**Data Preparation**

**Langkah-Langkah Proses Analisis dan Persiapan Data Machine Learning**

**Studi Kasus: kelulusan\_mahasiswa.csv**

**Dokumen ini menjelaskan alur kerja standar dalam mempersiapkan dataset mentah hingga siap digunakan untuk melatih model machine learning. Tujuannya adalah untuk mengubah data mentah menjadi data yang bersih, kaya akan fitur, dan terstruktur dengan baik.**

**Langkah 1: Pengumpulan Data (Data Collection)**

* **Tujuan: Memuat data mentah dari sumbernya (misal: file .csv) ke dalam lingkungan kerja (seperti Pandas DataFrame) untuk diolah.**
* **Aktivitas Utama:**
  + **Menggunakan fungsi pd.read\_csv() untuk membaca file.**
  + **Memeriksa informasi dasar dataset dengan df.info() untuk melihat tipe data dan jumlah data non-null.**
  + **Melihat beberapa baris pertama data dengan df.head() untuk memastikan data termuat dengan benar.**
* **Output yang Diharapkan: Sebuah DataFrame yang berisi seluruh data mentah dari file sumber.**

**Langkah 2: Pembersihan Data (Data Cleaning)**

* **Tujuan: Memastikan kualitas dan integritas data dengan mengidentifikasi dan menangani anomali seperti data kosong, duplikat, atau pencilan (*outlier*).**
* **Aktivitas Utama:**
  + **Memeriksa Data Kosong (*Missing Values*): Menggunakan df.isnull().sum() untuk menghitung jumlah data kosong di setiap kolom.**
  + **Menangani Data Duplikat: Menggunakan df.drop\_duplicates() untuk menghapus baris data yang identik.**
  + **Mengidentifikasi Pencilan (*Outliers*): Melakukan visualisasi data numerik menggunakan sns.boxplot() untuk melihat apakah ada nilai yang ekstrem.**
* **Output yang Diharapkan: DataFrame yang bersih, bebas dari data kosong dan duplikat, serta memiliki pemahaman awal tentang adanya pencilan.**

**Langkah 3: Analisis Data Eksploratif (Exploratory Data Analysis - EDA)**

* **Tujuan: Memahami karakteristik, pola, dan hubungan antar variabel dalam dataset melalui statistik deskriptif dan visualisasi.**
* **Aktivitas Utama:**
  + **Statistik Deskriptif: Menghitung statistik utama (rata-rata, median, standar deviasi, dll.) dengan df.describe().**
  + **Analisis Distribusi: Membuat histogram dengan sns.histplot() untuk memahami sebaran data pada kolom numerik (misal: IPK).**
  + **Analisis Hubungan: Membuat *scatterplot* dengan sns.scatterplot() untuk melihat hubungan antara dua variabel (misal: IPK vs Waktu Belajar).**
  + **Analisis Korelasi: Membuat *heatmap* dengan sns.heatmap(df.corr()) untuk mengukur dan memvisualisasikan kekuatan hubungan linear antar semua variabel numerik.**
* **Output yang Diharapkan: Wawasan (*insights*) penting mengenai data yang dapat membantu dalam pemilihan fitur dan model.**

**Langkah 4: Rekayasa Fitur (Feature Engineering)**

* **Tujuan: Menciptakan fitur-fitur baru dari fitur yang sudah ada untuk meningkatkan kekuatan prediktif model.**
* **Aktivitas Utama:**
  + **Membuat kolom baru berdasarkan operasi matematis atau logis dari kolom lain. Contoh:**
    - **df['Rasio\_Absensi'] = df['Jumlah\_Absensi'] / 14**
    - **df['IPK\_x\_Study'] = df['IPK'] \* df['Waktu\_Belajar\_Jam']**
* **Output yang Diharapkan: DataFrame yang "diperkaya" (*enriched*) dengan kolom-kolom fitur baru yang informatif.**

**Langkah 5: Pembagian Dataset (Data Splitting)**

* **Tujuan: Membagi dataset menjadi beberapa bagian (latih, validasi, dan uji) untuk melatih model dan mengevaluasi performanya secara objektif pada data yang belum pernah dilihat sebelumnya.**
* **Aktivitas Utama:**
  + **Memisahkan fitur (variabel independen, X) dari target (variabel dependen, y).**
  + **Menggunakan fungsi train\_test\_split() dari sklearn untuk membagi data.**
  + **Melakukan pembagian bertingkat (*stratified split*) dengan parameter stratify=y untuk memastikan proporsi kelas target tetap sama di setiap bagian (penting untuk masalah klasifikasi).**
* **Output yang Diharapkan: Set data terpisah: X\_train, y\_train, X\_val, y\_val, X\_test, dan y\_test.**

**Langkah 6: Penyimpanan Hasil (Saving the Result)**

* **Tujuan: Menyimpan dataset yang sudah bersih dan kaya fitur ke dalam file baru untuk digunakan pada tahap pemodelan.**
* **Aktivitas Utama:**
  + **Menggunakan fungsi df.to\_csv("processed\_data.csv", index=False) untuk menyimpan DataFrame ke file .csv.**
* **Output yang Diharapkan: Sebuah file .csv baru yang berisi data yang siap pakai.**









