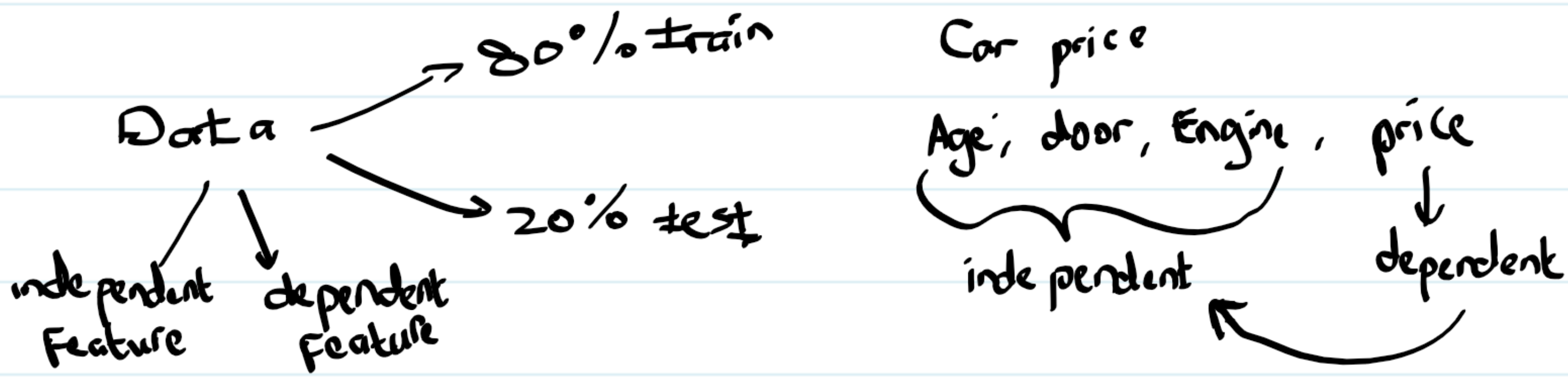


Machine Learning



Feature Scaling \longrightarrow standardization

normalization \downarrow

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

$[0, 1]$

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

$[-3, +3]$

X-train \rightarrow di standardisasi
can mean, sd,
dapat X'

X-test \rightarrow pakai mean dan sd
dari X-train

Feature scaling

e.g

Income	Age
70.000	45
60.000	44
52.000	40

data nya terlalu beragam range nya
Perlu dilakukan feature scaling.

after normalization

Income Age

$$\begin{array}{r} 1 \\ 0,444 \\ 0 \end{array}$$

1. Data Preprocessing

- split to independent feature and target
- take care missing values
 - delete rows data based on missing value 1%
 - Fill with mean → impute sklearn
- encoding categorical data (one-hot)

encoding categorical independent variable → encoding categorical dependent variable (misalkan ke angka)

- independent variable dependent variable
- Split → training Set sklearn.model
 ↘ test Set L_train_test_split
- linearization kolar normal distri

- Feature Scaling
 - normalization *kalo normal distribution*
 - standardization *padahal mayoritas featurenya*
 - will work well all the time
 - ↳ sklearn.preprocessing
 - ↳ StandardScaler

Library: • numpy → Matplotlib

- pandas → DataFrame

- Matplotlib → Visualisierung

- Scikit-learn → model, preprocessing

Kenapa Split data baru Feature scaling baru test set harus benar" baru
↳ test set harus benar" agar model optimal
Kalau FS data maka data sama dan
rata-rata mean sd → ter cupur dan
test set tidak benar" baru

encoding

Categorical independent feature

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

ada 3 category

one hot encoding

agar dapat dipahami mesin

Sklearn

compose \rightarrow Column transformer

→ preprocessing → onehot encoder

Maps

Biru 1 0 0

Merch 010

Kuang 001

Merah	0	1	0
Kuning	0	0	1

Why?

Categorical dependent feature \rightarrow sklearn preprocessing \rightarrow Label Encoder

eg Yes, No

Feature Scaling \rightarrow Normalization (Kalau mayoritas feature mempunyai Normal Distribution)

Standardization
(work well All the time)

$X_{\text{training}} \rightarrow \text{can mean, sd output } x' [-3, 3]$

X_{test} \rightarrow pakai mean, sd dari X_{training} untuk mencari x'

Standardization dioppy hanya untuk numerical feature (murni)
bukan encoding