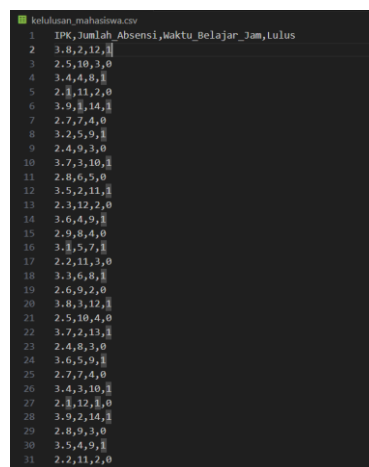


Nama : Rafly Sanjaya
NIM : 231011400875
Kelas : 05TPLE015
Mata Kuliah : Machine Learning

Laporan Data Preparation – Pertemuan 4

1. Buat Dataset

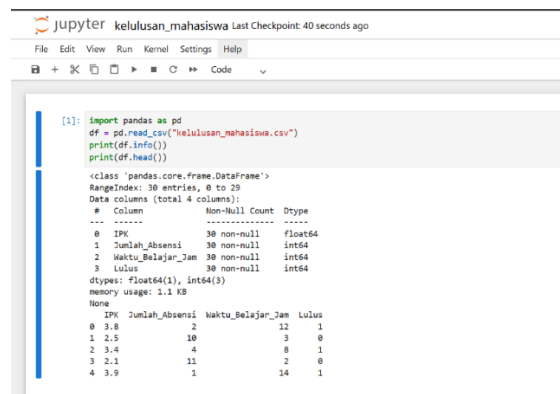
Dataset yang digunakan adalah dataset dummy atau fiktif yang saya acari di AI, kemudian diketik di vscode dengan format .csv



```
1 IPK,Jumlah_Absensi,Waktu_Belajar_Jam,Lulus
2 3.8,2,12,1
3 2.5,10,3,0
4 3.4,4,8,1
5 2.1,11,2,0
6 3.9,1,14,1
7 2.7,7,4,0
8 3.2,5,9,1
9 2.4,9,3,0
10 3.7,3,10,1
11 2.8,6,5,0
12 3.5,2,11,1
13 2.3,12,2,0
14 3.6,4,9,1
15 2.9,8,4,0
16 3.1,5,7,1
17 2.2,11,3,0
18 3.3,6,8,1
19 2.6,9,2,0
20 3.8,3,12,1
21 2.5,10,4,0
22 3.7,2,13,1
23 2.4,8,3,0
24 3.6,5,9,1
25 2.7,7,4,0
26 3.4,3,10,1
27 2.1,12,1,0
28 3.9,2,14,1
29 2.8,9,3,0
30 3.5,4,9,1
31 2.2,11,2,0
```

2. Collection

Dataset **kelulusan_mahasiswa.csv** dimuat menggunakan Pandas. Informasi awal menunjukkan dataset berisi 30 baris dan 4 kolom (**IPK**, **Jumlah_Absensi**, **Waktu_Belajar_Jam**, **Lulus**) dengan tipe data **float64** untuk IPK dan **int64** untuk kolom lainnya. Beberapa baris pertama (**df.head()**) menampilkan data mahasiswa dengan variasi IPK, jumlah absensi, waktu belajar, dan status kelulusan, sehingga dataset siap untuk tahap cleaning dan eksplorasi lebih lanjut.



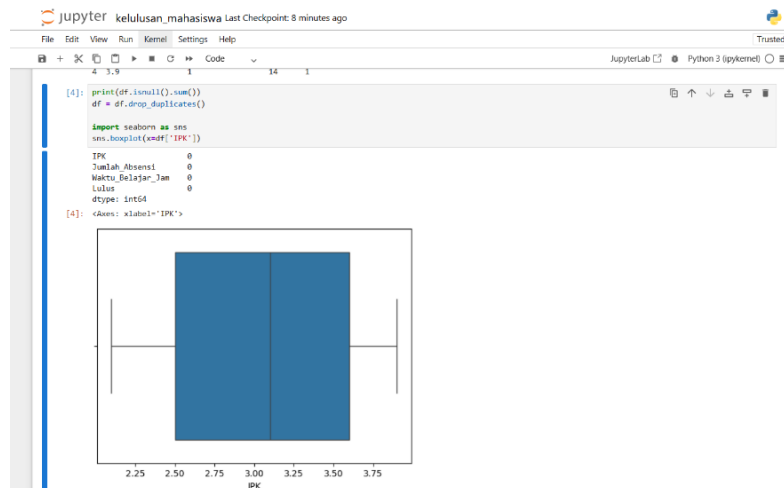
```
[1]: import pandas as pd
df = pd.read_csv("kelulusan_mahasiswa.csv")
print(df.info())
print(df.head())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype  
---  --
 0   IPK                 30 non-null    float64
 1   Jumlah_Absensi     30 non-null    int64  
 2   Waktu_Belajar_Jam  30 non-null    int64  
 3   Lulus              30 non-null    int64  
dtypes: float64(1), int64(3)
memory usage: 1.1 KB

None
IPK  Jumlah_Absensi  Waktu_Belajar_Jam  Lulus
0  3.8              2              12             1
1  2.5              10             3              0
2  3.4              4              8              1
3  2.1              11             2              0
4  3.9              1              14             1
```

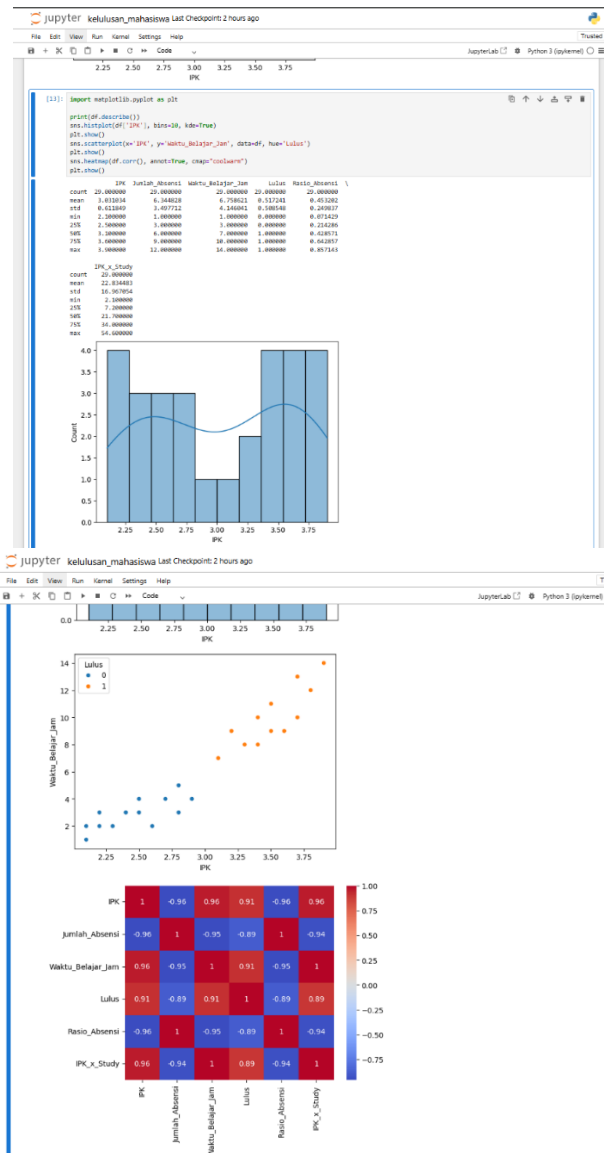
3. Cleaning

Dataset diperiksa untuk missing value menggunakan **df.isnull().sum()** dan tidak ditemukan nilai yang hilang. Data duplikat dihapus dengan **df.drop_duplicates()**. Selain itu, outlier pada kolom IPK diidentifikasi menggunakan boxplot dari Seaborn (**sns.boxplot**).



4. EDA (Exploratory Data Analysis)

Analisis deskriptif dilakukan menggunakan **df.describe()** untuk melihat ringkasan statistik setiap kolom. Distribusi IPK divisualisasikan dengan histogram (**sns.histplot**), hubungan antara IPK dan Waktu_Belajar_Jam divisualisasikan menggunakan scatterplot (**sns.scatterplot**) dengan warna berdasarkan label Lulus. Korelasi antar fitur ditampilkan menggunakan heatmap (**sns.heatmap**) untuk memahami hubungan antar variable



5. Feature Engineering

Dua fitur turunan dibuat untuk meningkatkan sinyal prediktif: **Rasio_Absensi**, yang merupakan perbandingan jumlah absensi terhadap total perkuliahan, dan **IPK_x_Study**, hasil perkalian antara IPK dan waktu belajar. Dataset hasil feature engineering disimpan ke file **processed_kelulusan.csv**.

```

processed_kelulusan.csv
1  IPK,Jumlah_Absensi,Waktu_Belajar_Jam,Lulus,Rasio_Absensi,IPK_x_Study
2  3.8,2,12,1,0.14285714285714285,45.599999999999994
3  2.5,10,3,0,0.7142857142857143,7.5
4  3.4,4,8,1,0.2857142857142857,27.2
5  2.1,11,2,0,0.7857142857142857,4.2
6  3.9,1,14,1,0.07142857142857142,54.6
7  2.7,7,4,0,0.5,10.8
8  3.2,5,9,1,0.35714285714285715,28.8
9  2.4,9,3,0,0.6428571428571429,7.199999999999999
10 3.7,3,10,1,0.21428571428571427,37.0
11 2.8,6,5,0,0.42857142857142855,14.0
12 3.5,2,11,1,0.14285714285714285,38.5
13 2.3,12,2,0,0.8571428571428571,4.6
14 3.6,4,9,1,0.2857142857142857,32.4
15 2.9,8,4,0,0.5714285714285714,11.6
16 3.1,5,7,1,0.35714285714285715,21.7
17 2.2,11,3,0,0.7857142857142857,6.6000000000000005
18 3.3,6,8,1,0.42857142857142855,26.4
19 2.6,9,2,0,0.6428571428571429,5.2
20 3.8,3,12,1,0.21428571428571427,45.599999999999994
21 2.5,10,4,0,0.7142857142857143,10.0
22 3.7,2,13,1,0.14285714285714285,48.1
23 2.4,8,3,0,0.5714285714285714,7.199999999999999
24 3.6,5,9,1,0.35714285714285715,32.4
25 3.4,3,10,1,0.21428571428571427,34.0
26 2.1,12,1,0,0.8571428571428571,2.1
27 3.9,2,14,1,0.14285714285714285,54.6
28 2.8,9,3,0,0.6428571428571429,8.399999999999999
29 3.5,4,9,1,0.2857142857142857,31.5
30 2.2,11,2,0,0.7857142857142857,4.4
31

```

6. Splitting Dataset

Dataset dibagi menjadi fitur (**X**) dan target (**y**), kemudian dilakukan stratified split untuk menjaga proporsi label Lulus. Hasil pembagian menunjukkan **Train** berjumlah 20 baris, **Validation** 4 baris, dan **Test** 5 baris,

```

[8]: from sklearn.model_selection import train_test_split

X = df.drop('Lulus', axis=1)
y = df['Lulus']

X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=42)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42)

print(X_train.shape, X_val.shape, X_test.shape)

(20, 5) (4, 5) (5, 5)

```

Kesimpulan: Tahap Data Preparation telah dilakukan mulai dari collection, cleaning, eksplorasi data (EDA), feature engineering, hingga splitting dataset. Dataset `kelulusan_mahasiswa.csv` berhasil dimuat, diperiksa, dan dibersihkan dari duplikasi serta missing value. Analisis deskriptif dan visualisasi membantu memahami distribusi, korelasi, dan pola antar fitur. Dua fitur turunan, yaitu `Rasio_Absensi` dan `IPK_x_Study`, berhasil dibuat untuk meningkatkan sinyal prediktif. Dataset akhirnya terbagi menjadi Train, Validation, dan Test dengan proporsi yang tepat, sehingga siap digunakan pada tahap pemodelan Machine Learning. Secara keseluruhan, proses ini memastikan data bersih, terdokumentasi, dan siap untuk evaluasi model yang akurat.