Tugas Kelompok
Data Preparation dari Sumber Open Source
23 Februari 2025

Deskripsi Tugas

Pada tugas ini, setiap kelompok akan bekerja dengan dataset dari sumber open source seperti Kaggle atau Hugging Face. Tujuan utama dari tugas ini adalah untuk memahami proses persiapan data sebelum digunakan dalam analisis atau pelatihan model machine learning.

Setiap kelompok (terdiri dari 10 kelompok) harus memilih satu dataset yang tersedia secara publik dan melakukan tahapan berikut:

1. **Data Description**: Jelaskan dataset yang dipilih, termasuk:
    - Nama dataset dan sumbernya
    - Deskripsi singkat tentang dataset
    - Jumlah data (jumlah sampel, fitur, dan label jika ada)
    - Format data (CSV, JSON, Images, Audio, dll.)
2. **Data Loading**:
    - Jelaskan cara memuat dataset ke dalam lingkungan pemrograman (Python menggunakan Pandas, NumPy, atau library lain yang relevan).
    - Tunjukkan contoh kode untuk memuat data.
    - Jelaskan tantangan dalam memuat data (jika ada) dan bagaimana mengatasinya.
3. **Data Understanding**:
    - Tampilkan statistik dasar dataset (misalnya jumlah missing values, distribusi data, korelasi antar fitur).
    - Gunakan visualisasi sederhana seperti histogram, boxplot, atau scatter plot untuk memberikan wawasan awal tentang data.
    - Jelaskan pola atau insight yang ditemukan dari eksplorasi awal data.
4. **Data Preparation**:
    - Lakukan langkah-langkah preprocessing yang diperlukan seperti:
        - Mengatasi missing values (imputasi, penghapusan, dll.)
        - Encoding categorical variables jika diperlukan
        - Normalisasi atau standardisasi data jika relevan
        - Feature selection atau extraction jika diperlukan
    - Tampilkan contoh kode untuk setiap langkah preprocessing.
    - Jelaskan keputusan yang diambil dalam preprocessing dan alasan di baliknya.

Output yang diharapkan:

- Laporan dalam format PDF (maksimal 10 halaman) yang menjelaskan keempat tahap di atas.
- Kode Python dalam format Jupyter Notebook (.ipynb)
- Dataset yang digunakan dalam tugas.
- Presentasi singkat (maksimal 5 slide) yang menjelaskan hasil utama tugas.

Rubrik Penilaian:

Aspek Penilaian	Bobot	Kriteria Penilaian
Pemilihan Dataset	10%	Relevansi dan kompleksitas dataset yang dipilih.
Data Description	20%	Kelengkapan deskripsi dataset (nama, sumber, jumlah data, format, dll.).
Data Loading	15%	Kemampuan memuat dataset dengan benar dan menjelaskan prosesnya.
Data Understanding	20%	Analisis eksplorasi yang jelas, visualisasi yang informatif, dan insight yang diperoleh.
Data Preparation	25%	Langkah preprocessing yang tepat, alasan yang jelas, dan implementasi kode yang baik.
Laporan & Presentasi	10%	Struktur laporan yang rapi, dokumentasi kode yang jelas, serta penyampaian presentasi yang efektif.

Ketentuan Tambahan:

- Setiap kelompok harus memastikan bahwa dataset yang dipilih tidak terlalu besar (maksimal 1GB) agar dapat dikelola dengan baik dalam lingkungan komputasi terbatas.
- Dilarang menggunakan dataset yang sudah pernah digunakan oleh kelompok lain.
- Tugas dikumpulkan melalui Github repository dengan nama repo
    - **Kelompok_<Nomor Kelompok>_Tugas01_Data_Preparation**
    - Tugas dikumpulkan oleh repository ketua kelompok saja

Batas Waktu Pengumpulan: 2 Maret 2025, Pukul 23:00 WIB