



HASIL **EXPLORATORY DATA ANALYSIS (EDA)**

Analisis Data Titanic Menggunakan Python (Google Colab



Faiz Haidar Halwi

dibimbing.id

TENTANG SAYA

Halo semuanya, saya Faiz Haidar Halwi. Saya seorang mahasiswa Teknik Informatika yang sedang mengikuti program dari dibimbing yaitu Data Science.

Saya melaksanakan tugas yang diberikan dan mari lihat hasil analisis saya sebagai Data Science



Latar Belakang Analisis Data Titanic

Tragedi tenggelamnya kapal Titanic pada tahun 1912 merupakan salah satu kecelakaan maritim paling terkenal dalam sejarah. Data penumpang Titanic telah menjadi dataset klasik dalam dunia data science dan machine learning karena mengandung berbagai fitur menarik seperti umur, jenis kelamin, kelas sosial, tarif, dan status keselamatan.

Tujuan Analisis

- Mengeksplorasi karakteristik data penumpang Titanic.
- Membersihkan dan memproses data (data cleaning & preprocessing).
- Menemukan pola dan insight awal melalui analisis statistik dan visualisasi.

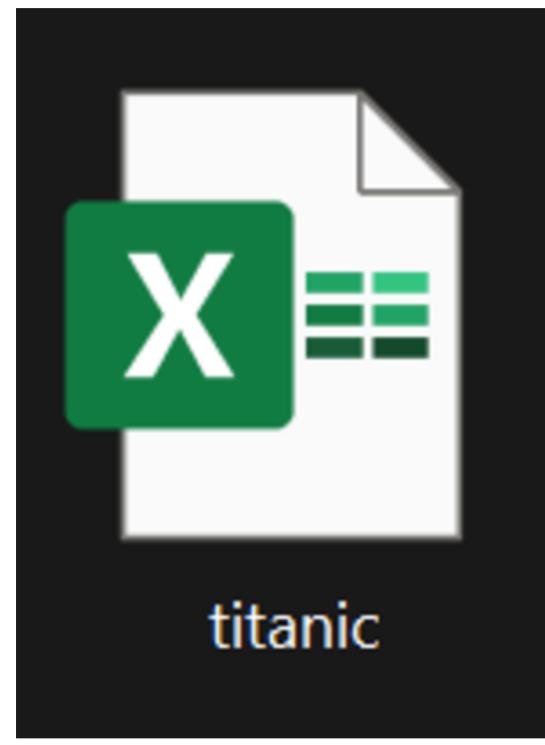
IMPORT LIBRARY DAN DATASET



```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)

# import data
df = pd.read_excel('titanic.xlsx')
df.head()
```

Import Library Python



Dataset

EKSPLORASI AWAL DATA

▼ Lihat head, tail, sample, dan info

```
# Lihat 5 baris pertama
print("== HEAD ==")
display(df.head())

# Lihat 5 baris terakhir
print("\n== TAIL ==")
display(df.tail())

# Ambil sample acak
print("\n== SAMPLE ==")
display(df.sample(5, random_state=42))

# Info lengkap
print("\n== INFO ==")
df.info()
```

== HEAD ==			
	survived	name	sex
0	1	Allen, Miss. Elisabeth Walton	female
1	1	Allison, Master. Hudson Trevor	male
2	0	Allison, Miss. Helen Loraine	female
3	0	Allison, Mr. Hudson Joshua Creighton	male
4	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female

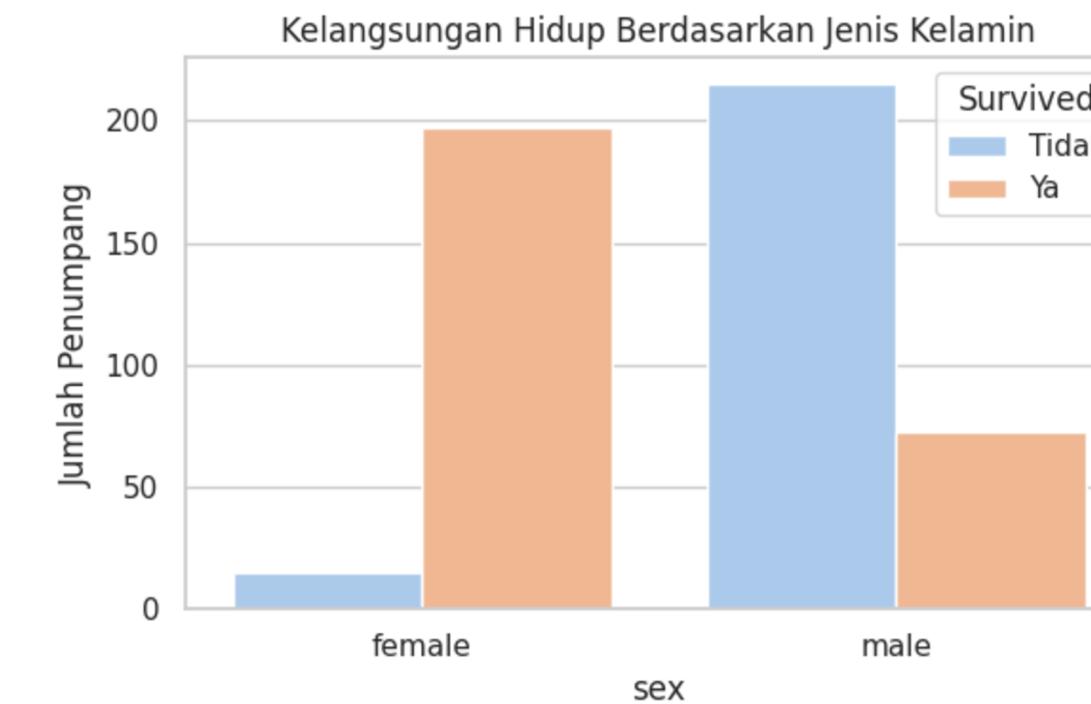
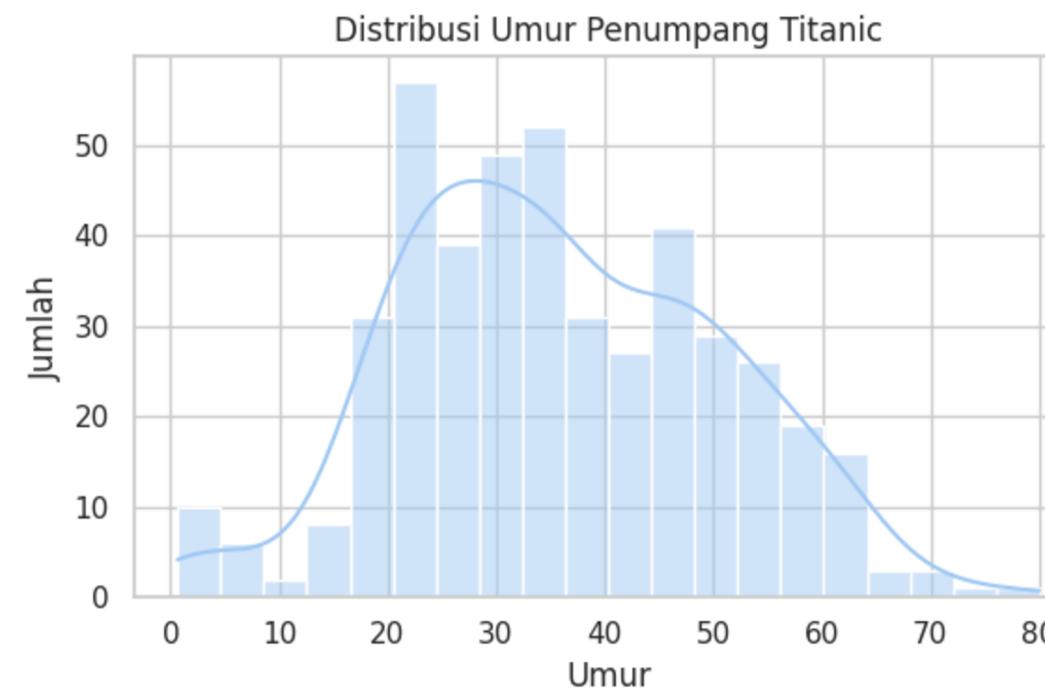
== TAIL ==			
	survived	name	sex
495	1	Mallet, Mrs. Albert (Antoinette Magnin)	female
496	0	Mangiavacchi, Mr. Serafino Emilio	male
497	0	Matthews, Mr. William John	male
498	0	Maybery, Mr. Frank Hubert	male
499	0	McCrae, Mr. Arthur Gordon	male

== SAMPLE ==			
	survived	name	sex
361	1	Caldwell, Mr. Albert Francis	male
73	1	Cleaver, Miss. Alice	female
374	1	Clarke, Mrs. Charles V (Ada Maria Winfield)	female
155	1	Hays, Mrs. Charles Melville (Clara Jennings Gr..	female
104	1	Eustis, Miss. Elizabeth Mussey	female

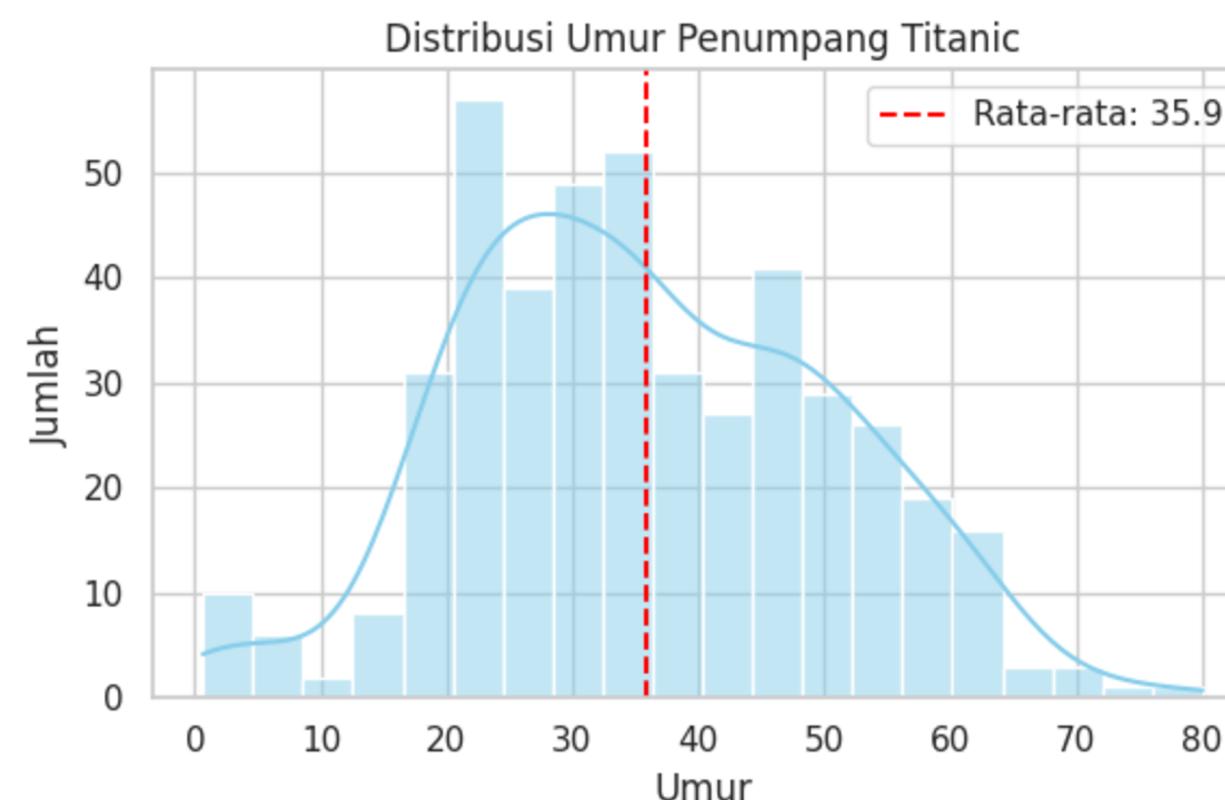
```
== INFO ==
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   survived  500 non-null   int64  
 1   name      500 non-null   object  
 2   sex       500 non-null   object  
 3   age       451 non-null   float64 
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```



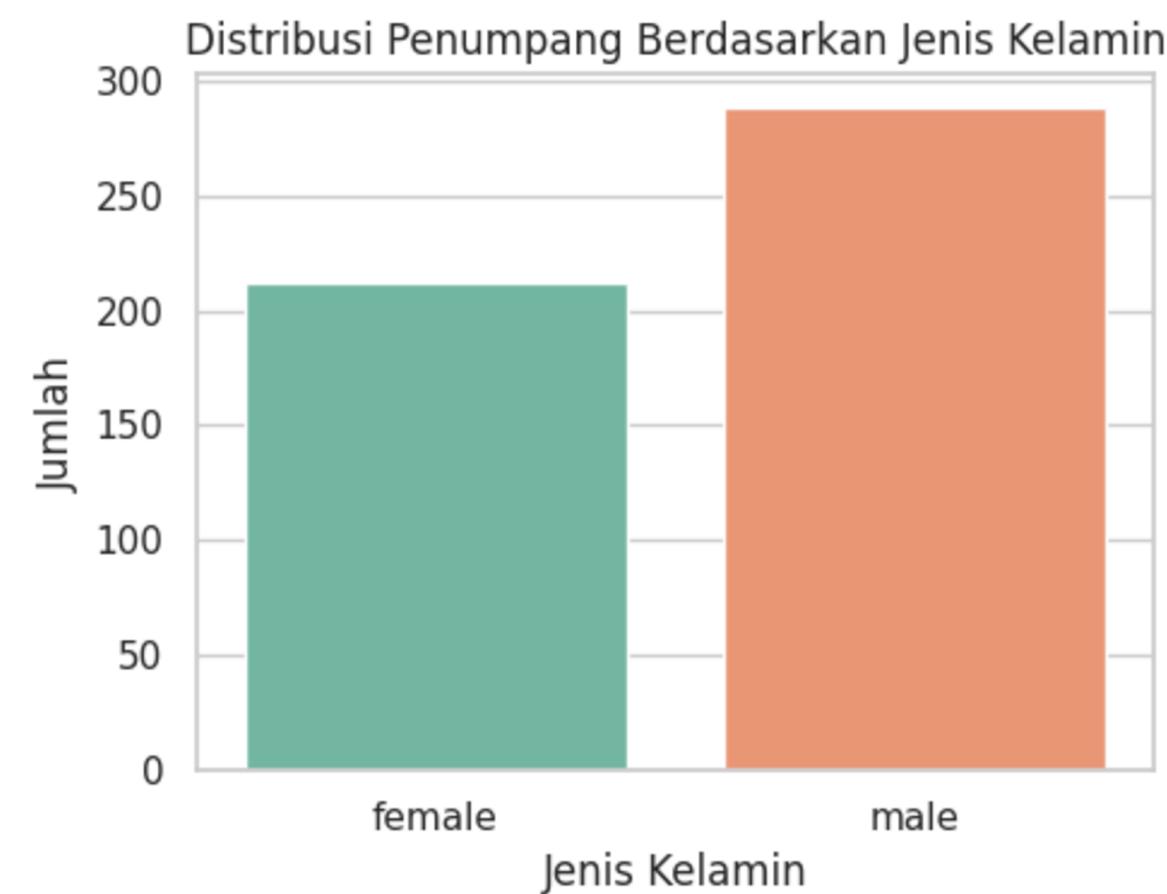
EKSPLORASI AWAL DATA BENTUK VISUALISASI



STATISTICAL SUMMARY



Umur Rata-rata: 35.92 tahun
Umur Termuda: 0.6667 tahun
Umur Tertua: 80.0 tahun



STATISTICAL SUMMARY



dibimbing.id

01.

Dari Kolom age (Umur)

- Rata-rata umur penumpang adalah sekitar 29.7 tahun
- Penumpang termuda berusia sekitar 0.42 tahun (masih bayi)
- Penumpang tertua berusia 80 tahun
- Mayoritas penumpang berada dalam rentang umur 20 hingga 40 tahun, yang ditunjukkan oleh bentuk distribusi umur yang miring ke kanan (right-skewed)

02.

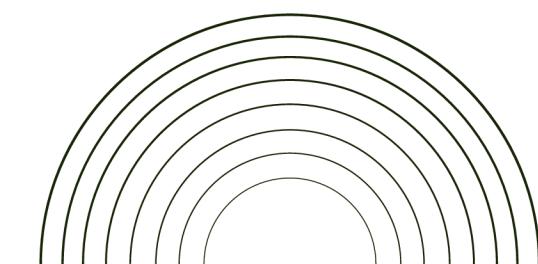
Dari Kolom sex (Jenis Kelamin)

- Jumlah penumpang laki-laki lebih banyak dibandingkan perempuan
- Visualisasi menunjukkan bahwa laki-laki mendominasi populasi penumpang

02.

Dari Kolom survived (Keselamatan)

- Proporsi penumpang yang selamat lebih sedikit dibandingkan yang tidak selamat
- Distribusi survived menunjukkan bahwa lebih dari 50% penumpang tidak selamat



PEMERIKSAAN DUPLIKAT

✓ Cek duplikat dan menghapusnya jika ada

```
[ ] # Cek jumlah duplikat
duplicates = df.duplicated().sum()
print(f"\nJumlah duplikat: {duplicates}")

# Jika ada, hapus duplikat
if duplicates > 0:
    df = df.drop_duplicates()
    print("Duplikat dihapus.")
```



Jumlah duplikat: 1
Duplikat dihapus.

Pemeriksaan Duplikat

- Menggunakan `df.duplicated().sum()` ditemukan:
Jumlah duplikat: 1

Penanganan

- Duplikat dihapus menggunakan `df.drop_duplicates()`
- Output: "Duplikat dihapus."



PEMERIKSAAN MISSING VALUES

dibimbing.id

```
▶ df.isnull().sum()
df.isnull().mean() * 100

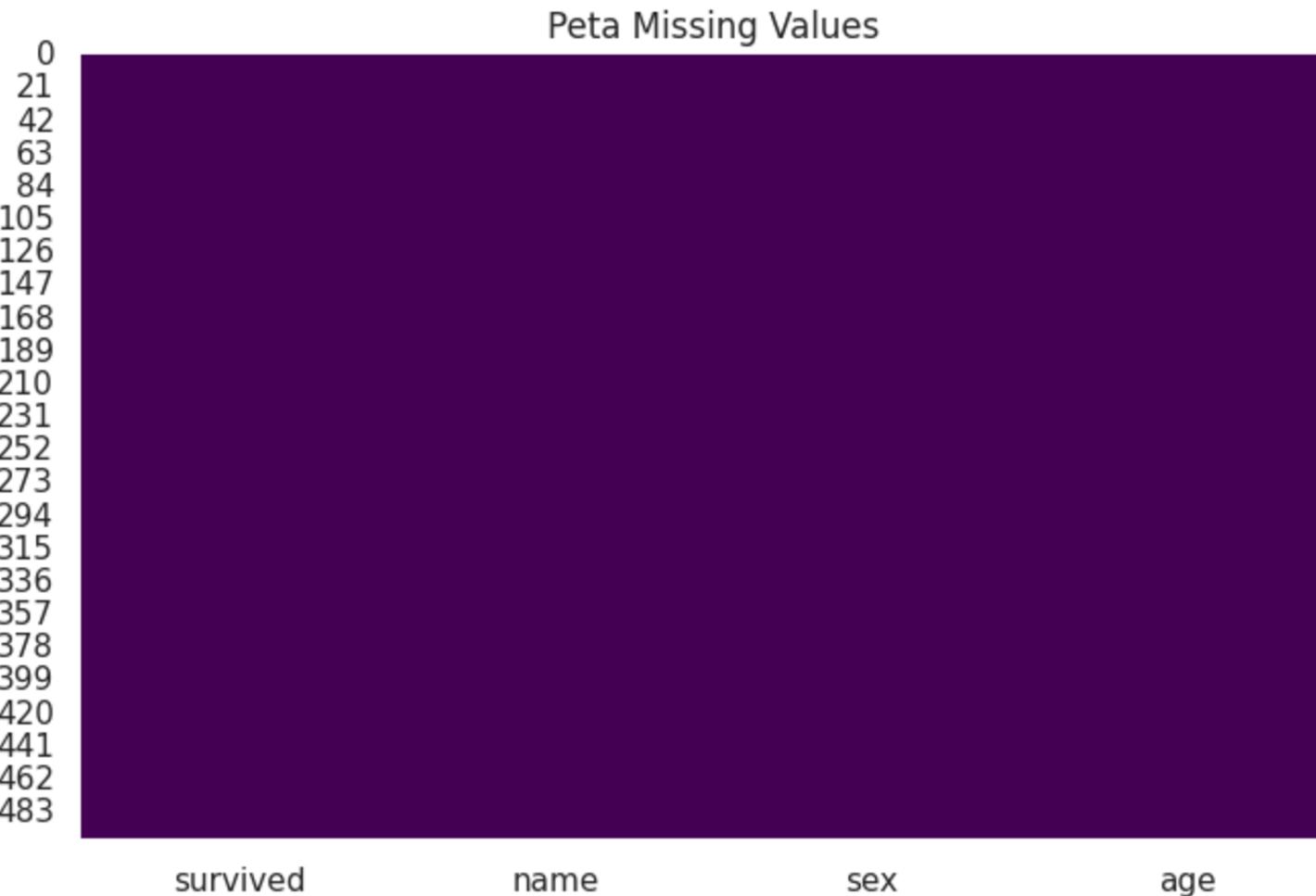
0
survived 0.0
name 0.0
sex 0.0
age 0.0

dtype: float64

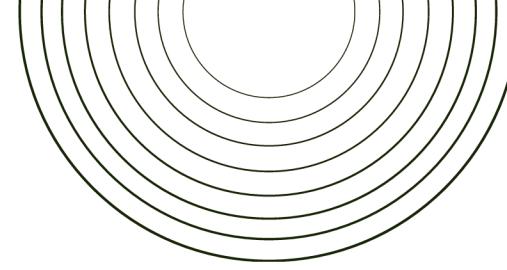
▶ import seaborn as sns
import matplotlib.pyplot as plt

sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
# Contoh imputasi dengan median
df['age'].fillna(df['age'].median(), inplace=True)

plt.title("Peta Missing Values")
plt.show()
```



- Tidak ditemukan missing value pada kolom survived, name, sex, dan age.
- Kolom age sebelumnya memiliki missing value, namun telah diimputasi dengan nilai median.
- Heatmap menunjukkan warna ungu tua, menandakan data sudah bersih dari nilai kosong.

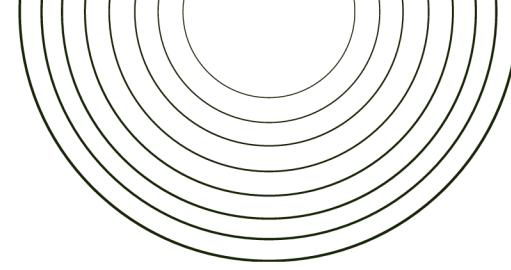


INSIGHT UTAMA

Kolom age sebelumnya memiliki banyak missing values, namun telah ditangani dengan imputasi median untuk menjaga distribusi data.

Mayoritas penumpang adalah laki-laki, menunjukkan potensi ketimpangan demografis pada data.

Tidak terdapat duplikat setelah pembersihan data, memastikan kualitas data yang baik.



REKOMENDASI ANALISIS LANJUT

Analisis multivariat: Menggabungkan faktor seperti jenis kelamin dan usia untuk melihat pengaruhnya terhadap keselamatan.

Penambahan kolom lain (seperti Pclass, Fare, Embarked) dari dataset lengkap akan memberikan insight yang lebih mendalam.

Penerapan model prediksi (machine learning) untuk memperkirakan kemungkinan selamat berdasarkan fitur-fitur tertentu.



SEKIAN TERIMA KASIH

Sekali lagi, terima kasih atas kesempatan ini. Mari kita lanjutkan diskusi ini di lain waktu. Sampai jumpa!

SELESAI

