Tugas 2: Praktikum & Praktikum Mandiri 2

Pandu Linggar Kumara - 0110221277,
Link GitHub - https://github.com/PanduLgg/M_Learning.git

[1] Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: pandulinggar1@gmail.com

Abstract. Kegiatan praktikum mencakup pembacaan dataset, perhitungan nilai-nilai statistik dasar (seperti mean, median, modus, variansi, standar deviasi, dan kuartil), serta analisis korelasi antarvariabel. Selain itu, dilakukan pula visualisasi data menggunakan boxplot, histogram, dan scatter plot untuk memahami distribusi serta hubungan antar variabel.

1. Connecting Google Colab & Drive

1.1 Menghubungkan lingkungan Google Colab dengan akun Google Drive

Sel ini berfungsi untuk menghubungkan lingkungan Google Colab dengan akun Google Drive

```
# menghubungkan colab dengan google drive
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

Gambar 1.1. Proses ini hanya perlu dilakukan satu kali per sesi.

1.2 Memanggil Data set dari Gdrive dan Membaca file .CSV menggunakan Pandas

Sel ini menggunakan library Pandas untuk membaca file data, yang diinginkan yaitu 'data/500_Person_Gender_Height_Weight_Index.csv'.

```
# memanggil data set lewat gdrive
path = "/content/gdrive/MyDrive/praktikum_ml/praktikum02/"
```

```
# membaca file csv menggunakan pandas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv(path + 'data/500_Person_Gender_Height_Weight_Index.csv')
df
```

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
...	...	...	...	...
495	Female	150	153	5
496	Female	184	121	4
497	Female	141	136	5
498	Male	150	95	5
499	Male	173	131	5

500 rows × 4 columns

Gambar 1.2. Proses ini hanya perlu dilakukan satu kali per sesi.

1.3 Mencari informasi data yang ada pada file

Sel ini menampilkan informasi yang ada di dalam file dari mulai tipe data nama kolom, dsb.

```
# Mencari info data pada file (tipe datanya, non nul count data, nama kolom)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Gender  500 non-null    object
 1   Height  500 non-null    int64
 2   Weight  500 non-null    int64
 3   Index   500 non-null    int64
dtypes: int64(3), object(1)
memory usage: 15.8+ KB
```

Gambar 1.3. Mencari info data pada file

1.4 Menghitung Ukuran Tendensi Sentral dan Ukuran Penyebaran Data

Sel ini menghitung mean, median, modus, variansi, dan standard deviasi pada file data

```
# Menghitung mean semua kolom numerik
df['Height'].mean()
```

```
np.float64(169.944)
```

```
# Menghitung Median semua kolom numerik
df['Height'].median()
```

```
170.5
```

```
# Mencari Modus bisa lebih dari satu
df['Height'].mode()
```

	Height
0	188

dtype: int64

```
# Menghitung Variansi & Standard Deviasi
df.var(numeric_only=True)
```

	0
Height	268.149162
Weight	1048.633267
Index	1.836168

dtype: float64

```
# Menghitung Standar Deviasi
df.std(numeric_only=True)
```

	0
Height	16.375261
Weight	32.382607
Index	1.355053

dtype: float64

Gambar 1.4. Menghitung Ukuran Tendensi Sentral dan Ukuran Penyebaran Data

1.5 Menghitung Qartil dan Membuat Deskripsi pada Data

Sel ini menghitung q1, q2, dan IQR, lalu membuat deskripsi pada type data Integer, lalu menghitung matriks korelasi dalam semua kolom numerik

```
# Menghitung Kuartil Pertama (Q1)
q1 = df['Height'].quantile(0.25)
print("Q1 : ", q1)

#Hitung Kuartil Ketiga (Q3)
q3 = df['Height'].quantile(0.75)
print("Q3 : ", q3)

#Hitung IQR (interquartile Range)
iqr = q3 - q1
print("IQR : ", iqr)
```

```
Q1 :  156.0
Q3 :  184.0
IQR :  28.0
```

```
# Untuk membuat statistika deskripsi pada type data int
df.describe()
```

	Height	Weight	Index
count	500.000000	500.000000	500.000000
mean	169.944000	106.000000	3.748000
std	16.375261	32.382607	1.355053
min	140.000000	50.000000	0.000000
25%	156.000000	80.000000	3.000000
50%	170.500000	106.000000	4.000000
75%	184.000000	136.000000	5.000000
max	199.000000	160.000000	5.000000

```
# Menghitung matriks korelasi untuk semua kolom numerik
correlation_matrix = df.corr(numeric_only=True)

# Menampilkan Matriks Korelasi
print("Matriks Korelasi:")
print(correlation_matrix)
```

```
Matriks Korelasi:
            Height    Weight     Index
Height  1.000000  0.000446 -0.422223
Weight  0.000446  1.000000  0.804569
Index  -0.422223  0.804569  1.000000
```

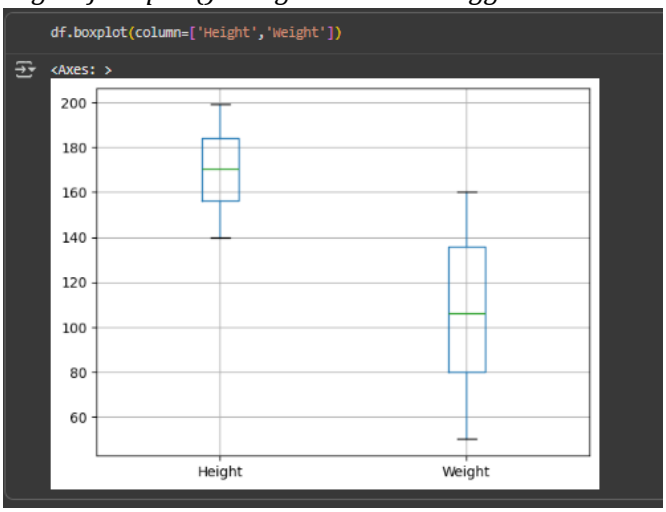
Gambar 1.5. Menghitung Qartil dan Membuat Deskripsi pada Data

2. Visualisasi Data

Memvisualisasikan Data menggunakan matplotlib, dengan metode Box Plot, Histogram, Scatter Plot

2.1 Visualisasi Data menggunakan Box Plot

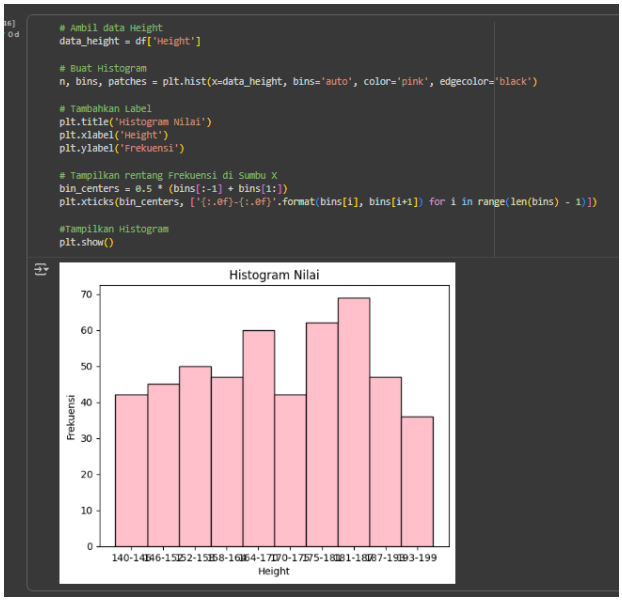
Sel ini menggunakan fungsi df.boxplot() mengambil data Tinggi dan Berat Badan



Gambar 2.1. Menampilkan Box Plot

2.2 Visualisasi Data menggunakan Histogram

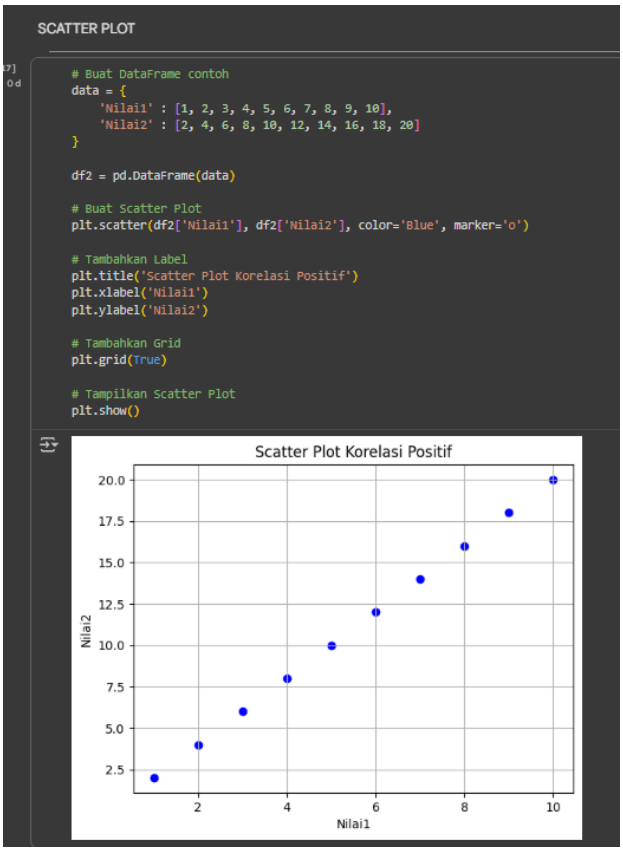
Sel ini membuat histogram untuk kolom 'Height'. Histogram adalah grafik yang menunjukkan distribusi frekuensi data



Gambar 2.2. Menampilkan Histogram

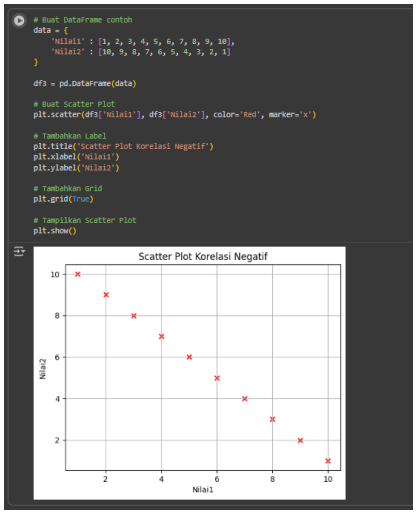
2.3 Visualisasi Data menggunakan Scatter Plot

Sel ini membuat scatter plot atau diagram pencar untuk memvisualisasikan hubungan antara dua variabel numerik. Berisi dua blok kode untuk membuat dua scatter plot terpisah. Kode pertama menggunakan data Nilai1 dan Nilai2 yang memiliki korelasi positif



Gambar 2.3. Menampilkan Scatter Plot Positif

sedangkan kode kedua memiliki korelasi negatif.



Gambar 2.4. Menampilkan Scatter Plot Negatif

Hasil:

1. Plot Korelasi Positif: Titik-titik data membentuk pola yang naik ke atas. Ini menunjukkan bahwa ketika Nilai1 meningkat, Nilai2 juga cenderung meningkat.
2. Plot Korelasi Negatif: Titik-titik data membentuk pola yang menurun. Ini menunjukkan bahwa ketika Nilai1 meningkat, Nilai2 justru cenderung menurun.

3. Praktikum Mandiri

Buat program untuk membagi dataset day.csv menjadi tiga bagian, yaitu:

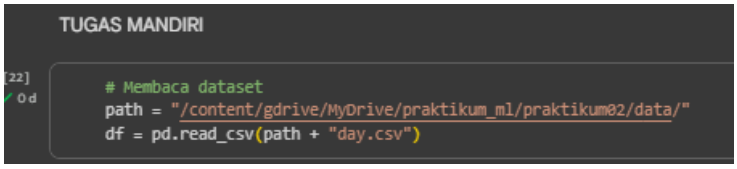
(a) Data Training: 80% dari total dataset

(b) Data Validation: 10% dari data training

(c) Data Testing: 20% dari total dataset

3.1 Membaca Dataset File

Sel ini membaca dataset yang diinginkan yaitu day.csv



Gambar 3.1. Proses ini hanya perlu dilakukan satu kali per sesi.

3.2 Membaca Dataset File

Sel ini membagi data Training 80%, Data Testing 20%, dan Data validation 10%, lalu menampilkan jumlah datanya

Gambar 3.2. Menampilkan Jumlah Data hasil pembagian persentase

3.3 Menampilkan 5 baris perta pada setiap Set

Sel ini menampilkan hasil dari 5 data teratas dari Data Training, Data Validation, dan Data Testing

```
[27]
✓ 0d
# Menampilkan 5 baris pertama dari setiap set
print("\n=== Data Training (5 teratas) ===")
display(train_data.head())

print("\n=== Data Validation (5 teratas) ===")
display(val_data.head())

print("\n=== Data Testing (5 teratas) ===")
display(test_data.head())
```

=== Data Training (5 teratas) ===

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
657	658	2012-10-19	4	1	10	0	5	1	2	0.563333	0.537896	0.815000	0.134954	753	4671	5424
163	164	2011-06-13	2	0	6	0	1	1	1	0.635000	0.601654	0.494583	0.305350	863	4157	5020
305	306	2011-11-02	4	0	11	0	3	1	1	0.377500	0.390133	0.718750	0.082092	370	3816	4186
111	112	2011-04-22	2	0	4	0	5	1	2	0.336667	0.321954	0.729583	0.219521	177	1506	1683
538	539	2012-06-22	3	1	6	0	5	1	1	0.777500	0.724121	0.573750	0.182842	964	4859	5823

=== Data Validation (5 teratas) ===

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
325	326	2011-11-22	4	0	11	0	2	1	3	0.416667	0.421696	0.962500	0.118792	69	1538	1607
410	411	2012-02-15	1	1	2	0	3	1	1	0.348333	0.351629	0.531250	0.181600	141	4028	4169
92	93	2011-04-03	2	0	4	0	0	0	1	0.378333	0.378767	0.480000	0.182213	1651	1598	3249
47	48	2011-02-17	1	0	2	0	4	1	1	0.435833	0.428658	0.505000	0.230104	259	2216	2475
508	509	2012-05-23	2	1	5	0	3	1	2	0.621667	0.584612	0.774583	0.102000	766	4494	5260

=== Data Testing (5 teratas) ===

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
703	704	2012-12-04	4	1	12	0	2	1	1	0.475833	0.469054	0.733750	0.174129	551	6055	6606
33	34	2011-02-03	1	0	2	0	4	1	1	0.186957	0.177878	0.437826	0.277752	61	1489	1550
300	301	2011-10-28	4	0	10	0	5	1	2	0.330833	0.318812	0.585833	0.229479	456	3291	3747
456	457	2012-04-01	2	1	4	0	0	0	2	0.425833	0.417287	0.676250	0.172267	2347	3694	6041
633	634	2012-09-25	4	1	9	0	2	1	1	0.550000	0.544179	0.570000	0.236321	845	6693	7538

Gambar 3.3. Menampilkan 5 Baris Data Teratas setiap Set

Referensi:

Munir, S., Seminar, K. B., Sudradjat, Sukoco, H., & Buono, A. (2022). The Use of Random Forest Regression for Estimating Leaf Nitrogen Content of Oil Palm Based on Sentinel 1-A Imagery. *Information*, *14*(1), 10. https://doi.org/10.3390/info14010010

Seminar, K. B., Imantho, H., Sudradjat, Yahya, S., Munir, S., Kaliana, I., Mei Haryadi, F., Noor Baroroh, A., Supriyanto, Handoyo, G. C., Kurnia Wijayanto, A., Ijang Wahyudin, C., Liyantono, Budiman, R., Bakir Pasaman, A., Rusiawan, D., & Sulastri. (2024). PreciPalm: An Intelligent System for Calculating Macronutrient Status and Fertilizer Recommendations for Oil Palm on Mineral Soils Based on a Precision Agriculture Approach. *Scientific World Journal*, *2024*(1). https://doi.org/10.1155/2024/1788726