Nama : FARHAN ALJIHADILLAH
KELAS : 05TPLE017
MATKUL: MACHINE LEARNING
PERTEMUAN PEKAN 4 : **Data Preparation**

---

1. Yang perlu pertama disiapkan file itu adalah data set dengan nama ‘kelulusan_mahasiswa.csv’ dengan isian sebagai berikut :

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

2. Lalu melakukan penggunaan code ini di maksudkan untuk mengimport panda sebagai pd dalam progam dan membaca file bernama “kelulusan_mahasiswa.csv” dan melakukan hasil print file tersebut

```
[2]
import pandas as pd
df = pd.read_csv("kelulusan_mahasiswa.csv")
print(df.info())
print(df.head())
```

dan hasil running nya menunjukkan data seperti berikut :

```
<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
```

3. Lalu kemudian penggunaan code (df.isnull().sum()) dan df.drop_duplicates untuk melakukan pengecekan apakah ada nilai kosong dalam data dan mengecek data agar tidak ada duplicate dan menampilkan nya dalam visualisasi statistik bentuk kotak ‘boxpolot’ dengan sns.boxplot

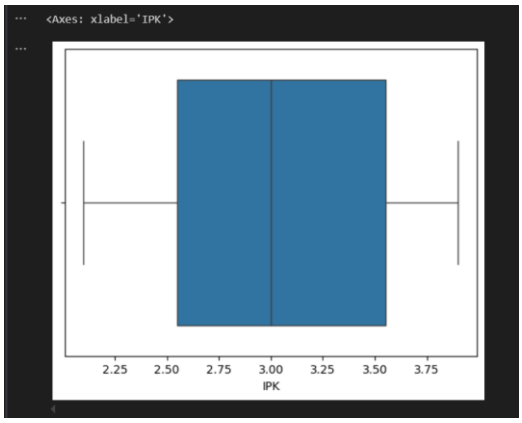
```
import pandas as pd
import seaborn as sns
df = pd.read_csv("kelulusan_mahasiswa.csv")

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


sns.boxplot(x=df['IPK'])
```
[5] ✓ 0.2s

Dan hasil yang ditampilkan dari kode tersebut seperti berikut :

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



Hasil dari (df.isnull().sum()) dan df.drop_duplicates

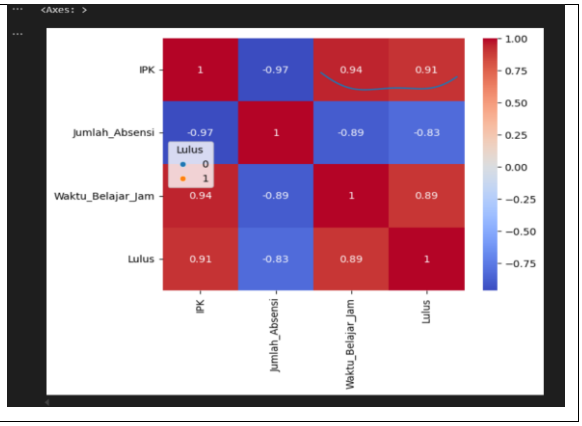
hasil dari sns.boxplot

4. Dan kemudian gunakan code print(df.describe())untuk menunjukkan deskripsi dari dataset yang berisi seperti count,mean,min,max
dan penggunaan heatmap untuk menampilkan visualisasi yang berdasarkan dari penggunaan hisplot dan scatterplot

```
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")
```
[6] ✓ 0.3s

Yang menghasilkan tampilan seperti dibawah ini :

	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



5. Lalu gunakan code berikut ini untuk menghasilkan file dataset baru yang bernama “processed_kelulusan.csv” dengan tambahan berisi rasio absensi,jumlah absensi, ipk x study, ipk dan waktu belajar

```
[7]  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)
     ✓ 0.0s
```

Jika berhasil maka akan ada file baru dalam folder seperti berikut :

6. Dan terakhir gunakan code berikut ini untuk melakukan test split

```
[13]  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=0.4, 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)
      ✓ 0.0s
```

Yang akan menghasilkan hasil data seperti berikut :

```
...  (6, 5) (2, 5) (2, 5)
```