Nama: Danuarta Silalahi

NIM: 231011401071

Kelas: 05TPLE017

Laporan Pertemuan 4 – Data Preparation

Pada pertemuan ini, tujuan utamanya adalah supaya data yang kita punya bisa benar-benar siap dipakai untuk proses *machine learning*. Jadi, sebelum dipakai buat bikin model, data harus dipastikan bersih, lengkap, dan sudah dibagi-bagi sesuai keperluan. Proses ini meliputi pengumpulan data, pembersihan dari error atau data aneh, eksplorasi data secara singkat, membuat fitur tambahan, dan membagi data menjadi beberapa bagian supaya evaluasi modelnya nanti lebih adil.

1. Collection (Pengumpulan Data)

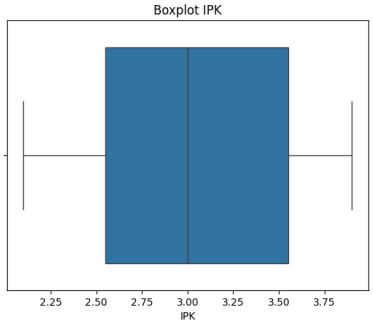
Data dikumpulkan secara manual dalam file **kelulusan_mahasiswa.csv**, yang terdiri dari 10 data mahasiswa dengan 4 kolom: **IPK**, **Jumlah_Absensi**, **Waktu_Belajar_Jam**, dan **Lulus**. Setelah itu, dataset dibaca menggunakan library **Pandas**, dan hasil perintah df.info() menunjukkan bahwa seluruh kolom tidak memiliki nilai kosong (*non-null*).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   IPK               10 non-null     float64
 1   Jumlah_Absensi    10 non-null     int64
 2   Waktu_Belajar_Jam  10 non-null     int64
 3   Lulus             10 non-null     int64
```

2. Cleaning (Pembersihan Data)

Proses pembersihan dilakukan untuk memastikan tidak ada data kosong, duplikasi, atau *outlier* dalam dataset. Dengan menggunakan perintah print (df.isnull().sum()) dan df = df.drop_duplicates(), hasil menunjukkan bahwa tidak ada nilai kosong maupun duplikat. Selain itu, boxplot digunakan untuk memeriksa potensi *outlier* pada kolom **IPK**, dan tidak ditemukan nilai ekstrem yang mencurigakan.

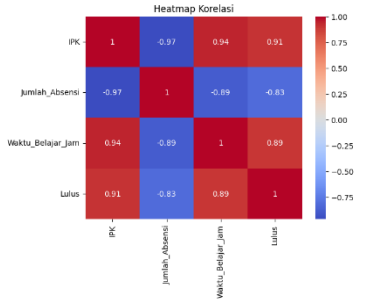
```
Missing values:
 IPK                0
Jumlah_Absensi      0
Waktu_Belajar_Jam   0
Lulus               0
dtype: int64
```

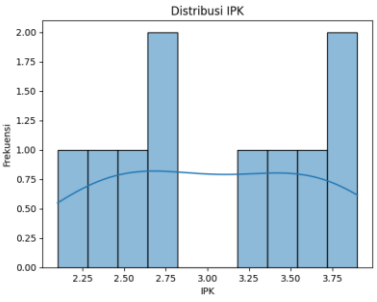


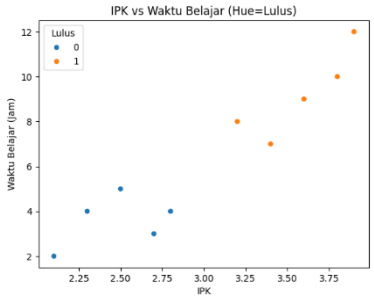
3. EDA (Exploratory Data Analysis)

Analisis eksploratif dilakukan untuk memahami pola dan distribusi data. Dengan perintah print(df.describe()) dan beberapa visualisasi, seperti histogram dan scatterplot, diperoleh informasi bahwa nilai **IPK** berkisar antara **2.1–3.9**, dengan rata-rata sekitar **3.2**. Juga ditemukan bahwa mahasiswa dengan **IPK tinggi dan waktu belajar lebih lama** cenderung **lulus (Lulus=1)**, serta adanya korelasi positif antara **IPK** dan **Waktu_Belajar_Jam**.

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
count	10.000000	10.00000	10.000000	10.000000
mean	3.030000	6.00000	6.400000	0.500000
std	0.639531	3.05505	3.306559	0.527046
min	2.100000	2.00000	2.000000	0.000000
25%	2.550000	4.00000	4.000000	0.000000
50%	3.000000	5.50000	6.000000	0.500000
75%	3.550000	7.75000	8.750000	1.000000
max	3.900000	12.00000	12.000000	1.000000







4. Feature Engineering

Dua fitur baru ditambahkan untuk meningkatkan kemampuan prediksi model. Fitur tersebut adalah **Rasio_Absensi** yang dihitung dari **Jumlah_Absensi** dibagi dengan 14, dan **IPK_x_Study** yang merupakan hasil kali antara **IPK** dan **Waktu_Belajar_Jam**. Penambahan fitur ini bertujuan untuk memberikan sinyal yang lebih baik bagi model dalam memahami hubungan antara ketidakhadiran, IPK, dan waktu belajar. Hasil dataset yang telah diproses disimpan dalam file **processed_kelulusan.csv**.

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus	Rasio_Absensi	IPK_x_Study
0	3.8	3	10	1	0.214286	38.0
1	2.5	8	5	0	0.571429	12.5
2	3.4	4	7	1	0.285714	23.8
3	2.1	12	2	0	0.857143	4.2
4	3.9	2	12	1	0.142857	46.8

processed_kelulusan.csv > data

```
IPK,Jumlah_Absensi,Waktu_Belajar_Jam,Lulus,Rasio_Absensi,IPK_x_Study
3.8,3,10,1,0.21428571428571427,38.0
2.5,8,5,0,0.5714285714285714,12.5
3.4,4,7,1,0.2857142857142857,23.8
2.1,12,2,0,0.8571428571428571,4.2
3.9,2,12,1,0.14285714285714285,46.8
2.8,6,4,0,0.42857142857142855,11.2
3.2,5,8,1,0.35714285714285715,25.6
2.7,7,3,0,0.5,8.100000000000001
3.6,4,9,1,0.2857142857142857,32.4
2.3,9,4,0,0.6428571428571429,9.2
```

5. Splitting Dataset

Dataset dibagi 70% training, 15% validasi, 15% testing, menjaga proporsi label seimbang. Data training: 7 baris, data validasi: 1 baris, dan data testing: 2 baris. Masing-masing memiliki 5 fitur(tanpa kolom Lulus).

```
(7, 5) (1, 5) (2, 5)
```