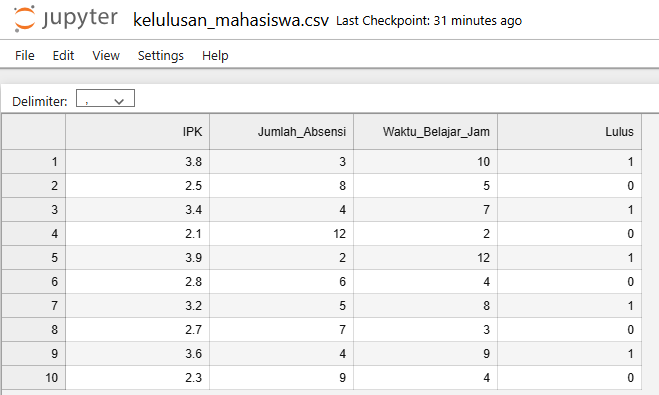
Laporan Lembar Kerja Pertemuan 4 — Data Preparation

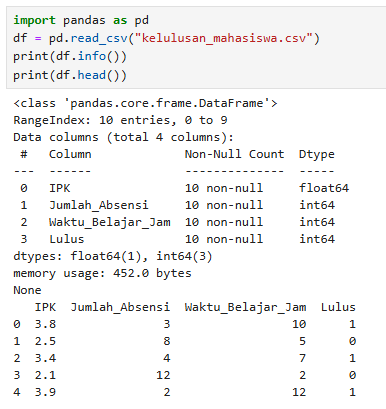
Notes : semua dilakukan dalam lembar kerja google colab

1. Langkah 1 — Buat Dataset CSV.

Berhasil membuat dataset CSV

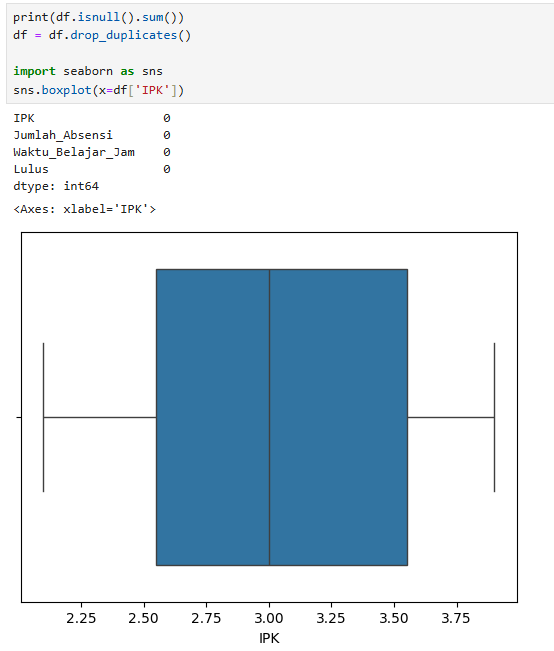


1. Langkah 2 — Collection (Membaca Data), membaca file CSV yang baru saja dibuat menggunakan library Pandas dan menampilkan informasi dasar serta beberapa baris pertama dari data tersebut.

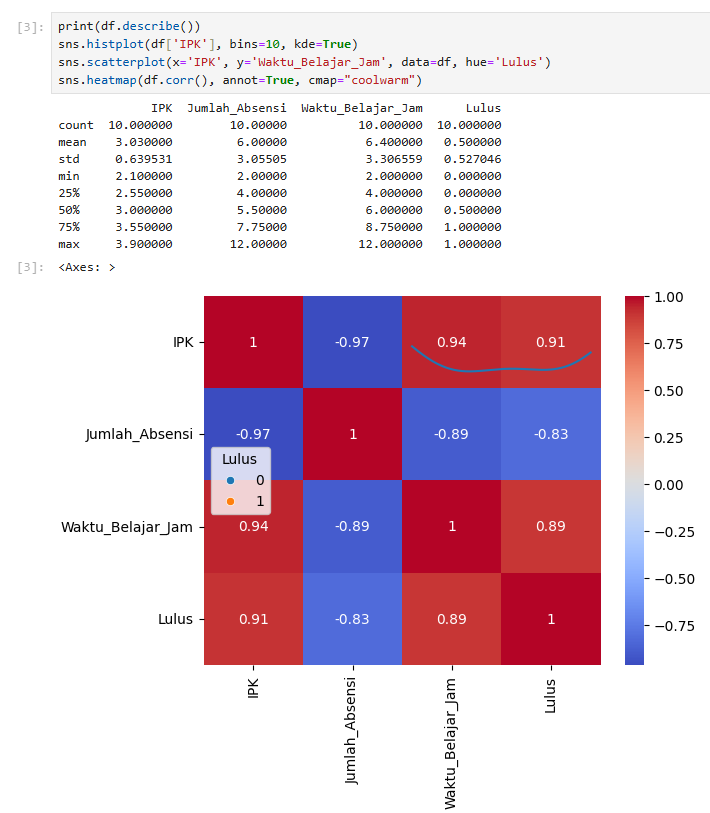


1. Langkah 3 — Cleaning (Pembersihan Data)

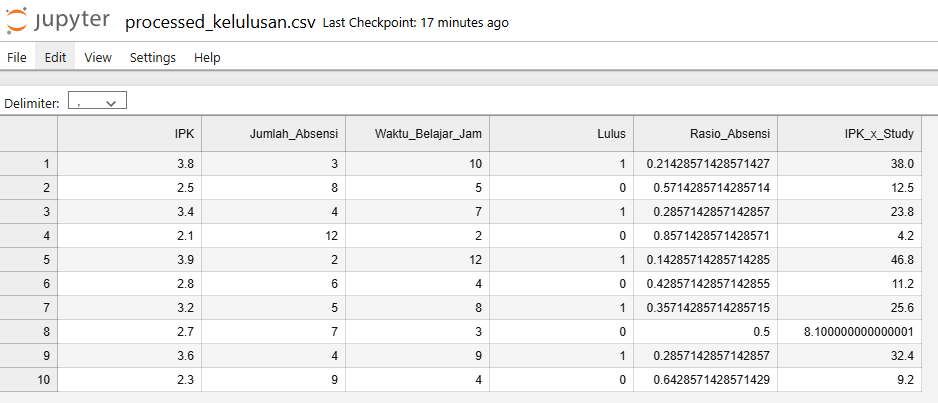
memeriksa apakah ada data yang hilang (missing values), menghapus data duplikat, dan mengidentifikasi outlier.



1. Langkah 4 - Exploratory Data Analysis (EDA), untuk memahami data lebih dalam melalui statistik deskriptif dan visualisasi.

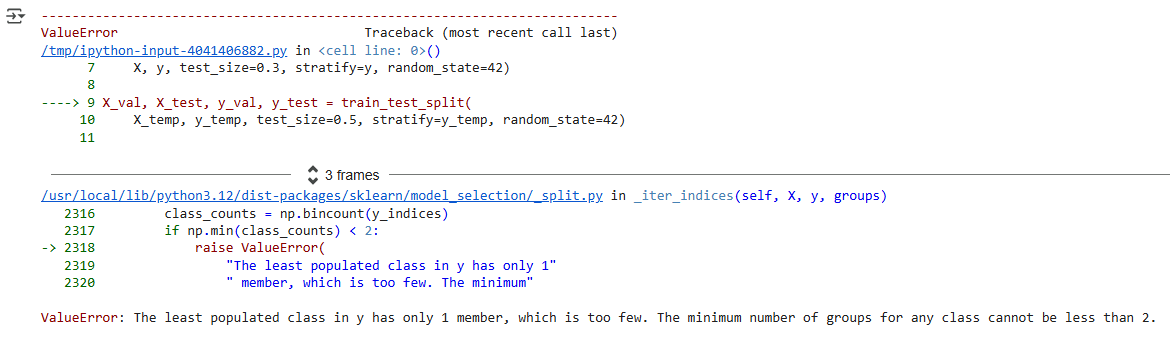


1. Langkah 5 - Feature Engineering, membuat fitur baru dari fitur yang sudah ada untuk meningkatkan performa model nantinya. Hasilnya akan disimpan ke file CSV baru.



1. Langkah 6 - Splitting Dataset, membagi dataset menjadi tiga bagian: data latih (train), data validasi (validation), dan data uji (test).

Pada langkah terakhir terdapat error yang seperti gambar dibawah ini :



Dari hasil error diatas, coba ditelusuri penyebab errornya ialah : ValueError: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less than 2.

Error terjadi karena penggunaan metode stratifikasi (stratify) pada dataset sementara yang ukurannya sangat kecil (hanya 3 baris). Akibatnya, salah satu kelas target hanya memiliki satu anggota, sehingga tidak memungkinkan untuk dipecah lebih lanjut ke dalam dua grup (validasi dan tes) secara proporsional.

Solusi yang Diterapkan: Kendala diatasi dengan menghapus parameter stratify pada fungsi pembagian dataset tahap kedua. Hal ini memungkinkan proses pembagian tetap berjalan pada data yang sangat terbatas tanpa memaksakan aturan proporsi kelas.

Setelah solusi diterapkan, proses pembagian berhasil dengan rincian ukuran sebagai berikut:

* Data Latih: 7 baris
* Data Validasi: 1 baris
* Data Uji: 2 baris

