

#### Hands-On

Hands-On ini digunakan pada kegiatan Microcredential Associate Data Scientist 2021

#### Pertemuan 5

Pertemuan 5 (lima) pada Microcredential Associate Data Scientist 2021 menyampaikan materi mengenai Mengumpulkan Data, Menelaah Data dengan metode Statistik

#### # Pengambilan Data dari API Kaggle

Salah satu portal yang menyediakan dataset untuk project Data Science adalah Kaggle (https://www.kaggle.com/). Pada latihan ini, silakan peserta mengunduh dataset mengenai bunga Iris dengan menggunakan kata kunci: "iris species" yang disediakan oleh UCI Machine Learning (UCIML)

#### 1. Install Modul kaggle:

```
# Install modul kaggle secara inline (di dalam notebook)
!pip install kaggle
```

```
Requirement already satisfied: kaggle in c:\users\user\anaconda3\lib\
site-packages (1.5.12)
Requirement already satisfied: certifi in c:\users\user\anaconda3\lib\
site-packages (from kaggle) (2020.12.5)
Requirement already satisfied: urllib3 in c:\users\user\anaconda3\lib\
site-packages (from kaggle) (1.26.4)
Requirement already satisfied: tqdm in c:\users\user\anaconda3\lib\
site-packages (from kaggle) (4.59.0)
Requirement already satisfied: python-slugify in c:\users\user\
anaconda3\lib\site-packages (from kaggle) (5.0.2)
Requirement already satisfied: six>=1.10 in c:\users\user\anaconda3\
lib\site-packages (from kaggle) (1.15.0)
Requirement already satisfied: requests in c:\users\user\anaconda3\
lib\site-packages (from kaggle) (2.25.1)
Requirement already satisfied: python-dateutil in c:\users\user\
anaconda3\lib\site-packages (from kaggle) (2.8.1)
Requirement already satisfied: text-unidecode>=1.3 in c:\users\user\
```

anaconda3\lib\site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: idna<3,>=2.5 in c:\users\user\
anaconda3\lib\site-packages (from requests->kaggle) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\user\
anaconda3\lib\site-packages (from requests->kaggle) (4.0.0)

# Install modul kaggle secara eksternal melalui anaconda prompt:

```
usage: kaggle datasets [-h]
                      {list, files, download, create, version, init, metadata, status} ...
optional arguments:
 -h, --help
                       show this help message and exit
commands:
  {list,files,download,create,version,init,metadata, status}
   list
                    List available datasets
   files
                     List dataset files
   download
                     Download dataset files
   create
                      Create a new dataset
   version
                     Create a new dataset version
   init
                     Initialize metadata file for dataset creation
                     Download metadata about a dataset
   metadata
                      Get the creation status for a dataset
   status
```

#### 2. Create Token API kaggle:

```
usage: kaggle datasets [-h]
                      {list, files, download, create, version, init, metadata, status} ...
optional arguments:
 -h, --help
                      show this help message and exit
 {list,files,download,create,version,init,metadata, status}
   list List available datasets
   files
                     List dataset files
   download
                    Download dataset files
   create
                      Create a new dataset
   version
                     Create a new dataset version
                     Initialize metadata file for dataset creation
   init
   metadata
                     Download metadata about a dataset
                      Get the creation status for a dataset
   status
```

- 1. Login Kaggle.com
- 2. Kemudian pada menu Profile --> Account
- 3. Klik Create New Api Token
- 4. Maka akan terdownload file kaggle.json

```
usage: kaggle datasets [-h]
                      {list, files, download, create, version, init, metadata, status} ...
optional arguments:
  -h, --help
                       show this help message and exit
commands:
  {list,files,download,create,version,init,metadata, status}
                     List available datasets
   files
                       List dataset files
                     Download dataset files
   download
   create
                      Create a new dataset
                       Create a new dataset version
   version
                       Initialize metadata file for dataset creation
   metadata
                       Download metadata about a dataset
    status
                       Get the creation status for a dataset
```

Kaggle API secara default mengasumsikan bahwa file kaggle.json tersebut berada di dalam folder:

- ~/.kaggle/ (Linux/Mac)
- C:\Users<Windows-username>.kaggle (Windows)

Jika folder tersebut belum ada:

- 1. Buat folder di direktori C:\Users<Windows-username>.kaggle
- 2. letakkan file kaggle.json kedalam folder tersebut

#### 3. Download Dataset dari Kaggle:

```
usage: kaggle datasets [-h]
                       {list, files, download, create, version, init, metadata, status} ...
optional arguments:
  -h, --help
                       show this help message and exit
commands:
  {list,files,download,create,version,init,metadata, status}
                      List available datasets
   list
    files
                       List dataset files
   download
                       Download dataset files
   create
                      Create a new dataset
   version
                      Create a new dataset version
    init
                       Initialize metadata file for dataset creation
    metadata
                       Download metadata about a dataset
                       Get the creation status for a dataset
    status
```

Dokumentasi Kaggle Commands selengkapnya Disini

```
# Mencari dataset yang tersedia di kaggle --> pilih data provider dari
UCIML
```

!kaggle datasets list -s Iris

	downloadCount voteCount usabi 	tle lityRating
uciml/iris 4KB 2016-09-27 07:38:05 arshid/iris-flower-dataset Dataset	226538 2680 0.7941 Ir 1010B 2018-03-22 15:18	is Flower
40617 371 0.823529 vikrishnan/iris-dataset 999B 2017-08-03 16:00:44 therohk/ireland-historical Waxy-Wany News	Ir 2931 26 0.764	rish Times -
2984 157 1.0 chuckyin/iris-datasets 1KB 2017-03-10 09:35:43 rtatman/iris-dataset-json-(JSON Version)	1772 14 0.7352	is Dataset
5637 43 0.75 parulpandey/palmer-archipe Archipelago (Antarctica) p 10053 114 0.970588	ago-antarctica-penguin-data Pa nguin data 11KB 2020-06-09	lmer 10:14:54
conorrot/irish-weather-hou (hourly data) 1864 40 0.8235294	ly-data	rish Weather .8
saurabh00007/iriscsv 1KB 2017-11-09 07:34:35 jillanisofttech/iris-datas uci	17151 57 0.4117	is dataset
37 12 1.0 fleanend/birds-songs-numer. Numeric Dataset		rds' Songs
706 25 0.9411765 kamrankausar/iris-data 1KB 2017-11-30 10:26:01 jeffheaton/iris-computer-v	1118 13 0.6470	ris_data 5884 ris Computer
Vision 306 9 0.875 styven/iris-dataset	5MB 2020-11-24 21:23:2	•
1KB 2017-11-04 14:10:12 arslanali4343/iris-species 2KB 2020-07-02 06:09:09	795 8 0.2941	.1766 ris Species
olgabelitskaya/flower-colo Images		ower Color
8363 161 0.75 naureenmohammad/mmu-iris-d dataset 645 19 0.5625	taset MM 30MB 2020-07-25 18	U iris ::38:33
rutujavaidya/iris-dataset	Ir	ris Dataset

```
      1KB
      2021-07-25
      17:37:14
      36
      6
      0.4117647

      shantanuss/iris-flower-dataset
      IRIS
      flower

      dataset
      1KB
      2020-01-18
      19:43:18

      197
      3
      0.9411765

      ashishs0ni/iris-dataset
      Iris_dataset

      1KB
      2018-08-05
      14:26:19
      601
      7
      0.64705884
```

# Download dan ekstrak dataset, secara default akan berada dalam satu direktori dengan notebook ini

!kaggle datasets download uciml/iris --unzip

Downloading iris.zip to C:\Users\dwiah\Pertemuan 5

```
0%| | 0.00/3.60k [00:00<?, ?B/s]
100%|######## | 3.60k/3.60k [00:00<00:00, 3.83MB/s]
```

Atau bisa juga menggunakan link dari kaggle

## Latihan (1)

Silahkan Download sebuah dataset menggunakan API Kaggle

```
#Latihan (1)
```

!kaggle datasets download jeffheaton/iris-computer-vision --unzip

Downloading iris-computer-vision.zip to C:\Users\dwiah\Pertemuan 5

```
0% | | 0.00/5.33M [00:00<?, ?B/s]

19% | #8 | 1.00M/5.33M [00:00<00:01, 2.80MB/s]

38% | ###7 | 2.00M/5.33M [00:01<00:02, 1.67MB/s]

56% | #####6 | 3.00M/5.33M [00:02<00:01, 1.34MB/s]

75% | #######5 | 4.00M/5.33M [00:03<00:01, 1.30MB/s]

94% | ######### | 5.00M/5.33M [00:04<00:00, 1.10MB/s]

100% | ######### | 5.33M/5.33M [00:05<00:00, 1.02MB/s]

100% | ######### | 5.33M/5.33M [00:05<00:00, 1.06MB/s]
```

#### PENGGUNAAN LIBRARY PANDAS dan NUMPY

Pada materi ini, peserta sudah mendapatkan pemahaman mengenai data dan dataset. Penggunaan library pada Python memberikan kemudahan dalam proses data understanding. Beberapa library yang digunakan adalah library Pandas dan Numpy.

# Latihan (2)

Lakukan import Library Pandas dan Library Numpy

```
#Latihan(2)
#Import Library Pandas
import pandas as pd
#Import Library Numpy
```

import numpy as np

#### **DATAFRAME**

DataFrame adalah struktur data 2 dimensi yang berbentuk tabular (mempunyai baris dan kolom). Hampir semua data tidak hanya memiliki 1 kolom tetapi lebih dari 1 kolom, sehingga lebih cocok menggunakan pandas DataFrame untuk mengolahnya.

Penggunaan dataframe pada Python dengan menggunakan syntaks: df.

# Latihan (3)

Panggil file (load dataset) dengan format .csv untuk dataset mengenai bunga Iris yang sudah peserta unduh dari Kaggle, dan akan disimpan di dalam dataframe df. Lalu tampilkan 5 baris awal dataset dengan function head()

```
#latihan(3)
#Panggil file (load file bernama Iris.csv) dan simpan dalam dataframe
Lalu tampilkan 5 baris awal dataset dengan function head()
```

```
bungairis ='Iris.csv'
df = pd.read_csv(bungairis)
df.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	
Spe	cie	S				
0	1	5.1	3.5	1.4	0.2	Iris-
set	osa					
1	2	4.9	3.0	1.4	0.2	Iris-
set	osa					
2	_	4.7	3.2	1.3	0.2	Iris-
set	osa					
3	4	4.6	3.1	1.5	0.2	Iris-
set	osa					
4	5	5.0	3.6	1.4	0.2	Iris-
set	osa					

#### **Telaah Data**

Pada telaah data, dapat dilakukan untuk mengetahui:

- tipe data dari setiap kolom
- · deskripsi statistik data

## Latihan (4)

Tampilkan tipe data dari kolom yang ada pada dataset

```
#latihan(4)
#Tampilkan tipe data dari kolom yang ada pada dataset
print(df.dtypes)
Id
                   int64
                 float64
SepalLengthCm
                 float64
SepalWidthCm
PetalLengthCm
                 float64
PetalWidthCm
                 float64
Species
                  object
dtype: object
```

## Latihan (5)

Apakah tipe Data dari kolom berikut ini: (silakan diisi pada cell di bawah ini)

```
#Latihan (5)
#Tipe Data dari kolom yang ada di dataset

#Kolom "Id" memiliki tipe data = <isikan jawaban di sini>
#Kolom "SepalLengthCm" memiliki tipe data = <isikan jawaban di sini>
#Kolom "Species" memiliki tipe data = <isikan jawaban di sini>
```

# Latihan (6)

Hitunglah ukuran (jumlah baris dan kolom) dari dataset. Dengan menggunakan method function

```
#Latihan (6)
#Hitung ukuran (jumlah baris dan kolom) dari dataset

df.shape
(150, 6)
```

#### Latihan (7)

Berapakah jumlah baris, dan jumlah kolom pada dataset? (silakan diisi pada cell di bawah ini)

```
#Latihan (7)
#Jumlah Baris pada dataset adalah = <isikan jawaban di sini>
#Jumlah kolom pada dataset adalah = <isikan jawaban di sini>
```

## Latihan (8)

Tampilkan data yang hanya berisi kolom "Id" dan kolom "Species" dalam bentuk dataframe.

```
#Latihan (8)
#Tampilkan data untuk kolom "Id" dan kolom "Species" dalam bentuk
dataframe
```

```
df[['Id', 'Species']]
      Ιd
                Species
0
      1
            Iris-setosa
1
      2
            Iris-setosa
2
      3
            Iris-setosa
3
      4
            Iris-setosa
4
      5
            Iris-setosa
145 146 Iris-virginica
146
    147 Iris-virginica
    148 Iris-virginica
147
148
    149 Iris-virginica
149
    150 Iris-virginica
```

```
[150 rows x 2 columns]
```

# Latihan (9)

Tampilkan data dengan dataframe, dan data yang ditampilkan adalah data pada baris dengan indeks 0 (nol) sampai dengan indeks 9 (sembilan)

```
#Latihan (9)
#Tampilkan data dengan dataframe, dan data yang ditampilkan adalah baris dengan indeks 0 (nol) sampai dengan indeks 9 (sembilan)

df.loc[:9]
```

I	d	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	
Spec	ies	5				
0		5.1	3.5	1.4	0.2	Iris-
seto	sa					
1 :	2	4.9	3.0	1.4	0.2	Iris-
seto	sa					
2		4.7	3.2	1.3	0.2	Iris-
seto	sa					
3 (		4.6	3.1	1.5	0.2	Iris-
seto	sa					
4	5	5.0	3.6	1.4	0.2	Iris-
seto	sa					
5	6	5.4	3.9	1.7	0.4	Iris-
seto	sa					
6	7	4.6	3.4	1.4	0.3	Iris-
seto	sa					
7	8	5.0	3.4	1.5	0.2	Iris-
seto	sa					
8	9	4.4	2.9	1.4	0.2	Iris-
seto	sa					
9 1	0	4.9	3.1	1.5	0.1	Iris-
seto	sa					

# Latihan (10)

Tampilkan data hanya kolom "Id" dan kolom "Species" dengan dataframe, dan yang ditampilkan adalah data pada baris dengan indeks 11 (sebelas) sampai dengan indeks 15 (limabelas)

## Latihan (11)

Pada DataFrame dapat menampilkan beberapa baris pertama/terakhir dari dataset yang di load. Gunakan Method head() dan tail().

Latihan: Tampilkan data pada 8 (delapan) baris pertama dari dataset, dengan dataframe.

#### #Latihan (11) #Tampilkan data pada 8 (delapan) baris pertama dari dataset, dengan dataframe

#### df.head (8)

Id		SepalWidthCm	PetalLengthCm	PetalWidthCm	
Specie					
0 1	5.1	3.5	1.4	0.2	Iris-
setosa	1				
1 2	4.9	3.0	1.4	0.2	Iris-
setosa	- 1				
2 3	4.7	3.2	1.3	0.2	Iris-
setosa		312	1.5	012	1113
3 4	4.6	3.1	1.5	0.2	Iris-
•	_	3.1	1.3	0.2	1112-
setosa					
4 5	5.0	3.6	1.4	0.2	Iris-
setosa					
5 6	5.4	3.9	1.7	0.4	Iris-
setosa	1				
6 7	4.6	3.4	1.4	0.3	Iris-
setosa		5	<b></b> .	0.5	1.13
7 8	5.0	3.4	1.5	0.2	Iris-
-		3.4	1.5	0.2	T1 T2-
setosa					

# Latihan (12)

Tampilkan data pada 3 (tiga) baris terakhir dari dataset, dengan dataframe.

```
#Latihan (12)
#Tampilkan data pada 3 (tiga) baris terakhir dari dataset, dengan
dataframe
```

## df.tail (3)

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	\
147	148	6.5	3.0	5.2	2.0	
148	149	6.2	3.4	5.4	2.3	
149	150	5.9	3.0	5.1	1.8	

Species

147 Iris-virginica

148 Iris-virginica

149 Iris-virginica

# **Deskripsi Statistik Data**

DataFrame method describe() menampilkan statistik dasar setiap kolom data yang bertipe numerik, mencakup banyaknya data (count), rerata aritmetik (mean), simpangan baku (std), nilai terkecil (min), kuartil pertama (25%), kuartil kedua/median (50%), kuartil ketiga (75%), dan nilai terbesar (max).

## Latihan (13)

Hitung korelasi dari dataset. Dengan menggunakan method function

```
#Latihan (13)
#Hitung korelasi dataset
```

df.corr()

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	\
Id	1.000000	0.716676	-0.397729	0.882747	
SepalLengthCm	0.716676	1.000000	-0.109369	0.871754	
SepalWidthCm	-0.397729	-0.109369	1.000000	-0.420516	
PetalLengthCm	0.882747	0.871754	-0.420516	1.000000	
PetalWidthCm	0.899759	0.817954	-0.356544	0.962757	

	PetalWidthCm
Id	0.899759
SepalLengthCm	0.817954
SepalWidthCm	-0.356544
PetalLengthCm	0.962757
PetalWidthCm	1.000000

# Latihan (14)

Berdasarkan pada perhitungan korelasi di Latihan (11), apakah yang dapat Bapak/Ibu simpulkan sementara? Silakan tuliskan simpulan sementara Bapak/Ibu pada cell di bawah ini.

```
#latihan (14)
#Simpulan Sementara Hasil Korelasi di latihan (13)

#Secara diagonal data memiliki nilai 1 artinya korelasinya sangat kuat (maksimal 1).
#Seperti matriks diagonal/matriks identitas,Korelasi nilai =1 pada kolom Id-Id, kolom SepalLengthCm-SepalLengthCm, dst.
#Terdapat korelasi yang hampir mendekati nilai = 11 yang artinya korelasinya sangat kuat, seperti kolom PetalLengthCM-PetalWidthCM.
#Korelasi hampir mendekati -1 artinya memiliki korelai yang sangat buruk, pada hasil korelasi diatas tampak ada nilai korelasi negatif,
```

# memperlihatkan korelasi lemah seperti pada kolom SepalWidthCm-PetalLengthWidthCm.

# Latihan (15)

Hitung korelasi untuk kolom berikut ini: PetalLengthCm, PetalWidthCm

```
#Latihan (15)
#Hitung korelasi dataset untuk kolom PetalLengthCm, PetalWidthCm
df[['PetalLengthCm', 'PetalWidthCm']].corr()
```

	PetalLengthCm	PetalWidthCm
PetalLengthCm	1.000000	0.962757
PetalWidthCm	0.962757	1.000000

#### Latihan (16)

Method "describe" secara otomatis melakukan komputasi statistik untuk semua continous variable. Secara default "describe" melakukan ignore terhadap variabel bertype objek.

Komputasi statistik yang dilakukan terdiri dari: count, mean, std, min, max, 25%, 75%, max.

Latihan: Gunakan method describe pada dataset yang sudah di load untuk semua continous variabel. (Dataset Iris.csv)

```
#Latihan (16)
# Penggunaan Metode describe untuk komputasi statistik
```

#### df.describe()

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm
PetalWidthCm		•	_
count 150.000000	150.000000	150.000000	150.000000
150.000000			
mean 75.500000	5.843333	3.054000	3.758667
1.198667			
std 43.445368	0.828066	0.433594	1.764420
0.763161			
min 1.000000	4.300000	2.000000	1.000000
0.100000			
25% 38.250000	5.100000	2.800000	1.600000
0.300000			
50% 75.500000	5.800000	3.000000	4.350000
1.300000			
75% 112.750000	6.400000	3.300000	5.100000
1.800000			

max 150.000000 7.900000 4.400000 6.900000 2.500000

# Latihan (17)

Gunakan method describe pada dataset yang sudah di load untuk data bertype objek. (Dataset Iris.csv)

#Latihan (17) #Gunakan method describe pada dataset yang sudah di load untuk data bertype objek

df.describe(include=['object'])

Species count 150 unique 3 top Iris-setosa freq 50

#### Latihan 18

Gunakan method describe pada dataset yang sudah di load untuk semua type data (continous variabel dan type object).

#Latihan (18) #Gunakan method describe pada dataset yang sudah di load untuk semua type data

df.describe(include='all')

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm
PetalWid	thCm \		•	_
count 3	150.000000	150.000000	150.000000	150.000000
150.00000	90			
unique	NaN	NaN	NaN	NaN
NaN				
top	NaN	NaN	NaN	NaN
NaN				
freq	NaN	NaN	NaN	NaN
NaN			2 27 4222	
mean	75.500000	5.843333	3.054000	3.758667
1.198667	42 445260	0.00000	0 422504	1 764420
std	43.445368	0.828066	0.433594	1.764420
0.763161	1 000000	4 200000	2 000000	1 000000
min	1.000000	4.300000	2.000000	1.000000
0.100000	20 250000	Г 100000	2 000000	1 (00000
25%	38.250000	5.100000	2.800000	1.600000

0.30000 50% 1.30000 75% 1.80000 max 2.50000	75.500000 0 112.750000 0 150.000000	5.800000 6.400000 7.900000	3.000000 3.300000 4.400000	4.350000 5.100000 6.900000
count unique top freq mean std min 25% 50% 75% max	Species 150 3 Iris-setosa 50 NaN NaN NaN NaN NaN NaN			

# Latihan (19)

Hitunglah nilai mean dari dataset.

```
#Latihan (19)
#Hitung nilai Mean dari dataset

df.mean()

Id 75.500000
SepalLengthCm 5.843333
SepalWidthCm 3.054000
PetalLengthCm 3.758667
PetalWidthCm 1.198667
dtype: float64
```

# Latihan (20)

Hitung nilai mean dari dataset untuk kolom PetalLengthCm.

```
#Latihan (20)
#Hitung nilai Mean untuk kolom PetalLengthCm

df[['PetalLengthCm']].mean()

PetalLengthCm     3.758667
dtype: float64
```

# Latihan (21)

Carilah nilai minimal dari dataset untuk kolom SepalWidthCm.

```
#Latihan (21)
#Cari nilai minimal untuk kolom SepalWidthCm

df[['SepalWidthCm']].min()
SepalWidthCm 2.0
dtype: float64
```

## **Method Groupby**

Method groupby memungkinkan analisis dilakukan secara per kelompok nilai atribut tertentu.

## Latihan (22)

Hitunglah nilai mean dari dataset untuk kolom SepalLengthCm per Species dengan menggunakan metode groupby.

#### **Method Value Count**

value\_counts() menghasilkan frekuensi setiap nilai unik di dalam kolom, dan yang tertinggi count-nya adalah merupakan modus pada kolom tersebut.

# Latihan (23)

Hitunglah frekuensi pada kolom 'Species' dengan menggunakan metode value\_counts().

```
#Latihan (23)
#Hitung frekuensi pada kolom 'Species' dengan menggunakan metode
value counts()
```

```
df['Species'].value_counts()
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
Name: Species, dtype: int64
```

#### Latihan (24)

Tampilkan perhitungan frekuensi pada kolom 'Species' dengan menggunakan metode value\_counts() dalam bentuk dataframe.

# Latihan (25)

4.7

4.8 1.7

Hitunglah frekuensi pada kolom 'PetalLenghCm' dengan menggunakan metode value\_counts() dan dalam bentuk dataframe.

```
#Latihan (25)
# Hitung frekuensi pada kolom 'PetalLenghCm' dengan menggunakan metode
value counts()
df['PetalLengthCm'].value counts().to frame()
     PetalLengthCm
1.5
                12
1.4
5.1
                 8
4.5
                 8
                 7
1.3
                 7
1.6
                 6
5.6
4.0
                 5
                 5
4.9
```

5

4

4.4 4.2 5.0 4.1 5.5 4.6 6.1 5.7 3.9 5.8 1.2 1.9 6.7 3.5 5.9 6.0 5.4 5.3 3.3 4.3 5.2 6.3 1.1 6.4 3.6 3.7 3.8 6.9 6.1 6.1 6.1 6.1 6.1 6.1 6.1 6.1	4443333333322222222221111111111
--	---------------------------------

#SELESAI, LATIHAN HANYA SAMPAI NOMOR 25