



# TITANIC DATASET

● FADHIL HAZZA ISWADINATA | DIGITAL SKIIL FAIR 30.0

---

Agustus 2024



# TITANIC DATASET

- PADA TAHUN 1912, TITANIC, KAPAL PENUMPANG TERBESAR SAAT ITU, TENGGELOM DALAM PELAYARAN PERDANANYA SETELAH MENABRAK GUNUNG ES, MENEWASKAN LEBIH DARI 1.500 DARI 2.224 PENUMPANG DAN AWAK KAPAL. DATASET TITANIC MEMUNGKINKAN KITA UNTUK MENGEKSPLORASI FAKTOR-FAKTOR YANG MEMPENGARUHI KELANGSUNGAN HIDUP PENUMPANG BERDASARKAN INFORMASI SEPERTI USIA, JENIS KELAMIN, KELAS, DAN UKURAN KELUARGA.

ANALISIS INI BERTUJUAN UNTUK MEMAHAMI KARAKTERISTIK PENUMPANG YANG SELAMAT DAN TIDAK SELAMAT, SERTA BAGAIMANA FAKTOR-FAKTOR TERTENTU MEMENGARUHI PELUANG KELANGSUNGAN HIDUP MEREKA. MELALUI EKSPLORASI DAN ANALISIS DATA, KAMI BERHARAP DAPAT MENGIDENTIFIKASI POLA-POLA PENTING YANG MEMBERIKAN WAWASAN LEBIH DALAM MENGENAI TRAGEDI INI.





# DESKRIPSI DATASET TITANIC

DATASET TITANIC YANG TERSEDIA DI KAGGLE ADALAH BAGIAN DARI KOMPETISI "TITANIC: MACHINE LEARNING FROM DISASTER" DAN DIGUNAKAN SECARA LUAS UNTUK MEMPELAJARI TEKNIK PREDIKSI DALAM MACHINE LEARNING. DATASET INI BERISI INFORMASI DEMOGRAFIS DAN DETAIL LAIN TENTANG PENUMPANG TITANIC, YANG BERTUJUAN UNTUK MEMPREDIKSI APAKAH SEORANG PENUMPANG SELAMAT ATAU TIDAK DARI TRAGEDI TERSEBUT.



# VARIABLE DATASET TITANIC

- PASSENGERID: ID UNIK UNTUK SETIAP PENUMPANG.
  - SURVIVED: VARIABEL TARGET YANG MENUNJUKKAN APAKAH PENUMPANG SELAMAT (1) ATAU TIDAK (0).
  - PCLASS: KELAS TIKET PENUMPANG, YANG MEREPRESENTASIKAN STATUS SOSIAL EKONOMI (1 = KELAS SATU, 2 = KELAS DUA, 3 = KELAS TIGA).
  - NAME: NAMA PENUMPANG, YANG JUGA MENCAKUP GELAR (MR., MRS., MISS., DLL.).
  - SEX: JENIS KELAMIN PENUMPANG (MALE = LAKI-LAKI, FEMALE = PEREMPUAN).
  - AGE: USIA PENUMPANG DALAM TAHUN; NILAI INI MEMILIKI BEBERAPA MISSING VALUES YANG PERLU DIIMPUTASI.
  - SIBSP: JUMLAH SAUDARA KANDUNG DAN PASANGAN PENUMPANG YANG IKUT DI KAPAL.
  - PARCH: JUMLAH ORANG TUA DAN ANAK PENUMPANG YANG IKUT DI KAPAL.
  - TICKET: NOMOR TIKET PENUMPANG.
  - FARE: TARIF YANG DIBAYAR UNTUK TIKET, DALAM POUND STERLING.
  - CABIN: NOMOR KABIN PENUMPANG; BANYAK NILAI YANG KOSONG DALAM KOLOM INI.
  - EMBARKED: PELABUHAN TEMPAT PENUMPANG NAIK KE KAPAL (C = CHERBOURG, Q = QUEENSTOWN, S = SOUTHAMPTON).
-



# TUJUAN DATASET TITANIC

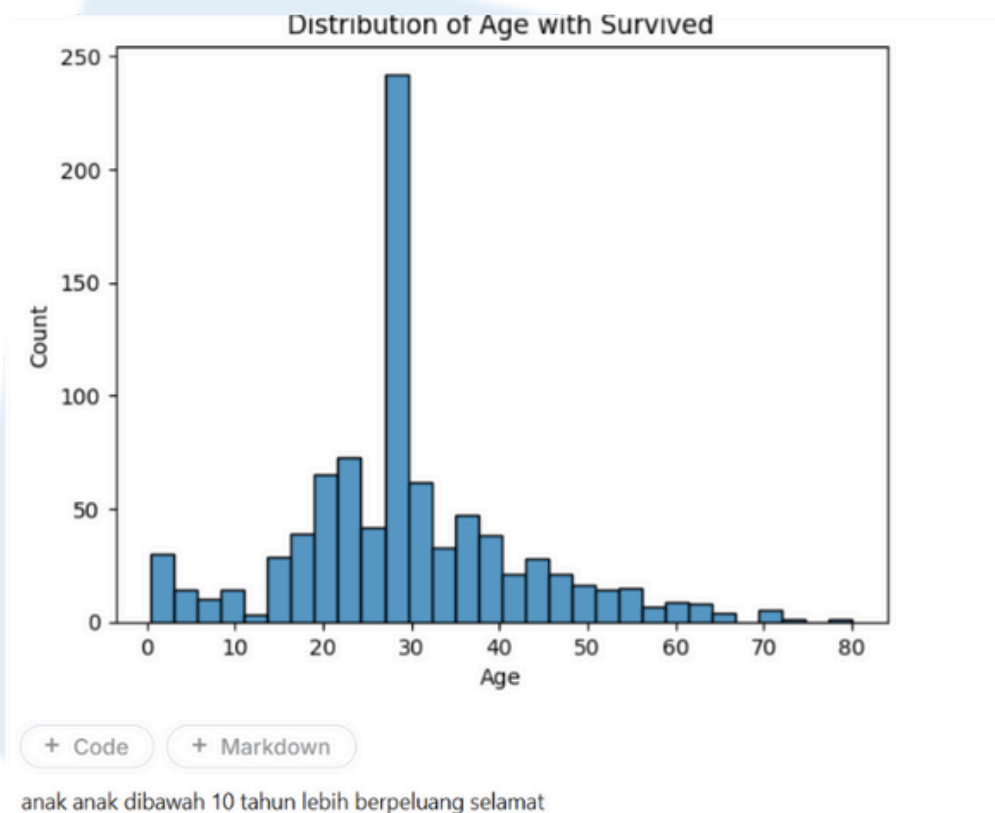
DATASET INI DIGUNAKAN UNTUK MEMPREDIKSI KELANGSUNGAN HIDUP PENUMPANG BERDASARKAN FAKTOR-FAKTOR SEPERTI KELAS SOSIAL, JENIS KELAMIN, USIA, DAN HUBUNGAN KELUARGA. INI ADALAH DATASET KLASIK UNTUK MEMPELAJARI TEKNIK SUPERVISED LEARNING, TERUTAMA DALAM KLASIFIKASI BINER, DAN DIGUNAKAN OLEH BANYAK PRAKTISI DATA UNTUK MELATIH MODEL PREDIKTIF SERTA MEMPRAKTIKKAN TEKNIK-TEKNIK DATA PREPROCESSING, FEATURE ENGINEERING, DAN EVALUASI MODEL.

---

# VISUALISASI DATA

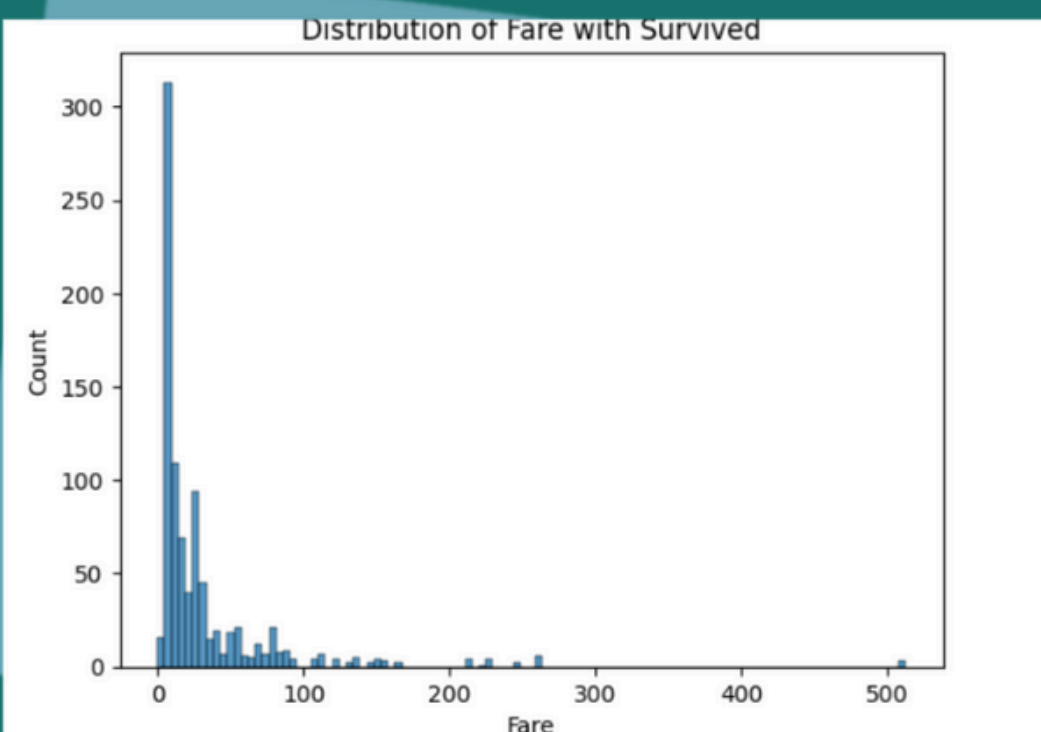
## DATA UMUR

Bisa dilihat Bahwa orang yang berumur dibawah 10 tahun Peluang selamatnya tinggi



## DATA PEMBAYARAN

Dalam data Tersebut dijelaskan bahwa orang yang dengan pembayaran/membayar lebih mahal peluang Selamatnya tinggi

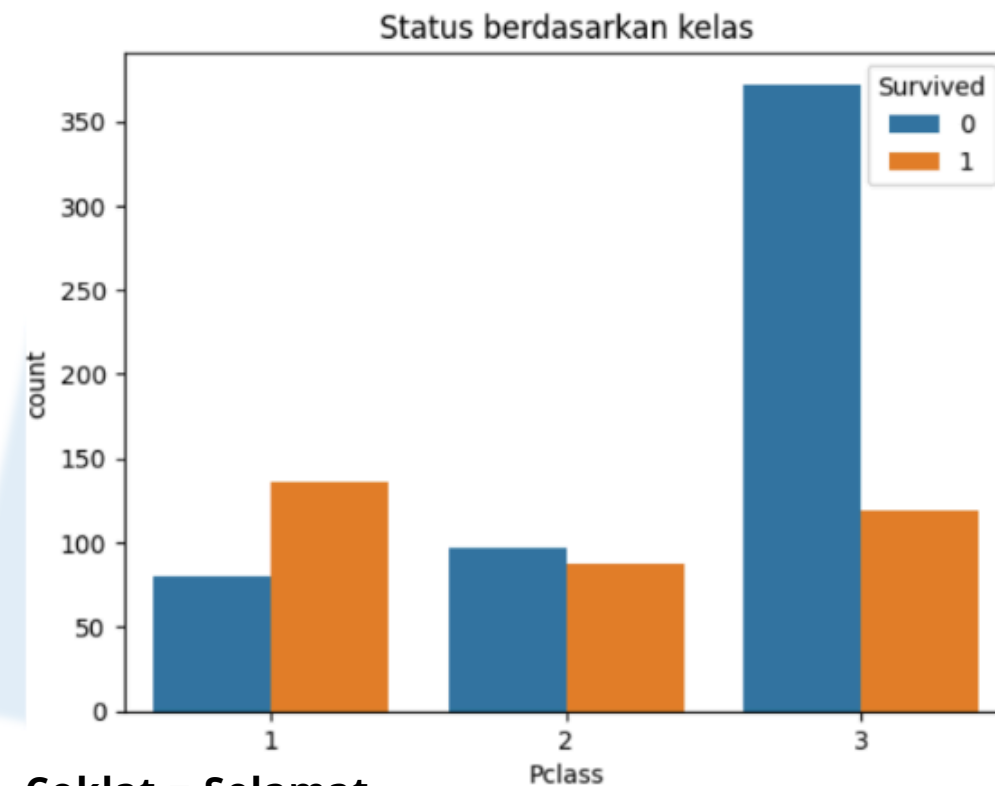


# VISUALISASI DATA

## DATA KELAS/PCLASS

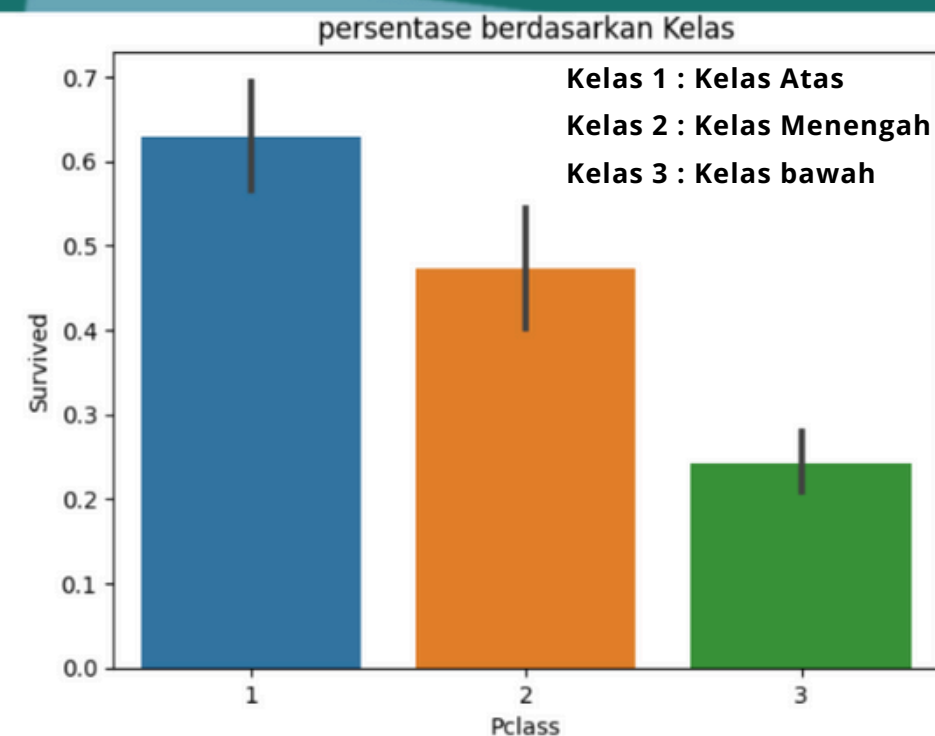
Ini adalah Perbandingan antara Penumpang Kelas 1, 2 & 3.

Terlihat



Coklat = Selamat

Biru = Tidak Selamat



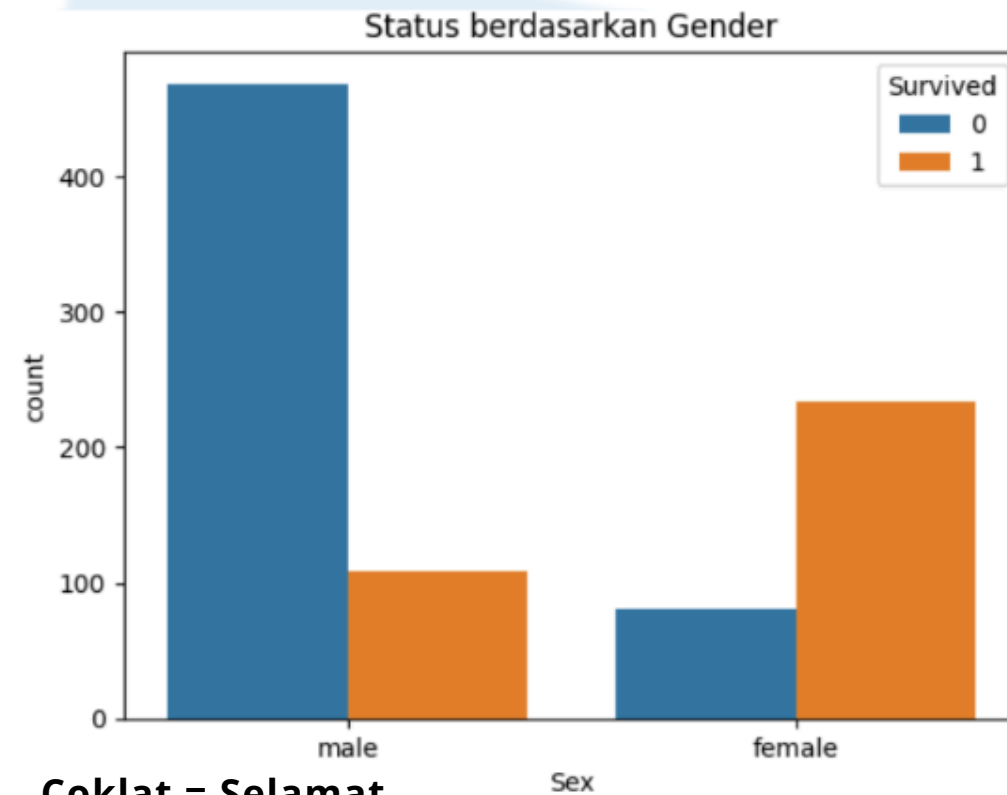
## DATA KELAS/PCLASS

Ini adalah Data Kelas dalam perahu Titanic, Terlihat Kelas 1 peluang selamatnya paling Tinggi dari Kelas 2 & 3

# VISUALISASI DATA

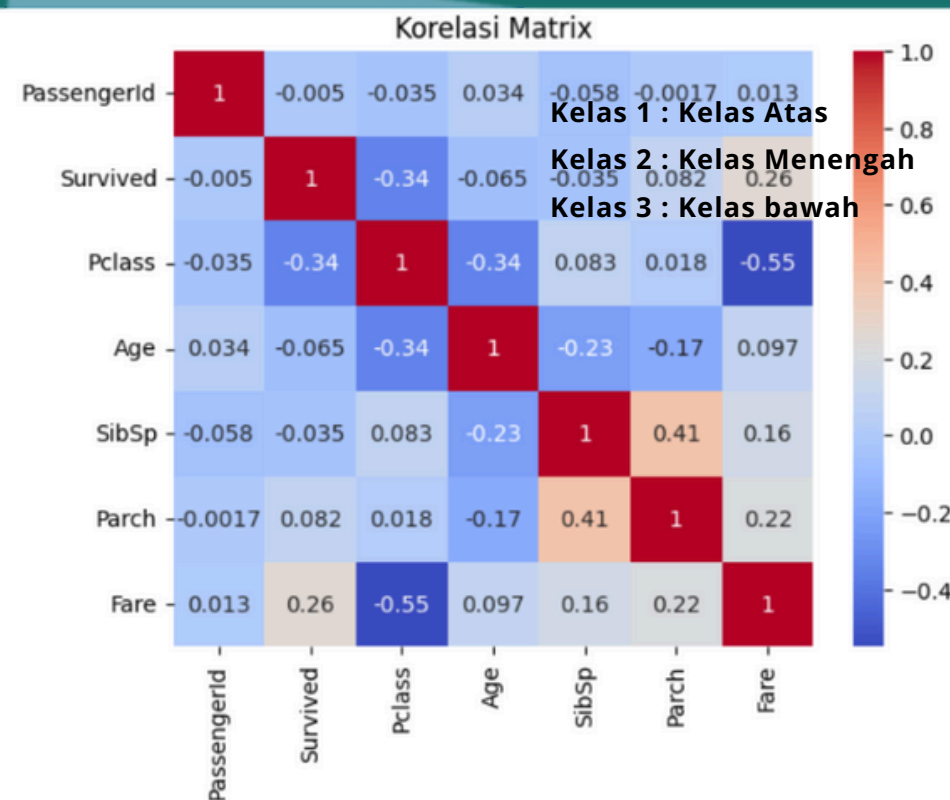
## DATA SEX/GENDER

Ini adalah Perbandingan antara Penumpang Laki Laki (Male) dan Perempuan (Female).  
Terlihat Perempuan Peluang Selamatnya Lebih Tinggi dari pada Perempuan



Coklat = Selamat

Biru = Tidak Selamat



Kelas 1 : Kelas Atas

Kelas 2 : Kelas Menengah

Kelas 3 : Kelas bawah

## KORELASI MATRIX

1. Pclass dan Fare memiliki korelasi Negatif, Semakin rendah Pclass maka semakin tinggi Fare nya
2. fitur yang paling berpengaruh terhadap survived adalah Pclass



[2]:

```
df = pd.read_csv("/kaggle/input/titanic/train.csv")
df
```

[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

# MENGINPUT DATA

```
df = pd.read_csv("/kaggle/input/titanic/train.csv")
```

df

Berfungsi untuk memanggil data, 'pd' artinya pandas dan 'read\_csv' itu untuk membaca atau menampilkan data dari data.

```
df.describe(include='all')
```

[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000	891.000000	891	891.000000	204	889
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	147	3
top	NaN	NaN	NaN	Braund, Mr. Owen Harris	male	NaN	NaN	NaN	347082	NaN	B96 B98	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	4	644
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008	0.381594	NaN	32.204208	NaN	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743	0.806057	NaN	49.693429	NaN	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000	0.000000	NaN	7.910400	NaN	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000	0.000000	NaN	31.000000	NaN	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN	NaN

# MENDESKRIPSIKAN DATA

```
df.describe(include='all')
```

Berfungsi untuk mendeskripsikan data dari sebelumnya yg sudah diinput 'df' artinya Data Frame dan 'describe' itu untuk menampilkan data secara deskriptif.

```
df.isnull().sum()
```

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2
dtype: int64	

## CROSS CHECK VALUE KOSONG

```
df.isnull().sum()
```

Berfungsi untuk mengecek data dari sebelumnya yang sudah diinput apakah ada yg kosong atau tidak.

## Data Processing

### Median

pada Data frame age atau umur yg kosong akan diisi oleh median dari data frame age itu sendiri

```
] : df['Age'] = df['Age'].fillna(df['Age'].median())
```

### Modus

Dimana embarked ini yg kosong akan diisi oleh modus dari data embarked itu sendiri

```
] : df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

# MEMPROSES DATA

```
df['Age'] = df['Age'].fillna(df['Age'].median())
```

Berfungsi jika semisal ada data yg kosong pada data Age akan diisi oleh mediannya Age.

```
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

Berfungsi jika semisal ada data yg kosong pada data Embarked akan diisi oleh modus dari data Embarked itu sendiri.

## Drop

jadi data frame Cabin itu sendiri tidak terlalu berguna untuk analisa data kali ini maka dari itu saya memanggil *drop* untuk menghapus atau menghilangkan dataframe Cabin

```
df = df.drop('Cabin', axis=1)
```

+ Code

+ Markdown

# MEMPROSES DATA

```
df = df.drop('Cabin', axis=1)
```

Berfungsi untuk Menghapus kolom dan baris.

Chek hasil setelah proses diatas menggunakan isnull().sum()

```
df.isnull().sum()
```

```
: PassengerId    0
   Survived      0
   Pclass       0
   Name         0
   Sex          0
   Age          0
   SibSp        0
   Parch        0
   Ticket       0
   Fare         0
   Embarked     0
dtype: int64
```

## MENGECEK HASIL ULANG


**df.isnull().sum()**

Berfungsi untuk mengecek apakah ada data yang kosong lagi atau tidak

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Emba
0	1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	0
2	3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	2
4	5	0	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	2

# FEATURE ENGINEERING

Label Encoder untuk mengubah data kategorik menjadi numerik



0	3	1	22.0	1	0	7.2500	2
1	1	0	38.0	1	0	71.2833	0
2	3	0	26.0	0	0	7.9250	2
3	1	0	35.0	1	0	53.1000	2
4	3	1	35.0	0	0	8.0500	2
...	...	...	...	...	...	...	...
886	2	1	27.0	0	0	13.0000	2
887	1	0	19.0	0	0	30.0000	2
888	3	0	28.0	1	2	23.4500	2
889	1	1	26.0	0	0	30.0000	0
890	3	1	32.0	0	0	7.7500	1

891 rows × 7 columns

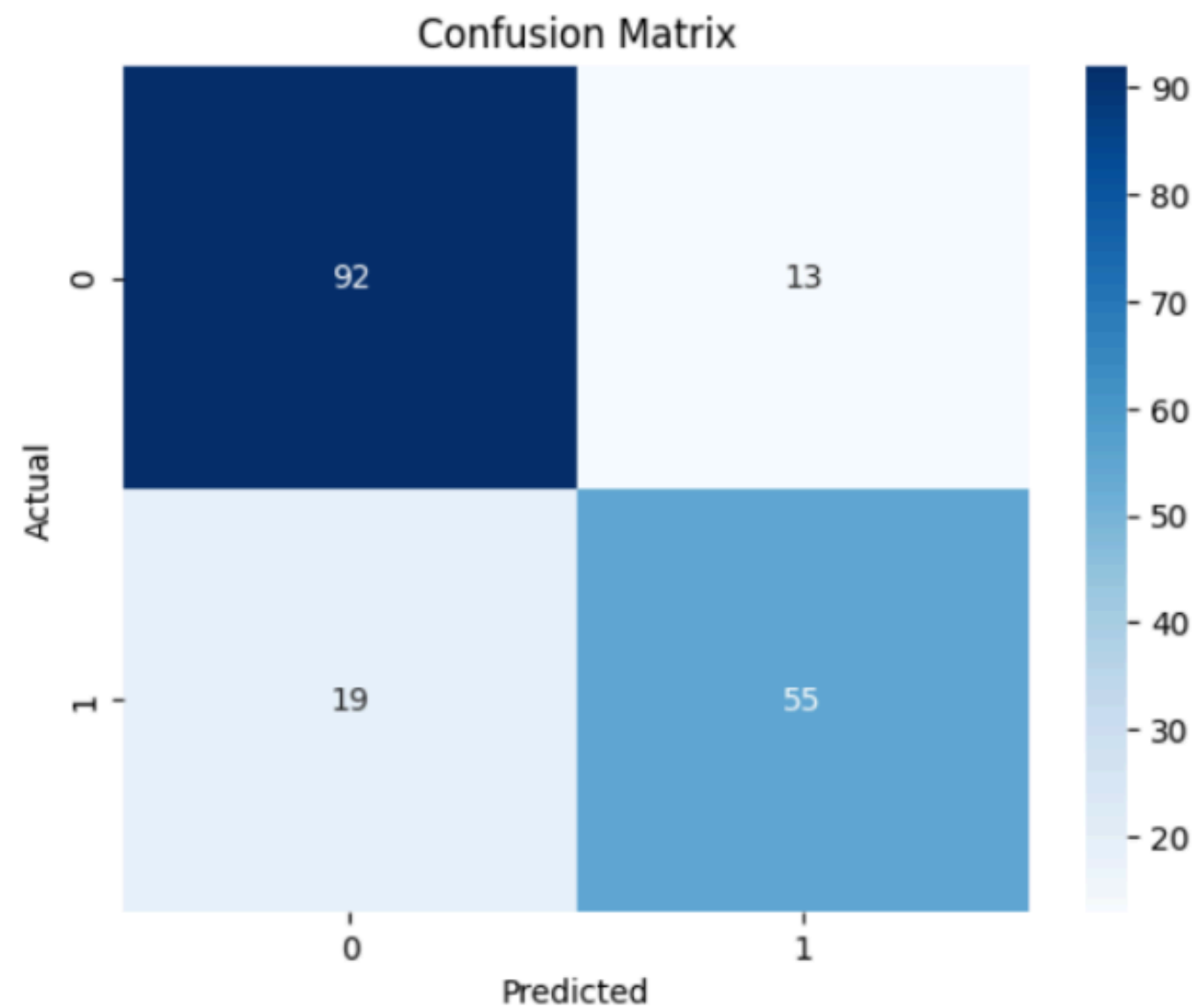
# FEATURE SELECTION



```
RandomForestClassifier  
  
RandomForestClassifier(random_state=42)
```

# DATA MODELLING

- `n_estimators` menentukan jumlah pohon (trees) yang akan dibuat dalam Random Forest. Ketika kita set `n_estimators=100`, ini berarti kita akan membuat 100 pohon keputusan (decision trees) yang berbeda. Setiap pohon akan membuat prediksi, dan hasil akhir adalah kombinasi dari semua prediksi pohon tersebut.
- `random_state` adalah angka yang digunakan untuk mengatur "benih" acak (random seed). Penggunaan `random_state` memastikan bahwa hasil yang kita dapatkan dari menjalankan model selalu konsisten. Misalnya, setiap kali kita menjalankan kode dengan `random_state=42`, hasilnya akan sama.



Hitung akurasi dari prediksi

```
] accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")
```

Accuracy: 0.82

# KONFUSION MATRIX

Dikonfusion Matrix ini adalah seperti ada beberapa data yg dipastikan selamat dan tidak selamat, dan juga ada data actualnya selamat tapi diprediksi tidak selamat dan juga ada actual tidak selamat tetapi diprediksi selamat

Ini adalah seberapa tepatnya data yg dianalisa



# KESIMPULAN

ADA BANYAK SEKALI FAKTOR FAKTOR YANG MEMENGARUHI DATA TITANIC INI,

- 1, FAKTOR UMUR, KARENA ANAK KECIL LEBIH DIPRIORITASKAN
2. FAKTOR GENDER, PEREMPUAN LEBIH DIPRIORITASKAN
3. FAKTOR PEMBAYARAN, KARENA STATUS SOSIAL PADA JAMAN ITU SANGAT DILIHAT SEKALI

Dataset Titanic yang digunakan dalam analisis ini diambil dari kompetisi 'Titanic: Machine Learning from Disaster' yang tersedia di Kaggle. Dataset ini mencakup informasi demografis dan detail lain tentang penumpang Titanic, yang memungkinkan prediksi kelangsungan hidup berdasarkan berbagai faktor seperti usia, jenis kelamin, dan kelas sosial.

---



# THANK YOU

● FOR YOUR NICE ATTENTION

