


## 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

 **jupyter** kelulusan\_mahasiswa.csv Last Checkpoint: 31 minutes ago

---

File Edit View Settings Help

---

Delimiter:

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
1	3.8	3	10	1
2	2.5	8	5	0
3	3.4	4	7	1
4	2.1	12	2	0
5	3.9	2	12	1
6	2.8	6	4	0
7	3.2	5	8	1
8	2.7	7	3	0
9	3.6	4	9	1
10	2.3	9	4	0

### 2. 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.

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

```
<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
dtypes: float64(1), int64(3)
memory usage: 452.0 bytes
None
```

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
0	3.8	3	10	1
1	2.5	8	5	0
2	3.4	4	7	1
3	2.1	12	2	0
4	3.9	2	12	1

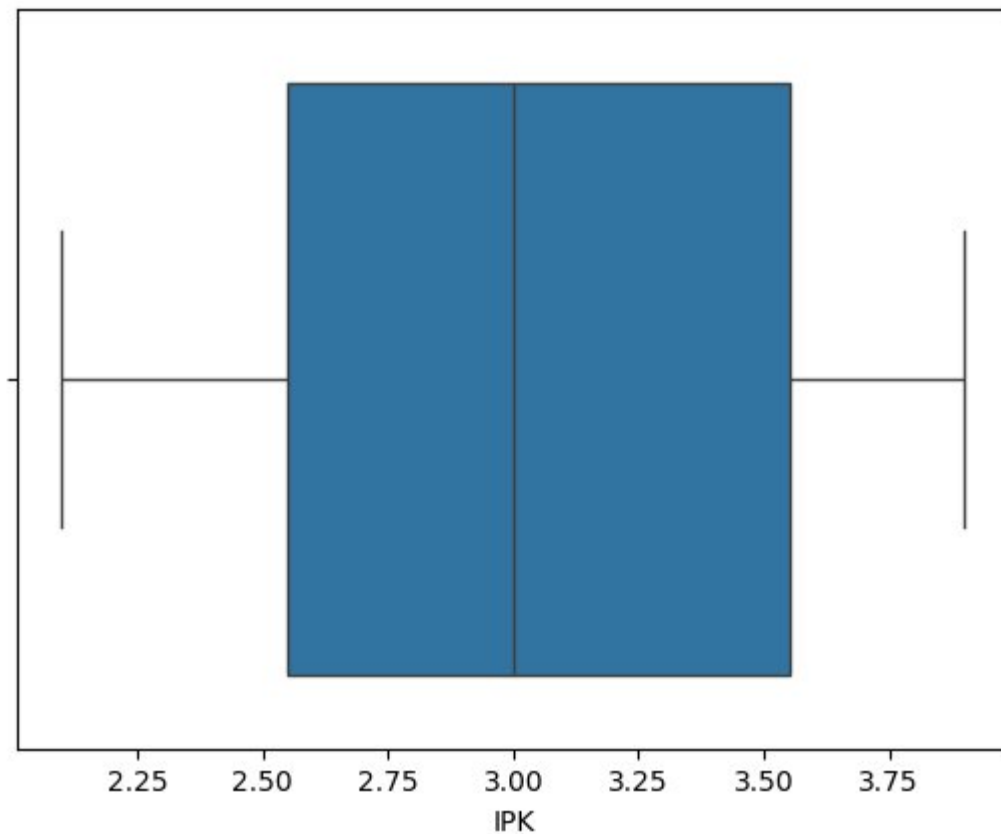
### 3. Langkah 3 — Cleaning (Pembersihan Data)

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

```
print(df.isnull().sum())  
df = df.drop_duplicates()
```

```
import seaborn as sns  
sns.boxplot(x=df['IPK'])
```

```
IPK      0  
Jumlah_Absensi  0  
Waktu_Belajar_Jam  0  
Lulus      0  
dtype: int64  
<Axes: xlabel='IPK'>
```

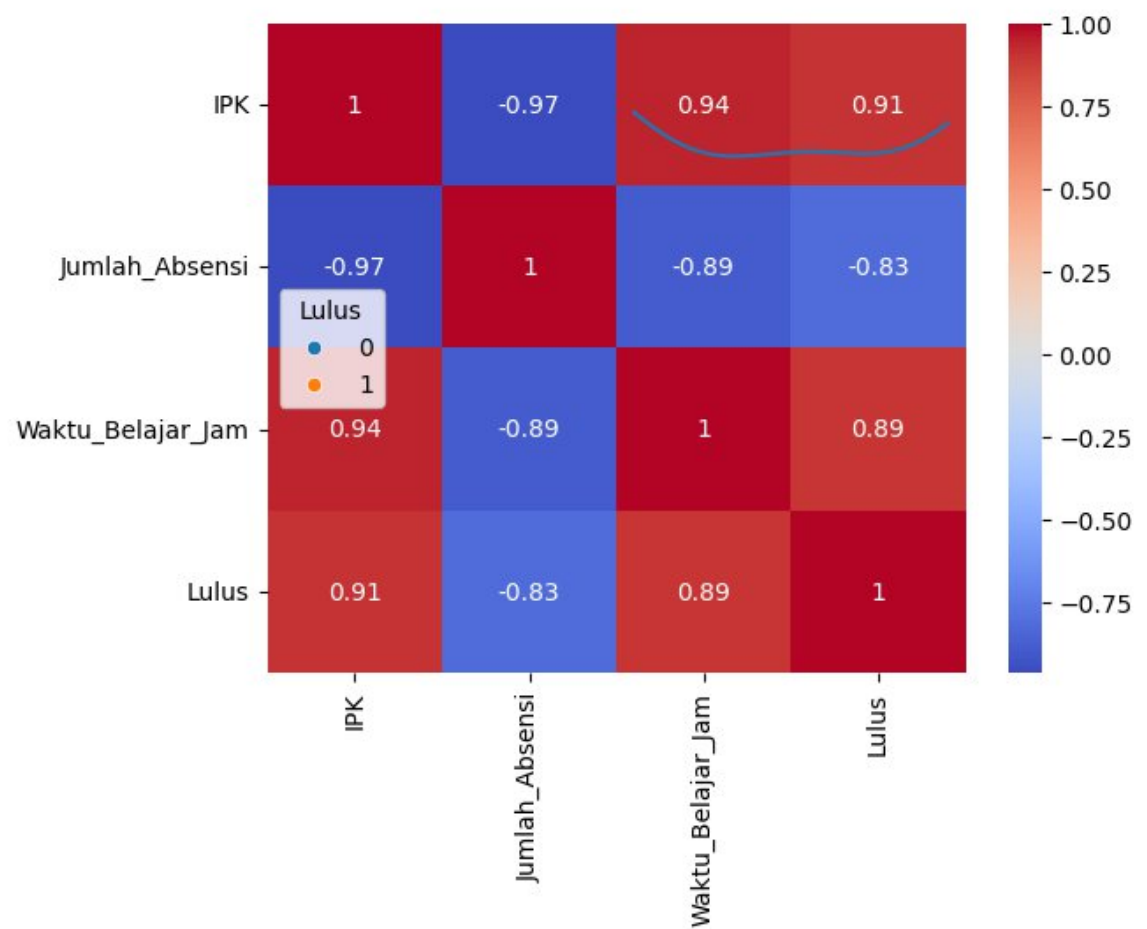


4. Langkah 4 - Exploratory Data Analysis (EDA), untuk memahami data lebih dalam melalui statistik deskriptif dan visualisasi.

```
[3]: print(df.describe())
sns.histplot(df['IPK'], bins=10, kde=True)
sns.scatterplot(x='IPK', y='Waktu_Belajar_Jam', data=df, hue='Lulus')
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
```

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
count	10.000000	10.000000	10.000000	10.000000
mean	3.030000	6.000000	6.400000	0.500000
std	0.639531	3.05505	3.306559	0.527046
min	2.100000	2.000000	2.000000	0.000000
25%	2.550000	4.000000	4.000000	0.000000
50%	3.000000	5.500000	6.000000	0.500000
75%	3.550000	7.750000	8.750000	1.000000
max	3.900000	12.000000	12.000000	1.000000

[3]: <Axes: >



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

```
df['Rasio_Absensi'] = df['Jumlah_Absensi'] / 14
df['IPK_x_Study'] = df['IPK'] * df['Waktu_Belajar_Jam']
df.to_csv("processed_kelulusan.csv", index=False)
```

Jupyter processed\_kelulusan.csv Last Checkpoint: 17 minutes ago

File Edit View Settings Help

Delimiter: ,

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

6. 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 :

```
-----
ValueError                                Traceback (most recent call last)
/tmp/ipython-input-4041406882.py in <cell line: 0>()
      7 X, y, test_size=0.3, stratify=y, random_state=42)
      8
----> 9 X_val, X_test, y_val, y_test = train_test_split(
     10 X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42)
     11

3 frames
/usr/local/lib/python3.12/dist-packages/sklearn/model_selection/_split.py in _iter_indices(self, X, y, groups)
    2316 class_counts = np.bincount(y_indices)
    2317 if np.min(class_counts) < 2:
-> 2318     raise ValueError(
    2319         "The least populated class in y has only 1"
    2320         " member, which is too few. The minimum"
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.
```

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

```
[7]: from sklearn.model_selection import train_test_split

# Memisahkan fitur (X) dan target (y)
X = df.drop('Lulus', axis=1)
y = df['Lulus']

# Pembagian pertama tetap menggunakan stratify (70% train, 30% temp)
X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=42)

# HAPUS stratify=y_temp dari pembagian kedua
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42) # <-- Perubahan di sini

print("--- Ukuran Setiap Set Data (Solusi 1) ---")
print(f"Ukuran X_train (data latih): {X_train.shape}")
print(f"Ukuran X_val (data validasi): {X_val.shape}")
print(f"Ukuran X_test (data uji): {X_test.shape}")

--- Ukuran Setiap Set Data (Solusi 1) ---
Ukuran X_train (data latih): (7, 5)
Ukuran X_val (data validasi): (1, 5)
Ukuran X_test (data uji): (2, 5)
```