Ahmad Naufal Luthfan Marzuqi

202210370311072

DIP B

Laporan Tugas 2

Link repository Github :

https://github.com/Naufallm/Assignment-2-DIP-Pre-Processing-data-.git

Link Kaggle dataset :
https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset

i. Pendahuluan

Laporan ini mendokumentasikan langkah-langkah Pre-processing data yang dilakukan pada Dataset Penilaian Mahasiswa. Pipeline Pre-processing ini dirancang untuk mempersiapkan data untuk analisis.

ii. Gambaran Dataset

Dataset asli berisi 5.000 catatan mahasiswa dengan 23 kolom, termasuk:

- Identifikasi mahasiswa (Student_ID)
- Informasi demografis (First_Name, Last_Name, Email, Gender, Age)
- Detail akademik (Department, Attendance, Scores, Grade)
- Kebiasaan belajar (Study_Hours_per_Week)

iii. Langkah Pre-processing yang dilakukan

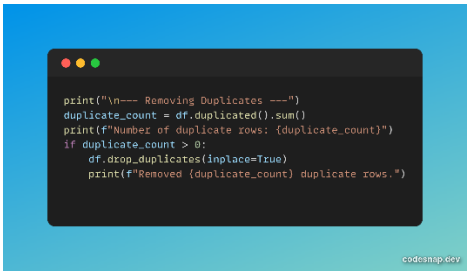
1. Penanganan nilai yang hilang

Nilai yang hilang ditangani menggunakan metode imputasi yang sesuai:

- Kolom numerik: Diimputasi dengan nilai median dari setiap kolom
- Kolom kategorikal: Diimputasi dengan modus (nilai yang paling sering muncul) dari setiap kolom

Pendekatan ini mempertahankan distribusi keseluruhan data sambil memastikan kelengkapan data.

2. **Penghapusan data duplikat**



Data duplikat diidentifikasi dan dihapus dari dataset. Langkah ini memastikan bahwa setiap observasi dalam dataset bersifat unik, mencegah bias dalam analisis selanjutnya.

3. **Cek konsistensi data**

Pemeriksaan konsistensi dilakukan pada variabel kategorikal untuk memastikan representasi yang seragam. Temuan penting meliputi:

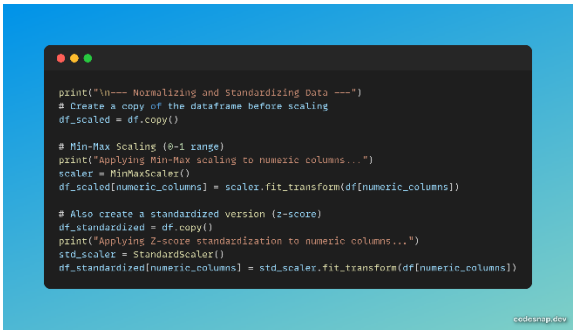
- Nilai Gender telah distandarisasi
- Nama Department diverifikasi untuk konsistensi
- Kategori Grade dikonfirmasi mengikuti format standar

4. Deteksi outlier

Metode statistik digunakan untuk mengidentifikasi outlier dalam kolom numerik:

- Metode Interquartile Range (IQR) diterapkan
- Boxplot dibuat untuk konfirmasi visual
- Outlier diidentifikasi tetapi dipertahankan dalam dataset untuk menghindari kehilangan informasi

5. **Normalisasi data**



Dua transformasi terpisah diterapkan pada fitur numerik:

- **Normalisasi Min-Max**: Menskalakan semua fitur numerik ke rentang [0,1]
- **Standardisasi Z-score**: Mentransformasi fitur untuk memiliki mean=0 dan standar deviasi=1

- **Hasil**
  1. **students_cleaned.csv**: Berisi dataset setelah pembersihan dasar (penanganan nilai hilang, penghapusan duplikat, pemeriksaan konsistensi)
  2. **students_normalized.csv**: Berisi dataset yang sudah dibersihkan dengan normalisasi Min-Max diterapkan
  3. **students_standardized.csv**: Berisi dataset yang sudah dibersihkan dengan standardisasi Z-score diterapkan
- **Perubahan dalam karakteristik data**

karakteristik	Sebelum pre-Processing	Setelah Pre-Processing
Dimensions	5,000 rows × 23 columns	5,000 rows × 23 columns
Missing Values	Terdapat di beberapa kolom	Tidak ada
Duplicates	Tidak terdapat duplikat data	Tidak ada Tindakan penghapusan
Scale	Skala campuran	Dinormalisasi [0,1]

- **Kesimpulan**

  Langkah-langkah pre-processing telah berhasil mengubah Dataset Penilaian Mahasiswa mentah menjadi format yang bersih, konsisten, dan terskala. Ketiga dataset yang telah diproses tersedia untuk berbagai jenis analisis:

  1. **students_cleaned.csv** untuk analisis data eksplorasi yang membutuhkan skala asli
  2. **students_normalized.csv** untuk algoritma yang sensitif terhadap rentang variable
  3. **students_standardized.csv** untuk algoritma yang mengasumsikan data terdistribusi normal