**MACHINE LEARNING**

**LAPORAN PERTEMUAN 4**

**Data Preparation dan Exploratory Data Analysis (EDA)**

**Nama : Haidar Reyhan**

**NIM : 231011400547**

**Kelas : 05TPLE016**

**Judul : Prediksi Kelulusan Mahasiswa**

**File : kelulusan\_mahasiswa.csv**

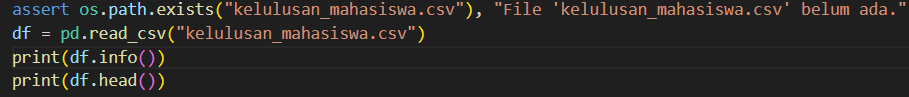
**Nama Output : processed\_mahasiswa.csv**

1. **Tujuan**

Tujuan kegiatan ini adalah melakukan proses **pembersihan data (data cleaning)**, **analisis eksploratif (EDA)**, dan **persiapan fitur (feature engineering)** sebelum data digunakan untuk pelatihan model Machine Learning.  
Proses ini bertujuan untuk:

* Mengidentifikasi missing values dan duplikasi data.
* Memahami pola distribusi setiap variabel.
* Membentuk fitur-fitur baru yang relevan terhadap target (kelulusan mahasiswa).
* Menyiapkan dataset bersih dan terstruktur untuk tahap modeling di pertemuan selanjutnya.

1. **Langkah-Langkah Pengerjaan**
   1. Langkah 1: Load Data



* Mengecek tipe data tiap kolom, jumlah baris, serta preview isi data
* Kolom utama yang digunakan: IPK, Jumlah\_Absensi, Waktu\_Belajar\_Jam, dan Lulus
  1. Langkah 2: Cleaning



* Mengecek apakah ada **missing value** (hasilnya tidak ada)
* Menghapus data duplikat agar hasil analisis tidak bias
  1. Langkah 3: Exploratory Data Analysis (EDA)

Visualisasi dan statistik digunakan untuk memahami pola data

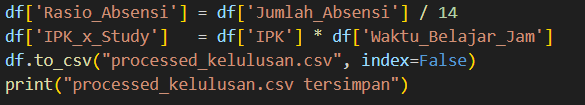


Temuan dari Visualisasi:

* **Boxplot:** mayoritas nilai IPK berada di kisaran 3.0–3.5; hanya sedikit outlier
* **Histogram:** distribusi IPK relatif normal
* **Scatterplot:** mahasiswa dengan IPK tinggi dan jam belajar lebih banyak umumnya berstatus “Lulus”
* **Heatmap Korelasi:** menunjukkan bahwa *IPK* dan *Waktu\_Belajar\_Jam* memiliki korelasi positif terhadap kelulusan

Semua grafik disimpan otomatis:

* p4\_boxplot\_ipk.png
* p4\_hist\_ipk.png
* p4\_scatter\_ipk\_study.png
* p4\_heatmap\_corr.png
  1. Langkah 4: Feature Engineering

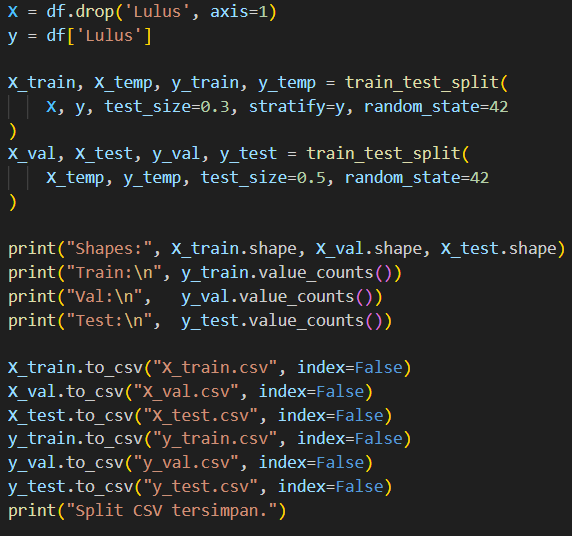


Dibuat dua fitur baru untuk memperkaya representasi data:

1. **Rasio\_Absensi**, proporsi kehadiran dari total 14 pertemuan.
2. **IPK\_x\_Study**, interaksi antara IPK dan jam belajar, yang bisa menangkap hubungan non-linear terhadap *Lulus*

Output akhir disimpan sebagai processed\_kelulusan.csv

* 1. Langkah 5: Split Data



**Penjelasan:**

* Data dibagi menjadi **Train 70%**, **Validation 15%**, dan **Test 15%**.
* Menggunakan stratify=y hanya sekali untuk menjaga proporsi kelas tetap seimbang.
* Proses split kedua tanpa stratify untuk menghindari error saat jumlah kelas terlalu kecil.

Semua subset data disimpan ke file:

*X\_train.csv, X\_val.csv, X\_test.csv*

*y\_train.csv, y\_val.csv, y\_test.csv*

1. **Hasil dan Temuan**

Berdasarkan analisis eksploratif, diperoleh bahwa IPK dan waktu belajar mahasiswa merupakan variabel yang paling memengaruhi kelulusan. Mahasiswa dengan absensi tinggi dan IPK stabil cenderung lulus tepat waktu.

|  |  |
| --- | --- |
| **Komponen** | **Hasil** |
| **Jumlah Data Awal** | (sesuai dataset asli, setelah pembersihan tidak ada duplikasi) |
| **Distribusi IPK** | Mayoritas antara 3.0–3.5 |
| **Outlier** | Sangat sedikit |
| **Fitur Baru** | Rasio\_Absensi, IPK\_x\_Study |
| **Korelasi Kuat** | IPK dan Waktu\_Belajar\_Jam ↔ Lulus |
| **Output Akhir** | processed\_kelulusan.csv dan file split train/val/test |

1. **Kesimpulan**

Dataset telah berhasil dibersihkan, divisualisasikan, dan disiapkan untuk modeling.  
Hasil eksplorasi menunjukkan bahwa:

*Mahasiswa dengan IPK tinggi, waktu belajar lebih banyak, serta tingkat absensi tinggi cenderung memiliki peluang kelulusan yang lebih besar*

Data kini siap digunakan untuk tahap modeling dan validasi (Pertemuan 5)

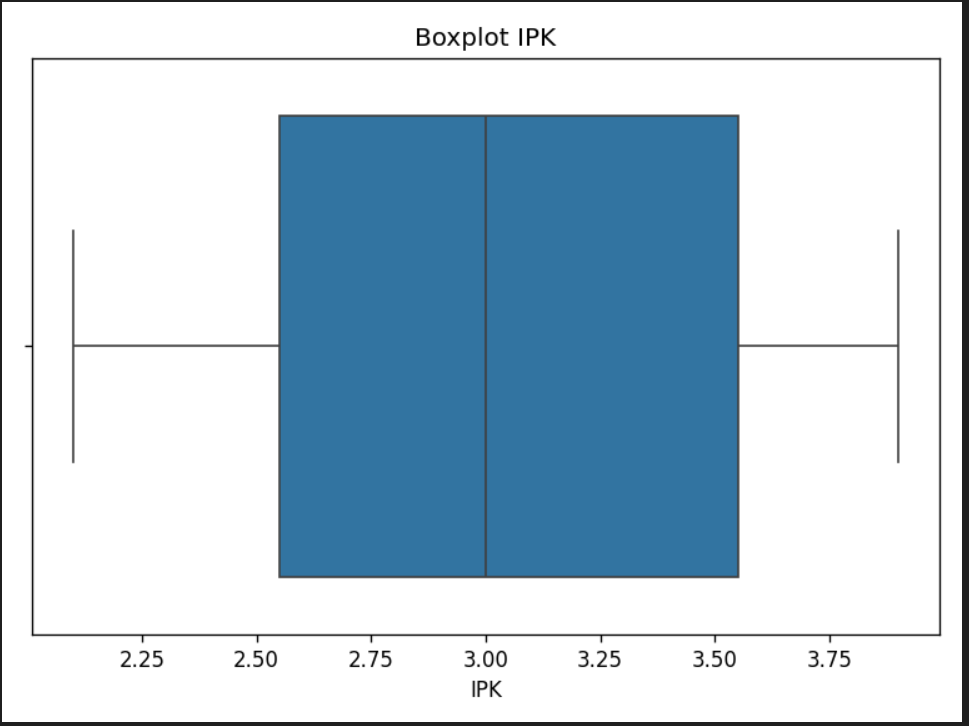
1. **Lampiran File yang dihasilkan**

File Output:

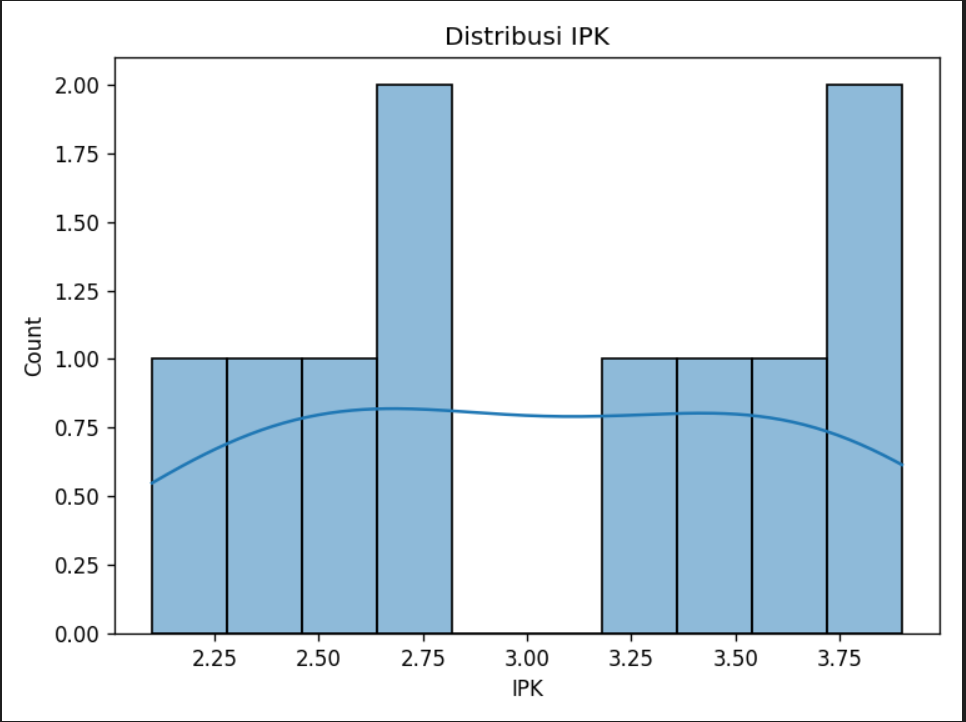
* processed\_kelulusan.csv
* X\_train.csv, X\_val.csv, X\_test.csv
* y\_train.csv, y\_val.csv, y\_test.csv

Gambar Visualisasi:

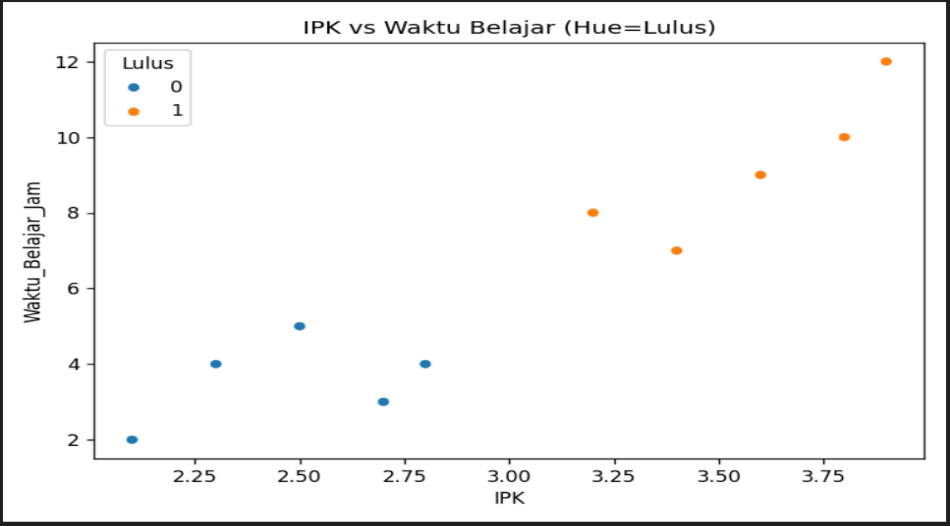
* p4\_boxplot\_ipk.png



* p4\_hist\_ipk.png



* p4\_scatter\_ipk\_study.png



* p4\_heatmap\_corr.png

