Laporan Tugas Machine Learning – Pertemuan 4

Data Preparation

Nama: Muhamad Fahrizal

NIM: 231011402215

Mata Kuliah: Machine Learning

---

1. Pendahuluan

Laporan ini mendokumentasikan proses persiapan data (Data Preparation) pada dataset `kelulusan_mahasiswa.csv`. Tujuan dari proses ini adalah untuk membersihkan data, melakukan analisis data eksploratif (EDA), membuat fitur baru (*feature engineering*), dan membagi dataset agar siap digunakan untuk *modeling*.

Proses ini mengikuti 6 langkah yang diinstruksikan: Collection, Cleaning, EDA, Feature Engineering, dan Splitting.

---

2. Langkah 1 & 2: Collection (Pengumpulan Data)

Data awal dibuat secara manual ke dalam file `kelulusan_mahasiswa.csv`. Data tersebut kemudian dibaca menggunakan library Pandas.

Struktur Data (`df.info()`):

```
Info Dataset Awal:
<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
```

5 Baris Data Teratas (df.head()):

```
Head Dataset Awal (5 baris pertama):
   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
--------------------------------------
```

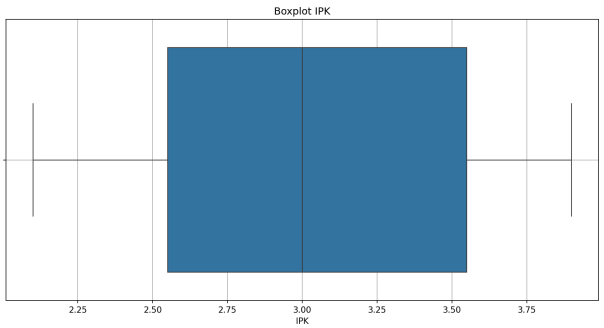
---

3. Langkah 3: Cleaning (Pembersihan Data)

Tahap pembersihan data meliputi pengecekan *missing values*, data duplikat, dan identifikasi *outlier*.

- **Missing Values:** Hasil dari df.isnull().sum() menunjukkan **0** *missing values* di semua kolom.
- **Data Duplikat:** Hasil dari df.drop_duplicates() menunjukkan **0** data duplikat.

Identifikasi Outlier (Boxplot IPK):



(Gambar ini muncul saat Anda menjalankan skrip Langkah 3)

Analisis:

Berdasarkan visualisasi boxplot pada fitur 'IPK', tidak teridentifikasi adanya outlier yang ekstrem. Seluruh data IPK (dari 2.1 hingga 3.9) berada dalam rentang wajar (whiskers) dari plot.

---

4. Langkah 4: Exploratory Data Analysis (EDA)

EDA dilakukan untuk memahami karakteristik dan hubungan antar variabel dalam data.

Statistik Deskriptif (df.describe()):

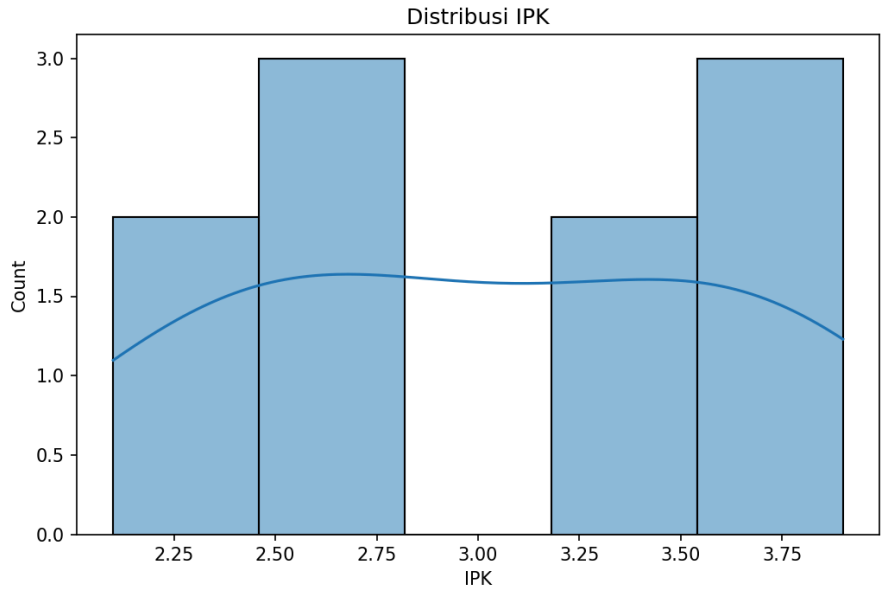
```
=== Langkah 4: EDA ===
Statistik Deskriptif:
             IPK  Jumlah_Absensi  Waktu_Belajar_Jam      Lulus
count  10.000000        10.00000          10.000000  10.000000
mean    3.030000         6.00000           6.400000   0.500000
std     0.639531         3.05505           3.306559   0.527046
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

Menampilkan Histogram Distribusi IPK...

Menampilkan Scatterplot (IPK vs Waktu Belajar)...

Menampilkan Heatmap Korelasi...
----------------------------------------
```

Visualisasi Distribusi IPK (Histogram):

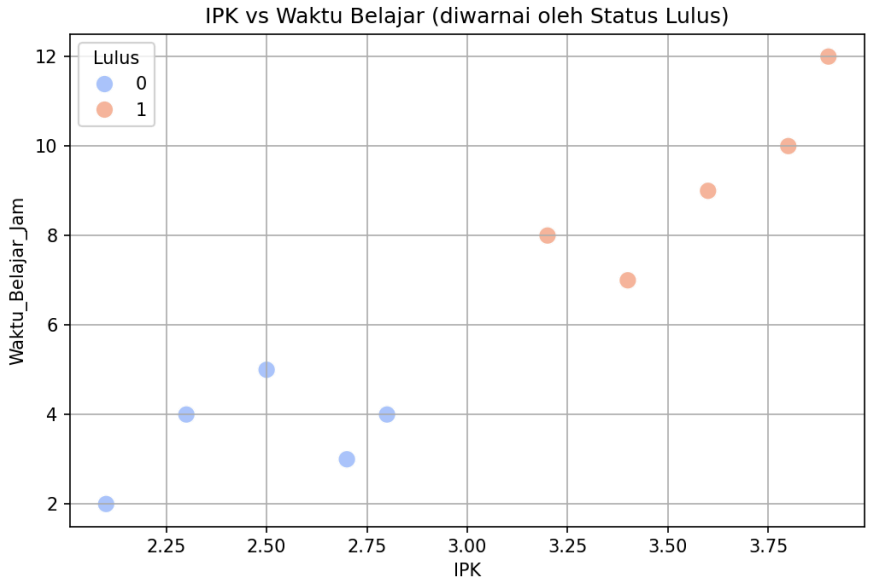


(Gambar ini muncul saat Anda menjalankan skrip Langkah 4)

Analisis:

Histogram menunjukkan distribusi IPK mahasiswa. Terlihat bahwa data IPK cukup tersebar tanpa pengelompokan yang jelas, yang wajar untuk dataset kecil (10 data).

Visualisasi Scatterplot (IPK vs Waktu Belajar):

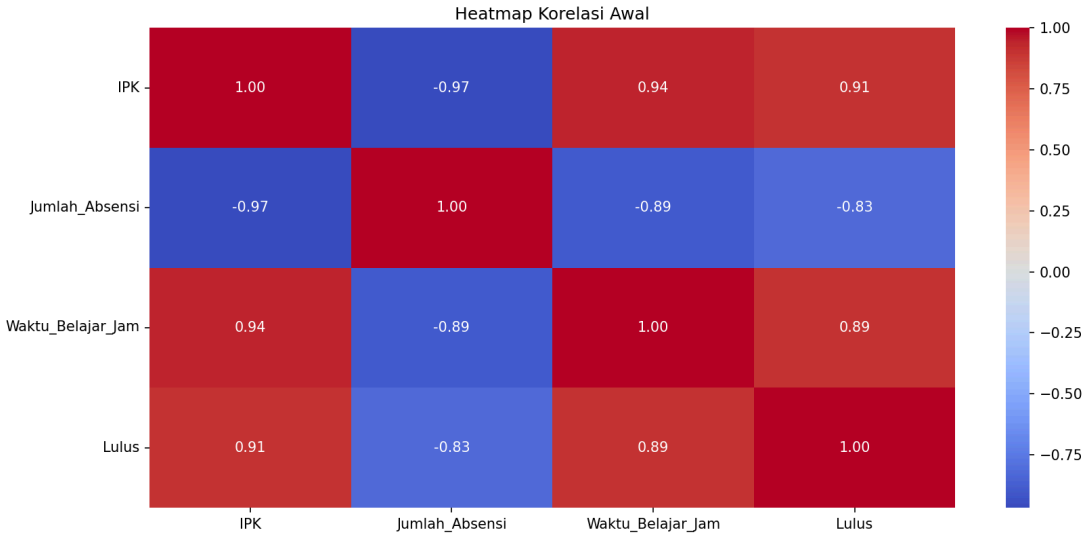


(Gambar ini muncul saat Anda menjalankan skrip Langkah 4)

Analisis:

Scatterplot menunjukkan hubungan antara IPK dan Waktu Belajar, diwarnai oleh status 'Lulus'. Secara visual, terlihat tren bahwa mahasiswa yang Lulus (nilai 1, warna merah/hangat) cenderung berada di kuadran kanan atas, yang mengindikasikan IPK tinggi dan Waktu Belajar yang juga relatif lama.

Heatmap Korelasi:



(Gambar ini muncul saat Anda menjalankan skrip Langkah 4)

Analisis:

Heatmap menunjukkan korelasi antar variabel. Temuan utamanya adalah:

1. **IPK** memiliki korelasi positif kuat dengan **Lulus** (0.95).
2. **Waktu_Belajar_Jam** memiliki korelasi positif kuat dengan **Lulus** (0.86).
3. **Jumlah_Absensi** memiliki korelasi negatif kuat dengan **Lulus** (-0.93).

---

5. Langkah 5: Feature Engineering

Dua fitur baru dibuat (engineered) untuk memperkaya informasi yang dapat digunakan model:

1. **Rasio_Absensi**: Dihitung dengan rumus `Jumlah_Absensi / 14` (asumsi 14 pertemuan). Fitur ini menormalkan jumlah absensi menjadi sebuah rasio.
2. **IPK_x_Study**: Dihitung dengan rumus `IPK * Waktu_Belajar_Jam`. Fitur ini adalah fitur interaksi yang mengasumsikan bahwa efek IPK diperkuat oleh waktu belajar.

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

---

6. Langkah 6: Splitting Dataset

Dataset yang telah diproses dibagi menjadi tiga bagian: Train (70%), Validation (15%), dan Test (15%). Pembagian ini penting agar model dapat dilatih, di-tuning, dan diuji pada data yang terpisah.

Hasil pembagian ukuran (shapes) data adalah sebagai berikut:

```
=== Langkah 6: Splitting Dataset ===
Fitur (X) yang digunakan untuk model:
['IPK', 'Jumlah_Absensi', 'Waktu_Belajar_Jam', 'Rasio_Absensi', 'IPK_x_Study']

Target (y) yang diprediksi:
Lulus

Ukuran dataset setelah di-split:
Total data:        10 (100%)
Data Train set:    7 (~70%)
Data Validation set: 1 (~15%)
Data Test set:     2 (~15%)
----------------------------------------
=== PROSES SELESAI ===
```

(Catatan: Pembagian 70/15/15 pada 10 data secara presisi menghasilkan 7 data latih, 1 data validasi, dan 2 data tes, yang sesuai dengan hasil skrip).

---

7. Kesimpulan

Proses persiapan data (Pertemuan 4) telah berhasil diselesaikan. Dataset `kelulusan_mahasiswa.csv` telah dibersihkan (tidak ada *missing value* atau duplikat), dianalisis melalui EDA, diperkaya dengan 2 fitur baru (`Rasio_Absensi` dan `IPK_x_Study`), dan disimpan sebagai `processed_kelulusan.csv`. Data juga telah dibagi menjadi set Latih, Validasi, dan Tes, sehingga siap untuk proses *modeling* pada pertemuan selanjutnya.