ac.id, web: umn.ac.id

MODUL 5
STATISTIC FOR DATA SCIENCE

DESKRIPSI TEMA

1. Variable
2. Random Sampling
3. Representative Sampling
4. Using Scikit-Learn and Matplotlib

CAPAIAN PEMBELAJARAN MINGGUAN (SUB-CAPAIAN PEMBELAJARAN)

Students are able to implement Statistics Methods for Data Science (C_3)

PENUNJANG PRAKTIKUM

1. Anaconda Navigator
2. Jupyter Notebook

(+ Perlengkapan Apd/Alat Pelindung Diri Yang Harus Digunakan, Jika Ada)

LANGKAH-LANGKAH PRAKTIKUM

Import Library

1. Lakukan langkah yang sama seperti minggu sebelumnya untuk meng-import library Numpy dan Pandas.

Scikit-Learn

Scikit-Learn (SKlearn) adalah salah satu library Python yang banyak digunakan untuk memproses data (processing data) machine learning dengan memanfaatkan statistical data modeling. Di dalam SKlearn sudah disediakan algoritma pemodelan data termasuk untuk klasifikasi, regresi, clustering, dan dimensionality. 3 Proses yang melibatkan SKlearn adalah mempersiapkan data, memilih model yang akan digunakan dan melakukan tinjauan ulang pada model algoritma yang dipilih.

Pada praktikum kali ini, kita akan menggunakan dataset Iris.

Dataset Iris merupakan dataset yang berisi 50 sample dari 3 tipe spesies Iris (Iris setosa, Iris virginica, Iris versicolor). Dataset ini biasa digunakan untuk melakukan klasifikasi tipe Iris berdasarkan panjang daun (Petal Length), lebar daun (Petal Width), panjang kelopak (Sepal Length), lebar kelopak (Sepal Width).

2. Import dataset Iris.

```
1 from sklearn.datasets import load_iris
```

3. data.shape()

```
1 iris = load_iris()
2 print(iris.data.shape)
```

function data.shape() adalah untuk melihat ukuran table dari dataset yang sudah di load. Cara membacanya adalah (rows, columns).

Catatlah, berapa jumlah barus dan kolom pada dataset tersebut?

4. Mengubah data ke dalam dataframe

Agar lebih mudah untuk melakukan memproses data,dataset ini juga dapat di load kedalam pandas dataframe dengan cara berikut.

```
1 dataset_iris = load_iris(as_frame = True)
```

```
1 dataset_iris.data.info()
```

5. Melihat deskripsi statistic data

Gunakan **data.descbribe()** untuk melihat deskripsi statistic data.

```
1 dataset_iris.data.describe()
```

6. Mengambil sample data (random sampling)

Gunakan **data.sample()** untuk mengambil sample data.

```
1 dataset_iris.data.sample(n=3, random_state=1)
```

Cara ini memanfaatkan *random sampling* dimana data yang dijadikan sample sebanyak 3 dan kontrol randomnya adalah sebanyak 1. Dengan memanfaatkan library Random, kontrol random juga bisa diatur sehingga data sample selalu mengeluarkan data yang berbeda setiap kali program di-run.

7. Mengambil sample data (representative sampling)

```
1 import random
2
3 dataset_iris.data.sample(n=10, random_state = random.randint(0,5))
```

Cara lain dari melakukan sampling adalah dengan melakukan *representative sampling*. Representative sampling adalah sample yang setiap barisnya adalah unik, artinya jika ada sample dengan nilai yang sama, sample tersebut akan ditimpa dengan ke sample yang lama sehingga jumlah sample akan menjadi lebih sedikit. Pada Pandas cara ini dapat dilakukan dengan menambahkan attribut replace = True ketika mengambil sample.

```
1 dataset_iris.data.sample(n=10, replace=True, random_state=1)
```

Melihat Penyebaran Data

Setelah melihat sample data, berikutnya adalah melihat penyebaran data bunga dari dataset Iris, dapat mengikuti contoh berikut. Penyebaran ini berdasarkan ukuran kelopak bunga (**sepal**).

8. Import Matplotlib

```
from matplotlib import pyplot as plt
```

9. Melihat penyebaran data berdasarkan Sepal

```
plt.scatter(dataset_iris.data['sepal length (cm)'],dataset_iris.data['sepal width (cm)'], c=dataset_iris.target)
```

10. Menambahkan FuncFormatter

    Pada gambar scatter, jenis bunga di bedakan dengan warna, namun belum ada penjelasan mengenai warna bunga tersebut. Gunakan formatter untuk memperlihatkan nama bunga.

    FuncFormatter akan melakukan looping untuk setiap nama pada array target_names dan function colorbar akan menandakan warna dengan urutan kolom dari formatter.

    Jangan lupa untuk menambahkan label agar grafik scatter dapat menampilkan data dengan informasi yang jelas.

```
formatter = plt.FuncFormatter(lambda i, *args: iris.target_names[int(i)])

plt.scatter(dataset_iris.data['sepal length (cm)'],dataset_iris.data['sepal width (cm)'], c=dataset_iris.target)
plt.colorbar(ticks=[0, 1, 2], format=formatter)
plt.xlabel('sepal length (cm)')
plt.ylabel('sepal width (cm)')
```

Pada grafik tersebut, dapat dilihat bahwa pada data ini, kelopak bunga setosa terlihat memiliki kelopak bunga yang sangat berbeda dibandingkan bunga yang lainnya. Sedangkan bunga versicolor dan virginica memiliki kelopak bunga yang hampir mirip.

11. Lakukan langkah 9 dan 10 untuk melihat penyebaran data berdasarkan **Petal**.

12. Membagi data training dan testing dengan Scikit-Learn

    Langkah berikutya adalah melakukan pembagian data menggunakan Scikit-Learn. Pada praktikum sebelumnya sudah dilakukan menggunakan x_train,y_train,x_test, dan y_test, dengan cara berikut.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(dataset_iris.data, dataset_iris.target, test_size=0.33, random_state=42)
```

    Attribute test_size adalah ukuran data yang kemudian akan dibagi (split) oleh sistem. Untuk test size 0.33, jumlah data yang akan di test adalah 1/3 dari keseluruhan data. Secara default sizenya adalah 0.25.

    random_state adalah atribut yang digunakan untuk melakukan pengacakan data yang akan dimasukkan kedalam test atau train. random_state berfungsi untuk melakukan kontrol pada tingkat pengacakan ketika melakukan split() atau fit(). Menggunakan angka seperti pada contoh, akan menghasilkan angka random yang sama, artinya pada percobaan 1, 2 ,3, dan seterusnya, sample data akan sama. Tetapi jika ingin benar-benar berbeda, dapat menggunakan function random dari numpy (secara default jika tidak diisi dengan angka juga akan menggunakan random).

PENGUMPULAN

1. File yang dikumpulkan terdiri dari:
    a. File project (.ipynb)
    b. File PDF berisi screenshot output dan jawaban (jika ada pertanyaan)
2. File di-compress (.zip) dan diberi nama **KODEMATAKULIAH_KELAS_NIM_NAMA_WEEK KE-XX.zip** (contoh: IS5411_A_13110310017_Monika Evelin Johan_Week-01.zip).

REFERENSI

Deitel, P., & Deitel, H. (2020). Intro to Python for Computer Science and Data Science. Pearson Education.