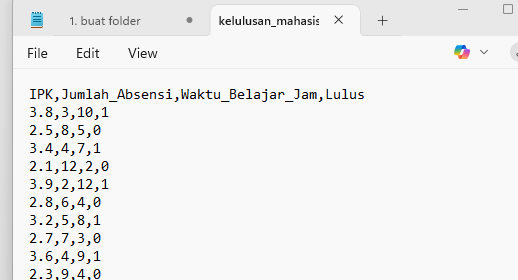
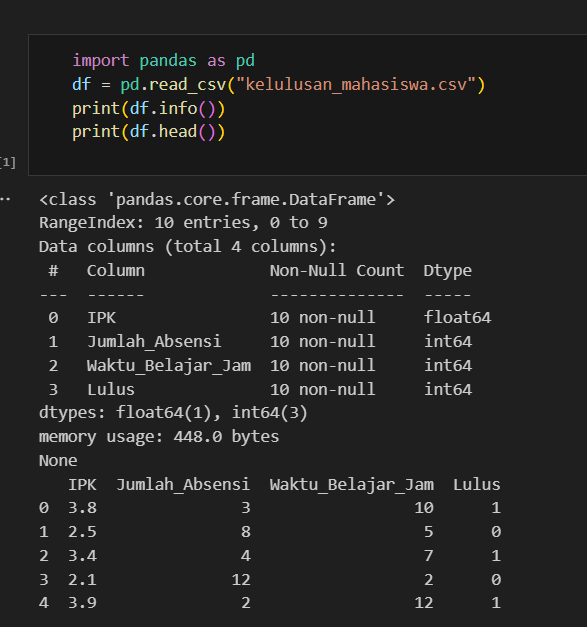
1. **Langkah 1 — Buat Dataset CSV**

Data set dapat dibuat melalui notepad, lalu save dengan format “.csv”.

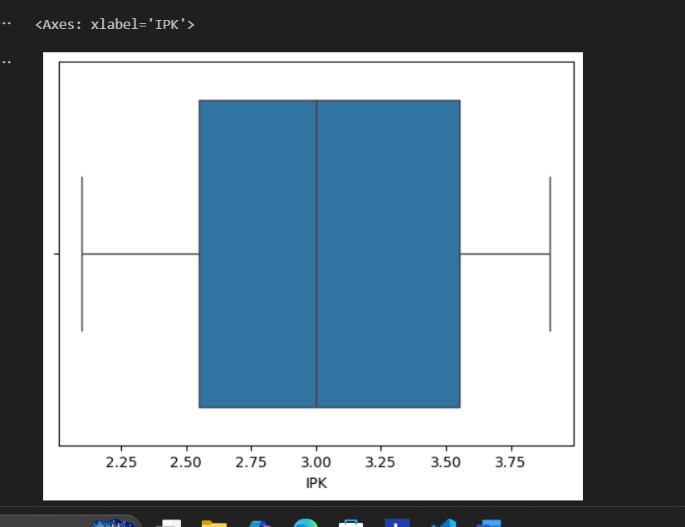


1. **Langkah 2 — Collection**

Dengan pandas.read\_csv(), file CSV bisa langsung dibaca jadi DataFrame (mirip tabel Excel) sehingga datanya lebih rapi dan mudah dibaca.



1. **Langkah 3 — Cleaning**

Menampilkan jumlah missing value (NaN) di setiap kolom fan data yang terindikasi duplikat. Pada gambar di bawah ini, tidak ada data yg terindikasi duplikat, sehingga hasilnya 0. Untuk sns.boxplot digunakan untuk menampilkan IPK.

A screen shot of a computer program

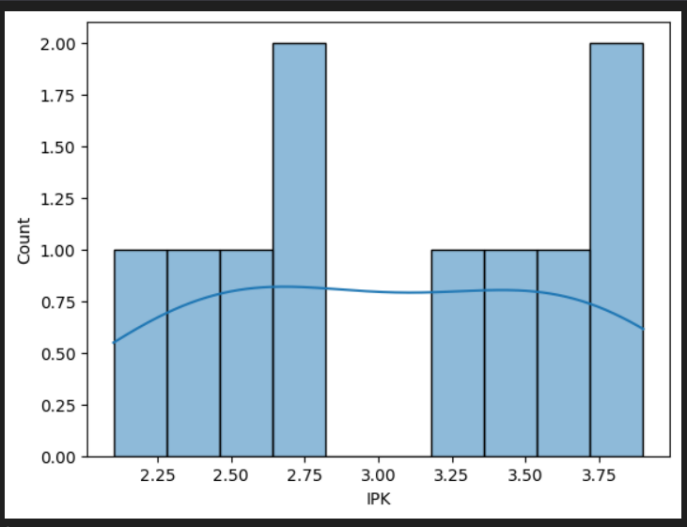
Description automatically generated

1. **Langkah 4 — Exploratory Data Analysis (EDA)**

* Ketiga kode di bawah digunakan untuk:  
  Histogram (sns.histplot) membantu melihat distribusi frekuensi data.
* Scatterplot (sns.scatterplot) menunjukkan hubungan antar dua variabel dengan warna sebagai pembeda kategori.
* Heatmap (sns.heatmap) memvisualisasikan korelasi antar variabel dengan warna yang menggambarkan kekuatan hubungan.

A computer code with colorful text

Description automatically generated

**A graph with numbers and dots

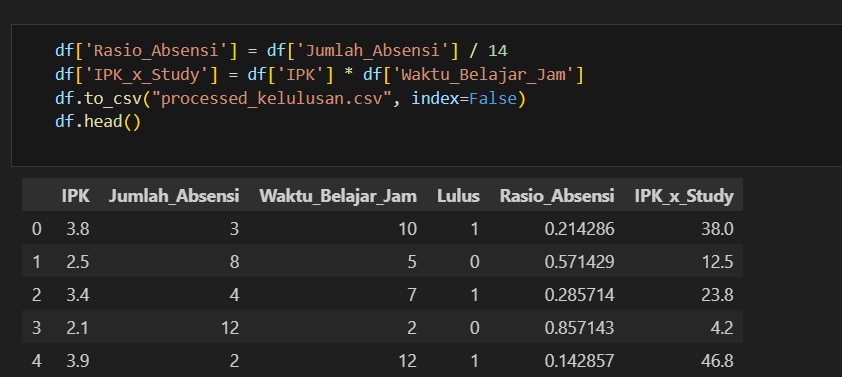
Description automatically generated**

**A red and blue squares with white text

Description automatically generated**

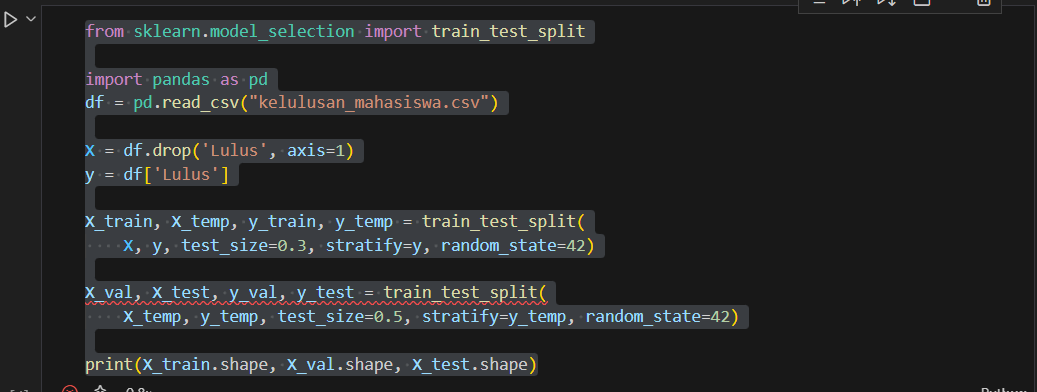
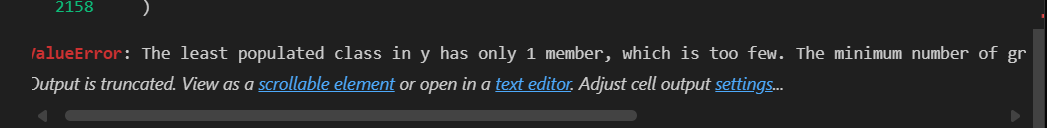
1. **Langkah 5 — Feature Engineering**

Feature engineering adalah proses membuat fitur (variabel) baru dari data yang sudah ada, supaya model bisa lebih mudah menemukan pola.

****

1. **Langkah 6 — Splitting Dataset**

Saat menjalankan script, terjadi error karena ukuran dataset terlalu kecil. Setelah dilakukan proses cleaning dan split pertama, hanya tersisa 1 data pada salah satu kelas. Akibatnya, proses split kedua gagal dijalankan karena stratifikasi membutuhkan minimal 2 data pada setiap kelas.

**** ****

Untuk mengatasi permasalahan tersebut, dilakukan menghapus stratify pada split kedua.

