

Introduction to Data Science



day 5



makassar coding
Everyone **can** Write the **Code**

Day 5

Outline

apa saja yang akan kita pelajari hari ini?

Outline

1. **Classification Model**

- Logistic Regression
- K-Nearest Neighbour
- Decision Tree

2. **Feature Engineering**

- Scaling
- Encoding
- Handling Missing Value
- Outlier
- Feature Selection

Supervised Learning

Classification

Supervised Learning

What Is Classification

- ❑ **Classification** digunakan untuk memprediksi data atau label yang sifatnya kategorik.
- ❑ Setiap kategori yang ada dapat juga disebut dengan kelas.
- ❑ Banyaknya kelas bisa dua atau bahkan lebih dari itu.
 - Y categorical - memiliki dua kelas (binary classification)
 - Y categorical - memiliki lebih dari dua kelas (multiclass classification)

x1 →
x2 →
...
xk →

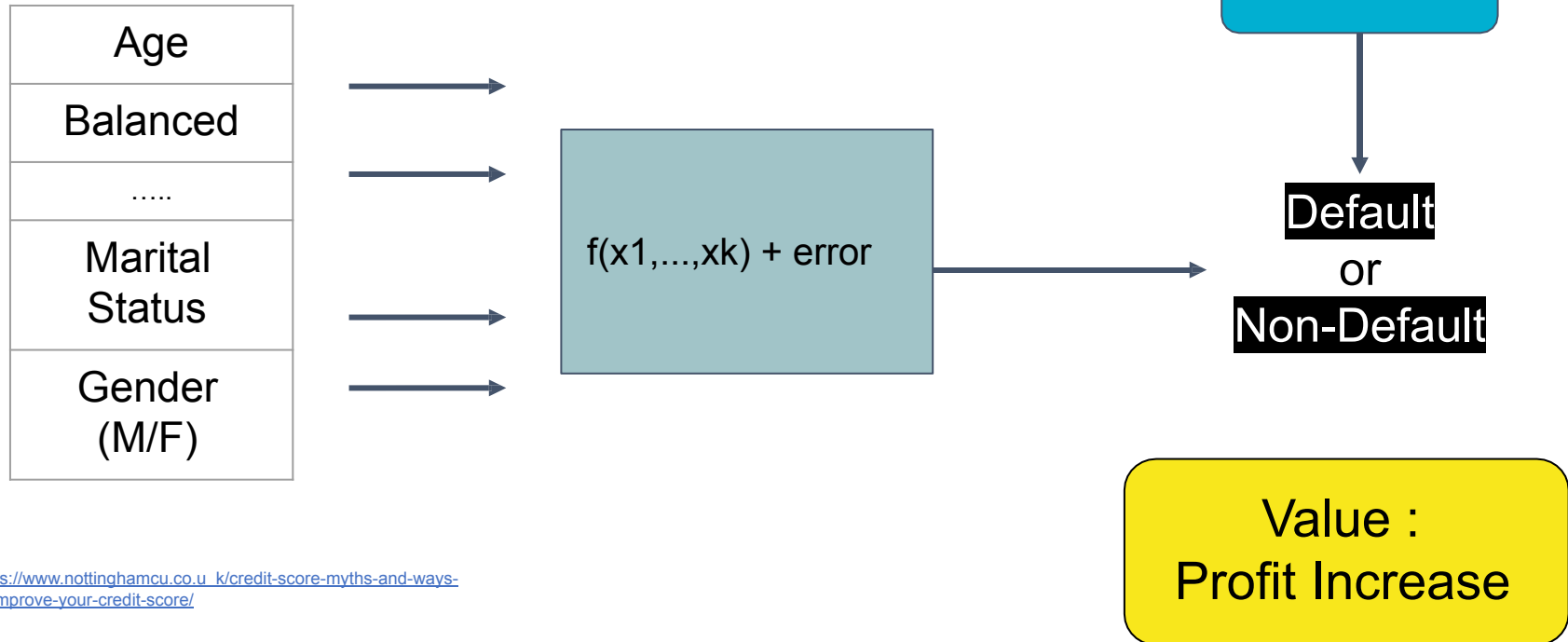
$f(x_1, \dots, x_k) + \text{error}$

What???

Y

Supervised Learning

Classification : Credit Scoring (Binary)



Supervised Learning

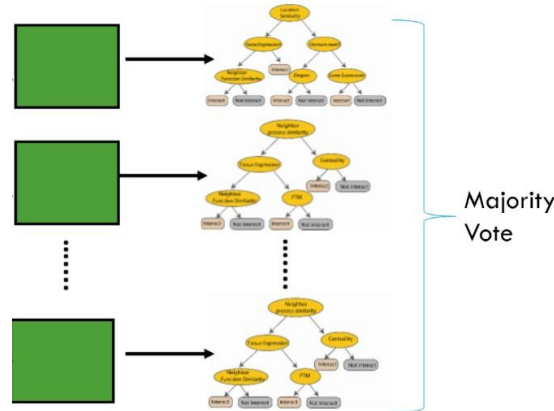
Classification : Some Method Usually Used In Classification

Logistic Regression

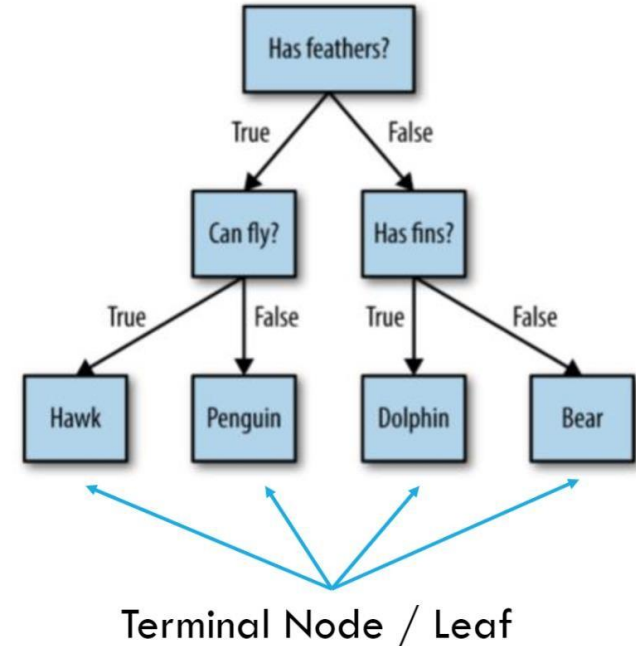
$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

Ensemble Method:

Note:
Other models:
K-Nearest
Neighbour (KNN),
Discriminant Analysis,
Support Vector
Machine (SVM),
Ensemble – Bagging,
Random Forest,
Boosting, etc



Decision Tree



Classification

Logistic Regression

Regresi logistik (Logistic Regression)

merupakan salah satu metode yang dapat digunakan untuk melakukan klasifikasi.

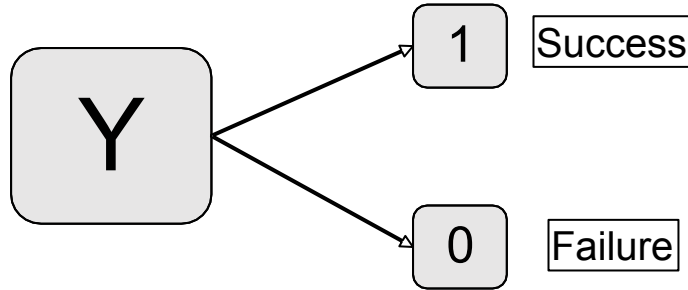
Logistic Regression

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

- **Binary Logistic Regression**, binary label
Ketika label memiliki skala pengukuran nominal dan hanya memiliki dua kategori atau biner regresi
- **Multinomial Logistic Regression**, multinomial label
Ketika label memiliki skala pengukuran nominal dengan lebih dari dua kategori
- **Ordinal Logistic Regression**, ordinal label
Ketika label memiliki skala pengukuran ordinal

Classification

Logistic Regression : What is Binary Logistic Regression ?



Binary logistic regression memodelkan peluang/probability suksesnya suatu kejadian.

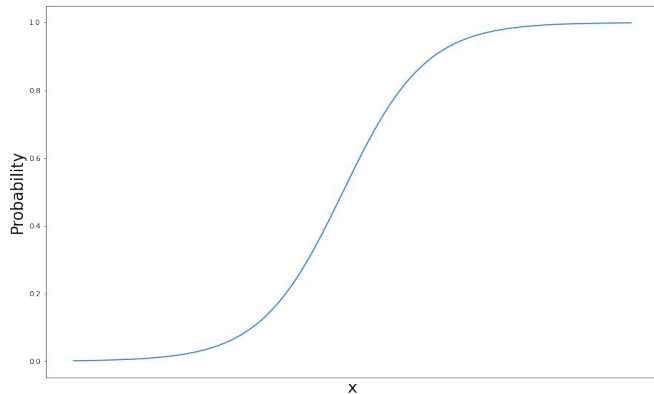
Has more interest in success event

Case	1	0
Credit scoring	Bad	Good
Churn Analysis	Turn Over	Stay
Propensity	Buy	Not Buy

Classification

Logistic Regression : Sigmoid Curve

$b > 0$, success rate increase when X increase

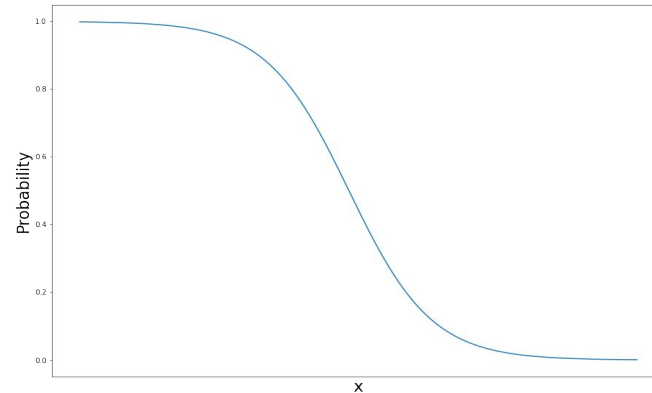


$$P(Y = 1) = \text{odd} / (1 + \text{odd}),$$

With

- $0 < P(Y = 1) < 1$
- Y = dependent variable, succes ($Y = 1$) failure ($Y = 0$)
- $\text{odd} = \exp(a + bx)$
- x = independent variable

$b < 0$, success rate decrease when X increase

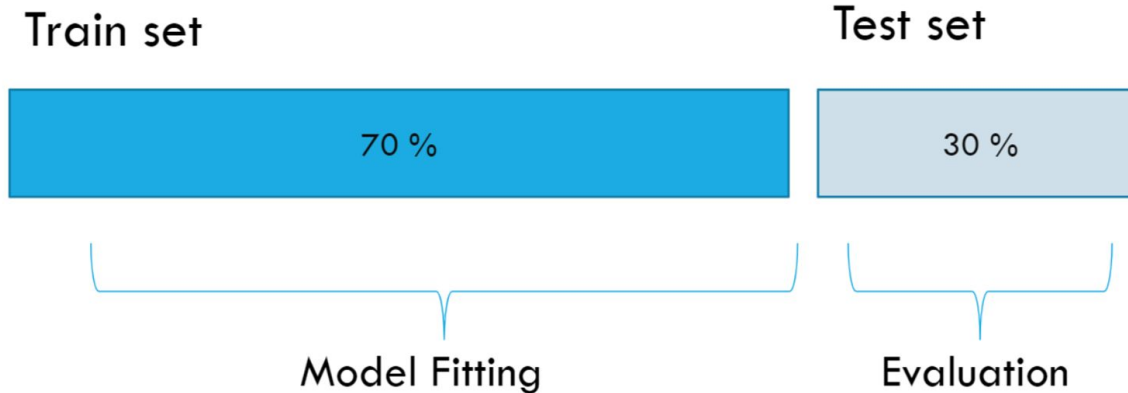


Odd merupakan rasio antara peluang gagal dan peluang sukses dalam kondisi tertentu.

Classification

Evaluation Method

Ketika melakukan pemodelan menggunakan machine learning, **model divalidasi nilai prediksinya menggunakan gugus data yang tidak terlibat sama sekali dalam pemodelan.**



Data dibagi menjadi dua bagian yaitu data training dan data testing dengan proporsi masing-masing misalkan 70% dan 30%. Gugus data training digunakan untuk membangun model sedangkan gugus data testing untuk validasi.

Classification

Measuring Performance of Classification Method

No	Prediction	Actual
1	1	1
2	1	0
3	0	1
..
499	0	0
500	0	1

Prediction	Actual	
	0	1
0	120	23
1	27	330

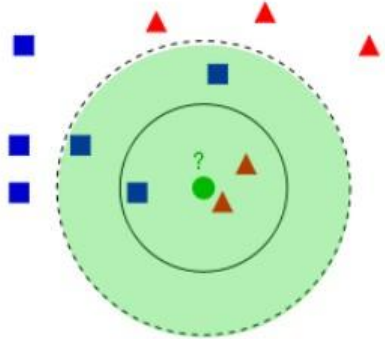
$$\begin{aligned}\text{Accuracy Of Prediction} &= (120+330)/500 \times 100\% \\ &= 90.0\%\end{aligned}$$

Artinya model dapat memprediksi dengan benar untuk 9 dari 10 orang.

Classification

K-Nearest Neighbour

Metode machine learning ini bekerja dengan memberikan hasil prediksi berdasarkan kelas mayoritas dari beberapa pengamatan yang serupa atau tetangga terdekat.

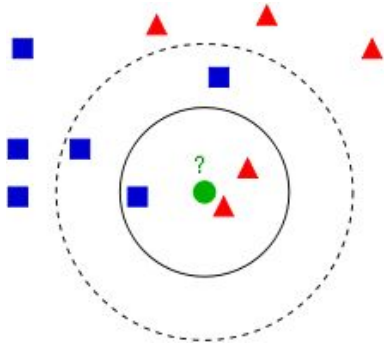


- KNN dapat diterapkan untuk **klasifikasi** maupun **regresi**.
- KNN merupakan suatu metode yang bersifat **non-parametrik** artinya KNN tidak menghasilkan persamaan seperti pada regresi linier dan regresi logistik.

Classification

K-Nearest Neighbour

- **KNN dalam penggunaannya perlu menyimpan data yang digunakan untuk training** sebagai bagian dari metode secara keseluruhan.
- Ketika memprediksi data poin yang baru, **KNN mencari sejumlah data poin yang memiliki kemiripan atau posisinya dekat dengan data poin baru yang ingin kita prediksi**. Hasil prediksi dari data poin yang baru diperoleh berdasarkan mayoritas kelas dari sejumlah data poin yang mirip.



■ ▲ Training data

● Test data (New observation)

Test data will be classified as ■ ? or ▲
(Ignore circle line for now)

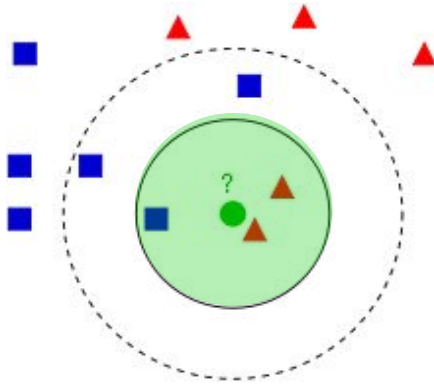
Classification

K-Nearest Neighbour

Banyaknya data terdekat yang digunakan, dinamakan faktor k.

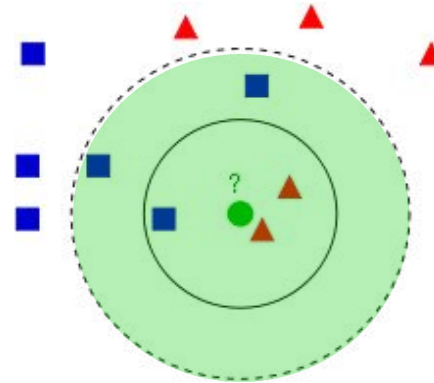
With **k=3** nearest neighbors

Test data classified into ▲



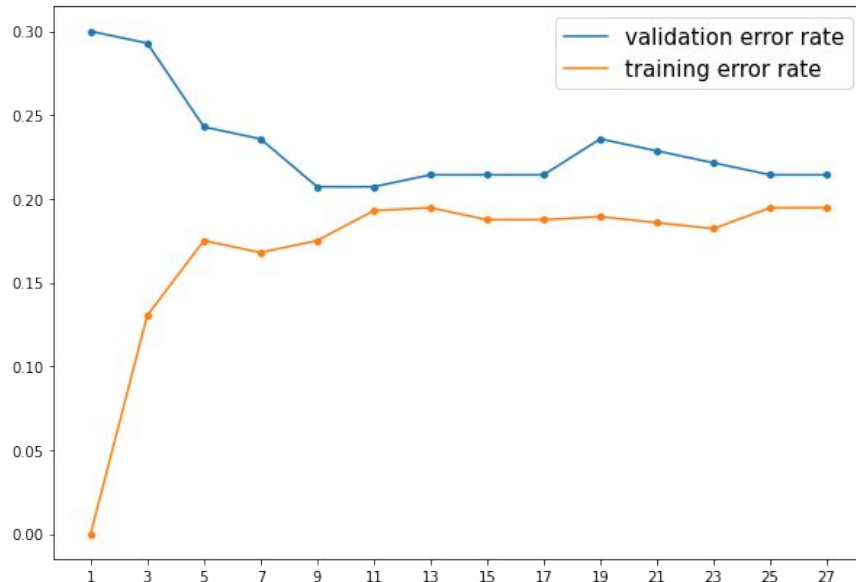
With **k=5** nearest neighbors

Test data classified into ■



Classification

K-Nearest Neighbour : How do we choose factor K?



- Tips 1: Use odd number of K
- Tips 2: Evaluate using validation data set

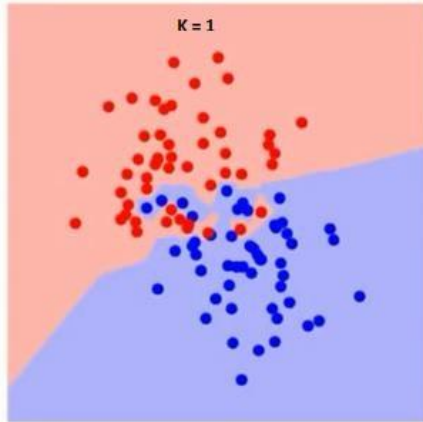
Best k = 9

- Error rate at K=1 can perfectly predict training sample, closest point to any data point is itself
- Our goal is to predict new data so we want good performance in validation data set
- Performance at K=1 not acceptable to predict new data
- Error rate in validation set generally decreases with increases K
- We choose k with minimum error rate in validation dataset

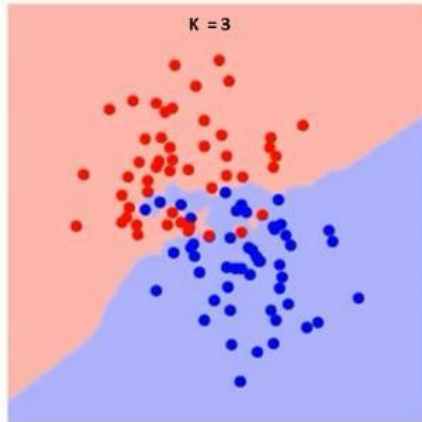
Classification

K-Nearest Neighbour : How do we choose factor K?

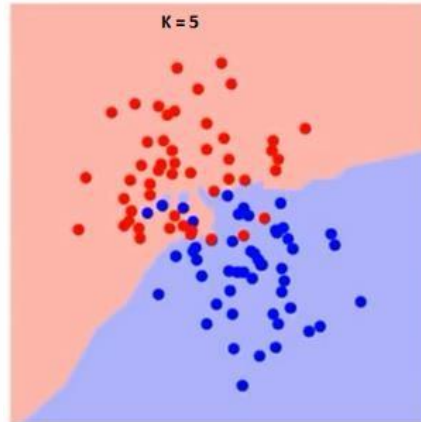
k=1



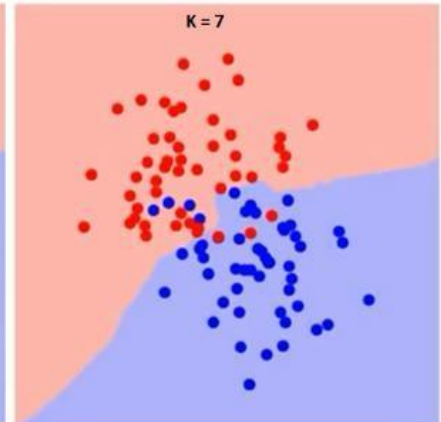
k=3



k=5



k=7



- Boundary becomes smoother with increase value of K
- With K increases to inf, finally becomes all-blue/all-red depending on total majority

Classification

K-Nearest Neighbour : Measuring distance

Measuring distance 1 Dimension



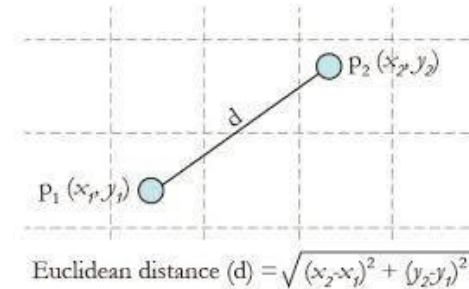
Simple Illustration
The distance is simply :

Distance	=	5 - 1
	=	4

Measuring distance > 2 Dimension

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \dots}$$

Measuring distance 2 Dimension



Illustration

$$(x_1, y_1) = (1, 2)$$

$$(x_2, y_2) = (5, 4)$$

Euclidean distance

$$= \sqrt{(5 - 1)^2 + (4 - 2)^2}$$

$$= \sqrt{(4)^2 + (2)^2}$$

$$= \sqrt{20}$$

$$= 4.47$$

Classification

K-Nearest Neighbour : Measuring distance

Yang sebenarnya terjadi dalam metode KNN, perhitungan jarak dilakukan untuk setiap data poin terhadap data poin yang lain.

data points	x1	x2	x3
1	12	15	16
2	12	16	17
3	20	13	18
4	9	14	18
5	17	15	20

data points	1	2	3	4	5
1	0				
2	1.414	0			
3	8.485	8.602	0		
4	3.741	3.7741	11.045	0	
5	6.403	5.916	4.123	8.306	0

Classification

K-Nearest Neighbour : The Closest Data Points

data points	1	2	3	4	5
1	0				
2	1.414	0			
3	8.485	8.602	0		
4	3.741	3.7741	11.045	0	
5	6.403	5.916	4.123	8.306	0



Distance Matrix

Data Points	1st closest	2nd closest	...
1	2	4	...
2	1	4	...
3	5	1	...
4	1	2	...
5	3	2	...

Classification

K-Nearest Neighbour : Issue with Euclidean Distance

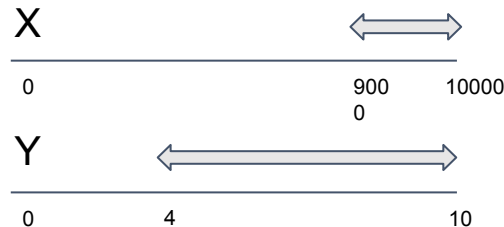
X	Y
4000	5
5000	4.5
2000	3
6000	4.5
7000	4.6
8000	4
9000	10

Misalkan ada suatu data set yang memiliki 2 variabel dengan skala yang berbeda jauh,

- X dalam hektar
- Y dalam kwintal

Perhatikan 2 data point terakhir:

- Selisih Area (X) $9000 - 8000 = 1000$ hektar
- Selisih Produksi (Y) $10 - 6 = 4$ kwintal



Perhatikan bahwa **X ribuan** dan **Y dalam satuan**. Jadi ketika jarak euclid digunakan begitu saja, variabel X akan lebih dominan merepresentasikan perbedaan jarak padahal variabel X lebih dominan memang hanya karena skalanya saja.

Classification

K-Nearest Neighbour : Solve Scale Issue

Solution to solve scale issue is **Normalization**.

Min-Max Scaling

Uses *MinMaxScaler*

Transform to defined range

$$y = \frac{x - \min x_i}{\max x_i - \min x_i}$$

Standardization

Uses *StandardScaler*

Transform to mean=0, sd=1

$$y = \frac{x - \bar{x}}{s}$$

Dimana

\bar{x} = mean/rata-rata

s = Standar Deviasi

Classification

K-Nearest Neighbour : Solve Scale Issue using Min-Max Scaling

Bigger Contribution
"area"

Look at the two last data points as example:

Variable	diff before	diff after
area	1000 hectare	0.143
production	6 Kw	0.858

Bigger Contribution
"production"

X1

0 900 1000
0 0

X2

0 4 10

The Process

	area	production
0	4000	5.0
1	5000	4.5
2	2000	3.0
3	6000	4.5
4	7000	4.6
5	8000	4.0
6	9000	10.0



	area	production
0	0.285714	0.285714
1	0.428571	0.214286
2	0.000000	0.000000
3	0.571429	0.214286
4	0.714286	0.228571
5	0.857143	0.142857
6	1.000000	1.000000

Contribution in reality

Classification

K-Nearest Neighbour : Kelebihan dan Kekurangan

Advantages ?

- Dapat memiliki performa model yang baik dalam berbagai macam kondisi.
- Mudah dipelajari
- Program algoritma KNN lebih mudah
- Waktu training yang terhitung cepat.

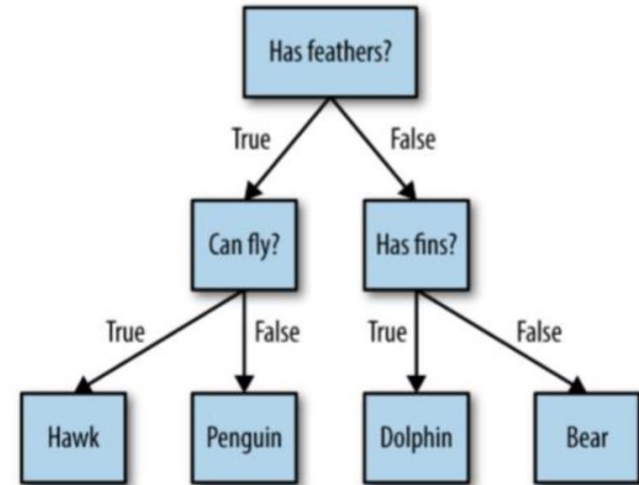
Disadvantages ?

- Metode KNN adalah membutuhkan memori yang lebih besar ketika data yang digunakan untuk training bertambah besar.
- Sulit untuk diinterpretasikan
- KNN tidak dapat membedakan fitur yang sebenarnya penting dalam prediksi.

Classification

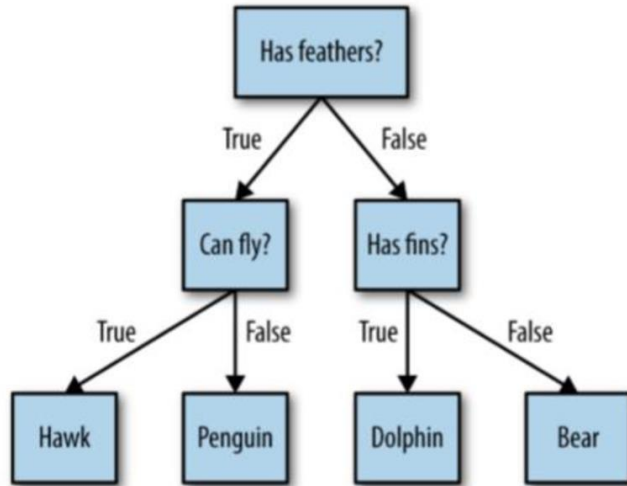
Decision Tree Classifier

- **Decision tree** adalah hierarki pertanyaan if/else, yang mengarah ke suatu keputusan. sifatnya non-parametrik seperti KNN.
- Decision tree termasuk ke dalam metode machine learning yang paling sering digunakan terutama untuk kasus klasifikasi. Selain untuk klasifikasi, decision tree juga dapat digunakan untuk masalah regresi.
- Dengan decision tree, kita tidak perlu membuat asumsi terkait bentuk dari model yang akan digunakan.
- Decision tree sangat fleksibel karena dapat menangkap segala jenis hubungan, linear maupun non linear.
- Decision tree juga baik digunakan ketika kita ingin model machine learning yang diperoleh cepat, fleksibel dan dapat diinterpretasikan dengan mudah.



Classification

Decision Tree Classifier : Decision Analogy

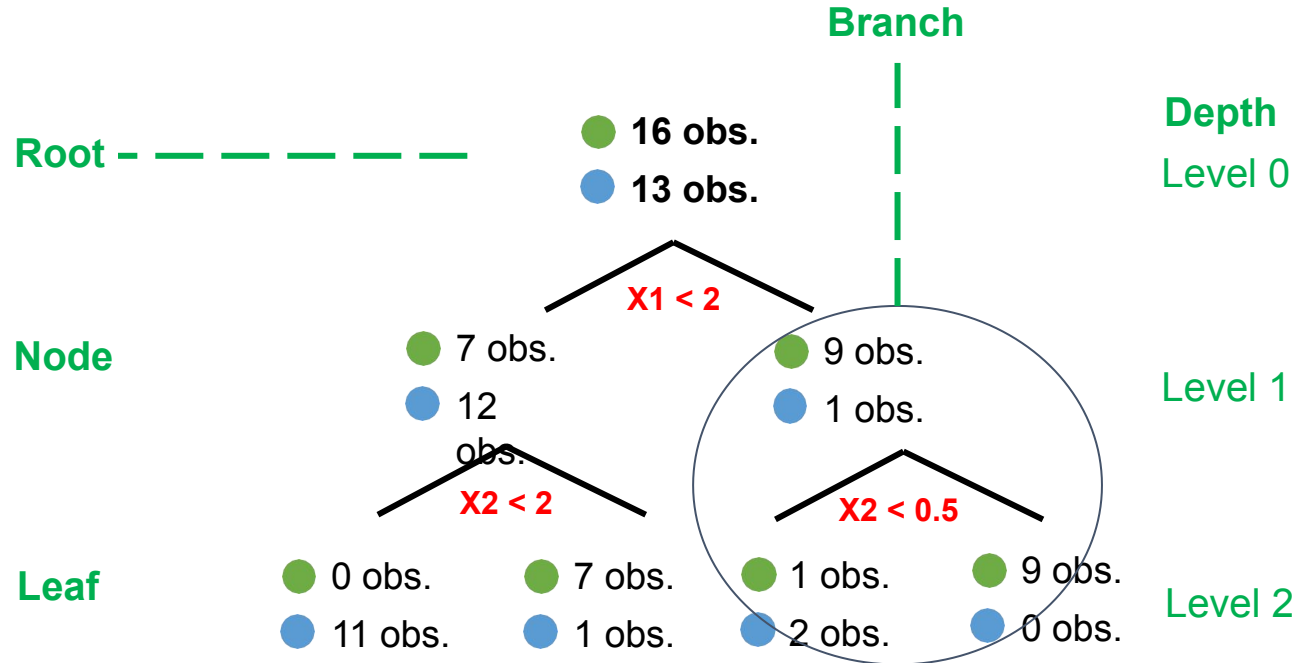


Analogies:

- Misalkan kita ingin mengklasifikasikan empat hewan berdasarkan ciri fisik atau karakteristiknya sesederhana mungkin. Hewan-hewan tersebut adalah **Elang, Penguin, Lumba-lumba, dan Beruang**.
- In term of classification:
 - Animals (Hawk, Penguin, Dolphin, Bear) → Target
 - Characteristics → feature
- We must be thinking what characteristics can differentiate them:
 - Among These animals, which one has feathers
 - feathers yes : Hawk and Penguin
 - feathers no : Dolphin and Bear
 - This is not enough we still need additional information
 - feather yes : add can fly or not
 - fly : Hawk
 - do not fly : Penguin
 - feather no : has fins or not
 - Has Fins : Dolphin
 - No Fins : Bear

Classification

Decision Tree Classifier : Terminologies



Classification

Decision Tree Classifier : Kelebihan dan Kekurangan

Advantages ?

- Mudah dipahami bagaimana bentuknya
- Sangat berguna dalam eksplorasi data
- Dapat bekerja untuk variabel numerik maupun kategorikal

Disadvantages ?

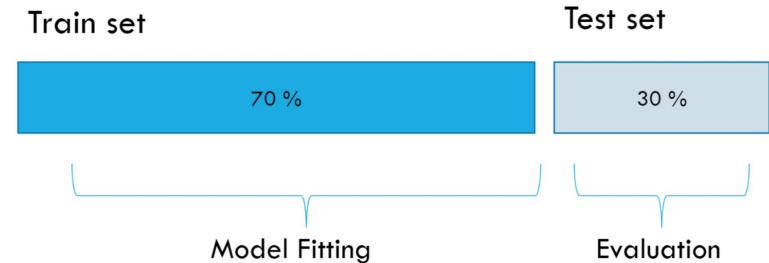
- Bentuk dari decision tree sangat tidak stabil bergantung pada sampel yang digunakan untuk membangun modelnya.
- Secara performa decision tree sulit bersaing dengan metode lainnya seperti bagging, random forest, dan boosting
- Semakin dalam pohon semakin sulit juga interpretasinya.

Generalization, Underfitting, Overfitting

Generalization, Underfitting, Overfitting

What is Generalization ?

- Dalam supervised learning, kita membangun metode menggunakan suatu dataset dengan harapan kita dapat melakukan prediksi yang akurat pada dataset yang baru.
- Untuk dapat mengetahui seberapa baik metode yang digunakan kita membagi data menjadi dua bagian secara random, yaitu **training set** dan **test set**.
- Training set digunakan untuk membangun machine learning sedangkan test set digunakan untuk melihat bagaimana gambaran performa dari machine learning itu sendiri.
- Test set tidak boleh terlibat sama sekali dalam pemodelan.
- Ketika metode machine learning yang digunakan mampu membuat prediksi yang akurat pada dataset yang baru, dapat dikatakan bahwa machine learning yang telah dibangun mampu menggeneralisasi test set menggunakan training set.



Generalization, Underfitting, Overfitting

Generalization Illustration

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Goal :

- Kita ingin memprediksi customer mana yang akan tertarik membeli suatu barang.

Let's build some rule:

- Jika customer memiliki usia lebih dari 45 tahun, memiliki kurang dari tiga anak dan tidak dalam kondisi bercerai maka customer tersebut dinyatakan akan membeli barang. Aturan ini 100 % akurat jika kita lihat pada data yang ada.

Generalization, Underfitting, Overfitting

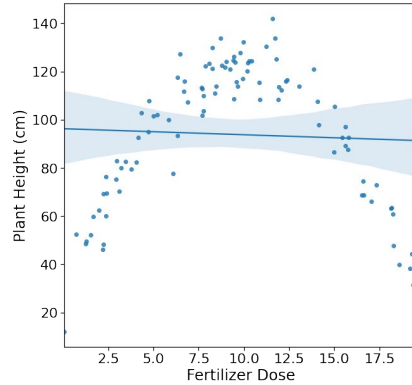
What are Underfitting and Overfitting ?

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

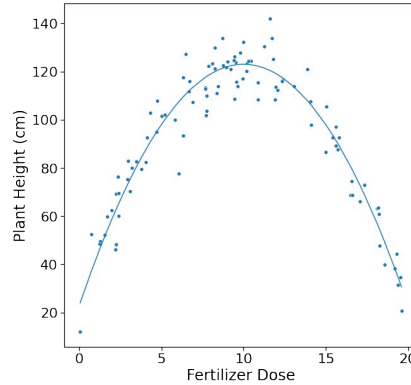
- Underfitting terjadi ketika metode ML yang dibangun masih terlalu sederhana
- Overfitting terjadi ketika metode ML yang dibangun terlalu kompleks.

Generalization, Underfitting, Overfitting

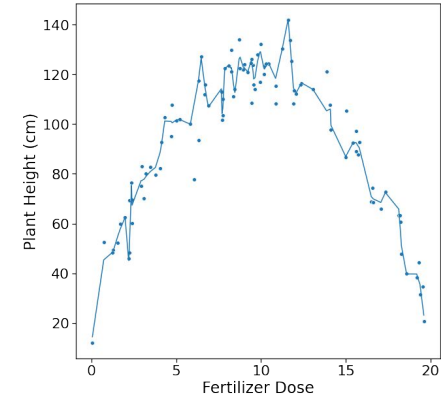
Capturing Underlying Data Trends



Underfitting
Model:
 $y = a + bx$



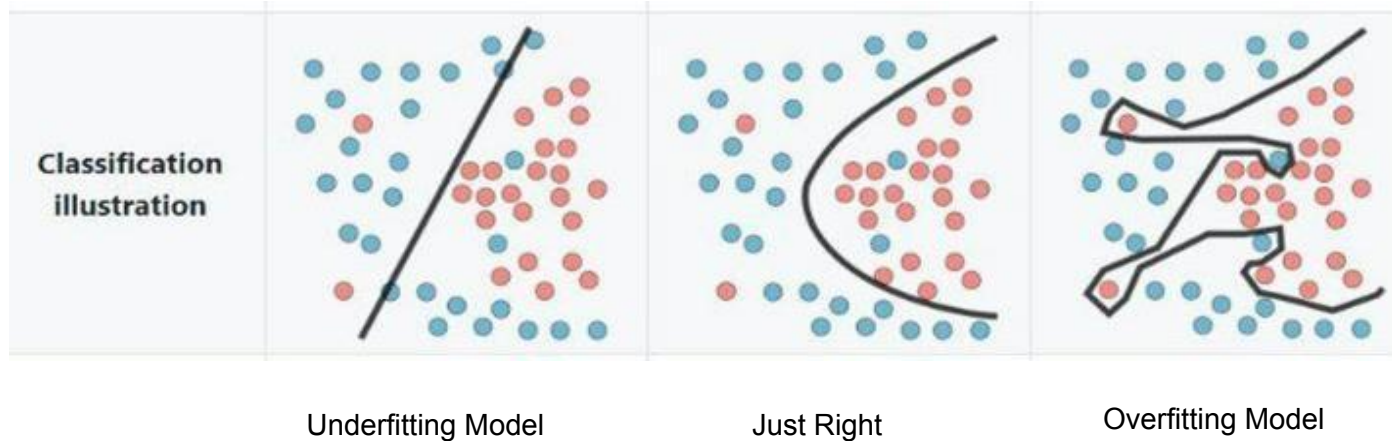
Just Right:
 $y = a + bx + cx^2$



Overfitting Model:
lowess regression

Generalization, Underfitting, Overfitting

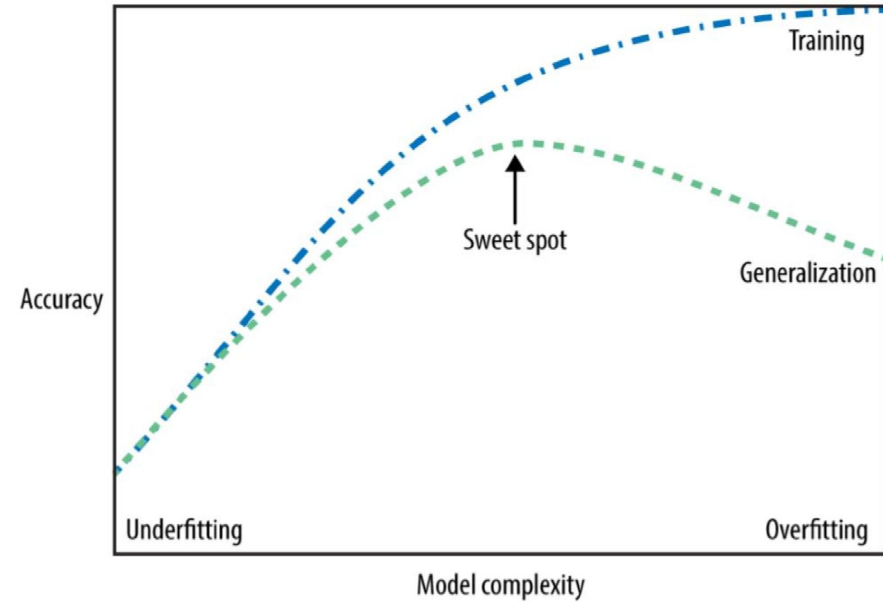
Capturing Underlying Data Trends



Generalization, Underfitting, Overfitting

Model Complexity vs Model Performance

- Suatu metode ML yang masih **underfitting** memiliki nilai prediksi dengan **bias yang tinggi** dan **variance yang rendah**
- Metode ML yang **overfitting** memiliki nilai prediksi dengan **bias rendah** dan **variance yang tinggi**.



Feature Engineering

Feature Engineering

- ❑ Dengan **feature engineering** kita dapat memberikan input yang lebih baik untuk machine learning.
- ❑ Input yang lebih baik tentu berpotensi memberikan hasil yang lebih baik juga.
- ❑ Dalam praktiknya, data tidak dapat langsung digunakan begitu saja. Ada banyak masalah-masalah yang dapat timbul seperti :
 - skala data yang berbeda
 - missing value
 - outlier
 - data yang tidak valid dan reliabel
 - covariate dan lainnya

Masalah-masalah tersebut perlu diatasi agar mendapatkan hasil yang optimal.

Each Model Optimize Differently

KNN

Scal
g

Decision Tree

~~Scal
g~~

Why Does It Matter ?



Scaling

Feature Engineering

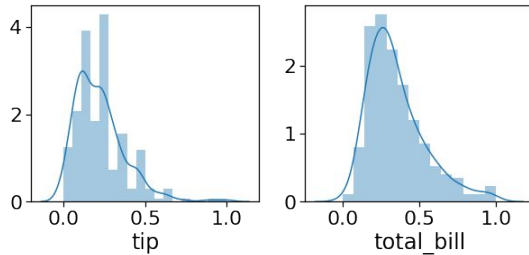
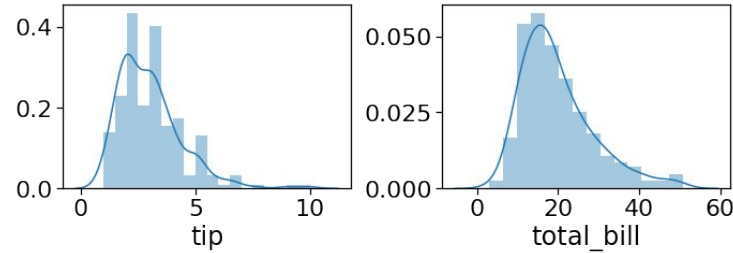
Scaling

- ❑ **Scaling** adalah metode untuk melakukan transformasi terhadap data numerik agar antar variabel memiliki skala yang sama.
- ❑ Metode scaling yang dapat digunakan ada berbagai macam, diantaranya :
 - MinMax Scaler
 - Standard Scaler
 - Robust Scaler
- ❑ Beberapa algoritma machine learning dapat memiliki performa yang lebih baik ketika skala yang digunakan sama, yaitu :
KNN, Neural Network, Linear Model

Feature Engineering

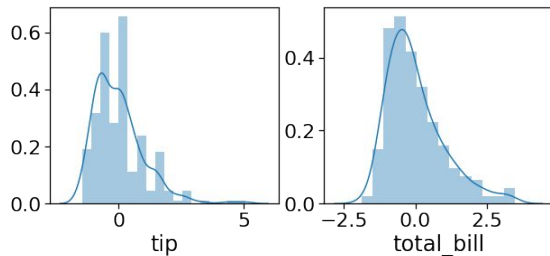
Scaling: MinMax Scaler & StandardScaler

Default Distribution



Transform To Range
0 - 1

$$y = \frac{x - \min x_i}{\max x_i - \min x_i}$$



Transform To mean
= 0 and sd = 1

$$y = \frac{x - \bar{x}}{s}$$

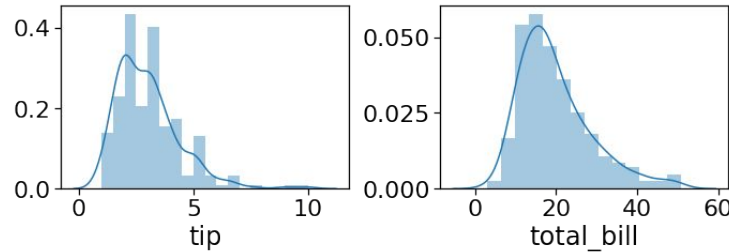
Where

\bar{x} = mean

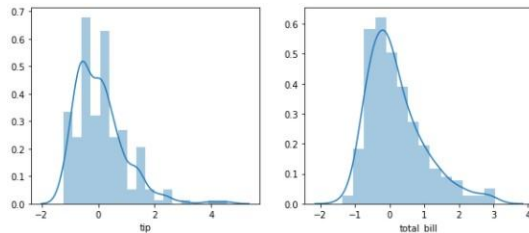
s = Standard deviation

Feature Engineering

Scaling : Robust Scaler



Default Distribution



Transform To Small
Range

Robust scaler ini dapat dijadikan sebagai alternatif karena min max scaler sangat sensitif terhadap outlier

$$z_i = \frac{x_i - Q_1(x_i)}{Q_3(x_i) - Q_1(x_i)}$$

Where:

$Q_1(x_i)$ = first quartile

$Q_3(x_i)$ = third quartile

Encoding

Feature Engineering

Encoding : What is Encoding ?

- ❑ **Encoding** adalah suatu metode yang dapat diterapkan untuk **merepresentasikan variabel kategorik dalam machine learning**.
- ❑ Ada berbagai macam jenis metode encoding diantaranya **one hot encoding**, **ordinal encoding** dan **binary encoding**.
- ❑ Kita dapat memilih metode encoding berdasarkan skala pengukuran datanya, yaitu nominal atau ordinal.

Scale of Measurement	Suggested Method		
	One Hot Encoding	Ordinal Encoding	Binary Encoding
Nominal	v	x	v
Ordinal	v	v	x

Encoding

One Hot Encoding

- ❑ Dalam **one not encoding** kita memecah suatu variabel kategorik menjadi beberapa variabel yang nilainya satu atau nol atau disebut juga dengan dummy variabel.

Gender
Male
Female
Female
Male
Female

Male	Female
1	0
0	1
0	1
1	0
0	1

City
Jakarta
Bogor
Bogor
Bekasi
Bekasi

Jakarta	Bogor	Bekasi
0	1	0
1	0	0
1	0	0
0	0	1
0	0	1

Encoding

One Hot Encoding For Linear Model

- ❑ Khusus untuk linear model, maksimal banyaknya dummy variabel yang perlu dibuat adalah banyaknya kategori dikurangi satu.

Gender
Male
Female
Female
Male
Female

Male
1
0
0
1
0

City
Jakarta
Bogor
Bogor
Bekasi
Bekasi

Jakarta	Bogor
0	1
1	0
1	0
0	0
0	0

Encoding

Ordinal Encoding

- ❑ Dalam **ordinal encoding** kita mentransformasi masing-masing kategori pada variabel ordinal menjadi nilai integer dan sesuai dengan urutannya.

Education
SD
SMP
SD
SMA
S1
S1



Education Encode
1
2
1
3
4
4

Value	Mapping
Other/None	0
SD	1
SMP	2
SMA	3
S1	4
Post-Grad	5

Encoding

Binary Encoding

CAR	Order	Binary Num	C1	C2	C3
Avanza	1	001	0	0	1
Xenia	2	010	0	1	0
Xenia	2	010	0	1	0
CR-V	3	011	0	1	1
Avanza	1	001	0	0	1
Calya	4	100	1	0	0
City	5	101	1	0	1
Calya	4	100	1	0	0
Jazz	6	110	1	1	0

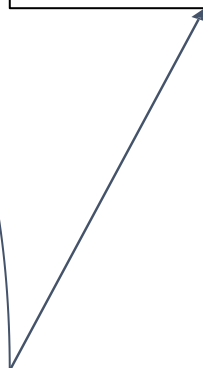
- ❑ **Binary encoding** dapat digunakan sebagai alternatif dari one hot encoding.
- ❑ **Binary encoding** digunakan untuk encoding variabel nominal yang memiliki terlalu banyak kategori.

Encoding

Binary Encoding

Number	Binary Number	Binary Number(alt.)
1	1	0001
2	10	0010
3	11	0011
4	100	0100
5	101	0101
6	110	0110
7	111	0111
8	1000	1000
9	1001	1001

Follow the largest digit



Encoding

.fit and .transform Method in preprocessing

Method	training set	test set or validation set
.fit	V	X
.transform	V	V

```
scaler = MinMaxScaler()  
scaler.fit(X_train)  
X_train_scaled = scaler.transform(X_train)  
X_test_scaled = scaler.transform(X_test)
```

- ❑ Perlu diketahui bahwa akan jauh lebih aman ketika .fit hanya diterapkan pada training set (Sejumlah metode preprocessing akan mengalami kekeliruan ketika .fit diterapkan kembali pada test set atau validation set)
- ❑ Contohnya adalah metode preprocessing binary encoding dan tf-idf

Missing Value

Missing Value

What is Missing Value ?

Gender	City	Income(IDR)
Male	Jakarta	-1
Female	Bogor	5,000,000
NaN	Unknown	2,500,000
Male	Bekasi	7,000,000
Female	Bekasi	12,000,000

Another value that might represent missing value :
“?”, 999999, “miss”, etc

Missing Value

Missing Value

Handling : Simple Technique

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
2	NaN	6.0	12	C	X	NaN
3	6.0	5.0	9	C	X	M
4	7.0	NaN	8	D	NaN	N
5	9.0	5.0	11	NaN	Y	NaN

Drop Row



	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
3	6.0	5.0	9	C	X	M

Simple Technique:

- Drop Column
- Drop Row
- Substitution with mean, median or mode.

Missing Value Handling : Simple Imputer

Library Pandas:

- fillna

Library Scikit-Learn:

- Mean
- Median
- Mode or new constant
- Multivariate feature imputation (equivalent to Expectation-Maximization)
- KNN-Imputer

Missing Value

Simple Imputer : Mean or Median

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
2	NaN	6.0	12	C	X	NaN
3	6.0	5.0	9	C	X	M
4	7.0	NaN	8	D	NaN	N
5	9.0	5.0	11	NaN	Y	NaN

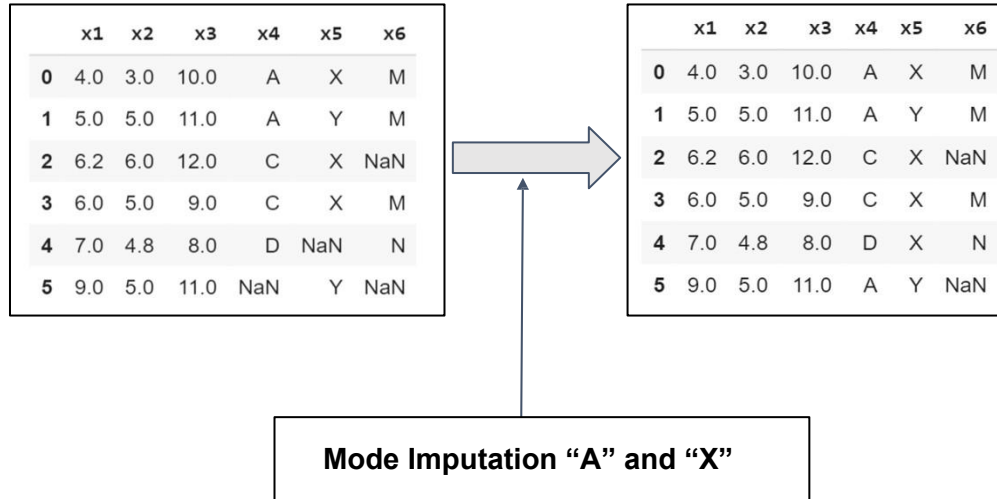


	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	NaN
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	NaN	N
5	9.0	5.0	11.0	NaN	Y	NaN

Mean Imputation

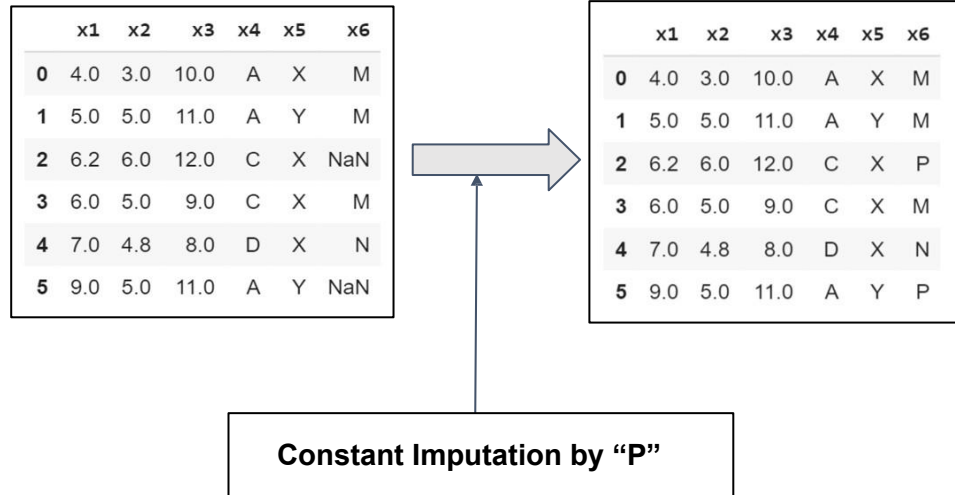
Missing Value

Simple Imputer : Mode



Missing Value

Simple Imputer : Constant



Missing Value

Handling : Iterative Imputer

- Iterative imputer merupakan metode yang dapat digunakan untuk mengisi variabel numerik secara iteratif.
- Dalam scikit-learn, metode ini hanya dapat diterapkan untuk mengisi variabel numerik. Metode ini memanfaatkan variabel lain untuk memprediksi missing value menggunakan regresi secara iteratif.

	x1	x2	x3	x4
0	4.3	2.9	9.0	A
1	5.1	5.1	11.1	A
2	NaN	6.3	NaN	C
3	6.3	4.9	8.9	C
4	7.4	NaN	9.1	D
5	9.1	5.4	11.0	D



	x1	x2	x3	x4
0	4.30000	2.900000	9.000000	A
1	5.10000	5.100000	11.100000	A
2	7.18363	6.300000	9.823389	C
3	6.30000	4.900000	8.900000	C
4	7.40000	5.073866	9.100000	D
5	9.10000	5.400000	11.000000	D

Iterative Imputer

Missing Value

Handling : KNN Imputer

- KNN imputer juga merupakan metode yang dapat digunakan untuk mengisi variabel numerik.
- Sama seperti iterative imputer, metode ini hanya dapat diterapkan untuk mengisi variabel numerik.
- Metode ini memprediksi missing value berdasarkan variabel lainnya menggunakan metode KNN.

	x1	x2	x3	x4
0	4.3	2.9	9.0	A
1	5.1	5.1	11.1	A
2	NaN	6.3	NaN	C
3	6.3	4.9	8.9	C
4	7.4	NaN	9.1	D
5	9.1	5.4	11.0	D



	x1	x2	x3	x4
0	4.3	2.90	9.00	A
1	5.1	5.10	11.10	A
2	7.1	6.30	11.05	C
3	6.3	4.90	8.90	C
4	7.4	5.15	9.10	D
5	9.1	5.40	11.00	D

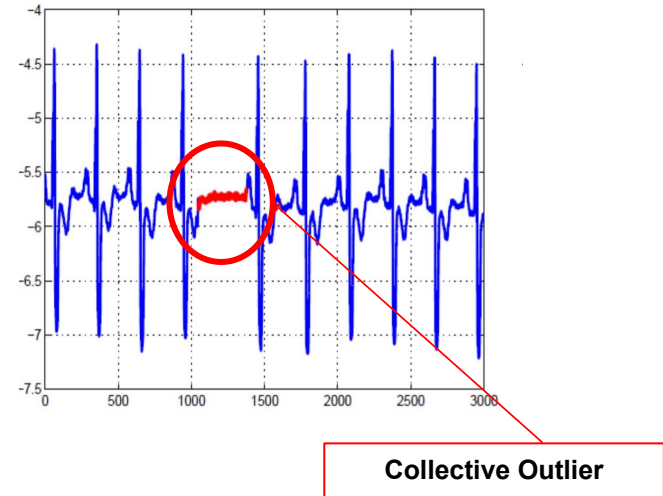
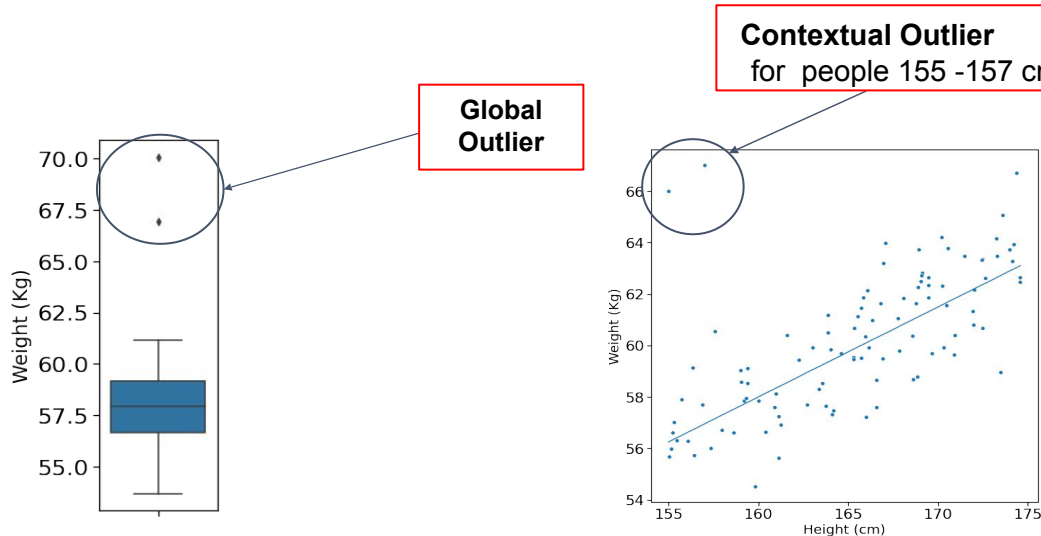
Iterative Imputer

Outlier

Outlier

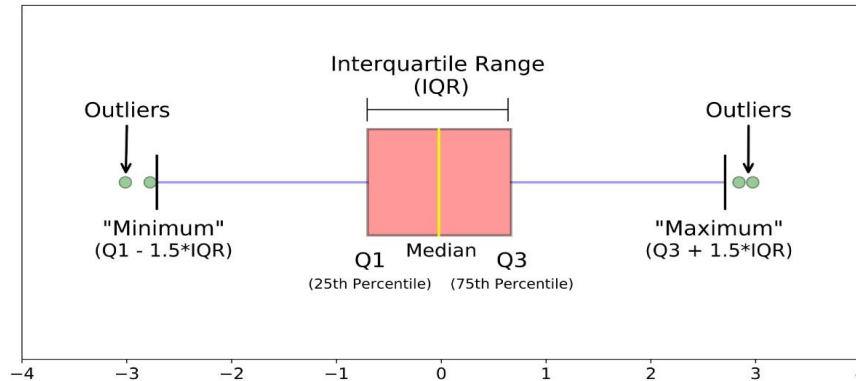
- Outlier merupakan observasi atau data poin yang nilainya berbeda atau jauh daripada observasi pada umumnya.
- Suatu outlier dapat mengindikasikan suatu nilai yang memang salah (experimental error) atau memang nilai yang disebabkan karena kondisi tertentu (variability in the measurement).

Outlier ada beberapa jenis, yaitu :



Outlier

Outlier in Univariate Variable



Kita dapat mendeteksi outlier pada univariate data menggunakan **IQR** dimana

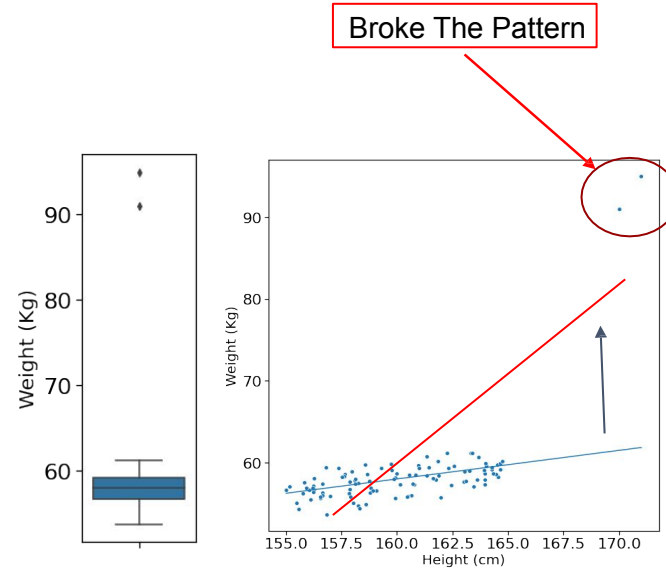
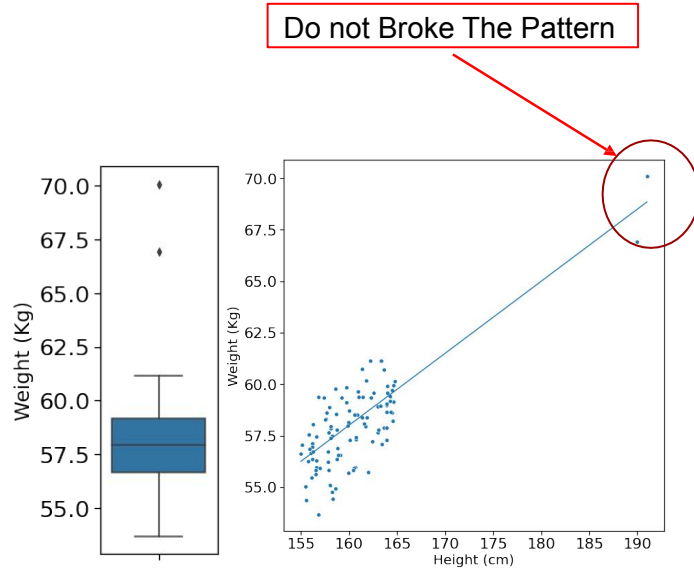
$$\text{IQR} = Q3 - Q1$$

Suatu data poin dikatakan outlier ketika

1. Nilai data $> Q3 + 1.5 \text{ IQR}$ atau lebih
2. Nilai data $< Q1 - 1.5 \text{ IQR}$.

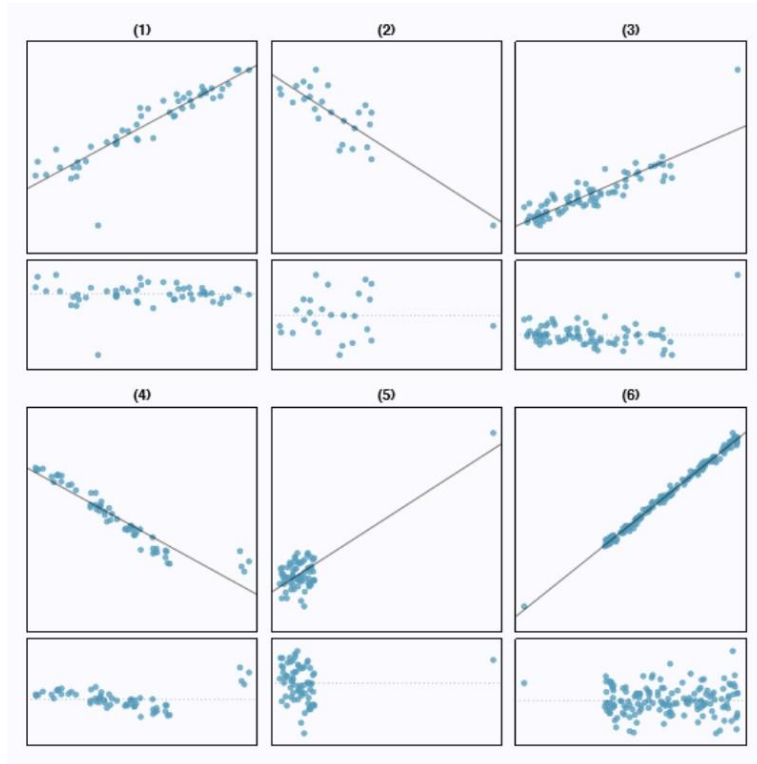
Outlier

Outlier in Linear Regression



Outlier

Outlier in Linear Regression



Bisa kita telusuri sejumlah skenario lain terkait bagaimana peran outlier dalam pemodelan dengan gambar-gambar berikut.

1. **Outlier slightly influence** the line
2. **Outlier do not much influence** the line
3. **Outlier slightly influence** the line
4. Line badly fitted because **outlier slightly influence** the line and each of the cluster data points may have interesting explanation
5. Actually there is no certain pattern but the line appeared to be linearly positive because of the outlier
6. **Outlier do not much influence** the line

Feature Selection

Feature Selection

What is Feature Selection ?

- Feature selection digunakan untuk melakukan seleksi terhadap fitur-fitur yang digunakan dalam pemodelan. Kita memilih fitur yang memang penting atau berpengaruh terhadap target variabel.
- Feature selection juga dapat dipandang sebagai suatu metode generalisasi.
- Ketika terlalu banyak fitur yang terlibat dalam pemodelan model akan cenderung **overfitting** dan jika fitur terlalu sedikit model menjadi **underfitting**.

X1	X2	X3	X4	X5	X6	X7	Y
3	10	11	32	0.5	100	54	12
4	13	12	30	0.5	99	56	10
6	12	15	33	0.1	87	57	13
...
6	10	12	12	1.9	81	78	16



X1	X4	X6	Y
3	32	100	12
4	30	99	10
6	33	87	13
...
6	12	81	16

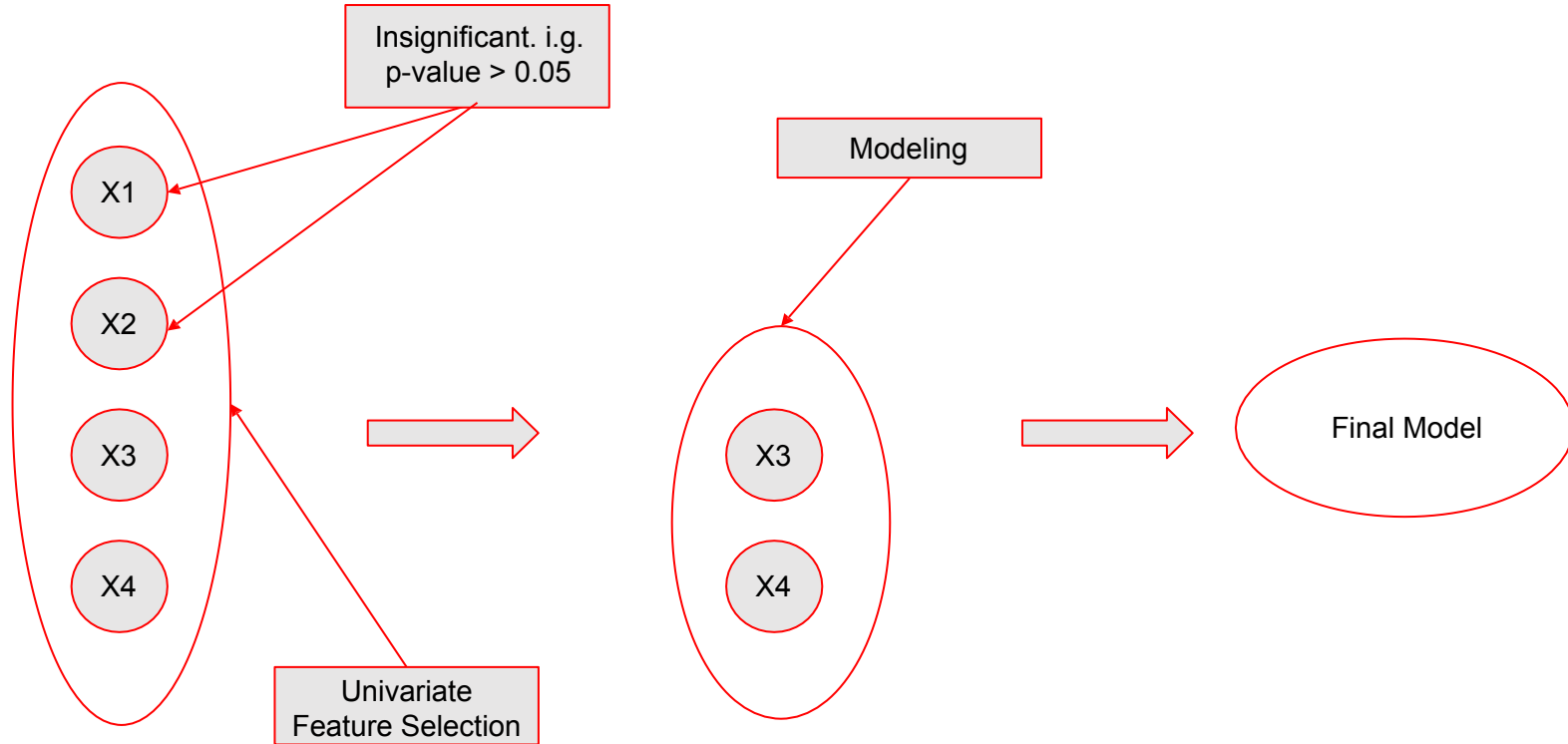
Feature Selection

Feature Selection Method

- Univariate Statistics Feature Selection
- Model Based Feature Selection
- Iterative Feature Selection

Feature Selection

Feature Selection Method : Univariate Statistics Feature Selection



Feature Selection

Feature Selection Method : Model Based Feature Selection

Dengan menggunakan metode model based feature selection, kita memilih fitur menggunakan model machine learningnya secara langsung.

- Contohnya pada **decision tree** atau tree based model lainnya, kita dapat gunakan nilai feature importance untuk melakukan feature selection.
- Untuk model linear regression atau logistic regression, kita dapat gunakan nilai absolut dari koefisiennya dengan catatan fitur yang digunakan memiliki skala yang sama atau sudah distandarisasi.
- Dengan metode ini, kita perlu membangun modelnya terlebih dahulu agar dapat melakukan feature selection.

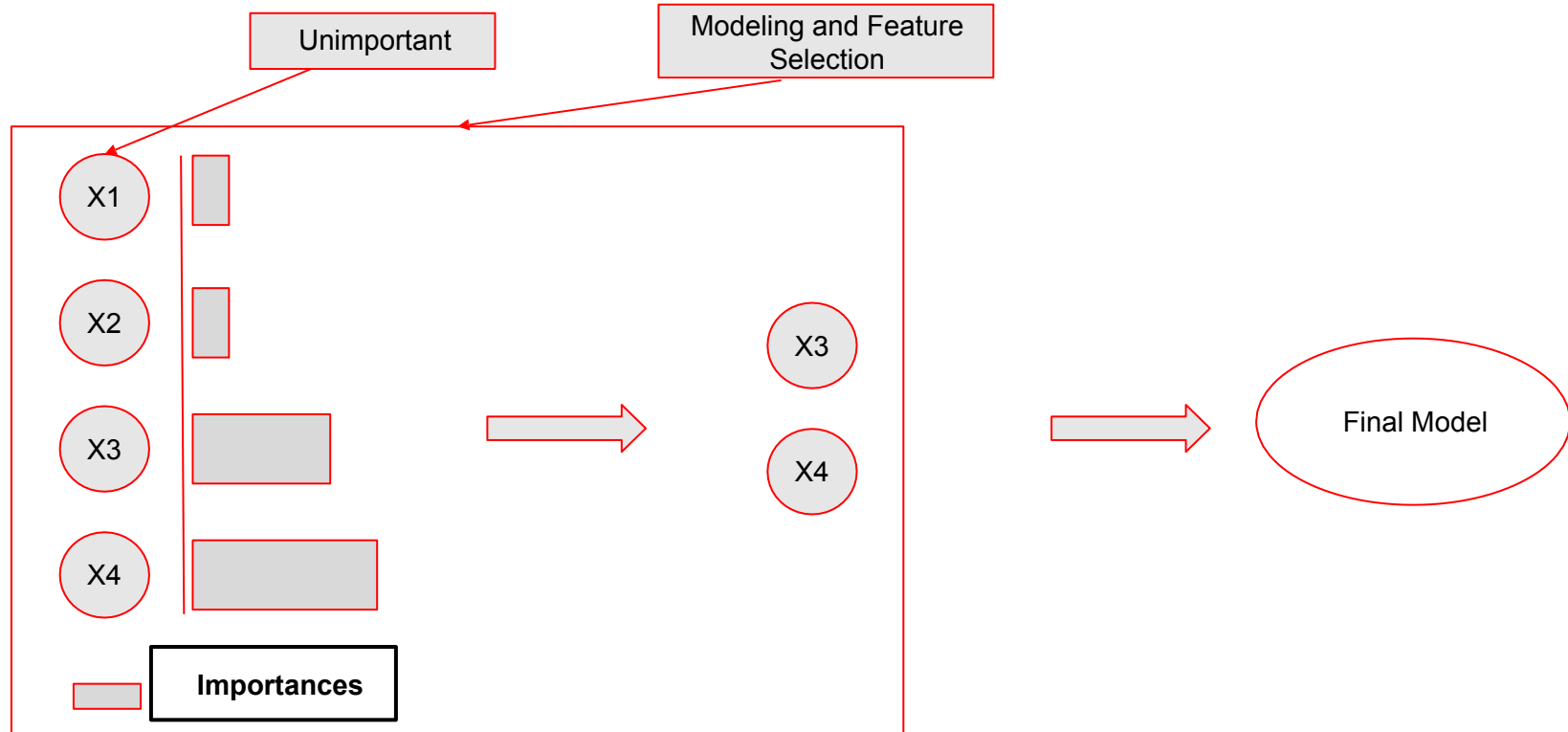
Pros : Hasil yang didapatkan bisa lebih optimal karena hasilnya sudah dapat menyesuaikan dengan cara kerja masing-masing model

Cons : Memerlukan waktu yang lebih lama dalam prosesnya.

- Keunggulan lain dari metode ini adalah kita melakukan seleksi secara sekaligus sehingga metode ini mampu mempertimbangkan aspek interaksi antara fitur.

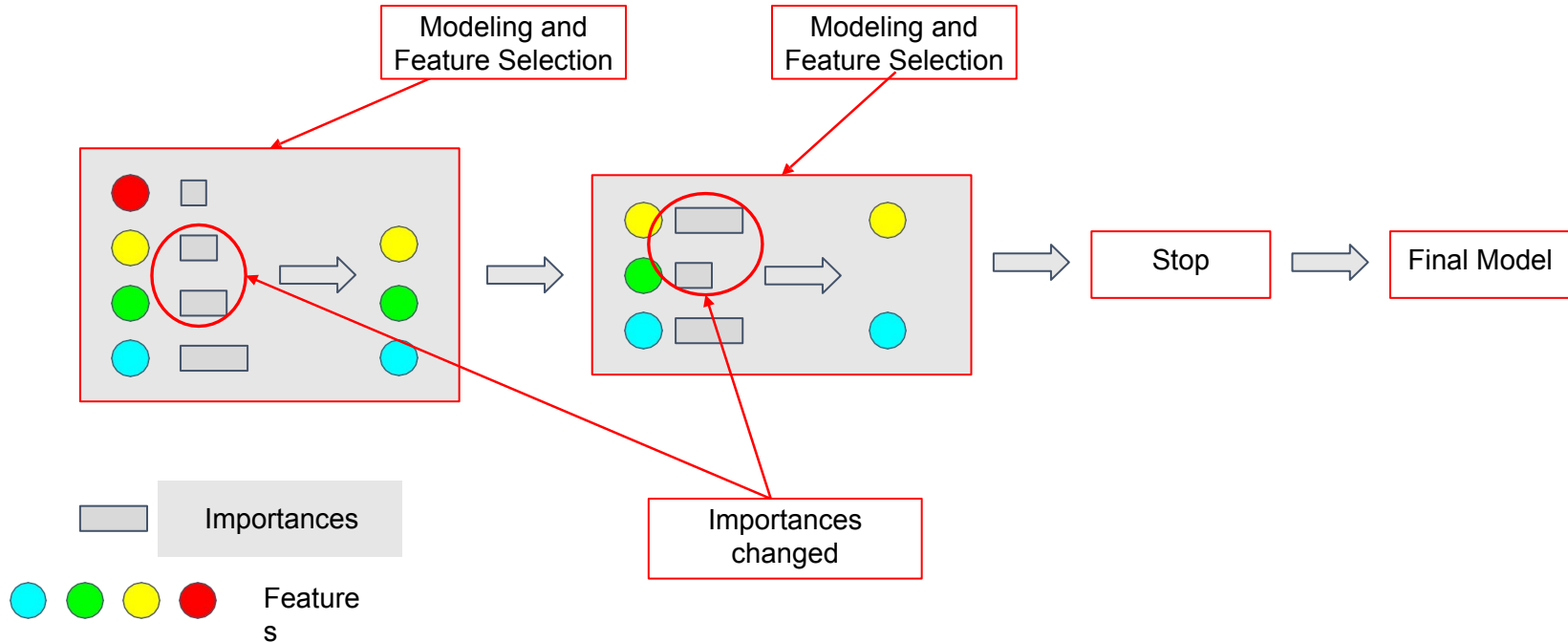
Feature Selection

Feature Selection Method : Model Based Feature Selection



Feature Selection

Feature Selection Method : Iterative Feature Selection



Feature Engineering Exercise

Kita akan mempraktekkan feature engineering menggunakan data “adult.csv”. Target variabel dari data ini adalah income. Kita hendak mengklasifikasikan income seseorang besar ataukah kecil menggunakan logistic regression.

Skenario preprocessing-nya adalah sebagai berikut:

- Missing value : simple imputer with constant
- one hot encoding : relationship, race, sex
- binary encoding : workclass, marital status, occupation, native country
- ordinal encoding : education
- no treatment untuk numerical variable.

Thankyou

ada pertanyaan?
