

EDA - Pacient Liver

Bootcamp Project

Next Slide



Who I am?

A recent graduate with a Bachelor's degree in Computer Systems from Universitas Sriwijaya, experienced as an IT Support and IT Assistant. Skilled in hardware and software troubleshooting, network installation and configuration, and familiar with tools such as Mikrotik, Cisco Packet Tracer, Winbox, and AnyDesk. Conducted a final project research on malware attack classification using machine learning methods. Strong communication skills, adaptive, responsive, and capable of working independently or collaboratively to support operational IT system efficiency.

MULKI PEDERSON, S.KOM

IT Support | Network Engineer | Data Scientist Enthusiast

BACKGROUND PROBLEM

Penyakit liver adalah masalah kesehatan serius dengan angka kematian tinggi. Dataset Indian Liver Patient berisi data pasien liver dan sehat dari India dengan berbagai fitur medis. Data ini memiliki nilai hilang dan terduplicat serta distribusi data tidak normal, sehingga perlu dilakukan EDA (Exploratory Data Analysis) terlebih dahulu. Proyek ini fokus pada EDA dan persiapan awal dataset pasien liver untuk mendukung tindakan selanjutnya.



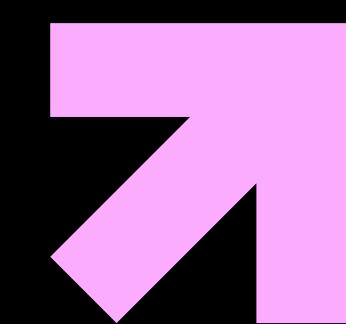
ABOUT DATASET

Dataset yang digunakan merupakan dataset yang diunduh dari Our backend team creates the logic and infrastructure that make applications function seamlessly. We leverage databases, server-side programming, and APIs to ensure data security, speed, and reliability. Dataset ini berisikan 583 data pasien di India dengan 11 kolom yang dapat menentukan pasien terkena sakit liver atau tidak sakit liver (sehat)

indian_liver_patient.csv

What will I do?

- Handling Missing Value
- Data Duplicate
- Handling Outlier?



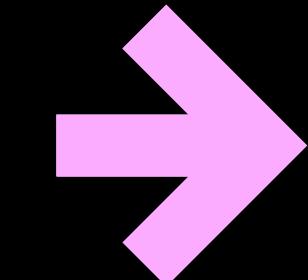
EDA (EXPLORATORY DATA ANALYSIS)



EDA - MISSING VALUE

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Age              583 non-null    int64  
 1   Gender            583 non-null    object  
 2   Total_Bilirubin  583 non-null    float64 
 3   Direct_Bilirubin 583 non-null    float64 
 4   Alkaline_Phosphotase 583 non-null  int64  
 5   Alamine_Aminotransferase 583 non-null  int64  
 6   Aspartate_Aminotransferase 583 non-null  int64  
 7   Total_Protiens    583 non-null    float64 
 8   Albumin           583 non-null    float64 
 9   Albumin_and_Globulin_Ratio 579 non-null  float64 
 10  Dataset            583 non-null    int64  
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ KB
```

Age	0
Gender	0
Total_Bilirubin	0
Direct_Bilirubin	0
Alkaline_Phosphotase	0
Alamine_Aminotransferase	0
Aspartate_Aminotransferase	0
Total_Protiens	0
Albumin	0
Albumin_and_Globulin_Ratio	4
Dataset	0



Age	0
Gender	0
Total_Bilirubin	0
Direct_Bilirubin	0
Alkaline_Phosphotase	0
Alamine_Aminotransferase	0
Aspartate_Aminotransferase	0
Total_Protiens	0
Albumin	0
Albumin_and_Globulin_Ratio	0
Dataset	0

Dari informasi di atas terdapat data yang hilang yaitu pada kolom Albumin_and_Globulin_Ratio

Handling missing value is done

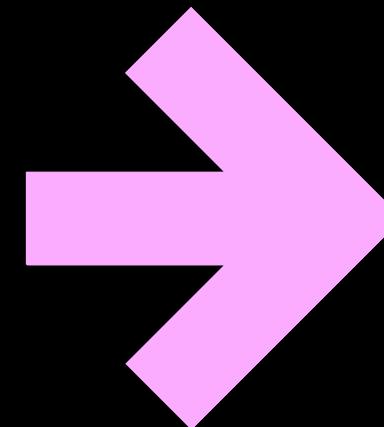
EDA - DATA DUPLICATE

Untuk melakukan pengecekan data duplicate :

```
data_duplicate = df.duplicated().sum()
```

Jika terdapat data duplicate bisa di drop/ hapus :

```
df = df.drop_duplicates()
```



```
# Mengecek apakah ada duplicate di seluruh kolom  
check_duplicate = df.duplicated().sum()
```

```
print(f"Jumlah data yang duplikat = {check_duplicate}")
```

```
Jumlah data yang duplikat = 13
```

```
# Handling duplicate
```

```
df = df.drop_duplicates()
```

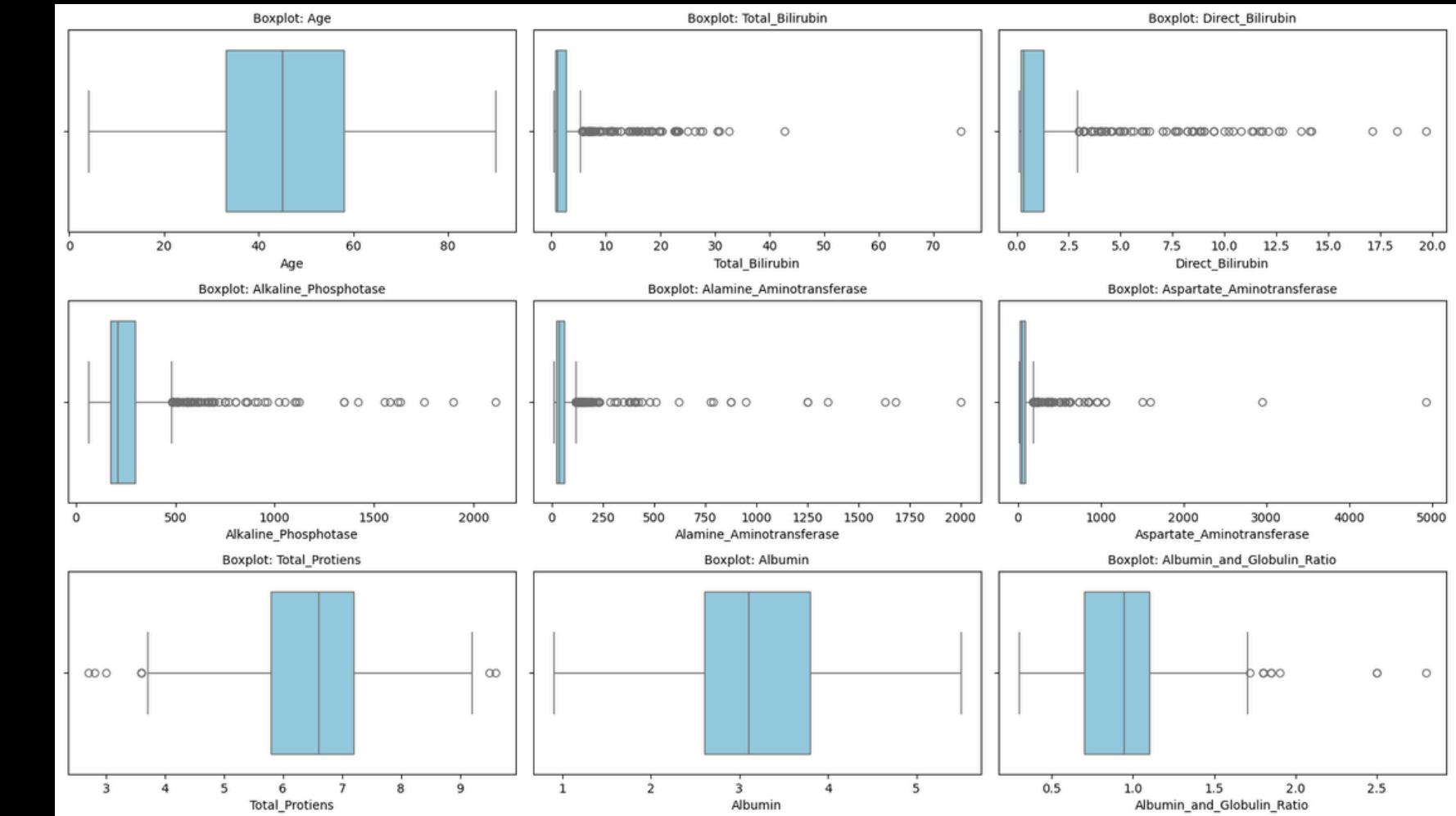
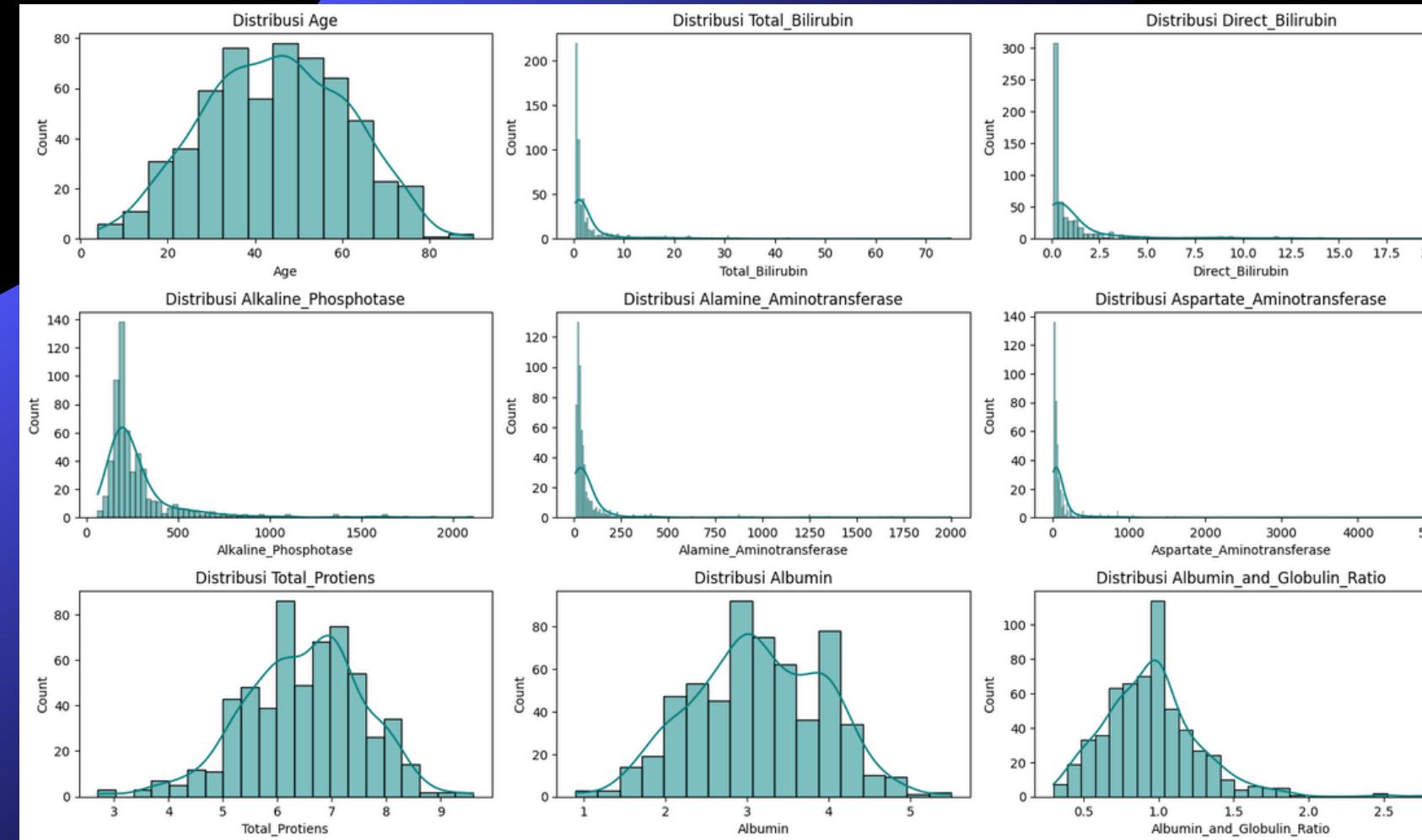
```
# Mengecek duplicate setelah di-handle
```

```
handle_duplicate = df.duplicated().sum()
```

```
print(f"Jumlah data yang duplikat = {handle_duplicate}")
```

```
Jumlah data yang duplikat = 0
```

EDA - OUTLIER



Data outlier pada data pasien liver (data medis) merupakan suatu yang wajar terjadi karena memang masih memungkinkan ada beberapa pasien yang sudah dalam kondisi yang akut, sehingga ini bukan suatu kesalahan yang tidak perlu dihapus/ dihilangkan yang bisa saja itu merupakan data yang penting. Pada Dataset ini juga terdapat beberapa data outlier pada masing - masing kolom sehingga data terlihat lebih ke right-skewed seperti gambar diatas. Gambar di peroleh dengan bantuan dari library matplotlib.pyplot dan seaborn.

INSIGHT DAN ADVISE



Insight

- Dataset memiliki data kosong (missing value) yang dapat ditangani, diisi dengan nilai rata-rata karena merupakan kolom numerik.
- Distribusi data tidak sepenuhnya normal (banyak yang skewed).
- Menampilkan visualisasi data dengan histogram dan boxplot
- Kolom target Dataset menunjukkan ketidakseimbangan data.
- Data yang duplikat dihapus, data menjadi bersih dan siapkan di analisis lebih lanjut

Advise

Perlu preprocessing lebih lanjut sebelum modeling seperti melakukan transformasi log/sqrt pada fitur skewed

Thank You

