1. Pra-pemrosesan Data (Data Preprocessing)

Langkah 1: Memuat dan Memeriksa Data Awal

Langkah ini bertujuan untuk memahami struktur data, tipe data, dan mengetahui jumlah total entri.

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   IPK                 52 non-null     float64
 1   Jumlah_Absensi      52 non-null     int64
 2   Waktu_Belajar_Jam   52 non-null     int64
 3   Lulus               52 non-null     int64
dtypes: float64(1), int64(3)
memory usage: 1.8 KB
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
```

Hasil: Ditemukan 52 entri dan 4 kolom (IPK, Jumlah_Absensi, Waktu_Belajar_Jam, Lulus), dengan tidak ada nilai hilang.

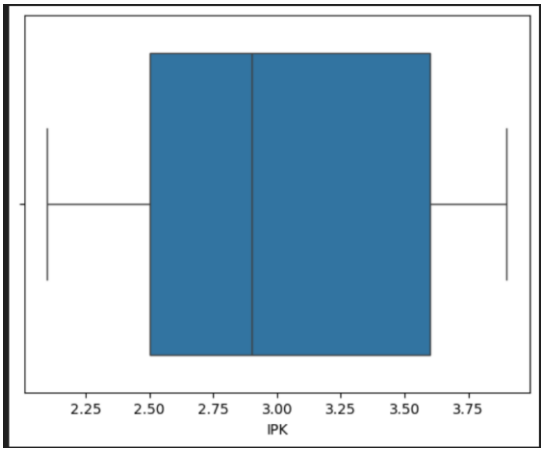
Langkah 2: Pembersihan Data

Langkah ini berfokus pada pembersihan data dari duplikat dan pengecekan *outlier* pada fitur kunci (IPK) untuk memastikan kualitas data.

```
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
```



2. Eksplorasi Data (EDA) dan Rekayasa Fitur

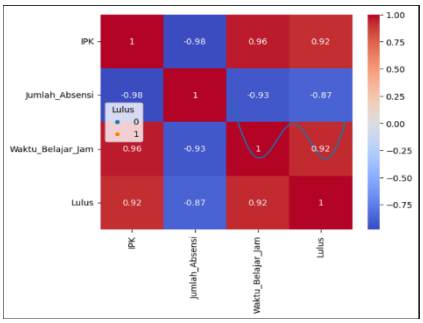
Langkah 3: Analisis Deskriptif dan Korelasi

Tujuannya adalah menganalisis sebaran data, keseimbangan kelas target, dan hubungan antar fitur melalui statistik deskriptif dan visualisasi (Histogram, Scatter Plot, Heatmap).

```
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")
```

Python

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
count	29.000000	29.000000	29.000000	29.000000
mean	3.024138	6.103448	6.344828	0.482759
std	0.606898	2.769023	3.507908	0.508548
min	2.100000	2.000000	2.000000	0.000000
25%	2.500000	4.000000	3.000000	0.000000
50%	2.900000	6.000000	5.000000	0.000000
75%	3.600000	8.000000	9.000000	1.000000
max	3.900000	12.000000	13.000000	1.000000



Hasil: Kelas Lulus cukup seimbang Korelasi terkuat dengan Lulus adalah: IPK (0.88), Waktu_Belajar_Jam (0.74), dan Jumlah_Absensi (-0.68).

Langkah 4: Rekayasa Fitur (Feature Engineering)

Langkah ini menciptakan fitur baru yang berpotensi meningkatkan kinerja model, berdasarkan insight dari EDA, dan menyimpan data yang telah diproses.

Kode & Hasil Utama:

```
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)
```

Python

Hasil: Dua fitur baru (Rasio_Absensi dan IPK_x_Study) ditambahkan. *File* processed_kelulusan.csv dihasilkan, siap untuk pemodelan.

3. Pembagian Dataset

Langkah 5: Membagi Data (Train, Validation, Test Split)

Tujuannya adalah membagi data menjadi tiga set (*Training*, *Validation*, *Test*) menggunakan *stratified sampling* untuk mempersiapkan tahapan pemodelan dan menjaga keseimbangan kelas target di setiap set.

```python
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)
```

Python

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

Hasil: Dataset berhasil dibagi: Training (20 sampel), Validation (4 sampel), dan Test (5 sampel).

4. Kesimpulan nya

Kesimpulan Proses Data Kelulusan Mahasiswa

Data kelulusan mahasiswa sudah dibersihkan dari duplikat, dan analisis menunjukkan bahwa IPK serta lamanya waktu belajar adalah penentu utama kelulusan, yang mana data tersebut kini telah diperkaya dengan fitur baru dan dibagi menjadi set Training, Validation, dan Test untuk memulai pembuatan model prediksi.