LEMBAR KERJA MAHASISWA (LKM)

**NPM :**

**Nama Mahasiswa :**

**Mata Kuliah :** Data Mining dan Bisnis Analitik  
**Jurusan :** Manajemen Informatika  
**Materi :** Data Preprocessing  
**Waktu :** 3 JP

## Tujuan Pembelajaran

Setelah menyelesaikan LKM ini, mahasiswa diharapkan mampu:

* Menjelaskan konsep dasar preprocessing data
* Melakukan data cleaning (mengatasi missing value, duplikasi, dan kesalahan kategori).
* Melakukan data transformation (normalisasi dan encoding).
* Menyimpan data yang sudah dibersihkan ke file baru.

## Petunjuk Pengerjaan

1. Baca dan pahami setiap instruksi.
2. Kerjakan menggunakan Google Colab atau Jupyter Notebook.
3. Simpan hasil preprocessing ke file baru.
4. Jawab pertanyaan refleksi di akhir.
5. Setiap kegiatan lakukan *screenshot* / copy kode kedalam LKM ini.
6. Download file di <https://github.com/otopet/belajar-git>
7. Tugas dikirim ke [sulistiyanto@polsri.ac.id](mailto:sulistiyanto@polsri.ac.id) / github dengan subjek **Tugas\_Data Cleaning**

\*\*\*\*\*\*\*\*

**Pendahuluan**

Data sering kotor, pembentuan model sangat bergantung pada kualitas data, "Garbage In, Garbage Out"

#### Data Cleaning

* Menghapus data duplikat
* Menangani missing values
  + **Metode**: Mean, Median, Mode, Drop Rows
* Mengoreksi nilai yang salah

## Langkah-Langkah Pengerjaan

### Upload Dataset

Upload salah satu file ke Google Colab. File dapat didownload (https://github.com/otopet/belajar-git)

Gunakan kode berikut:

from google.colab import files  
uploaded = files.upload()  
import pandas as pd  
df = pd.read\_excel("<nama dataset>")

### Tampilkan Informasi Dataset

Tampilkan 5 baris pertama dataset.  
Tampilkan jumlah missing value dan tipe data setiap kolom.  
Gunakan:

print(df.head())  
print(df.isnull().sum())  
print(df.info())

### Data Cleaning

Lakukan pembersihan data dengan metode yang sesuai dengan atribut yang dikerjakan.

### Simpan Dataset Hasil Preprocessing

Simpan ke file baru:  
df.to\_excel("dataset\_bersih.xlsx", index=False)

## Pertanyaan Refleksi

1. Apa pentingnya mengisi missing value daripada langsung menghapus data yang kosong?
2. Apa risiko jika tidak melakukan normalisasi data?
3. Dari dataset ini, apa insight awal yang bisa kamu simpulkan?

## Penilaian

|  |  |
| --- | --- |
| Komponen Penilaian | Skor Maksimal |
| Data berhasil dibaca dan ditampilkan  Data cleaning sesuai instruksi Data transformation sesuai instruksi File hasil disimpan dengan benar Jawaban refleksi lengkap | 20  30  30  10  10 |