M. Fikri Avishena Parinduri

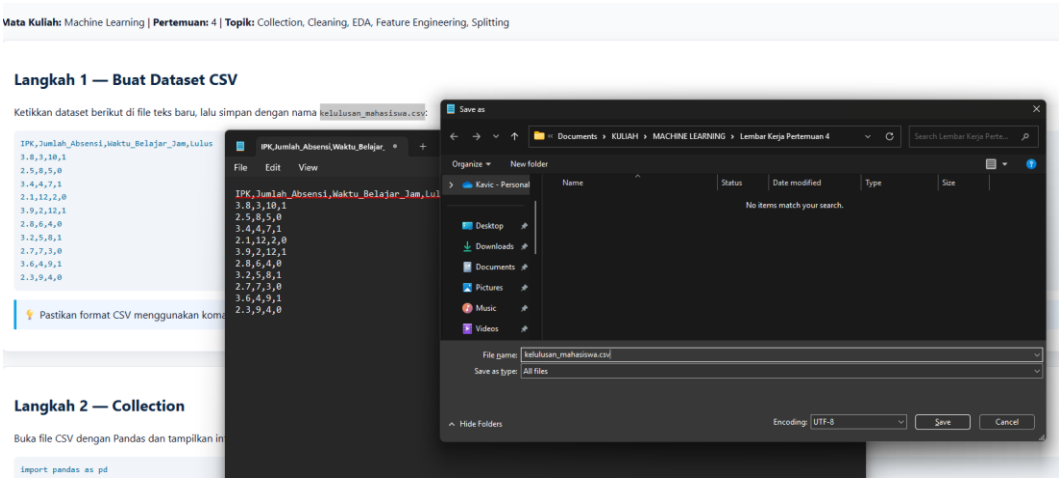
231011401029

05TPLE016

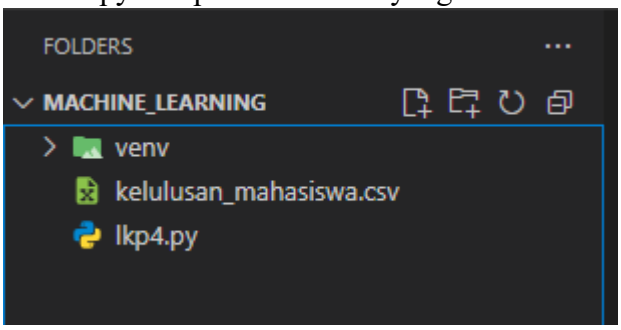
Machine Learning

Lembar Kerja Pertemuan 4

1. Langkah 1 – Buat Dataset CSV
   Simpen dulu di folder mana aja



Lalu copy dan paste ke folder yang sudah ada environment python nya di dalam nya



2. Langkah 2 – Collection
   Masukkan code sebagai berikut

```python
# langkah 2 - collection
import pandas as pd

df = pd.read_csv("kelulusan_mahasiswa.csv")

print(df.info())
print(df.head())
```

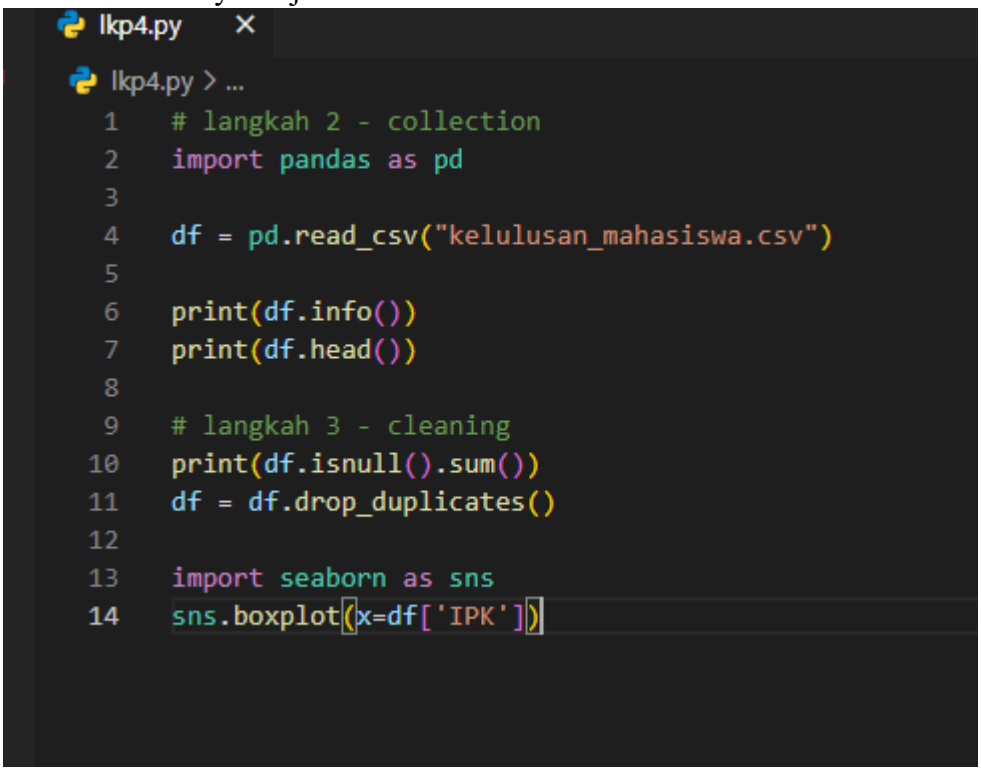
Lalu jalankan,
Hasil:

```
TERMINAL

PS C:\machine_learning> & C:\machine_learning\venv\Scripts\Activate.ps1
(venv) PS C:\machine_learning> python lkp4.py
<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: 448.0 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
(venv) PS C:\machine_learning>
```

Penjelasan:

- pandas.read_csv() membaca file CSV bernama kelulusan_mahasiswa.csv.
- df.info() menampilkan struktur DataFrame (jumlah kolom, tipe data, dan nilai null)
- df.head() menampilkan 5 baris pertama.

3. Langkah 3 – Cleaning
   Masukan kode nya lanjutkan dibawah



Jalankan dan hasil nya:

```
(venv) PS C:\machine_learning> python lkp4.py
<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: 448.0 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
IPK                 0
Jumlah_Absensi      0
Waktu_Belajar_Jam   0
Lulus               0
dtype: int64
(venv) PS C:\machine_learning>
```

Penjelasan:

- Mengecek apakah ada nilai kosong (NaN) per kolom.
- Menghapus baris duplikat, jika ada.
- Membuat boxplot dari kolom IPK untuk mendeteksi outlier.a.

4. Langkah 4 – Exploratory Data Analysis
   Masukan kode nya lanjutkan dibawah nya

```
15
16  # langkah 4 - exploratory data analysis
17  print(df.describe())
18  sns.histplot(df['IPK'], bins=10, kde=True)
19  sns.scatterplot(x='IPK', y='Waktu_Belajar_Jam', data=df, hue='Lulus')
20  sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
```

   Lalu jalankan dan hasil nya:

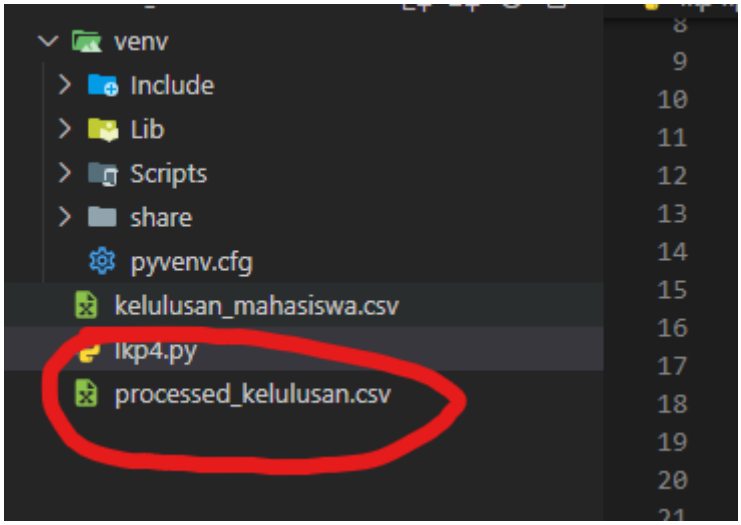
```
(venv) PS C:\machine_learning>

…
 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: 448.0 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
IPK                 0
Jumlah_Absensi      0
Waktu_Belajar_Jam   0
Lulus               0
dtype: int64
             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
(venv) PS C:\machine_learning>
```

   Penjelasan:
   - df.describe() → statistik ringkas (mean, std, min, max, quartiles).
   - sns.histplot() → distribusi nilai IPK.
   - sns.scatterplot() → hubungan antara IPK dan waktu belajar, diwarnai berdasarkan Lulus.
   - sns.heatmap() → menunjukkan korelasi antar variabel.

5. Langkah 5 – Feature Engineering
Masukan kode berikut, lanjutkan dibawah nya

```
21
22  # langkah 5 - feature engineering
23  df['Rasio_Absensi'] = df['Jumlah_Absensi'] / 14
24  df['IPK_x_Study'] = df['IPK'] * df['Waktu_Belajar_Jam']
25  df.to_csv("processed_kelulusan.csv", index=False)
```

Lalu jalankan, akan menghasilkan file “processed_kelulusan.csv” di dalam folder yang sama



Penjelasan:

- Membuat dua fitur baru:
    - Rasio_Absensi → rasio kehadiran mahasiswa terhadap total 14 pertemuan.
    - IPK_x_Study → hasil kali IPK dan waktu belajar (indikator kombinasi prestasi dan usaha).
- Menyimpan dataset baru ke processed_kelulusan.csv tanpa menulis index baris.

6. Langkah 6 – Splitting Dataset
   Masukan kode berikut, lanjutkan dibawahnya

```
df.to_csv('processed_kelulusan.csv', index=False)

# langkah 6 - splitting dataset
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)
```

Penjelasan:

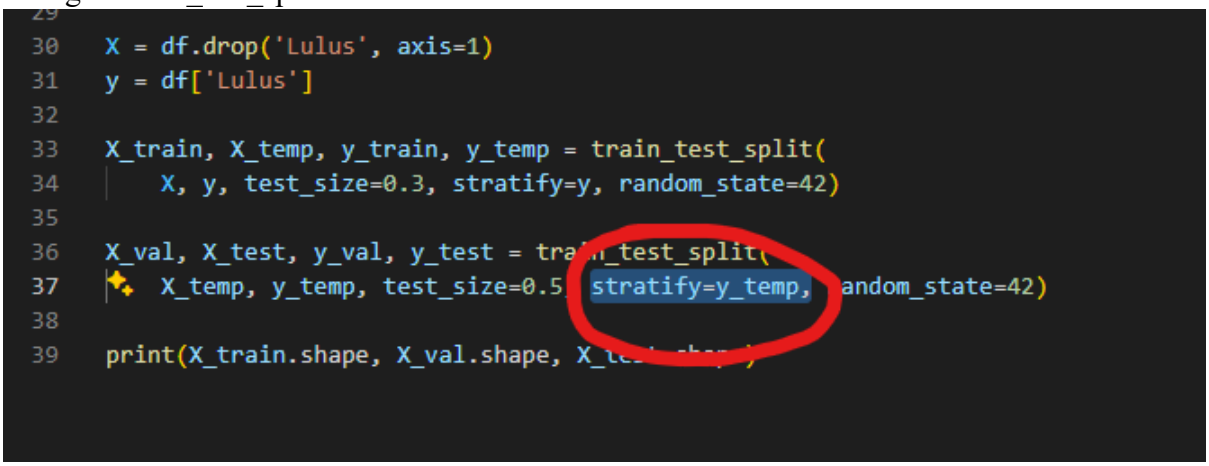
- Memisahkan fitur (X) dan target (y).
- Data dibagi menjadi:
  - 70% training,
  - 15% validation,
  - 15% testing.
- stratify=y menjaga proporsi kelas Lulus (1) dan Tidak Lulus (0) agar seimbang di semua subset.

Ketika dijalankan ini akan error:

```
(venv) PS C:\machine_learning> python lkp4.py
              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
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
Traceback (most recent call last):
  File "C:\machine_learning\lkp4.py", line 36, in <module>
    X_val, X_test, y_val, y_test = train_test_split(
  File "C:\machine_learning\venv\lib\site-packages\sklearn\utils\_param_validation.py", line 218, in wrapper
    return func(*args, **kwargs)
  File "C:\machine_learning\venv\lib\site-packages\sklearn\model_selection\_split.py", line 2940, in train_test_split
    train, test = next(cv.split(X=arrays[0], y=stratify))
  File "C:\machine_learning\venv\lib\site-packages\sklearn\model_selection\_split.py", line 1927, in split
    for train, test in self._iter_indices(X, y, groups):
  File "C:\machine_learning\venv\lib\site-packages\sklearn\model_selection\_split.py", line 2342, in _iter_indices
    raise ValueError(
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.
(venv) PS C:\machine_learning>
```

Error ini dikarenakan:
Kelas dengan populasi paling sedikit di y hanya memiliki 1 anggota, yang terlalu sedikit. Jumlah minimum grup untuk setiap kelas tidak boleh kurang dari 2.

Dibagian train_test_split ke-dua



Masalahnya murni karena jumlah data hanya 10 baris, sehingga saat dibelah dua kali, ada kelas target yang di sisa data hanya 1 baris—dan scikit-learn menolak melakukan stratify kalau suatu kelas < 2.

y_temp ternyata punya kelas yang jumlahnya hanya 1 data saja.
Dengan kata lain, saat pembagian tahap pertama (train_test_split awal) data yang sedikit menyebabkan salah satu kelas (misal Lulus = 1 atau 0) hanya tersisa 1 baris di y_temp, sehingga pembagian kedua gagal — karena stratify butuh minimal 2 data per kelas.

Disini saya akan melakukan penambahan data di file kelulusan_mahasiswa.csv yang awalnya 10 data, menjadi 16 baris data

I9

	A	B	C	D	E	F
1	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus		
2	3.8	3	10	1		
3	2.5	8	5	0		
4	3.4	4	7	1		
5	2.1	12	2	0		
6	3.9	2	12	1		
7	2.8	6	4	0		
8	3.2	5	8	1		
9	2.7	7	3	0		
10	3.6	4	9	1		
11	2.3	9	4	0		
12	3.5	3	11	1		
13	2.4	10	3	0		
14	3	6	7	1		
15	2.2	11	2	0		
16	3.7	5	9	1		
17	2.6	8	4	0		
18						
19						

Lakukan save, lalu jalankan lagi kode nya
Dan outputnya terlihat:

```
                IPK  Jumlah_Absensi  Waktu_Belajar_Jam      Lulu
count  16.000000       16.000000          16.000000  16.00000
mean    2.981250        6.437500           6.250000   0.50000
std     0.610157        3.010399           3.296463   0.51639
min     2.100000        2.000000           2.000000   0.00000
25%     2.475000        4.000000           3.750000   0.00000
50%     2.900000        6.000000           6.000000   0.50000
75%     3.525000        8.250000           9.000000   1.00000
max     3.900000       12.000000          12.000000   1.00000
(11, 5) (2, 5) (3, 5)
(venv) PS C:\machine_learning>
```