M. Fikri Avishena Parinduri

231011401029

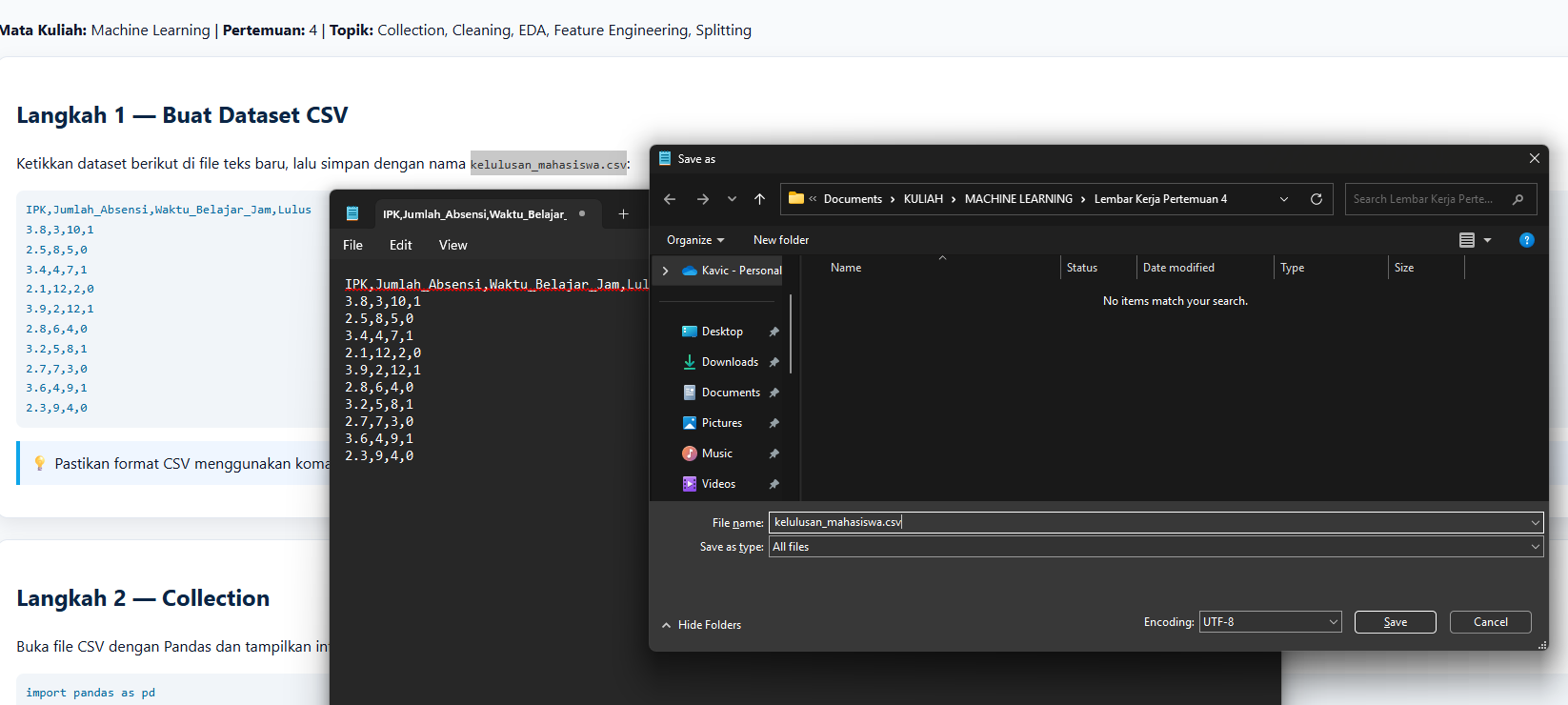
05TPLE016

Machine Learning

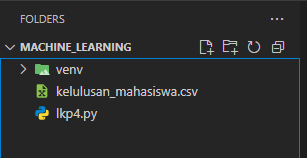
Lembar Kerja Pertemuan 4

1. Langkah 1 – Buat Dataset CSV

Simpen dulu di folder mana aja

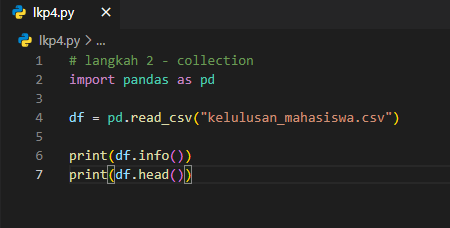


Lalu copy dan paste ke folder yang sudah ada environment python nya di dalam nya



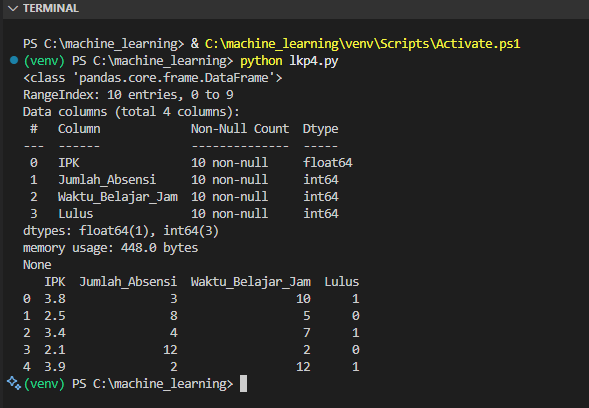
1. Langkah 2 – Collection

Masukkan code sebagai berikut



Lalu jalankan,

Hasil:

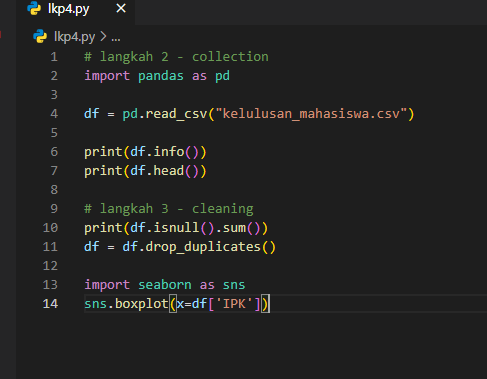


Penjelasan:

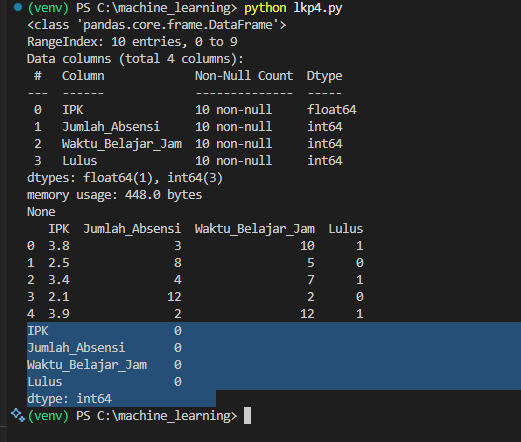
* pandas.read\_csv() membaca file CSV bernama kelulusan\_mahasiswa.csv.
* df.info() menampilkan struktur DataFrame (jumlah kolom, tipe data, dan nilai null)
* df.head() menampilkan 5 baris pertama.

1. Langkah 3 – Cleaning

Masukan kode nya lanjutkan dibawah



Jalankan dan hasil nya:

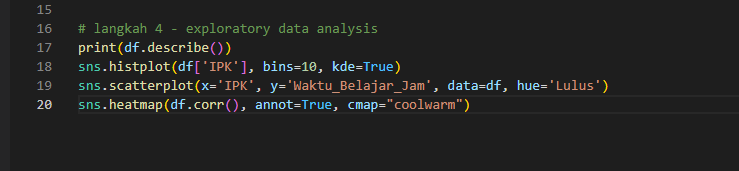


Penjelasan:

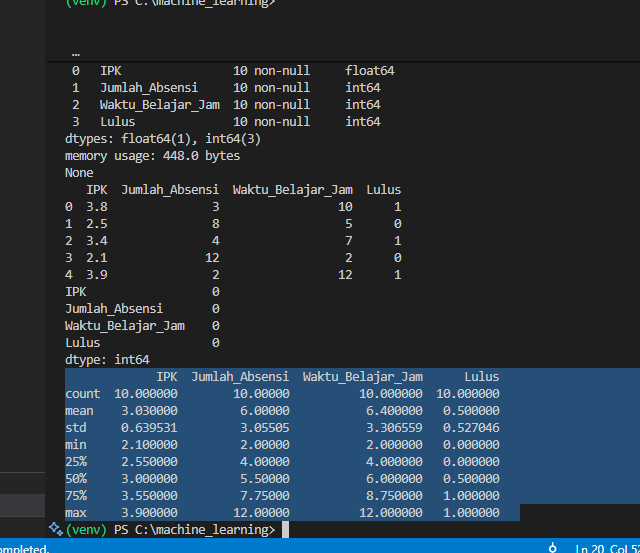
* Mengecek apakah ada nilai kosong (NaN) per kolom.
* Menghapus baris duplikat, jika ada.
* Membuat boxplot dari kolom IPK untuk mendeteksi outlier.a.

1. Langkah 4 – Exploratory Data Analysis

Masukan kode nya lanjutkan dibawah nya



Lalu jalankan dan hasil nya:

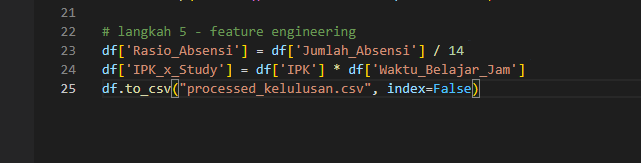


Penjelasan:

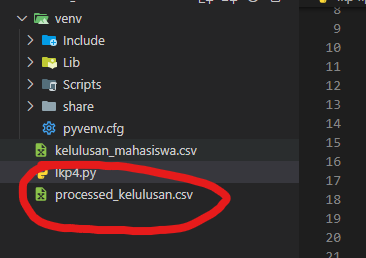
* df.describe() → statistik ringkas (mean, std, min, max, quartiles).
* sns.histplot() → distribusi nilai IPK.
* sns.scatterplot() → hubungan antara IPK dan waktu belajar, diwarnai berdasarkan Lulus.
* sns.heatmap() → menunjukkan korelasi antar variabel.

1. Langkah 5 – Feature Engineering

Masukan kode berikut, lanjutkan dibawah nya



Lalu jalankan, akan menghasilkan file “processed\_kelulusan.csv” di dalam folder yang sama

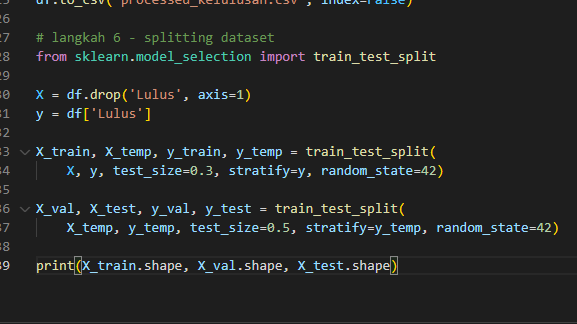


Penjelasan:

* Membuat dua fitur baru:
  + Rasio\_Absensi → rasio kehadiran mahasiswa terhadap total 14 pertemuan.
  + IPK\_x\_Study → hasil kali IPK dan waktu belajar (indikator kombinasi prestasi dan usaha).
* Menyimpan dataset baru ke processed\_kelulusan.csv tanpa menulis index baris.

1. Langkah 6 – Splitting Dataset

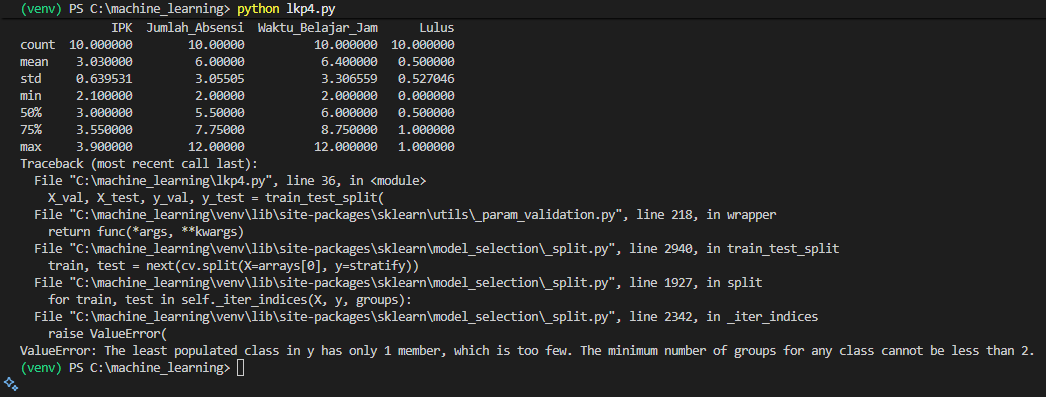
Masukan kode berikut, lanjutkan dibawahnya



Penjelasan:

* Memisahkan fitur (X) dan target (y).
* Data dibagi menjadi:
  + 70% training,
  + 15% validation,
  + 15% testing.
* stratify=y menjaga proporsi kelas Lulus (1) dan Tidak Lulus (0) agar seimbang di semua subset.

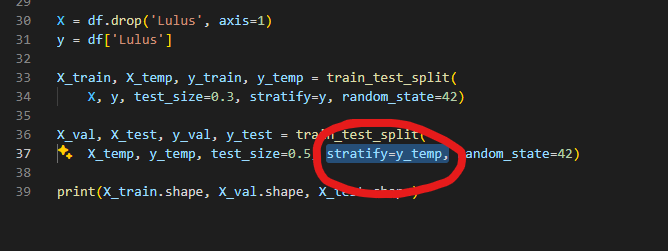
Ketika dijalankan ini akan error:



Error ini dikarenakan:

Kelas dengan populasi paling sedikit di y hanya memiliki 1 anggota, yang terlalu sedikit. Jumlah minimum grup untuk setiap kelas tidak boleh kurang dari 2.

Dibagian train\_test\_split ke-dua

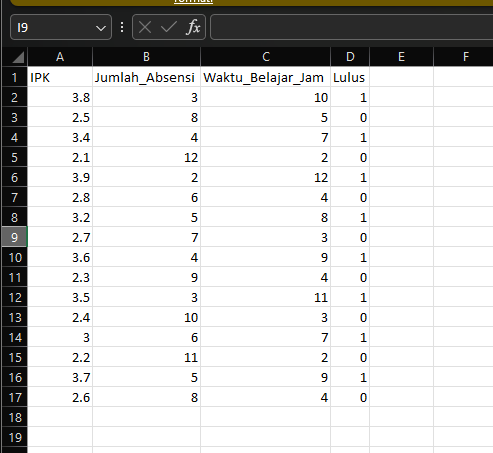


Masalahnya murni karena jumlah data hanya 10 baris, sehingga saat dibelah dua kali, ada kelas target yang di sisa data hanya 1 baris—dan scikit-learn menolak melakukan stratify kalau suatu kelas < 2.

y\_temp ternyata punya kelas yang jumlahnya hanya 1 data saja.

Dengan kata lain, saat pembagian tahap pertama (train\_test\_split awal) data yang sedikit menyebabkan salah satu kelas (misal Lulus = 1 atau 0) hanya tersisa 1 baris di y\_temp, sehingga pembagian kedua gagal — karena stratify butuh minimal 2 data per kelas.

Disini saya akan melakukan penambahan data di file kelulusan\_mahasiswa.csv yang awalnya 10 data, menjadi 16 baris data



Lakukan save, lalu jalankan lagi kode nya

Dan outputnya terlihat:

