Aura Indri Aprilia
231011403707
Machine Learning

PERTEMUAN 4

Sebelumnya, saya menggunakan Jupyter Notebook untuk menjalankan kode dan menghasilkan output berupa visualisasi gambar, seperti histogram, scatterplot, dan heatmap, guna mendukung proses analisis data secara eksploratif.

Saya telah mengikuti langkah 1 dengan membuat dataset dalam format CSV sesuai petunjuk.
Dataset tersebut saya ketik di file teks baru menggunakan pemisah koma (,), kemudian saya simpan dengan nama **kelulusan_mahasiswa.csv**.
Isi datasetnya adalah sebagai berikut:

	A	B	C	D	E
1	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus	
2	3.8	3	10	1	
3	2.5	8	5	0	
4	3.4	4	7	1	
5	2.1	12	2	0	
6	3.9	2	12	1	
7	2.8	6	4	0	
8	3.2	5	8	1	
9	2.7	7	3	0	
10	3.6	4	9	1	
11	2.3	9	4	0	
12					

Saya memastikan format CSV sudah benar, yaitu:

- Menggunakan **koma (,)** sebagai pemisah antar kolom.
- Baris pertama merupakan **header** (nama kolom).
- File disimpan dengan ekstensi **.csv** sesuai instruksi.

Setelah membuat file **kelulusan_mahasiswa.csv** pada langkah sebelumnya, saya melanjutkan ke **Langkah 2 (Collection)** untuk membaca dataset menggunakan library **Pandas** di Python. Sebelumnya disini saya akan

- Hitung statistik deskriptif.
- Buat histogram distribusi IPK.
- Visualisasi scatterplot (IPK vs Waktu Belajar).
- Tampilkan heatmap korelasi.

Saya memasukkan kode berikut

```
[1]
import pandas as pd

df = pd.read_csv("kelulusan_mahasiswa.csv")

print(df.info())
print(df.head())
```

Agar menghasilkan output

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   IPK                10 non-null     float64
 1   Jumlah_Absensi     10 non-null     int64
 2   Waktu_Belajar_Jam  10 non-null     int64
 3   Lulus              10 non-null     int64
dtypes: float64(1), int64(3)
memory usage: 448.0 bytes
None
   IPK  Jumlah_Absensi  Waktu_Belajar_Jam  Lulus
0  3.8               3                 10      1
1  2.5               8                  5      0
2  3.4               4                  7      1
3  2.1              12                  2      0
4  3.9               2                 12      1
```

Dari hasil tersebut dapat diketahui bahwa:

- Dataset memiliki 10 entri (baris) dan 4 kolom, yaitu: IPK, Jumlah_Absensi, Waktu_Belajar_Jam, dan Lulus.

- Semua kolom memiliki 10 nilai non-null, artinya tidak terdapat data kosong (missing value) dalam dataset.
- Tipe data terdiri dari float64 untuk kolom IPK, dan int64 untuk tiga kolom lainnya.
- Dataset memiliki ukuran memori sekitar 448 bytes, menunjukkan ukuran data yang relatif kecil.

Selain itu, ditampilkan juga beberapa data awal yang memperlihatkan isi dari dataset, di mana setiap baris merepresentasikan data mahasiswa yang berisi nilai IPK, jumlah absensi, waktu belajar, dan status kelulusan (1 = lulus, 0 = tidak lulus).

Pada tahap ini dilakukan proses pembersihan data (data cleaning) untuk memastikan dataset bebas dari kesalahan dan siap digunakan dalam proses analisis berikutnya.

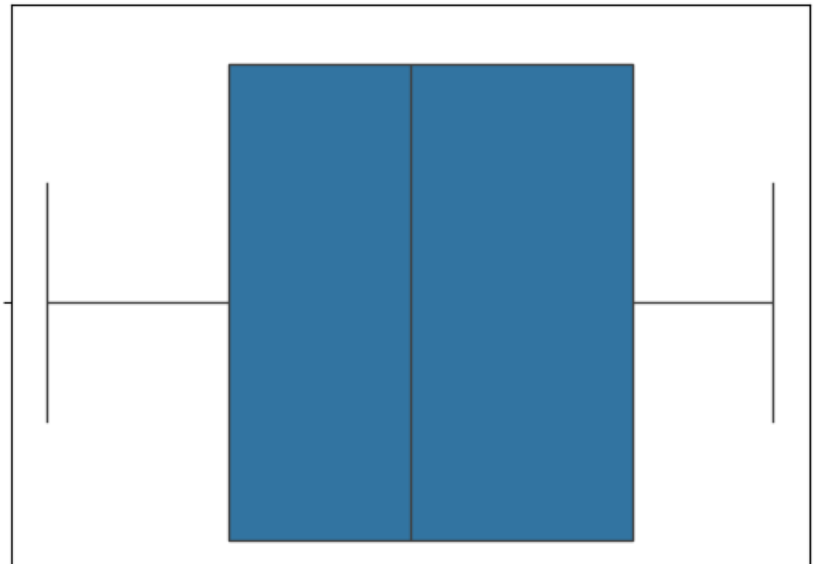
Adapun langkah-langkah yang dilakukan meliputi:

1. Pemeriksaan Missing Value
2. Menghapus Data Duplikat
3. Identifikasi Outlier

Saya menggunakan potongan kode berikut:

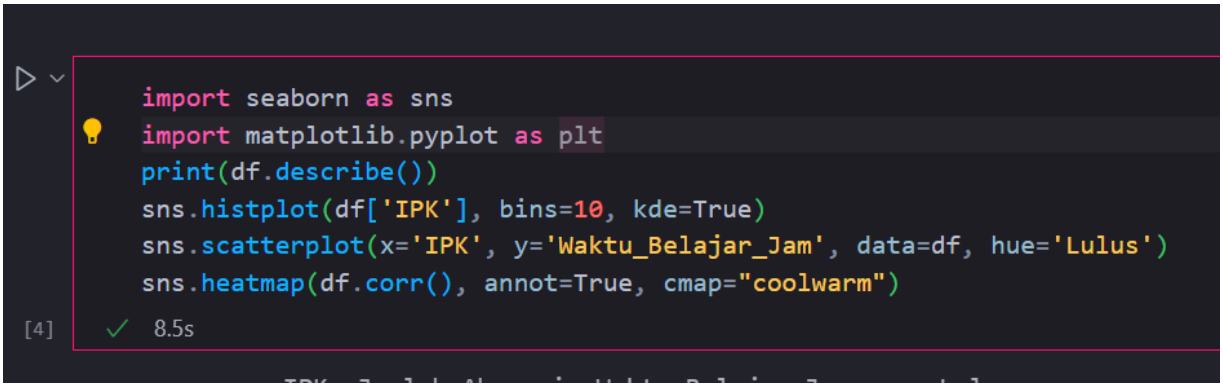
```
[2]
print(df.isnull().sum())
df = df.drop_duplicates()

import seaborn as sns
sns.boxplot(x=df['IPK'])
```

Dan menghasilkann output seperti ini



Setelah saya mengerjakan langkah ke-3, saya akan lanjutkan ke langkah ke-4 yaitu melakukan Exploratory Data Analysis (EDA) dengan menghitung statistik deskriptif, membuat histogram distribusi IPK, visualisasi scatterplot antara IPK dan waktu belajar, serta menampilkan heatmap korelasi antar variabel.

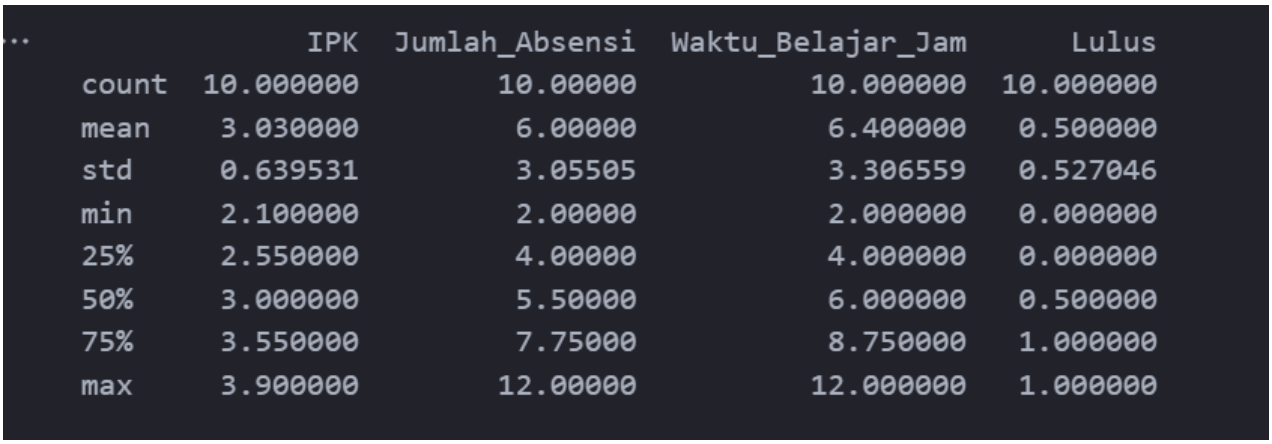
Saya menggunakan kode berikut



```
import seaborn as sns
import matplotlib.pyplot as plt
print(df.describe())
sns.histplot(df['IPK'], bins=10, kde=True)
sns.scatterplot(x='IPK', y='Waktu_Belajar_Jam', data=df, hue='Lulus')
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
```

[4] ✓ 8.5s

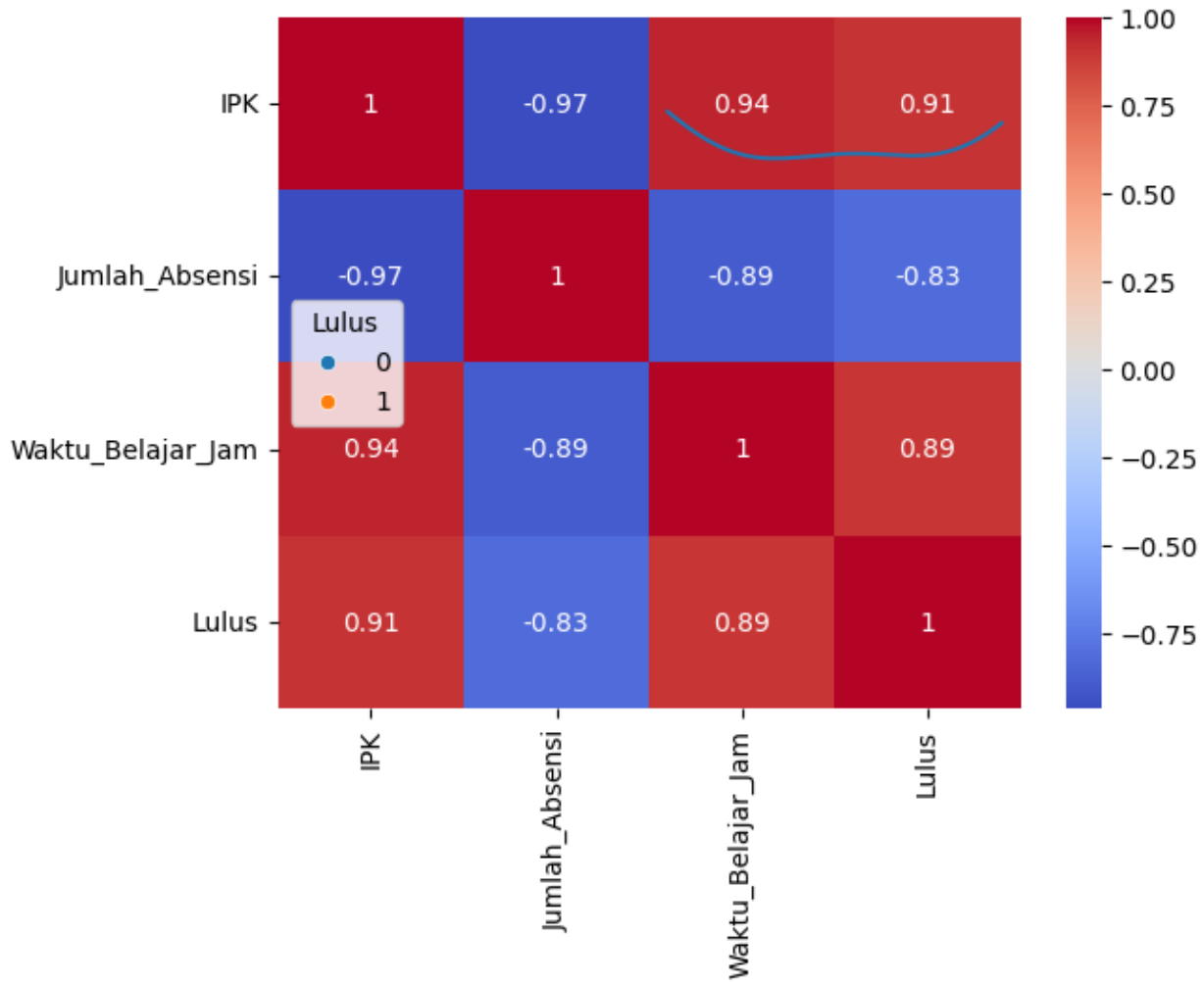
Dan output yang saya dapat dari kode tersebut



	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
count	10.000000	10.00000	10.000000	10.000000
mean	3.030000	6.00000	6.400000	0.500000
std	0.639531	3.05505	3.306559	0.527046
min	2.100000	2.00000	2.000000	0.000000
25%	2.550000	4.00000	4.000000	0.000000
50%	3.000000	5.50000	6.000000	0.500000
75%	3.550000	7.75000	8.750000	1.000000
max	3.900000	12.00000	12.000000	1.000000

Dari hasil yang ditampilkan, diperoleh informasi sebagai berikut:

- IPK memiliki rata-rata 3.03 dengan rentang nilai 2.1–3.9, menunjukkan performa akademik mahasiswa berada pada tingkat baik secara umum.
- Jumlah Absensi rata-rata sebanyak 6 kali, dengan variasi cukup tinggi antara 2 hingga 12 kali, menandakan adanya perbedaan tingkat kehadiran antar mahasiswa.
- Waktu Belajar (Jam) rata-rata 6.4 jam per minggu dengan rentang 2–12 jam, yang menunjukkan perbedaan kebiasaan belajar mahasiswa.
- Status Lulus memiliki nilai rata-rata 0.5, artinya 50% mahasiswa dinyatakan lulus dari total sampel.



Selain itu, kode juga membuat beberapa visualisasi, yaitu:

- Histogram untuk melihat distribusi nilai IPK.
- Scatterplot untuk melihat hubungan antara *IPK* dan *Waktu Belajar* dengan warna berdasarkan status kelulusan.
- Heatmap untuk menampilkan korelasi antar variabel dalam bentuk peta warna (*color map*) dengan skema “coolwarm”.

Setelah melakukan analisis awal pada dataset, langkah selanjutnya adalah melakukan **Feature Engineering** untuk menambahkan fitur-fitur baru yang dapat membantu dalam proses analisis dan pemodelan data.
Feature Engineering bertujuan untuk mengolah data mentah menjadi informasi yang lebih bermakna sehingga dapat meningkatkan kualitas hasil analisis.

Pada tahap ini, saya memasukkan kode berikut untuk membuat beberapa fitur baru pada dataset:

```
import pandas as pd

df = pd.read_csv("ke  (parameter) test_size: Float | None
df['Rasio_Absensi'] = df['Jumlah_Absensi'] / 14
df['IPK_x_Study'] = df['IPK'] * df['Waktu_Belajar_Jam']
df.to_csv("processed_kelulusan.csv", index=False)

print("✅ File processed_kelulusan.csv berhasil dibuat!")
```

[6]

```
✅ File processed_kelulusan.csv berhasil dibuat!
```

Setelah melakukan proses *Feature Engineering* dan mendapatkan dataset yang sudah diproses, langkah selanjutnya adalah melakukan **Splitting Dataset**.
Tahap ini bertujuan untuk membagi data menjadi beberapa bagian agar model dapat dilatih dan diuji secara seimbang, serta menghindari *overfitting*.

Pada tahap ini, dataset dibagi menjadi tiga bagian, yaitu:

- **Data Training (70%)** — digunakan untuk melatih model.
- **Data Validation (15%)** — digunakan untuk mengevaluasi performa model selama pelatihan.
- **Data Testing (15%)** — digunakan untuk menguji akurasi model terhadap data baru.

Proses pembagian dilakukan menggunakan metode **stratified split**, agar proporsi kelas pada variabel target tetap seimbang di setiap subset data.

```
from sklearn.model_selection import train_test_split

X = df.drop('Lulus', axis=1)
y = df['Lulus']

X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=42)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42)

print(X_train.shape, X_val.shape, X_test.shape)
```

[5]

```
(7, 5) (1, 5) (2, 5)
```