

# K E L O M P O K 5

TITANIC SURVIVAL PREDICTION DATASET

**Sadinal Mufti (2208107010007)**

**M. Agradika Ridhal Eljatin (2208107010020)**

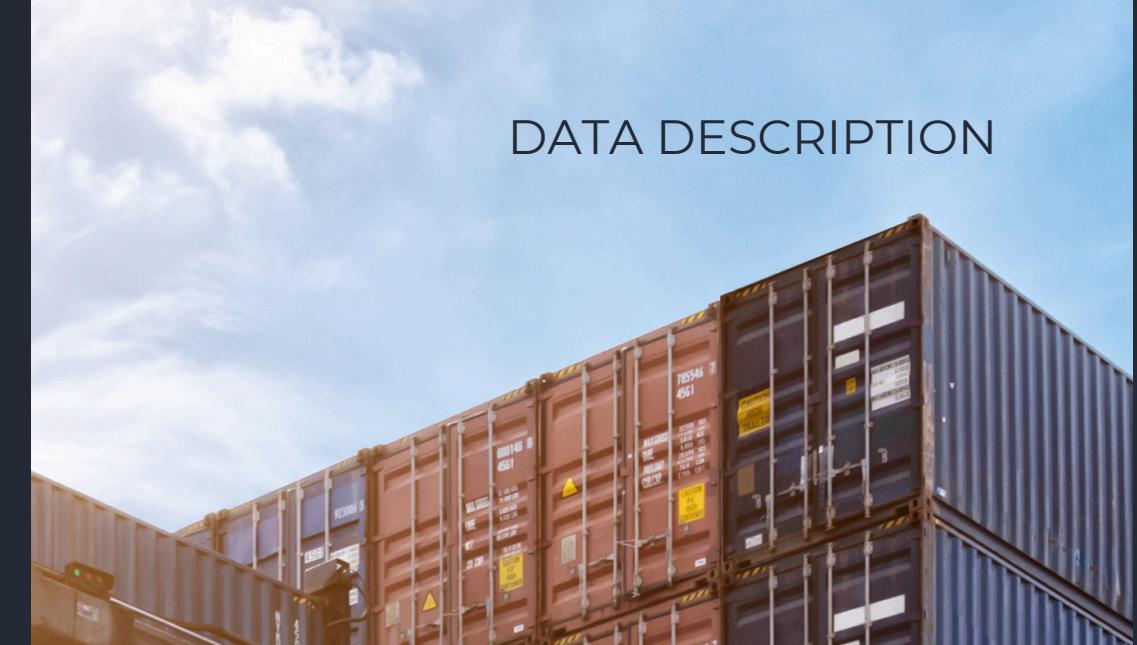
**Jihan Nabilah (2208107010035)**

**Firjatullah Afny Abus (2208107010059)**

**Athar Rayyan Muhammad (2208107010074)**



# Data Description



## DATASET SEPERTI APA YANG DIGUNAKAN?

Dataset yang digunakan dalam tugas ini adalah Titanic Dataset, yang diperoleh dari Kaggle. Dataset ini merupakan dataset klasik yang digunakan untuk menganalisis kelangsungan hidup penumpang kapal Titanic berdasarkan berbagai faktor seperti usia, jenis kelamin, kelas tiket, dan lainnya.

Struktur dataset ini terdiri dari 891 sampel data pelatihan dan 418 sampel data uji, dengan total 12 fitur. Label yang digunakan dalam analisis adalah kolom Survived, di mana nilai 0 menunjukkan bahwa penumpang tidak selamat, dan nilai 1 menunjukkan bahwa penumpang selamat. Dataset ini tersedia dalam format CSV (Comma-Separated Values).

Terdapat kolom Name yang berisi nama penumpang, sedangkan Sex dan Age masing-masing menunjukkan jenis kelamin serta usia penumpang. Kolom SibSp dan Parch menunjukkan jumlah saudara atau pasangan serta jumlah orang tua atau anak yang ikut dalam perjalanan. Selain itu, terdapat kolom Ticket yang berisi nomor tiket, Fare yang menunjukkan tarif tiket, serta Cabin yang menyatakan nomor kabin (dengan banyak data yang hilang). Terakhir, kolom Embarked menunjukkan pelabuhan embarkasi dengan nilai C untuk Cherbourg, Q untuk Queenstown, dan S untuk Southampton.





# Data Loading

- Metode Memuat Data: Dataset Titanic dalam format CSV dimuat menggunakan Python dengan library Pandas melalui pd.read\_csv(). Fungsi head() digunakan untuk melihat lima baris pertama dataset.
- Tantangan: Beberapa kolom memiliki missing values, terutama Age, Cabin, dan Embarked, yang perlu ditangani pada tahap berikutnya.

# Data Understanding

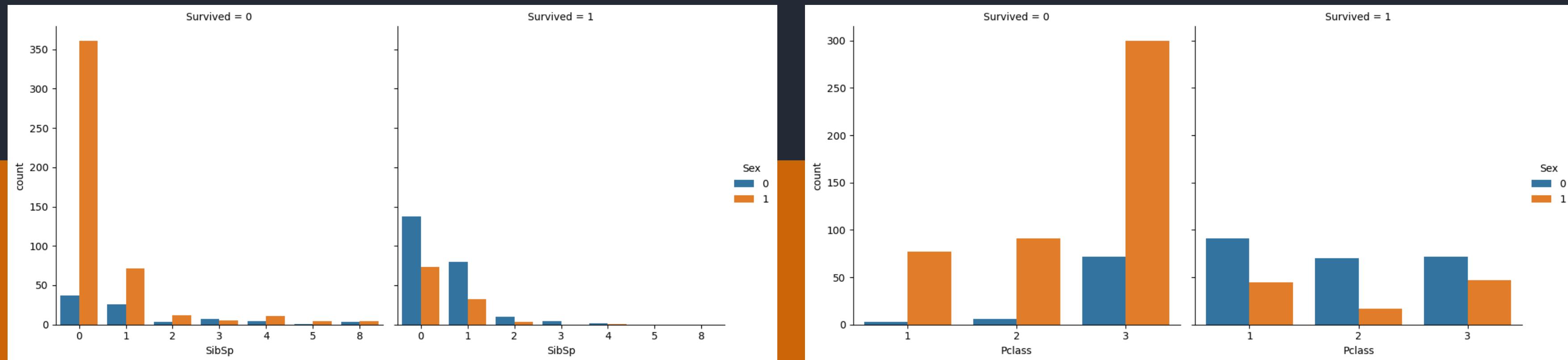
Analisis Statistik & Visualisasi:

- Missing Values: Banyak nilai kosong pada Age, Cabin, dan Embarked.
- Distribusi Gelar Penumpang: Gelar paling umum adalah Mr, Miss, dan Mrs.
- Kelompok Usia: Mayoritas penumpang adalah Adult (18-64 tahun), sementara anak-anak lebih sedikit tetapi memiliki tingkat keselamatan lebih tinggi.
- Korelasi Fitur: Sex, Pclass, dan Fare memiliki pengaruh signifikan terhadap keselamatan penumpang.





# Visualisasi Korelasi



HASIL ANALISIS MENUNJUKKAN BAHWA PENUMPANG DARI KELAS PERTAMA  
MEMILIKI TINGKAT KELANGSUNGAN HIDUP LEBIH TINGGI DIBANDINGKAN KELAS LAINNYA.



# Data Preparation

- Mengatasi missing values: Untuk mengatasi missing values, kolom "Title" diisi berdasarkan "Sex" (male → "Mr", female → "Mrs"). Kolom "Cabin" dihapus karena terlalu banyak missing values, sedangkan "PassengerId" dan "Ticket" dihapus karena tidak informatif.
- Encoding kategori: Kolom kategorikal seperti "Sex" dan "PassengerType" dikonversi ke bentuk numerik.
- Normalisasi dan Standardisasi: Normalisasi dan standardisasi tidak diterapkan karena sebagian besar fitur sudah dalam skala yang sebanding. Namun, jika diperlukan, metode seperti Min-Max Scaling atau Standardization dapat digunakan.
- Feature Selection/Extraction: Atribut "PassengerId", "Ticket", dan "Cabin" dihapus karena tidak signifikan terhadap prediksi "Survived". Dilakukan analisis korelasi serta hubungan antara "SibSp", "Pclass", dan "Survived" untuk memilih fitur yang berpengaruh.
- Justifikasi Keputusan Preprocessing: Missing values pada "Title" diisi berdasarkan jenis kelamin, sementara "Cabin" dihapus karena terlalu banyak missing values. "Sex" dan "PassengerType" dikonversi ke numerik, serta dilakukan feature selection untuk menghapus atribut yang tidak signifikan terhadap "Survived".