LAPORAN PERTEMUAN 4

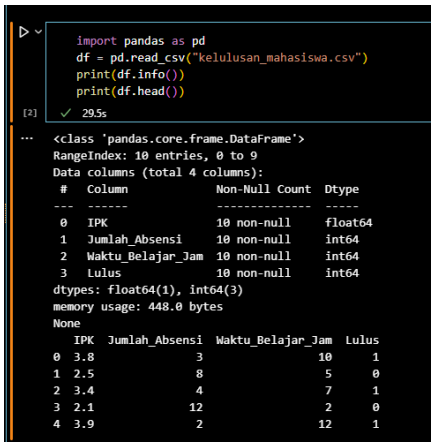
Nama : Muhammad Salman Imamwan Abdillah

Nim : 231011401032

Kelas : 05TPLE017

---

1. Kode ini digunakan untuk membaca data mahasiswa dari file CSV, menampilkan informasi dan isi datanya, serta memastikan datanya siap digunakan untuk analisis lebih lanjut.

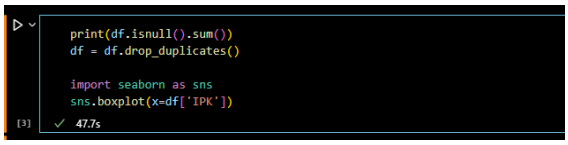


```
import pandas as pd
df = pd.read_csv("kelulusan_mahasiswa.csv")
print(df.info())
print(df.head())
```

```
[2]  ✓ 29.5s
```

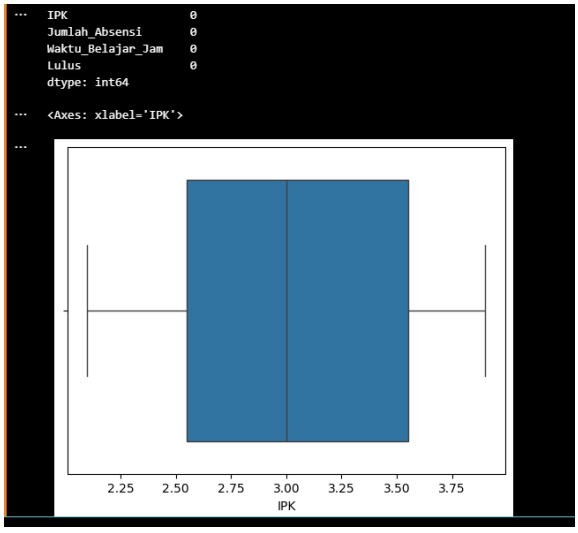
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   IPK               10 non-null     float64
 1   Jumlah_Absensi    10 non-null     int64
 2   Waktu_Belajar_Jam  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
```

2. Kode ini digunakan untuk memastikan data tidak memiliki nilai kosong maupun data ganda, serta untuk melihat distribusi nilai IPK menggunakan grafik boxplot. Hasilnya menunjukkan bahwa data bersih, lengkap, dan nilai IPK mahasiswa terdistribusi normal tanpa adanya outlier.



```
print(df.isnull().sum())
df = df.drop_duplicates()

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

```
[3]  ✓ 47.7s
```



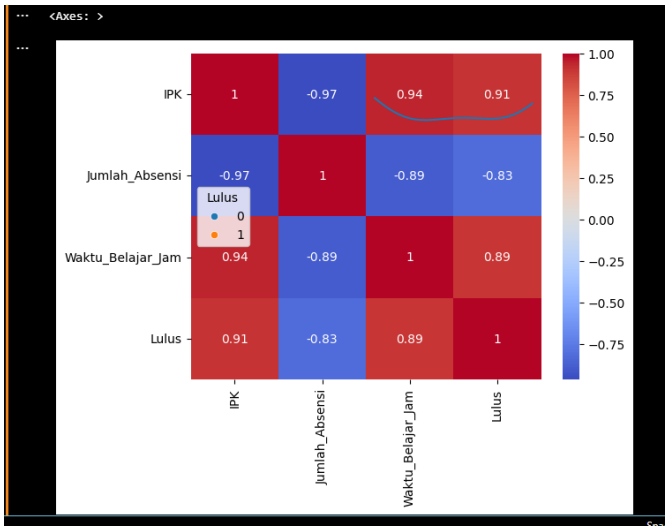
```
IPK               0
Jumlah_Absensi    0
Waktu_Belajar_Jam  0
Lulus             0
dtype: int64

<Axes: xlabel='IPK'>
```

3. Analisis dilakukan untuk melihat hubungan antara IPK, jumlah absensi, waktu belajar, dan kelulusan mahasiswa. Berdasarkan hasil statistik deskriptif dan visualisasi data, diketahui bahwa IPK dan waktu belajar memiliki korelasi positif yang kuat terhadap kelulusan, sedangkan jumlah absensi berkorelasi negatif. Hal ini menunjukkan bahwa mahasiswa dengan IPK tinggi dan waktu belajar lebih banyak cenderung lulus, sementara absensi yang tinggi menurunkan peluang kelulusan.

```
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] ✓ 1.1s

	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



4. Pada tahap ini dilakukan proses pembuatan fitur baru dan pembagian dataset untuk keperluan pelatihan model. Dua fitur tambahan, yaitu *Rasio_Absensi* (hasil pembagian jumlah absensi dengan total 14 pertemuan) dan *IPK_x_Study* (perkalian antara IPK dan waktu belajar), ditambahkan untuk memperkaya informasi dalam data. Selanjutnya, data disimpan ke file processed_kelulusan.csv dan dibagi menjadi tiga bagian menggunakan metode *train_test_split*, yaitu data latih (60%), data validasi (20%), dan data uji (20%). Pembagian dilakukan secara stratified agar proporsi kelas tetap seimbang.

```
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)
```

[6] ✓ 0.0s

```
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=0.4, stratify=y, random_state=42)

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

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

[7] ✓ 4.1s

```
(6, 5) (2, 5) (2, 5)
```