Laporan Praktikum Minggu 2: Data Wrangling & EDA

Informasi Mahasiswa:

Nama: Muhammad Ikhwan Manshur

NIM: 2411070053

Link W&B Project: https://wandb.ai/xxenonitee-stikomelrahma/titanic-eda-2026?nw=nwuserxxenonitee

1. Pendahuluan

Tujuan praktikum ini adalah memahami proses pembersihan data (*data wrangling*) dan eksplorasi data (EDA) sebelum masuk ke tahap *modeling*. Mahasiswa belajar cara mengambil data dari cloud, mengidentifikasi masalah seperti *missing values*, melakukan *feature engineering*, serta menyimpan snapshot data agar eksperimen dapat direproduksi. Dataset Titanic digunakan karena memiliki kompleksitas yang pas untuk pemula, termasuk data teks, angka, dan banyak nilai yang hilang.

2. Analisis Data Mentah (Inspeksi)

Berdasarkan hasil .info() dan .isnull().sum(), kolom yang memiliki *missing values* adalah:

- **Age:** 177 data kosong.
- **Cabin:** 687 data kosong (lebih dari 70% data).
- **Embarked:** 2 data kosong.

3. Strategi Pembersihan Data

 Imputasi Age: Menggunakan **Median** karena lebih tahan terhadap pencilan (*outliers*) dibandingkan Mean.

 Imputasi Embarked: Menggunakan **Modus** (nilai tersering) karena datanya berbentuk kategori lokasi.

 Penghapusan Cabin: Kolom ini dihapus karena lebih dari 70% datanya kosong. Mengisinya secara paksa akan menciptakan banyak data palsu (*noise*) yang merusak model.

4. Visualisasi & Temuan Utama

 Survival Rate by Gender: Wanita memiliki tingkat keselamatan lebih tinggi. Hal ini sesuai konteks historis "wanita dan anak-anak didahulukan" saat evakuasi.

 Distribusi Umur: Mayoritas penumpang berusia 20–40 tahun, dengan tren keselamatan yang bervariasi di tiap kelompok umur.

5. Feature Engineering

 Fitur FamilySize: Rumus yang digunakan adalah:

$$FamilySize = SibSp + Parch + 1$$

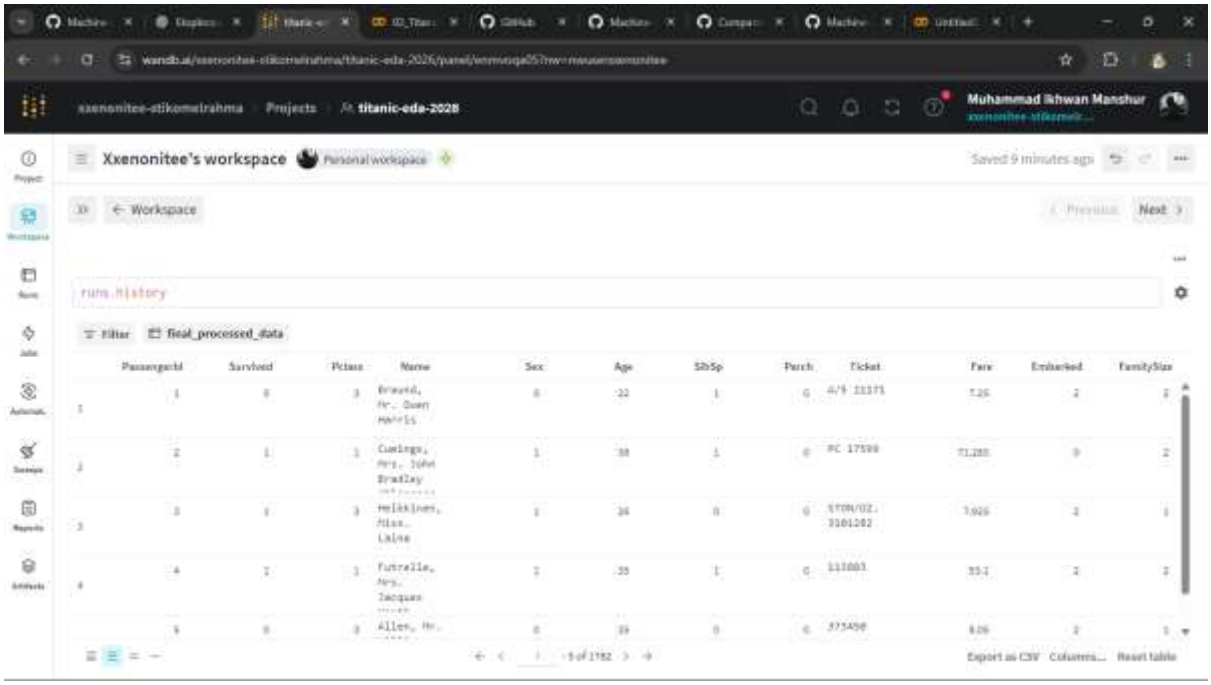
 Analisis: Ukuran keluarga berpengaruh pada peluang selamat. Penumpang dengan keluarga kecil (2-4 orang) cenderung lebih selamat dibanding yang bepergian sendiri atau keluarga sangat besar.

 Encoding: Data teks seperti 'Sex' diubah menjadi angka (0/1) karena algoritma Machine Learning adalah kalkulator matematis yang hanya bisa memproses input numerik.

Encoding: Mengapa kita perlu mengubah data teks (seperti 'Sex') menjadi angka (0/1) sebelum masuk ke model Machine Learning?

6. Integrasi Weights & Biases (W&B)

Di **W&B Tables**, saya bisa membandingkan raw_data dan final_processed_data secara interaktif. Fitur *filtering* membantu saya mencari penumpang spesifik tanpa kode tambahan, dan *sorting* membantu melihat korelasi antara harga tiket (*Fare*) dan keselamatan secara cepat dibandingkan hanya menggunakan .head().



7. Kesimpulan & Refleksi

Tantangan tersulit adalah memutuskan strategi imputasi yang paling tepat agar tidak merusak distribusi asli data. Kesimpulan utamanya adalah EDA sangat krusial; data yang kotor akan menghasilkan model yang buruk (*Garbage In, Garbage Out*).Instruksi Pengumpulan:

Isi template ini dengan lengkap.

Ekspor menjadi PDF atau simpan sebagai README.md di dalam folder tugas Anda.

Pastikan notebook .ipynb Anda sudah dijalankan hingga selesai (Run All) sebelum dikumpulkan.