Laporan Singkat: Persiapan Data (Pertemuan 4)

Tanggal: 25 Oktober 2025 **Disusun oleh:** [Nama Anda] **Mata Kuliah:** [Mata Kuliah Anda]

---

1. Pendahuluan

Laporan ini mendokumentasikan langkah-langkah yang diambil dalam Pertemuan 4 untuk persiapan dataset prediksi kelulusan mahasiswa. Tujuan dari tahap ini adalah untuk mengumpulkan, membersihkan, menganalisis (melalui EDA), dan melakukan *feature engineering* pada dataset awal untuk menghasilkan dataset bersih (`processed_kelulusan.csv`) yang siap digunakan untuk pemodelan *machine learning* pada pertemuan selanjutnya.

2. Langkah 1 & 2: Pengumpulan dan Pemuatan Data

2.1. Augmentasi Dataset

Dataset asli yang diberikan hanya terdiri dari 10 baris data. Ukuran ini terlalu kecil untuk melakukan *splitting* data (train/validation/test) dan *cross-validation* secara efektif, karena akan menimbulkan error atau model yang tidak stabil.

Untuk mengatasi ini, **40 baris data sintetis** yang realistis dibuat (menggunakan `np.random.seed(42)` untuk reproduktibilitas) dan digabungkan dengan 10 data asli. Hasilnya adalah dataset baru berisi **50 baris** yang disimpan sebagai `kelulusan_mahasiswa.csv`.

2.2. Pemuatan Data

Dataset dimuat ke dalam DataFrame Pandas. Hasil dari `df.info()` menunjukkan struktur data sebagai berikut:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   IPK                 50 non-null     float64
 1   Jumlah_Absensi      50 non-null     int64
 2   Waktu_Belajar_Jam   50 non-null     int64
 3   Lulus               50 non-null     int64
```

dtypes: float64(1), int64(3)
memory usage: 1.7 KB
None

Tinjauan awal data (`df.head()`) menunjukkan format yang sesuai:

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
0	3.047	5	6	1
1	3.667	3	14	1
2	2.169	14	6	0
3	3.842	2	11	1
4	2.100	12	2	0

3. Langkah 3: Pembersihan Data (Cleaning)

1. **Pengecekan Missing Values:** Perintah `df.isnull().sum()` dijalankan dan mengkonfirmasi **tidak ada nilai yang hilang** (missing values) di dalam dataset.
2. **Pengecekan Duplikat:** Perintah `df.drop_duplicates()` dijalankan untuk memastikan tidak ada baris data yang identik. Ukuran data tetap (50, 4), menunjukkan tidak ada duplikat.
3. **Identifikasi Outlier:** Sebuah *boxplot* untuk fitur `IPK` dibuat dan disimpan sebagai `p4_boxplot_ipk.png`. Visualisasi ini membantu mengidentifikasi adanya *outlier* atau pencilan, meskipun pada tahap ini tidak ada data yang dihapus.

4. Langkah 4: Exploratory Data Analysis (EDA)

Analisis data eksploratif dilakukan untuk memahami karakteristik dan hubungan antar variabel.

Statistik Deskriptif: `df.describe()` memberikan gambaran statistik dasar:

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus
count	50.000000	50.000000	50.000000	50.000000
mean	3.042562	6.360000	6.860000	0.500000
std	0.596041	3.784411	3.606385	0.505076
min	2.100000	0.000000	0.000000	0.000000
25%	2.559218	3.250000	4.000000	0.000000
50%	3.123689	5.000000	6.500000	0.500000
75%	3.539367	8.750000	9.750000	1.000000
max	3.931755	15.000000	15.000000	1.000000

1. *Catatan: Rata-rata IPK adalah 3.04 dan data target 'Lulus' seimbang (rata-rata 0.5).*
2. **Visualisasi Data:** Tiga plot utama dibuat dan disimpan:
    - **`p4_hist_ipk.png` (Histogram):** Menunjukkan distribusi IPK, yang tampak relatif normal.

- **`p4_scatter.png` (Scatter Plot):** Memvisualisasikan hubungan antara `IPK` dan `Waktu_Belajar_Jam`. Plot ini (diwarnai berdasarkan `Lulus`) secara visual mengkonfirmasi bahwa mahasiswa dengan IPK lebih tinggi dan Waktu Belajar lebih lama cenderung 'Lulus' (nilai 1).
- **`p4_heatmap.png` (Heatmap Korelasi):** Menunjukkan korelasi linear antar fitur. Terlihat korelasi positif kuat antara `Lulus` dengan `IPK` dan `Waktu_Belajar_Jam`, serta korelasi negatif kuat antara `Lulus` dengan `Jumlah_Absensi`.

5. Langkah 5: Feature Engineering

Untuk memperkaya model, dua fitur baru (fitur turunan) dibuat:

1. **`Rasio_Absensi`**: Dihitung dengan `Jumlah_Absensi / 14` (mengasumsikan 14 total pertemuan). Ini menormalisasi data absensi menjadi rasio.
2. **`IPK_x_Study`**: Dihitung dengan `IPK * Waktu_Belajar_Jam`. Fitur interaksi ini dibuat untuk menangkap efek gabungan dari IPK dan waktu belajar.

Dataset akhir dengan fitur-fitur baru ini kemudian disimpan sebagai **`processed_kelulusan.csv`**.