Tugas Proyek Mata Kuliah Sistem Informasi Cerdas

Eksplorasi dan Persiapan Data (Studi Kasus Pilihan Mahasiswa)

Dosen Pengampu Materi: Denny Sihombing
Universitas Katolik Indonesia Atma Jaya

Tanggal Pemberian: 9 April 2025
Batas Waktu Pengumpulan: Tanggal: 9 Mei 2025, Pukul 23:59 WIB

Tujuan Proyek

Tugas ini bertujuan agar mahasiswa dapat secara mandiri menerapkan konsep dan teknik Data Understanding serta Data Preparation pada dataset pilihan mereka. Mahasiswa diharapkan mampu:

1. Memilih dataset yang sesuai dari sumber publik (seperti Kaggle).
2. Melakukan eksplorasi data secara mendalam untuk memahami karakteristik, distribusi, hubungan antar variabel, dan kualitas data.
3. Mengidentifikasi dan menangani berbagai masalah kualitas data (tipe data salah, nilai hilang, duplikat).
4. Melakukan transformasi data yang diperlukan (encoding, scaling) agar siap untuk pemodelan.
5. Mendokumentasikan seluruh proses, temuan, dan justifikasi keputusan secara sistematis.

Pemilihan Dataset

- Mahasiswa **wajib memilih dataset sendiri** dari sumber publik yang kredibel. **Kaggle (https://www.kaggle.com/datasets) sangat direkomendasikan**.
- **Kriteria Dataset:**
  - Pilih dataset yang **tabular** (berbentuk tabel). Hindari dataset gambar, teks murni kompleks, atau time-series kompleks kecuali Anda yakin dapat menanganinya.
  - Sebaiknya memiliki **kombinasi fitur numerik dan kategorikal**.
  - Sebaiknya memiliki **variabel target yang jelas** (untuk konteks klasifikasi atau regresi).
  - Ukuran dataset sebaiknya **manageable** (tidak terlalu kecil atau terlalu besar untuk eksplorasi dan persiapan dasar).
  - **Sangat direkomendasikan** memilih dataset yang **memiliki beberapa nilai hilang (missing values)** untuk praktik penanganan.

- Pada awal laporan/notebook Anda, **wajib** mencantumkan:
  - Nama Dataset yang dipilih.
  - Sumber Dataset (URL lengkap).
  - Deskripsi singkat mengenai dataset dan konteks masalahnya.
  - Identifikasi variabel target.

Tools yang Digunakan

- **Bahasa Pemrograman:** Python 3.x
- **Library Utama:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Lingkungan Kerja:** Jupyter Notebook (`.ipynb`, sangat direkomendasikan) atau Google Colab.

Instruksi Pengerjaan

Kerjakan tugas ini secara individu. Buatlah sebuah Jupyter Notebook (`.ipynb`) yang berisi kode Python, output eksekusi kode, serta penjelasan, analisis, dan interpretasi Anda untuk setiap langkah. Notebook harus terstruktur dengan baik (gunakan Markdown untuk judul dan penjelasan) dan mudah diikuti. **Justifikasi** untuk setiap keputusan penting sangat ditekankan.

(Awal) Informasi Dataset Pilihan

- Cantumkan Nama, Sumber (URL), Deskripsi/Konteks, dan Variabel Target dataset pilihan Anda di bagian paling awal notebook.

Bagian 1: Data Understanding (Eksplorasi Data)

1. **Memuat Data:** Muat dataset, tampilkan head/tail, tampilkan dimensi.
2. **Inspeksi Awal:** Tampilkan `df.info()`, jelaskan temuan. Tampilkan nama kolom dan `df.dtypes`.
3. **Statistik Deskriptif:** Hitung dan interpretasikan `df.describe()` untuk numerik dan kategorikal.
4. **Pemeriksaan Kualitas Data:** Periksa dan laporkan jumlah/persentase nilai hilang per kolom. Periksa dan laporkan jumlah baris duplikat.
5. **Visualisasi Eksplorasi (EDA):**
   - Buat dan interpretasikan histogram fitur numerik relevan.
   - Buat dan interpretasikan countplot fitur kategorikal relevan (batasi kategori jika perlu).

- Buat dan interpretasikan visualisasi hubungan antara target dan minimal 4 fitur lainnya (2 numerik, 2 kategorikal).
- Buat dan interpretasikan heatmap korelasi antar fitur numerik.

Bagian 2: Data Preparation

1. **Penanganan Nilai Hilang:** Pilih, justifikasikan, dan implementasikan strategi penanganan nilai hilang. Verifikasi hasilnya.
2. **Penanganan Tipe Data & Inkonsistensi:** Perbaiki tipe data yang salah atau tangani inkonsistensi data jika ditemukan. Jelaskan prosesnya.
3. **Encoding Variabel Kategorikal:** Pilih, justifikasikan, dan terapkan metode encoding yang sesuai. Diskusikan dampaknya pada dimensi data. Tampilkan hasil (dimensi/head).
4. **Feature Scaling:** Identifikasi fitur numerik yang perlu diskalakan. Pilih, justifikasikan, dan terapkan metode scaling. Tampilkan hasil (head/describe) untuk verifikasi. (Ingat best practice terkait split data).
5. **Pemisahan Data (Train-Test Split):** Pisahkan X dan y. Bagi menjadi train/test set (misal 80:20), gunakan `random_state`, dan gunakan/jelaskan penggunaan `stratify`. Tampilkan dimensi semua set hasil split dan verifikasi proporsi target jika relevan.

Bagian 3: Laporan Singkat dan Kesimpulan

1. **Ringkasan Temuan:** Rangkum temuan penting dari Data Understanding dataset Anda.
2. **Ringkasan Persiapan:** Jelaskan ringkasan langkah Data Preparation dan justifikasi kunci. Sebutkan struktur data akhir.
3. **Refleksi & Tantangan:** Ceritakan tantangan utama dan bagaimana Anda mengatasinya.

Format Pengumpulan

- Submit file Jupyter Notebook (`.ipynb`) yang berisi seluruh kode, output, visualisasi, dan semua penjelasan/interpretasi/justifikasi Anda. Pastikan notebook terdokumentasi dengan baik dan dapat dijalankan ulang (*runnable*).
- **Nama file:** `NIM_NamaLengkap_ProjectSIC_DatasetPilihan.ipynb`

Kriteria Penilaian

- Kesesuaian pemilihan dataset dengan kriteria (**10%**).
- Kelengkapan dan kebenaran implementasi langkah Data Understanding (**25%**).
- Kelengkapan dan kebenaran implementasi langkah Data Preparation (**25%**).
- Kualitas analisis, interpretasi visualisasi, dan kedalaman pemahaman data (**20%**).

- Kejelasan dan kelogisan justifikasi untuk keputusan kunci (missing values, encoding, scaling) (**10%**).
- Struktur, kerapian, dan keterbacaan notebook (kode dan narasi) (**10%**).

Penalti dapat diterapkan untuk keterlambatan pengumpulan.

Catatan Tambahan

- Kirim melalui denny.jean@atmajaya.ac.id
- Kualitas justifikasi dan pemahaman proses sama pentingnya dengan kebenaran kode.
- Jelajahi dataset Anda secara mendalam; jangan hanya meniru langkah dari contoh lain jika tidak relevan.
- Junjung tinggi integritas akademik. Kerjakan tugas ini secara mandiri. Plagiarisme dalam bentuk apapun tidak akan ditoleransi.

– Selamat Mengerjakan –