Ahmad Naufal Luthfan Marzuqi

202210370311072

Data dan Informasi Pemodelan B

Laporan Tugas 3

Link Github: https://github.com/Naufallm/Assignment-3---DIP---Data-Cleaning-Integration.git

Link Colab: https://colab.research.google.com/drive/1jQ34oOh-kzScjR0K_efFdDvpYI60rE41?usp=sharing

Link Dataset:

1. https://www.kaggle.com/datasets/ricardobj/electric-vehicle-population
2. https://www.kaggle.com/datasets/shriyashjagtap/esg-and-financial-performance-dataset

Tujuan

Laporan ini merangkum proses analisis dan pembersihan data dataset. Proses meliputi eksplorasi awal, penanganan masalah data (missing values, outliers, noise, duplikat), dan pembersihan inkonsistensi nama untuk mempersiapkan data yang lebih bersih dan konsisten.

Proses dan Hasil

1. Eksplorasi Awal Dataset
   - Dataset kendaraan dianalisis menggunakan df.info(), df.describe(), dan df.head() untuk memahami struktur, tipe data, dan statistik dasar.
   - Jumlah baris dan kolom diperiksa dengan df.shape.
2. Identifikasi dan Penanganan Masalah Data
   - **Missing Values**: Diidentifikasi menggunakan df.isnull().sum(). Kolom seperti County, City, dan CAFV Eligibility diisi dengan nilai default ("Unknown" atau "Eligibility unknown").
   - **Outliers**: Diperiksa pada Model Year menggunakan metode IQR. Data di luar batas (misalnya, tahun sangat tua) dihapus.
   - **Noise**: Data dengan Model Year < 1900 dianggap noise dan dihapus.
   - **Duplicates**: Baris duplikat dihapus menggunakan df.drop_duplicates().

3. Hasil
   - Missing values berhasil ditangani tanpa menghapus data.
   - Outliers dan noise pada Model Year dihapus, memastikan data lebih realistis (tahun 2000-2025).
   - Jumlah duplikat yang dihapus tergantung pada data, tetapi langkah ini memastikan tidak ada baris berulang.

4. Penanganan Inkonsistensi Nama pada Dua Dataset
   Kolom Make (dataset kendaraan) dan CompanyName (dataset perusahaan) dibersihkan.
   Pembersihan:
   - Menghapus karakter khusus (termasuk underscore) menggunakan re.sub(r'[^a-zA-Z0-9]', '', ...).
   - Menghapus spasi berlebih dengan strip().
   - Menstandarisasi ke huruf kapital dengan upper().

Kesimpulan

Proses eksplorasi, penanganan masalah data, dan pembersihan inkonsistensi nama berhasil dilakukan. Dataset kendaraan kini lebih bersih untuk analisis lebih lanjut, meskipun tantangan penggabungan dengan dataset perusahaan tetap ada karena ketidaksesuaian isi data.