📑 Laporan Praktikum Minggu 2: Data Wrangling & EDA

Informasi Mahasiswa:

Nama: Mush'ab Abdurrahman Fathin

NIM: 2411070062

Link W&B Project: https://wandb.ai/mushab-a-fathin-stikomelrahma/titanic-eda-2026

1. Pendahuluan

Jelaskan secara singkat tujuan praktikum minggu ini. Apa itu **Data Wrangling** dan mengapa dataset Titanic digunakan sebagai standar pembelajaran?

1. **Jawaban:**

- Tujuan praktikum minggu adalah Mempelajari :
  Data Loading: Membaca data langsung dari Cloud (URL GitHub).

  Data Cleaning: Teknik imputasi cerdas untuk *Missing Values*.

  Interactive EDA: Menggunakan W&B Tables untuk eksplorasi data secara visual.

  Feature Engineering: Transformasi data mentah menjadi fitur siap latih.
- Data wrangling adalah proses mengubah dan memetakan data mentah (raw data) dari berbagai sumber ke dalam format yang terstruktur, bersih, dan konsisten agar siap digunakan untuk analisis, visualisasi, atau *machine learning*.
- dataset Titanic digunakan karena memiliki kompleksitas yang pas untuk pemula: data teks, angka, dan banyak data yang hilang.

2. Analisis Data Mentah (Inspeksi)

Berdasarkan hasil `.info()` dan `.isnull().sum()`, sebutkan kolom mana saja yang memiliki **missing values** dan berapa jumlahnya?

Kolom A: Age (177 data kosong)

Kolom B: Cabin (687 data kosong)

Catatan: Kolom Embarked juga memiliki 2 data kosong.

3. Strategi Pembersihan Data

Jelaskan alasan teknis di balik keputusan Anda dalam menangani data yang hilang:

Imputasi Age: Mengapa menggunakan Median daripada Mean?

Jawaban: Saya pakai median karena data umur punya beberapa nilai yang jauh (misalnya bayi dan lansia). Kalau pakai mean (rata-rata), nilainya bisa mudah “ketarik” oleh umur yang ekstrem, jadi kurang mewakili kebanyakan penumpang. Median lebih aman untuk data yang sebarannya tidak simetris.

Imputasi Embarked: Mengapa menggunakan Modus?

Jawaban: Embarked itu data kategori (huruf S/C/Q), jadi tidak bisa dihitung rata-ratanya. Karena yang kosong cuma sedikit, saya isi dengan modus (nilai yang paling sering muncul) supaya datanya tetap realistis dan tidak banyak mengubah pola.

Penghapusan Cabin: Mengapa kolom ini lebih baik dihapus daripada diisi?

Jawaban: Cabin kosongnya sangat banyak (lebih dari 70%). Kalau dipaksa diisi, hasilnya malah banyak “tebakan” dan bisa bikin data jadi tidak akurat. Selain itu, nilai Cabin juga unik-unik, jadi susah diimputasi dengan cara sederhana. Karena itu saya pilih drop kolomnya.

4. Visualisasi & Temuan Utama

Sertakan hasil temuan Anda dari visualisasi yang dilakukan:

Survival Rate by Gender: Berdasarkan grafik, gender mana yang memiliki tingkat keselamatan lebih tinggi? Jelaskan hubungannya dengan konteks historis?

Jawaban: Dari grafik, penumpang perempuan (female) punya tingkat keselamatan lebih tinggi dibanding laki-laki (male). Ini masuk akal karena pada kejadian Titanic ada prioritas evakuasi “women and children first”, jadi perempuan lebih dulu mendapat kesempatan naik sekoci.

Distribusi Umur: Jelaskan karakteristik umur penumpang Titanic yang Anda temukan.

Jawaban: Dari histogram umur, mayoritas penumpang berada di usia dewasa (sekitar 20-40 tahun). Tetap ada penumpang anak-anak dan juga beberapa yang sudah tua, jadi sebarannya tidak benar-benar rata. Setelah umur yang kosong diisi median, bentuk distribusinya tetap mirip (puncaknya tetap di umur dewasa).

5. Feature Engineering

Jelaskan proses pembuatan fitur baru dan transformasinya:

Fitur `FamilySize`: Tuliskan rumus yang Anda gunakan dalam LaTeX:

Jawaban (LaTeX): $FamilySize = SibSp + Parch + 1$
Penjelasan singkat: SibSp = jumlah saudara/pasangan, Parch = jumlah orang tua/anak, lalu +1 untuk dirinya sendiri.

Analisis: Apakah ukuran keluarga berpengaruh terhadap peluang selamat? Berikan bukti dari grafik yang Anda buat.

Jawaban: Dari heatmap korelasi yang saya buat, hubungan FamilySize dengan Survived terlihat sangat kecil (nilainya mendekati 0, sekitar 0.02). Artinya, di data ini ukuran keluarga tidak terlihat berpengaruh kuat secara langsung terhadap peluang selamat.

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

Jawaban: Model machine learning biasanya menghitung dengan angka, bukan teks. Jadi ‘male/female’ harus diubah jadi 0/1 supaya bisa diproses oleh model. Dengan begitu model bisa “membaca” perbedaan kategori sebagai fitur numerik.

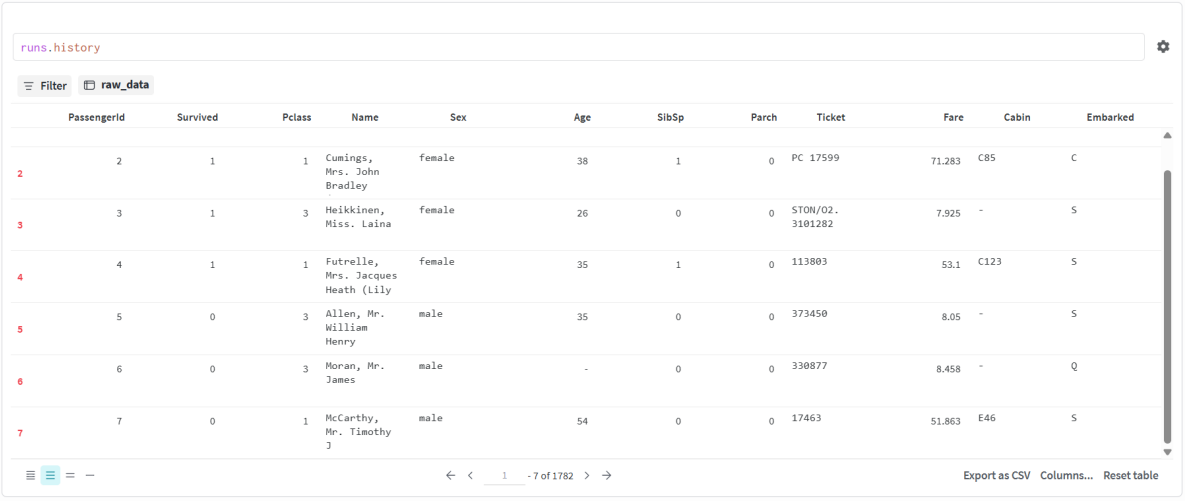
6. Integrasi Weights & Biases (W&B)

Lampirkan **screenshot** atau jelaskan apa saja yang Anda lihat di ****W&B Tables****. Bagaimana fitur **filtering** dan **sorting** di W&B membantu Anda memahami data secara interaktif dibandingkan hanya menggunakan `.head()` di notebook?

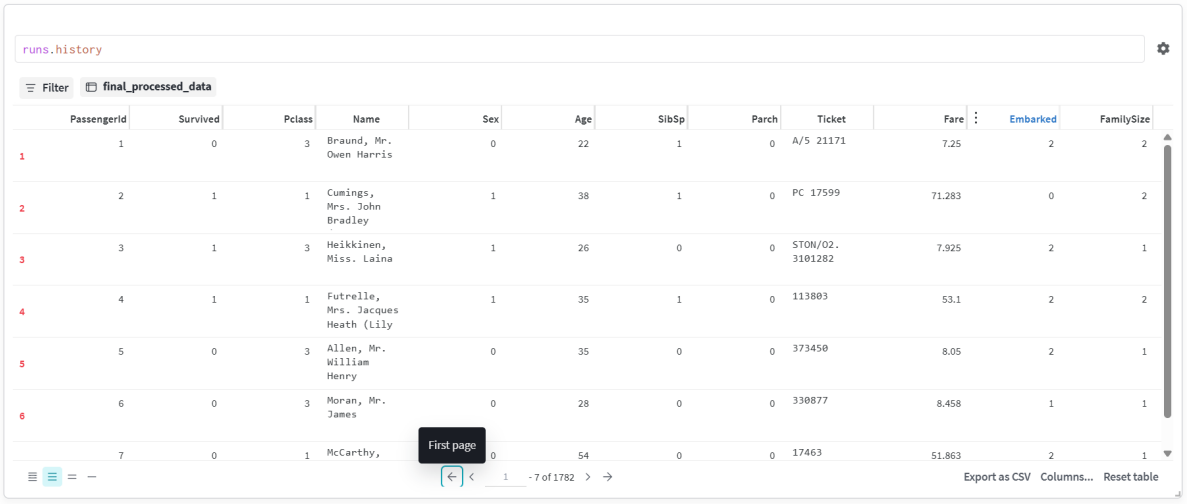
Jawaban: Di W&B Tables saya melihat dua tabel utama:
1) raw_data: data masih mentah (Sex masih teks male/female, masih ada Cabin, dan masih ada data Age yang kosong).
2) final_processed_data: data sudah dibersihkan (Cabin dihapus, Age dan Embarked sudah diisi, Sex sudah jadi 0/1, dan ada fitur baru FamilySize).

Menurut saya, fitur filtering dan sorting sangat membantu karena saya bisa langsung menyaring data (misalnya hanya Survived=1 atau hanya Sex=1) dan mengurutkan (misalnya Fare terbesar) tanpa harus menulis kode berulang-ulang. Kalau pakai .head(), kita cuma lihat baris awal dan tidak bisa interaktif.

Gambar 1. Tampilan W&B Tables - raw_data



Gambar 2. Tampilan W&B Tables - final_processed_data



7. Kesimpulan & Refleksi

Apa tantangan tersulit yang Anda hadapi dalam praktikum ini? Apa kesimpulan utama Anda mengenai pentingnya tahap EDA sebelum masuk ke tahap pemodelan (**Modeling**)?

Jawaban: Tantangan tersulit buat saya adalah menentukan cara menangani data yang hilang (missing values), terutama Cabin yang kosongnya banyak sekali. Saya juga perlu memahami kenapa beberapa kolom harus diubah dulu (encoding) sebelum dipakai model.

Kesimpulan saya: EDA itu penting karena membantu kita “kenalan” dulu dengan data. Kita jadi tahu kolom mana yang kosong, sebaran datanya seperti apa, dan fitur apa yang kelihatan berpengaruh. Kalau EDA dilewati, model bisa belajar dari data yang masih kotor dan hasilnya bisa menyesatkan.