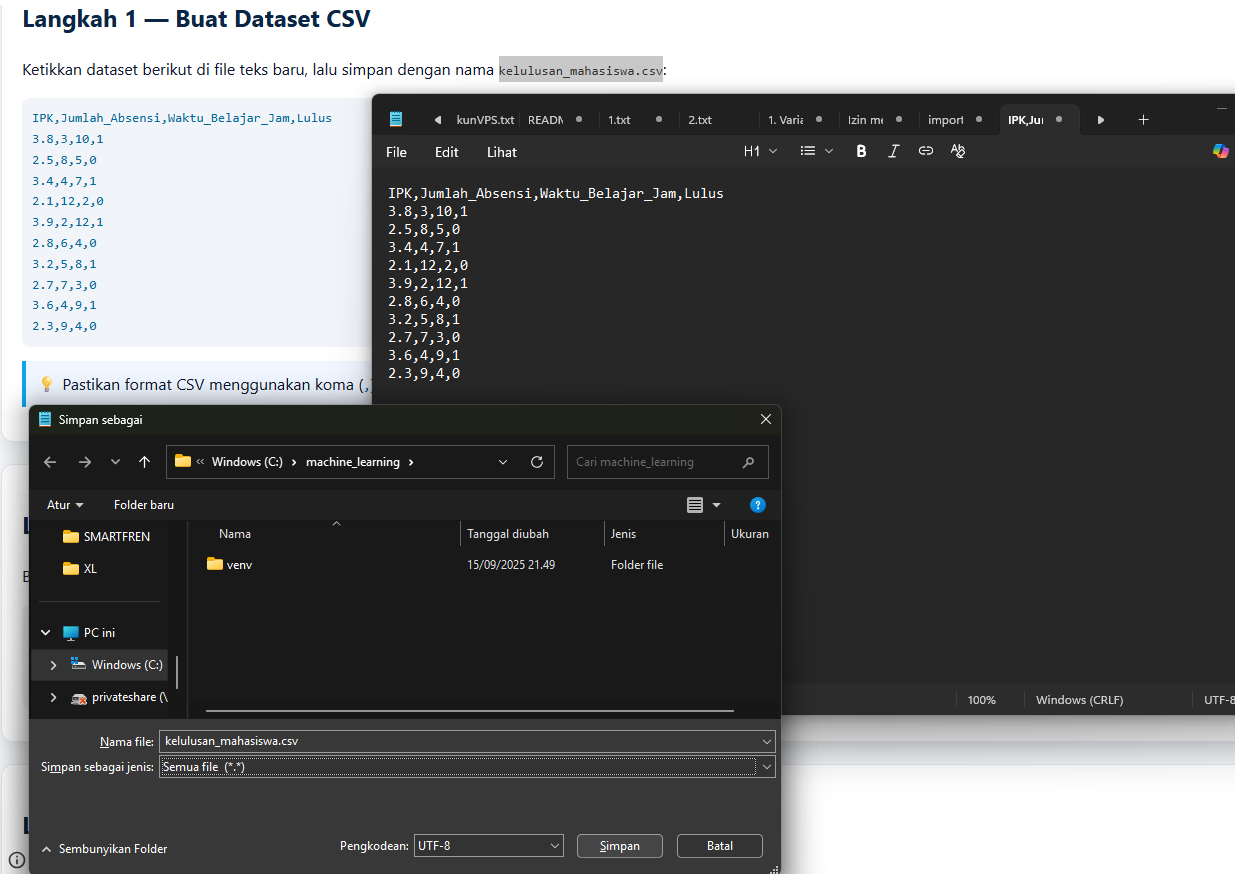
**NAMA : SEFTIA DELLA FIISYATIR RODHIAH**

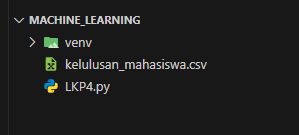
**NIM : 231011401012**

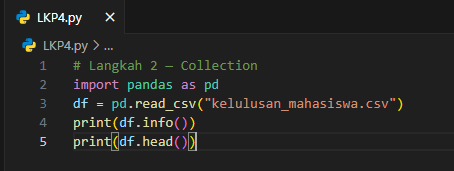
**KELAS : TI.05TPLE016**

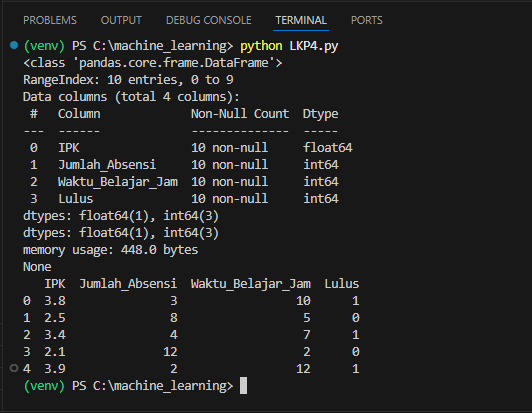
**Lembar Kerja Pertemuan 4** – **Machine Learning**

1. Langkah 1 – Membuat Dataset CSV  
   Simpan terlebih dahulu dataset dalam format CSV ke folder yang di pilih.



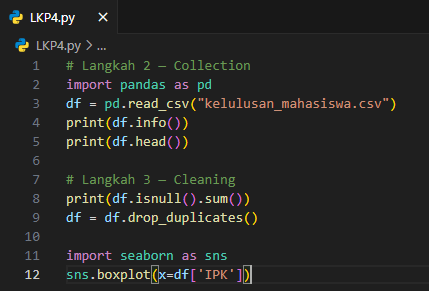
Kemudian salin file tersebut ke direktori yang sudah memiliki environment Python.

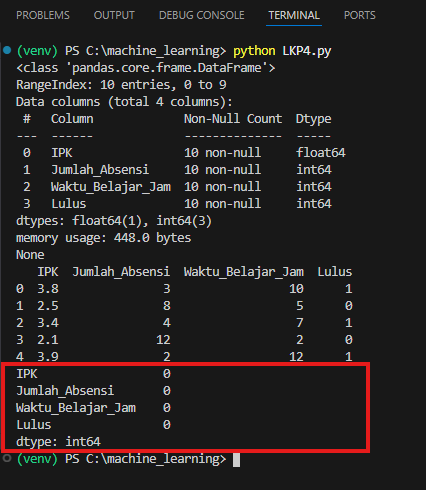
1. Langkah 2 – Collection  
   Masukkan potongan kode berikut, lalu jalankan.  
   

Hasil yang muncul akan terlihat seperti berikut:

**Penjelasan:**

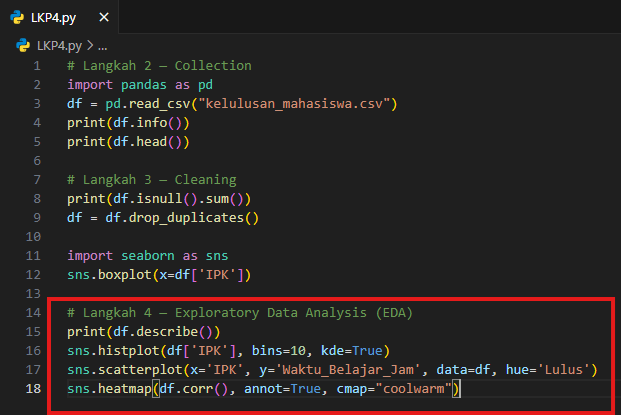
* Fungsi pandas.read\_csv() digunakan untuk membaca file CSV bernama **kelulusan\_mahasiswa.csv**.
* Perintah df.info() menampilkan informasi mengenai struktur DataFrame, seperti jumlah kolom, tipe data, serta apakah terdapat nilai kosong.
* Sedangkan df.head() digunakan untuk menampilkan **lima baris pertama** dari dataset.

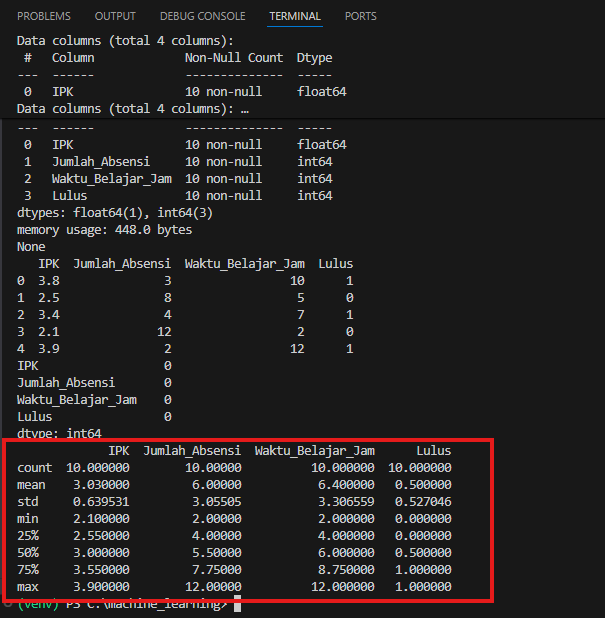
1. Langkah 3 – Data Cleaning  
   Masukkan kode lanjutan berikut untuk melakukan proses pembersihan data, lalu jalankan program tersebut.  
   

Output yang dihasilkan akan tampak seperti ini:

**Penjelasan:**

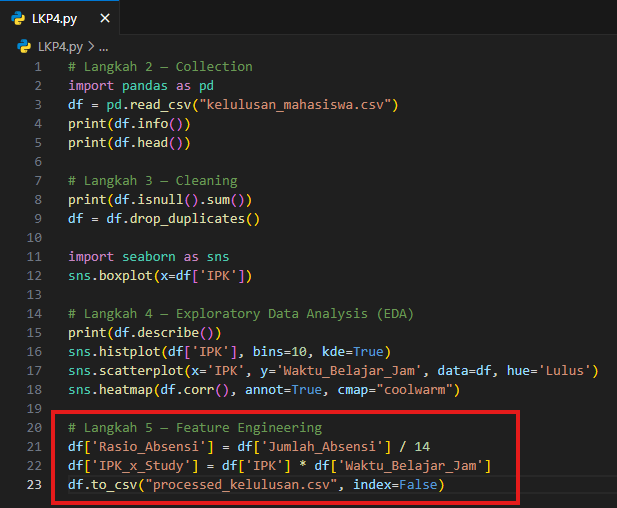
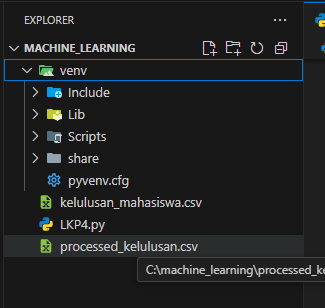
* Mengecek apakah terdapat nilai kosong (**NaN**) pada setiap kolom.
* Menghapus **baris duplikat** jika ditemukan data yang sama.
* Membuat **boxplot** pada kolom **IPK** untuk mendeteksi adanya **outlier**.

1. Langkah 4 – Exploratory Data Analysis (EDA)  
   Selanjutnya, tambahkan potongan kode berikut untuk melakukan analisis eksploratif terhadap data  
     
   Hasil analisis akan muncul setelah dijalankan.



**Penjelasan:**

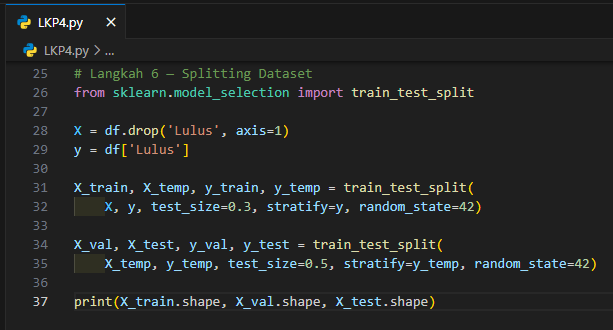
* df.describe() menampilkan **statistik deskriptif** seperti nilai rata-rata, standar deviasi, nilai minimum, maksimum, dan kuartil.
* sns.histplot() digunakan untuk melihat **distribusi nilai IPK**.
* sns.scatterplot() menampilkan **hubungan antara IPK dan waktu belajar**, dengan pewarnaan berdasarkan status kelulusan.
* sns.heatmap() menunjukkan **tingkat korelasi antar variabel** di dalam datase

1. Langkah 5 – Feature Engineering  
   Masukkan kode berikut untuk membuat fitur baru pada dataset, lalu jalankan.  
     
   setelah berhasil, akan terbentuk file baru bernama “**processed\_kelulusan.csv**” di folder yang sama  
   

**Penjelasan:**  
Dibuat dua fitur tambahan, yaitu:

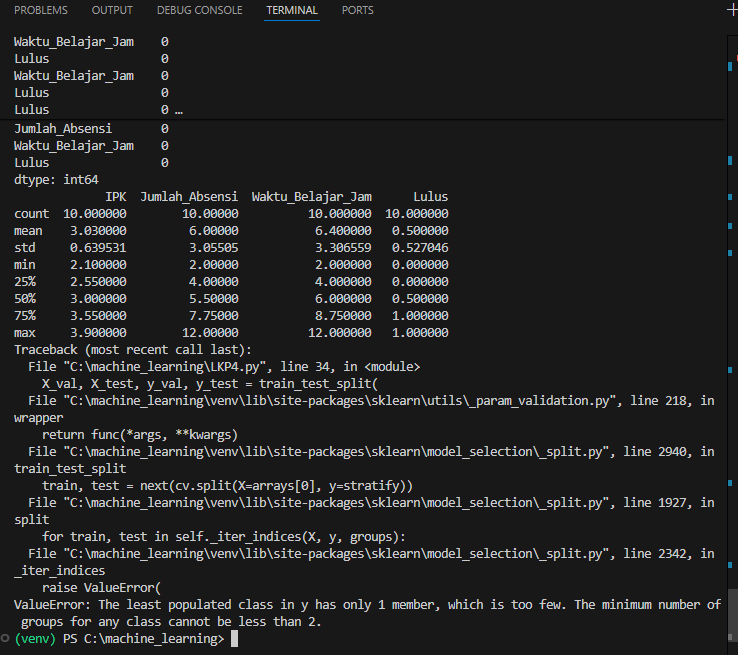
* **Rasio\_Absensi** → menunjukkan perbandingan antara jumlah kehadiran mahasiswa dengan total **14 pertemuan**.
* **IPK\_x\_Study** → hasil perkalian antara **IPK dan waktu belajar**, yang menggambarkan kombinasi antara prestasi dan usaha mahasiswa.

Dataset hasil modifikasi kemudian disimpan ke dalam file **processed\_kelulusan.csv**, dengan pengaturan agar **index baris tidak ikut disimpan**.

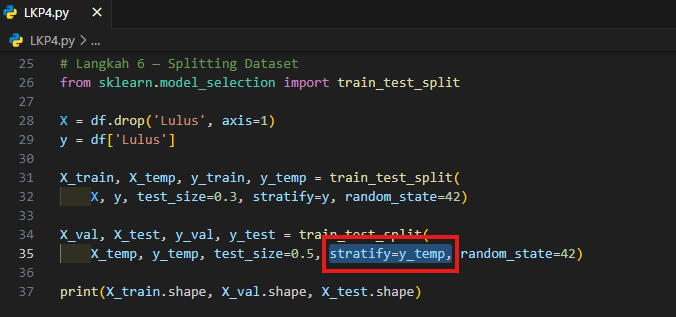
1. Langkah 6 – Splitting Dataset  
   tambahkan kode berikut untuk membagi dataset menjadi bagian pelatihan, validasi, dan pengujian, lalu jalankan.  
     
   **Penjelasan:**  
   Proses ini bertujuan untuk memisahkan:

* Fitur (X) dan target (y),
* Kemudian membagi data menjadi:
  + **70% untuk training**,
  + **15% untuk validation**,
  + **15% untuk testing**.

Parameter **stratify=y** berfungsi agar proporsi kelas *Lulus (1)* dan *Tidak Lulus (0)* tetap seimbang pada setiap subset data.

namun, ketika dijalankan, muncul pesan error.  


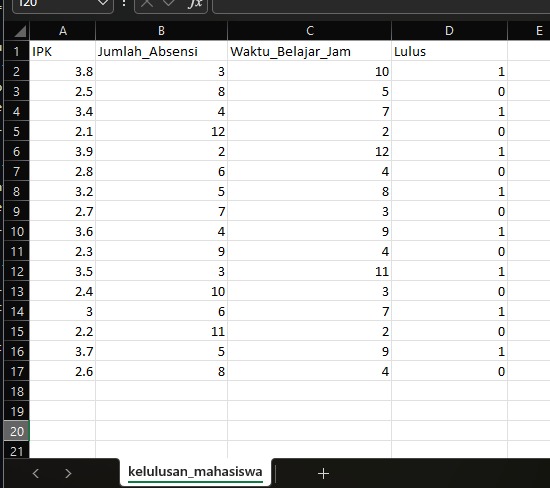
**Penyebab error:**  
Kelas dengan jumlah data paling sedikit di variabel **y** hanya memiliki satu anggota. Pada pembagian data dengan stratify, setiap kelas minimal harus memiliki dua data agar proses berjalan.



Masalah ini muncul di bagian **train\_test\_split kedua**, karena jumlah data total hanya **10 baris**. Saat proses pembagian dua tahap dilakukan, salah satu kelas di subset data (**y\_temp**) hanya memiliki **1 data saja**, sehingga scikit-learn tidak bisa melakukan stratifikasi.

Dengan kata lain, saat pembagian pertama dilakukan, salah satu label (misalnya Lulus = 1 atau Tidak Lulus = 0) tersisa satu baris saja, sehingga pembagian kedua gagal karena **stratify membutuhkan minimal 2 data per kelas**.

Untuk mengatasi hal tersebut, dilakukan **penambahan data baru** di file **kelulusan\_mahasiswa.csv**, dari **10 baris menjadi 16 baris**.



Setelah disimpan dan dijalankan kembali, program berhasil berjalan dengan **output yang sesuai**.

