📑 Laporan Praktikum Minggu 2: Data Wrangling & EDA

Informasi Mahasiswa:

- **Nama:** Nisa Agustin
- **NIM:** 2411070040
- **Link W&B Project:** titanic-eda-2026 Workspace – Weights & Biases

---

1. Pendahuluan

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

Jawaban: Praktikum minggu ini bertujuan untuk memahami tahapan awal dalam Machine Learning, khususnya proses Data Wrangling dan eksplorasi data sebelum dilakukan pemodelan. Data Wrangling adalah proses membersihkan, mengolah, dan mentransformasikan data mentah agar menjadi terstruktur dan siap dianalisis, seperti menangani missing values, mengubah tipe data. Dataset Titanic digunakan sebagai standar pembelajaran karena memiliki struktur yang sederhana, mengandung data numerik dan kategorikal, terdapat nilai yang hilang sehingga cocok untuk latihan pembersihan data

---

2. Analisis Data Mentah (Inspeksi)

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

- **Kolom Age :** 177
- **Kolom Cabin :** 687
- **Kolom Embarked :** 2

---

3. Strategi Pembersihan Data

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

- Alasan teknis penanganan data hilang adalah untuk menjaga kualitas data agar model tidak error dan hasil analisis tetap akurat. Median digunakan untuk data numerik agar tidak terpengaruh outlier, modus digunakan untuk data kategorikal, dan kolom dengan terlalu banyak nilai kosong dihapus agar tidak menimbulkan bias.

Imputasi Age: Mengapa menggunakan Median daripada Mean?

- Median dipilih karena lebih tahan terhadap outlier dibandingkan mean, sehingga lebih stabil untuk menggantikan usia yang hilang.

Imputasi Embarked: Mengapa menggunakan Modus?

- Karena Embarked adalah data kategorikal, maka nilai yang paling sering muncul (modus) paling tepat digunakan untuk mengisi data kosong.

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

- Kolom Cabin memiliki terlalu banyak data kosong, sehingga lebih baik dihapus agar tidak menimbulkan bias atau kesalahan dalam analisis.

---

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?
- Berdasarkan grafik, penumpang perempuan memiliki tingkat keselamatan lebih tinggi dibandingkan laki-laki. Hal ini sesuai dengan konteks historis tragedi Titanic, di mana diterapkan prinsip *"women and children first"*, sehingga perempuan dan anak-anak diprioritaskan untuk naik sekoci penyelamat.

Distribusi Umur: Distribusi umur menunjukkan bahwa sebagian besar penumpang berada pada rentang usia dewasa muda hingga dewasa (sekitar 20–40 tahun). Terdapat juga anak-anak dan lansia, namun jumlahnya lebih sedikit dibandingkan kelompok usia produktif.

---

5. Feature Engineering

Jelaskan proses pembuatan fitur baru dan transformasinya:

- Fitur *FamilySize* dibuat untuk mengetahui jumlah anggota keluarga yang ikut dalam perjalanan. Rumus yang digunakan:

$$\text{FamilySize} = \text{SibSp} + \text{Parch} + 1$$

  (+1 ditambahkan untuk menghitung penumpang itu sendiri).

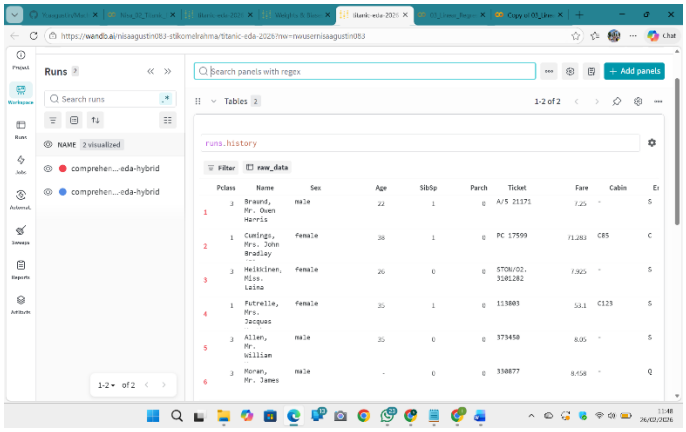
Fitur FamilySize: Tuliskan rumus yang Anda gunakan dalam LaTeX:

- **Analisis:** Berdasarkan grafik, penumpang dengan keluarga kecil (misalnya 1–3 orang) memiliki peluang selamat lebih tinggi dibandingkan yang bepergian sendiri atau dengan keluarga sangat besar. Hal ini menunjukkan bahwa ukuran keluarga berpengaruh terhadap peluang keselamatan.
- **Encoding:** Data teks seperti *Sex* perlu diubah menjadi angka (0/1) karena algoritma Machine Learning hanya dapat memproses data numerik. Encoding mengubah kategori menjadi format angka tanpa menghilangkan makna informasinya.

---

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?



Di W&B Tables saya bisa melihat data secara lengkap dan interaktif, tidak hanya beberapa baris seperti saat menggunakan .head(). Fitur **filtering** membantu menyaring data berdasarkan kondisi tertentu, sedangkan **sorting** memungkinkan mengurutkan data untuk melihat pola dengan lebih jelas. Ini membuat analisis lebih mudah dan mendalam dibandingkan tampilan statis di notebook.

---

7. Kesimpulan & Refleksi

Menurut saya Tantangan tersulit dalam praktikum ini adalah memahami cara menangani data yang hilang dan menentukan strategi yang tepat, seperti kapan harus mengisi data dan kapan harus menghapus kolom. Selain itu, membaca hasil visualisasi dan menarik kesimpulan dari grafik juga membutuhkan ketelitian agar tidak salah interpretasi.

Kesimpulan utama saya adalah tahap EDA sangat penting sebelum masuk ke modeling, karena dari EDA kita bisa memahami kondisi data, menemukan pola, serta mendeteksi masalah seperti missing values atau outlier. Jika langsung membuat model tanpa EDA, hasilnya bisa kurang akurat karena datanya belum benar-benar dipahami dan dibersihkan dengan baik.
