**Laporan Pertemuan 4 – Data Preparation.**

Nama : Khailla Yaman Putri

Nim : 231011402193

Kelas : 05TPLE015

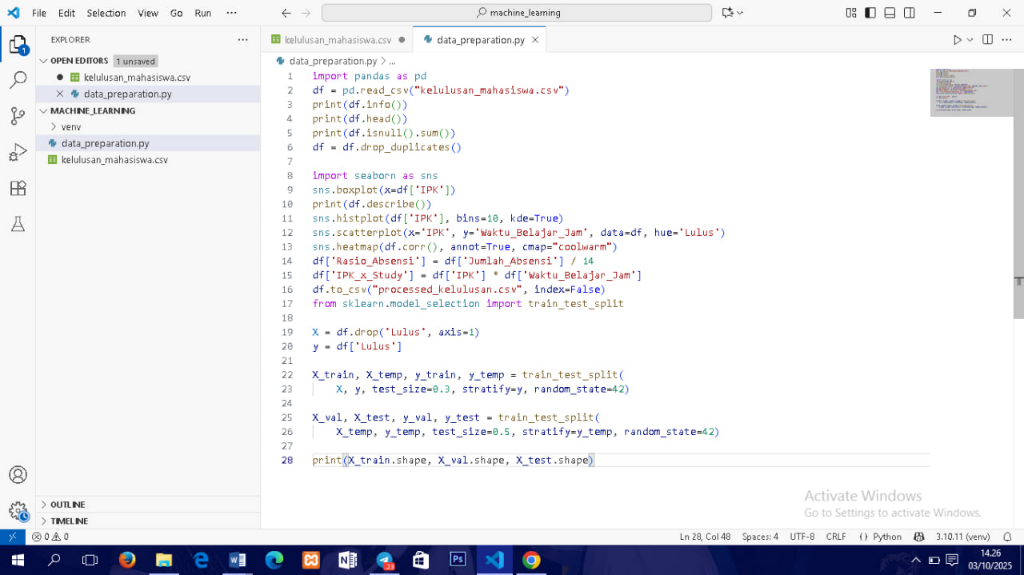
Mata Kuliah : Machine Learning

Dalam penelitian kali ini dilaukan analisis terhadap dataset kelulusan mahasiswa. Dataset berisi informasi IPK, jumlah absensi, waktu belajar per minggu, dan status kelulusan.

Hasil yang diharapkan:

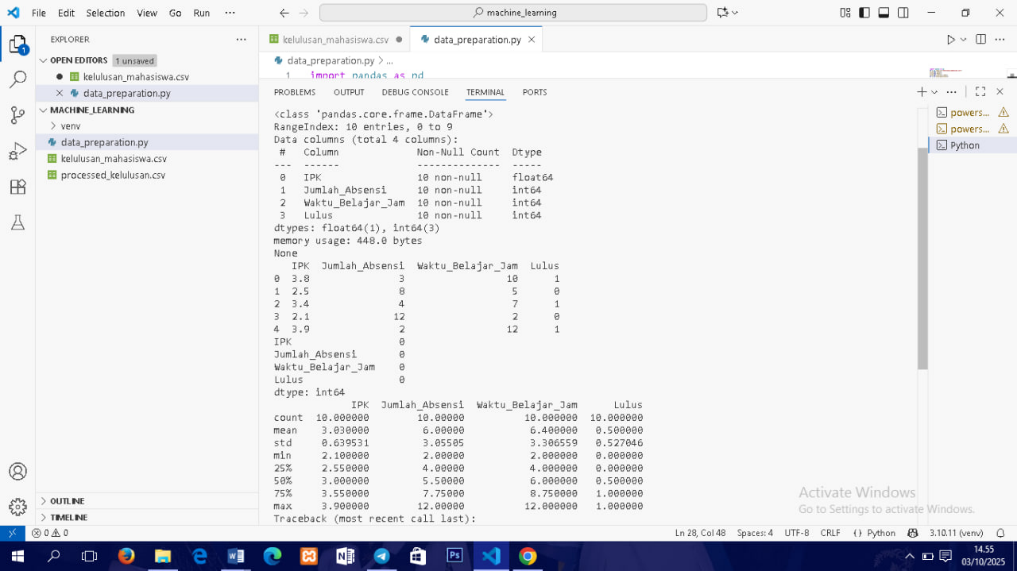
* File processed\_kelulusan.csv dengan data bersih + fitur baru.
* Statistik & visualisasi EDA yang lengkap.
* Dataset terbagi Train, Validation, Test.

Pada tahap awal, saya membuat dataset dalam bentuk csv dan menjalankan kode program yang sesuai pada web yang telah diberikan untuk membaca dan menganalisi dataset.



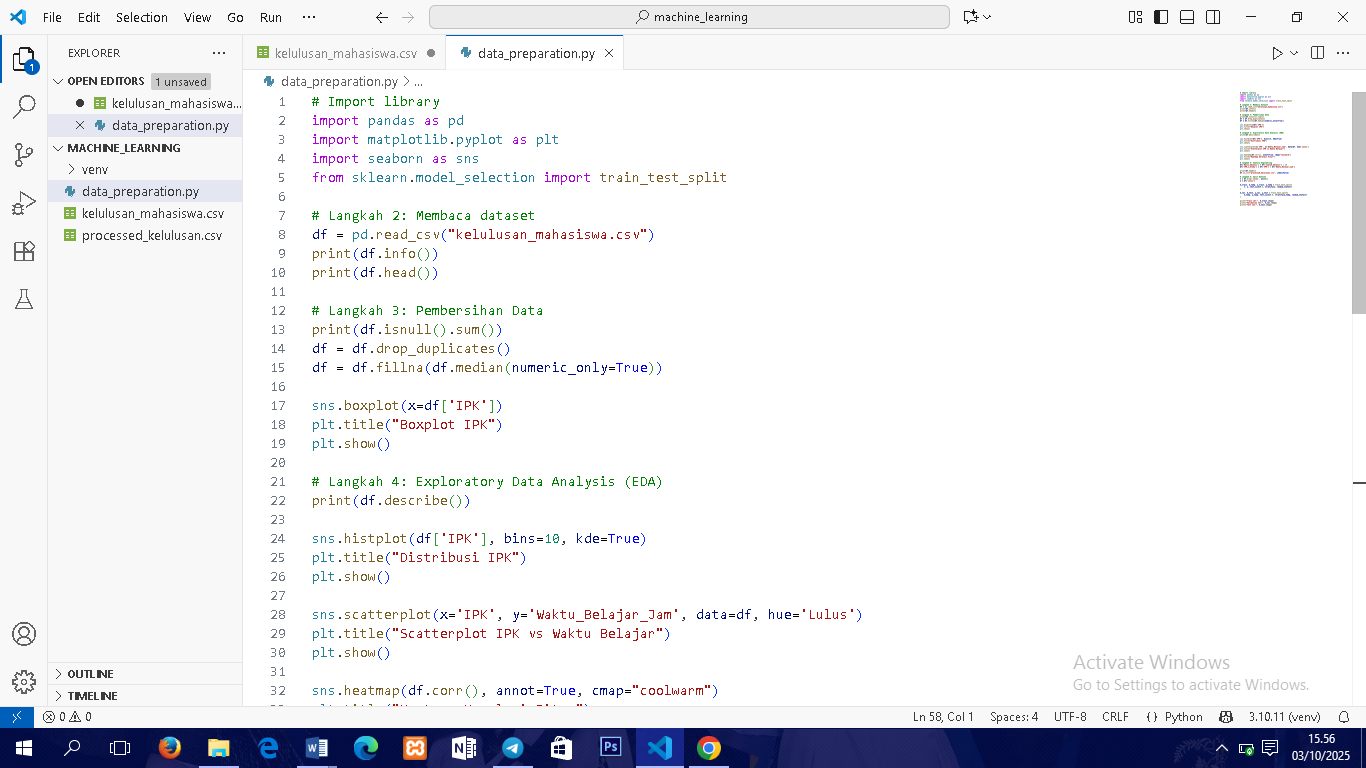
*Kode sebelum dirapihkan.*

Hasil yang muncul hanya sampai pada tampilan informasi dataset (info()), beberapa baris awal data (head()), serta statistik deskriptif (describe()) yakni langkah 2 sampai dengan langkah 4. Visualisasi data seperti histogram, scatterplot, dan heatmap belum ditampilkan, begitu juga dengan tahapan feature engineering dan pembagian dataset. Bagian ini ditampilkan sebagai dokumentasi awal sebelum dilakukan perbaikan dan penyempurnaan kode.



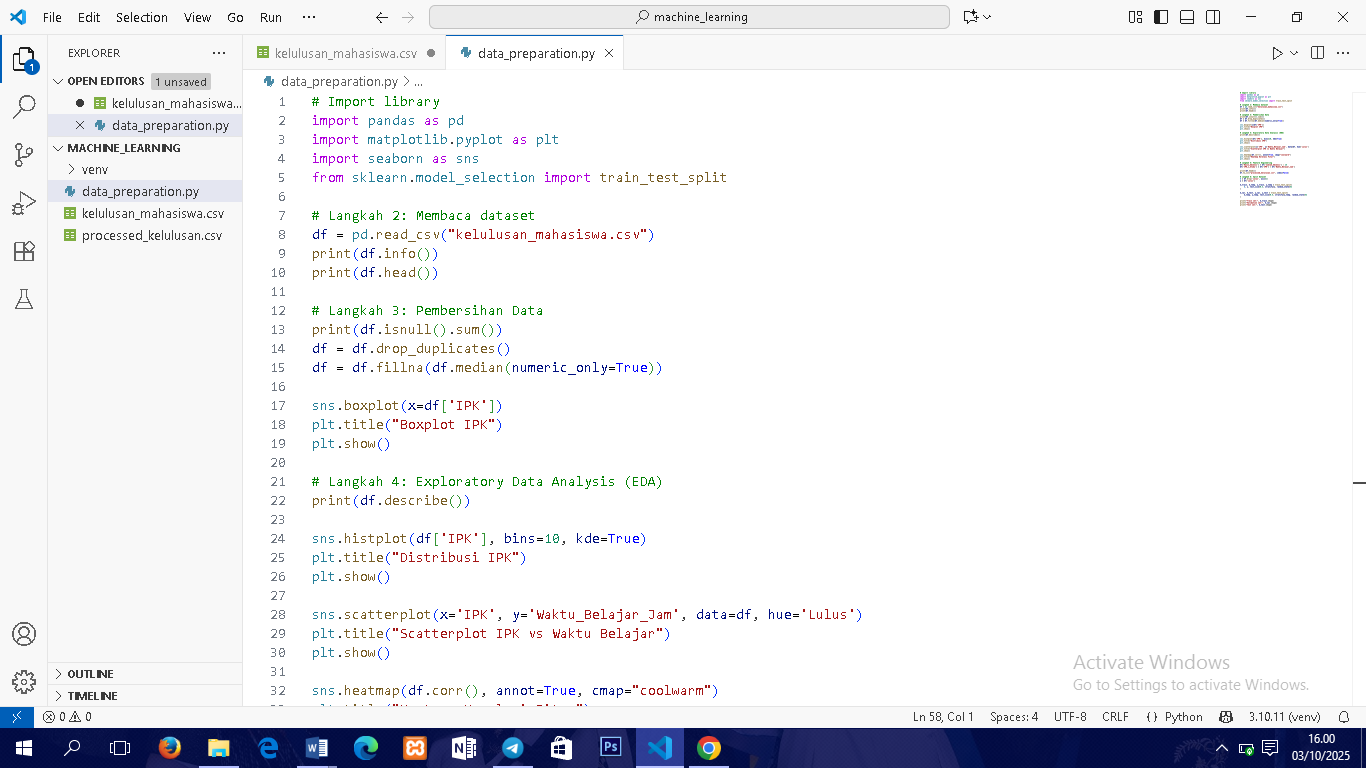
Analisa perlangkah:

**Langkah 1 (Dataset)** → dataset CSV sudah tersedia berisi IPK, Jumlah\_Absensi, Waktu\_Belajar\_Jam, dan Lulus

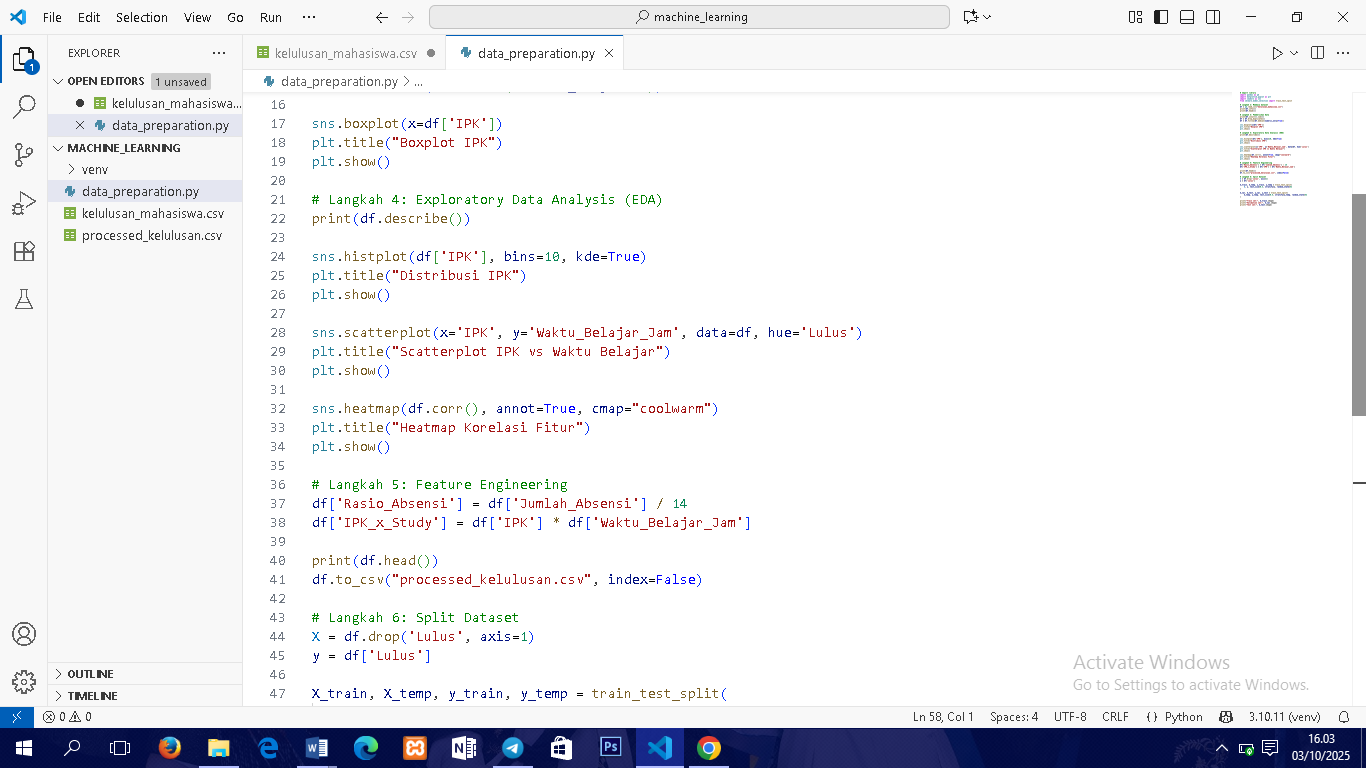
**Langkah 2 (Baca Dataset)** → berhasil baca data dengan pandas, tampil info() dan head(), tapi belum ada library visualisasi (matplotlib / seaborn) → jadi grafik tidak bisa maksimal. Pada perbaikan ditambahkan matplotlib.pyplot dan seaborn untuk keperluan visualisasi data.

**Langkah 3 (Pembersihan Data)** → hanya ada isnull().sum() dan drop\_duplicates(), belum ada handling data kosong (fillna) dan boxplot tidak muncul karena tidak ada plt.show().Pada perbaikan ditambahkan:

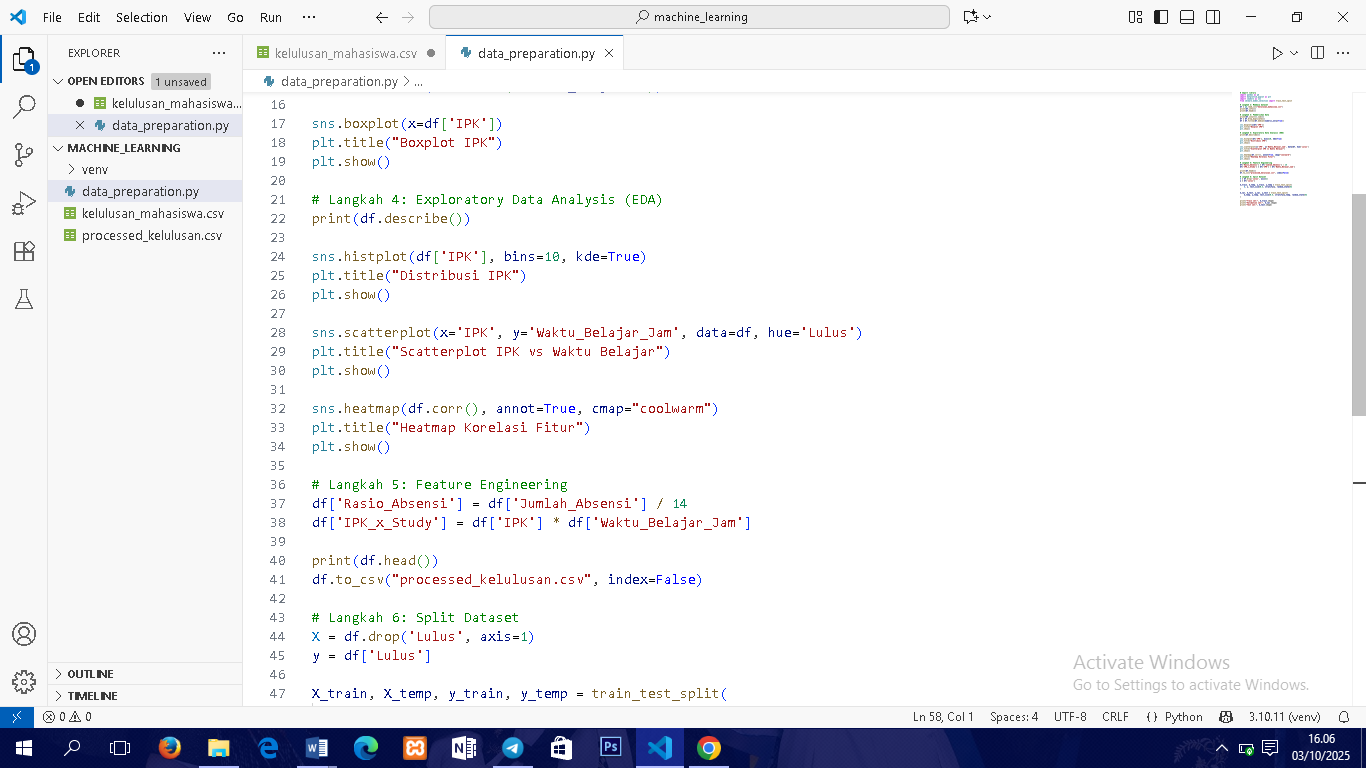
* Mengecek data kosong dengan isnull().sum()
* Mengisi nilai kosong dengan median (fillna)
* Menampilkan boxplot untuk mendeteksi outlier



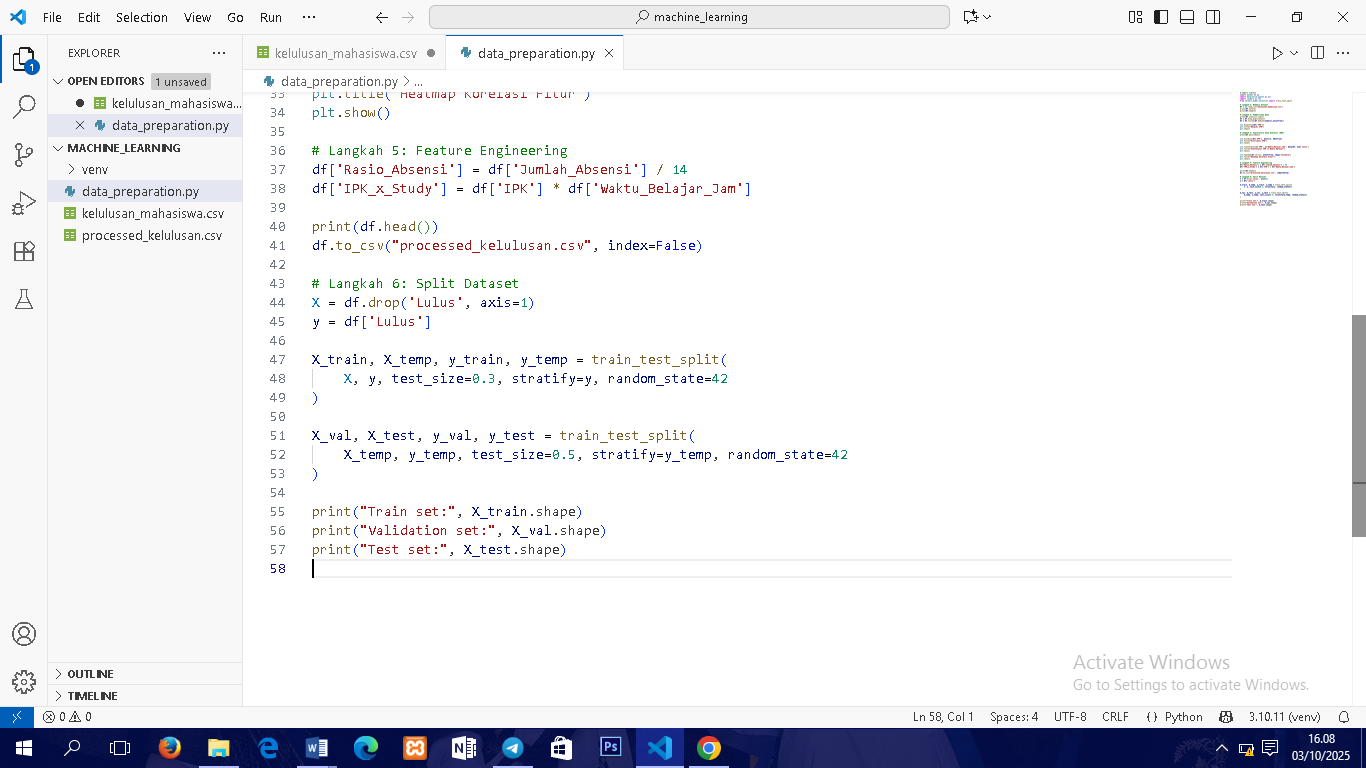
**Langkah 4 (EDA)** → hanya ada describe() serta pemanggilan histogram, scatterplot, dan heatmap → grafik tidak muncul karena belum disertakan plt.show().Perbaikan ditambahkan judul dan perintah plt.show() pada setiap grafik agar hasilnya tampil.

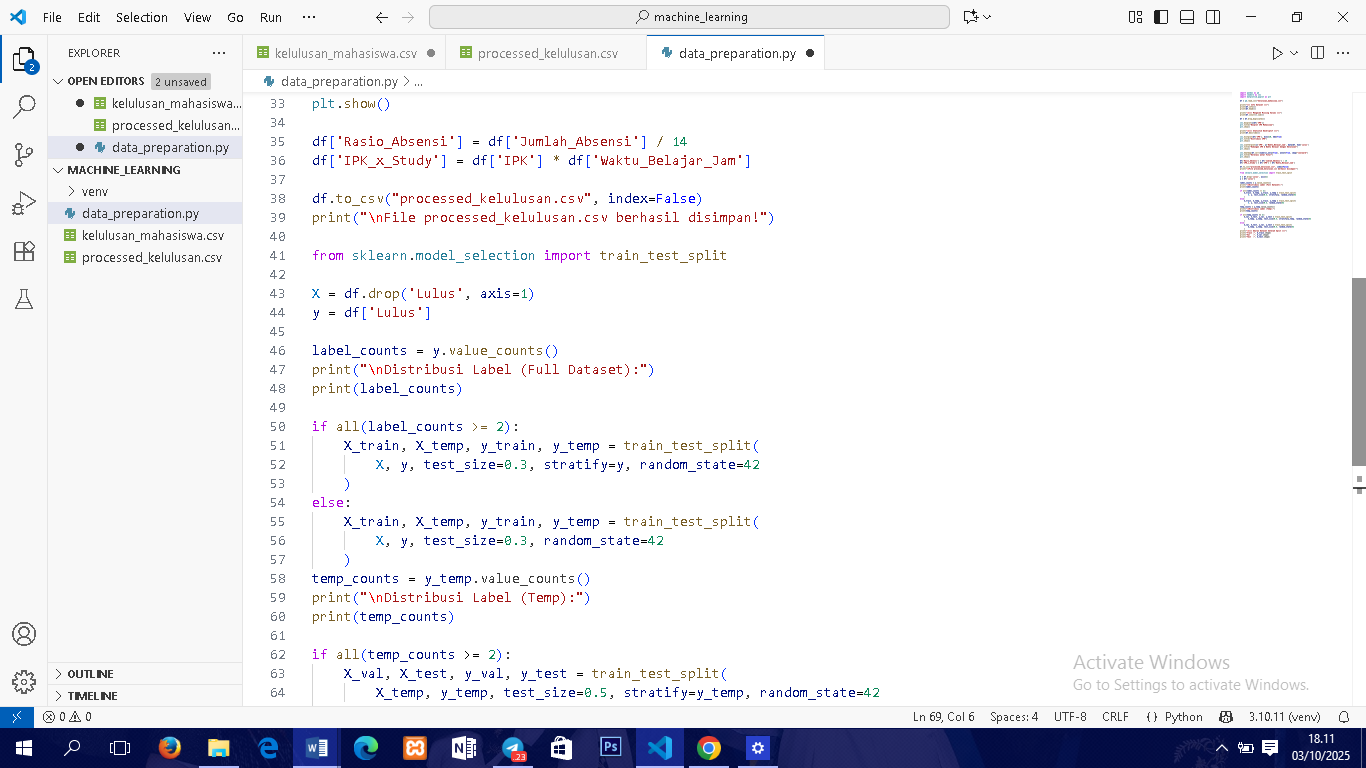
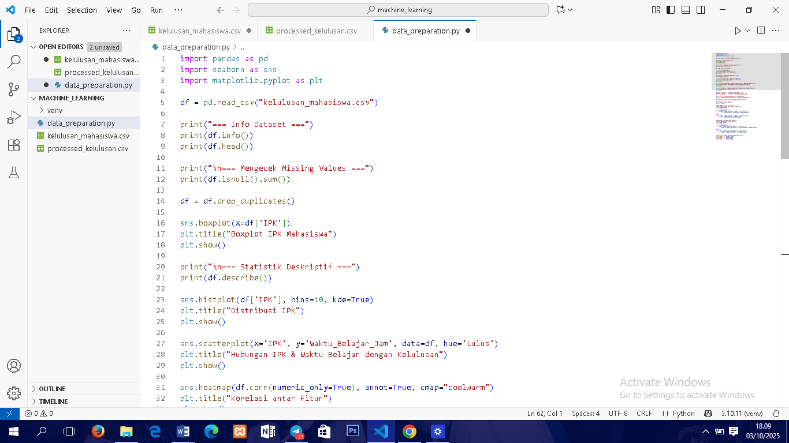


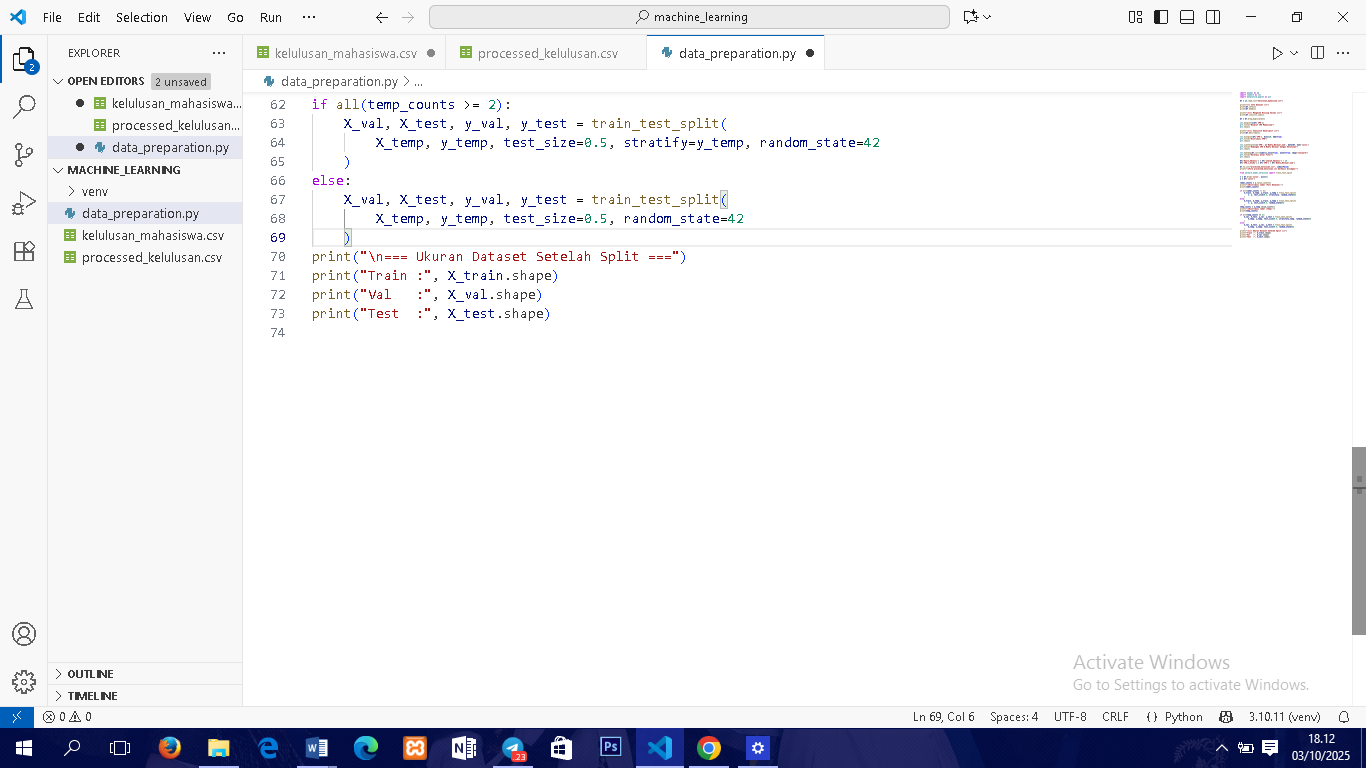
**Langkah 5 (Feature Engineering)** → berhasil menambahkan fitur Rasio\_Absensi dan IPK\_x\_Study, serta menyimpan file processed\_kelulusan.csv → tetapi belum ada pengecekan fitur baru dengan print(df.head()).Perbaikan ditambahkan print(df.head()) untuk mengecek hasil penambahan fitur.



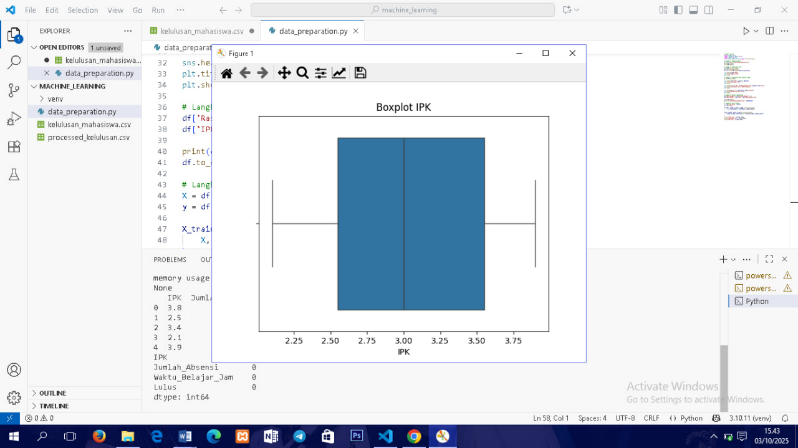
**Langkah 6 (Split Dataset)** → dataset sudah dibagi menjadi Train, Validation, dan Test menggunakan train\_test\_split, namun belum ada print ukuran dataset setelah dibagi. erbaikan ditambahkan print untuk menampilkan ukuran dataset Train, Validation, dan Test agar lebih jelas.

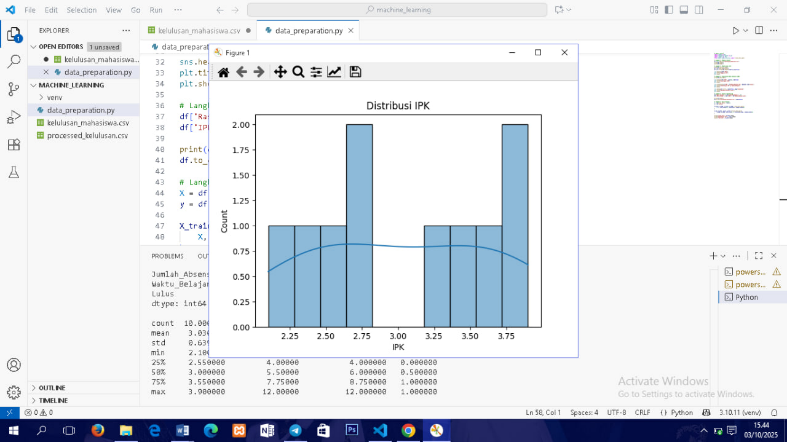


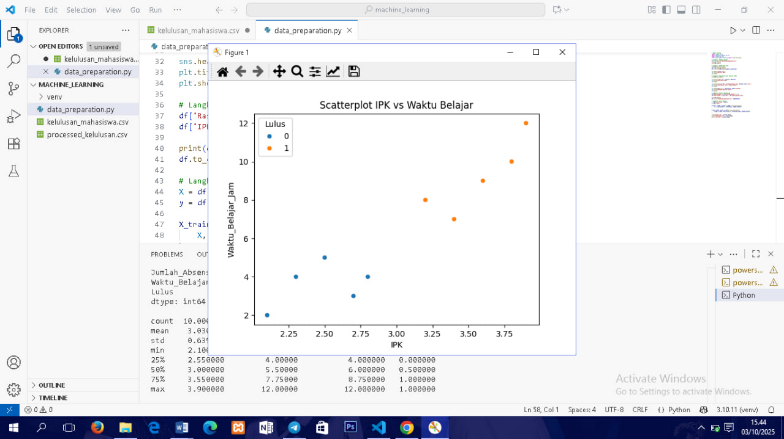
Kode dan Hasil:



*Kode yang sudah diperbaiki dan dirapihkan*

Boxplot memperlihatkan median IPK sekitar 3.0, dengan rentang nilai dari 2.2 hingga 3.9. Tidak ditemukan outlier yang ekstrem, sehingga distribusi IPK relatif merata dan stabil.

Grafik ini menunjukkan sebaran nilai IPK mahasiswa. Terlihat bahwa IPK berada pada kisaran 2.2 sampai 3.9, dengan kecenderungan ada dua kelompok utama yaitu mahasiswa dengan IPK rendah dan mahasiswa dengan IPK tinggi.

Scatterplot menunjukkan hubungan antara IPK dan waktu belajar berdasarkan status kelulusan. Mahasiswa yang lulus (warna oranye) cenderung memiliki IPK di atas 3.0 dan waktu belajar lebih dari 8 jam. Sebaliknya, mahasiswa yang tidak lulus (warna biru) umumnya memiliki IPK rendah dan waktu belajar yang sedikit.

Heatmap menampilkan hubungan antar variabel. Terlihat IPK berkorelasi positif kuat dengan waktu belajar (0.94) dan negatif kuat dengan absensi (-0.97). Artinya, semakin banyak waktu belajar IPK semakin tinggi, sedangkan semakin sering absen IPK semakin rendah. Faktor absensi dan waktu belajar juga berpengaruh besar terhadap kelulusan.