Machine Learning

Data → 80% train
Data → 20% test

independent Feature
dependent Feature

Car price

Age, door, Engine, price

price → dependent

Age, door, Engine → independent

Feature scaling → standardization

$$X' = \frac{X - \mu}{\sigma}$$

μ → mean
σ → sd

[-3, +3]

X_train → di standardisation cari mean, sd dapat X'

X_test → pakai mean dan sd dari X_train

Feature scaling → normalization

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

[0, 1]

Feature Scaling
Kenapa?

e.g

Income		Age	
70.000	10.000	45	1
60.000	8.000	44	4
52.000		40	

datanya terlalu berogam rangenya
Perlu dilakukan feature scaling.

after normalization

Income	Age
1	1
0,444	0,8
0	0

1. Data Preprocessing
   - split to Independent feature and dependent Feature
   - take care Missing values → delete (kalau data besar yg missing cuma 1%)
     → Fill with mean → impute sklearn
   - encoding categorical data (mengubah ke angka)
     - encoding categorical Independent variable
     - encoding categorical dependent variable
   - split → training set
     → test set
     sklearn.model_train_test_split
   - Feature Scaling → normalization kalau normal distribution pada mayoritas featurenya
     → standardization → will works well all the time
     → sklearn.preprocessing → StandardScaler

Library :
- Numpy → Matematis
- pandas → DataFrame
- MatplotLib → Visualization
- Scikit-learn → model, preprocessing

Kenapa
Split dulu baru Feature scaling
↳ test_set itu harus benar" baru agar model optimal
kalau FS dulu maka data semua pakai mean sd → test set tidak benar" baru

encoding

Categorical independent feature → sklearn
- compose → Column transformer
- preprocessing → OneHotEncoder

e.g Biru, Merah, Kuning, Merah, Kuning

ada 3 Category → ada 3 column

one hot encoding

agar dapat dipahami mesin

Biru	1	0	0
Merah	0	1	0
Kuning	0	0	1
Merah	0	1	0
Kuning	0	0	1

Categorical dependent feature → sklearn preprocessing → LabelEncoder

e.g Yes, No

Feature Scaling → Normalization (kalau mayoritas feature mempunyai Normal Distribution)
↓
Standardization
(work well All the time)

X_training → cari mean, sd dapat x' [-3,3]

X_test → pakai mean, sd dari X_training untuk mencari x'

Standardization diapply hanya untuk numerical feature (murni)
bukan encodingan