DATA PREPROCESSING

Asep Muhidin, S.Kom., M.Kom.

BACKGROUND

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results.

"How can the data be preprocessed in order to help improve the quality of the data ?

How can the data be preprocessed so as to improve the efficiency and ease of the mining process?

PREPROCESSING TASK

Data Cleaning

Data Integration

Data Reduction

Data Transformation

DATA CLEANING

Langkah-Langkah Data Cleaning

01

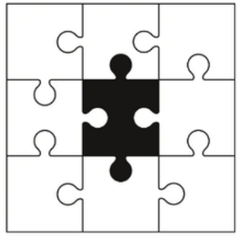
Hapus duplikat atau data tidak relevan

02

Perbaiki kesalahan struktural

03

Saring *outlier*



04

Kelola data yang hilang

05

Lakukan validasi dan *quality assurance* (QA)

Sumber: Tableau.com

HAPUS DUPLIKAT DATA

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

path="https://raw.githubusercontent.com/asepmuhidin/MK-Data-Mining-UPB/main/dataset/titanic.csv"
df=pd.read_csv(path)
df.duplicated().sum()
```

Untuk menghapus duplikat data :
df.drop_duplicates(inplace=True)

MISSING VALUE

Cek missing value

df.isna().sum()

```
df.isna().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
```

1 Sebutkan kolom yang missing value ?

2 Sebutkan tipe data kolom tersebut?

HANDLING MISSING VALUE

Solusi sederhana tapi mahal

Hapus semua data yang missing value **df.dropna()**

Jika memang terlalu besar missing valuenya(>50%), hapus kolom.

df.drop(columns=['Cabin'], inplace=True)

IMPUTASI MISSING VALUE

Tipe Data Numerik, isi dengan nilai rata-rata/ median / modus

Mencari nilai mean, median & modus :

```
df.Age.mean()
df.Age.median()
df.Age.mode()
```

Inpute dengan nilai mean/median/modus :

```
df.Age.fillna(df.Age.mean(),inplace=True)
```

Tipe Data Kategori, isi dengan nilai paling banyak muncul

```
df.Embarked.value_counts()
```

```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

```
df.Embarked.fillna('S',inplace=True)
```

THANK YOU