Nama : Sri Mashtufah Anjani

Nim : 231011401951

Mata Kuliah : Machine Learning

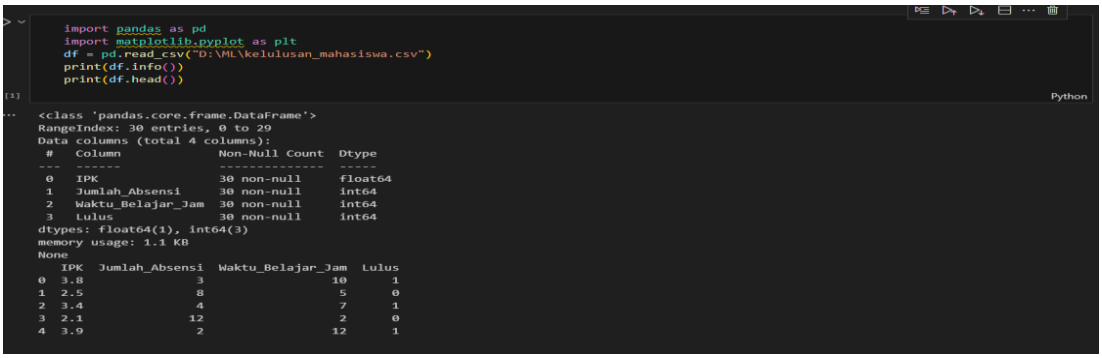
PERTEMUAN 4

Untuk menjalankan kode saya menggunakan Jupyter Netebook hingga menghasilkan output berupa visualisasi gambar

Saya membuat dataset kelulusan_mahasiswa.csv sesuai intruksi Langkah pertama

IPK,Jumlah_Absensi,Waktu_Belajar_Jam,Lulus
3.8,3,10,1
2.5,8,5,0
3.4,4,7,1
2.1,12,2,0
3.9,2,12,1
2.8,6,4,0
3.2,5,8,1
2.7,7,3,0
3.6,4,9,1
2.3,9,4,0

Selanjutnya mengikuti intruksi Langkah ke 2 Collection yaitu memuat dataset yang ada pada file kelulusan_mahasiswa.csv ke dalam DataFrame menggunakan pandas. Setelah dataset dimuat, kita perlu melihat gambaran umum dataset untuk memverifikasi struktur dan informasi yang ada.



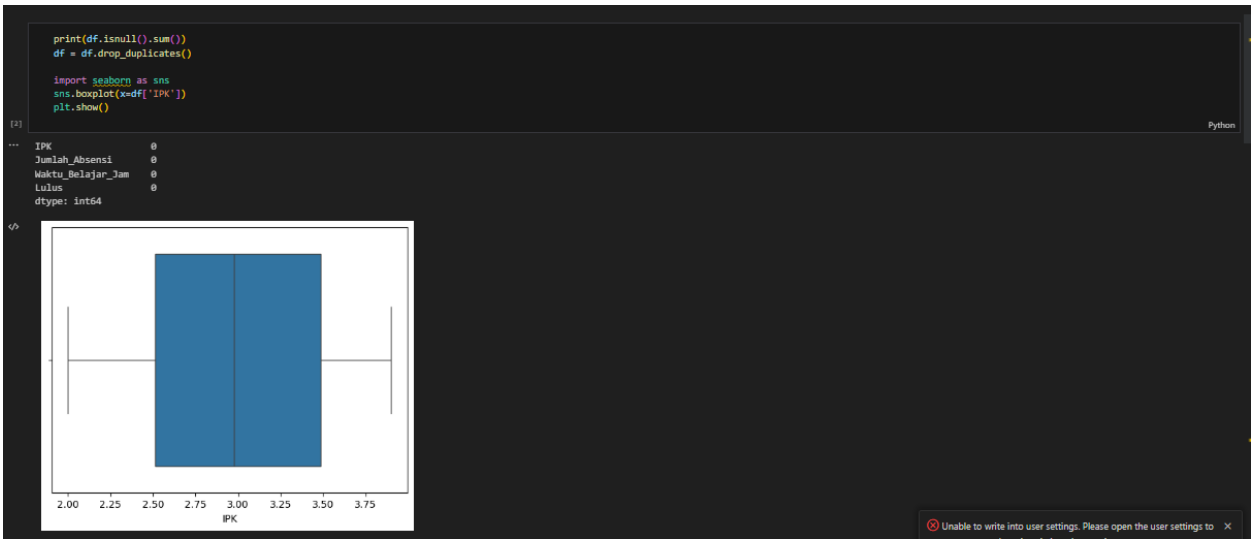
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("D:\ML\kelulusan_mahasiswa.csv")
print(df.info())
print(df.head())
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   IPK               30 non-null     float64
 1   Jumlah_Absensi    30 non-null     int64
 2   Waktu_Belajar_Jam  30 non-null     int64
 3   Lulus             30 non-null     int64
dtypes: float64(1), int64(3)
memory usage: 1.1 KB
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
```

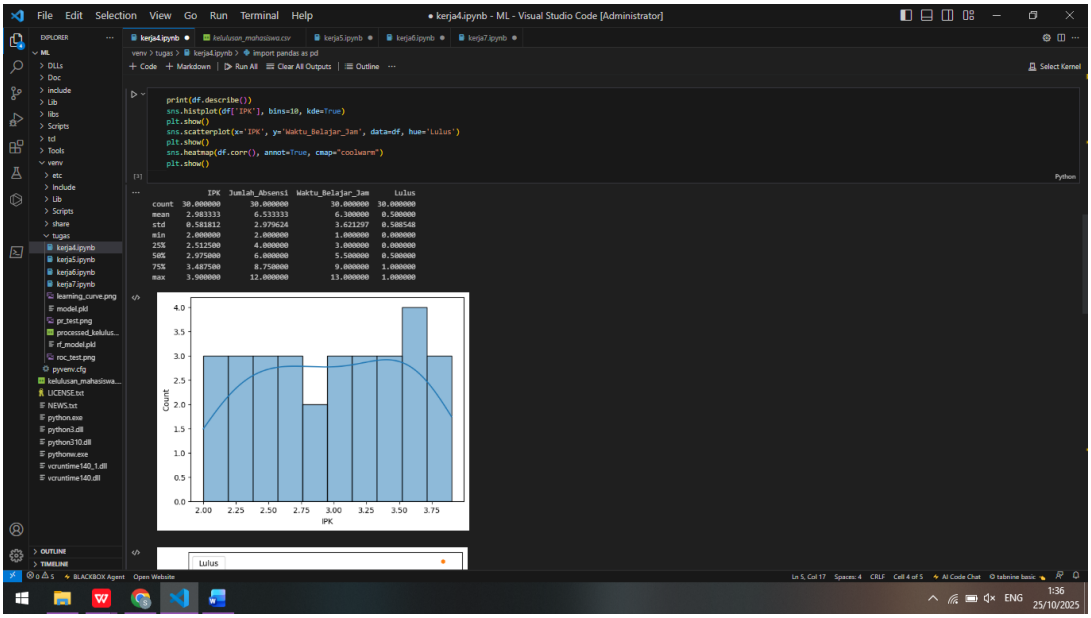
Lanjut Langkah ketiga Cleaning

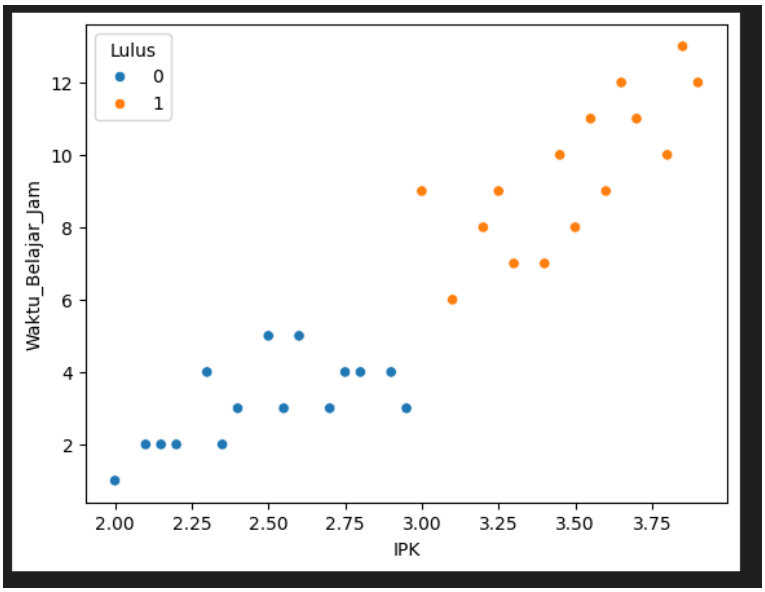
Di dalam dataset, kolom Lulus adalah target yang ingin kita prediksi, sedangkan fitur lainnya adalah variabel yang digunakan untuk memprediksi status kelulusan. Sebelum melakukan analisis lebih lanjut atau pelatihan model, kita perlu memastikan bahwa data tidak memiliki nilai yang hilang atau duplikat. Jika ada duplikat atau nilai yang hilang, kita bisa menghapus atau menggantinya.

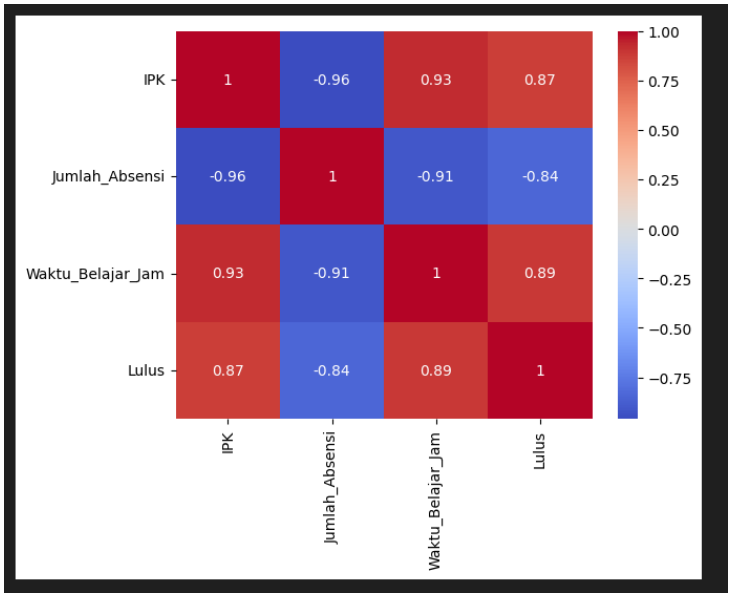


Lanjut Langkah 4 — Exploratory Data Analysis (EDA)

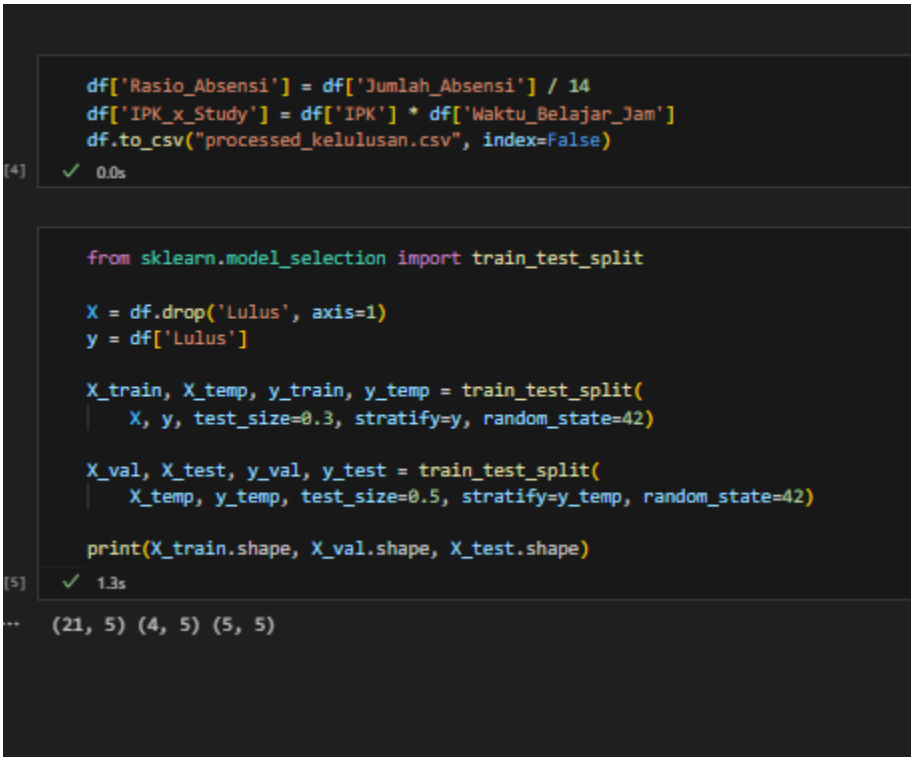
Dengan menghitung statistik deskriptif., membuat histogram distribusi IPK., memvisualisasi scatterplot (IPK vs Waktu Belajar).dengan warna yang menunjukan apakah mahasiswa lulus atau tidak dan menampilkanan heatmap korelasi untk memberi pemahaman yang lebih baik tentang bagaimana variabel-variabel tersebut saling berhubungan..







Langkah 5 — Visualisasi Learning Curve



- **X_train**: Data latih yang berisi 21 sampel dan 5 fitur.
- **X_val**: Data validasi yang berisi 4 sampel dan 5 fitur.
- **X_test**: Data uji yang berisi 5 sampel dan 5 fitur.

Ini menunjukkan bahwa data telah dibagi menjadi 3 subset: 70% untuk pelatihan, 15% untuk validasi, dan 15% untuk pengujian. Pembagian data ini penting untuk memastikan bahwa model diuji pada data yang tidak terlihat selama pelatihan, sehingga dapat menilai performa model dengan lebih akurat